Visualisation of Reduced-Dimension Microarray Data Using Gaussian Mixture Models

Julien Epps[†*] **and Eliathamby Ambikairajah**[†]

[†] School of Electrical Engineering and Telecommunications
The University of New South Wales, Sydney 2052, Australia

[*] National Information Communication Technology Australia (NICTA)
Australian Technology Park, Eveleigh 1430, Australia

julien.epps@nicta.com.au, ambi@ee.unsw.edu.au

Abstract

Dimensionality reduction, clustering and visualisation methods proposed in recent years have afforded new possibilities for the analysis of gene expression data. However, efficient, novel techniques for processing and representing microarray data are still required. We propose the use of the discrete cosine and sine transformations for dimensionality reduction of microarray data. These techniques have found powerful applications in the signal processing domain. Gaussian mixture models (GMMs) are then used for clustering and visualisation of the reduced-dimension data. Results on human fibroblast microarray data reveal that the discrete sine and cosine transforms can greatly reduce the dimensionality of gene expression data while preserving good clustering results. GMMs are shown to produce improved clustering results according to an intra-class cluster tightness criterion, in addition to a new two-dimensional representation whose axes afford the possibility of physical interpretations.

Keywords: Dimensionality reduction, discrete cosine transform, discrete sine transform, Gaussian mixture model, clustering, visualisation, microarray, gene expression.

1 Introduction

An important problem in the area of microarray analysis is the organization of large-dimensional gene expression data and its presentation in a format that can emphasize the similarities and differences between different gene expressions, thus facilitating their biological interpretation.

Literature in this area spans clustering techniques such as the *K*-means algorithm (Datta 2003), self-organising maps (Törönen *et al.* 1999), cellular neural networks (Zhang *et al.* 2003), and visualisation techniques such as image maps and dendrograms (Iyer *et al.* 1998), Sammon's algorithm (Törönen *et al.* 1999) and 2-D scatter plots (Datta 2003).

There have also been a number of contributions in microarray data projection and dimensionality reduction, such as the application of principal component analysis (PCA) (Raychaudhuri, Stuart and Altman 2000, Misra *et al*. 2002, Datta 2003), independent component analysis (ICA) (Liao *et al*. 2002) and singular value decomposition (Wall, Rechsteiner and Rocha 2003).

Gene expression data is usually generated in high dimensionality at laboratories, however often some of the dimensions are correlated, creating redundancy. Dimensionality reduction is a promising approach because it extracts the most important components of the data and allows for lower complexity in subsequent processing, and also facilitates visualisation of clustered data.

A further argument for dimensionality reduction arises from the fact that many microarray databases have a very large dimension *p* relative to the number of gene observations *n* to be clustered. Previous researchers (e.g. McLachlan, Bean and Peel 2002) have found that this situation, known as the "$p > n$" problem can often cause singular estimates of the within-cluster covariance matrices. Reducing the dimension, *p*, is one means of mitigating this problem with microarray clustering.

In this paper, we propose the use of the discrete cosine transform (DCT) and discrete sine transform (DST) for reducing the number of dimensions in microarray data, and Gaussian mixture models (GMMs) for clustering and visualisation of the reduced-dimension data. The DCT and DST provide decorrelation, ordering, and dimensionality reduction, while GMMs provide probability density function modeling. The use of DCTs as a front end and GMMs as a back end is well established in some signal processing applications, where they have made substantial contributions to the state of the art.

This paper is organized as follows. Section 2 provides a brief introduction to microarray measurements of gene expression and their analysis. In section 3 the DCT and DST are introduced, and their use in modelling gene expression profiles is explained. The application of Gaussian mixture models to the modelling of microarray data is described in section 4. In section 5, dimensionality reduction methods, clustering techniques and a measure of cluster tightness are presented, and the results of these experiments and visualisation of the reduced-dimension data are discussed in section 6.

2 Introduction to Microarray Analysis

To a large extent, the protein components of each cell dictate its function and response to various environmental changes. Each cell in an organism contains the information necessary to produce the entire repertoire of proteins the organism can specify, and their behaviour is in turn largely determined by the genes the cell is expressing.

