Bioinformatics Trends
A Journal of Bioinformatics and Its Applications

Genomics Signal Processing (GSP)

Abhishek Tiwari*[$], Vipin Wadhwa*
*Department of Biotechnology, Vellore Institute of Technology, Vellore, India
[$]Corresponding author (Email:abhishek_twr@yahoo.com)

Received - 04 May 2006 Published Online - 27 Sep 2006.

Abstract:
Application of Digital Signal/Image Processing techniques (*DSP & DIP*) to solve Genomics problems initiated the new field Genomics Signal Processing (*GSP*) which concentrates to encode the Genomics signals based on DSP/DIP framework. In this Genomics era, high throughput DNA sequencing and the use of DNA microarray to simultaneously conduct huge number of experiments has lead to many signal/image processing problems. There is an emergent need to develop signal/image-processing techniques to examine data and determine relationship between genes. In this paper, we will focus on application of DSP & DIP in Biomolecular Sequence Analysis, Genetic Network Modeling and DNA Microarray Image Analysis.

Key Words:
GSP (Genomics Signal Processing), DNA Microarray, DSP (Digital Signal Processing), DIP (Digital Image Processing), cDNA, Oligonucleotid, DFT (Discrete Fourier Transform).

Online Access:
http://www.bii.in/journal/BIIJOURNAL/LoginArticle1.aspx?ids=PDF/190

http://www.bii.in/journal/BIIJOURNAL/FinalPaperAbstract.aspx?ArticleID=190

The Genome:

As we know that genes are physically embodied within complex DNA macromolecules that lie within structures called chromosomes that are present in a living cell. Discovery of the structure of DNA by Watson and Crick in 1953 showed that a DNA molecule is a double helix consisting of two strands. Each helix is a chain of bases, chemical units of four types: thymine (T), cytosine (C), adenine (A), and guanine (G). Each base on one strand is joined by hydrogen bonds sto a complementary base on the other strand, where A is complementary to T, and C is complementary to G. Thus the two strands contain the same information. Certain segments within these chromosomal DNA molecules contain genes, which are the carriers of the genetic information and spell the names of the proteins. Thus the genetic information can be thought of as being encoded digitally, as strings over the four-letter alphabet {A, C, T, G}, much as information is encoded digitally in computers as strings of zeros and ones. In humans there are 23 pairs of chromosomes. All but two of these (the sex chromosomes) occur in pairs of "homologous" chromosomes. Two homologous chromosomes contain the same genes, but a gene may have several alternate forms called alleles, and the alleles of a gene on the two chromosomes may be different. The total content of the DNA molecules within the chromosomes is called the genome of an organism. Within an organism, each cell contains a copy of the genome. The human genome contains about 3 billion base pairs and about 35,000 genes. Genome size and no of chromosomes as well as Genomics complexity vary with organism to organism.

Central Dogma- From Genes to Proteins:

The Central dogma of molecular biology is that DNA codes for RNA and RNA codes for proteins. Thus the production of a protein is a two-stage process, with RNA playing a key role in both stages. An RNA molecule is a single-stranded chain of chemical bases of four types: A, U, C, and G. In the first stage, called transcription, a gene within the chromosomal DNA is copied base by base into RNA according to the correspondence A→U, C→G, T→A, G→C. The resulting RNA transcript of the gene is then transported within the cell to a molecular machine called the ribosome that has the task of translating the RNA into a protein. Translation takes place according to the genetic code, which maps successive triplets of RNA bases to amino acids. With minor exceptions, the 64 possible triplets of 4 bases (4^3) map to 20 amino acids for all organisms.

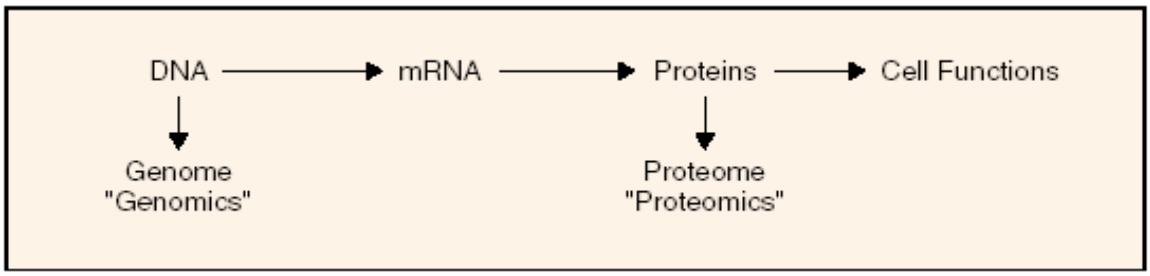


Figure1. Central Dogma of Molecular Biology (From Genes to Proteins)

Regulation of Gene Expression:

Gene expression can be viewed as a complex network of interactions involving genes, proteins, and RNA, as well as other factors such as temperature and the presence or absence of nutrients and drugs within the cell. It is clear that the expression of a gene within a cell (as measured by the abundance and level of activity of the proteins it produces) is ultimately regulated by the machinery of the cell. The transcription of a gene is typically regulated by proteins called transcription factors that bind to the DNA near the gene and enhance or inhibit the copying of the gene into RNA. Similarly, translation can be regulated by proteins that bind to the ribosome. Certain post-translational processes, such as the chemical modification of the protein or the transport of protein to a particular compartment in the cell, can also be regulated so as to affect the activity of the protein.

DSP/DIP Techniques in DNA Sequence Analysis

DSP Techniques in Sequence Comparisons and Classification

Wavelet analysis provides a useful DSP means for the visual description of inherent structure underlying DNA sequences. In, wavelet analysis is used to extract characteristic bands from protein sequences. In this research, the sequence-scale analysis with wavelet gave a multiresolution similarity comparison between protein sequences. This "similarity" expanded the traditional sequence similarity concept, which took into

account only the local pair-wise amino acid and disregarded the information contained in coarser spatial resolution. Also, this wavelet based method did not require the complex sequence alignment processing for sequences. Therefore, proteins with different sequence lengths could be compared easily. Other than sequence comparison, sequence classification is also a major problem in DNA signal analysis. The wavelet packet (WP) technique is used in for DNA sequence classification, i.e., to classify exons (a segment of DNA that is transcribed to RNA and specifies a portion of a protein) and introns (noncoding subregions in genes). After obtaining the energy distribution from WP coefficients, the energy map was used as a criterion for sequence classification. Digital signal processing (DSP) techniques offer more efficient ways to identify regions of the DNA exhibiting periodic behavior. In some cases, digital filters are employed to extract the T-3 component (the protein-coding regions of DNA demonstrate a period-3 (or T-3) performance due to codon structure. A codon is a sequence of three adjacent nucleotides constituting the genetic code that determines the insertion of a specific amino acid in a polypeptide chain during protein synthesis or the signal to stop/start protein synthesis.). In addition, digital filters are used to eliminate the background 1/f noise exhibited by nearly all DNA sequences. We will describe few interesting examples in details like *DNA Spectrograms* and *Color Maps*.

