

# Using repeated measurements to validate hierarchical gene clusters

Laurent BRÉHÉLIN<sup>a,\*</sup> Olivier GASCUEL<sup>a</sup> and Olivier MARTIN<sup>c,b,\*</sup>

<sup>a</sup> Méthodes et Algorithmes pour la Bioinformatique, LIRMM, CNRS - Univ. Montpellier II, France,

<sup>b</sup> INRA, Unité protéomique, 2 Place Viala, 34060 Montpellier Cédex 1, France, <sup>c</sup> INRA, Unité Biostatistique et Processus Spatiaux, 84914 Avignon Cédex 9, France

\* Both authors contributed equally to this work

## ABSTRACT

**Motivation:** Hierarchical clustering is a common approach to study protein and gene expression data. This unsupervised technique is used to find clusters of genes or proteins which are expressed in a coordinated manner across a set of conditions. Because of both the biological and technical variability, experimental repetitions are generally performed. In this work, we propose an approach to evaluate the stability of clusters derived from hierarchical clustering by taking repeated measurements into account.

**Results:** The method is based on the bootstrap technique that is used to obtain pseudo-hierarchies of genes from resampled datasets. Based on a fast dynamic programming algorithm, we compare the original hierarchy to the pseudo-hierarchies and assess the stability of the original gene clusters. Then a shuffling procedure can be used to assess the significance of the cluster stabilities. Our approach is illustrated on simulated data and on two microarray datasets. Compared to the standard hierarchical clustering methodology, it allows to point out the dubious and stable clusters, and thus avoids misleading interpretations.

**Availability:** The programs were developed in C and R languages. Supplementary Material and source code are available at address <http://www.lirmm.fr/~brehelin/Stability/>

**Contact:** brehelin@lirmm.fr, gascuel@lirmm.fr, olivier.martin@avignon.inra.fr

## 1 INTRODUCTION

The development of technologies to analyze transcriptomic and proteomic data has brought new perspectives in molecular biology. These technologies have also raised many challenging problems in experimental design and data analysis. One of the important drawbacks of these approaches is the experimental variability that can be divided into two categories: technical variability and biological variability. Technical variability is inherent to techniques used to quantify the transcriptome or the proteome. Biological variability corresponds to the variability which naturally exists between different individuals. In both cases, variability is a significant problem when biological mechanisms or processes are involved. To address this problem, experimental repetitions (technical or biological) are performed through experimental design.

Because of the number of genes or proteins and the complexity of genetic networks, clustering approaches have proven useful

to analyze expression profiles. Hierarchical clustering [6], self-organizing maps [17], K-means [18] and mixture models [21] are among the most used methods. Some authors proposed to combine several clustering methods to define consensus clustering [15]. Although these approaches have proven valuable in gene expression analysis, most studies do not consider variations in measured expression levels. The standard approach involves averaging repetitions for each gene or protein, and for each experimental condition. These averages are treated as if they were accurate measures of true expression levels and the measurement effects are neglected. This approach would be warranted if the number of repetitions for each measurement was sufficiently high to allow for reliable estimation of the expression level. In practice, the measurement cost does not allow for a high repetition number, which is usually limited to 3 or 4. In this case, and when variability is significant, the average is a poor estimate of the expression level and it is not apparent how these measurement variations might affect clustering. A better exploitation of the information brought by repetitions seems essential. Yeung et al. [23] evaluated several clustering algorithms that incorporate repeated measurements and showed that algorithms that take advantage of repeated measurements yield more accurate and more stable clusters.

Two approaches can be proposed to more suitably account for the experimental repetitions:

- The first one involves directly using the repetitions during the clustering procedure. For each variable, a gene is defined by its set of experimental repetitions. However, this approach remains difficult and a few studies followed this direction [2, 14].
- The second approach involves evaluating the effect on clustering induced by poor estimation of the expression level resulting from the average of repetitions.

In this paper, we explore this second approach by assessing the stability of clusters derived from hierarchical clustering using a bootstrap procedure.

Several papers use stability or bootstrap to deal with the optimal number of clusters. In [4] and [12], the authors propose a stability criterion based on supervised classification. In [4], a resampling based prediction method estimates the number of clusters by repeatedly and randomly dividing the original dataset into two non-overlapping sets. In [12], a stability measure is introduced

for supervised learning and is generalized to semi-supervised and unsupervised clustering. Yeung *et al.* [22] apply a clustering algorithm to all but one experimental condition in a dataset, and use the left-out condition to assess the predictive power of the clustering algorithm. Herrero *et al.* [7] propose a divisive hierarchical clustering algorithm, called SOTA, which stops tree growing thanks to a shuffling-based approach.

Several factors influence cluster stability: size of the cluster vs. total number of genes, proximity of the genes in the cluster vs. their distance to the other genes, variability of the repetitions for a given gene. For these reasons, clusters do not all have the same stability. Thus, in an overall unstable clustering, some stable clusters can nevertheless exist and can be interesting for further analysis. On the contrary, clusters of an overall stable clustering do not all have the same quality, and it could be wise to be wary of some of them. The approach we propose does not focus on the number of clusters. It is designed to identify stable gene clusters within a hierarchical clustering.

Some recent papers have addressed a similar problem, but in the inverse context, that is, classifying samples (e.g. tumors) using gene expression measurements. For example, Smolkin *et al.* [16] and Valentini [19] use perturbations based on space-dimension reduction. However, this elegant approach is of no use when the space dimension is reduced to a (few) dozen(s) or even less, as in the case here (e.g. 6 in our experiments, see below). In McShane *et al.* [13], data perturbations are achieved by adding independent normal errors to the original data, with the variance of these errors being equal to the variance of the experimental data.

