Simple approaches for evaluation of OTU quality based on dissimilarity arrays

Marie-Josée Cros, Jean-Marc Frigerio, Nathalie Peyrard, Alain Franc

Simple approaches for evaluation of OTU quality based on dissimilarity arrays

Marie-Josée Cros[1], Jean-Marc Frigerio[2,3], Nathalie Peyrard[1,*], and Alain Franc[2,3]

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

Running title: **OTU quality from dissimilarity arrays**

Abstract

An accurate and complete taxonomic description of the diversity present in an environmental sample is out of reach at this time. Instead metabarcoding is used today and it is expected that OTUs represent a category relevant for biodiversity inventories on a molecular basis. However artefacts in the production of OTUs can occur at different stages and may impact ecological conclusions. We propose to evaluate the quality of OTUs in a sample by characterising the deviation of each OTU's dissimilarity array from that of an ideal OTU where all sequences are at distances smaller than the barcoding gap. We consider two deviations: the creation of composed OTUs, corresponding to the artificial merging of several OTUs and the creation of noisy OTUs that contain some sequences that are loosely associated with the core sequence of the OTUs and that do not form a compact subgroup. We propose a simple and automatic 2-step method that successively categorises the OTUs of a sample as composed or single, and then identifies OTUs with noise among the simple ones. We applied the method on 32 samples of diatoms from Arcachon Bay (France) that represent contrasted environmental conditions and we obtained good agreement with expert categorisation of OTUs. We suggest that single OTUs without noise can be used as such for further ecological studies. Composed OTUs should be post-treated with classical clustering of community detection tools. The quality of single OTUs with noise remains to be further tested via supplementary studies on a diversity of organisms.

Keywords: metabarcoding, composed OTU, OTU with noise, diatoms, SVM, SBM

1 Introduction

Community ecology and macroecology aim at a better understanding of the diversity of life and its organisation patterns at various taxonomic levels and over space and time, (Mayr, 1982; Webb et al., 2002). To develop these studies, it is necessary to have reliable inventories of the diversity. Therefore, the concept of species, even if continuously debated, has emerged as a cornerstone. After a long history, it is currently addressed within the framework of evolutionary biology, especially with modern synthesis, and beyond (Mayr, 2004; de Queiroz, 2005). This has lead to molecular systematics (Hillis et al., 1996), which integrates statistical modelling of sequencesevolution and inference of phylogenetic trees between lineages (Nei and Kumar, 2000; Felsenstein, 2004). When no such tree is available, two molecular-based methods lead to the clustering or classification of unknown sequences of markers of taxonomic interest: building so-called OTUs with unsupervised clustering (Blaxter et al., 2005), and barcoding with supervised classification (Hebert et al., 2003). OTU stands for "Operational Taxonomic Unit". An OTU is a set of sequences that are ideally at a distance smaller than a given level referred to as barcoding gap (Blaxter et al., 2005). Exponential development of Next Generation Sequencing and High Throughput Sequencing has facilitated the industrial production of barcodes in environmental samples with metabarcoding (Hajibabaei et al., 2011; Taberlet et al., 2012; Kermarrec et al., 2013), produced in bulk, without knowing which organism they come from, especially in microbial communities. An environmental sample in metabarcoding is a set of readsthat are representative of the diversity of the community that has been sampled, and a sound basis for diversity studies. Being a set of sequences close to each other, it is expected that OTUs in an environmental sample represent a category relevant for biodiversity inventories on a molecular basis, where assemblages of OTUs mimic the organisation of communities as assemblages of species. It is expected that they represent building blocks of molecular diversity in communities, playing the same role as morphologically- or phylogenetically-based species. When possible, this can be validated by mapping some sequences in the OTU on taxonomically annotated reference databases. A difficulty is often encountered when doing so: not all species are available for learning in reference databases, because not all species are known or well represented in reference molecular databases even if they are morphologically well described. Note that this raises the question of qualifying and quantifying a correspondence (or not) between OTUs and the notion of species, which has been the subject of a long debate. In keeping with Blaxter et al. (2005), we adopt here the view that we are "*agnostic as to whether the taxa we can define using these barcode sequences [...] are species or not*". In our work, an OTU is defined as a set of sequences that are mutually close, and there is no attempt to make sense of an OTU, for example by naming it.

OTUs are building blocks of molecular-based inventories, and there are various protocols for building them from sets of sequences in an environmental sample. Artefacts in the production of OTUs can occur at different stages (see e.g. Bik et al., 2012). Moreover, as OTUs are used afterwards for computing diversity indices or performing statistical ecology, different delineations between OTUs may lead to different diversity indices or ecological profiles. The impact on diversity studies has been studied thoroughly, and some tools al-

ready exist to clean OTUs. For instance there may be more OTUs than expected from the expert knowledge about the diversity of the system studied. In Froslev et al. (2017) the authors propose a post-treatment method to identify and merge redundant OTUs, based on the identification of sequence similarities and of systematic co-occurrence of the OTUs in multiple samples. With metabaR (Zinger et al., 2021), it is possible to remove artefactual OTUs based on the analysis of their abundance across different samples. On the contrary, sequences of two distinct OTUs can be artificially grouped. For instance, it is known that single linkage clustering leads to chaining effects that may lead to the merging of two or more OTUs. In SWARM (Mahé et al., 2014, 2015), a post-treatment is proposed, referred to as the breaking phase, to split the potentially composed OTU. This is done by exploring the inner structure of OTUs which, for a composed OTU, is formed by picks of abundant amplicons with a valley in between (one pick per component). We propose to complete these tools for post-treating the OTUs of a sample, by using only the array of pairwise dissimilarities between sequences in each OTU.

To characterise the notion of quality of an OTU, we refer to an ideal OTU (where all distances within an OTU are smaller than the barcoding gap), and we identify possible deviations from the theoretical pattern of the corresponding distances array. Deviations, when they exist, are not random. We study two deviations leading to composed OTUs and OTUs with noise. As defined above, composed OTUs are the artificial merging of several OTUs, as opposed to single OTUs. We propose a new way to identify composed OTUs. Unlike the breaking phase in Mahé et al. (2014, 2015), it does not rely on a threshold parameter that must be fixed arbitrarily. Then, once composed OTUs have been split into single OTUs, we consider a second post-treatment to identify the presence of noise. We say that an OTU contains noise if it contains some sequences that are loosely associated with the core sequences and that do not form a compact subgroup of sequences. To the best of our knowledge, the identification and quantification of noise in OTUs has seldom been addressed. Our approach is a classification method based on simple statistics derived from the array matrix and on learning methods like a linear Support Vector Machine (SVM, Cortes and Vapnik 1995) and a Stochastic Block Model (SBM, Daudin et al. 2008).

We apply the approach on a dataset of diatoms from Arcachon Bay, kindly made available by Malabar project[1]. We believe that the fraction of the three OTU types (composed, single without noise and single with noise) present in an environmental sample can provide knowledge about on the ecology of the sample. As an illustration, we present results about the dependencies between the fraction of types and some known environmental variables describing the conditions under which the diatoms samples were collected.

2 Datasets, OTU picking and dissimilarity arrays

Data: The data sets we worked with to illustrate our method is a set of 32 fasta files of 32 environmental samples kindly provided by the Malabar project (Auby et al., 2022).

