Validation of systems biology models
Hasdemir Durmus, D.

Citation for published version (APA):
Validation of cluster analysis

Clustering is often regarded as a primary means to analyze genome-wide gene expression data. Co-clustering of genes can indicate co-regulation of functional associations. However, proper validation of cluster analysis often goes disregarded due to the lack of established guidelines. For the purpose of constructing a validation framework, resampling approaches to assess stability of cluster analysis seem promising. In this study, we evaluate the performance of certain cross validation and bootstrapping based measures on realistic, synthetic gene expression datasets consisting of highly overlapping clusters. Subsequently, we apply them on our real transcript dataset from cyanobacterium Synechocystis to test the validity of the cluster analysis we have performed. Having validated the analysis, we perform functional enrichment analysis on the resulting clusters and report biological results that shed light on the effect of the day/night rhythm on transcript regulation in Synechocystis.  

5.1. Introduction
5.1.1. A look into the literature of clustering validation
Clustering is extensively applied for the analysis of large scale biological data. Especially genome-wide transcriptome data analysis studies often incorporate a clustering

\footnote{This chapter is partly described in: S. Andreas Angermayr, Pascal van Alphen, Dicle Hasdemir, Gertjan Kramer, Muzamal Iqbal, Wilmar van Grondelle, Huub CJ. Hoefsloot, Young Hae Choi, Klaas J. Hellingwerf. Dynamics of the molecular composition and physiology of Synechocystis sp. PCC 6803 subject to a circadian regime relevant for mass culturing, \textit{manuscript in preparation}.}
step. Clustering allows to infer the dominant patterns in the data for both genes and individuals, depending on the choice of the variables in a study. As any statistical model, the outcome of a cluster analysis needs to be validated before deriving strong biological conclusions. However, there are no established guidelines for the validation of clustering analyses in the field of biological data analysis.

In a number of studies where the focus is on discriminating between different types of patients, clinical data can provide clues for the actual group labels of the patients. In such cases where information on the actual group labels exist, external validation can be performed. External validation is based on the comparison of the actual group labels with the cluster labels obtained by clustering the data. When actual labels are not available, a common practice is to apply different clustering algorithms and compare the results, namely performing a clustering comparison. In both external validation and clustering comparison, several metrics have been proposed as a measure of the overlap between the two different sets of labels for the clustered objects \([161]\). The metrics can either be based on mutual information (normalized mutual information \([5]\), variation of information \([108]\)), on the best matching groups in the two different sets of labels (F-measure \([93]\)) or on the number of pairs of objects that are in the same cluster/group according to the two different sets of labels (rand index \([124]\), adjusted rand index \([66]\)). However, the task of validation in the absence of actual labels is more abstract due to lack of a direct comparison.

Certain internal measures have been proposed for the validation of a cluster analysis \([58]\). These include measures based on the compactness and separation of the clusters and the combinations of the two such as the Dunn index \([40]\) and the Silhouette width \([126]\). These measures can perform well for quality check of the cluster analysis but they fail in being strict validation measures. This is because maximization of different quality measures depends primarily on the objective function used by the clustering algorithm itself. Therefore, other measures attract attention for validation such as stability based measures.

Stability of a cluster analysis against small perturbations in the data can be assessed by using a variety of techniques which are based on resampling of the data. Assessment of stability can be performed in four ways. Thinking that the data is comprised of expression of genes (rows) in a variety of different experimental conditions / individuals / time points (columns), according to the first technique, one column is left out in turn and the within-cluster similarity of genes’ expression levels are measured \([172]\). Figure of Merit (FoM) is defined as a measure of this similarity and can be plotted for a range of different number of clusters. Typically, when calculated for a variety of different clustering algorithms, FoM of certain algorithms
would dominate the others and the measure would help us to choose the algorithm with the most stable results.

The second technique is based on leaving out genes in turn and assessing stability on the remaining genes. In a variation of this method [17], two non-disjoint datasets (two subsamples) are created from the original dataset, by randomly leaving out 20% of the genes. Both subsamples are clustered and the similarity of the clustering is assessed by a measure based on counting pairs of genes clustered together. The average degree of the similarity between the two clusterings over a sufficient number of different subsamples is considered as a measure of stability. In another variation [98], the similarity of the clusterings on subsamples with the clustering on the original dataset is used to detect stability. These approaches assume that when only a small portion of the dataset is left out, the structure in the data would be maintained and therefore, clustering on the subsamples and on the original dataset has to be consistent with each other.

