Validation in the Cluster Analysis of Gene Expression Data

Jens Jäkel, Martin Nöllenburg

Forschungszentrum Karlsruhe GmbH, Institut für Angewandte Informatik
Postfach 3640, D-76021 Karlsruhe
Tel.: (07247) 825726, Fax: (07247) 825786, E-Mail: jens.jaekel@iai.fzk.de

1 Introduction

In the early 1990s, a new technology called *DNA microarrays* was developed that allows for a simultaneous measurement of the expression levels, i. e. activity levels, of several thousand genes at specific conditions. During the last years many genome-scale experiments for various species and conditions were conducted. Most of the resulting data is publicly available in internet databases.

There are several questions that can be tackled by analysing data from microarray experiments. In the research field of functional genomics one goal is to annotate genes with their respective functions. This can be approached by gene expression analysis using the hypothesis that co-expressed genes are often co-regulated and thus might share a common function or are at least involved in the same biological process. So identifying groups of co-expressed genes over a number of different experiments can give hints for the putative functions of genes whose function were previously unknown by looking at the known functions of genes in the same group. Another example is the inference of regulatory networks and the identification of regulatory motifs in the sequences from groups of co-expressed genes.

For all these questions it is necessary to identify genes with similar expression patterns, which is usually done using different clustering techniques known from the area of unsupervised machine learning.

After the publication of the first large scale cluster analysis by Eisen et al. [8] many different approaches for the clustering of gene expression data have been made and proven to be successful in their respective situations. Nevertheless, no single clustering algorithm, similarity measure or validation criterion has yet become accepted as being the optimal choice for clustering genes based on microarray data. Former general results on clustering in literature cannot be taken over in any case because many of the theoretical studies assume well separable data. This is not the case for microarray data since clusters often overlap and cannot be easily identified. Some of the problems in clustering gene expression data are discussed in [4]. Only few works systematically evaluate and compare different clustering methods and results. In [7] six clustering algorithms are compared, but the choice of the number of clusters or the dissimilarity measure is not addressed. [3] discusses three validation indices and evaluates them on Kohonen's Self Organizing Maps algorithm. [9] presents a framework for validation of clusterings using external biological information. In [22] an overview of several clustering algorithms and dissimilarities for microarray data is given.

In this article we show on an example how to select good clusterings step-by-step based on several validation criteria. This includes choice of the algorithm, the dissimilarity measure and the number of clusters. In Section 2 some basic background on microarray technology and the pre-processing of microarray data is given. Next, in Section 3 different dissimilarity measures for gene expression profiles are introduced followed by the description of two popular clustering algorithms and the discussion of several validation techniques. Section 4 finally gives the results of applying these clustering algorithms on a sample data set. The resulting clusterings are systematically compared and several candidates are selected for further visual and external validation using biological information about the clustered genes. Conclusions are presented in Section 5.

2 Measuring gene expression with microarrays

Gene expression is the process by which a gene's information is converted into a functional protein of a cell. It involves two main steps according to the central dogma of molecular biology: The section of the DNA corresponding to the gene is first transcribed into a single-stranded complementary messenger RNA (mRNA) molecule. Thereafter the mRNA is translated into a protein.

It is widely believed that regulation of gene expression is largely controlled at the transcriptional level [29], so studying the abundance of the mRNA can give insight into how the corresponding genes control the function of the cell. Microarrays are a new techniques to measure the mRNA abundance for several thousand genes in parallel.

Microarrays are based on the fact that two complementary DNA (or RNA) molecules can hybridize, i. e. they can bind together. When the sequence of the genome (entirety of all genes) of an organism is known, it is possible to synthesize millions of copies of DNA fragments of each gene. These are then able to hybridize with the corresponding complementary molecules.

Microarrays are small glass slides with thousands of spots printed on it in a grid-like fashion. Each spot corresponds to one gene and consists of thousands of identical and gene-specific single-stranded DNA sequences fixed to the glass surface. The mRNA abundance in a sample is measured indirectly by reverse transcribing the mRNA into cDNA[1]. Then the cDNA molecules are able to stick to the spots on the microarray that correspond to their respective genes. Although measuring transcript abundance is not exactly the same as measuring gene expression, it is very common to use both terms synonymously.

Usually the abundance of mRNA transcripts is measured relatively to a control sample. To this end the cDNA prepared from the mRNA of the experiment sample is labeled with a fluorescent dye (usually red Cy5) and the control cDNA is labeled with green-fluorescent dye (Cy3) or vice versa. When both samples are mixed at equal amounts and washed over the glass slide the target cDNA will hybridize on the spot with its complementary sequences (called probes).

Each dye can emit light at a specific wave length and thus, using a laser scanner, the intensity of fluorescence is measured for both dyes.

The construction of a microarray, sample preparation and scanning of the slides is illustrated schematically in Figure 1.

The resulting images can be overlayed and show whether a gene is over- or underexpressed relative to the control sample. Further, from these images an intensity value for each spot and both color channels, denoted by R_g and G_g, $g = 1, \ldots, n$, n the number of genes, can be extracted. Their log-ratio $M_g := \log_2(R_g/G_g)$ is related directly to the *fold change*, a common measure of differential expression. In case $R_g \geq G_g$ the fold change is simply R_g/G_g and otherwise the fold change is defined as $-G_g/R_g$. In that sense a fold change of 2 means that the corresponding gene is overexpressed by a factor 2 and a fold change of -2 means it is underexpressed by a factor 2.

In order to identify differentially expressed genes it is necessary to get rid off the systematic error which is present in microarray data. A well-known source of error is e. g. the different labelling efficiency of Cy3 and Cy5.

[1] Complementary DNA (abbreviated cDNA) denotes single-stranded DNA molecules that are complementary to their mRNA templates. cDNA is assembled by the enzyme reverse transcriptase.

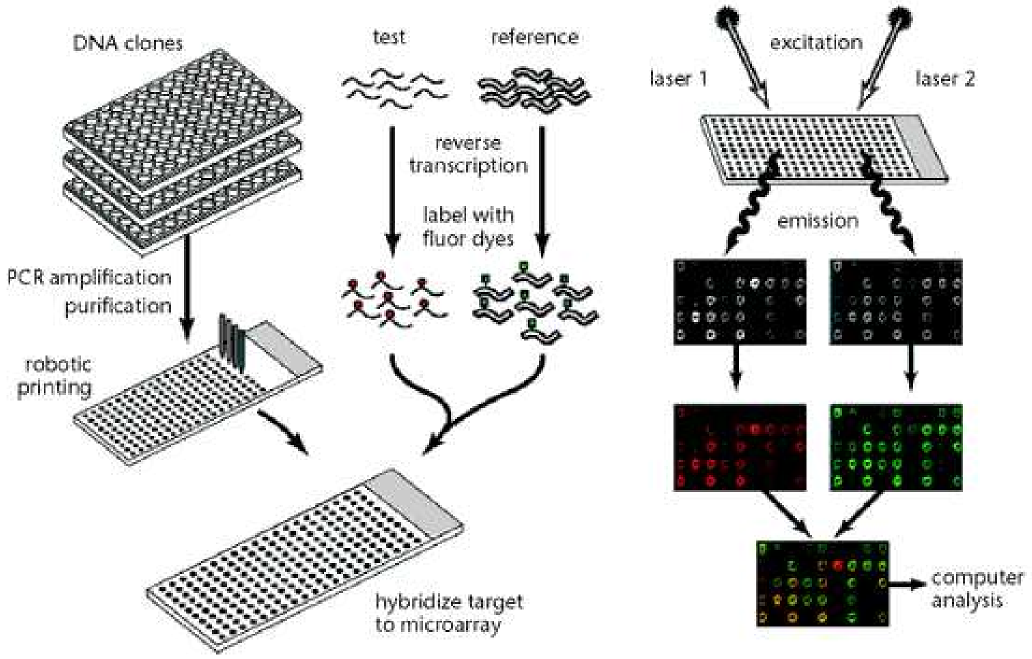


Figure 1: Schematic overview of a microarray experiment. Figure taken from [1].

One method to minimize the systematic variation in the data is by applying a global intensity-dependent normalization [27, 28] using local regression (e.g. performed by `loess` [18, 6]):

$$M_g^{norm} = M_g - c(A_g), \quad A_g := \log_2 \sqrt{R_g G_g}.$$

This regression method performs a robust locally linear fit and is not affected by a small number of outliers due to differentially expressed genes. Hence, it is applicable when the majority of the genes is assumed to be not differentially expressed.

