Does clustering of DNA barcodes agree with botanical classification directly at high taxonomic levels? Trees in French Guiana as a case study

Mohamed Anwar Abouabdallah[1,2], Nathalie Peyrard[3,*], and Alain Franc [1,2]

[1]Université de Bordeaux, INRAE, BIOGECO, 33612 Cestas, France
[2]Pleiade, EPC INRIA-INRAE-CNRS, Université de Bordeaux, 33405, Talence, France
[3]Université de Toulouse, INRAE, UR MIAT, 31320, Castanet-Tolosan, France
[*]corresponding author: nathalie.peyrard@inrae.fr

1 Agglomerative Hierarchical Clustering

Let us consider the problem of grouping N individuals into K groups, based on the values of the dissimilarity $d(i,j)$ between each pair (i,j) of individuals. Agglomerative Hierarchical Clustering (AHC) is a classical method to solve this problem (Hastie et al., 2009, sec. 14.3). It is initialised with a partition of the individuals into N groups, one per individual. Then at each step, two groups are merged. The choice of these two groups is made according to an aggregation criterion ℓ (also referred to as a linkage function), which measures the similarity between two groups. If we have already built a partition into k groups $c_1,\dots,c_k$, the partition into $k-1$ groups is obtained by merging the two groups c_a, c_b such that $\ell(c_a,c_b)$ is minimal. Several linkage functions have been proposed (Murtagh, 1983; Müllner, 2013). In this study we compared three choices: Single Linkage (ℓ_{SL}), Complete Linkage (ℓ_{CL}), and Ward (ℓ_W).
With SL, the similarity between two groups is defined as the smallest dissimilarity between two individuals of each group:

$$\ell_{SL}(c_a,c_b) = \min_{i\in c_a, j\in c_b} d(i,j)$$

With CL, it is the largest one:

$$\ell_{CL}(c_a,c_b) = \max_{i\in c_a, j\in c_b} d(i,j)$$

The Ward linkage function was originally defined for Euclidean distances and is equal to the difference between the inertia of the union of c_a and c_b (i.e., after merging) and the

sum of the inertia of c_a and c_b (i.e., before merging). The Smith-Waterman dissimilarity is not a Euclidean distance, but the definition of the Ward linkage function can be extended to dissimilarities (Chavent et al., 2017).
In our study we used the R package `cluster` and the Python package `scipy.cluster.hierarchy`, both using `fastcluster` (Müllner, 2013).

2 SBM

The idea underlying AHC is to form groups of individuals that are similar. The Stochastic Block Model (SBM, Holland et al., 1983; Daudin et al., 2008; Lee and Wilkinson, 2019) corresponds to a more general point of view: it builds groups such as individuals in a given group must have the same pattern of connections to the other groups and to their own group. Consequently, individuals in the same group can be dissimilar (but almost at the same distance from each other) if they share the same pattern of dissimilarities with the other groups at the same time. This happens, for instance, when the individuals are organised into a central hub and several peripheral individuals. The group formed by the hub has a pattern of small within-dissimilarities and intermediate dissimilarity with the peripheral group, while the peripheral group has a pattern of large within-dissimilarities and intermediate dissimilarities with the hub group. SBM relies on statistical modelling and latent variables. It was originally defined for a binary dissimilarity matrix (i.e when individuals are nodes of a graph, where an edge means similarity and absence of edge means dissimilarity), but we used its extension to dissimilarity matrices in our case. The observed variable is the dissimilarity matrix and the latent variables are the group memberships of each individual: $Z_i \in \{1, \ldots, K\}$ is the group of individual i. The model relies on two assumptions. First, the Z_i's are independent and their distribution is parameterised by the vector of probabilities $\alpha = (\alpha_1, \ldots, \alpha_K)$, such that $P(Z_i = k) = \alpha_k$. Second, the dissimilarity between i and j depends only on the groups of i and j. For the Smith-Waterman dissimilarity $P(d(i,j) \mid Z_i = k, Z_j = k')$ is modeled by a Poisson distribution with parameter $\lambda_{k,k'}$. For the distance based on kmers, it is modeled by a Gaussian distribution with parameters $\mu_{k,k'}$ and σ (the variance is the same for each couple (k, k')). In our study we used R package `blockmodels` with default settings.