DNA Spectrograms:

Spectrograms are powerful visual tools for biomolecular sequence analysis. It is well known that the appearance of spectrograms provides significant information about signals, to the extent that trained observers can figure out the words uttered in voice signals by simple visual inspection of their spectrograms. An important advantage of DSP-based tools is their flexibility. Spectrograms can be defined in many ways. For example, depending on the particular features that must be emphasized, we may wish to define spectrograms using certain values of parameters. Once a visual pattern appears to exist, we have the opportunity to interactively modify the values of these parameters in ways that will enhance the appearance of these patterns, thus clarifying their significance. It is hoped that visual inspection of spectrograms will establish links between particular visual features (like areas with peculiar texture or color) and certain yet undiscovered motifs of biological sequences. Figure 2.a shows a spectrogram using DFTs of length 60 of a DNA stretch of 4,000 nucleotides from chromosome III of *C. elegans* (GenBank Accession number NC000967). Figure 2.b shows the texture of a spectrogram coming from a sample of totally random DNA, i.e., in which each type of nucleotide appears with probability 0.25 and independent of the other nucleotides.

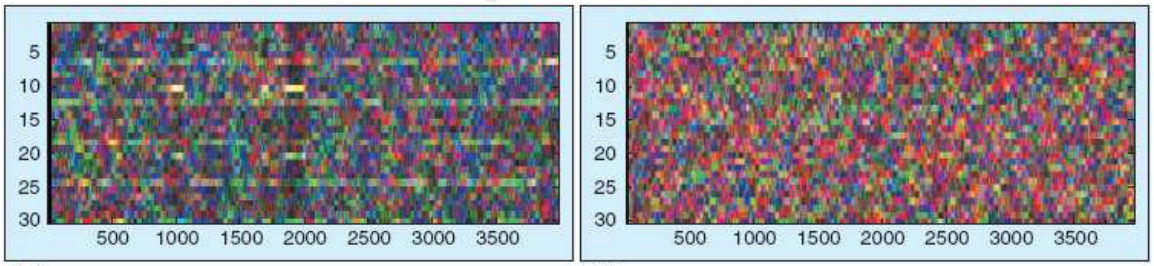


Figure2. (a)Color spectrogram of a DNA stretch (b) Color spectrogram of "totally random" DNA

Color Maps:

Color maps are mostly used for Reading Frame Identification and it is an excellent tool for sequence feature visualization. Because the number of primary colors (red, green, and blue) is the same as the number of possible forward coding reading frames, we can conveniently assign a color-coding scheme and based on that one can prepare color maps using Fourier transform. For example consider a DNA stretch from chromosome III of *S. cerevisiae* (GenBank accession number NC 001135).Note that there is no overlap with the collected statistics. The DNA stretch consists of 12,000 nucleotides starting from location 212041. It contains six genes (three forward coding and three reverse coding) at the locations shown in Table 1 relative to 212040.

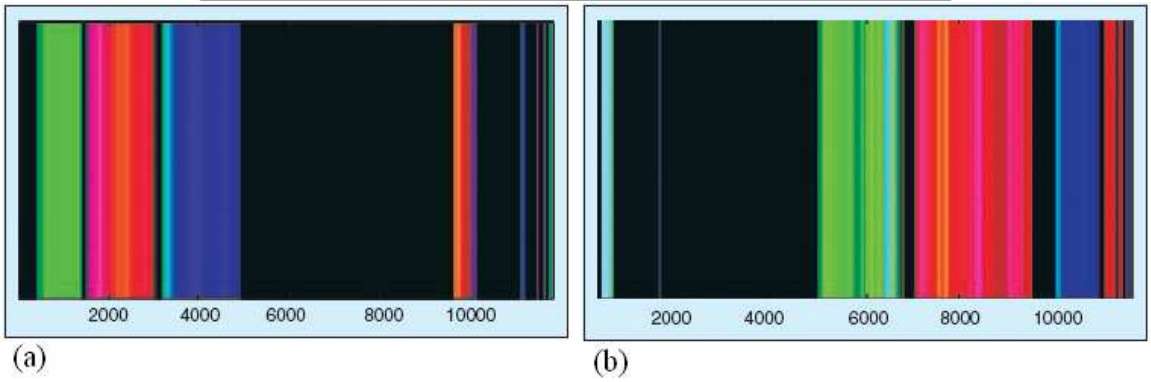
DSP Techniques in Exploring the Relationship between Sequence Structure and Function:

Before the wavelet method was applied in exploring the structural features within the sequences, conventional Fourier analysis had been used to elucidate the sequence structure information. However, the Fourier method could discover only "global" periodicities, and it could not extract hidden localized periodicities, which might provide hints about underlying construction rules [13]. A correlation function is constructed to compare each DNA base with its various neighbors [14]. After further Fourier or wavelet processing applied to the correlation function, their results readily showed some regular features in DNA

sequences. Wavelet is also used to search the DNA sequence construction rules. The salient spots in the final two-dimensional (2-D) analysis results revealed significant features in the DNA sequence. Their results demonstrated that while the noncoding sequences showed spectra similar to those from random sequences, coding sequences revealed specific periodicities of variable length and a common periodicity of three. Similarly, the method in quantifying symbolic sequence correlation to analyze DNA sequences is also used. The spectral density measurements of different base positions demonstrated the ubiquity of low frequency noise, long-range fractal correlation, and prominent short range periodicities. The results for several categories of DNA sequences also showed systematic changes in spectral exponent. This result provides a new technique for quantifying evolutionary changes in the information content of DNA. Wavelet transforms modulus maxima (WTMM) are used to analyze the fractal scaling properties in DNA sequences. The existence of long-range correlation is demonstrated in genes containing introns and noncoding regions, and this correlation is also quantified. The fluctuations in the DNA walk profiles were found to be homogeneous with Gaussian statistics. This result reveals useful information about the role of introns and noncoding intergenic regions in the nonequilibrium dynamic process that produced DNA sequences. Recently, it has been asserted that along with functional information, information about molecular evolution and relationships between organisms can also be derived. Since the evolution of genetic information and the principles through which nature produced the genetic information and genes are still not well understood, wavelet analysis for DNA sequences may provide some insights for these problems.

