A Rapid Technique for Classifying Phytoplankton Fluorescence Spectra Based on Self-Organizing Maps

ISMAEL F. AYMERICH,* JAUME PIERA, AURELI SORIA-FRISCH, and LLUÏSA CROS

Unidad de Tecnología Marina (UTM-CSIC), Pg. Marítim de la Barceloneta 37-49, E-08003 Barcelona, Spain (I.F.A., J.P.); Technical University of Catalonia (UPC), Barcelona, Spain (I.F.A.); Starlab S.L., C. de l'Observatori Fabra s/n, E-08035 Barcelona, Spain (A.S.-F.); and Institut de Ciències del Mar (ICM-CSIC), Pg. Marítim de la Barceloneta 37-49, E-08003 Barcelona, Spain (L.C.)

Fluorescence spectroscopy has been demonstrated to be a powerful tool for characterizing phytoplankton communities in marine environments. Using different fluorescence spectra techniques, it is now possible to discriminate the major phytoplankton groups. However, most of the current techniques are based on fluorescence excitation measurements, which require stimulation at different wavelengths and thus considerable time to obtain the complete spectral profile. This requirement may be an important constraint for several mobile oceanographic platforms, such as vertical profilers or autonomous underwater vehicles, which require rapid-acquisition instruments. This paper presents a novel technique for classifying fluorescence spectra based on self-organizing maps (SOMs), one of the most popular artificial neural network (ANN) methods. The method is able to achieve phytoplankton discrimination using only fluorescence emission spectra (single wavelength excitation), thus reducing the acquisition time. The discrimination capabilities of SOM using excitation and emission spectra are compared. The analysis shows that the SOM has a good performance using excitation spectra, whereas data preprocessing is required in order to obtain similar discrimination capabilities using emission spectra. The final results obtained using emission spectra indicate that the discrimination is properly achieved even between algal groups, such as diatoms and dinoflagellates, which cannot be discriminated with previous methods. We finally point out that although techniques based on excitation spectra can achieve a better taxonomic accuracy, there are some applications that require faster acquisition processes. Acquiring emission spectra is almost instantaneous, and techniques such as SOM can achieve good classification performance using appropriately preprocessed data.

Index Headings: **Self-organizing maps; SOMs; Phytoplankton discrimination; Classification; Fluorescence spectra; Derivative analysis.**

INTRODUCTION

Phytoplankton is one of the basic organic compounds of natural waters and its diagnosis is important for evaluating the ecological status of coastal seawater areas. Researchers are currently studying several rapid analysis techniques for measuring seawater properties directly and providing qualitative and quantitative information about phytoplankton. Among these techniques, the analysis of spectral fluoresence[1] is widely applied for characterizing the phytoplankton community in the marine environment. Measuring algae bio-optical properties is an efficient tool in high-frequency sensing of the algal community. The fluorescence method has been widely used, for example, to study the vertical distribution of chlorophyll *a* (chl *a*) concentration with high spatial resolution.[2] Moreover, this method is easy to perform and provides highly sensitive on-line information on the distribution of algae. It is also worth noting that changes in the phytoplankton community often take place with a high frequency and that this technique is fast enough to provide important information about them. Furthermore, the technique is nondestructive and requires little or no sample preparation.

Several studies have been carried out since Yentsch and Phinney[1] proposed an ataxonomic technique that utilized the spectral fluorescence signatures of major ocean phytoplankton to study their population structure in 1985. Kolbowski and Schreiber[3] showed in 1995 that with four excitation wavelengths using light emitting diodes (LEDs) they were able to discriminate between three groups of algae. Beutler[4] presented a free-falling depth profiler using five different excitation wavelengths and acquiring the fluorescence response at 680 nm (chl *a* emission). In this case, four phytoplankton groups can be distinguished (blue algae/cyanobacteria, green algae, brown algae, and cryptophyceae). The system is adaptable to new algae classes added to the measuring system, but diatoms and dinoflagellates cannot be distinguished from each other because they have similar fluorescence spectra. This is important because of their importance in bloom-forming algae. Zhang et al.[5] analyzed the discrimination between different phytoplankton classes using the information extracted from excitation–emission matrices (EEMs). They used processing methods such as singular value decomposition and Bayesian linear discriminant analysis to distinguish different algae from excitation fluorescence spectra, and they even achieved discrimination between diatoms and dinoflagellates. Using EEMs of this kind, Moore et al.[6] also presented an under-test prototype for *in situ* measurements and analysis.

However, the use of these methods involves some limitations. They offer good performance in terms of high taxonomic accuracy, but they all require nearly a second to acquire each sample, or even more in the case of the techniques using EEMs. The problem is that they need to stimulate at different excitation wavelengths. This time requirement means a limitation in the number of samples acquired and, in consequence, a low vertical resolution. Although low resolution is not a handicap for some studies, Cowles et al.[7,8] pointed out that some physiological and trophic processes may be constrained by physical processes operating over spatial scales of a few centimeters and temporal scales of seconds to minutes. The importance of detecting these processes, or what are called "thin layers", emphasizes the need to develop new techniques aimed at increasing the number of samples and the vertical resolution. Several studies and efforts have focused on this goal.[9] Another important aspect pointed out by Margalef[10,11] is that there is some evidence that pigment composition changes during the life of phytoplankton. This statement introduces another variable that makes the classification even more challenging.

The aim of the work presented herein was to evaluate the performance of a technique based on self-organizing maps

Received 25 March 2008; accepted 6 March 2009.

* Author to whom correspondence should be sent. E-mail: ismaelf@utm.csic.es.

0003-7028/09/6306-0716$2.00/0

(SOMs)[12] that is used to classify phytoplankton species. The SOM is a type of artificial neural network (ANN) that has been successfully applied for extracting interpretable patterns from large and complex data sets: for example, in satellite remote sensing.[13] This technique can be used as a rapid method for discriminating between different phytoplankton classes, avoiding the time requirement of the techniques mentioned above. It has not been widely applied to oceanography data, but in recent years several studies have shown its good performance in pattern recognition and classification.[14–16] In order to evaluate this method, two different approaches were followed and are presented in this study. Firstly, excitation spectra were used to achieve discrimination among different species; as mentioned, the main problem with this kind of acquisition is the time required to stimulate the sample with different excitations. Secondly, emission spectra were used in order to determine whether they could offer a faster method; this is possible because hyperspectral sensors acquire the whole emission spectrum almost instantly. Furthermore, in order to deal with changes in pigment composition during the life of phytoplankton, this work also evaluates the performance of SOM taking into account the age of the cultures in the training and test data sets.

The next section offers a brief description of ANN, paying special attention to the SOM method. The following sections present the materials and methods, the results and discussion, and the conclusions.

ARTIFICIAL NEURAL NETWORKS

Artificial neural networks are a processing technology that has been widely studied over the last few decades. Inspired by neuroscience, they are trained to behave like biological neural networks, emulating how they process the data. One of their advantages is that they are more robust in handling noisy and missing data than traditional methods. How neural networks work depends on the interconnectivity between the neurons. As Kohonen[12] mentions, there are three categories of neural networks, each one based on a different philosophy. In feedforward networks sets of input signals are transformed into sets of output signals, and this transformation is determined by externally adjusting several parameters. In feedback networks, the parameters are changed iteratively from an initial state until the desired outcome is obtained. Finally, in competitive, unsupervised, or self-organizing networks, neighboring cells in a neural network compete and interact to correctly match (represent) the input space. It is not the intention of this study to give a detailed theoretical description of the SOM[17] algorithm, but a brief description is presented below.