3 Normalised Mutual Information

Let us consider two classifications A and B of N individuals into K groups. The groups sizes are $n_1^A, \ldots, n_K^A$ for the first classification and $n_1^B, \ldots, n_K^B$ for the second one. The normalised vectors of the groups sizes are $p^A = (n_1^A/N, \ldots, n_K^A/N)$ and $p^B = (n_1^B/N, \ldots, n_K^B/N)$. Several indices exist to measure the similarity between p^A and p^B. We used a normalised version of the Mutual Information, referred to as NMI1 in Pfitzner et al. (2009), and that we refer to as NMI here. It is defined as the mutual information between A and B, $I(A, B)$, divided by the joint entropy of A and B, $H(A, B)$. The entropy $H(A)$ is

$$H(A) = -\sum_{k=1}^{K} p_k^A \log(p_k^A)$$

The joint entropy of A and B will then be

$$H(A, B) = -\sum_{k=1}^{K}\sum_{l=1}^{K} p_{kl}^{AB} \log(p_{kl}^{AB})$$

where N_{jk}^{AB} is the number of individuals that are both in class c_j^A and in class c_k^B and $p_{jk}^{AB} = N_{jk}^{AB}/N$. The mutual information between A and B is

$$I(A, B) = H(A) + H(B) - H(A, B)$$

Finally, $\mathrm{NMI}(A, B) = \frac{I(A,B)}{H(A,B)}$. It can be shown that $\mathrm{NMI}(A, B) \in [0, 1]$ where $\mathrm{NMI(A, B)} = 0$ if A and B are independent and $\mathrm{NMI}(A, B) = 1$ if A and B are identical.

4 Procedure to define thresholds for a quantitative analysis of the Normalised Mutual Information

We present here a procedure, based on simulated partitions, to define intervals of Normalised Mutual Information values corresponding to a very good, good, poor, very poor agreement between two classifications. Let us consider a partition A of a set of N individuals into K classes. We create a perturbed, or noised partition, $\tilde{A}$, by randomly reallocating a percentage p of the N individuals into a different class. Each individual is randomly chosen and its new class is also randomly chosen. We consider that if p is lower than 0.05, the agreement between A and $\tilde{A}$ is very good, between 0.05 and 0.15 it is good, between 0.15 and 0.3 it is average, between 0.3 and 0.5 it is poor and beyond 0.5 it is very poor. In order to convert these thresholds on the level of noise into thresholds on the Normalised Mutual Information scale, for each value of p (0.05, 0.15, 0.3, 0.5), we generated 5000 partitions. To generate a partition A, N and K were randomly generated according to a uniform distribution with bounds respectively $[10, 1400]$ and $[N, 55]$. The bounds correspond to the range of values of K and N observed on the subsets of the data set studied in the article. Then we generated a multinomial distribution, using the Dirichlet distribution with all parameters equal to 1, and the class of each of the N individuals was simulated according to this multinomial distribution. Then we computed $\mathrm{NMI}(A, \tilde{A})$. A violin representation of the probability density of $\mathrm{NMI}(A, \tilde{A})$ is represented on Figure 1. Each density is computed based on 5000 values. Thresholds on the Normalised Mutual Information scale are the median value of the 4 probability densities, leading to the following intervals:

NMI interval	Agreement
$]0.86, 1]$	very good
$]0.66, 0.86]$	good
$]0.45, 0.66]$	average
$]0.27, 0.45]$	poor
$[0, 0.27]$	very poor

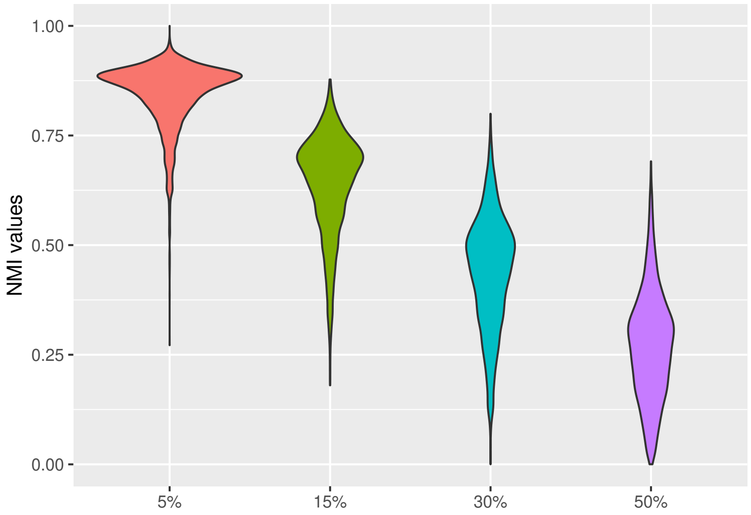

Figure 1: Distribution of Normalised Mutual Information values for increasing percentage of reallocation of individuals. The Normalised Mutual Information is computed between an original partition and a noisy one generated by randomly reallocating some individuals to another class. Each violin corresponds to a particular percentage of noise.

