On the (lack of) robustness of gene expression data clustering

JONAS GAMALIELSSON AND BJÖRN OLSSON
Department of Computer Science
University of Skövde
Box 408, 541 28 Skövde
SWEDEN
[bjorne,gam]@ida.his.se

Abstract: We assess the robustness of partitional clustering algorithms applied to gene expression data. A number of clusterings are made with identical parameter settings and input data using SOM and *k*-means algorithms, which both rely on random initialisation and may produce different clusterings with different seeds. We define a reproducibility index and use it to assess the algorithms. The index is based on the number of pairs of genes consistently clustered together in different clusterings. The effect of noise applied to the original data is also studied. Our results show a lack of robustness for both classes of algorithms, with slightly higher reproducibility for SOM than for *k*-means.

Keywords: Bioinformatics, Gene expression, Expression analysis, Clustering, Evaluation methods

1 Introduction

The reliability of cDNA microarray experiments has recently been questioned. There are many sources of variation [1], e.g. extracted mRNA samples being different due to variations in tissue and RNA degradation, reverse transcription from mRNA producing DNA species of different length, variations during the labeling process due to nucleotide composition, uncertainties in the PCR amplification of sample cDNA, different pins spotting different amounts of cDNA, problems related to hybridisation (e.g. temperature affecting the efficiency of the reaction), sample cDNA being unequally distributed over slides and the hybridisation reaction performing differently in different parts of the array, non-specific hybridisation, and uncertainties in the image analysis.

In addition to the factors pointed out in [1] the subsequent data analysis is also a source of uncertainty. The simplest form of analysis is to study if the difference in expression for a gene is significant for two different conditions [2,3]. Cluster analysis is a form of unsupervised classification and is a more complex way of studying expression data. Genes with similar expression profiles are believed to be functionally related, and clusters therefore give clues of co-regulation. A variety of clustering algorithms have been proposed for the analysis of gene expression data. It is common to discriminate between hierarchical and partitional algorithms [4]. Hierarchical algorithms derive a tree structure where the clustered objects are leaves, whereas partitional algorithms use a predefined number of clusters and assign each object to exactly one cluster. Clustering of gene expression data has been done both using hierarchical [5,6] and partitional algorithms, such as *k*-means [7] and SOMs [8,9]. More recent and problem specific algorithms include gene shaving [10], CLIFF [11] and CLICK [12]. However, different algorithms often produce different clusterings, and it is not known which algorithm is most appropriate for the task. Some researchers have systematically compared different classes of clustering algorithms. One empirical comparison [13] using 252 data sets with various levels of imperfections (dispersion, outliers, irrelevant variables, and nonuniform cluster densities), found that SOMs outperformed seven hierarchical clustering algorithms in accuracy and robustness. The study did not, however, tell us which one of *k*-means and SOMs, would perform better on gene expression data.

Both SOM and *k*-means algorithms iteratively assign objects in a dataset to different clusters. The number of clusters is predefined and a similarity measure is used to calculate how close an object is to a cluster. These algorithms often rely on randomised starting points and take a set of objects as input data. It was noticed in [14] that the order in which objects are processed affects the outcome of *k*-means. It was found that the effect is marginal when clusters are well-separated, and that it is principally objects falling between clusters that cause problems. Both [15] and [16] found that different starting partitions of optimisation methods can lead to different local optima but that well-structured data should minimise this problem. It was suggested in [17] that different groupings for different initial partitions and slow convergence rate are often signs of an incorrect number of clusters. Choosing the number of clusters is a notoriously hard problem and has been studied by numerous researchers. A detailed study investigating 30 methods for determination of the number of clusters was done in [18], while [19] is an example of more recent work on this problem in the context of gene expression data.

[20] describes a multi-stage procedure for assessing the reproducibility of hierarchical clusterings of expression data. First, the clustering tendency of the data is examined. Two indices of reproducibility are then calculated in case of inherent structure in the data. The reproducibility assessment procedure includes perturbation of the data in order to simulate data being collected again. Hence, [20] does not consider reproducibility in terms of random effects introduced by the algorithm itself. [21] used a bootstrap procedure for assessment of clusterings of microarray data. An analysis of variance model was used, containing factors accounting for experimental variation in mRNA samples, arrays, dyes and genes. The work in [21] relies on experimental replication.

We first discuss sources of uncertainty in gene expression experiments and issues related to the clustering of gene expression data. We then discuss the issue of robustness of clustering algorithms and present methods used for clustering and assessment of reproducibility. We describe a series of experiments, and present results and implications of our findings.

