RAIRO-Oper. Res. 49 (2015) 735–751
DOI: 10.1051/ro/2015002

RAIRO Operations Research
www.rairo-ro.org

THE USE OF THE HYPERBOLIC SMOOTHING CLUSTERING ALGORITHM IN TAXONOMY OF MACROALGAE

Maria Gardênia Sousa Batista[1], Francisca Lúcia de Lima[1], André Macedo Santana[2] and Adilson Elias Xavier[3]

Abstract. This work proposes a new methodological approach for grouping data in taxonomy. Macroalgae of the genus *Caulerpa* were selected as a study model on basis of their remarkable morphological plasticity, and of the difficulty in identifying those algae using the traditional systematical methods. The results obtained from the application of the hyperbolic smoothing algorithm demonstrate the feasibility of its use in biological taxonomy. The new methodology herein proposed may be used isolatedly or in association with other methodologies already proven, not only in phycology, but also in other areas of biology.

Keywords. Hyperbolic smoothing, biological taxonomy, macroalgae.

Mathematics Subject Classification. 92-XX, 92B10.

1. Introduction

Human beings have always attempted to classify the objects around, whether animated or inanimated. Classifying the objects into collective categories is a requisite for assigning them a name. Grouping objects is to recognise that such objects are sufficiently similar to be placed in the same group; this is also accomplished for identifying distinctions or separations between groups [43].

Received September 18, 2014. Accepted January 5, 2015.

[1] State University Piauí – UESPI, Brazil
batistamariagardenia@gmail.com; karnauba@gmail.com

[2] Federal University of Piauí – UFPI, Brazil andremacedo@ufpi.edu.br

[3] Federal University of Rio de Janeiro – UFRJ, Brazil adilsonxavier@gmail.com

Article published by EDP Sciences

Taxonomy or systematics is the science that deals with identification, nomenclature, and classification of objects of biological nature [40].

Biological systematics synthesises and ranks the diverse forms of life. Due to the mega biodiversity and the so-called "hot spots" and their interactions, studies on the field of systematics need to deal with the interaction between the different areas of biology and of computational mathematics. Knowledge of computational methods have a deep impact on the practices of systematics, not only for adding speeding, but also for making taxonomic data more sophisticated and effective, besides facilitating the dissemination and exchange of information, especially on the world computer network [2, 14].

The increase of knowledge of systematic data, especially of DNA, and the incorporation of new methodologies have leveraged new forms of analysing data and of practising the systematics. The fast-pace emergence of knowledge on new phylogenetic hypotheses enables the development of new methodologies and tools for inferring phylogenetic relationships, and provides the availability of new forms of evidence, enabling an increasingly accurate description of the evolutionary history [4, 25].

Systematics has evolved along with other sciences, including computation. However, there is still lack of knowledge to establish the quantity of vegetable and animal life existent on Earth, due to the shortage of taxonomists and the lack of relationship between biological knowledge and its mathematical and statistical language [4, 32, 51].

Since the pioneer work by Fisher [18] on the use of grouping methods based on metrics applied to plants of the genus *Iris* (*Iris versicolour, Iris setosa, Iris virginica*), the number of reports on the use of grouping algorithms with biological data has increased much along time.

Works such as those of [32, 42], which report on the efficiency of computational methods in systematic biology, have been frequent especially for studies of phylogeny [33, 37], conservation and biodiversity [26], particularly in phycology.

Several grouping techniques are described in the literature [28, 29, 38], and each researcher has to apply his own wisdom to choose the most suitable to his particular purpose, since different techniques may lead to different solutions [39].

One of the most used algorithms is the *k-means*. Although it has been proposed over 50 years ago, it is still traditionally used as a quick tool that can be easily understood and implemented [27, 45].

The clustering problem has applications in the most varied areas of research, including, for example: graphics and visual computation, medical computation, computational biology, communications networks, transport engineering, computer networks, manufacturing systems, among others [30, 44, 67].

In general, obtaining the solution to a clustering problem corresponds to the process of grouping the elements (objects) of a database (the set) in such a way that the groups formed, or *clusters*, represent a configuration where each element has more similarity with any element of that same *cluster* than it has with elements of other *clusters*. The purpose of such grouping process (clustering) is to

gather the units of the sample in groups, by any classification criterion, with basis on similarity [5, 6] in a way that there is homogeneity within the group and heterogeneity between groups [23, 52, 65].

Those clustering methods use diverse algorithms [19, 52], among them the hierarchical algorithms and those of partitioning [1, 17, 34, 65].

With traditional hierarchical clustering algorithms, the formation of clusters occur gradually through agglomerations or divisions of elements/*clusters*, generating a hierarchy of *clusters*, usually represented by a tree structure or dendrogram [15].

The partitioning methods have the set of elements divided into $k$subsets, $k$being known or not, and each configuration obtained is evaluated by means of an objective function. If the clustering evaluation indicates that the configuration is not appropriate for the problem in hand, a new configuration is obtained by migrating elements between *clusters*, and the process goes on in an iterative form until some criterion for stop is met. Thus, the *clusters* can be gradually improved, which do not occur with the hierarchical methods [69].

In the present work another algorithm will be used: one that is a novelty in the literature, but has been already showing an excellent performance, the hyperbolic smoothing clustering method (HSCM) [66]. This paper will expose the use of this new algorithm for classification of macroalgae with basis on biometric data of the genus *Caulerpa*. The clusters formed through the use of this new methodology will de analysed and discussed with basis on studies of the phylogeny of these algae.

2. Clustering problems

Let $S = \{s_1, \ldots, s_m\}$ denote a set of m patterns or observations from an Euclidean space with n components, to be clustered into a given number q of disjoint clusters. To formulate the original clustering problem as a min – sum – min problem, we proceed as follows. Let $x_i, i = 1, \ldots, q$ be the centroids of the clusters, where each $x_i \in \Re^n$. The set of these centroids will be represented by $x = (x_1, x_2, \ldots, x_q) \in R^{nq}$. Given a point s_j of S, we initially calculate the Euclidean distance from s_j to the nearest centroid. This is given by $z_j = \min_i \|s_j - x_i\|_2, j = 1, \ldots, m$. The most frequent measurement of the quality of a clustering associated to a specific position of q centroids is provided by the minimum sum of squares (MSSC) of these distances:

$$\text{minimize} \sum_{j=1}^{m} z_j^2 \tag{2.1}$$

subject to: $z_j = \min_i \|s_j - x_i\|_2, j = 1, \ldots, m.$

By performing an $\varepsilon \geqslant 0$ perturbation of problem (2.1) and by using an auxiliary function $\phi(y) = \max\{0, y\}$ we obtain the modified problem:

$$\text{minimize} \sum_{j=1}^{m} z_j^2 \tag{2.2}$$

subject to: $\sum_{i=1}^{q} \phi\left(z_j - \|s_j - x_i\|_2\right) \geqslant \varepsilon, \quad j = 1, \ldots, m.$

Since the feasible set of problem (2.1) is the limit of (2.2) when $\varepsilon \to 0_+$ we can then consider solving (2.1) by solving a sequence of problems like (2.2) with a sequence of decreasing values for ε approaching 0.

