# A NOVEL COMBINED ICA AND CLUSTERING TECHNIQUE FOR THE CLASSIFICATION OF GENE EXPRESSION DATA

*Amrish Kapoor*[1], *Thomas Bowles*[2], *Jonathon Chambers*[2]

1 Centre of Digital Signal Processing (DSP) Research, King's College London, U.K.
2 Centre of DSP, Cardiff University, Wales, CF24 OYF, U.K. {chambersj@cf.ac.uk}

## ABSTRACT

This study presents an effective method of blindly classifying large amounts of gene expression data into biologically meaningful groups using a combination of independent component analysis (ICA) and clustering techniques. Specifically, we show that the genes can be classified blindly into several groups based solely on their expression profiles. These groups have a very close correspondence with benchmarks obtained by studies using domain knowledge. These results suggest that ICA can be a very useful pre-processing tool in blind gene classification, rather than using the resulting sources as the final model profiles.

## 1. INTRODUCTION

DNA microarray technology has revolutionised the study of gene expression. DNA microarrays are capable of measuring the expression (specifically transcription) levels of thousands of genes simultaneously, and thus enable genome-scale analysis. In order to realise the potential power of microarray experiments, novel methods are required to accurately extract pertinent information from vast datasets. An enduring question is to determine how different sets of genes work together, i.e. gene pathways, under different conditions. To this end, we aim to classify the genes into biologically meaningful groups in an attempt to understand their interworking under different conditions. This study shows that using ICA as a pre-processing tool, followed by clustering is a promising approach to achieving this goal. We have used the gene expression data of yeast sporulation, which have been collected by Chu et al. [1] and are publicly available at http://cmgm.stanford.edu/pbrown/sporulation. The data consists of expression levels of 6118 genes, measured at seven different times during sporulation – at 0.0, 0.5, 2.0, 5.0, 7.0, 9.0 and 11.5 hours. Hence, the dataset is organised as a matrix of 6118 rows (genes) and 7 columns (sampling instants), with real-valued entries. Experiments have shown that during sporulation, specific genes are active at certain times, which is reflected in a corresponding significant change in their expression values, either positively or negatively. If the known genes can be classified into certain groups based on their expression profiles, then functions of previously unknown genes can be inferred from their proximity to one of these groups. Several approaches have been employed to perform this classification, including gene clustering, principal component analysis (PCA) [2], self-organizing maps and ICA [4]. In this study, we examine a blind classification method based on a combination of clustering of genes and ICA pre-processing. The following sections are organized as follows. Section 2 shows how the gene data are classified into several groups using domain knowledge, providing a benchmark for assessing blind classification methods. Section 3 shows how the direct gene clustering and ICA-only based methods estimate the model profiles. In Section 4, we present our method for gene classification, followed by a comparison of results in Section 5. Finally, Section 6 contains conclusions and discussions. The estimated induction patterns in this study are plotted using different scales where necessary (due to scale ambiguities associated with ICA) in order to accommodate them within the same figure.

## 2. CLASSIFICATION USING DOMAIN KNOWLEDGE

Chu et al. [1], classified the gene data into several meaningful groups using domain knowledge. They hand picked seven small sets of genes, which are known representatives of induction patterns of genes belonging to each category. The expression profiles of these genes were averaged to obtain seven model induction patterns over time (Fig 1). We will consider these model patterns as benchmark or correct patterns. Using these benchmark patterns, all other genes are classified into one of seven categories based on their correlation with each model

0-7803-8874-7/05/$20.00 ©2005 IEEE

pattern. The genes within each category are also ordered according to the relative magnitudes of their correlation.

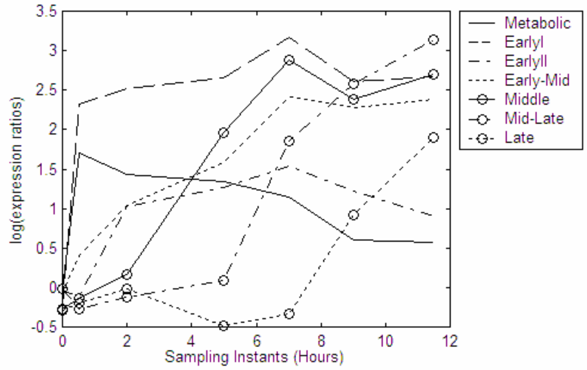

Fig. 1. Model (benchmark) profiles obtained by averaging the sets of representative genes

## 3. CLASSIFICATION USING CLUSTERS AND ICA

Here, we examine two known methods to classify the genes in our dataset, namely direct clustering of expression profiles and independent component analysis.