Self-Organizing Maps. The Kohonen self-organizing maps are a type of artificial neural network based on unsupervised learning, which means that the network learns only based on the input training data. In contrast, supervised learning needs the pairs of input/output training patterns in order to approximate the input data. The SOM projects high-dimensional input data, usually onto a two-dimensional map, a feature that is useful for the visualization and classification of high-dimensional data. Also, the algorithm is topology-preserving, which means that similar input data will be mapped to spatially close areas on the map, and elements which are spatially close on the map should have similar input data.

A SOM output map consists of neurons organized on a regular low-dimensional grid, each one holding a weight vector (W_{ij}) (Fig. 1). This weight vector has exactly the same length as the dimension of the input data, and the lattice of the nodes can be either hexagonal or rectangular. Once the weight vectors are initialized, the SOM training algorithm adapts them so that the neurons span across the data cloud. At the end of the training phase the map is organized such that neighboring neurons in the grid have similar weight vectors. Two auxiliary matrices are generated to help in the visualization of the resulting clusters, the **U** matrix and the hit-matrix. The training algorithm can be summarized in two steps:

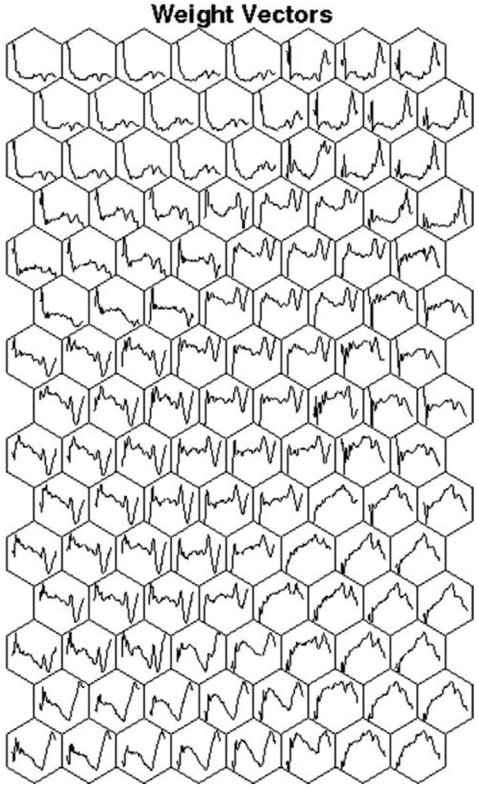

FIG. 1. Example of weight vectors, codebook, once the neural network has been trained. Neighbor neurons have a similar weight vector.

(1) Finding the Best Matching Unit. During each training step, one input sample x is randomly chosen from the training set. The distances between this input sample and the weight vectors of all neurons are thence computed (typically Euclidean distance). The neuron that has the minimum Euclidean distance between the input vector and its weight vector is the winning neuron and is called the best-matching unit (BMU).

(2) Adapting the Weight Vector. Once the BMU has been chosen and the input vector has been assigned to the winning neuron, it is time to learn. The BMU and its neighboring neurons update their weight vectors to make them similar to the input vector as follows (Eq. 1):

$$W_{ij}(t+1) = W_{ij}(t) + \alpha(t) \times h_c(t) \times [x(t) - W_{ij}(t)] \qquad (1)$$

where $x(t)$ is the input data vector, $h_c(t)$ is the learning neighborhood function (typically a Gaussian bell-shaped one), and $\alpha(t)$ is the learning rate. The neighboring function ($h_c(t)$) defines the region of influence that the input sample has on the SOM, and both α and h_c decrease with time, performing a fine-tuning at the end of the training. At each learning step, all the

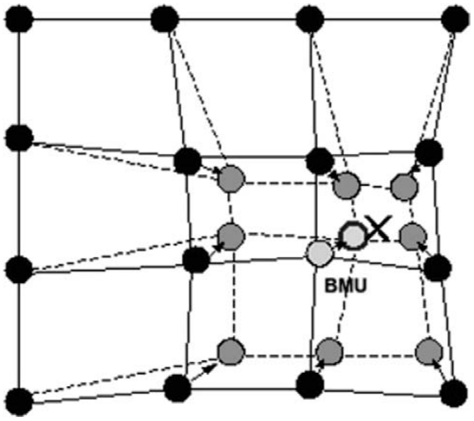

FIG. 2. Visual representation of the adaptive step, in which the BMU and its neighborhood learn and change their weight vectors. The data point of the training set driving the adaptation of neurons is represented as an X. The BMU moves into this position in the feature space. Due to the neighboring function definition, neighboring neurons are moved in the same direction.

neurons within the neighborhood (N_c) are updated, whereas cells outside N_c are left intact. The neighborhood function is often taken to be Gaussian (Eq. 2):

$$h(t) = \exp\left[-\frac{\rho^2(t)}{2\sigma^2}\right] \tag{2}$$

where σ^2 is the variance parameter specifying the spread of the Gaussian function, and $\rho(t)$ is the radius of the neighboring function centered at the BMU. The learning rate denotes the regularization parameter of the adapting procedure (Fig. 2).

Once the SOM training has finished, the **U** matrix is constructed, representing the distances between the neurons of the output map, for example, as gray values. For a network of **P** × **Q** neurons, the **U** matrix has $(2\mathbf{P} - 1) \times (2\mathbf{Q} - 1)$ distances between neurons or values.[18] It is used in order to obtain an initial idea of the cluster distribution.[19] Clusters are characterized in this representation as a homogeneous area of large gray values separated by edge-wise elongated areas of low gray values (an example of a **U** matrix is given in Fig. 3). Once the output map has been trained, the data set is applied once again in order to obtain the winning neuron for each sample. This information is accumulated, and the most-frequent winning values can be considered as the most representative ones. The result, presented in a two-dimensional histogram, is the so-called hit-matrix, used in the classification step explained below. In this study, the somtoolbox[20] for Matlab was used for the presented result computation.

MATERIALS AND METHODS

Fluorescence Measurements. Seven strains from different taxonomic groups of phytoplankton were selected for this study (Table I). The cultures were incubated at 20 °C in an *f*/2 medium[21,22] under a 12:12 h dark–light cycle. Fluorescence measurements were performed in the laboratory and three-dimensional (3D) excitation–emission matrices (EEMs) were obtained every day with an Aminco-Bowman Series 2 Spectrometer, a slit width of 4 nm, and a scan wavelength

