Chapter 5

Clustering of Gene Expression Data

Methods for clustering, or unsupervised classification, have been studied for many decades. Vast numbers of different algorithms have been proposed (Jain & Dubes 1988). Clustering methods generally aim to identify subsets (*cluster*) in the data based on the similarity between single objects. Similar objects should be assigned to the same cluster, while objects which are not similar to each other, should be assigned to different clusters.

Cluster analysis is applied to search for data patterns that may reveal relationships between individual examples. Frequently, the data structure detected by cluster analysis can give first insights into the data producing mechanisms. Clustering can, therefore, be seen as exploratory data analysis. It has become a popular technique for data mining and knowledge discovery. Clustering is especially useful if prior knowledge is little or non-existent, since it requires minimal prior assumptions.

This feature has made clustering to a tool that is widely applied in the analysis of microarray data, where knowledge of the underlying regulatory networks has been limited. An important discovery was that the expression patterns of genes of similar function tend to cluster together. The examination of gene clusters, therefore, can lead to new knowledge about functions of single genes as well as about the behaviour of whole genetic networks.

5.1 Introduction

Microarrays have revolutionized the study of complex genetic networks by measuring the activities of many thousands genes simultaneously. They have become very powerful techniques in the systematic analysis of gene regulation. A landmark experiment was the study of yeast using microarrays containing all genes of the yeast genome (DeRisi *et al.* 1997). It has revealed an unexpected richness of expression patterns. To reveal these structures, the first step in data analysis is often the application of clustering analysis. One of its main purposes is to infer the function of novel genes by grouping them with genes of well-known functionality. This is based on the observation that genes showing similar expression patterns (*coexpressed genes*) are often functionally related and are controlled by the same regulatory mechanisms (*coregulated genes*). Expression clusters are, therefore, frequently enriched by genes of certain functions e.g. DNA replication, or protein synthesis. If a novel gene of unknown function falls into such a cluster, it is likely to serve the same functions as other members of the cluster. This 'guilt-by-association' method enables assigning possible functions to a large number of genes by clustering of coexpressed genes (Chu *et al.* 1998). Analysis of cluster structure can further identify the underlying mechanisms of metabolic and regulatory networks in the cell (Tavazoie *et al.* 1999). It is especially valuable for organism and cell types where little previous knowledge about their biology exists.

Different cluster algorithms have been applied to the analysis of expression data: k-means, SOM and hierarchical clustering to name just a few (Tavazoie *et al.* 1999, Törönen *et al.* 1999, Eisen *et al.* 1998). They all assign genes to clusters based on the similarity of their expression patterns. The borders between clusters are hard i.e. genes are assigned to exactly one cluster even if their expression profile is similar to several cluster patterns. For several time-course experiments, however, it has been pointed out that there are no well-defined boundaries between classes of temporal patterns (Cho *et al.* 1998, Spellman *et al.* 1998, Chu *et al.* 1998). For example, Chu *et al.* noted that genes were often highly correlated with the patterns of more than one cluster (Chu *et al.* 1998). This might be expected, since genes products frequently participate in more than on regulatory mechanism to different degrees. The regulation of a gene is generally not in an 'on-off', but gradual manner which allows a finer control of the gene's functions. A cluster algorithm should reflect this finding by differentiating how closely a gene follows the dominant cluster patterns. Fuzzy clustering appears as a good candidate for this task since it can assign genes degrees of membership to a cluster. The membership values can vary between zero and one. This feature enables fuzzy clustering to provide more information about the structure of gene expression data.

A second reason for applying fuzzy clustering is the large noise component in microarray data due to various biological and experimental factors. A common procedure to reduce noise in microarray data is the setting of a minimum threshold for change in expression. Genes below this threshold are excluded from further analysis. However, the exact threshold value remains arbitrary due to the lack of an established error model. Additionally, filtering may ex-

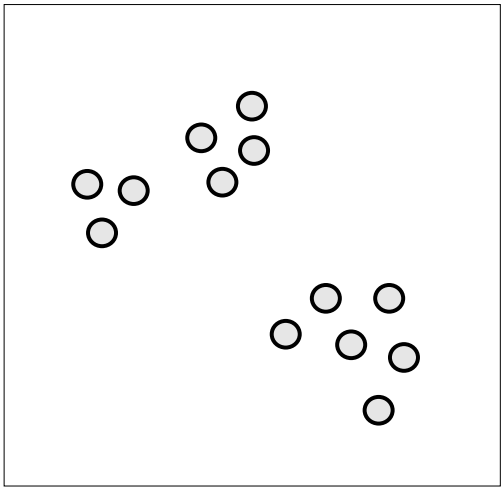


Figure 5.1: Are two or three clusters present? Both numbers of clusters seem to be correct depending on whether we consider the examples in the upper left to belong to one or two clusters. Thus, the scale of resolution determines the number of clusters. Ideally, a clustering method should give insights into the data structures on different scales.

clude interesting genes from further analysis. Fuzzy clustering isa valuable approach here since it is noise robust and may not require pre-filtering.

A crucial question is how many clusters can be found in the data? This is difficult to answer for gene expression data if clusters are not homogenous, but show sub-structures which can be interpreted as clusters themselves (Chu *et al.* 1998). A simplified example is presented in figure 5.1, where the number of clusters depends on the resolution scale. While hierarchical clustering indicates the different levels of clustering in the resulting dendrogram, partitional clustering algorithms lack the ability to indicate sub-structures in clusters. Additionally, no relationships between single clusters are indicated by most partitional clustering methods. Using fuzzy clustering, this drawback can be overcome, since it allows the definition of a coupling between single clusters.

In the next section, a review of methods for clustering gene expression data is given. Using fuzzy clustering, we re-analyse the yeast cell-cycle experiment by Cho *et al.* (section 2.2.1). After a description of the data pre-processing, the selection of clustering parameters is addressed. This is followed by a presentation of important features of fuzzy clustering and a comparison of the noise robustness of fuzzy and k-means clustering. A summary of the main results closes the chapter.

5.2 Review of Methods for Clustering Genes Based on Expression Data

This review focuses on methods applied to cluster genes. Clustering can also be applied to group tissue samples based on their overall gene expression profile.

