BIOINFORMATICS ORIGINAL PAPER
Vol. 23 no. 8 2007, pages 998–1005
doi:10.1093/bioinformatics/btm053

Gene expression

A multi-stage approach to clustering and imputation of gene expression profiles

Dorothy S. V. Wong*, Frederick K. Wong and Graham R. Wood
Department of Statistics, Macquarie University, NSW 2109, Australia

Received on July 30, 2006; revised on January 23, 2007; accepted on February 10, 2007
Advance Access publication February 18, 2007
Associate Editor: Martin Bishop

ABSTRACT

Motivation: Microarray experiments have revolutionized the study of gene expression with their ability to generate large amounts of data. This article describes an alternative to existing approaches to clustering of gene expression profiles; the key idea is to cluster in stages using a hierarchy of distance measures. This method is motivated by the way in which the human mind sorts and so groups many items. The distance measures arise from the orthogonal breakup of Euclidean distance, giving us a set of independent measures of different attributes of the gene expression profile. Interpretation of these distances is closely related to the statistical design of the microarray experiment. This clustering method not only accommodates missing data but also leads to an associated imputation method.

Results: The performance of the clustering and imputation methods was tested on a simulated dataset, a yeast cell cycle dataset and a central nervous system development dataset. Based on the Rand and adjusted Rand indices, the clustering method is more consistent with the biological classification of the data than commonly used clustering methods. The imputation method, at varying levels of missingness, outperforms most imputation methods, based on root mean squared error (RMSE).

Availability: Code in R is available on request from the authors.

Contact: dwong@efs.mq.edu.au

1 INTRODUCTION

Microarray experiments allow us to measure the expression of tens of thousands of genes simultaneously, thus having the potential to dramatically increase the efficiency of genome-wide studies. Following the conduct of a microarray experiment, a primary concern of the researcher is the appropriate grouping of similarly expressed genes. The biological motivation for performing clustering lies in the fact that many co-expressed genes are also co-regulated; clustering aids in functional annotation of novel genes, identification of transcription factor binding sites and discovery of complete biological pathways (Boutros and Okey, 2005). A secondary, but related, concern is the need for imputation of missing data. Gene expression profiles, especially those obtained from microarray chips, often include a substantial number of missing values.

*To whom correspondence should be addressed.

Techniques such as hierarchical clustering (Eisen *et al.*, 1998), *k*-means (Soukas *et al.*, 2000), Cluster affinity search technique (CAST) (Ben-Dor *et al.*, 1999), gene shaving (Hastie *et al.*, 2000), the use of self-organizing maps (SOM) (Tamayo *et al.*, 1999), self-organizing tree algorithms (SOTA) (Herrero *et al.*, 2001) and mixture models (McLachlan *et al.*, 2002; Yeung *et al.*, 2001) to name a few, have been used in the clustering of gene expression profiles. In practice, the most common clustering methods used by biologists for gene expression data (Knudsen, 2002) are hierarchical clustering, *k*-means and SOM. Hierarchical clustering links the genes, based on closest distance, to form a 'family tree'. The *k*-means method starts by randomly assigning each gene to one of *k* clusters. The distance between each gene and each cluster centre (or centroid) is calculated and used to assign genes to the closest centroid. The genes assigned to a centroid become a new cluster. The centroids are then recalculated and genes reassigned until the centroids converge. The SOM method is similar to *k*-means, the difference being that it is constrained to work on a 2D grid that provides information about the relationship between neighbouring clusters. SOTA is a hierarchical SOM, clustering using the hierarchical structure with the accuracy and robustness of a neural network. Most clustering techniques, however, are unable to deal with missing data. Samples containing missing data must be omitted or the values imputed.

de Brevern *et al.* (2004) have shown that the imputation method used affects the final clustering, even at a low rate of missingness. Therefore, choosing an appropriate imputation method is a crucial step in the analysis of gene expression data. Generally, we can categorize imputation methods into two classes: the first uses local information and the second uses global information. The two methods proposed initially by Troyanskaya *et al.* (2001), namely the *k*-nearest neighbour (KNN) and singular value decomposition (SVD) imputation methods are the respective pioneers in these two categories.

The KNN imputation method uses information from the *k*-nearest neighbours to estimate the missing value. Subsequent articles belonging to this category further developed this idea by either altering the gene selection process or the design of the estimation rule. The KNN method uses Euclidean distance for gene selection and a weighted average (with weights determined by gene similarity) for the estimation rule. Improvements in

gene selection include the use of Bayesian variable selection (Zhou *et al*., 2003), Gaussian mixture clustering (Ouyang *et al*., 2004) or correlation (Bo *et al*., 2004). Advancements in the estimation rule include the use of linear models (Scheel *et al*., 2005; Zhou *et al*., 2003), non-linear models (Zhou *et al*., 2003), the (Expectation-maximization) EM-algorithm (Bo *et al*., 2004; Ouyang *et al*., 2004) or least squares methods (Bo *et al*., 2004; Kim *et al*., 2005; Nguyen *et al*., 2004).

The SVD imputation method uses singular value decomposition to obtain mutually orthogonal expression patterns that can be linearly combined to approximate the expression of all genes in the dataset. Estimates of the missing values can be obtained by regressing against this set of genes. Here, a set of genes which can represent the entire dataset is selected and used to estimate the missing value. Further development in this category includes the introduction of Bayesian estimation into principal component analysis (Oba *et al*., 2003), partial least squares (Nguyen *et al*., 2004), a covariance-based method to rank genes (Sehgal *et al*., 2005) and support vector regression (Wang *et al*., 2006).