[1] see https://entrepot.recherche.data.gouv.fr/dataverse/malabar

They represent a sampling of the diversity of photosynthetic protists, mainly diatoms, in Arcachon Bay (France). Samples are allocated equally between the four seasons (autumn, winter, spring, summer), four locations (Bouée 13, Comprian, Jacquets, Teychan) and two water columns (pelagic high tide and benthic). This yields $4 \times 4 \times 2 = 32$ samples. Sample sizes range between 19 and 36 thousands reads (Table **??**). Reads are amplicons of a 312 bp region in the rbcL marker. For each sample, pairwise dissimilarities after dereplication between reads have been computed with the Smith-Waterman local alignment score. Because of the size and number of fasta files, we have used the distributed version of `disseq` called `mpidisseq` (see `https://gitlab.inria.fr/biodiversiton/disseq`), run on the cluster CURTA of the mésocentre of Nouvelle-Aquitaine. Hence, a $n \times n$ dissimilarity array is attached to each sample if it is composed of n reads.

Mapping reads on a reference database: A reference database for the rbcL marker for diatoms is available (Rimet et al., 2016). We mapped each sequence of the whole sample, regardless of the OTU it belongs to, on this reference database, with the `diagno-syst` (Frigerio et al., 2016) tool. Thanks to this mapping, we were able to explore the taxonomic profile of some of the OTUs typed as being composed or with noise. Not all sequences reached a match. For this study we had limited ourselves to OTUs with all sequences annotated in order to have a complete knowledge of the species present in the OTUs. This was the case for 180 OTUs, and the large majority (about 85 %) were monospecific. However, it is worth noticing that we have found one reference sequence of *Rhizosolenia fallax* which is present once and once only in several fully annotated OTUs belonging to several genera. Such a sequence has been disregarded.

OTU picking: The dissimilarity array of a sample is denoted D, and the dissimilarity between reads i and j is denoted $d(i, j)$ (term at row i and column j of D). In a second step, we computed OTUs from D for each sample. The numbers of sequences and OTUs per sample are given in Table 1 of Appendix 4. Two reads i and j belong to the same OTU if their dissimilarity is smaller than a selected barcoding gap (Blaxter et al., 2005). A problem with this definition is that such a property is not transitive (we can have $i \sim j$ and $j \sim k$ where $i \not\sim k$), whereas belonging to the same OTU is an equivalence relation, hence transitive. Knowing that, we implemented the following procedure on the dissimilarity array D after dereplication:

1. Select a barcoding gap g (here, $g = 9$, representing 3% of the marker length);
2. Create a graph $G = (V, E)$, where nodes $i \in V$ are the reads in the sample and $(i, j) \in E$ if and only if $d(i, j) \leq g$;
3. Compute all connected components of G.

An OTU is then defined as a connected component of G. We checked that our results are very close to the outputs of SWARM. It is not surprising because our procedure relies on building connected components at a given threshold and this is known to be equivalent to hierarchical aggregative clustering with Single Linkage (Gower and Ross, 1969). SWARM

relies on a bottom up algorithm (aggregate with seeds) equivalent to clustering with single linkage. When comparing our OTUs with SWARM, we noticed that SWARM OTUs were all included entirely within one of our OTU. The differences is due either to the breaking phase in SWARM which divides some of our OTUs, or to the production by SWARM of many very small sets of sequences. We currently investigate further this proximity between both approaches. For each sample, we extracted one dissimilarity array per OTU (dissimilarities between the reads within the OTU) denoted D_{otu}. The associated subgraph of G is denoted G_{otu}. It is connected by construction. Hence, we worked with 32 sets of dissimilarity arrays. We kept the OTUs with 15 reads or more only.

Ideal OTUs A drawback of the above procedure for derivation of OTUs is that two reads can have a dissimilarity larger than g and still belong to the same OTU. Is it possible to have a stronger property, i.e., to define an OTU by this procedure but with the guarantee that all dissimilarities within an OTU are below the barcoding gap? To answer positively, let us observe that the relationship $i \sim j$ if and only if $d(i,j) \leq \theta$ is transitive if and only if d is ultrametric. A distance d is said to be ultrametric if it fulfills the condition $d(i,j) \leq \max(d(i,k), d(j,k))$ for any read k, which is stronger than the classical triangular inequality. Dissimilarities do not satisfy either of these two properties. On the contrary, the evolutionary distance between two reads (the age of their Most Recent Common Ancestor in a dated phylogenetic tree) is ultrametric. If D is built with the evolutionary distance and steps 1 to 3 are applied, then all connected components of G are cliques, and an OTU is a clique (all pairs of reads within a given clique fulfil the condition $d(i,j) \leq g$).

Building OTUs by clustering from a matrix D of dissimilarities is a common approach. If all dissimilarities within the OTU are below the barcoding gap, the OTU is said to be "ideal". Our hypothesis is that the observed deviations from this ideal structure are not random, but are themselves structured. In what follows, we qualify and quantify two ways in which an OTU can diverge from being ideal: composed OTUs and OTUs with noise. We show that they represent the majority of the situations observed.

3 Method for identifying composed OTUs

It is well known that composed OTUs can be produced during the phase of clustering of the sample sequences by the well-known "chaining effect", especially with the single linkage aggregation method: if $d(i,j) < g$, $d(j,k) < g$ and $d(k,\ell) < g$, then at one aggregation step, i, j, k and ℓ will be grouped into the same OTU, while i and ℓ can be molecularly different, where $d(i,\ell)$ is large. We illustrate this effect on a sample by comparing the species and the OTU that each sequence belongs to. In Fig. 1, we show the same point cloud, the projection on the first two axes by Multidimensional Scaling (MDS, Cox and Cox 2001) of the dissimilarity array of a sample twice, once where dots (i.e., sequences) are coloured according to the OTU they belong to, and once according to the species they belong to (see Fig. 1). It is clear that the blue OTU (the largest) is composed because of the archipelago of isolated dots scattered between the three components, which leads to

chaining and a spurious OTU.

Here, we propose an automatic unsupervised method for sorting the OTUs of a sample into two groups: single OTUs and composed OTUs. They are distinct by the pattern of their dissimilarity array. We characterise the two categories as follows. A single OTU is close to what would be a theoretically ideal OTU, where all dissimilarities in D_{otu} are smaller than the barcoding gap. There may be a few exceptions for some sequences, but we will deal with that in a second step in Section 4. The corresponding graph G_{otu} is composed of a single large strongly-connected component with the possibility of some satellite nodes. Composed OTUs correspond to dissimilarity arrays with a structure of two or more blocks, with intra-block dissimilarities smaller than the barcoding gap, and most of the inter-block dissimilarities larger. This leads to a graph G_{otu} with several components, where the nodes in a component are strongly connected, and there are few connections between the components. In graph theory, such a graph is said to have a community structure (Girvan and Newman, 2002), and each component is part of the community. In Fig. 2 we provide an example of a graph of an ideal OTU, of a single OTU with satellite nodes, and of a composed OTU.

Therefore, for a single OTU, most dissimilarities in D_{otu} will be smaller than the barcoding gap. For a composed OTU, there will be a significant proportion of dissimilarities larger than the gap (due to the inter-component dissimilarities). This is the information we use to discriminate between single and composed OTUs. For a given OTU, we build G_{otu} from D_{otu}. We then define θ as the ratio between the number of missing edges in G_{otu} and the total number of possible edges. The number of missing edges corresponds to half the number of elements in D_{otu} that are larger than the barcoding gap. It is equal to $\sum_{i<j} \mathbb{1}_{D_{otu}(i,j)>g}$ where the sum is over all pair (i, j) of lines and columns of D_{otu} where $i < j$. The total number of possible edges is equal to $\frac{n_{otu}(n_{otu}-1)}{2}$, where n_{otu} is the number of sequences in the OTU. Therefore $\theta = 2\sum_{i<j} \mathbb{1}_{D_{otu}(i,j)>g}/(n_{otu}(n_{otu} - 1)$. It is clear that for single OTUs, θ will be small, because very few edges are missing. For composed OTUs, θ will be large. Indeed, let us take as an example an OTU with two balanced components. There will be few missing edges within each component, but many edges missing between both components. If each component has $n/2$ sequences, there are possibly $n_{otu}^2/4$ edges between both components, and as many potential missing edges. Hence $\sum_{i<j} \mathbb{1}_{D_{otu}(i,j)>g} \sim n_{otu}^2/4$ while $n_{otu}(n_{otu} - 1)/2 \sim n_{otu}^2/2$. Finally, $\theta \approx 1/2 \gg 0$.

