
# USO DO ALGORITMO DE SUAVIZAÇÃO HIPERBÓLICA EM TAXONOMIA DE MACROALGAS

Maria Gardênia Sousa Batista

Tese de Doutorado apresentada ao Programa de Pós-graduação em Engenharia de Sistemas e Computação, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Doutora em Engenharia de Sistemas e Computação.

Orientadores: Adilson Elias Xavier

Francisca Lúcia de Lima

Rio de Janeiro

Novembro de 2014

USO DO ALGORITMO DE SUAVIZAÇÃO HIPERBÓLICA EM TAXONOMIA DE MACROALGAS

Maria Gardênia Sousa Batista

TESE SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE DOUTORA EM CIÊNCIAS EM ENGENHARIA DE SISTEMAS E COMPUTAÇÃO.

Examinada por:

_____
Prof. Adilson Elias Xavier, D.Sc.

_____
Profa. Francisca Lúcia de Lima, Ph.D.

_____
Prof. Nelson Maculan Filho, D.Habil.

_____
Prof. André Macêdo Santana, D.Sc.

_____
Profa. Ariadne do Nascimento Moura, Ph.D.

_____
Profa. Márcia Helena Costa Fampa, Ph.D.

RIO DE JANEIRO/RJ – BRASIL

NOVEMBRO DE 2014

Batista, Maria Gardênia Sousa

Uso Do Algoritmo de Suavização Hiperbólica em Taxonomia de Macroalgas/ Maria Gardênia Sousa Batista. – Rio de Janeiro: UFRJ/COPPE, 2014.

XIII, 59 p.: il.; 29,7cm.

Orientador: Adilson Elias Xavier e Francisca Lúcia de Lima

Tese (doutorado) – UFRJ/COPPE/ Programa de Engenharia de Sistemas e Computação, 2014.

Referências Bibliográficas: p. 50 – 59.

1. Suavização Hiperbólica. 2. Taxonomia. 3. Algas I. Xavier, Adilson Elias *et al*. II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia de Sistemas e Computação. III. Título.

"*As fadigas que isto me causou e os esforços que me custaram, só Deus sabe. Quantas vezes desanimei e quantas voltei atrás e tornei a começar pelo desejo de saber; sei-o eu que passei por isso, e sabem-no também os que viviam na minha companhia. Agora dou graças ao Senhor, pois que colho os saborosos frutos das raízes amargas dos estudos*"

(SÃO JERÔNIMO)

*"Se, por acaso, espinhos ou sombras turvarem-te o caminho, exalta, porque somente os eleitos, são convidados ao testemunho, apenas os fortes são testados nos valores e, unicamente quem produz periodicamente passa pela avaliação que procede às promoções."*

(JOANNA DE ANGELIS)

*Dedico a ti meu amado Pai, Cícero Romão Batista (in memorian), este trabalho, meu eterno amor, gratidão e a certeza que está comigo sempre me dando forças.*

*A minha amada Mãe, Maria José, Souza Batista razão da minha existência e do meu sucesso.*

*E as minhas Avós Maria da Conceição e Benedita e suas irmãs Maria Rodrigues e Maria Eugracia exemplos de Mulheres Guerreiras.*

# Agradecimentos

Muito obrigada meu Deus, Pai de Infinita Bondade e Misericórdia!

"*Um só coração um só pensamento subirão até vós, como um grito de reconhecimento e de amor*".

Obrigada minha Senhora e também Minha Mãe, Maria Santíssima! *"Infinitas graças vos damos Soberana Rainha"*

Ao Mar onde eu encontro minha Paz, meu mundo, meu objeto de estudo! E que nos ensina que *quando somos apanhados na rebentação é preciso levantar-nos de imediato.*

A minha família que por ser extremamente grande prefiro não nomeá-los. Apenas agradecer por tudo!

As amigas e amigos, obrigada mesmo os que estiveram ausentes!

Porém, desejo especial agradecimentos a minha Mãe, Maria José, meu irmão Francisco Moisés & Nílvia seus filhos Jade e Júnior, obrigada pelo companheirismo e ajuda necessária a concretização de meus ideais.

Tias Duk e Bolota e Vó Conceição a vocês, muito obrigada.

Aos meus orientadores Adilson Elias Xavier e Francisca Lúcia de Lima pela oportunidade de realizar este sonho de concluir o doutorado. Acreditando e depositando em mim sua confiança.

Ao Dr. Raimundo Castro que prontamente auxiliou intermediando a participação do Dr. André Macedo, e a Florindo por esta iniciativa.

Ao apoio familiar de todos no Rio de Janeiro da família de Tia Carmem, Tio André, Tia Helena e Tio Joelton (*in memorian*) e seus familiares. Minha família carioca!

E a Alvarenga, Dona Zuza, Sibele e família; obrigada pelo apoio no Rio de Janeiro.

A Sociedade Brasileira de Ficologia – SBFic, aos meus amigos Ficólogos especialmente a Edisa Nascimento (USP), Sônia Barreto (UFRPE), Ariadne Moura (UFRPE), Diógina Barata (CEUNES), Pedrini (UERJ) entre outros mais meu obrigada pelo incentivo e apoio.

A Universidade Estadual do Piauí – UESPI, ao Centro de Ciências Biológicas – CCN.

Universidade Federal do Rio de Janeiro- UFRJ, Instituto Alberto Luís Coimbra de Pós - Graduação e Pesquisa de Engenharia – COPPE, ao Programa de Engenharia de Sistemas e Computação – PESC, a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES.

Aos meus estimados alunos obrigada pela confiança especialmente os que passaram pelo Laboratório de Ficologia e Limnologia – LABFIL/UESPI.

Ao meu querido Lucena e a eterna pergunta "*quando é que isso termina?"*.

A todos que em algum momento vibraram positivamente para que eu chegasse ao final do meu curso com êxito.

"*I don't need easy... I just need possible*" (Soul Surfer)

E se tivermos fé tudo é possível!

***Paz e Luz!***

Resumo da Tese apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Doutora em Ciências (Dra.Sc.)

# USO DO ALGORITMO DE SUAVIZAÇÃO HIPERBÓLICA EM TAXONOMIA DE MACROALGAS

Maria Gardênia Sousa Batista

Novembro/2014

Orientadores: Adilson Elias Xavier

Francisca Lúcia de Lima

Programa: Engenharia de Sistemas e Computação

Esta tese apresenta uma nova proposta metodológica para agrupamento de dados em Taxonomia. Macroalgas do gênero *Caulerpa* foram escolhidas como modelo de estudo para sua aplicação por apresentarem grande plasticidade morfológica e dificuldade em sua identificação por métodos sistemáticos tradicionais. Os resultados obtidos utilizando o algoritmo de suavização hiperbólica demonstram sua viabilidade de uso em taxonomia biológica. Podendo essa nova metodologia ser utilizada de forma isolada ou em associação a outras metodologias já consolidadas, não apenas em Ficologia, mas também em outras áreas da Biologia.

Abstract of Thesis presented to COPPE/UFRJ as a partial fulillment of the requirements for the degree of Doctor of Science (D.Sc.)

THE USE OF THE HYPERBOLIC SMOOTHING ALGORITHM IN TAXONOMY OF MACROALGAE

Maria Gardênia Sousa Batista

November/2014

Advisors: Adilson Elias Xavier

Francisca Lúcia de Lima

Department: Systems Engineering and Computer Science

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.

# Sumário

# Lista de Figuras

# Lista de Tabelas

# Lista de Gráficos

# Capítulo 1

# Introdução

O homem modifica constantemente seus critérios de classificação. No que se refere à classificação dos seres vivos, os critérios são modificados de acordo com o tipo de relação que o homem estabelece com a natureza.

O desenvolvimento de metodologias inovadoras vem ao encontro do atendimento das necessidades do taxonomista com vistas a resolver problemas na classificação, com o propósito de agrupar, tendo por base aspectos de semelhança entre os elementos classificados.

A Ciência da classificação ou sistemática é acima de tudo um desafio, uma incógnita, pois tenta dividir o indivisível, ou seja, a natureza de modo a torná-la inteligível. No propósito de contribuir modestamente para o avanço na sistemática é o que ora se apresenta uma nova metodologia para agrupamento de dados biológicos, através do uso de um algoritmo novo e mais eficiente. Nossa perspectiva é de poder ajudar ao sistemata a conhecer e entender o fechamento do ciclo, ou seja, daquele laborioso processo de identificação das espécies que produz uma grande quantidade de informações que servirá para uma interpretação adequada do ecossistema e consequentemente de novas abordagens cientificas.

Portanto, dentre as inúmeras áreas de interesse para pesquisa de doutorado, a otimização da sistemática biológica é uma área que permite o desenvolvimento de

trabalhos que possam, efetivamente, trazer benefícios significativos na preservação da biodiversidade.

A taxonomia é uma ciência intrinsecamente inter-relacionada com todos os ramos das ciências, igualmente importantes. Assim contribuir na construção de ferramentas, métodos e metodologias para análise de dados biológicos é que estão sendo concentrados os maiores esforços dos pesquisadores em biologia computacional. Pois é na análise que se encontram as grandes carências em soluções e é nesta etapa que é gerado o conhecimento necessário para pôr em prática tudo o que vem sendo prometido, como o desenvolvimento de fármacos e transgênicos, terapia gênica, dentre outros.

Esta tese visa acrescentar ao universo de pesquisa uma nova opção a ser explorada pelos taxonomistas, servindo como ferramenta durante a complexa tarefa de identificação vegetal.

Pode-se atribuir resumidamente, a motivação da pesquisa aqui apresentada pelas seguintes características do objeto de estudo:

- Complexidade do processo de identificação e classificação da diversidade biológica, e a necessidade de otimizar o trabalho do taxonomista;
- Necessidade de metodologias mais eficientes e práticas, proporcionando novas interpretações dos dados taxonômicos frente a novas abordagens de estudo, de modo a propiciar novas pesquisas;
- Aduzir o conceito de distância entre os elementos da classificação, a fim de permitir um novo desenho na representação dos caracteres taxonômicos definidores da classificação;

## 1.1 Objetivos e Contribuição

Buscou-se destacar como objetivo principal desse trabalho a apresentação de uma proposta metodológica heterodoxa dentro dos preceitos consuetudinários da taxonomia. Tem-se a veleidade, que não ostentamos uma proposta que sanou as fragilidades da metodologia hierárquica, mas que entende-se que seja uma ferramenta que pode oferecer uma alternativa para enriquecimento na interpretação taxonômica, podendo ser utilizada a partir da taxonomia morfológica.

As informações para aplicação do método foram extraídas da morfologia de três estruturas de macroalgas do gênero *Caulerpa*: ramos assimiladores (*frondes*), estolão (*estolão*) e rizóides (*rizóides*). Obtidas no trabalho sobre Taxonomia e Filogenia do Gênero *Caulerpa* J.V. Lamour. (Bryopsidales, Chlorophyta) no Brasil, de autoria de BARATA (2008).

Os dados da biometria das estruturas das algas, utilizados na aplicação do método desenvolvido, serviram para caracterização das espécies, e foram utilizados como vetores de características, obtidos da medida de altura mínima e máxima dos ramos assimiladores; do diâmetro mínimo e máximo do estolão e dos rizoides das algas.

Buscou-se propor esse novo paradigma na taxonomia com base na consciência de que para elaborar uma abordagem criativa, destarte, seria um estímulo para visão dinâmica, flexível das classificações vigentes na biologia, especialmente das algas.

Em combinação ao uso do algoritmo do tipo partição, HSCM, que adota um enfoque de suavização hiperbólica. Mostrar-se-á um resultado diferenciado quanto à representação taxonômica, pois se entende que a compreensão da taxonomia está ligada intimamente a sua forma de representação.