While both applications can formally be regarded as equal, the computational task differs. Clustering of tissue samples is based on the expression values of thousands of genes and frequently demands the pre-selection of differentially expressed genes. The challenges in discovering relationships between tissue samples are discussed in more detail in the next chapter. This is contrasted by the clustering of genes, which is based on a much smaller number of measurements, usually less than 100.

In the first time-course experiments measuring transcription on a genome-wide level in yeast, genes were assigned to pre-defined classes. Cho *et al.* divided the yeast cell cycle into five phases (early G_1, late G_1, S, G_2 and M) based on a set of genes which were previously detected to be periodically transcribed (Cho *et al.* 1998). Over 400 genes were identified to be cell cycle regulated and assigned to the five phases. Similarly, Chu *et al.* used genes previously studied to define seven expression classes for the analysis of yeast sporulation (Chu *et al.* 1998). Genes were classified based on their correlation with the expression profiles of the classes. Sequences of known or putative regulatory sites were found by sequence alignment of similarly expressed genes.

Systematic detection of periodically transcribed genes was the goal of the study by Spellman *et al.* (Spellman *et al.* 1998). They analysed gene expression in yeast cell cultures synchronised by different methods. To detect periodic changes, the data were Fourier-transformed and compared with profiles of known cell cycle regulated genes. Using this method, over 800 genes were identified as periodically transcribed and assigned to cell cycle phases. Spellman *et al.* noted that no clear boundary existed between categories defined by the cell cycle phases e.g. genes in the late G_2 and early M phase showed a very similar expression profile.

The study by Eisen *et al.* can be considered a milestone in the field of gene expression analysis (Eisen *et al.* 1998). They introduced cluster analysis to systematically identify groups of genes with similar expression patterns. Using hierarchical clustering, they found that genes of similar function cluster together in yeast and human expression data. The results by Eisen *et al.* opened the way for a large variety of clustering methods to be proposed for the analysis of microarray data.

Using k-means clustering, Tavazoie *et al.* re-analysed the data by Cho *et al.* (Tavazoie *et al.* 1999). Mapping of genes to functional categories showed that many expression clusters were significantly enriched by functionally related genes. Known or novel cis-regulatory motifs were identified by aligning the up-stream regions of genes assigned to the same cluster. The 'tightness' of expression clusters was shown to be correlated with the presence of significant sequence motifs. 'Tighter' clusters tended also to be more enriched by functionally related genes.

The use of self-organising maps (SOMs) for gene expression analysis was proposed in two studies (Tamayo *et al.* 1999, Törönen *et al.* 1999). SOMs consist of a grid with pre-defined geometry. The vertices of the grid represent the cluster centres and are iteratively adjusted based on their distance to the data objects. As adjacent vertices are moved simultaneously, the use of the grid structure ensures that clusters with similar expression patterns are mapped to vertices

which are close to each other on the grid. Tamayo *et al.* pointed out that this clustering structure is favourable for interpretation (Tamayo *et al.* 1999). Tamayo *et al.* used a 6x5 rectangular grid to cluster the yeast cell cycle data. They detected clusters that closely match those identified by Cho *et al.* No recommendation, however, was given on how to select the initial structure of the SOM. A similar clustering approach based on SOMs was presented by Törönen *et al.* analysing yeast gene expression (Törönen *et al.* 1999). For visualisation of the results, Sammon's mapping was used.

Heyer *et al.* criticised the use of k-means clustering and SOMs for gene expression analysis as these methods do not allow direct control of the within-cluster variation (Heyer *et al.* 1999). If the pre-specified number of clusters is too small, unrelated expression patterns are clustered together. If the number is to large, genes with similar expression may be assigned to different clusters. To overcome this difficulty, Heyer *et al.* introduced a quality measure defined by the diameter of the cluster. Based on this measure, an iterative method termed 'quality clustering' was proposed. After selecting a gene to define a candidate cluster, similar genes are assigned to this cluster until it surpasses the pre-chosen diameter. This is repeated for all genes. The largest cluster is determined and its genes excluded from further analysis. For the remaining genes, new candidate clusters are formed. The iterative process stops if the cluster size falls below a pre-specified threshold. To achieve a robust measure of similarity between expressed genes, a jackknife correlation was defined.

Sharon and Shamir introduced a graph-theoretical clustering method termed CLICK (Sharon & Shamir 1999). Gene expression vectors correspond to vertices in a connected weighted graph. The connecting weights are based on the similarity of expression vectors. CLICK consists of two steps: First, connections with low weights are iteratively removed. Second, remaining clusters are merged if they are similar. Using different measures for clustering quality, Sharon and Shamir showed that CLICK outperformed several other cluster methods.

Many clustering approaches treat all measurements equally for the grouping of genes. This is appropriate if groups of genes are coregulated in all measurements. It becomes problematic if genes are coregulated only for a subset of measurements. The remaining measurements contribute noise to the clustering process and should preferably be neglected by the clustering procedure. For this task, Getz *et al.* proposed coupled two-way cluster analysis (Getz *et al.* 2000). Genes and measurements were clustered simultaneously. The gene expression matrix was partitioned based on the detected clusters. For each resulting sub-matrix, a further two-way cluster analysis was performed independently. By this iterative procedure, clusters of highly correlated genes and measurements were identified. A related approach called 'biclustering' was presented by Cheng and Church (Cheng & Church 2000). They used a greedy search algorithm to exclude samples or genes which contribute most to within-cluster variance. Detected clusters were masked to allow for identification of further clusters. Masking was performed by replacing the clusters' expression values by random numbers.

A problem in cluster analysis is determining the correct number of clusters

in the data. For hierarchical clustering, the researcher has to decide how many clusters exist based on the dendrogram. For partitional clustering, the number of clusters is frequently a parameter which has to be specified *a priori*. Since both approaches pose difficulties for the analysis of gene expression data, several methods have been proposed to determine the optimal number of clusters.

Levine and Domany used a resampling scheme to find stable clustering (Levine & Domany 2001). Subsets were formed by random splits of the original data. Clusters detected in the subsets were then compared to the original clusters. Clustering stability was determined by calculating the percentage of genes which were clustered together for the full and resampled data sets. The number of clusters was considered optimal if it maximised the stability of the clustering.

