Microarray Image Segmentation Methodologies: A Review

2.1. Microarray Image Segmentation Methodologies

The segmentation of an image can generally be defined as the process of partitioning the image into different regions, [1] each having certain properties (e.g., color or texture) [2, 3, and 4]. In a microarray experiment, segmentation allows the classification of pixels as foreground (i.e. as corresponding to a spot of interest) or background, so that fluorescence intensities can be calculated for each spotted DNA sequence as measure of transcript abundance. Any segmentation method produces spot mask, which consists of the set of foreground pixels for a given spot. Spot intensity extraction refers to the calculation of the fluorescence signal's mean intensity for the foreground spot's. Extracted mean intensities correspond to gene expression levels that, in turn, are translated into biological conclusions from molecular biologists, by employing data mining techniques [5]. The objective of the segmentation inside of a grid cell is to find one segment that contains the foreground information. Existing segmentation methods for microarray
images can be categorized into four groups, according to geometry of the spots they produce:
1. Fixed circle segmentation.
3. Adaptive shape segmentation, and
4. Histogram segmentation. [2, 4, 6, 7, 8, 9, 10, 11, 12, and 13]

2.1.1. Fixed Circle Segmentation: In fixed circle segmentation fits a circle with a constant diameter to all spots in the image. This method is easy to implement and works nicely when all the spots are circular and of the same size. It was probably first implementation in the ScanAlyze software written by M.B.Eisen in 1999 [4, 6, 7, 8, 11, 12, 13, 14, 15, 16, and 17] and it is usually provided as an option in most software. Fixed diameter segmentation is clearly not satisfactory for all the spots. In Circle segmentation: after the addressing and given the extracted boundary of a spot, circle segmentation is performed locally using a simple algebraic algorithm. By outputting radius & center of a circle, pixels of the foreground are specified as belonging to the spot and can be used for density computations. [18]

2.1.2. Adaptive Circle Segmentation: In adaptive circle segmentation, the circle's diameter is estimated separately for each spot, the software, GenePix for the Axon scanner in 1999 implements such an algorithm [6, 7, 12, 13, 15, 16, 19, 20]. Note that ScanAlyze and other software do provide the user with the option to manually [21] adjust the circle diameter spot by spot. This can be very time consuming, since each array contains thousands of spots.

2.1.3. Adaptive Shape Segmentation: In adaptive shape segmentation, are used two common methods, watershed (Vincent and Soille in 1991 Beucher and Meyer [15 and 22] in 1993) and Seeded region growing (SRG) [1, 4] (Adams and Bichof in 1994. Both require the specification of starting points [23], or seeds. Out of these region-neighboring pixels, the algorithm selects the one whose pixel value is nearest to the average of the pixel values in the neighboring region. In microarray image analysis, however, we are in the rather unusual situation where the number of features (spot) is known exactly a priori and the appropriate locations of the spot centers are determined at the addressing stage. An obvious way to choose a seed for each spot is to choose a single pixel from the
intersections of the horizontal and vertical grid lines of the fitted foreground grid [24]. Microarray images are therefore well suited to such methods. However there are many problems with the flow, i.e. how to stop the growing, which makes this method very non robust and sensitive to input parameters. [25] And