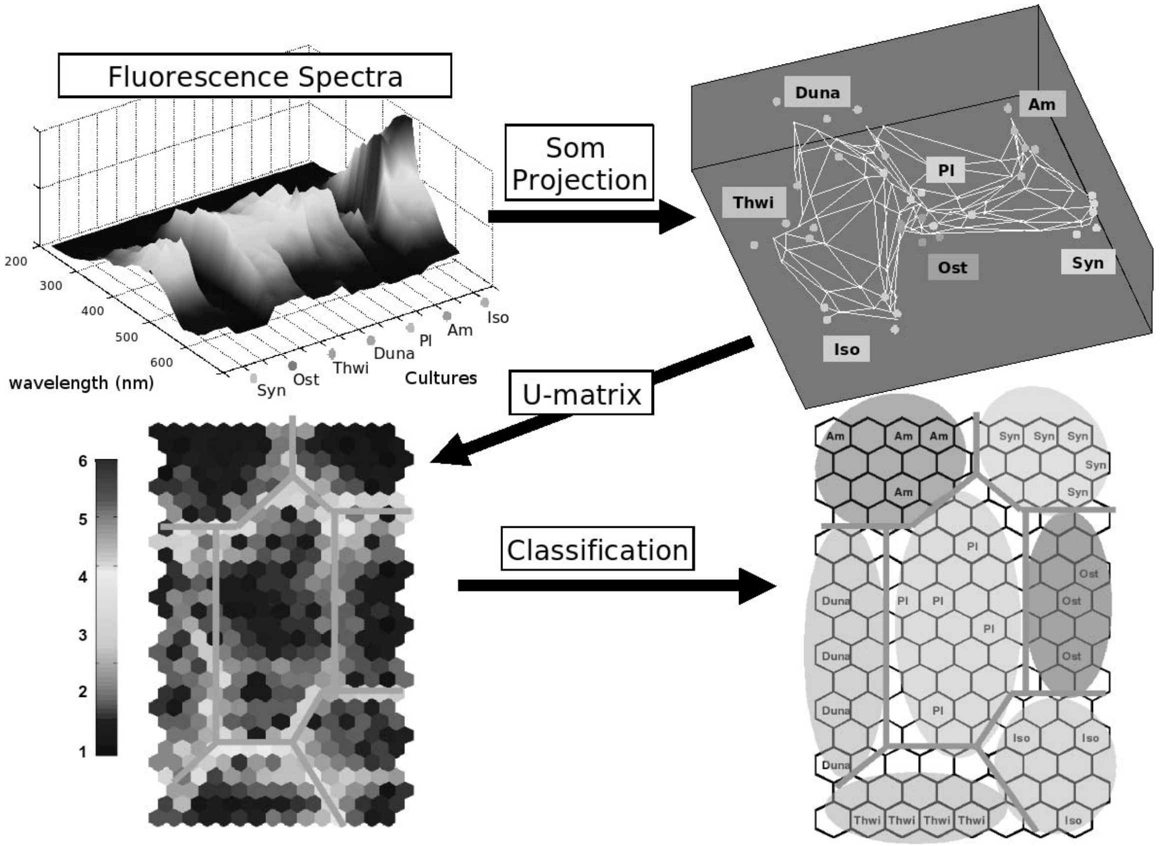

FIG. 3. Diagram of the steps followed by the SOM classification method used in this paper. Excitation spectra are used in this example, acquiring the different emission fluorescence spectra at 680 nm.

TABLE I. Phytoplankton cultures under test. The number of samples corresponds to the number of days the data were acquired. The differences in the number of samples are due to the different growth speeds of the cultures.

Species	Class	Abbreviation	No. of samples
Alexandrium minutum	Dinophyceae	Am	22
Thalassiosira weissflogii	Bacillariophyceae	Thwi	18
Dunaliella primolecta	Chlorophyceae	Duna	20
Isochrysis galbana	Prymnesiophyceae	Iso	10
Pleurochrysis elongata	Prymnesiophyceae	Pl	21
Synechococcus sp.	Cyanophyceae	Syn	15
Ostreococcus sp.	Prasinophyceae	Ost	15

speed of 20 nm/s. Due to the different growth speed between the cultures not all groups had the same number of samples. For this study the first two days were not taken into account due to their low culture concentration and fluorescence signal. The ranges of excitation and emission wavelength were 200–600 nm every 10 nm and 200–800 nm every 1 nm, respectively. The matrices measured had 41 rows and 601 columns. An example of one of these matrices is shown in Fig. 4, where the diagonal peaks correspond to the effects of Raman and Rayleigh scatters.

Self-Organizing Maps Classification. The classification using SOM was done following three steps.

First, the network was trained using the training data set (the selection of the training and test data sets is explained in the following section). The network adapted its properties to the input data and then the distances between neurons were calculated. The result of this process was the **U** matrix (Fig. 3), which shows the distances between neighboring neurons. The light gray is the edge of the clusters, where these distances are higher. Afterwards, a variant of the hit-matrix called the fuzzy hit-matrix[18] was computed. This matrix was obtained for each culture. The gray value of the matrices represents a particular neuron's membership of the culture class (Fig. 5). The matrices were normalized and labeled. Lastly, we wished to assign a label to each sample, so the best-matching unit of each sample was found. Then, the different membership values, which were extracted from the fuzzy hit-matrices, were compared. The winning class, whose label was then taken to classify the sample into one culture or another, was that with the largest BMU membership value.

In order to evaluate the classification performance, the confusion matrices for the test set were computed, and different indices were calculated: the percentage of true positives (true positive rate: TPR) and the percentage of false positives (false positive rate: FPR), as well as the Kappa index. Taking class 1 as a reference, TPR and FPR indices are computed as follows:

(1) TPR=(# test samples classified as class 1 actually corresponding to class 1)/(total of test samples actually corresponding to class 1)

(2) FPR=(# test samples classified as class 1 actually corresponding to other classes different from class 1)/(total of test samples actually corresponding to other classes different from class 1)

The Kappa index (Eq. 3) is also computed. It is a measure of confidence that considers all the elements in the confusion matrix, the diagonal elements as well as the errors of commission (classifying a sample into one class while it belongs to another one) and the errors of omission (classifying a sample belonging to one class into another class). Therefore, the Kappa value can be computed by applying the following expression:

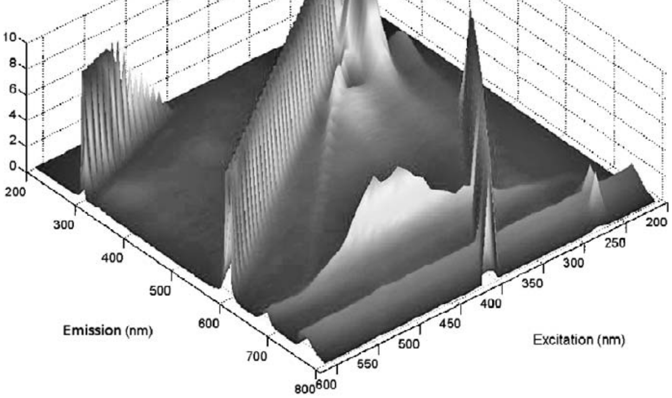

FIG. 4. Example of a 3D representation of an excitation–emission matrix. This matrix contains all the information about the excitation or emission fluorescence spectra, with ranges 200–600 and 200–800 nm, respectively.

$$K = \frac{n\sum_{k=1}^{r} X_{kk} - \sum_{k=1}^{r} X_{k+}X_{+k}}{n^2 - \sum_{k=1}^{r} X_{k+}X_{+k}} \tag{3}$$

where n is the total number of samples and X_{kk} is the correctly classified samples in class k. If the confusion matrix is considered line by line for class k, then X_{k+} is the user's accuracy, whereas a column-by-column analysis specifies the producer's accuracy as X_{+k}. Following the example shown in Fig. 5, in Table II there is an exemplary description on how to compute the Kappa index.

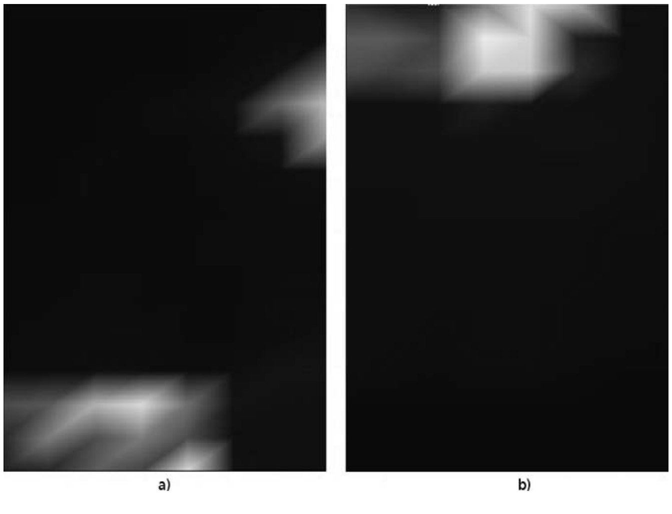

FIG. 5. Example of two fuzzy hit-matrices from different cultures, (**a**) *Thalassiosira weissflogii* (Thwi) and (**b**) *Alexandrium minutum* (Amin), extracted from excitation spectra analysis (Fig. 3). This information is used in the classification step.

