ELSEVIER

Analytica Chimica Acta 595 (2007) 160–169

CHIMICA ACTA

www.elsevier.com/locate/aca

Recognition of protozoa and metazoa using image analysis tools, discriminant analysis, neural networks and decision trees

Y.P. Ginoris [a,*], A.L. Amaral [b,c], A. Nicolau [b], M.A.Z. Coelho [a], E.C. Ferreira [b]

[a] *Departamento de Engenharia Bioquímica, Escola de Química/UFRJ, Centro de Tecnologia, E-203, Rio de Janeiro, Brazil*

[b] *Institute for Biotechnology and Bioengineering (IBB), Centre of Biological Engineering, Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal*

[c] *Instituto Politécnico de Bragança, Departamento de Tecnologia Química e Biológica, ESTIG, Apartado 1038, 5301-854 Bragança, Portugal*

Received 6 August 2006; received in revised form 11 December 2006; accepted 21 December 2006
Available online 13 January 2007

Abstract

Protozoa and metazoa are considered good indicators of the treatment quality in activated sludge systems due to the fact that these organisms are fairly sensitive to physical, chemical and operational processes. Therefore, it is possible to establish close relationships between the predominance of certain species or groups of species and several operational parameters of the plant, such as the biotic indices, namely the Sludge Biotic Index (SBI). This procedure requires the identification, classification and enumeration of the different species, which is usually achieved manually implying both time and expertise availability. Digital image analysis combined with multivariate statistical techniques has proved to be a useful tool to classify and quantify organisms in an automatic and not subjective way.

This work presents a semi-automatic image analysis procedure for protozoa and metazoa recognition developed in *Matlab* language. The obtained morphological descriptors were analyzed using discriminant analysis, neural network and decision trees multivariable statistical techniques to identify and classify each protozoan or metazoan. The obtained procedure was quite adequate for distinguishing between the non-sessile protozoa classes and also for the metazoa classes, with high values for the overall species recognition with the exception of sessile protozoa. In terms of the wastewater conditions assessment the obtained results were found to be suitable for the prediction of these conditions. Finally, the discriminant analysis and neural networks results were found to be quite similar whereas the decision trees technique was less appropriate.

Keywords: Discriminant analysis; Decision trees; Neural networks; Protozoa; Metazoa; Image analysis

1. Introduction

Activated sludge process can be briefly defined as a controlled aerobic biological treatment of wastewaters. A biomass, consisting of micro-organisms (mainly bacteria and protozoa, but also fungi and small metazoans) and some other solid particles, is aerated and maintained in a constant state of suspension to ensure the removal of organic matter and nutrients that are used by decomposers [1]. Protozoa and small metazoa predominate among eukaryotes in well-performing plants. Protozoa feed either by grazing on suspended particulate material including bacteria cells or by predation upon other protozoa, and have been shown to attain densities higher than 10^6 micro-organisms mL^{-1} in activated sludge systems [2]. Representatives of all the major taxa have been reported in several plant samples around the world [3–11]. Protozoan taxa in wastewater treatment plants (WWTP) can be classified in terms of flagellates, amoeba and, in particular high numbers, ciliates. The last group is normally divided into free swimming, crawling, carnivorous and stalked ciliates, according to their feeding and motion behavior [2].

Moreover, most samples of activated sludge biomass reveal the presence of larger organisms—small metazoan, like nematodes, rotifers and oligochaete worms, although limited to some simple forms with generation times shorter than the sludge age.

Since the operation in 1922 of the first industrial-scale wastewater treatment plant by activated sludge, a few studies have focused on the significance of protozoa and metazoa in biological wastewater treatment plants. Most of the literature support the view that protozoa play a vital direct role in reducing the numbers of freely suspended and loosely attached bacte-

* Corresponding author. Tel.: +55 21 2562 7622; fax: +55 21 2562 7622.
E-mail address: yovankaperez@yahoo.com.br (Y.P. Ginoris).

0003-2670/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2006.12.055

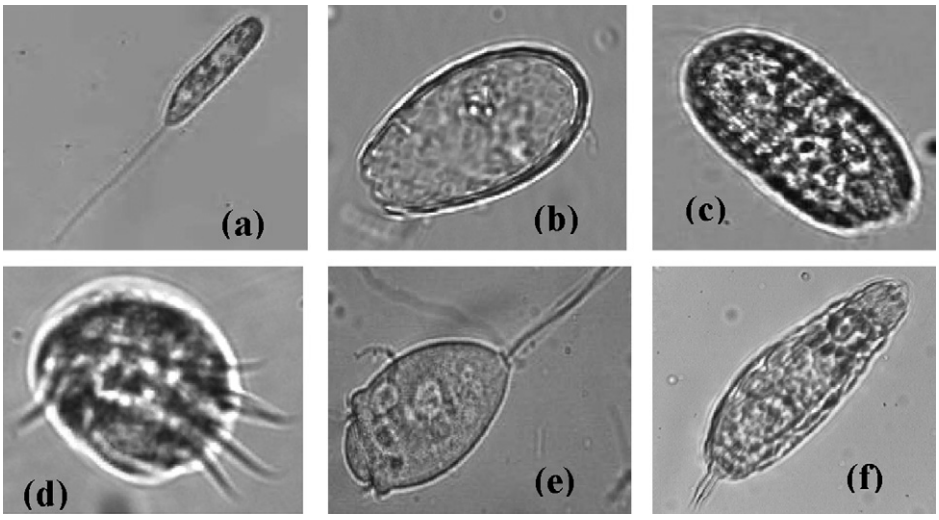

Fig. 1. Examples of protozoan and metazoan present in a WWTP. (a) *Peranema*, (b) *Euglypha*, (c) *Coleps*, (d) *A. cicada*, (e) *Vorticella*, and (f) Rotifer.

rial cells in the bulk liquid by the process of grazing, and thus improve the quality of the final effluent leaving the plant [2]. In the same way, metazoa play an important role in the plant operation as predators, consuming bacterial cells and protozoa. Metazoa seem to contribute to the flocculation process by excreting mucus in which new filamentous and flocculating bacteria may adhere and by breaking up bigger flocs when these organisms move in the medium. Some species representative of the main groups of protozoa and metazoa present in WWTP (flagellates, sarcodines, ciliates, sessile and metazoa) are illustrated in Fig. 1.