Other methods are either a variation on, or a combination of, the above categories. These include a sequential KNN method (Kim *et al*., 2004) which uses previously imputed values to impute subsequent missing values, use of a convex combination of existing methods (Jornsten *et al*., 2005), use of information about the quality of the spots (Tom *et al*., 2005) and use of information from gene ontology (Tuikkala *et al*., 2006). To date, the KNN approach is the most widely used imputation method due to its simplicity, efficiency and availability.

This article describes an alternative approach to the clustering of microarray data; the method accommodates missing data and also leads to an associated imputation method. This method is adapted from Godfrey *et al*. (2002), where it has been successfully used in a horticultural context for clustering in genotype-by-environment analyses with missing data. The method outperforms commonly used clustering methods while retaining their simplicity. The associated imputation method also produces promising results.

Section 2 describes the method by first detailing the derivation of the distance measures. This is followed by a modification of the distance measures to accommodate missing data. As an aside, the relationship between the distance measures and the experimental design used is presented. We then describe the clustering and imputation algorithms and introduce the 'jump factor' as a stopping criterion. In Section 3, the results of clustering and imputation using both two-stage and three-stage methods are presented. A short discussion is provided in Section 4 and a brief conclusion given in Section 5.

2 METHODS

The clustering method introduced is based on the simple idea of grouping in stages using a hierarchy of distance measures. The method captures the way in which the human mind sorts and thus groups items using a hierarchy of attributes, so increasing the probability of success. Consider, for example, how we tackle a jigsaw puzzle. It is common to sort the pieces into groups at the outset; sorting may require a number of stages, depending on the complexity of the puzzle. For example, we might first group the pieces based on shape into edge and non-edge pieces. Within these groups, we then sort based on colour. Similarly here, the more complex the design of the experiment, the greater the number of stages required for clustering. The situation can be modelled probabilistically, demonstrating that the probability of accurately grouping the items is always higher when done in stages than when done all at once.

2.1 Distance measures

The distance measures giving rise to the stages are the result of breaking Euclidean distance into a number of orthogonal components. For observations y_s, $s = 1, 2, \ldots, S$, we commonly have a decomposition of the total sum of squares into, say, n orthogonal components, as

$$\sum_{s=1}^{S} y_s^2 = C_1^2 + C_2^2 + \cdots + C_n^2$$

Let y_{is} denote the gene expression of the ith gene and sth sample. The squared Euclidean distance between the ith and jth gene, E_{ij}^2, is

$$E_{ij}^2 = \sum_{s=1}^{S} (y_{is} - y_{js})^2$$

Here we consider $y_{is} - y_{js}$ as our observation, so squared Euclidean distance is the total sum of squares of the given observations. Thus, we can commonly obtain an orthogonal decomposition of the squared Euclidean distance between the ith and jth gene, E_{ij}^2, as

$$E_{ij}^2 = D_1^2 + D_2^2 + \cdots + D_n^2$$

Each D_k, where $k = 1, 2, \ldots, n$, corresponds to a certain attribute of the data. Interpretation of D_k is on a case-by-case basis.

2.1.1 Two-stage decomposition Here we describe the simplest situation where there are only two components. Based on the very simple model

$$y_s = \mu + \epsilon_s$$

for $s = 1, 2, \ldots, S$, the total sum of squares can be expressed as

$$\sum_{s=1}^{S} y_s{}^2 = S\bar{y}^2 + \sum_{s=1}^{S} (y_s - \bar{y})^2$$

where $\bar{y}$ is the sample mean.

Paralleling this, the partitioned squared Euclidean distance between the ith and jth gene is given as

$$\begin{aligned} E_{ij}^2 &= \sum_{s=1}^{S} (y_{is} - y_{js})^2 \\ &= S(\bar{y}_{i\cdot} - \bar{y}_{j\cdot})^2 + (S-1)\frac{\sum_{s=1}^{S} \left((y_{is} - \bar{y}_{i\cdot}) - (y_{js} - \bar{y}_{j\cdot})\right)^2}{S-1} \end{aligned}$$

where $\bar{y}_{i\cdot}$ and $\bar{y}_{j\cdot}$ are the means of the y_{is} and the y_{js} respectively across all S samples.

Thus, we have partitioned the squared Euclidean distance between the ith and jth expression profiles into two squared distance measures,

$$E_{ij}^2 = D_1^2 + D_2^2$$

Here, D_1 is a multiple of a main effect distance M_{ij}, where

$$M_{ij} = |\bar{y}_{i\cdot} - \bar{y}_{j\cdot}|$$

and D_2 is a multiple of an interaction distance I_{ij}, where

$$I_{ij} = \sqrt{\frac{\sum_{s=1}^{S} \left((y_{is} - \bar{y}_{i\cdot}) - (y_{js} - \bar{y}_{j\cdot})\right)^2}{S-1}}$$

Table 1. Unbalanced one-way ANOVA

Source of variation	df	SS
Overall mean	1	$S\bar{y}_{..}^2$
Treatments	$T-1$	$\sum_{t=1}^{T} S_t(\bar{y}_{t\cdot} - \bar{y}_{..})^2$
Error	$S-T$	$\sum_{t=1}^{T}\sum_{s=1}^{S_t}(y_{ts} - \bar{y}_{t\cdot})^2$
Total	S	$\sum_{t=1}^{T}\sum_{s=1}^{S_t} y_{ts}^2$

