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**

Appendix 1: Stochastic Block Model

The Stochastic Block Model, in its weighted version, makes it possible to model a block structure into a dissimilarity array D between n individuals. There are two types of variables in this stochastic model: the observed variables, which are the elements $d(i,j)$ of the dissimilarity array, and the latent variables, which are the block memberships of each individual: $Z_i \in \{1,\dots,K\}$ is the group of individual i. K is the number of blocks. The model relies on two assumptions. First, the Z_i's are independent and their distribution is parameterised by a categorial distribution $\alpha = (\alpha_1,\dots,\alpha_K)$, such that $P(Z_i = k) = \alpha_k$. Second, the dissimilarity between i and j depends only on the blocks of i and j. In this study, we modelled $P(d(i,j) \mid Z_i = k, Z_j = k')$ with a Poisson distribution with parameter $\lambda_{k,k'}$. The K by K matrix Λ such that $\Lambda(k,k') = \lambda_{k,k'}$ is called the connectivity matrix of the model. The model parameters are α and Λ. In practice, only D is available and the objective is to infer α, Λ and the block memberships. Based on D the parameters can be estimated using the Variational EM algorithm (**?**). Each individual is then associated

with the block with the larger a posteriori probability. In our study, we used R package `blockmodels` with default settings for the function `BM_poisson`.

Appendix 2: Estimation of θ density for composed OTU identification

For each sample, the density of θ was estimated using the Kernel density estimation. Kernel density estimation is a non-parametric method to estimate a probability density function by smoothing the data. It relies on the choice of the kernel function and a bandwidth parameter that fixes the amount of smoothing. The bandwidth makes it possible to control the tradeoff between the bias and the variance of the estimated density. In `scikit-learn` Python's machine learning library, we used the KernelDensity estimator, with kernel='gaussian' and bandwidth=0.05.