Table 1.

Locations and Reading Frames of Six Genes.		
Relative Location	Gene Length	Reading Frame
761 → 1429	669	2
1687 → 3135	1449	1
3387 → 4931	1545	3
5066 ← 6757	1692	$\widetilde{2}$
7147 ← 9918	2772	$\widetilde{1}$
10143 ← 10919	777	$\widetilde{3}$



(a) (b)

Figure3.Color map after partition for the genes shown in Table 1. (a) For forward coding (b) For reverse coding

DSP Techniques in Sequence Structure Prediction:

Accurate prediction and detection of DNA regions or their underlying structural patterns is a constant source of difficulty for researchers. Traditional structure detection methods were primarily based on the average of DNA base contents within a fixed window. Therefore, the location accuracy depended on the chosen window length. The multiresolution analysis feature of wavelet transform is excellent in resolving this problem, allowing efficient extraction of basic components at different scales. In [24], discrete wavelet transform (DWT) is applied to find pathogenicity islands and gene mutation events in genome data. The

DWT is used to smooth G + C profiles to locate characteristic patterns in genome sequences, and a wavelet scalogram was obtained to compare the sequence profile among genomes and to separate the different components within a profile. Further a change-point based wavelet thresholding method (WCP) is used to predict transmembrane helix (HTM) locations and the topology of HTM segments in the primary amino acid sequences. Wavelet was applied to decompose the propensity profile, which was generated according to the frequency of residues in HTM sequences. With the wavelet coefficients, a data-dependent threshold was then used to choose the coefficients representing abrupt changes in the profile. The reported prediction results were comparable to other methods, such as hidden Markov models. Moreover, the computational task is simple. Similarly, a continuous wavelet transform (CWT) is used in [26] to predict the α-helix content from the secondary structure of protein using the information from its hydrophobicity profile and the amino acid composition. Models are designed to identify gene locations in human DNA, including the Markov model, the hidden Markov model, and a wavelet-based hidden Markov tree (HMT). In HMT processing, an adaptive wavelet model is designed to match individual CpG islands in a DNA sequence to optimize the location identification. A model-based method is introduced (and combined with wavelet) to depict the replacement rate variation in genes and proteins, in which the profile of relative replacement rates along the length of a sequence was defined as a function of the site number. Besides better performance in fitting the data, the model also provided an additional useful method for determining regions in genes and proteins that evolved significantly faster than the average sequence.

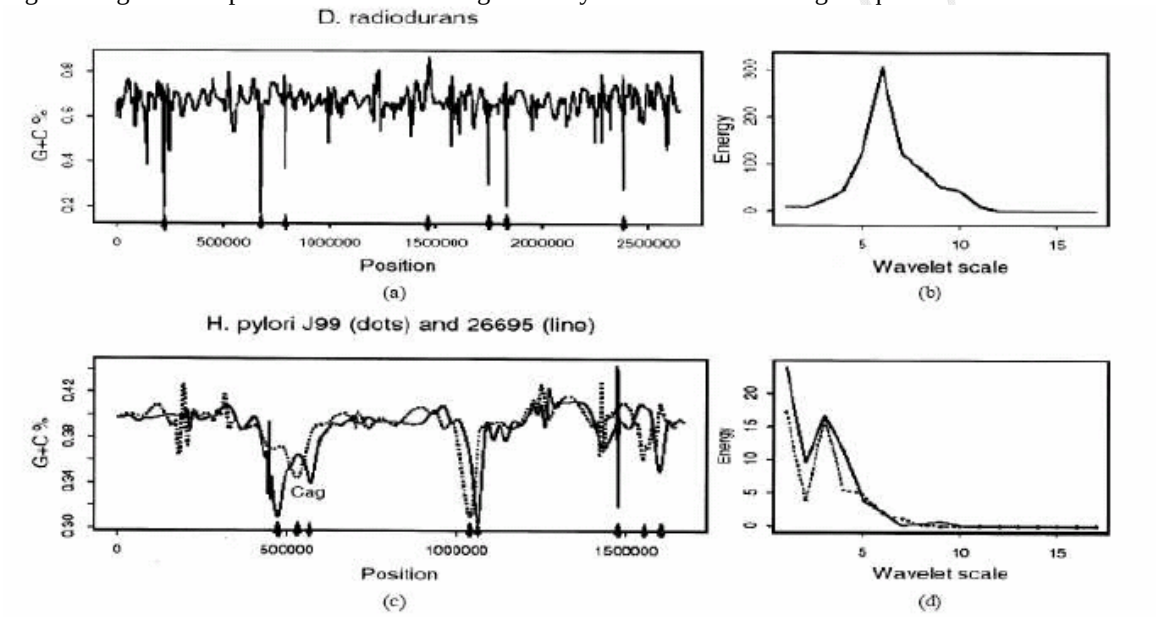


Figure4. (a) The G + C profiles of D. radiodurans chromosome I, (b) with its wavelet scalograms,(c) two H. pylori sequences, and (d) their relative scalograms.

DSP/DIP Techniques in Genetic Network Modeling

DSP Techniques in Modeling Genetic Regulatory Networks

The mathematical and computational modeling of genetic regulatory networks promises to uncover the fundamental principles governing biological systems. It also paves the way toward the development of systemic approaches for effective therapeutic intervention in disease. In [20], a Boolean formalism is presented as a building block to model complex, large-scale, and dynamical networks of genetic interactions. The role of Boolean networks is to understand cell differentiation and cellular functional states. These Boolean networks can also be related to nonlinear digital-filter design. In addition, the inference of Boolean networks from real gene expression data can be modeled using computational learning theory combined with nonlinear SP. To handle the uncertainty in Boolean networks, a Markov chain model is applied to analyze the probabilistic framework. The potential effect of individual genes on the global dynamical network behavior is also considered using stochastic perturbation analysis. This also leads to target identification for therapeutic intervention via the development of several computational tools based on first-passage times in Markov chains. In [21], DNA transition is modeled utilizing a Markov process, specifically Markov chains. If the Markov process does not capture the DNA transition process, then CpG "islands" can be used, where "p" simply indicates that "C" and "G" are connected by a

phosphodiester bond. (The CpG islands are aggregates of rich C-G pairs that are bunched in several hundred to several thousand nucleotides.) The CpG method can construct a rough probability distribution model. In [22], a different approach is taken by constructing a finite-state Markov chain whose transitions depend on state dependent multivariate conditional probabilities between gene-expression levels, based on microarray data. Mathematical modeling tools that allow estimation of steady-state behavior in biological systems would be useful for examining two ubiquitous forms of biological system behavior. The first is homeostasis, the ability of cells to maintain their ongoing processes within the narrow ranges compatible with survival, and the second is a switch-like functionality that allows cells to rapidly transition within limited process segments between metastable states.