2 Clustering methods evaluated

The *k*-means algorithm for assigning *m* objects to *k* clusters contains the following steps [22]:

1. Assign the first *k* objects to *k* separate clusters.
2. Assign each of the remaining (*m*-*k*) objects to the cluster having the centroid that is most similar to the object's data vector by some similarity measure. Recompute the centroid of the gaining cluster after each object assignment.
3. When all objects have been assigned, the cluster centroids are recalculated and each object is once again assigned to the nearest cluster.

In our work, an object is a vector of expression values where each value is measured at a specific timepoint. Commonly used similarity measures are Euclidean distance and Pearson correlation, and implementations of *k*-means often feature random initialisation and various convergence criteria. We used the Cleaver 1.0 implementation of *k*-means (classify.stanford.edu).

The neural network-based self-organizing map (SOM) [23] iteratively moves nodes representing clusters to a configuration adapted to the objects of a particular dataset. The GeneCluster 1.0 implementation [8] was used in our experiments. A SOM has an input layer containing *p*-dimensional observations **o** and an output layer of *k* nodes representing *k* clusters [24]. Each node has an associated *p*-dimensional weight vector **w**. The main steps of the SOM algorithm are:

1. Randomize each vector **w** in the interval [0,1].
2. Calculate similarity between randomly selected observation **o** and weight vector **w** of each node.
3. Update **w** of the node most similar to the observation as $\mathbf{w}_{\text{new}} = \mathbf{w}_{\text{old}} + \alpha(\mathbf{o} - \mathbf{w}_{\text{old}})$, where α is the learning rate, which decreases over time. A small neighborhood of nodes around the winning node are also updated but with a smaller α. The neighborhood also decreases over time.

An epoch consists of repeating steps 2 and 3 for all observations, and the algorithm usually runs for a large number of epochs.

3 Clustering assessment

The *N* by *M* gene expression matrix **G** contains continuous expression values, where *N* is the number of genes and *M* is the number of conditions. Expression profile $\mathbf{p}_k$ for gene *k* contains expression values for *M* conditions, where $p_{ki}=G_{\text{ki}}$ for gene *k* and condition *i*. Clustering vector **C** contains a set of clusterings, where an individual clustering C_x contains a pre-defined number of clusters η where each cluster contains a set of genes. The clustering reproducibility assessment algorithm used in our work calculates reproducibility index *R*, which reflects how consistently pairs of genes appear in the same cluster in different clusterings of a dataset. *R* varies in the interval [0,1], where *R*=1 means perfect reproducibility. The algorithm is defined as:

```
$n \leftarrow 0$
for all 2-combinations of clusterings $C_i$ and $C_j$ do
    for all 2-combinations of profiles $\mathbf{p}_k$ and $\mathbf{p}_l$ do
        if $\mathbf{p}_k$ $\mathbf{p}_l$ belong to same cluster in $C_i$ do begin
            $n \leftarrow n+1$
            $\kappa_n \leftarrow corr(m^{C_j,\mathbf{p}_k}, m^{C_j,\mathbf{p}_l})$
        end
$R \leftarrow |\{\kappa_x \mid \kappa_x=1 \wedge x = 1,..,n\}|/n$
```

where $m^{C_j,\mathbf{p}_k}$ is the mean profile (centroid) of the cluster in which $\mathbf{p}_k$ appears in clustering C_j, while $corr(m^{C_j,\mathbf{p}_k}, m^{C_j,\mathbf{p}_l})$ is Pearson correlation between the centroids of the clusters in which profiles $\mathbf{p}_k$ and $\mathbf{p}_l$ appear in C_j. Pearson correlation between $\mathbf{p}_k$ and $\mathbf{p}_l$ is:

$$corr(\mathbf{p}_k,\mathbf{p}_l) = \frac{\sum_{i=1}^{M}(p_{ki}-\overline{\mathbf{p}}_k)\cdot(p_{li}-\overline{\mathbf{p}}_l)}{S(\mathbf{p}_k)\cdot S(\mathbf{p}_l)} \quad (1)$$

where

$$S(\mathbf{p}_x) = \sqrt{\sum_{i=1}^{M}(p_{xi}-\overline{\mathbf{p}}_x)^2} \quad (2)$$

