

Analysis of Microarray Gene Expression Data

Tuan D. Pham^{*1,2}, Christine Wells^{3,4} and Denis I. Crane^{3,4}

¹Bioinformatics Applications Research Centre, ²School of Information Technology, James Cook University, Townsville, QLD 4811, Australia

³Cell Biology Group, Eskitis Institute for Cell and Molecular Therapies, ⁴School of Biomolecular and Biomedical Science, Griffith University, Nathan, QLD 4111, Australia

Abstract: Microarrays provide the biological research community with tremendously rich, sensitive and detailed information on gene expression profiles. Gene expression profiling and gene expression patterns have been found useful for solving a wide variety of important biological and biomedical problems, including the study of metabolic pathways, inference of the functions of unknown genes, diagnosis of diseased states, as well as facilitating the development of individualized drug treatments through pharmacogenomics. Given the significant impact of microarray gene expression data in biological and biomedical research, this breakthrough technology urgently needs the assistance of advanced computational methods for interpreting and utilizing the raw information. This paper reviews several main research directions and methods in the analysis of microarray gene expression data.

Keywords: Microarrays, gene expression, data analysis, image processing, signal processing, pattern classification, machine learning, computational techniques.

1. INTRODUCTION

The advent of genomics has witnessed an explosion of large datasets into the public domain; not least of these the predictions of tens of thousands of genes, hundreds of thousands of transcripts, and a combinatorial number of regulatory regions. Microarrays are a relatively new technology that allows interrogation of the breadth of this dataset, providing novel insights into gene expression and gene regulation. Microarray technology has been applied in diverse areas ranging from genetics and drug discovery to disciplines such as virology, microbiology, immunology, endocrinology, and neurobiology. Microarray-based methods are the most widely used technology for large-scale analysis of gene expression because they allow simultaneous study of mRNA abundance for thousands of genes in a single experiment [1].

Microarrays present unique opportunities to analyze gene expression and regulation in a global cellular context. Equally, the generation of large datasets present unique challenges in the acquisition, annotation, analysis and warehousing of that data. There are numerous platforms available for the interrogation of gene expression, and the expansion of gene expression data available in the public databases is testimony to the ready adoption of this technology. Yet questions still remain as to the reliability of information generated by microarray expression profiling, particularly when comparing information collated on different technology platforms.

Gene expression data analysis has undergone rapid development since the review carried out by Brazma and Vilo five years ago [2]. This review article builds on this and

other earlier reviews to report the recent developments in computational methods for analyzing and interpreting microarray data. We first address the importance of understanding the design, annotation and normalization of probes and probesets, which is the key to generating meaningful insights from microarray data. We follow this by a discussion of the various computational methods for the analysis of microarray data.

2. MICROARRAY PLATFORM OVERVIEW

A microarray is a series of probes ordered on a fixed surface such that the identity of each probe can be matched through its 'address' or position on the array. While almost any biological molecule can be arrayed, the probes of gene expression arrays consist of a string of nucleotides complementary to the sequence of the gene or gene product being investigated. A typical oligonucleotide probe is single stranded and between 25 and 70 bases long, whereas cDNA probes are double stranded and may be as long as a full-length gene product (on average 2kb). Genomes may be represented on a microarray by arraying BAC clones – large genomic fragments that may contain many genes – or probe sets targeting particular genomic features, such as coding regions (for expression profiling), spanning splice junctions (for the identification of alternatively spliced variants) or regulatory regions such as promoters.. This wide variety of potential microarray probe sets offers a versatile platform to analyze genomes from gene expression, chromatin regulation, to gene dosage and mutation analysis.

cDNA arrays offer an attractive probe set for incompletely sequenced genomes. Typically derived from Expressed Sequence Tags (ESTs), these probes are captured from gene products expressed in an organism of interest, so representing a set of likely 'hits' when probing for gene expression patterns. The disadvantages of a cDNA array include lack of annotations associated with probes – they may be annotated only as a position on a freezer plate where

*Address correspondence to this author at the Bioinformatics Applications Research Centre, School of Information Technology, James Cook University, Townsville, QLD 4811, Australia; Tel: +61-7-4781 6903; Fax: +61-7-4781 4029; E-mail: tuan.pham@jcu.edu.au

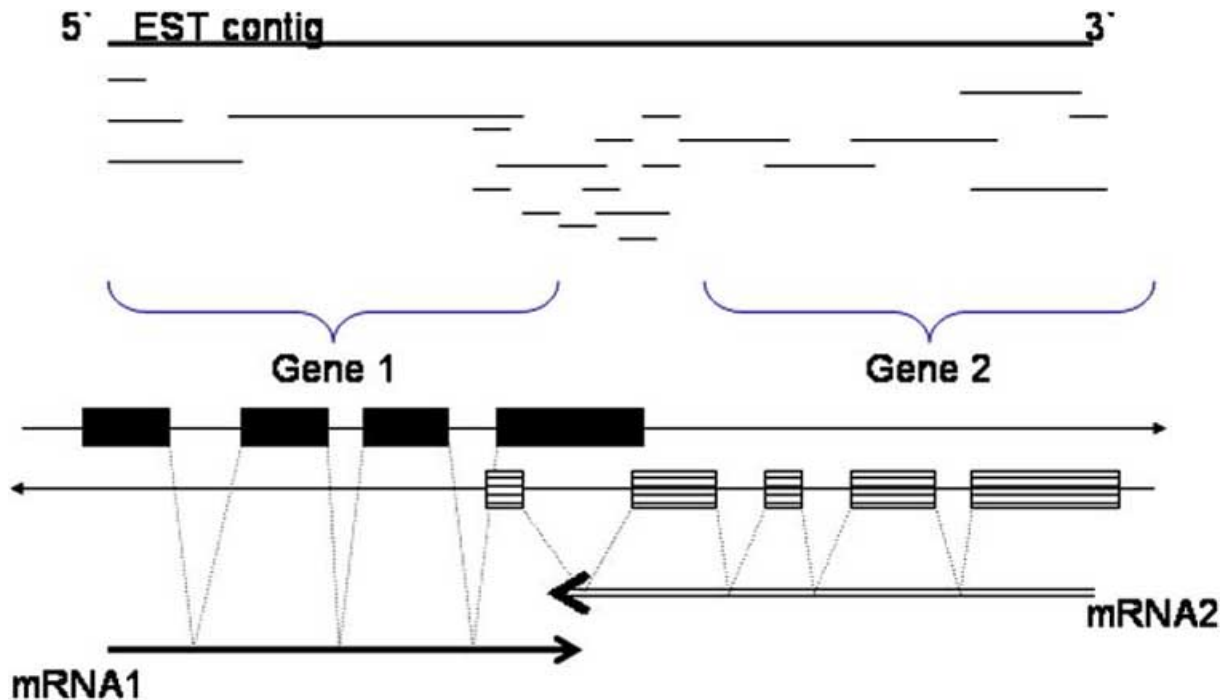


Fig. (1). Contig of expressed sequence tags.

the clones are stored – as well as redundancy of gene products represented on the array, cross-hybridization of sequences common to different probes and sub-optimal variation in hybridization efficiencies (determined by probe length and composition) across the probe set. Nevertheless, cDNA arrays offer a rapid and successful entry into microarray technology for those organisms not well represented in the public sequencing databases.

Oligonucleotide (oligo) arrays are the probe of choice for well-sequenced and annotated genomes. The reproducibility of data derived from a variety of oligo platforms is reasonably robust given the variation between laboratories [3,4], and this permits a critical comparison of data derived from different array platforms (discussed below). The design of each oligo is critical to the robustness of the array as a whole. Most oligos are designed to represent the most common transcripts expressed from a gene, and are biased to the 3' end of the gene. Probe design is constrained by meeting the most consistent hybridization parameters across the entire set – as the size and concentration of each oligo is predetermined, the most critical factors are GC content (in determining melting temperature), predicted secondary structures within the probe itself, as well as the uniqueness of the probe sequence, particularly at the 5' stretch [5-7].

Commercial platforms such as Affymetrix GeneChip [8] or Illumina bead stations [9] use a combinatorial probe for each target gene – that is each gene on the array is represented by a series of smaller oligonucleotides that span different parts of the gene. The Illumina bead technology synthesizes 50mer probes onto small beads immobilized into high-density fiber optic arrays. In the Affymetrix setup, 25mer oligos are synthesized directly onto the array substrate using photolithography, producing a barcode effect across

the chip rather than the discrete spot associated with contact-printing. This strategy incorporates mismatch oligos to give a measure of hybridization specificity across the series for each gene.

The choice of platform determines the type of experimental design possible (two color direct comparisons, or single color indirect comparisons), the type of normalization and filtering strategies employed (across multiples of the same probe, or across sets of probes for each gene), but most fundamentally impacts on strategies for maintaining and updating the identity and annotations of probes on the array.

3. IDENTITY CRISIS: THE SHIFTING NATURE OF PROBE ANNOTATIONS

Regardless of the platform used all probe annotations suffer from periodic loss-of-identity crises. Oligonucleotide probes are generally designed from 'gene' sequences that have been assembled from EST contigs – UniGene [10] and TIGR tentative consensus (TC) [11] sequences are examples of genes that have been assembled from a contig of expressed sequence tags (see Fig. 1). Likewise, many cDNA probe sequences are identified by their similarity to a UniGene contig. EST contigs offer the advantage of assembling many small, and individually uninformative sequences (ESTs) into longer, overlapping contiguous sequences that can be mapped back to the same genomic region, and that together predict some gene structures (such as exon boundaries) and predicted protein products where an open reading frame (ORF) can be detected. Such contigs are by their very nature fluid – as new ESTs are added to the database, previously disjointed contigs assemble together (merger). Conversely, as genome sequences are updated,

previously aligned contigs are broken up into separate entities (split) [12]. This means that the association of any individual probe (particularly a short oligo probe) with a contig-driven annotation will change as those contigs are updated.

The RefSeq sequences curated by NCBI are the most stable consensus gene sequences (usually accompanied with functional or empirical validation), however even these are problematic. Information about transcript variation (such as tissue specific alternate splicing events) are not reliably captured, and sometimes even merged, in the RefSeq dataset. New full-length cDNA (FL-cDNA) data emerging from annotation pipelines such as FANTOM (mouse) and the mammalian gene collection team (human) are highlighting the transcriptional complexity of the genome, and challenging our notions about genes as discrete heritable units [13,14]. The FLcDNA projects assemble transcripts into frameworks (Transcriptional Units, TU or Transcriptional frameworks, TK) that demonstrate the overlapping structure of genes. The exons of two functionally distinct frameworks are most commonly shared by the overlap of transcripts from different strands of a genomic neighborhood (sense-antisense transcription) – this is estimated as occurring as frequently as 70% of gene neighborhoods [15]. Exons may also be shared by transcriptional frameworks lying adjacent on the same strand – this type of promiscuous splicing event is less common than sense-antisense transcription, but is nevertheless a well validated occurrence (see [16] for example).

