Introduction to Microarray

"Start by doing what's necessary, then do what's possible and suddenly you are doing the impossible" – Saint Francis of Assisi.

1. What is Microarray?
Microarray also called gene chip or biochip [1, 2]. To study how large number of genes with each other & how a cell's regulatory network control vast bacteria's of genes simultaneously. Microarray is a very small chip. It is a matrix [3] like structure rows & columns [2, 4]. It contains thousands of spots; [4, 5] each spot contains store different type of genetic information & each spot contains many, copies of a single DNA sequence such as gene [6]. And these slides after scanning gives image that contain colored spots each spot intensity represents the amount of the expression level of gene. Microarray provides an easy way to compare gene expression profiles between biological samples, by detecting their differential expression [7]. Generally DNA spots are left on the slide using a robotic arm, as their dimension does not allow a hand-made positioning. Microarray is one of the most recent and important technologies for exploring the genome. In fact, this objective can be accomplished only through a multidisciplinary approach to data analysis in which biological knowledge are integrated by bioinformatic skills and advanced statistical applications. [8]
In general, microarray is mainly two categories: Spotted DNA microarrays, and oligonucleotide gene chips. Spotted DNA microarrays: DNA probe is delivered onto the array by robotic machines. Oligonucleotide gene chips: DNA probe is in-situ synthesized on the surface of the DNA chip. [9]

1.1. Role of Microarray in Bioinformatics:
Microarray is new techniques to investigate the expression levels of thousands of genes simultaneously [5,10]. Microarray present new statistical problems because the data is very high dimensional with very little replication methods of adjustment for multiple testing therefore becomes extremely important. Microarray offers an exciting entry point for statistics into the modern areas of computational biology & bioinformatics.

1.2. Goal of Microarray:
- **Find the genes:** that change expression between experimental and control samples
- **Classify samples:** based on a gene expression profile
- **Find patterns:** Groups of biologically related genes that change expression together across samples/treatments.

1.3. Types of Microarray:

1.3.1. DNA Microarray: A high-density grid of tiny DNA (or oligonucleotide) spots. [11] or A DNA microarray is a collection of microscopic DNA spots attached to a solid surface, such as glass, plastic or silicon chip forming an array. Scientists use DNA microarrays to measure the expression levels of large numbers of genes simultaneously. The affixed DNA segments are known as probes, thousands of spot, which can be used in a single DNA Microarray. [12]

1.3.2. cDNA Microarray: Also known, as biochip, DNA chip, or gene array, cDNA microarray technology allows for identification of gene expression levels in a biologic sample. cDNAs or oligonucleotides, each representing a given gene, are immobilized on a small chip or nylon membrane, tagged, and serves as probes that will indicate whether
they are expressed in biologic samples of interest. Thus, the simultaneous expression of thousands of genes can be monitored simultaneously. [5, 10, 13]

1.3.3. **Microarray**: Arrangement of oligonucleotides (a linear sequence of nucleotides joined by phosphodiester bonds) or DNA fragments on a chip to eventually study many genes or entire genomes simultaneously. [14]

1.3.4. **Tissue-Microarray**: Used to analyze the expression of genes of interest simultaneously in multiple tissue samples, tissue microarrays consist of hundreds of individual tissue samples placed on slides ranging from 2 to 3 mm in diameter. Using conventional histochemical and molecular detection techniques, tissue microarrays are powerful tools to evaluate the expression of genes of interest in tissue samples. [15]

1.3.5. **Genome Tiling Microarray**: A type of microarray in which overlapping oligonucleotides are designed to cover a genomic region of interest. [16]

1.3.6. **Conventional Microarray**: A 2D array of discrete spots of different molecules. Each spot is made up of millions of identical molecules, within the limits of practical purity. [17]

1.3.7. **Protein Microarray**: A protein microarray is a piece of glass on which different molecules of protein have been affixed at separate locations in an ordered manner thus forming a microscopic array. These are used to identify protein-protein interactions, to identify the substrates of protein kinases, or to identify the targets of biologically active small molecules. [18]

1.3.8. **Microarray Analysis**: The function of cells, compare the differences between healthy & diseased tissue & observe changes with the applications of drugs [19]. Or a powerful tool for identifying genes that is associated with complex biological phenomena. Microarrays contain human cDNAs of known and unknown sequences, which are printed on glass slides using high-speed robotics. These DNA 'chips' are used to quantitatively monitor differential expression of the human genes using a highly sensitive two-color hybridization assay. [20]

1.3.8.1. **Visualizing Microarray Data**: Create a figure to visualize Microarray & get the data ready for analysis.