This reproducibility index is similar to the Rand-index, originally proposed in [25]. The Rand-index is based on pairwise object comparisons for two different clusterings C_i and C_j, where a similarity between C_i and C_j occurs when a pair of objects are placed either in the same cluster in both C_i and C_j, or in different clusters in both C_i and C_j, whereas *R* only considers objects that are consistently clustered together. We argue that *R* is more relevant than the Rand-index for the assessment of gene expression clusterings, since it is normally only co-clustered genes that are of interest in the biological analysis of clustering results, whereas the fact that two genes occur in separate clusters is nearly always ignored in the biological analysis. Also, in many cases, the number of gene pairs that are clustered apart is much higher than the number of gene pairs that are clustered together. Thus, the Rand-index is influenced more by the irrelevant cases than by the relevant cases, which can make the assessment results misleading. In addition, we propose a "softer" means of assessment, where the average value of the vector η containing correlation values between compared objects in the different clusterings, is used. The reader is referred to, for example, [26], for a comparison of other alternative reproducibility indices (among them, the Jaccard statistic, which is similar to R).

A measure of compactness of clusters is average within-cluster distance d^w_{avg} (eq. 3), where d^e is Euclidean distance (eq. 4), η is the number of clusters, N_i is the number of profiles in cluster i. D^w_{avg} is defined as average d^w_{avg} for all clusterings in the experiment:

$$d^w_{avg} = \frac{1}{\eta}\sum_{i=1}^{\eta}\frac{2}{N_i(N_i-1)}\sum_{k=1}^{N_i-1}\sum_{l=k+1}^{N_i} d^e(\mathbf{p}_k,\mathbf{p}_l) \quad (3)$$

$$d^e(\mathbf{p}_k,\mathbf{p}_l) = \sqrt{\sum_{i=1}^{M}(p_{ki}-p_{li})^2} \quad (4)$$

Maximum within-cluster distance $d^w_{\max}$ for a clustering is defined as:

$$d^w_{\max} = \max(\{d^e(\mathbf{p}_k,\mathbf{p}_l) \mid i=1,..,\eta; k=1,..,N_i-1; l=k+1,..,N_i\}) \quad (5)$$

$D^w_{\max}$ is defined as average $d^w_{\max}$ for all clusterings in an experiment. Average between-cluster distance d^b_{avg} shows how separated the clusters are, and is defined as:

$$d^b_{avg} = \frac{1}{\Gamma}\sum_{i=1}^{\eta}\sum_{j=1}^{\eta}\sum_{k=1}^{N_i}\sum_{l=1}^{N_j} d^e(\mathbf{p}_k,\mathbf{p}_l) \quad (6)$$

$$\Gamma = \sum_{i=1}^{\eta}\sum_{j=1}^{\eta}\sum_{k=1}^{N_i}\sum_{l=1}^{N_j}\gamma_{ij} \quad (7) \qquad \gamma_{ij} = \begin{cases} 1 & if\ i \neq j \\ 0 & otherwise \end{cases} \quad (8)$$

D^b_{avg} is defined as average d^b_{avg} for all clusterings in an experiment. Minimum between-cluster distance $d^b_{\min}$ for a clustering indicates the smallest distance between two genes in different clusters and is defined as: (9)

$$d^b_{\min} = \min(\{d^e(\mathbf{p}_k,\mathbf{p}_l) \mid i=1,..,\eta; j=1,..,\eta; k=1,..,N_i; l=1,..,N_j; i \neq j\})$$

$D^b_{\min}$ is defined as average $d^b_{\min}$ for all clusterings in an experiment. The distance measures introduced here are similar to compactness and isolation [4], but are easier to implement. Compactness measures the internal cohesion of objects in a cluster, whereas isolation measures the separation between clusters.

Two public datasets were used. One is the dataset of the HL-60 model for hematopoietic differentiation used in [8], and we applied the same variation filter and normalisation to the dataset using Genecluster 1.0. The variation filter removed genes with expression values below 20 and above 20000, genes where the quotient between the maximum and minimum value is below 3, and where the difference between the maximum and minimum values is below 100. Linear normalisation was performed after the variation filter, according to:

$$\mathbf{p} = \frac{\mathbf{x}-\bar{\mathbf{x}}}{\sigma_{(\mathbf{x}-\bar{\mathbf{x}})}} \quad (10)$$

where $\mathbf{x}$ is the expression vector before normalisation, which gives $\bar{\mathbf{p}} = 0$ and $\sigma^2_{\mathbf{p}} = 1$. In [8] was reported that 567 genes of 7229 passed the variation filter, but 585 genes passed it in our work despite following the description in [8] in our implementation. The dataset contains 4 conditions. Hence, **G** is sized 585 by 4.