In summary, squared Euclidean distance can be written as

$$E_{ij}^2 = SM_{ij}^2 + (S-1)I_{ij}^2$$

2.1.2 Three-stage decomposition Using a more elaborate model, we can extend the two-stage method to a three-stage method. Suppose we have T treatments within a gene and S_t repetitions (samples) within treatment t. For a given gene, we can express this in the model

$$y_{ts} = \mu + \tau_t + \epsilon_{ts}$$

where $t = 1, 2, \ldots, T$ and $s = 1, 2, \ldots, S_t$. Letting $S = \sum_{t=1}^{T} S_t$, analysis of this model can be summarized using the unbalanced one-way analysis of variance (ANOVA) table shown in Table 1. Thus, the total sum of squares of the y_{ts} can be partitioned into three orthogonal components as

$$\sum_{t=1}^{T}\sum_{s=1}^{S_t} y_{ts}^2 = S\bar{y}_{..}^2 + \sum_{t=1}^{T} S_t(\bar{y}_{t\cdot} - \bar{y}_{..})^2 + \sum_{t=1}^{T}\sum_{s=1}^{S_t}(y_{ts} - \bar{y}_{t\cdot})^2$$

Letting y_{its} and y_{jts} denote the expression values for the ith and jth genes in the tth treatment and sth sample, the squared Euclidean distance between the ith and jth expression profiles can be expressed as

$$\begin{aligned} E_{ij}^2 &= \sum_{t=1}^{T}\sum_{s=1}^{S_t}(y_{its} - y_{jts})^2 \\ &= S(\bar{y}_{i\cdot\cdot} - \bar{y}_{j\cdot\cdot})^2 \\ &\quad + (T-1)\frac{\sum_{t=1}^{T} S_t\left((\bar{y}_{it\cdot} - \bar{y}_{i\cdot\cdot}) - (\bar{y}_{jt\cdot} - \bar{y}_{j\cdot\cdot})\right)^2}{T-1} \\ &\quad + (S-T)\frac{\sum_{t=1}^{T}\sum_{s=1}^{S_t}\left((y_{its} - \bar{y}_{it\cdot}) - (y_{jts} - \bar{y}_{jt\cdot})\right)^2}{S-T} \end{aligned}$$

where $\bar{y}_{i\cdot\cdot}$ is the mean of the y_{its} across all S samples and $\bar{y}_{it\cdot}$ is the mean of the y_{its} for the samples within treatment t. In summary, we have partitioned the squared Euclidean distance into three squared distance measures,

$$E_{ij}^2 = D_1^2 + D_2^2 + D_3^2$$

Here, D_1 is a multiple of a main effect distance M_{ij}, where

$$M_{ij} = |\bar{y}_{i\cdot\cdot} - \bar{y}_{j\cdot\cdot}|,$$

D_2 is a multiple of a treatment effect distance T_{ij}, where

$$T_{ij} = \sqrt{\frac{\sum_{t=1}^{T} S_t\left((\bar{y}_{it\cdot} - \bar{y}_{i\cdot\cdot}) - (\bar{y}_{jt\cdot} - \bar{y}_{j\cdot\cdot})\right)^2}{T-1}}$$

and D_3 is a multiple of an interaction effect distance I_{ij}, where

$$I_{ij} = \sqrt{\frac{\sum_{t=1}^{T}\sum_{s=1}^{S_t}\left((y_{its} - \bar{y}_{it\cdot}) - (y_{jts} - \bar{y}_{jt\cdot})\right)^2}{S-T}}$$

Thus, squared Euclidean distance can be decomposed as

$$E_{ij}^2 = SM_{ij}^2 + (T-1)T_{ij}^2 + (S-T)I_{ij}^2$$

2.2 Distance measures accommodating missing values

The distance measures can be modified to accommodate missing values. This involves calculating the squared Euclidean distance over samples common to both genes. We let s_{ij} denote the indices of samples where values for genes i and j are present.

2.2.1 Two-stage decomposition accommodating missing values Let p_{ij} be the number of samples common to genes i and j and $\bar{y}_{i\cdot}^{(j)}$ be the mean of the y_{is} across these p_{ij} common samples. The orthogonal partition of the squared Euclidean distance using only common samples will be

$$\begin{aligned} E_{ij}^2 &= \sum_{s_{ij}}(y_{is} - y_{js})^2 \\ &= p_{ij}(\bar{y}_{i\cdot}^{(j)} - \bar{y}_{j\cdot}^{(i)})^2 \\ &\quad + (p_{ij}-1)\frac{\sum_{s_{ij}}\left((y_{is} - \bar{y}_{i\cdot}^{(j)}) - (y_{js} - \bar{y}_{j\cdot}^{(i)})\right)^2}{p_{ij}^{-1}} \end{aligned}$$

Thus, we have a main effect distance

$$\frac{M_{ij} = |\bar{y}_{i\cdot}^{(j)} - \bar{y}_{j\cdot}^{(i)}|}{p_{ij}^{-1}}$$

and an interaction distance

$$I_{ij} = \sqrt{\frac{\sum_{s_{ij}}\left((y_{is} - \bar{y}_{i\cdot}^{(j)}) - (y_{js} - \bar{y}_{j\cdot}^{(i)})\right)^2}{p_{ij}-1}}$$

that accommodate missing values.