The consequence of these overlapping transcriptional frameworks on EST contigging approaches to gene annotation is immediately apparent – many of these contigs (UniGene and TC) summarize transcription from complex regions of the genome into a single consensus sequence. This occurs because ESTs often have little or no information about the strand from which they were transcribed (sense-antisense contigs); equally small ESTs that map to an exon shared between two transcriptional frameworks contain insufficient information to discriminate between the TK, and instead merge the two separate transcription events.

A common error in microarray annotations is to map the array probe to an EST contig once (such as a UniGene accession), and then inherit (and update) the annotations associated with that contig. A more satisfactory approach is to retain the original sequence of a microarray probe, and use this to identify and annotate each time. Even better is the mapping of each probe to the base genome sequence [17]. The annotations assembled through RefSeq and UniGene will still contribute, however information on exon specificity is retained. In an array where some redundancy exists – that is several probes are associated with one transcriptional framework – differences in signal across those probes might provide information on transcript variants arising from that framework.

For single oligo or cDNA probes the match between probe sequence and annotation can be updated automatically. More difficult is updating annotations on probes that summarize the signal from several oligos, as these consist of a series of short oligo sequences tiled across a gene. This approach suffers particularly from historical inaccuracies in

EST contigs and annotations as the identity of an individual oligo is much less transparent, and the signal from the probe series is a compilation that may not relate to the same transcriptional events. Mecham and colleagues have shown discrepancies with the mapping and identification of Affymetrix oligos [18]. Depending on the genome, between 20-60 % of probesets contained sequences that did not match the gene they were annotated as representing. These studies highlight the problems faced by annotating probesets on summary gene information – the probeset is designed to a contig that becomes outdated, probes within the probeset may belong to different transcription events but annotations are made through historical association rather than the actual sequence of the probe.

4. FUNCTIONAL ANALYSIS AND BIOLOGICAL INTERPRETATION OF MICROARRAY DATA

Researchers are demanding much more from a microarray experiment than a list of up- or down-regulated genes. If microarrays are to offer a truly horizontal approach to profiling cellular events, then functional associations between those genes must be made. The ultimate goal is to identify pathways-signaling events or genetic networks that regulate the phenotype of that cell. Good gene annotation requires the assembly of information about protein products (and variants) and protein motifs, including active domains. Functional information can be inferred from information gathered about a protein family or specific protein domains, and structural motifs [19,20] shared between functional classes. Subcellular localization can be predicted for some classes of secreted molecules and membrane proteins [21,22]. Gene Ontologies (GO) are a hierarchy of standardized nomenclature used to capture and describe such annotations. They provide broad biological classifications, specific molecular interactions, and subcellular localizations, based on the domains or motifs present in the predicted ORF (see <http://www.geneontology.org/> for more details).

The genomes of plants and higher eukaryotes (man and mouse, for example) express large numbers of noncoding RNAs – that is a full-length transcript from which no ORF can be predicted. Traditional GO pipelines are unable to effectively annotate these RNAs, although new ontologies are being developed, and are attempting to define putative functional classes (such as microRNAs) from this emerging category of transcripts [23]. UniGene and TIGR TC sequences are a common entry point for automated annotation tools such as NIH DAVID [24], Stanford SOURCE (<http://source.stanford.edu/cgi-bin/source/sourceSearch>) [25], TIGR Resourcer (<http://www.tigr.org/tigr-scripts/magic/r1.pl>) [26], EBI GOA (<http://www.ebi.ac.uk/GOA>) [27] or ProbeLynx for dynamic sequence-contig mappings [17]. Automated gene annotation pipelines rely heavily on gene ontologies for functional inference. These tools provide functional information, such as GO, predicted from the protein sequence associated with an EST contig that could not be predicted from an isolated EST. The mapping of individual genes into pathways is somewhat *ad hoc*, depending on external links to preassembled pathways such as those assembled at the Kyoto encyclopedia of genes and genomes (Kegg: <http://www.genome.ad.jp/kegg/>) [28,29] or BIOCARTA (<http://www.biocarta.com/index.asp>).

The continual annotation of probes on a microarray platform must be considered when warehousing microarray data, and must be curated in a transparent manner, preferably relying on genomic rather than EST mappings. This allows for the updating of annotations that take advantage of summary gene function, as well as information unique to the exons spanned by a probe or a set of probes. The developments of new normalization strategies that analyze every probe as an individual entity, combined with probe-based annotations rather than contig based annotations, are improving the reproducibility of information derived from these array platforms, and increasing the reliability of cross-platform comparisons.

5. MICROARRAY EXPERIMENTAL DESIGN: ASKING THE RIGHT QUESTIONS

Microarrays have been used to interrogate a wide range of biological questions, and the experimental design is crucial when considering what kind of information can be obtained from a microarray experiment. The simplest experiment captures a snapshot of genes expressed in a population of cells. These experiments do not rely on cross-array analysis; instead replicated data indicates the probability that a gene has been significantly detected in that population of transcripts. Comparison of gene expression across a series of such one-color hybridizations relies on adequate normalization (addressed below), and strategies that rank the expression of genes from highest to lowest within like experimental conditions. These strategies might include gross expression thresholds (i.e., a present call in one experimental condition but an absent call in all other conditions), alternatively the average hybridization intensity of each probe is ranked by comparing expression in one experimental condition relative to the average hybridization intensity across all experimental conditions. Finally, a 1-tailed t-test is also commonly employed to rank probes into 'high' and 'low' expression.

Several large datasets have been compiled to compare different tissues across the same experimental platform (see SymAtlas: <http://symatlas.gnf.org/SymAtlas/> and MGPD: <http://mgpd.med.utoronto.ca/> and READ: <http://read.gsc.riken.go.jp/>) [31-33] for examples. These databases can be queried at the level of tissue (specifying genes expressed at a particular level in that tissue), or at the level of an individual gene – interrogating the pattern of expression of that gene across many tissues. These databases employ similarity metrics, commonly pairwise comparisons with significantly scoring correlation coefficients (such as Pearson correlation) to identify other genes with statistically similar expression patterns in the dataset. This type of co-expression might imply co-regulation (eg tissue specificity) and/or a functional association (pathway specificity) between the gene products. Comparison of genes expressed in these different tissues is permitting the identification of cell lineage markers [31-33]. Similarly, the correlation of gene expression from tumor sample with clinical outcomes has revealed new markers for metastases, hormone responsiveness, response to drug therapies, and has even been used to predict the aggressiveness of the tumor (reviewed in [34]).

While comparing data generated as part of the same experiment, or even on the same experimental platform, is

relatively intuitive, a more difficult comparison is that of gene expression profiles generated on different experimental sources. Successful strategies for cross-platform analyses treat microarray data as qualitative rather than strictly quantitative, and ensure that data extracted from different datasets are related to a comparable baseline [18,35].

Although DNA microarray data have been used for many important applications in the life sciences, they are often represented with missing values which are caused by spurious image signals that can arise due to insufficient image resolution, precipitated probe, other hybridization artifacts, or dust on the surface of the slide. Missing values in microarray data make it difficult for the analysis and classification of the data sets by other computational methods, which require complete gene expression matrices. Therefore effective methods for estimating missing values are needed to make the matrix of gene array values complete to be ready for the study of gene expression.

Microarray gene expression data are conventionally presented in the form of matrix format where the number of rows stands for the number of genes, and the number of columns is the number of different experimental conditions. Missing elements of a microarray matrix are simplistically replaced with zeros [64] or the row average expression levels for the corresponding genes. However, Troyanskaya *et al.* [65] showed that these methods are not effective because replacing missing values with zeros is very heuristic and far from optimal, whereas the row average is only based on simple statistics which ignore the information of other genes. These authors carried out a comparative study on three methods for estimating missing values in gene microarray data: the singular value decomposition based method, the weighted k -nearest neighbor method (KNNimpute), and the row average method. They found that the weighted k -nearest neighbour method was superior to the other two methods and the method of filling missing values with zeros.

Bø *et al.* [66] proposed two least-squares based methods for estimating missing values using the correlations between genes and between arrays, and then combining these two methods by taking the weighted averages of the estimates from the two methods. After testing their proposed methods (LSimpute) and the KNNimpute with three data sets, they concluded that the LSimpute methods produced more accurate estimates than the KNNimpute and its accuracy was at least equivalent to the expectation maximization approach (EMimpute). Nguyen *et al.* [67] evaluated the accuracy of the mean imputation and k -nearest neighbor (KNN) imputation on predicting gene expression missing values, and proposed two regression methods for estimating missing values. The first method is the imputation using repeated ordinary least squares (OLS) regression which takes pairwise gene information into account; whereas the second imputation method is the partial least squares (PLS) regression which incorporates global gene structure by using all gene expression values to estimate the missing values. After testing different methods on a variety of cDNA and oligonucleotide microarray data sets, the results were found to be consistent with those reported by Troyanskaya *et al.* [65] for yeast cDNA microarray data, in that the performance of the mean imputation was much poorer than the KNN imputation, which showed high performance on the

average of the experiments. In addition, these authors found that both OLS and PLS provided better accuracy of the estimation than the KNN over some ranges of the gene expression values. Kim *et al.* [68] proposed an imputation approach based on local least squares criteria (LLSimpute). Missing gene expression data were estimated by LLSimpute as a linear combination of similar genes. Based on the experimental results, these authors reported that their proposed approach is competitive with KNNimpute, and a Bayesian estimator – BPCA [69].

In analyzing the variability of a particular data set which can be spatially related, classical statistical methods make no use of this type of information; whereas the theory of geostatistics considers the spatial information of the data set in its regression analysis for estimating missing observations or unknown data. Pham [123] proposed a framework for estimating missing observations which incorporates the modeling of fuzzy prototypes in the cokriging system of geostatistics in order to improve the accuracy of the estimates and alleviate the computational complexity of cokriging. Geostatistics is thought as a special theory of applied statistics. The formulation of geostatistics is based on the abstract theory of the regionalized variable, which is spatially continuous and has the properties of being intermediate between truly random and completely deterministic. In other words, geostatistics considers the observed values of a variable as the realizations of a stochastic process in space. For each location in a spatial domain, there exists a measurable quantity, which is called a regionalized variable. This regionalized variable is considered to be a particular outcome or realization of a random variable. The set of these auto-correlated random variables constitutes a random function.