However, analyzing problem (2.2), the definition of the function ϕ turns it into an extremely rigid non-differentiable structure, which makes its computational solution very hard. In view of this, the numerical method we adopt for solving problem (2.2) takes a smoothing approach. From this perspective, let us define the function $\phi(y,\tau) = (y+\sqrt{y^2+\tau^2})/2$, which has the following asymptotic property $\lim_{\tau\to 0} \phi(y,\tau) = \phi(y)$. So, by using the function ϕ in the place of the function ϕ, the problem turns into

$$\text{minimize} \sum_{j=1}^{m} z_j^2 \tag{2.3}$$

$$\text{subject to: } \sum_{i=1}^{q} \phi\left(z_j - \|s_j - x_i\|, \tau\right) \geqslant \varepsilon, \quad j = 1,\ldots,m.$$

To obtain a differentiable problem, it is still necessary to smooth the Euclidian distance $\|s_j - x_i\|_2$. So, let us define the approximation function $\theta_2(s_j, x_i, \gamma) = \sqrt{\sum_{l=1}^{n} \left(s_j^l - x_i^l\right)^2 + \gamma^2}$, which has the asymptotic property $\lim_{\tau\to 0} \theta_2(s_j, x_i, \gamma) = \|s_j - x_i\|_2$. By using the approximation smooth function θ of the Euclidian distance $\|s_j - x_i\|_2$ we obtain the following completely differentiable problem:

$$\text{minimize} \sum_{j=1}^{m} z_j^2 \tag{2.4}$$

$$\text{subject to: } \sum_{i=1}^{q} \phi\left(z_j - \theta(s_j, x_i, \gamma), \tau\right) \geqslant \varepsilon, \quad j = 1,\ldots,m.$$

Since $z_j \geqslant 0$, $j = 1,\ldots,m$, the objective function minimization process will work for reducing these values to the utmost. On the other hand, given any set of centroids $x_i, i = 1,\ldots,q$, each term $\phi\left(z_j - \theta(s_j, x_i, \gamma), \tau\right)$, $j = 1,\ldots,m$, is an increasing convex C^∞ function in variable z_j. So, these constraints will certainly be active and problem (2.4) will finally be equivalent to problem:

$$\text{minimize} \sum_{j=1}^{m} z_j^2 \tag{2.5}$$

$$\text{subject to: } h_j(z_j, x) = \sum_{i=1}^{q} \phi\left(z_j - \theta(s_j, x_i, \gamma), \tau\right) - \varepsilon = 0, j = 1,\ldots,m.$$

Problem (2.5) is completely differentiable and due the properties of functions ϕ and θ we can seek a solution to problem (2.1) by solving a sequence of sub-problems like problem (2.5), produced by decreasing the parameters $\gamma \to 0$, $\tau \to 0$, $\varepsilon \to 0$.

The dimension of the variable domain space of problem (2.5) is $(nq + m)$. As, in general, the value of the parameter m, the cardinality of the set S of the observations, is large, problem (2.5) has a large number of variables. However, it has a separable structure because each variable z_j appears only in one equality constraint. Therefore, as the partial derivative of $h_j(z_j, x)$ with respect to z_j, $j = 1, \ldots, m$ is not equal to zero, it is possible to use the Implicit Function Theorem to calculate each component z_j, $j = 1, \ldots, m$ as a function of the centroid variables $x_i, i = 1, \ldots, q$. This way, the unconstrained problem

$$\text{minimize} \sum_{j=1}^{m} z_j(x)^2 \tag{2.6}$$

is obtained, where each $z_j(x)$ results from the calculation of a zero of each equation

$$h_j(z_j, x) = \sum_{i=1}^{q} \phi\left(z_j - \theta\left(s_j, x_i, \gamma\right), \tau\right) - \varepsilon \;=\; 0, \quad j = 1, \ldots, m. \tag{2.7}$$

Again, due to the Implicit Function Theorem, functions $z_j(x)$ have all derivatives with respect to the variables $x_i, i = 1, \ldots, q$ and therefore it is possible to calculate the gradient of the objective function of unconstrained problem (2.6)

$$\nabla F(x) = \sum_{j=1}^{m} 2\, z_j(x)\; \nabla z_j(x) \tag{2.8}$$

where

$$\nabla z_j(x) = -\nabla h_j(z_j, x) \;/\; \frac{\partial\, h_j(z_j, x)}{\partial\, z_j}, \tag{2.9}$$

while $\nabla h_j(z_j, x)$ and $\partial\, h_j(z_j, x) \;/\; \partial\, z_j$ are obtained from equation (2.7) and from definitions of functions ϕ and θ. In this way, it is easy to solve the unconstrained problem (2.6) by making use of any method based on first order derivative information. At last, it must be emphasized that problem (2.6) is defined on a (nq)-dimensional space, so it is a small problem, since the number of clusters, q, is usually very small for real-world applications.

The solution of the original clustering problem can be obtained by using the general Hyperbolic Smoothing Clustering Method, HSCM, described below in a simplified form:

The HSCM Method

Initialization Step:

Choose initial values: x^0, γ^1, τ^1, ε^1

Choose a reduction factor $0 < \rho < 1$, let $k = 1$.

Main Step: Repeat until an arbitrary stopping rule is attained

Solve problem (2.6) with $\gamma = \gamma^k$, $\tau = \tau^k$ and $\varepsilon = \varepsilon^k$, starting at the initial point x^{k-1} and let x^k be the solution obtained.

Let $\gamma^{k+1} = \rho\,\gamma^k$, $\tau^{k+1} = \rho\,\tau^k$, $\varepsilon^{k+1} = \rho\,\varepsilon^k$, $k = k + 1$.

3. The problem of classification of the algae

The algae form a heterogeneous set of organisms that vary in size, from the unicellular and microscopic ones to the gigantic marine macroalgae. The algae are also major contributors to biodiversity, with an estimated number of 10 million species. The science that studies the algae is named phycology [21, 41].

While it is not possible to describe the organisation of primitive life with absolute surety, fossil records strongly indicate that organisms similar to the present algae were alive more than three billion years ago. This does not permit us to assert categorically that the algae were the earliest living beings, for fossil records are always incomplete, but there are signs that the algae, along with bacteria and certain fungi, are extremely ancient organisms which, due to the photosynthesis process, are responsible for the composition and structure of the earth atmosphere as we know it today, enabling life on the surface of the Planet [24, 35, 58].