The fundamental aim in studies focusing on the third technique is to construct a cross validation scheme for clustering [21, 48, 92, 147]. In this approach, dataset is randomly split into two parts: the training and the test parts. Both parts are clustered separately. At the next step, genes in the test part are assigned to the clusters obtained by clustering on the training part. This assignment obviously can be done in different ways. The most important criterion is, however that the assignment of the test genes has to mimic the clustering of the test genes. For a k-means clustering, a nearest centroid classifier can work which means that the test genes will be assigned to the clusters with the closest centroids. The two set of test set labels obtained by clustering and assignment to the training centroids can then be compared. This is repeated for a high number of different training and test sets and the average similarity is used as a measure of stability.

The fourth technique is based on bootstrapping. In this approach, bootstrap datasets are created by residual bootstrapping either in a nonparametric way using the error obtained from a fitted ANOVA model [79] or in a parametric way assuming normally distributed noise on the log-ratios of the expression values [18]. The bootstrap datasets are then clustered and the similarity between the clustering on different bootstrap datasets can be used as a measure of stability.

We should note that not all the four techniques are conceptually similar to each other. Thinking that the gene expression data has two modes, the genes and the samples, these approaches apply clustering and resampling in varying modes. Table 5.1 can be seen for a visual summary of this phenomenon.

In the first part of this study, we focus on the last two methods we mentioned above, namely cross validation and bootstrapping. Cross validation based measures
5. Validation of cluster analysis

Table 5.1: Modes of resampling

<table>
<thead>
<tr>
<th>Resampling Mode</th>
<th>Gene</th>
<th>Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering / Classification Mode</td>
<td>CV for gene clustering [92, 147], similarity of clusterings on subsets of genes [17, 98]</td>
<td>FoM [172], Bootstrapping [18, 79]</td>
</tr>
<tr>
<td>Sample</td>
<td>CV coupled to traditional supervised classification [113], CV for sample clustering [92, 147]</td>
<td></td>
</tr>
</tbody>
</table>

have been proposed primarily for determining the optimal parameters of a clustering algorithm such as the number of clusters in k-means clustering. Here, we rather apply them for stability assessment and validation of cluster analysis based upon stability. To our knowledge, the evaluation of their performance upon clustering of high dimensional, complex and large synthetic gene expression datasets are missing in the literature. Datasets with significant covariance between the variables and high degrees of overlap in the clusters have been neglected and application on real datasets have been performed only for clustering in the sample mode, not in the gene mode. Similarly, performance evaluation of bootstrapping based approaches using synthetic data has not been carried out. Therefore, before applying them on a real dataset, we perform an evaluation of the methods on a realistic synthetic dataset. Later, we apply these methods on our gene expression dataset obtained in time series experiments of Synechocystis in the second part of the study about which we give more details in Section 5.1.2.

5.1.2. Circadian dynamics in Synechocystis

Cyanobacteria, such as Synechocystis sp. PCC 6803 (hereafter: Synechocystis) are oxygenic phototrophic microorganisms that can be used for various biotechnological applications. It can utilize CO$_2$ as the carbon substrate for a sustainable source of CO$_2$-neutral energy sources and commodity chemicals such as bioplastics and biofuels [39, 168].

In this study, cells were grown in a continuous culture in a turbidostat-controlled lab-scale photobioreactor, sparged with N$_2$ and CO$_2$ [117]. The culture was subjected to a circadian rhythm of 12 hours light followed by 12 hours darkness, resulting in oxygen depletion during the dark phase. These are conditions that mimic the conditions in a large scale photobioreactor that can be equipped with a degassing
system and that can be sparged with off-gases originating from the combustion of fossil fuel [73]. Cyanobacteria have adapted to the changing levels of light availability by evolving regulation systems governed by a circadian clock. The clock provides the cells with the means to react in a timed manner to use energy efficiently, i.e. to anticipate a synchronization of their physiology with the availability of light [114] and allows synchronizing between cells in a population [70].

We set out to investigate the transcriptome of such cells from samples taken at time points close to the change of light-availability, to study the immediate response and subsequent adaption to the oscillating light conditions. For an educated biotechnological application it is important to understand the regulation of *Synechocystis* at the transcriptome level. The results of this study contribute to this understanding.