Appendix 3: Supplementary figures

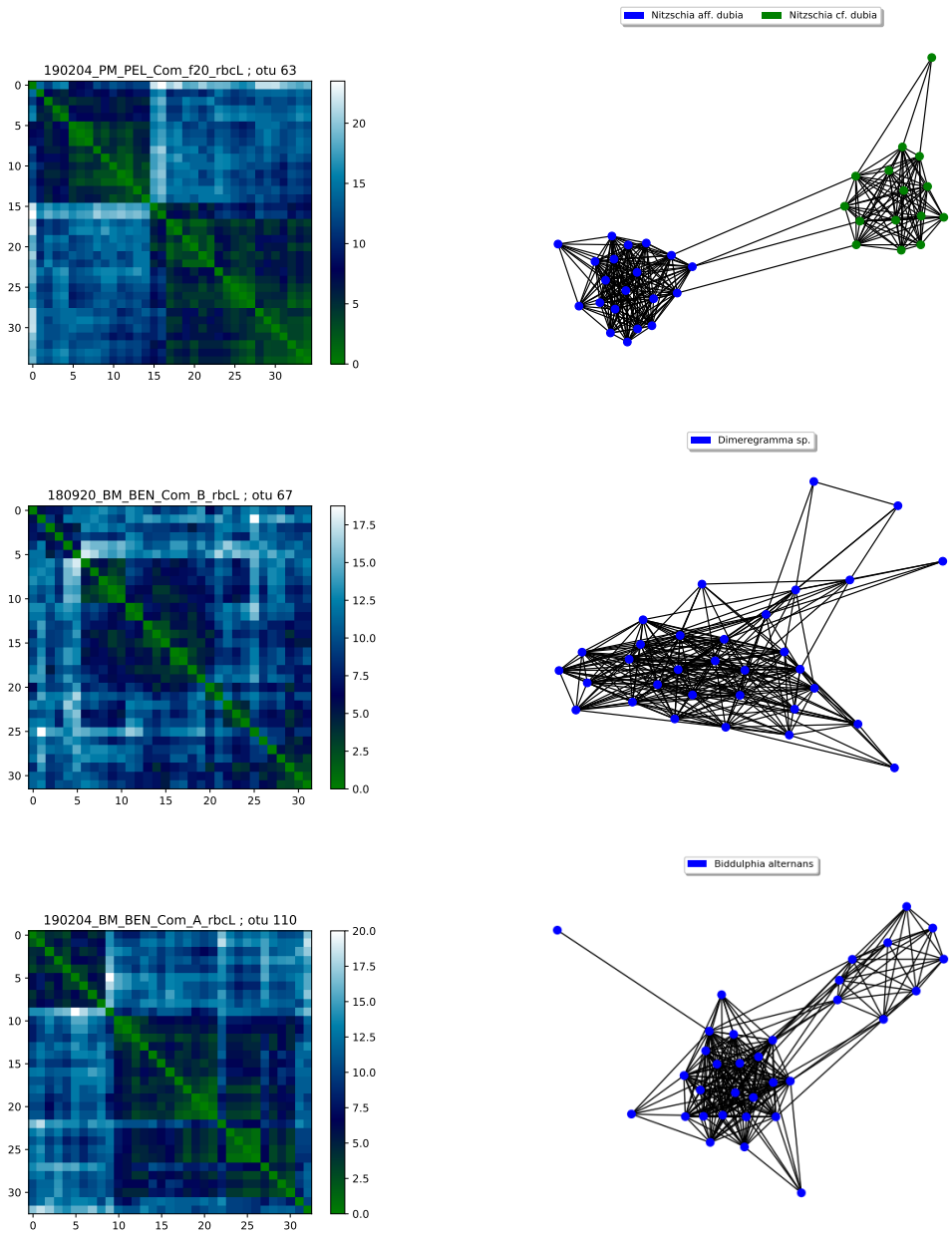

Figure 1: Examples of OTUs fully annotated and categorised as composed. Left column: heatmap of the dissimilarity array D_{otu}, Right column: associated graph G_{otu}. The top line corresponds to the typical situation targeted when identifying composed OTUs, the middle line corresponds to a monospecific OTU with loosely connected sequences, and the bottom line corresponds to a monospecific OTU with distances structured into two blocks.

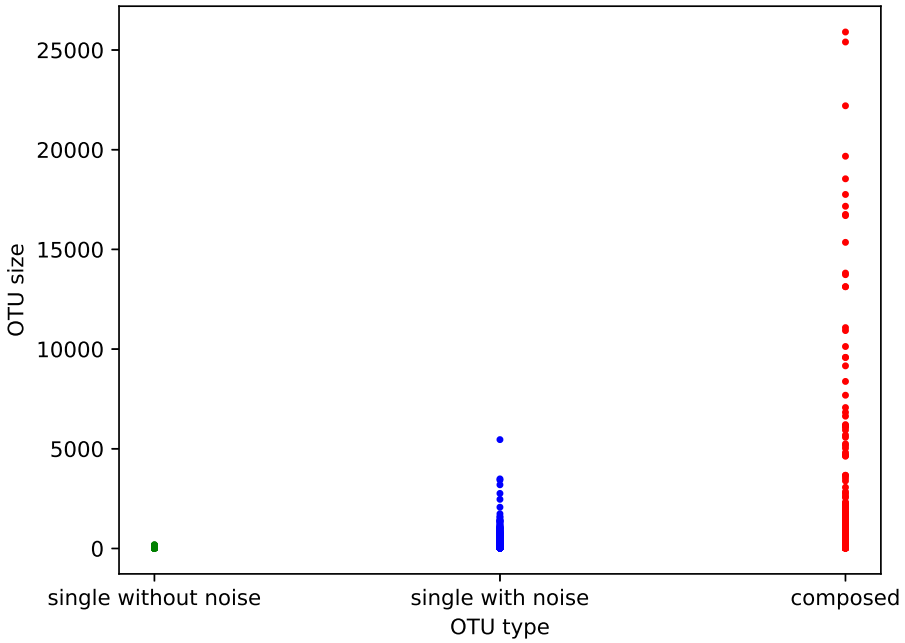

Figure 2: OTU sizes for the three types of OTUs: green dots for single OTUs without noise, blue dots for single OTUs with noise, and red dots for composed OTUs.