DSP/DIP Techniques in DNA Microarray Analysis

DNA Array Technology:

The principle of a microarray experiment, as opposed to classical analysis, is that mRNA from a given cell line or tissue is used to generate a labeled sample, sometimes termed the target, which is hybridized in parallel with a large number of DNA sequences, and immobilized on a solid surface in an ordered array. Tens of thousands of transcript species can be detected and quantified simultaneously. During recent years, DNA microarray technology has been advancing rapidly. The development of more powerful robots for arraying, new surface technology for glass slides, and new labeling protocols and dyes, together with increasing genome-sequence information for different organisms, including humans, will enable us to extend the quality and complexity of microarray experiments. Although academic groups and commercial suppliers have developed many different microarray systems, the most commonly used systems today can be divided into two groups, according to the arrayed material:

▲ 1) complementary DNA (cDNA)

▲ 2) oligonucleotide microarrays

as shown in Figure 4.

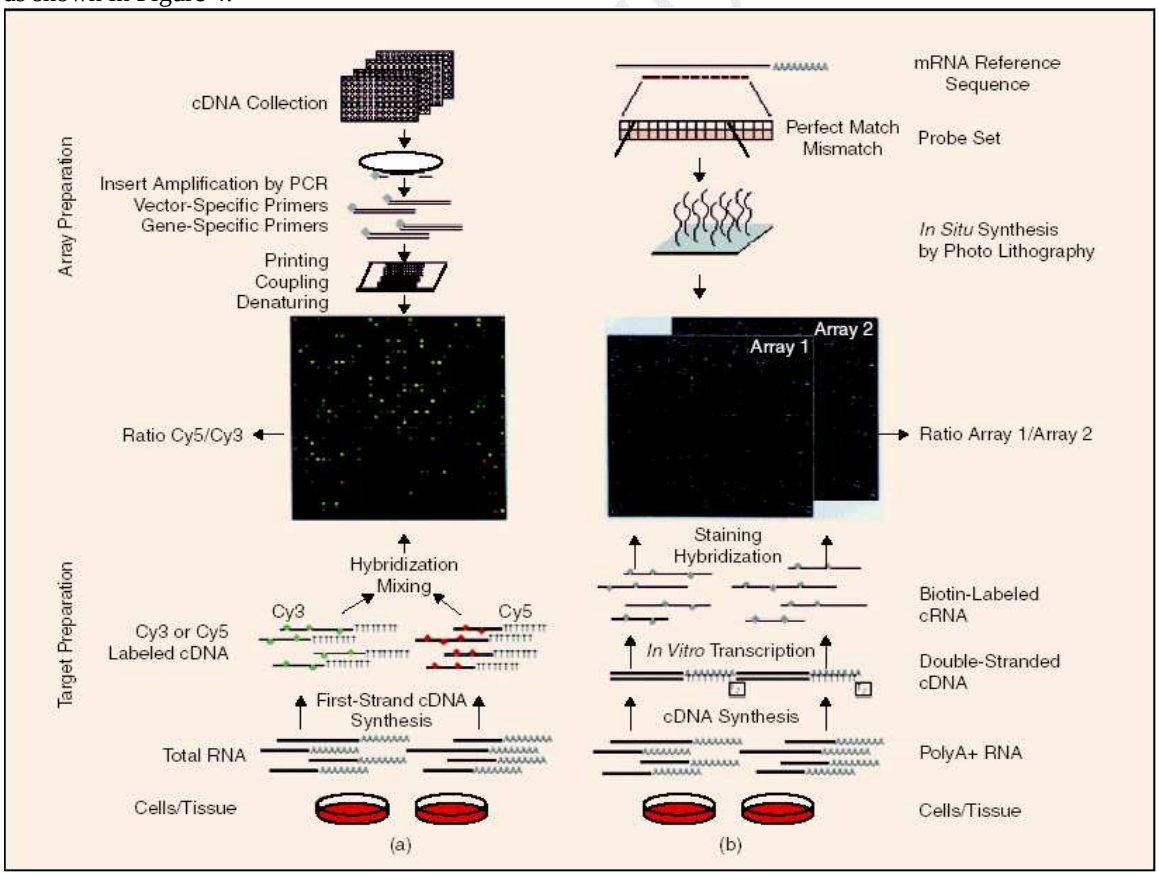


Figure5 Schematic overview of probe array and target preparation for spotted cDNA microarrays and high-density oligonucleotide microarrays :(a) cDNA microarrays and (b) high-density oligonucleotide.

The fabrication of the DNA microarray can be done with the spotting method or on-chip synthesis. The target genes are normally required to be amplified and labeled with fluorescence. Multicolored

fluorescent techniques are often used in comparative hybridization detection (Figure 4 gives an example of a typical fluorescence image). Fluorescent detection is a conventional method for the detection of hybridization results. The most important characteristic to draw from a fluorescence image is the assessment of the hybridization degree (e.g., whether or not, and at which quantitative level, they hybridize with a given nucleic acid sequence), which is proportional to the intensity of each color spot. Ratios of spot intensities in both dyes in comparative hybridization can then be used to compute the differential expression of the gene or the expressed sequence tags between the two samples. The analysis of data from fluorescence images and the gene expression database is necessary in the treatment of such large amounts of information obtained from a DNA microarray. Since the fluorescent image obtained from microarray hybridization contains nearly all the gene expression levels for detected DNA or RNA sequences, the performance of image-processing methods used for fluorescent images has a potential impact on subsequent analysis such as clustering or the identification of differentially expressed genes. Many software tools have been developed for microarray image processing. The basic goal is to transform an image of spots of varying intensities into a matrix, called a gene expression matrix, with a measure of the intensity (or, for multicolored fluorescent images, the ratio of intensities) for each spot. Although it seems to be a relatively straightforward goal, the variation, noise, and large number of pixels on a microarray image make it a complex process. The major issues these software tools must address are how to reduce noise to improve the accuracy and how to realize automation for the processing procedure. Implementing real-time processing for a microarray image is becoming increasingly critical because of the increasing number of microarrays that must be analyzed.

