Available online at www.sciencedirect.com

SCIENCE DIRECT®

Pattern Recognition 38 (2005) 2055–2073

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Pattern recognition techniques for the emerging field of bioinformatics: A review

Alan Wee-Chung Liew[a,*], Hong Yan[b,c], Mengsu Yang[d]

[a]Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong
[b]Department of Computer Engineering and Information Technology, City University of Hong Kong, Kowloon, Hong Kong
[c]School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia
[d]Department of Chemistry and Biology, City University of Hong Kong, Kowloon, Hong Kong

Received 8 October 2004; received in revised form 28 February 2005; accepted 28 February 2005

Abstract

The emerging field of bioinformatics has recently created much interest in the computer science and engineering communities. With the wealth of sequence data in many public online databases and the huge amount of data generated from the Human Genome Project, computer analysis has become indispensable. This calls for novel algorithms and opens up new areas of applications for many pattern recognition techniques. In this article, we review two major avenues of research in bioinformatics, namely DNA sequence analysis and DNA microarray data analysis. In DNA sequence analysis, we focus on the topics of sequence comparison and gene recognition. For DNA microarray data analysis, we discuss key issues such as image analysis for gene expression data extraction, data pre-processing, clustering analysis for pattern discovery and gene expression time series data analysis. We describe current methods and show how computational techniques could be useful in these areas. It is our hope that this review article could demonstrate how the pattern recognition community could have an impact on the fascinating and challenging area of genomic research.

Keywords: Bioinformatics; DNA sequence analysis; DNA sequence comparison; Gene recognition; DNA microarray; Gene expression data analysis; Gene expression clustering; Gene expression time series; Gene regulation

1. Introduction

Recent advancement in molecular biology and genomic research, such as high throughput sequencing methods and cDNA microarray technology, has generated an unprecedented amount of data. The completion of the Human Genome Project in sequencing the complete human genome has also spurred great interest in the research community to utilize such wealth of information in different areas of biological and medical sciences. Efficient analysis of this massive amount of data by computational methods is fast becoming a major challenge [1–3].

Two technologies that play dominant roles in elucidating the relations, structure and function of genes are DNA sequence analysis and DNA microarray data analysis. DNA sequence analysis has been around for over two decades, even before the availability of mass scale sequencing techniques. Nevertheless, it has regained momentum in recent years due to the advent of fast computers and algorithms, and the availability of up-to-date, public domain online databases holding massive amount of sequence data (see Table 1). These databases also enable researchers to share their works or to access the works of others in the most up to date manner.

[*] Corresponding author. Tel.: +852 2609 8419; fax: +852 2603 5024.
E-mail address: wcliew@cse.cuhk.edu.hk (A.W.-C. Liew).

0031-3203/$30.00 © 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2005.02.019

Table 1
Three major public domain online DNA sequence databases

(1)	EMBL (http://www.ebi.ac.uk/embl/index.html) The EMBL database is maintained by the European Bioinformatics Institute (EBI) and is Europe's primary collection of nucleotide sequences. The current release is EMBL Release 82 (24 February 2005), which contains 49.5 million sequence entries comprising 85.1 billion nucleotides.
(2)	GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) The GenBank database is maintained by the National Center for Biotechnology Information (NCBI), USA. The current release is Release 147 (20 April 2005), and contains similar number of sequence entry and nucleotide bases as in EMBL.
(3)	DDBJ (http://www.ddbj.nig.ac.jp/Welcome-e.html) The DDBJ database is maintained by DNA Data Bank of Japan. The current release in DDBJ is Release 61 (28 April 2005), which contains 43.12 million sequence entries comprising 47.1 billion nucleotides.

In the first part of this review article, we present an overview on some of the major research areas in DNA sequence analysis. To set the scene, we first give a brief description of the biology background required for a proper understanding of the material. Next, we describe a very important area in DNA sequence analysis, namely, sequence comparison. When a molecular biologist is presented with an unknown DNA sequence, his/her first task would be to search for similar annotated sequences in the major public sequence databases. Doing so would allow the biologist to make use of the prior knowledge accumulated through the efforts of many researchers to infer the possible function or structure of the unknown sequence, which in term lead to more specific and targeted analysis or experimentation later on. The other problem of DNA sequence analysis we describe is gene prediction, which has been an area of active research in bioinformatics in recent years. The basic idea in gene prediction is to look for characteristic features that are present in the coding regions of a DNA sequence. One well-known characteristic of coding regions is the 3-base periodicity due to the unequal usage of the codons. Many coding measures that aim to detect such periodicity have been proposed. Some of the most successful gene prediction algorithms are based on classical signal processing techniques such as the discrete Fourier transform, as well as the detection of specific sub-sequence patterns in the upstream and down stream regulatory regions.

In the second part of this review article, we present an overview of the DNA microarray technology and gene expression analysis. DNA Microarray technology, which allows massively parallel, high throughput profiling of gene expression in a single hybridization experiment, has recently emerged as a powerful tool for genomic research. A critical aspect of DNA microarray technology is to extract expression data from the microarray images accurately. Due to the nature of the microarray images, innovative image processing techniques are required to locate the spots in the images and to measure the resulting expression ratio reliably. We outline a robust method of spot extraction, and describe how data pre-processing is performed on the expression data. Once the expression data are obtained, they usually undergo cluster analysis to detect groups of genes that are similarly expressed. Some of our recent works on clustering of gene expression data are outlined. Very often, one is interested in seeing the dynamic behavior of genes under different stimuli or environment variables. Whole-genome expression time-series data can describe a dynamic biological process such as the cell cycle or metabolic process. They allow one to determine the causal relationships between the expressions of different genes and to infer gene regulatory information in a cell. We outline various approaches for studying time series expression data. Our recent work on time series expression data analysis using autoregressive (AR) modeling for frequency spectral estimation indicates that many useful and biologically relevant regulatory information that are otherwise hidden could be uncovered.

2. DNA sequence analysis

2.1. Biology background

DNA is the basis of heredity. It is a polymer made up of small molecules called nucleotides, which can be distinguished by the four bases: adenine (A), cytosine (C), guanine (G) and thymine (T). A DNA sequence is therefore specified completely by a sequence consisting of the four alphabets {A, C, G, T}. DNA usually occurs in double strands, and the bases in the two strands are complementary to each other, i.e., A pairing with T and G pairing with C by hydrogen bonds. The double-stranded DNA forms the well-known double helix in space (see Fig. 1). The pairing mechanism allows one strand of DNA to serve as template for producing the reverse complement strand, thus explaining how DNA can duplicate.

DNA carries the genetic information required by an organism to function. The flow of information within a cell is summarized by the diagram in Fig. 2. In the schematic, we see that the intermediate step from DNA to protein synthesis is the process called transcription. Transcription copies information in the DNA into copies called RNA. If a segment in the DNA sequence encodes a protein (corresponds to a coding region in the DNA sequence), the RNA is called a messenger, or mRNA. In transcription, the DNA nucleotides A, C, G, T are respectively transcribed into RNA nucleotides U (uracil (U) replaces thymine (T) in RNA molecules), G, C, A. The final step of information flow is the translation process. The information encoded in the mRNA is used

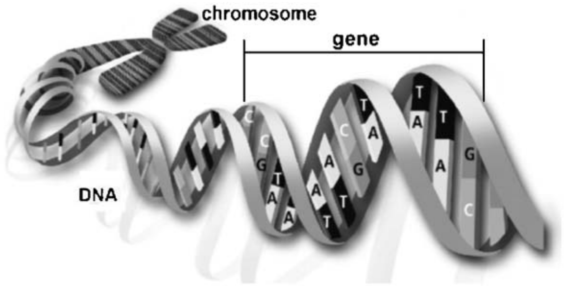

Fig. 1. The double helix of DNA sequence with a gene in the sequence delimited. Genes are specific sequences of bases that encode instructions on how to make proteins. (Courtesy of U.S. Department of Energy Human Genome Program, http://www.ornl.gov/hgmis).

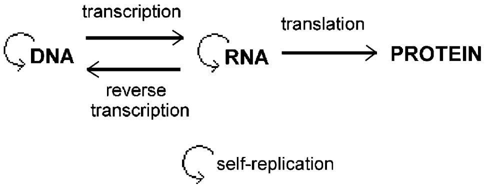

Fig. 2. Flow of information within a cell.

to specify the precise ordering of the amino acids, which form proteins. Proteins are polypeptide chains composed of an alphabet of 20 different amino acids. The genetic code is a triplet bases code, where successive *codons* consisting of three successive RNA nucleotides encode one of the 20 amino acids or the signal to stop translation.

The transcription process is different in prokaryotes (i.e., simple bacteria) and eukaryotes (non-bacteria, possess a nucleus, e.g., fungi, unicellular paramecia, all plants and animals). In prokaryotes, the RNA polymerase produces an mRNA transcript directly from the DNA template. In eukaryotes, genes in a DNA sequence are not continuous, but instead are broken up into coding regions (exons, which code for proteins) and non-coding regions (introns). The RNA is transcribed in the nucleus and then undergoes post-transcriptional modification (i.e., pre-mRNA splicing), where the introns are spliced out and the remaining exons are joined to form the final mRNA (see Fig. 3), which is then used for protein synthesis during translation. Thus, embedded within the DNA sequence are specific sub-sequences that control the initiation or termination of transcription. These sub-sequences, such as promoters, enhancers, silencers, terminators, are regulators of gene expression. Other sequences of interest within eukaryote DNA sequence are coding regions (exons), non-coding regions (introns and intergenic regions), splice signals or splice sites, and the location of the open reading frames (ORFs).

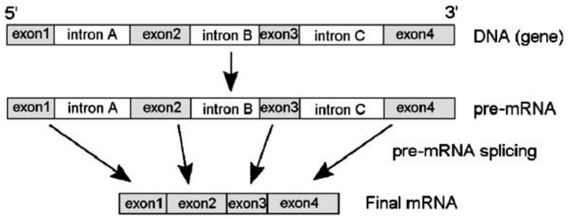

Fig. 3. Transcription process in eukaryotes cell. With alternative splicing, some exons may be omitted when forming the final mRNA.

2.2. Sequence comparison

Given a new DNA sequence, one would like to study the functional and structural information encoded in the sequence. In general, the first step taken by a biologist would be to compare the new sequence with sequences that are already well studied and annotated. Sequences that are similar would probably have the same function, in terms of a functional role (i.e., ORFs coding for similar proteins), regulatory role (i.e., similar regulatory or biochemical pathways) or structural properties in the case of proteins. Additionally, if two sequences from different organisms are similar, there may be a common ancestor sequence, and the sequences are then said to be homologous. Relationship between homologous sequences has important implications in speciation study and phylogenetic analysis.

One method for sequence comparison is by sequence alignment. This is similar to the string matching problem that has been studied extensively in pattern recognition. Sequence alignment is the procedure of comparing two (pairwise alignment) or more (multiple sequence alignment) sequences by searching for a series of individual characters or character patterns that are in the same order in the sequences. The standard pairwise alignment method is based on dynamic programming [4–6]. The method compares every pairs of characters in the two sequences and generates an alignment and a score, which is dependent on the scoring scheme used (i.e., a scoring matrix for the different base–pair combinations, match and mismatch scores, and a scheme for insertion or deletion (gap) penalties).