5 Measures of the relative differences between within group and between group dissimilarities

Let us consider a set of N sequences organised into K groups for a given taxonomic level, and let us denote the dissimilarity between two sequences i and j by $d(i, j)$. We then consider $M(k)$ the maximal dissimilarity value between two elements of group c_k, and $m(k, k')$ the minimal dissimilarity value between an element of group c_k and an element of group k':

$$M(k) = \max_{i,j \in c_k} d(i, j)$$

and

$$m(k, k') = \min_{i \in c_k,\ j \in c_{k'}} d(i, j)$$

We expect that the automatic recovery of the K groups based on the dissimilarities will be easier if the $M(k)$ values are much smaller than the $m(k, k')$ values. Therefore, based on these quantities, we defined three measures of the relative difference between within group and between group dissimilarities:

$$\begin{aligned} r_{minmax} &= \frac{\min_k M(k)}{\max_{k,k'} m(k,k')} \\ r_{maxmin} &= \frac{\max_k M(k)}{\min_{k,k'} m(k,k')} \\ r_{mean} &= \frac{\frac{1}{K}\sum_{k=1}^{K} M(k)}{\frac{K(K-1)}{2}\sum_{k=1}^{K-1}\sum_{k'=k+1}^{K} m(k,k')} \end{aligned}$$

The ratio r_{minmax} corresponds to the smallest observed ratio between $M(k)$ and $m(k,k')$, whereas the ratio r_{maxmin} corresponds to the largest one, and r_{mean} captures a mean behaviour. Intuitively when the dissimilarity314matrix is well structured into several groups each with a small within-class dissimilarity then r_{mean} will be lower than 1. On contrary, when there are no clearly delimited groups of similar individuals then r_{mean} will be larger than 1. This is illustrated on Figure 2.

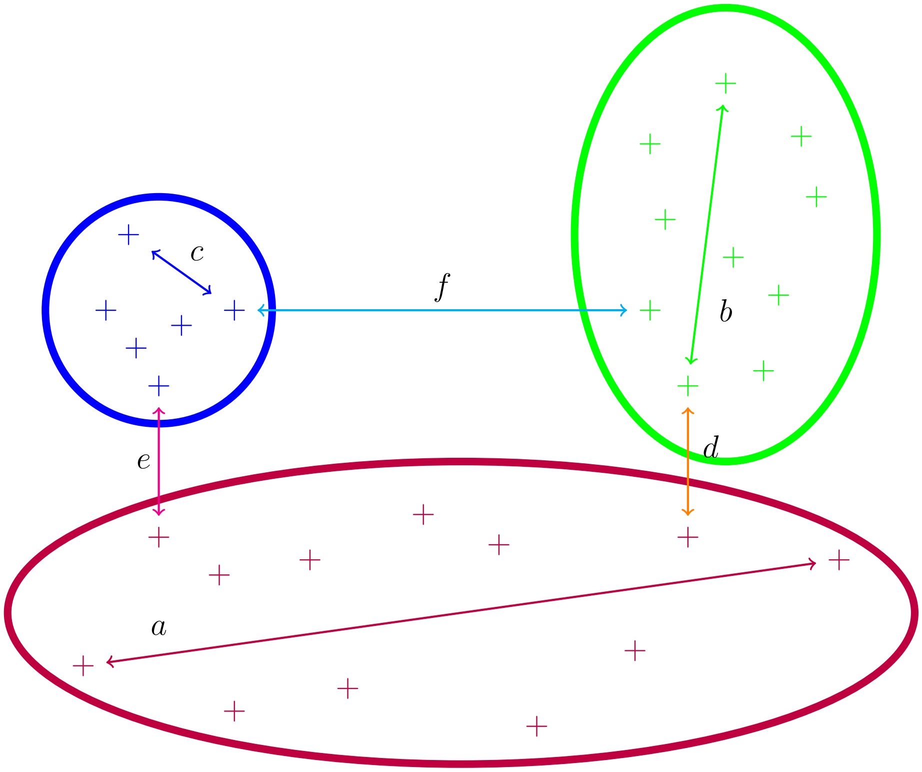

Figure 2: Illustration of a situation where r_{mean} is larger than 1. The partition is composed of 3 groups. Distances of interest to compute r_{mean} are: $M(1) = a$, $M(2) = b$, $M(3) = c$, $m(1,2) = d$, $m(1,3) = e$ and $m(2,3) = f$. Since $a + b + c$ is larger than $d + e + f$, r_{mean} is larger than 1.

We expect that the quality of the automatic recovery of the K groups increases when any of these ratios decreases. In our experiments, no pattern was found between the NMI value and r_{minmax}, probably because this ratio is a too optimistic measure of the level of structure in the dissimilarity matrix. Correlation patterns between r_{maxmin} and NMI were less obvious than with r_{mean}.

6 Neighbor-joining

We ran neighbor joining (Saitou and Nei, 1987) using `ape` library in R on the dissimilarity array between the 1387 sequences of the 11 orders with 15 sequences or more (first row of Table 1 in the manuscript). We drew the tree with a color pattern as follows: (*i*) for each order, we selected the leaves with this order as a label (*ii*) we computed the Most Recent Common Ancestor of this set of leaves (*iii*) we colored (with one color per order) all the paths between the Most Recent Common Ancestor and each leaf labelled with this order. The graphics is shown as figure 8 here in SI. We can see a couple of things: (*i*) several orders, like the Ericales (in purple) are monophyletic (*ii*) for some pairs of orders, like Laurales (red) and Magnoliales (orange), one order (here Laurales) is monophyletic and in the descent of the Most Recent Common Ancestor of the second one (which is paraphyletic) (*iii*) some others like Malpighiales (in light green) and Sapindales (in green) are paraphyletic, with a break-up into several clades with several orders in between. Hence, the adequacy between orders and clades in the tree is not excellent for the neighbor-joining tree.