Moreover, two approaches have been proposed to assess cluster validity in the context of gene classification. Zhang *et al.* [24] introduce a parametric bootstrap approach to assess the reliability of gene clusters identified by hierarchical methods. Kerr *et al.* [11] proposes a "residual bootstrapping" approach that utilizes an analysis of variance model. The ANOVA model provides an estimate of the relative expression of the genes. In addition, residuals from the fitted model provide an empirical estimate of the error distribution, which is used in a sampling procedure to create new datasets. The above two references [24, 11] were designed at a time where replicated microarray experiments were rare. They thus use various assumptions and models to estimate the error (e.g. ANOVA, gene independence, homoscedasticity) and simulate new datasets. In this paper, we propose a non-parametric bootstrap approach, which uses the experimental repetitions to perturb the data without any error distribution assumption.

## 2 METHOD

Briefly, our method is as follows. First, an original hierarchical clustering is computed in the standard way by using the average of experimental repetitions for each gene. Next, the method involves disturbing the data by resampling the repetitions of each measurement with a bootstrap procedure; one again computes the averages using the bootstrap samples, and carries out a new hierarchical clustering that is compared with the original one, using a natural stability criterion and a fast dynamic programming algorithm. Repeating this procedure a number of times enables us to evaluate the stability of each cluster of the original hierarchy. The general idea is that if resampling disturbance substantially changes the elements of a cluster, it seems risky to take this one into account

for further analysis. On the contrary, if this cluster is identified in the new hierarchy with only small differences and in spite of disturbances, this means that the approximation made by the average of the repetitions does not have a significant impact on this cluster and hence that it can be selected.

In the following, we assume that clustering is performed on  $N$  genes or proteins measured for  $T$  biological variables (e.g. a kinetic with  $T$  time points). It is assumed that  $R$  repetitions of the  $N$  measurements are carried out for each of the  $T$  biological conditions (our procedure is easily extended to the case where the number of repetitions varies among conditions). The dataset is denoted as  $\mathcal{D}(N, T, R)$ .

### 2.1 Stability criterion

We define  $\mathcal{T}_0$  as the original hierarchical clustering, obtained by averaging repetitions for each gene and for each condition. Typically, we use the Euclidean distance to estimate the similarity between gene expression profiles, and infer the hierarchy using the Ward algorithm [20]. We use this approach in the experiments described below, but our method is independent of these choices and could be used with other components (e.g. with linear correlation coefficient and average linkage algorithm). Then, for each condition,  $R$  samplings with replacement are carried out in the  $R$  experimental repetitions. We thus obtain a new dataset, denoted as  $\mathcal{D}_1(N, T, R)$ . Repetitions of this new dataset are averaged and a new clustering  $\mathcal{T}_1$  is computed. This procedure complies with standard bootstrap theory; the  $R$  pseudo-repetitions provide a fair (asymptotically unbiased) view of the variability within the  $R$  original measures for each condition [5].

The stability criterion we define aims to compare the two clusterings  $\mathcal{T}_0$  and  $\mathcal{T}_1$ . A hierarchical clustering tree of  $N$  genes involves a total of  $(2N - 1)$  nodes. For each node  $i = 1, \dots, (2N - 1)$  of  $\mathcal{T}_0$ ,  $\mathcal{T}_0(i)$  denotes the cluster (set of genes) associated with  $i$ , and  $|\mathcal{T}_0(i)|$  is the cardinal number of this cluster. As there is a one-to-one correspondence between the nodes and clusters defined by a hierarchy, we shall use both terms indifferently, depending on the context. The same holds for the genes and the tree leaves. For each node  $i$  of  $\mathcal{T}_0$ , we use the following score function derived from the Jaccard index:

$$\mathcal{S}(i, \mathcal{T}_1) = \max_{j \in \{1, \dots, 2N-1\}} \frac{|\mathcal{T}_0(i) \cap \mathcal{T}_1(j)|}{|\mathcal{T}_0(i) \cup \mathcal{T}_1(j)|}. \quad (1)$$

We obviously have  $0 < \mathcal{S}(i, \mathcal{T}_1) \leq 1$  for each node  $i$  of  $\mathcal{T}_0$ . The score is equal to 1 when there is a node  $j$  of  $\mathcal{T}_1$  that covers the same genes as those covered by  $i$  in  $\mathcal{T}_0$ , that is  $\mathcal{T}_0(i) = \mathcal{T}_1(j)$ . The criterion tends to 0 when the number of genes in common tends to 0.

To obtain reliable stability estimates, this procedure is repeated  $B$  times (typically  $B = 30$ ).  $B$  datasets  $\mathcal{D}_1(N, T, R), \dots, \mathcal{D}_B(N, T, R)$  are generated from the original dataset with the previously described sampling procedure, and  $B$  scores are computed for each node  $i$  of  $\mathcal{T}_0$ . Let  $\mathcal{S}(i, \mathcal{T}_1), \dots, \mathcal{S}(i, \mathcal{T}_B)$  be the  $B$  scores associated with node  $i$ . We define the stability criterion for node  $i$  of tree  $\mathcal{T}_0$  as the average of the scores for the different resampled datasets:

$$\mathcal{S}(i) = \frac{1}{B} \sum_{b=1}^B \mathcal{S}(i, \mathcal{T}_b). \quad (2)$$

Given the properties of the score function  $\mathcal{S}(i, \cdot)$ , we have  $0 < \mathcal{S}(i) \leq 1$  for every node  $i$ .