2.1.4. Histogram Segmentation: In Histogram Segmentation, [2, 4, and 23] this type of method uses a target mask, which is chosen to be larger than any spot. For each spot, foreground and background intensity estimates are determined in some fashion from the histogram of pixel values for pixels within the masked area. These methods therefore do not use any local spatial information. A circular target mask and computes a threshold value based on a Mann-Whitney test [4, 12, 13, and 26]. Pixels are classified as foreground if their value is greater than the threshold and as background otherwise. This method is implemented in the QuantArray software for the GSI Lumonics in 1999 scanner [2, 4, 6, 7, 12, 13, 15, and 16]. The main advantage of this method is their simplicity & Disadvantage is that quantification is unstable when a large target mask is set to compensate for spot size variation. Kashif I. Siddiqui, et al. proposed Morphological Watershed Segmentation which has the advantages: 1) no seeding within spot boundaries is necessary, 2) the watershed region provide partition which is used to isolate local noise background for each spots, and 3) the implementation is fully automatic, elimination need for gridding [18, 24] or other manual pre-processing [21]. In this way he classifies broadly the various background methods implemented in software package into four categories. These are: 1.Local background: Background intensities are estimate by focusing on small regions surrounding the spot mask. Usually, the background estimate is the median of pixel values within these specific regions. 2. Morphological opening: Morphological opening [4] is applied to the original images R and G using a square structuring element with side length at least twice as the spot separation distance. This operation removes all the spots and generates an image, which is an estimate of the background for the entire slide; hence it is expected to have very slow spatial variation. This results in smaller background estimates than other simpler methods. 3) Constant background: This is a global method, which subtracts a constant background for all spots. 4) No Adjustment: Finally, consider the possibility of no background adjustment at all. [15]
Recently, Weng Guirong (2008), have to use the mathematical morphology to do microarray image processing has no requirements to spot shape, and it can be more accurate and effective. This method is to achieve spot gridding, denoising, enhancement, segmentation, [2, 18, and 24] & calculating the area and centroid intensity. Have also compare three methods result, ScanAlyze, method for Hongwei Li, & Angulo J. [19, 23]. Some clustering methods have been used for image processing of microarray data (Bozinov and Rahnenfuhrer in 2002) they have used k-mean [9, 13, and 23.] & partitioning around methods (PAM) on the two dimensional vectors of the intensities, and improve on this by considering only the pixels within the average spot shape, which turns out to be almost exactly a circle. Luis Rueda & Li Qin, present an optimized clustering based method, this method is used more than one feature to traditional clustering based method, and does not need a post-processing stage to eliminate the noisy pixel (Nagarajan in 2003) used the same method, but only on the intensities from the green channel. (Glasbey and Ghazal in 2003) considered a Gaussian mixture model for the two dimensional vector of the square root of the intensities. All of these methods consider only two clusters. Wu et al. used a k-means clustering algorithm [11, 12, 31], which is referring to as single-feature k-means clustering microarray image segmentation (SKMIS). They attempt to cluster the pixels into two groups, one for foreground, and the other for background [9, 23]. Steinfath et al (2001) fitted a scaled bivariate Gaussian distribution to pixel value, but using a robust fitting method. Brandle et al. (2003) described a robust fitting for gaussian spot model using an M-estimator. Schadle et al (1999) proposed an adaptive pixel selection algorithm to remove pixels contaminated by noise [15]. Kim et al (2001) used an edge detection method. They were aware of the problem of inner holes and used a threshold of intensity to decide the eligibility of pixels as foreground. And the application of edge detection technology on separating spots from the background decreases the probability of the errors and gives more accurate information about the states of spots such as the pixel number, degree of fragmentation, width and height of spot, and circumference of spot [27]. Hirata et al (2002) and Angulo and Serra (2003) used mathematical morphology [19]. Their methods can deal with blank spots, but not with spots with inner holes. Anders Bengtsson (Dec 2003) also used mathematical morphology this important to for low intensity spots where even a small
bias has great influence [28]. Glasbey and Ghazal (2003) used a combinational way to consider a variety of methods, including fixed circle, properties of histogram, and k-mean clustering with different preprocessing and different parameters. O'Neill et al (2003) recreate the background slide and subtract it. Their method is deal effectively with global artifacts that involve a substantial number of spots, but not with inner holes or more local artifacts. Software package, Spot [11, 12, and 29], is a prototype system for the analysis of microarray data. Spot is built on "R", an environment for data analysis, which is available as free software under the GNU Public License (GPL). As well as providing a wide range of graphical and statistical tools, R supports a well-designed and efficient programming environment. The result of an analysis by spot of microarray image data is returned as a number of ways. Spot is actually a specialized version of another R package called VOIR, which is currently being developed by the CSIRO image analysis group, and provides a more general image analysis environment [30]. Luis Rueda, they proposed Adaptive ellipse method, and show various advantages when compared to the adaptive circle method. Due to fact that most of the spots in a microarray have the form of circles or ellipses, this method utilizes the process of digitalization. Adaptive ellipse is capable of extracting information from the images, which is ignored by the traditional adaptive circle method, and here showing more flexibility. Due to fact that most of the spots in a microarray have the form of circles or ellipses, this method utilizes the process of digitalization. After the digitalization the process is applied, the coordinates of the pixels are real numbers. To computing the radius that determines the edge that separates the spot from the background region, As pointed out earlier, the pixel coordinates, in most of the cases, become real numbers, and so it is not possible to apply traditional edge detection techniques, such as the Laplacian transform. This is to compute the radius of the foreground region [6]. Software’s that use statistical parameter based on the t-test to analyze array data has been described (Long et al. 2001). Another basic statistical approach to compare data set is the analysis of variance ANOVA and its modifications MAANOVA (Kerr et al. 2000 and Churchill 2001) [11]. Yue Wang, et. al. has described the application of drug of high–density microarrays to examine the effect of drug on gene expression in yeast as a model system. A similar system applied to human breast cancer
cells and tissues would have direct utility in the identification and validation of novel therapeutics. [31, 32, 33]

2.1.5. Grid Alignment Methods: There are two types. First Automation & Second is under the category of image analysis approaches. First, Automation as 1) Manual grid alignment method, [21] 2) Semiautomated grid alignment method, 3) Fully automated grid alignment method. Second, is under the category of image analysis approaches: 1) Template based approaches is the most of previous literature or existing software like ScanAlyze, GenePix, QuantArray, Dapple, etc. And 2) Data-driven approaches: It involves a) finding grid lines, b) processing multiple channels, c) estimating grid rotation, and d) finding multiple grids. Also present the algorithmic issues related to I) tradeoffs between speed and accuracy, II) repeatability and parameter optimization, and III) incorporating prior knowledge about grids. [24]

2.1.6. Foreground Separation Methods: Foreground separation using 1) spatial templates, 2) intensity-based clustering, 3) intensity based segmentation, and 4) spatial and intensity information (hybrid method). Hybrid method is to consist of clustering image partitions, spatial template image partition, statistical testing, and foreground/background trimming. [24]