### 3.1. Direct Gene Clustering

One approach to generating model patterns is to directly cluster the genes' expression profiles. We present here the results obtained using this method. After removing the means of the expression profiles of all the 6118 genes, we applied k-means clustering to the entire data to obtain seven clusters. From Fig. 2, we can see that up to four clusters represent reasonable matches to some of the benchmark patterns shown earlier, namely those representing metabolic, earlyI, early-mid and middle categories. However, the other clusters do not seem to present any biologically meaningful results. Thus, we can estimate four of the seven model patterns, though only prior knowledge enables us to know which four patterns are meaningful.

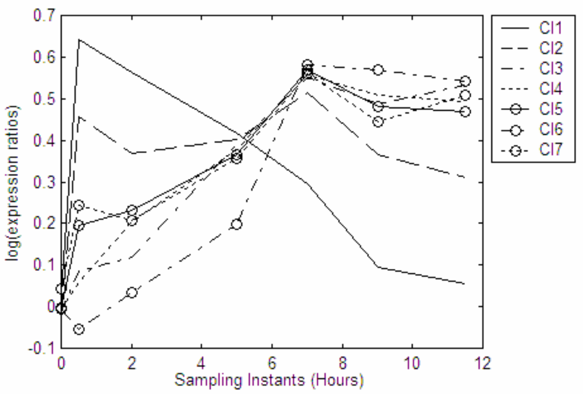

Fig 2. The clusters obtained after direct clustering of the gene data

### 3.2. Independent Component Analysis

In an attempt to improve upon the results obtained using direct gene clustering, the method of ICA was also applied to the same problem by Hori et al. [4]. This section partially uses their results. We represent the data matrix of size 6118x7 as A. In this method, ICA was applied to the expression data, using the following de-mixing,

$$Y = Wx$$

where 'x' is a 7-dimensional vector sampled from the transposed data matrix $A^T$. Hori et al. [4] used the JADE algorithm to obtain the 7x7 de-mixing matrix, W, and used the columns of the inverse of the de-mixing matrix as the estimates of the model induction patterns. The inverted relation is

$$x = W^{-1}Y$$

Hence, the columns of the inverse of the de-mixing matrix represent the required independent sources. Note here that in a normal ICA formulation, to estimate the sources as a matrix of size 7x7 would require calculation of a de-mixing matrix of size 7x6118. Instead Hori et al. only estimated an unmixing matrix of size 7x7, and used its inverse as the required matrix of source signals. Thus, they estimated the independent sources without calculating the large de-mixing matrix Y of size 7x6118. This is equivalent to the 'scalp map' in the case of EEG signals de-mixing. A consequence of estimating the independent components as described is that they must be linearly independent, since the de-mixing matrix W must necessarily have full row rank. The columns of the inverse de-mixing matrix were then used as estimates of the benchmark patterns, and are shown in Fig. 3. Hori et al. [4] used only three independent components to classify the genes to eliminate any ill effects from the components with small magnitudes. Hence, the calculated de-mixing matrix is a 3x7 matrix, and the columns of its 7x3 pseudo inverse matrix were used as the model induction patterns. Thus, even the ICA-only method can reproduce just three of the seven model patterns.

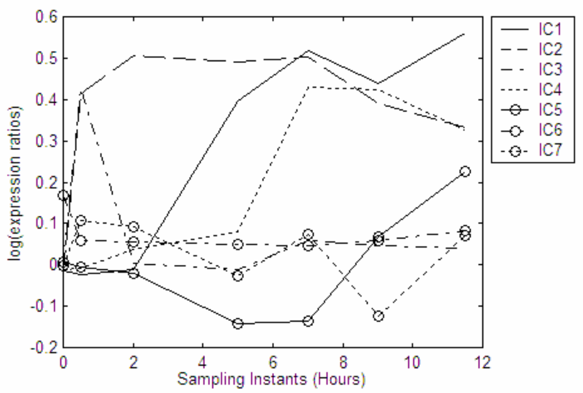

Fig. 3. Columns of the inverse of de-mixing matrix W

## 4. CLASSIFICATION USING CLUSTERING OF ICA PRE-PROCESSED DATA

The ICA-only based approach implicitly assumes that the observed expression data are generated by linearly combining the different independent components (these components are taken as estimates of the model induction patterns). This model does not accurately reflect the relation between the benchmark patterns and the observed data. Chu et al. [1] had calculated the model induction patterns by averaging the expression profiles of a small set of representative genes. This would naturally suggest a clustering based approach as the preferred model for the relation between the benchmark patterns and the observed expression data matrix. But the clustering based approach also does not yield much better results in this case, as shown earlier. This is because the data is very widely spread out (has large variance from the benchmark patterns), and hence the distinctions between the different categories are very hazy, and difficult to determine. We propose to use ICA as an important pre-processing step, after which we perform clustering to extract estimates of model induction patterns.