To sort the OTUs of a sample into composed and single ones, we use θ, which can be computed directly from D. We define a critical value θ_c as follows. We compute θ for each OTU and we build a smoothed version of the histogram of the θs using a Gaussian kernel (see Appendix 2). This estimated density always shows a first large mode around low values of θ, followed by one or several other less important modes. We define θ_c as the value of θ for which the minimum of the estimated density is reached between the first and the second mode. If $\theta < \theta_c$ the OTU is classified as single, otherwise it is classified as composed (see Fig. 3).

We compared the results of this automatic method with an expert classification of the OTUs, which works as follows. If a clustering method is applied to the dissimilarity array of the OTU, components are expected to be two distinct clusters (or more) because of the split of dissimilarities between intra-component and inter-component dissimilarities. First, heatmaps of the dissimilarity array of each OTU were drawn, with sequences ordered according to the leaves of a dendrogram (Aggregative Hierarchical Clustering with Ward criteria, Müllner 2013). Examples of a heatmap for one component only and for two components are given in Fig. 2. Second, we saw at the beginning of this section that the graph G attached to a composed OTU is organised as a set of connected communities, one per component. Such graphs were drawn for each OTU of the sample. Finally, we attributed the character "single" or "composed" for each OTU by visual inspection of the heatmap and the graph. Most of the cases were unambiguous, clearly belonging to one type or the other. However, the transition is smooth rather than discrete (for example, θ varies continuously). It may happen that intermediate cases occur, for example if there are two components with highly unbalanced sizes, like a dominant one and a small satellite one. In such a case, the OTU was labelled as "uncertain".

4 Method for identifying OTUs with noise from among single ones

We focus now on OTUs identified as single. In Section 5 we will discuss possible tools to split OTUs identified as being composed in order to obtain a clustering of the sample's sequences formed only of single OTUs. For these OTUs, i.e., with a majority of sequences that belong to the same entity, there could still be some sequences that are loosely associated with the OTU: for such a sequence i, dissimilarities $d(i, j)$ are below the barcoding gap for only a small number of sequences j in the OTU. These sequences are far from the core sequences of the OTU and they do not form a second entity (as in a composed OTU) since they can be far from each other (see Fig. 2, centre). We qualify these sequences as part of the noise. In order to determine if a single OTU contains noise sequences or not, we propose a fully automatic supervised classification method whose input variables are features derived from the dissimilarity array D_{otu}. Namely, we use a linear Support Vector Machine (SVM) to discriminate between the two types of single OTUs. To derive the features, we estimate the parameters of a Stochastic Block Model (SBM, Holland et al., 1983; Daudin et al., 2008; Lee and Wilkinson, 2019) with a Poisson distribution on dissimilarities, and with two blocks (see Appendix 1 for a description of the SBM model). The reason for choosing two blocks is that in the presence of noise, we expect that the core sequences of the OTU will be grouped into one block, and the atypic sequences into another. We use the block parameters as features. More precisely, the SBM makes it possible to cluster individuals based on their pairwise dissimilarities. Individuals in the same block share the same pattern of connectivity. A specificity of SBM lies in its plasticity: a block is not necessarily assortative (i.e., small within-block dissimilarities); it may also be dissortative. Our argument for choosing SBM is that if there are some noise sequences in an OTU, they will be grouped into a dissortative block. If there is no noise, the two blocks will be assorta-

tive. These two different patterns can be identified using the connectivity matrix Λ of the SBM. In the case of a two-block SBM this is a 2×2 symmetric matrix. The two diagonal elements, $\Lambda(1,1)$ and $\Lambda(2,2)$, correspond to the mean intra-block dissimilarity, and the non-diagonal element $\Lambda(1,2)$ corresponds to the mean inter-block dissimilarity. If a single OTU contains noise sequences, they will be grouped into one of the two blocks, let us say, block 2, with a large value for $\Lambda(2,2)$ and for $\Lambda(1,2)$. If the OTU is without noise, all the elements of Λ should be small. We chose the two values, $\Lambda(1,2)$ and $max(\Lambda(1,1), \Lambda(2,2))$, as features for the linear SVM. We considered other combinations of the elements of Λ but they did not improve the performance of the classification and this choice is easier to interpret in terms of presence/absence of noise.

In practice, we assigned an 'expert' label to each OTU of a training set, among 'with noise', 'uncertain', and 'without noise'. To do this, we computed the normalised degree δ_{seq} of each sequence of the OTU, defined as the percentage of dissimilarities smaller than the barcoding gap in the line corresponding to this sequence in the dissimilarity array D_{otu}: $\delta_{seq} = 100 * \sum_{j \neq i} \mathbb{1}_{\{d(i,j)<g\}} / (n_{otu} - 1)$. If the minimum of δ_{seq}, over the OTU sequences is lower than 20%, the OTU is labelled as 'with noise'; if it is larger than 70%, it is labelled as 'without noise'; otherwise, the OTU is labelled as 'uncertain'. Only OTUs labelled as 'with noise' or 'without noise' are used to learn the SVM. Note that this method could be directly envisaged as a candidate for identifying OTUs with noise. However, it is not fully automatic since it relies on two thresholds that were manually defined, and some OTUs remain unclassified ('uncertain' type). We refer to it as the degree-based classifier below.

As opposed to the method to identify composed and single OTUs, the method to identify OTUs with noise is a supervised method that requires a training set to learn the SVM. The most discriminant factors, when studying diversity are the season and the water column. This is the reason why we built the training set on one location (Teychan) and the test set on the other three locations. Both sets contain samples associated with different values for the season and the water column. We will refer to our training set as the Teychan dataset, i.e., the set of the eight samples located at Teychan (two samples per seasons: one for benthic and one for pelagic). It is composed of 654 OTUs.

5 Identification of composed OTUs in diatom samples

Analysis of the Teychan dataset. For the Teychan dataset, an expert classification of each OTU of each sample into one of the three categories - composed, single or uncertain - is available (based on the expert procedure described in Section 3). We use this dataset to illustrate the behaviour of the automatic method for identifying composed and single OTUs. The contingency table built from the 654 OTUs of the Teychan dataset (Table 1) shows a very good agreement between the expert classification and the automatic one. In particular, single OTUs are very well identified: only 12 false negatives (among 104 composed OTUs) and only nine false positives (among 527 single OTUs). Uncertain OTUs are evenly distributed between composed and single categories by the automatic method.

In Appendix 4, we provide the individual contingency matrices and θ_c values for each of the eight samples composing the Teychan dataset.

Analysis of the whole dataset. We then applied the procedure to the whole data set (the 32 samples). We do not have the expert classification that would require a visual inspection of 2529 dissimilarity arrays.