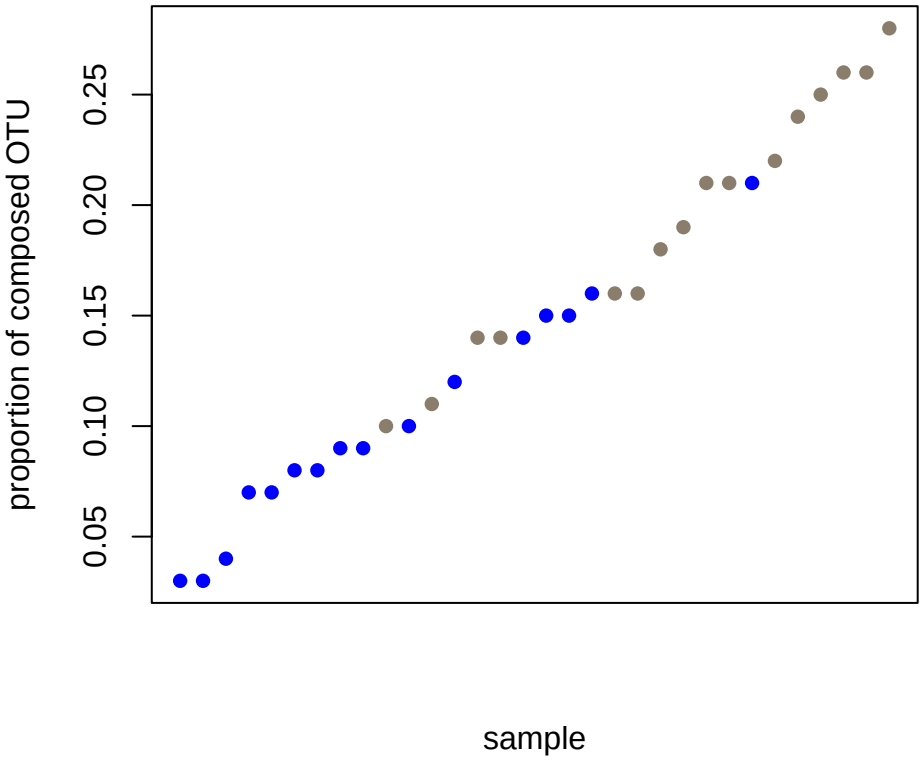

Figure 3: The 32 samples ordered by increasing values of their fraction of composed OTUs. Samples from a benthic environment tend to be associated with larger fractions.