Lukashin and Fuchs proposed an alternative approach to selecting the optimal number of clusters (Lukashin & Fuchs 2000). It is based on the analysis of the distribution of distances between pairs of genes. First, a baseline distribution for a randomised version of the data was calculated. The value for the lower 5% of the distances was then determined. This gives the threshold for the maximal distance between gene pairs in the same cluster for the original data. Finally, the cluster number is gradually increased until only a pre-specified percentage e.g. 5% of the distances within clusters surpasses the threshold. Lukashin and Fuchs used stimulated annealing for clustering gene expression data.

To find the statistical significance of dendrograms produced by hierarchical clustering, Hughes *et al.* used a bootstrapping techniques (Hughes *et al.* 2000). The test statistic was the overall similarity of genes within a cluster. Each bifurcation of the cluster tree was assigned a p-value by comparing the test statistic for the original clustering with the ones obtained for randomised data.

A difficulty related to determining the number of clusters is the assessment of the clustering reliability. This problem was studied by Kerr and Churchill using bootstrapping cluster analysis (Kerr & Churchill 2001). They based their analysis on an ANOVA model of the data. The residuals of the model provided an estimate of the error distribution in the experiment. By resampling the residuals, simulated data sets were created and their cluster profiles compared with the original profiles. Variation within the replicated clusterings indicates the stability of the clusters.

Clustering methods based on probability models offer an alternative to heuristically motivated methods such as k-means. Model-based clustering assumes that the data values are generated by a mixture of probabilistic distributions. Yeung *et al.* used Gaussian mixture models to determine clusters in gene expression data (Yeung *et al.* 2001). The model parameters were estimated by the expectation-maximisation algorithm. To determine the number of clusters, the Bayesian Information Criterion was used. It indicates how well the data are fitted by the model. Yeung *et al.* showed that model-based clustering performs well for the data sets analysed, which were, however, of low complexity.

5.3 Data Pre-Processing and Normalisation

In the yeast cell cycle experiment by Cho *et al.*, 6178 genes were monitored at 17 time points over a span of 160 minutes using Affymetrix chips (Cho *et al.* 1998). This gave a total of over 100000 measurement values. The expression values for the time point $t = 90$ minutes were excluded in our analysis as these data were considered erroneous (Tavazoie *et al.* 1999). Further, genes with less than 75% of the measurements present were excluded. This reduced the number of genes for the cluster analysis to 6101. To convert the Affymetrix data into ratios, the measured intensities of each gene were divided by their average values. The arrays were globally normalised i.e. the total intensity of each array was linearly scaled to have the same value. To treat positive and negative fold changes equally, the data were $\log_2$-transformed.

Missing values Artifacts such as printing errors, dust and scratches on the array frequently lead to missing values in microarray experiments. The data set used by Cho *et al.* contained over 6000 missing measurement values i.e. ca. 6% of the expression values were missing. Fuzzy clustering, like many other cluster algorithms, does not allow for missing values. Different strategies to overcome this problem exist. Genes might be excluded if some of their expression values are missing. This, however, would have lead to a dramatic reduction of the number of genes included in our analysis, since all expression values were present for less than a third of the genes. Alternatively, we can estimate the missing values based on present expression data. The following knn method was applied:

A missing value of gene i at time point t is estimated by the average values for time t of the 10 nearest neighbouring genes j. The distance was calculated by

$$d(\mathbf{g}_i, \mathbf{g}_j)^2 = \frac{n}{n-m} \sum_k (g_{ik} - g_{jk})^2$$

where $\mathbf{g}_i$ is the gene expression vector for gene i, $\mathbf{g}_j$ is the gene expression vector for neighbouring gene j, n is the number of arrays in the time-course experiment and m is the number of measurements which are missing for gene i or j or both. The sum includes only measurements for which both gene expression values (g_{ik}, g_{jk}) are present.

This procedure exploits the high correlation between genes in expression data. It assumes that genes which are well correlated for existing measurements are also correlated for missing measurements. In a recent comparison the knn method for estimating missing values proved to be superior to other methods (Trojanskaya *et al.* 2001).

Filtering Most cluster analyses include a filtering step to remove genes which are expressed at low levels or show only small changes in expression. Different filtering procedures have been proposed for the analysis of the expression data analysed here. Heyer *et al.* excluded all genes with a mean or variance in the lower 25% of the data (Heyer *et al.* 1999). Tavazoie *et al.* included only 3000

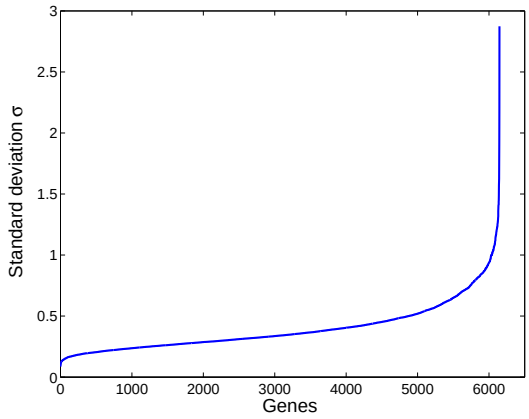


Figure 5.2: Standard deviation of gene expression values before normalisation. The genes were ordered by their standard deviation.

genes, which showed the largest variation (Tavazoie *et al.* 1999). Tamayo *et al.* reduced the number of genes for analysis to as few as 823 by setting thresholds for the relative and the absolute change (Tamayo *et al.* 1999). Inspection of the different measures proposed, however, revealed that no obvious threshold for filtering existed. For example, figure 5.2 shows the standard deviation of genes in the experiment. The transition between low and high values for variation is smooth and no particular cut-off point is indicated. Thus, the value of a filtering threshold remains arbitrary. Alternatively, an error model might be introduced to select significantly expressed genes (section 4.6). This, however, is difficult since the experiment by Cho *et al.* lacks replicates.

As no stringent filtering procedure currently exists, we avoided any prior filtering of gene data. This prevents the loss of biologically important information, as many genes show only small changes in transcription (Hughes *et al.* 2000). The inclusion of all genes in the analysis demands, however, a clustering method which is robust against noise. We demonstrate that this is the case for fuzzy clustering in section 5.4 and 5.6.