A dataset showing the variation in expression during the central nervous system (CNS) development in rats was also used [6]. This dataset contains 112 genes and 9 conditions. In [6] the genes were clustered by Euclidean distance using the FITCH software [27] and identified four different groups, where most of the genes appeared. Two other clusters containing genes with estimated constant or diverging expression profiles were also identified. In [6], no filtering or normalisation was used. In our work, only the 85 genes found in the four waves were used (i.e. **G** is sized 85 by 9). We used the same normalisation procedure as for the HL-60 data.

Randomised datasets of the same sizes as the HL-60 and CNS datasets were also used. Each element of **G** was assigned a random number from a uniform distribution with the interval [0,1] and each expression profile was normalised (eq. 10). Noise was also applied to the HL-60 and CNS datasets in some of the experiments by adding a random number from a normal distribution with mean 0 and variance σ^2 to each element in **G**. This approach was also used in [28] and [20]. No further normalisation was performed after the application of noise.

Five experiments were performed:

1) *Reproducibility of SOM.* Repeated clusterings were performed using the HL-60 and CNS datasets and GeneCluster 1.0. Reproducibility values and cluster distances were calculated, and the effect of different initialisation- and neighbourhood-functions was investigated. In addition, randomised datasets with the same size as the HL-60 and CNS datasets were used as a comparison. As in [8], we used 12 clusters (4 rows by 3 columns). The number of epochs was set to 1000

to ensure convergence. The default settings were used for the other parameters (Num seeds=1, $\alpha_i = 0.1$, $\alpha_f = 0.005$, $\sigma_i = 5$ and $\sigma_f = 0.2$). 10 independent clusterings were performed for each combination of dataset, initialisation function, and neighbourhood function (16 combinations and 160 clusterings).

2) *Reproducibility of k-means*. Except for the use of *k*-means, this experiment is similar to the first experiment. Euclidean distance was used, as well as 100 iterations, which is the maximum number of iterations allowed. Four clusters were used in our experiments. Since *k*-means uses random initialisation and has no neighbourhood function, only 10 clusterings were needed per dataset (40 in total).

3) *Reproducibility of random clustering.* Random clustering is a baseline to compare SOM and *k*-means with. Each expression profile in **G** is assigned to a cluster by generating a random cluster number (integer) from a uniform distribution in the interval [1,η] where η is the total number of clusters. 10 clusterings were performed per dataset (40 in total).

4) *Reproducibility of SOM using noisy data.* Same as experiment 1 except that that each clustering is performed on a dataset with noise added. Hence, it is the clustering reproducibility in the presence of measurement noise that is tested. The noise variance was set to 0.1 which is of similar magnitude as the noise used in [28]. It has also been claimed that microarray data can contain as much as 30-50% measurement noise [29], which further motivates this experiment. The random variants of the HL-60 and CNS datasets were not used. For each combination of dataset, initialisation function, and neighbourhood function, we performed 10 independent clusterings, each time adding random noise to the original data.

5) *Reproducibility of k-means using noisy data.* This experiment is similar to experiment four, except that the *k*-means algorithm and two different noise variance values were used.

4 Results

Looking first at the differences between the reproducibility indices, table 1 shows that the average Rand-index varies in the interval [0.93,0.96], whereas the interval for *R* is [0.80,0.91]. This shows that the number of similarities from co-clustered pairs are few, i.e. the high Rand-index results are largely influenced by similarities of object pairs that are not co-clustered. A similar difference between the Rand-index and *R* are found in table 2. For random clustering (table 3) it is evident that the number of consistently co-clustered genes is particularly low, and the advantage of *R* becomes particularly apparent. The Rand-index is heavily influenced by the fact that many genes are consistently clustered apart even in a random clustering, simply because the probability of two genes ending up in different clusters twice is very high. Thus, the Rand-index shows as high values as [0.59,0.84], which can mislead the user into thinking that the reproducibility is high. *R*, taking only consistently co-clustered genes into account, reveals the the low reproducibility by giving as low values as [0.09,0.29].

Table 1. Results for SOM. The **D** column is for dataset (H = HL-60, C = CNS, Hr = randomized HL-60, Cr = randomized CNS). **I** is initialisation method (V = Random vectors, D = Random datapoints), **N** is neighbourhood function (B = Bubble, G = Gaussian), **R** is reproducibility index, $\overline{\kappa}$ is the mean of the correlation vector κ.