2.2.2 Three-stage decomposition accommodating missing values Let s_{ijt} be the indices of samples in treatment t where values for genes i and j are present, p_{ijt} be the number of samples common to genes i and j in treatment t, $\bar{y}_{it\cdot}^{(j)}$ be the mean of the y_{its}, using only the p_{ijt} common samples and $\bar{y}_{i\cdot\cdot}^{(j)}$ be the overall mean of the y_{its}, using only common samples between gene i and gene j. Letting $p_{ij} = \sum_{t=1}^{T} p_{ijt}$, we can express $y_{it\cdot}^{(j)}$ and $y_{i\cdot\cdot}^{(j)}$ as

$$\bar{y}_{it\cdot}^{(j)} = \frac{\sum_{s_{ijt}} y_{its}}{p_{ijt}}$$

$$\bar{y}_{i\cdot\cdot}^{(j)} = \frac{\sum_{t=1}^{T}\sum_{s_{ijt}} y_{its}}{p_{ij}}$$

The squared Euclidean distance can then be expressed as

$$\begin{aligned} E_{ij}^2 &= \sum_{t=1}^{T}\sum_{s_{ijt}}(y_{its} - y_{jts})^2 \\ &= p_{ij}(\bar{y}_{i\cdot\cdot}^{(j)} - \bar{y}_{j\cdot\cdot}^{(i)})^2 \\ &\quad + (T-1)\frac{\sum_{t=1}^{T} p_{ijt}\left((\bar{y}_{it\cdot}^{(j)} - \bar{y}_{i\cdot\cdot}^{(j)}) - (\bar{y}_{jt\cdot}^{(i)} - \bar{y}_{j\cdot\cdot}^{(i)})\right)^2}{T-1} \\ &\quad + (p_{ij}-T)\frac{\sum_{t=1}^{T}\sum_{s_{ijt}}\left((y_{ts} - \bar{y}_{it\cdot}^{(j)}) - (y_{jts} - \bar{y}_{jt\cdot}^{(i)})\right)^2}{p_{ij}-T} \end{aligned}$$

As a result, we have a main effect distance, M_{ij}, where

$$M_{ij} = |\bar{y}_{i\cdot\cdot}^{(j)} - \bar{y}_{j\cdot\cdot}^{(i)}|$$

a treatment effect distance, T_{ij}, where

$$T_{ij} = \sqrt{\frac{\sum_{t=1}^{T} p_{ijt}\left((\bar{y}_{it\cdot}^{(j)} - \bar{y}_{i\cdot\cdot}^{(j)}) - (\bar{y}_{jt\cdot}^{(i)} - \bar{y}_{j\cdot\cdot}^{(i)})\right)^2}{T-1}}$$

and an interaction effect distance, I_{ij}, where

$$I_{ij} = \sqrt{\frac{\sum_{t=1}^{T}\sum_{s_{ijt}}\left((y_{its} - \bar{y}_{it\cdot}^{(j)}) - (y_{jts} - \bar{y}_{jt\cdot}^{(i)})\right)^2}{p_{ij} - T}}$$

that accommodate missing values.

2.3 Relationship between the distance measures and the gene expression model

In this section, we show how the distance measures obtained from the two-stage decomposition are related to a model for gene expression data. This can be extended to the distance measures obtained from higher order decompositions and the associated models.

We consider a two-factor factorial design with no replicates. Let Y_{is} denote the gene expression for the ith gene and sth sample. An appropriate model for this design is

$$Y_{is} = \mu + G_i + S_s + (GS_{is} + \epsilon_{is})$$

where G_i denotes the gene effect, S_s denotes the sample effect, GS_{is} is the gene-by-sample ($G\times S$) interaction and ϵ_{is} is the error term assumed to be independently and normally drawn with mean zero and variance σ^2. Since there are no replicates, GS_{is} and ϵ_{is} are confounded. The squared Euclidean distance between the expression profiles for genes i and j is

$$E_{ij}^2 = \sum_{s=1}^{S}\left((G_i - G_j) + (GS_{is} - GS_{js}) + (\epsilon_{is} - \epsilon_{js})\right)^2$$

This is a combination of the main effect difference between genes i and j, the $G\times S$ interaction difference between genes i and j and the error difference between genes i and j.

It can be shown that $SM_{ij}^2/2\sigma^2$ follows a non-central χ^2 distribution with one degree of freedom and non-centrality parameter $S(G_i - G_j)^2/2\sigma^2$. Therefore, the expected value of M_{ij}^2 is $(2\sigma^2/S) + (G_i - G_j)^2$. This is a translation of the squared difference in gene expression level. Moreover, the translation value, $2\sigma^2/S$, is usually small relative to $(G_i - G_j)^2$. Consequently, M_{ij}^2 serves as a satisfactory measure of difference in gene level.