Each of the random variables has the same probability law; and at all locations the random variable has the same statistical expectation. The joint distribution of any pair of random variables depends on the lag distance between the two variables and not on their locations. Thus, the problem of characterizing the spatial variability of a regionalized variable reduces to that of characterizing the correlations of set of the random variables, which constitute the random function [124]. Using the properties of regionalized variables, kriging and cokriging systems are developed for estimating missing or unknown values in the ways that i) the estimated value is the weighted linear combination of the available data, ii) the estimate is unbiased (average error is zero), and iii) the error variance is minimized. To achieve such purposes, the estimation methods of geostatistics utilize a random function model to express the estimate error, its mean value, and its variance; and then work out how to weight the neighbor data so that the estimates satisfy the above criteria. The fuzzy cokriging estimator was tested using the heredity breast cancer microarray-based data set [72], which consists of twenty-two breast tumor samples from 21 patients: 7 BRCA1, 8 BRCA2, and 7 sporadic. This data set comprises 3226 genes for each tumor sample, and the complete gene expression matrix of the combined 22 breast cancer samples consists of 3226 cDNAs (3226 x 22 matrix). Each of 3226 genes can be considered as a random variable, which has 22 realizations. The ordinal one-dimensional representations of the sampled values of the genes allow us to model the spatial relationships of the gene

variables. To compare the fuzzy cokriging with the conventional cokriging, 10 sets of 20 genes were randomly selected and introduced the numbers of missing observations to be 2, 4, 6, 8, and 10 in each of the 10 subsets. The number of nearby available samples for the estimate was chosen to be 4, where the position of the unknown value is at the center of the nearby samples. The accuracy is measured using the average root mean square (RMS) errors, and the resultant RMS error was then taken as the average of all the test sets. While the errors of both methods are found equivalent for 2, 4, and 6 missing values, the errors of the fuzzy cokriging are smaller than those of the cokriging in the cases of 8 and 10 missing values. The fuzzy cokriging method was further tested with the whole data set of 3226 genes, and used the same numbers of missing values on 10 as well as other parameters as introduced in the first test. Given the large number of genes, we selected the number of cluster centers to be 20. Due to a significant large number of genes, it becomes impractical to apply the cokriging in a straightforward procedure. The proposed method was compared with the row average (mean) method. The results obtained from the proposed method were significantly superior to those obtained from the row average method.

6. DATA PRE-PROCESSING AND NORMALIZATION

Microarrays generate a large amount of horizontal data (from thousands of individual hybridization events on each array), but tend not to generate a large amount of vertical (replicated) data for any single probe in the course of an experiment. This causes unique issues when dealing with data normalization and statistical analysis. The data generated from each probe on each array must be transformed so that it is comparable to the other hybridization events on the same array (per-chip normalization) as well as the hybridization events of that probe across a series of arrays (per-gene normalization). Regardless of the platform used, a median-normalized per-gene and per-chip strategy assumes Gaussian distribution of the signal intensities between and across experiments [36].

Microarrays fall into several categories when considering normalization strategies. The first, best typified by the Affymetrix platform, and arguably most widely adopted category, is the hybridization of a single RNA population to each array, and the comparison of multiple RNA populations across multiple arrays. These arrays use the average of a series of probes to represent one gene. The variability of signal generated within the probe series has been well documented [37], not least because of the design problems highlighted above. A recent advance in normalization of Affymetrix arrays by Speed's laboratory has been widely and successfully adopted [38]. Instead of summarizing the signal from a set of gene-specific probes before normalizing the data generated from any one GeneChip, the RMA normalization strategy treats each probe as a unique entity and performs the per-gene and per-chip normalizations before summarizing the geneset.

The second category is associated with spotted cDNA and oligo arrays, and permits the hybridization of two (or more) RNA populations to a single array, where the origin of each RNA molecule is identified by a fluorescent tag. The basic principle of producing microarrays in an experiment is

based on the two-color fluorescence imaging labeling method. For such an experimental platform, mRNAs from the sample population and control cells are converted to cDNAs and labeled with two different fluorescent dyes: red dye Cy5 for the RNA from the sample population, and green dye Cy3 for the RNA from the control population. The two labeled extracts are then hybridized over the microarray, where labeled cDNA sequences bind to their respective probes on the array. Since the two differently labeled cDNAs are placed on one array, some measure of hybridization can be observed for each probe, which is called the measure of DNA abundance or gene expression. If the gene from the sample source is highly expressed then the spot on the array will be predominantly red. Each spot is normalized relative to the two fluorescent signals, generating a ratio. However the dyes available for labeling have different incorporation efficiencies and fluorescent intensities, and these differences vary as signal intensity increases. A second confounding factor in the normalization of printed arrays are regional differences in the morphology and concentration of probe spotted depending on the physical tip used to deposit the probe onto the surface of the slide. Print-tip lowess normalization is an intensity-dependent, regional normalization tool that uses local-linear regression to normalize signal intensities across an array (per chip), as well as between arrays (per gene) [39]. Lowess normalization was important because it recognized that the data generated across an array was not linear. It is equally important to remember that the data generated by a microarray experiment is also constrained by the linear range of the equipment used to detect the signal – and no amount of normalization, filtering or analysis can alter the lower (signal-noise ratio) or upper (saturation point) of the data once it has been collected.

A common problem in the analysis of microarray data is the correction of the background image intensity. The generation of DNA microarray spots involves the hybridization of two probes labelled with a fluorescent red dye Cy3 or a fluorescent green dye Cy5. The relative image intensity values of the red dye and the green dye on a particular spot of the arrays indicate the expression ratio for the corresponding gene of the two samples from which the mRNAs have been extracted.

Kooperberg *et al.* [40] pointed out that there are two problems associated with the measure of gene expression values. Corrections of these problems would enhance the accuracy in the estimation of the gene expression ratio. The first is known as the background intensity problem and arises when a probe is applied to the array even when cDNA is not available. The second problem is that not only a probe corresponding to the correct gene hybridizes with the cDNA but some probe corresponding to some other gene may also do so, leading to cross hybridization. This is a profound problem, as it is not consistent across an array, but rather caused by local hybridization events (buffer precipitation, slide drying out, unevenness of slide substrate coating, poor blocking etc.). The standard approach for background image correction is to subtract the background image from the spot image. However, this approach is not effective when the spot image is low or similar to the intensity levels of the background image. Thus, the estimate of the expression ratio becomes unreliable. Kooperberg and coworkers dealt

particularly with the problem of background correction and proposed a Bayesian method to account for the background noise. Summary statistical properties, which include mean, median, and standard deviation of each spot, were used as the features for the background correction. These authors were not aware of any methods for correcting spotted glass array data for cross hybridization, and compared the estimation of the expression ratios based on glass spotted arrays with those obtained using Northern blot analysis. They pointed out that the Northern blot analysis does not suffer from cross hybridization the way that glass spotted arrays do, providing an explanation of the main inconsistency between these two results. However, this simply may not be true – Northern blots most certainly do suffer from cross hybridization events. In both a Northern and an array experimental conditions such as buffer stringency (salt), hybridization temperature, % match between probe and target, composition of probe, secondary structure of probe etc., all affect nonspecific events in an array or a Northern. The main inconsistencies between an array and a Northern include sensitivity, isoform information, and probe misannotation on an array.

Earlier research work on the estimation of gene expression ratios for glass spotted arrays includes an algorithm developed by Chen *et al.* [41], which used a Wilcoxon statistics to identify the background area as the pixels having significantly lower image intensity values than those of other pixels. Theilhaber *et al.* [42] proposed a Bayesian approach for calculating the expression ratios with an assumption that the background image has been corrected. Another Bayesian method for calculating the log-ratio after the background correction was proposed by Newton *et al.* [43] and based on the assumption that the foreground pixel intensity values for the red and green channels obey a Gamma distribution. Although several methods have been developed for dealing with the correction of background image to enhance the estimate of gene expression ratios, spots of low intensity values remain a difficult problem, which also greatly affects particular subsequent tasks of microarray analysis involving image segmentation methods and clustering algorithms. It is important to reiterate here that low signal is not simply a computational problem – below a certain threshold one can reach a physical limit of the equipment, the data collected does not reflect the signal produced, and the scanner is not sensitive enough to pick it up.

7. IMAGE PROCESSING OF MICROARRAY DATA

7.1. Gridding

To quantify the fluorescence-based spot images, one of the first steps is to identify the location of each image block which contains the spots as well as the location of each image spot. This task refers to grid fitting or spot addressing. After this task the process of spot fitting or quantification of spots is carried out to compute the intensity values of the spot and the background images. To automate this process, Jung and Cho [44] proposed an automatic block and spot indexing system based on a k -nearest neighbors graph. For the gridding problem, the results of the indexing system are the expression profiles which include the average or median value of the image intensities in the region of each spot. These authors addressed three problems of gridding: block

indexing which finds the block index of a particular image spot; spot indexing which identifies the element of a particular spot in an indexed block; and intensity computation which estimates the image intensity of a particular spot. The authors discussed a method for quality control of the shape and position of the spot proposed by Buhler *et al.* [47]. However, this method is not effective for the full automation of image processing of microarrays when there are large amounts of images produced by heterogeneous microarrays.

Jung and Cho [44] stated that the ideal microarray image should have all the blocks of the same size, uniform spacing between two blocks, uniform distance between two spots, spots of the same size and perfectly circular shape, fixed location of the grids for a given type of slides, a slide which is free from dust or other contamination, and uniform background intensity across the image. This ideal is difficult to realize for printed arrays, and even arrays produced *in situ* (Affy) suffer from regional artifacts. So filtering of imperfect spots and contaminations is an important computational problem. These authors also discussed other methods for automating the image analysis of microarrays proposed by Steinfath *et al.* [45] and Jain *et al.* [46]. The former method was developed to automate the microarray hybridization experiments from image analysis to clustering but did not address the problem of image block identification. The latter method is concerned with the automated quantification of microarray image data by locating and segmenting the spots as well as estimating their expression ratios. However, this system works well if the expression rate is more than 80%; if it is less than 70% then manual analysis is needed.

7.2. Spot Extraction

In order to ensure the accuracy of gene expression data, the extraction of image spots must be carefully carried out in the presence of noise, artifacts, low contrast, and irregular shapes of microarray images. Several image analysis methods have been developed to address this requirement of microarray data as a preprocessing step. The most popular approach for this task is known as image segmentation. Most commercial software, including ScanAlyze [48], GenePix [49] and QuantArray [50], assume the spot shape to be circular and of fixed diameter to facilitate the extraction procedure and require some manual analysis. Vincent and Soille [51] as well as Beucher and Meyer [52] developed a watershed-based shape segmentation methods using adaptive procedures; whereas Adams and Bischof [53] developed a method for spot segmentation using the region-growing approach which was also adopted by the software Spot [54] for segmenting microarray images. More recently, Bozinov and Rahnenfuhrer [55] proposed an unsupervised method for segmenting microarray spots using an adaptive procedure based on two clustering methods, the *k*-means and medoids algorithms, to process the microarray image data. Three main phases of image processing of microarray data, which contains the intensity values for fluorescence green (*Cy3*) and fluorescence red (*Cy5*) components, and the ratio of *Cy5/Cy3* for a single spot, are the identification, segmentation, and estimation of relative abundance. The identification phase involves the finding of the target area, which includes the spot and the background pixels of the spot area; the segmentation tries to separate the target area

into two distinct regions of background and foreground; whereas the third phase is to compute the representative intensity values of the red and green fluorescence dyes and the corresponding ratio to express the relative abundance of each spot. Bozinov and Rahnenfuhrer [55] addressed the second and the third phases of the image processing, and concluded that their method was effective to deal with various types of artifacts, and superior to some other methods in handling weak gene spots. However, spots having low intensity values and large artifacts tend to make it difficult for the clustering analysis proposed by these authors.