We made use of a custom made Agilent chip for *Synechocystis* sp. PCC 6803 for the purpose of detecting the gene expression levels [158]. The gene expression data consists of 3263 genes measured at 5 time points during a 24 hour time course (TP1: 1 hour after the light onset, TP2: 3 hours after the light onset, TP3: 15 minutes before the dark onset, TP4: 1 hour after the dark onset and TP5: 3 hours after the dark onset) with two biological replicates.

In the second part of this study, we analyze the gene expression data by clustering and perform the validation of the cluster analysis using a comparative approach which takes into account the validity of clustering on a dense single cluster synthetic dataset. Following the validation, we report the results we obtain from a functional enrichment analysis we carried out on the clusters of the real data.

### 5.2. Methods

#### 5.2.1. K-means clustering

For k-means clustering of genes, we used the 'kmeans' function implemented in Matlab [101]. The function uses k-means++ algorithm for initializing the cluster centers [9]. We repeated the algorithm with 200 different starting points and selected the clustering that gives the smallest sum of distances between the genes and the cluster centroids.

We used the mean of biological replicates for the clustering of the real transcriptome data. Prior to clustering, we transformed the data to account for the differences in the absolute values of gene expression. For this purpose, we used z-transformation (see Equation 5.1).
\[ x_z = \frac{x - \bar{x}}{\sigma_x} \quad (5.1) \]

where \( x, x_z, \bar{x} \) and \( \sigma_x \) represent the gene expression at a time point, the standardized value of gene expression, the mean of gene expression across all five time points and the standard deviation of gene expression across all five time points, respectively. For both synthetic and real data clustering, we use an euclidean distance metric. Performing k-means on z-transformed real data by using an euclidean distance metric allows us to focus only on the relative temporal changes of the genes.

5.2.2. Generation of synthetic data
We mimicked the real data while generating the synthetic data. For this purpose, we applied 6-means clustering on the z-transformed real dataset. We calculated the covariance matrix for each of the 6 clusters. Later, we generated Gaussian clusters that are centered around the centroids of the real data clusters and are of equal sizes to the real data clusters. To obtain datasets with increasing degree of overlap between the clusters, we drew the synthetic gene expression values from a multivariate normal distribution with a covariance matrix equal to 0.1%, 5%, 20% and 100% of the covariance matrices of the real data clusters, respectively. On top of every gene expression value, we added relative noise drawn from normal distribution with varying standard deviation.

In order to apply a bootstrapping based stability assessment on the synthetic data, we used 100 different noise realizations of the data as 100 different parametric bootstrap samples. The first noise realization was selected as the original dataset and the other 99 realizations constituted the bootstrap samples.

5.2.3. Generation of bootstrap datasets for real data
For the real data, we applied a non-parametric bootstrapping approach. In this approach, we take the mean of the two biological replicates and calculate the residual of each replicate at each time point. We standardize the residuals through normalization by their mean values. We store the standardized residuals in the matrix \( Q \in \mathbb{R}^{2N \times M} \). For each gene expression value at a certain time point, we draw the standardized residual from the respective column in \( Q \) with replacement, multiply it with the mean value at that time point (\( \bar{Y}_{ij} \)) and add on the mean value. Each bootstrap dataset \( (Y^*) \) then consisted of gene expression values that are averaged across the two replicates \( (Y^{**}) \). Equation 5.2 shows the generation of replicate and average bootstrap datasets.
5.2. Methods

\[
Y_{ij} = \bar{Y}_{ij} + \bar{Y}_{ij} \times \text{error}
\]

\[
Y_{ij}^* = \frac{\sum_{k=1}^{2} Y_{ijk}^*}{2}
\]

where \( \text{error} \in Q_{\star j} \)

\( k=1:2 \) index for the replicates

\( i=1:N \) index for genes

\( j=1:M \) index for time points

(5.2)

5.2.4. Assessment of stability by bootstrapping

We performed k-means clustering on the original dataset and each of the 99 bootstrap samples \((C_U)\) to the clustering on the original dataset \((C_V)\) using the adjusted rand index [66] which is a pairwise similarity metric. Adjusted rand index \((ARI)\) is a variation on the raw Rand index \((RI)\) which is adjusted for chance (see Equation 5.3). The adjusted rand index takes values in the \([-1,1]\) interval where higher values indicate better similarity of the clustering analyses on different datasets.

\[
RI = \frac{a + b}{\binom{N}{2}}
\]

\[
ARI = \frac{RI - E[RI]}{\max(RI) - E[RI]}
\]

