Functional Analysis and Classification of Phytoplankton Based on data from an Automated Flow Cytometer

Anthony Malkassian, David Nerini, M. A. van Dijk, Melilotus Thyssen, Claude Manté, Gérald Grégori

Functional Analysis and Classification of Phytoplankton Based on Data from an Automated Flow Cytometer

Anthony Malkassian,[1]* David Nerini,[1] Mark A. van Dijk,[2] Melilotus Thyssen,[3] Claude Mante,[1] Gerald Gregori[1]

[1]Universite de la Mediterranee Aix-Marseille II, Laboratoire de Microbiologie, de Geochimie et d'Ecologie Marines, UMR 6117 CNRS - Observatoire des Sciences de l'Univers (OSU), Centre d'Oceanologie de Marseille, France

[2]Netherlands Institute of Ecology (NIOO-KNAW), Department of Microbial Ecology, Rijksstraatweg 6, 3631 AC Nieuwersluis, The Netherlands

[3]Laboratoire d'Oceanologie et de Geosciences UMR 8187 Maison de la Recherche en Environnements Naturels Avenue Foch 62930 Wimereux, France

*Correspondence to: A. Malkassian, Laboratoire de Microbiologie, de Geochimie et d'Ecologie Marines, Universite de la Mediterranee Aix-Marseille II, UMR 6117 CNRS - Observatoire des Sciences de l'Univers (OSU), Centre d'Oceanologie de Marseille, France

• Abstract

Analytical flow cytometry (FCM) is well suited for the analysis of phytoplankton communities in fresh and sea waters. The measurement of light scatter and autofluorescence properties of particles by FCM provides optical fingerprints, which enables different phytoplankton groups to be separated. A submersible version of the CytoSense flow cytometer (the CytoSub) has been designed for in situ autonomous sampling and analysis, making it possible to monitor phytoplankton at a short temporal scale and obtain accurate information about its dynamics. For data analysis, a manual clustering is usually performed a posteriori: data are displayed on histograms and scatterplots, and group discrimination is made by drawing and combining regions (gating). The purpose of this study is to provide greater objectivity in the data analysis by applying a nonmanual and consistent method to automatically discriminate clusters of particles. In other words, we seek for partitioning methods based on the optical fingerprints of each particle. As the CytoSense is able to record the full pulse shape for each variable, it quickly generates a large and complex dataset to analyze. The shape, length, and area of each curve were chosen as descriptors for the analysis. To test the developed method, numerical experiments were performed on simulated curves. Then, the method was applied and validated on phytoplankton cultures data. Promising results have been obtained with a mixture of various species whose optical fingerprints overlapped considerably and could not be accurately separated using manual gating.

• Key terms

phytoplankton; automated flow cytometry; functional data analysis; multivariate statistics; clustering

IN the euphotic layer of the ocean, oxygenic photosynthesis is responsible for virtually all biochemical production of organic matter, resulting in an annual flux of 4 $\times$ 10^{15} moles of carbon (1). This biological pump constitutes the most important carbon sink at the oceanic scale, keeping the atmospheric carbon dioxide concentration 150 to 200 ppmv lower than it would be without phytoplankton in the ocean (2). Marine primary production represents 45% of the bulk primary production on Earth (1) whereas the marine phytoplankton biomass only accounts for 2% of the global photosynthetic biomass. The high productivity shown by this taxon can be explained by high potential growth rates and short life cycles (3). Biological absorption of carbon is almost entirely realized by small-sized phytoplankton communities ($<$10 μm) under the control of light, nutrients (4), grazing, and viral lysis.

Because of the complex origin of the chloroplast, the phytoplankton is a polyphyletic taxon (5,6). This deep taxonomic diversity induces a highly functional diversity (7): as the result of evolutionary processes that have led to the optimization of light harvesting, different sets of chlorophyll and accessory pigments (carotenoids, phycobiliproteins, etc.) can now be observed (8). Phytoplankton communities are also morphologically diverse, varying in shape and size, as a result of adaptation to

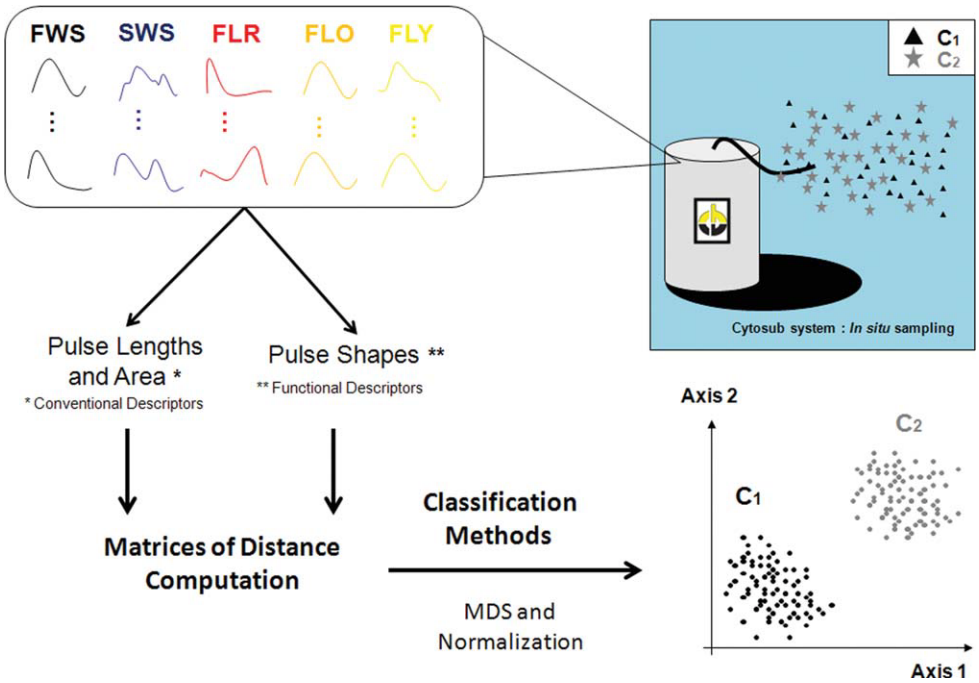

Figure 1. General scheme of the proposed method. (MDS: Multidimensional scaling). Data collected by the CytoSub (top right) lead to a large and complex set of data (top left): five pulse shape signals (forward and sideward light scatter, FWS and SWS, respectively; red, orange, and yellow fluorescences, FLR, FLO, and FLY, respectively). From the raw signals pulse lengths and amplitudes (conventional descriptors) and pulse shapes (functional descriptors) are computed. Based on distance matrices computation, classification methods are then applied in order to find the various clusters (bottom right). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

physical processes (such as hydrodynamics, irradiance), grazing (formation of colonies, extracellular spikes), nutrient uptake (variation of the volume/surface ratio) (9–13).

To understand the complex dynamics within the phytoplankton community and how the biotic and abiotic factors control them, it is necessary to obtain accurate information at various spatial (from the cell to the ocean) and temporal (from hours to years) scales. Taxonomic analysis by optical microscopy has reached its limit as it is time consuming and requires experienced people (14). Consequently, high frequency analysis (typically several times per hour) is still out of reach. Therefore, other faster techniques such as high pressure liquid chromatography (HPLC) or spectrofluorimetry have been developed and successfully applied to aquatic environment studies. However, they only provide a bulk measurement. Analytical flow cytometry (FCM) has become an attractive alternative as it can perform measurements at high frequency and at the single particle level. For each particle passing through a light source (typically one or several laser beams), a set of real values related to light scattering and fluorescence (natural or induced) are recorded.

Although being an ataxonomical method, FCM allows the discrimination of particle clusters within an aquatic sample based on their optical fingerprints (fluorescence signatures and scattering properties). In the last 20 years, flow cytometers have been designed to marine applications (10). This is the case for the CytoSense instrument (Cytobuoy B.V.). A particular feature of this instrument is its capacity to record the full pulse shape along each particle for both scatter and fluorescence signals (15). This way of scanning cells sequentially provides more information on the morphological variability within the phytoplankton community. By monitoring the phytoplankton clusters at high frequency, unexpected dynamics have been revealed, with respect to strong wind events and physicochemical conditions (16). Additionally, studies by Thyssen et al. demonstrated the capability of this flow cytometer to identify groups that were not discernable using more conventional instruments (16,17).