TABLE II. Confusion matrix. Classification behavior using excitation spectra. Samples from the stable growth stage used for training, and samples from the exponential growth stage used for testing.

		Predicted class									
		Ost	Syn	Thwii	Duna	Pl	Am	Iso	Sum	TPR	FPR
True class	Ost	4	0	0	2	0	0	4	10	0.4	0.013
	Syn	0	3	0	0	0	7	0	10	0.2	0
	Thwi	1	0	9	2	1	0	0	13	0.692	0
	Duna	0	0	0	11	1	3	0	15	0.666	0.028
	Pl	0	0	0	1	2	13	0	16	0.187	0.086
	Am	0	0	0	0	0	17	0	17	1	0.348
	Iso	0	0	0	1	0	0	4	5	0.8	0.049

[a] Using this table, the Kappa index is computed as follows:
$n\ \Sigma^{r}_{k=1}\ X_{kk} = 86 \times (4 + 3 + 9 + 11 + 2 + 17 + 4) = 86 \times 50 = 4300$
$\Sigma^{r}_{k=1}\ X_{k+}X_{+k} = (10 \times 5) + (10 \times 3) + (13 \times 9) + (15 \times 17) + (16 \times 4) + (17 \times 40) + (5 \times 8) = 1236$
$K = (n\ \Sigma^{r}_{k=1}\ X_{kk} - \Sigma^{r}_{k=1}\ X_{k+}X_{+k})/(n^2 - \Sigma^{r}_{k=1}\ X_{k+}X_{+k}) = (4300 - 1236)/(86^2 - 1236) = 0.4974.$

Data Preparation. Two different analyses were carried out in this study. The phytoplankton discrimination was studied first using excitation spectra, and second using only emission spectra and a shorter acquisition time. Furthermore, as stated above, there is some evidence that pigment composition changes during the life of phytoplankton. This is why we also wished to evaluate the performance of this technique taking into account the age of the cultures, just to focus on the capability of this technique for dealing with these pigment changes. For this reason, a parallel study in which the neural network was trained with the fluorescence response from adult cultures was carried out. Because the stable stage presents a higher regularity than the dynamic stage, it was thought to be advisable to train the algorithm with stable data. The testing data, i.e., in this case the dynamic stage, was expected to present similar trends as the training one. These trends were expected to be rightly classified thanks to the generalization capability of the classification algorithm.

In order to work with these approaches, several data sets were prepared:

Excitation Spectra. In the first case, the data consisted of several excitation spectra. Each culture was excited at different wavelengths (excitations between 200 and 600 nm every 10 nm), and its emission fluorescence was recorded at 680 nm for each excitation. This procedure was repeated over several days, and the resulting training matrix consisted of 59 rows (training samples) and 41 columns. The training and test data sets were chosen by carrying out repeated random sub-sampling validation. The percentage of samples for both was almost the same, and the results of ten different classifications were averaged.

Still working with excitation spectra, as stated above, another experiment was carried out. This time, the training data set chosen contained five spectra from each culture, but they belonged to the stable growth stage. Figure 6 shows an example of a culture growth curve, and it is important in this case to give the percentage of training and test data. As can be clearly seen, the samples taken for training correspond to the stable stage, while the samples chosen for the test belong to the exponential growth stage.

The effect of the Rayleigh scattering peak was studied. Although some studies[5] set it to zero, other approaches such as leaving the peak were evaluated, but the results of the SOM are slightly better using an interpolation with the two neighboring samples to avoid this effect.

Emission Spectra. In this case, emission spectra were used instead of excitation spectra. A preliminary study was carried out in order to choose the best excitation wavelength, containing as much information as possible to achieve the best classification. To this end, an exploratory classification for each excitation wavelength was made, and the results are shown in Fig. 7. The 490 nm wavelength was the most discriminating wavelength, and from then on it was the excitation wavelength chosen to perform this study.

The training and test data sets were constructed again by performing repeated random sub-sampling from emission spectra excited at 490 nm (almost the same number of samples for training and testing). It is important to mention that the emission spectra range was chosen between 535 and 735 nm to reduce the dimensionality of the data, because there is no fluorescence emission at wavelengths below the excitation. The results with these data were also averaged between ten different classifications.

Following the procedure explained above, the second part of this approach focused on evaluating the performance of this technique when the SOM network was trained with samples taken from the stable growth stage. In this case, the last five spectra from each culture, corresponding to the stable stage,

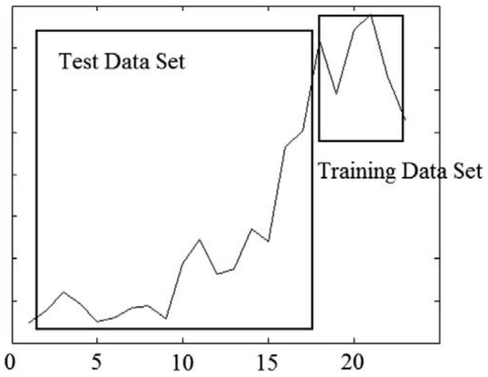

FIG. 6. As an example, *Alexandrium minutum*'s growth curve is shown. It has been made computing the fluorescence emission at 680 nm every day with 470 nm excitation wavelength.

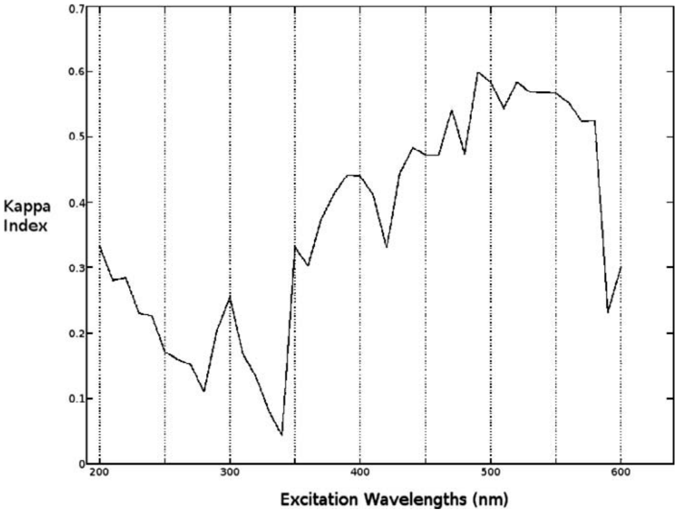

FIG. 7. Kappa index value as a function of excitation wavelength. The classification performance improves with the excitation, with maximum around 490 nm.

were used as training samples. Then, the training matrix had 35 training vectors with 201 components.

RESULTS AND DISCUSSION

Classification Using Excitation Spectra. As mentioned above, excitation–emission matrices were acquired, but in this step only the responses at 680 nm emission wavelength with a range of 200–600 nm excitations were used. In this case, randomly selected training and test data sets were constructed to evaluate the method. An example of a training data set is shown in Fig. 8. Once the neural network has been trained, a label can be assigned to each neuron in the SOM by first computing the hit-matrix over the whole training set. The label of each neuron will correspond to the label of the class with a maximal number of hits in each neuron. In this label representation, the discrimination of the method can be appreciated (Fig. 9a). The neurons have changed their properties to better characterize the input data. The different cultures should appear classified, grouping all the samples of the same culture, but there are some mixed samples. The samples that have a similar spectrum appear closer. For instance, *Alexandrium minutum* and S*ynechococcus sp.* have similar excitation spectra, whereas *Thalassiosira weissflogii* differs a lot from them.