DNA microarrays rely on the hybridization properties of nucleic acids to monitor DNA or RNA abundance in different cells over time, and these abundance levels give a quantitative description of the extent to which gene expression is occurring. In microarray experiments, there are certain systematic sources of variation, usually due to specific features of the microarray measurement technology, which should be corrected prior to further analysis. This normalization is often performed by subtracting the background (or average value) from the signal for each gene. The pre-processed data can be visualised as a matrix, with each row consisting of expression values at different time instants or for different experimental conditions for a single gene. Most experiments typically contain between 4000 to 8000 rows (genes) and between 4 and 80 columns (gene expression values).

Eisen *et al.* (1998) found that larger groups of clustered genes tend to share common roles in cellular processes. Hence, a major use of microarray data is to classify genes with similar expression profiles into groups in order to investigate their biological significance. A clustering tree (Iyer *et al.* 1998), for example, can show how genes form groups. A wide variety of clustering algorithms are employed for these purposes. Regardless of how clustering is performed, however, the purpose of such classification is to provide summarized information about gene similarity to biological practitioners. Judicious use of statistical techniques and an appropriate visualization can help to achieve this objective.

3 The Discrete Cosine and Sine Transforms

3.1 Principal Component Analysis

Principal Component Analysis (PCA) is a classical statistical technique that can be used to find structure in multidimensional data sets. PCA uses eigenvalue decomposition to estimate a series of components that are ordered according to what proportion of the variance of the K original variables is contained in each component. These principal components are mutually uncorrelated and orthogonal. The principal components are defined by the eigenvectors of the $K \times K$ covariance matrix of the microarray data, and from these, the projection of the ith gene along the jth principal component is calculated as

$$a_{ij} = \sum_{k=1}^{K} a_{ik} v_{jk} \ , \qquad (1)$$

where a_{ik} is the kth element of the ith gene expression profile, and v_{jk} is the kth element of the jth principal component. Since the first few principal components contain more of the variance than the later components, a_{ik} can be reasonably well approximated by including only the first few components $\mathbf{v}_j$ in (1), i.e. by summing from k=1 to some reduced dimension $K'<K$. This method of dimensionality reduction has previously been used with some success on various microarray databases (Raychaudhuri, Stuart and Altman 2000, Misra *et al.* 2002, Datta 2003), and hence is employed for the purpose of comparison in the experiments of sections 5 and 6 below.

3.2 The Discrete Cosine Transform

The purpose of the discrete cosine transform is to transform a data sequence into another domain, in order to take advantage of some characteristics of the data so that the energy of the transformed data is localized into a small number of coefficients. The discrete cosine transform is real valued, orthonormal, has near-optimal properties for energy compaction of highly correlated data, and can be computed more efficiently than other similar transforms (e.g. PCA, and ICA).

The DCT represents a data sequence $x(n)$ in terms of its cosine series expansion with coefficients C_k, calculated as

$$C_k = \alpha_k \sum_{n=0}^{K-1} x(n) \cos\left[\frac{(2n+1)k\pi}{2K}\right], \qquad (2)$$

where $k = 0 , \ldots , K$-1, n is the sequence sample index, K is the length of the input sequence $x(n)$, $\alpha_0 = 1/\sqrt{K}$ and $\alpha_k = \sqrt{2/K}$ for $k = 1 , \ldots , K$-1. The first two DCT coefficients in particular have the following interpretations:

C_0 is the arithmetic mean of the data sequence $x(n)$, and thus corresponds to the average gene expression ratio.

C_1 is the amplitude of a cosine wave of period $2K$, and practically behaves as an approximation to the gradient of the data sequence $x(n)$. Thus, C_1 gives a rough approximation to the overall shape of the expression pattern for a gene.

Possibly the main advantage of the DCT and DST for microarray analysis is that the gene expression profiles are always effectively projected onto the same axes. In the case of low-order coefficients such as C_0 and C_1, these axes allow consistent interpretation from a biological, rather than purely statistical, perspective as suggested above. This is in contrast to PCA, where the axes onto which the gene expression profiles are projected are data-dependent, and will therefore be different from one data set to another. This is a particular advantage for visualization purposes, where it is desirable for axes to be consistent and easily interpreted by an observer.

3.3 The Discrete Sine Transform

The discrete sine transform is similar to the DCT, and is real, orthonormal, has excellent compaction properties for uncorrelated data sequences, and can be computed as efficiently as the DCT. The DST is defined as

$$S_k = \sqrt{\frac{2}{K+1}} \sum_{n=0}^{K-1} x(n) \sin\left[\frac{(n+1)(k+1)\pi}{K+1}\right], \qquad (3)$$

for $k = 0 , \ldots , K$-1. An example of the application of the DCT and DST is given in Figure 1. Note that most of the energy in C_k and S_k is concentrated in the first few coefficients, hence the use of the DCT and DST in speech and audio compression applications. Higher order coefficients tend to model detailed variations in the data sequence $x(n)$, possibly including noise (removal of which is desirable for analysis purposes).

