Azuaje F, and N. Bolshakova. "Clustering Genome Expression Data: Design and Evaluation Principles", in *Understanding and Using Microarray Analysis Techniques: A Practical Guide*, Berrar D, Dubitzky W and Granzow M, editors, London: Springer Verlag, 2002.

Chapter 13

CLUSTERING GENOMIC EXPRESSION DATA: DESIGN AND EVALUATION PRINCIPLES

Francisco Azuaje, Nadia Bolshakova
University of Dublin – Trinity College, *Department of Computer Science, Dublin 2, Ireland,*
e-mail: {Francisco.Azuaje,Nadia.Bolshakova}@cs.tcd.ie

1. INTRODUCTION: CLUSTERING AND GENOMIC EXPRESSION ANALYSIS

The analysis of expression data is based on the idea that genes that are involved in a particular pathway, or respond to a common environmental stimulus, should be co-regulated and therefore should exhibit similar patterns of expression. Thus, a fundamental task is to identify groups of genes or samples showing similar expression patterns.

Clustering may be defined as a process that aims to find partitions or groups of similar objects. It can be seen as an unsupervised recognition procedure whose products are known as *clusters*. In a genomic expression application, a cluster may consist of a number of samples (or genes) whose expression patterns are more similar than those belonging to other clusters. Figure 13.1 depicts a situation, in which two types of genes, each one associated with a different biological function, are clustered based on their expression profiles. The clusters are represented by circles, and the genes that are linked to each cluster are depicted randomly within the correspondent circle.

Azuaje F, and N. Bolshakova. "Clustering Genome Expression Data: Design and Evaluation Principles", in *Understanding and Using Microarray Analysis Techniques: A Practical Guide*, Berrar D, Dubitzky W and Granzow M, editors, London: Springer Verlag, 2002.

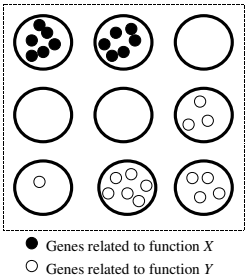

Figure 13.1. Clustering of genes according to their expression patterns

Clustering has become a fundamental approach to analysing genomic expression data. It can support the identification of existing underlying relationships among a set of variables such as biological conditions or perturbations. Clustering may represent a basic tool not only for the classification of known categories, but also (and perhaps most importantly) for the discovery of relevant classes. The description and interpretation of its outcomes may also allow the detection of associations between samples or variables, the generation of rules for decision-making support and the evaluation of experimental models. In the expression domain it has provided the basis for novel clinical diagnostic and prognostic studies (Bittner et al., 2000), and other applications using different model organisms (Ideker et al., 2001).

Several clustering methods have been proposed for expression analysis, and many other options will surely be applied in the future. Moreover, post-genome scientists deal with highly complex and diverse biological problem domains. Therefore, it would not be reasonable to expect the existence of universal clustering solutions. This chapter provides an overview of the major types of clustering problems and techniques for genomic expression data. It focuses on crucial design and analytical aspects of the clustering process. We hope that this chapter will guide our readers to address questions such as: Which clustering strategy should I use? How many clusters should it find? Is this a good partition? Is there a better partition?

Section 2 introduces important concepts for the effective application of clustering techniques. It overviews some of the major types of clustering algorithms for genomic expression data: their advantages, limitations and applications. It provides the reader with some important criteria for the selection of clustering methods. Section 3 approaches the systematic evaluation of clustering results based on their relevance and validity (both computational and biological). Two evaluation models will be presented: Cluster validity strategies based on the *Dunn's index*, and the *silhouette method*. As a way of illustration these methods are implemented using two expression data sets, which were obtained from different clinical diagnostic

studies. The results demonstrate that such validity frameworks may represent a useful tool to support biomedical knowledge discovery. Section 4 concludes with a discussion of the results, future work and recommendations.

2. DESIGN PRINCIPLES FOR CLUSTERING STUDIES

Typical clustering algorithms are based on the optimisation of a partition quality measure. Generally these measures are related to the following factors: a) the heterogeneity of the clusters, also known as the cluster cohesion or compactness; and b) their separation from the rest of the data, also known as cluster isolation. Thus, a basic clustering approach may aim to search for a partition that a) minimize intra-cluster distances, and b) maximize inter-cluster distances.