Furthermore, protozoa have proved to be excellent tools in assessing and predicting the performance of WWTP and of the final effluent quality [12–21].

The relationship between the clarity and quality of the liquid effluent from the clarifiers and the occurrence of these protozoa and metazoa has led to the suggestion that the presence of particular classes may be used to indicate the overall plant performance [22,23,5,24]. The community structure of protozoa and metazoa rapidly changes as a response to different operating conditions [14], and each plant has a particular faunal structure, so regular monitoring of the plants is important for accurately predicting day-to-day performance. The optimal efficiency occurs when a correct balance between crawling and sessile ciliates and metazoa is achieved. An overpopulation of flagellates, sarcodines or free-swimming ciliates reflects high organic loads (high F:M ratio), whereas the dominance of sessile (stalked) ciliates and metazoa reveals the opposite [25].

Frequent microscopic examination of the biomass may provide a fast, simple and convenient method for indicating sudden changes in the plant performance. Methods based on protozoan population structure have been used to assess activated sludge plant performance. Among them, the Sludge Biotic Index, (SBI), developed by Madoni [2] and inspired in the "Extended Biotic Index" of Woodiwis [26], is the most common method to accomplish the diagnosis of the WWTP operation state using the activated sludge microfauna as biological indicator. This method is based on the abundance and specific diversity of the microfauna community and in the different sensibilities revealed by some protozoan classes to the prevalent physiochemical factors in the system. However, up to date, the identification and micro-organisms enumeration to estimate the SBI have been determined manually, through microscopic observations of activated sludge samples which are highly time-consuming and needs considerable skills in protozoology.

Therefore, automatic image analysis could be seen as a useful tool for performing the taxonomic classification and quantification of organisms in an automatic and non-subjective way. Some studies have already been done using these technique combined with multivariable statistical analysis through techniques such as neural network, discriminant analysis, and principal components analysis in order to perform the recognition of protozoa and metazoa commonly present in the aeration basins of WWTP such as the works of Amaral et al. [27,28] and da Motta et al. [29]. The results of these studies have shown to be promissory for the classification and identification of these classes in a semi-automatic way.

In this work, an image analysis procedure was developed in *Matlab* for the recognition of protozoa and metazoa. Discriminant analysis, neural network and decision trees techniques were employed to allow the recognition of the main protozoa and metazoa found in the WWTP activated sludge.

2. Materials and methods

The protozoa and metazoa classes studied in this work were collected from aeration tanks of WWTPs of Nancy (France) and Portugal treating domestic and industrial effluents.

A total of 22 classes of protozoa and metazoa belonging to several species, genera, orders and sub-classes were included in the study and are presented in Table 1. In all cases the maximum period between the samples collection and the images acquisition did not exceed 3 h, and aeration was provided to the sludge samples during this period.

After the mixed liqueur collection, a drop of the samples was deposited carefully in a slide and covered with a cover

Table 1
Protozoa and metazoa included in this study

Protozoa	
Flagellate	*Peranema* sp.
Sarcodine	*Arcella* sp.
	Euglypha sp.
Ciliate	
Free swimming	*Trachelophyllum* sp.
Carnivorous	*Coleps* sp.
	Litonotus sp.
	Suctoria (sub-class)
Crawling	*Aspidisca* cicada
	Euplotes sp.
	Trithigmostoma sp.
	Trochilia sp.
Sessile	*Carchesium* sp.
	Epystilis sp.
	Opercularia sp.
	V. aquadulcis
	V. microstoma
	V. convallaria
	Zoothamnium sp.
Metazoa	
Rotifer	Digononta (order)
	Monogononta (order)
Gastrotrichia	Nematoda (sub-class)
Oligotrichia	*Aelosoma* sp.

slip (without addition of dyes) for visualization and acquisition of the images using the bright field microscopic technique. The total magnification for visualizing and acquiring each protozoa and metazoa class was dependent from the size of the species organisms and was as follows: *Aelosoma* sp. (25 and 100 times); Nematoda (100 and 250 times); Digononta, Monogonta, *Arcella* sp. and *Euglypha* sp. (250 and 400 times); *Aspidisca cicada*, *Carchesium* sp., *Epistylis* sp., *Euplotes* sp., *Litonotus* sp., *Coleps* sp., *Opercularia* sp., *Peranema* sp., Suctoria, *Trachellophyllum* sp., *Trithigmostoma* sp., *Trochilia* sp., *V. aquadulcis*, *V. microstoma*, *Vorticella* sp. and *Zoothamnium* sp. (400 times). The dimensions of metric units (μm) were correlated with the corresponding pixels using a micrometric slide.

Among the evaluated classes two species of *Epistylis* and *Trachellophylum* were analyzed. Moreover, an additional class of micro-organisms with similar morphological characteristics of *Epystilis* sp. and *Opercularia* sp. was included due to the fact that when these organisms occur with the buccal apparel closed it is quite difficult distinguish one class from the other. Finally, the frontal and lateral views of *Arcella* sp., *Aspidisca cicada* and *Trithigmostoma* sp. were also analyzed, on cause of their lack of axial similitude.