Although dynamic programming for sequence alignment is mathematically optimal, it is far too slow for comparing a large number of bases. Typical DNA database today contains billions of bases, and the number is still increasing rapidly. To enable sequence search and comparison to be performed in a reasonable time, fast heuristic local alignment algorithms have been developed. The trade-off in mathematical optimality in this case is more than compensated for by the gain in speed and efficiency. The most widely used heuristic database search tool is *BLAST* [7,8] and is freely available in many websites around the world, such as NCBI (National Center for Biotechnology Information, http://www.ncbi.nlm.nih.gov/BLAST) and the EBI

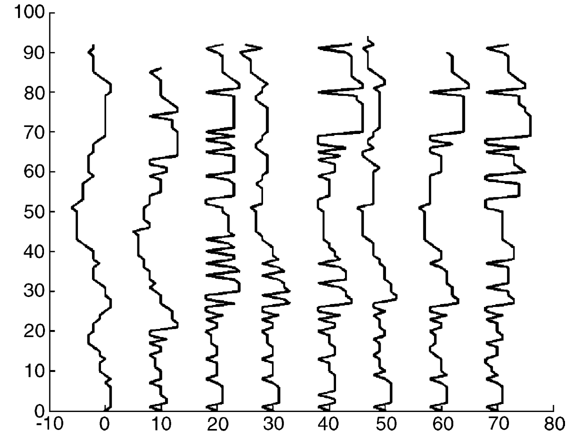

Fig. 4. The AC DB-curve of the DNA sequences of the first exons of beta-globin genes for eight different species (Human, Goat, Opossum, Gallus, Lemur, Mouse, Rabbit, and Rat).

(European Bioinformatics Institute, http://www.ebi.ad.uk/blastall). Many variants of *BLAST* have been developed to search for different type of databases and for different applications. *BLAST* has now become the standard for sequence alignment and database searching.

For relatively short sequences, sequence comparison can also be done visually by sequence visualization techniques. One such 2D visualization technique is the DB-curve (dual-base curve) [9]. In DB-curve DNA sequence visualization, two out of the four bases are considered at a time. The two bases are assigned a +45° and −45° vectors, respectively, whereas the remaining bases are assigned a +90° vector. Fig. 4 shows the AC DB-curve of the DNA sequences of the first exons of beta-globin genes for eight different species. Similarities and differences in the sequences can be readily observed from the plots.

Although most sequence comparison methods to date are based on string matching, other comparison techniques based on more traditional signal processing approach are possible. The later methods usually require the string of alphabets of DNA sequence be first mapped to a numeric sequence prior to analysis. Signal processing techniques such as the fast Fourier transform (FFT) and correlation could then be applied to the numeric sequences [10–12]. Recently, multiresolution signal processing techniques such as wavelet transform have started to be applied to DNA sequences too [13,14].

The DNA sequence comparison problems are described here as a sort of string alignment and matching problems; a problem well-studied in the PR community. However, comparison and alignment of biological sequences is different from the general string matching problem in computer science in at least several aspects. First, the number of bases to be compared can be huge. For example, just for the human genome, the number of nucleotide bases is around 3×10^9. This places huge demand on the computational efficiency and speed of the processing algorithms. Secondly, the percentage identity between two sequences can vary greatly; a score of 30% identity can still be considered biologically relevant. Thirdly, one should have some knowledge about the biological nature of the problem to be solved. A result that is mathematically sound may be highly implausible and might not reflect what is known about the biological process. For example, consider the problem of alignment of two protein-coding DNA sequences. It is not very sensible to align the DNA sequences of protein-coding genes. Instead, it is much more sensible to translate the sequences to their corresponding amino acid sequences and then put the gaps into the DNA sequence alignment according to where they are found in the amino acid alignment. To illustrate, consider the alignment of two protein-coding DNA sequences ATGCTGTTAGGG and ATGCTCGTAGGG. An alignment algorithm might give the solution below as the preferred alignment:

A T G C T G T T A G G G

A T G C T C G T A G G G

However, the alternative alignment below, although less mathematically optimal (i.e., with a smaller similarity score due to the penalties imposed for gaps), may be much more plausible biologically:

A T G C T - G T T A G G G

A T G C T C G T - A G G G

2.3. Gene prediction

Gene prediction, which is also widely known as gene recognition, has been an area of active research in bioinformatics [15]. In prokaryotes, gene finding is made simpler by the fact that coding regions are not interrupted by intervening sequences such as introns. Still, especially for short open reading frames, it is highly non-trivial to distinguish between sequences that represent true genes and those that do not. Eukaryotic gene typically consists of exons interrupted by non-coding regions such as introns or intergenic regions. Prediction of eukaryotic gene is therefore a much more difficult problem.

One way to analyze a sequence for regions of high coding potential is by an examination of various coding statistics. A coding statistic describes the likelihood that a DNA sequence is coding for a protein. Such approach has the advantage that no similar sequence is needed as the information to predict the protein coding genes in the sequence is mined from the sequence itself. Many coding statistics have been proposed by various researchers. Some of these coding statistics are: codon usage bias, base compositional bias between codon positions and periodicity in base occurrence [16–19]. It is clear that the sensitivity and the accuracy of the prediction depend on the statistics used. In [22], we have performed a statistical study on the effective coding features for coding/non-coding DNA sequence classification for yeast, C. elegans and human. A total of 22 features, divided into six groups, are analyzed. We compared the discriminative power of the 22 coding features based on their information content. We observed that the information content of different coding features vary greatly for the three species and not all features are equally effective for different species. We found that features that are effective for yeast and C. elegans are generally not very effective for human and vice versa. The study indicated that a careful selection of coding features tailored to the species of interest is important to ensure good classification performance.

Recognition of coding regions or ORFs in human genome based on coding statistics is a difficult problem due to the short exon length, where the average length of exons of vertebrate gene is only 137 bp [23]. Although good recognition rate can be achieved in the recognition of coding and non-coding regions in yeast genome (> 95% accuracy) [24], the strengths of the statistical features alone are generally not sufficient to identify human exons due to their limited average length [20]. In Refs. [21,22], we are able to identify a small subset of features that has high discriminative power (with classification accuracy of up to 90% for human) while at the same time is complementary in their information content.

It is well known that coding regions in a DNA sequence usually exhibit a characteristic 3-base periodicity. It results from the fact that coding sequences consist of codons and these codons are not equally used (see, for example, the relative frequency of codon usage in Homo sapiens given in http://www.kazusa.or.jp/codon/). This characteristic has been exploited in the recognition of coding regions in a DNA sequence using spectral analysis [25–27]. Given a DNA sequence consisting of an alphabet of four characters {A, T, G, C}. Let $u_A(n)$, $u_T(n)$, $u_G(n)$, $u_C(n)$ be the binary indicator sequences which take the value of either 1 or 0 at location n, depending on whether the corresponding characters exist at location n. Then the discrete Fourier transform (DFT) sequences $U_A(k)$, $U_T(k)$, $U_G(k)$ and $U_C(k)$ provide a four-dimensional representation of the frequency spectrum of the DNA string. In the spectral analysis of DNA sequence, the 3-base periodicity due to codon usage bias will show up as a distinct peak at the frequency index $k = N/3$, where N is the length of the sequence. Usually no such peak is apparent for non-coding sequences. If we define the following normalized DFT coefficients at $k = N/3$:

$$A = \frac{1}{N}U_A(N/3), \quad T = \frac{1}{N}U_T(N/3),$$
$$G = \frac{1}{N}U_G(N/3), \quad C = \frac{1}{N}U_C(N/3). \tag{1}$$

Then the quantity $|A|^2 + |T|^2 + |G|^2 + |C|^2$, computed over a sliding window, can be used as an exon predictor in a DNA sequence [25,27]. In [26], Anastassiou used the optimized measure $W = |aA + tT + gG + cC|^2$ as a superior predictor of exons, where the weights a, t, g, c are optimized with respect to some cost function that maximally separate the coding sequences from the non-coding sequences in the training set. Fig. 5 shows the quantity W of the 351-point DFT for a DNA stretch of C. elegans (GenBank accession number AF099922), containing 8000 nucleotides starting from location 7021. The DNA stretch contains a gene (F56F11.4) with five exons (indicated by the rectangular boxes overlay on the plot), all identified by the peaks of the plot.

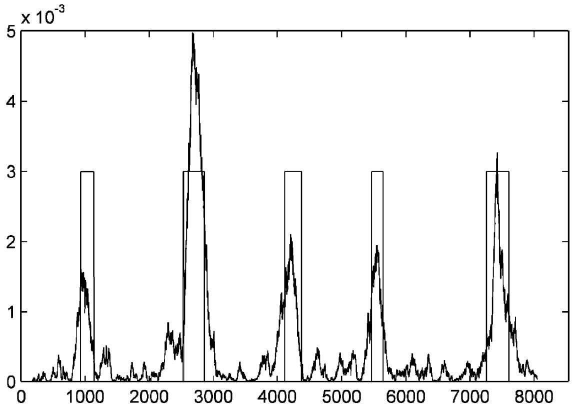

Fig. 5. Plot of $W = |aA + tT + gG + cC|^2$ using the optimized weights given in Ref. [26] for a DNA stretch of C. elegans containing 8000 nucleotides starting from location 7021.

The spectral analysis technique can be applied to different representation of the DNA sequence. Clearly, the result would be dependent on the representation chosen. For example, we have applied the spectral analysis technique to the DB curve representation. Preliminary results have indicated that sometime better positioning of the exons can be obtained. One parameter that affects the exon prediction result is the choice of window length. Appropriate window length is currently found through experimentation. A multiresolution approach using technique such as wavelet analysis could be an interesting avenue of research [28,29]. Spectral analysis technique can also be used to detect other latent periodicities and features of biological interest. Interested readers are referred to [30].

Besides consideration of coding statistics, gene prediction could also involve the identification of specific sub-sequence patterns such as splice sites, promoter regions, transcription factors binding sites and polyA sites [1]. These specific regulatory elements play important roles in the transcription and translation process of a gene and their existence is strong indication of gene presence. Unfortunately, their identification is no trivial task [31,32]. Pattern recognition and machine learning techniques can be very useful in such area. Some of the latest gene prediction algorithms have used powerful machine learning techniques (such as neural networks, pattern recognition methods, and rule-based methods) and probabilistic learning models (such as hidden Markov models (HMM)) to achieve better prediction results [33]. Some examples of these algorithms are: GRAIL [34], GeneScan [35], Glimmer [36], GeneMark.hmm [37], MZEF [38], GeneFinder [http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html], MORGAN [39].

The problem of gene prediction is far from being solved. Take for example the estimation of the number of genes in the human genome. A few years ago, it was estimated that the human genome contains around 100,000 genes. That number drops to around 30,000 when the human genome sequence was completed. It was recently suggested that the number could possibly be as low as 20,000. In the summer of 2003, gene prediction researchers admitted that they were nowhere near establishing a final count and have decided to call the winner to be at 24,500 for now [40]. The main problem with gene prediction lies in what actually defines a gene. Molecular biologists are finding that some genes are shorter than anybody expected a gene to be. It is also hard to tell sometime whether a piece of code is a single gene or two that overlap. The community is also not quite sure how to classify genes that code for multiple proteins or gene-like sequences that code only for RNA. Added to this complication are the so-called dark matters, which are seemingly geneless regions in a genome that might contain hidden coding sequences. It is clear that better prediction performance will require better knowledge about why genes have the characteristics they do, and to be able to exploit that biological knowledge in the prediction algorithms.