There are several types of metrics to assess the distance or similarity, between samples and between clusters (Everitt, 1993). A clustering algorithm commonly requires the data to be described by a matrix of values, x_{ij} (i = 1, …, m) (j = 1, …, n). Where x_{ij} refers to the value of the jth feature associated with the ith sample. In an expression data application x_{ij} may represent, for instance, the expression value of gene i during a perturbation j.

Other techniques require a matrix of *pairwise* values, p_{ij} (i, j = 1, …, m), where p_{ij} represents the similarity (or disimilarity) value between the ith and jth objects to be clustered. In an expression data application p_{ij} may represent, for instance, the similarity or dissimilarity between the ith and jth genes under a biological condition.

Some basic measures for heterogeneity or compactness assessment are the *sum of squares*, L_1 *measures*, *intra-cluster diameter metrics* and the *sum of distances* (Everitt, 1993). Isolation may be measured by, for example, calculating the minimum distance between clusters, or the sum of dissimilarities between samples in a particular cluster and samples belonging to other clusters. The reader is referred to (Hansen and Jaumard, 1997) for a more detailed description on heterogeneity and isolation measures for clustering processes.

The second part of this section will introduce relevant clustering systems for expression data applications. This overview addresses three major types of clustering systems: a) hierarchical clustering, b) techniques based on iterative relocation, and c) adaptive solutions and other advances.

Azuaje F, and N. Bolshakova. "Clustering Genome Expression Data: Design and Evaluation Principles", in *Understanding and Using Microarray Analysis Techniques: A Practical Guide*, Berrar D, Dubitzky W and Granzow M, editors, London: Springer Verlag, 2002.

2.1 Key Clustering Approaches For Expression Data

2.1.1 Hierarchical Clustering

Hierarchical clustering is perhaps the best-known clustering method for expression data analyses. Chapter 14 discusses its implementation and applications in more detail. The main objective of this technique is to produce a tree like structure in which the nodes represent subsets of an expression data set. Thus, expression samples are joined to form groups, which are further joined until a single hierarchical tree (also known as dendrogram) is produced. There are different versions of hierarchical clustering, which depend, for example, on the metric used to assess the separation between clusters.

Several studies on the molecular classification of cancers and biological modelling have been based on this type of algorithms. Pioneering studies include an investigation by Eisen et al. (1998), which found that hierarchical clustering may be used to group genes of known similar function in *Saccharomyces cerevisiae*. Dhanasekaran et al. (2001) illustrates how dendrograms can reveal the variation in gene expression pattern between distinct pools of normal prostate samples. Perou et al. (2000) measured the variation in the expression of 1,753 genes in 84 experimental breast cancer samples "before and after" chemotherapy. This study shows how these patterns provide a distinctive molecular portrait of each tumour. Moreover, the tumours could be classified into subtypes based on the differences of their gene expression patterns.

2.1.2 Models Based On Iterative Relocation

This type of clustering algorithms involves a number of "learning" steps to search for an optimal partition of samples. Such processes may require: a) the specification of an initial partition of objects into a number of classes; b) the specification of a number of clustering parameters to implement the search process and assess the adequacy of its outcomes; c) a set of procedures to transform the structure or composition of a partition; and d) a repetitive sequence of such transformation procedures.

Some techniques included in this category are the *k-means* or *c-means* algorithms, and the *Kohonen Self-organising Map* (SOM). The k-means method categorises samples into a fixed number (k) of clusters, but it requires a priori knowledge on the number of clusters representing the expression data under study. SOMs have been applied to analyse expression profiles in several biomedical and *systems biology* studies (Quackenbush, 2001). This is a clustering approach based on hypothetical neural structures called feature

maps, which are adapted by the effect of the input expression samples to be classified. Thus, users may use SOMs to find and visualise clusters of similar expression patterns. The SOM-based model was one of the first machine learning techniques used to illustrate the molecular classification of cancer. Golub and colleagues (1999) reported a model to discover the distinction between acute myeloid leukaemia and acute lymphoblastic leukaemia. To illustrate the value of SOMs Tamayo and coworkers applied it to hematopoietic differentiation data (Tamayo et al., 1999). In this research SOMs organized samples into biologically relevant clusters that suggest, for example, genes involved in differentiation therapy used in the treatment of leukemia. Ideker and colleagues (2001) used SOMs to support an integrated approach to building and refining a cellular pathway model. Based on this method they identified a number of mRNAs responding to key perturbations of the yeast galactose-utilization pathway. Chapter 15 illustrates the application of SOMs in expression data.