Our method has several relevant features:

- The bootstrap sampling is structured: one independent sampling is done for each of the  $T$  variables. This preserves the correlation structure between genes, and simulates the variability that should be obtained when performing new experimental measurements.
- Despite the small number of repetitions, the number of datasets that can be obtained by bootstrap is  $R^{R \times T}$ . Thus, even if  $R$  and  $T$  are relatively low, our bootstrap procedure samples from a very large population of pseudo-datasets that mimic the variability of the original data.
- In contrast with previous works [11, 24] on gene cluster validation, the proposed bootstrap procedure is not based on a statistical model with assumptions such as homoscedasticity (next relaxed in ref.[10]), or gene independence.
- The use of the Jaccard index instead of the classical stability measure used in phylogenetic studies (and also proposed in ref. [11, 24]). In phylogenetics, the stability of a cluster is measured by the "bootstrap proportion", that is, the proportion of times this cluster is exactly found in the bootstrap samples [9]. With (highly variable) gene expression data this approach does not give satisfactory results, because most of the clusters get stability around 0. This fact is illustrated in the Experiments below.

## 2.2 Computing the stability criterion

Our approach requires to calculate, for each node  $i$  of  $\mathcal{T}_0$  and each node  $j$  of  $\mathcal{T}_b$ , the value of the ratio  $\frac{|\mathcal{T}_0(i) \cap \mathcal{T}_b(j)|}{|\mathcal{T}_0(i) \cup \mathcal{T}_b(j)|}$ . Computing the cardinal number of the intersection/union of two sets involves a number of operations linear into the sum of their cardinal numbers. Thus, the computation of the intersection/union cardinal of node  $i$  with every node  $j$  takes  $O(N^2)$  operations, which leads to a total time complexity of  $O(N^3)$  to compute Expression (1) for all nodes  $i$ . As these computations are carried out at each iteration of the bootstrap procedure, a more efficient algorithm is required. Fortunately, this is allowed by the tree structure of hierarchical clustering. We use a dynamic programming approach to compute, for each cluster  $i$  of  $\mathcal{T}_0$  and each cluster  $j$  of  $\mathcal{T}_b$ , the value of variables  $I_{ij}$  and  $C_{ij}$ , which represents the number of genes that are in  $i$  and in  $j$  (i.e.  $|\mathcal{T}_0(i) \cap \mathcal{T}_b(j)|$ ), and the number of genes which are in  $j$  and not in  $i$  (i.e.  $|\mathcal{T}_b(j) - \mathcal{T}_0(i)|$ ), respectively. At the end of the algorithm, one computes the value of  $\frac{|\mathcal{T}_0(i) \cap \mathcal{T}_b(j)|}{|\mathcal{T}_0(i) \cup \mathcal{T}_b(j)|}$  using equation

$$\frac{|\mathcal{T}_0(i) \cap \mathcal{T}_b(j)|}{|\mathcal{T}_0(i) \cup \mathcal{T}_b(j)|} = \frac{I_{ij}}{|\mathcal{T}_0(i)| + C_{ij}}. \quad (3)$$

The algorithm used to compute the  $I_{ij}$  and  $C_{ij}$  values is as follows. First, the values (0 or 1) associated with each leaf  $j$  of  $\mathcal{T}_b$  and each node  $i$  of  $\mathcal{T}_0$  are computed. This is done by a post-order traversal of  $\mathcal{T}_0$ : for each leaf  $i$  of  $\mathcal{T}_0$ , the pair  $(I_{ij}, C_{ij})$  is (1,0) or (0,1) depending on whether  $i$  is equal to  $j$  or not (remember that the tree leaves are labelled by the genes). Then, for an internal node  $i$ , the values are obtained by applying a Boolean recurrence on the values computed on the children  $i'$  and  $i''$  of  $i$ , that is:  $I_{ij} = I_{i'j}$  OR  $I_{i''j}$  and  $C_{ij} = C_{i'j}$  AND  $C_{i''j}$  (assuming, as usual, TRUE=1 and FALSE=0). Next, a numerical recurrence is used to compute the  $I_{ij}$  and  $C_{ij}$  values associated with each internal node  $j$  of  $\mathcal{T}_b$ . This is done by a post-order traversal of  $\mathcal{T}_b$  that

uses values computed on leaves during the previous step: for every cluster  $i$  of  $\mathcal{T}_0$ , we have  $I_{ij} = I_{ij'} + I_{ij''}$  and  $C_{ij} = C_{ij'} + C_{ij''}$ , with  $j'$  and  $j''$  the child nodes of  $j$ . During this tree traversal, we also compute the maximum over  $j$  of the stability criterion (3).

In summary, we have two kinds of tree traversals: the first is performed on  $\mathcal{T}_0$  for every leaf  $j$  of  $\mathcal{T}_b$ , the second is performed on  $\mathcal{T}_b$  for every internal node of  $\mathcal{T}_0$ . The total time complexity is therefore  $O(N^2)$ . Note that this complexity is not higher than that of a hierarchical clustering and does not constitute a handicap for the bootstrap application. With Ward and average linkage algorithms (in  $O(N^2T)$ ), application of our bootstrap procedure thus requires  $O(N^2TB)$  time, which can be achieved with most datasets.