DNA Microarray Image Processing:

DNA microarray image processing is one of the information extraction problems occurring in molecular biology and bioinformatics. Molecular biologists and bioinformaticians are using microarray technology for identifying a gene in a biological sequence and predicting the function of the identified gene within a larger system (although there is still an active debate about how to define the bioinformatics discipline). Microarray technology is based on creating DNA microarray that are typically composed of thousands of DNA sequences, called probes, fixed to a glass or silicon substrate. Usually, samples from two sources are labeled with different fluorescent molecules (emitting at red and green wavelengths) and hybridized together on the same array. The array is then scanned by activation with lasers at the appropriate wavelength to excite each dye. The relative fluorescence between each dye on each spot is then recorded and a composite image may be produced. The relative intensities of each channel represent the relative abundance of the RNA or DNA product in each of the two samples. 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 analyses are usually part of a microarray data processing workflow that includes, grid alignment, spot segmentation, quality assurance, data quantification and normalization, identification of differentially expressed genes and their significance testing, and data mining. An example of microarray data processing workflow is illustrated in Figure 3. The subset of image processing steps is enclosed with a dashed line in Figure 5.

The major tasks of DNA microarray image processing are to identify the array format including the array layout, spot size and shape, spot intensities, distances between spots, and background fluorescence, and to extract spot descriptors, as well as the uncertainty of the descriptors that represent the underlying microarray experiment. Biological conclusions are then drawn based on the results from data mining and statistical analysis of all extracted descriptors. The reliability of spot descriptors depends on many different factors. For example, one could list basic factors, such as microarray technology components, and protocols for array production, sample labeling, hybridization and image acquisition. Printing parameters, such as pin size and shape, printing speed, temperature and humidity, printing buffers and deposition surface, will all affect the size and morphology of the individual spots. The type of glass and coating, blocking agents, hybridization and wash buffers will affect background fluorescence. Any DNA array image analysis programs must be easily adapted to these varying parameters. In order to choose an appropriate image processing approach and automate DNA microarray image analysis, one has to understand variations of input microarray images in terms of (1) the image content including foreground and background morphology (e.g., grid layout, spot location, shape and size), and intensity information (e.g., spot descriptors derived from foreground and background intensities), (2) the computer characteristics of input digital images (e.g., number of channels, number of bytes per pixel, file format).

Ideal Microarray Image:

First, let us define an “ideal” cDNA microarray image in terms of its image content. The image content would be characterized by deterministic grid geometry, known background intensity with zero uncertainty,

pre-defined spot shape (morphology), and constant spot intensity that (a) is different from the background, (b) is directly proportional to the biological phenomenon (up- or -down regulation), and (c) has zero uncertainty for all spots. While finding such an ideal cDNA image is probably a pure utopia, it is a good starting point for understanding image variations and possibly simulating them.

Another aspect of an "ideal" cDNA microarray image can be expressed in terms of statistical confidence. If one could not possibly acquire an ideal microarray image, then a high statistical confidence in microarray measurements would be obtained with a very large number of pixels per spot (theoretically it would reach infinity). However, the cost of experiments, the limitations of laser scanners in terms of image resolution, storage of extremely high resolution images and other specimen preparation issues are the real world constraints that have to be taken into account.

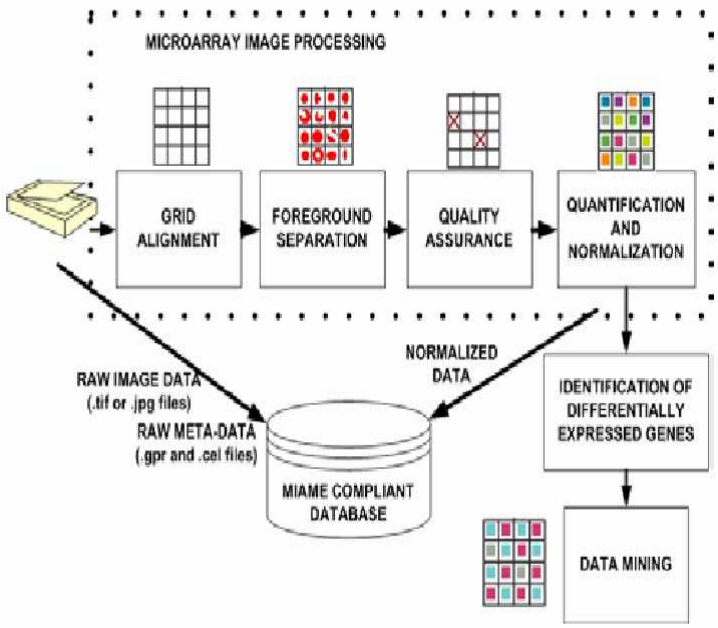


Figure 6.Microarray data processing workflow. The diagram stresses the requirement to archive both raw and processed data.

Image-Processing Methods:

The automation of this process is complicated by the variation of size and position of spots, the relative placement of the adjacent grid, and the overall position of the array image. Many existing software packages for microarray image analysis require some degree of user intervention in this step. Allowing user intervention may increase there liability and ensure accuracy of the whole quantization process; however, this may make the process unacceptably slow. Semiautomated packages for grid generation do exist [36], but all require some limited degree of user intervention, either in locating spots, setting thresholds, or making unrealistic assumptions about the regularity of data (e.g., assuming only circular spots). A novel operator independent and reproducible method is proposed for the automated analysis of gene microarray images [37] and the algorithm is based on the regular structure of the images and uses Fourier methods to extract this periodic structure as an initial approximation of spot locations. This initial addressing is then refined by an iterative method that produces accurate locations of all spots on the array. In addition, a classifier by spot quality is employed in some software packages to automate detection of spot-finding errors and spots of poor quality for further improvement of effectiveness. Many current implementations require the user to specify explicit thresholds of various attributes, such as brightness, that separate acceptable from unacceptable spots. Choosing good thresholds manually for multiple attributes through an extended process of trial and error is time consuming and may not achieve the desired result. In [38], a novel example- based classifier is implemented to decide whether candidate spots have been found correctly and are usable for further analysis. Machine learning techniques are introduced in the classifier to provide a convenient and powerful way for an investigator to specify complex concepts of spots without explicitly determining classification thresholds for image attribute values. According to the test, the automated classification matched their manual classification for more than 95% of candidate spots [38].