2.1.3 Adaptive Systems And Other Advances

Some of these clustering solutions, unlike the methods introduced in Section 2.1.2, may not require the specification of an initial partition or knowledge on the underlying class structure. That is the case of some adaptations of the original SOM, such as Growing Cell Structures (GCS), which has been applied for the discovery of relevant expression patterns in biomedical studies (Azuaje, 2001a). Chapter 15 introduces the design and application of GCS-based clustering models.

Recent advances for expression data analysis include *Biclustering*, which consists of a one-step process to find direct correlations between a subset of features (genes or perturbations) and a subset of samples (genes or tissues) (Cheng and Church, 2000). From a biological perspective this is a useful approach because it allows the simultaneous clustering of genes and conditions, as well as the representation of multiple-cluster membership.

Other contributions have demonstrated how a supervised neural network can be used to perform automatic clustering or discovery of classes. A model based on a supervised neural network called Simplified Fuzzy ARTMAP (Kasuba, 1993) has been used to recognise relevant expression patterns for the classification of lymphomas (Azuaje, 2001b). From a user's point of view this type of models also offers a number of computational advantages. For example, the user only needs to specify a single clustering parameter, and the clustering process can be executed with a single processing iteration.

2.2 Basic Criteria For The Selection Of Clustering Techniques

Even when one would not expect the development of universal clustering solutions for genomic expression data, it is important to understand fundamental factors that may influence the choice and performance of the most appropriate technique. This section provides readers with basic criteria to select clustering techniques. These guidelines address questions such as: Which clustering algorithm should I use? Should I apply an alternative solution? How can results be improved by using different methods? This discussion does not intend to offer a formal framework for the selection of clustering algorithms, but to highlight important dimensions that may have to be taken into account for improving the quality of clustering-based studies.

Choosing "the best" algorithm for a particular problem may represent a challenging task. There are multiple clustering techniques that can be used to analyse expression data. Advantages and limitations may depend on factors such as the statistical nature of the data, pre-processing procedures, number of features etc. Moreover, it is not uncommon to observe inconsistent results when different clustering methods are tested on a particular data set. In order to make an appropriate choice is important to have a good understanding of:

a) the problem domain under study, and
b) the clustering options available.

Knowledge on the underlying biological problem may allow a scientist to choose a tool that satisfies certain requirements, such as the capacity to detect overlapping classes. Knowledge on the mathematical properties or processing dynamics of a clustering technique may significantly support the selection process. How does this algorithm represent similarity (or dissimilarity)?, how much relevance does it assign to cluster heterogeneity?, how does it implement the process of measuring cluster isolation?. Answers to these questions may indicate crucial directions for the selection of an adequate clustering algorithm.

Empirical studies have defined several *mathematical criteria of acceptability* (Fisher and Van Ness, 1971). For example, there may be clustering algorithms that are capable of guaranteeing the generation of partitions whose cluster structures do not intersect. Such algorithms may be called *convex admissible*. There are algorithms capable of generating partition results that are insensitive to the duplication of data samples. These techniques may be called *point proportion admissible*. Other clustering algorithms may be known as *monotone admissible* or *noise-tolerant* if their clustering outcomes are not affected by monotone transformations on the data.

It has been demonstrated, for instance, that both single-linkage and complete-linkage hierarchical clustering should be characterised as non-convex admissible, point proportion admissible and monotone admissible. The reader is referred to Fisher and Van Ness (1971) for a review on these and other mathematical criteria of acceptability.

Several algorithms indirectly assume that the cluster structure of the data under consideration exhibits particular characteristics. For instance, the k-means algorithm assumes that the shape of the clusters is spherical; and single-linkage hierarchical clustering assumes that the clusters are well separated. Unfortunately, this type of knowledge may not always be available in an expression data study. In this situation a solution may be to test a number of techniques on related data sets, which have previously been classified (a reference data set). Thus, a user may choose a clustering method if it produced consistent categorisation results in relation to such reference data set.