We tested the hypothesis of a link between the OTU type (single or composed) and its size. Fig. 2 of Appendix 3 visually presents the link between the size of an OTU and its type. It can be seen that single OTUs (green and blue dots) have small to medium sizes, and that composed OTUs (red dots) have larger sizes. This was quantified by a Wilcoxon Mann-Whitney test (function `mannwhitneyu()` in Python library `scipy.stats`) between single and composed categories. The results (see Table 2) show a strong evidence for a link between the OTU size and its type (composed or single).

Among the 180 OTUs that were fully annotated, eight were categorised as composed. We observed three situations. For two of them, there are two or three species present in the OTU, and the distance array D_{otu} and graph G_{otu} are clearly structured into two blocks separating one species from the other(s). This is the typical situation that we target when identifying composed OTUs. Three other OTUs are monospecific and there is no obvious structure in D_{otu} or G_{otu}. However, they have the particularity that sequences are loosely connected to the others, leading to a large value of θ, larger than θ_c. Finally, the last three OTUs are monospecific (or nearly), and D_{otu} and G_{otu} are nevertheless structured into two blocks. An example of each situation is given in Fig. 1 of Appendix 3.

6 Identification of OTUs with noise in diatom samples

Training on the Teychan dataset. We use the Teychan dataset as a training set to learn the SVM classifier for identifying OTUs with noise among the one categorised as single in the first step. This training step is performed using only OTUs that have been categorised as with or without noise by the degree-based classifier (uncertain OTUs cannot be used here).

For each choice of features (pair of coefficients of the Λ matrix), we ran a 10-fold cross validation to estimate the error of prediction. We obtained the best Area Under Curve value (AUC = 0.951) with the features $f_1 = max(\Lambda(1,1), \Lambda(2,2))$ and $f_2 = \Lambda(1,2)$. The feature f_1 represents the mean dissimilarity between two sequences of the SBM block with the larger mean intra dissimilarity. If there are noise sequences, they should be in this block. The feature f_2 represents the mean inter-block dissimilarity in the SBM model. The SVM classifier frontier is defined by the expression $y = -9.452 + 0.569 * f_1 + 0.876 * f_2$. Contingency Table 3 reports the comparison between the two classification methods, now including the OTUs categorised as uncertain (the eight contingency matrices, one per sample in the Teychan dataset, are provided in Appendix 4). The SVM classifier very efficiently detects the OTUs with noise (only six missed among 381). It is a bit less efficient to detect OTUs without noise (6 missed among 48). The majority of uncertain OTUs are classified as being with noise by the SVM classifier.

Results on the test set. The SVM classifier obtained on the training set is applied to the OTUs of the 24 samples of the test set. Since the expert method (see Section 4) can also be automatised, we can compare the results of the two classifiers. They are reported in contingency Table 4. It can be observed that for both methods, there are much fewer OTUs classified as without noise than with noise. The SVM classifier identifies all the OTUs with noise. However, only 64 % of the OTUs without noise are identified. We can see (Fig. 4) a pattern in the values of the two features, f_1 and f_2 that are used by the SVM classifier, depending on the OTU type (with or without noise). We recall that f_1 represents the mean intra dissimilarity of the SBM block with larger intra dissimilarity. f_2 is the mean inter block dissimilarity in the infered SBM model. Single OTUs identified as 'without noise' are associated with a low value of f_1 (between 4 and 8) and a low value of f_2 (between 4 and 8 as well). On the contrary, single OTUs identified as 'with noise' are associated with large values of f_1 (almost always between 6 and 16) and with large values of f_2 (between 6 and 25). Furthermore these two parameters of the inferred SBM model increase simultaneously, showing a gradient of noise intensity among the single OTUs.

Link between OTU size and OTU type. For OTUs categorised as single, we test the hypothesis of a link between the OTU size and its category (with or without noise). The same test as in Section 5 is performed (based on the single OTUs of the 32 samples) and the results show that there is strong evidence for such a link (see Table 5).

Among the 180 OTUs that were fully annotated, 153 were categorised as single with noise and 23 as single without noise. Ignoring the artefactual presence of sequences of *Rhizosolenia fallax* species, almost all were monospecific (only two exceptions).

7 Link between sample composition and environmental conditions

Having applied the two procedures to each of the 32 samples for identification of composed, single with noise and single without noise OTUs, we then computed the proportion of each type per sample. In Fig. 5 we show a visualisation by ternary plot of these proportions. Globally, all samples have a low proportion of OTUs without noise, and we can observe that the proportion of single OTUs with noise and composed OTUs vary from one sample to another.

The central ternary plot of Fig. 5 suggests a potential link between these proportions and the water column of the sample. This is also suggested by the plot of the fraction of composed OTUs for each of the 32 samples as displayed in Fig. 3 of Appendix 3. To test this link, we first considered two sets of 16 values: the list of percentages of composed OTUs in the benthic samples and in the pelagic samples. We applied a Wilcoxon rank test and obtained a p-value of 1.6×10^{-4}. The mean fraction of composed OTUs in a benthic sample (resp. a pelagic sample) is 0.19 (resp. 0.10). Consequently, there is strong evidence that the fraction of composed OTUs is larger in benthic samples than in pelagic ones.

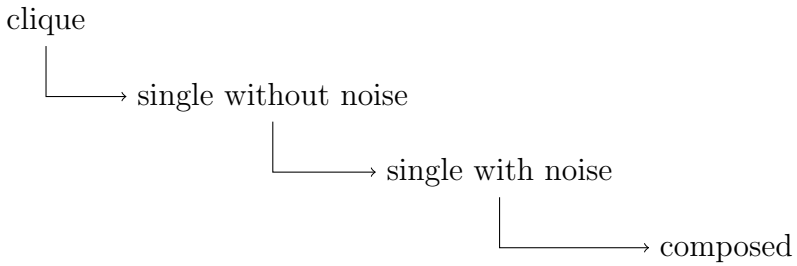
We then considered two other sets of 16 values: the list of percentage of OTUs with noise (among the single OTUs) in the benthic samples and in the pelagic samples. We also applied a Wilcoxon rank test and we obtained a p-value of 0.04763. We concluded that there is no evidence that the fraction of OTUs with noise (among the single OTUs) in a sample is different for benthic and pelagic condition.

We did not test whether the other environmental conditions (season, location) have or don't have an influence on the composition in the sample since the number of observations per condition would be too small (8).

8 Discussion

Recent advances in massively parallel sequencing technology has led to the rapid production of millions of reads. This has opened the way to the analysis of many environmental communities, leading to further exploration of their diversity and ecology, at a pace that was unimaginable beforehand. The building blocks of such studies are sets of OTUs obtained by clustering the reads of a given sample. In parallel, clustering methods have diversified considerably, leading to several benchmark studies (see e.g. Sun et al. 2011; Fahad et al. 2014). Here, we propose a tool to make progress in assessing the quality of an OTU once a clustering method has been selected, by developing the comparison between its inner structure (pairwise dissimilarity array) and an ideal one, and by characterising two ways in which the structure of an OTU can deviate from the ideal situation: first, we distinguish composed vs. single OTUs. Second, among the single OTUs, we distinguish OTUs with and without noise.