Hirata *et al.* [56] proposed a method based on mathematical morphology for microarray spot segmentation which automatically performs the task of spot segmentation based on the object shapes. The image processing procedure of the method consists of image composition, rotation correction, subarray gridding, spot gridding, automatic correction of spot gridding, and spot segmentation. The method was tested with several microarray images from different microarray spotters and scanners. The authors reported that the experimental results indicated good visual inspection on the separation of the hybridized intensity from the background, and good segmentation of spots having irregular shapes or weak signals. However, poor segmentation was observed when the intensity values of the spots were very close to the background image. Unfortunately, the proposed method was not compared with other techniques. As another mathematical-morphology based method, Angulo and Serra [57] proposed a non-supervised set of algorithms for spot extraction from DNA microarrays using morphological operators. This method works as follows: it carries out the segmentation of the image into spot quadrants then performs the analysis of the spot quadrant images by the quantification of spot distribution, extraction of background noise by morphological filtering, spot orthogonal gridding, spot segmentation or spot boundary detection by watershed transform, and extraction of spot intensity values. The authors compared their proposed method with two microarray imaging packages known as ScanAlyze [48] and GenePix [58] and concluded that their method was superior to the others in terms of robustness and precision.

Liew *et al.* [59] developed an adaptive method for microarray spot segmentation. This method, which is based on adaptive thresholding and statistical intensity modeling, generates the grid structure and then performs the image segmentation to extract the spots within the subregions of the image grid. These authors compared their method with GenePix [58] and concluded that their method was superior in terms of robustness and suitability for high throughput analysis of microarray data. Another adaptive method recently proposed by Damiance *et al.* [61] is based on dynamic system modeling for extracting the microarray spots. This method adopts a network model and a similarity measure to assign the pixels to their corresponding clusters, without having to predefine the number of clusters and without having to perform the gridding process. However, again, this method has not been tested against other methods for microarray image segmentation and spot extraction.

Some other recently developed non-segmentation based methods for microarray data analysis include the combinatorial image analysis of DNA microarray features [62] and compression of microarray images [63]. The first method uses median filters to suppress noise and top-hat filters to perform the background correction. In addition to solving the gridding problem, the second method addressed the issue of compressing microarray images due to the massive amount of image data. This method compresses the pixels of the microarray spots without the loss of data while allowing some controlled loss of the background pixels. The authors concluded that their method produced significant compression rate without affecting the regions of interest of images.

8. CLASSIFICATION OF MICROARRAY DATA

The delivery of reliable microarray data for downstream study is the result of several experimental and analytical processes: from the experimental platform for performing gene hybridization, to image processing, and estimation of missing values. Classification of gene samples into appropriate groups or categories is an important application of gene expression microarray data. DNA microarray data consist of a large number of genes and a relatively small number of experimental samples. The number of genes on an array is in the order of thousands, and because this far exceeds the number of samples, dimension reduction is needed to allow efficient analysis of data classification techniques. Based on the motivation that conventional statistical methods for pattern classification break down when there are more variables (genes) than there are samples, Nguyen and Rocke [70] proposed an approach for classifying human tumor samples based on microarray gene expression data. This approach attempted to reduce the gene dimension using a partial least squares (PLS) method, followed by classification of the data using quadratic discriminant analysis (QDA). The authors tested their proposed methods with five microarray data sets collected from various human tumor samples. These five data sets included normal versus ovarian tumor, acute myeloid leukemia (AML) versus acute lymphoblastic leukemia (ALL), diffuse large B-cell lymphoma (DLBCL) versus B-cell chronic lymphocytic leukemia (BCLL), normal versus colon tumor, and non-small-cell lung carcinoma versus renal samples. They compared the PLS with the method of principal component analysis (PCA), found their method was superior and concluded that their proposed classification procedure was able to distinguish between normal and tumor as well as between two types of tumors from the five microarray data sets with high accuracy.

Zhou *et al.* [71] proposed a Bayesian approach for selecting the strongest genes based on microarray gene expression data and the logistic regression model for classifying and predicting cancer genes. These authors used the Gibbs sampling and Markov-chain Monte-Carlo methods to identify important genes, which were then analysed *via* the logistic regression model for cancer classification and prediction. The proposed method was tested against three large microarray data sets that included hereditary breast cancer, small round blue-cell tumors, and acute leukemia. The authors concluded that their proposed method could identify important genes whose biological functions were

consistent with the disease sets [72], and could classify the data sets with higher accuracy than other methods [72,73].

Statnikov *et al.* [74] carried out a comprehensive evaluation of classification methods for cancer diagnosis based on microarray gene expression data [75]-[89]. Based on the motivation that cancer diagnosis is one of the most important emerging clinical applications of gene expression microarray technology, these authors developed a computer model for use in microarray-data-based cancer diagnosis and carried out comprehensive evaluation of several major algorithms for multiclassification. They concluded that multiclassification support vector machines (MC-SVMs) are the most effective classifiers in performing accurate cancer diagnosis from gene expression data, outperforming other popular machine-learning algorithms including *k*-nearest neighbors, backpropagation, and probabilistic neural networks. They found that gene selection techniques could significantly improve the classification performance of both MC-SVMs and other non-SVM learning algorithms, and that ensemble classifiers generally did not improve performance of the best non-ensemble models. Based on these results, these authors built a software system, GEMS (Gene Expression Model Selector), that automates high quality model construction and enforces sound optimization and performance estimation procedures.

As an alternative approach, we report in detail here a very recent theoretical development – an optimally weighted fuzzy *k*-NN decision rule, which may be useful for the classification of microarray data. The nearest neighbor rule [91] is a non-parametric approach and has been widely used for pattern classification. The *k*-nearest neighbor (*k*-NN) rule assigns crisp memberships of samples to class labels; whereas the fuzzy *k*-NN neighbor rule [92] replaces crisp memberships with fuzzy memberships. The membership assignment by the conventional fuzzy *k*-NN algorithm has a disadvantage in that it depends on the choice of some distance function, which is not based on any principle of optimality. To overcome this problem, we have introduced a computational scheme for determining optimal weights to be combined with different fuzzy membership grades for classification by the fuzzy *k*-NN approach.

This version of the *k*-NN algorithm [90] assigns the fuzzy membership for an unknown sample x_u to class label y as an optimally weighted linear combination of the fuzzy membership grades of k nearest samples:

$$\mu_{yu} = \sum_{i=1}^k w_i \mu_{yi} \quad (1)$$

where μ_{yu} and μ_{yi} have been previously defined, $\{w_i, i=1, \dots, k\}$ are the optimal weights which indicate the relationship between x_i and x_u , and to be determined. It should be noted that normalization is not needed in (1) because the sum of w_i is equal to one.

The set of optimal weights expressed in (1), which quantify the relationships between the unknown and available samples, can be equivalently derived from the estimate of the value of the unknown sample x ; this results in the set of optimal weights for the linear combination of the available samples:

$$\hat{x}_u = \sum_{i=1}^k w_i x_i \quad (2)$$

where x_u is the estimate of x_u , and x_1, \dots, x_k are available sample data.

There are different approaches for determining the weights to the available or neighbor data with respect to the unknown value, and different approaches lead to different computational schemes. One particular approach for optimally computing these weights is to minimize the average error of estimation. Thus, if one lets r_j denote the error between any particular estimated x_u value and the true value x_u :

$$r_j = \hat{x}_j - x_j \quad (3)$$

then the average error, denoted as r_a , of k estimates is

$$r_a = \frac{1}{k} \sum_{j=1}^k r_j \quad (4)$$

However, minimizing r_a is unrealistic because the true values x_1, \dots, x_k are not known. One possible solution to this problem is the use of an ordinary kriging computational scheme that considers the unknown values as the outcome of a random process and solves the problem by statistical procedures. In other words, it is not possible to minimize the variance of the actual errors, but it is possible to minimize the variance of the modeled error, which is defined as the difference between the random variables modeling the estimate and the true value. As the result of statistical and analytical analysis, kriging computes the set of optimal weights by solving the following system of equations:

$$Cw = D \quad (5)$$

where

$$C = \begin{bmatrix} c_{11} & \dots & c_{1k} & 1 \\ \cdot & \dots & \cdot & \cdot \\ \cdot & \dots & \cdot & \cdot \\ \cdot & \dots & \cdot & \cdot \\ c_{k1} & \dots & c_{kk} & 1 \\ 1 & \dots & 1 & 0 \end{bmatrix}$$

$$w = [w_1 \dots w_k \beta]^T$$

and

$$D = [c_{1u} \dots c_{ku} \ 1]$$

where c_{ij} is the covariance of x_i and x_j , w_1, \dots, w_k are kriging (optimal) weights, and β is a Lagrange multiplier.

The values of the kriging weights can be obtained by solving

$$w = [C^{-1}D] \quad (6)$$

where C^{-1} is the inverse of the covariance matrix C .

It is known that the solution of a kriging system can result in negative weights that should be avoided in order to ensure the robustness of the estimation. One can adopt the simple and effective procedure for correcting negative weights proposed by Journel and Rao [119]. This method

determines the largest negative weight and adds an equivalent positive constant to all weights, which are then normalized:

$$W_i^* = \frac{W_i + \alpha}{\sum_{i=1}^k (W_i + \alpha)}, \forall i \quad (7)$$

where w_i^* is the corrected weight of w_i and

$$\alpha = -\min_i w_i \quad (8)$$

The derivation of the kriging system expressed by (5) is based on the assumption that the probabilistic model employed by kriging is a stationary random function. This stationary function consists of several random variables, one for each of the available values and one for the unknown value. Let $V(x_1), \dots, V(x_k)$ be the random variables for k samples x_1, \dots, x_k , respectively; and $V(x_u)$ be the random variable for x_u . These random variables are assumed to have the same probability distribution, and the expected value of the random variables at all locations is $E\{V\}$. Thus, the estimate of x_u is also a random variable and expressed by a weighted linear combination of the random variables at k locations:

$$\hat{V}(x_u) = \sum_{i=1}^k W_i V(x_i) \quad (9)$$

Thus, the error of estimation is

$$R(x_u) = \sum_{i=1}^k W_i V(x_i) - V(x_u) \quad (10)$$

The expected value of the error of estimate is

$$E\{R(x_u)\} = \sum_{i=1}^k W_i E\{V(x_i)\} - E\{V(x_u)\} \quad (11)$$

Based on the assumption that the random function is stationary, Eq. (11) becomes

$$E\{R(x_u)\} = \sum_{i=1}^k W_i E\{V\} - E\{V\} \quad (12)$$

To satisfy the unbiased condition, $E\{R(x_u)\}$ must be set to zero:

$$E\{R(x_u)\} = 0 = \sum_{i=1}^k W_i E\{V\} - E\{V\} \quad (13)$$

which leads to

$$E\{V\} \sum_{i=1}^k W_i = E\{V\} \quad (14)$$

Therefore

$$\sum_{i=1}^k W_i = 1 \quad (15)$$

The variance of the random variable $V(x_u)$ is given by

$$\text{Var}\left\{\sum_{i=1}^k w_i v_i = \sum_{i=1}^k \sum_{j=1}^k w_i w_j \text{Cov}\{v_i v_j\}\right\} \quad (16)$$

Given that $R(x_u) = V(x_u) - V(x_u)$ and using (16), the error variance is defined as

$$\text{Var}\{R(x_u)\} = \text{Cov}\{\hat{V}(x_u)\hat{V}(x_u)\} - 2\text{Cov}\{\hat{V}(x_u)V(x_u)\} + \text{Cov}\{V(x_u)V(x_u)\} \quad (17)$$

which can be written as

$$\sigma_R^2 = \sigma^2 \sum_{i=1}^k \sum_{j=1}^k w_i w_j C_{ij} - 2 \sum_{i=1}^k w_i C_i \quad (18)$$

which defines the variance of error as a function of w_1, \dots, w_k .