After collecting data with the CytoSense, the usual approach is to reduce each pulse into classical descriptors (inertia, fill factor, asymmetry, number of peaks, length, etc) using the Cytoclus© software (Cytobuoy B.V., The Netherlands). Data are displayed by means of scatterplots and histograms that facilitate the visualization and identification of particle clusters defined by similar optical properties. The clusters are usually created by manually drawing and combining regions (gating). This way of defining arbitrary groups is not always objective and can lead to errors, in particular when clusters overlap, shift positions or when different pulse shapes lead to similar classical descriptors. The aim of this study is to provide a observer-independent and consistent method to automatically analyse the data and define clusters (Fig. 1). Despite the large quantity of approved tools available for multivariate analysis, few researchers have worked on the automation of FCM data processing. The major advances have been obtained with Artificial Neural Networks (18–23), mixture

models approach (24,25,26) or discriminant analysis (27,28). As longitudinal information related to the particle morphology clearly appears through the pulse shapes, one of our goals is to verify to what extent the statistical analysis of functions (29) i.e., the shapes of the full raw pulses can offer an advantage over using only usual descriptors. The shape, length and area of the various recorded curves have therefore been chosen as descriptors and then used in this study. Several tests have been performed on simulated pulses to test the efficiency of the clustering method. The model has then been validated on biological data collected from phytoplankton cultures.

Materials and Methods

The Autonomous Flow Cytometer

The flow cytometer used in this study is a CytoSub operating in bench top mode (Cytobuoy B.V., The Netherlands). It has been designed to analyze phytoplankton in situ, over a large size range (up to 800 μm in width and a few mm in length). The sample is pumped by a peristaltic pump at a flow rate between 0.4 and 9.6 μL s^{-1}. The sheath fluid is made of 0.2 μm filtered seawater supplemented by ~1% formaldehyde solution in order to prevent any bacteria development. The sheath fluid and the sample do not mix together as they behave as laminar flow until the outflow of the flow cell. After this point they are mixed together and then filtered over 0.45 and 0.2 μm porosity filters to recycle the sheath fluid for the next analyses. In the flow cell, each particle is intercepted by a blue laser beam produced by a solid-state laser (Coherent Saphyre, 488 nm, 15 mW). The forward angle light scatter signal (FWS) is collected via a PIN photodiode. The sideward angle light scatter (SWS), the red (FLR, 734–668 nm), orange (FLO, 668–601 nm) and yellow (FLY, 601–536 nm) fluorescence signals are separated by a concave holographic grating and collected on Hybrid Photomultiplier tubes. FWS was used as trigger signal for data recording. All signals are recorded at a frequency of 4 MHz (i.e., four times per microsecond) and stored in a data grabber before being transferred to the computer. Particles flow through the 5 μm wide focal point of the laser beam at a flow rate of 2 m s^{-1}, and therefore the pulses of a 1 μm particle approximately contain 12 points. In order to monitor the stability of the instrument, several fluorescent microsphere solutions have been used for quality control.

Programming of the algorithms (statistical analysis) and displaying of the data have been performed using R software (30). For the convenience of readers who would like to repeat this work, the R codes and datasets are available on the website http://www.com.univ-mrs.fr/~malkassian/ (Anthony Malkassian Home Page). The clustering methods have been adapted to handle the fingerprints, considering the most important features: shape, length and area under the pulse (AUP). The method (Fig. 1) entails computing the distance matrices for the three descriptors. The three resulting matrices are then combined into a single one, called the global distance matrix. From the patterns of similarity thus obtained, clustering methods (31) are applied. In addition, the data can also be represented in a reduced dimensional space, by using multidimensional scaling methods. This enables data to be visualized, which wasn't possible before as each individual event was defined by five curves.

Computation of the Distance Between the Conventional Descriptors

For each particle (individual), extraction of AUP and Length matrices is performed with CytoClus© software and the values are logarithmically transformed. For a sample composed of n individuals, both the curve length (i.e., width of the signal) and AUP are recorded as real and positive values on p channels, forming a $n \times 2p$ table of observations (here $p = 5$ real variables). The curve length and AUP are separated in two blocks: block A with p real variables for AUP values, and block L with p real variables for the length of the curve. Similarities within each block are measured by means of the Euclidean distance. For two individuals i and j the distance between AUP values and between lengths (respectively A and L) is defined as:

$$d^2(\mathbf{x}_i^k, \mathbf{x}_j^k) = \left\|\mathbf{x}_i^k - \mathbf{x}_j^k\right\|^2 = \sum_{l=1}^{p} (x_{il}^k - x_{jl}^k)^2, \; k = A, L. \qquad (1)$$

For a sample one can then build distance matrices D_L and D_A, which are $n \times n$ matrices where entries are respectively the distances between the lengths and the AUPs of two individuals as defined in equation (1).

The Distance Between Functional Descriptors

Let us consider, for instance, a collection of FWS signals $E = \{y_1(t),\ldots,y_n(t)\}$ collected in a marine sample. Each curve is a sampled function where the argument t varies in a compact interval τ of $\mathbb{R}$. The function takes values in a Hilbert space $\mathcal{H}$ of functions on τ where $\langle .,. \rangle_{\mathcal{H}}$ denotes its inner product. The distance between two random curves y_i and y_j is the Hilbertian distance between the functions:

$$d^2(y_i, y_j) = \|y_i - y_j\|_{\mathcal{H}}^2 = \langle y_i - y_j, y_i - y_j \rangle_{\mathcal{H}} = \int_\tau (y_i - y_j)^2 dt. \qquad (2)$$

However, stretched by the fluid acceleration, particles are supposed to become orientated along their longest axes, parallel to the flow direction (32). It is therefore possible for a non-symmetrical particle to cross the laser beam both ways, which leads to the recording of two opposite fingerprints. In this way, similar particles randomly lined up in the fluid stream can provide different pulse shapes while having the same optical properties. It is necessary to describe this process in terms of distance computation, as applied in Khelil et al. (33). A distance called invariant to orientation is then computed, considering that the rotational effect comes down to a 180° rotation of the pulse with respect to the ordinate axis.

$$D_{\text{inv}}^2(y_i, y_j) = \min\{d^2(y_i, y_j), d^2(y_i, y_j^*)\}$$

where y_j^* denotes the symmetrical version of y_j with respect to the ordinate axis. For a sample composed by n objects, one

can build the $n \times n$ pulse shape distance matrices for the five channels: $D_{\text{inv}}^{\text{FWS}}$, $D_{\text{inv}}^{\text{SWS}}$, $D_{\text{inv}}^{\text{FLR}}$, $D_{\text{inv}}^{\text{FLO}}$, $D_{\text{inv}}^{\text{FLY}}$.
The previous distance matrices can then be combined to form the $n \times n$ global distance matrix:

$$D_{\text{global}} = \Pi_1 D_L + \Pi_2 D_A + \Pi_3 D_{\text{inv}}^{\text{FWS}} + \Pi_4 D_{\text{inv}}^{\text{SWS}} + \Pi_5 D_{\text{inv}}^{\text{FLR}} + \Pi_6 D_{\text{inv}}^{\text{FLO}} + \Pi_7 D_{\text{inv}}^{\text{FLY}}$$

where $\Pi_1,\ldots,\Pi_7$ are arbitrary positive weights.

From the Raw Pulse to the Functional Pulse Shape Descriptors

Two problems remain unsolved: (i) How to generate the functional pulse shape descriptors from the raw pulses (output data of the CytoSense) to compute distances between pulse shapes and (ii) how to explicit the way to compute a distance invariant to orientation.

The distance in equation (2) involves integral computation. A comfortable way to approximate this distance is to consider that any function can be expressed in terms of linear combinations of known basis functions (31,34) such that:

$$y(t) = \sum_{k=0}^{K} c_k \phi_k(t) = \mathbf{c}'\Phi(t)$$

where K denotes a fixed number of basis functions, $\Phi(t) = (\phi_0(t),\ldots,\phi_K(t))'$ the K-vector of basis functions, and $\mathbf{c} = (c_0,\ldots,c_K)'$ the K-vector of associated coefficients. We have chosen a Fourier basis decomposition, because the pulses are periodic functions.

Moreover, the Fourier functions form an orthonormal basis such that $\langle \phi_k, \phi_l \rangle_{\mathscr{H}} = \delta_{kl} = \begin{cases} 1 & \text{if} \quad k = l \\ 0 & \text{else} \end{cases}$, and $||\phi_k||^2_{\mathscr{H}} = 1$, $k = 1,\ldots,K$, $l = 1,\ldots,K$, which simplifies the distance computation.

In practice, a raw pulse y is recorded in the form of a discrete set of m points $\{y(t_j), j = 1, \ldots, m\}$. The coefficients estimation is then performed by least squares regression when minimizing the following criterion:

$$\text{SSE}(c_1, \ldots, c_K) = \sum_{j=1}^{m} \left[y(t_j) - \sum_{k=1}^{K} c_k \phi_k(t_j) \right]^2.$$

The matrix form is given by:

$$\text{SSE}(\mathbf{c}) = (\mathbf{y} - \Phi\mathbf{c})'(\mathbf{y} - \Phi\mathbf{c})$$

where $\Phi = \{\phi_k(t_j), k = 1, \ldots, K, j = 1, \ldots, m\}$. This criterion is minimized by making the first derivative equal to zero:

$$\frac{\partial \text{SSE}(\mathbf{c})}{\partial \mathbf{c}} = 2\Phi'\Phi\mathbf{c} - 2\Phi'\mathbf{y} = 0.$$

The least squares estimate $\hat{\mathbf{c}}$ of $\mathbf{c}$ is solution of the latter equation: $\hat{\mathbf{c}} = (\Phi'\Phi)^{-1}\Phi'\mathbf{y}$.

