USE OF THE DISCRETE COSINE TRANSFORM FOR GENE EXPRESSION DATA ANALYSIS

Julien Epps and Eliathamby Ambikairajah

School of Electrical Engineering and Telecommunications
The University of New South Wales, Sydney 2052, Australia
j.epps@unsw.edu.au; ambi@ee.unsw.edu.au

ABSTRACT

Analysis of microarray gene expression data using signal processing and statistical techniques has received considerable interest in recent years, however the problem of organizing and visualizing these large data sets is still a pressing issue. This paper proposes the use of the discrete cosine transformation (DCT) to reduce the large number of dimensions of the microarray data, thereby simplifying the subsequent clustering problem, aiding visualization and hence decision-making. Results of the application of the DCT and three clustering techniques to yeast sporulation data show that, similarly to the use of principal components analysis (PCA), the dimensionality of gene expression data can be reduced by a factor of three while still achieving good clustering.

1. INTRODUCTION

Microarray data are characterized by both large dimensions and non-trivial relationships between the various constituent gene expressions. A major challenge in the area of microarray analysis is to organize the gene expressions so as to emphasize their similarities and differences thus facilitating their biological interpretation, while simultaneously preserving the complex relationships existing between them.

Much of the existing literature in this area has focused on techniques for clustering the gene expression data, including the *K*-means algorithm [2], Sammon's algorithm [2, 7], self-organising maps [7] and cellular neural networks [8].

Research effort has also been directed towards data projection and dimensionality reduction techniques such as principal component analysis (PCA) and independent component analysis (ICA) [2, 5] and the combination of the two. Dimensionality reduction is a promising approach because it extracts the important characteristics of the data and allows for lower complexity in subsequent processing.

In this paper, we propose the use of the discrete cosine transform (DCT) for reducing the number of dimensions in microarray data. The motivation for applying the DCT arises from its extensive successful usage in the areas of speech and image processing for decorrelation, ordering and dimensionality reduction purposes.

This paper is organized as follows. Section 2 provides a brief introduction to the DCT and its application to modeling gene expression profiles. In section 3, dimensionality reduction methods, clustering techniques and a measure of cluster tightness are presented, and the results of these experiments are discussed in section 4.

2. THE DISCRETE COSINE TRANSFORM

The discrete cosine transform (DCT) is an approximation to the Karhunen-Loeve Transformation, and is used to orthogonalize a given vector and reduce its dimensionality. The DCT represents a data sequence $x(n)$ in terms of its cosine series expansion with coefficients c_k, calculated as follows:

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

where $k = 0, \ldots, K\text{-}1$, n is the sequence sample index and K is the length of the input sequence $x(n)$.

Typically, the first few DCT coefficients contain most of the energy of the data sequence, and hence the DCT is often used for data compression applications such as speech and image coding. As seen in the example of Fig. 1, a data sequence (obtained at various time instants), such as a gene expression profile, can be modeled with reasonable accuracy using only the first two or three DCT coefficients. Hence, the DCT is a good tool for dimensionality reduction in this context, and it is this property that is exploited in the ensuing sections of this paper.

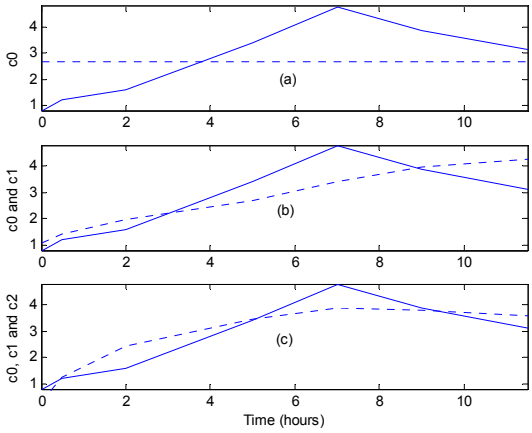

Figure 1. An example gene expression profile from the yeast sporulation data [1] (solid), and its reconstructions using the first one (a), two (b) and three (c) DCT coefficients (dashed).

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 slope of the data sequence $x(n)$. Thus, c_1 gives a rough approximation to the overall shape of the expression pattern for a gene.

The higher order DCT coefficients model detailed variations in the data sequence $x(n)$, which may include noise (removal of which is preferable for clustering purposes).

3. DIMENSIONALITY REDUCTION AND CLUSTERING

3.1. Yeast sporulation data

The techniques discussed in this paper were evaluated on the data set collected from experiments examining the sporulation of budding yeast by Chu *et al.* [1]. The data were derived from DNA microarrays containing nearly every yeast gene, with measurements of the mRNA levels made at non-equal intervals of 0, 0.5, 2, 5, 7, 9, and 11 hours.

Of the 6118 known and predicted genes in the original data set, more than 1000 showed significant changes in mRNA levels during sporulation, and of these about half were induced during that time [1]. A subset of only 512 genes were considered in this paper (similarly to [2]), comprising genes which are positively expressed and whose expression levels met a pre-determined threshold [1]. Thus the data set used in this work consisted of N = 512 genes with temporal profiles measured at 7 different time instants, i.e. K = 7 dimensions.