Samples from two sites, Braga in Portugal, and Nancy in France, were used. The image acquisition system used in Nancy was composed by a *Leitz Dialux 20* optic microscope (Leitz, Wetzlar) coupled to a grey scale video camera *Hitachi CCTV HV-720E(F)* (Hitachi, Tokyo). The images were grabbed to the computer in 768 × 576 pixels and 8 bit format (256 grey levels) by a *Matrox Meteor* frame grabber (Matrox, Montreal) using the *Visilog 5* commercial software (Noesis, S.A., les Ulis). In Braga, the acquisition system was composed by an optic microscope *Zeiss Axioscop* (Zeiss, Oberkochen) coupled to a *Sony CCD ACV D5CE* grey scale video camera (Sony, Tokyo) and connected to a PC through the *Data Translation DT 3155* frame grabber (Data Translation, Marlboro), in order to obtain 8-bit digital 768 × 576 pixels images through commercial software *Image-Pro® Plus* (Media Cybernetics, Silver Spring).

A smaller set of images was acquired during the present work using an acquisition system consisting of a *Leitz Laborlux S* optic microscope (Leitz, Wetzlar) coupled to a *Zeiss Axion Cam HR* video camera (Zeiss, Oberkochen). The images acquisition was performed in 1300 × 1030 pixels and 8 bit format through the commercial software of acquisition *Axion Vision* 3.1 (Zeiss, Oberkochen).

3. Image processing programme

The procedure to process the acquired images and determine the morphological parameters was adapted from the *ProtoRec* v.4 program previously developed by Amaral et al. [28] in Visilog (*Noesis*, *Les Ulis*, *France*) and now converted into the *Matlab* 7.0 (*The MathWorks Inc.*, *Natick*, *USA*) own language.

The first step of the image processing procedure consists on the grey-level images pre-treatment by applying a local histogram equalization (8 × 8 pixels) in order to enhance the contrast of each region in the image, followed by the use of the median filter to perform a noise reduction and the *Bottom hat* filter to enhance the organisms borders. The resulting images are then combined for a better differentiation between the organism's borders and the background. After the pre-treatment step, a polygonal region of interest (ROI) around the selected organism is defined by the user with the aid of the mouse. Once defined the ROI, the image is segmented by thresholding the organism's borders, by a threshold value defined either manually or automatically through the Otsu or Entropy methods. It should be emphasized that the manual threshold procedure is used only for the cases where the automatic threshold procedure fails to fully recognize the micro-organisms borders.

In the subsequent stage of the image processing procedure, in order to eliminate debris material (small artifacts and other materials that may interfere with the analysis), a series of morphological operations are applied to the binary images to facilitate the determination of the morphological parameters, including morphological closing, filling and opening operations. Fig. 2 represents the main steps of the image processing procedure.

4. Morphological parameters determination

The determination of the protozoan and metazoan morphological parameters is performed in two stages. In the first stage the parameters are computed to the whole of the organism's body including their external structures such as flagella, cilia, cirri and stalk. In the second stage the parameters are determined for the organism's body core, i.e., after the removal of all external structures, performed after an empirical automatic determination of the number of erosions necessary to remove each of these structures.

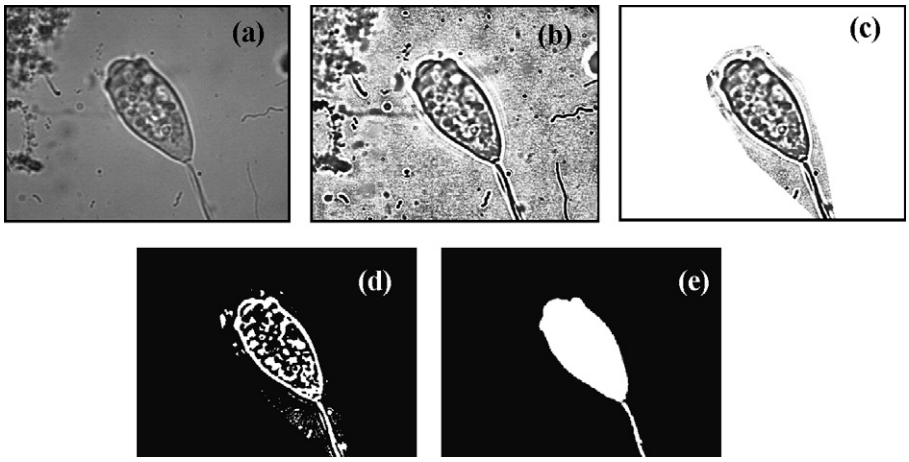

Fig. 2. Main steps of the program: original image (a), pre-treated image (b), region of interest (c), binary image after segmentation (d), and final image (e).

The morphological parameters that were determined for both the whole of the organism and the organism's body are as follows: area, perimeter, length, width, compactness, solidity [30], convexity, roundness, eccentricity [31], robustness, concavity index, concavity ratio [32], shape factor, Feret shape, average width, width ratio, mass fractal dimension, corrected mass fractal dimension, surface fractal dimension, area versus perimeter fractal dimension, mass ratio fractal dimension (overall, lower values, upper values), and Euclidean distance map fractal dimension (overall, lower values, upper values) [33].

Additionally, the following parameters were also determined: cilia presence, stalk presence, stalk width, stalk average width per body average width ratio, tentacles presence, flagella presence [33].

These descriptors were subsequently studied and organized in a manner that allowed the isolation and identification of each species, genus, order or sub-class. Bearing this purpose in mind, the multivariate statistical techniques discriminant analysis, neural networks and decision trees were performed using the Matlab 7.0 platform (The *MathWorks, Natick*).

5. Data processing

5.1. Data organization

The studied micro-organisms were split before the application of the multivariate statistical analysis into two easily recognizable groups: stalked and non-stalked classes. The separation is performed by the user, relying on the fact that the stalk is a well recognizable structure resulting in a simplified and faster multivariate analysis methodology.

Initially, a training set of each of the 22 micro-organisms was used for the determination of the discriminant functions, neural network architecture and decision trees. As in some cases it was not possible to determine if a given individual was an *Epistylis* or an *Opercularia* (closed buccal apparel) an additional class was used resulting thus in 23 classes.