By construction, the number of coefficients cannot be higher than the number of sampled points. However, all the functions are sampled on a mesh of equally spaced points (at a fixed frequency of 4 MHz). The maximum number of coefficients (i.e., the dimension of the whole basis) is conditioned by the length of the shortest particle crossing the laser beam. In most cases, this number, while controlling the global smoothness of the curve, is not sufficient to describe the whole variation of complex particle shapes. For this reason, before estimating the coefficients, we have proposed a regularization step provided by a cubic smoothing spline (35) in order to obtain a smooth version of y that can be valued at any $t \in \tau$.

A pulse can then be considered such that: $y(t) = g(t) + \varepsilon(t)$, $t \in \tau$ where $g(t)$ is the smooth version of $y(t)$ for which the expression

$$\frac{1}{m}\sum_{j=1}^{m}(g(t_j) - y(t_j))^2 + \lambda \int_{\tau} (g''(u))^2 \, du$$

is minimum. The residual variation $\varepsilon(t)$ can be referred to as noise and will be considered as negligible. The smoothing parameter λ controls the tradeoff between the smoothness of the solution as measured by the norm of the second derivative of g: $\int_\tau (g''(u))^2 du$ and the empirical mean squares error of the data computed by $\frac{1}{m}\sum_{j=1}^{m}(g(t_j) - y(t_j))^2$. The parameter λ is commonly estimated by cross-validation (31). Once the function $g(t)$ has been found, the sampling mesh can be refined since the spline can be valued at any $t \in \tau$. This allows to increase the number of basis functions for curves having the most complex shape variations. This also insures that variations in shapes for less complex pulses will be handled as well. Moreover, one can use more points for regression than the sampled data values ($K > m$) and still achieves a good fit.

As defined by Dryden (29), the shape is all the geometrical information that remains when location, scale and rotational effects are filtered out from an object. To measure the similarities between pulse shapes, length is normalized, and it is imposed that $t_1 = 0$ and $t_m = 1$ (i.e., $\tau = [0,1]$) with $y(t_1) = 0$ and $y(t_m) = 0$. The curve intensity is then normalized dividing all the c_k by c_0. This coefficient represents the area under the pulse: $c_0 = \frac{1}{T}\int_0^T y(t)dt$, here $T = 1$.

By following these steps, we get a registered version of each curve. The distance between shapes of y_i and y_j is then easily computed by comparing their coefficients:

$$d^2(y_i, y_j) = \left\| y_i - y_j \right\|^2_{\mathscr{H}} = (\mathbf{c}_i - \mathbf{c}_j)'\Phi'\Phi(\mathbf{c}_i - \mathbf{c}_j) = \left\| \mathbf{c}_i - \mathbf{c}_j \right\|^2_{\mathscr{H}} \quad (3)$$

where $\mathbf{c}_i$ and $\mathbf{c}_j$ are the K-vectors of the normalized coefficients.

The Computation of the Distance Invariant to Orientation

In order to explicit the invariance to orientation, we remind that the Fourier basis decomposition provides a sum of sines and cosines:

$$y(t) = \sum_{k=0}^{K} c_k \phi_k(t) = c_0 + c_1 \sin(\omega t) + c_2 \cos(\omega t) + c_3 \sin(2\omega t) + \cdots + c_K \cos(K\omega t).$$

The formula of $y^*(t)$, the symmetrical version of $y(t)$ with respect to the ordinate axis, is straightforwardly obtained by inverting the sign of each sine function:

$$y^*(t) = \sum_{k=0}^{K} c_k \phi_k^*(t)$$

where

$$\phi_k^*(t) = \begin{cases} \phi_0^*(t) = 1 \\ \phi_{2r-1}^*(t) = -\sin(r\omega t) \\ \phi_{2r}^*(t) = \cos(r\omega t) \end{cases}$$

Once the decomposition into the Fourier basis has been achieved for every observation, it is easy to compute the overall distance matrix D_{global}.

The Classical Multidimensional Scaling

Considering the relationship between n individuals through the global distance matrix D_{global} computation, we seek to represent the set of observations in a reduced dimensional space $\mathbb{R}^d$, with typically $d = 2$ or 3. To obtain the coordinates of the individuals, a multidimensional scaling method (36), also known as Torgerson scaling, has been used. The relationships between individuals is specified using distance computation. Supervised classification methods are applied to define the various groups.

The Classification Method

The clustering objective is to classify the data in q groups $\{C_1, \ldots, C_q\}$ with a fixed value of q and $q \ll n$ (with n the number of objects), so there is the strongest similarity between objects belonging to the same group. Following the work of Kaufmann and Rousseeuw (37) two clustering methods have been tested and compared: Partition Around Medoids[1] and Fuzzy analysis[2] algorithms.

[1] The partition around medoids (PAM) algorithm proceeds in two steps: In the Build step q objects are sequentially selected, in order to be used as initial medoids. In the Swap step, the aim is to reduce an objective function (for instance, the intergroup variance) by interchanging an initial medoid with an unselected object. This step is recursively repeated until a stopping rule is applied (the objective function can no longer be decreased).

[2] In opposition to hard clustering method, where each individual is assigned to one class (i.e., a clear-cut decision), in FANNY (Fuzzy analysis) method (37), each individual can belong to more than one class. The degree of belonging to different classes is quantified by means of membership coefficients, ranging from 0 to 1 for each class and summing to one over the whle set of classes.

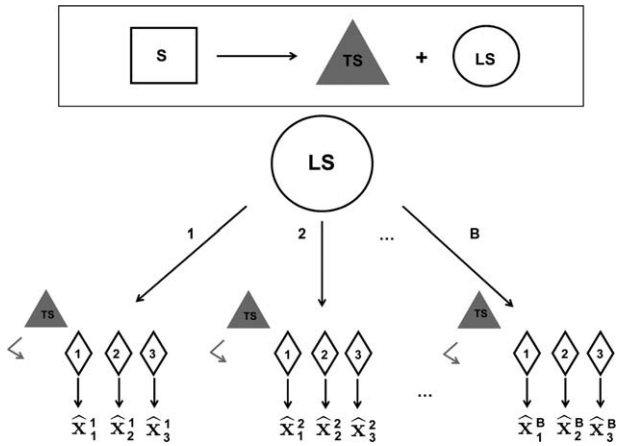

Figure 2. The bootstrap aggregated predictor method (bagging). The sample S is splitted in: a test set (TS) containing about 1/3 of the data and a learning set (LS) with the remaining data. The bootstrap samples LS^b, $b = 1,\ldots,B$ are replicated datasets each consisting of card (LS) individuals randomly drawn from LS with replacement. The partition in j classes is constructed over each bootstrap sample and the centroids (i.e., representative objects): $\{\hat{\mathbf{x}}_j^b, j = 1, \ldots, p\}$ are computed for each class. A class predictor is constructed and the classification success is evaluated on the test sample TS. In this shematic example, the number of classes j is fixed to $j = 3$.

The Optimal Partition Estimator

The major questions to address are: how to determine the optimal number of clusters and what is the best clustering method? It is often difficult to identify clusters that are overlapping and with various sizes and shapes (25). However, when the number of classes is unknown, it is necessary to get a criterion that evaluates the partition validity and enables the selection of an appropriate number of groups. The silhouette coefficient (SC) is a measure of the amount of clustering structure that has been discovered by the classification algorithm (37). This is a dimensionless value computed over all possible partition numbers. The highest value provides an appropriate partition number for the given data set.

The Test of Partition Robustness

The robustness of the clusters constituting the sample S has been tested through the construction of a bootstrap aggregated predictor (called bagging, Fig. 2). This method consists of combining multiple versions of the classification model based on bootstrap samples of the data to test the effects of sampling changes on the structure of clusters (38,39).