Cabe ressaltar que qualquer pessoa perceptiva notará que há necessidade de um entendimento na compreensão da taxonomia ligada à forma de representação visual.

Portanto, em contraposição as metodologias hierárquicas que ensejam a mínima possibilidade de oferecer esta visão analítica de que o dendograma é limitado por ser planar, enquanto a realidade taxonômica é multidimensional, nesse propósito, suscitou-se a produção de desenhos não lineares, de acordo com os designíos específicos da aplicação do especialista.

No presente trabalho utilizaremos outro algoritmo, que é novidade na literatura, porém tem se mostrado com excelente desempenho, o Hyperbolic Smoothing Clustering Method (HSCM), XAVIER (2010). Será exposto o uso deste novo algoritmo para classificação de macroalgas, tem como base dados biométricos do gênero *Caulerpa*. Os clusters formados com a utilização desta nova metodologia serão analisados e discutidos a partir de estudos de filogenia destas algas.

# Capítulo 2

# Revisão Bibliográfica

Os seres humanos sempre tentaram classificar os objetos animados e inanimados que os cercam. Classificar objetos em categorias coletivas é um pré-requisito para nomeá-los. Agrupar é reconhecer que os objetos são suficientemente semelhantes para serem colocados no mesmo grupo e também para identificar distinções ou separações entre os grupos (LEGENDRE & LEGENDRE, 2012).

Taxonomia ou Sistemática é a ciência que trata da identificação, nomenclatura e classificação de objetos de natureza biológica (LAWRENCE, 1973).

A classificação e a identificação dos organismos foram reconhecidas desde a Grécia Antiga, sendo considerada uma das primeiras atividades do homem. Surgindo como ciência com Aristóteles e evoluído bastante com *Linnaeus*. Após Charles Darwin, a biologia estabeleceu perspectivas modernas, através da demonstração da história evolutiva – filogenia – dos organismos como objetivo maior da sistemática. O aumento no conhecimento de dados sistemáticos, especialmente do DNA e a incorporação de novas metodologias alavancaram novas formas de analisar os dados e praticar a sistemática. É interessante perceber que uma ciência tão antiga como a sistemática,

apresente um conhecimento sempre provisório, evoluindo constantemente para refletir as novas descobertas. O surgimento a passos largos do conhecimento de novas hipóteses filogenéticas remete-se ao desenvolvimento de novas metodologias e ferramentas para inferir relações filogenéticas, bem como a disponibilidade de novas formas de evidência, oportunizando uma descrição cada vez mais acurada da história evolutiva.

Em virtude da megabiodiversidade e suas interações, os estudos de sistemática baseiam-se em dados provenientes de todos os ramos das ciências biológica, a fim de sintetizar e enquadrar hierarquicamente as diversas formas de vida. A atual situação do estado de conservação da diversidade genética que é em si a base de toda nossa megabiodiversidade, face à crescente destruição dos *habitats* e ameaças de extinção, torna urgente que se estabeleça outro nível de informação em análise de dados, para maximização dos conhecimentos adquiridos em estudos de sistemática. Tornando imprescindível uma interface da atuação do sistemata com pesquisas em outras áreas, como no incremento do uso de dados, em utilização de algorítmos de agrupamento.

Em Biologia, existe uma grande necessidade de métodos eficientes para organizar a biodiversidade e posicionar taxonomicamente as espécies em grupos ou "*clusters*" que consigam reter informações similares segundo as características avaliadas.

## 1.1 Evolução da Sistemática

O aumento no conhecimento de dados sistemáticos, especialmente do DNA e a incorporação de novas metodologias alavancaram novas formas de analisar os dados e praticar a sistemática. O surgimento de novas hipóteses filogenéticas nos remete ao desenvolvimento de novas metodologias e ferramentas para inferir relações filogenéticas, bem como a disponibilidade de novas formas de evidência, oportunizando

uma descrição cada vez mais acurada da história evolutiva (AMORIM, 2002; HICKMAN *et al.*,2004).

Uma das mais básicas e essenciais habilidades das criaturas vivas é o agrupamento de objetos similares produzindo uma classificação. É esta habilidade que permitirá, por exemplo, a uma criatura viva determinar se algo é nocivo ou não à sua existência No processo de classificação são considerados objetos e atributos. Objeto é tudo aquilo que se quer classificar e atributos são as informações a respeito do objeto que serão consideradas no processo de classificação.

Ao longo da história, o homem aprendeu que a prática de classificar seres e objetos facilita a manipulação e a compreensão das entidades classificadas, além de permitir que seu estudo seja compartilhado entre pessoas, constituindo um eficiente método de comunicação (DING *et al.*, 2009; ALIGULIYEV, 2009).

Aristóteles, filósofo grego, considerado "o pai da biologia", foi o primeiro a classificar os organismos tendo como base suas similaridades estruturais. Um sistema de classificação mais abrangente foi introduzido por John Ray (1627 a 1705), naturalista inglês, seguidor da renascença europeia. Somente no Século XVIII, a sistemática culminou com o trabalho de Carolus Linnaeus (1707 a 1778), através da publicação *Systema Naturae*. Utilizando-se da morfologia para organizar espécimes em coleções, Linnaeus, teve um *insight* que possibilitou um avanço muito grande, propôs a classificação binomial e foi considerado "*pai da sistemática*" (HICKMAN *et al.*,2004; CALIJURI *et al.*,2006).

O termo *sistema*, no final do século XVIII, era utilizado para denominar classificações baseadas em um único caráter, hoje, o objetivo central da Sistemática, além da descrição da diversidade é elaborar um sistema geral de referência, e contribuir para a compreensão dessa diversidade, através do estudo das relações de parentesco

entre as espécies. As classificações resultantes devem refletir a história filogenética e, assim, possibilitar a previsão das características dos organismos atuais, além de recuperar as informações indexadas (AMORIM, 2002).

A Sistemática é a parte da Biologia que trata do estudo dos seres vivos, classificando-os em grupos ordenados (os táxons ou categorias hierárquicas), e estabelecendo um sistema natural de classificação. O reconhecimento destes grupos está baseado na clareza, notoriedade e utilidade para o observador. O uso das classificações como auxiliar da memória era fundamental antes do advento da informática e em período em que os livros não eram tão comuns. Atualmente, a sistemática ocupa uma posição central na biologia evolutiva e está desempenhando uma função cuja importância tem aumentado gradativamente para outras disciplinas como a informática.

Os sistemas de classificação podem dividir-se em artificiais, naturais e filogenéticos. Os sistemas artificiais consideram os caracteres independentemente de sua origem e sem se preocupar com as possíveis afinidades e parentescos entre os indivíduos classificados. Visando a praticidade, através da identificação rápida e inequívoca, na atualidade, representam a maior parte dos sistemas de classificação. Os sistemas naturais, de posse de todas as informações disponíveis, sobre as espécies (morfologia, fisiologia, bioquímica, genética, citologia, ultraestrutura...), propiciando uma base mais ampla de conhecimento, não considera também as relações de parentesco (JUDD *et. al.*, 2009).

A difusão rápida e a pronta aceitação das teorias de Darwin (1809-1882) demonstraram a insatisfação que existia entre os pesquisadores a respeito dos diversos sistemas de classificação até então existentes. O sistema filogenético surgiu então, baseado na variabilidade das espécies. Este cuida de suas relações genéticas, levando em consideração os atuais seres, como aqueles de outras eras geológicas. Em síntese, o

sistema filogenético se firma na teoria evolutiva, classificando os organismos com base nas modificações de seus caracteres, eles registram o grau e a quantidade de diversidade e complexidade, sem violar a ordem de ramificação assim como a cronologia de suas ramificações (BICUDO & MENEZES, 2006).

A incorporação da teoria Darwiniana da evolução alterou radicalmente o propósito dos sistemas de classificação. Os autores dos sistemas naturais procuravam obter grupos morfologicamente consistentes; nos sistemas de classificação evolutivos (sistemas filogenéticos) passou a ser prioritário que os *taxa* refletissem relações de parentesco (relações filogenéticas), através da proximidade evolutiva.

O método cladístico foi originalmente proposta pelo entomólogo alemão Willi Hennig, em 1950, trabalha com a inferência filogenética, é um método desenvolvido para gerar hipóteses sobre as relações de parentesco entre organismos ou grupos de organismos. Baseia-se num pressuposto fundamental: os grupos de organismos têm de reunir todos, e apenas, os descendentes de um ancestral comum.

O desenvolvimento da Sistemática Filogenética ou Cladismo, demonstrando que a história evolutiva de ancestralidade-descendência dos organismos pode ser reconstruída e representada mediante um diagrama hipotético denominado *cladograma*. Nesta representação o compartilhamento de características evolutivas (sinapomorfias), permite relacionar o parentesco, entre a diversidade biológica, expressa na ramificação de uma linhagem para outra no decorrer da evolução. Assim, somente os agrupamentos de organismos cuja realidade histórica apresente pelo menos um caráter no estado derivado (*grupos monofiléticos*) podem ser utilizados na classificação (JUDD *et al*., 2009).

Os cladogramas são diagramas que expressam graficamente uma hipótese sobre as relações de parentesco de um dado conjunto de organismos ou grupos de organismos. Os cladograma são, portanto, um resumo da sua história evolutiva.

Nos cladogramas cada ramo apenas se pode cindir noutros dois ramos (ramificação dicotómicas); os nós (pontos onde ocorrem ramificações) e as extremidades dos ramos representam, respectivamente, eventos de divergência evolutiva e um grupo monofilético de organismos. Dois clados situados lado a lado num cladograma dizem-se irmãos (grupos irmãos, ing. *sister groups*). Um cladograma pode ser cortado em qualquer ponto, o ramo resultante – o clado – inclui necessariamente todos os descendentes de um dado ancestral (SANTOS, 2008).

Como referido anteriormente, os cladogramas são o produto final de uma análise cladística. Quanto maior o número de caracteres e de estados de carácter envolvidos numa análise, maior o número de soluções possíveis. Por outro lado, diferentes algoritmos produzem diferentes resultados, o mesmo acontecendo quando se eliminam ou adicionam caracteres na matriz original. Determinar qual o melhor cladograma, isto é, qual o cladograma que melhor retrata a história evolutiva de um dado grupo de plantas, é uma questão chave em cladística. Dois princípios são usados para a resolver: o princípio da congruência e o princípio da parcimónia. O princípio da congruência baseia-se numa ideia simples: se o mesmo resultado – o mesmo cladograma – é obtido com dois ou mais conjuntos de caracteres, então a probabilidade da filogenia obtida ser verdadeira cresce. O cladograma que minimiza o número de transições entre estados de carácter é o mais parcimonioso. O princípio da parcimónia é crítico porque sendo um princípio filosófico (epistemológico), produzido pela mente humana, nada obriga a que seja seguido nos processos evolucionários. Por outras palavras, a natureza não é necessariamente parcimoniosa, embora tendencialmente o

seja. As homologias e, implicitamente, as analogias, são determinadas *a posteriori* pela análise da partilha de caracteres ao longo do "melhor" cladograma.(SCHUH, 2000; AMORIM, 2002).

Os princípios da congruência e da parcimónia conjugam-se na chamada reamostragem por *bootstrap* (ou em métodos similares, e.g. *jackknife*). Este processo inicia-se com a construção de peudoreplicações (cladogramas parciais) a partir de uma amostra (parcial) aleatória de caracteres da matriz original de caracteres (mantendo a dimensão da matriz original). Em cada pseudoreplicação é selecionado o cladograma mais parcimonioso. Depois de repetir o processo um determinado número de vezes (e.g. mil repetições) o resultado é sumarizado num cladograma de consenso (árvore de consenso) sendo possível aferir a incerteza associada a cada clado. Uma percentagem de *bootstrap* de 95% significa que o clado em causa surgiu em 95 de 100 pseudoreplicações (KITCHING *et al*., 1998).