Standardisation The expression values measured for a gene define its gene expression vector. For cluster analysis, these vectors have to be standardised, as coexpressed genes frequently show similar changes in expression but may differ in the overall expression rate. This corresponds to expression vectors of similar directions, but different length. Since the clustering is performed in Euclidian space, coexpressed genes may thus be wrongly assigned to different clusters. Therefore, the expression values of genes were standardised to have a mean value of zero and a standard deviation of one to ensure that vectors of genes with similar changes in expression are close in Euclidean space. The Euclidean distance d is then closely linked to the correlation coefficient r ($d = \sqrt{2 - 2r}$). The minimal distance $d = 0$ is achieved for $r = 1$; the maximal distance $d = 2$ for $r = -1$.

5.4 Determination of the Fuzzy Clustering Parameters and Cluster Validation

To use fuzzy c-means (FCM) for cluster analysis, several parameters have to be specified. Besides the number of clusters c and the fuzzification parameter m, users must choose values of the minimal change ϵ in the objective function for termination and the maximal number T_{max} of iterations (section 3.3.1). For termination of the clustering process, we specified the minimal change $\epsilon = 0.001$ and the maximal number of iterations $T_{max} = 100$. Different choices can be made for the distance metric ($\| \cdot \|_A$). Euclidean, diagonal and Mahalanobis distances are frequently used. In this study we applied the Euclidean metric, since the observed overall variances for the different time points were similar and we wanted to treat each time point equally.

The fuzzification parameter m is a crucial clustering parameter since it determines the influence of noise genes on the cluster analysis. For $m \to 1$, it can be shown that the clustering becomes hard (Bezdak 1981). The FCM algorithm is then equivalent to the k-means clustering. The membership values are either one or zero. All genes of a cluster are treated equally for the calculation of the cluster centre. Increasing the parameter m reduces the influence of genes with low membership values as can be seen in equation 3.6. Gene expression vectors with large noise content generally have a low membership value, since the corresponding genes are not well represented by a single cluster, but rather are partially assigned to several clusters. Selection of the fuzzification parameter m determines the influence of noise on the clustering process. It also allows investigation of the stability of clusters. We define stable clusters as clusters which show only minor changes in their structure with variation of the parameters c and m. Stable clusters are generally isolated and compact. This is contrasted by weak clusters which lose their internal structure or disappear if m was increased (section 5.5). For $m \to \infty$, the partition approaches maximal fuzziness. A gene i is assigned to all clusters equally and the partition matrix becomes uniform.

Monitoring the clustering results for increasing m therefore gives insights into the structures of the data. We will use this feature to prevent the detection of clusters in random data.

Construction of a baseline distribution A major problem with hard clustering algorithms such as k-means or SOMs is that they always assign objects to a pre-selected number of clusters. Even if the data are random, distinct clusters are formed. This is illustrated in figure 5.3 which shows clusters detected by k-means for randomised yeast cell cycle data. The randomisation was achieved by random permutation of the time order of every gene independently. Data structures occurring by chance were identified as clusters by k-means. This feature of hard clustering is problematic, as it can easily lead to false results.

This drawback in hard clustering can be overcome by fuzzy clustering. Since the fuzzification parameter m controls the sensitivity of the clustering process to noise, we can adjust m to prevent the detection of clusters in the randomised

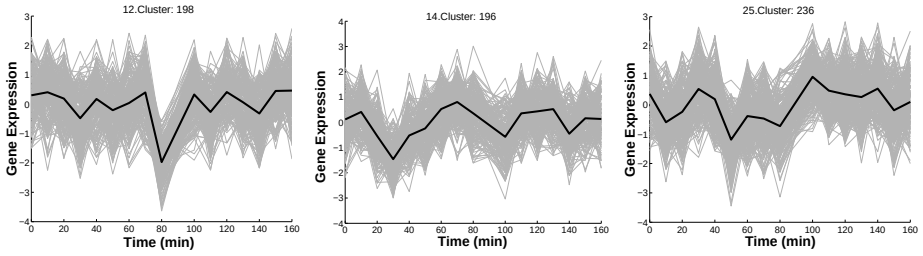


Figure 5.3: Example clusters for k-means (k = 30) clustering of randomised expression data. The cluster centres are indicated by solid lines.

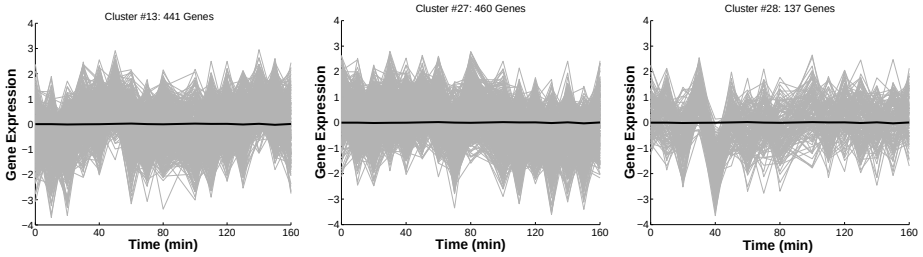


Figure 5.4: Example clusters for FCM ($m = 1.25$, $c = 30$) clusters of randomised expression data. The cluster centres were indicated by solid lines. The illustrated cluster structures are somewhat misleading, since genes were assigned to clusters for which they had a maximal membership value. The membership values showed, however, only minimal variation. The mean membership value was $0.033 \approx 1/30$ with a standard deviation $1.3 \cdot 10^{-5}$. Essentially, all genes were equally included for the calculation of each cluster centre.

data. For the determination of parameter m, randomised data was clustered with parameter values between 1.05 and 3.05. The number of clusters was varied between 2 and 60. Inspection of the FCM clustering results showed that no clusters are detected for $m \geq 1.25$. The partition matrices became uniform i.e. every gene was approximately equally assigned to all clusters. All genes were included equally for the calculation of each cluster centre. Therefore, the cluster centres derived were approximately zero vectors i.e. vectors with all coordinates equaling zero. Examples are shown in figure 5.4. The fuzzification parameter was therefore set to $m \geq 1.25$ in the following analyses.

