Analysis of gene expression: case study for bacteria

Journal of Physics: Conference Series **128** (2008) 012030 doi:10.1088/1742-6596/128/1/012030

Analysis of gene expression: case study for bacteria

Maia Angelova and Chris Myers

Intelligent Modelling Lab, School of CEIS, Northumbria University, Pandon Building, Newcastle NE2 1HE, UK

E-mail: maia.angelova@unn.ac.uk, chris.myers@unn.ac.uk

Abstract. The analysis of gene expression allow to study the functions of genes and their roles in different processes in the cell of a living system, including the cell cycle. Clustering is widely used in the analysis of high-throughput gene expression data to find patterns of similarity that enable related gene groups and functions to be identified. Clustering algorithms are very sensitive to the choice of initial conditions and optimal number of clusters. In this paper, we investigate the impact of metrics and cluster parametrisation for three clustering models and propose a method for optimisation of cluster parameters based on cluster compactness and separation. A case study presents the analysis of gene expression data for *E.coli* bacteria.

1. Introduction

The world of biological sciences has undergone an information revolution in recent years due to the development of rapid DNA sequencing techniques, the progress in computer based technologies and mathematical and computer modelling. The development of DNA microarray technology has made it possible to measure the expression levels of thousand of genes simultaneously under controlled experimental conditions. The completion of the genome sequence for human and many model organisms has provided powerful biological data of richness and scale that require a data driven approach. Moreover, the atomic description of molecular structure and functions, as well as molecular explanation of cellular behaviour, have become increasingly possible. The growing amount of detailed information about biological systems and the complexity of these systems is a challenge that needs modelling and computational tools, adaptable to large data sets, and an integrative modelling approach [1, 2].

The central dogma of molecular biology is that the information is stored in DNA, transcribed to messenger RNA (mRNA) and then translated into proteins (see for example [1]), i.e. once the gene has been activated it is expressed in the cell. This justifies the premise that the information about the functional state of an organism is to a great extent determined by the information on gene expression and is the motivation for the analysis of large-scale gene expression data. Among many powerful automatic techniques for analysing high-throughput gene expression data from microarray experiments, clustering is widely used [2, 3]. It is accomplished by finding similarity patterns within gene expression data thus enabling related gene groups and functions to be identified. Clustering results are used to learn about gene functions, gene regulation, cellular processes and subtypes of cells. The genes with similar expression patterns, known as co-expressed genes, can be clustered together. Co-expressed genes are likely to have similar cellular functions, or to be involved in the same cellular processes [4, 5, 6, 7]. Further, the regulatory modules of biological function of unknown genes can be discovered by associating them with

other genes that have similar expression patterns and known regulatory elements or functions. However, most clustering algorithms are very sensitive to the initial choice of parameters and optimal selection of clusters.

In this paper, we investigate the impact of metrics and cluster parametrisation for three clustering models and propose a method for optimisation of cluster parameters based on cluster compactness, separation and stability. A case study for bacteria is used to illustrate the main concepts and results. The rest of the paper is organised as follows. In Section 2 we consider the effect of similarity metrics and compare the optimisation criteria in three clustering models, K-means, EM and minimal entropy models . In Section 3 we discuss the quality and stability of the clusters and propose a method for optimisation of cluster parameters. In Section 4 the analysis of clustering results for genome-wide expresion of *Escherichia coli (E.coli)* bacteria is presented, the discussion and conclusions are given in Section 5.

2. Effect of similarity metrics in clustering models

Clustering models are very popular in the analysis of high-throughput gene expression data from microarray experiments [2, 3]. Clustering is the exploratory, unsupervised process of grouping data objects into a set of disjoint classes, called clusters, so that objects within one class have high similarity to each other and are dissimilar to objects in another class [8].

Microarray gene expression data can be presented by a real value gene expression matrix,

$$M = \{w_{ij} | 1 \leq i \leq n, 1 \leq j \leq m\} \tag{1}$$

where the rows $G = \{\boldsymbol{g_1}, \boldsymbol{g_2}, ..., \boldsymbol{g_n}\}$ form the expression patterns of the genes, the columns $S = \{\boldsymbol{s_1}, \boldsymbol{s_2}, ..., \boldsymbol{s_m}\}$ represent the expression profiles of the samples (experiments), and each cell of the expression matrix, w_{ij}, is the measured expression level of gene i in sample j, $i = 1, 2, ...n, j = 1, 2, ..., m$. In what follows, the vector $\boldsymbol{r}$ indicates a gene expression data object, which can represent a gene $\boldsymbol{g_i}$ in the n-dimensional gene space or a sample $\boldsymbol{s_j}$ expression profile in the m-dimensional array space. The similarity pattern can be established by comparing genes or experimental conditions (samples). The original gene expression matrix contains noise, missing values, and systematic variations arising from the experimental procedure. In most cases, data preparation and normalization is necessary before any data analysis can be performed.