Specific user requirements may also influence a selection decision. For example, a scientist may be interested in observing direct relationships between classes and subclasses in a data partition. In this case, a hierarchical clustering approach may represent a basic solution. But in some studies hierarchical clustering results could be difficult to visualise because of the number of samples and features involved. Thus, for instance, a SOM may be considered to guide an exploratory analysis of the data.

In general the application of two or more clustering techniques may provide the basis for the synthesis of accurate and reliable results. A scientist may be more confident about the clustering experiments if very similar results are obtained by using different techniques. This approach may also include the implementation of voting strategies, consensus classifications, clustering fusion techniques and statistical measures of consistency (Everitt, 1993).

3. CLUSTER VALIDITY AND EVALUATION FRAMEWORKS FOR EXPRESSION DATA

Several clustering techniques have been applied to the analysis of expression data, but fewer approaches to the evaluation and validation of clustering results have been studied.

Once a clustering algorithm has been selected and applied scientists may deal with questions such as: Which is the best data partition?, which clusters should we consider for further analysis?, what is the right number of clusters?.

Answering those questions may represent a complex and time-consuming task. However, it has been shown that a robust strategy may consist of estimating the correct number of clusters based on validity indices (Azuaje, 2002a).

Such indices evaluate a measure, $Q(U)$, of quality of a partition, U, into c clusters. Thus, the main goal is to identify the partition of c clusters for which $Q(U)$ is optimal.

Two such cluster validity approaches are introduced and tested on expression data sets: The *Dunn's based indices* (Bezdek and Pal, 1998) and the *silhouette* method (Rousseeuw, 1987).

3.1 Assessing Cluster Quality With Dunn's Validity Indices

This index aims at identifying sets of clusters that are compact and well separated. For any partition $U \leftrightarrow X$: $X_1 \cup ... X_i \cup \ldots X_c$, where X_i represents the *ith* cluster of such partition, the Dunn's validation index, V, is defined as:

$$V(U) = \min_{1 \le i \le c} \left\{ \min_{\substack{1 \le j \le c \\ j \ne i}} \left\{ \frac{\delta(X_i, X_j)}{\max_{1 \le k \le c} \{\Delta(X_k)\}} \right\} \right\} \qquad (1)$$

$\delta(X_i,X_j)$ defines the distance between clusters X_i and X_j (intercluster distance); $\Delta(X_k)$ represents the intracluster distance of cluster X_k; and c is the number of clusters of partition U. The main goal of this measure is to maximise intercluster distances whilst minimising intracluster distances. Thus, large values of V correspond to good clusters. Therefore, the number of clusters that maximises V is taken as the optimal number of clusters, c (Bezdek and Pal, 1998).

In this study eighteen validity indices based on equation (1) were compared. These indices consist of different combinations of intercluster and intracluster distance techniques. Six intercluster distances, δ_i, $1 \le i \le 6$; and 3 intracluster distances, Δ_j, $1 \le j \le 3$ were implemented. Thus, for example, V_{13}, represents a validity index based on an intercluster distance, δ_1, and an intracluster distance Δ_3. The mathematical definitions of these intercluster and intracluster distances are described in Tables 13.1 and 13.2 respectively.

Azuaje F, and N. Bolshakova. "Clustering Genome Expression Data: Design and Evaluation Principles", in *Understanding and Using Microarray Analysis Techniques: A Practical Guide*, Berrar D, Dubitzky W and Granzow M, editors, London: Springer Verlag, 2002.

Table 13.1. Intercluster distances used to implement the Dunn's index. *S* and *T* are clusters from partition *U*; *d*(*x*,*y*) defines the distance between any two samples, *x* and *y*, belonging to *S* and *T* respectively; $|S|$ and $|T|$ provide the number of samples included in clusters *S* and *T* respectively.