1.3.8.2. **Analyzing Gene Expression Data**: Analyze Microarray data for patterns & plots the results. [20]
1.4. Microarray Process:

Figure 1.1 shows the whole process of microarray technology. Firstly, biologists the RNA/probe preparation & target spotting is done by using DNA chip. In RNA preparation use the red color & for the DNA target spotting use the blue color. Secondly, by using dye fluorescence label (green) does hybridization. Then, the next step this hybridized slide is scanned by using microarray scanner. After that, slide scanning is to found the raw microarray image. This image is also need for computer analysis, such as to remove the noise from the image, enhance the image, to segment the image, etc. for microarray spot recognition.
1.5. Microarray Quality Controls:
Table 1.1 shows the properties and main utilities of four common control types. Controls give us useful information about quality of printing, efficiency of target preparation, specificity and sensitivity of hybridization, dynamic range of fluorescence signal and also serve as "landing lights" for analytical software. [9]

Table: 1.1: Microarray Controls:

<table>
<thead>
<tr>
<th>Type of Control</th>
<th>Composition</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Pooled genomic DNA</td>
<td>Control for labeling and hybridization success</td>
</tr>
<tr>
<td>Negative</td>
<td>DNA fragments derived from unrelated species</td>
<td>Specificity of hybridization detection limit</td>
</tr>
<tr>
<td>Ratio</td>
<td>Two identical sequences spiked into each sample before labeling at different amounts</td>
<td>Success of labeling and hybridization color discrimination</td>
</tr>
<tr>
<td>Dynamic Range</td>
<td>Different sequences spiked into samples before labeling at different molar amounts</td>
<td>Success of labeling and hybridization color balance dynamic range of detection limit and saturation of signal.</td>
</tr>
</tbody>
</table>

1.6. Steps of Construction of DNA Microarrays:
1. Probe DNA preparation: In DNA probe preparation including these steps: primer design, oligonucleotide design, sensitivity and specificity, primer and oligonucleotide modifications, PCR amplification, PCR product purification, PCR product quantification and quality assurance.
2. Printing probe DNA: In DNA probe printing includes: a) array substrate, b) spotting: spotting is also perform different ways: i.e. Mechanical microspoting, Ink Jetting, Spotting solutions, etc., c) DNA probe fixation, & d) quality control of the spotting procedure. [9]

1.7. Steps of Preparation of Target DNA:
- 1. RNA isolation.
- 2. RNA labeling.
- 3. Target purification.
- 4. Target and RNA amplification.
- 5. DNA labeling.

Hybridization
Washing
Image Capture
Data analysis
Image processing, and
Normalization. [9]

1.8. Microarray Image Processing:

1.8.1. Introduction:
The task is “How can one draw biologically meaningful conclusions based on microarray image data and information extracted about gene expression levels?” Since the invention of microarray technology in 1995, researchers developed several microarray image processing methods, statistical models and data mining techniques that are specific to DNA microarray analysis these are Grid Alignment, Foreground separation (segmentation), Spot quality assessment, Quantification, Normalization, Identification, and Data Mining [21, 22, 23]. Gridding is the process of assigning coordinates to each cell; the latter is a square ROI containing the pixels of both the spot and its background. Segmentation classifies cell-pixels as foreground or background. Intensity extraction calculates ratios of red and green fluorescence intensities for foreground and background respectively. The difference of two samples can be the spots of strong red intensity shows that their genes have higher expression level, and genes represented by those of green intensity have the lower one. When two genes have similar expression levels, their spots are yellow [1]. Normalization is a process of minimizing variations. This performs by the software SNOMAD. [18] Data mining is employed to group genes so that molecular biologists may extract meaningful biological information or make assumptions regarding unknown genes. Gene quantification is confounded by number of technical factors, which operate during the fabrication, target labeling, and hybridization stage. [24]
1.8.2. Microarray Technologies:
In microarray technologies understanding cellular processes & the relationship between cells of differing function and metabolic pathways is essential for the understanding of the life sciences. The end of the last century the ability to measure gene expression or DNA polymorphisms were restricted to individual genes throughout the traditional separation and hybridization methods. [21]

1.8.3. Requirements:
Firstly, to understand the variations of microarray images such as, Input image including foreground and background morphology (e.g. grid layout, spot location, shape and size) [23, 25], Intensity information (e.g. spot descriptors derived from foreground and background intensities) [26], geometric measures (volume, area, and length) [27]. These variations have to be compensated by microarray image processing algorithms so that the processing performance meets excepted accuracy and speed requirements [28].