## 2.3 Effect of cluster size on the stability criterion

A high criterion value for a node of the tree indicates that the corresponding cluster is stable. Nevertheless, one issue concerns the effect of the size of the cluster on the computed stability. Very large clusters are more likely to have high stability. For example, the criterion is always equal to 1 at the root of the tree. In the same manner, small clusters also tend to have high stability (leaves of the tree have stability 1).

This can be assessed by computing the stability criterion for different cluster sizes under the hypothesis  $H_0$  that there is no structure in the data. This is achieved with the following procedure. For each gene separately, we randomly permute the  $T$  biological conditions, preserving all repetitions in each condition. Then we perform a new clustering  $\mathcal{T}_0^{(1)}$  of the permuted data, and compute for each node of  $\mathcal{T}_0^{(1)}$  the stability criterion with the above bootstrap procedure. This shuffling procedure is repeated  $S$  times (e.g.  $S = 5000$ ) and stability criterion values are stored as a function of the cluster size. In this way, we build the empirical distribution of the stability criterion for each cluster size under the  $H_0$  hypothesis. From these empirical distributions, we derive the critical values  $t_k^\alpha$ , at significance level  $\alpha$  (typically  $\alpha = 1\%$ ) and for each cluster size  $k$  ( $2 \leq k \leq N - 1$ ). Once all the different estimations have been done for each cluster size, the  $t_k^\alpha$  curve is smoothed. This aims to reduce the variability in  $t_k^\alpha$  estimation. This smoothing is achieved using a simple algorithm based on a sliding window of variable size [3]. This size is enlarged if the number of observations under  $H_0$  is too small to give a reliable estimation of  $t_k^\alpha$ . In this way, we obtain the confidence region as a function of the cluster size.

The above procedure is time consuming ( $O(N^2TBS)$ ) and hence cannot be applied in an exploratory analysis. However, as we will see in the experiments of the next section, only very small ( $< 5$  genes) and very large (hundreds of genes) clusters tend to artificially have high stability. As these clusters are generally not considered in the analysis, the shuffling procedure is not required in practice if the stability threshold used to select the clusters is sufficiently high (say  $> 0.8$ ). It can be reserved to the case where clusters with medium (e.g. in the 0.6 – 0.8 range) stability have to be analysed.

Another problem may arise for clusters made up of genes that have, by chance, low variability among repetitions for all biological variables. When this happens, these clusters appear stable (by chance). However, this may only occur for very small clusters. When the number of genes increases, it is highly unlikely to find a cluster mainly composed of genes that have (by chance) low variability for all conditions.

Component	Component proportion	Mean vector $\beta^k$	std. dev. $\sigma^k$	Class symbol
$k = 1$	1/4	(0, 2, 4, 6)	5	★
$k = 2$	1/4	(0, 2, 4, 0)	5	
$k = 3$	1/2	(8, 4, 2, 0)	30	○

**Table 1.** Parameters of the mixture model for the simulated dataset.

### 3 NUMERICAL EXPERIMENTS

In this section, numerical experiments on simulated data are reported, with  $N = 80$  genes,  $T = 4$  variables and  $R = 4$  repetitions. Any observation is denoted as  $y_{itr}$ , where  $i$ ,  $t$  and  $r$  stand for the gene, the variable and the repetition, respectively. Hierarchical clustering is achieved using Euclidean distance and Ward algorithm [20].

#### 3.1 The simulated data

To simulate observations, we use a Gaussian mixture with  $K = 3$  components. Observations for a gene  $i$  arise from one of the  $K$  components. This component is denoted  $k$ . The observations of the gene  $i$  in component  $k$  define a random vector  $\mathbf{y}_i^k = (y_{i11}^k, \dots, y_{itr}^k, \dots, y_{iTR}^k)$  of size  $TR$ . The vector  $\mathbf{y}_i^k$  verifies equation  $y_{itr}^k = \beta_t^k + \epsilon_{itr}^k$  where  $\beta^k = (\beta_1^k, \dots, \beta_T^k)$  is the mean vector for component  $k$  and  $\epsilon_{itr}^k$  is a random error such that  $\epsilon_{itr}^k \sim \mathcal{N}(0, (\sigma^k)^2)$ . Parameter values we used in the simulations are displayed in Table 1.

#### 3.2 Results using the proposed approach

Figure 1-a shows the hierarchical clustering that is achieved with one simulated dataset. Symbols at the tree leaves denote the class of the observations (see Table 1). Note that the observations are well classified, except two genes from class 3 which are assigned to class 2.

A standard approach in analyzing hierarchical clustering is to cut the long branches to define the clusters of interest. Based on this approach, 4 clusters would be distinguished, dividing class 3 (circles) into two sub-clusters. However, when looking at the stabilities, we see that the two sub-clusters of class 3 are doubtful, with stabilities around 0.6, and hence can be discarded.

We next compute the stability criterion under the  $H_0$  hypothesis using the shuffling procedure of section 2.3. In Figure 1-b, the curve defines the confidence region for  $\alpha = 1\%$ . Four points clearly appear significant (above the curve), corresponding to the 3 classes we generated, plus the union of class 1 and 2. These two classes are close (see Table 1) and are consistently grouped together.

We also note that the boundary of the confidence region markedly decreases for cluster sizes between 1 and 10, illustrating the fact that only very small clusters (less than 5 genes) can have high stability under  $H_0$ . In other words, for reasonable cluster size and sufficiently high stability threshold (say above 0.8), stability alone is sufficient to detect good clusters. However, the smaller the stability threshold, the greater the probability of selecting clusters with non-significant stability.