3.2. Dimensionality reduction

Analysis of the gene expression raw data set described in section 3.1 was then performed on the raw data itself, on the DCT coefficients of the raw data, and on the principal components of the raw data.

DCT coefficients were obtained by applying equation (1) to each temporal profile (vector) of the raw data, and forming four transformed data sets: one comprising the first coefficient c_0 of all vectors, another comprising the first two coefficients c_0 and c_1, another comprising the first three coefficients c_0, c_1 and c_2, and another comprising all seven DCT coefficients.

Another transformed data set was also obtained by taking the first two principal components of the raw data, following PCA.

Although the 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 with much larger length N and dimension K will benefit even more greatly from dimensionality reduction techniques such as these.

3.3. Clustering techniques

Clustering was then performed on the raw and transformed data sets described in section 3.2. In keeping with the original manual classification of the yeast sporulation data into seven temporal classes, comprising Metabolic, Early I, Early II, Early Middle, Middle, Mid Late and Late [1], the number of clusters M was chosen as seven.

In order to remove any dependency of this comparison upon a particular clustering technique, three techniques commonly used in the digital signal processing field (especially speech processing) were chosen:

- The LBG algorithm [6], with splitting performed one cluster at a time (clusters with largest variance are split first)
- Self-organising maps [4]
- The K-means algorithm [3]

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 3.4.

3.4. 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. Since the data will occupy different ranges depending on the type of dimensionality reduction technique employed, this measure was normalized according to the global standard deviation along each dimension. Thus,

$$CTM = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{K}\sum_{k=1}^{K}\frac{\sigma_k^m}{\sigma_k^G}\right), \qquad (2)$$

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.

Of the three clustering methods employed in section 3.3, the LBG algorithm is the only one that does not use random initialization. In order to improve the accuracy of *CTM* for self-organising maps and the *K*-means algorithm, clustering was repeated five times for these methods, and the cluster tightness measures averaged to provide the final estimate of *CTM*.

4. RESULTS

4.1. Comparison of DCT and PCA

The cluster tightness measurements described in section 3 yielded similar results for the reduction of the dimension from 7 to 2 using both DCT and PCA, as seen in Table 1. Clustering based on 2-dimensional DCT or PCA 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.

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

	Cluster tightness measure *CTM*		
Clustering features	LBG	SOM	*K*-means
Gene expression raw data	0.286	0.285	0.316
DCT coefficient c_0	0.567	0.436	0.546
DCT coefficients c_0, c_1	0.315	0.296	0.328
DCT coefficients c_0, c_1, c_2	0.300	0.294	0.333
All DCT coefficients	0.282	0.288	0.317
Two principal components	0.318	0.296	0.315

Experiments using one (c_0), two (c_0 and c_1), three (c_0, c_1 and c_2) and all DCT coefficients as the clustering features showed that while a single DCT coefficient is inadequate for clustering purposes, the use of c_0 and c_1 achieved approximately 10% of the cluster tightness obtained by clustering based on all DCT coefficients. This confirms the hypothesis from section 2 that the majority of the data sequence energy resides in the first two DCT coefficients for this particular data set.

A further result from Table 1 is that the LBG algorithm and self-organizing maps produce slightly tighter clusters for these data than the *K*-means algorithm.

4.2. Graphical comparison of DCT and PCA

From the results of section 4.1, it is reasonable to characterize the gene expression profiles using two DCT or PCA coefficients, and hence also to plot them in two dimensions. Figure 2 shows the first two principal components of the gene expression data clustered using *K*-means, similarly to Fig. 2B in [2].

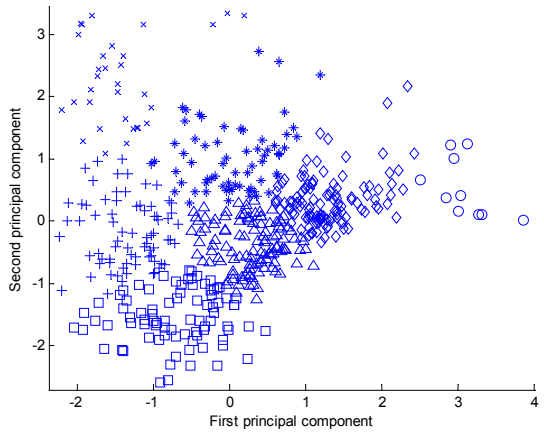

Figure 2. Gene expression data for yeast sporulation transformed using the first two principal components and clustered using the *K*-means algorithm.