Because the algae do not constitute a definite taxonomic category, but a group of diversified categories, so contrasting that fit into three different kingdoms (Monera, Protista and *Plantae*), they even bear different denominations in other systems based on molecular biology data [8]. The essentially negative nature of the definition of algae demonstrates that their classification is extremely complex and remains in full evolution [11, 47, 53, 56].

Several forms of classifying the organisms are known, and three of them deserve being highlighted with respect to classification of the algae: morphological taxonomy, molecular taxonomy, and chemical taxonomy or chemotaxonomy [16, 44, 54, 60].

- morphological taxonomy is the classification based on traditional morphology criteria, *via* observation of the external characteristics. It has been traditionally the most frequently used form of classifying an organism;
- molecular taxonomy uses molecular data for the taxonomic studies based on the DNA or protein sequences.
- chemical taxonomy or chemotaxonomy is the classification based on the chemical constituents of the organisms, that is, on the production of the natural products (metabolites) by those organisms.

Nevertheless, the classification of the algae is still a much discussed subject, since the number of divisions and the classes that include the algae vary from one system of classification to another, from author to author, even in consequence of the relative simplicity of the majority of the algae.

It is not our intent to discuss here the pros and against of dozens of systems of classification of the algae, but to emphasise the importance of the use of a new methodology that will contribute to the interpretation of the information obtained along the process of identification of the measures, so as to be able to precisely evaluate the metric limits of the population under study, and simultaneously be able to observe multiple aspects during the formation of the groupings.

With regard to morphological identification, it is important to stress that even experienced taxonomists may have difficulty in differentiating species that are very

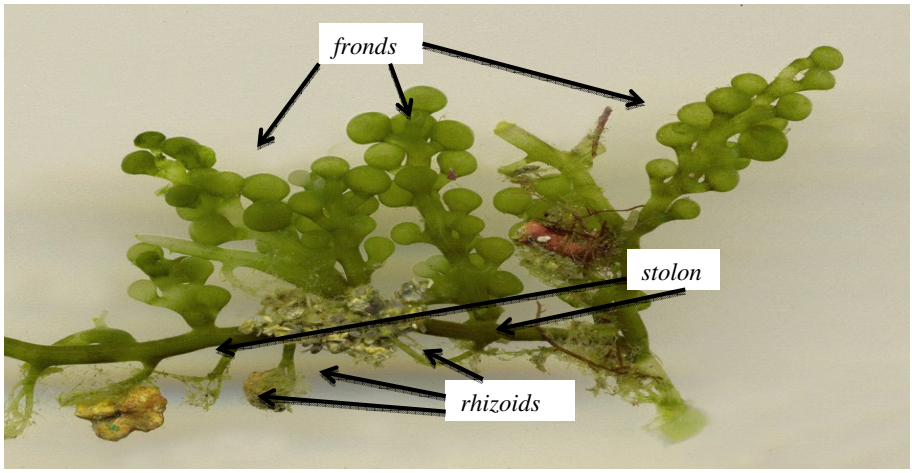


FIGURE 1. Photo of a specimen of the genus *Caulerpa*, with identification of its morphological structures (*fronds*, *stolon*, *rizoids*) as used in classical taxonomy.

similar to one another, and that morphological taxonomy still constitutes the basis for the biological classification in current use. However, the addition of molecular studies on *Caulerpa* has clarified some of the problems of morphological plasticity and of the phylogenetic consideration of certain species. Such morphological variation is reflected on the great confusion still existent in the identification of their varieties and forms [7, 46].

One solution to these obstacles resides in the alliance between molecular biology and morphological taxonomy. Therefore, the use of molecular tools should be associated with taxonomic analysis, whenever possible.

This work proposes the use of an alternative methodology, one that uses morphological data in the classification, such data having been discussed in the light of molecular studies for the genus *Caulerpa.*

4. Results and discussion

Following, we present a new computational experience in order to demonstrate the performance of HSCM (Hyperbolic Smoothing Clustering Method), in particular its capability to solve problems involving biometric data.

The data used were obtained from the morphological measures of three structures used in the traditional taxonomy of the macroalgae of the genus *Caulerpa* (*fronds, stolon* and *rhizoids*) (Fig. 1).

The measures of the variables were extracted from [7] and are presented in Table 1. The first column lists the algae under study. The second and third columns show the measures of the minimum and maximum heights of the *fronds,*

TABLE 1. Measures of the variables of *Caulerpa* species used in this work.

1st	2nd	3rd	4th	5th	6th	7th	8th	9th
1. *C. ashmeadii*	50	120	10	2.2	2	2.6	1.7	2
2. *C. brachypus*	5	22	2	4	0.55	1.04	355	500
3. *C. cupressoides*	20	11.5	1.7	9	1.1	3	0.2	2
4. *C. fastigiata*	0.3	12	0.2	1	0.09	0.3	50	660
5. *C. kempfii*	4.2	13	1	4.1	0.2	1	110	360
6. *C. lanuginosa*	50	160	4.8	9.4	0.4	6	0.58	1.47
7. *C mexicana*	6	75	4	16	0.4	2.2	0.1	1.02
8. *C. microphysa*	2.7	9.2	2.3	5.7	0.7	1	0.13	0.48
9. *C. murrayi*	1.4	2	1.51	2.31	0.1	0.29	0.01	0.02
10. *C. prolifera*	13	18	6	20	0.39	1.88	0.1	1
11. *C. pusilla*	1.6	8.4	0.8	2.75	0.07	0.42	0.05	0.17
12. *C. racemosa*	10	70	6	16	1.35	5.25	0.3	2.1
13. *C. scalpelliformis*	25	252	8	20	0.9	9.6	0.1	1.6
14. *C. serrulata*	5	23	2.1	3.1	1.6	2.9	1.3	2.2
15. *C. sertularioides*	12	92	3.5	14.2	0.4	3.3	0.1	2.1
16. *C. taxifolia*	65	80	3.3	4.6	1.3	1.5	0.5	0.7
17. *C. verticillata*	5.4	13.2	3.3	7.8	0.3	0.8	0.06	0.24
18. *C. webbiana*	3.5	11.3	1.2	1.5	0.12	0.2	0.01	0.36

respectively (2nd min., 3rd max.); Minimum and maximum widths of the *fronds* are shown on the fourth and fifth columns, respectively (4th min., 5th max.); The measures of the *stolon* diameter are shown on the sixth and seventh columns, respectively (6th min., 7th max.); and those of the *rhizoids* are shown on the eighth and ninth column, respectively (8th min., 9th max.). The experiments were executed in a notebook Intel Core i7- 2620M Windows with 2.70 GHz and 8 GB RAM.