2.1.7. TIGR Software: TIGR spotfinder is image-processing software: Using this software the quantification of expression levels is possible. Another use is the data converting scanned image into numerical data. There are four types: 1) TM4, 2) MIDAS, 3) MeV, & 4) MADAM. MIDAS is to perform Normalization; MeV is an application of the identification of genes & expression pattern of interest. [34, 35]

2.1.8. Spotfinder: It requires four separate 16-bit or 8-bit grayscale tiff images. Spots on array are divided into contiguous sections or subarrays or grids. It requires number of parameters to build the grid: The spot spacing, Number of printing pins in horizontal and vertical direction, & Number of rows & column [4] of grid automatically. During the process pass the program automatically searches grid element to identify the spots, measure the signal & local background. The output can be exported to tab-delimited MeV file for whole slide. The MeV file can be passed to other TM4 program such as TIGR Viewer (MeV), or MIDAS. This program is limited to microarray image & problem we identified & targeted during development. The software requires multiple images
generated by scanning a single slide in both the references and queries channels using identical definitions of the scan area. The sensitivity and gain settings can be for all channels. It uses data from all channels (all located images) to define spot boundaries. Non-overlapping images or those of different sizes will cause the program to fail. [35]

2.1.9. Semantic Analysis: High-speed camera, Optical- instrumentation, sophisticated microscope optics. In semantic analysis in the context of biological imaging refers to the development and use of image processing and knowledge extraction tools and algorithm for automated image interpretation. This interpretation require different levels of processing, including low-level image processing like segmentation and feature extraction & high-level processing for extracting representing, & modeling Spatio-Temporal semantics of biological discovery and drug development. Different biological applications have different requirements in terms of semantic analysis. While some require only fluorescence intensity quantification other requires cell division and motility. Traditionally been on extracting simple image and object features like size, intensity, etc.

In Object Detection and Tracing Layer, It has two major functions. Firstly, it preprocesses the data to remove undesirable artifacts, and separates objects of interest from the background using segmentation algorithms. Secondly, it tracks the movement of biological objects in time- lapse images. Preprocessing and Segmentation: Imaging instruments are composed of different components. There includes microscopes, cameras, filters and lenses, all of these components add noise to biological images. Illumination may also vary from one image to another when a large population of cells is imaged. Images need to be processed to remove this noise. Variable illumination problem for microscopic imaging is addressed by illumination normalization method. Microscope optics blurs the image as a result of the limited aperture of the microscopic objective. Deconvolution algorithm attempts to remove this blur and improve the contrast in microscopic images. Cell and nuclei segmentation algorithm based on the watershed transform have also been proposed. [36]

2.1.10. Compression: Compression is necessary for efficient distribution and storage. There are two types: lossy and lossless compression. An object based coding technique to compress the segmented images using modified block coding with optimized truncation (EBCOT) [37, 38]. By extending EBCOT to arbitrary shaped objects, our scheme can
realize lossless coding of foreground and loss-to-lossless coding of background separately, a feature EBCOT does not offer, and achieve better compression results than other popular coding schemes like JPEG-2000, JPEG-LS and LZW. In mean time, microarray images contains large (huge) amount of data saved at a resolution of 16 bits per pixel, each image set is typically above 30MB in size, which demands highly efficient compression method [38]. Currently, the common method to archive microarray images is to store them losslessly in TIFF format with LZW compression but such approach does not exploit 2D correlation of data between pixels and does not support lossy compression. Due to this reason the huge data size microarray images require efficient compression algorithm, which support lossy compression, but also lossy compression with graceful degradation of image quality for downstream data analysis at low bit rates. Another method is segmented LOCO (SLOCO) [39] employs a two tier coding structure, this exploits the possibility of lossy-to-lossless compression for microarray images, which has been incorporated in the lossless compression standard of JPEG-LS, and coding is conducted on the foreground and background separately. In BASICA, we also incorporate lossy-to-lossless compression of microarray images. The aims are: 1) To generate progressive bit streams that can fulfill the requirements of signal processing & data analysis at low bit rates for data sharing and transmission applications & 2) To deliver competitive lossless compression performance to data archiving applications with a progressive bit stream, also they adopt distortion measure based on the log ratios and the log products because they are the most used transforms in common applications. [4, 8, 29]

2.1.11. Data Acquisition and Pre-processing: Mohsen Abbaspour developed new APEX technology. APEX probe ologonucleotide were printed onto specific grid positions [24 and 29] on microarray slides. The arrays were imaged a biochip reader, fitted with filter sets that allow four channels of data to be collected of each sample. The average spot diameter varies for different imaging resolutions from about 10 pixels for 10-micron resolution to 33 pixels for 3-micron resolution scanning and size of each channel varies from 8 MBs for low resolution data, to about 80 MBs for that of high resolution. Microarray channel images for gridding and segmentation, the former has a higher SNR. This is because many of noisy channels are saturated in each channel. The
summation increases signal level, without making any changes in noise level. The channels contain different background levels, caused by the varying response of dyes to different excitations. [23, 24, 25 and 29]