The choice of similarity measure used within the clustering algorithm is very important as it influences the output and interpretation of the results. It has received much discussion [2, 3, 9] but there is still no agreement over the best metric to use and little work has been done on assessing the impact of different measures in the analysis of gene expression [10, 11].

The similarity is defined as a function, Sim, that measures association, usually distance or correlation, between data objects, representing genes or samples in the expression matrix. The distance measures the proximity between data objects $\boldsymbol{r_i}$ and $\boldsymbol{r_j}$ and represents the dissimilarity or unlikeness between them. A typical distance metric is the Euclidean distance,

$$D(\boldsymbol{r_i}, \boldsymbol{r_j}) = \sqrt{\sum_{d=1}^{m} (w_{id} - w_{jd})^2}, i, j = 1, 2, ..., n \tag{2}$$

Other distance measures used in this work are Manhattan and Minkowski distance.

The correlation measures the similarity or alikeness between two objects $\boldsymbol{r_i}$ and $\boldsymbol{r_j}$. It is different from the distance as a measure of the relationship between gene expression profiles. A typical correlation function is the Pearson correlation coefficient,

$$P(\boldsymbol{r_i}, \boldsymbol{r_j}) = \frac{\sum_{d=1}^{m} (w_{id} - \mu_i)(w_{jd} - \mu_j)}{\sqrt{\sum_{d=1}^{m} (w_{id} - \mu_i)^2} \sqrt{\sum_{d=1}^{m} (w_{jd} - \mu_j)^2}} \tag{3}$$

where μ_i and μ_j are the means for the objects $\boldsymbol{r_i}$ and $\boldsymbol{r_j}$ respectively. The correlation coefficient takes a value between -1 and $+1$. For example, a value of 0 represents perfectly uncorrelated variables, values close to ± 1 represents strong positive (negative) correlation between genes and could point to association of their functions. Cosine correlation coefficient is another example of correlation function. The correlation metric can be converted to a distance memetric [3]. Other examples of similarity metrics, are Jaccard and dice similarities, which take values between 0 and 1. We have investigated the impact of generalised similarity metrics using K-means, EM and entropy based models.

2.1. K-means model

K-means is an iterative partitioning method in which objects are moved among a pre-specified number of clusters, K, until an optimal solution is reached. It is an example of hard clustering algorithm in which each gene is assigned to a single cluster. The basic K-means clustering algorithm can be summarised with the following steps (see for example [8]): the number of clusters K is specified; the initial values for the cluster centres are chosen; each data object $\boldsymbol{r}$ is assigned to a cluster $C_i, i = 1, ..., K$, using the selected Sim function; the cluster centres $\boldsymbol{m_i}, i = 1, ..., K$, are re-calculated using the mean of all objects in each cluster; the objects are re-assigned; the last two steps are repeated until the cluster membership is stable.

The algorithm minimises a global error criterion, known as cost function [2], which depends on the preliminary selection of number of cluster K, cluster centres $\boldsymbol{m_i}$ and similarity function Sim. Although there is no universally accepted definition and the cost function could be tailored to the problem, it is usually defined as "within-cluster" sum of the squared distances between each data object $\boldsymbol{r}$ belonging to the cluster C_i and its cluster centre $\boldsymbol{m_i}$,

$$CF_2 = \sum_{i=1}^{K} \sum_{\boldsymbol{r} \in C_i} \mid \boldsymbol{r} - \boldsymbol{m_i} \mid^2 \tag{4}$$

and represents the total error. It is well-known that the algorithm minimises CF_2 but converges to a local rather than a global minimum depending on the choice of initial parameters, such as number of clusters K and cluster centres. Thus, the automatic selection of an optimal number of clusters for the clustering algorithm is a complex task and is a common problem for all partitioning algorithms. An optimisation procedure of these parameters is discussed in Section 3.

The impact of similarity measures has been investigated using microarray data for *E.coli* gene expression. The case study is considered in more details in Section 4. Here we present a summary of the metric's analysis for a reduced data set of 264 genes.