Embora a importância dos dados moleculares em cladística seja inquestionável, a morfologia externa, permanece essencial, talvez ainda mais importante do que no passado, no esclarecimento das afinidades evolutivas. A escassez de caracteres morfológicos e a abundância de convergências evolutivas que os caracteriza é mais do que compensada pela informação filogenética útil que transportam. Ao contrário do que ocorre com muitos caracteres moleculares, os caracteres morfológicos são funcionalmente relevantes tendo, por essa razão, sido moldados pela seleção natural.

Os sistemas de classificação cladísticos apresentam várias vantagens frente aos sistemas de classificação evolutivos tradicionais:

a) Robustez – à medida que as relações filogenéticas são clarificadas a circunscrição e a nomenclatura dos *taxa* tende a estabilizar;

b) Reprodutibilidade – diferentes investigadores obtêm os mesmos resultados se utilizarem os mesmos dados iniciais;

c) Objetividade – envolvem menos assunções intuitivas.

As classificações cladísticas têm, porém, uma enorme desvantagem prática. Ao produzirem a melhor estimativa das relações evolutivas podem dar origem a grupos morfologicamente inconsistentes, pouco intuitivos, que dificultam a sua apreensão pelos não especialistas.

O procedimento de análise dos dados em matrizes de caracteres, gerando cladogramas, utilizando-se recursos computacionais, é a base para o desenvolvimento de métodos numéricos de análise filogenética. Esse método é executado, utilizando-se de um táxon terminal inicial e adicionando novos táxons, um de cada vez, considerando o conjunto dos caracteres de cada táxon. Observa-se ai um padrão meramente dicotômico, nem sempre, viável para refletir o grau de conhecimento atual das relações de parentesco. Cladogramas geralmente colocam dúvidas, mas do que as respostas que apresentam. Segundo AMORIM (2002), nenhum sistema de classificação que se fundamenta em um conhecimento que evolui gradualmente é estável, assim, pode-se adotar uma das filogenias propostas por determinado autor ou pode-se gerar uma filogenia e consequente classificação.

A atual situação do estado de conservação da diversidade genética que é em si a base de toda nossa megabiodiversidade, face à crescente destruição dos habitats e ameaças de extinção, torna urgente que se estabeleça outro nível de informação em analise de dados, para maximização dos conhecimentos adquiridos em estudos de sistemática. Tornando imprescindível a integração crescente em interfaces da atuação do sistemata com pesquisas em outras áreas, como no incremento do uso de dados, em utilização de algoritmos de agrupamento. A necessidade de buscar por uma forma

sistemática para encontrar grupos em dados levou ao desenvolvimento de técnicas para resolver este problema e são conhecidas como “agrupamento de dados” (KAUFMAN & ROUSSEEUW, 1990; GORDON 1998; EVERITT *et al.* 2001), “taxonomia numérica” ou, ainda, “classificação automática de dados”.

## 2.2 Discussão Sucinta dos Métodos de Agrupamento

Desde o trabalho pioneiro de FISHER (1936) no uso de métodos de agrupamento baseados em métrica, utilizando-se espécies de *Iris* (*Iris versicolour, Iris setosa, Iris virginica*). Os relatos do uso de algoritmos de agrupamento com dados biológicos têm-se intensificado muito ao longo dos tempos.

Trabalhos como de LEGENDRE & ROGERS (1972), KAANDORP & KUBLER (2001), que fazem relato da eficiência de métodos computacionais em sistemática biológica, têm sido comuns especialmente para estudos de filogenia (KAPRAUN, 2005; LAM & ZECHMAN, 2006), conservação e biodiversidade (HILL *et al*., 1998).

Diversas técnicas de agrupamento são descritas na literatura (JAIN & DUBES, 1988; JAIN *et al*., 1999; LASZLO & MUKHERJEE, 2007), levando o pesquisador a ter sapiência de escolher o mais adequado ao seu propósito, uma vez que as diferentes técnicas podem levar a diferentes soluções (LAVESSON, 2006).

Um dos algoritmos mais utilizado é o *k-means*, embora tenha sido proposto a mais de 50 anos, ainda é tradicionalmente utilizado como uma ferramenta rápida de fácil entendimento e implementação (MACQUENN, 1967; JAIN, 2010).

O problema de agrupamento possui aplicações nas mais variadas áreas de pesquisa incluindo, por exemplo: computação visual e gráfica, computação médica,

biologia computacional, redes de comunicações, engenharia de transportes, redes de computadores, sistemas de manufatura, entre outras (JAJUGA *et al.*, 2002; XU & WUNS, 2005).

De uma forma geral, consiste em agrupar os elementos (objetos) de uma base de dados (conjunto) de tal forma que os grupos formados, ou *clusters*, representem uma configuração em que cada elemento possua uma maior similaridade com qualquer elemento do mesmo *cluster* do que com elementos de outros *clusters*. Tem por finalidade reunir, por algum critério de classificação, as unidades amostrais em grupos, baseado na similaridade (BAGIROV & YEARWOOD, 2006; BAGIROV, 2008) de tal forma que exista homogeneidade dentro do grupo e heterogeneidade entre grupos (HAN & KAMBER, 2001; PARK & JUN, 2009; XAVIER & XAVIER, 2011).

Estes métodos utilizam diversos algoritmos (GARAI & CHAUDHURI, 2004; PARK & JUN, 2009) entre eles os algoritmos hierárquicos e de partição (FASULO, 1999; XAVIER, 2010; KARABOGA & OZTURK, 2011; ACKERMAN & BEN-DAVID, 2013).

Nos algoritmos tradicionais para agrupamento hierárquico as formações dos *clusters* ocorrem de forma gradativa através de aglomerações ou divisões de elementos/*clusters*, gerando uma hierarquia de *clusters*, normalmente representada através de uma estrutura em árvore ou dendograma (ESTER *et al*., 1998).

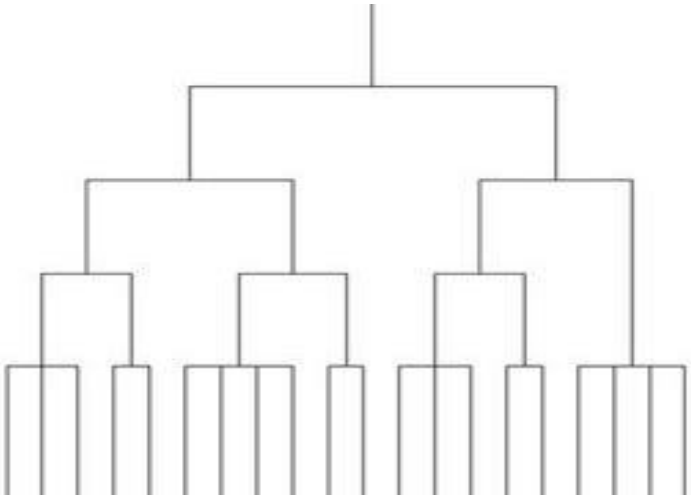


**Figura 2.2:** Exemplo de dendograma.

Pode-se assim caracterizar os métodos hierárquicos, como frágeis, pois gera árvore dicotômica e assim expressa uma decisão irreversível em torno dos grupos formados. Favorecendo uma análise filogenética meramente monofilética. Além disso, em cladogramas, a representação da ausência de conhecimento filogenético, é feita através de uma politomia, favorecendo uma multiplicidade nos resultados a serem formados em decorrência da fragilidade intrínseca do método. Segundo AMORIM (2002), *cladogramas* "são inferências (hipóteses) permanentemente sujeitas à transformação".

Métodos de partição: No método de partição o conjunto de dados é dividido em *q* grupos, os quais juntos satisfazem os requisitos de uma partição:

Cada grupo deve conter no mínimo um objeto;

Cada objeto deve pertencer exatamente a um grupo.

Essas condições demonstram existir no máximo tantos grupos quanto são os objetos. A segunda condição designa que dois grupos diferentes não podem possuir objetos em comum e que os *q* grupos juntos devem conter todos os objetos (Fig. 2.3).

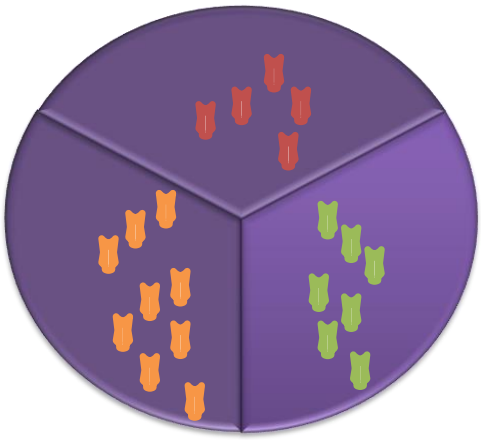


**Figura 2.3:** Exemplo de partição.

Um método de partição produz uma única partição do conjunto de dados sem nenhuma estrutura hierárquica, tal como em um dendograma produzido por uma técnica hierárquica.

Os métodos de particionamento tem o conjunto de elementos dividido em *k* subconjuntos, podendo *k* ser conhecido ou não, e cada configuração obtida é avaliada através de uma função-objetivo. Caso a avaliação da clusterização indique que a configuração não atende ao problema em questão, nova configuração é obtida através da migração de elementos entre os *clusters*, e o processo continua de forma iterativa até que algum critério de parada seja alcançado. Assim, os *clusters* podem ser melhorados gradativamente, o que não ocorre nos métodos hierárquicos (ZADEGAN *et al.*, 2013).

O problema de cluster é de interesse em diversas áreas que necessitam agrupar dados, assim, nos últimos anos, tem-se intensificado o número de trabalhos publicados utilizando-se da análise de agrupamentos. A importância, do estudo dos métodos de agrupamento, na comunidade científica, pode ser observada pela publicação da temática

em diversos jornais de extrema importância tais como, são apresentados por XU & WUNSCH (2005).

A aplicação dos métodos de análise de *cluster* tem produzido uma dinâmica de melhoramento e produção de novos métodos BAGIROV & YEARWOOD (2006), BAGIROV (2008), XAVIER (2010), a fim de viabilizar sua utilização e desempenho em diversas áreas.

# Capítulo 3

# As Algas

O termo *alga*, usado a partir de 1753 e introduzido por Linaeus (1707-1778), é aplicado a uma variedade tão grande de organismos que hoje não se pode atribuir um significado taxonômico e não corresponde, portanto a nenhuma categoria nomenclatural. Sendo simplesmente um termo coletivo para um grupo de plantas extremamente heterogêneo e que torna difícil e também problemática sua definição. É difícil especificar quais são suas características positivas comuns a todos os indivíduos, usualmente conhecidas como *algas* (REVIERS, 2006; LEE, 2008).

Segundo BICUDO & MENEZES (2006), "alga" é um termo de uso popular, como palmeira ou grama, utilizado para designar um verdadeiro universo de organismos tão diferentes quanto sua morfologia, reprodução, fisiologia e ecologia, o que se torna praticamente impossível sua definição.

Embora não se possa descrever a organização da vida primitiva com absoluta certeza, os registros fósseis indicam firmemente que "cianobactérias" viveram há mais de três bilhões de anos. Isso não permite afirmar categoricamente que as algas foram os seres vivos mais antigos, pois os registros fósseis são sempre incompletos, mas há indícios de que as algas, juntamente com as bactérias e certos fungos, são organismos extremamente antigos, os quais, devido ao processo da fotossíntese, são responsáveis

pela estruturação da atmosfera terrestre como se conhece, possibilitando a vida sobre a superfície do planeta (BOLD, 1972; HAN & RUNNEGAR, 1992; SCHOPF, 1993; KASTING, 1993; ALLÈGRE & SCHEINEIDER, 1994; KRISHNER, 1994; DUVE, 1996).