With the purpose of comparing the taxonomic result obtained through the method with the results obtained through the use of molecular techniques for taxonomy of the genus *Caulerpa,* research was carried out through documental data such as primary sources, as well as bibliographic research involving works specialised in phylogeny of the genus under study.

The genus *Caulerpa* was recognised by Lamouroux in 1809. These algae are found in tropical marine environments and are traditionally recognised with basis on their morphological characteristics [12, 59, 63].

Caulerpas have both economic and ecological importance. Some species are used as food, *in natura* in salads or in preparation of other foods, and are cultivated in small scale. They produce substances used against high blood pressure and also as source of vitamins and mineral salts [62]. Biological properties like antiviral and anticoagulant activity have been reported [20, 57].

These algae have attracted much attention in the latest years due to their potential for substituting native vegetation, thereby altering the structure and function of marine landscape. Because of their great facility of self-adaptation to

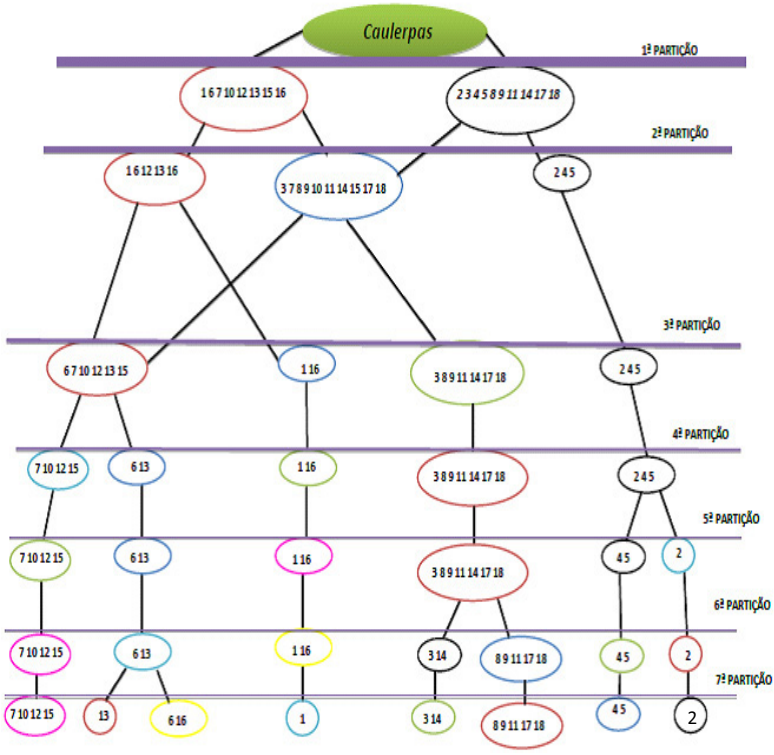


FIGURE 2. Diagram of the division of the groups with the use of HSCM.

environmental variations, besides their fast-pace growth and propagation, they are viewed by many researchers as being invasive algae that cause ecological imbalance, mainly on Asian and Mediterranean coastlines [10]. They are in fact a plague. Cases of bioinvasion have also been reported: they can dominate on a substratum and compete against the native organisms. They have a remarkable spreading capability, and possibly do not find natural predators, maybe because of their production of *caulerpenyne,* which has a toxic effect, or because of environmental conditions that limit the expansion of predators.

This genus presents considerable difficulty for taxonomic identification of species, due to phenotypical plasticity in diagnostic characters [48], which may explain the fact that, of 359 species (including forms and varieties) of the genus *Caulerpa*, only 85 are taxonomically valid [22]. Thus, some researchers have often been led to misdescribe different varieties and forms of one species [9, 13].

The application of the methodology presented herein enabled the formation of groups, from successive partitions realised with the use of algorithm HSCM. A diagram of such process can be seen in Figure 2.

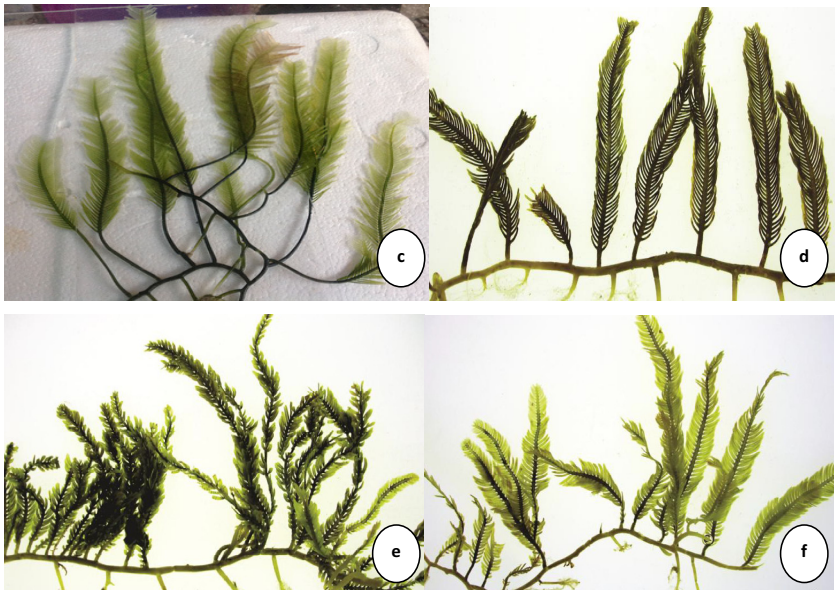


FIGURE 3. Photo of specimen of *C. ashmeadii* (c); *C. taxifolia* (d); *C. mexicana* (e) and *C. sertularioides* (f).

The results obtained with the application of the new methodology presented herein reaffirms that, in taxonomy, the physical reality is the form, the format, and the function of the organ (or of the organs). This is accepted in the taxonomists' activity, because it can be verified by anyone.

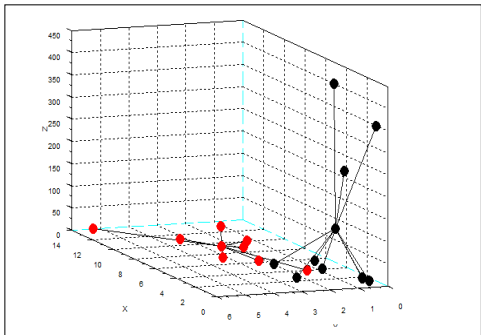
Studies by [50] reported that in a universe of 241 samples of the genus *Caulerpa*, 12,7% were morphologically classified wrongly by experienced phycologists. Species of *C. ashmeadii*, *C. taxifolia*, *C. mexicana,* and *C. sertularioides* are morphologically similar, and one can often be taken for another.