3. DNA microarray gene expression profiling

3.1. cDNA microarray technology

Important insights into gene function can be gained by gene expression profiling. Gene expressing profiling is the process of determining when and where particular genes are expressed. For example, some genes are turned on (expressed) or turned off (repressed) when there is a change in external conditions or stimuli. In multi-cellular organisms, gene expressions in different cell types are different during different developmental stages in life. Even within the same cell type, gene expressions are dependent on the cell cycle the cells are in. DNA mutation may alter the expression of certain genes, which causes illness such as abnormal tumor growth or cancer. Furthermore, the expression of one gene is often regulated by the expression of another gene. A detail analysis of all these information will provide an understanding about the inter-networking of different genes and their functional roles.

Microarray technology, which allows massively parallel, high throughput profiling of gene expression in a single hybridization experiment, has recently emerged as a powerful tool for genomic research [41–43]. The technique allows the simultaneous study of tens of thousands of different DNA nucleotide sequences on a single microscopic glass slide. Besides the enormous scientific potential of cDNA microarrays in the fundamental study of gene expressions, gene regulations and interactions, they also have very important applications in pharmaceutical and clinical research. For example, by comparing gene expressions in normal and disease cells, microarrays can be used to identify disease genes for therapeutic drugs or for assessing the effect of a treatment.

The cDNA microarray holds hundreds or thousands of spots, each of which contains a different known DNA sequence called a probe. These spots are printed onto a glass slide by a robotic arrayer. In a microarray experiment, two samples of cRNA, which are reversed transcribed from mRNA purified from cellular contents, are labeled with different fluorescent dyes (usually Cy3 and Cy5, which have different emission wavelengths) to constitute the cDNA targets. The two cDNA targets are then hybridized onto the microarray. If a target contains a cDNA whose sequence is complementary to the DNA probe on a given spot, that cDNA will hybridize to the spot, where it will be detectable by its fluorescence. Spots with more bound targets will have more fluorescent dyes and will therefore fluoresce more intensely.

Once the cDNA targets have been hybridized to the array and any loose target has been washed off, the array is scanned by a laser scanner to determine how much of each target is bound to each spot. The hybridized microarray is scanned for the red wavelength (at approximately 635 nm for the cyanine5, Cy5 dye) and the green wavelength (at approximately 530 nm for the cyanine3, Cy3 dye), which produces two images typically in 16-bit Tiff format. The ratio of the two fluorescence intensities at each spot indicates the relative abundance of the corresponding DNA sequence in the two cDNA samples that are hybridized to the DNA sequence on the spot. By examining the expression ratio of each spots in the Cy3 and Cy5 images, gene expression study can be performed. Fig. 6 shows a schematic of the cDNA microarray technique and the steps in performing a cDNA microarray experiment.

The large amount of data in the microarray images necessitates the use of computer analysis. In general, analysis of microarray data can be categorized into two parts: image analysis for data extraction and data analysis on the gene expression ratio [44]. Automatic and reliable analysis of microarray images has proved to be difficult due to the poor contrast between spots and background, and the many contaminations or artifacts arising from the hybridization procedures such as irregular spot shape and size, dust on the slide, large intensity variation within spots and background, and nonspecific hybridization.

3.2. Image processing

The task in microarray image analysis involves computing the expression ratio for each spot giving information about the relative extent of hybridization of the two cDNA

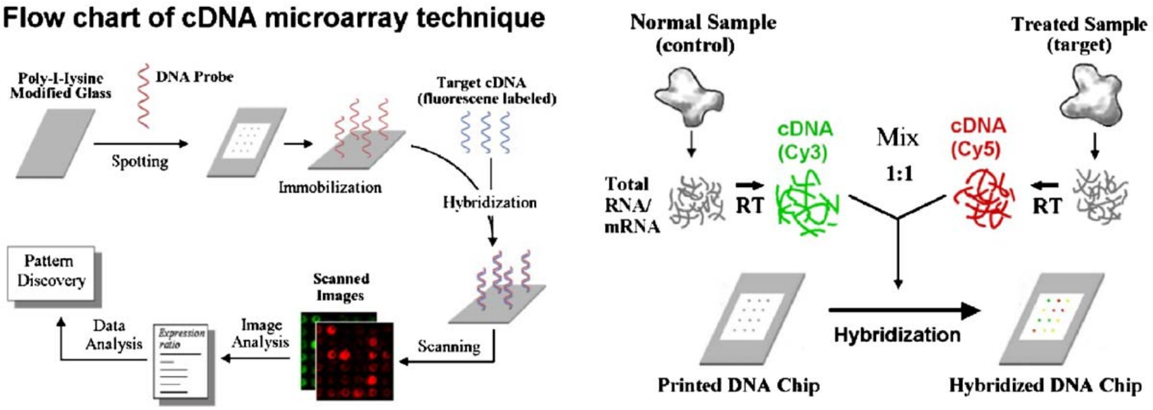

Fig. 6. Left: A schematic of the cDNA microarray technique. Right: The steps involved in a cDNA microarray experiment.

samples. It typically involves the following steps: (i) identify the location of all blocks on the microarray image, (ii) generate the grid within each block which subdivides the block into $p \times q$ subregions, each containing at most one spot, and (iii) segment the spot, if any, in each subregion. We give a brief account of each of the steps in our image analysis algorithm [45] below.

The microarray image analysis algorithm starts by generating a single gray level image from the two TIFF images. The two TIFF images could undergo image registration if needed. The resultant gray level image could undergo image smoothing to reduce the effect of image noise. One way to obtain the gray level image X is to use the following equation:

$$X = \left\lfloor 0.5 * \left(G' + \left(\frac{median(G')}{median(R')} \right) R' \right) \right\rfloor, \qquad (2)$$

where G refers to the Cy3 image, R refers to the Cy5 image, $G' = \sqrt{G}$, $R' = \sqrt{R}$, and $\lfloor \ \rfloor$ denotes rounding to the nearest integer in the range [0–255].

The blocks in a microarray image are arranged in a rigid pattern due to the printing process, and each of the blocks in a microarray image is surrounded by regions void of any spots. Hence, an effective way for block segmentation is through an analysis of the vertical and horizontal image projection profiles. In our image analysis algorithm, the projection profiles are obtained from an adaptively binarized image. By performing analysis on the projection profiles, accurate block segmentation can be achieved (see Fig. 7, left). To locate the individual spots in a block, we perform the gridding operation. Our gridding strategy consists of first locating the good quality spots (we called them guide spots). To account for the variable background and spot intensity, a novel adaptive thresholding procedure and morphological processing are used to detect the guide spots. The geometry of the grid is then inferred from these spots (see Fig. 7, right).

Spot segmentation is then performed in each of the subregions defined by the grid. The segmentation involves finding a circle that separates out the spot, if any, from the background. When a spot is present, the intensity distribution of the pixels within the subregion is modeled using a 2-class Gaussian-Mixture model to find the optimum threshold. Once the sub-region is thresholded and segmented, a best-fit circle is computed for the final spot segmentation. Although the spot shape is constrained to be circular to ensure robustness to poor quality segmentation, adaptive shape segmentation can be easily adopted for good quality spots. Fig. 8 presents some spot segmentation examples for blocks of different spot density and quality from different microarray images.

3.3. Data extraction and processing

Once the spots in a microarray image are extracted, the intensity value of each spot can be obtained and the log ratio, i.e., $M = \log_2 R/G$, which indicates the differential expression of the two DNA samples, can be computed. However, due to contaminations and experimental errors, some preprocessing of the raw intensity value is needed before the expression data can be subjected to further analysis. The preprocessing steps usually involve (i) background correction, (ii) data normalization, and (iii) missing values estimation.

The motivation for background correction is the belief that a spot's measured intensity includes a contribution not due to the specific hybridization of the target to the probe. This could arise from non-specific hybridization and stray fluorescence emitted from other chemicals on the glass slide. Such contribution should be removed from the spot's measured intensity to obtain a more accurate quantification of hybridization. Different approaches, ranging from simple subtraction of local background intensity [46,47] to sophisticated statistical correction have been proposed [48].

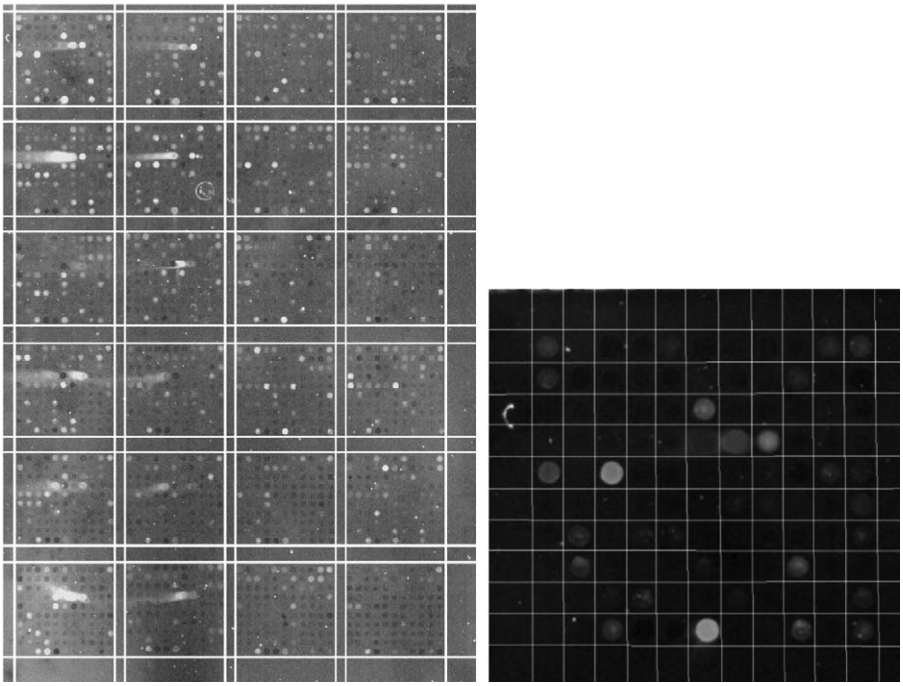

Fig. 7. Left: The segmentation of a microarray image into blocks. Right: Gridding in a block.

