Computers and Electronics in Agriculture 71 (2010) 163–169

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Automatic image analysis and spot classification for detection of pathogenic *Escherichia coli* on glass slide DNA microarrays

Ron P. Haff[a], Beatriz Quiñones[b], Michelle S. Swimley[b], Natsuko Toyofuku[a,*]

[a] Plant Mycotoxin Research Unit, U.S. Department of Agriculture-Agricultural Research Service, Western Regional Research Center, Albany, CA 94710, USA
[b] Produce Safety and Microbiology Research Unit, U.S. Department of Agriculture-Agricultural Research Service, Western Regional Research Center, Albany, CA 94710, USA

ARTICLE INFO

Article history:
Received 26 September 2009
Received in revised form 13 January 2010
Accepted 25 January 2010

Keywords:
DNA microarray
Algorithm
Spot location
Escherichia coli
Food safety

ABSTRACT

A computer algorithm was created to analyze and quantify scanned images from DNA microarray slides developed for detecting pathogenic *Escherichia coli* isolates recovered from agricultural food products. The algorithm computed centroid locations for signal and background pixel intensities in RGB space and defined a plane perpendicular to the line connecting the centroids as a decision boundary. The algorithm was tested on 1534 potential spot locations which were visually classified depending on the strength of the signal. Three other standard measures of SNR (SSR, SBR, and SSDR) were also performed for each potential spot location. The number of errors as compared to visual classifications was computed for each of the four measures. SSR and SSDR, which depend on pixel intensity standard deviations, performed poorly with high false positive results, while the current algorithm and SBR, which were independent of standard deviations, performed much better. Overall error rates were 1.4% for the reported algorithm, 2.0% for SBR, 14.2% for SSDR, and 16.8% for SSR.

Published by Elsevier B.V.

1. Introduction

Bacterial contamination of agricultural products continues to be a serious health threat, and recent increases in the reported occurrence of outbreaks have led to an increased emphasis on the development of food safety programs in the United States. A leading cause of foodborne illness is considered to be *Escherichia coli* (*E. coli*), which is thought to contribute to more than 73,000 cases of human infection in the United States per year (Rangel et al., 2005). Over the past years, there has been a rise in *E. coli* outbreaks due to the consumption of leafy vegetables, and many of these *E. coli* outbreaks were traced to the Salinas Valley region of California (Centers for Disease Control and Prevention, 2006; Cooley et al., 2007). In September 2006, a multi-state outbreak of *E. coli* infections was linked to baby spinach grown in San Benito County near the Salinas Valley region in California and resulted in 205 confirmed illnesses and 3 deaths (Centers for Disease Control and Prevention, 2006).

The rise in outbreaks linked to the consumption of agricultural food products has heightened the importance of developing better methods to rapidly detect and characterize pathogenic *E. coli* strains. Established culturing methods are very labor-intensive and time-consuming and are limited in the number of samples to be analyzed (Bettelheim and Beutin, 2003). Current methods for automated detection such as sorting based on X-ray, visible light, or near infrared have not been optimized for identifying single bacterial cells. Thus, optimization of procedures for pathogen surveillance is needed with sufficient sensitivity, cost-effectiveness and suitability for routine testing. Recently, methods have been developed using glass slide DNA microarrays for the rapid and economical identification of pathogenic *E. coli* recovered from food products (Quiñones et al., 2009). This information can be vital in guiding subsequent contamination control procedures and preventing contaminated food products from reaching the consumer.

The original glass slide microarrays were produced in 1995 (Schena et al., 1995), and their use as a tool in genomic research has expanded enormously. The development of commercially available printing devices that can precisely situate printing pins over glass slides has accelerated the adoption of this technology. One of the greater challenges has been the extraction, storage, and analysis of the huge amount of generated data (Holloway et al., 2002; Heller, 2002). While printing devices can rapidly produce thousands of spots, extracting the desired data from the microarrays can be time-consuming, slowing research progress. Glass slide microarrays are scanned into image files, and the data analysis then becomes an image processing exercise. Typical software requires selecting and saving pixel value data from multiple regions of interest for each spot on the slide, which can number in the thousands. Consequently, there is a considerable demand for the development of algorithms that can standardize and simplify the extraction of data from the scanned images (Heller, 2002).

* Corresponding author at: USDA-ARS-WRRC, Plant Mycotoxin Research Unit, 800 Buchanan Street, Albany, CA 94710, USA. Tel.: +1 510 559 5868; fax: +1 510 559 5684.
E-mail address: Natsuko.Toyofuku@ars.usda.gov (N. Toyofuku).

0168-1699/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.compag.2010.01.002

There are a number of different glass slide formats used in the field of genomics for which analysis algorithms have been developed (Heller, 2002; Bhandarkar et al., 2004). Most microarray images are generated by using precise robotic controls, resulting in a grid of spots that are scanned and saved into an image file for subsequent analysis. As part of this process, one important goal is to determine some measure of the signal-to-noise ratio (SNR) between each spot in the array and the background. However, there is no consensus on how SNR should be determined, particularly in terms of determining background levels that are measured locally in the neighborhood of each spot or globally at a point outside of the grid. Some arguments have been made that only the local background estimate is adequate (Angulo and Serra, 2002). There has also been disagreement over the formula for computing SNR given the background and signal pixel intensities (Holloway et al., 2002; He and Zhou, 2008). Two common formulas for computing the SNR in the neighborhood of a spot are the signal to standard deviation (σ) ratio (SSR) given by