The data set was analysed using seven different metrics such as distance: Euclidean(E), Manhattan (Mh), Minkowski (Mk), correlation: Cosine(C), Pearson correlation coefficients (P) and other similarity: Dice (D) and Jaccard (J). The results are given in Table 1. The table compares the cluster sizes (number of genes per cluster) for different metrics and values of K. The similarity metric is indicated in the first column, the first row gives the number of clusters $K = 4, 5, 6, 7$. The size of each cluster for each value of K is given in the remaining columns for each similarity measure. The table shows that the clustering results are dependent on both, the choice of similarity measure and the value of K. Whilst the cluster sizes naturally decrease as K increases, there is some evidence of stability between the small clusters which tend, in the case of this data set, to contain the most highly expressed genes. The table illustrates the difference, for each value of K, between cluster sizes obtained using distance, correlation or other metrics. However, the results are similar for the three distance measures E, Mh and Mk, same applies for the correlation metrics C and P, and other similarity metrics D and J.

Table 1. Comparison of cluster sizes for different similarity metrics

Number of clusters Metric	4	5	6	7
E	247,9,4,4	159,89,9,4,3	160,87,9,4,3,1	92,84,71,9,4,3,1
Mh	247,9,4,4	158,90,9,4,3	158,89,8,5,3,1	93,91,63,9,4,3,1
Mk	160,88,13,3	160,88,9,4,3	160,87,8,5,3,1	160,87,7,4,3,2,1
D	159,88,13,4	157,90,10,4,3	91,87,69,10,4,3	96,76,71,7,6,5,3
J	159,88,13,4	157,90,10,4,3	156,86,8,6,5,3	94,75,71,7,7,6,4
C	98,96,59,11	117,69,51,17,10	101,67,50,19,17,10	98,57,40,31,23,8,7
P	87,81,67,29	110,61,38,29,26	88,61,56,28,2,8,3	83,53,43,38,28,16,3

2.2. EM model

In the probabilistic models, data is assumed to be drawn from a series of probability distributions, usually multivariate Gaussian distributions. These models use the Expectation-Maximization (EM) algorithm [12] to produce the best fit between the data and a series of Gaussian distributions. The EM algorithm uses the likelihood as a similarity measure instead of distance or correlation. The algorithm takes into account that each object can belong to each cluster with a certain probability and finds the maximal log-likelihood. The log-likelihood is given by

$$\mathcal{L} = \sum_{i=1}^{n} \log(\sum_{k=1}^{K} \lambda_k p_k(\boldsymbol{r}_i|M_k)) \tag{5}$$

where λ_k is the probability that data object $\boldsymbol{r_i}$ belongs to cluster C_k, $\lambda_k \geq 0$, $\sum_k \lambda_k = 1$, each cluster C_k is represented by a model M_k, $p(\boldsymbol{r_i}|M_k)$ is the probability density of $\boldsymbol{r_i}$ in $M_k, k = 1, ..., K$ and K is the number of clusters. Each model M_k can be represented by a multivariate d-dimensional Gaussian distribution with mean μ_k and covariance Σ_k. The EM algorithm finds the maximal log-likelihood $\mathcal{L}_M$ for a given data model $M\{M_1, ..., M_K\}$.

It can be shown that when the cost function corresponds to an underlying probabilistic model, K-means can be regarded as an approximation of the classical EM algorithm on a spherical Gaussian mixture model [2]. Like K-means, the EM algorithm requires the number of clusters to be specified in advance.

2.3. Minimal entropy model

The entropy-based models use the entropy of the clusters as a similarity metric. The entropy measures the uncertainty of a random variable. In Shannon's information theory [13], the entropy of a random variable X is defined as

$$H(X) = -\sum_{i} p(\boldsymbol{r_i}) \log(p(\boldsymbol{r_i})) \tag{6}$$

In thermodynamics, the entropy is a measure of the disorder in the system. Applied to clustering, the concept of entropy means that each cluster should have a low entropy as objects in the same cluster are similar. Thus, the search for clusters with minimal entropy can be used as a clustering criterium.