### 3.3 Comparison with SOTA

We analysed these simulated data with the clustering approach of Herrero *et al.*[7], implemented in the SOTA web server<sup>1</sup>. This approach uses a divisive scheme: clustering is performed from top (one cluster with all genes) to bottom (each gene is a cluster). A variability criterion is used to guide and stop tree growing. Cluster variability is measured by the pairwise comparisons of the expression profiles of the genes within the cluster. When cluster variability is low (gene profiles are all similar), cluster division is stopped and the corresponding clusters are outputted by the program. Else, divisions continue until homogeneous clusters are found. A shuffling procedure is used to estimate the confidence level of the variability criterion (called variability threshold by the program). Cluster hierarchies found by SOTA are provided in Supplementary Material for various confidence values.

By using the default value of the stopping criterion (*variability threshold equal to 90%*), twenty clusters are identified (see Supplementary Material). This is far from the three simulated clusters. Cluster 1 and 2 can be recovered by empirically decreasing the stopping criterion (see Supplementary Material, *variability threshold equal to 30%*), but in this case cluster 3 is still divided into two sub-clusters. Moreover, four genes from class 3 are assigned to class 2. With a lower value (*variability threshold equal to 20%*), cluster 1 and 2 are grouped in a single cluster and cluster 3 remains composed of two sub-clusters. So it seems impossible to recover the simulated clusters, even by empirically changing the threshold of the SOTA criterion. This comparison illustrates how variability problems may lead to wrong interpretations of the actual cluster structure, and shows the usefulness of the information brought by the repetitions.

## 4 APPLICATIONS TO TRANSCRIPTOMIC DATASETS

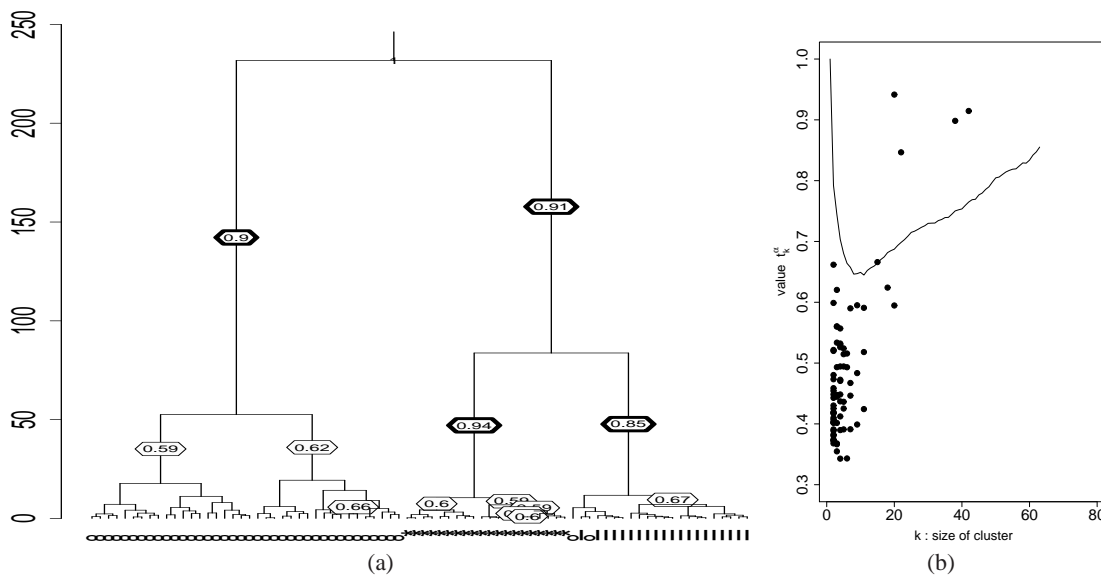
In this section, our method is applied to two transcriptomic datasets. Just as above, hierarchical clusterings are performed using the Euclidean distance and Ward algorithm [20].

### 4.1 Transcriptomic study of wood formation

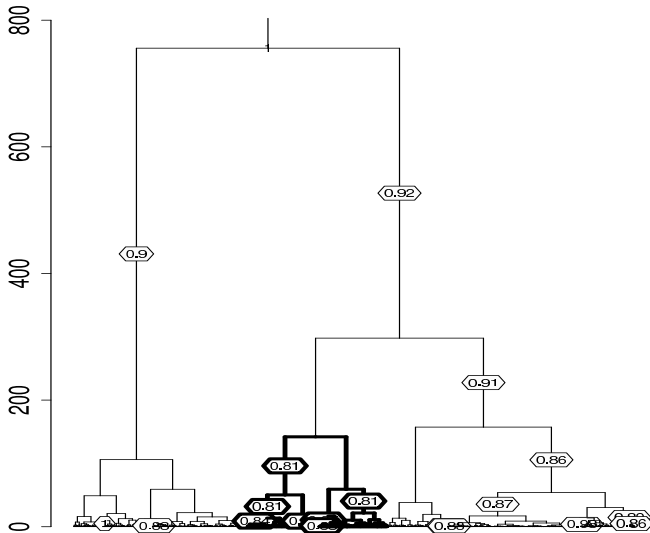
**4.1.1 Data** Hertzberg *et al.* [8] studied wood formation in poplar by analyzing the profiles of 2995 expressed sequence tags (EST) with cDNA-microarrays. The high organization of secondary xylem revealed 6 different developmental zones, which are ordered from the exterior of the trunk (phloem) to the core, and are denoted as Phl, A, B, C, D and E.  $R = 4$  repetitions were measured for each EST to compare the  $T = 6$  conditions to the control sample. The expression ratios (ratio between the condition and the control) were computed for all genes at the  $T = 6$  conditions. Only genes with a ratio greater than 2 or less than 1/2 for at least one of the  $T = 6$  conditions were selected. This procedure reduced the size of the dataset to  $N = 870$ . We clustered these genes using so-defined log-ratios.