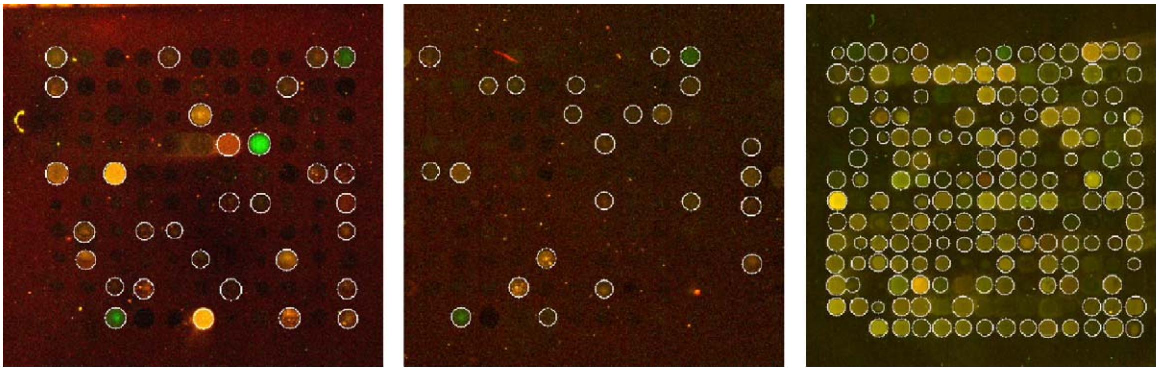

Fig. 8. cDNA microarray spot segmentation results (contrast is enhanced for visual display).

The purpose of normalization is to adjust for any bias that arises from variation in the microarray process rather than from biological differences between the RNA samples. Position variation on a slide may arise due to differences between the print-tips, variation over the course of the print-run or non-uniformity in the hybridization. Differences between slides may arise from differences in ambient conditions when the slides were prepared. Another common variation is the red–green bias due to the differences between the labeling efficiencies, the scanning properties of the two fluors, and the scanner settings. It is necessary to normalize the spot intensities before any subsequent analysis is carried out.

The most widely used within-slide normalization method assumes that the red–green bias is constant on the log-scale across the slide. The log-ratios are corrected by subtracting a constant c to get the normalized values $M_{\text{norm}} = \log_2(R/G) - c$. The global constant c is usually estimated from the mean or the median log-ratios value over a subset of the genes assumed to be not differentially expressed [49,50]. However, the imbalance in the red and green intensities is usually not constant across the spots within and between slides, and can vary according to overall spot intensity, location on the slide, slide origin, and possibly other variables. Other more sophisticated normalization methods are available to account for these dependencies [51]. Additionally, housekeeping genes can be used as control spots for normalization.

Microarray gene expression experiments usually suffered from the missing values problem. Missing values occur due to various reasons, including artifacts on the microarray image, insufficient resolution, image corruption, etc. The unreliable spots on the microarray image are usually manually flagged and excluded from subsequent analysis, resulting in the missing of data on those locations. The existence of missing values has important implication for subsequent data analysis. For example, the inability of many cluster algorithms to process the missing values means that profiles containing missing values are often discarded. However, instead of ignoring gene expression profiles containing missing values (thus throwing away useful information), such missing values can often be estimated based on available knowledge and assumptions about the data.

Reliable estimation of missing values is important since it greatly affects subsequent data analysis. Common methods to deal with missing values are replacements by zeros or by the average of the expression profile. Such techniques, however, made very crude use of the available knowledge within the data. Other more advanced techniques, such as the K-nearest neighbor method (*KNNimpute*) or the singular value decomposition (SVD) method (*SVDimpute*), have recently been proposed [52]. We have recently proposed a missing value imputation technique based on projection onto convex sets (POCS) and SVD [53]. Our algorithm uses two convex sets derived from performing SVD on the expression matrix. Let the gene expression data be tabulated as a matrix A of size $M \times N$, where M denotes the number of genes being studied and N denotes the number of arrays produced under N different experimental conditions. If we perform SVD to matrix A, we get

$$A_{M\times N} = U_{M\times M}\Sigma_{M\times N}V^{\mathrm{T}}_{N\times N}. \tag{3}$$

Let $L = \min\{M, N\}$, matrix V^{T} now contains L eigengenes, and matrix U contain L eigenarrays. Unlike *SVDimpute,* our method makes use of information in both the eigengenes and eigenarrays for missing value imputation. Moreover, we allow uncertainties in the estimated values by modeling them as convex sets and use the POCS algorithm to iteratively refine the estimated values. Using the new algorithm, we were able to obtain a normalized root mean squared error reduction of between 15–20% compared to *KNNimpute* and *SVDimpute* on the gene expression datasets of yeast cell-cycle [54].

3.4. Pattern discovery by cluster analysis

A standard tool in gene expression data analysis is cluster analysis. Cluster analysis aims at finding groups in a given data set such that objects in the same group are similar to each other while objects in different groups are dissimilar. Since genes with related functions are expected to have similar expression patterns, clustering of genes may suggest possible roles for genes with unknown functions based on the known functions of some other genes that are placed in the same cluster. Many clustering algorithms developed in pattern recognition, for example, K-means, Self-Organizing Maps (SOM), Hierarchical clustering, Self-Organizing Tree Algorithm, Principal Component Analysis (PCA), and Multi-Dimensional Scaling, have all been applied to the study of high-dimension gene expression data [55–62]. Clustering of gene expression data has been applied to the study of temporal expression of yeast genes in sporulation [63], the identification of gene regulatory networks [64], and the study of cancer [65].

Traditional clustering techniques can generally be classified into two categories, hierarchical and partitional. Hierarchical clustering algorithm transforms a pairwise dissimilarity matrix of patterns into a sequence of nested partitions, called a dendrogram. Partitional clustering, on the other hand, performs a partition of patterns into K clusters, such that patterns in a cluster are more similar to each other than to patterns in different clusters. Both categories of clustering algorithms have their merits and weaknesses and both have been used extensively in gene expression data study.

We have recently proposed a novel hierarchical-partitioning framework that combines the features of both categories of algorithms, which we called binary hierarchical clustering (BHC) [66]. The BHC is inspired by the idea of hierarchical binary subdivision of data proposed in [67]. In essence, the algorithm performs a successive binary subdivision of the data in a hierarchical manner, until further splitting of a partition into two smaller partitions is insignificant anymore (see Fig. 9). The hierarchical structure is manifested in the binary tree structure of the clustering result, where a parent node gives rise to two children nodes if the projection onto the optimal Fisher discriminant axis is greater than a certain threshold. The tree structure allows the relationship between adjacent clusters and the variation within each cluster to be observed easily. The partitioning behavior of our algorithm is incorporated in the cluster splitting process, where a variant of the fuzzy C-means clustering algorithm is used to split a parent cluster into two children clusters. The main advantages of the BHC clustering algorithm are: (i) The number of clusters can be estimated from the data directly using a binary hierarchical framework; (ii) No constraint about the number of samples in each cluster is required, and (iii) No prior assumption about the class distribution is needed.

In traditional partition-based clustering algorithms, if the number of prototypes is less than that of the natural clusters in the dataset, a prototype could win patterns from more than one cluster. This behavior is called one-prototype-take-multiple-clusters (OPTMC) (see Fig. 10a). Thus, a natural cluster might be erroneously divided into two or more classes, or several natural clusters or part of them are erroneously grouped into one class. In view of the above

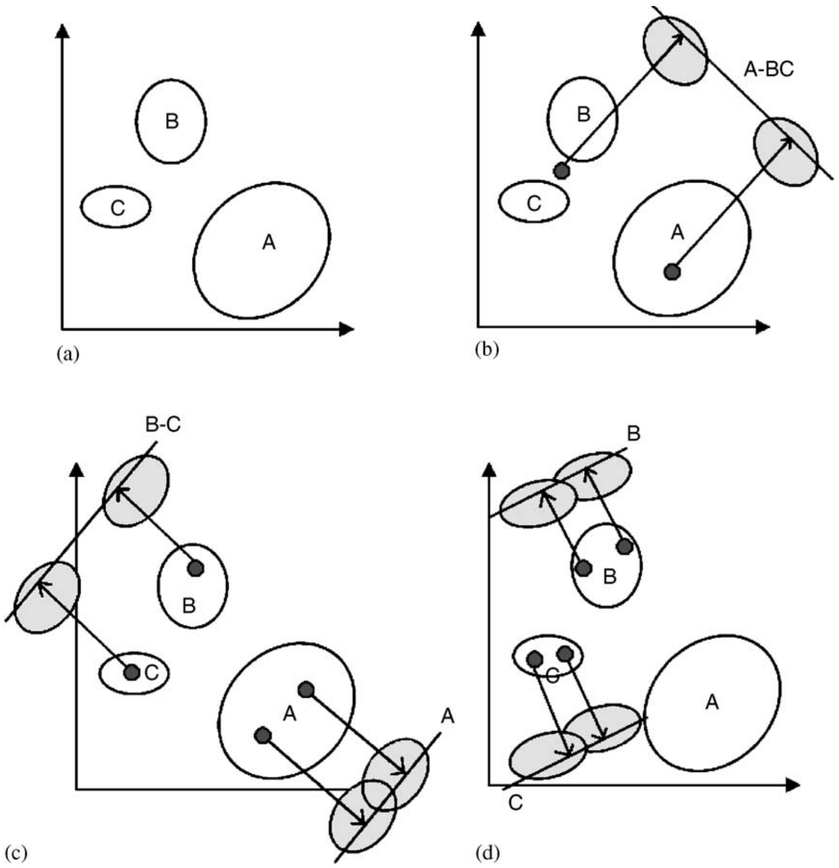

Fig. 9. The binary hierarchical clustering framework. (a) Original gene expression data treated as one class. (b) Split the class into two clusters, A and BC. (c) Cluster A cannot be split further, but cluster BC is split into two clusters, B and C. (d) Both cluster B and C cannot be split any more, and we have three clusters A, B, and C (figure adapted from Ref. [67]).

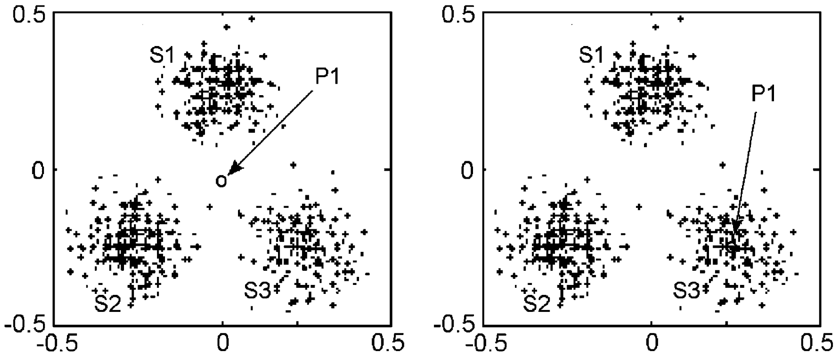

Fig. 10. Two learning methods: OPTMC versus OPTOC. Left: One prototype takes the center of three clusters (OPTMC). Right: One prototype takes one cluster (OPTOC) and ignores the other two clusters. See Ref. [70].

shortcoming, we recently proposed a new partition-based clustering framework called Self-Splitting and Merging Competitive Learning Clustering (SSMCL) [68,69]. The new algorithm is able to identify the natural clusters through the adoption of a new competitive learning paradigm called the one-prototype-take-one-clusters (OPTOC) [70]. The OPTOC learning paradigm allows a cluster prototype to focus on just one natural cluster, while minimizing the