We follow the ICA method outlined in the previous section, proposed by Hori et al. [4], to generate the three most representative independent components (which are also linearly independent as described earlier, and hence can be interpreted as a basis for the seven-dimensional vector space containing all the 6118 gene expression patterns as data points – we use this next, in our study). Unlike in [4], we do not use these components as the estimates of the model induction patterns. Instead we use the independent components to define a new vector space, which is important in the pre-processing of our data. Let us denote by M, the 7x3 pseudo-inverse of the de-mixing matrix, W, described in the previous section. We postulate that the columns of this matrix, M, form a basis for a space that very accurately contains the benchmark induction patterns. Thus, each model induction pattern can be represented as a linear combination of the columns of M. Now each gene is projected onto the space spanned by the columns of M. This is achieved by computing the product:

$$M\, x_{LS}$$

where $x_{LS}$ is a 3-component vector, calculated as

$$x_{LS} = (M^TM)^{-1}M^Tb$$

and b is the gene being projected.

Note that projection onto the column space of M requires the computation of its pseudo-inverse, which results again in the original de-mixing matrix W, i.e. $(M^TM)^{-1}M^T = W$.

We now perform clustering of the projected data. Instead of clustering the projected genes themselves (the product $M\, x_{LS}$), we perform clustering on the 3-component feature vectors, $x_{LS}$, obtained when projecting each gene. The new vector space is spanned by the three main (linearly) independent components of the dataset - it is a space that captures the 'essential characteristics' of the data. The benchmark patterns are an indication of the main 'shapes' present in the data. Hence, since the new space captures the essential characteristics of the data, it should accurately contain the main variants in the data. When the rest of the data is mapped into this space, the large variants or 'outliers' that make direct clustering difficult, tend to fall into the nearest model pattern. This makes the separation sharper, thus making the data more amenable to clustering. It should be noted that as more and more independent components are used to form the vector space, it results in poorer performance since the space becomes more 'general'. The use of ICA here is a crucial pre-processing step. A similar basis for a new space can also be obtained by computing the singular value decomposition (SVD) of the expression data. But the basis of the space formed using the SVD yields inferior results to those obtained using ICA. This is possibly because ICA returns components that are 'as independent as possible', without requiring them to be orthogonal, and hence captures the underlying nature of the data very effectively in only a few components. Another popular pre-processing method is PCA, but this also results in poor estimates. In fact, Yeung et al. [3] showed empirically, working on this same dataset, that clustering with PCs instead of the original variables does not necessarily improve, and often degrades cluster quality. The clustering of the mapped genes results in patterns that are much more biologically meaningful than those obtained by earlier methods. The genes can then be classified into one of seven groups according to the correlation coefficient between the gene and each group. All seven cluster patterns thus obtained, shown in Fig 4, appear to be very close to the benchmark patterns shown in Section 1, as compared to only three and four obtained using ICA-only and direct clustering respectively. This is an impressive result for a purely blind method.

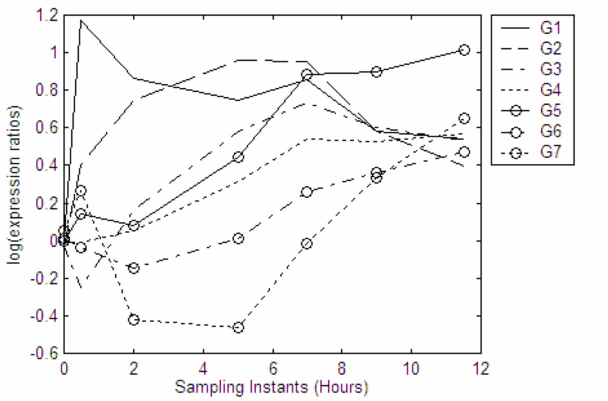

Fig. 4. The seven cluster patterns obtained using clustering of ICA-pre-processed data

## 5. COMPARISON OF RESULTS

In this section we visually demonstrate the superior separation obtained when using ICA as the pre-processing method prior to clustering. As explained in the previous section, the three-dimensional projection co-ordinates are calculated for the top 200 genes from each model class as identified by Chu et al. [1]. Similarly, projection co-ordinates are also computed along axes obtained using principal component analysis (PCA), as in [2] and SVD. The resultant plots are shown in Fig 5.