**4.1.2 Analysis** The results of our approach are shown in Figure 2. Only clusters with high ( $> 0.8$ ) stability and more than 5 genes are indicated. A large number of nodes are selected, which indicates that this dataset is of good quality, with low experimental

<sup>1</sup> <http://bioinfo.cnio.es/sotarray>



**Fig. 1.** Analysis on simulated data. (a) Hierarchical clustering obtained using the Euclidean distance and Ward algorithm; nodes with stability greater than 0.6 are indicated on the tree; nodes with significant stability are in bold. (b) Confidence region inferred by shuffling; horizontal axis represents the cluster size, and vertical axis the stability criterion. Points above the curve correspond to clusters with significant stability ( $\alpha = 1\%$ ), and points under the curve to clusters with non-significant stability.



**Fig. 2.** Hierarchical clustering and cluster stability with the wood dataset. Clusters with high ( $> 0.8$ ) stability are indicated; the bold subtree is shown in more detail in Figure 3.

and biological variability (compared with the results of the next dataset). However, some clusters appear less stable and are likely to be less relevant (e.g. all the clusters under the left hand cluster with stability 0.9).

Let us illustrate the interest of the stability criterion by focusing on two relatively small clusters (36 and 73 genes), which cannot be divided into stable sub-clusters and have a stability of 0.84 and 0.81, respectively. Because of these features, we assume that these two clusters correspond to homogeneous groups of coregulated genes. Both are sub-clusters of the bold subtree in Figure 2, and

are shown in more detail in Figure 3-a, along with the profiles of the corresponding genes (Figure 3-b). We see from these profiles that genes in Cluster 1 are under-expressed in the phloem (PhI) and are not differentially expressed in the other developmental zones. Cluster 2 correspond to genes that are underexpressed in the most internal zone. Overall, it is easy to interpret the two selected clusters in terms of expression level in the different wood zones, which should provide a starting point for more in-depth transcriptional and functional studies.

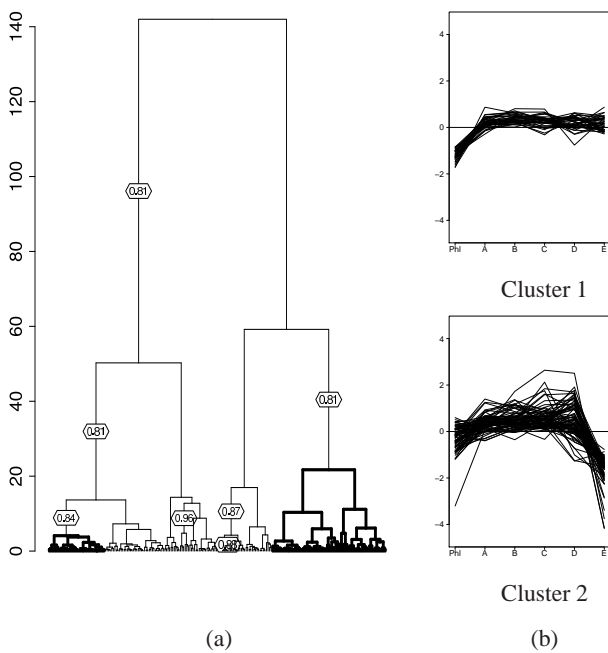
Finally, the same analysis was carried out using the classical phylogenetic bootstrap proportion (see Supplementary Material). Most of the clusters have very low stability (below 0.1), and only some very small clusters get moderate stability, which illustrates the inappropriateness of this measure when searching for stable clusters with microarray data.

## 4.2 Transcriptomic study on iron stress for *A. thaliana*

**4.2.1 Data** Data have been obtained with DNA chips to study the response of *Arabidopsis thaliana* to an iron excess. The expression levels of 24,960 probes for a kinetic of  $T = 6$  time points (5 min, 15 min, 30 min, 60 min, 6 h and 24 h) were measured. After a filtering step similar to that of the previous study,  $N = 733$  probes were selected for clustering. The number of repetitions was  $R = 4$ .

**4.2.2 Analysis** Figure 4 provides the results of our approach. Only a few clusters are stable ( $> 0.8$ ). The hierarchy in Figure 4 thus appears globally unstable and the clusters have to be considered with care. Few very small ( $< 5$  genes), one large (128 genes), and one very large ( $> 400$  genes) clusters have stability above 0.8. Thus, apart from the large cluster, it seems that this dataset cannot provide much useful information. We then searched for over-represented Gene Ontology<sup>2</sup> terms among genes within several clusters selected on the basis of their branch length (indicated by letters A-H on

<sup>2</sup> <http://www.geneontology.org/>



**Fig. 3.** Two stable clusters which cannot be divided into smaller stable clusters. (a) position of these two (bold) clusters in bold subtree of Figure 2; Cluster 1 (left) has a stability of 0.84; cluster 2 (right) has a stability of 0.81. (b) Profiles of the corresponding genes.