Regarding the stalked group two different analyses were performed: an analysis containing 10 classes with the two *Epistylis* species as two different classes and a second analysis containing 9 classes with the two species represented in a single class. The two values in the *Epistylis* column in Tables 2 and 3 are reported to the cases where the two *Epistylis* species were analyzed as a single class or two different classes, respectively. In the final results however the results of the two *Epistylis* classes were fused together in order to appear only one *Epistylis* class.

For the non-stalked set a total of 18 classes were analyzed due to the fact that two different *Trachellophylum* species were studied and for the *A. cicada*, *Arcella* and *Trithigmostoma* species both front and side views (in separate classes) were treated. For validation purposes a different set of individuals (test set) of each of the 23 classes was used with a third of individual cells number of the training set and the same number of classes. The number of individual cells used in each case is presented in Tables 2 and 3. In the final results however the results of the two *Trachellophylum, A. cicada*, *Arcella* and *Trithigmostoma* classes were fused together in order to appear only one class of each of the above-mentioned micro-organisms.

In this work, the different protozoa and metazoa are represented by: *A. cicada* (acic), *Aelosoma* sp. (aelo), *Arcella* sp. (arce), *Carchesium* sp. (carc), *Coleps* sp. (cole), Digononta order (digo), *Epistylis* sp. (epis), *Euglypha* sp. (eugl), *Euplotes* sp. (eupl), *Litonotus* sp. (lito), Monogononta order (mono), Nematoda sub-class (nema), *Opercularia* sp. (oper), *Peranema* sp. (pera), Suctoria sub-class (suct), *Trachelophyllum* sp. (trac), *Trithigmostoma* sp. (trit), *Trochilia* sp. (troc), *V. aquadulcis* (vaqu), *V. convallaria* (vcon), *V. microstoma* (vmic) and *Zoothamnium* sp. (zoot). When it was not possible to determine if a given cell was an *Epistylis* or an *Opercularia* (closed buccal apparel) the term ep/op was adopted.

5.2. Discriminant analysis

This technique is used to determine which variables discriminate between two or more naturally occurring classes. In a manner similar to factor analysis and principal components analysis, discriminant analysis (DA) defines new variables (discriminant functions) as linear combinations of the original descriptors, increasing the inter-class variability and obtaining

Table 2
Number of individual cells present in the training set

acic	aelo	arce	carc	cole	digo	epis	ep/op	eugl	eupl	lito	mono
134	46	108	67	67	57	67/96	67	67	67	67	67

nema	oper	pera	suct	trac	trit	troc	vaqu	vcon	vmic	zoot
37	47	67	38	86	78	46	67	67	67	67

Table 3
Number of individual cells present in the test set

acic	aelo	arce	carc	cole	digo	epis	ep/op	eugl	eupl	lito	mono
66	23	54	33	33	29	33/47	33	33	33	33	33

nema	oper	pera	suct	trac	trit	troc	vaqu	vcon	vmic	zoot
20	23	33	18	43	39	22	33	33	33	33

thus an increase separation between the studied classes. Additionally, this technique allows the classes of data to be modeled in order to reclassify the given object with a minimum error and classify new objects using the new discriminant functions [34]. The objects coordinates in the new discriminant functions space are determined with the parameters raw data for each object.

Discriminant analysis was performed for the whole set of the micro-organisms training set. The performed discriminant analysis was of a linear type, i.e., the multivariate normal (MVN) density function used was a relative log posterior density function (D) with a pooled estimate of variance. The value of the MVN density function was therefore determined for each of the individual cells regarding all the studied classes for both the training and the test sets.

In the validation process, and in order to determine each micro-organism class, the MVN density function value was determined for all the individual cells on the test set and for each class. Each cell was then attributed to the class where it presented the highest MVN density function value (D) provided that:

$$D < (\bar{D}_g - f\delta_g^D) \tag{1}$$

where $\bar{D}_g$ is the mean value of the MVN density function value for class g, δ_g^D the standard deviation and f is a factor ranging from 0.25 to 5 in 0.25 step values. Micro-organisms that do not conform Eq. (1) were classified as non-identified. For the stalked group the best f value was 5, whereas for the non-stalked group was 2.75, for the totality of the classes.

5.3. Neural networks

Neural networks (NN) have been successfully employed in numerous applications, ranging from data forecasting to medical diagnosis. Generally, NN are constructed of many processing elements (PEs) that are linked in a certain way. Fig. 3 shows a typical structure of the multi-layer perceptron neural network. The detailed description on a multi-layer perceptron is given in Ref. [35].

Essentially, a multi-layer perceptron is comprised of three different layers: input layer, hidden layer and output layer. The hidden layer maps the input pattern x with output pattern y through a series of interconnected weights.

$$u = \sum_{i=0}^{n} x_i w_i \tag{2}$$

where u is the aggregated input signal; w_i is the weight of the ith input vector which is connected to ith PE. Following this equation, learning in NN implies that the PE changes the output in response to the input change by adjusting the connecting weights. In so doing, u must be further processed by the activation function, resulting in the final neuron's output signal such as:

$$f_0 = f_a(u) \tag{3}$$

The activation function determines the processing inside the neuron. It can be linear or non-linear function depending on the network topology. In this work, the logistic sigmoidal activation function was selected. This function is given by,

$$f_a = \frac{1}{1 + e^{-n}} \tag{4}$$

In order to obtain a model with validation capabilities, data is split into two data sets: a training set used to obtain the neural model, and a validation set used to test the behavior of the model with data different from the training set used to obtain the model.

Two different approaches can be carried out to train the network. The first approach is based on updating the synaptic weights for each pattern of the data set (on-line approach) whereas the other approach is based on updating the synaptic weights once for all the training data set (batch approach).