The entropy of the clusters can be written as,

$$H = \sum_{j=1}^{K} p_j H(X|C_j) \tag{7}$$

where $H(X|C_j)$ is the entropy of the cluster C_j, p_j is the probability of the cluster C_j such that $\sum_j p_j = 1$ and K is the number of clusters. A clustering algorithm that minimises (7) has been developed in [14]. The entropy of the cluster $H(X|C_j)$ can be measured using the actual relationship between data objects and clusters. Li *et al* [14] have evaluated this relationship by using *posteriori* probabilities $p(C_j|\boldsymbol{r_i})$ of object $\boldsymbol{r_i}$ in cluster C_j. However, the choice of particular data distribution (such as Gaussian distribution) can lead to a poor representation of the data. An alternative method is based on the actual density of data objects using the Parzen density approach [15]. The probability $p(C_j|\boldsymbol{r_i})$ is evaluated using the Parzen density estimation for the clustering problem as [14],

$$p(C_j|\boldsymbol{r_i}) = \frac{n_{ij}}{n_i} \tag{8}$$

where n_{ij} is the number of samples $\boldsymbol{r_i}$ from cluster C_j and n_i is the number of all samples located in a selected region $R(\boldsymbol{r_i})$. The entropy clustering criterion can be written as

$$H = -\sum_{i=1}^{n}\sum_{j=1}^{K}\frac{n_{ij}}{n_i}\log(\frac{n_{ij}}{n_i}). \tag{9}$$

The algorithm minimises (9) to find the minimal entropy of the clusters. Like the other clustering functions discussed in the paper, H has local minima rather then a global minimum. We have used (9) with K-means and EM to optimise and improve the clustering results.

3. Quality, stability and optimisation of clusters

Clustering algorithms, like K-means, EM and minimal entropy, require the number of clusters to be given in advance. This is often very difficult as biologists may not know the exact number of functional categories as some of the genes may have unknown functions or belong to groups with unknown functional categories. The optimal selection of the number of clusters in the clustering algorithm and the stability of the clusters is important as it impacts upon the clustering solution (see Table 1). A good clustering result should produce tightly packed clusters which are stable and well separated [2, 9].

3.1. Quality of clusters

The quality of the clusters can be measured in terms of "intra cluster" homogeneity and "inter cluster" separateness. The term homogeneity [9] is used to represent the sameness of data points within a cluster. The corresponding function can be defined as,

$$Hom(C_i) = \frac{1}{\|C_i\|}\sum_{\boldsymbol{r}\in C_i} Sim(\boldsymbol{r}, \boldsymbol{m_i}) \tag{10}$$

where $\|C_i\|$ is the number of points allocated to cluster C_i and Sim is the similarity function. To illustrate these ideas in K-means, we have chosen Sim as Euclidean distance (2). The term separateness is used to estimate the separability between clusters and can be measured by the function,

$$S(C_i, C_j) = Sim(\boldsymbol{m_i}, \boldsymbol{m_j}), \;\; i \neq j, \tag{11}$$

giving the distance between two cluster centres.

The average cluster homogeneity measures the "intra cluster" sameness and represents the average distance between each data object and its cluster centre. The average cluster separation measures "inter cluster" separateness and gives the average (weighted) distance between cluster centres. When considered in terms of distance between objects, our aim is to find solutions based around compact, well separated clusters, with low homogeneity (high density of packing)

and high separation. Based on these concepts, we evaluate the quality of the clusters by defining two cost functions and two quality functions.

The error function CF_1 is the sum of distances between each data object and its cluster centre,

$$CF_1 = \sum_{i=1}^{K} \sum_{\boldsymbol{r} \in C_i} \mid \boldsymbol{r} - \boldsymbol{m_i} \mid \tag{12}$$

It represents the mean error per data point and measures average cluster homogeneity. The cost function CF_2 (4) is the sum of squared distances between each data point and its cluster centre. The quality function QF_1 is the difference between the mean distance between cluster centres and the mean error per data point,

$$QF_1 = \frac{1}{K(K-1)} \sum_{i,j,i \neq j} \mid \boldsymbol{m_i} - \boldsymbol{m_j} \mid - \frac{CF_1}{n} \tag{13}$$

It represents the balance between cluster separateness and cluster compactness. QF_2 is the difference between the mean cluster separation and the mean error per data point,

$$QF_2 = \frac{1}{\sum_{i,j,i \neq j} \|C_i\| \|C_j\|} \sum_{i,j,i \neq j} \|C_i\| \|C_j\| S(C_i, C_j) - \frac{CF_1}{n}. \tag{14}$$

We have optimised the number of clusters within K-means using the functions CF_i, $QF_i, i = 1, 2$. A low cost, high density and high separation solution corresponds to a minimum of CF_i and a maximum of $QF_i, i = 1, 2$, for the same values of the parameters.