7 Supplementary figures

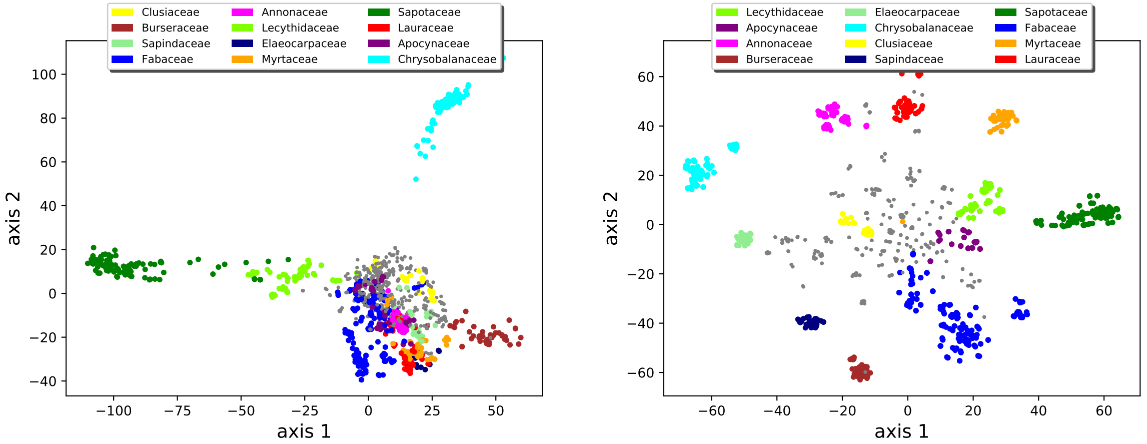

Figure 3: Visualisation of the sequences of the whole data set, as a point cloud. One dot is one sequence. Points of the twelve more numerous families are coloured while the others are in grey. Dissimilarities are computed with the Smith-Waterman algorithm. Left: MDS, projected on axis 1 and 2. Right, t-SNE.

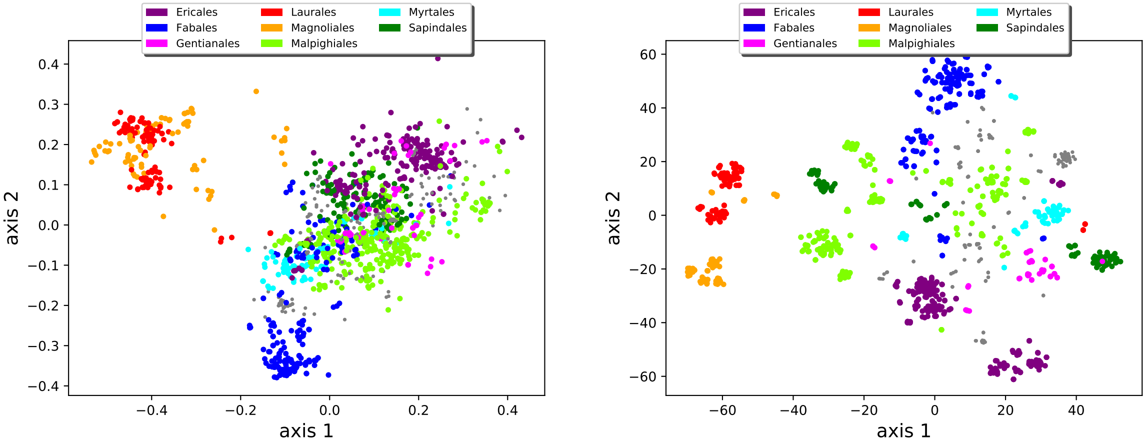

Figure 4: Visualisation of the sequences of the whole data set, as a point cloud. One dot is one sequence. The points of the eight more numerous orders are coloured, while the others are in grey. Dissimilarities are computed using 4-mers. Left: MDS, projected on axis 1 and 2. Right, t-SNE.