Once the neural network has been trained, the labeled output map can be used to classify the test data set. Based on this classification, the confusion matrix is computed in order to evaluate the performance of the classification methodology. The Kappa indices, the TPR, and the FPR are calculated. Ten different validation runs, in which the training and test data sets had been randomly constructed iteratively, were undertaken. Thence the results were averaged in order to make the performance evaluation as independent as possible from the selected training set. An example of a confusion matrix is presented in Table III. Even though the discrimination does not seem to be good enough, the average TPR and the average FPR over cultures and validation runs were 0.7344 and 0.0508, respectively. Also, the Kappa index value was quite good, 0.6839, with 1 denoting a perfect classification without any mistake, and 0 denoting the result of a random classification. The greatest problems arose with *Pleurochrysis elongata*, which is not properly classified. This means that the spectra from this culture are very similar, in this case, to those of *Alexandrium minutum*.

In the second part of this approach, the performance of the method using the stable samples as training data is evaluated. In this case, where the pigment composition variations play an important role in the classification, the resulting **U** matrix is shown in Fig. 3. The label representation (Fig. 9b) shows a good performance at first sight. Testing samples from the

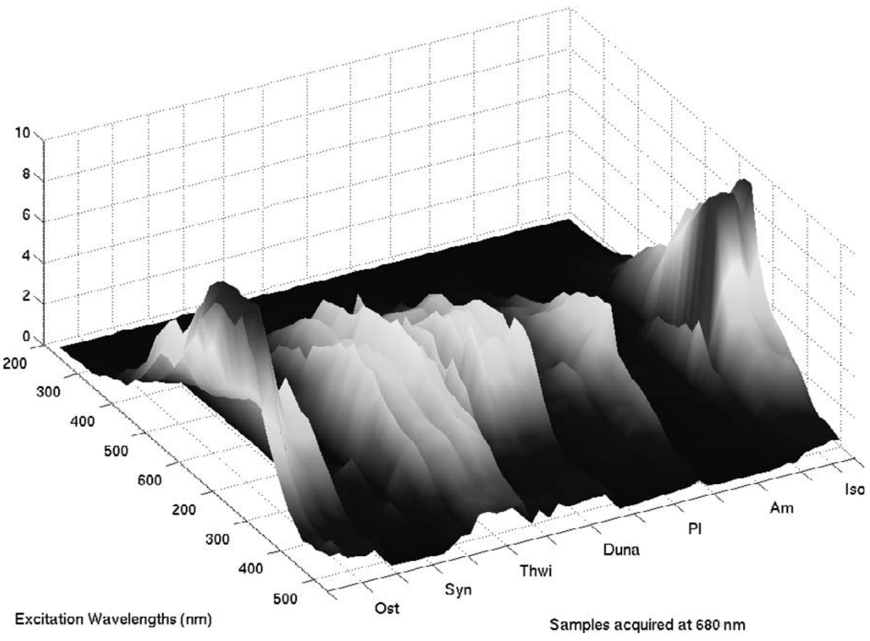

FIG. 8. An example of a training data set working with excitation spectra. Fluorescence excitation spectra acquired at 680 nm. 60 samples randomly selected representing the seven cultures. The excitation range is 200–600 nm, every 10 nm.

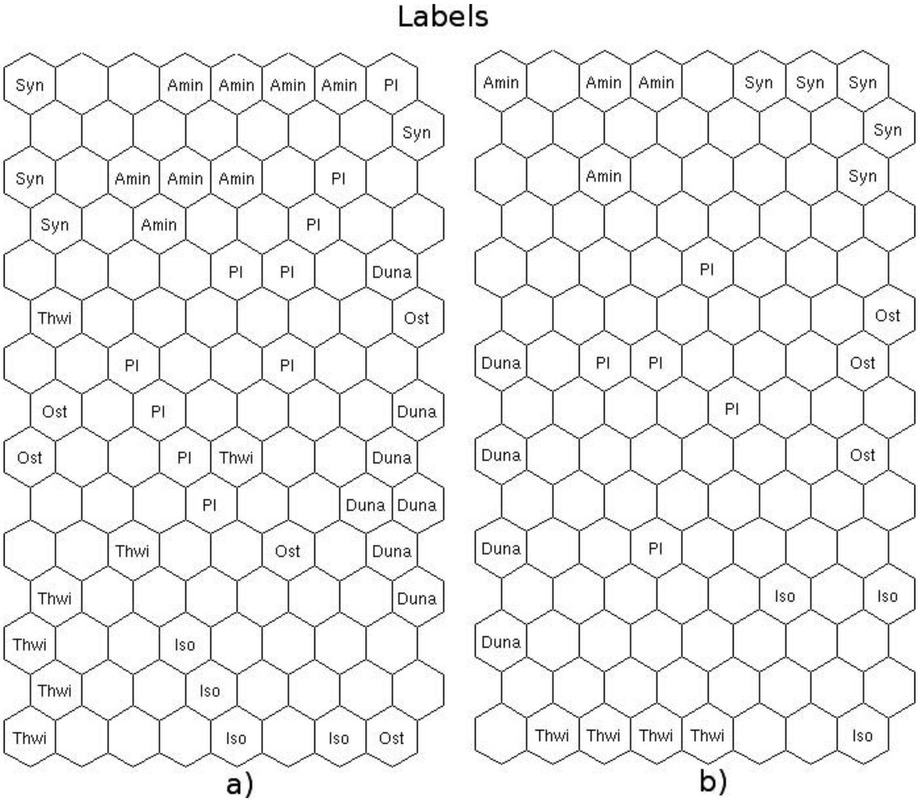

FIG. 9. Label representation once the network has been trained using excitation spectra: (**a**) an example of the results obtained from randomly chosen training and test samples, (**b**) result obtained using samples from the stable growth stage to train the network.

exponential growth stage were classified to form the confusion matrix (Table II). The Kappa index obtained was 0.4974 (TPR = 0.5787, FPR = 0.0981). Again, the greatest classification problems arose with *Pleurochrysis*. Although initially the **U** matrix seems to be better, the Kappa value was below 0.5. This result could be a consequence of the pigment changes that phytoplankton cultures suffer during their growth stage.

Classification Using Emission Spectra. Using the emission spectra, *in situ* acquisition would be faster and a higher vertical and horizontal resolution could be achieved. The main problem is that fluorescence information is not as high in emission spectra as in excitation spectra. This means that differences between fluorescence responses of different phytoplankton classes will be less, and discrimination will be more difficult. Having established the good performance of the SOM with excitation spectra, its performance using only emission spectra was evaluated. The results are described in the following paragraphs.

The EEMs acquired were used again, but now using the training and test data sets as described above: emission spectra (535–735 nm) excited at 490 nm and randomly sub-sampled. An example of training samples for this case is presented in Fig. 10. The discrimination can be observed in Fig. 11a, and Table IV represents the confusion matrix obtained. The averaged TPR index over ten validation runs is 0.7046, the average FPR is 0.0588, and the Kappa index is 0.6343. Although a good classification performance is obtained, there are again some classes that appear mixed. For example, *Synechococcus sp.* is clearly distinguishable, while the other classes appear closer and mixed. The reason is that these classes have very similar spectra and they are more difficult to discriminate from emission fluorescence.

Repeating the same procedure as in the previous approach, we now focus our attention on the evaluation of the performance using the stable samples for training and then classifying samples from the growth stage. Surprisingly, using emission spectra, the results (Fig. 11b and Table V) are better than using excitation spectra stable samples; the Kappa index is equal to 0.5985. From this result and that obtained with excitation spectra, it seems that

TABLE III. Example of a confusion matrix. Classification of excitation spectra from a random selection of training and test samples.