In our results, the algae *C. ashmeadii* and *C. taxifolia* are grouped in a same cluster (Graph 1). In the work by [7], these species have 100% sequence homology when compared through cpDNA *tuf* A analyses, a fact that demonstrates high phylogenetic affinity as well as results that are compatible with the method proposed.

Also, in our results, the groups obtained are similar to those of the studies by [36, 50], which shows that *C. taxifolia and C. mexicana* form clades that were placed separately in phylogenetic trees, that is, in different groups according to our method (Graph 2).

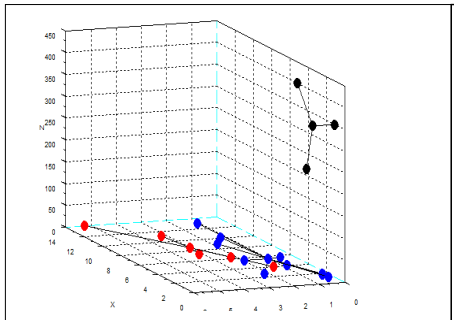
By observing species *C. kempfii, C. verticillata* and *C. pusilla,* it is possible to recognise similar morphological characters, such as the rather small delicate thallus with assimilating branches covered by verticils of ramuli dichotomically branched. According to [61], the main characteristic differing *C. verticillata* from *C. pusilla* is the absence of pilosity on the stolon of the former (Fig. 4).

FIGURE 4. Photo of specimen of *C. verticillata* (a) and *C. pusilla* (b).



1st PARTITION
GROUP 1 (red)
SPECIES 1 ***C. ashmeadii***
SPECIES 6 *C. lanuginosa*
SPECIES 7 *C. mexicana*
SPECIES 10 *C. prolifera*
SPECIES 12 *C. racemosa*
SPECIES 13 *C. scalpelliformis*
SPECIES 15 *C. sertularioides*
SPECIES 16 ***C. taxifolia***
GROUP 2 (black)
SPECIES 2 *C. brachypus*
SPECIES 3 *C. cupressoides*
SPECIES 4 *C. fastigiata*
SPECIES 5 *C. kempfii*
SPECIES 8 *C. microphysa*
SPECIES 9 *C. murrayi*
SPECIES 11 *C. pusilla*
SPECIES 14 *C. serrulata*
SPECIES 17 *C. verticillata*
SPECIES 18 *C. Webbiana*

Graph 1: Graphic representation of the clusters formed by the first partition using HSCM.

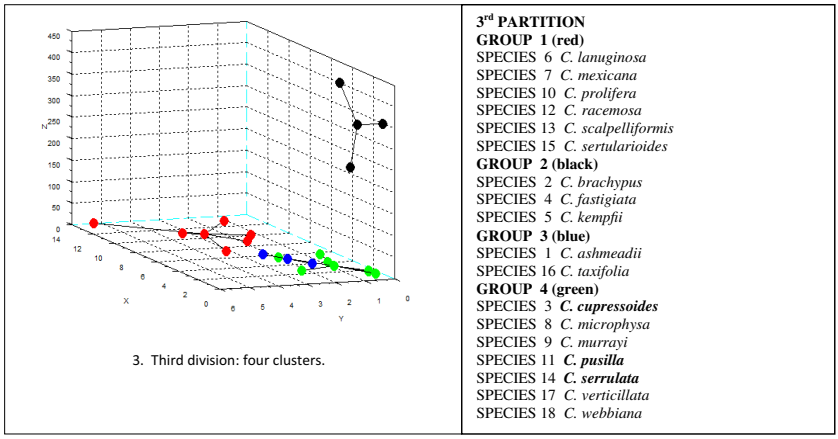


2nd PARTITION
GROUP 1 (red)
SPECIES 1 *C. ashmeadii*
SPECIES 6 *C. lanuginosa*
SPECIES 12 *C. racemosa*
SPECIES 13 *C. scalpelliformis*
SPECIES 16 ***C. taxifolia***
GROUP 2 (black)
SPECIES 2 *C.brachypus*
SPECIES 4 *C. fastigiata*
SPECIES 5 *C. kempfii*
GROUP 3 (blue)
SPECIES 3 *C. cupressoides*
SPECIES 7 ***C. mexicana***
SPECIES 8 *C. microphysa*
SPECIES 9 *C. murrayi*
SPECIES 10 *C. prolifera*
SPECIES 11 *C. pusilla*
SPECIES 14 *C. serrulata*
SPECIES 15 *C. sertularioides*
SPECIES 17 *C. verticillata*
SPECIES 18 *C. webbiana*

Graph 2: Graphic representation of the clusters formed by the second partition using HSCM.

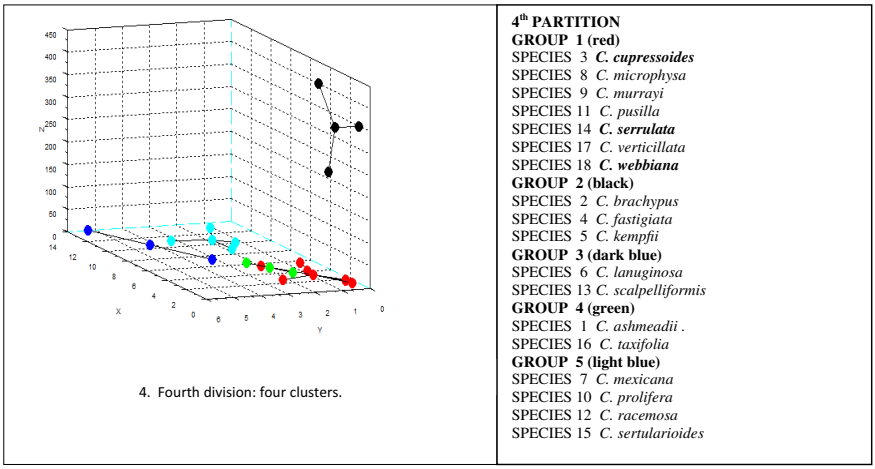
According to [7], the positioning of *C. pusilla* in the cladogram did not correspond to the one found with basis on its morphological characters, for *C. pusilla*

had been positioned along with *C. cupressoides* and *C. serrulata*, which are species that have a robust thallus reaching beyond 20 cm in length (Graph 3).