2.1.12. Separating Sub-Grids: The goal of segmentation technique is to apply our technique to each cell separately, making automated and fast. To solve this problem by using robust noise removal step. There are two types of noise are of concern: blobs & scratches. The background subtraction method applied removes the blobs to some extent, but not completely. The noise removal strategies are two. Blobs by applying closing followed by opening morphological filter, with disk shape structuring element having diameter is 2. The scratches effect is removed with the morphological closing and opening filter applied in four directions. Before computing the projection profiles, a contrast adjustment is applied to the image. [4, 25]

Microarray image processing is typically categorized to achieve the following tasks: Addressing (gridding), [2, 4, 5, 12, 19, 24] Spot Segmentation, intensity extraction, Background Subtraction and Normalization [11, 18, 27, 34]. Addressing is the process of assigning coordinates to individual spots on the microarray image, in order to facilitate individual spot segmentation and mapping of spot intensities to their respective genetic information. [25] The geometry of addressing can vary in a number of ways: 1.Basic structure: arrangement of spots within grids, 2.Translation: shift in all spot positions from image to image, 3.Rotation: rotation of the image with respect to a template [26]. Spot segmentation is the process of classifying the pixels of cells obtained through the gridding [4, 18, 19, 23, 24, 29] (is found in GenePix, ScaAlyze & Koadarray) [40] process as either foreground, which contains genetic information, background. Background area detection is important for subsequent quantification. Spot segmentation is affected by various image degradations such as different noise types and printing inaccuracies, as well as other detects including spot shape irregularities, such as donut-shape spot [41], background variations. Segmentation based on active contour models has been used but still fails to solve the problem of non-homogenous spots such as donut-shaped spots containing an inner hole. A wavelet based edge detection method was presented. This is sensitive to contamination of noise according to the reported results [42]. Yue Wang, et al two major data processing operations are involved: background
correction and interexperiment normalization [33, 34]. In background correction, local sampling of background can be used to specify a threshold that a true signal must exceed. It is even possible to accurately detect weak signals and extract a mean intensity of background of the target. The novel feature is to unify the tasks of estimating normalization coefficients and identifying control gene sets. Unification is realized by constructing a window function over the scatter plot defining the subset of constantly expressed genes and by affecting optimization using an iterative procedure. The recovery of normalization regression and control genes selection are interleaved and are realized by applying coupled operations to the mean square error function [39]. In general normalization is the process of minimizing variations. This is usually performed software is SNOMAD. Normalization is divided in four categories: global, non-global, linear, & non-linear. Global normalization means the parameters are estimated using the data from all features of the array. Non-global normalization means parameters are estimated using data only from a certain set of genes or they are applied to only part of the array i.e. printing tip group. Many linear normalization methods are based on the assumption of equally of the intensities of two channels, this balancing of the dyes can be utilizing a constant scaling factor to the other channel i.e. linear regression, ANOVA, etc. Non-linear is the most effects of variation are not linear in nature, e.g. the bias where ratio values have dependence on spot intensity (lowpass) [11].

Component extraction and shape analysis: An eight-connected morphological connectivity labeling stage is applied to extract objects constituting a connected component. [42]

2.1.13. DNASER I: Layout and Data Analysis for real time acquisition and elaboration of images from fluorescent DNA microarrays. DNASER (equipped with the ORCA II Hamamatsu CCD camera) for the analysis of traditional and innovative DNA microarrays. The spot analysis algorithm is fully automated and does not require any traditional information about the DNA microarray geometry. According to the relative position, the spot feature is stored in a 3-D structure that is appealing to build an efficient database, for classification and further processing purposes. It can be used to characterize the cellular differences between different tissue types such as between normal cells and cancers with
different responses to treatment, or between control cells and cells treated with a particular drug. [43]

2.1.14. BASICA: It to provide processing, background adjustment, compression and analysis of cDNA microarray images. It uses fast Mann-Whitney [12, 13, and 26] test-based algorithm to segment cDNA microarray images and performs postprocessing to eliminate the segmentation irregularities. Jianping Huai, introduce a new distribution measurement for cDNA microarray image compression and device a coding scheme by modifying the EBCOT algorithm (Taubman 2000) to achieve optimal rate-distribution performance in lossy coding while still maintaining outstanding lossless compression performance, which is now incorporated in the JPEG2000 standard. In segmentation, the background adjustment component estimates the each spot foreground and background intensities and calculates the log ratios [4, 29] values based on the background-subtracted intensities. [8]

2.1.15. Edge Detection for Signal Extraction: Generally, many chances to make errors during the manufacture of cDNA microarray chip, hybridization of mRNA extracted from the samples, & scanning of chips. The main remaining problems however are sensitivity of detection, reproductivity and processing. During processing of microarray images, especially irregularities of spot position [23] and shape could generate significant errors: small regions of signal spot can be mis-included into background area. Application of this method on separating spots from the background decreases the probability of the errors and gives more accurate information about the state of spots such as pixel number, degree of fragmentation, width, height of spot. Such information can be used for the quality control of cDNA microarray and filtering of low quality spot [40]. HDG, like Melanie II, identifies candidate spots by tracing their edges. Spot morphology on membranes is quite variable, so edges may be of arbitrary shape and may even be locally concave. HDG does not use the geometry of the array to direct its search for spots. The lack of positional bias in spot finding is appropriate for membrane-based arrays because it is robust to nonlinear deformations of the membrane, which commonly occur during image acquisition. [17, 29]