3 Background on cluster analysis

Cluster analysis is a data mining technique concerned with grouping a set of objects having certain attributes into subsets called *clusters*. The objective is to arrange the groups such that all objects within the same cluster are, based on their attribute values, similar to each other and objects assigned to different clusters are less similar. So clustering is about revealing a hidden structure in a given set of data, usually without any external knowledge about the objects.

Not always well-separated groups are present in a data set. In this case clustering is sometimes referred to as segmentation [10]. For such problems it is considerably more difficult to assess the results of cluster analysis and decide which is the correct number of clusters.

The following notation is used throughout the paper: Let $S := \{s_1, \ldots, s_n\}$ be a set of objects s_i. For each s_i there are p attributes $A_1, \ldots, A_p$ observable. Here we can assume all attributes to be real-valued. The object-attribute matrix $M := (m_{ij})$ contains the values m_{ij} of attribute A_j for object s_i, where $i = 1, \ldots, n$, $j = 1, \ldots, p$. A clustering $\mathcal{C}$ of size k of the set S is then a partition of S into pairwise disjoint non-empty clusters $C_1, \ldots, C_k$.

3.1 Dissimilarity measures

There are many different dissimilarities for cluster analysis listed in the literature (see for example [13]). The choice of a dissimilarity measure is highly dependent on the data that is to be analysed. In the context of gene expression profiles we will only consider Euclidean distance and a dissimilarity based on Pearson's correlation coefficient.

3.1.1 Euclidean distance

Probably the most widely used dissimilarity measure for numeric attributes is the Euclidean distance. The objects s_i are considered as points in the p-dimensional space and the standard Euclidean metric is used:

$$d_{ij} := d(s_i, s_j) := \sqrt{\sum_{l=1}^{p} (m_{il} - m_{jl})^2}$$

In case of missing values all incomplete attribute pairs are discarded in the above sum and the sum is scaled by $\frac{p}{p_c}$ where p_c is the number of complete attribute pairs.

3.1.2 Pearson Correlation

Another common dissimilarity measure, especially for gene expression data, based on the Pearson correlation, is

$$d_{ij} := d(s_i, s_j) := 1 - \left| \frac{\sum_{l=1}^{p} (m_{il} - \overline{m}_i)(m_{jl} - \overline{m}_j)}{\sqrt{\sum_{l=1}^{p} (m_{il} - \overline{m}_i)^2 \sum_{l=1}^{p} (m_{jl} - \overline{m}_j)^2}} \right|,$$

where $\overline{m}_i := \frac{m_{i1} + \cdots + m_{ip}}{p}$ is the arithmetic mean of s_i's attributes and $\overline{m}_j$ for s_j respectively. Note that this measure treats positively and negatively correlated objects equally.

Attribute pairs with at least one missing value are discarded from the calculation of the correlation coefficient.

3.2 Characteristics of clusterings

Given a clustering $\mathcal{C} = (C_1, \ldots, C_k)$ of S and the underlying dissimilarity measure d, two characteristic values can be defined as in [10]:

$$W(\mathcal{C}) := \frac{1}{2} \sum_{l=1}^{k} \sum_{i,j \in C_l} d(s_i, s_j),$$

the so-called (total) *within cluster point scatter* and the (total) *between cluster point scatter*

$$B(\mathcal{C}) := \frac{1}{2} \sum_{l=1}^{k} \sum_{\substack{i \in C_l \\ j \notin C_l}} d(s_i, s_j).$$

$W(\mathcal{C})$ characterizes the internal cohesion as it measures the pairwise dissimilarities within each cluster, whereas $B(\mathcal{C})$ characterizes external isolation of clusters.

They both are related through $T = W(\mathcal{C}) + B(\mathcal{C})$. T, which is nothing else but the sum of all pairwise dissimilarities, is called *total point scatter* and is constant for all clusterings of S given d. Because the natural aims of clustering are to produce well-isolated and internally similar clusters, this task can be seen as minimizing $W(\mathcal{C})$ or maximizing $B(\mathcal{C})$.

Therefore W and B will play a role in assessing the quality of different clusterings in Section 3.4.

3.3 Clustering methods

There is a large number of clustering algorithms known in pattern recognition and data mining. For this work we restricted ourselves to two widely used combinatorial algorithms. Combinatorial is used here in the sense of [10], i. e. each observation s_i is uniquely assigned to a cluster C_j based solely on the data without making any assumption about an underlying probabilistic model.

Combinatorial clustering algorithms are often divided into partitioning and hierarchical methods [14]. The former construct k clusters for a given parameter k whereas the latter construct a hierarchy covering all possible values for k at the same time. In each group of methods we picked one algorithm, namely partitioning around medoids (PAM) and agglomerative clustering using Ward's method.

3.3.1 Partitioning around medoids

Partitioning methods cluster the data into k groups, where k is a user-specified parameter. It is important to know that a partitioning method will find k groups in the data for any k provided as a parameter, regardless of whether there is a "natural" clustering with k clusters or not. This leads to different criteria for choosing an optimal k. Some are discussed in Section 3.4.

Partitioning around medoids (PAM), also called k-medoids clustering[2], is an algorithm described by Kaufman and Rousseeuw [14] and is implemented in the R package `cluster` [18]. Its objective, seen as an optimization problem, is to minimize the within cluster point scatter $W(\mathcal{C})$. The resulting clustering of S is usually only a local minimum of $W(\mathcal{C})$.

The idea of PAM is to select k representative objects, or medoids, among S and assign the remaining objects to the group identified by the nearest medoid. Initially, in the medoids can be chosen arbitrarily, although the R implementation of PAM distributes them in a way that S is well covered. Then, all objects $s \in S$ are assigned to the nearest medoid. In an iterative loop as a first step a new medoid is determined for each cluster by finding the object with minimum total dissimilarity to all other cluster elements. Next, all $s \in S$ are reassigned to their clusters according to the new set of medoids. This loop repeats until no more changes of the clustering appear.

Because the R implementation of PAM assigns the initial clustering deterministically the results of PAM will always be identical and repeated runs to cope with random effects are not necessary.

3.3.2 Agglomerative clustering

Agglomerative methods are very popular in microarray data analysis. For example the first genome-wide microarray clustering study [8] used agglomerative hierarchical clustering.

Hierarchical clustering methods do not partition the set S into a fixed number k of clusters but construct a tree-like hierarchy that encodes implicitly all possible values of k. At each level $j \in \{1, \dots, n\}$ there are j clusters encoded. The lowest level consists of the n singleton clusters and at level one there is just one cluster containing all objects. However, hierarchical clustering imposes a nested tree-like cluster structure on the data regardless of whether the data really have this property. Therefore one has to be careful when drawing conclusions from hierarchical clustering.

[2] There is a close relationship to the popular k-means clustering algorithm. The advantage of k-medoids is that it can be used with any dissimilarity and not only with Euclidean distance.

In agglomerative clustering, at each level j the two "closest" clusters are merged to form level $j-1$ with one less cluster. Agglomerative methods vary only in terms of the dissimilarity measure between clusters. In any case the dissimilarity $d : S \times S \to \mathbb{R}_{\geq 0}$ must be extended to $D : \mathcal{P}(S) \times \mathcal{P}(S) \to \mathbb{R}_{\geq 0}$, such that $D(\{s\},\{t\}) = d(s,t)$ for $s,t \in S$. Note that the dissimilarity matrix for D would be of exponential size, but only few values are actually needed. Therefore, in practice the values of D will be computed on-demand. The general agglomerative hierarchical clustering algorithm proceeds as follows.

Initially, the set A is the trivial partition of S into singleton sets. Then in an iteration from $j = n$ down to 1 the current partition A is assigned to the clustering $\mathcal{C}^{(j)}$. Further, the two clusters with the smallest dissimilarity are determined, merged and replaced in A by their union.

There are several dissimilarity measure between clusters, e. g.

- single linkage, leading mainly to large, elongated clusters,
- complete linkage, yielding rather compact clusters,
- average linkage, being a compromise between these two extremes, or
- the dissimilarity measure used by Ward's method [25].

The latter method uses the increment in the within cluster point scatter, which would result from merging two clusters, as the dissimilarity between them. When defining the within cluster point scatter

$$W(G) := \frac{1}{2} \sum_{i,j \in G} d(s_i, s_j)$$

for a single cluster G analogously to the total within cluster point scatter (see Section 3.2), the dissimilarity is given as

$$D(G,H) = \frac{2}{n_G + n_H} W(G \cup H) - \frac{2}{n_G} W(G) - \frac{2}{n_H} W(H).$$

Since the within cluster point scatter is a measure for the internal cohesion of clusters, Ward's method tends to create compact clusters with very similar objects. In the case of squared Euclidean distance this method is also known as incremental sum of squares. Merging clusters that minimize D is equivalent to minimizing the within cluster variance.