As algas por não constituírem uma categoria taxonômica definida, mas sim um grupo de categorias díspares, tão diversas que se enquadram em três reinos diferentes (Monera, Protistas e *Plantae*), ou ainda com diferentes denominações em outros sistemas apoiados em dados de biologia molecular (SOGIN *et.al.*, 1989; BHATTACHARYA & MEDLIN, 1998). A classificação das algas é extremamente complexa e em plena evolução. (DAWES, 1997; REVIERS, 2003, 2006; CALIJURI *et. al.* 2006; LEE, 2008; GRAHAM, 2009).

## 3.1 O Problema de Classificação das Algas

São conhecidas várias formas de classificar os organismos, merecendo destaque, na classificação das algas: taxonomia morfológica, taxonomia molecular e taxonomia química ou quimiotaxonomia e a ultraestrutura (TEIXEIRA, 2010; OLIVEIRA & MILSTEIN, 2010).

- A taxonomia morfológica é a classificação baseada em critérios tradicionais da morfologia, via observação das características externas, é tradicionalmente a forma mais utilizada para classificar um organismo;
- A taxonomia molecular utiliza dados moleculares para os estudos taxonômicos a partir das sequências de DNA ou proteínas

- Taxonomia química ou quimiotaxonomia é a classificação baseada nos constituintes químicos dos organismos, ou seja, na produção dos produtos naturais (metabólitos) desses organismos.

Não cabe aqui, discutir os prós e os contras dos diversos sistemas de classificação das algas, mas enfatizar a importância do uso de uma nova metodologia que irá contribuir para a interpretação das informações obtidas no processo de identificação das medidas a fim de que se possa precisamente avaliar os limites métricos da população em estudo, permitindo observar simultaneamente, múltiplos aspectos na formação dos agrupamentos.

## 3.2 O Gênero *Caulerpa* J.V. Lamouroux (1809)

O nome do gênero *Caulerpa* vem do grego e significa *caulus* "tronco" e *erpos* "crescer ao longo do solo" (BARATA, 2008). As espécies incluídas no gênero *Caulerpa* são caracterizados por serem macroalgas marinhas, ocorrendo no médio e infralitoral, geralmente em regiões tropicais e subtropicais, possuem uma estrutura semelhante a folhas para cima (*frondes e pínulas*), sustentada por um estolão cilíndrico, horizontal, que origina os rizóides (Figura 3.2) (DAWES, 1997). Seus ramos eretos podem atingem algumas dezenas de centímetros em altura, mas cada organismo é constituído por uma única célula multinucleada, apresentando o nível de organização do talo conhecido como cenocítico. (TAYLOR, 1960).

Em se tratando de algas há uma enorme variedade de dificuldades de classificação que poder-se-ia ter escolhido para trabalhar, no entanto foi escolhida o gênero *Caulerpa* J.V. LAMOUROUX (1809), as evidências experimentais demonstram que a morfologia dos ramos eretos do gênero *Caulerpa* apresenta grande plasticidade,

podendo variar dentro da mesma espécie dependendo das condições ambientais (GRAHAM, 2009). Este fato levou alguns pesquisadores a descrever diferentes variedades e forma de uma espécie, muitas vezes de maneira equivocada (BELTON *et al*, 2014). Segundo TAYLOR (1960) *"esta famosa espécie tropical é ainda a mais variável em seu gênero variável"*. BRAYNER *et al.*, (2008), destaca que em virtude da representatividade do gênero, há necessidade de estudos para obtenção de maiores informações taxonômicas e ecológicas sobre as *Caulerpa*.

Os problemas de identificação são devidos a grande plasticidade do talo, possuindo assim uma controvérsia quanto ao número de espécies, podendo variar em torno de 350 espécies, das quais 85 são consideradas válidas (GUIRY & GUIRY, 2013). Para o Brasil, segundo BARATA (2008), são citadas 19 espécies de *Caulerpa*, e mais 26 variedades e formas num total de 45 táxons infraespecíficos conhecidos.


**Figura 3.2**: Aspectos gerais das *caulerpas*.

As espécies de *Caulerpa* são comuns em águas rasas e águas profundas dos mares tropicais e subtropicais (UKABI *et al*, 2012). *Caulerpa* são comumente encontradas na costa brasileira (RODRIGUES *et al*, 2010).

A grande variedade das morfologias em *Caulerpa*, segundo a literatura, tem sido provada como sendo muito influenciadas pelas alterações do *hábitat*,

principalmente pelo tipo de substrato, exposições às ondas, correntes, profundidade, intensidade de luz, estação do ano e pressão de predação. Sendo assim, a plasticidade fenotípica, juntamente com o tipo de propagação clonal, fazem as espécies neste gênero apresentarem grande flexibilidade de resposta às mudanças ambientais rápidas, pois não há necessidade de adaptação. Essa variação morfológica é refletida na grande confusão que ainda existe na identificação de suas variedades e formas (VERLAQUE *et. al*., 2003; MADL & YIP, 2003).

O gênero *Caulerpa* é importante econômica e ecologicamente, pois algumas espécies são utilizadas na alimentação em saladas, *in natura* ou ainda na preparação de alimentos. Sendo cultivadas em pequena escala. Estas algas também produzem substâncias utilizadas no tratamento de pressão alta e, também, como fontes de vitaminas e sais minerais (GHOSH *et al*.,2004). São relatadas propriedades biológicas, tais como antiviral e anticoagulante (RODRIGUES & FARIAS, 2005).

*Caulerpa* tem atraído muita atenção nos últimos anos devido o seu potencial de substituir vegetação nativa, alterando assim a estrutura e função da paisagem marinha. Estas algas apresentam grandes facilidades de adaptação às variações ambientais e crescimento e propagação rápida, tida por muitos pesquisadores como sendo algas invasoras, provocando desequilíbrio ecológico, principalmente na costa da Ásia e Mediterrâneo (BULLERI *et al.* 2010). Também são relatados eventos de bioinvasão, podem dominar sobre um substrato e competir com os organismos nativos. Elas apresentam uma grande capacidade de expansão, e possivelmente, não encontram predadores naturais, talvez pela produção de um composto químico denominado de *caulerpenina*, que produz ação tóxica, ou as condições ambientais que limitem a sua expansão (TEIXEIRA, 1991; VALENTIM & SOUZA, 2011; SOUZA, 2013).

# Capítulo 4

# O Algoritmo de Suavização Hiperbólica

## 4.1 O Problema de Agrupamento como um Problema *min-sum-min*

Seja $S = \{s_1, \ldots, s_m\}$ um conjunto de $m$ padrões ou observações pertencentes a um espaço Euclideano de dimensão $n$ que deve ser dividido em número pré-fixado $q$ de grupos disjuntos.

Para formular o problema de agrupamento original como um problema *min-sum-min*, procede-se como segue. Sejam $x_i, i = 1, \ldots, q,$ os centroides dos grupos, onde cada $x_i \in \Re^n$. O conjunto das coordenadas destes centroides será representado por $X \in \Re^{nq}$. Dado um ponto $s_j$ de $S$, inicialmente é calculada a distância deste ponto ao centroide em $X$ mais próximo. Isto é expresso por

$$z_j = \min_{x_i \in X} \left\| s_j - x_i \right\|_2. \tag{4.1.1}$$

Uma medida de qualidade do agrupamento associado a uma posição específica de $q$ centroides é fornecida pela soma dos quadrados destas distâncias.

$$D(X) = \sum_{j=1}^{m} z_j^2 \tag{4.1.2}$$

A localização ótima dos centroides deve fornecer o melhor valor desta medida de qualidade. Então, se $X^*$ denota uma localização ótima, o problema é

$$X^* = \arg\min_{X \in \Re^{nq}} D(X), \tag{4.1.3}$$

onde $X$ é o conjunto de todas as localizações dos $q$ centroides. Usando (4.1.1)-(4.1.3), finalmente obtêm-se

$$X^* = \arg\min_{X \in \Re^{nq}} \sum_{j=1}^{m} \min_{x_i \in X} \left\| s_j - x_i \right\|_2^2. \tag{4.1.4}$$

## 4.2 Transformação do Problema

O problema anterior (4.1.4) pode ser substituído pelo seguinte problema equivalente

$$\begin{aligned} &\textit{minimizar} \ \sum_{j=1}^{m} z_j^2 \\ &\textit{sujeito a}: \ z_j = \min_{i=1,\ldots,q} \left\| s_j - x_i \right\|_2, \quad j = 1,\ldots,m. \end{aligned} \tag{4.2.1}$$

Considerando (4.1.1), $z_j$ deve necessariamente satisfazer o seguinte conjunto de desigualdades:

$$z_j - \left\| s_j - x_i \right\|_2 \le 0, \quad i = 1,\ldots,q. \tag{4.2.2}$$

Substituindo as igualdades do problema (4.2.1) pelas desigualdades (4.2.2), obtêm-se o problema relaxado

$$\begin{aligned} &\textit{minimizar} \ \sum_{j=1}^{m} z_j^2 \\ &\textit{sujeito a}: \ z_j - \left\| s_j - x_i \right\|_2 \le 0, \quad j = 1,\ldots,m, \ i = 1,\ldots,q \,. \end{aligned} \tag{4.2.3}$$

Desde que as variáveis $z_j$ não são limitadas inferiormente, é fácil ver que a solução do problema relaxado será $z_j = 0,\ j = 1,\ldots,m$. Portanto, o problema (4.2.3) não é equivalente ao problema (4.2.1). Para obter a equivalência desejada, deve-se, portanto, modificar o problema (4.2.3).

Nesse ponto é introduzida a função auxiliar

$$\varphi(y) = max\{0, y\}\,. \tag{4.2.4}$$

É fácil observar que, se as desigualdades (4.2.2) são válidas, necessariamente deve ser observada a restrição:

$$\sum_{i=1}^{q} \varphi(z_j - \left\| s_j - x_i \right\|_2) = 0, \quad j = 1,\ldots,m\,. \tag{4.2.5}$$

A figura 4.1 ilustra o gráfico das três primeiras parcelas componentes das equações (4.2.5) como função de $z_j$, onde $d_i = \left\| s_j - x_i \right\|_2$ e também é suposto que as distâncias $d_i$ estejam ordenadas em ordem crescente segundo os índices.