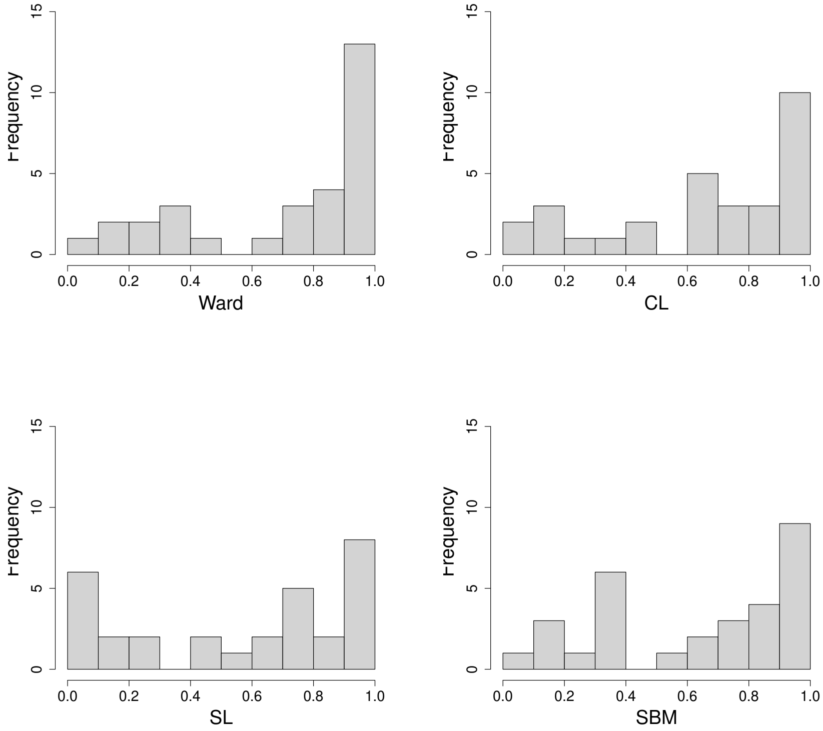

Figure 5: Histograms of the Normalised Mutual Information between each molecular-based clustering and the botanical classification, for the 30 replicates. Results obtained using the Smith-Waterman dissimilarity.

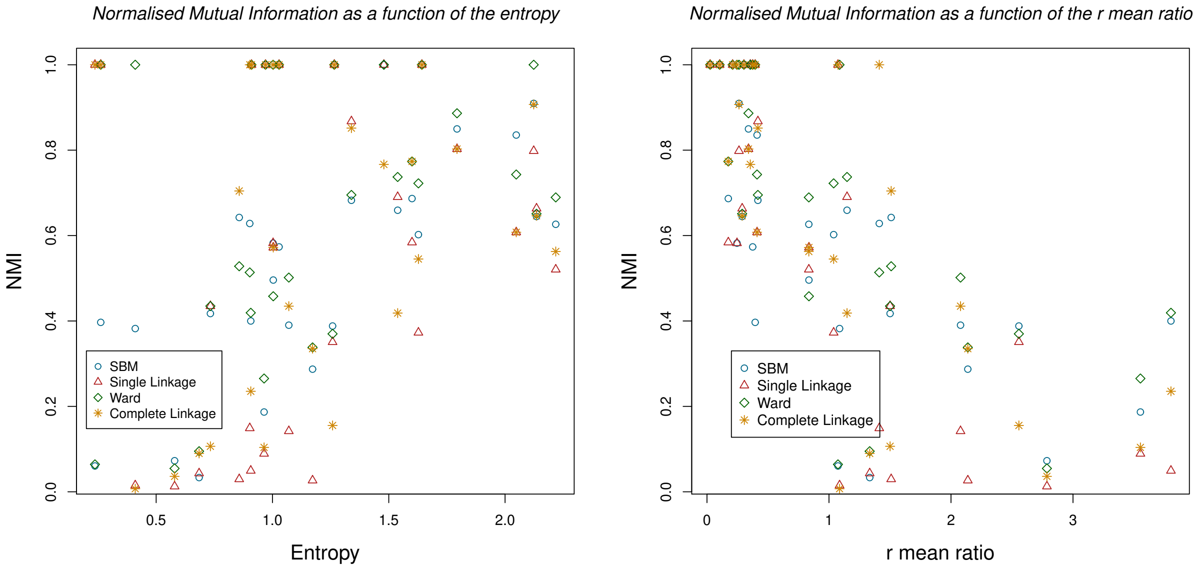

Figure 6: Values of the Normalised Mutual Information as a function of the entropy (left) and the ratio r_{mean} (right) computed of the botanical classification. Each point corresponds to one of the four molecular-based clustering methods applied to one of the 30 replicates. The x-axis is the value of the entropy or ratio r_{mean} computed on the botanical classification, the y-axis is the Normalised Mutual Information between the botanical classification and the molecular-based one. Clustering is made using the tetramer-based distances.