A problem with Ward's method is that it minimizes the objective function locally so that decisions taken at lower levels of the hierarchy do not necessarily mean optimality at higher levels. The other agglomerative methods suffer from this fact as well. After two clusters have been merged on a certain level there is no way of reversing this decision at a later step of the algorithm although it might be favorable. Especially when rather few clusters are sought this might be disadvantageous because many previous merging decisions are influencing the shape of higher level clusters.

When comparing the results of different agglomerative methods using the function `hclust` from the R package `mva` for our gene expression data, Ward's method is performing best. Several comparative studies, mentioned in [12], also suggest that Ward outperforms other hierarchical clustering methods. Thus for our experiments in Section 4 we chose to compare the PAM algorithm and agglomerative clustering using Ward's method.

3.4 Cluster validation

In most applications of clustering techniques it is impossible to speak of *the* correct clustering and therefore it is necessary to use some validation criteria to assess the quality of the results of cluster analysis. These criteria may then be used to compare the adequacy of certain algorithms and dissimilarity measures or to choose the best number k of clusters. This is especially important when the correct number of clusters is unknown a-priori as it is the case in this study. When using PAM, the algorithm is run with different values for the parameter k and when using agglomerative clustering, this refers to finding the optimal level in the hierarchy.

Following [12], validation measures are grouped into *internal*, *relative* and *external criteria*. Internal criteria assess the quality of a given clustering based solely on the data themselves or on the dissimilarity used. Four internal criteria are introduced in Section 3.4.1.

Relative criteria are used to directly compare the agreement between two clusterings, for example to examine how similar two k-clusterings resulting from different algorithms or dissimilarities are. Section 3.4.2 describes two relative criteria.

External criteria are measuring the quality of a clustering by bringing in some kind of external information such as a-priori class labels when available. Here we denote criteria based on visualization of the cluster data as external criteria too. Visualization of clusters can help the user to assess the adequacy of a given clustering. Despite this is less objective than an internal criterion, visualization is often preferred in the final evaluation of the clustering results. Further, gene clusters based on expression profiles can be evaluated by looking at the functional annotations for their constituent genes. If certain properties are shared by many genes this might be a sign for a “good” cluster. Because external criteria are more domain specific they are described in Section 4 when the data and experiments are covered.

3.4.1 Internal validation

Silhouettes Rousseeuw [20] proposed the *silhouette* statistic which assigns to each object a value describing how well it fits into its cluster. Let $a(s_i)$ denote the average dissimilarity of s_i to all points in its own cluster and let $b(s_i)$ denote the minimum of all average dissimilarities to the other clusters, i. e. the average dissimilarity to the second best cluster for s_i. Then

$$sil(s_i) = \frac{b(s_i) - a(s_i)}{max(a(s_i), b(s_i))}$$

is the silhouette value for s_i. Object s_i matches its cluster well if $sil(s_i)$ is close to one and poorly matches it if $sil(s_i)$ is close to zero or even negative. Negative values only occur when an object is not assigned to the best fitting cluster.

A natural measure for the quality of the whole clustering is

$$sil(C) = \frac{1}{n} \sum_{s_i \in S} sil(s_i), \tag{1}$$

the average silhouette for all objects in S. According to this criterion choose the number of clusters $\hat{k}$ as the value maximizing the average silhouette.

Measure of Calinski and Harabasz In [17] the authors compare 28 validation criteria and found that in their experiments the measure by Calinski and Harabasz [5] performed best. It assesses the quality of a clustering with k clusters via the index

$$CH(k) = \frac{BSS(k)/(k-1)}{WSS(k)/(n-k)} \tag{2}$$

where $WSS(k)$ and $BSS(k)$ are the within and between cluster sums of squares defined analogously to W and B in Section 3.2 but using squared dissimilarities. The optimal value $\hat{k}$ for the number of clusters is again the value k maximizing the criterion.

The idea is to choose clusterings with well isolated and coherent clusters but at the same time keeping the number of clusters as small as possible.

Originally this index was meant for squared Euclidean distance. Because our optimization criterion is not based on squared dissimilarities we used W and B instead of WSS and BSS in the definition of CH. This still follows the same motivation as using squared dissimilarities and is more robust against outliers.

Measure of Krzanowski and Lai Krzanowski and Lai [15] defined an index based on the decrease of the within cluster sums of squares. First, they defined

$$DIFF(k) = (k-1)^{2/p}\, WSS(k-1) - k^{2/p}\, WSS(k)$$

and then the index

$$KL(k) = \left| \frac{DIFF(k)}{DIFF(k+1)} \right| \tag{3}$$

which should be maximized again.

Let g be the correct number of groups in the data. Then the idea of KL is based on the assumption that $WSS(k)$ decreases rapidly for $k \leq g$ and it decreases only slightly for $k > g$. This is justified by the hypothesis that for $k \leq g$ at every successive step large clusters that do not belong together are separated resulting in a strong decrease of WSS. Conversely for $k > g$ good clusters are split resulting in a very small decrease of the within cluster sum of squares. Thus one can expect that $DIFF(k)$ is small for all k but $k = g$ and consequently $KL(k)$ is largest for the optimal k.

The same intuition holds when replacing WSS by W as in the case of Calinski and Harabasz' measure. Thus we used W in the definition of $DIFF$ for our experiments because again the measure was originally designed for use with squared Euclidean distance.

Prediction strength A more recent approach to assessing the number of clusters is the measure of *prediction strength* proposed by Tibshirani et al. [24]. It uses cross-validation of the clustering process and determines how well the clusters formed in the training set agree with the clusters in the test set. More precisely, the set S is divided into a test set S_{te} and a training set S_{tr}. Then both sets are clustered individually into k clusters yielding two clusterings $\mathcal{C}_{te}$ and $\mathcal{C}_{tr}$. For each cluster a suitable representative element is chosen, e.g. the cluster medoid. Finally the elements in S_{te} are assigned to the training set clusters by minimizing the dissimilarity to the representative elements of the clusters in $\mathcal{C}_{tr}$.

Then for each pair of elements belonging to the same cluster in $\mathcal{C}_{te}$ it is checked whether they fall into the same cluster again when using the training clusters. If this is the case for most pairs the prediction strength should be high otherwise it should be lower. So formally the prediction strength is defined as

$$ps(k) = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{n_{C_j}(n_{C_j}-1)} \sum_{\substack{i,i' \in C_j \\ i \neq i'}} D[\mathcal{C}_{tr}, S_{te}]_{ii'}, \tag{4}$$

where $D[\mathcal{C}_{tr}, S_{te}]$ is a matrix with the (i,i')th entry equal to one if the elements s_i and $s_{i'}$ fall into the same cluster when assigned to the training clusters as described above and zero otherwise. Thus, $ps(k)$ is the average proportion of test pairs correctly classified by the training clusters.

[24] shows that 2-fold cross-validation has no disadvantages compared to higher order cross-validation. Therefore we used 2-fold cross-validation to measure the quality of a clustering given the parameter k.

3.4.2 Relative validation

Rand index The Rand index [19] for comparing two clusterings $\mathcal{C}$ and $\mathcal{C}'$ of the same data S is based on four counts of the pairs (s_i, s_j) of objects in S.

N_{11} number of pairs that are in the same cluster both in $\mathcal{C}$ and in $\mathcal{C}'$
N_{00} number of pairs that are in different clusters both in $\mathcal{C}$ and in $\mathcal{C}'$
N_{10} number of pairs that are in the same cluster in in $\mathcal{C}$ but not in $\mathcal{C}'$
N_{01} number of pairs that are in the same cluster in in $\mathcal{C}'$ but not in $\mathcal{C}$

The Rand index is defined as the relative proportion of identically classified pairs

$$R(\mathcal{C},\mathcal{C}') = \frac{N_{11}+N_{00}}{N_{11}+N_{00}+N_{10}+N_{01}}$$

and lies between zero and one.

It has the disadvantage that its expected value is not equal to zero for two random partitions. Therefore [11] introduced the adjusted Rand index

$$R'(\mathcal{C},\mathcal{C}') = \frac{R(\mathcal{C},\mathcal{C}') - E(R(\mathcal{C},\mathcal{C}'))}{1 - E(R(\mathcal{C},\mathcal{C}'))} \tag{5}$$

as a normalized form of Rand's criterion. It is explicitly given in [12]. The maximum value is still one and reached only if the two clusterings are identical.