2.1.16. Optimized Clustering: In this method is to consider the shape of the spot, the pixel whose distance to the center of the spot is smaller is more likely to be foreground
pixels. And also thus can be take this spatial information about the pixels into account, and construct different features, i.e. the distance from the pixel to the center of the spot in the x-direction and in the y-direction [11]. Alternatively this can take Euclidean distance from the pixel to the center of spot as a feature and to calculate the coordinates of the spot center. When considering more than one feature, the normalization [34, 43] process is very important, not only by scaling, but also analyzing the correlation between each pair of features. Another point is that principle component analysis (PCA) is a widely used method in future. Manuele Bicego, et al., used PCA technique, for investigation of dimensionality reduction. [44]

2.1.17. Dapple: Is a new program for finding and quantitating spots on microarray images, which depart from previously, described implementations in two ways. First, Dapple's spot finder takes advantage of consistent spot morphology (i.e. circularity) to increase its robustness both to image noise and to variability in spot position and size. Second, the program learns the investigators concept of spot quality by example, using a classifier, which can be trained, on manually classified examples of the spot finder's output. Solving the spot finding problem is crucial to making accurate expression measurement because errors in spot finding, the first, step in processing microarray data, propagate to all subsequent analysis. Three constraints combine to make the problem difficult. First the spot finder must be robust to substantial uncertainties in spot size and position [4, 23, and 29] caused by variations in the amount of DNA on each spot and in the location where it is spotted. Second, the finder must cope with both diffuse image noise and discrete image artifacts arising from airborne particles or non-uniform washing of the array surface. Third the spot finder must be efficient and effective when applied to large number of spots. It performs parameterization: the user manually [21] classifies a training set of candidate spots qualitatively according to their perceived accuracy, after which the software adjust its classifier to best reproduce the user's judgement as expressed on the training set. Parameterization by basic example is machine learning [17, 29, 43]. This software finds spot by detecting edges of spot & calculates the negative second derivative of the image (laplacian).

2.1.18. K-means Driven Adaptive Image Restoration: K-means is nonhierarchical clustering [11, 12]. This is combined 1/a gridding for locating individual cell images, 2/a
clustering for assessing local noise from the spot's background, and 3/a wiener restoration filter, for enhancing individual spots. The effect of the proposed technique quantified using a well-known boundary detection method. The application of this method on cDNA microarray images resulted in noise suppression and facilitated spot edge detection. [23, 31, 45]

**2.1.19. SOM:** Self-Organizing Map [11] is ideally suited for explorative data analysis where prior information about the distribution of the data is not available. SOM is a method for producing ordered low-dimensional representations of an input data space. Typically such input data is complex and high dimensional with data elements being related to each other in a nonlinear fashion. A multiple dimension DNA microarray data matrix is constructed, each node representing a point on the DNA microarray. Then random node is selected for according to the pattern of expression. By using this model the results are very easy to visualize and interpret. [31].

**2.1.20. MAIA:** This software is to develop algorithm ensuring higher level of intelligence and robustness. Eugene Novikov, present a fully functional software package for automatic processing of the two-color microarray images including spot localization, quantitation and quality control. This aims at making ratio estimates more resistant to array contamination and after automatic tools to evaluate spot quality control. The spot localization module (i) identifies the position [29] of each spot on the array to associate it with the spotted clones; and (ii) establishes the borders between the neighboring spots to allow further independent data processing for each spot. MAIA deals with the most widespread, orthogonal, spot localization pattern. In this pattern, the spots are aligned horizontally and vertically and can be arranged in block containing different numbers of spots rows and columns [4]. The spot quantification module generates two sets of quantitative parameters for each spot: ratio estimates and spot quality characteristics. Two algorithms for ratio estimate first, a direct arithmetic ratio of the background-corrected fluorescence intensity estimates in the two color channels. This approach requires the identification of both the foreground and background typically the level of non-specific hybridization. The second ratio estimates is the slope of the linear regression plotting of the pixel intensities in one-color channel verses another one. This approach does not require spot segmentation. Spot quality module provides a value of spot quality
reflecting the level of confidence in the obtained ratio estimate at each spot. The unique spot quality value is derived from a set of nine marginal quality parameters characterizing certain features of the spot. Image simulation: this software component for Monte-Carlo simulation of the microarray images. It takes into account the statistical noise and different types of contamination like non-specific hybridization and dust, the true values of the ratios are precisely known, it allows us to evaluate, to test and to compare different algorithms of microarray image analysis. [23, 46]

2.1.21. UIMA: Unstructured Information Management Architecture. It requires domain expert to implement Analysis Engine which enable computers to extract useful information from unstructured contents, such as annotations in a common analysis structure, an object-based data structure, for representing and sharing the structured information in the framework. [47]