The discussion of the quality of the different types of OTUs is organised along a gradient of complexity of the structure of the OTUs, as follows:



Cliques: The expected structure of the graph G_{otu} built from dissimilarity array D_{otu} is a clique if the dissimilarities are evolutionary distances (the age of the Most Recent Common Ancestor, MRCA). In such an ideal case, the OTU is obviously reliable. However, in practice, we work with genetic distances computed from local alignment scores. The discrepancy between evolutionary distances and genetic distances within a set of sequences increases with the age of the MRCA. It can therefore be expected that cliques represent clusters with a relatively young MRCA, where genetic distances are close to evolutionary

distances. This allows us to postulate that cliques built from genetic distances are OTUs of good quality. There are four cliques over all of the 32 samples of diatoms. Three of them have no annotated read. This may mean that they represent species that are absent from the reference database. One of them is partially annotated, always with the same species. The fact that some reads in the clique are not recovered probably means that mapping reached its limit in terms of quality, because if a query maps on references with different taxa, the mapping is said to be ambiguous and the read is not annotated.

Single OTUs without noise: Let us recall that the noise (or the absence of noise) in a single OTU is detected based on the value of two features f_1 and f_2, where f_1 represents the mean dissimilarity between two sequences of the SBM block with the larger mean intra dissimilarity, and f_2 represents the mean inter-block dissimilarity. A single OTU is typed as "without noise" if the parameters f_1 and f_2 are both small, as illustrated in Fig. 4 for one sample. In section 6 we showned that both features f_1 and f_2 are always lower than 8 for single OTUs without noise. This implies that for those OTUs in the SBM modelling of D_{otu}, all distances are realisations of a Poisson distribution with a mean lower than the barcoding gap (nine in our study). Therefore, the graph G_{otu} is close to a clique. It is tempting to extrapolate and derive the conclusion that single OTUs without noise can be considered of good quality for use in further studies.

In order to provide indications about the quality of an OTU (which means it can be accepted as an OTU for further studies) that is not a clique or a single OTU without noise, we referred to an external expert evaluation. Although we are agnostic as to whether an OTU has does not have a taxonomic meaning, we used the mapping of reads on a reference database as external information. If all the reads in an OTU are annotated, and assigned to the same species, then OTU picking and taxonomy converge, suggesting that the OTU can be considered of good quality. Otherwise it is questionable. Hence, we focused on fully annotated OTUs in the discussion.

Single OTUs with noise: A single OTU is typed as "with noise" if features f_1 and f_2 are both large. Such an OTU displays a minority of satellite sequences (the halo), which are close to (at a distance smaller than the gap) a small fraction of the remaining sequences only (the core, the main densely connected component). In the subsample of fully annotated OTUs, almost all of the single OTUs with noise are monospecific ones. All of them are monospecific, regardless of the quantity or intensity of noise. Whether such a conclusion can be extended beyond fully annotated OTUs is an open question and deserves further studies on a diversity of organisms to progress along this line. Indeed, a partial covering only by mapping can be due to the fact that uncovered reads either belong to another species absent in the reference database, lowering the acceptability of the OTU, or that they belong to the same species but are labelled as unknown due to imperfections and errors in the mapping or the reference database.

Composed OTUs: Composed OTUs are very likely to be large OTUs and to be composed of two or more components each of which is a candidate to be a more reliable OTU.

However, in the subsample of fully annotated OTUs, we observed some composed OTUs with a different profile: either monospecific OTUs with, overall, a low level of connections in G_{otu}, or monospecific OTUs with a clear structure divided into two blocks. Both cases lead to large values of missing edges and the OTUs are therefore typed as composed. In the latter case, one possible reason for the block pattern of the distance array may be a structure in the intraspecific molecular diversity. However, the number of specimens in one OTU is often too small to check with population genetics indices (see Phillips et al. 2018 for a discussion about the sample size necessary for assessing molecular intraspecific diversity). Regarding large composed OTUs, the production of spurious clusters by a chaining effect in aggregative clustering with single linkage is well known (see, e.g., Kopp 1978), and can lead to composed OTUs. Programmes like SWARM (Mahé et al., 2014, 2015) have identified this issue and provide a way to solve it by breaking the chains in the amplicon space. Here, we suggest that chaining can occur because of the non-universality of the barcoding gap. Some structures of the dissimilarity array of a sample are more likely to lead to chaining. Indeed, depending on the sample studied, positions of OTUs or their components can vary along a gradient from well isolated to densely connected in networks, over loosely connected two by two. These situations refer to the separability of OTUs. We suggest that this is an issue for the quality of OTUs. Some examples of the diversity of those situations are given in Fig. 4 of Appendix 3. Such a variability can be understood keeping in mind that the barcoding gap is not universal. We refer the reader to Phillips et al. (2022), the running title of which is "Is the Barcoding Gap Real?" for a thorough discussion and critique of the notion of barcoding gap. It can vary between clades in a sample. Indeed, let us assume that for a given small gap, we have a set of well delineated OTUs corresponding each to a taxon. Then increasing the gap to build connected components will lead to new edges between former OTUs and, possibly, to a network of connected entities, with a weakening of the possibility to discriminate them, as in the top graph of Fig. 4 in Appendix 3. We have shown that composed OTUs are the largest ones with very high significativity. This means that medium size and small size OTUs are single and likely to correspond to taxa. This also means that the selected barcoding gap is relevant for delineating most OTUs (the middle and small size ones), but inadequate for most of largest ones: the existence of spurious composed OTUs reflects the inadequacy of the selected barcoding gap to delineate OTUs among those sequences. This may explain the difference in the ratio of single / composed OTUs between benthic (with several composed large OTUs) and pelagic samples (with fewer ones): the structure of the molecular diversity of pelagic and benthic diatom flora differ. Indeed most species are preferentially present either in the benthic samples or in the pelagic samples (see Fig. 5 of Appendix 3). We can hypothesise from this observation a pattern where distances between a significant fraction of species in benthic flora is smaller than in pelagic flora. This deserves further investigation. The large spurious OTUs, automatically detected by $\theta_{\text{otu}} > \theta_c$, should be reshaped as sets of new and smaller OTUs. Two ways to do this are to build them as outcomes of either unsupervised clustering of the dissimilarity array of the composed OTU or of community detection (see Fortunato 2010 for a review of this approach) in the graph induced by the array.

The approach developed here to type OTUs requires the dissimilarity array D_{otu} of each

OTU in a sample. OTUs in a sample are classically derived by clustering methods like UCLUST (Edgar, 2010, 2013), or some standard aggregative clustering methods available through mothur (Schloss et al., 2009) or QIIME (Caporaso et al., 2010) classical softwares (see also Abouabdallah et al. 2022 for a comparison of different aggregative clustering methods on amplicons for barcoding). However, not all distances are computed in UCLUST (only dissimilarities between queries and some seeds). Several methods are available in mothur to compare pairwise distances, e.g. from multiple or global alignments. Among all of the methods available, we recommend the choice of exact methods to compute the dissimilarities on pairwise alignments, like Neddleman-Wunsch when sequence lengths are very close, or Smith-Waterman when they are more variable, both being an implementation of an edit distance (see Gusfield 1997, Chapter 11). Needleman-Wunsch-based dissimilarities are available in mothur with `pairwise.seqs`. Other pairwise dissimilarities, less costly in time, based on heuristics like kmer distances can also be used for building OTUs. Smith-Waterman and Needleman-Wunsch based dissimilarities can be built from a fasta file with the `disseq` programme dedicated to this, and kmer (short and long)-based distances with the programme `jelly_diskm.py`. Both are publically available from gitlab: `https://gitlab.inria.fr/biodiversiton/disseq`. The choice of a dissimilarity can indeed be an option when applying our method for identifying the type of OTUs in a sample. However, whether such an outcome depends or not on the choice of the dissimilarities deserves further investigations.

9 Data accessibility

The codes for learning the noise classifier and for determining the type of OTUs are available on figshare (Cros et al., 2022). The figshare project also contains the dissimilarity arrays of the OTUs for the 32 samples of the diatom studies and the results of the study. The figshare reference link is `https://doi.org/10.6084/m9.figshare.20764690`.