Variation of information A recent work by [16] introduces an information theoretical criterion called *variation of information*. It is defined as

$$VI(\mathcal{C},\mathcal{C}') = H(\mathcal{C}) + H(\mathcal{C}') - 2I(\mathcal{C},\mathcal{C}'), \tag{6}$$

where H is the entropy of a clustering and I is the mutual information of two clusterings. The probabilities needed for these terms are estimated by the relative cluster sizes. [16] shows that VI is a metric on the space of all clusterings and gives upper bounds for its value. $VI(\mathcal{C},\mathcal{C}')$ can be seen as a measure for the uncertainty about the cluster in $\mathcal{C}$ of an element s knowing its cluster in $\mathcal{C}'$.

4 Experiments and results

4.1 Yeast cell-cycle data

A very popular data source for comparative gene expression studies is the cell cycle data set for the yeast *Saccharomyces cerevisiae* published by Spellman et al. [21]. *S. cerevisiae* is the most studied eukaryotic model organism and has about 6300 potential genes. It has the advantage that many details about its individual genes are available helping to verify the results of a cluster analysis using this external information.

Spellman et al. conducted three different time series microarray experiments. We selected the *cdc15* series consisting of 24 measurements of the expression levels for 6283 open reading frames[3] (ORFs). The yeast strain was grown and then arrested at a certain

[3] An open reading frame is a DNA sequence that has the potential to encode a protein or polypeptide. It does not necessarily correspond to a gene.

state of the cell cycle by incubating it at 37°C. The synchronized sample was then released from the arrest and measurements were taken every 10 or 20 minutes over a total period of 300 minutes corresponding to about three cell cycles. Part of the same original culture was grown unsynchronized and served as the control sample in the microarray experiments.

The raw data for our analyses were retrieved from the Stanford Microarray Database[4] and consist of absolute intensity values for both color channels as well as a quality flag for each gene on the array. Then the data were transformed into log-ratios and normalized as described in Section 2.

A subset of 238 genes was preselected for further processing based on several criteria. First, all genes having three or more missing values were excluded. Next, genes showing not enough differential expression over time were discarded. We chose to set a threshold of at least two time points showing an absolute fold change[5] of two or higher. Finally, we only considered genes that exhibit a periodic behavior and thus are likely to be cell cycle dependent. This was done using a statistical method introduced in [26] and implemented in the R package `GeneTS`.

4.2 Clustering and internal validation

The general procedure is to compute the pairwise Euclidean distances (see Section 3.1.1) and the correlation based dissimilarities (see Section 3.1.2). Note that the Euclidean distances are computed from standardized profiles[6]. Then we apply PAM clustering and agglomerative clustering using Ward's method.

When clustering gene expression data, in most cases the number of clusters k is unknown in advance. Thus determining the number of clusters is a very important task. Evaluating the internal validation measures discussed in Section 3.4.1 for a range of possible k-values we can choose a couple of good candidate clusterings. By visualizing the clusterings, the homogeneity and separation of the clusters can be assessed. Differences between the candidate clusterings are quantified using the relative validation measures of Section 3.4.2. Only the best clusterings from these candidates are kept and evaluated further with external methods, which is discussed in Section 4.3.

The subset to be clustered contained 238 genes as given in Section 4.1. Therefore we chose to compare clusterings with 2 to 40 clusters. Further increasing the number of clusters would only result in a growing number of singletons clusters which is not desirable.

We used both Euclidean distance and the correlation based dissimilarity measure. The main difference between the two is that two expression profiles that are strongly negatively correlated have a high Euclidean distance but a very low correlation based dissimilarity. Hence, the genes will almost never be in the same cluster using Euclidean distance and they are very likely to end up in the same cluster using the correlation dissimilarity. The biological justification for using the correlation based measure is that genes that regulate biological processes can either be activating or repressing. So for the initiation of such a process the activator genes must be highly expressed whereas the expression of the repressor genes must be scaled back. Nevertheless it makes sense to put both groups of genes into the same cluster because they launch the same process.

Figure 2 shows the internal validation measures applied to the four different types of clusterings. The general tendency is that prediction strength and silhouette propose rather

[4] `http://genome-www5.stanford.edu`

[5] See definition in Section 2.

[6] Standardization means that for each profile the profile's mean value is subtracted and then it is divided by its standard deviation. Thus the standardized profiles have zero mean and unity standard deviation.

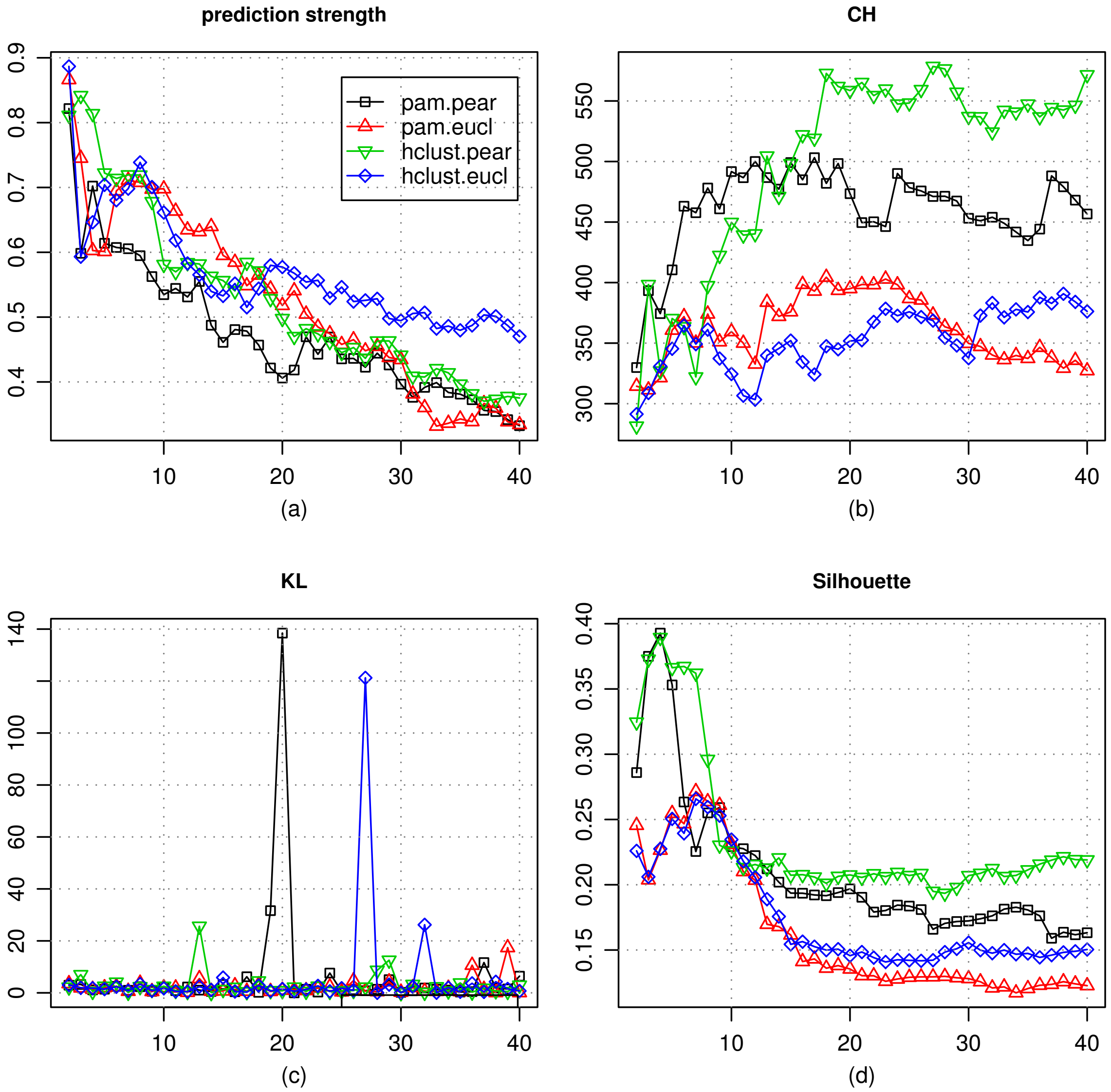


Figure 2: Plots of the internal validation measures for clusterings with $k = 2, \ldots, 40$ clusters. Each plot shows the respective values for one of four validation measures applied to PAM and hierarchical clustering (hclust) using both Euclidean and Pearson correlation based dissimilarities. (a) prediction strength, (b) measure of Calinski and Harabasz, (c) measure of Krzanowski and Lai and (d) silhouette.

small cluster numbers whereas the CH criterion prefers clusterings with more clusters. The plot (c) shows strong peaks at certain positions and has a much smaller range for the rest of the values. When looking closer at how these peaks originate from eq. (3) one can see that they don't always reflect a good clustering. For example the peak at $k = 20$ with a KL value of about 140 arises from dividing $DIFF(20) = 1.636$ by $DIFF(21) = 0.012$. This is of course a *relatively* high decrease in the $DIFF$-values. But in comparison to $DIFF$-values in the order of 10^4 for the first $k = 2,3,4,5$ considered here, these small $DIFF$-values for larger k do not mean any true improvement in the clustering at all. This suggests that the KL index is not very useful for poorly separated data such as gene expression profiles.