Also, $(S-1)I_{ij}^2/2\sigma^2$ follows a non-central χ^2 distribution with $S-1$ degrees of freedom and non-centrality parameter $(\sum_{s=1}^{S}(GS_{is} - GS_{js})^2)/2\sigma^2$. Thus, the expected value of I_{ij}^2 is $2\sigma^2 + (\sum_{s=1}^{S}(GS_{is} - GS_{js})^2)/(S-1)$. Let $GS_{is} - GS_{js}$, the difference between the $G\times S$ interaction of the ith and jth gene, be considered as an observation. Note that $GS_{is} - GS_{js}$ across the S samples has mean zero, whence $(\sum_{s=1}^{S}(GS_{is} - GS_{js})^2)/(S-1)$ is the sample variance of the $GS_{is} - GS_{js}$. The value of this variance gives us a good indication of the extent of the difference in $G\times S$ interaction between the ith and jth genes. The expected value of I_{ij}^2 is a translation of this variance by $2\sigma^2$. Consequently, I_{ij}^2 serves as a satisfactory measure of difference in $G\times S$ interaction.

2.4 Clustering

The idea underlying the clustering method proposed is to group in stages using a hierarchy of distance measures. The hierarchy begins with the most dominant attribute and progresses through to finer attributes. This mimics the way mails are sorted, for example firstly by country then by state, down to postcode and so on. We can summarize the clustering algorithm as follows:

Stage 1: Cluster using D_1

Stage 2: Cluster within each first-stage cluster using D_2

Stage 3: Cluster within each second-stage cluster using D_3.

In general, to cluster in n stages, first cluster using D_1 then cluster within each stage $i-1$ cluster using D_i, for i from 2 to n. For all stages, we use hierarchical agglomerative clustering with Ward's linkage method. Ward's linkage method combines the two clusters which minimize the increase in total error sum of squares (ESS). The ESS of a cluster is the sum of squares of the deviations from the mean value.

2.4.1 Two-stage clustering We begin by describing the two-stage clustering method, with distance measures of main effect distance and interaction distance. This is summarized as follows:

First stage

(1) Calculate all main effect distances M_{ij}
(2) Cluster genes using these main effect distances (this produces level-similar clusters).

Second stage

(1) Calculate interaction distances I_{ij} for all gene pairs i and j within first-stage clusters
(2) Cluster genes within each first-stage cluster using the interaction distances (this produces level-similar and shape-similar gene clusters).

2.4.2 Three-stage clustering When we have three distance measures, namely the main effect, treatment and interaction distances, a three-stage clustering algorithm can be summarized as follows:

First stage

(1) Calculate all main effect distances M_{ij}
(2) Cluster genes using these main effect distances (this produces level-similar clusters).

Second stage

(1) Calculate treatment distances T_{ij} for all gene pairs i and j within first-stage clusters
(2) Cluster genes within each first-stage cluster using the treatment distances (this produces level-similar and treatment shape-similar gene clusters).

Third stage

(1) Calculate interaction distances I_{ij} for all gene pairs i and j within second-stage clusters
(2) Cluster genes within each second-stage cluster using the interaction distances (this produces level-similar, treatment shape-similar as well as interaction shape-similar gene clusters).

2.5 Imputation

For imputation, we cluster using interaction distance modified to accommodate missing values. We use information from the genes in the cluster of the gene with missing data to find an imputed value.

2.5.1 Two-stage imputation

(1) Perform clustering using only interaction distance
(2) For each missing value, identify the gene and the sample to which it corresponds (we call these the target gene and the target sample)
(3) Identify the genes that belong to the same interaction cluster as the target gene (we call these the parent genes)
(4) For each parent gene with an expression value in the target sample, calculate the corresponding overall mean
(5) Find the difference between the expression value in the target sample and the calculated overall mean
(6) Calculate the mean of all values obtained in Step 5
(7) The imputed value is the overall mean of the target gene plus the value calculated in Step 6.

2.5.2 Three-stage imputation

(1) Perform clustering using only interaction distance
(2) For each missing value, identify the gene, the sample and the treatment to which it corresponds (we call these the target gene, the target sample and the target treatment)
(3) Identify the genes that belong to the same interaction cluster as the target gene (we again call these the parent genes)
(4) For each parent gene with an expression value in the target sample, calculate the corresponding target treatment mean
(5) Find the difference between the expression value in the target sample and the calculated target treatment mean
(6) Calculate the mean of all values obtained in Step 5
(7) The imputed value is the target treatment mean of the target gene plus the value calculated in Step 6.

2.6 Stopping criterion

A critical challenge is to determine an appropriate number of clusters when no prior knowledge is available. To identify the appropriate number of clusters, we plot the height of the new cluster to be formed against the current number of clusters. Height here corresponds to the criterion used to determine which two clusters are to be merged to form a new cluster. For example, we used Ward's linkage method where height is the total ESS after merging two clusters. Since a hierarchical agglomerative technique starts with each data point a cluster, then at each iterative step joins the two closest clusters, the height of the new cluster will be the largest height calculated so far. We propose that clustering should stop when the height increases markedly. As a quantitative measure of this, we use a 'jump factor' defined as

$$\frac{\text{current height increase}}{\max\{\text{previous height increases}\}}$$

The appropriate number of clusters is that just before the maximum jump factor occurs. Intuitively, this ensures that we stop just before we merge strongly resistant clusters.

3 RESULTS

In this section, we first illustrate the performance of the two-stage and the three-stage clustering methods. We then demonstrate the performance of the two-stage and the three-stage imputation methods.