$$SSR = \frac{(\bar{S} - \overline{BG})}{\sigma_{BG}}, \quad (1)$$

and the signal to background ratio (SBR) given by

$$SBR = \frac{\bar{S}}{\overline{BG}}, \quad (2)$$

where $\bar{S}$ is the mean intensity for signal pixels, $\overline{BG}$ is the mean intensity for background pixels, and σ_{BG} is the standard deviation of the background. Recently a new measure for SNR has been reported that addresses the problem of standard definitions not taking into account the non-uniformity of the intensities of signal pixels (He and Zhou, 2008). The signal to both σ's ratio (SSDR) has been defined as

$$SSR = \frac{(\bar{S} - \overline{BG})}{\sigma_s + \sigma_{BG}} \quad (3)$$

where σ_s represents the σ of signal pixel intensities. SNR is not always the method used for determining spot presence. A simple threshold on the pixels in the predominant color channel or a threshold on the grayscale conversion is sometimes used. A threshold on an intensity histogram has also been used to separate background pixels from signal pixels (Steinfath et al., 2001).

Most automated algorithms reported to date attempt to correlate the pixel intensity (or SNR) at the spot location to the concentration of the sample being measured, such as expression levels and DNA copy number in biological samples. Hypothetically, spot intensity can be correlated with the amount of probe at that location of the grid, and some algorithms have been developed to estimate sample concentrations from the arrays (Lopez et al., 2004). In some cases a simple binary decision is the objective of spot analysis; either there is a spot or there is not (Lazo et al., 2005).

Previously reported automated algorithms for analyzing DNA microarray images have been concerned with detecting and analyzing many thousands of spots in a single image. The majority of the developed algorithms follow the same basic blueprint, which involves determining the grid layout and orientation, the spacing between spots and hence the locations of individual spots, measuring signal pixel and background pixel intensities in the neighborhood of each spot, and performing the desired statistics, *e.g.* the SNR, for each spot (Jain et al., 2002). The grid layout is often determined by simply summing pixels both vertically and horizontally and looking for the peaks in the resulting arrays corresponding to the rows and columns of the grid (Rueda and Vidyadharan, 2006). Proper orientation (or lack thereof) of the grid is also a concern for algorithms and there have been a variety of methods, including Hough transforms (Audic and Zanetti, 1995) to detect and measure the rotation of the grid as compared to the edges of the image.

Recently, researchers have begun to develop techniques using photo-polymerization for rapid and economical identification of toxin producing bacteria. One such study has an objective for the detection of pathogenic *E. coli* on glass slide DNA microarrays (Quiñones et al., 2009). The system generates bitmap images of the scanned glass slides. For the rapid detection of *E. coli*, it is necessary to automatically analyze the images. In the present study, the main objective was to develop an automatic computer algorithm for the analysis of images of glass slide DNA microarrays generated by a novel detection system that allows for the rapid and economical detection of pathogenic *E. coli* (Quiñones et al., 2009). A second objective was to demonstrate a new technique, not related to SNR, for the determination of the presence or absence of spots from scanned microarray images.

2. Materials and methods

2.1. Image parameters

The initial images generated by the scanner were 9 MB bitmap in 24 bit color format with a width of 2048 pixels and a height of 1536 pixels (Fig. 1). The region of interest was a smaller rectangle of 664 × 656 pixels. Fig. 2a shows the layout of the grid, which consisted of an array of 8 columns and 9 rows of potential spot locations, where the presence of a spot indicated a positive result. The 500 μm pins used for printing yielded spots of 400–450 μm (37–42 pixels) in diameter. For this algorithm, the spot diameter was defined as 40 pixels. Center-to-center spacing was 700 μm (66 pixels). The first, fifth, and ninth rows were controls and were expected to always show a positive result. These control spots divided the 24 distinct targets (*E. coli* strains to be detected) into four groups, with each target spotted in duplicate (Fig. 2a). The location of the grid within the overall image was consistent to within plus or minus ten pixels, and any rotation of the grid as compared to the edges of the image was small, less than about three degrees.

2.2. Image processing

Red, green, and blue pixel intensity values were read into separate arrays of 664 columns by 656 rows covering the region of interest, while the rest of the image was discarded. The array was flipped bottom to top to compensate for the reverse order of data

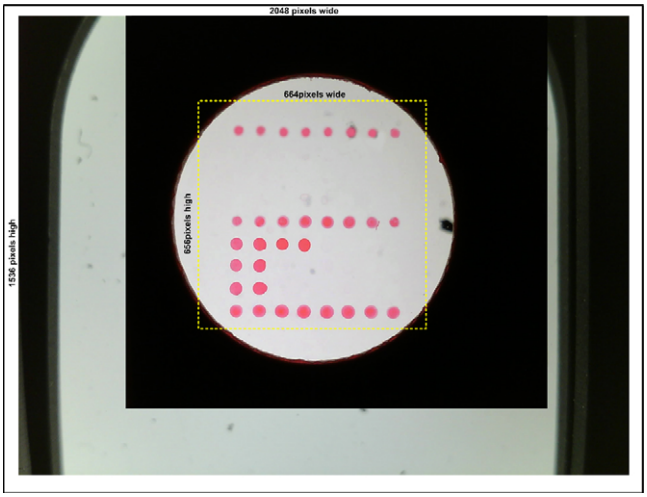

Fig. 1. Bitmap image generated by scanning microarray slides to rapidly detect and identify *E. coli* strains from agricultural products. The initial image generated by the scanner is a 9 MB bitmap in 24 bit color format with a width of 2048 pixels and a height of 1536 pixels. The region of interest is a much smaller rectangle of approximately 664 × 656 pixels.