10 Acknowledgements

We thank the participants of the Malabar project for their authorisation to use the data produced in this project, especially the Laboratoire Environnement-Ressources of l'IFREMER at Arcachon for the field campaign, Emilie Chancerel and Franck Salin at INRAE BioGeCo for the production of DNA sequences. Computer time for the preparation of the data in this study was provided by the computing facilities of the MCIA (Mésocentre de Calcul Intensif Aquitain). Data have been produced in the Malabar project, supported by "Cote Labex" Call for Research Projects, Year 2017.

11 Author contributions

A.F. and N.P. designed the study. J.-M.F. prepared the data. A.F., N.P. and M.-J.C. performed the research and analysed the data. M.-J.C. developed the R and python codes.

The paper was written by A.F. and N.P. All authors commented on and approved the final manuscript.

References

Abouabdallah, A.-M., Peyrard, N., and Franc, A. (2022). Does clustering of DNA barcodes agree with botanical classification directly at high taxonomic levels? Trees in French Guiana as a case study. *Molecular Ecology Resources*, 22(5):1746–1761.

Auby, I., Méteigner, C., Rumebe, M., Chancerel, E., Salin, F., Aluome, C., Barraquand, F., Carassou, L., Del Amo, Y., Meleder, V., Petit, A., Picoche, C., Frigerio, J.-M., and Franc, A. (2022). Malabar datasets used in study "OTU quality from dissimilarity arrays". Recherche Data Gouv, V1, DOI: 10.57745/7T2UCB.

Bik, H. M., Porazinska, D. L., Creer, S., Caporaso, J. G., Knight, R., and Thomas, W. K. (2012). Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology and Evolution,*, 27:233–243.

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., and Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B*, 360:1935–1943.

Caporaso, J., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E., Fierer, N., Peña, A., Goodrich, J., Gordon, J., Huttley, G., Kelley, S., Knights, D., Koenig, J., Ley, R., Lozupone, C., McDonald, D., Muegge, B., Pirrung, M., Reeder, J., Sevinsky, J., Turnbaugh, P., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, 7:335–336.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cox, T. and Cox, M. A. A. (2001). *Multidimensional Scaling - Second edition*, volume 88 of *Monographs on Statistics and Applied Probability*. Chapman & al.

Cros, M.-J., Frigerio, J.-M., Peyrard, N., and Franc, A. (2022). Code, dataset and results for the study "OTU quality from dissimilarity arrays". Figshare, https://doi.org/10.6084/m9.figshare.20764690.

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

de Queiroz, K. (2005). Ernst Mayr and the modern concept of species. *PNAS*, 102(suppl. 1):6600–6607.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26:2460–2461.

Edgar, R. C. (2013). UPARSE: Highly accurate OTU sequences from microbial amplicon reads. *Nature Methods*.

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S., and Bouras, B. (2014). A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. *IEEE Transactions on Emerging Topics on Computing*, 2(4):267–279.

Felsenstein, J. (2004). *Inferring phylogenies*. Sinauer.

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, 486:75–174.

Frigerio, J. M., Rimet, F., Bouchez, A., Chancerel, E., Chaumeil, P., Salin, F., Thérond, S., Kahlert, M., and Franc, A. (2016). diagno-syst: a tool for accurate inventories in metabarcoding. *arXiv, https://arxiv.org/abs/1611.09410*.

Froslev, T., Kjoller, R., Bruun, H., Ejrnaes, R., Brunbjerg, A., Pietroni, C., and Hansen, A. (2017). Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. *Nature Communications*.

Girvan, M. and Newman, M. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12).

Gower, J. C. and Ross, G. J. S. (1969). Minimum Spanning Trees and Single Linkage Cluster Analysis. *Journal of the Royal Statistical Society, Series C*, 18(1):54–64.

Gusfield, D. (1997). *Algorithms on Strings, Trees and sequences*. Cambridge University Press.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G. A. C., and Baird, D. J. (2011). Environmental barcoding: a next generation sequencing approach for biomonitoring applications using river benthos. *PLoS One*, 6(4):e17497.

Hebert, P. D. N., Cywinska, A., Ball, S. L., and deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London*, 270:313–321.

Hillis, D. M., Moritz, C., and Mable, B. (1996). *Molecular Systematics*. Sinauer, Sunderland, Mass.

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

Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Humbert, J.-F., and Bouchez, A. (2013). Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. *Molecular Ecology Resources*, 13:607–619.

Kopp, B. (1978). Hierarchical Classification I: Single Linkage Method. *Biometrical Journal*, 20(5):495–501.

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

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2014). Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, 2:e593:https://doi.org/10.7717/peerj.593.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2015). Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ*, 3:e1420;:https://doi.org/10.7717/peerj.1420.

Mayr, E. (1982). *The Growth of Biological Thought: Diversity, Evolution and Inheritance.* Harvard University Press.

Mayr, E. (2004). 80 years of watching the evolutionary scenery. *Science*, 305(5680):46–47.

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

Nei, M. and Kumar, S. (2000). *Molecular Evolution and Phylogenetics*. Oxford University Press, Inc., NY.

Phillips, J. D., Gillis, D. J., and Hanner, R. H. (2018). Incomplete estimates of genetic diversity within species: Implications for DNA barcoding. *Ecology and Evolution*, 9:2996–3010.

Phillips, J. D., Gillis, D. J., and Hanner, R. H. (2022). Lack of Statistical Rigor in DNA Barcoding Likely Invalidates the Presence of a True Species' Barcode Gap. *Frontiers in Ecology and Evolution*, 10:859099.

Rimet, F., Chaumeil, P., Keck, F., Kermarrec, L., Vasselon, V., Kahlert, M., Franc, A., and Bouchez, A. (2016). R-Syst::diatom: an open-access and curated barcode database for diatoms and freshwater monitoring. *Database*, page baw016.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., and Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, 75:7537–7541.

Sun, Y., Cai, Y., Huse, S. M., Knight, R., Farmerie, W. G., Wang, X., and Mai, V. (2011). A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. *Briefings in Bioinformatics*, 13(1):107–121.

Taberlet, P., Coissac, E., Hajibabaei, M., and Rieseberg, L. (2012). Environmental DNA. *Molecular Ecology*, 2:1789–1793.

Webb, C. O., Ackerly, D. D., PcPeek, M. A., and Donoghue, M. J. (2002). Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, 33:475–505.

Zinger, L., Lionnet, C., Benoiston, A.-S., Donald, J., Mercier, C., and Boyer, F. (2021). metabaR: An R package for the evaluation and improvement of DNA metabarcoding data quality. *Methods in Ecology and Evolution*, 12(4):586–592.