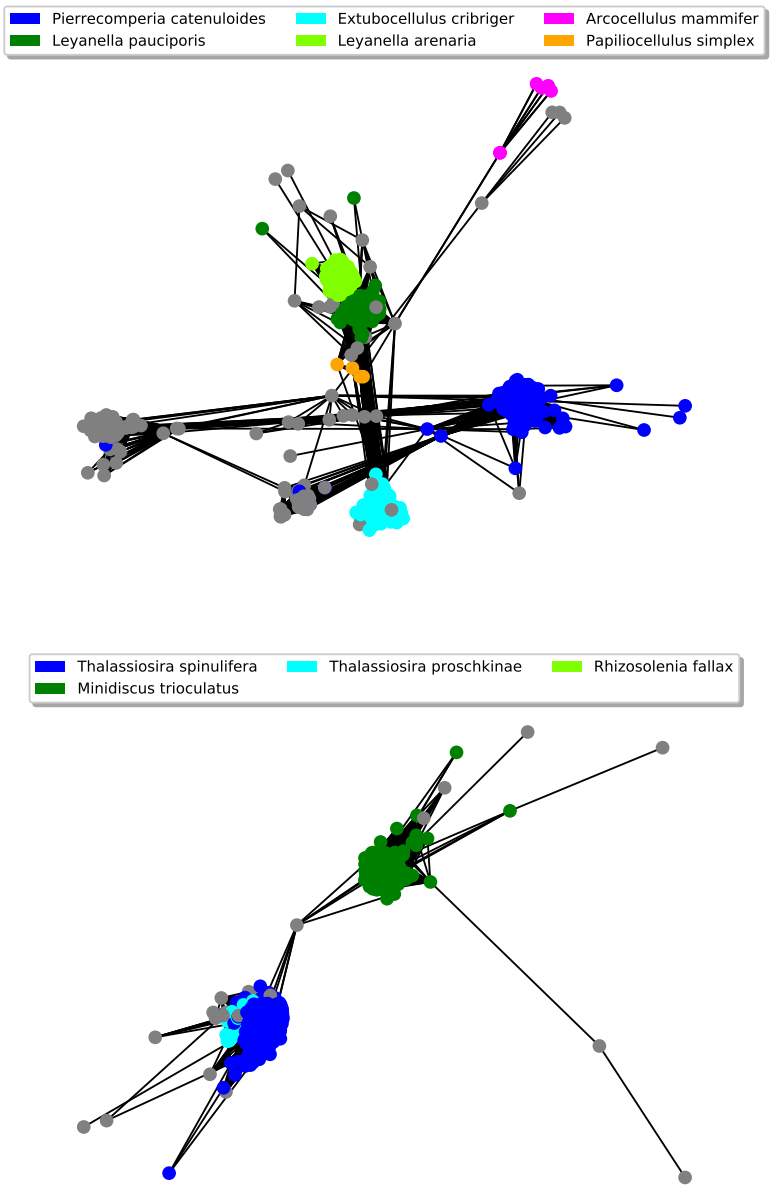

Figure 4: Two different examples of connections between OTUs composed each of two or more components, and taxonomic annotation of sequences (when available). The relationship is not one-to-one between taxa and components. It seems easy in the bottom graph to cut a small number of edges to recover two components, each monogeneric (*Thalassiosira and Mindiscus*, *Rhizosolenia* is not visible because it is behind dots of other genera), whereas such a simple operation seems complicated in the top graph where components made up of different genera are connected by bundles, themselves made up of a large number of links (like *Leyanella* in green and *Extubocellulus* in cyan).