Determination of the number of clusters After the selecting the fuzzification parameter m, the correct number of clusters has to be determined. For this task, we gradually increased the cluster number c in the FCM algorithm and examined the results of the clustering. We observed that the membership values of genes tend to spread more between clusters as the detected clusters become more similar for increasing c. Especially for less isolated clusters, the number of genes with a membership value larger than 0.5 decreased. Finally

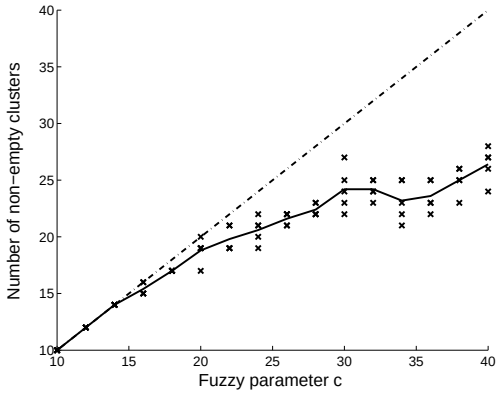


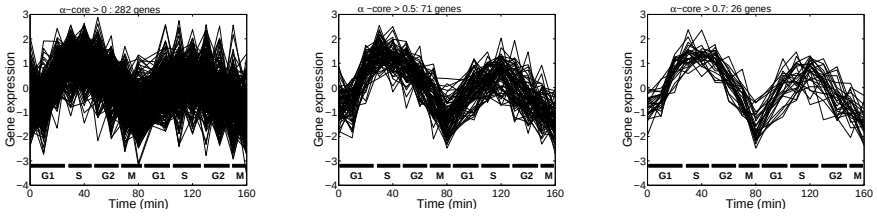
Figure 5.5: Appearance of empty clusters for FCM clustering for increasing parameter c. Five repeated clustering with random initiations were performed and the number of non-empty clusters monitored. The fuzzification parameter m was set to 1.25. The dotted line shows the maximum possible number of non-empty clusters.

clusters are detected for which none of the genes surpasses the membership value of 0.5. We call these clusters *empty* clusters as no gene is primarily assigned to them. This allows the setting of the parameter c.

We repeatedly clustered the expression data and monitored the number of non-empty clusters. Figure 5.5 shows that none of the repeated clusterings produced empty clusters for $c \leq 15$. Increasing c leads to the appearance of empty clusters for some clusterings. For $c = 20$, at least one clustering did not result in empty clusters.

A further increase of c always produced empty clusters, although the number of non-empty clusters also increased. Since empty clusters can be easily detected, the setting of the cluster number is less problematic for fuzzy clustering compared with hard clustering which does not indicate the quality of clusters. Clusters identified by FCM can be analysed separately for their stability and the number of genes with membership values less than 0.5. Increasing the cluster number to $c > 20$ may even be favourable for the study of local structures as we discuss in section 5.7.

For the following analysis, we selected $c = 20$ to prevent the appearance of empty clusters. The same number of clusters was found by Luskashin and Fuchs using stimulated annealing (Lukashin & Fuchs 2000). For $c = 20$ and $m = 1.25$, an average of 1560 genes was assigned maximal membership values less than 0.5 i.e. over 25% of the genes were not primarily assigned to a single cluster.



(a) $\alpha = 0$ (b) $\alpha = 0.5$ (c) $\alpha = 0.7$

Figure 5.6: Different α-cores for FCM cluster ($c = 20$ and $m = 1.25$): Sub-figure (a) shows only those genes which were primarily assigned to the cluster.

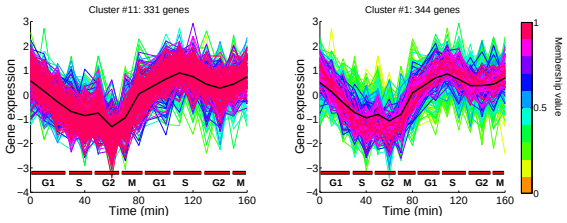
5.5 Analysis of Information Rich Structures of Fuzzy Clustering

Differentiation in cluster membership and profiling of cluster cores For FCM, the cluster centres result from the weighted sum of all cluster members and show the overall expression patterns of clusters. The membership values μ_{ik} indicate how well the gene expression vector $\mathbf{g_i}$ is represented by the cluster centre $\mathbf{c_k}$. Low values μ_{ik} point to a poor representation of gene i by $\mathbf{c_k}$. Large values μ_{ik} point to a high correlation of $\mathbf{g_i}$ with $\mathbf{c_k}$. Membership values can also indicate the similarity of vectors to each other. If two gene expression vectors have a high membership value for a specific cluster, they are generally similar to each other. This is the basis for the definition of the *core* of a cluster. We define that genes with membership values larger than a chosen threshold α belong to the α-*core* of the cluster. This overcomes the limitations of hard partitional clustering which does not define any relationship between genes within a cluster. Similarly to hierarchical clustering, the internal structures of clusters become accessible.

Figure 5.6 shows different α-cores for an expression cluster. Genes can be differentiated by examining wether they are included in a certain α-core. Figure 5.6a shows all genes which were primarily assigned to the cluster. This cluster structure is equivalent to hard clustering. The within-cluster variation of the gene expression values is large. Local peaks indicated a high background noise. Setting the α-threshold to 0.5 decreased the variation within the cluster. Genes which were poorly correlated with the overall cluster pattern were excluded. The periodicity of the remaining genes became more clearly visible.

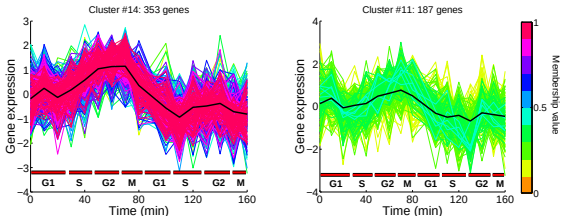
Increasing the α-threshold to 0.7 led to a decreased number of genes included in the α-core. Only 28 genes of 282 originally assigned to the cluster remained. Simultaneously, the average within-cluster variation was reduced from 0.78 for $\alpha = 0$ to 0.40 for $\alpha = 0.7$.

Analysing the α-cores facilitates the identification of underlying networks as genes can be ranked based on their membership values. The use of the α-threshold can therefore act as *a posteriori* filtering of genes. This contrasts with previously discussed procedures which demand the problematic setting of



(a) $m_1 = 1.15$ (b) $m_2 = 1.25$

Figure 5.7: Stable clusters maintain their core for increasing m. An example of a stable cluster is shown for two FCM clusterings ($c = 20$, $m_1 = 1.15$, $m_2 = 1.25$). The color bar on the right indicates the colour coding of the membership values.