$\delta_1(S,T) = \min\left\{ d(x,y) \right\}_{x \in S, y \in T}$	$\delta_4(S,T) = d(vs,vt)$ $vs = \frac{1}{\|S\|}\sum_{x \in S} x$ $vt = \frac{1}{\|T\|}\sum_{y \in T} y$
$\delta_2(S,T) = \max\left\{ d(x,y) \right\}_{x \in S, y \in T}$	$\delta_5(S,T) = \frac{1}{\|S\|+\|T\|}\left(\sum_{x \in S} d(x,vt) + \sum_{y \in T} d(y,vs) \right)$
$\delta_3(S,T) = \frac{1}{\|S\|\|T\|}\sum_{\substack{x \in S \\ y \in T}} d(x,y)$	$\delta_6(S,T) = \max\{\delta(S,T), \delta(T,S)\}$ $\delta(S,T) = \max_{x \in S}\left\{ \min_{y \in T}\{d(x,y)\} \right\}$ $\delta(T,S) = \max_{y \in T}\left\{ \min_{x \in S}\{d(x,y)\} \right\}$

Table 13.2. Intracluster distances used to implement the Dunn's index. *S* is a cluster from partition *U*; *d*(*x*,*y*) defines the distance between any two samples, *x* and *y*, belonging to *S*; $|S|$ represents the number of samples included in cluster *S*.

$\Delta_1(S) = \max_{x,y \in S}\{d(x,y)\}$
$\Delta_2(S) = \frac{1}{\|S\| \cdot (\|S\|-1)}\sum_{\substack{x,y \in S \\ x \neq y}} d(x,y)$
$\Delta_3(S) = 2\left(\frac{\sum_{x \in S} d(x,\bar{v})}{\|S\|} \right)$ $\bar{v} = \frac{1}{\|S\|}\sum_{x \in S} x$

As a way of illustration, this validation process is tested on expression data from a study on the molecular classification of lymphomas. Clustering is performed using the SOM algorithm. The expression levels from a number of genes with suspected roles in processes relevant in diffuse large B-cell lymphoma (DLBCL) were used as the features for the automatic clustering of a number of B-cell samples. The data consisted of 63 cases (45 DLBCL and 18 normal) described by the expression levels of 23 genes. These data were obtained from an investigation published by Alizadeh and colleagues (2000), who identified subgroups of DLBCL based on the analysis of the patterns generated by a specialized cDNA microarray technique. A key goal of this study was to distinguish two categories of DLBCL: Germinal Centre B-like DLBCL (GC B-like DLBCL) (22 samples) and Activated B-like DLBCL (23 samples) (Alizadehn et al., 2000). The full data and experimental methods are available on the Web site of Alizadeh et al. (http://llmpp.nih.gov/lymphoma).

Table 13.3 shows the values of the 18 validity indices and the average index at each number of clusters, c, for $c = 2$ to $c = 6$. The shaded entries correspond to the highest values of the indices, and $d(x,y)$ was calculated using the Euclidean distance. Fifteen of the indices indicated the correct value $c = 2$ while the remaining favour $c = 5$.

An examination of these partitions confirms that the case $c = 2$ represents the most appropriate prediction from a biomedical point of view. This partition accurately allows the identification of the two DLBCL subtypes: GC B-like and activated B-like. Table 13.4 describes the clusters obtained using the optimal value $c = 2$. Cluster 1 may be referred to as the cluster representing activated B-like DLBCL, while Cluster 2 recognises the subclass GC B-like DLBCL.

A more robust way to predict the optimal value for c may consist of: a) implementing a voting procedure, or b) calculating the average index value for each cluster configuration. Table 13.3 indicates that based on such criteria the best partition consist of two clusters.

Table 13.3. Predicting the correct number of clusters: Validity indices for expression clusters originating from B-cells. The entries represent the Dunn's values using 3 types of intracluster measures and 6 types of intercluster measures. Shaded entries represent the optimal number of clusters, *c*, predicted by each index.

Validity index	***c* = 2**	***c* = 3**	***c* = 4**	***c* = 5**	***c* = 6**
V_{11}	0.29	0.29	0.29	0.31	0.26
V_{21}	1.46	0.98	0.77	0.86	0.69
V_{31}	0.72	0.60	0.53	0.54	0.50
V_{41}	0.50	0.37	0.30	0.30	0.27
V_{51}	0.62	0.50	0.45	0.44	0.41
V_{61}	0.83	0.71	0.58	0.62	0.52
V_{12}	0.51	0.51	0.51	0.52	0.45
V_{22}	2.57	1.76	1.36	1.47	1.20
V_{32}	1.27	1.08	0.94	0.93	0.87
V_{42}	0.88	0.66	0.54	0.51	0.47
V_{52}	1.09	0.90	0.79	0.76	0.71
V_{62}	1.47	1.27	1.02	1.05	0.91
V_{13}	0.37	0.37	0.37	0.38	0.34
V_{23}	1.86	1.28	0.99	1.08	0.90
V_{33}	0.92	0.79	0.69	0.68	0.65
V_{43}	0.64	0.48	0.39	0.37	0.35
V_{53}	0.79	0.66	0.58	0.56	0.54
V_{63}	1.06	0.93	0.75	0.77	0.68
Average	0.99	0.79	0.66	0.68	0.60

Table 13.4. A relevant partition for a study on lymphoma data.