2.1.22. MIA: This system for solving two existing critical problems in microarray image analysis and data management. MIA system is designed to offer a flexible, scalable, and extensible environment, and to provide accurate microarray analysis results without the need of manually correcting [21] the image processing results. Such as accurately addresses each spot, and handles uneven background and noise in slide. Noise removal is one of the major contributions as compare to GenePix is fail. A simple or effective method for spot segmentation is proposed and applied to each cell in grids. This method MIA system is robust and outperforms GenePix. This system includes four major modules: 1) slide information module, 2) slide blocking module, 3) slide gridding module, 4) slide segmentation module. The success rate of donut-shaped spot by using this system is 89.84% & 90.00% for noisy spots, while GenePix can’t handle donut-shaped spots; the success rate of noisy spots is only 54.00 %. [47]

2.1.23. Lucidea Microarray ScoreCard: This s/w provides an integrated system of controls and analysis s/w for the assessment of data quality from microarray experiments. The quality metrics provide measurement for validating and standardizing microarray results. This system includes two-color normalization & provides quality metrics for validation, statistical analysis, &significance testing for gene expression data. In quality metrics allow for data extraction, normalization, and hybridization. This method is also
allowing easy comparison of the normalized values with the target values, & calculates Standard deviation & log ratios. [34]

2.1.24. CLAHE: The contrast limited adaptive histogram equalization, technique is beneficial for image enhancement, for enhancing individual cell-images and thus, for facilitating accurate cell-spot detection. [5]

2.1.25. Image Denoising Using Two Stage Multiresolutions Technique: Microarray image analysis, which takes into account the effect of local spot image noise in microarray images for improving spot segmentation and subsequently gene quantification. Microarray images are affected by inherent noise. A two-stage noise removal approach that the processes additive and multiplicative noise component. This approach first decomposes the signal by a multiresolution transform and then accounts for both the multiscale correlation of the subband decompositions and their heavy-tailed statistics. Change in the measured transcript values in the samples renders the clustering [11] of genes into functional groups and the classification of samples difficult. A major challenge in microarray analysis is to eliminate the effect of the noise and recover the gene expression measurements & repetitions can increase the significance of conclusions from gene array experiment. The importance of including both additive and multiplicative measurement specific noise in an error model for gene arrays are already exists. The major source of noise is often overlooked by researcher, is caused by the microarray image generation process. Usually involves the collection of fluorescence of the labeled samples, the amplification of the analog signal and the conversion to digital through dedicated imaging devices. The microarray area is divided into equally sized pixels and the imaging devices produce a digital map of the fluorescence intensities for each pixel in the form of 16-bit TIFF image files [4]. By increasing the detection settings, the source noise, which includes photon noise and dust on the slides, remains unaffected. On the other hand, the detector noise, which includes feature of the amplification and digitization process, is increased. In high throughput whole-genome approaches, applying a two-stage “correlation and coring” approach enhances the dynamic range of exiting microarray imaging technology, which is very important in order to identify the most significant genes with increased accuracy and robustness. This method is to producing multiple three images of the same microarray at different detection settings because it
yields better results for the image obtained at high detection settings [48]. Hirata, Nina S.T. suggests that the main idea is inspired from stacked generalization and consists in, at each training level, combining the outcomes of previous level operators. The final operator is a multilevel operator that ultimately depends on a large neighborhood than of the individual operators that have been combined. They also show that iterative two-level operators are an effective multi-level approach to obtain better results [49]. Yuan-Kai Wang, has solved the problem of image rotation without manual adjustment, [26, 43] the detection is robust with respect to irregular spot gaps, It is not affected by variations of color and size of spots, [44] this is also performs well on different number of input channels. But it cannot only handle multiple spots, but also find more precise contour of spots. Wavelet based profile analysis is performed for building orthogonal grid system of the image. The rotation angle can be found by searching for the maxima of the standard deviation. Both vertical-horizontal axes perform wavelet. [50]

2.1.26. Feature Extraction: Feature extraction is depends on these factors: spot intensity, spot size, spot morphology, pixel intensity distribution, and spot quality classification. Yihui Liu, perform wavelet analysis on high dimensional microarray data. Two methods of feature extraction are to characterize the features of gene profiles on microarray data, using wavelet approximation coefficients and detail coefficient respectively. Approximation coefficients compress the gene profiles and reduce the dimensionality of microarray data. The features projected wavelet bases are used to feed into support vector machine [23, 44] for classifying the different diagnostic classes. Researchers focus on three different supervised machine-learning techniques in cancer classification, decision tree, and bagged and boosted decision trees. [51]

2.1.27. Machine Learning: Man Wai Mak, suggest various approaches of machine learning to this are challenge also. First, clustering of gene expression data: This is popular approach is to identifying group of genes or samples of similar expression patterns [11]. In Teng and Chen, discovering biclusters by iteratively sorting with weighted correlation coefficient in gene expression data. Lam et al., A regularized clustering algorithm based on calculus of variations. This is useful fuzzy C-means algorithm. Second, gene selection method is Gan et al. propose effectively gene selection method using Bayesian Discriminant based criteria and genetic algorithms. Third, spot
segmentation, this is a comparison of fuzzy clustering spot-segmentation approach for quantification of microarray gene expression, Wang et al., propose a fuzzy-clustering spot-segmentation method that can handle array spots with complex shapes such as donuts and scratches. Fourth, is dynamic of gene expression. Fifth, inferring gene networks, recently there has been much interest in inferring gene and protein interaction networks. [43].