\( a = \) number of element pairs that are in the same cluster according to both \(C_U\) and \(C_V\)

\( b = \) number of element pairs that are not in the same cluster according to both \(C_U\) and \(C_V\)

(5.3)

5.2.5. Assessment of stability by cross validation

In cross validation based stability assessment, we split the set of genes into two subsets, the training and the test sets. For 100 different splits of the data, we cluster the training set genes and assign the test set genes to the nearest training set cluster centroid. We denote the group labels of the test set genes that we obtained in this manner with \(C_U\). We also cluster the test set genes and denote their cluster labels with \(C_V\). Then, we compute either a pairwise or a unary similarity metric to assess the similarity between \(C_U\) and \(C_V\). The main assumption here is that due to high number of genes and relatively very low number of clusters, the structure and the location of the clusters would be preserved even when only half of the data is used for clustering.
In the case of the 'Stability Index' measure introduced in [92], we compute a unary dissimilarity metric whose smaller values indicate higher stability. Each cluster in \( C_V \) is mapped to a cluster in \( C_U \) by minimizing a cost function which is the Hamming distance between the clusters mapped to each other in \( C_U \) and \( C_V \), normalized by the number of genes in the clusters. The value of the minimized cost function is then further normalized by that obtained from two random partitionings. The purpose of this normalization is to correct for the bias that this measure tends to decrease with smaller number of clusters.

In the case of the 'Prediction Strength' measure introduced in [147], we calculate a pairwise similarity metric similar to the adjusted rand index. In this metric, we calculate for each cluster in \( C_V \), the number of gene pairs that are in that cluster in \( C_V \) and are also in the same cluster in \( C_U \). We normalize this similarity measure by the total number of gene pairs in that cluster in \( C_V \). The minimum of the normalized measure across all clusters in \( C_V \) gives us the 'prediction strength'. In [147], the authors argue that clustering with a prediction strength greater than 0.9 can be accepted as stable.

5.2.6. Functional enrichment analysis of real data
For the enrichment analysis, we used the gene function categories provided in the Cyanobase database [116]. We computed the p-values by using the cumulative hypergeometric distribution function, 'hygecdf' in Matlab following Equation 5.4.

\[
p - \text{value} = 1 - \sum_{i=0}^{x-1} \frac{K \cdot (M-K)}{N \cdot (M-N)}
\]

where \( K \) is the number of the genes in a specific category, \( N \) is the total number of genes in the cluster, \( x \) is the number of genes in the cluster that are associated with the specific category in question and \( M \) is the total number of genes in the background set. We used a background set of 3263 genes which is the set of all genes on the microarray. We used Bonferroni correction to correct for the multiple testing problem that occurred due to the testing of several gene function categories at the same time.

5.3. Results and discussion
5.3.1. Application on synthetic data
The first cross validation based measure we evaluate in this study is defined in [92] and is named as 'stability index' (See the first quarter in Table 5.1). However, it
actually is a measure of dissimilarity between the group labels of the test objects obtained by clustering them and those obtained by assigning the test objects to the training clusters with the nearest centroid. Therefore, the lower the measure is, the higher the consistency between the two different sets of group labels is. Figure 5.1 shows the behavior of this measure when synthetic data with varying degrees of overlap between the 6 clusters are used. For each cluster number, the stability index values obtained at 100 different splits to the test and training parts are plotted where the red line shows their median. The overlap between the clusters is the smallest in Figure 5.1a and is the highest in Figure 5.1d. In Figures 5.1a-c, we see that the stability index is at its lowest value when 6-means clustering is applied on the data which is the true number of clusters in the data. However, the measure is very low also at some of the smaller number of clusters selected. This means that clustering by using also smaller number of clusters can be regarded as valid by using this measure. However, the index starts to increase after k=6 which means that clustering with k>6 becomes less favorable compared to k≤6. In Figure 5.1d, we see that the index starts to increase steadily already after k=2. This implies that the clusters in the data are not well separable anymore and selecting k above 2 makes the cluster analysis less favorable. A visual inspection of the data (data not shown) also confirms that the clusters are not separable anymore due to very high overlap. We should also note that the particular worsening of stability at k=3 is specific to the structure of the dataset and we do not observe it for some other generated datasets (data not shown).