Phytoplankton Cultures

For a first experiment, various phytoplankton cultures from the Culture Collection Yerseke (CCY, NIOO Centre for Estuarine and Marine Ecology, Yerseke, The Netherlands) have been used to apply the method described above on real data. They are obviously not species that normally would be found together in a natural sample as they originate from fresh, brackish, and marine waters. Actually, these strains have been chosen as (i) their optical properties lead to different degrees of overlap according to the conventional flow cytometry

Table 1. Classification results on a sample of 2,000 individuals composed with a mixture of 20 phytoplankton cultures (100-fold bagging with fuzzy clustering method)

	CLASSIFICATION SUCCESS RATE				
	USUAL DECRIPTORS		USUAL DESCRIPTORS AND PULSE SHAPES		
SPECIES	MEAN	SD (%)	MEAN	SD (%)	*T*-TEST (T)
Anabaena cylindrica	0.490	0.396	0.685	0.383	−1.994*
Ankistrodesmus acicularis	0.996	0.017	0.902	0.240	2.313**
Aphanizomenon sp.	0.970	0.079	0.928	0.194	1.240
Chaetoceros muelleri	0.991	0.042	0.951	0.160	1.342
Chlorella sp.	1.000	0.000	1.000	0.000	–
Ditylum brightwellii	0.149	0.105	0.177	0.069	1.314
Emiliania huxlyei	0.982	0.111	0.895	0.278	1.864
Gloeothece sp.	0.077	0.080	0.197	0.248	−2.715**
Hemiselmis sp.	0.000	0.000	0.855	0.325	−14.184***
Isochrysis sp.	0.889	0.311	0.946	0.223	−0.917
Melosira sp.	0.450	0.429	0.597	0.452	−1.332
Monoraphidium sp.	1.000	0.000	1.000	0.000	–
Nodularia sp.	0.077	0.131	0.078	0.138	−0.0134
Pavlova sp.	0.000	0.000	0.000	0.000	–
Porphyridium sp.	0.294	0.223	0.386	0.481	−1.0203
Pseudanabaena sp.	1.000	0.000	0.925	0.238	2.029*
Pediastrum sp.	1.000	0.000	1.000	0.000	–
Rhodomonas sp.	1.000	0.000	1.000	0.000	–
Skeletonema costatum	0.411	0.281	0.278	0.406	1.624
Thalassiosira pseudonana	1.000	0.000	1.000	0.000	–

Comparison between classification successes obtained using usual descriptors versus usual descriptors and pulse shapes (Student t-test, *: $p < 0.05$, **: $p < 0.01$, ****: $p < 0.001$).

descriptors (see Supporting Information), (ii) they were available in the collection of the NIOO. Mixing these cultures allowed not only to get a mixture of sizes and shapes, but also various pigment contents as the species belong to several taxonomic groups. A total of 20 different strains of phytoplankton (Table 1) were selected and grown as mono-cultures on their corresponding nutrient-rich media in the lab at room temperature under a 14:10-h light dark cycle, Fresh culture material was analyzed with the CytoSub flow cytometer, working in bench-top mode in the laboratory (i.e., CytoSense). In a second experiment, we focused on the flow cytometry data of two specific strains with very similar fingerprints. One being toxic and the other not:

- *Amphidinium carterae*, a dinoflagellate well known for his toxicity and is responsible for a foodborne disease called Ciguatera (also known as CFP[3]).
- *Tetraselmis tetrathele*, a eurythermic Prasinophyceae found in the temperate/tropical regions (40).

Results

The Numerical Experiments: Simulation of Cytometric Pulses

The choice to work on simulated curves as a preliminary step was driven by the need to test the efficiency of the method. The numerical experiments on simulated cytometric pulses allow creating controlled samples where the variability and abundance of each group can be easily tuned. We aim to classify functions for which the shape, or at least the family of shapes, is known in advance. We propose to construct a learning sample L composed of n curves belonging to p classes. The number p of classes is firstly fixed to the value $p = 6$. In the same way, the proportion of individuals belonging to class j into the sample L is fixed to 1/6th, but other ratios will be tested later on. A class C_j, $j = 1, \ldots, 6$ is characterized with a reference function f_j, $j = 1, \ldots, 6$, which is a density:

$$f_j(x) \geq 0, \quad \int f_j(x)dx = 1, \quad j = 1, \ldots, 6 \quad \forall x \in \mathbb{R}.$$

Choosing probability densities provides the opportunity to deal with positive curves such as those recorded by the CytoSub flow cytometer.

These curves are already normalized. Finally, the construction of a random sample of such curves is easily conducted thanks to the strong connections between a sample of realizations of a random variable and the associated density. The construction of class C_j containing $n_j = n/6$ individuals is achieved by kernel density estimation of n_j random samples of size m drawn from the reference density f_j associated to that class. Figure 3 shows how the kernel density estimate approximates a reference density (here a centered Gaussian curve).

[3]Ciguatera Fish Poisoning.

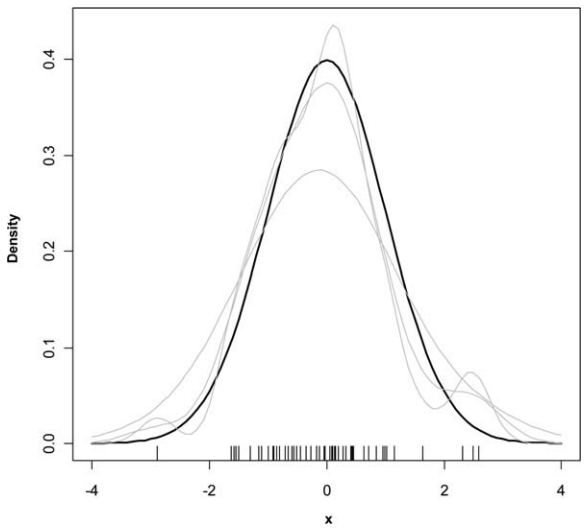

Figure 3. Example of three kernel estimates (grey bolded curves) of a centered Gaussian density (black bolded curve). The roughness of the estimates is controlled by a smoothing parameter. Each density is estimated on the same dataset composed of 50 values of *x* randomly drawn from the Gaussian distribution.

Figure 4 displays an example of reference densities (black lines) and their realizations (gray lines) for $p = 6$ classes. The roughness of the estimated curve is controlled both by the number m of sampled points and by a bandwidth parameter which in our case can be chosen by cross-validation. Thus, changing the sample of points randomly drawn according to a reference density provides a new estimated curve which can be considered as a random version of that reference function. This is how the different classes are constructed (see Appendix for more details).

From the mix of these six families of curves, we expected to find four real classes. The optimal SC is equal to 0.68 for a partition in 6 groups when the simple distance is computed (Fig. 5, left panel). When the distance invariant to orientation was computed, the optimal partition provides four groups, with a SC value of 0.8 (Fig. 5, right panel).

The fuzzy clustering method was performed on the latter distance matrix and four groups of interest were found without error. The bagging test of robustness provides a classification success of 100%. However, this example can be considered too simple as each family contains the same number of individuals. Moreover, within each class the individuals present similar fingerprints (i.e., high homogeneity intra classes). This case is unlikely to occur in natural samples, thus the next step consists of altering the test sample by modifying the variability among the pulse shapes and the relative abundance of each class.

In Figure 6, we started with close relative abundances of 16.5% for each class. At the same time the intra class variability was kept weak with a high number of observations m. A destabilization was created in 50 steps by successively increasing the relative abundance of C_6 and decreasing the other five classes. This finally led to a proportion of 75% for C_6 and 5% for the others. The classification success that has been measured at each step with the bagging method was found to decrease during the alteration from about 100% to about 80%. This was due to the clustering method which did not always converge to the right partition, but rather introduced a splitting within the most abundant class C_6. As the number of wrong splits increased at each step the misclassification rate also increased. At the end of the destabilization, only the individuals belonging to the class C_6 are classified with success.

The second alteration consisted of adding noise to the curves. Figure 7 displays an example with either a low or high degree of heterogeneity. The results of the clustering methods have been compared. On the two-dimensional dotplots created from the distance matrices by multidimensional scaling the data structure can be observed. An increase in the intra class heterogeneity resulted in an increase of their overlap and a decrease in the classification success regardless the clustering method. The shapes become less detectable. However, the clustering methods did not react with the same intensity at this overlap. The fuzzy method presented a better classification success when the overlap was high.

Experiments Using Phytoplankton Cultures

For the first experiment, datasets coming from 20 different strains have been selected and artificially mixed. Particles with a very low red fluorescence intensity (total FLR below 50 mV) were removed from each dataset. These particles are considered as background noise (cellular debris, dead cells or bacteria contamination of the culture). The mixture dataset was constructed by randomly selecting 2×10^3 individuals from each of the 20 individual datasets. The classification results are presented in Table 1. Please note that these results can present slight variations due the construction of the bagging test sample which depends on the random sampling. A first analysis was performed with the usual descriptors only (length and AUP) and in a second analysis the shape descriptors were also taken into account. The classification method, based on conventional descriptors only, often mixed up *Hemiselmis* sp. with *Porphiridium* sp. Considering pulse shape descriptors led to a better distinction. The same was true for *Skeletonema costatum*, *Melosira* sp., and *Ditylum brightwellii* clusters. Thanks to the discriminant information included in the shape, in most cases the method was able to distinguish strains for which clusters solely defined by conventional descriptors overlapped considerably. For instance, the classification success strongly increased for *Hemiselmis* sp. (from 0 to 85.5 %). However the classification success for species *Nodularia* sp., *Pavlova* sp. and *Gloeothece* sp. remained very weak even if the shape decriptor was considered. A slight degradation of the classification success has been sometimes observed for *Ankistrodesmus acicularis* and *Pseudanabaena* sp.