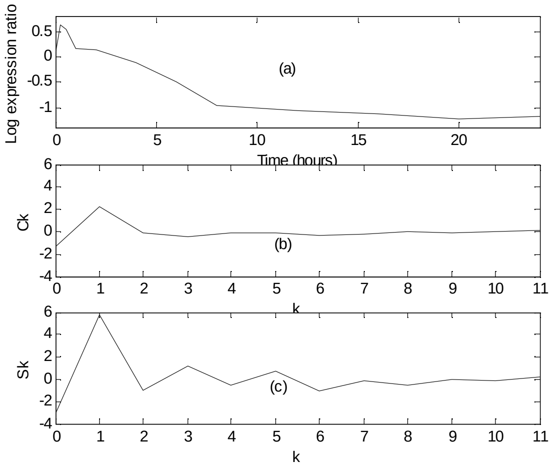

Figure 1: An example gene expression profile (a) and its DCT (b) and DST (c) coefficients. In this example, the DCT provides a more compact representation than the DST.

4 Gaussian Mixture Models

Gaussian Mixture Models (GMMs) comprise a weighted sum of Gaussian probability density functions, or mixtures, and are used in applications with probability density functions that are generated by more than one 'source'. In microarray analysis, GMMs can be used to estimate the relative contributions of each source (or cluster) to the overall probability density function, so that the mean of each Gaussian mixture an estimate of the cluster centroid. GMMs are usually parameterized by their mean vectors $\boldsymbol{\mu}_m$, covariance matrices $\boldsymbol{\Sigma}_m$ and mixture weights w_m, and the overall probability density function is given as

$$p(\mathbf{x}) = \sum_{m=1}^{M} w_m \exp\left[-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_m)'\boldsymbol{\Sigma}_m^{-1}(\mathbf{x}-\boldsymbol{\mu}_m)\right]. \qquad (4)$$

The number of Gaussian mixtures M determines the detail with which the probability density function of the sequence $\mathbf{x} = \{x_1 ,\ldots, x_K\}$ is modelled. The parameters of GMMs are often estimated using the EM algorithm (Dempster, Laird and Rubin 1977).

A one-dimensional example of distribution modelling is shown in Figure 2, where a mixture model (solid line) is the sum of four individual Gaussian probability density functions (dashed lines) with different means, variances and weights. The individual mixtures can be considered to represent possible clusters in the gene expression data. In this example, three of the probability density functions have very similar means and overlap substantially, indicating either that the clusters are not easily separable or that all three are modelling the same cluster.

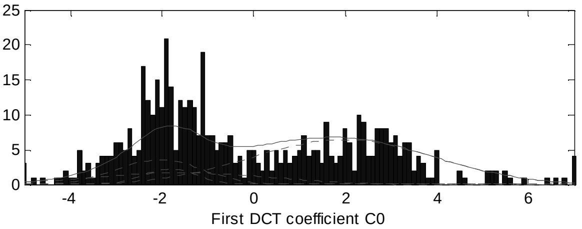

Figure 2: Histogram of the distribution of C_0 for fibroblast microarray data and its approximation using a one-dimensional GMM with M = 4 mixtures.

In this paper, two-dimensional Gaussian mixture models are estimated for human fibroblast data that has been transformed using the DCT and then truncated to just two coefficients (C_0 and C_1). Thus, peaks in this probability density function correspond to regions of the data where there are a relatively large number of genes with similar expression profiles. The location of such peaks corresponds to the cluster centroids $\boldsymbol{\mu}_m$, while the 'sharpness' $\boldsymbol{\Sigma}_m$ of such peaks corresponds to the spread of gene data about the centroids, or the cluster tightness.

5 Dimensionality Reduction and Clustering

5.1 Human Fibroblast Data

The techniques discussed in this paper were evaluated on the data set collected from experiments exploring the response of fibroblasts to serum, performed by Iyer *et al.* (1998). The data were derived from microarrays with temporal measurements of the mRNA levels made at K = 12 non-equal intervals of 0, 0.25, 0.5, 1, 2, 4, 6, 8, 12, 16, 20, and 24 hours. The same subset of N = 517 genes used in (Iyer *et al.* 1998) were considered in this paper, and the $\log_2$ expression ratio was used during clustering.