The neural network used in this paper was a feed forward neural network (back propagation algorithm) consisting of a two-layer network with no hidden layers. Levenberg–Marquardt optimization and resilient optimization functions were tested for the stalked micro-organisms training whereas for the non-stalked only the resilient optimization function was studied. The

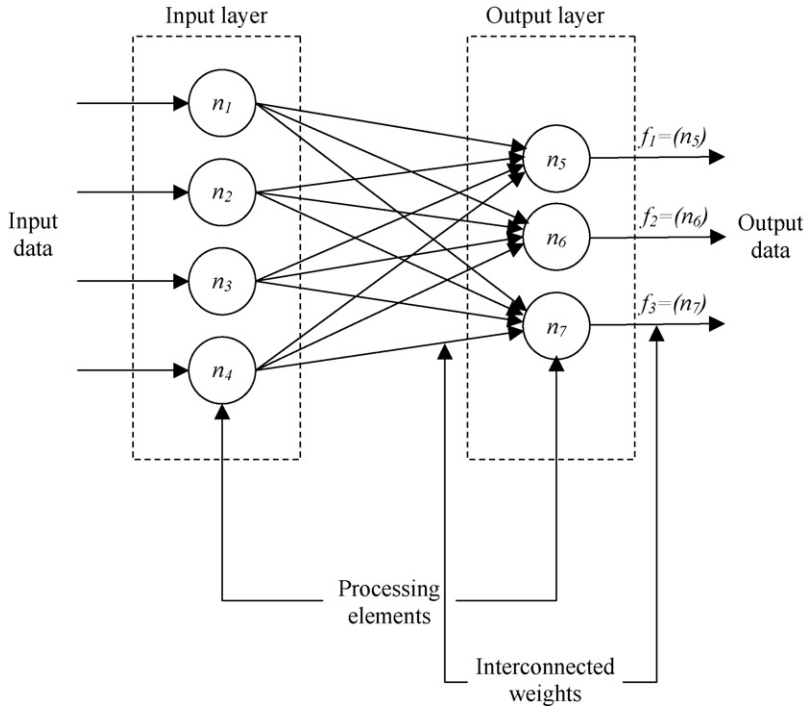

Fig. 3. Multi-layer perceptron neural network with an input layer and an output layer.

chosen learning function in both cases was the gradient descent algorithm and the mean squared error was used as a performance (error) function with its goal set to zero. For each one of these configurations 100 tests with different initial values for a maximum of 500 iterations for each test were carried out. Additionally, two configurations of layer nodes were tested for the stalked organisms: 10/10 and 15/10 (when the two *Epistylis* species were analyzed as different classes) and 9/9 and 14/9 (when the two *Epistylis* species were analyzed as a single class) for the first/second layer, respectively. For the non-stalked organisms, the number of nodes tested were of 18/18 and 11/18, respectively, for the first/second layer.

The neural network was performed for the whole set of the micro-organisms training set. The application of NN aimed to obtain an output value of 1 for the correct microbial class and 0 for all the other classes. Therefore, each micro-organism was attributed to the class with a single higher output value larger than 0.01, and micro-organisms with more than a single maximum class output were classified as non-identified.

5.4. Decision trees

In order to classify each protozoa and metazoa dedcision trees were also tested. In the case being, and although it was intended to ultimately classify the micro-organisms, the decision trees (DT) that were used were not classification trees but instead regression trees. That was so to prevent the classification into one of the existing classes of a micro-organism regardless of belonging to any of the classes current database or being a non-studied class. By using regression trees and establishing threshold values for a given micro-organism to be classified into a class this model is able to classify new micro-organisms as not belonging to any of the classes current database. A regression tree is a predictive model based on the ability to submit the input data matrix with a series of consecutive yes or no questions, and accurately predict a given response vector. Each question evaluates a given condition (either continuous or discrete) and, depending on the answer proceeds to a new question or arrives at the fitted response value. However, one should be careful to avoid over fitting. In fact, a decision tree might be trained to fit so perfectly the data set that would not be appropriated for predicting new values. That is so when the tree has too much branches and the lower ones are strongly affected by outliers and other artefacts on the data set. One way to determine the best tree size is by cross-validation, which determines a resubstitution estimate of the error variance and a cross-validation estimate for the tree and for a series of pruned trees. Then the best tree is chosen as the tree presenting the residual variance that is no more than one standard error above the minimum value along the cross-validation line [36].

6. Results and discussion

The results obtained for the studied neural network procedures allowed to determine small to negligible differences between the 18/18 and 11/18 non-stalked neural architectures although for the reduced and normalized results slight improvements in the 18/18 were observed. However, and given the higher computing speed in the 11/18 configuration it was found that this architecture complies with this work objectives. With respect to the stalked micro-organisms the configuration 15/10 allowed for better results and therefore, proved to be a real advantage over the 10/10 architecture.

Table 4
Recognition percentages, misclassification error and overall recognition performances of the non-stalked classes

	Discriminant analysis			Neural networks			Decision trees		
	Rec.%	Misc.%	Overall%	Rec.%	Misc.%	Overall%	Rec.%	Misc.%	Overall%
acic	86.4	0.2	86.2	87.9	0.5	87.9	92.4	0.5	92
aelo	100.0	0.0	100	95.7	0.0	95.7	91.3	0.0	91.3
arce	96.3	1.1	95.2	92.6	0.5	92.2	88.9	0.5	88.5
cole	97.0	0.2	96.8	97.0	0.7	96.3	90.9	0.4	90.5
digo	82.8	0.9	82	86.2	0.9	85.5	82.8	0.9	82
eugl	97.0	1.5	95.5	93.9	1.1	92.9	84.8	0.7	84.3
eupl	78.8	1.1	77.9	87.9	0.4	87.5	90.9	0.9	90.1
lito	97.0	0.2	96.8	97.0	1.3	95.7	87.9	0.7	87.3
mono	81.8	1.3	80.8	84.9	0.9	84.1	87.9	2.6	85.6
nema	95.0	0.0	95	90.0	0.0	90	95.0	0.0	95
pera	100.0	0.2	99.8	97.0	0.2	96.8	100.0	0.4	99.6
trac	100.0	0.0	100	97.7	0.7	97	97.7	0.4	97.2
trit	89.7	0.4	89.3	79.5	1.1	78.6	89.7	1.5	88.4
troc	100.0	0.6	99.4	95.5	0.4	95.1	100.0	0.0	100
NI	0.2			1.2			0		

NI—not identified.