1.8.4. Ideal Microarray Image:
Microarray image contains grid geometry, known background intensity with zero uncertainty, pre-defined spot shape (morphology). Another aspect is statistical measurements would be obtained with a very large number of pixels per spot. [28]

Fig. 1.2: Ideal Microarray Image
1.8.5. Creation of Scanned Microarray Image:

Robotic Arrayer: cDNA microarrays consist of individual DNA sequence printed in a high-density array on a glass microscope slide using robotic arrayer. For mRNA samples, the two samples or target are reverse-transcribed into cDNA, labeled using different fluorescent dyes, then mixed and hybridized with the arrayed DNA sequence or probes. After this the slides are imaged using a scanner, which makes fluorescence measurement for each dye [21]. The special optomechanical instrumentation and the wide-band ORCA II camera allow acquiring DNA microarray images at different frequencies simply changing two different interference filters. Complex confocal microscope (it is to focus the laser beam on the sample), multicolor analysis system (to compare samples with reference data), image processing routines (to extract samples characteristics), static analysis routines (to extract the gene characteristics), and database (to store the categories of genetic expression) [6], Laser scanning confocal microscope [29] or scanner & CCD camera [30 and 31]. A microarray scanner performs an area scan of a slide and produces for each dye a digital map, or image, of the fluorescent dye molecules is converted into digital signals. A typical microarray laser scanner operates with the following parts, user can set the Scanner settings- scan rate, laser power, user can adjust PMT voltage. There are Functions- excitation emitted light collection, spatial addressing, excitation/emission discrimination, and detection. Another part is the sources of non-circularity includes the printing process (e.g. insufficient time of dehydration) [21, 28]. DNASER is (equipped with the ORCA II Hamamatsu CCD camera), for the analysis of traditional and innovative DNA microarrays. [6]

1.8.6. Sources of Microarray Image Variations: Slide fabrication, mRNA preparation, sample handling, and difference in fluorescence dye labeling, gene hybridization, robotic spotting, and green & red fluophores excitation by lasers, imaging using optics, slide scanning, analog to digital conversion using charge - coupled devices or photomultiplier tube, slide-to-slide variation, and finally storage of image. [2, 9, 21, 23]

1.8.7. Variations of Microarray Image Channels: microarray image data can consist of any number of channels. It is possible to foresee the use of more than two or three channels in future, for example, by using hyperspectral imaging. Another variation of
microarray image channels is the storage file format, data compression and data accuracy (number of bytes per pixel). [21, 23]

1.8.8. Variations of Grid Geometry: If a spotting machine with several dipping pins prints multiple 2D arrays of spots, then the dipping pins might bend over time and cause irregularly in a 2D arrangement of the printed spots. Other sources of variations are the slide material, such as nylon filters; glass slides, and probes type, such as radioactively labeled probes and fluorescently labeled probes. These variations can be occurred by mechanical strain (nylon filters), or by low discrimination power for small surface area (glass slide), strong background signal (fluorescently labeled probes) or strong signal interference of neighboring spots (radioactively spots). It is important to understand the strain extreme cases in order to limit the search space of grid geometry. [21, 23]

1.8.9. Variations of Background & Noise: Contamination is another major issue for microarray, such as dust, impurities, etc. [2] Background variation occurs due to the microarray slide preparation (spotting errors and hybridization), acquisition procedures (presence of dust), and image acquisition instruments (non linearity). It contains unknown background noise such as spike noise [21, 23]. The uncertainties in spot finding on microarrays are variable spot size and position [25, 26], discrete image artifacts. The solution containing transferring drops from the pin tips to the slide surface. The relative placement of adjacent grids is therefore determined by the spacing between adjacent pins, which may vary [23, 33].

1.8.10. Variations of Spot Morphology: The majority of current cDNA microarray imagery is produced with circular spots as shape primitives; one can find the use of other primitive shapes than a round spot shape will be used in microarray technology in the future. It also corrupted by irregularities in the shape, size, and position of the spot [7,12].

1.8.11. Microarray Image File Formats:
Image file format can be separated into three broad categories: lossy and lossless compression formats, and uncompressed formats. The lossless file formats are PNG (Portable network graphics), BMP (Windows Bitmap), GIF (Graphics interchange format), and TIFF (Tagged image file format) & Lossy file format is JPEG (Joint photographic expert group). JPG/JPEG file format using the compression algorithm
based on discrete cosine transform (DCT). It is recommended to use microarray image file formats without lossy image compression.

Laser scanning of a cDNA or oligo microarray slide generates two 16-bit TIFF files. This is suitable for saving 1-bit (binary), 4-bit, 8-bit (byte) and 16-bit (short) data. The choice of 16-bit per pixel is based on the dynamic range of fluorescence measurements and sensitivity of laser scanner. The TIFF file format specification version 6.0 is publicly available it also includes image compression option. It is not recommended to use any lossy compression in order to prevent spot information loss, and to avoid increased uncertainty of extracted spot statistics.