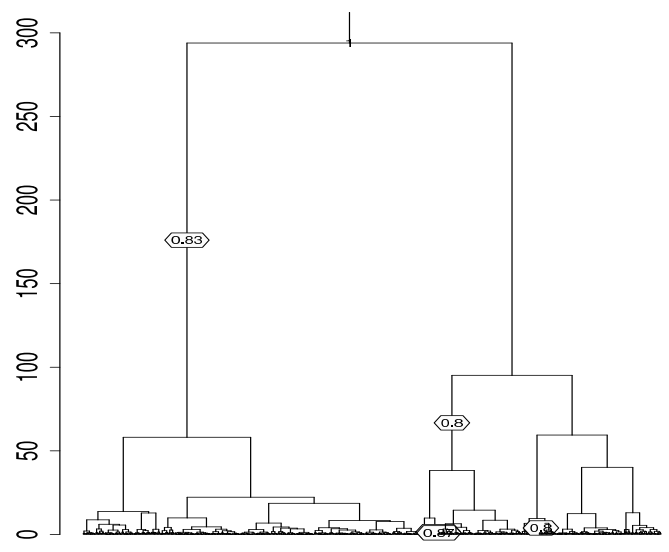
Figure 5). We used the Gostat<sup>3</sup> program of Beissbarth *et al.* [1] for this analysis, with default parameters (p-value 0.01, and Benjamini procedure for multiple testing). Interestingly, no cluster apart from cluster C (the large stable cluster) exhibit over-represented GO terms, which is in accordance with our stability analysis.

In such a case, a natural approach is to use a lower stability threshold, e.g. 0.7 instead of 0.8. However, this has to be done with care, to avoid selecting non-significant clusters. Figure 5-a shows the clusters with stability above 0.7, while Figure 5-b displays the confidence region and the stability of the different clusters. According to the latter, we see that all clusters with reasonable size ( $> 5$ ) and stability above 0.7 are actually significant, and can be considered for further analysis. A few significant clusters are thus added when dropping the stability threshold from 0.8 to 0.7: some very small and uninteresting, and one cluster with 12 genes and stability 0.77, which is included in the large cluster C with stability 0.8 discussed previously. A detailed analysis of this cluster and of cluster C is presented in Supplementary Material. Basically, results indicate that both cluster C and this sub-cluster contain over represented GO terms; C terms correspond to heat and light stresses, while sub-cluster terms correspond to heat stress only.

## 5 SOFTWARE

The programs used for these analyzes were developed in C and R (<http://cran.r-project.org>) languages. Source code is available at <http://www.lirmm.fr/~brehelin/Stability/>. The figures presented in this article were obtained directly with this code.

<sup>3</sup> <http://gostat.wehi.edu.au/>



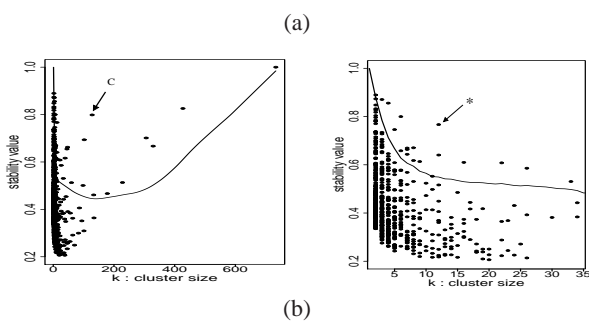
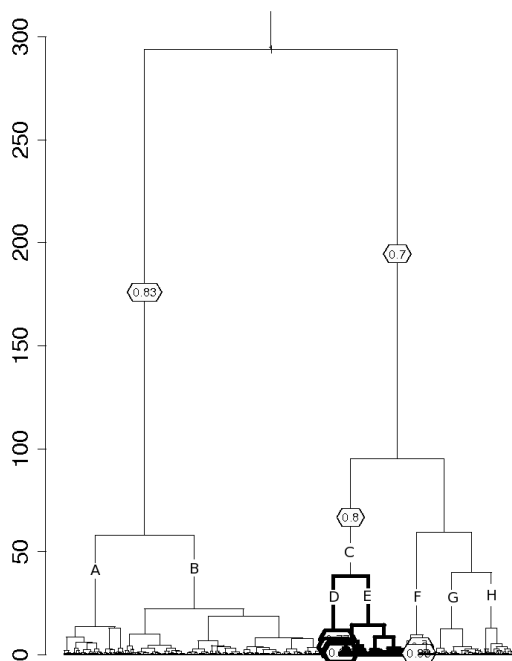
**Fig. 4.** Results with Arabidopsis transcriptomic data. Clusters with high ( $> 0.8$ ) stability are indicated.

## 6 CONCLUSION AND DISCUSSION

In this work, we have dealt with the validation of clusters derived from hierarchical clustering. In most cases, clustering is based on averages from several experimental repetitions. Such approaches are too direct due to the variability of experimental measurements. The approach we propose uses experimental repetitions, which provide information on the variability of measurements across genes and conditions. Our approach is non-parametric and based on bootstrap sampling to measure cluster stability. Moreover, data shuffling can be used to assess the significance of this measure. Experiments with simulated data show that our approach is able to recover the true cluster structure. Moreover, we illustrated its capabilities on two different situations. First, when the whole clustering is globally stable, the procedure points out the doubtful clusters. Second, when the clustering is globally unstable, it enables the discovery of some interesting clusters.