The second experiment consisted of focusing on two specific strains with very similar fingerprints (Fig. 8): the toxic *Amphidinium carterae* and nontoxic *Tetraselmis tetrathele*. The goal was obviously to distinguish the toxic species from the other one. The classical cytometric analysis performed with the usual descriptors showed a high level of overlap between

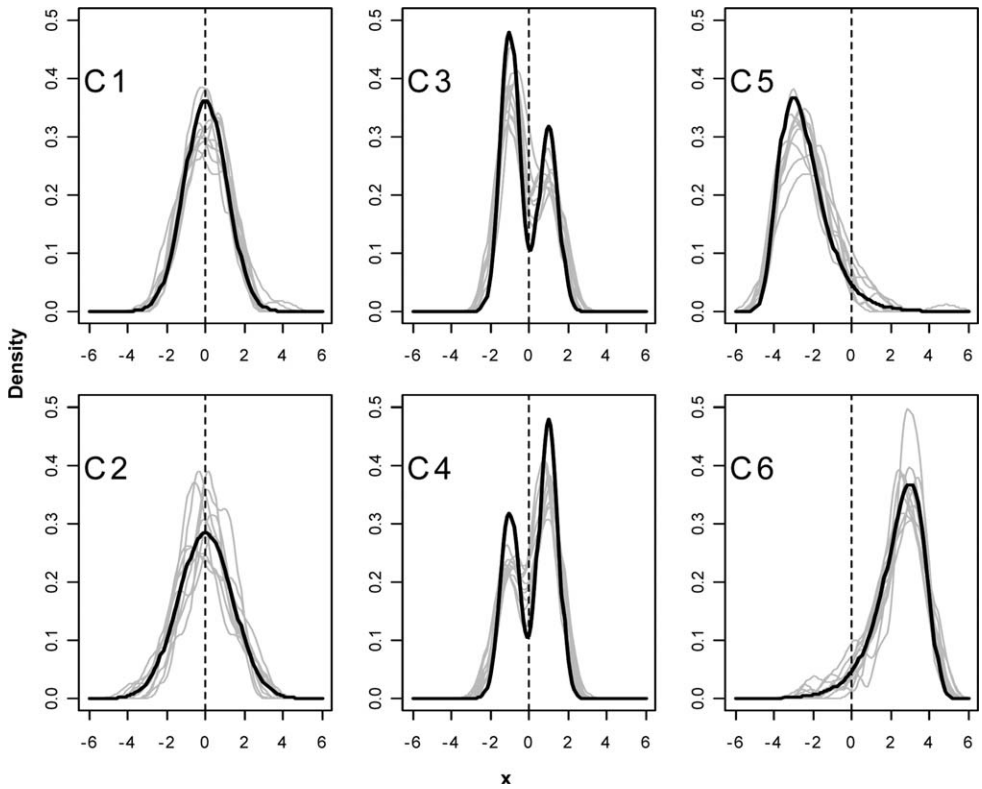

Figure 4. The six families of curves including the reference density for each class (black bolded curves) and a sample of kernel density estimates. Each random function in class C_j is estimated using 50 points randomly drawn from the reference density f_j.

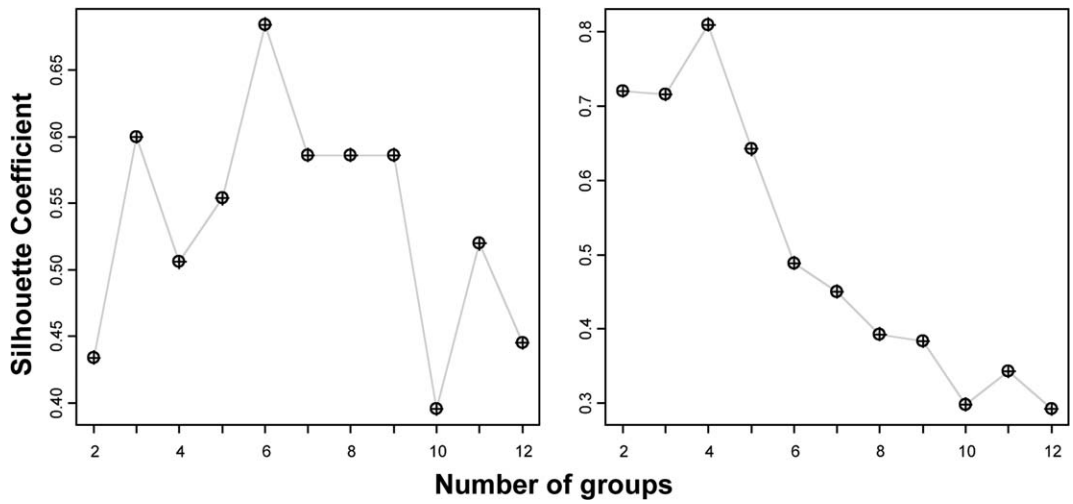

Figure 5. The Silhouette Coefficient (SC) plots, where the maximum value indicates the optimal number of clusters. On the left panel, the invariance by symmetry is not applied to compute the distance matrix. On the right panel, the invariance by symmetry is taken into account for the distance matrix computing. The real number of classes (4) is retrieved in that case.

their respective flow cytometric clusters (Fig. 8). The two species were not manually discernable with the CytoClus software. Nevertheless, their pulse shapes are slightly different and could contain some discriminatory information. Three analyses were therefore performed: (i) First by taking into account the usual descriptors only, then (ii) taking into account the pulse shape only and finally (iii) combining both descriptors.

The results obtained from 200 bagging samples are presented in boxplot form (Fig. 9). When the analysis was based on the length and AUP descriptors only, about 71% of the

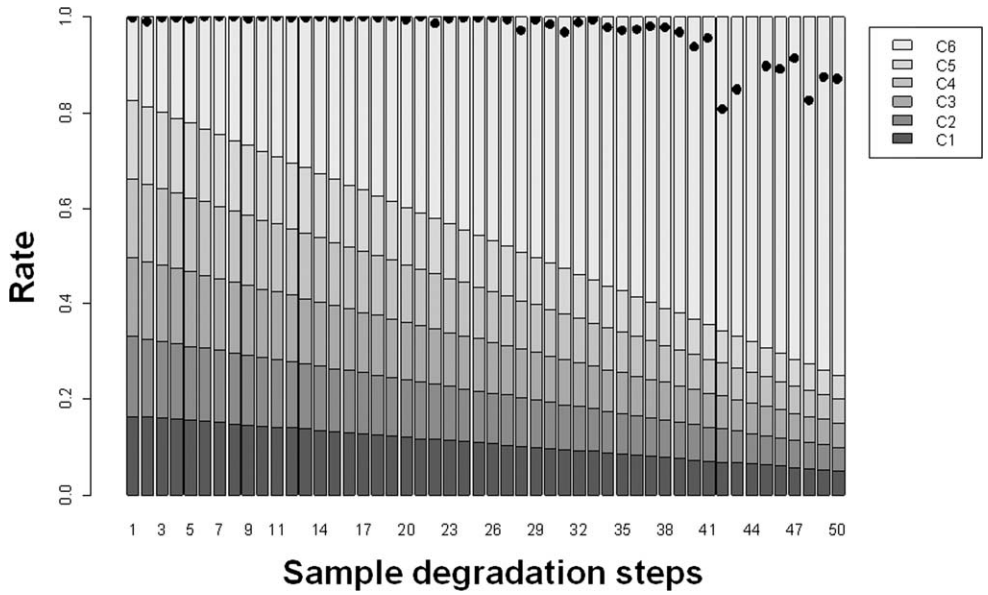

Figure 6. The modification of abundances experiment: The black dots, representing the mean classification success over 20-fold bagging, show the decrease of the classification success by modification of the number of individuals in each class. These results are obtained in 50 steps, the alteration of the relative abundance of class C_1 to C_6 is represented by the barplots.