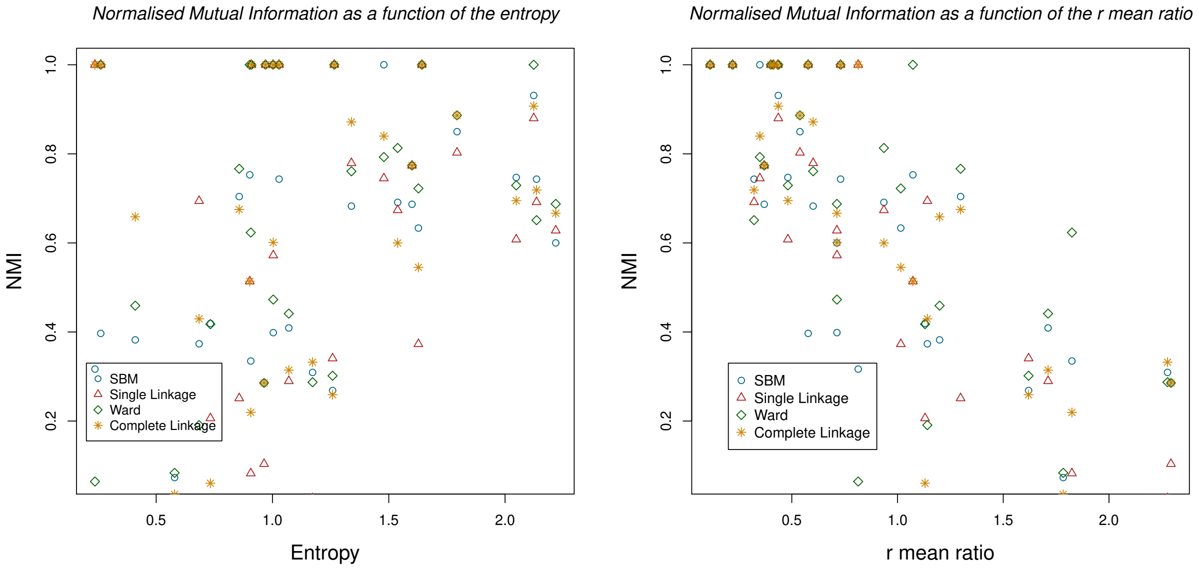

Figure 7: Values of the Normalised Mutual Information as a function of the entropy (left) and the ratio r_{mean} (right) computed of the botanical classification. Each point corresponds to one of the four molecular-based clustering methods applied to one of the 30 replicates. The x-axis is the value of the entropy or ratio r_{mean} computed on the botanical classification, the y-axis is the Normalised Mutual Information between the botanical classification and the molecular-based one. Clustering is made using the hexomer-based distances.

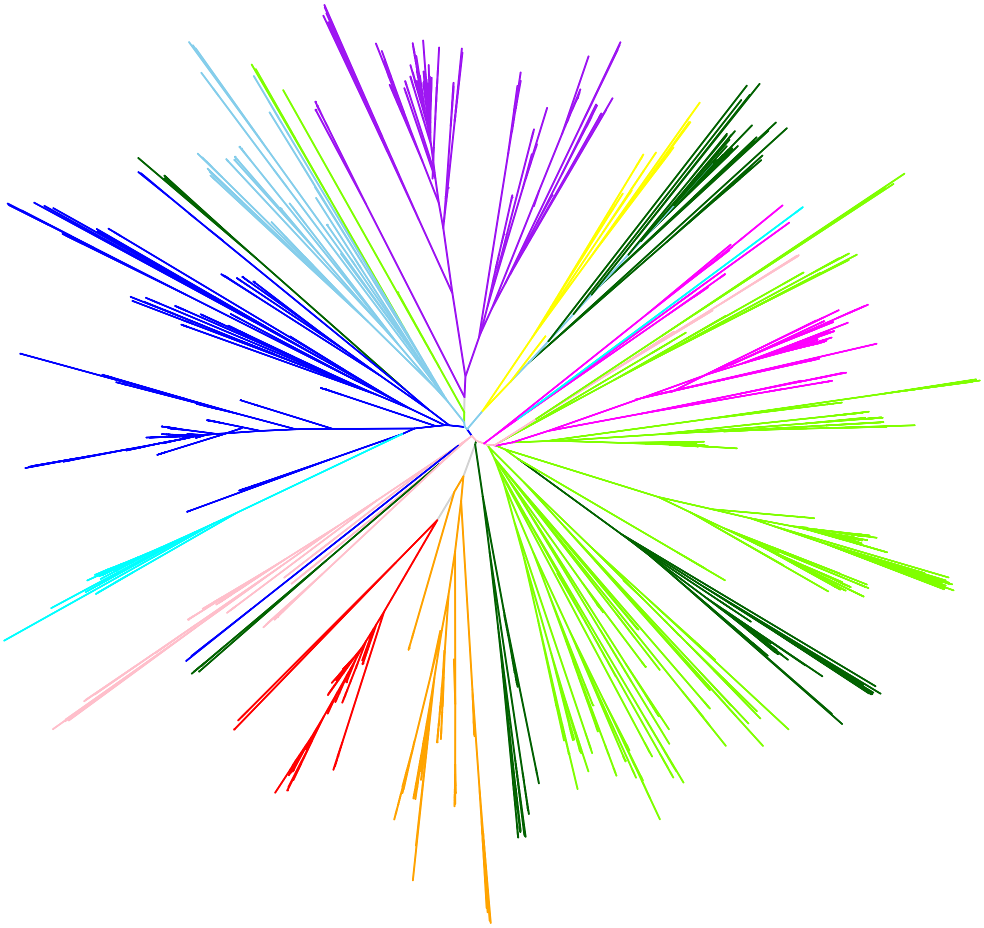

Figure 8: Unrooted phylogenetic tree built with neighbor-joining on the subset of sequences belonging to orders with 15 or more sequences. For each order, the paths joining their Most Recent Common Ancestor in the tree and each leaf labelled with this order has been colored with a same color. Ericales (purple) is a monophyletic order. Laurales (red) is a monophyletic order in the descent of Magnoliales (orange). Malphighiales (light green) and Sapindales (green) are excessively paraphyletic.