D	**I**	**N**	**R**	$\overline{\kappa}$	**Rand**	D^w_{avg}	D^w_{max}	D^b_{avg}	D^b_{min}
H	V	B	0.78	0.95	0.95	2.10	3.46	2.35	0.15
H	V	G	0.98	1.00	0.96	2.10	3.46	2.34	0.13
H	D	B	0.81	0.97	0.96	2.10	3.46	2.35	0.14
H	D	G	0.94	0.99	0.99	2.10	3.46	2.34	0.13
		Avg.	0.88	0.98	0.97	2.10	3.46	2.35	0.14
Hr	V	B	0.78	0.89	0.96	2.31	3.46	2.44	0.13
Hr	V	G	0.77	0.85	0.95	2.31	3.46	2.44	0.13
Hr	D	B	0.83	0.91	0.97	2.31	3.46	2.44	0.14
Hr	D	G	0.83	0.91	0.97	2.31	3.46	2.44	0.13
		Avg.	0.80	0.89	0.96	2.31	3.46	2.44	0.13
C	V	B	0.84	0.91	0.92	3.22	5.44	3.93	0.90
C	V	G	1.00	1.00	1.00	3.22	5.44	3.97	0.87
C	D	B	0.81	0.90	0.91	3.22	5.44	3.95	0.93
C	D	G	1.00	1.00	1.00	3.22	5.44	3.97	0.87
		Avg.	0.91	0.95	0.96	3.22	5.44	3.96	0.89
Cr	V	B	0.78	0.76	0.89	3.93	5.51	4.12	1.85
Cr	V	G	0.78	0.76	0.89	3.93	5.51	4.12	2.02
Cr	D	B	0.97	0.97	0.98	3.94	5.51	4.12	1.99
Cr	D	G	0.92	0.92	0.96	3.93	5.51	4.12	2.01
		Avg.	0.86	0.85	0.93	3.93	5.51	4.12	1.97

In the reproducibility results for SOM (tab. 1) the most important finding is that reproducibility is sub-optimal (*R*<1) in almost all cases, the only exceptions being when the Gaussian neighbourhood function is used on the CNS data. Interesting is also that reproducibility is almost as high for the randomised datasets as for the original datasets. The initialisation functions gave similar average reproducibility on non-randomised data ($\overline{R} = 0.90$ and 0.91 for random vectors and random datapoints initialisation, respectively). For randomised data, however, random datapoints initialisation gave clearly better $\overline{R}$ (0.89) than random vectors initialisation (0.78). The Gaussian neighbourhood function seems considerably better (0.98) than Bubble (0.81) for the original datasets. Bubble is the default setting in GeneCluster, but our results indicate that this may be unfortunate. On the other hand, $\overline{R}$ is almost equal for Gaussian and Bubble on the randomised versions of the datasets (0.82 and

0.84, respectively). The difference between D^w_{avg} and D^b_{avg} is smaller for the randomised datasets, reflecting that these are less structured than real datasets. Variation is generally absent or negligible in standard deviation for all distance measures: for 40 of the 64 distance values in tab. 1 standard deviation over the 10 clusterings was 0.005 or lower, and in 60 of the 64 cases it was 0.05 or lower. The highest relative standard deviation (15%, or 0.27 for $D^b_{min} = 1.85$) was found for dataset C when using random vectors initialisation and the Bubble neighbourhood function.

In most cases, $\overline{\kappa} > R$. This is reasonable since R is based on cases where $\kappa = 1$ and since $\overline{\kappa}$ is a "softer" reproducibility measure. However, R is a more relevant measure for partitional clustering algorithms, since (in contrast to hierarchical clustering) the information that two profiles have been assigned to neighbouring clusters is not used when interpreting the clustering.

Tab. 2 shows reproducibility results for *k*-means. *R*-values for both HL-60 and CNS datasets are worse than the corresponding *R*-values for SOM. R = 0.67 for the HL-60 dataset can be compared with $\overline{R} = 0.88$ for the four different parameter settings used for SOM. The corresponding values for the CNS data are 0.74 vs 0.91. R is also worse for *k*-means than for SOM on random datasets. Overall, the two experiments indicate higher reproducibility of SOM than of *k*-means. The values of the four different distance measures are approximately the same as those for the SOM experiment, except for the Hr dataset where all distance values besides D^w_{max} are lower. As in the SOM experiments, the standard deviation of almost all distance measures was very low. The highest relative standard deviation (0.13 of 0.97 for D^b_{min}) was found when using the CNS data.