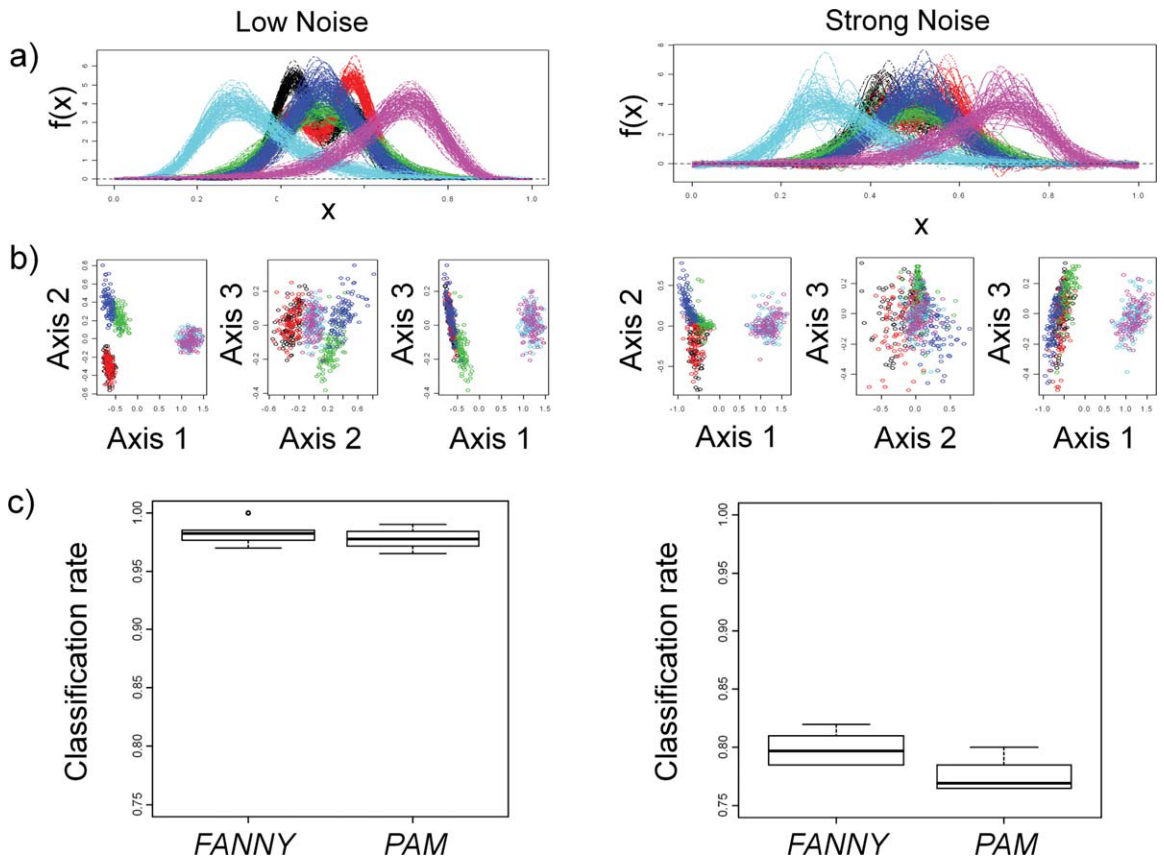

Figure 7. Effect of noise over the classification success. Starting from a sample of six families of curves corrupted by noise (**a**) and its 3D display obtained from the distance matrices with multidimensional scaling (**b**), the classification are realized using PAM and FANNY methods (**c**). The results show the classification success over 10-fold bagging. FANNY gives better results for a higher noise. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

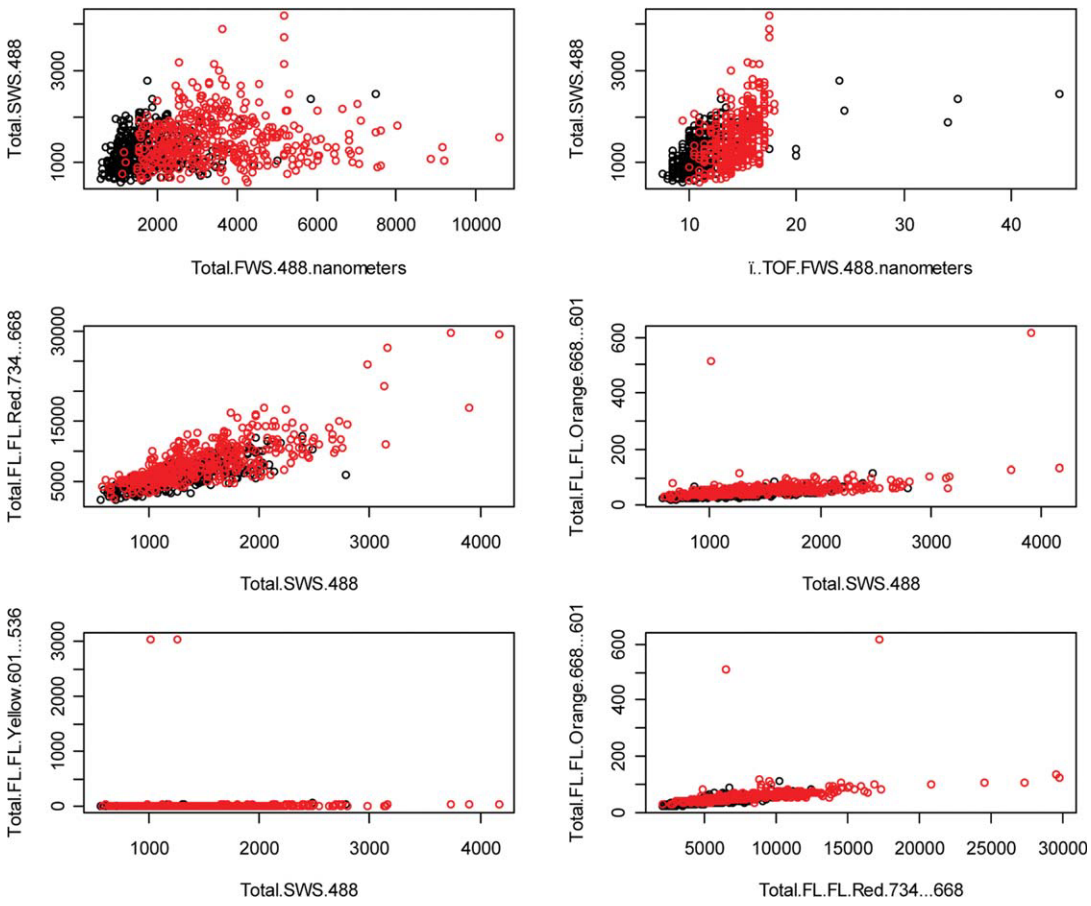

Figure 8. Scatterplots of two phytoplankton culture datasets, *Amphidinium carterae* (red symbols) and *Tetraselmis tetrathele* (black symbols), artificially mixed. The display shows a strong overlap between culture datasets (SWS: sideward scatter, FWS: forward scatter, FL: fluorescence, TOF: time of flight). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

individuals were successfully classified on average with a small standard deviation. When the pulse shape descriptor was considered, the classification success ranged from 60 to 66%. Finally, by combining the descriptors, the average range of classification efficiency reached 78% and varied only slightly. Combining both descriptors arose to a gain of about 10 points, demonstrating that there is discriminating information in the shape descriptors.

Discussion and Conclusion

Analysis of aquatic microorganisms performed by flow cytometry is currently used to address their abundance, diversity, and dynamics (10). Data analysis for conventional flow cytometers is based on a set of real values (peak, area, pulse width) corresponding to the light scatter and fluorescence signals recorded for each single particle as it is intercepted by the light source. The clusters are drawn from various histograms and dotplots. The interpretation of these clusters is based on the operator expertise. This way of analysis is particularly well suited for specific observations in samples with known groups (cultures, previously analyzed samples). As far as aquatic environmental studies are concerned, the main purpose of conventional flow cytometric analysis is to define these groups, count the cells, and get information at the group level (basic statistics for light scatter and fluorescence signals: mean, median, mode, and standard deviation for instance). In aquatic environments, phytoplankton diversity is huge, gathering thousands of species with various shapes and covering four decades in size. Some of the species are harmful and need to be monitored at high frequency to detect as fast as possible any sanitary risk.

The advances in electronics and computing contribute to the development of more compact instruments able to record a growing number of variables (10). Some instruments are even able to collect pictures of the particles as they flow through the flow cytometer. Particular models such as the CytoSub (15) and the Flow Cytobot (Heidi Sosik and Rob Olson, WHOI) have been especially developed for the marine field (41). Once deployed in situ (moored or in a buoy) these instruments can perform automated analyses of the phytoplankton cells at a scheduled sampling frequency. With the CytoSub, up to six analyses per hour can be scheduled, quickly generating a huge quantity of data. This high frequency analysis opens the way to new information out of reach when using

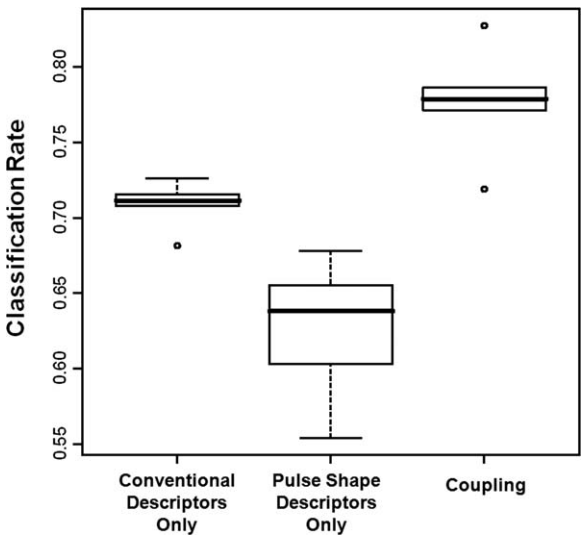

Figure 9. Classification success for three different analyses (200-fold bagging). Coupling conventional descriptors and functional pulse shapes gives accurates classification.

the classical methods (16). The automation of analysis therefore becomes critical. To some extent, performing such analyses by an operator would become impossible (time consuming, lack of objectivity in the clustering, etc).