Cluster	**Description**
1 (Activated B-like DLBCL)	23 samples belonging to subtype Activated B-like DLBCL, 1 sample belonging to subtype GC B-like DLBCL 9 Normal samples
2 (GC B-like DLBCL)	21 samples belonging to subtype GC B-like DLBCL 9 Normal samples

Table 13.5. Validity indices for expression clusters originating from a study on DLBCL. The entries represent the average Dunn's values based on the distances shown in Tables 13.1 and 13.2, and using three measures for $d(x,y)$. Shaded entries represent the optimal number of clusters, c, predicted by each method. *E.dist.*: Euclidean distance; *M.dist.*: Manhattan distance; *C.dist.*: Chebychev distance.

Index based on	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$
E.dist.	0.99	0.79	0.66	0.68	0.60
M.dist.	1.57	1.21	1.02	1.04	0.92
C.dist.	0.97	0.79	0.70	0.69	0.63

The results shown in Table 13.3 were obtained when $d(x,y)$ was calculated using the well-known Euclidean distance (Tables 13.1 and 13.2). However there are several ways to define $d(x,y)$ such as the *Manhattan* and *Chebychev* metrics (Everitt, 1993). Therefore, an important problem is to know how the choice of $d(x,y)$ may influence the prediction process. Table 13.5 summarises the effects of three measures, $d(x,y)$, on the calculation of the Dunn's cluster validity indices. This analysis suggests that the estimation of the optimal partition is not sensitive to the type of metric, $d(x,y)$, implemented.

3.2 Assessing Cluster Validity With Silhouettes

For a given cluster, X_j ($j = 1,\ldots, c$), this method assigns to each sample of X_j a quality measure, $s(i)$ ($i = 1,\ldots, m$), known as the *silhouette width*. The silhouette width is a confidence indicator on the membership of the ith sample in cluster X_j.

The silhouette width for the ith sample in cluster X_j is defined as:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad (2)$$

where $a(i)$ is the average distance between the ith sample and all of samples included in X_j, 'max' is the maximum operator, and $b(i)$ is implemented as:

$$b(i) = \min_{X_k \neq X_j} (d(i, X_k)) \qquad (2.1)$$

where $d(i, X_k)$ is the average distance between the ith sample and all of the samples clustered in X_k; and 'min' represents the minimum value of $d(i, X_k)$ ($k = 1,\ldots, c; k \neq j$). It is easily seen from (2) that $-1 \leq s(i) \leq 1$.

When a $s(i)$ is close to 1, one may infer that the ith sample has been "well-clustered", i.e. it was assigned to an appropriate cluster. When a $s(i)$ is close to zero, it suggests that the ith sample could also be assigned to the nearest neighbouring cluster, i.e. such a sample lies equally far away from both clusters. If $s(i)$ is close to –1, one may argue that such a sample has been "misclassified".

Thus, for a given cluster, X_j ($j = 1,\ldots, c$), it is possible to calculate a cluster silhouette S_j, which characterises the heterogeneity and isolation properties of such a cluster:

$$S_j = \frac{1}{m}\sum_{i=1}^{m} s(i) \qquad (3)$$

It has been shown that for any partition $U \leftrightarrow X$: $X_1 \cup ... X_i \cup \ldots X_c$, a *global silhouette value*, GS_u, can be used as an effective validity index for U (Rousseeuw, 1987).

$$GS_u = \frac{1}{c}\sum_{j=1}^{c} S_j \qquad (4)$$

Furthermore, it has been demonstrated that equation (4) can be applied to estimate the most appropriate number of clusters for U. In this case the partition with the maximum S_u is taken as the optimal partition.