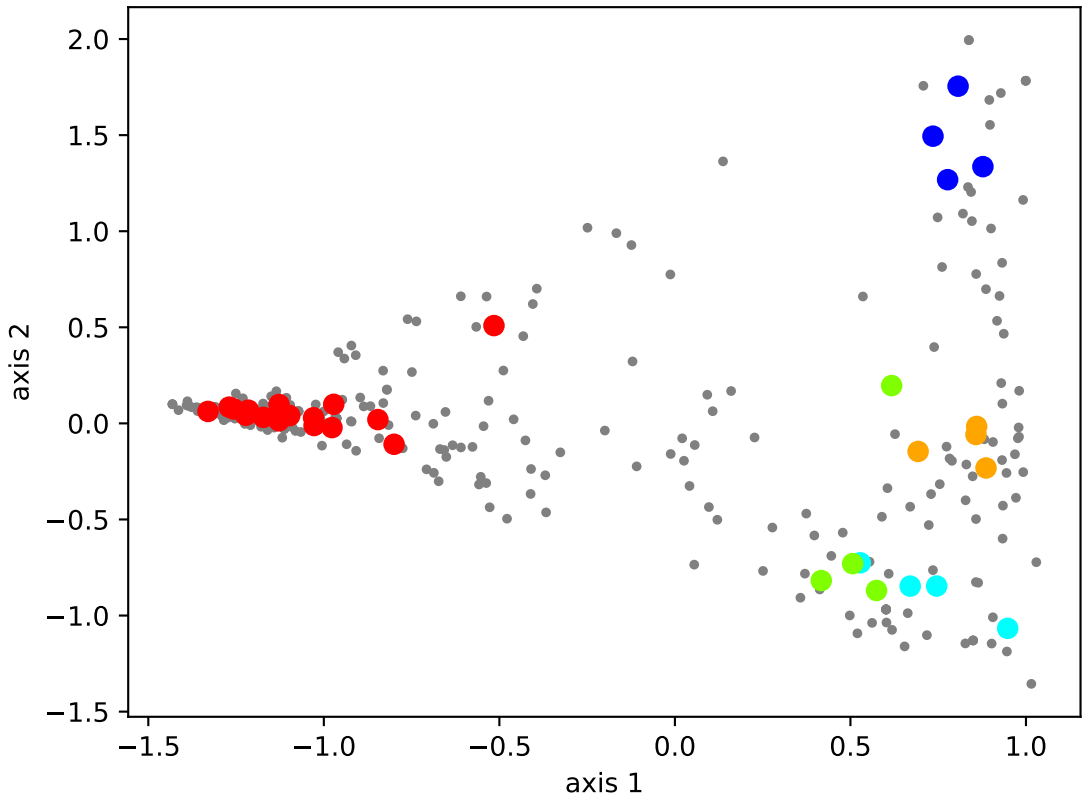

Figure 5: Result of a Correspondence Analysis performed on the contingency table between the 216 species identified by the mapping of at least one sequence and the 32 samples. Small grey dots represent species and large coloured dots represent the samples. The benthic samples are in red. The pelagic samples are coloured by season. We observe that the principal factor which drives the composition in a sample clearly is the opposition between benthic and pelagic samples.

Appendix 4: Supplementary tables

Location	Season	Water column	Number of sequences	Number of OTUs
Teychan	benthic	summer	30372	107
		autumn	29229	106
		winter	20355	85
		spring	28203	94
	pelagic	summer	23593	68
		autumn	23571	52
		winter	25893	92
		spring	26529	50
Bouée 13	benthic	summer	30056	109
		autumn	29022	84
		winter	27670	104
		spring	25176	85
	pelagic	summer	34550	66
		autumn	29956	55
		winter	25244	67
		spring	35729	69
Comprian	benthic	summer	26044	81
		autumn	26766	83
		winter	34011	145
		spring	25643	82
	pelagic	summer	28814	27
		autumn	21462	38
		winter	19224	93
		spring	31965	77
Jacquet	benthic	summer	28490	95
		autumn	27984	121
		winter	32131	125
		spring	27301	115
	pelagic	summer	26719	52
		autumn	30210	27
		winter	20398	27
		spring	31304	48

Table 1: Description of the dataset of 32 environmental samples of diatoms from Arcachon Bay.

Contingency table model

		Automatic	
		Composed OTUs	Single OTUs
Expert	Composed OTUs		
	Uncertain OTUs		
	Single OTUs		

Contingency tables per sample

sample zone	sample season	θ_c	contingency table	sample zone	sample season	θ_c	contingency table
BEN	Summer	0.5454	10 7 1 4 0 85	PEL	Summer	0.3292	8 0 1 0 2 57
BEN	Autumn	0.3880	24 2 1 2 2 75	PEL	Autumn	0.6300	2 1 0 2 0 47
BEN	Winter	0.3180	15 0 2 2 2 64	PEL	Winter	0.2871	12 0 2 1 0 77
BEN	Spring	0.3728	17 0 4 0 3 70	PEL	Spring	0.4648	4 2 0 1 0 43

Table 2: Top: contingency table model. Bottom: Critical θ (θ_c) and contingency tables for composed OTU typing considering the eight samples from the Teychan location.

Contingency table model

		Automatic	
		OTUs with noise	OTUs without noise
Expert	OTUs with noise		
	uncertain OTUs		
	OTUs without noise		

Contingency tables per sample

sample		contingency table	sample		contingency table
zone	season		zone	season	
BEN	Summer	74 0 15 2 0 5	PEL	Summer	38 0 12 2 0 5
BEN	Autumn	66 0 12 0 1 0	PEL	Autumn	42 0 7 0 0 1
BEN	Winter	35 3 15 4 2 7	PEL	Winter	28 3 8 15 2 22
BEN	Spring	55 0 12 1 1 1	PEL	Spring	37 0 6 2 0 1

Table 3: Top: contingency table model. Bottom: contingency tables for the task of identifying single OTUs with and without noise, considering the eight samples from the Teychan location.