For agglomerative clustering the dendrograms also give hints for good clusterings. In Figure 3 the dendrograms for both dissimilarities are shown. In plot (a) up to five clusters can be easily distinguished and about 13 clusters still have a reasonable dissimilarity when

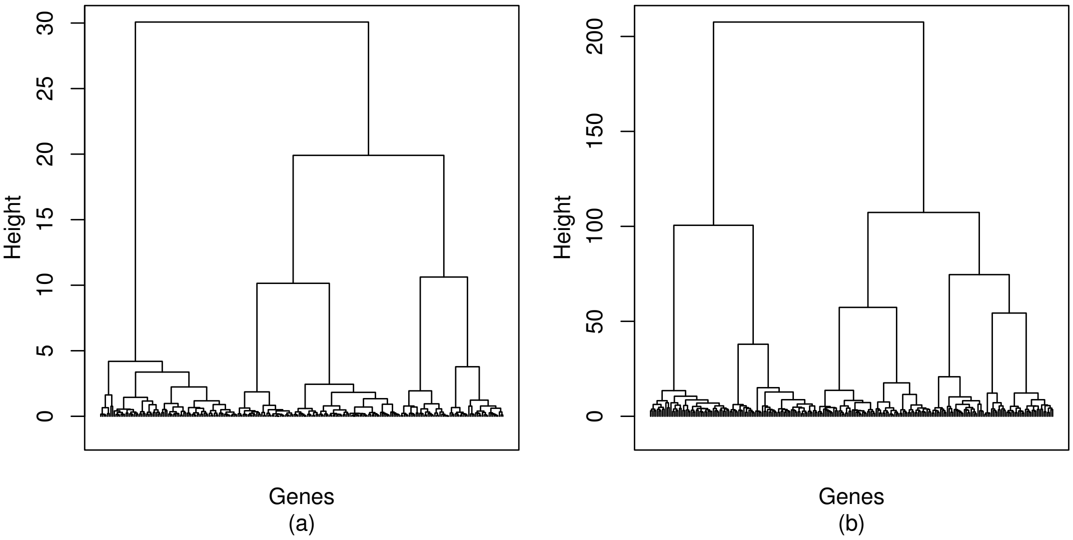


Figure 3: Dendrograms from agglomerative clustering using Ward's method and (a) correlation based dissimilarities or (b) Euclidean distances.

they are merged, suggesting that they really represent different groups of genes. Plot (b), showing the dendrogram for Euclidean dissimilarity, indicates slightly more clusters than plot (a) when using similar arguments. This follows the motivation for the correlation based dissimilarity given at the beginning of this section: One can expect that a cluster combining positively and negatively correlated expression profiles in (a) corresponds to at least two clusters in (b).

Taking into account all four validation measures and their local behavior we identified for each combination of algorithm and dissimilarity a set of candidate values for further investigation. The selected values are given in Table 1. As an example $k = 8$ for agglomerative clustering with Euclidean dissimilarity was chosen because in Figure 2(a) the prediction strength at $k = 8$ forms a local maximum with a good value of 0.73. Further *CH* in part (b) of the figure has a local maximum with a high value at $k = 8$ and the silhouette in Figure 2(d) is still in a stable range before falling for $k \geq 10$. For these candidate values we displayed the clusterings as *Eisenplots* and cluster profile plots. Here only two examples are given in figures 4 and 5.

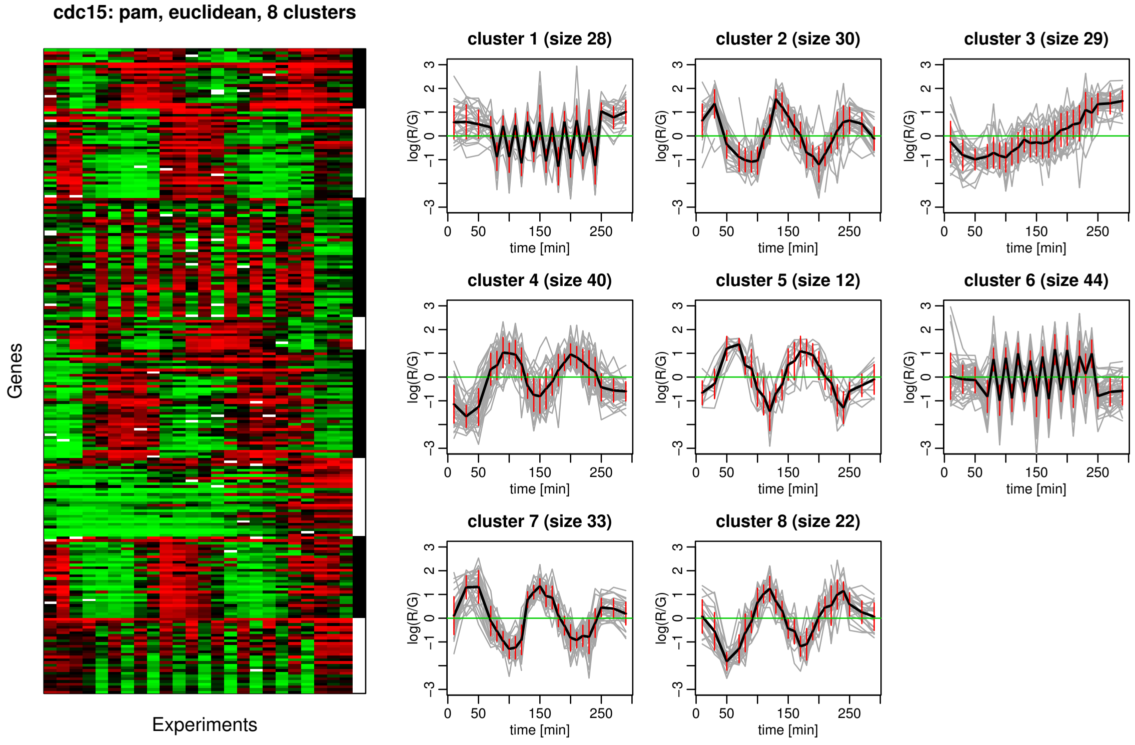
algorithm	dissimilarity	candidate values for k
PAM	correlation based	**6**, 8, **10**, 12, 15, 17, 20
PAM	Euclidean	6, **8**, **10**, 13, 18
agglomerative	correlation based	5, **6**, **10**, 13
agglomerative	Euclidean	6, **8**, **15**

Table 1: Candidate values for the number of clusters. The bold values are selected as "good" by visual examination.

Eisenplots (see Figure 4(a)) are named after M. B. Eisen who introduced this type of display in [8]. The expression data contained in the matrix M (see Section 2) are plotted as a table where row i and column j encodes the expression value for gene g_i at time point t_j by a color similar to the original color of its spot on the microarray. This means that high expression is coded as red and low expression as green. If the value is zero, meaning no differential expression, it is displayed in black. Note that in case of correlation based dissimilarity, profiles negatively correlated in respect to the medoid profile are multiplied by -1 because otherwise the plots would become messy. The rows are ordered such that the elements of each cluster appear next to each other. To help identifying the cluster boundaries we included an additional column on the right hand side of the plot where alternating black and white bars mark the individual clusters.

The cluster profile plots show for each cluster the profile of the cluster medoid together with a vertical bar giving the standard deviation within the cluster at each time point. Further, in light grey all profiles of genes in the cluster are plotted and the size of the cluster is given. Note that again in case of correlation based dissimilarity, profiles negatively correlated with respect to the medoid are inverted before calculating the standard deviations. However, these profiles are plotted without change in a very light grey in the background. This enables the viewer to distinguish between positively and negatively correlated profiles.