3.1 Clustering

We compared the two-stage and three-stage clustering methods to commonly used methods, namely, the hierarchical, *k*-means, SOM, SOTA and model-based clustering (Yeung *et al.*, 2001) methods. All codes were obtained from R packages (cluster, stats, som, mclust) downloadable from the comprehensive R archive network (CRAN) except for SOTA, which we ran on GEPAS, a web-based server for SOTA. The Rand and adjusted Rand indices were used to measure performance. The Rand index (Rand, 1971) is the number of agreements (pairs that are either in the same cluster or in different clusters in both clusterings) divided by the total number of pairs. The adjusted Rand index proposed by Hubert and Arabie (1985), adjusts the score so that its expected value for random clustering is zero. The maximum value for the Rand and adjusted Rand indices is one; a high index indicates a high level of agreement between the clusterings. The jump factor criterion was used to detect the number of clusters for hierarchical, two-stage and three-stage methods. The model-based method and SOTA have a built-in criteria while the number of clusters for *k*-means and SOM are user specified.

3.1.1 Two-stage clustering To test the two-stage clustering method, we used a simulated dataset placed by Michaud *et al.* (2003) at http://www.che.udel.edu/eXPatGen/paper/example2.out and the yeast cell cycle data with MIPS criterion (extracted from Cho *et al.* (2001)) made available by Yeung *et al.* (2001) at http://faculty.washington.edu/kayee/cluster/. The simulated dataset contains 100 genes and 36 samples and is generated based on known biological features of expression complexity, diversity and interconnectivity. There are 10 clusters in this dataset, with each cluster containing 10 genes. Close examination of this dataset, however, shows that the first two clusters contain genes that are neither repressed nor induced at any point of the experiment. We treat these clusters as identical and so the dataset contains only nine true clusters. The yeast cell cycle dataset contains 237 genes and 17 samples. These genes corresponding to four categories in the MIPS database (DNA synthesis and replication, organization of centrosome, nitrogen and sulphur metabolism, and ribosomal proteins); we assume these to be the true clusters. Table 2 shows the Rand and adjusted Rand indices for the two-stage method and the other

Table 2. Rand and adjusted Rand indices for the simulated data and the yeast cell cycle data

Index	Simulated data		Yeast cell cycle data	
	Rand	Adjusted Rand	Rand	Adjusted Rand
Two-stage	0.9961	0.9806	0.7087	0.3697
Hierarchical	0.9760	0.8873	0.6709	0.2984
k-means	0.9748	0.8794	0.6892	0.3197
SOM	0.8588	0.5363	0.6715	0.3159
SOTA	0.9596	0.8235	–	–
Model-based	0.9961	0.9806	0.5702	0.1472

Table 3. Rand and adjusted Rand indices for the CNS data

Index	Rand	Adjusted Rand
Two-stage	0.7479	0.0685
Three-stage	0.8797	0.0899

commonly used methods, against the true clusters for the simulated data and the yeast cell cycle data.

For the simulated data, both the two-stage and model-based method did equally well, with nine clusters detected and only one gene misclassified. Hierarchical clustering detected only eight clusters, SOTA detected seven and the *k*-means method had the number of clusters pre-specified as nine. All these methods had slightly lower Rand and Adjusted Rand indices compared to the two-stage method. We were unable to force SOM to produce nine clusters; six clusters was the optimal choice. This method produced the lowest Rand and adjusted Rand indices.

For the yeast cell cycle data, the two-stage method detected four clusters (two in the first stage and two in each first-stage cluster in the second stage). Two-stage clustering has the highest Rand and adjusted Rand indices. Using hierarchical clustering, three clusters were detected. If we pre-specify four as the number of clusters in hierarchical clustering, it performed slightly better but not as well as the two-stage method. The *k*-means and SOM methods, despite having the advantage of four being pre-specified as the number of clusters, have a lower adjusted Rand index than the two-stage clustering method. The model-based method detected only two clusters; this could be the reason behind its having the lowest Rand and adjusted Rand indices. We did not include SOTA in the results because it produced too many clusters (up to 50).

3.1.2 Three-stage clustering The central nervous system (CNS) development gene expression data (Wen *et al*., 1998) made available by Yeung *et al*. (2001) at http://faculty.washington.edu/kayee/cluster/ was used to test the three-stage method. There are 112 genes known to belong to major gene families deemed important for spinal cord development. There are nine samples in this experiment, measured using embryonic days 11, 13, 15, 18 and 21, postnatal days 0, 7 and 14 and adult (postnatal day 90).

We divided the data into three groups, early embryonic days (E11, E13, E15), late embryonic days (E18, E21) and all data after birth (P0, P7, P14, P90). Table 3 shows the Rand and adjusted Rand indices for the two-stage and three-stage methods. Wen *et al*. (1998) classified the genes into 14 general functional classes. The two-stage clustering resulted in six clusters, while three-stage clustering yielded 21 clusters. In this dataset, three-stage clustering performed better based on both the Rand and adjusted Rand indices.

3.2 Imputation

Missing values were generated by randomly removing from 1 to 20% of the data. The root mean squared error (RMSE) was calculated to measure the performance of the imputation.