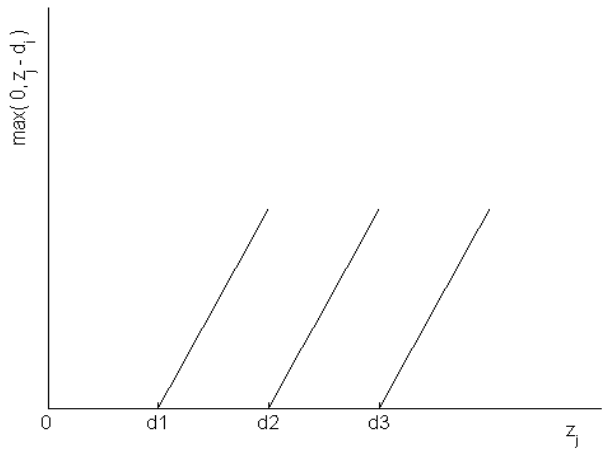



**Figura 4.2:** Três primeiras parcelas componentes das equações (4.2.5).

Com a substituição das desigualdades em (4.2.3) pelas equações (4.2.5), seria obtido um problema equivalente mantendo, portanto, a propriedade indesejável que $z_j,\ j=1,\ldots,m$ ainda são livres inferiormente. Entretanto, considerando que a função objetivo do problema (4.2.3) forçará os valores $z_j,\ j=1,\ldots,m$, a assumir os menores valores, pode-se pensar em limitar inferiormente essas variáveis ao considerar ">" no lugar de "=" em (4.2.5), resultando no seguinte problema "não canônico".

$$\begin{aligned} &\textit{minimizar}\ \sum_{j=1}^{m} z_j^2 \\ &\textit{sujeito a}:\ \sum_{i=1}^{q} \varphi(z_j - \left\| s_j - x_i \right\|_2) > 0, \quad j=1,\ldots,m\,. \end{aligned} \tag{4.2.6}$$

Para recuperar a formulação canônica, as desigualdades em (4.2.6) são perturbadas, obtendo-se o problema modificado:

$$\begin{aligned} &\textit{minimizar}\ \sum_{j=1}^{m} z_j^2 \\ &\textit{sujeito a}:\ \sum_{i=1}^{q} \varphi(z_j - \left\| s_j - x_i \right\|_2) \geq \varepsilon, \quad j=1,\ldots,m \end{aligned} \tag{4.2.7}$$

para $\varepsilon > 0$. Desde que o conjunto viável do problema (4.2.6) é o limite do conjunto viável do problema (4.2.7) quando $\varepsilon \to 0_+$, pode-se, então, pensar em resolver (4.2.6) através da resolução de uma sequência de problemas iguais a (4.2.7) para uma sequência de valores decrescentes de $\varepsilon$ que se aproximam de $0$.

Como mostrado a seguir, o problema (4.2.6) definido numa região aberta, ou alternativamente o problema (4.2.7) no limite quando $\varepsilon \to 0_+$, possui uma importante propriedade.

**Teorema**: O valor da solução ótima do problema (4.2.1) está arbitrariamente próximo do valor da solução do problema (4.2.6).

**Prova**: Seja $x^*, z_j^*, j = 1,\ldots,m$ a solução ótima do problema (4.2.1). Verifica-se que esse ponto está arbitrariamente próximo da região viável do problema (4.2.6).

Seja $I_j^* = \left\{ i \mid z_j^* = \left\| s_j - x_i^* \right\|_2 \right\}$. O cumprimento das igualdades de (4.2.1) implica que $I_j^* \neq \varnothing$ para $j = 1,\ldots,m$. Considerando as desigualdades de (4.2.6), para todo $j = 1,\ldots,m$, para as componentes do somatório associadas a $i \notin I_j^*$ tem-se trivialmente:

$$\varphi(z_j^* - \left\| s_j - x_i^* \right\|_2) = 0, i \notin I_j^*. \tag{4.2.8}$$

Todavia, as componentes do somatório associadas a $i \in I_j^*$, as funções

$$\varphi(z_j^* - \left\| s_j - x_i^* \right\|_2), i \in I_j^*, \tag{4.2.9}$$

são avaliadas exatamente no seu ponto de descontinuidade de sua derivada. Destarte, pela mais completa continuidade das funções intervenientes, existe no espaço viável do problema (4.2.6) um ponto arbitrariamente próximo do ponto ótimo do problema (4.2.1).

Portanto, o ponto da solução ótima de (4.2.1) está arbitrariamente próximo da região viável do problema (4.2.6). Deve ser observado que esses problemas possuem a mesma função objetivo. Logo, os valores das soluções ótimas desses problemas estão arbitrariamente próximos entre si.

## 4.3 Suavização do Problema

Analisando o problema (4.2.7), a definição da função $\varphi$ impõe a ele uma estrutura não diferençável muito rígida, que o torna sem qualquer utilidade prática. Em vista disso, o método numérico de resolução do problema (4.2.7), adotado no presente

trabalho, se fundamenta na ideia de suavização do mesmo. Dentro dessa perspectiva, define-se a função

$$\phi(y,\tau) = \left(y + \sqrt{y^2 + \tau^2}\right)/2 \qquad (4.3.1)$$

para $y \in \Re$ e $\tau > 0$.

A função $\phi$ possui, trivialmente, as seguintes propriedades:

(a) $\phi(y,\tau) > \varphi(y), \quad \forall \tau > 0$;

(b) $\lim_{\tau \to 0} \phi(y,\tau) = \varphi(y)$;

(c) $\phi(.,\tau)$ é uma função convexa crescente que pertence à classe de funções $C^{\infty}$.

Então, a função $\phi$ se constitui em uma aproximação da função $\varphi$ definida pela equação (4.2.4). Adotando-se as mesmas convenções especificadas na apresentação da Figura 4.1, as três primeiras parcelas componentes de (4.2.5) e a correspondente suavização, dada por (4.3.1), são mostradas lado a lado na Figura 4.2.

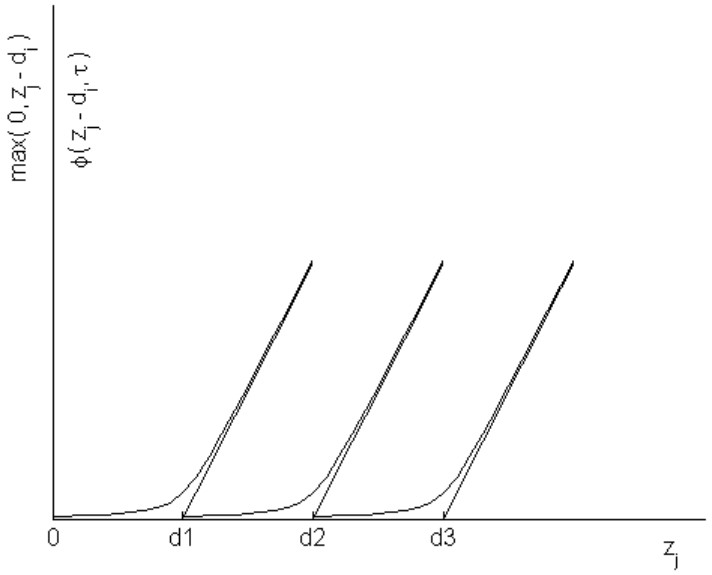



**Gráfico 4.3:** Gráfico das funções $\varphi$ e $\phi$

Ao substituir a função $\varphi$ pela função $\phi$ no problema (4.2.7), é obtido o seguinte problema

$$\begin{aligned} &\textit{minimizar} \ \sum_{j=1}^{m} z_j^2 \\ &\textit{sujeito a}: \ \sum_{i=1}^{q} \phi(z_j - \left\| s_j - x_i \right\|_2, \tau) \geq \varepsilon, \quad j = 1, \ldots, m . \end{aligned} \tag{4.3.2}$$

No sentido de se obter uma formulação completamente diferenciável, faz-se ainda necessária a suavização das distâncias Euclideanas $\left\| s_j - x_i \right\|_2$ do problema anterior. Com esse objetivo, define-se a função

$$\theta(s_j, x_i, \gamma) = \sqrt{\sum_{l=1}^{n} (s_{jl} - x_{il})^2 + \gamma^2} \ . \tag{4.3.3}$$

para $\gamma > 0$.

A função $\theta$ possui, trivialmente, as seguintes propriedades:

(a) $\lim_{\gamma \to 0} \theta(s_j, x_i, \gamma) = \left\| s_j - x_i \right\|_2$;

(b) $\theta$ é uma função que pertence à classe de funções $C^{\infty}$.

Ao substituir as distâncias Euclideanas $\left\| s_j - x_i \right\|_2$ do problema (4.3.2), é obtido, agora, o problema completamente diferenciável

$$\begin{aligned} &\textit{minimizar} \ \sum_{j=1}^{m} z_j^2 \\ &\textit{sujeito a}: \ \sum_{i=1}^{q} \phi(z_j - \theta(s_j, x_i, \gamma), \tau) \geq \varepsilon, \quad j = 1, \ldots, m . \end{aligned} \tag{4.3.4}$$

As propriedades das funções $\phi$ e $\theta$ permitem buscar uma solução para o problema (4.2.6) através da resolução de uma sequência de subproblemas da forma (4.3.4), produzida pela redução dos parâmetros $\gamma \to 0$, $\tau \to 0$ e $\varepsilon \to 0$.

## 4.4 Resolução do Problema

Analisando-se as condições Karush-Kuhn-Tucker (KKT) para o problema (4.3.4), será mostrado abaixo que todas as suas desigualdades serão certamente ativas.

Seja o problema geral de programação não linear sujeito a restrições de desigualdade:

$$\begin{aligned} &\textit{minimizar} \quad f(x) \\ &\textit{sujeito a}: \; g_j(x) \geq 0, \quad j = 1,\ldots,m\,. \end{aligned} \tag{4.4.1}$$

As condições KKT para o problema (4.4.1) são:

$$\nabla f(x) - \sum_{j=1}^{m} \mu_j \nabla g_j(x) = 0\,, \tag{4.4.2}$$

$$g_j(x) \geq 0\,,\; j = 1,\ldots,m\,, \tag{4.4.3}$$

$$\mu_j g_j(x) = 0\,,\; j = 1,\ldots,m\,, \tag{4.4.4}$$

$$\mu_j \geq 0\,,\; j = 1,\ldots,m\,. \tag{4.4.5}$$

Aplicando-se as equações (4.4.2) para o caso específico do problema (4.3.4), $f(x)$ e $g_j(x),\; j = 1,\ldots,m$, serão substituídas, respectivamente, por $f(x,z) = \sum_{j=1}^{m} z_j^2$

(4.4.6) e $g_j(x,z_j) = \sum_{i=1}^{q} \phi(z_j - \theta(s_j, x_i, \gamma), \tau) - \varepsilon,\; j = 1,\ldots,m\,.$ (4.4.7)

Os gradientes, em relação a *x* e *z*, para as funções (4.4.6) e (4.4.7) são, respectivamente, iguais a $\nabla f(x,z) = \left(0, 2z_1, \ldots, 2z_m\right)^T$ (4.4.8)

e $\nabla g_j(x,z_j) = \left( \nabla_x g_j(x,z_j), 0, \ldots, \frac{\partial g_j(x,z_j)}{\partial z_j}, \ldots, 0 \right)^T,\; j = 1,\ldots,m\,,$ (4.4.9)

onde $\nabla_x g_j(x,z_j),\; j = 1,\ldots,m$ representa as componentes do gradiente em relação ao vetor *x* dos centroides.

Agora substituindo-se (4.4.8) e (4.4.9) nas equações (4.4.2), tem-se que as últimas *m* equações serão:

$$2z_j - \mu_j \frac{\partial g_j(x, z_j)}{\partial z_j} = 0,\ j = 1, \ldots, m. \tag{4.4.10}$$

Observando que $z_j > 0,\ j = 1, \ldots, m$, devido às desigualdades de (4.3.4), para cumprimento das igualdades acima, deve-se necessariamente ter

$$\mu_j \frac{\partial g_j(x, z_j)}{\partial z_j} > 0,\ j = 1, \ldots, m. \tag{4.4.11}$$

Finalmente, considerando as condições de complementariedade (4.4.4), pode-se concluir que todas as *m* desigualdades de (4.3.4) serão ativas porque $\mu_j > 0,\ j = 1, \ldots, m$.