Figure 1 represents the cost functions CF_1 and CF_2 plotted against the number of clusters, K, for the reduced set of 264 genes. The quality functions QF_1 and QF_2 are given on Figure 2. The functions CF_1 and CF_2 do not exhibit a global minimum but a series of local minima with respect to K. The function QF_2, however, has a global maximum. Thus, the optimal solution corresponds to a local minimum of CF_2 and a global maximum of QF_1, which for the set of 264 genes is at $K = 6$.

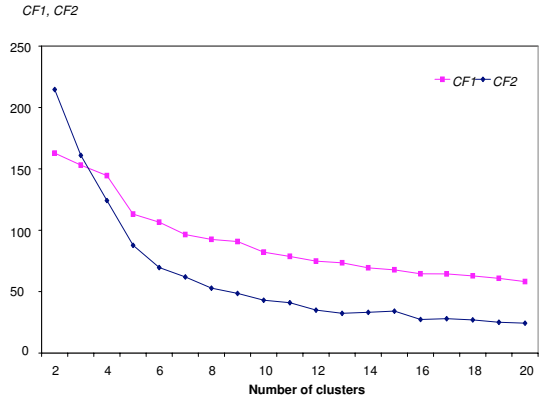

Figure 1. Cost functions CF_1, CF_2 vs number of clusters.

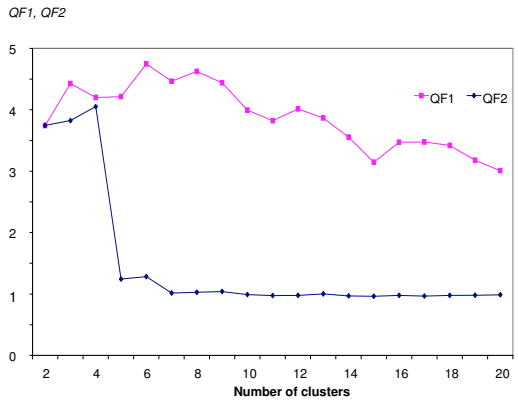

Figure 2. Quality functions QF_1, QF_2 vs number of clusters.

Smet *et al* [16] have used an alternative quality based approach in which clusters are defined sequentially and can only contain genes lying within a specified volume. Other approaches to the optimisation are associated with the statistics of the data [17].

We have analysed the results of the EM model using the Bayesian Information Criterion, BIC (see for example [18] and references therein),

$$BIC = 2\mathcal{L}_M - m_M \log(n) \tag{15}$$

where $\mathcal{L}_M$ is the maximised mixture log likelihood (5) for the data model M, m_M is the number of independent parameters to be estimated and n is the number of data objects. BIC has been used in [18] to estimate the optimal number of components within a mixture model. Whilst the maximised log likelihood will increase as the number of components increases, the second term in (15), which is based on the number of parameters, compensates this increase. Equation (15) can be considered as an alternative of the cluster balance (13). We have used (15) to search for optimal solution in EM with maximised log likelihood and maximal BIC. Figure 3 presents BIC as a function of the number of clusters for the set of 264 genes. The optimal solution is at $K = 5$ where BIC has a global maximum.

We have estimated the entropy of the clusters in K-means and EM-model using the criterium (9). An illustration for the set of 264 genes in K-means is given on Figure 4. We have combined the intra cluster sameness CF_2 and the minimal entropy H to identify the optimal solution. It corresponds to a global minimum of the function $CF_2 + H$, which for the set of 264 genes is at $K = 7$.

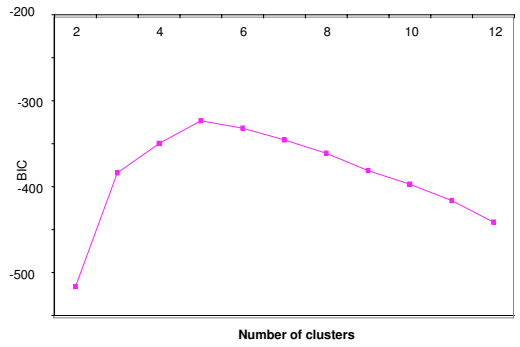

Figure 3. BIC vs number of clusters.