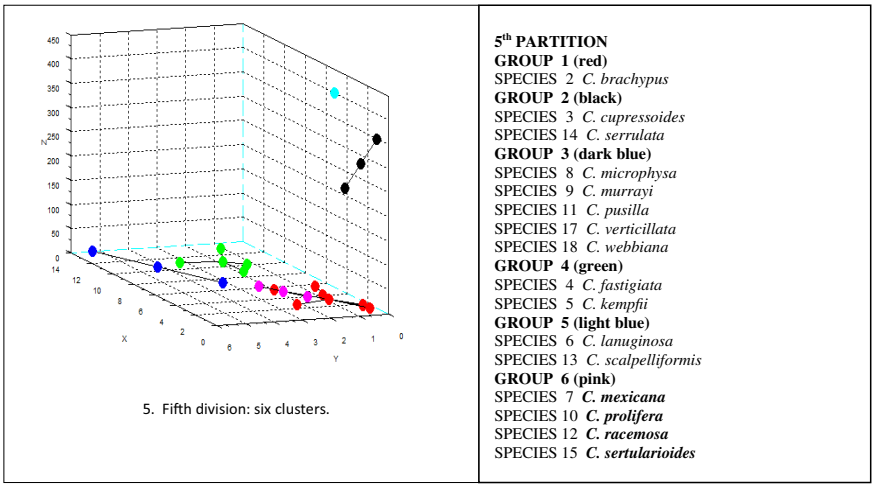
Graph 3: Graphic representation of the clusters formed by the third partition using HSCM.

This result is similar to the one found with the application of the method presented herein. It has been also verified that the results were consistent with the molecular studies by [68] on the *C. webbiana* phylogenetic pattern, in studies of phylogenetic trees with 90% and 100% of *bootstrap*, reinforcing that *C. webbiana* is closer to the group containing the varieties of *C. cupressoides and C. serrulata* (Graph 4).



Graph 4: Graphic representation of the clusters formed by the fourth partition using HSCM.

Even not presenting morphological similarities, *C. mexicana, C. prolifera, C. racemosa, and C. sertulararioides* (Graph 5) form a group in the fourth partition, similar to the one proposed by [31].



Graph 5: Graphic representation of the clusters formed by the fifth partition using HSCM.

The group formed by *C. cupressoides* and *C. serrulata* (Graph 6) is similar to the result obtained by [7] when comparing analyses of the tufA genetic sequences with those of cpDNA and obtained evidence of high phylogenetic affinity between them. Thus, no consistent pattern has been observed in the relationship between morphological characters and placement on the phylogenetic tree of taxonomic units based on molecular markers. Similarly, [36] observed in their study on *C. cupressoides* and *C. serrulata* that these are clearly different in morphological characteristics, but present paraphyletic lineage.

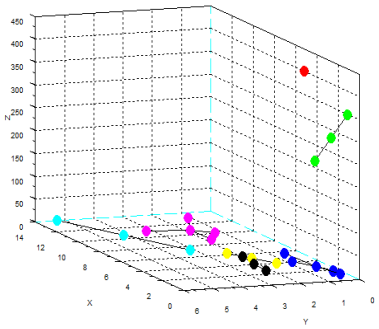
Also isolated is *C. brachypus* (Graph 7), demonstrating that the species, although similar to *C. prolifera,* is distinct from all other *Caulerpa* species under study. In consequence of phylogenetic studies reported by [64], this species was proposed to constitute a phylogenetically separate group, which corroborates the results found in our analysis.

According to [55], *C. scalpelliformis* is a species aside from the others, remaining in an isolated terminal clade, in agreement with our results.

5. Conclusion

The systems of classification, in their generality, are diverse, emphasise their authors personal points of view, and for such reasons are questionable and susceptible to corrections and modifications, exactly because they deal with living organisms which are subject to continuous alterations and the influence of the environment.

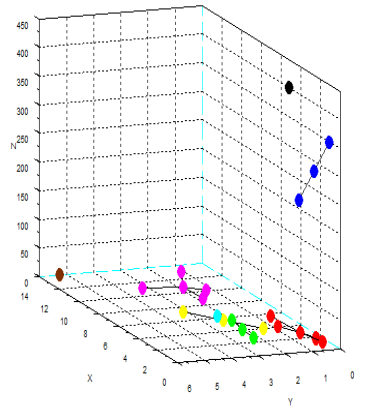
In Taxonomy, information systems based on morphological characters are still the most acknowledgeable, in spite of the constant development and improvement of the numerous systems of classification based on an ample range of morphological and non-morphological information.



6. Sixth division: seven clusters.

6th PARTITION
GROUP 1 (red)
SPECIES 2 *C. brachypus*
GROUP 2 (black)
SPECIES 3 ***C. cupressoides***
SPECIES 14 ***C. serrulata***
GROUP 3 (dark blue)
SPECIES 8 *C. microphysa*
SPECIES 9 *C. murrayi*
SPECIES 11 *C. pusilla*
SPECIES 17 *C. verticillata*
SPECIES 18 *C. webbiana*
GROUP 4 (green)
SPECIES 4 *C. fastigiata*
SPECIES 5 *C. kempfii*
GROUP 5 (light blue)
SPECIES 6 *C. lanuginosa*
SPECIES 13 *C. scalpelliformis*
GROUP 6 (pink)
SPECIES 7 *C. mexicana*
SPECIES 10 *C. prolifera*
SPECIES 12 *C. racemosa*
SPECIES 15 *C. sertularioides*
GROUP 7 (yellow)
SPECIES 1 *C. ashmeadii.*
SPECIES 16 *C. taxifolia*

Graph 6: Graphic representation of the clusters formed by the sixth partition using HSCM.



7. Seventh division: eight clusters.

7th PARTITION
GROUP 1 (red)
SPECIES 8 *C. microphysa*
SPECIES 9 *C. murrayi*
SPECIES 11 *C. pusilla*
SPECIES 17 *C. verticillata*
SPECIES 18 *C. webbiana*
GROUP 2 (black)
SPECIES 2 ***C. brachypus***
GROUP 3 (blue)
SPECIES 4 *C. fastigiata*
SPECIES 5 *C. kempfii*
GROUP 4 (green)
SPECIES 3 *C. cupressoides*
SPECIES 14 *C. serrulata*
GROUP 5 (light blue)
SPECIES 1 *C. ashmeadii*
GROUP 6 (pink)
SPECIES 7 *C. mexicana*
SPECIES 10 *C. prolifera*
SPECIES 12 *C. racemosa*
SPECIES 15 *C. sertularioides*
GROUP 7 (yellow)
SPECIES 6 *C. lanuginosa*
SPECIES 16 *C. taxifolia*
GROUP 8 (brown)
SPECIES 13 ***C. scalpelliformis***

Graph 7: Graphic representation of the clusters formed by the seventh partition using HSCM.