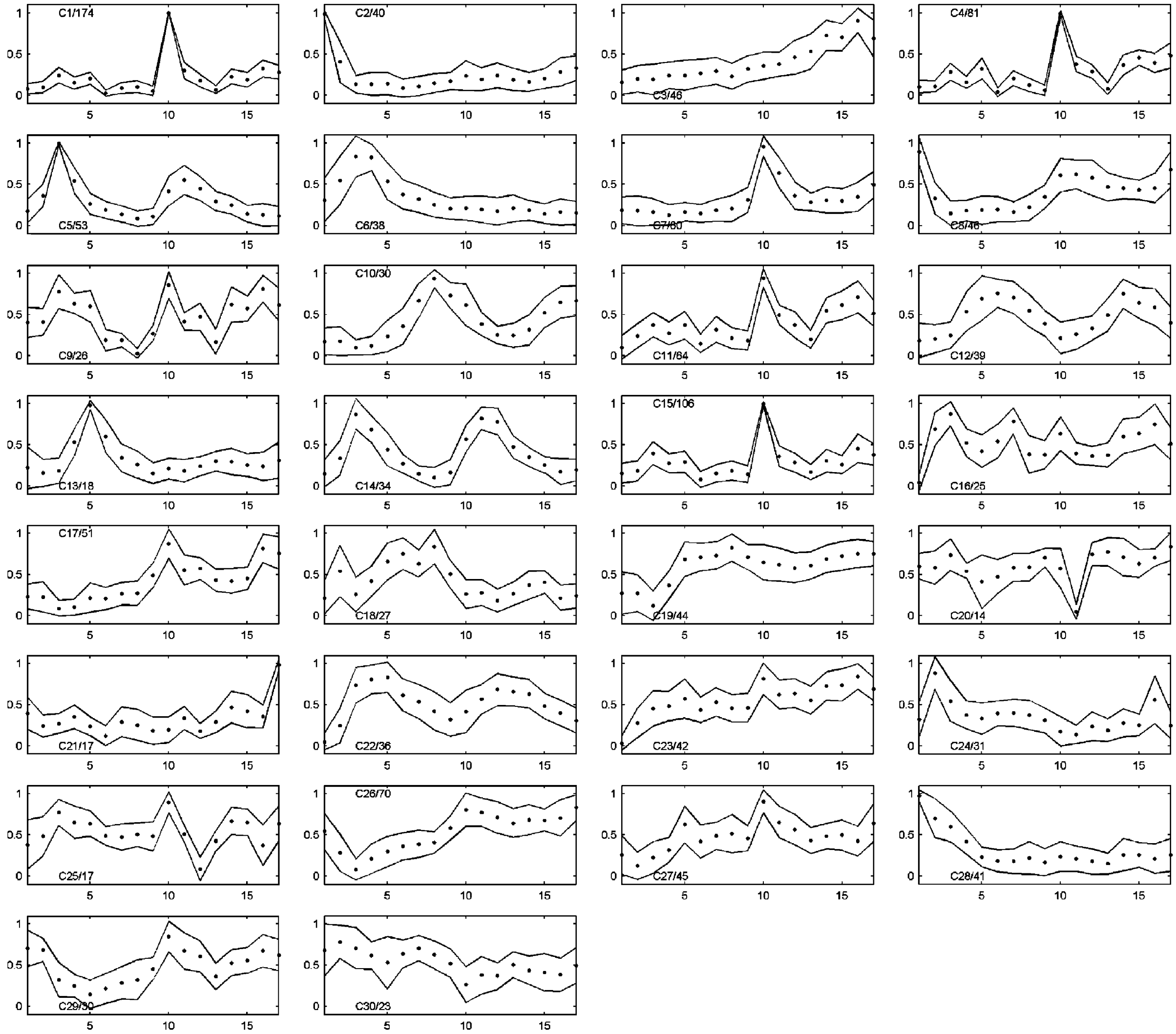

Fig. 11. The clustering results for the yeast cell cycle data [54]. The number of clusters is set to 30.

competitions from other natural clusters (see Fig. 10b). The OPTOC behavior of a cluster prototype is achieved through the use of a dynamic neighborhood, which causes the prototype to eventually settle at the center of a natural cluster, while ignoring competitions from other clusters.

Since it is very difficult to estimate reliably the correct number of natural clusters in a complex high dimension dataset, an over-clustering and merging strategy was used to estimate the number of *distinct clusters* in the dataset. The over-clustering and merging strategy can be viewed as a top-down (divisive clustering), followed by a bottom-up (agglomerative clustering) process. In the top-down step, loose clusters (as measured by their variances) are successively split into two clusters until a pre-specific number of clusters, set to be larger than the true number of clusters in the data, are obtained. The over-clustering minimizes the chance of missing some natural clusters in the data. The merging step then attempts to merge similar clusters together, until finally all remaining clusters are distinct from each other. The SSMCL algorithm was used to cluster the yeast cell cycle data [54]. The data was first over-clustered into 30 clusters (see Fig. 11). Cluster-merging is then performed on the result until 22 clusters remained (see Fig. 12). The final clusters are all visually distinct from each other.

3.5. *Temporal profile analysis and gene regulation*

Whole-genome expression time-series data are a particularly valuable source of information because they can describe a dynamic biological process such as the cell cycle or metabolic process [54,71]. They allow the determination of causal relationships between the expressions of

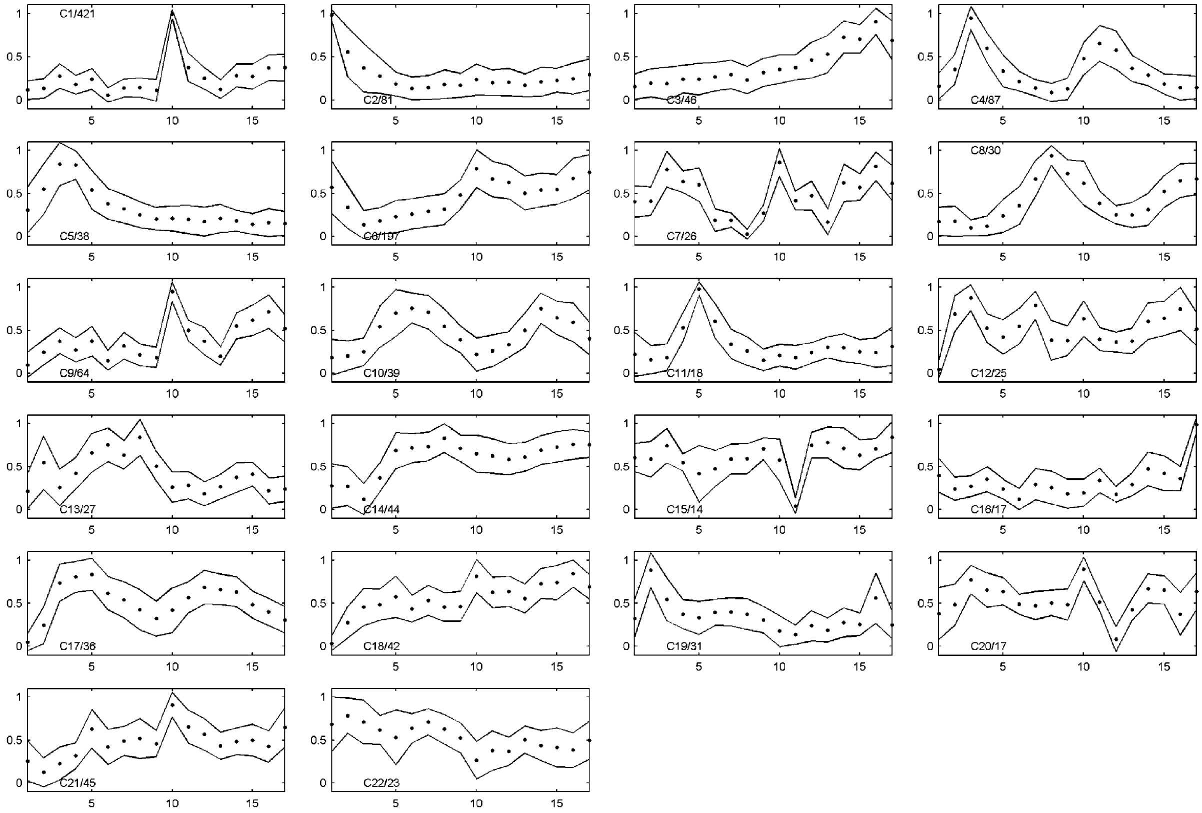

Fig. 12. The final clustering results for the yeast cell cycle data after cluster merging. 22 distinct clusters are obtained.

different genes. Such causal relationship allows the extraction of gene regulatory information, and ultimately leads to a better understanding of the complicated gene networking process within a cell.

Several methods were proposed to extract the significant modes of variation from the large array of expression time series data. Holter et al. [72] used SVD to extract the "characteristic modes" of gene expression. They showed that the behavior of the widely disparate gene systems analyzed in their work is dominated by a small subset of the characteristic modes and that a linear combination of just a few modes provides a good approximation to the behavior of the entire system in most cases. Alter et al. [73] used a similar analysis on two cell cycle time series and found that the first two modes for the cell cycle time series are approximately sinusoidal and 90° out of phase. The temporal nature of time series gene expression data was explicitly modeled by Dewey and Galas [74] using a dynamic model. They modeled the entire set of gene expression data using a first order Markov model which is equivalent to a first order autoregressive (AR) model. The construction of a genetic network consisting of "dynamic classes" based on the transition matrix was also demonstrated.

All the methods mentioned above analyzed the gene expression data as a whole, and attempt to summarize the dataset by a few dominant components. These methods can therefore be considered as global. Another class of algorithms attempt to perform pairwise comparison of gene expressions to identify pairs of genes that have direct regulatory relationships from the set of gene expression profiles. Several algorithms that perform such pairwise analysis for extracting regulatory information from microarray time-series data are the simple correlation analysis method [75], the edge detection method [76], the event method [77], and our spectral component correlation method [78,79].

Among the various pairwise comparison approaches, correlation-based method is perhaps the most popular one. This method determines whether or not two genes have a regulatory relationship by checking the global similarity between their expression profiles using the Pearson correlation measure. However, it does not take into account the fact that there is often a time delay before the regulator gene product can exert its influence on the target gene. Such time delay can significantly degrade the performance of the method. Correlation method also strongly favours global similarity over more localized similarities arising from conditional regulatory relationships.

The edge detection method and the event-based method are specifically designed to overcome the shortcomings of the correlation-based analysis. The edge detection method scans through each gene expression curve to determine where major changes in expression level (edges) occur. To produce a score, the edge detection method sums up the number of edges in two gene expression curves that share the same direction and are within reasonable distances of each other. Gene pairs that are likely to have an activation relationship would give high scores. Similar to the edge method, the event method also examine the slope of the expression profile at each time interval. Depending on the slope value, the algorithm marks each event as either rising (R), constant (C), or falling (F), thus resulting in a string of events for each expression profile. A pairwise sequence alignment of the event strings is then performed to obtain a numerical score that reflect the regulatory relationship between two genes.

If two genes, A and B, are co-regulated, the expression of gene A and gene B should vary more or less at the same frequency. This frequency of variation, however, may not be easily seen from the two time-series expression profiles due to noise and other factors. In addition, if gene B is under the influence of both gene A and gene C ("two-regulating-one" situation), and the expression profiles of these influencing genes are varying at different frequencies, then the relationship between gene A and gene B may not be easily seen from their time-series profiles. This would cause problem for correlation-based similarity comparison, as well as the edge detection method and the event-based method.