Cluster visualization allows the viewer to assess the adequacy of the selected clusterings. One can make statements about the compactness and the isolation of the given clusters by verifying how similar the plotted profiles within each cluster and how dissimilar the medoid profiles are. If multiple clusters have very similar medoids the clustering parameter *k* might be chosen too large. In contrast, if clusters show a large within cluster variance it might be better to increase *k* in order to split those clusters.



(a) (b)

Figure 4: Visualization of PAM clustering with Euclidean dissimilarity and eight clusters. In the Eisenplot in (a) the individual clusters are marked on the right side of the plot. Cluster 1 is on the bottom and Cluster 8 on the top of the plot. In (b) the corresponding cluster profile plots are shown.

It is impossible to discuss the visualization for all candidate clusterings, so we just give two illustrative examples. In Figure 4 the result of PAM clustering with Euclidean dissimilarity and eight clusters is displayed. Three different types of clusters can be identified in the Eisenplot in (a) and the cluster profiles in (b):

Clusters 1 and 6 group together profiles that oscillate with a period of 20 minutes. Note that the first four time points and the last three time points are 20 minutes apart

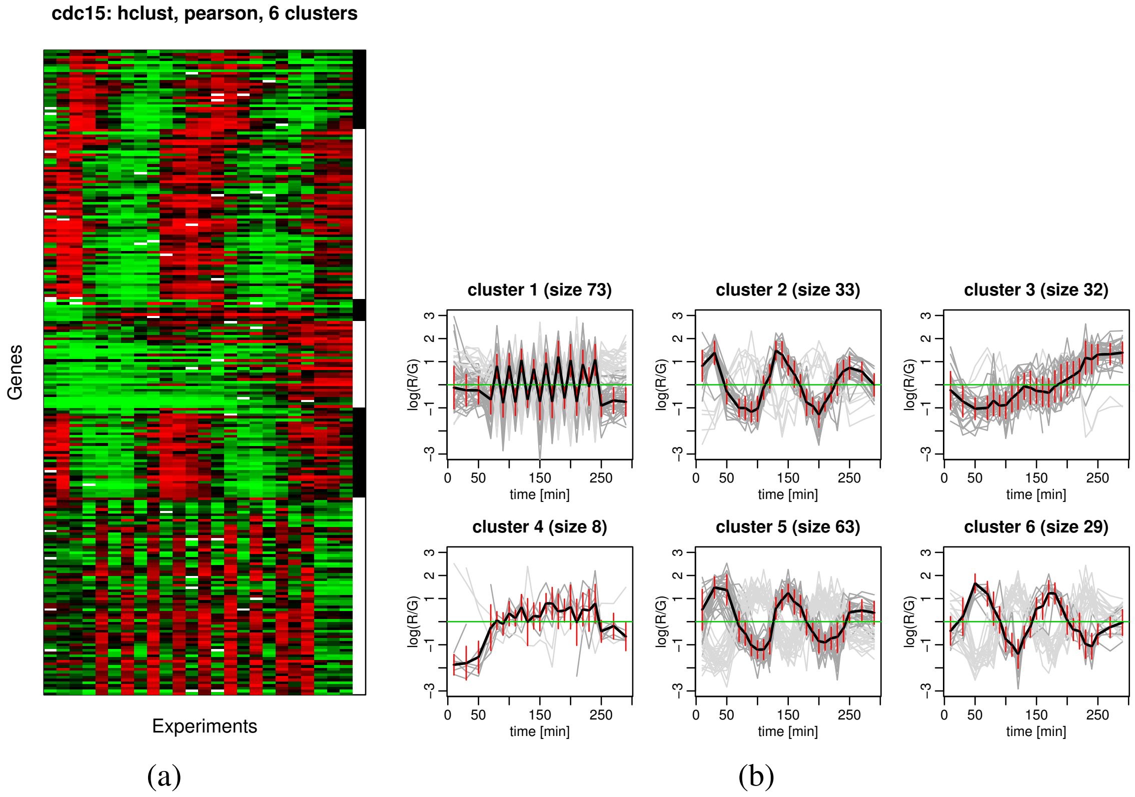


Figure 5: Visualization of agglomerative clustering with correlation based dissimilarity and 6 clusters. In the Eisenplot in (a) the individual clusters are marked on the right side of the plot. Cluster 1 is on the bottom and Cluster 6 on the top of the plot. In (b) the corresponding cluster profile plots are shown.

instead of 10 minutes for the rest of the time points. Thus the profiles look different in the beginning and in the end. The clusters are not merged because the profiles are strongly negatively correlated between both clusters resulting in large Euclidean distances.

Cluster 3 is distinct from the others by grouping profiles that have a slowly increasing behavior over the full range of measurements. Possibly these profiles have been identified as periodic with a period longer than 300 minutes.

Finally clusters 2, 4, 5, 7 and 8 show a cyclic behavior matching the approximately 2.5 cell cycles studied in the experiment. They can be subdivided into clusters 4 and 8, which consist of profiles that start off with low expression. The peaks in Cluster 4 appear about 30 minutes before those in Cluster 8 and therefore it seems reasonable not to merge them. The second group consists of clusters 2, 5 and 7, which have profiles increasing first. Here the profiles in Cluster 2 take their maxima about 20 minutes earlier than those in Cluster 7. Cluster 7 in turn is left-shifted in comparison to Cluster 5 by about 20 minutes.

The second example is given in Figure 5. To cover both algorithms and dissimilarities this figure shows agglomerative clustering with correlation based dissimilarity and 6 clusters. In contrast to the previous example, profiles that can be transferred into each other by mirroring them on the x-axis are likely to end up in the same cluster because they are strongly negatively correlated and thus have a small dissimilarity now.

As expected, Cluster 1 represents all the “zigzag” profiles that were in different clusters beforehand. Cluster 3 contains the slowly increasing profiles as before whereas Cluster 4, the smallest one with only eight elements, cannot be found in the previous example.

Clusters 2, 5 and 6 group the cell cycle dependant profiles. They can again be distinguished by the positions of the peaks over time. But here the clusters group together both positively and negatively correlated genes, which is shown by the grey-shaded curves in the background of the plots in Figure 5(b).

The eight clusterings that we finally selected after visual examination are given in bold face in Table 1. The next step is to apply the adjusted Rand index and the variation of information criterion given in Section 3.4.2.

The goal of applying these relative validation measures is to identify the clustering that is most similar to all other clusterings. This clustering is a good candidate for further external validation discussed in the next section. Moreover the clustering being least similar to the others can be used for external validation as well. This second clustering should have discovered a structure in the data that is very different from the one discovered with the first clustering. Therefore it will be interesting to see how these two extremes perform in external validation.

clusterings			PAM				agglomerative			
			correlation		Euclidean		correlation		Euclidean	
			6	10	8	10	6	10	8	15
PAM	correlation	6	0.000	0.226	0.246	0.278	0.195	0.261	0.259	0.331
		10	0.226	0.000	0.308	0.292	0.228	0.249	0.303	0.307
	Euclidean	8	0.246	0.308	0.000	0.075	0.197	0.265	0.097	0.190
		10	0.278	0.292	0.075	0.000	0.216	0.257	0.115	0.174
aggl.	correlation	6	0.195	0.228	0.197	0.216	0.000	0.102	0.191	0.263
		10	0.261	0.249	0.265	0.257	0.102	0.000	0.263	0.280
	Euclidean	8	0.259	0.303	0.097	0.115	0.191	0.263	0.000	0.108
		15	0.331	0.307	0.190	0.174	0.263	0.280	0.108	0.000
cumulative			1.796	**1.913**	1.378	1.407	1.392	1.677	**1.336**	1.653

Table 2: Values of the variation of information criterion for all pairs of the eight selected clusterings. The smaller the value the more similar are the two clusterings with the minimum 0 reached only if two clusterings are equal. The bottom row gives the cumulative values over all columns. The smallest and largest value are given in bold face.

Both criteria, adjusted Rand index and variation of information, suggest to use the same two clusterings for the final external validation. Therefore only the variation of information is shown in Table 2. The last row contains the column sums. The smallest value, indicating the most central clustering, is found in the column of agglomerative Euclidean clustering with eight clusters. Further the largest value is the one for correlation based PAM with 10 clusters.

4.3 External validation

External validation is the final step in evaluating clusterings of gene expression data. It involves information about the genes that have not been used in the cluster analysis itself and aims at evaluating the biological relevance of the clusters. As mentioned briefly in Section 3.4 one possibility of assessing the biological meaning of a cluster is by looking at the functional annotations of its constituent genes. Based on the assumption that genes with similar functions or genes involved in the same biological processes are also expressed similarly, we expect meaningful clusters to group exactly these genes. In other words an optimal cluster would reflect all those and only those genes in a data set having the same function or participating in the same biological process.