BMP file format are generally larger without compression so not optimal for much. It handles the graphic files within the Microsoft Windows operating system. The advantage of this file format is their simplicity. [21, 33]

1.9. Applications of Microarray:
organism's metabolism, [1] Pharmacology, molecular biology research and genomic clinical diagnosis [5] and type1 type2 diabetes diagnosis, HIV infection, [6] Medical diagnosis such as M. tuberculosis (Gingearas in 1998) HIV-1, (Cronin in 1996) [9], drug discovery (drug development) [22], Toxicology, Biotechnology, [27] environmental engineering, biological sciences [34], discovery of genes [35] and Tumor classification, cancer classification, [36] simulated images may be useful for testing and evaluating image analysis tools. [37]

1.10. Properties of Microarray Image:
1. All the subgrids are of the same size.
2. The spacing between subgrids is regular.
3. The location of the spots is centered on the intersections of the lines of the subgrid.
4. The size and shape of the spots are perfectly circular and it is the same for all the spots.
5. The location of the grids is fixed in images for a given type of slides.
6. No dust or contamination is on the slide.
7. There is minimal and uniform background intensity across the image. [4, 23, 25, and 27]

1.11. The Current Status of DNA Microarrays:
A new and advanced intelligent image processing techniques for both eliminates the noise sources inherent in the DNA microarray process and also the development of tailor-
made image processing methodologies for speeding up the real-time diagnosis and implementation procedures of the next generation of system-on-a-chip devices. In this special issue, some of these challenges in enhanced imaging processing techniques and the new methodologies addressed above. O’Neill et al. discuss an image reconstruction technique to eliminate the noise sources in microarray images. The paper proposes a processing technique that can take a mask of the artifact and then recreate the image of the noise. This technique can reduce the noise dramatically. X. H. Wang et al. discusses an image enhancing technique by denoising using stationary wavelet transform. This technique can overall eliminate the random noise to improve the quality of gene expression data [15]. DNA microarray technology that allows simultaneous assay of thousands of genes [29, 34] in a single experiment has steadily advanced to become a mainstream method used in research, and has reached a stage that envisions its use in medical applications and personalized medicine. Many different strategies have been developed for manufacturing DNA microarrays. Shi Leining et al, (in 2009) has suggests the manufacturing characteristics of seven microarray platforms that were used in a recently completed large study by the MicroArray Quality Control (MAQC) consortium, which evaluated the concordance of results across these platforms. The platforms can be grouped into three categories: 1) in situ synthesis of oligonucleotide probes on microarrays (Affymetrix GeneChip® arrays based on photolithography synthesis and Agilent's arrays based on inkjet synthesis); 2) spotting of presynthesized oligonucleotide probes on microarrays (GE Healthcare's CodeLink system, Applied Biosystems' Genome Survey Microarrays, and the custom microarrays printed with Operon's oligonucleotide set); and 3) deposition of presynthesized oligonucleotide probes on bead-based microarrays (Illumina's BeadChip microarrays). [38]

1.12. Main Stages of Microarray Image Analysis:

Microarray analysis involves three basic stages namely experimental design, image processing, and gene quantitation. The spot analysis can be logically divided in the following three steps: image segmentation, spot recognition, and spot feature extraction. [1, 24, 31, 38]
1.12.1. **Microarray Image Segmentation:** The process of identifying the spots and separating the foreground from the background is known as microarray image segmentation. [6, 25, 26, 30, 39]

**Segmentation:** Segmentation subdivides an image into its constituent regions or objects [6]. The segmentation of an image can generally be defined as the process of partitioning the image into different regions, each having certain properties. 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 a spot mask, which consists of the set of foreground pixels for a given spot. [28]

1.12.2. **Spot Recognition:** the spot recognition algorithm, working on the segmented image, identifies the spots present in the microarray image. In this process can be categorized by two steps: Spot preliminary identification: the pixels in the foreground set are partitioned, including in the same subset all the pixels that belong to the same spot. Spot identification refining: the features and of the spots [4, 26] are used to check their consistency. Also, if this is the case, lost spots are recovered from the background set.