		Predicted class									
		Ost	Syn	Thwii	Duna	Pl	Am	Iso	Sum	TPR	FPR
True class	Ost	4	0	0	2	0	0	2	8	0.5	0
	Syn	0	4	0	0	2	2	0	8	0.5	0
	Thwi	0	0	8	1	0	0	0	9	0.88	0.018
	Duna	0	0	0	8	1	0	1	10	0.8	0.057
	Pl	0	0	0	0	5	6	0	11	0.45	0.058
	Am	0	0	0	0	0	11	0	11	1	0.157
	Iso	0	0	1	0	0	0	4	5	0.8	0.052

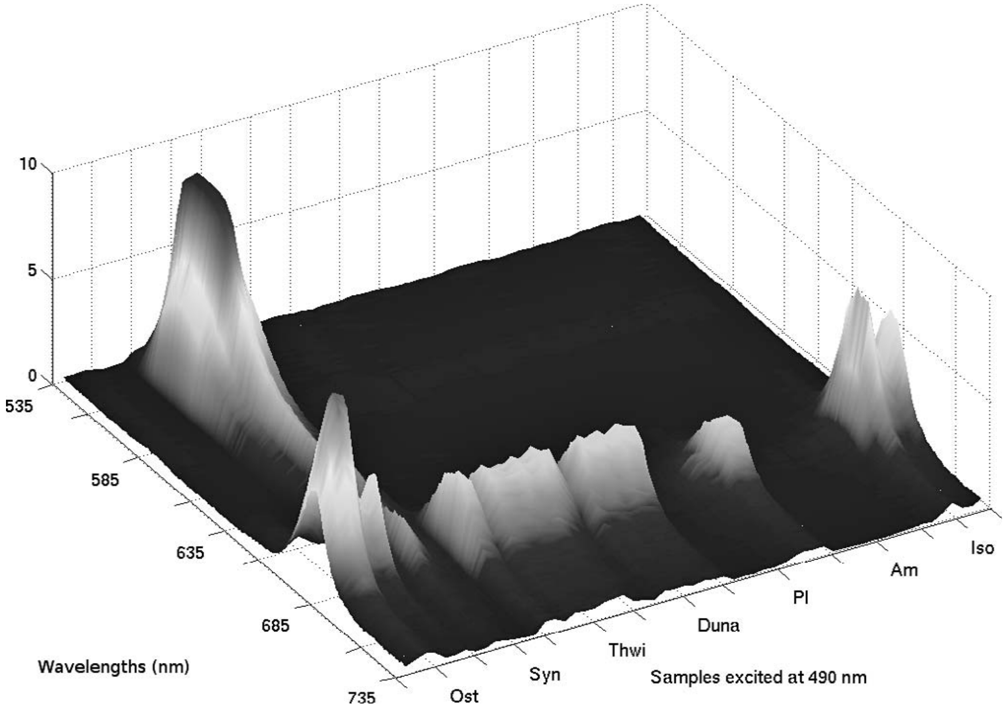

FIG. 10. An example of a training data set working with emission spectra. Fluorescence emission spectra excited at 490 nm. 60 samples randomly selected representing the seven cultures. The emission range is 535–735 nm, every 1 nm.

the pigment changes have a greater effect on the excitation spectra ($K = 0.4974$) than on the emission spectra ($K = 0.5985$), making emission spectra more robust to these changes.

Several studies use derivative techniques to enhance minute differences between similar signals.[23,24] These techniques have proven to be a powerful tool that is commonly used, for example, in the analysis of hyperspectral data. However, the derivative spectroscopy used to explore these minute features

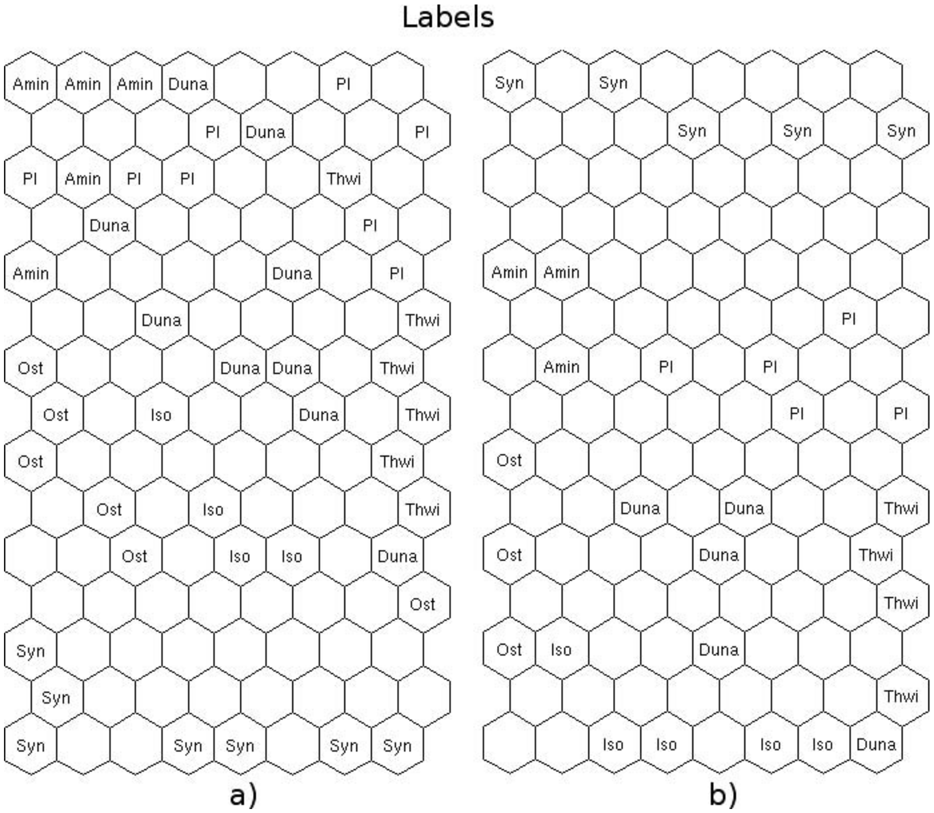

FIG. 11. Label representation once the network has been trained using emission spectra: (**a**) an example of the results obtained from randomly chosen training and test samples, (**b**) result obtained using samples from the stable growth stage to train the network.

TABLE IV. Example of a confusion matrix. Classification of emission spectra from a random selection of training and test samples.

		Predicted class									
		Ost	Syn	Thwii	Duna	Pl	Am	Iso	Sum	TPR	FPR
True class	Ost	6	0	0	0	0	0	2	8	0.75	0
	Syn	0	8	0	0	0	0	0	8	1	0
	Thwi	0	0	9	0	0	0	0	9	1	0.037
	Duna	0	0	1	5	4	0	0	10	0.5	0.038
	Pl	0	0	1	0	7	3	0	11	0.63	0.176
	Am	0	0	0	0	5	6	0	11	0.54	0.059
	Iso	0	0	0	2	0	0	3	5	0.6	0.035

TABLE V. Confusion matrix. Classification behavior using emission spectra. Samples from the stable growth stage used for training, and samples from the exponential growth stage used for testing.