One way to check this property is by scanning the Gene Ontology terms associated with the genes in a cluster. Gene Ontology (GO) [23] is a widely accepted approach for a unified vocabulary to describe molecular functions, biological processes and cellular components of genes or gene products. The terms used for the description of genes and gene products are organized as a directed acyclic graph (DAG) and become more precise at the lower levels of the graph. It is possible that a single gene has multiple functions, takes part in different processes and appears in several components of the cell. Further, each term in the ontology can have several (less specialized) parent terms and the terms themselves follow the "true path rule". This means that a gene product described by a child term is also described by all parent terms.

Many tools exist for accessing the Gene Ontology. For our purpose of evaluating the genes in a cluster relative to a reference set FatiGO [2] is suitable. It is a web-based application (`http://fatigo.bioinfo.cnio.es`) that extracts the GO terms for a query and a reference group of genes and further computes several statistics for the query group. We used FatiGO to access the biological process annotations for each cluster C_i in the clusterings selected in the last section. As reference set we used the union of the corresponding complementary clusters, i. e. all of the clustered genes that do not fall into this cluster C_i. The GO level to be used in the analysis has to be fixed in advance between 2 and 5. We used level 5 because most of the genes under study actually have annotations at this level[7]. As a consequence only the subset of a cluster consisting of genes that have level 5 annotations can be evaluated this way.

We used two criteria for validating clusters externally. First the cluster selectivity is assessed. This means that the proportion of genes with a certain annotation in the cluster relative to all genes in the data having this annotation is determined. A high selectivity thus indicates that the clustering algorithm is able to distinguish these genes well, based on their expression profiles, among all genes.

The second criterion is the cluster sensitivity, the proportion of genes with a certain annotation relative to all genes within the same cluster. If the sensitivity of a cluster is high then most genes in the cluster have the same annotation, in this case they participate in the same process. This is important for annotating previously unknown genes. The putative biological process for an unknown gene found in a very sensitive cluster can be given with a higher confidence compared to unknown genes in a cluster representing genes from many different processes.

For the cell cycle data we have selected the two clusterings "agglomerative with Euclidean distance and eight clusters" and "PAM with correlation dissimilarity and 10 clusters". It is not possible to give the validation results for each cluster. Rather we give only selected results which have some interesting properties.

It must be stated that most clusters are neither very selective nor very sensitive. This may be caused on the one hand by using GO annotations from a too high level. When the level is too high, the categories are too coarse so that genes participating in subprocesses with rather different expression properties still have the same annotation from the common ancestor node in the GO tree. Of course this results in a rather low selectivity because the clustering algorithm will not group genes from these subprocesses together due to their different expression profiles. On the other hand when the level is too low, meaning that the annotations are very specific, only few genes actually have a annotation at this level and therefore only a few can have a common annotation. In this case the sensitivity of a cluster is generally low unless the cluster sizes are very small and consequently the number of clusters is undesirably large. This shows that there is a trade-off between cluster selectivity, sensitivity and the number of clusters.

Figure 6 shows the results of FatiGO[8] for Cluster 3 of the hierarchical Euclidean

[7] Actually almost all genes not being annotated at level 5 have the annotation "molecular function unknown" at level 2.

[8] Note that the three p-values given in the figure are computed by FatiGO to assess the significance of the differences between query and reference set. The first value is the unadjusted p-value, the second and

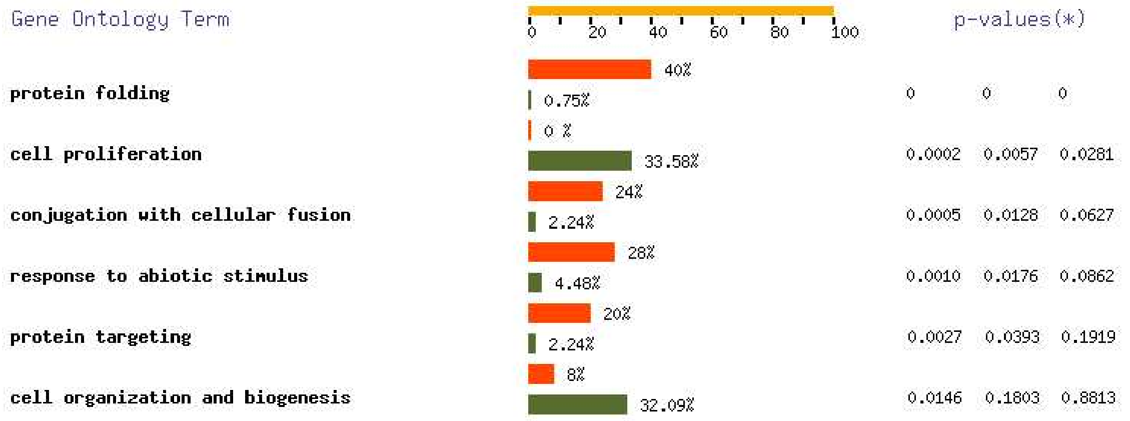


Figure 6: Part of the output of FatiGO for Cluster 3 of hierarchical Euclidean clustering with eight clusters. The six most significant differences between Cluster 3 and the reference set are given. For each GO term the upper bar gives the percentage of genes in the query cluster and the lower bar in the reference set.

clustering with eight clusters. Cluster 3 contains 25 annotated genes and the reference set has 134 annotated genes. The figure shows that this cluster is very selective for genes involved in protein folding. When looking at the absolute numbers it groups 10 out of 11 genes having this annotation. However, it is not very sensitive for protein folding, since 60 percent of the cluster is constituted by genes not using this term.

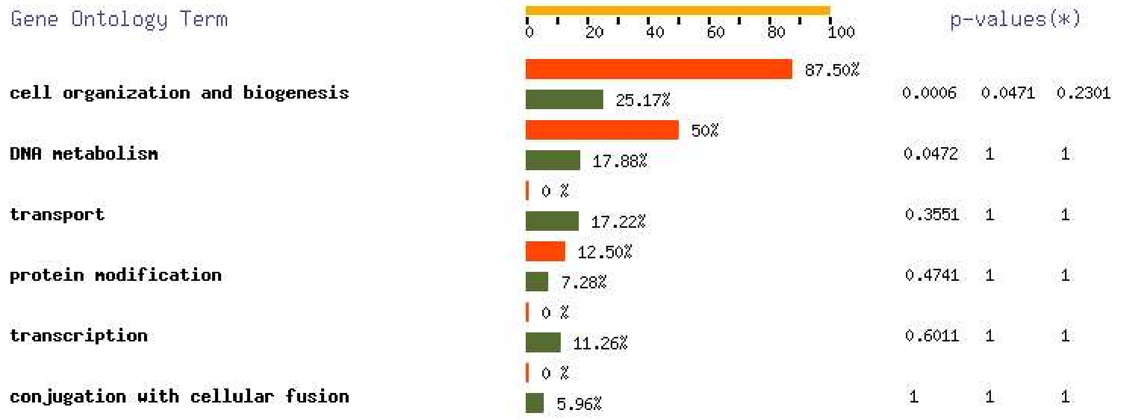


Figure 7: The six most significant differences between Cluster 5 and the reference set for the same clustering as used in Figure 6.

In Figure 7 Cluster 5, a very sensitive cluster, is shown. Seven of the eight annotated genes are labeled with cell organization and biosynthesis. However, it does not have a high selectivity for this feature because only 7 out of 45 genes involved in this process have been selected.

When considering the second clustering, PAM with correlation dissimilarity and 10 clusters, the clusters are generally less selective and sensitive, probably caused by the different dissimilarity measure and its properties. Nevertheless for example Cluster 10 shown in Figure 8 is both selective and sensitive for protein folding. As in the example of Figure 6, it contains 10 out of 11 protein folding genes. But since it is made up by

third value are computed using the *false discovery rate* adjustment procedure by Benjamini and Hochberg assuming independence and arbitrary dependence between GO terms respectively.