Table 4. Mean, minimum and maximum RMSE for 1000 runs of the simulated dataset and the yeast cell cycle data, with 10% data missing (chosen at random in each run), using the different imputation methods

RMSE	Simulated data			Yeast cell cycle data		
	Mean	Min	Max	Mean	Min	Max
Two-stage	0.1225	0.0915	0.1866	0.2743	0.2182	0.3374
Zero	1.031	0.9248	1.1489	6.5875	6.3737	6.7703
Row mean	0.7053	0.6485	0.7530	0.3837	0.3230	0.4522
KNN	0.1325	0.097	0.1325	0.2860	0.2332	0.3500
LLSimpute	0.1849	0.1202	0.3005	0.4600	0.2787	8.2121
CMVE	–	0.7216	–	–	1.8897	–
BPCA	0.0916	0.0738	0.1333	0.2493	0.2032	0.3058

To assess the consistency of results, 1000 runs (each with different values removed randomly) were performed and the mean, minimum and maximum RMSE across all runs were obtained.

3.2.1 Two-stage imputation To illustrate the performance of the two-stage imputation method, the simulated dataset and the yeast cell cycle data with MIPS criterion was used. We compared the two-stage imputation method to the commonly used imputation methods of zero, row mean and KNN imputation. Current methods that are more sophisticated such as the local least squares imputation method (LLSimpute) (Kim *et al*., 2005), Bayesian principal component method (BPCA) (Oba *et al*., 2003) and collateral missing value imputation (CMVE) (Sehgal *et al*., 2005) were also compared to give more credibility to the comparison between methods. The code for KNN imputation was obtained from the R package, called 'impute', downloadable from CRAN. LLSimpute, BPCA and CMVE were obtained via downloadable Matlab code available at the associated author website. An example of the results, with 10% of data missing, is shown in Table 4.

For the simulated dataset, we used $k = 10$ for the KNN, LLSimpute and CMVE method since each cluster contains 10 genes, and this value is reported to work well for all three methods (Sehgal *et al*., 2005). For the yeast cell cycle data, we used $k = 8$ for the KNN method and for two-stage imputation we used 30 clusters; this corresponds to approximately 8 genes in a cluster, assuming that they are uniformly distributed. This choice produces a markedly low RMSE for both methods. The same *k* value is used for LLSimpute and CMVE to provide an equitable comparison.

Based on RMSE, the two-stage method outperformed all methods except BPCA. BPCA is a more sophisticated method than the two-stage method and far more computationally intensive. A single run using the simulated dataset with 10% missing values takes BPCA approximately 40 s while the two-stage imputation method takes ~1 s. Furthermore, Bayesian approaches are highly dependent on the chosen prior distribution; a wrong choice of prior distribution would result in poor performance. In our investigation, BPCA outperforms LLSimpute and CMVE, while Kim *et al*. (2005) and Sehgal *et al*. (2005) have reported that their respective methods outperform BPCA. LLSimpute is reported to perform well when

Table 5. Mean, minimum and maximum RMSE for 1000 runs of the CNS data, with 10% missing data (chosen at random in each run), using the different imputation methods

RMSE	Mean	Minimum	Maximum
Two-stage	0.1720	0.1371	0.2321
Three-stage	0.1770	0.1276	0.2458

Table 6. Mean RMSE of 1000 runs for the CNS data with differing percentages of missing values, using the two-stage and three-stage imputation methods

	1%	5%	10%	15%	20%
Two-stage	0.1702	0.1693	0.1720	0.1752	0.1789
Three-stage	0.1635	0.1679	0.1770	0.1866	0.1980

k is large; based on the 1000 runs, we found that the optimal k varies extensively. CMVE performed extremely badly on a number of runs, giving an infinitely large RMSE. Moreover, even its minimum RMSE is high in comparison with other methods. This could be due to bugs in the Matlab code or sensitivity to the distribution of the missing data.

3.2.2 Three-stage imputation To test the performance of the three-stage imputation method, the CNS data was used. The main focus was to compare the three-stage method with the two-stage method. An example of the results, with 10% of data missing, is shown in Table 5.

For both two-stage and three-stage imputation, we used 11 clusters. This number of clusters was chosen because it produces a low RMSE. For 10% of values missing, on average, two-stage imputation performed slightly better than three-stage imputation. The three-stage method, however, has a lower minimum compared to two-stage imputation. Table 6 shows the difference between the performance of the two-stage and the three-stage imputation methods as the percentage of missing values increases from 1 to 20%.

On average, three-stage imputation performed better than two-stage imputation when the percentage of missing values was low (i.e. at 1 and 5%). As the percentage of missing values increases, the performance of three-stage imputation drops. This is because the replication within treatment is very small (i.e. three replications within each treatment). When the percentage of missing values is high, there is a high probability that all replications within a treatment are missing and therefore, the imputed value is the overall mean of the gene expression profile. This reduces the three-stage imputation method to row mean imputation and thus its performance falls.

4 DISCUSSION

Our results suggest that decomposing the profiles into orthogonal components and clustering in stages is a useful approach. Geometrically, the multi-stage approach breaks the S-dimensional space in which the data lies into orthogonal subspaces. The advantage of this approach is most apparent when the scale of the dominant attribute is considerably larger than that of the others. In this case, clustering is largely based on the dominant attribute when using Euclidean distance. The multi-stage approach, however, allows the subtlety of all components to be acknowledged. Therefore, extra information is available when performing the clustering.

In this article, for two-stage clustering, we have used main effect distance first then interaction distance second; we refer to this as the 'top-down' order. Altering the order can give very different clustering results, especially when the clusters are indistinct. It is possible that a 'bottom-up' approach could produce better results. The question here is not to determine which order is better, rather it is to determine when a certain order is better. Since the process is not reversible, we always begin with the attribute which produces the most distinct clusters. If the separation is unclear in the first stage, the misclustering that occurs is carried on to subsequent stages. Information on the separation of the clusters in each stage is usually unattainable; we resolve this difficulty by assuming that attributes with larger values tend to have more distinct clusters. Thus, a top-down approach should be the default option.