By way of example, this technique is tested on expression data originating from a study on the molecular classification of leukemia*s* (Golub et al., 1999). Clustering is again performed using SOM. The analysed data consisted of 38 bone marrow samples: 27 acute lymphoblastic leukemia (ALL) and 11 acute myeloid leukemia (AML), whose original descriptions and experimental protocols can be found on the *MIT Whitehead Institute* Web site (http://www.genome.wi.mit.edu/MPR).

Table 13.6. Silhouette values for expression clusters originating from leukemia samples. The entries represent the global silhouette values, GS_u, for each partition, and the silhouette values, S, for each cluster defining a partition. Shaded entries highlight the optimal number of clusters, c, predicted by this method.

c	GS_u	S_1	S_2	S_3	S_4	S_5	S_6
2	0.43	0.17	0.57				
3	0.14	0.11	0.35	0.11			
4	0.25	0.15	0.31	0.31	0.26		
5	0.19	0.07	0.45	0.23	0.23	0.21	
6	0.23	0.28	0.23	0.28	0.42	0.14	0.14

Table 13.6 shows the global silhouette values, GS_u, for each partition, and the silhouette values, S, at each number of clusters, c, for $c = 2$ to $c = 6$. The shaded entries correspond to the optimal values for this validation method. It predicts $c = 2$ as the best clustering configuration. Table 13.7 describes the clusters obtained using $c = 2$, which adequately distinguish ALL from AML samples.

Table 13.7. An optimal partition of leukemia samples which distinguishes ALL from AML samples.

Cluster	Description
1 (AML class)	11 AML samples 2 ALL samples
2 (ALL class)	25 ALL samples

Table 13.6 suggests that the partition consisting of 4 clusters may also be considered as a useful partition, because it generates the second highest GS_u. An examination of such a partition confirms that it represents relevant information relating to the detection of the ALL subclasses, B-cell and T-cell, as demonstrated by Golub and colleagues (1999). The composition of this alternative partition is described in Table 13.8.

Table 13.8. Predicting appropriate partitions in a leukemia study: distinction of subtypes of ALL samples.

Cluster	Description
1 (AML class)	10 AML samples
2 (Unlabeled class)	2 B-ALL samples 1 T-ALL samples 1 AML sample
3 (T-ALL subclass)	7 T-ALL samples 2 B-ALL samples
4 (B-ALL subclass)	15 B-ALL samples

The results shown in Table 13.6 were obtained using the well-known Euclidean distance. Alternative measures include, for example, the *Manhattan* and the *Chebychev* metrics. Table 13.9 summarises the effects of three distance measures on the calculation of the highest global silhouette values, GS_u. These results indicate that the estimation of the optimal partition is not sensitive to the type of distance metric chosen to implement equation (2).

Table 13.9. Prediction of the optimal partition based on silhouettes and different distance metrics for leukaemia data. The entries represent the global silhouette values, GS_u, for each partition. Shaded entries highlight the optimal number of clusters, *c*, predicted by each method. *E.dist.*: Euclidean distance; *M.dist.*: Manhattan distance; *C.dist.*: Chebychev distance.

GS_u based on	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$
E.dist.	0.43	0.14	0.25	0.19	0.23
M.dist.	0.43	0.14	0.25	0.19	0.23
C.dist.	0.43	0.14	0.25	0.19	0.23

4. CONCLUSIONS

This chapter has introduced key aspects of clustering systems for genomic expression data. An overview of the major types of clustering approaches, problems and design criteria was presented. It addressed the evaluation of clustering results and the prediction of optimal partitions. This problem, which has not traditionally received adequate attention from the expression research community, is crucial for the implementation of advanced clustering-based studies. A cluster evaluation framework may have a major impact on the generation of relevant and valid results. This paper shows how it may also support or guide biomedical knowledge discovery tasks. The clustering and validation techniques presented in this chapter may be applied to expression data of higher sample and feature set dimensionality.

A general approach to developing clustering applications may consist of the comparison, synthesis and validation of results obtained from different algorithms. For instance, in the case of hierarchical clustering there are tools that can support the combination of results into *consensus trees* (Bremer, 1990). However, additional methods will be required to automatically compare different partitions based on validation indices and/or graphical representations.

Other problems that deserve further research are the development of clustering techniques based on the direct correlation between subsets of samples and features, multiple-membership clustering, and context-oriented visual tools for clustering support (Azuaje, 2002b). Furthermore there is the need to improve, adapt and expand the use of statistical techniques to assess uncertainty and significance in genomic expression experiments.