1.12.3. **Spot Feature Extraction:** To extract for each spot a set of features. The following features are considered to microarray: spot foreground area: the number of pixels belonging to the corresponding blob of the image, spot background area: the number of background pixels lying in the spot chess box. Spot intensity: the average intensity of the correspondent blob. Spot background intensity: the average intensity of the background pixels in the spot chess boxes it to reduce noise, only a proper subset of these pixels is considered [28]. Spot circularity: a measure of the circularity [26] of the correspondent in the image. Spot position error: the difference between the position of the correspondent blob in the image and the position of the spot chess box center. These are stored on 3-D structure that is appealing to build an efficient database, for classification and further processing purposes. [6]

1.13. **Objective:** Several researchers have done in the past but all the methods or softwares are not fully automated, it requires human interaction & preprocessing time. At the minimum, this softwares requires the user to specify geometry of the array, such as
shape of spot, size of spot, number of grids, number of rows and columns [23, 26], etc. i.e. SPOT, IMAGENE, DAPPLE & SRG [32]. Microarray images contain large number of spots need to be analyzed in a very fast automated way and this analysis is very important in discovering the cause of diseases. To develop segmentation technique for microarray images, to cater the need of microarray imaging to solve the problem specified. This technique may also improve the most of the mention drawbacks and to develop the new algorithm for microarray images. These results can easily be understood and used to make interpretation Bioinformatics/Biotechnologist. 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.

1.14. Significance of the Thesis:
The application of microarray image segmentation is one of the recent areas in medical diagnosis. The application of microarray image segmentation techniques has also used in other research community such as agriculture, medicine, biomedical engg., cancer detection, etc. Results have recently been reported for image segmentation using different types of techniques such as, fixed circle segmentation, adaptive circle segmentation, adaptive shape segmentation, & histogram segmentation. This thesis presents different methods of comparison.

Motivation: In the selection of microarray image segmentation topic for research it consists of review of previous work done in the microarray image segmentation field. Lot of research work has been done in this area but still the problems are exists related to fixed circle segmentation, watershed, SRG, etc. In fixed circle segmentation, it could not segment the image properly because it estimates as a same area for all spots. In this proposed work we have developed various intelligent segmentation techniques using novel four shapes of morphology. Each two shapes of combination are in one system by using mathematical morphology. A few of challenges are mentioned in earlier section.

The major contributions of this thesis are listed below:
The new technology for recognize microarray spots for medical imaging purposed is providing as regards as reliable but efficient recognition is depending on the quality of input image. Recognition of microarray image becomes complex computer problem
while dealing with noisy and low quality images. The proposed morphology based methods, is used for microarray image enhancement, segmentation, & spot recognition. Also morphological operations, opening, closing and tophat, bottom hat transforms used. We have applied proposed method on Stanford Microarray (20 images) Database, MAGIC tool websites (10 images) Database, and other different websites microarray (80 images) databases and experimental results shown that, specific width & periodic line shaped method is more good & effective than other existing techniques. Finally we have compared results by using the image quality measures. Most of microarray images are different circle like complex spots, & two spots are merged. Most of segmentation techniques can improve the spot quality but while dealing with boundary opening and closing. In this research work we have used the mathematical morphology to remove the noise information and merging a two spot information is separated through tophat and bottom-hat transform and measure the spots before and after by using one of the quality measure i.e. object count is to count number of objects or spots in whole microarray image, this is the novelty. The Advantage of this, the counting process is automatic or very fast.

1.15. Outline of the Thesis:
This thesis is divided into seven chapters: Chapter1 describes introduction of microarray, some history, types, applications, properties, sources and development of microarray technology. Chapter2 provides the literature review of various microarray segmentation methodologies and their principles. Chapter3 provides general segmentation methodologies and their comparisons. Chapter4 provides general morphological methodologies, which is useful to microarray images. Chapter5 provides the evaluation of microarray image segmentation using morphology-based techniques, which is useful for microarray image segmentations & automatic spots counting. It has gives about how morphology is superior approach than other microarray analysis tools i.e. fixed circle segmentation, watershed segmentation, etc. Chapter6 has given the various performance evaluation methods and their results after segmentation of image data; it is often necessary to evaluate the quality of the microarray image. Chapter7 provides conclusion and future direction.
1.16. Summary: In this chapter we have introduced about the microarray and their types, applications, properties, and steps of construction of DNA microarray, quality control, requirements of microarray, etc. Another major issue was creation of microarray image due to the problems, which occur at the time of acquiring and storing and define the microarray. This problems were efficiently handled by using different image analysis techniques such as image enhancement, segmentation, etc which is covered in the next chapter “Various Methodologies of microarray image segmentation.”

References:
[13] www.jco.org/cgi/glossarylookup
[16] www.nature.com/nrm/journal/v6/n5/glossary/nrm1645_glossary.html
[17] www.solexa.co.uk/Glossary/g.htm
[18] en.wikipedia.org/wiki/Protein_microarray
[19] www.mathwork.com