Our approach shares several features with the approaches proposed by Zhang *et al.* [24] and Kerr *et al.* [11], but also bears several differences. First, as already discussed, our method uses non-parametric bootstrap, without error modelling (e.g. ANOVA) and strong assumptions, such as the independence of the error measurements among genes. Second, the use of the Jaccard index avoids the low stabilities obtained with standard phylogenetic bootstrap proportions, due to the high variability of expression data. Finally, contrary to [11] which takes place in a general clustering framework, our approach is designed for hierarchical clustering. Our dynamic programming algorithm enables to rapidly compute the stability of all clusters of a hierarchy (e.g., with the wood experiment, less than 2 minutes are required on a standard laptop), which allows the method to be used in the context of exploratory analysis.

Concerning the p-value computation, from a practical standpoint, the shuffling procedure is too time consuming to allow an intensive and systematic application. However, considering only clusters with high stability and reasonable size allows this shuffling procedure to



**Fig. 5.** Results for Arabidopsis transcriptomic data. (a) Whole hierarchy; cluster stabilities above 0.7 are indicated. (b) Decision boundary under  $H_0$  ( $\alpha = 1\%$ ); each point corresponds to one cluster in hierarchy (a); points above the curve are significant; the right hand plot zooms on the small clusters. Arrows indicate the large cluster C and the small cluster of 12 genes (analysed in Supplementary Material).

be avoided. The use of the shuffling procedure should be reserved to cases requiring a study of clusters with moderate stabilities.

Several directions deserve further investigations. The main improvements would concern the statistical significance of our stability criterion. An efficient procedure is clearly needed, in place of our time-consuming shuffling method. Testing multiple and (hierarchically) correlated class stabilities can also be an important issue when the number of nodes selected for p-value computation is large.

## ACKNOWLEDGEMENTS

The authors thank Dr. Rossignol, Director of the INRA proteomic laboratory of Montpellier, for providing the *Arabidopsis* dataset and for his help on the biological comments. They also thank Pr. Caraux for his helpful remarks.

## REFERENCES

- [1] T. Beissbarth and T. P. Speed. Gostat: Find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, 20(9), 2004.
- [2] G. Celeux, O. Martin, and C. Lavergne. Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling*, 5:243–267, 2005.
- [3] W.S. Cleveland. Lowess: a program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, 35:54, 1981.
- [4] S. Dudoit and J. Fridlyand. A prediction-based resampling method for estimating the number of cluster in a dataset. *Genome Biology*, 3, 2002.
- [5] B. Efron and R. Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986.
- [6] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95:14863–14868, 1998.
- [7] J. Herrero, A. Valencia, and J. Dopazo. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17(2):126–136, 2001.
- [8] M. Hertzberg, H. Aspeborg, J. Schrader, A. Andersson, R. Erlandsson, K. Blomqvist, R. Bhalerao, M. Uhlen, T.T. Teeri, J. Lundeberg, B. Sundberg, P. Nilsson, and G. Sandberg. A transcriptional roadmap to wood formation. *PNAS*, 98(25):14732–14737, 2001.
- [9] J. Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39:783–791, 1985.
- [10] M.K. Kerr, C.A. Afshari, L. Bennett, P. Bushel, J. Martinez, N.J. Walker, and G.A. Churchill. Statistical analysis of a gene expression microarray experiment with replication. *Statistica Sinica*, 12:203–217, 2002.
- [11] M.K. Kerr and G.A. Churchill. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of Sciences*, 98:8961–8965, 2001.
- [12] T. Lange, M.L. Braun, V. Roth, and J.M. Buhmann. Stability based model selection. In *In Advances in Neural Information Processing Systems*, volume 15, 2003.
- [13] L.M. McShane, M.D. Radmacher, B. Freidlin, R. Yu, M.-C. Li, and R. Simon. Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics*, 18(11):1462–1469, 2002.
- [14] M. Medvedovic, K.Y. Yeung, and R.E. Bumgarner. Bayesian mixture model based clustering of replicated microarray data. *Bioinformatics*, 20(8):1222–1232, 2004.
- [15] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning*, 52(1-2):91–118, 2003.
- [16] M. Smolkin and D. Ghosh. Cluster stability scores for microarray data in cancer studies. *BMC Bioinformatics*, 4(36), 2003.
- [17] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, and T.R. Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *PNAS*, 96:2907–2912, 1999.
- [18] S. Tavazoie, J.D. Hughes, M.J. Campbell, R.J. Cho, and G.M. Church. Systematic determination of genetic network architecture. *Nature Genetic*, 22:281–285, 1999.
- [19] G. Valentini. Clusterv: a tool for assessing the reliability of clusters discovered in dna microarray data. *Bioinformatics*, 22(3):369–370, 2006.
- [20] J.H. Ward. Hierarchical clustering to optimise an objective function. *Journal of the American Statistical Association*, 58:238–244, 1963.
- [21] K.Y. Yeung, C. Fraley, A. Murua, A.E. Raftery, and W.L. Ruzzo. Model-based clustering and data transformations for gene expression data. *Bioinformatics*, 17:977–987, 2001.
- [22] K.Y. Yeung, D.R. Haynor, and W.L. Ruzzo. Validating clustering for gene expression data. *Bioinformatics*, 17:309–318, 2001.
- [23] K.Y. Yeung, M. Medvedovic, and R.E. Bumgarner. Clustering gene-expression data with repeated measurements. *Genome Biology*, 4(5):R34, 2003.
- [24] K. Zhang and H. Zhao. Assessing reliability of gene clusters from gene expression data. *Functional and Integrative Genomics*, 1:156–173, 2000.