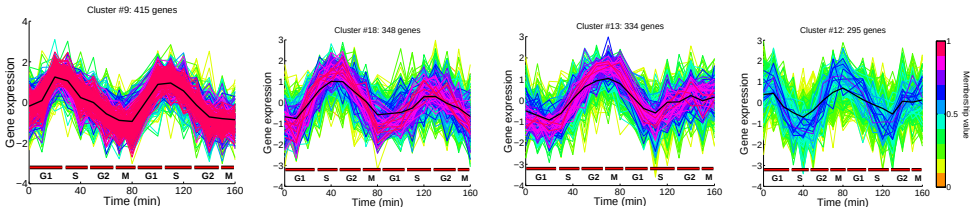
(a) $m_1 = 1.15$ (b) $m_2 = 1.25$

Figure 5.8: Weak clusters lose their core for increasing m. The FCM parameters were $c = 20$, $m_1 = 1.15$, $m_2 = 1.25$. The membership values as well as the number of genes are reduced.

a threshold *a priori* to the cluster analysis. Fuzzy clustering thus avoids the exclusion of possibly important genes from cluster analysis.

Stability of clusters Besides determining the global fuzziness of data partitions, variation of the fuzziness parameter m can be used to gain insight into the internal structure of single clusters. The choice of the parameter m controls the fuzziness of the cluster i.e. the distributions of the cluster membership values. Small m led to a large cluster core with little variation of the membership functions. When m approached one, the cluster membership values became either one or zero. Increasing m yielded partitionings with more distributed membership values. The α-core of the clusters became more differentiated e.g. a larger α-threshold resulted in smaller cluster cores. By adjusting the parameter m, the distribution of membership values can therefore be fine-tuned and internal cluster structures can be examined. To facilitate this examination, we colour-coded the α-core of clusters (figures 5.7). Temporal patterns can readily be detected.

Variation of the fuzzification parameter m also indicate the stability of clusters. Clusters in figures 5.7 and 5.8 illustrate this procedure. Both clusters seemed to have a well-defined α-core for $m = 1.15$. Differences appeared, how-



(a) G_1 (b) S (c) G_2 (d) M

Figure 5.9: Clusters with periodic gene expression patterns: The labelling of the clusters is based on the peak of gene expression. Periodic clusters had generally large α-cores. The FCM parameters were $c = 20$ and $m = 1.25$.

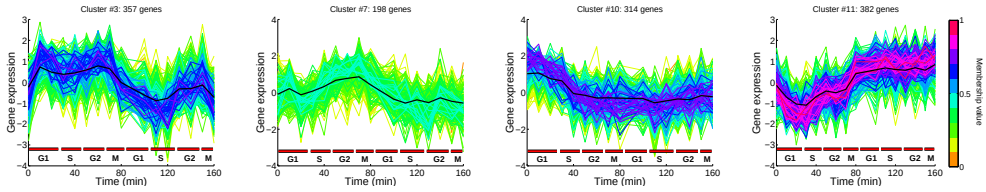


Figure 5.10: Clusters with aperiodic expression patterns: These clusters were generally weaker than the periodic clusters detected. FCM parameters were $c = 20$ and $m = 1.25$.

ever, for $m = 1.25$. For the stable cluster in figure 5.7, most genes retained their large membership values, whereas the membership values decreased considerably for the weak cluster in figure 5.8. Additionally, the number of genes assigned to the strong cluster stayed approximately the same. This is contrasted by the weak cluster which lost genes for larger m.

By continually increasing m, it is possible to rank clusters according to their stability. Biologically, this may give indications of how strongly genes are coregulated in the underlying genetic networks.

Periodicity of clusters The clustering process yielded two types of clusters. The first type consisted of clusters with periodic expression patterns (figure 5.9). The clusters are usually labelled according to their peak time in the phases of the cell cycle (G_1, S, G_2, M). Generally, these clusters showed a high stability and were enriched with genes required for functioning of the cell cycle.

The second group of clusters shows a variety of aperiodic patterns (figure 5.10). The occurrence of these clusters may be related to initial conditions in the time-course experiment. For synchronisation of the cell culture, temperature-sensitive cells were arrested in the late G_1 phase. Lowering the temperature to a permissive range might have led to specific response patterns of genes which were revealed by clustering. Several of the aperiodic clusters showed only low stability. This may point to a high background noise or weak coregulation of genes in the clusters.

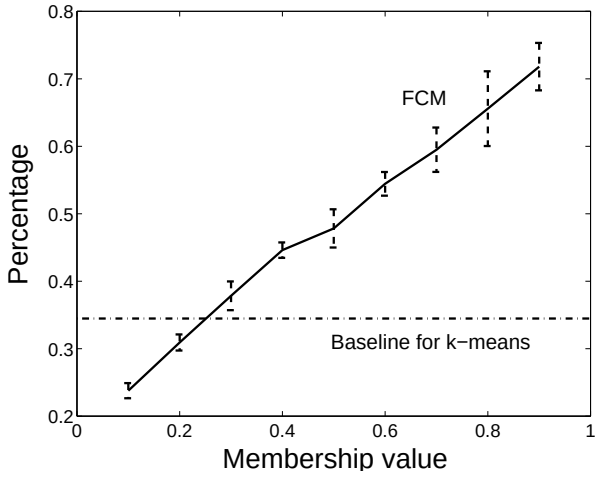


Figure 5.11: Percentage of gene pairs which clustered together for original and 'noisy' data. Derivation of mean values and error interval were based on five clusterings of independently generated 'noisy' data sets. The percentages for a membership value x give the fraction of gene pairs clustered together with both genes having a maximal membership value $x \pm 0.05$ for the original clustering.

5.6 Noise Robustness

Filtering procedures aim to reduce noise, but may lead to a loss of valuable information. We therefore included all genes in the clustering analysis. Fuzzy clustering is a favourable method for such an approach, as it indicates gene expression vectors which should be considered as noisy.

The previous sections showed that FCM assigns membership values to genes based on the similarity of their expression to the overall pattern of the cluster. Expression vectors with low membership values may thus considered as noisy. To test this hypothesis, we analysed the stability of fuzzy clustering against increased noise. For this task, random Gaussian noise was added to the gene expression data. The following formula was used:

$$\tilde{g}_i = g_i + 0.5 \cdot N(0,1)$$

where g_i is the original expression vector of gene i, $\tilde{g}_i$ is the expression vector with added noise and $N(0,1)$ is the standard normal distribution.