At this point we would like to stress that we are not trying to optimize the number of clusters (k) using the presented measures. For our validation purposes, assessment of the stability only at a fixed, chosen 'k' where the analysis has been performed is important. However, determining the optimal number of clusters is not a straight-forward task, especially for complex biological data that are usually dealt with in the field. Therefore, we evaluate the stability at an extended set of 'k's that might possibly be selected as optimal for our datasets.

The prediction strength measure by [147] is also based on using cross validation for stability assessment. However, unlike the stability index, it is not a measure of dissimilarity but it is a measure of similarity. In Figure 5.2, we see the behavior of the measure obtained by clustering data with different degree of cluster overlap as we have mentioned also before. A prediction strength over 0.9 suggests a good cluster analysis as mentioned in [147]. We observe similar changes with the stability index measure with respect to the number of clusters selected. Clustering on data with low to medium cluster overlap using k>6 and on data with very high cluster overlap using k>2 results in instability.
Figure 5.1: **Stability Index proposed by** [92]. Boxplots show SI obtained at 100 different splits of the data. The outliers are denoted by the crosses. Red lines show the median of each distribution. The overlap between the clusters in the dataset is the lowest in a and is the highest in d.

Figure 5.2: **Prediction Strength proposed by** [147]. Boxplots show PS obtained at 100 different splits of the data. The outliers are denoted by the crosses. Red lines show the median of each distribution. The overlap between the clusters in the dataset is the lowest in a and is the highest in d.
5.3. Results and discussion

Figure 5.3: Stability assessed by bootstrapping. Boxplots show the ARI obtained in 99 comparisons. Each comparison is performed between the original data and one of the bootstrap samples. Stability is shown for cluster numbers between 4 and 8. The overlap between the clusters in the dataset is the lowest in a-b and is the highest in g. The datasets used in a,c,e and g include 1% noise and the ones in b,d and f include 5% noise in average.

We also use a stability based validation method employing bootstrapping. In Figure 5.3 we see how clustering consistency between different bootstrap samples and the original data changes with respect to two factors. The first factor is the degree of cluster overlap and the second factor is the added experimental noise in the data. The degree of overlap increases from top to the bottom in Figure 5.3 and the noise increases from left to right. We investigate the stability of the clustering at 5 different numbers of clusters that approach the true number of clusters (k=6) from above and below. Each box plot shows the adjusted rand index distribution across all 99 comparisons made between the bootstrap samples and the original dataset. We see that the stability of the clustering using the true number of clusters (k=6) decreases with increasing cluster overlap. Pairwise comparison of Figures 5.3a-b,5.3c-d and 5.3e-f reveals that it decreases also with increasing noise. Additionally, inspection of Figures 5.3a-f shows that the stability starts to decrease when the number of clusters is chosen as more than the true number of clusters (k≥6). However, Figure 5.3g shows an exceptional case in which the original clusters in the data are not well separated anymore. In such a case, the stability of the clustering using the true number of clusters is not exceptionally higher than the stability of clustering obtained by using more than needed cluster numbers. The stability is the highest at k=4 meaning that it starts to decrease slowly after k=4.
As a summary, we conclude that cross validation and bootstrapping based techniques support one another although they are conceptually different (see Table 5.1). Clustering on well structured data is favored due to high stability. Clustering using the true number or lower than the true number of clusters in the data is very stable on these datasets. With increasing overlap between the clusters in the data, the stability, however decreases. When the structure in the data starts to get lost (e.g. when there is high overlap between the clusters in the data) clustering at lower number of clusters than the true number of clusters shows higher stability. Additionally, the average stability that can be achieved by using any number of clusters decreases visibly compared to well separated data.

Usually, in biology we deal with biological data where the clusters in the data are not well separated. Therefore, we find it very important to observe the behavior of the proposed measures on synthetic data with overlapping clusters to assess the validity of cluster analysis on real biological data.

5.3.2. Application on real data
Prior to clustering, we performed pre-processing on the data. To compensate between-chip differences in our experimental setting, we applied the Least Oscillatory Set (LOS) normalization, following the advice of Lehmann et al [97]. Lehmann et al. present an extensive study on normalization methods that can be used for oscillatory gene expression data without disturbing the oscillatory behavior of different genes. The results from their study show that the LOS normalization is the most appropriate method and it achieves patterns that agree best with biological knowledge. We refer the readers to [97] for more details on the method and the implementation in R programming language.