The results obtained in the identification of the stalked and non-stalked organisms after applying the three statistical multivariate techniques are presented in Tables 4 and 5.

The recognition percentage (Rec.%) of the protozoa and metazoa classes was estimated as the ratio between the number of micro-organisms correctly classified in a given class and the total number of analyzed micro-organisms in that class. The misclassification error (Misc.%) was determined as the ratio between the total number of micro-organisms incorrectly classified in a given class and the total number of micro-organisms belonging to all the other classes. The overall recognition performance (Overall%) for each class was determined by multiplying the corresponding recognition percentage by the factor (100 − Misc.%) and then dividing by 100.

As observed in Table 4, in general, for all of the non-stalked organisms the class recognition percentages were above 81%, except for *Euplotes* (78.8%) with discriminant analysis and *Trithigmostoma* (79.5%) with neural networks. On the other hand, it was evidenced that the misclassification error was negligible (bellow 2% except for Monogononta in decision trees) for all the non-stalked organisms included in the study. This results yielded good overall recognition performances (higher than 80% except for *Euplotes* in DA and *Trithigmostoma* in NN). Additionally, less than 2% of the non-stalked organisms were not identified in all three techniques used to analyze the experimental data descriptors.

However, the class recognition percentage reached for the stalked organisms (Table 5) were quite poor for the three multivariate analysis techniques when compared with the recognition percentages observed for the non-stalked organisms, with values lower than 83% (except for *Suctoria*). Also the misclassification error presented high values up to 9.5% for the stalked micro-organisms, and the overall recognition performance did not surpass 82% in all cases (except for *Suctoria*).

The global recognition percentage for each group was determined as the ratio between the number of micro-organisms correctly classified in a given group and the total number of analyzed micro-organisms in that group. The overall recognition performance, for the global stalked and non-stalked groups (Table 6), was dependent, as expected, on the individual class values and was found better for the non-stalked organisms, with values higher than 83% in all cases. At the same time, the global recognition percentages remained above 91% and the global misclassification error below 9% for the non-stalked

Table 5
Recognition percentages, misclassification error and overall recognition performances of the stalked classes

	Discriminant analysis			Neural networks			Decision Trees		
	Rec.%	Misc.%	Overall%	Rec.%	Misc.%	Overall%	Rec.%	Misc.%	Overall%
carc	54.5	3.2	52.8	60.6	3.6	58.5	48.5	4.4	46.4
epis	83.0	6.3	77.8	72.3	4.6	69	59.6	1.3	58.8
ep/op	54.6	2.8	53	66.7	3.2	64.6	78.8	7.1	73.2
oper	82.6	1.5	81.4	82.6	3.8	79.5	65.2	3.0	63.2
suct	88.9	0.4	88.6	94.4	0.4	94.1	88.9	0.4	88.6
vaqu	66.7	2.8	64.8	69.7	3.2	67.5	69.7	7.5	64.5
vcon	81.8	3.6	78.9	66.7	2.8	64.8	57.6	1.2	56.9
vmic	69.7	6.3	65.3	72.7	5.1	69	45.5	8.3	41.7
zoot	69.7	4.7	66.4	63.6	6.7	59.4	60.6	9.5	54.9
NI	0.7			0			0		

Table 6
Global recognition, misclassification percentages and overall recognition performances for the stalked and non-stalked groups

	Discriminant analysis			Neural networks			Decision trees		
	Rec.%	Misc.%	Overall%	Rec.%	Misc.%	Overall%	Rec.%	Misc.%	Overall%
Stalk	71.7	27.6	51.9	70.6	29.4	49.9	62.2	37.8	38.7
Non-stalk	92.5	7.3	85.8	91.3	7.5	84.5	91.3	8.7	83.3

organisms in all three multivariate techniques. In terms of the statistical techniques, all of the techniques yield small differences regarding the overall recognition performance achieved for the non-stalked group, although the decision trees technique resulted in the lowest overall recognition performance (83.3%) and highest misclassification error (8.7%). These results can be considered fairly well with respect to the overall recognition of protozoa and metazoa, even more considering the complexity of the image acquisition of some microbial groups such as crawling ciliates, partly due to their movement on the flocs surfaces which makes difficult the subsequent identification of the external structures.

On the other hand, the overall recognition performances, obtained for the identification of the stalked group was quite inferior (bellow 52% in all cases) to the non-stalked group, reflecting the poor global recognition percentages and high misclassification percentages of the classes. It was also evidenced that the decision trees data processing resulted in the lowest overall recognition performance (38.7%) due to the lowest global recognition percentage (62.2%) and highest misclassification error (37.8%). These results were considerably inferior to those obtained for non-stalked organisms due to the fact some species of stalked organisms such as *V. convallaria*, *V. microstoma*, *Carchesium* and *Zoothamnium*, for instance, are morphologically very similar on their projected form and, thus, hard to distinguish from each other.

The global recognition percentages, misclassification error and overall recognition performances for the DA, NN and DT techniques regarding the protozoa and metazoa groups (sarcodines, cilates, flagellates and metazoa) and protozoan ciliates (carnivorous, free-swimming, crawling, sessile and non-ciliates) identification are presented in Table 7.

It could be observed that for the three multivariate statistical techniques applied to the morphological data, the identification of the main protozoa and metazoa groups as well as for the recognition of the protozoan ciliates, the global recognition percentages were comparable and higher than 95% for all of the analyzed groups. In the same way, the misclassification error was not significant for all the applied techniques (bellow 4%) resulting in overall recognition performances higher than 92% in all cases.