Table 2. Results for *k*-means clustering.

D	R	$\overline{\kappa}$	Rand	D^w_{avg}	D^w_{max}	D^b_{avg}	D^b_{min}
H	0.67	0.93	0.92	2.10	3.46	2.37	0.13
Hr	0.58	0.83	0.93	2.07	3.46	2.18	0.00
C	0.74	0.85	0.85	3.27	5.43	4.06	0.97
Cr	0.50	0.36	0.74	3.91	5.50	4.12	1.69

Table 3. Results for random clustering.

D	R	$\overline{\kappa}$	Rand	D^w_{avg}	D^w_{max}	D^b_{avg}	D^b_{min}
H	0.09	0.71	0.84	2.09	3.46	2.18	0.00
Hr	0.09	0.11	0.84	2.31	3.46	2.31	0.02
C	0.27	0.84	0.60	3.25	5.43	3.49	0.45
Cr	0.29	0.32	0.59	3.93	5.46	3.92	0.97

Comparing tab. 1 and 2 shows that both algorithms have higher average reproducibility on the CNS data than on the HL-60 data. There may be a number of reasons for this. The inherent structure in the data, the number of conditions, the number of genes and the number of clusters are all likely to affect reproducibility. The effect of dimensionality was tested by reducing the number of conditions of the CNS data using principal component analysis (PCA). Three dimensions accounting for 87% of the total variation in the dataset were chosen, and the new dataset clustered 10 times by *k*-means, which gave R = 0.91 compared to R = 0.74 when all nine conditions were used, indicating that the dimensionality of datasets affects reproducibility. The PCA results in fig. 1 do not show any distinct groupings in the CNS data. For the HL-60 data (fig. 2), the three most important dimensions accounted for >99% of the variation in the data, but distinct groupings are again missing.

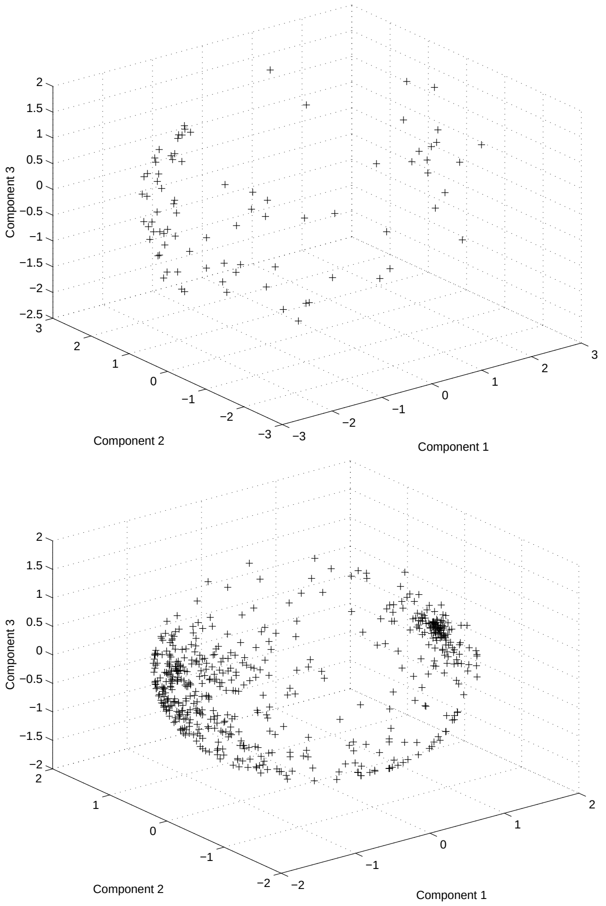

Figure 1. PCA for CNS (upper) and HL-60 (lower), obtained by the Cleaver (1.0) software using three components accounting for, respectively, 87% and 99% of the variation in the data.

Reproducibility results for random clustering are shown in tab. 3. *R*-values are close to 1/η, which is reasonable since R = 1/η is what we would expect from randomly assigning profiles to clusters. In addition, $\overline{\kappa}$ is considerably lower for the randomised datasets than for the original datasets. This fact confirms that profiles are generally less similar in the randomised datasets. It can also be noticed that D^b_{avg} is smaller for datasets HL-60 and CNS when comparing with SOM and *k*-means, indicating that the clusters are not as well separated anymore. For the randomised datasets the difference between D^{w}_{avg} and D^{b}_{avg} is zero, showing that there is no structure in neither the data itself nor the clustering of the data. Values for D^{b}_{min} are also lower in most cases when comparing with the SOM and *k*-means algorithms, indicating less separation between clusters.