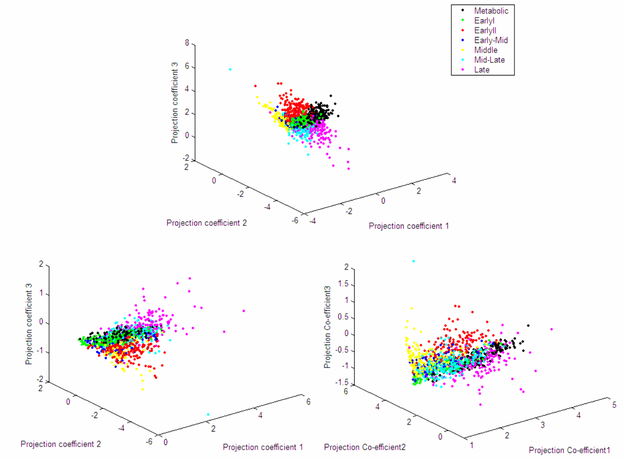

Fig. 5. Projection co-ordinates of top 200 genes from each model class, projected along ICA axes (top), PCA axes (bottom left) and SVD axes (bottom right)

Figure 5 clearly shows the improved separation between the classes obtained using ICA as a pre-processing method as compared to PCA or SVD (these plots are best viewed in full colour). Earlier, we have already established the suitability of employing clustering as the preferred physical model for data generation, and now we have also shown how clustering can be improved by significantly enhancing the class distinctions using ICA.

Finally, we also compare the seven clusters obtained using this method with the benchmark patterns. The correlation coefficient between each benchmark pattern and its nearest cluster estimate were computed. Once the patterns have been estimated, they can be used to classify all the genes into different categories. A number of methods have been employed to perform this classification, and is, in itself another research topic. We restrict ourselves to trying to blindly estimate the benchmark patterns as closely as possible. The best matches among the clusters, along with the correlation coefficients were:

Metabolic (0.9538), Early I (0.8295), Early II (0.9253), Early-Mid (0.9478), Middle (0.9863), Mid-Late (0.9788), Late (0.7670 – this increases to 0.9231 if only the biologically important points, i.e. the positive values representing gene induction, are considered.)

Not only do the clusters represent excellent matches for the benchmark patterns but, crucially, each cluster is mapped to a distinct model pattern, i.e. all seven clusters are biologically meaningful. In contrast, using only ICA or direct clustering produces just three and four estimates:

ICA only: Meta (0.73), EarlyI (0.79), Mid (0.98).

Direct Clustering: Meta (0.94), EarlyI (0.93), Mid (0.99), Early-Mid (0.99).

## 6. CONCLUDING REMARKS

This study has shown that ICA is a useful technique, but the sources cannot be taken as the final model estimates in classification of genes based on their expression patterns. Such a model does not accurately reflect the underlying physical model of data generation. However, it can form a very effective pre-processing step, producing a vector space onto which the data can be projected so as to enhance separability. This blind classification method uses no domain knowledge and results in profiles that are impressively close to those obtained by handpicking.

We note that this method has only been tested on yeast sporulation data. Though the results are very promising, its performance on other datasets needs to be examined. Secondly, we have only used k-means clustering to obtain the model induction patterns, using the Euclidean distance as the criterion for grouping data in the projected space. It would be interesting to try other clustering methods, such as hierarchical clustering or dynamic k-means, and also using other criteria for grouping.

In conclusion, we emphasize that our method reflects the underlying model for data generation, and significantly outperforms the other blind classification methods.

## 7. REFERENCES

[1] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D.Botstein, P.O. Brown and I. Herskowitz, "The transcriptional program of sporulation in budding yeast", *Science, Vol. 282*, pp. 699-705, 1998.

[2] S. Raychaudhuri, J.M. Stuart and R.B. Altman, "Principal Component Analysis to summarize microarray experiments: Application to sporulation time series", *Pacific Symposium on Biocomputing, 5*, pp. 452-463, 2000.

[3] K.Y. Yeung and W.L. Ruzzo, "An empirical study of PCA for clustering gene expression data", *Bioinformatics 2001., Vol. 17*, pp. 763-774, 2001.

[4] G. Hori, M. Inoue, S. Nishimura and H. Nakahara, "Blind gene classification based on ICA of microarray data", *ICA2001*, San Diego, USA, pp. 332-336, 2001.