		Predicted class									
		Ost	Syn	Thwii	Duna	Pl	Am	Iso	Sum	TPR	FPR
True class	Ost	5	0	1	0	0	0	4	10	0.5	0
	Syn	0	10	0	0	0	0	0	10	1	0
	Thwi	0	0	10	2	1	0	0	13	0.77	0.041
	Duna	0	0	1	9	0	5	0	15	0.6	0.028
	Pl	0	0	0	0	2	14	0	16	0.125	0.014
	Am	0	0	0	0	0	17	0	17	1	0.275
	Iso	0	0	1	0	0	0	4	5	0.8	0.049

in spectral data is notoriously sensitive to noise.[25] To remove this noise from the hyperspectral data, smoothing techniques are commonly used.[26] It is worth noting that there must be a trade-off between noise removal and the ability to resolve fine spectral details.[27] The following section is devoted to the analysis of the SOM method classification, using as input data the derivative of the spectra.

Classification Applying Derivative Analysis to Emission Spectra. In an attempt to increase the performance of the classification using emission fluorescence spectra, derivative analysis was applied to the emission spectra in order to enhance the differences between the fluorescence spectra of each algae. Previous to this process, the noise of the signals was reduced using a wavelet denoising technique.[28] Each spectrum was processed and a training data set with the first-order derivatives was obtained (Fig. 12).

First of all, the samples were also randomly sub-sampled in order to evaluate the classification indices for different training and test data sets. An example of the results obtained using the derivative training data is presented in Fig. 13a. The discrimination this time was higher. The neurons of the network were properly distributed and the indices obtained (Table VI) were slightly better than those obtained using excitation spectra: TPR = 0.7574, FPR = 0.0481, and Kappa = 0.7109.

In the case of using stable samples for training, the results

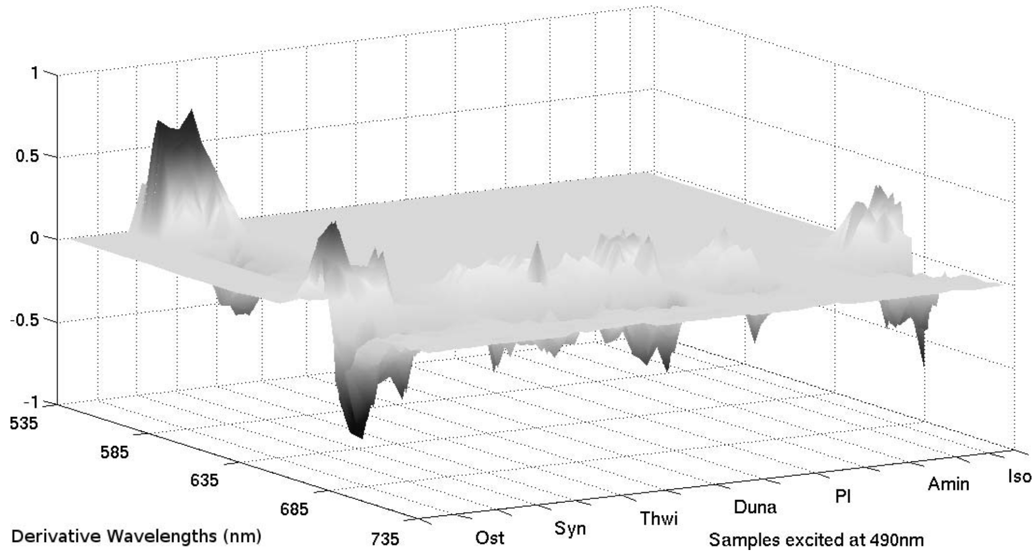

FIG. 12. An example of a training data set working with derivative emission spectra. Derivative fluorescence emission spectra excited at 490 nm. 60 samples randomly selected representing the seven cultures. The emission range is 535–735 nm, every 1 nm.

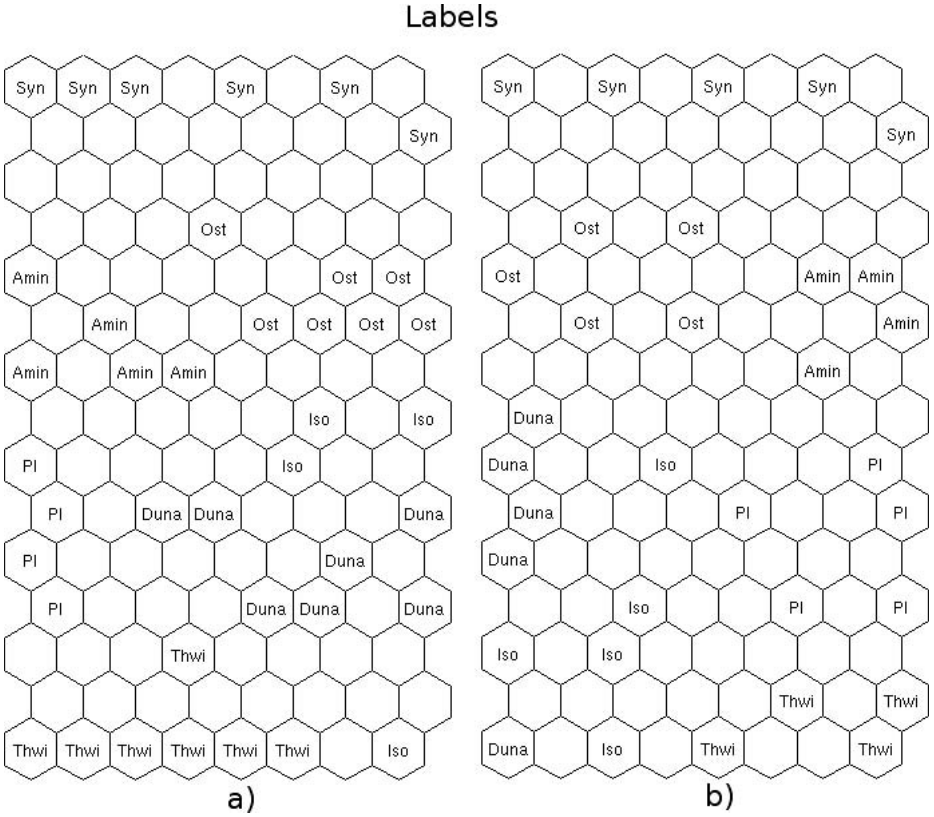

FIG. 13. Label representation once the network has been trained using derivative emission spectra: (**a**) an example of the results obtained from randomly chosen training and test samples, (**b**) result obtained using samples from the stable growth stage to train the network.

obtained were TPR = 0.7638, FPR = 0.0611, and Kappa = 0.6803 (Fig. 13b and Table VII).

As can be clearly seen in the confusion matrices, the greatest problems arose with *Pleurochrysis elongata* (Pl). The fluorescence properties of *Alexandrium minutum* and Pl are very similar for this method. Both species like coastal areas where they can form blooms;[29–32] it has been noticed that coastal species in coccolithophores share pigments unexpected by their phylogeny,[33] and if we join these two cultures into one group, the performance increases considerably. For example, for the

TABLE VI. Example of a confusion matrix. Classification of first-derivative emission spectra from a random selection of training and test samples.

		Predicted class									
		Ost	Syn	Thwii	Duna	Pl	Am	Iso	Sum	TPR	FPR
True class	Ost	8	0	0	0	0	0	0	8	1	0
	Syn	0	8	0	0	0	0	0	8	1	0
	Thwi	0	0	6	2	1	0	0	9	0.66	0.019
	Duna	0	0	0	9	1	0	0	10	0.9	0.057
	Pl	0	0	1	1	6	3	0	11	0.54	0.059
	Am	0	0	0	0	1	10	0	11	0.9	0.059
	Iso	0	0	0	0	0	0	5	5	1	0

TABLE VII. Confusion matrix. Classification behavior using first-derivative emission spectra. Samples from the stable growth stage used for training, and samples from the exponential growth stage used for testing.