The 'noisy' gene expression data was clustered and the results compared to the clustering of the original data. To evaluate the stability of fuzzy clustering, we calculated the percentage of pairs of genes which fall in the same clusters for both clusterings. Ideally, gene pairs for the original clustering should also be found for the clustering of the 'noisy' data. This is, however, only partially the case for the data analysed here. K-means clustering, which was used for comparison, assigns on average only 34% of the gene pairs to the same cluster. This can be improved by fuzzy clustering, since it can differentiate gene pairs based on their similarity of expression in the original clusters. Pairs with high membership values for the original cluster are more likely to be clustered to-

gether for noisy data as shown in figure 5.11. Less than 30% of the pairs are clustered together again if both genes had membership values of less than 0.3 for the original cluster. Over 70% of pairs, however, were clustered together if both genes had membership value of 0.85 or higher. Cluster cores obtained by high α-threshold are therefore more likely to reflect the 'noise-free' cluster structure.

5.7 Global Clustering Structures

An interesting feature of fuzzy clustering is the overlap or *coupling* between clusters. Coupling V_{kl} between cluster k and cluster l can be defined by

$$V_{kl} = \frac{1}{N} \sum_{i=1}^{N} \mu_{ik} \mu_{il}$$

where N is the total number of expression vectors. The coupling indicates how many genes are shared by two clusters. Clusters which have a low coupling show distinct overall patterns. If the coupling is large, clusters are more similar. The coupling V defines thus a similarity measure for pairs of fuzzy clusters.

This allows the analysis of global clustering structures obtained by FCM, since relationships between clusters can be examined. Figure 5.12 shows the overall clustering structure for three different settings of cluster number ($c = 12, 18, 24$). For $c = 12$, the coupling between clusters was generally weak (figure 5.12-I). Several isolated clusters were produced. An example of such an isolated cluster is the G_1 cluster in figure 5.12d. It can be considered a stable cluster, since it remains isolated with increasing c (figure 5.12e). This is contrasted by the G_2 cluster which was split into two sub-cluster (figure 5.12a,b,c). Both clusters are strongly coupled, as the overall pattern is similar. Note that two G_2 sub-clusters show also some differences. The sub-cluster in figure 5.12b has a dominant expression peak during the first cell cycle, whereas peaks in both cell cycle are of similar amplitude for the cluster in 5.12c. The membership values for the second cluster were lower, so it can be considered as a weaker cluster. This demonstrates that subtle differences can be revealed by sub-clustering of fuzzy partitions.

If the cluster number was increased further ($c = 24$), genes tended to be assigned to several clusters (figure 5.12-III) The coupling between the clusters, thus, became stronger and no isolated cluster remained. Additionally, a larger parameter c led to empty clusters which can be, however, easily detected. Empty clusters usually showed strong coupling to many other clusters, as they were poorly isolated and less compact.

5.8 Summary

Fuzzy clustering includes advantages of partitional and hierarchical clustering. As with partitional clustering method, fuzzy clustering is based on the overall structure allowing for more robust clustering. Additionally internal cluster