We used the aforementioned measures for the validation of clustering on the time series data from *Synechocystis*. Since there is no specific threshold of stability that can be set to regard a cluster analysis as valid, we made use of an empirical threshold. For this purpose, we compared the values we obtained for real data to a reference ‘meaningless’ cluster analysis. The dataset we used as reference was a synthetic dataset that consisted of only one big cluster and was the same size as the real dataset.

Furthermore, we wanted to assess the contribution of a common practice to clustering, namely filtering out stationary genes and including only the genes that show a significant change across time. For this purpose, we evaluated the stability of the clustering also on a shorter list of 2183 genes. This shorter list consists of genes which show significant time dependent change with an ANOVA p-value < 0.1. We performed repeated measurements ANOVA with a fixed factor (time) and
a random factor (bioreactor) to account for a probable bioreactor effect since we have two biological replicates at two different bioreactors in the experiment.

![Box plots showing Stability Index with and without ANOVA filtering](image)

Figure 5.4: **Stability Index in the real data with and without ANOVA filtering.** Red and black box plots show the SI obtained in 100 different splits of the real data and the single cluster synthetic dataset, respectively.

In Figures 5.4 and 5.5, we see how the cross validation (CV) based measures behave when the real data is clustered by using different numbers of clusters. In the figures, red box plots show the distribution of the measures across 100 different splits of the real data and the black boxes show the same for the single cluster synthetic data. Clustering both the filtered and the unfiltered real data by using only two clusters seems very favorable since the highest difference between the real and the single cluster data occurs at k=2. Clustering at k=3 seems very problematic since also the clustering on the single cluster data using 3 clusters shows enhanced stability. When filtering of genes by ANOVA is not carried out, we observe the same phenomenon also for the k=5. However, in general, we do not see great differences between the clustering of the filtered and the unfiltered data. This suggests that we can not show the added value of the filtering step by comparing solely the stability of the clustering.

Here we should mention a possible drawback of using CV based measures for clustering validation. To explain this, we should revisit the basic assumption behind the measures. When half of the genes are left out as test set, there have to be enough genes in the training set to form all the clusters to which the test set genes actually belong. Some of the true clusters in the data might be consisting of too few genes and possible heterogeneous distribution of these genes into the training and test sets leads to a decrease in stability. However, assuming that the ratio of the total number
5. Validation of cluster analysis

Figure 5.5: **Prediction Strength in the real data with and without ANOVA filtering.** Red and black box plots show the PS obtained in 100 different splits of the real data and the single cluster synthetic dataset, respectively.

...of genes to the true number of clusters in the data is fairly high and therefore, there are no clusters which are critically small, we accept these measures as a good means of detecting stability.

Figure 5.6 shows the stability of the clustering on the real data assessed by bootstrapping using different number of clusters. The black color in the figure indicates the stability of clustering performed on single cluster data at 1.6 % noise level. The noise level in the single cluster synthetic data is comparable to the noise in the real data (around 1.5 %) which can be calculated by using the biological replicates. Therefore, the behavior of stability obtained by bootstrapping on the two datasets can be compared.

Our results indicate enhanced stability of clustering on the real data compared to the single cluster synthetic dataset, for a variety of possible cluster numbers in the data. All three measures agree on the cluster number that gives the most stable clustering which is k=2, showing great similarity with the highly overlapping synthetic dataset (see Figures 5.1 and 5.2). For our real dataset, we apply filtering by ANOVA and perform 10-means clustering. The results from our stability assessment indicate enhanced stability at this setting, compared to single cluster data. Having validated our cluster analysis based on stability measures, we perform functional enrichment analysis on the resulting clusters.
Figure 5.6: **Stability assessed by bootstrapping in the real data with and without ANOVA filtering.** Red and black box plots show the ARI obtained in 99 different comparisons between the real data and the bootstrap samples and between the single cluster synthetic dataset and its bootstrap samples, respectively.

### 5.3.3. Biological results

We used the elbow method complemented with biological knowledge to determine the optimal number of clusters in the data. According to this method, the decrease in the variance within the clusters with increasing number of clusters is not sharp anymore, after a certain threshold which is the actual number of clusters in the data. For our real dataset, 6 to 10 clusters were reasonable to pick (data not shown) since no sharp elbow could be observed. Functional enrichment analysis of 10 clusters results in more significant enrichment of gene function categories compared to lower number of clusters. Therefore, we performed k-means clustering on the ANOVA filtered set of genes, with \( k = 10 \). The results from the functional enrichment analysis on the ten clusters can be seen in Table 5.2.