Therefore, the most adequate classification tool will depend on the characteristics the taxonomist has at hand, and on the best analytical treatment for those

characteristics. So, "each case is a case" and the user must always study his problem carefully, because with the emergence of new grouping methods it became ever more important to use judgment in order to choose a method that in fact solves the problem, or at least helps to meet the user's need for classification. A grouping method that satisfies the requisites of a group of users may not satisfy the requisites of another, since the appropriate grouping is in the specialist's perspective. Indeed, data grouping should involve the needs of the user or the application of such data.

This work major contribution was to present a new algorithm that may be used unequivocally in several areas, especially in taxonomy. This new view broadens the scope of utilisation of this tool, serving as inspiration for future works.

The algorithm proposed is functional, robust, and highly promising in its diversity of application.

References

[1] M. Ackerman and B.-D. Shai, A characterization of linkage-based hierarchical clustering. *J. Mach. Learn. Res.* **31** (2013) 66–74.

[2] R.M. Aliguliyev, Performance evaluation of density-based clustering methods. *Inform. Sci.* **179** (2009) 3583–3602.

[3] C.J. Allegre and H.S. Stephen, The evolution of the Earth. *Sci. Am.* **271** (1994) 44–51.

[4] Dalton de Souza. Amorim, Fundamentos de sistemática filogenética, in *Fundamentos de sistemática filogenética*. Holos (2002).

[5] A.M. Bagirov, Modified global k-means algorithm for minimum sum-of-squares clustering problems. *Pattern Recognition* **41** (2008) 3192–3199.

[6] A.M. Bagirov and J. Yearwood. A new nonsmooth optimization algorithm for minimum sum-of-squares clustering problems. *Eur. J. Oper. Res.* **170** (2006) 578–596.

[7] D. Barata, Taxonomia e filogenia do gênero *Caulerpa* J.V. Lamour (Bryopsidales, Chlorophyta). Brasil Tese, Instituto de Botânica, São Paulo (2008).

[8] D. Bhattacharya and L. Medlin, Algal phylogeny and the origin of land plants. *Plant Physiol.* **116** (1998) 9–15.

[9] S. Brayner, S.M. Barreto Pereira and M. Elizabeth Bandeira-Pedrosa. Taxonomia e distribuição do gênero *Caulerpa* Lamouroux. *Acta Botanica Brasilica* **22** (2008) 914–928.

[10] F. Bulleri, D. Balata, I. Bertocci, L. Tamburello and L. Benedetti-Cecchi, The seaweed *Caulerpa racemosa* on Mediterranean rocky reefs: from passenger to driver of ecological change. *Ecology* **91** (2010) 2205–2212.

[11] M. do Carmo Calijuri, A. Cordeiro Alves Dos Santos and M. Suely Adriani Alves. *Cianobactérias e cianotoxinas em águas continentais*. RiMa (2006).

[12] E. Coppejans and T. Beeckman, *Caulerpa* section Sedoideae (Chlorophyta, Caulerpales) from the Kenyan coast. *Nova Hedwigia* **49** (1989) 381–393.

[13] E.W. Coppejans and W.F. Prud'homme van Reine, Seaweeds of the Snellius-II Expedition (E. Indonesia): the genus *Caulerpa* (Chlorophyta-Caulerpales). *Buil. Séanc. Acad. r. Sei. Outre-Mer* **37** (1992) 667–712.

[14] Wei Ding, T.F. Stepinski, R. Parmar, D. Jiang and C.F. Eick, Discovery of feature-based hot spots using supervised clustering. *Comput. Geosci.* **35** (2009) 1508–1516.

[15] M. Ester, H.P. Kriegel, J. Sander and X. Xu, Clustering for mining in large spatial databases. *KI* **12** (1998) 18–24.

[16] D.E. Fairbrothers, T.J. Mabry, R.L. Scogin and B.L. Turner, The bases of angiosperm phylogeny: chemotaxonomy. *Ann. Missouri Bot. Garden* (1975) 765–800.

[17] D. Fasulo, *An analysis of recent work on clustering algorithms*. Department of Computer Science & Engineering, University of Washington (1999).

[18] R.A. Fisher, The use of multiple measurements in taxonomic problems. *Annals Eugenics* **7** (1936) 179–188.
[19] G. Garai and B.B. Chaudhuri, A novel genetic algorithm for automatic clustering. *Patt. Recog. Lett.* **25** (2004) 173–187.
[20] P. Ghosh, U. Adhikari, P.K. Ghosal, C.A. Pujol, M.J. Carlucci, E.B. Damonte and B. Ray, *In vitro* anti-herpetic activity of sulfated polysaccharide fractions from *Caulerpa racemosa*. *Phytochemistry* **65** (2004) 3151–3157.
[21] L.E. Graham and L.W. Wilcox, *Algae*. Prentice-Hall do Brasil, Rio de Janeiro (2000).
[22] M.D. Guiry and G.M. Guiry. AlgaeBase. *AlgaeBase* (2008).
[23] J. Han, M. Kamber and K.H. Tung, Spatial clustering methods in data mining: A survey, in *Geographic data mining and knowledge discovery*, edited by H.J. Miller and J. Han. Taylor and Francis (2001).
[24] Tsu-Ming Han and B. Runnegar, Megascopic eukaryotic algae from the 2.1-billion-year-old Negaunee Iron-Formation, Michigan. *Science* **257** (1992) 232–235.
[25] C.P. Hickman Jr., L.S. Roberts and A. Larson, *Princípios integrados de zoologia* (2004).
[26] D. Hill, *et al.* An algorithmic model for invasive species: Application to *Caulerpa taxifolia* (Vahl) C. Agardh development in the North-Western Mediterranean Sea. *Ecol. Model.* **109** (1998) 251–266.
[27] A.K. Jain, Data clustering: 50 years beyond K-means. *Patt. Recog. Lett.* **31** (2010) 651–666.
[28] A.K. Jain and R.C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc. (1988).
[29] A.K. Jain, M. Narasimha Murty and P.J. Flynn, Data clustering: a review. *ACM computing surveys (CSUR)* **31** (1999) 264–323.
[30] K. Jajuga, A. Sokolowski and H.H. Bock, *Classification, clustering, and data analysis: Recent advances and applications* (studies in classification, data analysis, and knowledge organization) (2002).
[31] O. Jousson, J. Pawlowski, L. Zaninetti, A. Meinesz and C.F. Boudouresque, Molecular evidence for the aquarium origin of the green alga *Caulerpa taxifolia* introduced to the Mediterranean Sea. *Mar. Ecol. Prog. Ser.* **172** (1998) 275–280.
[32] J.A. Kaandorp and J.E. Kübler, *The algorithmic beauty of seaweeds, sponges and corals*. Springer (2001).
[33] D.F. Kapraun, Nuclear DNA content estimates in multicellular green, red and brown algae: phylogenetic considerations. *Ann. Bot.* **95** (2005) 7–44.
[34] D. Karaboga and C. Ozturk. A novel clustering approach: Artificial Bee Colony (ABC) algorithm. *Appl. Soft Comput.* **11** (2011) 652–657.
[35] J.F. Kasting, Earth's early atmosphere. *Science* **259** (1993) 920–926.
[36] M.A. Kazi, C.R.K. Reddy and B. Jha. Molecular phylogeny and barcoding of *Caulerpa* (Bryopsidales) based on the tufA, rbcL, 18S rDNA and ITS rDNA Genes. *PloS One* **8** (2013) e82438.
[37] D.W. Lam and F.W. Zechman, Phylogenetic analyses of the Bryopsidales (Ulvophyceae, Chlorophyta) based on Rubisco large subunit gene sequences. *J. Phycol.* **42** (2006) 669–678.
[38] M. Laszlo and S. Mukherjee, A genetic algorithm that exchanges neighboring centers for means clustering. *Patt. Recog. Lett.* **28** (2007) 2359–2366.
[39] N. Lavesson, *Evaluation and analysis of supervised learning algorithms and classifiers*. Blekinge Institute of Technology (2006).
[40] G. Lawrence and M. Hill, *Taxonomia das plantas vasculares*. Fundação Calouste Gulbenkian (1973).
[41] Lee, Robert Edward. *Phycology*. Cambridge University Press (2008).
[42] P. Legendre and D.J. Rogers, Characters and clustering in taxonomy: a synthesis of two taximetric procedures. *Taxon* (1972) 567–606.
[43] P. Legendre and L.F.J. Legendre, *Numerical ecology*. Elsevier (2012).
[44] L.-J., Hermann and C. Weihs, *Classification as a Tool for Research: Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft Für Klassifikation EV, Dresden, March 13-18, 2009.* Springer (2010), Vol. 11.