2.2. Comparison Study of Existing Techniques:

**Table 2.1:** Description of microarray image analysis methods used in the comparison study: name of the technique, software’s. [2]

<table>
<thead>
<tr>
<th>Name</th>
<th>Software</th>
<th>Segmentation</th>
<th>Background</th>
</tr>
</thead>
<tbody>
<tr>
<td>S.nbg</td>
<td>Spot</td>
<td>Segmented region growing</td>
<td>None</td>
</tr>
<tr>
<td>GP.nbg</td>
<td>GenePix</td>
<td>Proprietary algorithm that result in adaptively sized circles</td>
<td>None</td>
</tr>
<tr>
<td>SA.nbg</td>
<td>ScanAlyze</td>
<td>Fixed circles, 10 pixels in diameter</td>
<td>None</td>
</tr>
<tr>
<td>QA.fix.nbg</td>
<td>QuantArray</td>
<td>Spot intensity is the mean of pixel values between the 45th and 85th percentile within a fixed circle of 9 pixels in diameter</td>
<td>None</td>
</tr>
<tr>
<td>QA.hist.nbg</td>
<td>QuantArray</td>
<td>Spot intensity is the mean of pixel values between the 80th and 95th percentile of a 11-by-11 pixels square</td>
<td>None</td>
</tr>
<tr>
<td>QA.adp.nbg</td>
<td>QuantArray</td>
<td>Chen's method, with a circular target mask of 10 pixels in diameter and 1 0.001 p-value cut-off.</td>
<td>None</td>
</tr>
<tr>
<td>Algorithm</td>
<td>Software</td>
<td>Methodology</td>
<td>Description</td>
</tr>
<tr>
<td>-------------</td>
<td>----------</td>
<td>------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>S.valley</td>
<td>Spot</td>
<td>Seeded Region Growing</td>
<td>Median from &quot;valley of spot&quot;</td>
</tr>
<tr>
<td>GP</td>
<td>GenePix</td>
<td>Proprietary algorithm that results in adaptively sized circles.</td>
<td>Median from &quot;valley of spot&quot;</td>
</tr>
<tr>
<td>SA</td>
<td>ScanAlyze</td>
<td>Fixed circles, 10 pixels in diameter.</td>
<td>Median value in local square region</td>
</tr>
<tr>
<td>QA.fix</td>
<td>QuantArray</td>
<td>Spot intensity is the mean of pixel values between the 45th and 85th percentile within a circle of 9 pixels in diameter.</td>
<td>The mean of pixel values between the 5th and 55th percentile of the background mask. The background mask is the region between two circles with diameter of 11 and 13 pixels, and concentric with the spot mask.</td>
</tr>
<tr>
<td>QA.hist</td>
<td>QuantArray</td>
<td>Spot intensity is the mean of pixel values between the 80th and 95th percentile of a 11-by-11 pixels square.</td>
<td>The mean of pixel values between the 5th and 20th percentile of a 11-by-11-pixel square.</td>
</tr>
<tr>
<td>QA.adp</td>
<td>QuantArray</td>
<td>Chen's method with a circular target mask of 10 pixels in diameter and a 0.001 p-values cut-off.</td>
<td>The mean of the median 8 background pixels in the background mask.</td>
</tr>
<tr>
<td>S.morph</td>
<td>Spot</td>
<td>Seeded Region Growing</td>
<td>Based on morphological opening. The structuring element is a square region with sides of length 2.5 times the approximate spot to spot separation.</td>
</tr>
</tbody>
</table>
2.3. Drawbacks of Existing Microarray Image Segmentation Methodologies:

Kashif I. Siddiqui, suggest that the challenging problems of gene clustering, feature extraction, and data mining. A major issue in gene microarray data analysis is to accurately quantify spot shapes and intensities of microarray images [1]. Luis Rueda, proposed method is facing the problem to compute the slope of the probability density function for each pixel, and then find the radius that generates the largest slope average [6]. Recently, Weng Guirong, mention the mathematical morphology method usually neglects the spot of low intensity, which results in its low stability. Warps come from the mechanical framework of laser scanner and anmorphic spots cannot be extracted accurately [19]. Some existing methods include manual processing of DNA spot images using a generic image-processing tool, such as NIH image. Using such a tool a user visually locates each DNA spot image in a micro-array image, and moves a display pointer to each spot image, and manually defines a small area around the spot image. The image processing tool then reports image intensity values within the small area. The user then manually records the intensity values and continues this process for other visually located DNA spot images in the micro-array image. So, such manual systems are impractical for micro-arrays with more than a handful of spot images, tedious and repetitive, requiring considerable time and effort. For example, with a micro-array image having about 600 DNA image spots, such manual methods can take about 8 hours of work, and resulting in quantification of only a limited number of image spots which visually seem to have a “good” expression level. As the micro-array density increases and becomes more complex, use of such methods becomes even more prohibitive. For example, current micro-array sizes range from several hundred to 1,200 genes, arrayed in a 1.8×1.8 cm area. As tip fabrication has improved, arrays with greater than 50,000 genes are available. Such methods are also prone to various errors, including errors in manually