Regarding the WWTP performance, as mentioned above, and according to Madoni [24], Canler et al. [37] and Curds [22], a relationship between the protozoa and metazoa classes included in the present study can be observed with the operational conditions of the WWTP. According to these authors the species belonging to the genus *Opercularia* sp., *Trachelophyllum* sp., and *V. microstoma* are indicators of low effluent quality; whereas *Aelosoma* sp., *Arcella* sp., *Carchesium* sp., *Epistylis* sp., *Euglypha* sp., *Euplotes* sp., order Monogononta, *Peranema* sp., *Trithigmostoma* sp., *Trochilia* sp., *V. aquadulcis* and *Zoothamnium* sp., are indicators of high quality of the treated effluent.

On the other hand, the organisms belonging to the Nematoda sub-class, *Opercularia* sp. and *V. microstoma* point towards poor aeration of the activated sludge (below 0.2–0.5 mg O_2 L^{-1}), while *Aelosoma* sp., *Carchesium* sp., *Euglypha* sp., *Arcella* sp., Monogononta order, *Trochilia* sp., *V. aquadulcis* and *Zoothamnium* sp. are indicators of a satisfactory aeration (above 1–2 mg O_2 L^{-1}).

The nitrification process can be inferred by the occurrence in the aerated tank of micro-organisms belonging to the *Aelosoma* sp., *Arcella* sp., *Carchesium* sp., *Coleps* sp., *Epistylis* sp., *Euplotes* sp., *Trochilia* sp. and the Monogononta order.

Furthermore, the presence in the aeration tanks of *Peranema* sp. and *V. microstoma* is an indication of fresh sludge (few days), while *Aelosoma* sp., *Arcella* sp., *Euglypha* sp., and the Digononta and Monogononta orders have been pointed out as indicators of old sludge (20 days or more).

The values obtained for the overall plant operating conditions assessment is showed in Table 8 where it is possible to observe that the DA and NN techniques yield comparable results for the recognition percentage, misclassification and overall performance. Meanwhile, the overall results were generally more affected when the DT techniques was used on the operating conditions assessment. Nevertheless, these results can be considered fairly reasonable for the effluent quality, aeration and nitrification evaluation and quite good on the sludge age determination.

Table 7
Global recognition percentages, misclassification error and overall recognition performances for the protozoa and metazoa groups and protozoan ciliates

	Discriminant analysis			Neural networks			Decision trees		
	Rec.%	Misc.%	Overall%	Rec.%	Misc.%	Overall%	Rec.%	Misc.%	Overall%
Groups	97.4	2.2	95.3	97.0	2.2	94.9	97.3	2.7	94.7
Ciliates	97.2	2.4	94.8	95.4	3.8	92.2	96.4	3.6	92.9

Table 8
Global recognition percentages, misclassification error and overall recognition performances of the groups indicating of final effluent quality, aeration, nitrification and sludge age assessment

	Discriminant analysis			Neural networks			Decision trees		
	Rec.%	Misc.%	Overall%	Rec.%	Misc.%	Overall%	Rec.%	Misc.%	Overall%
Effluent quality	89.6	10.0	80.7	89.8	9.4	80.5	88.3	11.7	78.0
Aeration	90.9	8.7	83.0	91.0	8.2	82.6	89.6	10.4	80.3
Nitrification	90.1	9.5	81.6	90.5	8.7	82.1	89.2	10.8	79.6
Sludge age	94.2	5.4	89.2	94.0	5.2	89.3	92.6	7.4	85.7

Table 9
Global recognition percentages, misclassification error and overall recognition performances for the critical conditions applying the three multivariate techniques

	Discriminant analysis			Neural networks			Decision trees		
	Rec.%	Misc.%	Overall%	Rec.%	Misc.%	Overall%	Rec.%	Misc.%	Overall%
Low effluent quality	89.9	2.3	87.8	90.9	3.1	88.1	82.8	3.1	80.3
Low aeration	85.5	2.3	83.6	86.8	2.6	84.6	77.6	2.7	75.5
Fresh sludge	84.8	2.4	82.8	84.8	2.0	83.2	72.7	3.2	70.4

Close related to the overall results, the assessment of critical conditions such as low effluent quality, low aeration, and fresh sludge shown in Table 9, attained reasonable to good recognition percentage, misclassification error and overall recognition performance levels for the three studied statistical multivariate techniques. It could also be observed that the recognition and misclassification percentages attained the best results for the neural networks and discriminant analysis techniques while the lowest values were again obtained for the Decision Trees technique. Thus, the NN and DA techniques seem to be the most appropriate to perform the WWTP critical operational conditions assessment and the DT the less appropriate.

7. Conclusions

It was possible to include in the analyzed classes two different *Trachellophylum* and *Epistylis* species and both front and side views of *A. cicada*, *Arcella* and *Trithigmostoma* species (in separate classes), which improved the image acquisition process as well as the recognition and identification of such species.

The developed procedure proved to be robust to distinguish the non-stalked organisms with overall recognition performances above 85% for most of the species, which could be considered quite reasonable regarding the complexity in the acquisition of the images of several species, mainly the crawling and free-swimming protozoa. Furthermore, the overall recognition performance for the entire non-stalked group was above 83%, which can be considered fair. Regarding the stalked microorganisms, the overall recognition performance levels were lower than 75% for most species using the three multivariate techniques, resulting in an overall recognition performance for the entire stalked group bellow 52%. These results were markedly lower than those achieved for non-stalked organisms, due mainly to the similar morphology of some classes like *V. convallaria*, *V. microstoma*, and *V. aquadulcis* and also for *Carchesium* and *Zoothamnium*, which made these organisms hard to distinguish among each other.

With respect to the identification of the main protozoa and metazoa groups (flagellate, ciliate, sarcodine and metazoa) as well as the ciliated protozoa groups (carnivorous, crawling, free-swimming and sessile), the results could be considered quite good. The overall recognition performances were above 94% for the overall protozoa and metazoa groups and above 92% for the overall ciliated protozoa groups and for all three multivariate techniques. These results are important from the point of view of the WWTP operation and the quality of the final effluent assessment as the ciliates organisms are very important indicators of those. Hence, and regarding the assessment of operational WWTP conditions, the overall recognition performances were fairly reasonable (above 80% in most cases) for all three multivariate techniques. The assessment of critical or mediocre operational conditions such as low effluent quality, low aeration and fresh sludge proved to be likewise satisfactory in terms of the overall recognition levels.