Então, o problema (4.3.4) será equivalente ao problema:

$$\begin{aligned} &\textit{minimizar} \ \sum_{j=1}^{m} z_j^2 \\ &\textit{sujeito a}: \ h(x, z_j) = \sum_{i=1}^{q} \phi(z_j - \theta(s_j, x_i, \gamma), \tau) - \varepsilon = 0, \quad j = 1, \ldots, m. \end{aligned} \tag{4.4.12}$$

O domínio das variáveis do problema (4.4.12) é definido num espaço com $(nq + m)$ dimensões. Como, em geral, o valor do parâmetro *m*, a cardinalidade do conjunto *S* das observações, é muito grande, o problema (4.4.12) possui um número muito grande de variáveis. Contudo, o mesmo possui uma estrutura separável, e assim reúne todas as condições desejáveis para a aplicação do Teorema da Função Implícita. Todas as funções desse problema pertencem à classe de funções $C^{\infty}$ em relação às variáveis $(x, z)$. Cada variável $z_j$ aparece somente em uma restrição de igualdade e a derivada parcial de $h(x, z_j)$ em relação a $z_j,\ j = 1, \ldots, m$ é diferente de zero. Portanto, é possível usar o Teorema da Função Implícita para calcular cada $z_j,\ j = 1, \ldots, m$, como

uma função das variáveis dos centroides $x_i, i = 1, \ldots, q$. Deste modo, obtêm-se o problema irrestrito

$$minimizar \quad f(x) = \sum_{j=1}^{m} z_j(x)^2 \tag{4.4.13}$$

onde cada $z_j$ é determinado através do cálculo da única raíz de cada equação

$$h(x, z_j) = \sum_{i=1}^{q} \phi(z_j - \theta(s_j, x_i, \gamma), \tau) - \varepsilon = 0, \quad j = 1, \ldots, m, \tag{4.4.14}$$

sendo a unicidade da raiz decorrente da propriedade da função $\phi$ acima ser estritamente crescente com $z_j$.

Novamente, devido ao Teorema da Função Implícita, as funções $z_j(x)$ possuem todas as derivadas em relação às variáveis $x_i, i = 1, \ldots, q$. Então, é possível calcular com facilidade o gradiente da função objetivo do problema (4.4.13)

$$\nabla f(x) = \sum_{j=1}^{m} 2\, z_j(x)\, \nabla z_j(x) \tag{4.4.15}$$

onde

$$\nabla z_j(x) = -\nabla h(x, z_j) / \frac{\partial h(x, z_j)}{\partial z_j}. \tag{4.4.16}$$

A abordagem acima não é nada mais que a ideia básica usada por ABADIE e CARPENTIER (1969) para o desenvolvimento do algoritmo do gradiente reduzido generalizado, com o objetivo de resolver o problema geral de programação não linear sujeito a restrições de igualdade.

Deste modo, é muito fácil resolver o problema (4.4.13), através do uso de qualquer método baseado na informação da derivada de primeira ordem. Por último, deve-se destacar que o problema (4.4.13) é definido num espaço com $(nq)$ dimensões,

portanto muito menor que o espaço do problema (4.4.12), que tem $(nq+m)$ dimensões. Isto é, o número de variáveis do problema (4.4.13) não depende do número de observações *m*. Deve-se observar que, em muitas aplicações reais, o número de observações *m* é muito maior do que o número de atributos *n*, como salienta BAGIROV e YEARWOOD (2006).

A solução do problema de agrupamento original pode ser obtida pelo uso do Algoritmo de Suavização Hiperbólica, descrito a seguir em uma forma simplificada.

**Algoritmo Simplificado**

**Passo de Inicialização:** Escolha valores $0 < \rho_1 < 1$, $0 < \rho_2 < 1$, $0 < \rho_3 < 1$; seja $k = 1$ e escolha valores iniciais: $x^0$, $\gamma^1$, $\tau^1$, $\varepsilon^1$.

**Passo Principal:** Repita indefinidamente

Resolva o problema (4.4.13) com $\gamma = \gamma^k$, $\tau = \tau^k$, $\varepsilon = \varepsilon^k$, iniciando em um ponto inicial $x^{k-1}$, e seja $x^k$ a solução obtida.

Seja $\gamma^{k+1} = \rho_1 \gamma^k$, $\tau^{k+1} = \rho_2 \tau^k$, $\varepsilon^{k+1} = \rho_3 \varepsilon^k$, $k = k+1$. □

Como em outros métodos de suavização, a solução para o problema de agrupamento é obtida através da resolução de uma sequência infinita de subproblemas de minimização irrestritos ($k = 1, 2, \ldots$ no Passo Principal).

Note que o algoritmo faz $\tau$ e $\gamma$ se aproximarem de zero, logo as restrições dos subproblemas que ele resolve, como dado em (4.3.4), tendem àquelas de (4.2.7). Adicionalmente, o algoritmo faz $\varepsilon$ se aproximar de zero, portanto, em um movimento simultâneo, o problema resolvido (4.2.7) gradativamente aproxima-se do problema (4.2.6).

Implicitamente é assumido que o algoritmo encontra, $x^k$, uma solução global do *k*-ésimo subproblema suavizado.

Sob essas hipóteses, e devido às propriedades de continuidade de todas as funções envolvidas, a sequência $x^1, x^2, \ldots$ de valores ótimos tende ao valor ótimo de (4.1.1).

# Capítulo 5

# Utilização do Algoritmo de Suavização Hiperbólica em Taxonomia de Macroalgas

A seguir, apresentamos uma nova experiência computacional, a fim de demonstrar o desempenho do HSCM (Hyperbolic Smoothing Clustering Method), em particular, para demonstrar a sua capacidade para resolver problemas envolvendo dados biométricos.

Os dados utilizados foram obtidos a partir da medida da morfologia de três estruturas utilizadas na taxonomia tradicional das macroalgas do gênero *Caulerpa* (*frondes, estolão e rizoides*) (Figura 5).

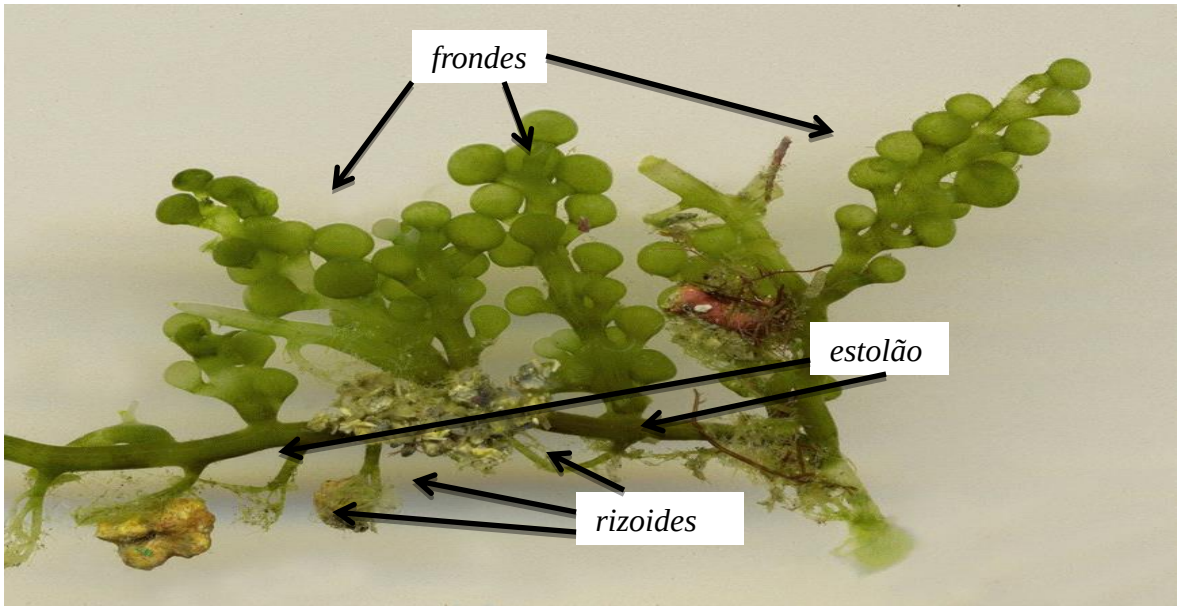



**Figura 5:** Foto de um exemplar do gênero *Caulerpa*, identificando suas estruturas morfológicas (*frondes*, *estolão*, *rizoides*) utilizados na taxonomia morfológica.

## 5.1 Análise do Uso do HSCM em Taxonomia de Algas do Gênero Caulerpa

Com o objetivo de comparar o resultado taxonômico obtido pelo método, *vis-à-vis*, com os resultados obtidos com o uso de técnicas moleculares para taxonomia do gênero *Caulerpa*, foram feitas pesquisas a dados documentais tais como, fontes primárias, bem como por meio de pesquisa bibliográfica de trabalhos especializados em filogenia do gênero estudado.

As medidas das variáveis foram extraídas de BARATA (2008) e estão dispostas na Tabela 5.1. Os experimentos foram realizados em um Notebook Intel Core i7-2620M Windows com 2.70GHz e 8 GB RAM.

Os dados de biometria das algas, antes da entrada no algoritmo de *clusters*, são normalizados de maneira que cada componente tenha média igual à zero ($\bar{x} = 0$) e desvio padrão igual a um ($\sigma = 1$). Em cada uma das oito componentes.

Desta forma através da normalização todas as componentes tem a mesma influencia no tratamento dos dados no algoritmo de clusterização.

**Tabela 5.1:** A primeira coluna é relação nominal das algas em estudo. Nas segunda e terceira coluna são listadas respectivamente as médias das medidas (cm) dos comprimentos mínima e máxima das *frondes* (2ª min. e 3ª máx.); As larguras das *frondes* são apresentadas respectivamente nas quarta e quinta colunas (4ª min. e 5ª máx.); As medidas do diâmetro (cm) do *estolão* estão dispostas nas colunas seis e sete (6ª min. e 7ª máx.); e dos *rizoides* na oitava e nona coluna (8ª min. e 9ª máx.).

| 1ª | 2ª | 3ª | 4ª | 5ª | 6ª | 7ª | 8ª | 9ª |
|---|---|---|---|---|---|---|---|---|
| **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 |

O gênero *Caulerpa* foi reconhecido por Lamouroux em 1809, estas algas são encontradas em ambientes marinhos tropicais e tradicionalmente são reconhecidas com base em suas características morfológicas (WEBER-VAN, 1898; COPPEJANS & BEECKMAN, 1989).

Hoje a identificação rápida e correta dessa alga é motivo de estudos em todo mundo pelo grande impacto econômico que tem gerado em vários ecossistemas, principalmente do Mar Mediterrâneo, onde sua reprodução tem alterado populações de outras espécies (BARATA, 2008).

Esse Gênero apresenta uma dificuldade considerável em identificação taxonômica em nível de espécie, devido à plasticidade fenotípica em caracteres diagnósticos (MEUSNIER *et al*., 2004). Isto pode ser fundamentada pelo fato de que, de 359 espécies (incluindo formas e variedades) do gênero *Caulerpa,* apenas 85 são taxonomicamente válido GUIRY & GUIRY (2013). Este fato levou alguns pesquisadores a descrever diferentes variedades e forma de uma espécie, muitas vezes de maneira equivocada (VAN REINE *et al*., 1996; BRAYNER *et al*., 2008).

A aplicação da metodologia aqui apresentada permitiu a formação de grupos, a partir de sucessivas partições realizadas com o uso de algoritmo HSCM, onde o arranjo pode ser visualizado na Figura 5.1.