Work in progress involves the implementation of the multi-stage idea into a model-based method. Classical and Bayesian model-based approaches to clustering of gene expression profiles will be studied and compared at a future date.

5 CONCLUSION

We have introduced an alternative approach to the clustering of gene expression profiles, an approach that involves clustering in a number of stages using a hierarchy of distance measures. This enables the clustering method to deal with large datasets in a systematic way.

We have shown that the distance measures are related to the design of experiment employed and reflect different attributes of the data. The multi-stage approach enhances the distinguishing power of the distance measures, because it allows subtle differences to not be masked by a more dominant attribute of the gene expression profiles. Thus, the precision of clustering is improved, as seen in the results displayed.

This clustering method is modified to accommodate missing values and leads to an associated imputation method. The multi-stage imputation method is simple and robust. It also outperforms imputation methods within its league.

The multi-stage approach is not only theoretically grounded but also biologically supported. It achieves this by putting emphasis on shape similarity, so taking into account the fact that co-expressed genes are co-regulated. Furthermore, it can be used on incomplete datasets and brings with it the ability to estimate the missing values.

Conflict of Interest: none declared.

REFERENCES

Ben-Dor,A. *et al.* (1999) Clustering gene expression patterns. *J. of Comput. Biol.*, **6**, 281–297.

Bo,T.H. *et al.* (2004) LSimpute: accurate estimation of missing values in microarray data with least squares methods. *Nucleic Acids Res.*, **32**, e34.

Boutros,P.C. and Okey,A.B. (2005) Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data. *Brief. Bioinform.*, **6**, 331–343.

Cho,R.J. *et al.* (2001) Transcriptional regulation and function during the human cell cycle. *Nat. Genet.*, **27**, 48–54.

de Brevern,A.G. *et al.* (2004) Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering. *BMC Bioinformatics*, **5**, 114.

Eisen,M.B. *et al.* (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci. USA*, **95**, 14863–14868.

Godfrey,A.J.R. *et al.* (2002) Two-stage clustering in genotype-by-environment analyses with missing data. *J. Agric. Sci.*, **139**, 67–77.

Hastie,T. *et al.* (2000) 'Gene shaving' as a method of identifying distinct sets of genes with similar expression patterns. *Genome Biol.*, **1**, research0003.1-0003.21.

Herrero,J. *et al.* (2001) A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, **17**, 126–136.

Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classification*, **4**, 193–218.

Jornsten,R. *et al.* (2005) DNA microarray data imputation and significance analysis of differential expression. *Bioinformatics*, **21**, 4155–4161.

Kim,H. *et al.* (2005) Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics*, **21**, 187–198.

Kim,K.Y. *et al.* (2004) Reuse of imputed data in microarray analysis increases imputation efficiency. *BMC Bioinformatics*, **5**, 160.

Knudsen,S. (2002) *A Biologist's Guide to Analysis of DNA Microarray Data*, John Wiley and Sons, Inc., New York.

McLachlan,G.J. *et al.* (2002) A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 413–422.

Michaud,D.J. *et al.* (2003) eXPatGen: generating dynamic expression patterns for the systematic evaluation of analytical methods. *Bioinformatics*, **19**, 1140–1146.

Nguyen,D.V. *et al.* (2004) Evaluation of missing value estimation for microarray data. *J. Data Sci.*, **2**, 347–370.

Oba,S. *et al.* (2003) A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, **19**, 2088–2096.

Ouyang,M. *et al.* (2004) Gaussian mixture clustering and imputation of microarray data. *Bioinformatics*, **20**, 917–923.

Rand,W.M. (1971) Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.*, **66**, 846–850.

Scheel,I. *et al.* (2005) The influence of missing value imputation on detection of differentially expressed genes from microarray data. *Bioinformatics*, **21**, 4272–4279.

Sehgal,M.S. *et al.* (2005) Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data. *Bioinformatics*, **21**, 2417–2423.

Soukas,A. *et al.* (2000) Leptin-specific patterns of gene expression in white adipose tissue. *Genes Dev.*, **14**, 963–980.

Tamayo,P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Nat. Acad. Sci. USA*, **96**, 2907–2912.

Tom,B.D. *et al.* (2005) Quality determination and the repair of poor quality spots in array experiments. *BMC Bioinformatics*, **6**, 234.

Troyanskaya,O. *et al.* (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics*, **17**, 520–525.

Tuikkala,J. *et al.* (2006) Improving missing value estimation in microarray data with gene ontology. *Bioinformatics*, **22**, 566–572.

Wang,X. *et al.* (2006) Missing value estimation for DNA microarray gene expression data by support vector regression imputation and orthogonal coding scheme. *BMC Bioinformatics*, **7**, 32.

Wen,X. *et al.* (1998) Large-scale temporal gene expression mapping of central nervous system development. *Proc. Natl. Acad. Sci. USA*, **95**, 334–339.

Yeung,K.Y. *et al.* (2001) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**, 977–987.

Zhou,X. *et al.* (2003) Missing-value estimation using linear and non-linear regression with Bayesian gene selection. *Bioinformatics*, **19**, 2302–2307.