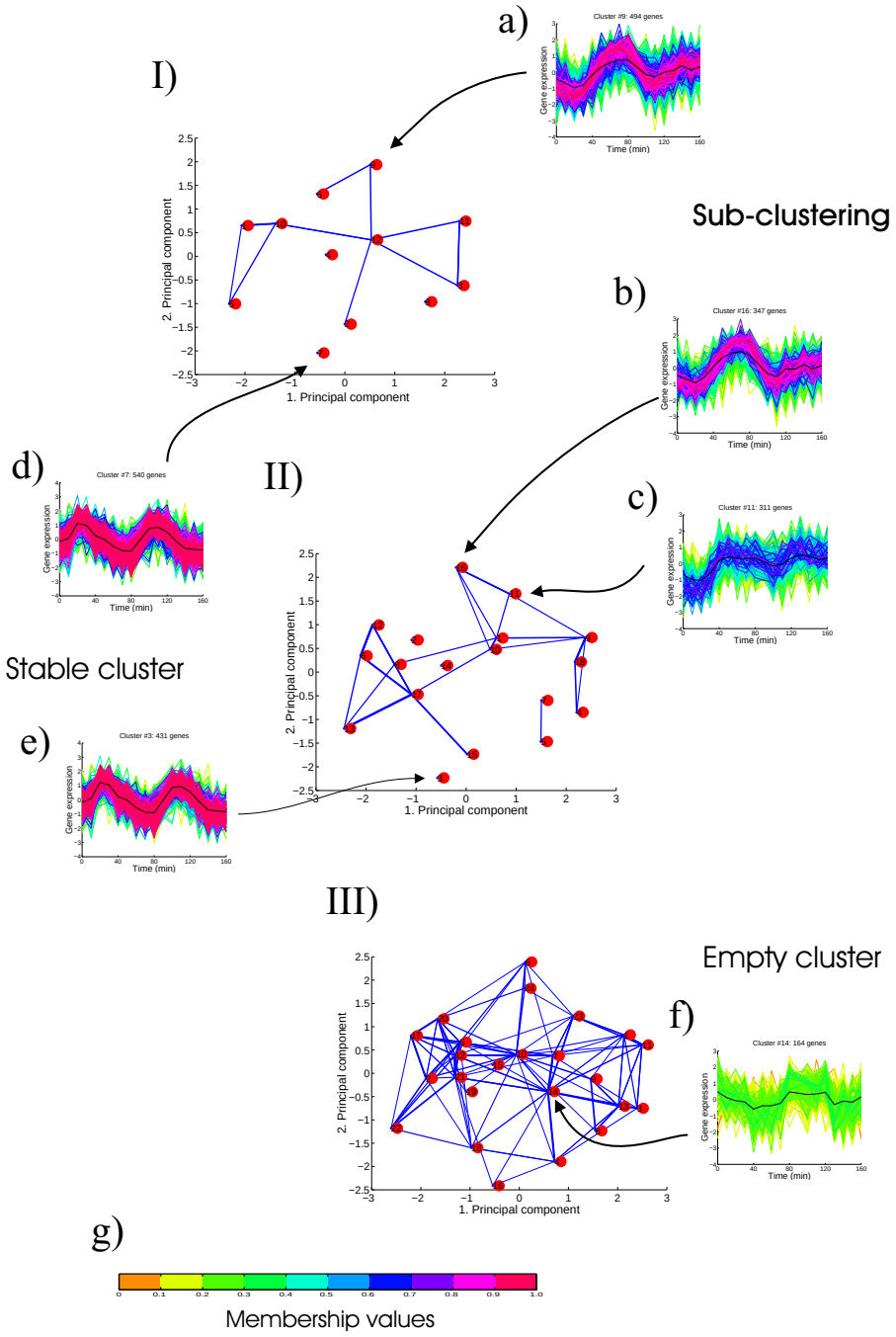


Figure 5.12: FCM sub-clustering: Three clusterings (I,II,III) for $c = 12, 18, 24$ were analysed. To visualise the global structure of the cluster space, we used PCA to project the cluster centres in two-dimensional space. The principal components were derived for clustering III. The strength of coupling of two clusters is represented by the width of connecting line between them. Sub-figures (a)-(e) shows the core structures for several example clusters. Sub-figure (g) shows the colour encoding of the membership values.

structures are produced. Relationships between genes within a cluster as well as between clusters can be defined and visualised. This facilitates the discovery of knowledge, as more information about the data structure is obtained. K-means clustering lacks these features, since no relationships between clusters are given by the k-means method. Hierarchical clustering defines relationships between clusters in a dendrogram, but it is based on local data features and may not give the best representation of the overall data structure.

We applied fuzzy c-means clustering to yeast cell cycle data. Although we did not filter the data for background noise, fuzzy clustering was able to assign genes to known periodic and aperiodic clusters. These clusters display an internal structure: Expression vectors of a large membership value were generally found near the centre of the clusters. Expression vectors with a large noise level received a small membership value and contributed less to the determination of the cluster centres. Noise can therefore be reduced *a posteriori* by determining the α-cores of clusters to exclude 'noisy' genes.

Noise robust clustering methods such as FCM are especially desirable, if changes of expression are small or restricted to a subset of genes. For example, we applied fuzzy clustering to study the genome-wide response to the induced expression of a single gene (Dunbier *et al.* 2001). Although changes in expression were generally small, fuzzy clustering was able to detect cores of subtle clusters while neglecting most genes which contributed noise. Subtle data structures can also be expected for experiments with limited amount of target RNA, which is often the case for studies of human tissue samples. These data frequently contain a high background noise.

Many cluster analyses are based on a specific setting for clustering parameters (section 5.2). Here we followed an more flexible approach: After determining the range of appropriate parameter values, we explored the dependency of the results on the parameter settings. Different settings of FCM parameters c and m can reveal different features of the data structure. The analysis showed that variation of c leads to clustering on different resolution levels. If a small parameter c is chosen, clusterings indicate the global data structure i.e. only the main clusters are detected. Larger c results in the detection of local structures, but it may also produce artifacts such as empty clusters.

Variation of the fuzzification parameter m offered easy identification of stable clusters. Alternatively, resampling techniques can be used (Levine & Domany 2001, Kerr & Churchill 2001). These approaches are, however, computationally intensive. For example, Kerr and Churchill reported that their bootstrapping analysis of yeast sporulation data required several hours on a conventional lab computer. Although this procedure might deliver valuable information about the clustering reliability, it is prohibitive for many current research procedures. Biologists frequently require interactive tools for data analysis. Different parameter settings and pre-processing methods may need to be tested to discover new biological knowledge. Such procedures are facilitated by computationally inexpensive analyses of cluster stability based on the variation of fuzzy parameter m.

For the cluster analysis presented here, we used Euclidean distance for assessing the similarity between genes. The expression values derived from differ-

ent measurements are treated equally. However, genes may be correlated only for a subset of measurements. For this case, the use of more complex distance measures is preferable. One possibility is the replacement of the Euclidean by the Mahalanobis distance derived from the full covariance matrix (equation 7.2). Such approaches are intended to detect ellipsoid clusters (Gustafson & Kessel 1979). This may be favourable for gene expression analysis as it indicates the correlation of genes within a cluster for each measurement. High correlation corresponds to a small cluster diameter for the measurement. Low correlation leads to large diameters. This correspondence may facilitate the identification of experimental conditions for which genes are coregulated and, thus, the identification of underlying regulatory processes.

One major challenge in cluster analysis is determining the number of cluster in data sets. In this study, we explored the internal cluster structures to select the number of clusters. Alternatively, objective functions can be used. These functions should reach an optimum if the correct number of clusters is chosen. We applied this approach to a smaller version of the gene expression data set analysed here (Futschik & Kasabov 2002). Several objective functions were compared using original and model-based data. The results can be found in appendix B.2. Pre-specification of the numbers of clusters can be avoided altogether if evolving clustering methods are used. Examples of such methods are evolving self-organising maps (ESOMs). During iterative clustering, new cluster centres are created and connected to neighbouring centres. ESOMs were successfully applied to identify known and putative novel regulatory sequences in the yeast genome (Futschik *et al.* 2000).

The clustering methods used to date have been restricted to a one-to-one mapping: one gene belongs to exactly one cluster. While this principle seems reasonable in many fields of cluster analysis, it might be too limited for the study of microarray data. Genes can participate in different genetic networks and are frequently controlled by a variety of regulatory mechanisms. For the analysis of microarray data, we can therefore expect that single genes will belong to several clusters. This can be accommodated by fuzzy clustering, since it can assign a gene to several clusters. A difficulty is, however, the distinction between genes which belong to several clusters and 'noisy' genes which are generally distributed between several clusters by fuzzy clustering. For this distinction, a filtering step *a priori* to the clustering might be necessary, although it would lead to a loss of information.

A major challenge in the future is the analysis of complex data sets. Sequence information and function annotations of genes might be included in the analysis. Simultaneous clustering of genes based on these different types of information may reveal new relationships between genes. One difficulty to overcome is the definition of a similarity measure. Sequence information or categorical data such as functional annotations have to be treated differently to continuous gene expression data. Additionally, information about the samples may be included in the analysis of gene expression data. An example of such incorporation is supervised classification of tissue samples, which is presented in the following chapters.