Cellular Neural Network (CNN) Methods:

Currently, with the software package mentioned above, the time spent on quantitatively processing a typical microarray fluorescent image is in the order of minutes. Some researchers think that such throughput is acceptable for laboratory purposes. However, it seems rather slow for the very high-output user, such as a pharmaceutical company, that might produce tens of thousands of arrays per year. Furthermore, DNA microarrays are predicted to be a normal diagnostic tool in clinics in the future, much like today's blood test. A low-efficiency analysis approach is sure to impede the progress of microarray adoption in this area. Enhancing the processing speed, or ideally realizing real-time processing, is desirable. Microarray analysis by a traditional computer, which sequentially processes images pixel by pixel, not only is very time consuming, but also destroys the parallel nature of microarray techniques itself. Fortunately, the CNN was recently introduced into microarray analysis, which promises to provide a breakthrough in DNA microarray parallel processing and obtain the gene expression profile in real time. A CNN is an analog dynamic processor array that reflects this property: the processing elements interact directly within a finite local neighborhood. Due to its architecture, a two-dimensional CNN array is widely used to solve image processing and pattern recognition problems; moreover, the parallelism characteristic of this structure allows one to perform the most computationally expensive image analysis tasks three orders of magnitude faster than a classical CPU-based computer. This approach, thanks to the supercomputing capabilities of the CNN architecture, makes the whole DNA chip methodology fully parallel and also makes the processing phase, until now very time consuming, a real-time step. Overall CNN scheme is described in Figure 7 and 8.

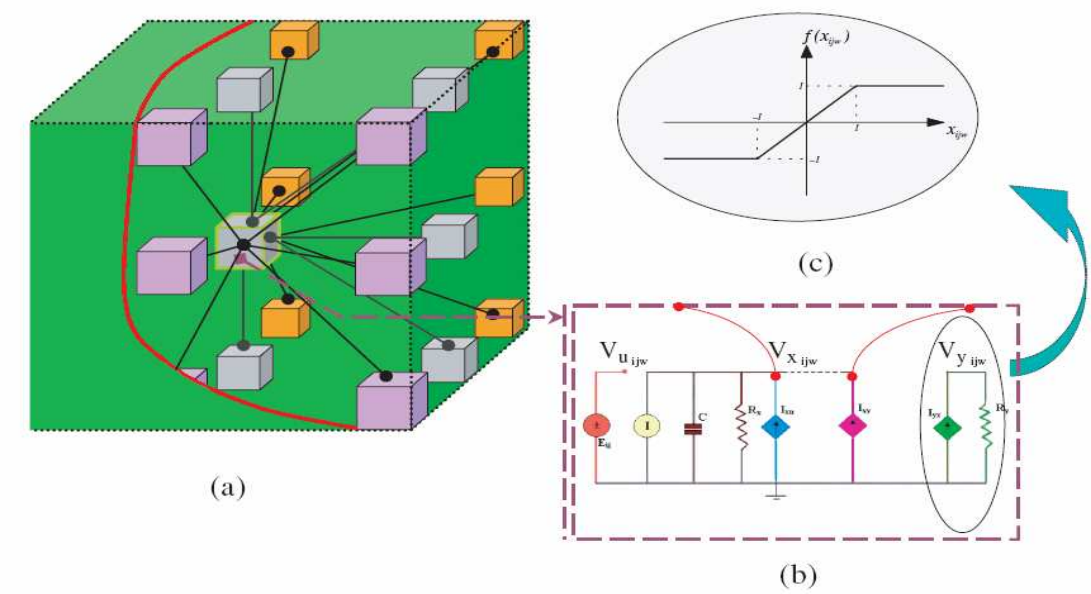


Figure 7.The CNN scheme: (a) The spatial grid showing local connections (b) The circuit of each cell(c) The classical output nonlinear function for each cell

Biomodel-Based Technique:

With the exciting results achieved from Genomics signal analysis, there also exist several open problems in this field. For instance, the prediction accuracy of a characteristic Genomics location is only around 70–80%. The fixed-shape segmentation for microarray images does not reflect the reality of the system, and an adaptive technique is needed to more reliably detect spot location and its size in microarray images. Therefore, more efficient methods are being applied to solve these problems, such as singular value decomposition and template technique, etc. Biomodel based approaches will play an important role in the future of GSP, since modeling methods come from the actual physiological and Genomics processes, and they simulate the underlying dynamic mechanism for the signal generation. Also, these models have rich mathematical structures and form the theoretical basis for the applications. In [42], a stochastic model is used to interpret DNA microarray images and to unravel their underlying physical process. The nonlinear modeling method is applied to analyze the DNA sequence. These model-based methods have provided new

viewpoints in Genomics information exploration and will also accelerate this process. Bionic wavelet transform (BWT) is another example of such a biomodel-based method for time-frequency analysis. BWT can realize adaptive signal-dependent 2-D resolution adjustments over the time-frequency plane rather than the fixed frequency-dependent resolution in the wavelet case. Further research shows that BWT also possesses many other features, such as a more concentrated signal presentation over the time-frequency plane, and a better robustness to noise, etc.