5.2 Dimensionality Reduction

Analysis of the log gene expression data set was then performed on the raw data itself, and on the DCT and DST coefficients of the raw data. Thus, two further data sets were obtained by applying the DCT (C_0 and C_1 from equation (2) were used) and DST (S_0 and S_1 from equation (3) were used) to the raw data set. The objective of this comparison was to determine whether the clustering accuracy obtained from the raw data could be preserved even when the data dimensionality was reduced from 12 to 2.

The motivation for selecting only the first two DCT and DST coefficients stems from previous experimental work on yeast sporulation data (Epps and Ambikairajah 2004), in which various combinations of DCT coefficients were compared, revealing that the statistical spread of classes clustered using C_0 and C_1 was only 10% greater than that of classes clustered using the original data.

A further third data set was generated by applying principal component analysis to the raw data set. The objective of this comparison was to determine whether the DCT and DST would provide comparable clustering accuracy to an existing dimensionality technique from the literature. PCA was chosen as a standard for comparison because of its frequent use as a dimensionality reduction technique in prior microarray analysis literature (e.g. Raychaudhuri, Stuart and Altman 2000, Misra *et al.* 2002, Datta 2003).

Although the relatively small dimensions of the raw data set used in this work make the use of a wide range of analysis techniques feasible without requiring any reduction in the number of dimensions, other applications and data sets with much larger length *N* and dimension *K* may benefit even more greatly from dimensionality reduction techniques such as these.

5.3 Clustering Techniques

Clustering was then performed on the raw and transformed data sets described in section 5.2. In keeping with the original classification of the fibroblast data into ten classes (clusters A to J) (Iyer *et al.* 1998), the number of clusters was chosen as $M = 10$. In order to remove any dependency of this comparison upon a particular clustering technique, four techniques commonly used in pattern recognition were selected:

- The LBG algorithm (Linde, Buzo and Gray 1980), with splitting performed one cluster at a time
- Self-organizing maps (Kohonen 1980)
- The K-means algorithm (Gersho and Gray, 1993)
- The EM algorithm (Dempster, Laird and Rubin 1977) applied to 10-mixture GMMs

Where clustering was performed on transformed data sets, the cluster membership index of each vector was retained. The raw data set was then clustered based upon these indices, in preparation for use by the cluster tightness criterion described in section 5.4.

5.4 Cluster Tightness Measure (CTM)

In order to measure the efficacy of the clustering, a measure based upon the standard deviations of each cluster along each dimension was devised. This measure was normalized according to the global standard deviation along each dimension, so that

$$CTM = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{K}\sum_{k=1}^{K}\frac{\sigma_k^m}{\sigma_k^G}\right), \tag{5}$$

where σ_k^m is the standard deviation of the *m*’th cluster along the *k*’th dimension, σ_k^G is the standard deviation across all data along the *k*’th dimension, *K* is the length of the input sequence and *M* is the number of clusters. If *CTM* is zero, this implies that all data lies on the cluster centroids, while larger values of *CTM* imply that clusters are spread widely and may overlap.

The cluster tightness measure is a within-class estimate of classification effectiveness, however it is possible to devise inter-class measures also, to better measure the separation between the various classes. Recent literature (Shah 2004) has also shown that a useful and more biologically-motivated criterion based on Significance Analysis of Microarray (SAM), used widely in the bioinformatics area, can also provide good insight into the effects of dimensionality reduction and clustering.

Of the four clustering methods employed in section 5.3, the LBG algorithm is the only one that does not use random initialization. In order to improve the accuracy of CTM for self-organizing maps, the *K*-means algorithm and GMMs, clustering was repeated five times for these methods, and the cluster tightness measures averaged to provide the final estimate of CTM.

6 Results

6.1 Comparison of Dimensionality Reduction Techniques

Clustering based on 2-dimensional DCT or DST coefficients generally achieved within 10% of the cluster tightness attained by clustering the raw data directly. From this evidence that the accuracy of clustering is not significantly impeded by dimensionality reduction, clustering based upon reduced-dimension data appears to be a powerful tool for efficiently analyzing large dimension microarray data sets. The cluster tightness measurements yielded similar results for both DCT and DST, as given in Table 1.