In [78,79], we propose a spectral component correlation approach for measuring the correlation between time-series expression data, and use the results to infer the potential regulatory relationships between genes. The technique summarizes the essential features of an expression pattern by means of a frequency spectrum estimated by AR modeling [80]. This method has been studied extensively in magnetic resonance spectroscopy [81]. The idea behind our technique is to decompose a time-series expression profile $x[n]$ into a set of discrete-time damped sinusoids of various frequencies,

$$x[n] = \sum_{i=1}^{M} x_i[n] = \sum_{i=1}^{M} \alpha_i \exp(\sigma_i n) \cos(\omega_i n + \phi_i). \tag{4}$$

The parameters in this model, α_i, σ_i, ω_i, and ϕ_i ($i = 1, 2, 3, \ldots, M$), are the amplitude, damping factor, normalized frequency and phase angle respectively of component i. The correlation of $x[n]$ with another sequence $y[n]$ can then be reformulated as a sum of scaled component-wise correlations,

$$x[n] \circ y[n] = \sum_i \sum_j \sqrt{\frac{E_{x_i} E_{y_j}}{E_x E_y}} x_i[n] \circ y_j[n], \tag{5}$$

Table 2
Results for the two correlation methods applied to all 439 known regulatory pairs

	Traditional correlation < 0.5	Traditional correlation > 0.5	Total
(a) Statistics for the 343 activation pairs			
Component-wise correlation < 0.5	111	9	120
Component-wise correlation > 0.5	196*	27	223
Total	307	36	343

	Traditional correlation < -0.5	Traditional correlation > -0.5	Total
(b) Statistics for the 96 inhibition pairs			
Component-wise correlation < 0.5	1	40	41
Component-wise correlation > 0.5	4	51*	55
Total	5	91	96

where the symbol $\circ$ denotes correlation operation and each term with letter E represents either total energy of a sequence or energy of a particular component. This equation shows how a correlation of two sequences can be separated into a set of scaled component-wise correlations between each spectral component. Such component-wise correlation could provide more insights into the regulatory relationship. For instance, for the "two-regulating-one" situation, correlation between the expression profiles of gene A and gene B may not be strong enough to suggest their relationship due to the presence of spectral components in gene B induced by gene C. However, the spectral components of gene B due to gene A would exhibit strong correlations to gene A's expression profile.

We use the spectral component correlation algorithm to analyze the alpha-synchronized yeast cell-cycle dataset [54]. We were able to detect many regulatory pairs that were missed by the traditional correlation method due to weak correlation value. For those regulations with strong oscillatory but time-shifted expression pairs, we can easily identify them by using only the spectral magnitude information. When the component-wise correlation analysis is applied to all 439 known regulations, the results indicated that 223 out of 343 activations and 55 out of 96 inhibitions have their component-wise correlations score greater than 0.5 (see Table 2). We found that a large number of visually dissimilar expression pairs do have very similar dominant frequency components. For example, among those 307 pairs having traditional correlation coefficients of less than 0.5, 196 of them have greater than 0.5 component-wise correlation coefficients. Furthermore, 60 out of this 196 pairs have their component-wise correlation coefficients greater than 0.9 and the expression patterns in each of these pairs strongly oscillate at almost identical frequencies.

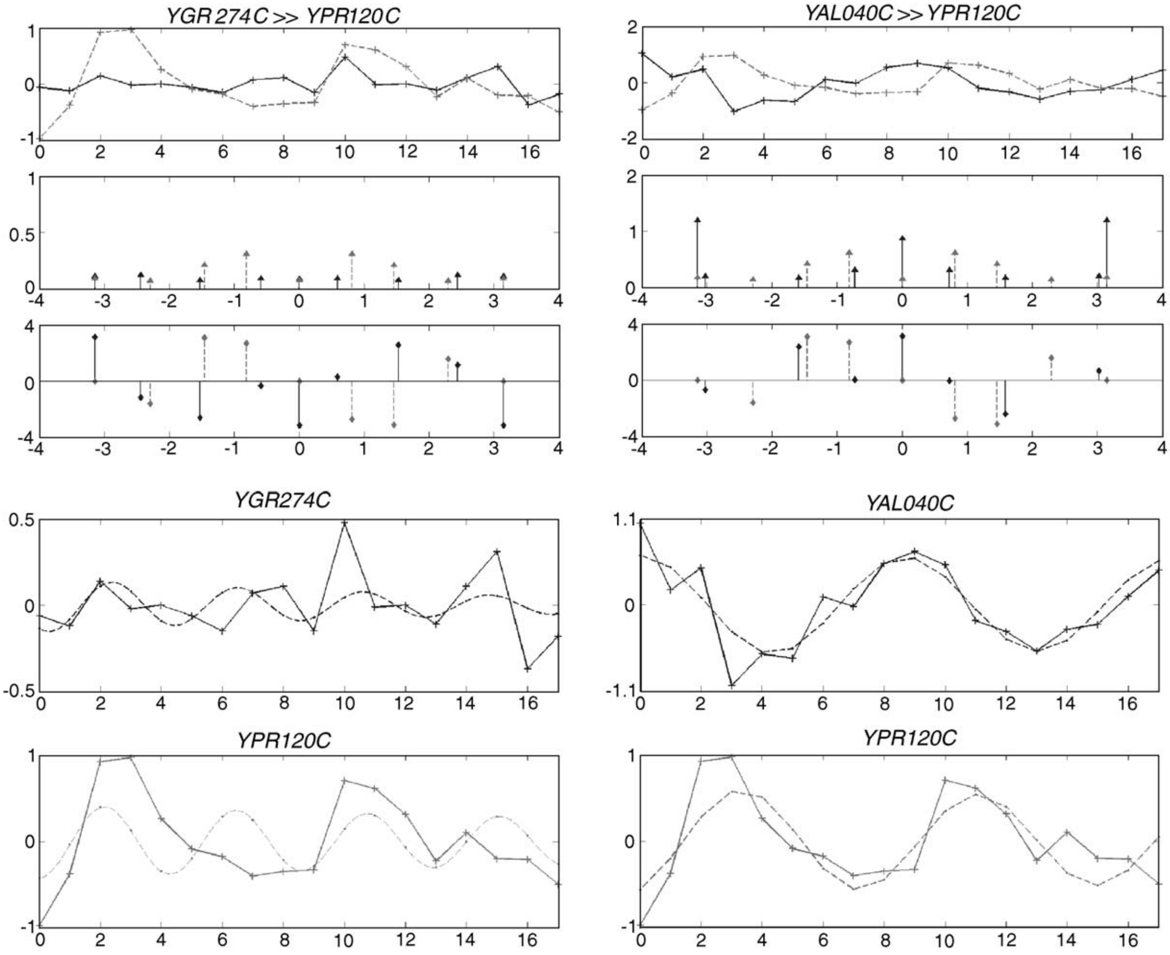

Fig. 13. Two activation regulations with gene *YPR120C* as an activatee. (a) Activation regulation with gene *YGR274C* as an activator. (b) Activation regulation with gene *YAL040C* as an activator. (c) Correlated frequency components for the first pair. (d) Correlated frequency components for the second pair.

The spectral component correlation method therefore allows the hidden component-wise relationships between two expression profiles to be revealed, which are otherwise hidden in the traditional correlation method.

For those regulations involving a single gene being simultaneously regulated by two or more genes with different expression frequencies, it could be possible to identify them by checking for the existence of regulators' frequencies from the expression profile of the gene being regulated. Fig. 13 shows two known activation regulations with a common gene *YPR120C* as an activatee. The figure reveals that the first regulation has its expression profiles correlated at frequency of around 1.48 rad/s, whereas the second regulation has its profiles correlated at around 0.76 rad/s.

To see how causal relationship can be inferred from the algorithm, we choose the genes *YBR240C* and *YAL040C* as references and find all other genes in the Filkov's dataset [76] which has a component-wise correlation coefficient of greater than 0.7. There are 55 out of 288 genes for *YBR240C* and 59 out of 288 genes for *YAL040C* that satisfy this threshold. These two sets of genes with their scores are shown in Fig. 16 and their oscillatory properties are clearly revealed when they are arranged such that their phase is in descending order. Within these genes, one known activation regulation of gene *YBR240C* is contained in the first set and three for gene *YAL040C* are contained in the second set. Note that genes below the reference gene have their phases lag by 0–180° relative to the reference gene's phase, and they can be considered as potential activated candidates. On the other hand, genes above have their phases lead by 0–180°, and they can be considered as potential inhibited candidates. If we look at the known activatees for the two examples shown in Fig. 14, we see that they are all located below their corresponding activators. The spectral

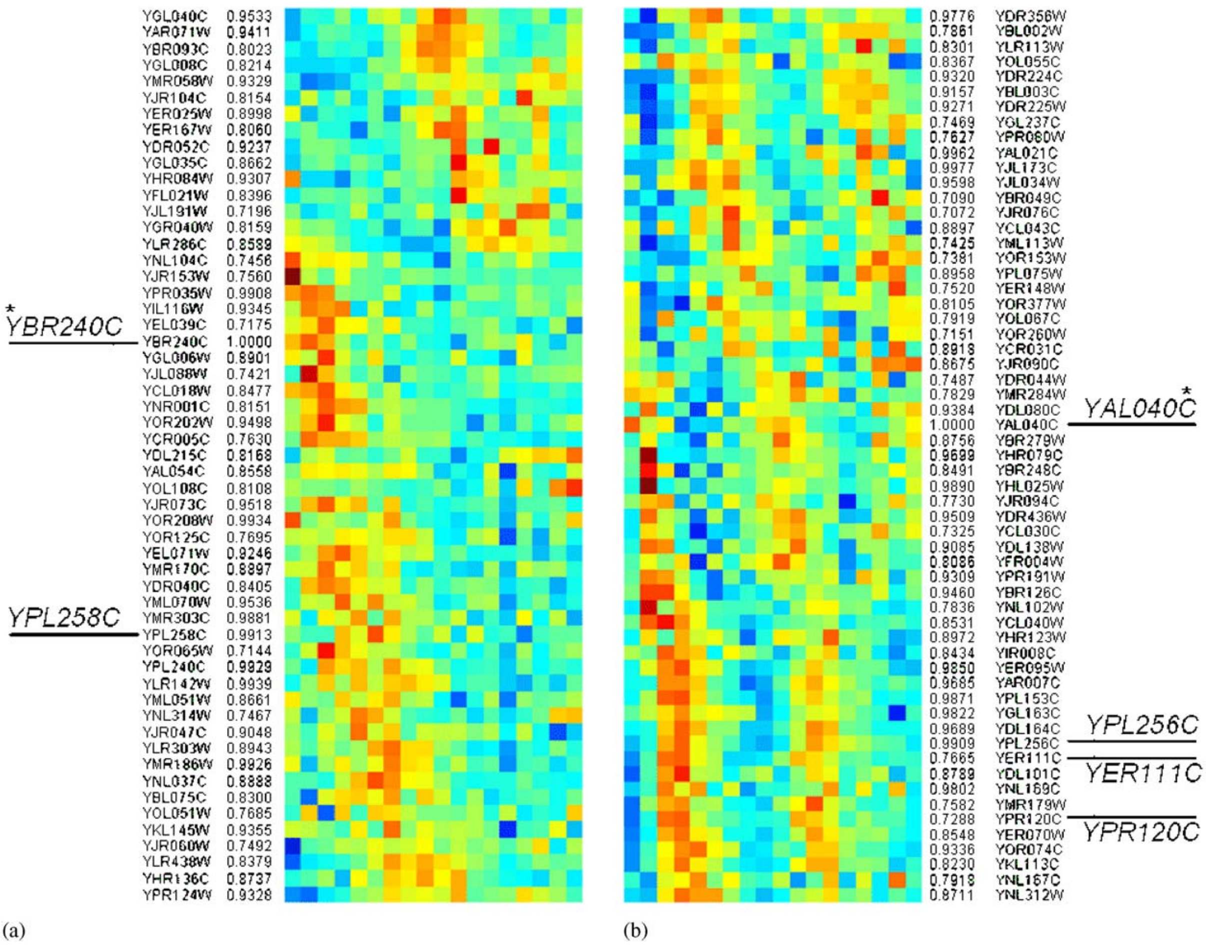

Fig. 14. Genes in the Filkov's dataset have their component-wise correlation coefficients, relative to gene (a) *YBR240C* and gene (b) *YAL040C*, greater than 0.7. The known activation regulations with genes *YBR240C* and *YAL040C* as activators are highlighted.

component correlation method allows such casual relationship to be observed.