Tab. 4 shows reproducibility of SOM when noise from a normal distribution (σ^2 = 0.1) was added to the original datasets. The results show that $\bar{R}$ drops from 0.88 without noise to 0.70 with noise for the HL-60 dataset. For the CNS dataset $\bar{R}$ drops from 0.91 to 0.84. This behaviour seems plausible because the noise disorganises the data. The noise also causes a slight increase in the relative standard deviation of all distance measures except D^{w}_{max} for the HL-60 data.

Tab. 5 shows results from the reproducibility tests with noise added to the data and using *k*-means. When σ^2 = 0.1, the difference in *R*-value due to noise application is not evident: *R* = 0.66 compared to *R* = 0.67 (without noise) for the HL-60 dataset and *R* = 0.70 compared to *R* = 0.74 (without noise) for the CNS dataset. The difference in reproducibility is more obvious when σ^2 = 0.3 as *R* drops to 0.42 for the HL-60 dataset, and to 0.63 for the CNS dataset.

Table 4. Results for SOM on noisy data, σ^2=0.1.

D	**I**	**N**	**R**	$\bar{\kappa}$	**Rand**	D^{w}_{avg}	D^{w}_{max}	D^{b}_{avg}	D^{b}_{min}
H	V	B	0.69	0.93	0.93	2.10	3.46	2.35	0.12
H	V	G	0.71	0.94	0.93	2.10	3.46	2.34	0.11
H	D	B	0.68	0.93	0.93	2.10	3.46	2.36	0.12
H	D	G	0.72	0.94	0.94	2.10	3.46	2.34	0.11
Avg.			0.70	0.94	0.93	2.10	3.46	2.35	0.13
C	V	B	0.82	0.89	0.90	3.22	5.44	3.94	0.99
C	V	G	0.86	0.92	0.93	3.21	5.43	3.97	0.96
C	D	B	0.81	0.89	0.90	3.23	5.44	3.93	0.95
C	D	G	0.85	0.91	0.92	3.21	5.43	3.97	0.96
Avg.			0.84	0.90	0.91	3.22	5.44	3.95	0.96

Table 5. Results for *k*-means using noisy data.

D	σ^2	**R**	$\bar{\kappa}$	**Rand**	D^{w}_{avg}	D^{w}_{max}	D^{b}_{avg}	$D^{b}_{\min}$
H	0.1	0.66	0.92	0.92	2.10	3.46	2.36	0.12
H	0.3	0.42	0.83	0.88	2.11	3.46	2.37	0.11
C	0.1	0.70	0.85	0.83	3.28	5.43	4.01	0.98
C	0.3	0.63	0.79	0.81	3.36	5.47	4.00	1.14

5 Discussion

Our results show that results from partitional clustering vary considerably when performing multiple runs with the same input data and identical parameter settings. One reason for this behaviour is the stochastic nature of the tested algorithms where randomised starting points are used. Another possible reason is the fact that expression data can contain as much as 30-50% measurement noise [29]. Clustering relies on the presence of inherent groups in the microarray data, but these groups may be concealed by such high levels of noise. This motivates using clustering tendency methods in order to determine whether the data is suitable for clustering or not, something which [20] does. Weak inherent groupings in the data can be observed in our results as the average within-cluster distance is close to the average between-cluster distance. The presence of measurement noise was also found to further aggravate the robustness problems. Our results show that clustering algorithms should be used with caution. Unreliable clustering results increase the risk of errors in downstream analysis stages, e.g. gene regulatory network derivation. If the reproducibility problem can not be solved, it should at least be taken into consideration when interpreting and using clustering results in downstream analysis.

Our work was limited to two common clustering algorithms. There may be other algorithms suffering from reproducibility problems, but there are also reproducible algorithms. Hierarchical algorithms are typically deterministic, and therefore by definition reproducible. It is important to keep in mind, however, that reproducibility does not imply validity. A deterministic algorithm may have perfect reproducibility while reproducing a clustering that does not agree with the inherent structure of the data.

Results from the experiments using random datasets also show that the clustering algorithms are able to find patterns also in data without inherent groups.