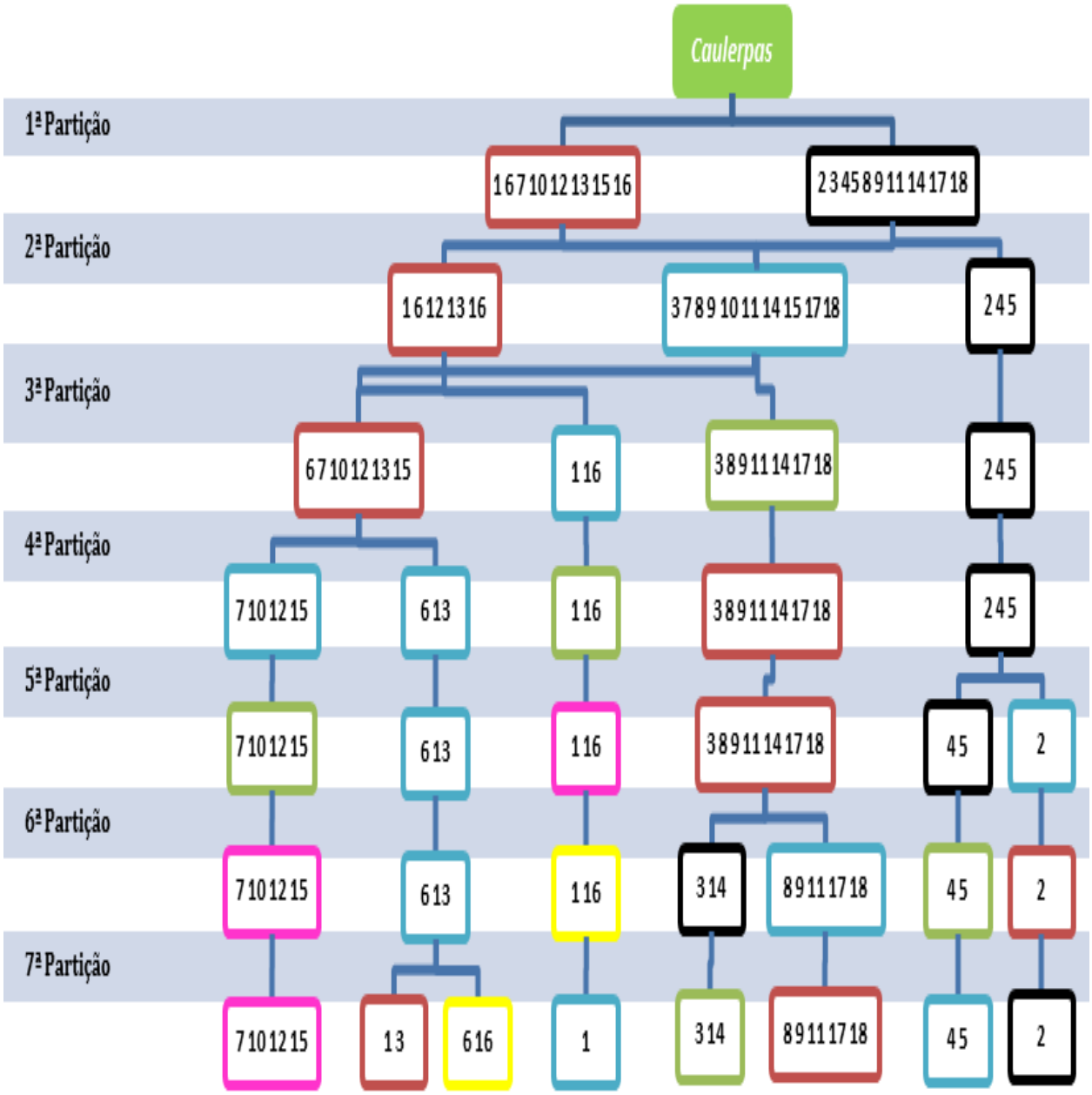
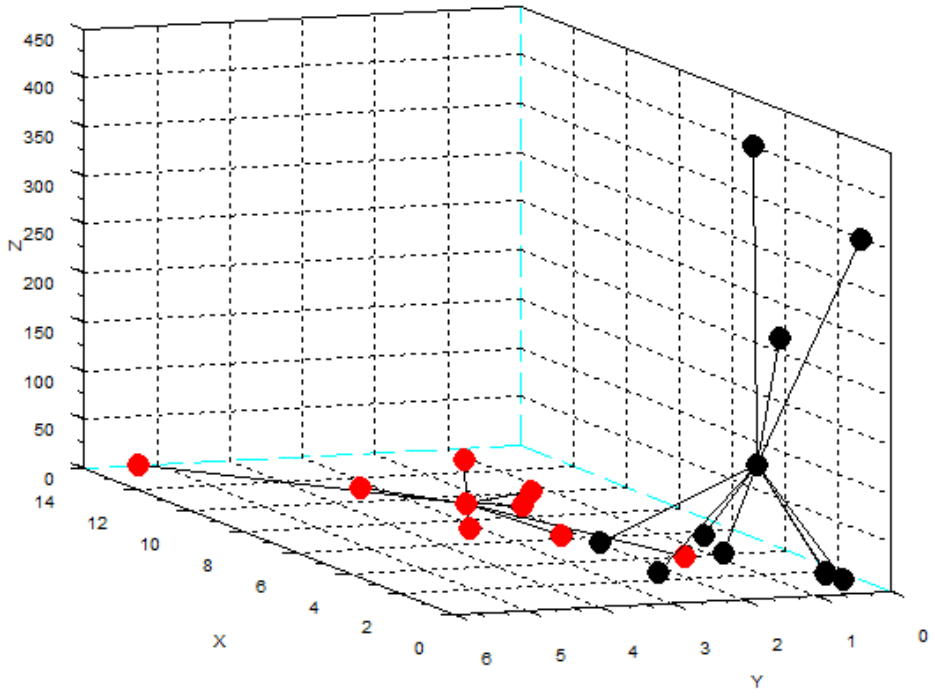



**Figura 5.1:** Diagrama da divisão dos grupos com o uso do HSCM.

Estudos relataram que em um universo de 241 amostras do Gênero *Caulerpa*, 12,7% foram classificados morfologicamente de forma errada por ficologistas experientes. As espécies de *C. ashmeadii*, *C. taxifolia*, *C. mexicana* e *C. sertularioides*, são morfologicamente semelhantes e podem ser confundidas (OLSEN *et al*., 1998).


**Figura 5.2:** Foto de exemplares de *C. ashmeadii* **(c)**; *C. taxifolia* **(d)**; *C. mexicana* **(e)** e *C. Sertulararioides* **(f)**.

As algas *C. ashmeadii* e *C. taxifolia*, estão agrupadas em um mesmo cluster em nossos resultados (gráfico 5.1). No trabalho de BARATA (2008), essas espécies possuem 100 por cento de homologia das sequências comparadas das análises do *tuf*A cpDNA, o que demonstram alta afinidade filogenética e resultados compatíveis com o método proposto.



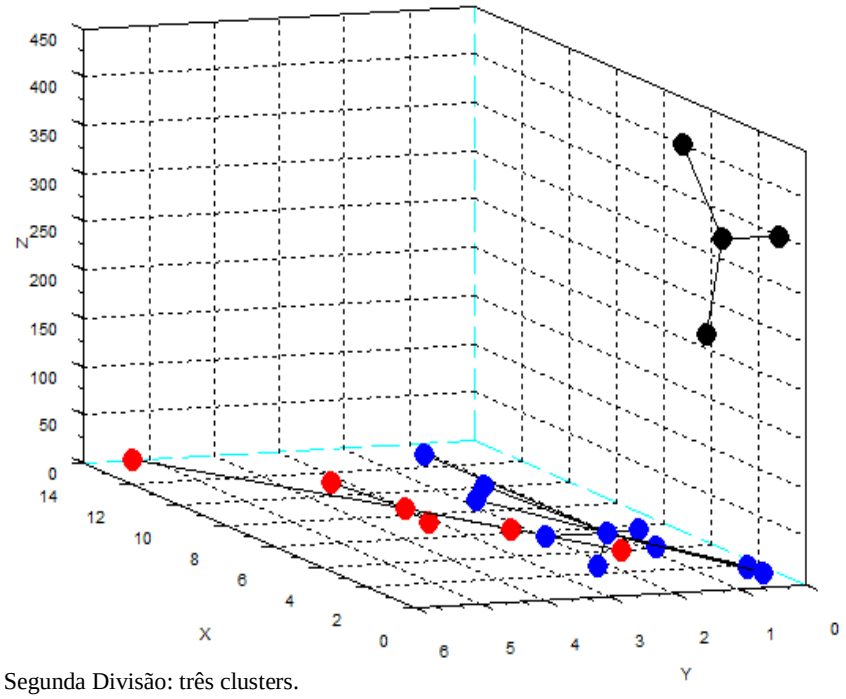

Primeira divisão: dois clusters.

**1ª PARTIÇÃO**
**GRUPO 1 (vermelho)**
ESPECIE 1 ***C. ashmeadii***
ESPECIE 6 *C. lanuginosa*
ESPECIE 7 *C. mexicana*
ESPECIE 10 *C. prolifera*
ESPECIE 12 *C. racemosa*
ESPECIE 13 *C. scalpelliformis*
ESPECIE 15 *C. sertularioides*
ESPECIE 16 ***C. taxifolia***

**GRUPO 2 (preto)**
ESPECIE 2 *C. brachypus*
ESPECIE 3 *C. cupressoides*
ESPECIE 4 *C. fastigiata*
ESPECIE 5 *C. kempfii*
ESPECIE 8 *C. microphysa*
ESPECIE 9 *C. murrayi*
ESPECIE 11 *C. pusilla*
ESPECIE 14 *C. serrulata*
ESPECIE 17 *C. verticillata*
ESPECIE 18 *C. Webbiana*

**Gráfico 5.1:** Representação Gráfica dos clusters formados após a primeira partição utilizando HSCM.

Em nossos resultados foram obtidos, também grupos similares aos dos estudos de OLSEN *et al.*, (1998) e KAZI *et al.*, (2013), que mostram que *C. taxifolia* e *C. mexicana* formam clados colocados separadamente em árvores filogenéticas, ou seja, em agrupamentos diferentes segundo o nosso método (gráfico 5.2).




Segunda Divisão: três clusters.

**2ª PARTIÇÃO**

**GRUPO 1 (vermelho)**
ESPECIE 1 *C. ashmeadii*
ESPECIE 6 *C. lanuginosa*
ESPECIE 12 *C. racemosa*
ESPECIE 13 *C. scalpelliformis*
ESPECIE 16 ***C. taxifolia***

**GRUPO 2 (preto)**
ESPECIE 2 *C.brachypus*
ESPECIE 4 *C. fastigiata*
ESPECIE 5 *C. kempfii*

**GRUPO 3 (azul)**

ESPECIE 3 *C. cupressoides*
ESPECIE 7 ***C. mexicana***
ESPECIE 8 *C. microphysa*
ESPECIE 9 *C. murrayi*
ESPECIE 10 *C. prolifera*
ESPECIE 11 *C. pusilla*
ESPECIE 14 *C. serrulata*
ESPECIE 15 *C. sertularioides*
ESPECIE 17 *C. verticillata*
ESPECIE 18 *C. webbiana*

**Gráfico 5.2:** Representação Gráfica dos clusters formados após a segunda partição utilizando HSCM.

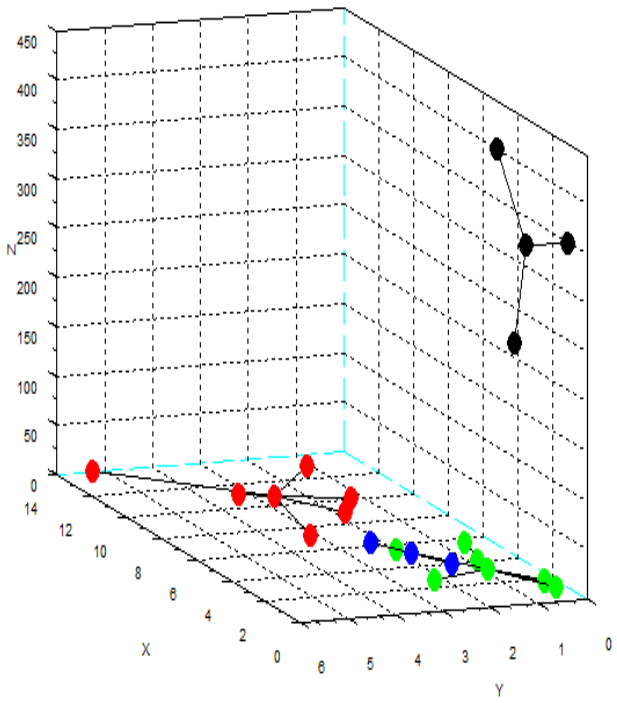
Observando-se as espécies *C. kempfii*, *C. verticillata* e *C. pusilla*, possuem caracteres morfológicos semelhantes, tais como talo delicado de tamanho reduzido e com ramos assimiladores cobertos por verticilos de râmulos dicotomicamente ramificados. Segundo TORRANO-SILVA *et al* (2013), a principal característica que difere *C. verticillata* de *C. pusilla*, é a ausência de pelos no estolão da primeira (Figura 5.3).