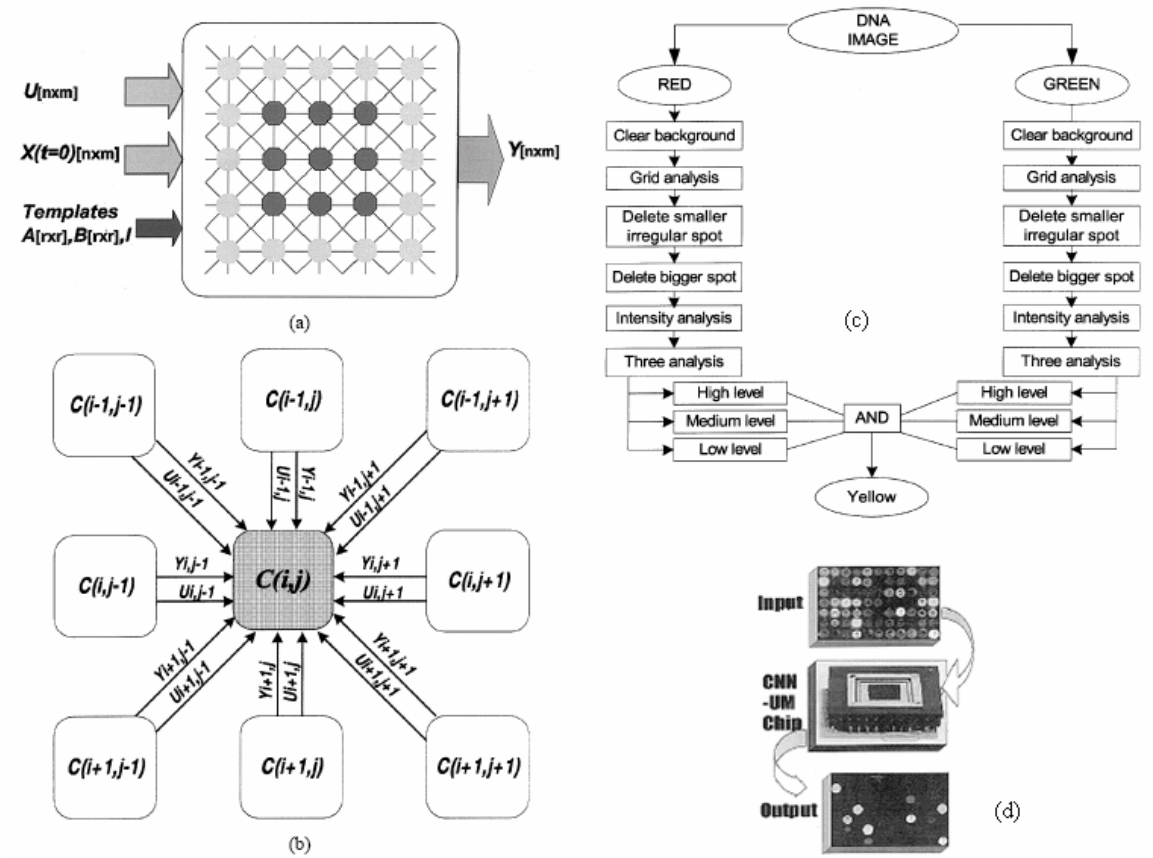


Figure 8. (a) Block scheme of the CNN architecture and operation principle (b) CNN local interaction between cells (c) Flowchart of the CNN algorithms for microarray analysis (d) CNN-UM chip for DNA microarray processing

Usage of Multidimensional DSP/DIP Technique:

Until now, two independent strategies have been developed for microarray data analysis. One treats the array-organized data as a one-dimensional (1-D) time-domain signal and uses classical 1-D SP methods to analyze these signals. The second one treats the overall gene expression spots as a contour plot or color mapped image and analyzes the data in a 2-D large-scale pattern. Statistical analysis on the color distribution of the microarray image is currently a major interest and tool in microarray data processing. Multidimensional (beyond 2-D) analysis will be another research trend in the Genomics research field. For instance, in protein engineering research, it is of great interest to reveal the complex 3-D structure of the Genomics sequence. Currently, most Genomics SP techniques are still static. The four-dimensional technique will help to detect the real-time dynamic Genomics structure change during future drug experiments. With the application of these multidimensional techniques, it will be possible in the future for us to reveal the underlying Genomics structure and function, and their relationship in dynamic situations.

Noise Reduction:

Noise reduction is a pervasive issue in DNA microarray analysis. If a DNA microarray image were an array of spots with a precise size and position located on a uniform low background, it would constitute a very simple problem in terms of image processing. It is the variation and noise in fluorescent images that complicate this problem. It is important for us to analyze the source of noise and to seek effective solutions.

The major source of variation and noise on microarray images originates from microarray fabrication machines, the treatment of glass slides, and fluorescence detectors. Despite the use of precise fabrication machines, spots vary significantly in size and position owing to variations in the amount of DNA on each spot and in the location where it is placed. Detector noise includes that from the amplification and digitization process, such as photon noise, electronic noise, laser light refection, and background fluorescence. In practice, the natural fluorescence of the glass and any nonspecifically bound DNA or dye molecules add a substantial noise floor to the image. This diffuse noise exhibits considerable variability in intensity both within and between small rectangles containing individual spots. Microarrays are also afflicted with discrete image artifacts such as highly fluorescent dust particles, unattached dye, salt deposits from vaporated solvents, and fibers or other airborne debris. Such artifacts appear in the vicinity of 10–15% of spots at random, even after a thorough cleaning of the slide, and can easily be brighter and sometimes larger than nearby useful spots [38]. Their heterogeneous brightness, shape, and size make them hard to detect and remove automatically, especially in the presence of spots that are themselves of variable size and brightness.

Conclusion

Many DSP and DIP techniques are effectively used and analyzed for the purpose of Genome Analysis and DNA Microarrays within last few years. More research work and efforts are required to use DSP technologies for solving biological problems. Some of the major areas of investigation and future challenges that can be tailored for future research directions in this emerging area are high throughput Genome Analysis especially in search of structural elements of genome and the development of new and advanced intelligent image processing techniques for both eliminating 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.