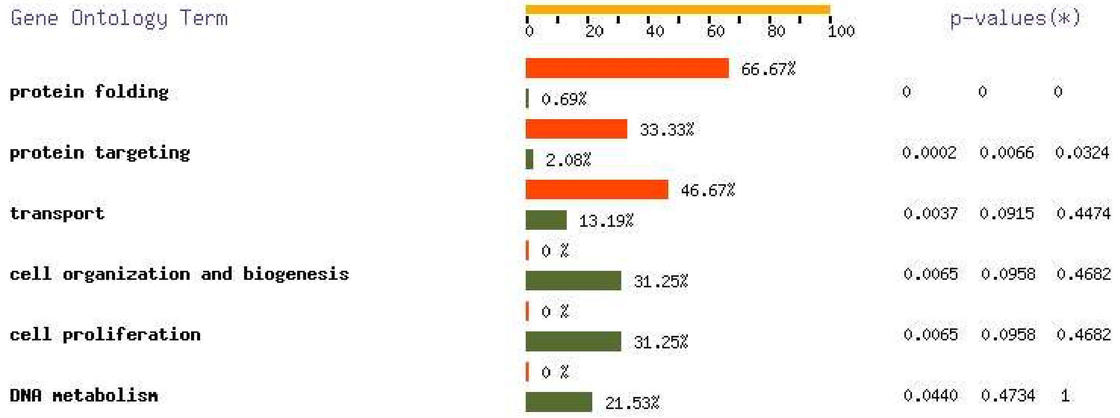


Figure 8: The six most significant differences between Cluster 10 and the reference set, now for PAM clustering with correlation based dissimilarity and 10 clusters.

only 15 annotated genes it also has a good sensitivity of 66 percent for protein folding. Another property of this cluster is that it is negatively selective for GO terms like “cell proliferation” and others. Although there are a total of 45 genes with this term none of them falls in Cluster 10.

5 Conclusions

The results presented in the previous section show once again that it is impossible to recommend a single algorithm or dissimilarity. All the clusterings evaluated – whether they came from hierarchical or PAM clustering and whether they used Euclidean or correlation based dissimilarities – have some clusters of good quality.

But still many clusters are neither very selective nor very sensitive for certain biological processes. This may have many different reasons. First, the clustering quality is highly dependent on the experimental design underlying the data. Not all experiments have the same power to extract certain groups of genes by analysing their expression profiles. Genes whose expression does not depend on the conditions tested in the experiment will most likely not show specific behaviors since they will behave asynchronously and thus not much differential expression can be expected. This means that genes with the same biological annotation do not necessarily have similar expression profiles and hence the selectivity for these genes will be low. Further, often the assumption that genes with similar functions or biological processes actually have similar expression patterns does not hold, in particular for higher and more unspecific levels in the GO tree. But this assumption is crucial for example to reveal the function of unknown genes by gene expression analysis.

In contrast, this is not a problem when the goal is to find regulatory motifs in the sequences of co-expressed genes or to discover regulatory networks. Here the relationship between co-regulation and co-expression is much closer and external validation must follow different approaches from what is described in Section 4. However, in this case it is more important that the dissimilarity measures take into account for instance that time-shifted profiles might still belong to the same group if the temporal distance is not too large. This is just one example for the importance of the choice of a dissimilarity measure. How similar should genes be when their profiles are scaled by a positive factor or shifted horizontally or vertically? What about scaling with negative factors resulting in negatively correlated genes? These questions have to be answered before clustering.

A second reason for clusters of low quality is that it is especially difficult to cluster microarray data because the expression profiles tend to fill the feature space in a way that the data points are not well separated. This leads to the absence of “natural” clusters and clustering becomes segmentation. The known difficulties with the measuring precision of microarrays are partly overcome by sophisticated normalization methods. Still, the precision could be greatly improved by repeated measurements at the same conditions and a better temporal resolution of time series experiments.

The bottom line is that clustering gene expression data from microarrays is a powerful tool in bioinformatics and can reveal biologically relevant information. But it is important to compare multiple clusterings and not run one algorithm with one parameter setting and then take the results as the true structure of the data. Only the comparison of carefully chosen clusterings can result in reliable conclusions drawn from cluster analysis.

In this article we have shown how to choose and compare clusterings from different algorithms and parameter settings. After selecting meaningful dissimilarity measures a set of candidate clusterings can be chosen by evaluating several internal validation criteria. This includes specifying the number of clusters. Next, relative validation indices can help to determine the differences between clusterings and identify relatively stable clusters that appear in several clusterings. These clusters are likely to be more reliable as their structure is extracted from the data by several algorithms or dissimilarity measures. Finally, if possible suitable external biological information should be used to assess the quality of the clusterings. The Gene Ontology annotations used in this work are of course just one example of a source of external information.

However, even conclusions drawn from “good” clusters can only be seen as indications of biological meaning. The power of cluster analysis of gene expression data is that it can greatly reduce the search space and thus can lead biologists towards promising presumptions which are worth further biological examination. The verification of these presumptions by biological experiments is not replaceable.

References

[1] Chipping Forecast. *Nature Genetics Supplement* (1999).

[2] Al-Shahrour, F.; Diaz-Uriarte, R.; Dopazo, J.: FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20 (2004) 4, pp. 578–580.

[3] Bolshakova, N.; Azuaje, F.: Cluster validation techniques for genome expression data. *Signal Processing* 83 (2003) 4, pp. 825–833.

[4] Bryan, J.: Problems in gene clustering based on gene expression data. *J. Multivariate Analysis* 90 (2004) 1, pp. 44–66.

[5] Calinski, R. B.; Harabasz, J.: A dendrite method for cluster analysis. *Communications in statistics* 3 (1974), pp. 1–27.

[6] Cleveland, W. S.; Grosse, E.; Shyu, W. M.: *Statistical Models in S*, chap. 8. Wadsworth & Brooks/Cole. 1992.

[7] Datta, S.; Datta, S.: Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics* 19 (2003) 4, pp. 459–466.

[8] Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D.: Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95 (1998), pp. 14863 – 14868.

[9] Gat-Viks, I.; Sharan, R.; Shamir, R.: Scoring clustering solutions by their biological relevance. *Bioinformatics* 19 (2003) 18, pp. 2381–2389.

[10] Hastie, T.; Tibshirani, R.; Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer. 2001.

[11] Hubert, L.; Arabie, P.: Comparing partitions. *Journal of Classification* 2 (1985), pp. 193–218.

[12] Jain, A. K.; Dubes, R. C.: *Algorithms for clustering data*. Prentice-Hall, Inc. 1988.

[13] Janowitz, M. F.: *Short Course: A Combinatorial Introduction to Cluster Analysis*. Classification Society of North America. URL `http://www.pitt.edu/~csna/reports/janowitz.pdf`. 2002.

[14] Kaufman, L.; Rousseeuw, P. J.: *Finding groups in data*. John Wiley & Sons. 1990.

[15] Krzanowski, W. J.; Lai, Y. T.: A criterion for determining the number of groups in a data set using sum of squares clustering. *Biometrics* 44 (1985), pp. 23–44.

[16] Meilă, M.: Comparing Clusterings. Tech. Rep. 418, University of Washington. 2002.

[17] Milligan, G. W.; Cooper, M. C.: An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50 (1985), pp. 159–179.

[18] R Development Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. 2003.

[19] Rand, W. M.: Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association* 66 (1971), pp. 846–850.

[20] Rousseeuw, P. J.: Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65.

[21] Spellman, P. T.; Sherlock, G.; Zhang, M. Q.; Iyer, V. R.; Anders, K.; Eisen, M. B.; Brown, P. O.; Botstein, D.; Futcher, B.: Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Molecular Biology of the Cell* 9 (1998), pp. 3273–3297.

[22] Sturn, A.: *Cluster analysis for large scale gene expression studies*. Master's thesis, Technische Universität Graz and The Institute for Genomic Research (TIGR) Rockville. 2001.

[23] The Gene Ontology Consortium: Gene ontology: tool for the unification of biology. *Nature Genetics* 25 (2000), pp. 25–29.

[24] Tibshirani, R.; Walther, G.; Botstein, D.; Brown, P.: Cluster validation by prediction strength. Tech. rep., Stanford University. 2001.

[25] Ward, J. H.: Hierarchical grouping to optimize an objective function. *J. Amer. Statist. Assoc.* 58 (1963), pp. 236–244.

[26] Wichert, S.; Fokianos, K.; Strimmer, K.: Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics* 20 (2004) 1, pp. 5–20.

[27] Yang, Y. H.: *Statistical methods in the design and analysis of gene expression data from cDNA microarray experiments*. Ph.D. thesis, University of California, Berkeley. 2002.

[28] Yang, Y. H.; Dudoit, S.; Luu, P.; Lin, D. M.; Peng, V.; Ngai, J.; Speed, T. P.: Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucl. Acids. Res.* 30 (2002) 4, pp. e15–.

[29] Zhang, M. Q.: Large-Scale Gene Expression Data Analysis: A New Challenge to Computational Biologists. *Genome Research* 9 (1999) 8, pp. 681–688.