4. Discussions and conclusions

This article presents a review of applications of pattern recognition techniques to the emerging field of bioinformatics. In particular, we focus on two key areas: DNA sequence analysis and microarray gene expression analysis, which have been topics of intense research recently. In DNA sequence analysis, sequence comparison with the annotated sequences in the online databases is an important problem. Currently, the most commonly used sequence comparison tools are those based on string pattern matching. Algorithms that are efficient and computationally fast are of paramount importance in view of the enormous size of current DNA sequence databases. For relatively short sequences, other sequence comparison and visualization techniques are possible. We briefly described the DB curve sequence visualization method for short sequence comparison. One inadequacy of string matching technique is that structural or functional contexts are not readily observable in the matching sequences. For example, in string matching two sequences, we would not know explicitly where the segments correspond to, say, coding region, are located in the matching sequence. Such contexts may be more readily observable if similarity matching is performed in the transform domain.

A major area of research in DNA sequence analysis is that of gene prediction. For gene prediction, finding effective features is a major task. The problem is somewhat similar to conventional pattern recognition problems, where finding a set of discriminative features that adequately describe the various classes is an important first step. We described how one can exploit the 3-base periodicities in a coding sequence by a classical signal processing technique, i.e., the discrete Fourier transform. The windowed FFT algorithm has recently been applied successfully to detect exons and introns, to identify correct reading frame, and to reveal other biologically relevant periodicities in the DNA sequence. Although substantial work has been done in the area of gene prediction, the problem is far from been solved. Further work is required to improve the detection rates and to decrease the level of falsely predicted genes. Such improvements may come from the incorporation of new and better sub-models, of promoters or initial and terminal exons, as well as other physical properties and signals present in the DNA, such as bendability or nucleosome positioning, and better classifier design.

In the second part of this review, we described some key issues in microarray gene expression analysis. We described how to extract the gene expression data from the microarray images. We also discussed the pre-processing steps necessary, i.e., background correction, normalization, missing value estimation, to prepare the gene expression data for subsequent analysis. Cluster analysis is an important tool for the discovery of interesting patterns and structures in gene expression data. We presented some of our latest work on cluster analysis of gene expression data which overcome some of the short comings of current clustering techniques. Further work in this area could be the investigation of alternative feature space obtained through some coordinate transformations, such that the salient structures in the data become more accessible to clustering. Gene expression time series data could be used to study the dynamic biological process such as the cell cycle or metabolic process. Spectral estimation technique is particularly useful in this area. We discussed current approaches in the study of expression time series data and point out some of their shortcomings. We then described our recent work for expression time series data analysis using the technique of spectral component correlation. The dominant spectral components, estimated through AR modeling, allow many otherwise weak but biologically significant correlations to be detected successfully. The technique also allows the inference of complex regulatory relationships, such as multiple-to-one regulations between multiple genes, as well as the causal relationships between regulators and regulatees.

In conclusion, we note that the field of bioinformatics is multi-discipline in nature. Collaboration between biologists, computer scientists and engineers is indispensable to solving many of the fundamental biological problems. As many of the problems in bioinformatics involve identifying and analyzing specific features and patterns in the data, we believe that the pattern recognition community could make significant contribution to this challenging but fascinating emerging research area. Having said that, we certainly do not want to convey the impression that biological problems are just like normal computer science problems. Instead, the readers are reminded that for any computer algorithms to be useful in bioinformatics, they should make biological sense. Domain knowledge about the underlying biological process should always be taken into account if possible when performing any computation and during the result interpretation phase.

Acknowledgements

The work described in this paper was supported by an interdisciplinary grant (Project 9010003) and a strategic research grant (Project 7001706) from City University of Hong Kong. We like to thank Lawrence C-H Tzang from the Department of Biology and Chemistry at City University of Hong Kong for doing the artworks in Fig. 6.

References

[1] D.W. Mount, Bioinformatics—Sequence and Genome Analysis, Cold Spring Harbor Laboratory Press, New York, 2001.

[2] A.W.C. Liew, H. Yan, M. Yang, Data Mining for Bioinformatics, in: Yi-Ping Phoebe Chen (Ed.), Bioinformatics Technologies, Springer, 2005, pp. 63–116, (Chapter 4).

[3] A.W.C. Liew, H. Yan, M. Yang, Phoebe Chen, Microarray Data Analysis, in: Yi-Ping Phoebe Chen (Ed.), Bioinformatics Technologies, Springer, 2005, pp. 353–388, (Chapter 12).

[4] S.B. Needleman, C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, J. Mol. Biol. 48 (1970) 443–453.

[5] T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197.

[6] T.F. Smith, M.S. Waterman, Comparison of biosequences, Adv. Appl. Math. 2 (1981) 482–489.

[7] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.

[8] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucl. Acids Res. 25 (1997) 3398–3402.

[9] Y. Wu, A.W.C. Liew, H. Yan, M. Yang, DB-Curve: a novel 2D method of DNA sequence visualization and representation, Chem. Phys. Lett. 367 (2003) 170–176.

[10] D. Anastassiou, Genomic Signal Processing, IEEE Signal Processing Magazine, July 2001, pp. 8–20.

[11] H.T. Chang, N.W. Lo, W.C. Lu, C.J. Kuo, Visualization and comparison of DNA sequences by use of three-dimensional trajectories, Proceedings of the First Asia-Pacific Bioinformatics Conference APBC2003, February 2003, Adelaide, Australia.

[12] E.A. Cheever, G.C. Overton, D.B. Searls, Fast Fourier transform-based correlation of DNA sequences using complex plane encoding, Comput. Appl. Biosci. 7 (2) (1991) 143–154.

[13] J. Ning, C.N. Moore, J.C. Nelson, Preliminary wavelet analysis of genomic sequences, Proceedings of the Second IEEE Computer Society Bioinformatics Conference CSB2003, August 2003, Stanford, CA, USA, pp. 509–510.

[14] N. Kawagashira, Y. Ohtomo, K. Murakami, K. Matsubara, J. Kawai, P. Carninci, Y. Hayashizaki, S. Kikuchi, Wavelet profiles: their application in Oryza sativa DNA sequence analysis, Proceedings of the First IEEE Computer Society Bioinformatics Conference CSB2002, August 2002, Stanford, CA, USA, pp. 343–344.

[15] C. Mathe, M.F. Sagot, T. Schiex, P. Rouze, Current methods of gene prediction, their strengths and weakness—survey and summary, Nucl. Acids Res. 30 (2002) 4103–4117.

[16] J.W. Fickett, C.S. Tung, Assessment of protein coding measures, Nucl. Acids Res. 20 (1992) 64412–64450.

[17] J.W. Fickett, Finding genes by computer: the state of the art, Trends Genet. 12 (1996) 316–320.

[18] M. Burset, R. Guigo, Evaluation of gene structure prediction programs, Genomics 34 (1996) 353–367.

[19] R. Guigo, DNA composition, codon usage and exon prediction, in: M.J. Bishop (Ed.), Genetic Databases, Academic Press, 1999, pp. 53–80, (Chapter 4).

[20] Y. Wu, A.W.C. Liew, H. Yan, M. Yang, Classification of short human exons and introns based on statistical features, Phys. Rev. E 67 (6) (2003) Art. No. 061916.

[21] A.W.C. Liew, Y. Wu, H. Yan, Selection of Statistical Features Based on Mutual Information for Classification of Human Coding and Non-coding DNA Sequences, Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, United Kingdom, 23–26 August 2004.

[22] A.W.C. Liew, Y. Wu, H. Yan, M. Yang, Effective statistical features for coding and non-coding DNA sequence classification for yeast, C. elegans and Human, International Journal of Bioinformatics and Application, to appear.

[23] J.D. Hawkins, A survey on intron and exon lengths, Nucl. Acids Res. 16 (1988) 9893–9908.

[24] C.T. Zhang, J. Wang, Recognition of protein coding genes in the Yeast genome at better than 95% accuracy based on the Z curve, Nucl. Acids Res. 28 (2000) 2804–2814.

[25] B.D. Silverman, R. Linsker, A measure of DNA periodicity, J. Theor. Biol. 118 (1986) 295–300.

[26] D. Anastassiou, Frequency-domain analysis of biomolecular sequences, Bioinformatics 6 (12) (2000) 1073–1082.

[27] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, R. Ramaswamy, Prediction of probable genes by Fourier analysis of genomic sequences, Comput. Appl. Biosci. 13 (1997) 263–270.

[28] J. Song, T. Ware, S.L. Liu, M. Surette, Comparative genomics via wavelet analysis for closely related bacteria, EURASIP J. Appl. Signal Process. (1) (2004) 5–12.

[29] X.Y. Zhang, F. Chen, Y.T. Zhang, S.C. Agner, M. Akay, Z.H. Lu, M.M.Y. Waye, S.K.W. Tsui, Signal processing techniques in genomic engineering, Proc. IEEE 9 (12) (2002) 1822–1833.

[30] D. Sussilo, A. Kundaje, D. Anastassiou, Spectrogram analysis of genomes, EURASIP J. Appl Signal Process. (1) (2004) 29–42.

[31] J.W. Fickett, A.G. Hatzigeorgiou, Eukaryotic promoter recognition, Genome Res. 7 (9) (1997) 861–878.

[32] T. Werner, Models for prediction and recognition of eukaryotic promoters, Mammalian Genome 10 (2) (1999) 168–175.

[33] S. Rogic, A.K. Mackworth, F.B.F. Ouellette, Evaluation of gene-finding programs on mammalian sequences, Genome Res. 11 (5) (2001) 817–832.

[34] E.C. Uberbacher, Y. Xu, R.J. Mural, Discovering and understanding genes in human DNA sequence using GRAIL, Methods Enzymol. (Series title: Computer methods for macromolecular sequence analysis) 266 (1996) 259–281.