Table 1 also reveals that similar levels of cluster tightness can be achieved using the DCT, the DST or PCA. It is suggested that the DCT or DST are to be preferred over PCA since they are faster and their axes may be more readily interpreted physically. This is because the axes resulting from PCA are determined by the data, and thus vary from one data set to the next.

A further result from Table 1 is that GMM-based clustering with the EM algorithm generally produces tighter clusters for these data than the LBG algorithm, self-organizing maps and the *K*-means algorithm.

	Cluster tightness measure *CTM*			
Clustering features	LBG	SOM	*K*-M	EM
Raw data	0.330	0.325	0.331	0.316
DCT coeff’ts $C_{0,} C_1$	0.362	0.354	0.358	0.340
DST coeff’ts $S_{0,} S_1$	0.363	0.345	0.364	0.351
Two principal comp’ts	0.357	0.328	0.355	0.348

Table 1: Cluster tightness for raw and transformed data sets, with clustering performed by the LBG algorithm, self-organizing maps (SOM), the K-means algorithm (*K*-M) and GMMs with the EM algorithm.

6.2 GMM-Based Visualisation

For the purposes of visualizing the gene expression data, GMMs have the advantage that the overall density function is estimated, rather than simply the cluster means or centroids, as in the LBG, SOMs or *K*-means approaches. Thus, for two-dimensional GMMs, a two-dimensional visualisation based upon the probability density function can be generated by calculating the probabilities at all positions on a two-dimensional grid using (3), and plotting the contours of equal probability, as seen in Figure 3. In this figure, the cluster centroids from (Iyer *et al*. 1998) are marked for reference, although clearly the use of a different clustering method here has produced different centroids. Ultimately, the evaluation of which clustering technique produces the most informative centroids is subjective, and requires specialist biological interpretation.

Peaks in Figure 3 represent regions where there are large concentrations of genes with similar temporal expression profiles. This allows a molecular biologist to put the similarity of a pair of gene expression profiles into the context of the overall distribution of all profiles. A Euclidean distance measure is often used in microarray analysis, however the GMM-based representation of human fibroblast data, which reveals large variations in density (compare the peak near cluster B to that near cluster J in Fig. 3), demonstrates that the similarity of a pair of gene expression profiles depends on more than just the Euclidean distance.

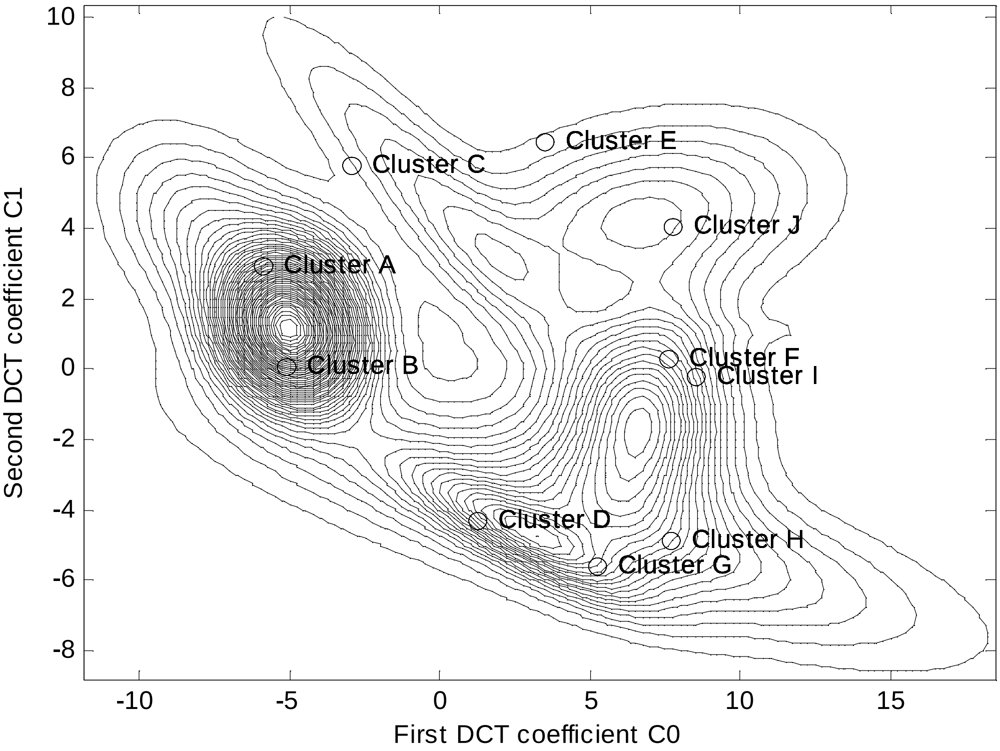

Figure 3: 2-D visualisation of the GMM-based probability density function estimate based upon the first two DCT coefficients of fibroblast log expression data. The original classification into ten clusters (A to J) from Iyer *et al.* (1998) is shown for the purpose of comparison. As mentioned in section 3.2, the horizontal and vertical axes can be biologically interpreted as the average within-gene expression value and the average within-gene expression gradient.