[45] J. MacQueen, Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (1967), Vol. 1. No. 14.
[46] P. Madl and M. Yip, Literature review of *Caulerpa taxifolia*. HTTP:http://www.sbg.ac.at/ipk/avstudio/pierofun/ct/caulerpa.htm (accessed 12 February 2004) (2003).
[47] L. Margulis and Karlene V. Schwartz, *Cinco reinos: um guia ilustrado dos filos da vida na Terra*. Editora Guanabara Koogan (2001).
[48] I. Meusnier, M. Valero, J.L. Olsen, W.T. Stam, Analysis of rDNA ITS1 indels in *Caulerpa taxifolia* (Chlorophyta) supports a derived, incipient species status for the invasive strain. *Eur. J. Phycol.* **39** (2004) 83–92.
[49] M.C. Oliveira and D. Milstein, Taxonomia molecular. In *Macroalgas: uma introduçã o à axonomia, Technical Books Editora, Rio de Janeiro,* edited by A. de G. Pedrini (Og.) (2010) 71–82.
[50] J.L. Olsen, M. Valero, I. Meusnier, S. Boele-Bos and W.T. Stam, Mediterranean *Caulerpa taxifolia* and *C. mexicana* (Chlorophyta) are not conspecific. *J. Phycology* **34** (1998) 850–856.
[51] N. Papavero, *Fundamentos práticos de taxonomia zoológica*. Unesp (1994).
[52] H.-S. Park and C.-H. Jun, A simple and fast algorithm for K-medoids clustering. *Exp. Syst. Appl.* **36** (2009) 3336–3341.
[53] O.O. Parra and C.E. Bicudo, *Introducción a la biología y sistemática de las algas de aguas continentales*. Universidad de Concepción (1996).
[54] A. de G. Pedrini, Macroalgas; uma introdução à taxonomia. Rio de Janeiro: Technical Books (2010).
[55] A. Pillmann, G.W. Woolcott, J.L. Olsen, *et al.* Inter-and intraspecific genetic variation in *Caulerpa* (Chlorophyta) based on nuclear rDNA ITS sequences. *Eur. J. Phycol.* **32** (1997) 379–386.
[56] B. de Reviers, *Biologia e filogenia das algas*. Artmed (2006).
[57] J.A.G. Rodrigues, *et al.* Polissacarídeos sulfatados isolados das clorofíceas Caulerpa racemosa e *Caulerpa cupressoides*-extração, fracionamento e atividade anticoagulante. *Acta Sci. Biol. Sci.* **32** (2010) 113–120.
[58] J.W. Schopf, Microfossils of the Early Archean Apex chert: new evidence of the antiquity of life. *Science* **260** (1993) 640–646.
[59] W.R. Taylor, *Marine algae of the eastern tropical and subtropical coasts of the Americas* (1960).
[60] Teixeira, V.L. Taxonomia química. In *Macroalgas: uma introdução à taxonomia,* Technical Books Editora, Rio de Janeiro. edited by A. de G. Pedrini (2010) 83–97.
[61] B.N. Torrano-Silva, C.E. Amancio and E.C. Oliveira, Algas de aquarios ornamentales en Brasil: previsión de las introducciones. *Latin Amer. J. Aquat. Res.* **41** (2013) 344–350.
[62] Jr. Trono and C. Gavino, Diversity of the seaweed flora of the Philippines and its utilization. *Hydrobiologia* **398** (1999) 1–6.
[63] A. Weber-van Bosse, *Monographie des Caulerpes* (1898).
[64] M.J. Wynne, V. Heroen and L.A. Dror, The recognition of *Caulerpa integerrima* (Zanardini) comb. et stat. nov. (Bryopsidales, Chlorophyta) from the Red Sea. *Phycologia* **48** (2009) 291–301.
[65] A.E. Xavier and V.L. Xavier, Solving the minimum sum-of-squares clustering problem by hyperbolic smoothing and partition into boundary and gravitational regions. *Pattern Recognition* **44** (2011) 70–77.
[66] Xavier, Adilson Elias. The hyperbolic smoothing clustering method. *Pattern Recognition* **43** (2010) 731–737.
[67] Xu, Rui and D. Wunsch, Survey of clustering algorithms. *Neural Netw. IEEE Trans.* **16** (2005) 645–678.
[68] Wen-Ji Yeh and G.-Y. Chen, Nuclear rDNA and internal transcribed spacer sequences clarify *Caulerpa racemosa* vars. from other *Caulerpa* species. *Aq. Bot.* **80** (2004) 193–207.
[69] R.S.M. Zadegan, M. Mehdi and F. Sadoughi, Ranked medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets. *Knowledge-Based Systems* **39** (2013) 133–143.