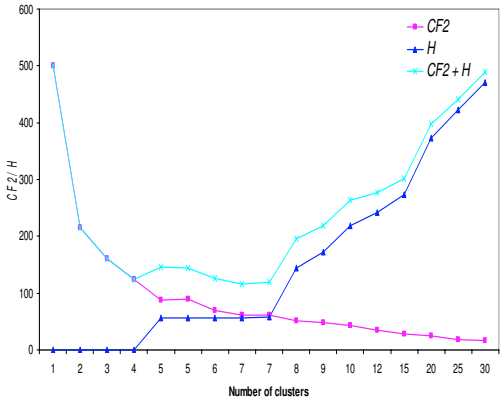

Figure 4. Cost function CF_2 and entropy H vs number of clusters

3.2. Stability of clusters

The Rand index is a useful way of comparing different clustering results from a given dataset. Given a set of data objects $G = \{\boldsymbol{r_1}, \boldsymbol{r_2}, ..., \boldsymbol{r_n}\}$ and two different clustering solutions $C^1 = \{C_1^1, C_2^1, ..., C_K^1\}$ and $C^2 = \{C_1^2, C_2^2, ..., C_P^2\}, K \neq P$, the Rand index, R, can be defined as,

$$R = \frac{a+b}{a+b+c+d}, \tag{16}$$

where a is the number of pairs of data objects in G that are in the same cluster in both C^1 and C^2, b is the number of data objects in G that are not in the same cluster in both C^1 and C^2, c is the number of data objects in G that are in the same cluster in C^1 but not in the same cluster in C_2 and d is the number of pairs of data objects in G that are not in the same cluster in C^1 but are in the same cluster in C^2. R has a value between 0 and 1 with 0 indicating that

the two clustering solutions do not agree on any pair of data objects and 1 indicating that the two solutions are identical.

Rand index was used to compare different solutions in one model obtained by varying the number of clusters, a reliable solution would be expected to exhibit some level of stability as the number of clusters is varied. We have also used Rand index calculations to compare the effect of similarity metrics. For the set of 264 genes, the optimal results in K-means are for six clusters. Rand index calculations confirm that distance metrics lead to similar clustering results, R(E/Mh)=0.99, R(E/Mk)=0.99, R(Mh/Mk)=0.97. For correlation metrics, R(P/C)=0.79, R(J/D)=0.74,R(P/D)=0.73; R(P/E)=0.7 indicates some difference in clustering results when Pearson coefficient or Euclidean distance is used as metric (see also Table 1).

Rand index calculations were used to compare results from K-means and EM model. 6 clusters in both K-means (Euclidean metric) and EM give R=0.67, 6 clusters in K-means and 8 clusters in EM give R=0.65. The relatively low value of R indicates that the two models produce to an extent different results, one of the reason being that cluster sizes are different. Table 2 gives the number of genes per cluster in K-means and EM for a solution with 6 clusters for set of 264 genes.

Table 2. Comparison of cluster sizes in K-means and EM

Cluster Model	1	2	3	4	5	6
K-means	160	87	9	4	3	1
EM	67	54	47	43	35	18

A stability based method is developed in [19] by measuring the overlap between clusters as the number of clusters is varied. The optimal number of clusters is determined by selecting the solution with the highest stability. The method is applied to a set of clustering results with the number of clusters varying from 2 to m. Let $C_{c,l}$ be a set of data objects in cluster l resulting from a clustering solution with c clusters, $2 < c < m$. Let k, $0 < k < m - c$, be the threshold at which the stability calculation will stop. The cluster stability [19] of cluster l is,

$$S_{c,l} = Min_{i=c+l}^{c+k}\{Max_{j=1}^{i}\{\frac{\|C_{c,j}\bigcap C_{i,j}\|}{\|C_{c,l}\|}\}\}. \tag{17}$$

Equation (17) calculates the values of k for the maximum number of overlapping objects between cluster l and all other clusters in the clustering solution $c, c < i < c + k$. Then it takes the minimum of the k maximum values as the stability of the cluster l. The stability value is normalised to range from 0 to 1 by dividing it with the number of objects in cluster l. The average stability of all clusters in this solution is,

$$< S_c >= \frac{1}{c}\sum_{l=1}^{c} S_{c,l} \tag{18}$$

We have used Rand index (16) together with equations (17) and (18) to evaluate the stability of the clustering results in K-means and the EM model. The values of $< S_c >$ for K-means and EM models are given in Table 3 for $K = 4, .., 7$ for the set of 264 genes. The table indicates that the most stable solution is obtained for 6 clusters. Alternative approaches to the stability of the clustering results are investigated in [20, 21].