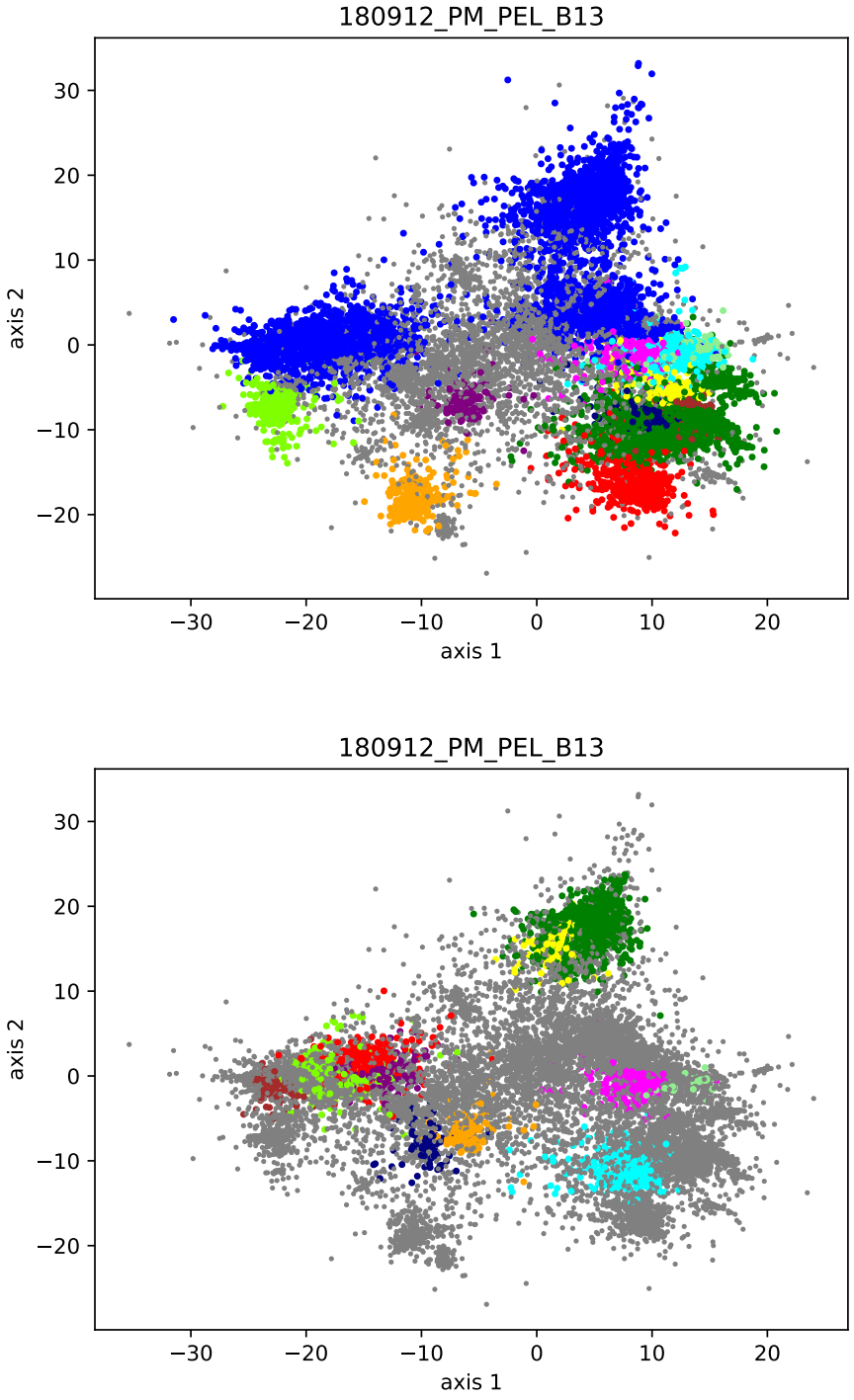


Figure 1: Illustration of the chaining effect. Both figures display the same scatter plot of sample `180912_PM_PEL_B13` (high tide, pelagic, summer, Bouée 13), where one dot is a sequence (there are $37,036$ dots), with the first MDS component on the x axis and the second one on the y axis. The two plots differ by the way dots are coloured. In the top plot, dots are coloured according to the OTU they belong to. In the bottom plot, they are coloured according to the species they have been assigned to. Only the species and OTUs with the 12 largest sizes have been coloured; the remaining ones are coloured in grey (if not, many colours would have been indistinguishable).

		Automatic		
		Composed OTUs	Single OTU	Total
Expert	Composed OTUs	92	12	104
	Uncertain OTUs	11	12	23
	Single OTUs	9	518	527
Total		112	542	654

Table 1: Comparison of the expert classification and the automatic classification of the OTUs into the composed and single categories, for the Teychan dataset.

Number of OTUs	$2,529$
Mean rank for single OTUs	$1,163.5$
Mean rank for composed OTUs	$1,778.5$
$p-$value	1.535×10^{-55}

Table 2: Link between OTU size and its classification as composed or single. Statistics of the ranks (the ranks are ordered from smallest to largest size) and the p-value of the Wilconxon Mann-Whitney test.

		Automatic		
		OTUs with noise	OTUs without noise	Total
Expert	OTUs with noise	375	6	381
	Uncertain OTUs	87	26	113
	OTUs without noise	6	42	48
Total		468	74	542

Table 3: Comparison of the degree-based classification and the SVM classification of the single OTUs into the 'with noise' and 'without noise' categories, on the Teychan dataset (training set).

		Automatic		
		OTUs with noise	OTUs without noise	Total
Expert	OTUs with noise	1228	0	1228
	Uncertain OTUs	277	24	301
	OTUs without noise	16	29	45
Total		1521	53	1574

Table 4: Comparison of the degree-based classification and the SVM classification of the single OTUs into the 'with noise' and 'without noise' categories, on the test set.

Number of OTUs	$2,116$
Mean rank for single OTUs without noise	552.5
Mean rank for single OTUs with noise	$1,089.7$
$p-$value	3.7×10^{-22}

Table 5: Link between OTU size and its classification as single with or without noise. Statistics of the ranks (the ranks are ordered from smallest to largest size) and the p-value of the Wilconxon Mann-Whitney test.