7 Conclusion

This paper has demonstrated a novel, efficient technique for dimensionality reduction and data representation using the DCT or DST and GMMs. Using a measure of cluster tightness to evaluate clustered human fibroblast serum data, the DCT and DST were shown to provide similar clustering performance to the raw data, despite a 6-fold reduction in dimensionality. The DCT and DST have the additional benefits of lower computational complexity and biologically meaningful projection axes. Clustering using GMMs yielded tight clusters and a probability density function that can be visualized in two dimensions to distinguish patterns of similarity in gene expression data.

Future work will focus on broader comparisons with other dimensionality reduction techniques such as ICA and Sammon’s algorithm across several databases, and also on applications of this approach to visualisation of various other kinds of data.

8 References

Datta, S. (2003): Statistical techniques for microarray data: A partial overview. *Communications in Statistics - Theory and Methods*, **32**:263-280.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977): Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. series B*, **39:** 1-38.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998): Cluster analysis and display of genome-wide expression patterns. *Proc. of the National Academy of Sciences*, **95**:14863–14868.

Epps, J., and Ambikairajah, E. (2004): Use of the discrete cosine transform for gene expression data analysis. In *Proc. Workshop on Genomic Signal Processing and Statistics*, Baltimore, USA, I-13.

Gersho, A., and Gray, R. M. (1993): *Vector Quantisation and Signal Compression*. Boston, Kluwer Academic Publishers.

Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C. F., Trent, J. M., Staudt, L. M., Hudson, J., Boguski, M. S., Lashkari, D., Shalon, D., Botstein, D., and Brown, P. O. (1998): The transcriptional program in the response of human fibroblasts to Serum. *Science*, **283:**82-87. Data available from http://genome-www.stanford.edu/serum/

Kohonen, T. (1984): *Self-organization and associative memory*. Berlin, Springer-Verlag.

Liao, X., Dasgupta, N., Lin, S. M., and Carin, L. (2002), ICA and PLS modelling for functional analysis and drug sensitivity for DNA microarray signals. *Proc. Workshop on Genomic Signal Processing and Statistics*, CP1-11.

Linde, Y., Buzo, A., and Gray, R. M. (1980): An algorithm for vector quantiser design. *IEEE Trans. Commun.*, **COM-28**(1):84-95.

McLachlan, G. J., Bean, R. W. and Peel, D. (2002): A Mixture Model-based Approach to the Clustering of Microarray Expression Data. *Bioinformatics*, **18**(3):413-422.

Misra, J., Schmitt, W., Hwang, D., Hsiao, L., Gullans, S., Stephanopoulos, G., and Stephanopoulos, G. (2002): Interactive Exploration of Microarray Gene Expression Patterns in a Reduced Dimensional Space. *Genome Res*, **12**(7):1112-1120.

Raychaudhuri, S., Stuart, J. M., and Altman, R. B. (2000): Principal components analysis to summarize microarray experiments: application to sporulating time series. *Pacific symposium on Biocomputing*, 452-463.

Shah, R. (2004): Genome Signal Processing. Honours thesis. University of New South Wales, Australia.

Törönen, P., Kolehmainen, M., Wong, G., and Castrén, E. (1999): Analysis of gene expression data using self-organising maps. *FEBS Letters*, **451:**142-146.

Wall, M. E., Rechtsteiner, A., Rocha, L. M. (2003) Singular Value Decomposition and Principal Component Analysis. In *.A Practical Approach to Microarray Data Analysis*. 91-109. Berrar, D.P., Dubitzky, W., Granzow, M., (eds). Norwell, MA , Kluwer.

Zhang, X.-Y., Chen, F., Zhang, Y.-T., Agner, S. C., Akay, M., Lu, Z.-H., Waye, M. M. Y., and Tsui, S. K.-W. (2003) Signal processing techniques in genomic engineering. *Proceedings of the IEEE*, **90(**12):1822-1833.