An optimal choice for the kriging weights is to minimize the variance of error. This can be done by the Lagrangean method:

$$\sigma_R^2 = \sigma^2 + \sum_{i=1}^k \sum_{j=1}^k w_i w_j C_{ij} - 2 \sum_{i=1}^k w_i C_i + 2\beta \left(\sum_{i=1}^k w_i - 1\right) \quad (19)$$

where β is a Lagrange multiplier.

After differentiating (19) with respect to all w_i and β and setting each one to zero, we obtain

$$\sum_{j=1}^k w_j C_{ij} + 2\beta = C_{iu}, \forall i = 1, \dots, k \quad (20)$$

Expressions (20) and (15) define the ordinary kriging system of equations expressed in (5), which is represented in the form of matrix notation. If the data are spatially related then the covariance can be calculated as [118]

$$C(h) = \frac{1}{N(h)} \sum_{(i,j) \in h} x_i x_j - \left(\frac{1}{n} \sum_{k=1}^n x_k\right)^2 \quad (21)$$

in which the covariance is a function of the lag distance h , $N(h)$ is the number of pairs that x_i and x_j are separated by h , and n is the total number of data.

Alternatively, the covariance function $C(h)$ can be replaced by the variogram function, denoted as $\gamma(h)$, which is half the average squared difference between the paired data values:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{(i,j) \in h} (x_i - x_j)^2 \quad (22)$$

It can be noted that the computation of the kriging weights that are used to make inference about the fuzzy membership grade of an unknown sample with respect to a particular class is not restricted to the sense that the data are spatially related. Both conventional and spatial covariance values can be used in the computation of the kriging system to derive the set of optimal weights for the proposed fuzzy k -NN algorithm.

The data used to test the optimally weighted k -NN algorithm is the microarray-based hereditary breast cancer data, first studied by Hedenfalk *et al.* [72]. The data consist of 22 cDNA microarrays comprising 3226 genes. The twenty-two breast tumor samples were collected from the biopsy specimens of 7 patients with germ-line mutations of BRCA1, 8 patients with germ-line mutations of BRCA2, and

7 patients with sporadic cases. The ratio data was truncated below 0.1 and above 20. Log of the ratio data were used to classify BRCA1, BRCA2, and sporadic. The microarray data can be represented in matrix notation as $X = [x_{ij}]$, $i = 1, \dots, N$, $j = 1, \dots, M$, where N and M are the numbers of tumor samples and genes, respectively.

To determine the fuzzy membership grades for sample data, the fuzzy c -means algorithm (FCM) [96] was applied to partition the data set into three fuzzy prototypes according to the three classes. The FCM performs the partition based on the following objective function

$$J_m = \sum_{i=1}^N \sum_{y=1}^c (\mu_{iy})^m d_{yi}^2 \quad (23)$$

where

$$d_{yi}^2 = \|x_i - v_y\|_A^2 = (x_i - v_y)^T A (x_i - v_y) \quad (24)$$

in which c is the number of clusters or fuzzy prototypes, m is the weighting exponent, $1 < m < \infty$, $v = (v_1, v_2, \dots, v_c)$, the vector of cluster centers, $v_y = (v_{y1}, \dots, v_{yM})$, $\|\cdot\|_A$ is the A -norm which is positive-definite ($M \times M$) weight matrix, and if A is the identity matrix then it becomes the Euclidean norm.

The FCM tries to minimize J_m by iteratively updating the partition matrix using the following equations:

$$v_y = \frac{\sum_{i=1}^N (\mu_{yi})^m x_i}{\sum_{i=1}^N (\mu_{yi})^m}, 1 \leq y \leq c \quad (25)$$

and

$$\mu_{yi} = \frac{1}{\sum_{z=1}^c (d_{zi}/d_{yi})^{2/(m-1)}}, 1 \leq i \leq N; 1 \leq y \leq c. \quad (26)$$

Ten subsets of the cancer data set were randomly selected, each consisting of 22 tumor samples and 100 genes, to test the optimally weighted fuzzy k -NN method and compare its results with those obtained by the k -NN and fuzzy k -NN algorithms. The leave-one-out method was used to evaluate the classification performances of the k -NN (KNN), fuzzy k -KNN (FKNN), and the optimally weighted fuzzy k -NN (OWFKNN) classifiers. The numbers of nearest neighbors for the classification were: $k = 5, 10$, and 15 . The weighting exponent m expressed in (23) was taken to be 2. The fuzzy prototypes obtained from the FCM were used as the mean values for calculating the covariances which were included in the computation of the kriging system.

For $k = 5$, the total average percentage of classification accuracy for the KNN, FKNN, and OWFKNN was determined to be 89.5%, 91.4%, and 92.9% respectively. For $k = 10$, the total average percentage of classification accuracy for the KNN, FKNN, and OWFKNN was 90.5%, 92.4%, and 94.4%, respectively. For $k = 15$, the total average percentage of classification accuracy for the KNN, FKNN, and OWFKNN was 91.9%, 94.8%, and 97.7%, respectively. It can be seen that the OWFKNN outperformed the other two classifiers in all test cases. It can be seen that the

classification results for all algorithms improve when k is increased. The performance of OWFKNN was particularly enhanced when more nearest neighbor samples were considered in the sense of statistical correlation. In terms of the computational aspect of each algorithm, the KNN is the simplest and fastest method, whereas the OWFKNN requires the most computational effort – this is due to the computations of fuzzy prototypes, covariance matrix, and kriging system of equations.

9. CLUSTER ANALYSIS OF MICROARRAY DATA

Cluster analysis plays an essential role in the study of microarray data, through the grouping of genes into sets with similar expression patterns or biologically linked functions. Many clustering methods have been applied and developed for studying gene expression data. Different clustering models using different optimal criteria and similarity measures lead to different partitions of the data sets. Dougherty *et al.* [94] discussed a model-based clustering toolbox that evaluates cluster accuracy. These authors studied five algorithms for clustering microarray data: k -means, fuzzy k -means, self-organizing maps, hierarchical Euclidean-distance-based clustering, and correlation-based clustering. In some cases, the fuzzy k -means and the hierarchical Euclidean-distance-based clustering algorithms showed better performance than the others; whereas the k -means algorithm was found to be inconsistent on several tests.

The k -means or self-organizing maps assign each gene to a single cluster but do not provide information about the influence of a given gene on the overall shape of clusters. Dembélé and Kastner [95] applied the fuzzy c -means (FCM) algorithm [96] to assign cluster membership values to genes. These authors pointed out that a problem in applying the FCM method for clustering microarray data is the choice of the exponent weight (parameter of fuzziness) and showed that the commonly used value of two for this weight is not appropriate for some data sets, and that this value may vary widely from one data set to another. They proposed an empirical method, based on the distribution of distances between genes in a given data set, to determine an adequate value for this parameter. The authors tested the FCM, where the number of clusters were identified by an algorithm developed by Sharan and Shamir [97], with a yeast cell cycle data set, in which the expression profiles of 6200 yeast genes were measured every 10 min during two cell cycles in 17 hybridization experiments [98], but used the same selection of 2945 genes made by Tavazoie *et al.* [99]. They showed that an appropriate selection of the exponent weight of the FCM increased the overall biological significance of the genes within the cluster. Their proposed procedure was also tested against the serum data [100], which consists of 517 genes whose expression varies in response to serum concentration in human fibroblasts, and human cancer data (<http://discover.nci.nih.gov/nature2000/>), which represents gene expression patterns of 9703 genes in 60 human cancer cell lines. The authors concluded that the FCM is a convenient method for selecting genes exhibiting tight association to given clusters, whereas conventional clustering methods force all genes into clusters and assign to each cluster some genes which may only be marginally relevant for the biological significance of the cluster.

Although the FCM being reported by Dembélé and Kastner [95] is an effective method for the clustering analysis of microarray hybridization data to identify biologically relevant groups of genes, its reliability in the analysis of microarray data has not yet been evaluated. Asyali and Alci [101] pointed out that a serious limitation in microarray analysis is the unreliability of the data generated from low signal intensities. Such data may produce erroneous gene expression ratios and cause unnecessary validation or post-analysis follow-up tasks. Therefore, the elimination of unreliable signal intensities would enhance reproducibility and reliability of gene expression ratios produced from microarray data. These authors applied the FCM and normal mixture modeling (NMM) based classification methods to separate microarray data into reliable and unreliable signal intensity populations. They compared the results of FCM classification with those of classification based on NMM. Both approaches were validated against reference sets of biological data consisting of only true positives and true negatives. They used data from three independent experiments of microarray gene expression from the same cell system [102,103] in order to test and compare different classification approaches. The cDNA microarrays they used consisted of about 2000 cDNA distinct probes and a total of approximately 4000 elements [104], and a publicly available data set recently studied by Chang *et al.* [105]. Based on experimental results, they observed that both methods performed equally well in terms of sensitivity and specificity. Although a comparison of the computation times indicated that the fuzzy approach is computationally more efficient, other considerations support the use of NMM for the reliability analysis of microarray data.

Sturn *et al.* [125] recognized that there have been several clustering models applied to the analysis of microarray data, but there was no single tool that combines clustering and visualization methods and allows comparisons using different clustering methods. These authors therefore developed a computer program for large-scale gene expression analysis which integrates various computational models, including hierarchical clustering, self organizing maps, k -means, principal component analysis, and support vector machines to explore and visualize various relationships obtained from different methods. However, this work did not either compare the effectiveness or combine the results given by different methods.

Li *et al.* [126] proposed a clustering algorithm based on the minimum entropy criteria, which is the conditional entropy of clusters given the observations. These authors generalize the entropy criterion by replacing the well-known Shannon's entropy [127] with the Havrda-Charvat's structural α -entropy [128]. They tested their proposed clustering algorithm with gene expression data and reported that their algorithm was superior to the k -means algorithm, hierarchical clustering, self organizing map, and expectation-maximization clustering algorithm in terms of adjusted Rand index [129]. However, again, the authors did not extend their investigations to compare their methods with the other clustering algorithms in the context of biological meaning as the result of gene expression analysis.