Table 3. Average cluster stability for K-means and EM models

Number of clusters Model	4	5	6	7
K-means	0.8484	0.9288	0.9203	0.8968
EM	0.8081	0.9057	0.9471	0.9340

4. Case study: *E.coli*

Ferenc *et al* [22] have studied the effects of knocking out the methionine repressor gene, $metJ$, on the *E.coli* transcriptome. Genome-wide expression data has been obtained where strains of LU106(pFM26) and LU106(pFM20) have been compared using oligonucleotide based array of 4288 *E.coli* genes. The study has confirmed that repression is largely restricted to known genes involved in the biosynthesis and uptake of methionine. The number of additional genes that are up-regulated in the absence of the repressor has been identified. Several other recently characterised genes in the methionine regulon have been identified and previously unknown potentially regulated loci highlighted.

We have used raw data from these experiments, normalised to eliminate background noise and systematic error. The *E.coli* genes have been clustered by hierarchical clustering (with GeneSpring [23]), K-means, EM algorithm and entropy-based algorithms. Data was filtered to select differentially expressed genes and a subset of 265 genes was obtained and analysed. In this subset, $metE$ gene has the highest expression level and always forms a cluster by itself. This is due to the biological setting of the experiments. The reduced set of the remaining 264 genes is used in the paper to illustrate the results. The genes in the reduced data set were clustered using different clustering models. For comparison, genes in the complete set were also clustered.

The clustering was performed using our Java-based clustering tool with embedded entropy algorithm to refine the results of K-means and EM algorithms. K-means was executed multiple times with $K = 1, ..., 20$ with a random starting values for selecting the cluster centres. EM-algorithm was executed in a similar way with a number of Gaussians varying from 2 to 10. For each clustering model, the quality and stability of the clusters were examined and optimal solution was chosen based on the criteria described in Section 3. The entropy-based algorithm was executed with K-means and EM to provide a refined set of initial conditions. The solutions of the different models were compared. The optimal solution in K-means was chosen with 6 clusters for the 264 genes. The data was analysed with seven different similarity metrics. Although different metrics can lead to different clustering patterns (as illustrated by Rand index calculations and Table 1), we have established that a number of highly expressed genes are always clustered together. Table 4 shows highly expressed genes consistently clustered together in K-means with different similarity metrics. The genes with the highest expression levels are presented in the first column of the table, distance metrics are given in the second column, other similarity metrics in third column and correlation metrics in the last column. If the genes in the given row belong to the same cluster for the corresponding group of metrics, this is indicated by "yes", otherwise by "no".

K-means clusters of the 20 genes with the highest expression level are given in Table 5. Euclidean distance is used as a metric. Each row gives the genes belonging to the same cluster. genes lit and $b1240$ are clustered together if K=5 but split when K=6. The optimal solution in EM model for the 264 genes showed that genes with the highest expression level are distributed in 3 clusters (Table 6). The refinement with the minimal entropy algorithm showed that the two genes lit and $bi240$ split in a separate cluster. Genes $metI$ and $metN$, grouped in the same EM cluster, are also clustered together in K-means when Pearson correlation coefficients are used

Table 4. Highly expressed genes consistently clustered together

Genes	E, Mh, Mk	D, J	C, P
$yaeS, metI$	yes	yes	yes
$metA, metR$	yes	yes	yes
$narV, yaeO$	yes	no	no
$metK, metF, b0539, yi82_1$	yes	no	no
$polB, ydcN, metN, prpD$	yes	yes	no
$folE, metK$	no	no	yes
$lit, b1240$	no	no	yes
$metF, metN$	yes	yes	yes
$polB, prpD$	yes	yes	yes

Table 5. K-means clusters of genes with the highest expression levels

$metE$
$metA, metR$
$narV, yaeO$
$metI, metB, yeaS, folB$
$metF, metK, metN, cspA, polB, ydcN, prpD, b0539, yi82_1$
$b1240$
lit

as a metric. $metE$ gene forms a cluster by itself in both models and is included in the tables for completeness.

Table 6. EM clusters of genes with the highest expression levels

$metE$
$metA, metR, narV$
$metI, metB, metF, metK, metN, folE, yeaS, cspA, polB, ydcN, prpD, b0539, yi82_1$ $lit, b1240$
$yaeO$

Co-expressed genes in the same cluster could have similar functions or indicate co-regulation. This has to be investigated further by using meaningful biological criteria. We have used a consensus sequence to identify genes belonging to the *met* box and our results are in agreement with the biological findings [22].

5. Conclusion and further work

Clustering techniques are used frequently in the analysis of gene expression data from microarray experiments. The identification of co-expressed genes allows to infer the function of unknown genes by comparing the co-regulated genes to the genes with known functions. Co-expressed genes in the same cluster are probably involved in the same cellular process and strong expression correlation between those genes indicates co-regulation.