References

1. Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eikchoff, H., Lehrach, H., Herzel, H. (2000) Normalization strategies for cDNA microarrays, *Nucl. Acids Res.* **28**: e47.
2. DeRisi, J.L., Iyer, V.R., Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science* **278**: 680-686.
3. Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., Davis, R.W. (1996) Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes, *Proc. Natl. Acad. Sci. USA* **93**: 10614-10619.
4. Jain, J.K., Dubes, R.C. (1988) *Algorithms for clustering data*, Prentice-Hall, New Jersey.
5. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein D. (1998) Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA* **95**: 14863-14868.
6. Wen, X., Fuhrman, S., Michaels, G.S., Carr, D.B., Smith, S., Barker, J.L., Somogyi, R. (1998) Large-scale temporal gene expression mapping of central nervous system development, *Proc. Natl. Acad. Sci. USA* **95**: 334-339.
7. Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M. (1999) Systematic determination of genetic network architecture, *Nature Genet.* **22**: 281-285.
8. Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., Golub, T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation, *Proc. Natl. Acad. Sci. USA* **96**: 2907-2912.

9. Mody, M., Cao, Y., Cui, Z., Tai, K.-Y., Shyong, A., Shimuzu, E., Pham, K., Schultz, P., Welsh, D., Tsien, J.Z. (2001) Genome-wide gene expression profiles of the developing mouse hippocampus, *Proc. Natl. Acad. Sci. USA* **98:** 8862-8867.
10. Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Botstein, D., Brown P. (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns, *Genome Biology* **1**: 1-21.
11. Xing, E.P., Karp, R.M. (2001) CLIFF: Clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts, Bioinformatics **17** Suppl. 1: 306-315.
12. Sharan, R., Shamir, R. (2000) CLICK: A clustering algorithm with applications to gene expression analysis, in: P. Bourne, M. Gribskov, R. Altman, N. Jense, D. Hope, T. Lengauer, J. Mitchell, E. Scheeff, C. Smith, S. Strande, H. Weissig (eds.), *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB)*, AAAI Press, 307-316.
13. Mangiameli, P., Chen, S.K., West, D. (1996) A comparison of SOM neural network and hierarchical clustering methods, *European Journal of Operational Research* **93**: 402-417.
14. MacQueen, J.B. (1967) Some methods for classification and analysis of multivariate observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, 281-297.
15. Rubin, J. (1967) Optimal classification into groups: An approach for solving the taxonomy problem, *Journal of Theoreticl Biology* **15**: 103-144.
16. Blushfield, R.K. (1976) Mixture model tests of cluster analysis. Accuracy of four agglomerative hierarchical methods, *Psychological Bulletin* **83**: 377-385.
17. Marriott, F.H.C. (1982) Optimization methods of cluster analysis, *Biometrica* **69**: 417-421.
18. Milligan, G.W., Cooper, M.C. (1985) An examination of procedures for determining the number of clusters in a data set, *Psychometrica* **50**: 159-179.
19. Lukashin, A.V., Fuchs, R. (2001) Analysis of temporal gene expression profiles: Clustering by simulated annealing and determining the optimal number of clusters, *Bioinformatics* **17**: 405-414.
20. McShane, L.M., Radmacher, M.D., Fredilin, B., Yu, R., Li, M.C., Simon, R. (2002) Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data, *Bioinformatics* **18**(11): 1462-1469.
21. Kerr, M.K., Churchill, G.A. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments, *Proc. Natl. Acad. Sci. USA* **98**: 8961-8965.
22. Anderberg, M.R. (1973) *Cluster analysis for applications*, Academic Press, New York.
23. Kohonen, T. (1990) The self-organizing map, *Proceedings of the IEEE* **78**: 1464-1480.
24. Everitt, B.S., Landau, S., Leese, M. (2001) Cluster analysis, Arnold, London.
25. Rand, W.M. (1971) Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association* **66**: 846-850.
26. Milligan, G.W., Soon, S.C., Sokol, L.M. (1983) The effect of cluster size, dimensionality, and the number of clusters on recovery of true cluster structure, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **5**(1): 40-47.
27. Felsenstein, J. (1993) *PHYLIP (PHYLogeny Inference Package), version 3.5c*. Department of Genetics, Univ. of Washington, Seattle.
28. Hörnquist, M., Hertz, J., Wahde, M. (2002) Effective dimensionality of large-scale expression data using principal component analysis, *Biosystems* **65**: 147-156.
29. Szallasi, Z. (1999) Genetic network analysis in light of massively parallel biological data acquisition, *Pacific Symposium on Biocomputing* **4**: 5-16.