10. METHODS FOR SELECTING DIFFERENTIALLY REGULATED GENES

Another important task in microarray data analysis is to determine which genes are differentially expressed across two kinds of tissues or samples obtained under two experimental conditions. Thus, the term differentially expressed genes, or discriminator genes, are genes which have significantly different expression in two defined groups of microarray experiments. Pan [106] reported that several statistical methods have been proposed to deal with this problem when there are replicated samples under each condition. The author applied and compared three methods: the *t*-test, a regression modeling approach [107] and a mixture model approach [107] with particular attention to their different modeling assumptions. The data set used was the acute leukemia data [73] which consists of 27 ALL samples and 11 AML samples, with the task being to find genes with differential expression between ALL and AML. After the experimental study, the author concluded that all three methods gave similar results in terms of the test statistics, but differed in the level of statistical significance and the numbers of genes identified. This author also pointed out that both *t*-test and the regression analysis are dependent on strong parametric assumptions for small sample sizes, which are likely violated in practice; whereas the mixture model approach can estimate the null distribution directly based on some reasonable assumptions. In contrast, the regression approach is flexible and can be easily extended to model more complex biological problems.

Troyanskaya *et al.* [109] addressed the problem of robust identification of differentially expressed genes from microarray data. These authors compared three model-free approaches: nonparametric *t*-test, the Wilcoxon or Mann–Whitney rank sum test, and a heuristic method based on high Pearson correlation to a perfectly differentially expressed gene. They evaluated the performance of each method based on simulated and biological data under varying noise levels and *p*-value cutoffs. They concluded that all methods exhibited very low false positive rates, and identified a large fraction of the differentially expressed genes in simulated data sets with noise level similar to that of actual data. The rank sum test appeared to be most conservative and may be of advantage when the computationally identified genes need to be tested biologically. However, if a more inclusive list of markers is desired, a higher *p*-value cutoff or the nonparametric *t*-test may be appropriate. Using the data from lung tumor and lymphoma data sets, the methods identified biologically relevant, differentially expressed genes that show clear separation between groups. Thus these methods provide a convenient and robust way to identify differentially expressed genes for further biological and clinical analysis.

Abul *et al.* [110] discussed the finding of differentially expressed genes in a microarray-based breast cancer data of BRCA1 versus BRCA2 tumor types [72]. These authors developed two methods, which are mainly based on the *q*-values approach. The first is a direct extension of the *q*-values approach, while the second uses two approaches: *q*-values and maximum-likelihood. They presented two algorithms for the second method, one for error minimization and the other for confidence bounding. These

authors demonstrated that the number of down-regulated genes in their data set was larger than the number of up-regulated genes using both methods, results consistent with those reported by Hedenfalk *et al.* [72] and Storey [111].

For further information, interested readers are encouraged to refer to the recent work by Pan [106], in which fair descriptions and comparisons of many statistical methods for the analysis of differentially expressed genes in replicated microarray data were reported; and by Troyanskaya *et al.* [109], in which several nonparametric methods for discriminator genes in microarray data were described and assessed.

11. SPECTRAL ANALYSIS OF MICROARRAY DATA

Kluger *et al.* [112] considered the coclustering of genes and experimental conditions in which genes are clustered if they exhibit similar expression patterns across conditions. Their proposed model can be reduced to the analysis of the same eigensystem derived in Dhillon's formulation for the problem of coclustering of words and documents [113]. To apply the Dhillon method to microarray data, these authors constructed a bipartite graph, where one set of nodes in this graph represents the genes, and the other represents experimental conditions. An arc between a gene and condition represents the level of over-expression or under-expression of this gene under this condition. They pointed out that the bipartite approach is limited in that it can only divide the genes and conditions into the same number of clusters, which is often impractical. Therefore these authors formulated a procedure that allows the number of gene clusters to be different from the number of condition clusters. The Dhillon's optimal partitioning eigenvector has a hybrid structure containing both gene and condition entries, whereas their approach searches for separate piecewise constant structure of the gene and corresponding sample eigenvectors. The authors applied their proposed spectral biclustering methods to five groups of cancer microarray data sets: lymphoma (microarray and Affymetrix), leukemia, breast cancer, and central nervous system embryonal tumors, and concluded that their spectral biclustering methods are superior to the Dhillon method.

Yeung *et al.* [114] introduced a spectral component analysis of time-series microarray data for the identification of genes that are subjected to common transcriptional regulation. Based on the motivation that the most commonly used approach to determine if the two genes have a potential regulatory relationship is to measure their expressional similarity using the Pearson correlation coefficient, but recognizing that this approach has many limitations, these authors instead proposed an autoregressive (AR)-based technique. In this work, they used the well-known AR modeling technique to characterize temporal gene expression data from the Spellman's a-synchronized yeast cell-cycle experiment [115], and time-series expression profiles were decomposed into spectral components and correlations between profiles computed. The full Spellman's a-synchronized data set consists of 18-point temporal mRNA level measurements sampled at every 7 minutes time interval for all 6178 ORFs in yeast. The test samples in the experiment were synchronized by a-factor method such that all the cells would be at the same stage in their cell cycle,

and the reported data are log ratios of the test sample expression over control sample expression level measurements. The authors used a subset studied by Filkov *et al.* [116], which contains 439 pairs of known transcriptional regulations of which 343 pairs are activations and 96 pairs are inhibitions, and comprising 288 involved genes. These authors reported that their method, when applied to data of known transcriptional regulations, was able to identify many of those missed by the traditional correlation method.

12. POTENTIAL SPECTRAL ANALYSIS OF MICROARRAY DATA

To date there are very few published research articles reporting the application of spectral analysis of microarray data. We believe that spectral methods with a particular reference to spectral distortion measures have potential application to the analysis of microarray data. Being analogous to speech signals, microarray data can be represented by the time series of spectral vectors, with the spectral difference or spectral distortion between the pair of spectra measured for the purpose of pattern comparison. Brazma and Vilo [2] pointed out, as we have already discussed in the previous sections, that there are two ways for the analysis of a microarray gene expression matrix. The first way is to compare expression profiles of genes by comparing the rows of the expression matrix; whereas the second way is to compare expression profiles of samples by comparing the columns of the expression matrix. The comparison of either rows or columns can be used to determine the similarities or dissimilarities between the data pairs. If two rows (genes) are found to be similar then it can be said that the respective genes are co-regulated and have similar functions. By comparing columns (samples), one can determine which genes are differentially expressed and then study the affects of various compounds on this expression. All of these involve the concept of pattern comparison in which the measures of similarity play the central role. Most measures applied to pattern comparison or cluster analysis are metric, distance or correlation based functions. There is rarely any research work on microarray data using the concept of distortion measures for pattern comparison. Basing on this motivation we briefly present herein the concept of spectral distortion measures, which have been well developed and successfully applied to the field of speech recognition [121], as a potential approach for the analysis of microarray data.

12.1. Linear Predictive Coding

Let $s(n)$, $n = 1, 2, \dots, N$ be a sequence of length N whose elements are represented by the microarray values of a particular row or column of the gene expression matrix. The estimated microarray value at position n , denoted by $\hat{s}(n)$, can be calculated as a linear combination of the past p microarray samples. This linear prediction can be expressed as

$$\hat{s}(n) = \sum_{k=1}^p a_k s(n-k) \quad (3)$$

where the terms $\{a_k\}$ are called the linear prediction coefficients.

The prediction error $e(n)$ between the observed sample $s(n)$ and the predicted value $\hat{s}(n)$ can be defined as

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^p a_k s(n-k) \quad (4)$$

The prediction coefficients $\{a_k\}$ can be optimally determined by minimizing the sum of squared errors

$$E = \sum_{n=1}^N e^2(n) = \sum_{n=1}^N \left[s(n) - \sum_{k=1}^p a_k s(n-k) \right]^2 \quad (5)$$

To solve (5) for the prediction coefficients, we differentiate E with respect to each a_k and equate the result to zero:

$$\frac{\partial E}{\partial a_k} = 0, \quad k = 1, \dots, p \quad (6)$$

The result is a set of p linear equations

$$\sum_{k=1}^p a_k r(m-k) = r(m), \quad m = 1, \dots, p \quad (7)$$

where $r(m)$ is the autocorrelation of $s(n)$, that is

$$r(m) = \sum_{n=1}^N s(n)s(n+m) \quad (8)$$

Equation (7) can be expressed in matrix form as

$$\mathbf{R}\mathbf{a} = \mathbf{r} \quad (9)$$

where \mathbf{R} is a $p \times p$ autocorrelation matrix, \mathbf{r} is a $p \times 1$ autocorrelation vector, and \mathbf{a} is a $p \times 1$ vector of prediction coefficients. Thus

$$\mathbf{a} = \mathbf{R}^{-1} \mathbf{r} \quad (10)$$

where \mathbf{R}^{-1} is the inverse of \mathbf{R}

12.2. LPC Cepstral Distortion Measure

Based the derivation of the linear predictive coefficients expressed in Eq. (10), a new similarity or dissimilarity measure between two microarray-data vectors can be computed. The calculation of vector similarity is based on various developments of distance and distortion measures. Before proceeding to the mathematical description of a distortion measure, we wish to point out the difference between distance and distortion functions, where the latter is more restricted in mathematical sense.

Let x , y and z be the vectors defined on a vector space V . A metric or distance d on V is defined as a real-valued function on the Cartesian product $V \times V$ if it has the following properties:

- (1) Positive definiteness: $0 \leq d(x, y) < \infty, x, y \in V$ and $d(x, y) = 0$ iff $x = y$;
- (2) Symmetry: $d(x, y) = d(y, x)$ for $x, y, z \in V$;
- (3) Triangle inequality: $d(x, y) \leq d(x, z) + d(z, y)$ for $x, y, z \in V$.

If a measure of dissimilarity satisfies only the property of positive definiteness, it is referred to as a distortion measure, which is considered very common for the vectorized representations of signal spectra [120]. In this sense, what

we will describe next is the mathematical measure of distortion, which relaxes the properties of symmetry and triangle inequality. From now on, the term d will be used to denote a distortion measure. There are several measures of distortion developed for speech recognition such as the Itakura-Saito distortion, the likelihood-ratio distortion, the log-likelihood-ratio distortion, and the LPC cepstral distortion measure [121]. However, the LPC cepstral distortion is the most widely used distortion measure for speech recognition.

Let $S(\omega)$ be the power spectrum (magnitude-squared Fourier transform) of a signal. The complex cepstrum of the signal is defined as the Fourier transform of the log of the signal spectrum:

$$\log S(\omega) = \sum c_n e^{-jn\omega} \quad (35)$$

where $c_n = -c_{-n}$ are real and referred to as the cepstral coefficients.

Consider $S(\omega)$ and $S'(\omega)$ to be the power spectra of the two (protein) signals and apply the Parseval's theorem [122], the L_2 -norm cepstral distance between $S(\omega)$ and $S'(\omega)$ can be related to the root-mean-square log spectral distance as [120]

$$\begin{aligned} d_2^2 &= \int_{-\pi}^{\pi} |\log S(\omega) - \log S'(\omega)|^2 \frac{d\omega}{2\pi} \\ &= \sum (c_n - c'_n)^2 \end{aligned} \quad (36)$$

where c_n and c'_n are the cepstral coefficients of $S(\omega)$ and $S'(\omega)$ respectively.