However, clustering models are very sensitive to the choice of initial conditions and optimal selection of clusters. In this paper we have investigated the impact of initial conditions in three

clustering models, K-means, EM and entropy-based model. The effect of similarity metrics is investigated in K-means. We have used cluster quality, stability and balance to optimise the clustering solutions. The entropy of the clusters is explored for improvement of the clustering results. We have clustered gene expression data of *E.coli* obtained in experiments investigating the effect of knocking out the methionine repressor gene. Work is in progress to compare automatically clustering results with known gene functionalities using the Gene Ontology (GO) and published literature. The visualisation of high dimensional gene expression data is an essential part of the analysis as it facilitates the discovery of structures, features, patterns and relationships, and enables human exploration and communication of the data. Our recently published method for targeted projection [25] and the tool for exploration and visualisation of high dimensional data [26] can be used to improve classifications of clustering results following the requirements of the user.

Acknowledgements: The authors thank the Astbury Centre for Molecular Biology at Leeds University, UK, for providing raw data for the genome-wide expression of *E.coli*.

References

[1] Higgs P G and Atwood T K 2006 *Bioinformatics and Molecular Evolution* (Oxford: Blackwell Publishing).
[2] Baldi P and Hatfield G W 2003 *DNA Microarrays and Gene Expression* (Cambridge: CUP).
[3] Stekel D 2003 *Microarray Bioinformatics* (Cambridge:CUP).
[4] DeRisi J L, Iyer V R and Brown P O 1997 *Science* **278** 680–686.
[5] M.B. Eisen MB , Spellman P T, Brown P O and Botstein D 1998 *Proc. Nat. Acad. Sci. USA* **95** 14863–14868.
[6] J. Khan,J.S. Wei,M. Ringnr,L.H. Saal,M. Ladanyi, F. Westermann, F.Berthold, M. Schwab, C.R. Antonescu, C. Peterson and P.S. Meltzer 2001 *Nature Medicine*, **7** 673–679.
[7] Golub T R, Slonim D K, Tamayo P, Huard C, Gaasenbeek M, Mesirov J P, Coller H, Loh M L, Downing J R, Caligiuri M A, Bloomfield C D and Lander E S 1999 *Science* **286** 531–5377.
[8] Dunham M 2003 *Data Mining. Introductory and Advanced Topics* (New Jersey: Prentice Hall).
[9] Jiang D, Tang C and Zhang A 2004 *IEEE Trans Knowledge Data Engin***16** 1370–1386.
[10] Brazma A and Vilo J 2000 Fed. Eur. Biochem. Soc. Lett. **480** 17–24.
[11] Verdoucci J S, Mefi V F, Lin S, Wang Z, Roy S and Sen Ch 2006 *Physiol Genomics* **25** 355–363.
[12] Dempster A P, Laird N M andRubin D B 1970 *Proc. Roy. Stat. Soc.* **B39** 1–38.
[13] Shannon C E 1948 *Bell System Tech. J.* **27** 389–423 ibid 623–656.
[14] Li H, Zhang K and Jiang T 2004 *IEEE Trans. Knowledge data Engineering* **16** 1370–1386.
[15] Parzen E 1962 *An. Math. Stat.* **33** 1065–1076.
[16] Smet F D, Mathys J, Marchal K , Thijis G, Moor B D and Moreau Y 2002 *Bioinformatics* **18** 735–746.
[17] L. Kaufman and P.J. Rousseeuw 1990 *Finding Groups in Data* (New York: Wiley).
[18] Fraley C and Raftery A E 1988 *The Computer Journal* **41** 578–588.
[19] Famili A F, Liu G and Liu Z 2004 *Bioinformatics* **20** 1535–1545.
[20] Datta S and Datta S 2002 *Bioinformatics* **19** 459–466.
[21] Bolshakova N, Azuaje F and Cunningham P 2005 *Bioinformatics* **21** 2546–2547.
[22] Marincs F, Manfield I, Stead J, McDowall K and Stockley P 2006 *Biochem. J.* **396** 227–234.
[23] Angelova M 2006 *Bulg. J. Phys.* **33** 876–883.
[24] Misra J, Schmitt W, Hwang D, Hsiao L L, Gullans S, Stephanopoulos Ge and Stephanopoulos Gr 2002 *Genome Research* **12** 1112–1120.
[25] J. Faith, R. Mintram and M. Angelova 2006 *Bioinformatics* **22** 2667–2673.
[26] J. Faith, M.Brockway 2006 *J. Integrative Biology* **3** 43–50.