Acknowledgements

This contribution was partly supported by the *Enterprise Ireland Research Innovation Fund* 2001.

REFERENCES

Alizadeh A.A., Eisen M.B., Davis R.E., Ma C., Lossos I.S., Rosenwald A., Boldrick J.C., Sabet H., Tran T., Yu X., Powell J.I., Yang L., Marti G.E., Moore T., Hudson J., Lu L., Lewis D.B., Tibshirani R., Sherlock G., Chan W.C., Greiner T.C., Weisenburger D.D., Armitage J.O., Warnke R., Levy R., Wilson W., Grever M.R., Bird J.C., Botstein D., Brown P.O., Staudt M. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000; 403:503-511

Azuaje F. An unsupervised neural network approach to discovering gene expression patterns in B-cell lymphoma. Online Journal of Bioinformatics 2001a; 1:23-41

Azuaje F. A computational neural approach to support the discovery of gene function and classes of cancer. IEEE Transactions on Biomedical Engineering 2001b; 48:332-339

Azuaje F. A cluster validity framework for genome expression data. Bioinformatics 2002a; 18:319-320

Azuaje F. In silico approaches to microarray-based disease classification and gene function discovery. Annals of Medicine 2002b; 34

Bezdek J.C., Pal N.R. Some new indexes of cluster validity. IEEE Transactions on Systems, Man and Cybernetics, Part B 1998; 28:301-315

Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, Radmacher M, Simon R, Yakhini Z, Ben-Dor A, Sampas N, Dougherty E, Wang E, Marincola F, Gooden C, Lueders J, Glatfelter A, Pollock P, Carpten J, Gillanders E, Leja D, Dietrich K, Beaudry C, Berens M, Alberts D, Sondak V, Hayward N, Trent J. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 2000; 406:536-540

Bremer K. Combinable component consensus. Cladistics 1990; 6:69-372

Cheng Y., Church G.M. Biclustering of expresssion data. Proceedings of ISMB 8th International Conference on Intelligent Systems for Molecular Biology; 2000 August 19 - 23; La Jolla. California, 2000.

Dhanasekaran S.M., Barrete T., Ghosh ., Shah R., Varambally S., Kurachi K., Pienta K., Rubin M., Chinnaiyan A. Delineation of prognostic biomarkers in prostate cancer. Nature 2001; 412:822-826

Eisen M.B., Spellman P., Brown P.O., Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. USA 1998; 95:14863-14868

Everitt, Brian, *Cluster Analysis*. London: Edward Arnold, 1993.

Fisher L., Van Ness J.W. Admissible clustering procedures. Biometrika 1971; 58:91-104

Azuaje F, and N. Bolshakova. "Clustering Genome Expression Data: Design and Evaluation Principles", in *Understanding and Using Microarray Analysis Techniques: A Practical Guide*, Berrar D, Dubitzky W and Granzow M, editors, London: Springer Verlag, 2002.

Golub T.R., Slonim D.K., Tamayo P., Huard C., Gassenbeck M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., Lander E.S. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999; 286:531-537

Hansen P., Jaumard B. Cluster analysis and mathematical programming. Mathematical Programming 1997; 79:191-215

Ideker T., Thorsson V., Ranish J.A., Christmas R., Buhler J., Eng J.K., Bumgarner R., Goodlett D.R., Aebersol R., Hood L. Integrated genomic and proteomic analyses of a systematically perturbated metabolic network. Science 2001; 292:929-933

Kasuba T. Simplified fuzzy ARTMAP. AI Expert 1993; 8:19-25

Perou C.M., Sorlie T., Eisen M.B., Van de Rijn M., Jeffrey S.S., Rees C.A., Pollack J.R., Ross D.T., Johnsen H., Aksien L.A., Fluge O., Pergamenschikov A., Williams C., Zhu S.X., Lonning P.E., Borresen-Dale A.L., Brown P.O., Botstein D. Molecular portraits of human breast tumours. Nature 2000; 406:747-752

Quackenbush J. Computational analysis of microarray data. Nature Reviews Genetics 2001; 2:418-427

Rousseeuw P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics 1987; 20:53-65

Tamayo P., Slonim D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E.S., Golub R. Intepretating patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA 1999; 96:2907-2912