Since the cepstrum is a decaying sequence, the infinite number of terms in (36) can be truncated to some finite number $L \geq p$, that is

$$d^2(L) = \sum_{m=1}^L (c_m - c'_m)^2 \quad (37)$$

The cepstral coefficients can be directly derived from the LPC parameters using the following recursive procedure.

$$c_0 = \ln(G) \quad (38)$$

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad 1 \leq m \leq p \quad (39)$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad m > p \quad (40)$$

where G is the LPC gain, whose squared term is given as [117]

$$G^2 = r(0) - \sum_{k=1}^p a_k r(k) \quad (41)$$

Thus we have presented a potential spectral distortion measure as a potential approach for computing the similarity of time-series microarray data.

13. CONCLUSIONS

Gene microarray technology has become an integral component of the arsenal available to biologists for

addressing gene expression changes in a wide range of biological systems. This technology has advanced considerably in recent years to keep pace with the wealth of data being generated from genome sequencing and annotation projects.

One identifiable challenge for the future of this field that would considerably enhance its usefulness is to be better able to couple gene clustering capability with annotation information implying functional association between genes, or more accurately their expressed products. This would advance the goal of identifying changes in gene expression at the level of entire gene product networks, such as for metabolic and signaling pathways and functional protein complexes. From the biologist's perspective, this represents the next level of biological information needed to approach a paradigm of whole cell, and intercellular function. Developments in the microarray field have necessitated the introduction of technologies not conventionally associated with the biologist, and include those from disciplines such as engineering, mathematics, computer science, information technology, and information sciences. As discussed in this review, however, the utility of microarray data is still limited by the technology underpinning its capture and analysis. Thus, this field is highly amenable to the introduction of better analytical methodologies, and in this context we have introduced here the concept of spectral distortion measures as potentially useful approaches for improving on computing the similarities of microarrays or discriminating their differences in a different computational point of view. With advances in microarray technology new approaches for handling the new data will have yet to be identified.

In conclusion, microarray technology has represented the cornerstone for the analysis of gene expression in biology. To date, the promise of this technology has not been fully realized, but the wealth of genome information that this technology can potentially tap into both currently and in the future augurs well for the future of this technology and provides impetus for further developments in the mathematical and computer bases of this technology.

REFERENCES

- [1] Kellam P, Liu X, Experimental use of DNA arrays, In: Orenco CA, Jones DT, and Thornton JM Eds, *Bioinformatics: Genes, Proteins & Structures*, Bios, Oxford 2003.
- [2] Brazma A, Vilo J, Gene expression data analysis. *FEBS Lett* **2000**; 480:17-24.
- [3] Irizarry RA, Warren D, Spencer F, *et al.* Multiple-laboratory comparison of microarray platforms. *Nat Methods* **2005**; 2: 345-350.
- [4] Bammler T, Beyer RP, Bhattacharya S, *et al.* Standardizing global gene expression analysis between laboratories and across platforms. *Nat Methods* **2005**; 2: 351-356.
- [5] Mrowka R, Schuchhardt J, Gille C. Oligoddb – interactive design of oligo DNA for transcription profiling of human genes. *Bioinformatics* **2002**; 18: 1686-1687.
- [6] Li F, Stormo GD. Selection of optimal DNA oligos for gene expression arrays. *Bioinformatics* **2001**; 17: 1067-1076.
- [7] C, Carta R, Zhang L. Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic Acids Res* **2005**; 33: e84.
- [8] Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. *Nat Genet* **1999**; 21 (1 Suppl): 20-24.
- [9] Kuhn K, Baker SC, Chudin E, *et al.* A novel, high-performance random array platform for quantitative gene expression profiling. *Genome Res* **2004**; 14: 2347-2356.

- [10] Boguski MS, Schuler GD. ESTablishing a human transcript map. *Nat Genet* **1995**; 10:369-371.
- [11] Quackenbush J, Liang F, Holt I, Perlea G, Upton J. The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res* **2000**; 28: 141-145.
- [12] Murray CG, Larsson TP, Hill T, Bjorklind R, Fredriksson R, Schioth HB. Evaluation of EST-data using the genome assembly. *Biochem Biophys Res Commun* **2005**; 331: 1566-1576.
- [13] Okazaki Y, Furuno M, Kasukawa T, et al. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **2002**; 420: 563-573.
- [14] Strausberg RL, Feingold EA, Grouse LH, et al. Mammalian Gene Collection Program Team. Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences. *Proc Natl Acad Sci USA* **2002**; 99: 16899-16903
- [15] Katayama S, Wells C, Lipovich L, Engström, PG, and the FANTOM3 consortium. Antisense transcription in the mammalian transcriptome (submitted).
- [16] Holmes R, Williamson C, Peters J, Denny P, Wells C. A comprehensive transcript map of the mouse Gnas imprinted complex. *Genome Res* **2003**; 13: 1410-1415.
- [17] Roche FM, Hokamp K, Acab M, Babiuk LA, Hancock RE, Brinkman FS. ProbeLynx: a tool for updating the association of microarray probes to genes. *Nucleic Acids Res* **2004**; 32 (Web Server issue):W471-474.
- [18] Mecham BH, Klus GT, Strovel J, et al. Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. *Nucleic Acids Res* **2004**; 32: e74.
- [19] Kanapin A, Batalov S, Davis MJ, et al. RIKEN GER Group, GSL Members. Mouse proteome analysis. *Genome Res.* **2003**; 13: 1335-1344.
- [20] Gough J. The SUPERFAMILY database in structural genomics. *Acta Crystallogr D Biol Crystallogr* **2002**; 58: 1897-1900.
- [21] van Vliet C, Thomas EC, Merino-Trigo A, Teasdale RD, Gleeson PA. Intracellular sorting and transport of proteins. *Prog Biophys Mol Biol* **2003**; 83:1-45.
- [22] Grimmond SM, Miranda KC, Yuan Z, et al. The mouse secretome: functional classification of the proteins secreted into the extracellular environment. *Genome Res* **2003**; 13:1350-1359.
- [23] Pang KC, Stephen S, Engstrom PG, et al. RNAdb – a comprehensive mammalian noncoding RNA database. *Nucleic Acids Res* **2005**; 33(Database issue): D125-130.
- [24] Dennis G Jr, Sherman BT, Hosack DA, et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **2003**; 4:P3.
- [25] Diehn M, Sherlock G, Binkley G, et al. SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res* **2003**; 31:219-223.
- [26] Tsai J, Sultana R, Lee Y, et al. Resourcerer: A database for annotating and linking microarray resources within and across species. *Genome Biology* **2001**; 2: software 0002.1-2.4.
- [27] Camon E, Magrane M, Barrell D, et al. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res* **2004**; 32 (Database issue):D262-266.
- [28] Kanehisa M. The KEGG database. *Novartis Found Symp* **2002**; 247:91-101, discussion 101-103, 119-128, 244-252.
- [29] Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **2000**; 28: 27-30
- [30] Grimmond S, Van Hateren N, Siggers P, et al. Sexually dimorphic expression of protease nexin-1 and vanin-1 in the developing mouse gonad prior to overt differentiation suggests a role in mammalian sexual development. *Hum Mol Genet* **2000**; 9: 1553-1560.
- [31] Su AI, Wiltshire T, Batalov S, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* **2004**; 101: 6062-6067.
- [32] Bono H, Kasukawa T, Hayashizaki Y, Okazaki Y. READ: RIKEN Expression Array Database. *Nucleic Acids Res* **2002**; 30:211-213.
- [33] Zhang W, Morris QD, Chang R, et al. The functional landscape of mouse gene expression. *J Biol* **2004**; 3: 21.
- [34] Liu ET Classification of cancers by expression profiling *Curr Opin Genet Dev* **2003**; 13: 97-103
- [35] Kim RD, Park PJ. Improving identification of differentially expressed genes in microarray studies using information from public databases. *Genome Biol* **2004**; 5: R70.
- [36] Brody JP, Williams BA, Wold BJ, Quake SR. Significance and statistical errors in the analysis of DNA microarray data. *Proc Natl Acad Sci USA* **2002**; 99: 12975-12978.
- [37] Li C, Hung Wong W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. *Genome Biol* **2001**; 2: RESEARCH0032.
- [38] Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **2003**; 31: e15.
- [39] Smyth GK, Speed T. Normalization of cDNA microarray data. *Methods* **2003**; 31: 265-273.
- [40] Kooperberg C, Fazio TG, Delrow JJ, Tsukiyama T. Improved background correction for spotted DNA microarrays. *J Computational Biology* **2002**; 9:55-66.
- [41] Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J Biomed Optics* **1997**; 2:364-367.
- [42] Theilhaber J, Bushnell S, Fuchs R. Bayesian estimation of fold-changes in the analysis of gene expression: The PFOLD algorithm. *Nature Genet* **1999**; 23:78.
- [43] Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. On differential variability of expression ratios: Improving statistics inference about gene expression changes from microarray data. *J Comp Biol* **2000**; 8:37-52.
- [44] Jung H-Y, Cho H-G. An automatic block and spot indexing with k -nearest neighbors graph for microarray image analysis. *Bioinformatics* **2002**; 18: S141-S151.
- [45] Steinfath M, Wruck W, Seidel H, Lehrach H, Radelof U, O'Brien J. Automated image analysis for array hybridization experiments. *Bioinformatics* **2001**; 17: 634-641.
- [46] Jain AN, Tokuyasu TA, Snijders AM, Segraves R, Albertson DG, Pinkel D. Fully automated quantification of microarray image data. *Genome Res.* **2002**; 12: 325-332.
- [47] Buhler J, Ideker T, Haynor D Dapple. Improved techniques for finding spots on DNA microarrays. Technical Report UWTR 2000-08-05. University of Washington.
- [48] Eisen MB. ScanAlyse **1999** (<http://rana.stanford.edu/software/>).
- [49] Axon Instruments Inc. *GenePix 4000A User's Guide* **1999**.
- [50] GSI Luminomics, QuantArray Analysis Software, *Operator's Manual* **1999**.
- [51] Vincent L, Soille P. Watersheds in digital spaces: an efficient algorithm based on immersion simulations. *IEEE Trans Patt Anal Mach Intel* **1991**; 13:583-598.
- [52] Beucher S, Meyer F. The morphological approach to segmentation: the watershed transformation. In *Mathematical Morphology in Image Processing*, Volume 34th of Optics Engineering, chapter 12, Marcel Dekker, New York 1992; pp.433-481.
- [53] Adams R, Bischof L. Seeded region growing. *IEEE Trans Pattern Analysis Machine Intelligence* **1994**; 16: 641-647.
- [54] Buckley JM. *The spot user's guide*. CSIRO Mathematical and Information Sciences **2000**, <http://www.cmis.csiro.au/IAP/Spot/spotmanual.htm>.
- [55] Bozinov D, Rahnenfuhrer J. Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering. *Bioinformatics* **2002**; 18: 747-756.
- [56] Hirata R, Barrera J, Hashimoto RF, Dantas DO, Esteves GH. Segmentation of microarray images by mathematical morphology. *Real-Time Imaging* **2002**; 8: 491-505.
- [57] Angulo J, and Serra J. Automatic analysis of DNA microarray images using mathematical morphology. *Bioinformatics* **2003**; 19:553-562.
- [58] Axon Instruments Inc. *GenePix Pro 4.0, Documentation* 2002.
- [59] Liew AW-C, Yan H, Yang M. Robust adaptive spot segmentation of DNA microarray images. *Pattern Recognition* **2003**; 36:1251-1254.
- [60] Axon Instruments Inc. *GenePix Pro 3.0, Technical manual* **2001**.
- [61] Damiance APG, Zhao L, Carvalho ACPLF. A dynamical model with adaptive pixel moving for microarray images segmentation. *Real-Time Imaging* **2004**; 10:189-195.
- [62] Glasbey CA, Ghazal P. Combinatorial image analysis of DNA microarray features. *Bioinformatics* **2003**; 19:194-203.