Phytoplankton analysis with the CytoSub flow cytometer is innovative in the way that it is based on the pulse shape recording along each particle. It is a compromise between the taxonomical complexity and conventional flow cytometry. It provides information on phytoplankton diversity without fully addressing the complexity of the taxonomical identification. Shape analysis becomes relevant when the recognition and the differences between different shapes are surrounded by mathematical laws.

This study purposes data processing automation in order to provide efficient tools for objective analysis of the full fingerprints of phytoplankton. To test and validate the clustering methods, we have started with some numerical experiments to work on simulated objects known a priori: the classes are defined in advance and the dataset is tunable (heterogeneity within classes; number and relative abundance of each class). Moreover, an infinite number of situations can then be generated (from easy cases to more complex ones). After test and validation of the clustering methods, experiments have been performed with real data collected from the flow cytometry analysis of several cultures (20 different strains belonging to various taxa). Important features can be discussed about the results obtained with the numerical experiments and the phytoplankton cultures. The distance invariant to orientation acts as a deformation of the functional space, gathering shapes similar by symmetry. The experiments carried out with the six simulated families have proven the effectiveness of the distance invariant to orientation computation. When classes present great deviations in their relative abundance (Fig. 6), the partition occurs within the group presenting the predominant abundance. In this case, the clustering does not converge to the proper cutting. In other words, the predominance of a group in a natural sample could prevent identification of other groups in lower abundances.

In aquatic environments, natural samples contain a smaller abundance of large phytoplankton cells (i.e., >20 μm or chain-forming species) and a larger abundance of small phytoplankton (42,43). It will be essential to consider this phenomenon. Thanks to the modification experiments on the variability within families, it was possible to get different results for the tested clustering methods. A gain was induced by testing several methods and comparing them. However, one approach cannot be considered to be better than another, but more or less adapted to a particular case. In this study, the fuzzy clustering fitted better with the type of data generated by the CytoSub, providing higher classification success than the K-medoids method. This result is due to the specificity of the fuzzy method, which enables a better separation of overlapping groups.

To handle the complex data collected with the CytoSub (i.e., the optical fingerprints corresponding to the five raw pulses), it was necessary to find a way to deal with descriptors of different types such as length, AUP, and functional shape of the various optical fingerprints. The distance matrices of each descriptor were first computed individually and then successfully combined. While looking for the most efficient clustering method, our primary focus was to find out whether using the functional shape could be more efficient than the classical method (i.e., based on real numbers). To do so, two particular datasets of phytoplankton cultures (*Amphidinium carterae* and *Tetraselmis tetrathele*) were selected and artificially mixed into a single data file. By analyzing both species with the CytoClus software, i.e., the software dedicated to the CytoSub data analysis using the classical method with conventional descriptors, the toxic and nontoxic species could not be adequately distinguished. Their optical fingerprints were too similar to form distinct clusters in the classical two-dimensional dotplots preventing any efficient manual separation. On the contrary, the autonomous clustering method was clearly efficient (Fig. 9). The classification success reached about 78%, and the two species were well discriminated. Another aim was to test the contribution of the shape related information compared to the classical descriptors. In this case the gain was about 10 points between the classical descriptors and the combination of functional shape descriptors and classical descriptors, a weak improvement but significant. The shape related information appeared useful when particles presented morphological modification or typical features such as the repetition of a similar pattern (for instance chain-forming cells), or the presence of appendages usually linked to an environmental adaptation.

Adversely, shape related information was less efficient for small particles because their shape tends typically to a sphere and thus the corresponding optical fingerprints are dominated by a Gaussian shaped curve (16). However, the use of full pulse shapes is surprisingly applicable for cells that are smaller than the height of the focused laser (5 μm). From the analysis of very small particles (2 to 6 μm in size) the following state-

ments can be made: (i) By considering observations as "curves" (actually "densities" would be more appropriate) one takes into account all moments of all orders and not only mean and variance, (ii) most of the signals look like bell shaped but there is a great variability between the signal shapes due to the difference in skewness (data not shown), (iii) moreover the position of the maximum is not always central leading to asymmetrical curves and this is potentially linked to cell morphology, (iv) considering the entire optical fingerprint (i.e., the whole five variables) these slight variations in signal shapes induce a decoupling between signals. This constitutes an additional information with regards to classical method handling only with length, height or area under the signal.

Through all experiments described in this study, with numerical simulations and real data from more than 20 cultures, a new method of analysis has been validated. It is a new method as it combines conventional descriptors with the pulse shapes. This is complementary to the previous works by Boddy and collaborators who considered the peak integrated values and pulse widths (18). The main known difficulty with unsupervised classification methods is to choose the number of clusters: thanks to the Silhouette coefficient computation, the optimal number of groups can also be found without any human interference. The robustness or consistency of the associated partition is also provided by the maximum of the Silhouette coefficient values. It provides also a visual display of the data with a rational criteria proposed to select splits. But this display is limited by the initial number of observations which must be reasonable. That is not the case when dealing with datasets coming from CytoSub and its large number of observations (several thousands of cells). It is however possible to use subsampling methods to evaluate the number of final clusters with clearer displays. Another original and interesting feature of the described method is that the analysis remains flexible due to the system of weights that can be associated with the distance matrices of each descriptor. The operator can tune the weight applied to the various variables depending on their respective interest and therefore decide to adjust the method to any particular case. By handling the raw pulse shape as a functional descriptor, the potential of the CytoSub flow cytometer is fully utilized. It is true that this study does not present any results of an analysis on natural sample, needed to consider all the complexity that can occur in the field (various clusters, large biodiversity, background noise, etc). The major reason therefore is that to test the efficiency of the clustering methods, it was necessary to have a knowledge of the sample composition. It was mandatory to control the clustering efficiency by comparing the results with what was expected. The work with natural samples is ongoing and will be addressed in other studies.

The automation of sampling acquisition as well as the data analysis and clustering open the way to the spatiotemporal analysis at high frequency, which has previously been out of reach because of physical constraints (need for operator(s), work onboard depending on the ship availability and meteorology, etc). Oceanographic cruises, for instance, are characterized by their limits both in space, whether or not their track covers a long distance, and mainly in time, failing to provide the spatial coverage and temporal resolution required to determine a realistic picture of the marine environment and detect changes within it. To face such challenges, many efforts have been dedicated to the automation of measurements and the autonomy of instruments in order to produce monitoring systems delivering sufficient online data. This is the impetus of the Global ocean observing system (GOOS) endorsed by the United Nations (UNESCO) and in Europe by the European GOOS initiative EuroGOOS. The International Council for the Exploration of the Sea (ICES) and the Mediterranean Science Commission (CIESM) are also developing such activities (see TRANSMED: http://www.ciesm.org/marine/programs/transmed.htm, CIESM pilot project). The high frequency survey should bring new information, which is essential to better understanding the complex dynamics of phytoplankton communities in relation to their environment.

Acknowledgments

A.M. is a recipient of a fellowship from the Council of the region PACA (Provence Alpes Côte d'Azur). We thank the technical support and expertise of the PRECYM flow cytometry platform of the COM-OSU (http://precym.com.univ-mrs.fr)