The resulting ten clusters allow, by visual inspection, a rough discrimination into four main patterns, (i) genes that are up-regulated after the shift into the dark phase (clusters 1, 5, and 6), (ii) genes that are down-regulated in the dark phase (clusters 2, 3, 4, 9, and 10), and (iii) genes that are up-regulated in the dark phase, however show an up-regulation already before the shift at TP3 (cluster 8) and (iv) genes that are down-regulated in the dark phase, however are so already before TP3 (cluster 7). Closer inspection allows to refine the coarse patterns, e.g. discriminating clusters 1 and 6 from cluster 5 based on the response time of the respective genes. Similarly, light phase clusters (those with genes that are down-regulated in the dark phase) can be discriminated by the response time after the shift from the dark phase to the light phase.
Cluster 1 which shows a steep increase between the last two time points after a very stable temporal profile during the light phase, is enriched in transposon related functions (enrichment p-value < 1e−16). As hypothesized already by Labiosa et al. [91] transposon activity during the dark phase, when cells do not grow and there is no DNA damaging light, might be beneficial to avoid the rapid loss of a new genotype through selection during growth. It has been shown before that DNA replication and cell division in cyanobacteria are separated over the course of a circadian cycle, placing genome replication mainly into the dark phase [114]. Cluster 6 shows an even more uniform behavior with respect to stable expression in the light phase followed by an increase in the dark phase. It is enriched with genes that are involved in protein degradation. Another strong enrichment, found in cluster 4, shows genes coding for the ribosomal proteins (p-value < 1e−15). Consequently, this is accompanied for an enrichment of the biological function one level higher, translation. This cluster shows an increase between the first two time points and a sharp decrease between the last two time points, indicating enhanced translational activity during the day. In addition, cluster 4 is enriched in the phycobilisome and photosystem I categories, both associated with photosynthesis, light energy acquisition, and needed for growth. This is consistent also with the growth rate observations from a parallel physiological study [7]. Cluster 3 also shows a similar
behavior as cluster 4. This cluster is enriched in respiratory terminal oxidases. *Synechocystis* thylakoid-localized terminal oxidases have recently been shown to be essential if cultures are treated with high light in square-wave cycles similar to the light regime used here [95]. Cluster 10 shows transcripts up-regulated in the light. However, the genes in this cluster reach their peak expression already at the second time point and later, slightly decrease during the day. The most significant category associated with this cluster is photosystem II (p-value < 1e − 16). Photosystem I, phycobilisome and ATP synthase are similar subcategories that are also associated with this cluster.

Table 5.2: **Functional enrichment analysis results**

<table>
<thead>
<tr>
<th>Cluster No</th>
<th>Gene function category</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>O:Other categories</td>
<td>&lt;1e-16</td>
</tr>
<tr>
<td></td>
<td>O8:Transposon-related functions</td>
<td>&lt;1e-16</td>
</tr>
<tr>
<td>2</td>
<td>H5:NADH dehydrogenase</td>
<td>2.68e-4</td>
</tr>
<tr>
<td></td>
<td>M:Translation</td>
<td>6.64e-4</td>
</tr>
<tr>
<td>3</td>
<td>H10:Respiratory terminal oxidases</td>
<td>4.99e-4</td>
</tr>
<tr>
<td>4</td>
<td>M5:Ribosomal proteins: synthesis and modification</td>
<td>7.77e-16</td>
</tr>
<tr>
<td></td>
<td>M:Translation</td>
<td>8.45e-11</td>
</tr>
<tr>
<td></td>
<td>H8:Phycobilisome</td>
<td>2.10e-6</td>
</tr>
<tr>
<td></td>
<td>H:Photosynthesis and respiration</td>
<td>4.05e-4</td>
</tr>
<tr>
<td></td>
<td>H6:Photosystem I</td>
<td>4.21e-4</td>
</tr>
<tr>
<td>6</td>
<td>P:Hypothetical</td>
<td>1.16e-4</td>
</tr>
<tr>
<td></td>
<td>M2:Degradation of proteins, peptides, and glycopeptides</td>
<td>8.86e-4</td>
</tr>
<tr>
<td>8</td>
<td>F10:Pentose phosphate pathway</td>
<td>2.02e-4</td>
</tr>
<tr>
<td>9</td>
<td>C2:Murein sacculus and peptidoglycan</td>
<td>1.26e-6</td>
</tr>
<tr>
<td></td>
<td>C:Cell envelope</td>
<td>3.42e-6</td>
</tr>
<tr>
<td></td>
<td>L2:RNA synthesis, modification, and DNA transcrip</td>
<td>3.99e-5</td>
</tr>
<tr>
<td></td>
<td>L:Transcription</td>
<td>6.99e-4</td>
</tr>
<tr>
<td>10</td>
<td>H:Photosynthesis and respiration</td>
<td>&lt;1e-16</td>
</tr>
<tr>
<td></td>
<td>H7:Photosystem II</td>
<td>&lt;1e-16</td>
</tr>
<tr>
<td></td>
<td>H6:Photosystem I</td>
<td>1.02e-6</td>
</tr>
<tr>
<td></td>
<td>H8:Phycobilisome</td>
<td>3.87e-5</td>
</tr>
<tr>
<td></td>
<td>H1:ATP synthase</td>
<td>1.29e-4</td>
</tr>
</tbody>
</table>