- [63] Lonardi S, Luo Y. Gridding and compression of microarray images. *Proc 2004 IEEE Computational Systems Bioinformatics Conf (CSB 2004)*.
- [64] Alizadeh AA, Eisen MB, Davis RE, *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **2000**; 403: 503-511.
- [65] Troyanskaya O, Cantor M, Sherlock G, *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001**; 17: 520-525.
- [66] Bø TH, Dysvik B, Jonassen I. LSImpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Res* **2004**; 32:3 e34.
- [67] Nguyen DV, Wang N, Carroll RJ. Evaluation of missing value estimation for microarray data. *J Data Sci* **2004**; 2: 347-370.
- [68] Kim H, Golub GH, Park H. Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics* **2005**; 21: 187-198.
- [69] Oba S, Sata M, Takesama I, Monden M, Matsubara K, Ishii S. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* **2003**; 19: 2088-2096.
- [70] Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics* **2002**; 18:39-50.
- [71] Zhou X, Liu K-Y, Wong STC. Cancer classification and prediction using logistic regression with Bayesian gene selection. *J Biomedical Informatics* **2004**; 37:249-259.
- [72] Hedenfalk I, Duggan D, Chen Y, *et al.* Gene expression profiles in hereditary breast cancer. *New Eng J Med* **2001**; 344: 539-548.
- [73] Golub TR, Slonim DK, Tamayo P, *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **1999**; 286:531-537.
- [74] Statnikov A, Aliferis CF, Tsamardinos I, Hardin D, Levy S. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* **2005**; 21: 631-643.
- [75] Armstrong SA, Staunton JE, Silverman LB, *et al.* MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat Genet* **2002**; 30: 41-47.
- [76] Bhattacharjee A, Richards WG, Staunton J, *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci USA* **2001**; 98: 13790-13795.
- [77] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc* **2002**; 97: 77-87.
- [78] Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* **2000**; 16: 906-914.
- [79] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning* **2002**; 46: 389-422.
- [80] Khan J, Wei JS, Ringner M, *et al.* Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* **2001**; 7: 673-679.
- [81] Lee Y, Lee C-K. Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics* **2003**; 19: 1132-1139.
- [82] Mukherjee S. *Classifying Microarray Data Using Support Vector Machines, Understanding And Using Microarray Analysis Techniques: A Practical Guide*. Kluwer Academic Publishers, Boston, MA, 2003.
- [83] Nutt CL, Mani DR, Betensky RA, *et al.* Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res* **2003**; 63: 1602-1607.
- [84] Pomeroy SL, Tamayo P, Gaasenbeek M, *et al.* Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **2002**; 415: 436-442.
- [85] Ramaswamy S, Tamayo P, Rifkin R, *et al.* Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci USA* **2001**; 98: 15149-15154.
- [86] Romualdi C, Campanaro S, Campagna D, *et al.* Pattern recognition in gene expression profiling using DNA array: a comparative study of different statistical methods applied to cancer classification. *Hum Mol Genet* **2003**; 12: 823-836.
- [87] Singh D, Febbo PG, Ross K, *et al.* Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell* **2002**; 203-209.
- [88] Shipp MA, Ross KN, Tamayo P, *et al.* Diffuse large B-cell lymphoma outcome prediction by gene expression profiling and supervised machine learning. *Nat Med* **2002**; 8: 68-74.
- [89] Welsh JB, Sapinoso LM, Kern SG, *et al.* Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res* **2001**; 1: 7388-7393.
- [90] Pham TD. An optimally weighted fuzzy *k*-NN algorithm. Singh S, Singh M, Apte C, Perner P. (Eds.): ICAPR 2005, Lecture Notes in Computer Science 2005; 3686: 239-247.
- [91] Duda R, Hart P. *Pattern Classification and Scene Analysis*. John Wiley & Sons, New York, 1973.
- [92] Keller JM, Gray MR, Givens JA. A fuzzy *k*-nearest neighbor algorithm. *IEEE Trans Systems, Man and Cybernetics* **1985**; 15: 580-585.
- [93] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* **1998**; 95: 14863-14868.
- [94] Dougherty ER, Barrera J, Brun M, *et al.* Inference from clustering with application to gene-expression microarrays. *J Computational Biology* **2002**; 9: 105-126.
- [95] Dembélé D, Kastner P. Fuzzy *C*-means method for clustering microarray data. *Bioinformatics* **2003**; 19: 973-980.
- [96] Bezdek JC. *Pattern Recognition With Fuzzy Objective Function Algorithms*. Plenum Press, New York, 1981.
- [97] Sharan R, Shamir R. CLICK: a clustering algorithm with application to gene expression analysis. *Proceedings of AAAI-ISMB 2000*; pp.307-316.
- [98] Cho RJ, Campbell MJ, Winzeler EA, *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* **1998**; 2: 65-73.
- [99] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. *Nat Genet* **1999**; 22: 281-285.
- [100] Iyer VR, Eisen MB, Ross DT, *et al.* The transcriptional program in the response of human fibroblast to serum. *Science* **1999**; 283: 83-87.
- [101] Asyali MH, Alci M. Reliability analysis of microarray data using fuzzy *c*-means and normal mixture modeling based classification methods. *Bioinformatics* **2005**; 21:644-649.
- [102] Suzuki T, Hashimoto S, Toyoda N, *et al.* Comprehensive gene expression profile of LPS-stimulated human monocytes by SAGE. *Blood* **2000**; 96: 2584-2591.
- [103] Murayama T, Ohara Y, Obuchi M, *et al.* Human cytomegalovirus induces interleukin-8 production by a human monocytic cell line, THP-1, through acting concurrently on AP-1- and NF-kappaB-binding sites of the interleukin-8 gene. *J Virol*. **1997**; 71: 5692-5695.
- [104] Frevel MA, Bakheet T, Silva AM, Hissong JG, Khabar KS, Williams BR. p38 Mitogen-activated protein kinase-dependent and -independent signaling of mRNA stability of AU-rich element-containing transcripts. *Mol Cell Biol* **2003**; 23: 425-436.
- [105] Chang HY, Sneddon JB, Alizadeh AA, *et al.* Gene expression signature of fibroblast serum response predicts human cancer progression: similarities between tumors and wounds. *PLoS Biol* **2004**; 2: E7.
- [106] Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* **2002**; 18: 546-554;
- [107] Thomas JG, Olson JM, Tapscott SJ, Zhao LP. An efficient and robust statistical approach to discover differentially expressed genes using genomic expression profiles. *Genome Res* **2001**; 11:1227-1236.
- [108] Pan W, Lin J, Le C. A mixture model approach to detecting differentially expressed genes with microarray data. *Technical Report*, Division of Biostatistics, University of Minnesota 2001.
- [109] Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics* **2002**; 18: 1454-1461.
- [110] Abul O, Alhaji R, Polat F, Barker K. Finding differentially expressed genes for pattern generation. *Bioinformatics* **2005**; 21: 445-450
- [111] Storey JD. False discovery rates: theory and applications to DNA microarrays. *PhD Thesis*, Department of Statistics, Stanford University 2002.

- [112] Kluger Y, Basri R, Chang JT, Gerstein M. Spectral Biclustering of Microarray Data: Co-clustering Genes and Conditions. *Genome Research* **2003**; 13:703-716.
- [113] Dhillon IS. Co-clustering documents and words using bipartite spectral graph partitioning. Proc. *Seventh ACM 2001, Special Interest Group on Knowledge Discovery in Data and Data Mining Conference*, San Francisco, CA.
- [114] Yeung LK, Szeto LK, Liew AW-C, Yan H. Dominant spectral component analysis for transcriptional regulations using microarray time-series data. *Bioinformatics* **2004**; 20: 742-749.
- [115] Spellman P, Sherlock G, Zhang M, *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* **1998**; 9: 3273-3297.
- [116] Filkov V, Skiena S, Zhi J. Analysis techniques for microarray time-series data. *J Comput Biol* **2002**; 9:317-330.
- [117] Ingle VK, Proakis JG. *Digital Signal Processing Using Matlab V.4*. Boston, PWS Publishing, 1997.
- [118] Isaaks EH, Srivastava RM. *An Introduction to Applied Geostatistics*. Oxford University Press, New York 1989.
- [119] Journel AG, Rao SE, Deriving conditional distribution from ordinary kriging. *Stanford Center for Reservoir Forecasting. Stanford University Report No. 29* (1996), 25p.
- [120] L. Rabiner, B.H. Juang. *Fundamentals of Speech Recognition*. New Jersey, Prentice Hall 1993.
- [121] Nocerino N, Soong FK, Rabiner LR, Klatt DH. Comparative study of several distortion measures for speech recognition. *IEEE Proc Int Conf Acoustics, Speech and Signal Processing* **1985**; 11.4.1, pp. 387-390.
- [122] O'Shaughnessy D. *Speech Communication – Human and Machine*. Reading, Massachusetts, Addison-Wesley, 1987.
- [123] Pham TD. Integration of fuzzy and geostatistical models for estimating missing multivariate observations. *WSEAS Trans Systems* **2005**; 4: 233-237.
- [124] Journel AG, Huibregts CJ. *Mining Geostatistics*. Academic Press, London, 1978.
- [125] Sturn A, Quackenbush J, Trajanoski Z. Genesis: cluster analysis of microarray data. *Bioinformatics* **2002**; 18: 207-208.
- [126] Li H, Zhang K, Jiang T. Minimum entropy clustering and applications to gene expression analysis. In *Proceedings of the 3rd IEEE Computational Systems. Bioinformatics Conference*, pages 142 - 151. Stanford, CA, 2004.
- [127] Cover TM, Thomas JA. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.
- [128] Havrda J, Charvat F. Quantification method of classification processes: Concept of structural α -entropy. *Kybernetika* **1967**; 3:30-35.
- [129] Rand WM. Objective criteria for evaluation of clustering methods. *J Am Statist Assoc* **1971**; 66: 846-850.