[35] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, R. Ramaswamy, Prediction of probable genes by Fourier analysis of genomic sequences, Comput. Appl. Biosci. 13 (1997) 263–270.

[36] S.L. Salzberg, A.L. Delcher, S. Kasif, O. White, Microbial gene identification using interpolated Markov models, Nucl. Acids Res. 26 (1998) 544–548.

[37] A.V. Lukashin, M. Borodovsky, GeneMark.hmm: new solutions for gene finding, Nucl. Acids Res. 26 (1998) 1107–1115.

[38] M.Q. Zhang, Identification of protein coding regions in the human genome by quadratic discriminant analysis, Proc. Natl. Acad. Sci. USA 94 (1997) 565–568.

[39] S.L. Salzberg, A.L. Delcher, K.H. Fasman, J. Henderson, A decision tree system for finding genes in DNA, J. Comp. Biol. 5 (1998) 667–680.

[40] Elizabeth Pennisi, A low number wins the GeneSweep pool, Science 300 (5625) (2003) 1484.

[41] M. Schena, D. Shalon, R.W. Davis, P.O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, Science 270 (1995) 467–470.

[42] S.K. Moore, Making Chips to probe genes, IEEE Spectrum (2001) 54–60.

[43] D.J. Lockhart, E.A. Winzeler, Genomics, gene expression and DNA arrays, Nature 405 (2000) 827–846.

[44] A.W.C. Liew, L.K. Szeto, S.S. Tang, H. Yan, “A computational approach to gene expression data extraction and analysis”, Special issue on Genomic Signal Processing, Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology, Vol. 38, Issue 3, November 2004, pp. 237–258.

[45] A.W.C. Liew, H. Yan, M. Yang, Robust adaptive spot segmentation of DNA microarray images, Pattern Recognition 36 (5) (2003) 1251–1254.

[46] M. Eisen, ScanAlyze User Manual. Stanford University, 1999. http://.rana.lbl.gov/EisenSoftware.htm.

[47] Axon Instruments Inc. GenePix Pro 3.0, 2001.

[48] C. Kooperberg, T.G. Fazzio, J.J. Delrow, T. Tsukiyama, Improved background correction for spotted DNA microarrays, J. Comp. Biol. 9 (1) (2002) 55–66.

[49] Y. Chen, E.R. Dougherty, M.L. Bittner, Ratio-based decisions and the quantitative analysis of cDNA microarray images, J. Biomed. Opt. 2 (1997) 364–374.

[50] R.D. Wolfinger, G. Gibson, E.D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, R.S. Paules, Assessing gene significance from cDNA microarray expression data via mixed models, J. Comp. Biol. 8 (2001) 625–637.

[51] Y.H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, T.P. Speed, Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, Nucl. Acids Res. 30 (4) (2002) e15.

[52] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, R.B. Altman, Missing values estimation methods for DNA microarrays, Bioinformatics 17 (2001) 520–525.

[53] X. Gan, A.W.C. Liew, H. Yan, Missing value Estimation for Microarray Data Based on Projection onto Convex Sets method”, Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, United Kingdom, 23–26 August 2004.

[54] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray

hybridization, Molecular Biol. Cell 9 (1998) 3273–3297 (http://cellcycle-www.stanford.edu).

[55] A. Brazma, J. Vilo, Minireview: gene expression data analysis, European Molecular Biology Laboratory, Outstation Hinxton—the European Bioinformatics institute, Cambridge CB10 ISD UK, 2000.

[56] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, Gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, Proc. Nat. Acad. Sci. USA 96 (1999) 6745–6750.

[57] C.M. Perou, S.S. Jeffrey, M. van de Rijn, C.A. Rees, M.B. Eisen, D.T. Ross, A. Pergamenschikov, C.F. Williams, S.X. Zhu, J.C.F. Lee, D. Lashkari, D. Shalon, P.O. Brown, D. Botstein, Distinctive gene expression patterns in human mammary epithelial cells and breast cancers, Proc. Nat. Acad. Sci. USA 96 (1999) 9212–9217.

[58] K.P. White, S.A. Rifkin, P. Hurban, D.S. Hogness, Microarray analysis of Drosophila development during metamorphosis, Science 286 (1999) 2179–2184.

[59] K.Y. Yeung, W.L. Ruzzo, Principal component analysis for clustering gene expression data, Bioinformatics 17 (9) (2001) 763–774.

[60] C. Tang, L. Zhang, A. Zhang, Interactive visualization and analysis for gene expression data, IEEE Proceedings of the Hawaii International Conference on System Sciences. Big Island, HI, January 2002, vol. 6, pp. 143–166.

[61] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Nat. Acad. Sci. USA 95 (1998) 14863–14868.

[62] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, T.R. Golub, Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation, Proc. Nat. Acad. Sci. USA 96 (1999) 2907–2912.

[63] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P.O. Brown, I. Herskowitz, The transcriptional program of sporulation in budding yeast, Science 282 (1998) 699–705.

[64] T. Chen, V. Filkov, S.S. Skiena, Identifying gene regulatory networks from experimental data, Proceedings of the Third Annual International Conference on Computational Molecular Biology RECOMB99, Lyon, France, March 1999, pp. 94–103.

[65] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.

[66] L.K. Szeto, A.W.C. Liew, H. Yan, S.S. Tang, Gene expression data clustering and visualization based on a binary hierarchical clustering framework, J. Vis. Lang. Comput. 14 (2003) 341–362.

[67] D.A. Clausi, K-means Iterative Fisher (KIF) unsupervised clustering algorithm applied to image texture segmentation, Pattern Recognition 35 (2002) 1959–1972.

[68] S. Wu, A.W.C. Liew, H. Yan, Cluster analysis of gene expression data based on self-splitting and merging competitive learning, IEEE Trans. Inform. Technol. Biomed. 8 (1) (2004) 5–15.

[69] A.W.C. Liew, H. Yan, S. Wu, A novel OPTOC-based clustering algorithm for gene expression data analysis, Proceedings of the Fourth International Conference on Information, Communications & Signal Processing and Fourth IEEE Pacific-Rim Conference on Multimedia, ICICS-PCM2003, 15–18 December 2003, Singapore.

[70] Y.J. Zhang, Z.Q. Liu, Self-Splittng competitive learning: a new on-line clustering paradigm, IEEE Trans. Neural Networks 13 (2002) 369–380.

[71] J.L. DeRisi, V.R. Lyer, P.O. Brown, Exploring the metabolic and genetic control of gene expression on a genomic scale, Science 278 (1997) 680–686.

[72] N.S. Holter, M. Mitra, A. Maritan, M. Cieplak, J.R. Banavar, N.V. Fedoroff, Fundamental patterns underlying gene expression profiles: simplicity from complexity, Proc. Nat. Acad. Sci. USA 97 (2000) 8409–8414.

[73] O. Alter, P.O. Brown, D. Botstein, Singular value decomposition for genome-wide expression data processing and modeling, Proc. Nat. Acad. Sci. USA 97 (2000) 10101–10106.

[74] T.G. Dewey, D.J. Galas, Dynamic models of gene expression and classification, Funct. Integr. Genomics 1 (2001) 69–278.

[75] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proc. Nat. Acad. Sci. USA 95 (1998) 14863–14868.

[76] V. Filkov, S. Skiena, J. Zhi, Analysis Techniques for microarray time series data, J. Comp. Biol. 9 (2) (2002) 317–330.

[77] A.T. Kwon, H.H. Hoos, R. Ng, Inference of transcriptional regulation relationships from gene expression data, Bioinformatics 19 (8) (2003) 905–912.

[78] L.K. Yeung, H. Yan, A.W.C. Liew, L.K. Szeto, M. Yang, R. Kong, Measuring correlation between microarray time-series data using dominant spectral component, Proceedings of the 2nd Asia-Pacific Bioinformatics Conference APBC2004, 18–22 Jan, 2004, Dunedin, New Zealand, pp. 309–314.

[79] L.K. Yeung, L.K. Szeto, A.W.C. Liew, H. Yan, Dominant spectral component analysis for transcriptional regulations using microarray time-series data, Bioinformatics 20 (5) (2004) 742–749.

[80] S. Marple, Digital Spectral Analysis with Applications, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1987.

[81] H. Yan, Signal Processing for Magnetic Resonance Imaging and Spectroscopy, Marcel Dekker, New York, 2002.

About the Author—ALAN WEE-CHUNG LIEW received his B.Eng. with first class honors in Electrical and Electronic Engineering from the University of Auckland, New Zealand, in 1993 and Ph.D. in Electronic Engineering from the University of Tasmania, Australia, in 1997. He is currently an Assistant Professor in the Department of Computer Science and Engineering, The Chinese University of Hong Kong. His current research interests include computer vision, medical imaging, pattern recognition and bioinformatics. He has served as a technical reviewer for a number of international conferences and journals in IEEE Transactions, IEE proceedings, bioinformatics and computational

biology. Dr. Liew is a member of the Institute of Electrical and Electronic Engineers (IEEE), and his biography is listed in the 2005 Marquis Who's Who in the World.

About the Author—HONG YAN received a B.E. degree from Nanking Institute of Posts and Telecommunications in 1982, an M.S.E. degree from the University of Michigan in 1984, and a Ph.D. degree from Yale University in 1989, all in electrical engineering. In 1982 and 1983 he worked on signal detection and estimation as a graduate student and research assistant at Tsinghua University. From 1986 to 1989 he was a research scientist at General Network Corporation, New Haven, CT, USA, where he worked on design and optimization of computer and telecommunications networks. He joined the University of Sydney in 1989 and became Professor of Imaging Science in 1997. He is currently Professor of Computer Engineering at City University of Hong Kong. His research interests include image processing, pattern recognition and bioinformatics. He is author or co-author of one book and over 200 refereed technical papers in these areas. Professor Yan is a fellow of the International Association for Pattern Recognition (IAPR), a fellow of the Institution of Engineers, Australia (IEAust), a senior member of the Institute of Electrical and Electronic Engineers (IEEE) and a member of the International Society for Computational Biology (ISCB).

About the Author—MENGSU YANG received his BSc degree in chemistry from Xiamen University, China (1984), MSc degree in organic chemistry from Simon Fraser University, Canada (1989), and PhD degree in analytical chemistry from University of Toronto, Canada (1993). He obtained his postdoctoral training in molecular biology in The Scripps Research Institute, USA (1993–1994). He joined City University of Hong Kong in 1994 and is currently Professor of Chemistry and Director of the Applied Research Centre for Genomics Technology at City University of Hong Kong. Professor Yang has published over 90 peer-reviewed scientific papers on the development of novel analytical techniques for biomedical applications and the studies of biomolecular interactions in cellular processes. His work has been recognized by the Best Paper Awards in the Eurasia Chemical Conference (1996) and the Asia-Pacific Conference of Tumor Biology (2001). Professor Yang was awarded the K. C. Wong Education Foundation Scholar Award in 2003. He is a member of the International Advisory Committee of "The Analyst" (a Royal Society of Chemistry publication) and a member of the Editorial Board of Life Science Instruments (a Chinese Society of Chemistry publication). He holds honorary professorships in The University of Hong Kong and Zhejiang University, China.