Literature Cited

1. Field C, Behrenfeld M, Randerson J. Primary production of the biosphere: Integrating terrestrial and oceanic components. Science 1998;281:237–240.
2. Falkowski P, Scholes RJ, Boyle E, Canadell J, Canfield D, Elser J, Gruber N, Hibbard K, Högberg P, Linder S, Mackenzie FT, Moore B III, Pedersen T, Rosenthal Y, Seitzinger S, Smetacek V, Steffen W. The global carbon cycle: A test of our knowledge of earth as a system. Science 2000;290:291–296.
3. Furnas M. Net in situ growth rates of phytoplankton in an oligotrophic, tropical shelf ecosystem. Limnol Oceanogr 1991;36:13–29.
4. Li W, Rao DS, Harrison W, Smith J, Cullen J, Irvin B, Platt T. Autotrophic picoplankton in the tropical ocean. Science 1983;219:292–295.
5. Baldauf S. The deep roots of eukaryotes. Science 2003;300:1703–1706.
6. Boudouresque C, Ruiton S, Verlaque M. Anthropogenic impacts on marine vegetation in the Mediterranean. In: Proceedings of the second Mediterranean symposium on marine vegetation. United Nations Environment Programme MAP, for Specially Protected Areas publication RAC, Athens, 12–13 December 2003, 2006. pp34–54.
7. Weithoff G. The concepts of plant functional types and functional diversity in lake phytoplankton — A new understanding of phytoplankton ecology? Freshwater Biol 2003;48:1669–1675.
8. Larkum A, Douglas S, Raven J. Photosynthesis in algae, advances in photosynthesis and respiration. Urbana Illinois: Kluwer Academic Publishers; 2003.500 p.
9. Dokulil M, Donabaum K, Teubner K. Modifications in phytoplankton size structure by environmental constraints induced by regime shifts in an urban lake. Hydrobiologia 2007;578:59–63.
10. Dubelaar GBJ, Casotti R, Tarran GA, Biegala I. Phytoplankton and their analysis by flow cytometry. In: Dolezel J, Greilhuber J, Suda J, editors. Flow Cytometry with Plant Cells. Weinheim, Germany: Wiley-VCH. 2007. pp287–322.
11. O'Farrell I, de Tezanos Pinto P, Izaguirre I. Phytoplankton morphological response to the underwater light conditions in a vegetated wetland. Hydrobiologia 2007; 578:65–77.
12. Naselli-Flores L, Barone R. Pluriannual morphological variability of phytoplankton in a highly productive Mediterranean reservoir (lake Arancio, southwestern Sicily). Hydrobiologia 2007;578:87–95.
13. Sournia A. Form and function in marine phytoplancton. Biol Rev 1982;57:347–394.
14. Uthermohl H. Toward the improvement of the quantitative phytoplankton method. Mitt Int Vereiningung für Limnonogie 1958;9:1–38.
15. Dubelaar G, Gerritzen P. Cytobuoy : A step forward towards using flow cytometry in operational oceanography. Sci Marina 2000;64:255–265.
16. Thyssen M, Mathieu D, Garcia N, Denis M. Short-term variation of phytoplankton assemblages in mediterranean coastal waters recorded with an automated submerged flow cytometer. J Plankton Res 2008;30:1027–1040.
17. Thyssen M, Tarran GA, Zubkov M, Holland RJ, Grégori G, Burkill PH, Denis M. The emergence of automated high frequency flow cytometry: Revealing temporal and spatial phytoplankton variability. J Plankton Res 2008;30:333–343.
18. Boddy L, Morris C, Wilkins M, Al-Haddad L, Tarran G, Jonker R. Identification of 72 phytoplankton species by radial basis function neural network analysis of flow cytometric data. Mar Ecol Prog Ser 2000;195:47–59.

19. Boddy L, Wilkins M, Morris C. Pattern recognition in flow cytometry. Cytometry 2001;44:195–209.
20. Frankel D, Olson R, Frankel S, Chisholm S. Use of a neural net computer system for analysis of flow cytometric data of phytoplankton populations. Cytometry 1989; 10:540–550.
21. Godavarti M, Rodriguez J, Yopp T, Lambert G, Galbraith D. Automated particle classification based on digital acquisition and analysis of flow cytometric pulse waveforms. Cytometry 1996;24:330–339.
22. Wilkins M, Hardy S, Boddy L, Morris C. Comparison of five clustering algorithms to classify phytoplankton from flow cytometry data. Cytometry 2001;44:210–217.
23. Wilkins M, Boddy L, Dubelaar G. Identification of Marine Microalgae by Neural Network Analysis of Simple Descriptors of Flow Cytometric Pulse Shapes, 1st ed. Berlin: Springer; 2003.
24. Fraley C, Raftery A. How many clusters? which clustering methods? answers via model-based cluster analysis. Comput J 1998;41:578–588.
25. Lo K, Brinkman R, Gottardo R. Automated gating of flow cytometry data via robust model-based clustering. Cytometry Part A 2008;73A:321–332.
26. Pyne S, Hu X, Wang K, Rossin E, Lin T, Maier L, Baecher-Allan C, McLachlan G, Tamayo P, Hafler D, Jager PD, Mesirov J. Automated high-dimensional flow cytometric data analysis. PNAS 2009;106:8519–8524.
27. Collins G, Krzanowski W. Nonparametric discriminant analysis of phytoplankton species using data from analytical flow cytometry. Cytometry 2002;48:26–33.
28. Demers S, Kim J, Legendre P, Legendre L. Analyzing multivariate flow cytometric data in aquatic sciences. Cytometry 1992;13:291–298.
29. Dryden I, Mardia K. Statistical Shape Analysis. Chichester, UK: Wiley series in probability and statistics; 1998. 376 p.
30. R: A Language and Environment for Statistical Computing. R Development Core Team. 2010. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. http://www.R-project.org.
31. Ramsay J, Silverman B. Functional Data Analysis. New York: Springer; 2005. 430 p.
32. Kachel. Hydrodynamic properties of flow cytometry instruments. In: Melamed MR, Lindmo T, Mendelsohn ML, editors. Flow Cytometry and Sorting, 2nd ed. New York: Wiley-Liss; 1990. pp 27–44.
33. Khelil A, Mante C, David P. Statistical methods for localization and discrimination of acoustical signals backscattered by air bubbles. Trait Du sig 1997;14:151–159.
34. Nerini D, Ghattas B. Classifying densities using functional regression trees: Applications in oceanology. Comput Stat and Data Anal 2007;51:4984–4993.
35. Wahba G. Spline Models for Observational Data. Montpelier, Vermont: Society for Industrial and Applied Mathematics; 1990. 180 p.
36. Kruskal J, Wish M. Multidimensional Scaling. Beverly Hills USA, London UK: SAGE Publications; 1978. 96 p.
37. Kaufman J, Rousseeuw P. Finding Groups in Data: An Introduction to Cluster Analysis. 9th ed. Hoboken, NJ: Wiley Series in Probability and Statistics; 2005. 368 p.
38. Breiman L. Bagging predictors. Mach Learn 1996;24:123–140.
39. Hastie T, Tibshirani R, Friedman J. The Elements of Statisticals Learning. New York: Springer-Verlag; 2001. 552 p.
40. Ronquillo J, Matias J, Saisho T, Yamasaki S. Culture of tetraselmis tetrathele and its utilization in the hatchery production of different penaeid shrimps in Asia. Hydrobiologia 1997;358:237–244.
41. Olson RJ, Sosik HM. A submersible imaging-in-flow instrument to analyze nano and microplankton: Imaging flowcytobot. Limnol Oceanog: Methods 2007;5:195–205.
42. Sieburth J, Smetacek V, Lenz J. Pelagic ecosystem structure: Heterotrophic compartments of the plankton and their relationship to plankton size fractions. Limnol Oceanogr 1978;23:1256–1263.
43. Azam F, Fenchel T, Gray J, Meyer-Reil L, Thingstad T. The ecological role of watercolumn microbes in the sea. Mar Ecol Progr Ser 1983;10:257–263.

Appendix

Reference densities f_1 and f_2 are Gaussian densities $N(\mu_j, \sigma_j^2)$, $j = 1, 2$ with mean population parameter μ and variance σ^2. These functions get similar shapes with one maximum, but they differ in their spread. Reference densities f_3 and f_4 for classes 3 and 4 are Gaussian mixture densities of the form

$$f_j = \gamma_j N(\mu_{1j}, \sigma_{1j}^2) + (1 - \gamma_j) N(\mu_{2j}, \sigma_{2j}^2), \quad j = 3, 4$$

where γ_j denotes a mixture coefficient, μ_j are mean parameters and σ_j^2 are variance parameters. These functions possess two maxima whose ordinates differ according to the value of the mixture coefficient. For well-chosen mixture coefficients, f_3 is the symmetrical version of f_4 with respect to the ordinate axis. Reference density f_5 is a Gumbel density $G(\alpha, \beta)$ and f_6 its symmetrical version with respect to the ordinate axis. The location parameter α controls the position of its single maximum and scale parameter β controls the distribution spread. These functions get an asymmetrical shape. Here are the settings of parameter values according to each class:

$\mathbf{C_1}$ $\mu_1 = 0,\ \sigma_1^2 = 1.5$

$\mathbf{C_2}$ $\mu_2 = 0,\ \sigma_2^2 = 3$

$\mathbf{C_3}$ $\gamma_3 = 3/5,\ \mu_{31} = -1,\ \mu_{32} = 1,\ \sigma_{31}^2 = 1/2,\ \sigma_{32}^2 = 1/2$

$\mathbf{C_4}$ $\gamma_4 = 2/5,\ \mu_{41} = -1,\ \mu_{42} = 1,\ \sigma_{41}^2 = 1/2,\ \sigma_{42}^2 = 1/2$

$\mathbf{C_5}$ and $\mathbf{C_6}$ $\alpha = -3,\ \beta = 1$

The choice of the above reference densities has been conducted in order to mimic simple shapes (Figure 4). These densities also present the advantage of covering the main problems encountered when trying to classify curves on cytometry signals. Following the different families of curves described above, the sample of curves S is constructed as follows:

1. Do for $j = 1, \ldots, p$
2. Select a class C_j with reference density f_j
3. Draw randomly m points from reference distribution f_j
4. From this sample of points, compute a random curve by kernel density estimation, choosing the bandwith by cross-validation
5. Repeat n_j times steps 3) and 4) to form a sample of n_j random curves which belong to class C_j
6. Go to step 1)

Once these operations have been achieved, merge classes C_1 to C_6 to form the sample S.