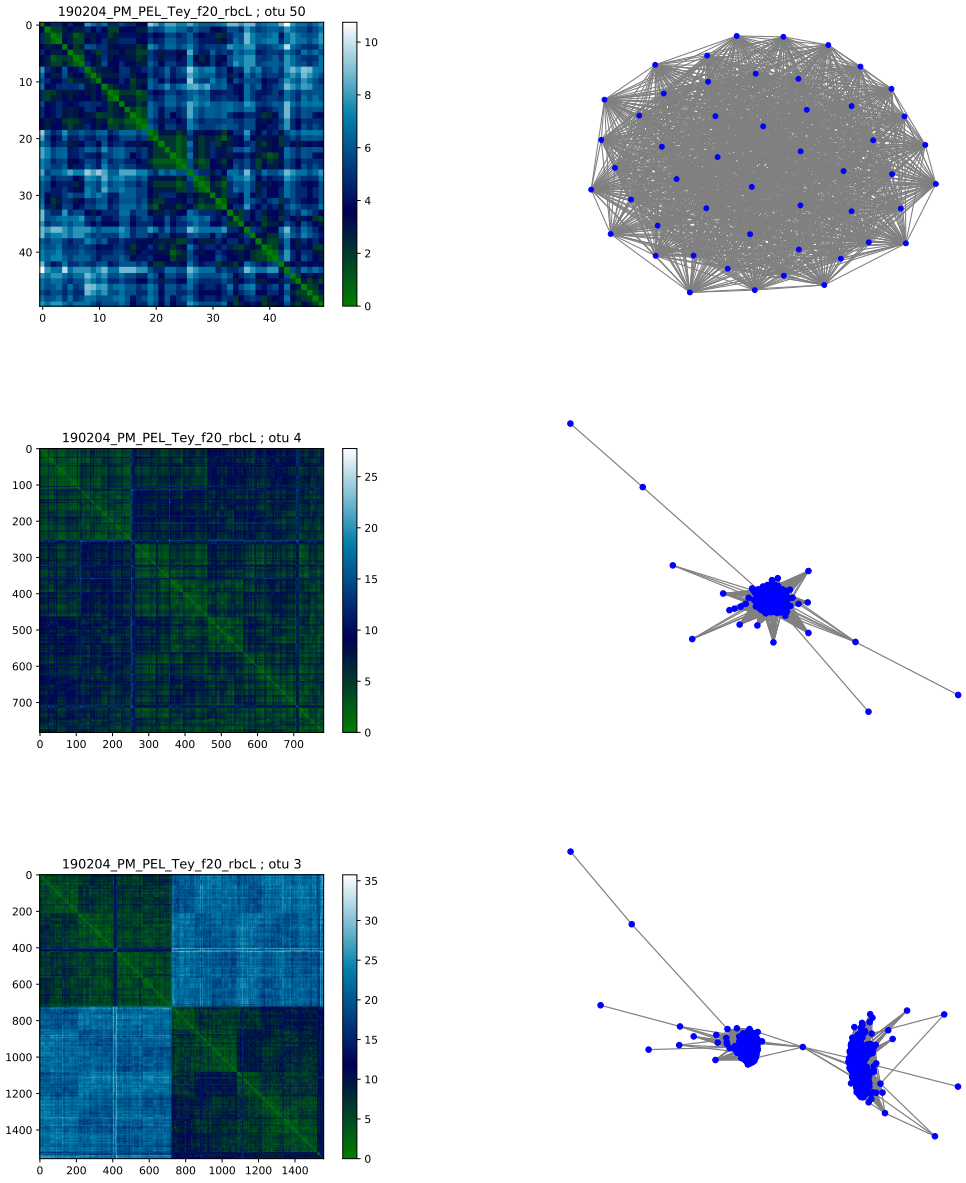


Figure 2: Examples of graph G_{otu} for three types of OTUs, from top to bottom: (*i*) ideal OTU, which is single, and a clique (each sequence has a dissimilarity smaller then the bacoding gap with all the other sequences of the OTU; (*ii*) a single OTU with a large strongly connected core component, and some satellite nodes; (*iii*) a composed OTU, consisting of several components with high intra-component connections rates and low between components connection rates (and some satellite nodes as well).

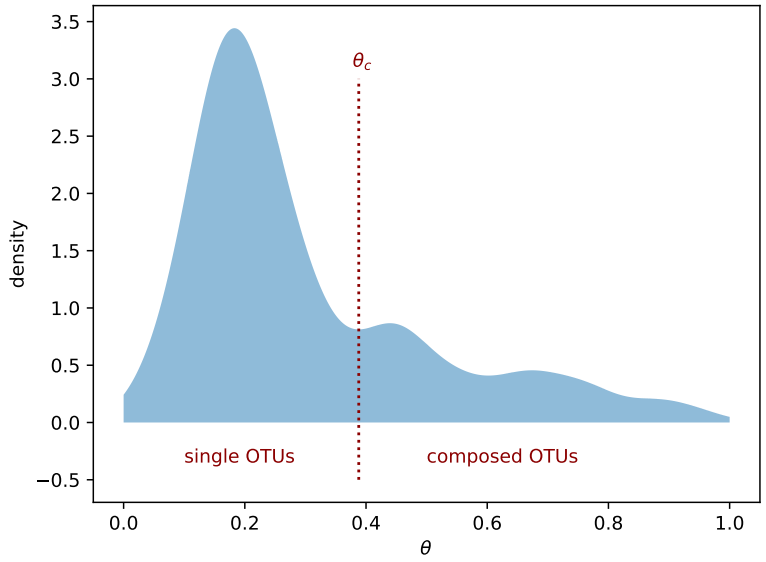


Figure 3: Principle of the method for sorting OTUs of a sample into composed and single ones. Example of a smoothed version of the histogram of θ values (ratio between the number of missing edges in G_{otu} over the total number of possible edges in the OTU): θ_c is the first local minimum after the first mode, OTUs with $\theta < \theta_c$ are singles, and OTUs with $\theta > \theta_c$ are composed.

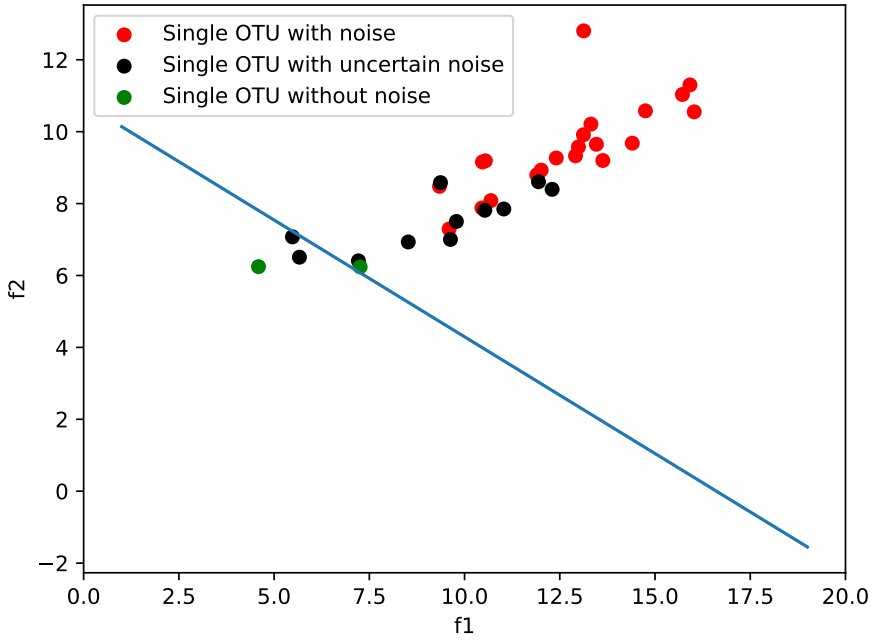


Figure 4: SVM frontier for the Comprian-Pelagic-Autumn sample. The two axes represent the values of the two features used for classification. A dot represents an OTU and is coloured according to the expert classification.

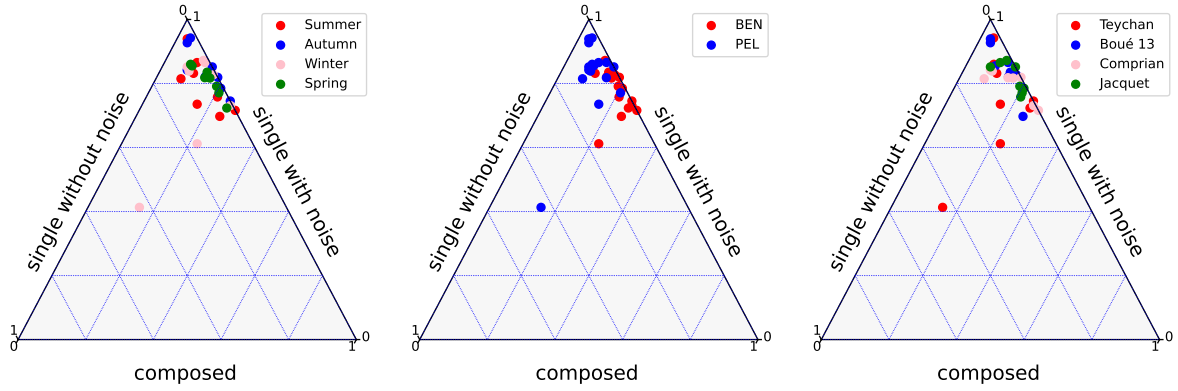


Figure 5: Visualisation of the proportion of composed, single with noise and single without noise OTUs for each sample. Left: dots coloured by seasons, centre: dots coloured by zone, right: dots coloured by location.