		Predicted class									
		Ost	Syn	Thwii	Duna	Pl	Am	Iso	Sum	TPR	FPR
True class	Ost	10	0	0	0	0	0	0	10	1	0
	Syn	0	10	0	0	0	0	0	10	1	0
	Thwi	0	0	9	0	4	0	0	13	0.692	0
	Duna	0	0	0	10	0	5	0	15	0.666	0.014
	Pl	0	0	0	0	3	13	0	16	0.187	0.057
	Am	0	0	0	0	0	17	0	17	1	0.26
	Iso	0	0	0	1	0	0	4	5	0.8	0

confusion matrices found in the last two approximations studied above (using derivative emission fluorescence spectra), if these two species are grouped the Kappa indices are 0.8105 and 0.8439, respectively.

CONCLUSION

This study shows the feasibility of using self-organizing maps (SOM) as a method for processing fluorescence data in order to classify phytoplankton cultures. For this purpose seven different cultures were cultivated and their excitation–emission matrices were acquired.

The results presented in this paper show that the fluorescence signals of the different cultures are very similar and are difficult to discriminate. Based on the results presented here, classification using excitation fluorescence spectra is better than classification using only emission fluorescence spectra. However, it is worth mentioning that the use of adequate preprocessing techniques such as derivative analysis can help to improve the results of the latter.

From the results of this work, it can be concluded that SOM shows good performance using excitation spectra and also using only emission spectra. Furthermore, the performance of the method improves when emission spectra are combined with derivative analysis to enhance spectra singularities. This combination is a powerful method for obtaining a good discrimination among different cultures. Moreover, *Thalassiosira weissflogii* can be distinguished from *Alexandrium minutum* using only one excitation, an important result for distinguishing diatoms from dinoflagellates. However, *Pleurochrysis elongata* cannot be differentiated from *Alexandrium minutum* at all, which means that their fluorescence properties are very similar. One possibility would be to group them for consideration together.

The preliminary results are encouraging. Working with emission spectra obviates the need to use different excitation sources, and thus reduces the acquisition time. However, further work is necessary in order to better adjust the parameters of the method, and different preprocessing techniques could even be tested in future work. One of these could be to use an average of the input data (i.e., an average of normal and derivative spectra) for training, which would perhaps provide better discrimination between the strains.

Although this technique has been tested only with cultures, we wish to present SOM as a powerful technique for determining major phytoplankton compounds present in the water column by direct fluorescence measurement of seawater.

ACKNOWLEDGMENTS

This work was supported by the project ANERIS (PIF-015-1), funded by the Spanish National Research Council (CSIC). We would also like to thank the reviewers for many helpful comments that have made this a better paper and A. Lounds for his work correcting the English style.

1. C. S. Yentsch and D. A. Phinney, J. Plankton Res. **7,** 617 (1985).
2. T. J. Cowles, R. A. Desiderio, and S. Neuer, Marine Biol. **115,** 217 (1993).
3. J. Kolbowski and U. Schreiber, Proceedings Xth Photosynthesis Congress (Montpellier, France, 1995).
4. M. Beutler, K. H. Wiltshire, B. Meyer, C. Moldaenke, C. Lüring, M. Meyerhöfer, U.-P. Hansen, and H. Dau, Photosynth. Res. **72,** 39 (2002).
5. Q.-Q. Zhang, S.-H. Lei, X.-L. Wang, L. Wang, and C.-J. Zhu, Spectrochim. Acta, Part A **63,** 361 (2006).
6. C. C. Moore, J. Da Cunha, B. Rhoades, M. Twardowski, and J. Zaneveld, "A new in-situ measurement and analysis system for excitation-emission fluorescence in natural waters", Ocean Optics XVII (Freemantle, Australia, October, 2004).
7. T. J. Cowles and R. A. Desiderio, Oceanography **6,** 105 (1993).
8. T. J. Cowles, R. A. Desiderio, and M.-E. Carr, Oceanography **11,** 4 (1998).
9. R. A. George, L. Gee, A. W. Hill, J. A. Thomson, and P. Jeanjean, "High-Resolution AUV Surveys of the Eastern Sigsbee Escarpment", Offshore Technology Conference (Houston, TX, May 6–9, 2002).
10. R. Margalef, Rapports et procès-verbaux des reunions CIESMM **15,** 277 (1960).
11. R. Margalef, *Perspectives in Ecological Theory* (University of Chicago Press, Chicago, 1968).
12. T. Kohonen, Proc. IEEE **78,** 1464 (1990).
13. A. J. Richardson, C. Risien, and F. A. Shillington, Prog. Oceanogr. **59,** 223 (2003).
14. E. J. Ainsworth and S. F. Jones, IEEE Trans. Geosci. Remote Sens. **37,** 1645 (1999).
15. A. J. Richardson, M. C. Pfaff, J. G. Field, N. F. Silulwane, and F. A. Shillington, J. Plankton Res. **24,** 1289 (2002).
16. N. F. Silulwane, A. J. Richardson, F. A. Shillington, and B. A. Mitchell-Innes, African J. Marine Sci. **23,** 37 (2001).
17. T. Kohonen, *Self-Organizing Maps* (Springer, Berlin, 2001), 3rd ed., Springer Series in Information Sciences, vol. 30.
18. A. Soria-Frisch, Int. J. Approx. Reasoning **41,** 23 (2006).
19. J. Vesanto and E. Alhoniemi, IEEE Trans. Neural Networks **11,** 586 (2000).
20. J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas, Proceedings of the Matlab DSP Conference, 35 (Finland, 1999).
21. R. R. L. Guillard and J. H. Ryther, Can. J. Microbiol. **8,** 229 (1962).
22. R. R. L. Guillard, "Culture of phytoplankton for feeding marine invertebrates", in *Culture of Marine Invertebrate Animals,* W. L. Smith and M. H. Chanley, Eds. (Plenum Press, New York, 1975), pp. 26–60.
23. W. L. Butler and D. W. Hopkins, Photochem. Photobiol. **12,** 439 (1970).
24. T. H. Demetriades-Shah, M. D. Steven, and J. A. Clark, Remote Sens. Environ. **33,** 55 (1990).
25. F. Tsai and W. D. Philpot, Remote Sens. Environ. **66,** 41 (1998).
26. C. Vaiphasa, ISPRS J. Photogrammetry Remote Sens. **60,** 91 (2006).
27. E. Torrecilla, I. F. Aymerich, S. Pons, and J. Piera, "Effect of spectral resolution in hyperspectral data analysis," IEEE International Geoscience and Remote Sensing Symposium (Barcelona, Spain, July 23–27, 2007).
28. J. Piera, R. Quesada, A. Manuel-lazaro, J. Del Rio, S. Shariat Panahi, and G. Olivar, Sens. Proc. IEEE **3,** 1468 (2004).
29. A. G. Saez, I. Probert, J. R. Young, B. Edvardsen, W. Eikrem, and L. K. Medlin, *in Coccolithophores: From Molecular Processes to Global Impact*, H. R. Thierstein and J. Young, Eds. (Springer, Berlin, 2004), pp. 251–269.
30. A. G. Sáez, A. Zaldivar-Riverón, and L. K. Medlin, J. Plankton Res. **30,** 559 (2008).
31. E. Balech, Phycologia **28,** 206 (1989).
32. M. Delgado, M. Estrada, J. Camp, J. J. Fernandez, M. Santmarti, and C. Lleti, Scientia Marina **54,** 1 (1990).
33. K. van Lenning, I. Probert, M. Latasa, M. Estrada, and J. R. Young, in *Coccolithophores: From Molecular Processes to Global Impact,* H. R. Thierstein and J. R. Young, Eds. (Springer, Berlin, 2004), pp. 51–73.