<table>
<thead>
<tr>
<th>S.const</th>
<th>Spot</th>
<th>Seeded Region Growing</th>
<th>Constant subtraction; the constant value is the 3rd percentile of all the foreground spot intensities.</th>
</tr>
</thead>
</table>
recording the intensity values. Further such methods provide inconsistent quantification of intensity values, both for different spot images measured by a single individual, and for multiple individuals making measurements from the same micro-array image. Some existing methods automate the process of locating DNA spot images from micro-array images and quantifying corresponding expression values. Such methods utilize a computer to manually position a cell grid on an area of the micro-array image containing an array of DNA spot images. The grid used in such methods is either completely fixed in shape, or has limited global flexibility (e.g., resizing and rotating the entire grid). The printing tips are difficult to fabricate and many do not work uniformly. Existing methods are unable to cope with irregular micro-array pattern, search for DNA image spots, and accurately quantify specific signals while accounting for the local background. Another existing techniques do not use a grid at all but apply a “spot” filter to detect locations in the microarray image which “look-like” DNA spot images, using such methods it is difficult to define what a spot should look like. Another thing is that, extensive noise and variations in the spot shape, due to the processing and scanning mechanisms, significantly reduce the signal to noise ratio (SNR) of the spot images. The detection scheme misses many real spots and processes many false patches in the image as real DNA spot images. Another drawback is their inability to display micro-array image pixel intensities, corresponding to gene expression values in related DNA spots. There is a need for automatically segmenting DNA array images into individual DNA spot images for quantification. There is also a need for such method to process irregular micro-array patterns, search for DNA image spots, and accurately quantify, and intuitively display, specific signals while accounting for the local background [21].

Peter Bajcsy, suggest it might be resolved without any accuracy loss by using either supercomputers or distributed parallel computing with grid-based technology, it might still be beneficial to design image-processing algorithms that could incorporate such resource limitations [10, 24]. Mohsen Abaspour suggests future research flagging spots, distribution analysis, data modeling and statistical study of pixel intensities across each spot [25]. H. Samartzidou, L. Turner, & T. Houts has mention problems in scorecard software with data quality, some of which are not easily detected by simple visual examination of microarray images, it can improve the process of microarray data analysis
by allowing only validated data to proceed to archiving, data visualization, and data mining [34]. By using Watershed transform have also been proposed watershed algorithm generally suffer from the over segmentation problem & require a post processing step for improving segmentation accuracy. Accuracy and fully automated segmentation still remains a challenge [36]. Hirata, Nina S.T. suggests that the design of morphological operators that are translation-invariant and locally defined by a finite neighborhood window correspond to the problem of designing Boolean functions. This operator designed from training sample also suffers from over fitting [49]. Yihui Liu, suggest the problem can be regarded as a classification problem in machine learning. Generally microarray expression experiments allow the recording of expression levels of thousands of genes simultaneously [12, 43, 51]. Robert A Lotufo, suggest the segmentation of the nucleus in the reference image is the main challenge of the image analysis [52].

The proposed work will emphasize to develop the segmentation technique for microarray images. To cater the need of microarray imaging to solve the problem specified above. This technique may also improve the most of the above mention drawbacks and identify dense and sparse patterns, correlation among attributes and overall distribution patterns and identify outliers therefore useful to detect anomaly (irregularity). These results can easily be understood and used to make interpretation Bioinformatics/Biotechnologist.

2.4. Summary: In this chapter we have discussed various microarray image segmentation methodologies, & did the compartative analysis. In this we concentrated on microarray image segmentation in detail considering each and every methods advantages and dissadvantages. Useful methods out of all for our work were fixed circle, adaptive circle, adaptive shape, & Histogram segmentation. Fixed circle segmentation is very easy to implement but it applies only with constant diameter to all spots in the image, which is inadequate to read all spots. In Adaptive circle method, the circle's diameter is estimated separately for each spot, which resolves the problem of varying diameter but it consumes more computable time. The adaptive shape methods are uses two methods SRG & watershed. Both require the specification of starting points or seeds which is very helpful in our experimental work thus major study was done on it before actually implementing it and tried with different variation which is mention in the next chapter.
References:


[23] Sharon Grrenblum, Max Krucoff, Jacob Furst, Daniela Raicu, "Automatic image analysis of noisy microarrays", Reference link=> facweb.cti.depaul.edu/research/VC/…/VISAPP-4-pages.pdf


Semantic analysis of biological imaging data: challenges and opportunities.


Jianping Hua, Zixiang Xiong, Qiang Wu, and Kenneth Castleman, “Fast segmentation and lossy-to-lossless compression of DNA microarray images”,


Mohsen Abaspour, Rafeef Abugarbieh, Mohua Podder, and Scott J. Tebbutt, “Hybrid spot segmentation in four-channel microarray genotyping data”, proceedings of IEEE international symposium on signal processing and IT, 2006