Another subcategory of photosynthesis and respiration that can be linked to our clusters is the NADH dehydrogenase term in cluster 2 which show a decrease during the transition to the dark phase. Cluster 2 differs from 4 and 10 in the fact that here the genes are already up-regulated at TP1. This either means that NADH dehydrogenase is up-regulated very fast (within the first 1 hour of the light phase), or has been up-regulated before, in the dark-phase which would suggest an anticipatory regulation. Either way it seems necessary as soon as there is light. *Synechocystis* NADH dehydrogenases are large and dynamic protein complexes in-
volved in the prime function of electron transport, including them in the adaptation to input-changes subsequently responsible for efficient energy conversion. Thus, their functionality in electron distribution from the transport chain [14] has to react immediately to the availability and intensity of light. Cluster 9 also shows, next to an increase from TP1 to TP2, a similar level at TP3, a very sharp decrease during the transition to the dark phase (TPs 4 and 5). This cluster is enriched with genes from the murein sacculus and peptidoglycan category, a machinery that is needed for synthesis and remodeling of cell envelope. Additionally genes involved in RNA synthesis, RNA modification, and DNA transcription are enriched, again a machinery needed for growth. The pattern of genes grouped in cluster 8 is similar to the clusters 3, 4 and 9, with the distinct difference of continuous increase during the light phase. This indicates a regulation through light availability.

5.3.4. Notes on the determination of the optimal number of clusters

The cross validation (CV) based stability measures were originally proposed for optimizing the parameters of a cluster analysis such as the number of clusters to be selected. According to [92], the cluster number that gives the highest stability is optimal. However our results indicate that the stability can be equally high for lower cluster numbers. This bias towards low number of clusters makes it difficult to determine the optimal number of clusters when the complexity of the data is increased. We have shown this by using synthetic data with highly overlapping clusters and co-varying variables mimicking real gene expression data.

The problem explained above shows itself also for our real dataset from *Synechocystis*. We would have chosen $k=2$ for this dataset as the optimal number of clusters using CV based measures. The two clusters in the data correspond to the groups of genes that increase or decrease during the night time points. However, that provides a very coarse cluster analysis. Meaningful biological interpretation of the clusters require finer clustering of the data, thus more compact clusters. Therefore, we conclude that the optimal number of clusters should not be determined solely based on the stability measures. Biological information should be incorporated for this task, considering a trade-off between stability and biological interpretability of the clustering performed.

5.4. Conclusions

CV and bootstrapping based approaches provide good means to detect stability of the cluster analysis against small perturbations in the data. However, a reference
point is needed to arrive at strict conclusions about the validity of the analysis. Comparison of the stability measures from real data to the values obtained on ‘meaningless’ clustering as a reference point seems promising. It gives an idea of how these measures behave on datasets which actually do not have more than one cluster. However, a more detailed and dedicated study has to be performed to propose established reference points for more proper validation.

Acknowledgements
We are grateful to Klaas J. Hellingwerf (University of Amsterdam) for discussions regarding the biological interpretation of the results and Sandra Waaijenborg for performing the normalization of the microarray data. We thank also Robert Lehmann (Humboldt University Berlin) and Ilka Axmann (University of Düsseldorf) for sharing the R code for the LOS normalization and the members of the RNA Biology & Applied Bioinformatics group at the SILS for performing the microarray experiments.