References

[1] R.M. Karp, “Mathematical challenges from Genomics and molecular biology,” Notices AMS, vol. 49, pp. 544–553, May 2002.
[2] Gen Bank. Available: http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html
NCBI, http://www.ncbi.nlm.nih.gov/
[3]SWISS-PROT. Available: http://www.ebi.ac.uk/swissprot/
[4] N.M. Luscombe, D. Greenbaum, and M. Gerstein,“What is bioinformatics? A proposed definition and overview of the field,” Method Inform. Med., pp. 346–358, Apr. 2001.
[5] The Protein Data Bank. Available: http://www.rcsb.org/pdb/
[6] FASTA Programs at the U. of Virginia. Available:http://fasta.bioch.virginia.edu/
[7] Available: http://pauling.mbu.iisc.ernet.in/~pali/tree.html
[8] NCBI Blast homepage. Available: http://ncbi.nih.gov/BLAST
[9] C.H. Trad, Q. Fang, and I. Cosic, “Protein sequence comparison based on the wavelet transform approach,” Protein Eng., vol. 15, pp.193–203, Mar. 2002.
[10] J. Zhao, X.W. Yang, J.P. Li, and Y.Y. Tang, “DNA sequences classification based on wavelet packet analysis,” in Proc. Wavelet Analysis and Its Applications, 2nd Int. Conf., WAA, 2001, pp.424–429.
[11] D. Anastassion, “Genomic signal processing,”IEEE Signal Processing Mag., pp. 8–20, July 2001.
[12] In silico Cloning. Available: http://www.hgmp.mrc.ac.uk/ESTBlast/Tutorial
[13] A.A. Tsonis, P. Kumar, J.B. Elsner, and P.A.Tsonis, “Wavelet analysis of DNA sequences,” Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdsip. Top., vol. 53, pp. 1828–1834, Feb.1996.
[14] G. Dodin, P. Vandergheynst, P. Levoir, C. Cordier, and L. Marcourt, “Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences,” J. Theor. Biol., vol. 206, pp. 323–326, Oct. 2000.
[15] R.F. Voss, “Evolution of long-range fractal correlations and 1/f noise in DNA base sequences,” Phys. Rev. Lett., vol. 68, pp. 3805–3908, June 1992.
[16] A. Arneodo, E. Bacry, P.V. Graves, and J.F. Muzy, “Characterzing long-range correlations in DNA sequences from wavelet analysis,” Phys. Rev. Lett., vol. 74, pp. 3293–3296, Apr. 1995.
[17] Swiss-Protein. Available: http://www.ebi.ac.uk/swissprot/index.html
[18] The Drosophila melanogaster ADH sequence. Available: http://www.flybase.org
[19] Center for automated learning and discoveries. Available:http://www.cs.cmu.edu/cald/research.html

[20] I. Shmulevich, E.R. Dougherty, and W. Zhang, "From Boolean to probabilistic Boolean networks as models of genetic regulatory networks," Proc.IEEE, vol. 90, pp. 1778–1792, Nov. 2002.
[21] J. White, J. Chen, J. Wang, and K.J. Ray Liu, "Modeling DNA transition using Markov random network," submitted for publication.
[22] S. Kim, H. Li, E.R. Dougherty, N. Cao, Y.D.Chen, M. Bittner, and E.B. Suh, "Can Markov chain models mimic biological regulation?" J.Biol. Syst., vol. 10, pp. 337–357, Nov. 2002.
[23] Critical Assessment of Structure Prediction. Available:http://www.ncbi.nlm.nih.gov/ Structure/RESEARCH/casp3/index.shtml
[24] P. Lio and M. Vannucci, "Finding pathogenicity islands and gene transfer events in genome data," Bioinformatics, vol. 16, pp. 932–940, Oct. 2000.
[25] P. Lio, "Wavelet change-point prediction of transmembrane proteins," Bioinformatics, vol. 16, pp. 376–382, Apr. 2000.
[26] L. Pattini, L. Riva, and S. Cerutti, "A wavelet based method to predict the alpha helix content in the secondary structure of globular proteins," in Proc. IEEE Special Top. Conf. Mol. Cell, Tissue Engineering, 2002, pp. 142–143.
[27] N. Dasgupta, S. Lin, and L. Carin, "Sequential modeling for identifying gene locations in human genome," Dept. Elec. Comput. Eng., Duke Univ., Durham, NC, Tech. Rep., Dec. 2001.
[28] P. Morozov, T. Sitnikova, G. Churchill, F.J. Ayala, and A. Rzhetsky, "A new method for replacement rate variation in molecular sequences: Application of the Fourier and wavelet models to Drosophila and mammalian proteins," Genetics, vol. 154, pp. 381–395, Jan. 2000.
[29] R.E. Green and S.E. Brenner, "Bootstrapping and normalization for enhanced evaluations of pairwise sequence comparison," Proc. IEEE, vol. 90, pp. 1834–1847, Dec. 2002.
[30] E. Pirogova, Q. Fang, M. Akay, and I. Cosic,"Investigation of the structural and functional relationships of oncogene proteins," Proc. IEEE, vol. 90, pp. 1859–1867, Dec. 2002.
[31] M. Peleg, I.S. Gabashvili, and R.B. Altman, "Linking genetic polymorphisms of tRNA to their functional sequelae," Proc. IEEE, vol. 90, pp. 1875–1886, Dec. 2002.
[32] J. Cheng and L.J. Kricka, Biochip Technology. New York: Hardwood Academic, 2001.
[33] A. Schulze and J. Downward. (2001, August).Navigating gene expression using microarrays—A technology review. Nature Cell Biol. [Online]. Vol 3. Available: http://cellbio.nature.com
[34] V.G. Cheung, M. Morley, F. Aguilar, A. Massimi, R. Kucherlapati, and G. Childs, "Making and reading microarrays," Nature Genetics Suppl., vol. 21, pp. 15–19, Jan. 1999.
[35] R.J. Lipshutz, S.P.A. Fodor, T.R. Gingeras, and D.J. Lockhart, "High density synthetic oligonucleotide arrays," Nature Genetics Suppl. vol. 21, pp. 20–24, Jan. 1999.
[36] Media Cybernetics, Inc. Array-pro analyzer. Available: http://www.mediacy.com/ arraypro.htm
[37] C. Bowman, R. Baumgartner, and S. Booth, "Automated analysis of gene microarray images," in Proc. IEEE Can. Conf. Elect. Comput. Eng. 2002, pp. 1140–1144.
[38] J. Buhler, T. Ideker, and D. Haynor, "Dapple: Improved techniques for finding spots on DNA microarray," Comput. Sci. Eng., Univ. Washington, Seattle, Tech. Rep. UWTR 2000-08-05, Aug. 2000.
[39] P. Arena, L. Fortuna, and L. Occhipinti, "A CNN algorithm for real time analysis of DNA microarrays," IEEE Trans. Circuits Syst. I, vol. 49, pp. 335–340, Mar. 2002.
[40] L. Fortuna, P. Arena, D. Balya, and A. Zarandy, "Cellular neural networks: A paradigm for nonlinear spatio- temporal processing," IEEE Circuits Syst. Mag., vol. 1, pp. 6–21, Apr. 2001.
[41] L.O. Chua and L. Yang, "Cellular neural networks: Theory," IEEE Trans. Circuits Syst. I, vol. 35, pp. 1257–1272, Oct. 1988.
[42] D. Seale and S.W. Davies, "Stochastic model of DNA microarray," in Proc. IEEE Special Top. Conf. Mol., Cell, Tissue Engineering, 2002, pp. 113–114.