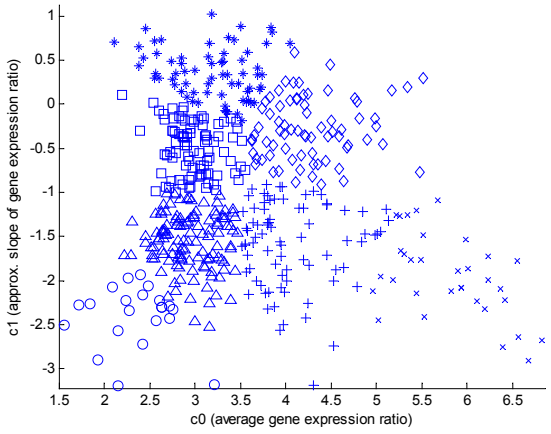

Figure 3. Gene expression data for yeast sporulation transformed using the first two DCT coefficients and clustered using the LBG algorithm.

By way comparison, Fig. 3 shows the first two DCT coefficients of the same data clustered using the LBG algorithm. Subjectively, the data organized using the DCT has classes at least as distinctly separate as in Fig. 2. Further, the DCT-based two-dimensional data visualization in Figure 3 offers axes that have physical interpretations (approximate slope of gene expression ratio vs. average gene expression ratio), unlike that of Figure 2.

4.3. Reordering of gene expression data using DCT coefficients

As a final demonstration of the application of the DCT to gene expression data, the original raw data set (Fig. 4(a)) was reordered using c_0 and c_1. The resulting clusters were then arranged in order of their temporal classes as illustrated in Fig. 4(b), showing the Metabolic, Early I, Early II, Early Middle, Middle, Mid Late and Late temporal classes from top to bottom. This figure should be compared with Fig. 5A in [1], which shows a similar result.

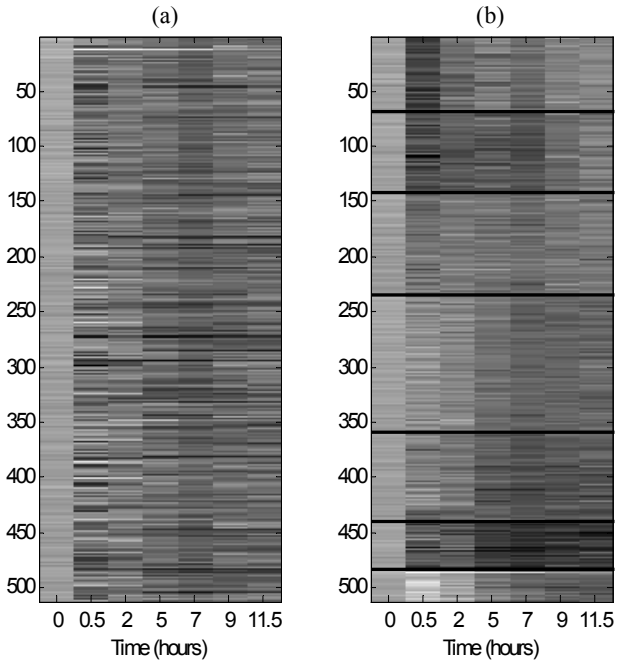

Figure 4. Genes induced during sporulation – (a) original unordered data set and (b) genes clustered using first two DCT coefficients, with clusters arranged by time of first induction. Solid black horizontal lines indicate cluster boundaries.

5. CONCLUSION

Using a measure of cluster tightness to evaluate clustered yeast sporulation gene expression data, this paper has shown that substantial reduction of the dimensionality is possible without adverse effects to the tightness of the clustering. In particular, this paper has shown that use of the DCT can provide similar clustering performance to PCA, with the additional benefits of lower computational complexity and the possibility of physical interpretation of the transformed data. Future work will concentrate on examining whether the demonstrated good performance of the DCT as a dimensionality reduction technique generalizes to other microarray data sets. Additionally, the applicability of other speech and image processing techniques such as GMMs and HMMs to microarray data will also be investigated.

6. REFERENCES

[1] Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., Herskowitz, I., "The transcriptional program of sporulation in budding yeast", *Science*, vol. 282, no. 5389, October 1998, pp. 699-705. Data available from http://cmgm.stanford.edu/pbrown/sporulation

[2] Datta, S., "Statistical techniques for microarray data: A partial overview", *Communications in Statistics-Theory and Methods*, vol. 32, 2003, pp. 263-280.

[3] Gersho, A., and Gray, R. M., *Vector Quantization and Signal Compression,* Kluwer Academic Publishers, Boston, 1993.

[4] Kohonen, T., *Self-organization and associative memory*, Springer-Verlag, Berlin, 1984.

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

[6] Linde, Y., Buzo, A., and Gray, R. M., "An algorithm for vector quantizer design", *IEEE Trans. Commun.*, vol. COM-28, no. 1, January 1980, pp. 84-95.

[7] Törönen, P., Kolehmainen, M., Wong, G., and Castrén, E., "Analysis of gene expression data using self-organising maps", *FEBS Letters*, vol. 451, 1999, pp. 142-146.

[8] 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., "Signal processing techniques in genomic engineering", *Proc. IEEE*, vol. 90, no. 12, December, 2003, pp. 1822-1833.