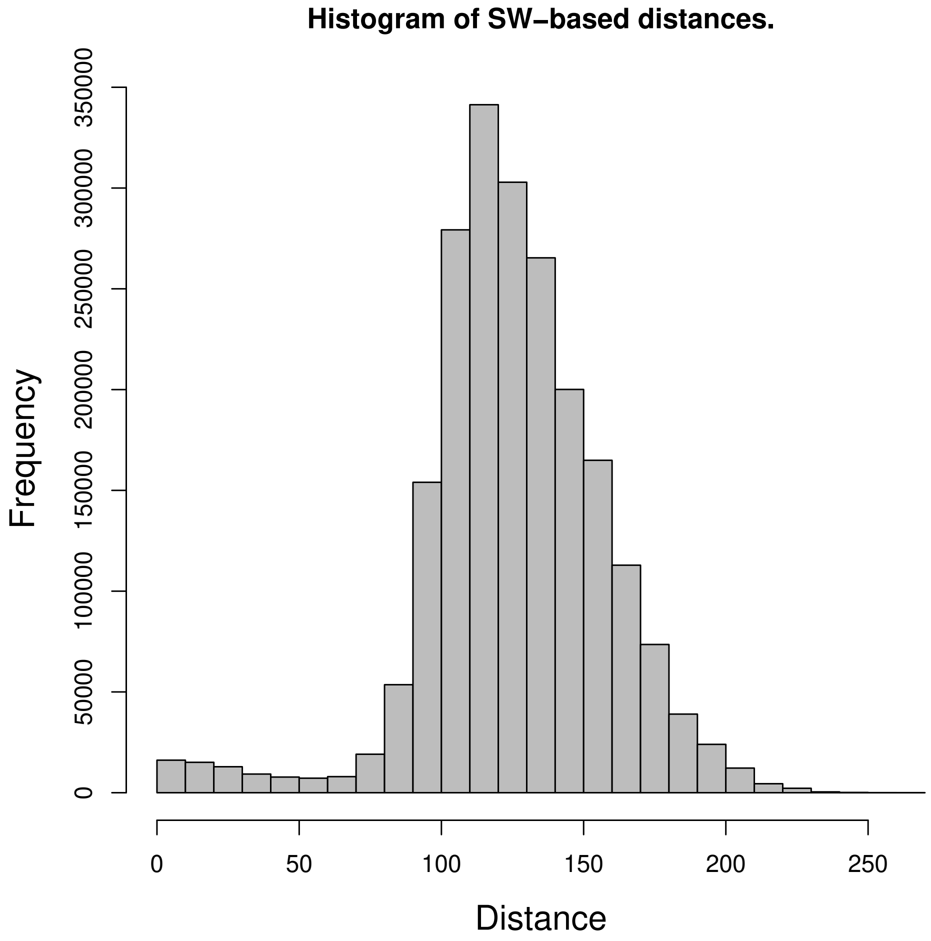

Figure 9: Histogram of Smith-Waterman dissimilarities.

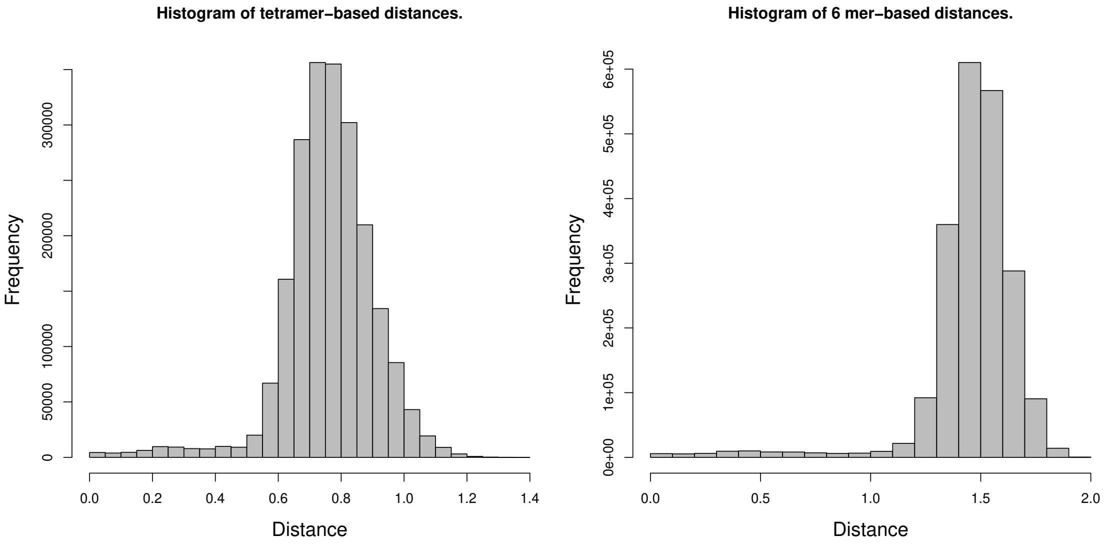

Figure 10: Histogram of 4mer-based (left) and 6mer-based (right) distances.