Finally, the results proved that the image analysis coupled with the multivariate statistical techniques is a promising tool for assessing and monitoring protozoa and metazoa populations in a WWTP. Of the three multivariate techniques evaluated in this study the decision trees was found to be the less appropriate to be used in the identification of the protozoa and metazoa, whereas the discriminant analysis and neural networks attained close results.

Acknowledgements

The authors are grateful to the National Council of Scientific and Technological Development of Brazil (CNPq), the BI-EURAM III ALFA co-operation project (European Commission), the POCI/AMB/57069/2004 project supported by the Fundação para a Ciência e a Tecnologia (Portugal). Data

from Nancy plant made available by Prof. Maurício da Motta (UFPE, Recife, Brasil) is also acknowledged.

References

[1] H. Hawkes, in: K. Chater, H. Somerville (Eds.), The Oil Industry and Microbial Ecosystem, Heyden, London, 1978.
[2] P. Madoni, Water Res. 28 (1994) 67.
[3] W. Foissner, Hydrobiology 166 (1988) 1.
[4] G. Esteban, C. Tellez, L.M. Bautista, Environ. Technol. 12 (1990) 381.
[5] P. Madoni, D. Davoli, E. Chierici, Water Res. 27 (1993) 1485.
[6] H. Salvadó, M.P. Gracia, J.M. Amigó, Water Res. 29 (1995) 1041.
[7] M. Martin-Cereceda, S. Serrano, A. Guinea, FEMS Microbiol. Ecol. 21 (1996) 267.
[8] T. Sasahara, T. Ogawa, Monatsschrift für Brauwissenschaft 11 (1983) 443.
[9] Macek, Int. Rev. Ges. Hydrobiol. 74 (1989) 643.
[10] S. Yun-Fen, A.L. Buikema Jr., W.H. Yongue Jr., J.R. Pratt, J. Cairns Jr., J. Protozool. 33 (1986) 146.
[11] S. Chen, M. Xu, H. Cao, J. Zhu, K. Zhou, J. Xu, X. Yang, Y. Gan, W. Liu, J. Zhai, Y. Shao, Eur. J. Protistol. 40 (2004) 147.
[12] S.M. Al-Shahwani, N.J. Horan, Water Res. 25 (1991) 633.
[13] G. Esteban, C. Tellez, L.M. Bautista, Water Res. 25 (1991) 967.
[14] J. Fried, G. Mayr, H. Berger, Water Sci. Technol. 41 (2000) 309.
[15] P. Madoni, Hydrobiology 111 (1984) 201.
[16] P. Madoni, Eur. J. Protistol. 36 (2000) 465.
[17] P. Madoni, P.F. Ghetti, Hydrobiology 83 (1981) 207.
[18] P. Madoni, D. Davoli, G. Gorbi, Water Res. 30 (1996) 135.
[19] J.E.P. Poole, Water Res. 18 (1984) 281.
[20] J.E.P. Poole, J.C. Fry, J. Water Pollut. Control Fed. 79 (1980) 19.
[21] H. Salvadó, M.P. Gracia, Water Res. 27 (1993) 891.
[22] C.R. Curds, Annu. Rev. Microbiol. 36 (1982) 27.
[23] C.R. Curds, H.A. Hawkes (Eds.), Biological Activities and Treatment Processes. Ecological Aspects of Used-Water Treatment, Academic Press, London, 1983 (Chapter 2).
[24] P. Madoni, Wat. Sci. Technol. 29 (1994) 63.
[25] M. Richard, Activated Sludge Microbiology, second ed., The Water Pollution Control Federation, Virginia, 1991.
[26] Commission of the European Communities, in: F. S. Woodiwis (Ed.), Biological monitoring of surface water quality, Summary Report, ENV/787/80-EN, 1980, 45 pp.
[27] A.L. Amaral, C. Baptiste, M.N. Pons, A. Nicolau, N. Lima, E. Ferreira, M. Mota, H. Vivier, Biotechnol. Tech. 13 (1999) 111.
[28] A.L. Amaral, M. da Motta, M.N. Pons, H. Vivier, N. Roche, M. Mota, E.C. Ferreira, Environmetrics 15 (2004) 381.
[29] M. da Motta, M.N. Pons, H. Vivier, A.L. Amaral, E.C. Ferreira, M. Mota, Braz. J. Chem. Eng. 18 (2001) 103.
[30] C.R. Russ, The Image Processing Handbook, CRC Press, Boca Raton, 1995.
[31] C.A. Glasbey, G.W. Horgan, Image Analysis for the Biological Sciences, John Wiley and Sons, Chichester, 1995.
[32] M.N. Pons, H. Vivier, J. Dodds, Part. Syst. Charact. 14 (1997) 272.
[33] A.L. Amaral, Image analysis in biotechnological processes: applications to wastewater treatment, Ph.D. Thesis, 2003.
[34] J.W. Einax, H.W. Zwazinger, S. Geiss (Eds.), Chemometrics in Environmental Analysis, VHC Verlagsgesellshaft, Weinheim, 1997.
[35] S. Haykin, Neural Networks: A Comprehensive Foundation, Prentice Hall, Englewood Cliffs, NJ, 1999.
[36] L. Breiman, J. Friedman, R. Olshen, C. Stone (Eds.), Classification and Regression Trees, Wadsworth, Belmont, 1984.
[37] J.P. Canler, J.M. Perret, F. Duchene, E. Cotteaux, in: Aide au Diagnostic des Stations d'Èpuration par l'observation Microscopique des Boues Activeés, Cemagref Editions.