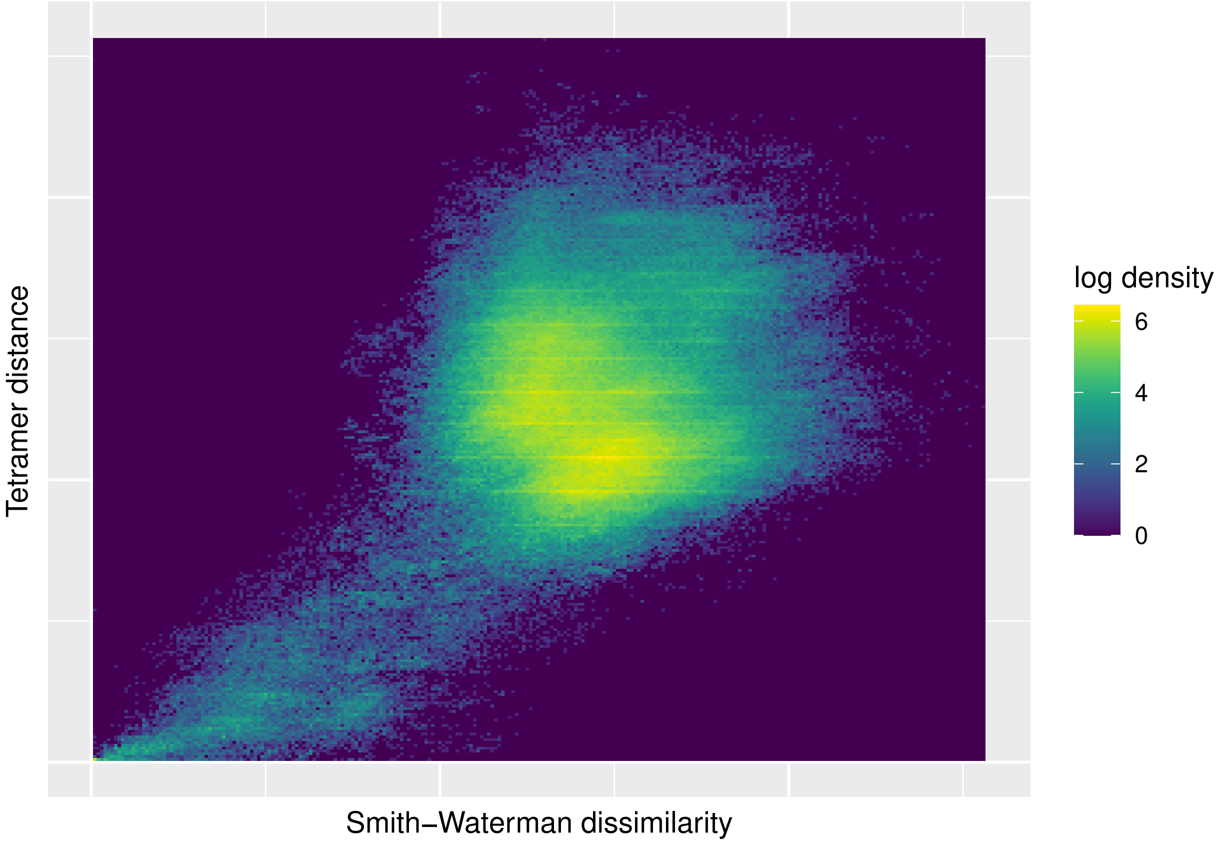

Figure 11: Comparison between Smith-Waterman and kmer-based dissimilarities (length $k = 4$). Density heatmap with logarithmic scale. x axis: kmer-based distance; y axis: Smith-Waterman dissimilarity. The color at a given pixel represents the logarithm of the number of pairs of sequences.

References

Chavent, M., Kuentz-Simonet, V., Labenne, A., and Saracco, J. (2017). Clustgeo2: an r package for hierarchical clustering with spatial constraints. *Computational Satistics*, 33(4):1–24.

Daudin, J.-J., Picard, F., and Robin, S. (2008). A mixture model for random graphs. *Statistics and Computing*, 18:173–183.

Hastie, T., Tishibani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, second edition.

Holland, P., Laskey, K., and Leinhardt, S. (1983). Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137.

Lee, C. and Wilkinson, D. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science*, 4(122).

Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359.

Müllner, D. (2013). fastcluster: Fast Hierarchical, Agglomerative Clustering Routines for R and Python. *Journal of Statistical Software*, 53(9):1–18.

Pfitzner, D., Leibbrandt, R., and Powers, D. (2009). Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems*, 19(361).

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, (4):406–425.