**Figura 5.3:** Foto de exemplares de *C. verticillata* **(a)** e *C. pusilla* **(b).**

Segundo BARATA (2008), *C. pusilla* apresentou um posicionamento no cladograma que não correspondeu ao encontrado baseado em seus caracteres morfológicos. Pois *C. pusilla* posicionou-se no clado juntamente com *C. cupressoides* e *C. serrulata*, que são espécies que apresentam talo robusto, chegando a mais de 20 cm de comprimento (gráfico 5.3).

Esse resultado é similar ao encontrado com a aplicação do método aqui apresentado. Verificou-se também que os resultados eram consistentes aos estudos moleculares de YEH & CHEN (2004), para o comportamento da filogenia de *C. webbiana*, em estudos de árvores filogenéticas com 90 e 100% de *bootstrap* reforçando

que *C. webbiana* é mais próxima do grupo que contém as variedades de *C. cupressoides e C. serrulata* (gráfico 5.4).

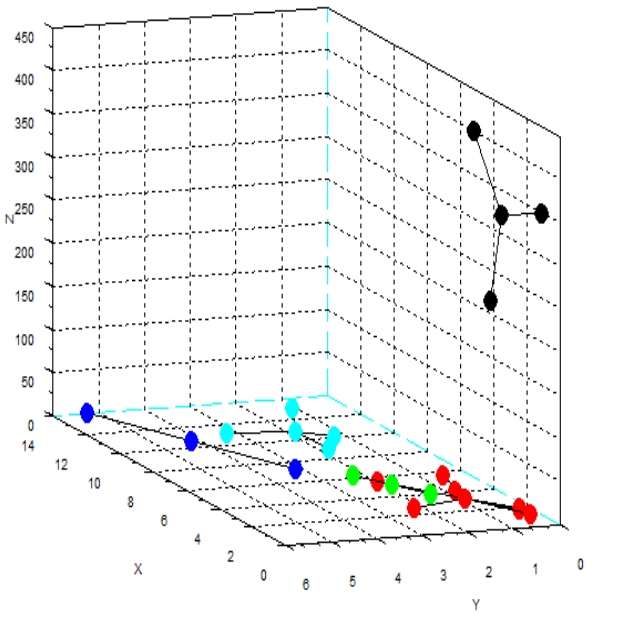



Terceira Divisão: quatro clusters.

**3ª PARTIÇÃO**
**GRUPO 1 (vermelho)**
ESPECIE 6 *C. lanuginosa*
ESPECIE 7 *C. mexicana*
ESPECIE 10 *C. prolifera*
ESPECIE 12 *C. racemosa*
ESPECIE 13 *C. scalpelliformis*
ESPECIE 15 *C. sertularioides*
**GRUPO 2 (preto)**
ESPECIE 2 *C. brachypus*
ESPECIE 4 *C. fastigiata*
ESPECIE 5 *C. kempfii*

**GRUPO 3 (azul)**
ESPECIE 1 *C. ashmeadii*
ESPECIE 16 *C. taxifolia*
**GRUPO 4 (verde)**
ESPECIE 3 ***C. cupressoides***
ESPECIE 8 *C. microphysa*
ESPECIE 9 *C. murrayi*
ESPECIE 11 ***C. pusilla***
ESPECIE 14 ***C. serrulata***
ESPECIE 17 *C. verticillata*
ESPECIE 18 *C. webbiana*

**Gráfico 5.3:** Representação Gráfica dos clusters formados após a terceira partição utilizando HSCM.

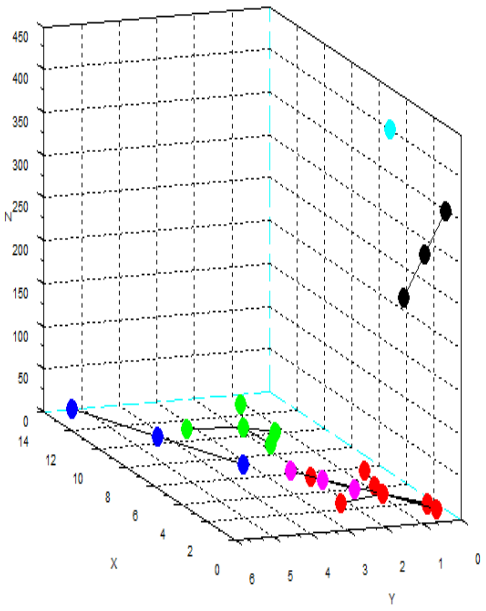



Quarta Divisão: quatro clusters.

**4ª PARTIÇÃO**
**GRUPO 1 (vermelho)**
ESPECIE 3 ***C. cupressoides***
ESPECIE 8 *C. microphysa*
ESPECIE 9 *C. murrayi*
ESPECIE 11 *C. pusilla*
ESPECIE 14 ***C. serrulata***
ESPECIE 17 *C. verticillata*
ESPECIE 18 ***C. webbiana***
**GRUPO 2 (preto)**
ESPECIE 2 *C. brachypus*
ESPECIE 4 *C. fastigiata*
ESPECIE 5 *C. kempfii*
**GRUPO 3 (azul escuro)**
ESPECIE 6 *C. lanuginosa*
ESPECIE 13 *C. scalpelliformis*
**GRUPO 4 (verde)**
ESPECIE 1 *C. ashmeadii* .
ESPECIE 16 *C. taxifolia*
**GRUPO 5 (azul claro)**
ESPECIE 7 *C. mexicana*
ESPECIE 10 *C. prolifera*
ESPECIE 12 *C. racemosa*
ESPECIE 15 *C. sertularioides*

**Gráfico 5.4:** Representação Gráfica dos clusters formados após a quarta partição utilizando HSCM.

O grupo composto por *C. mexicana, C. prolifera, C. racemosa* e *C. sertulararioides* (gráfico 5.5) apresentado na quarta partição, mesmo sem apresentarem semelhanças morfológicas, formam um agrupamento, semelhante ao grupo proposto por JOUSSON *et al.,* (1998).




Quinta Divisão: seis clusters.

**5ª PARTIÇÃO**
**GRUPO 1 (vermelho)**
ESPECIE 2 *C. brachypus*

**GRUPO 2 (preto)**
ESPECIE 3 *C. cupressoides*
ESPECIE 14 *C. serrulata*

**GRUPO 3 (azul escuro)**
ESPECIE 8 *C. microphysa*
ESPECIE 9 *C. murrayi*
ESPECIE 11 *C. pusilla*
ESPECIE 17 *C. verticillata*
ESPECIE 18 *C. webbiana*

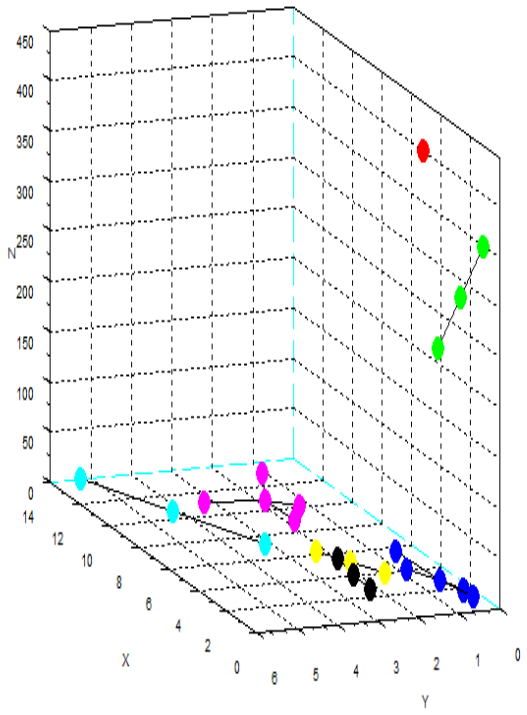
**GRUPO 4 (verde)**
ESPECIE 4 *C. fastigiata*
ESPECIE 5 *C. kempfii*
**GRUPO 5 (azul claro)**
ESPECIE 6 *C. lanuginosa*
ESPECIE 13 *C. scalpelliformis*

**GRUPO 6 (rosa)**
ESPECIE 7 ***C. mexicana***
ESPECIE 10 ***C. prolifera***
ESPECIE 12 ***C. racemosa***
ESPECIE 15 ***C. sertularioides***

**Gráfico 5.5:** Representação Gráfica dos clusters formados após a quinta partição utilizando HSCM.

O agrupamento formando por *C. cupressoides* e *C. serrulata* (gráficos 5.6), é similar ao resultado obtido por BARATA (2008) quando comparou as análises das sequências genéticas do tufA as cpDNA, e obteve a comprovação de que ambas apresentam alta afinidade filogenética. Assim, pode-se observar que não há um padrão consistente observado na relação entre caracteres morfológicos e colocação na árvore filogenética de taxa com base nos marcadores moleculares. Da mesma forma, que KAZI *et al.,* (2013), observou em seu estudo sobre *C. cupressoides* e *C. serrulata* que são

claramente diferentes em características morfológicas, mas apresentam linhagens parafiléticas.

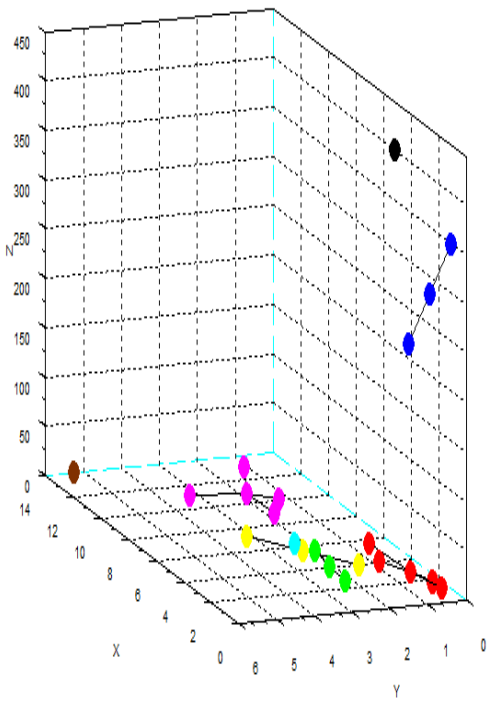



Sexta Divisão: sete clusters.

**6ª PARTIÇÃO**
**GRUPO 1 (vermelho)**
ESPECIE 2 *C. brachypus*
**GRUPO 2 (preto)**
ESPECIE 3 ***C. cupressoides***
ESPECIE 14 ***C. serrulata***
**GRUPO 3 (azul escuro)**
ESPECIE 8 *C. microphysa*
ESPECIE 9 *C. murrayi*
ESPECIE 11 *C. pusilla*
ESPECIE 17 *C. verticillata*
ESPECIE 18 *C. webbiana*
**GRUPO 4 (verde)**
ESPECIE 4 *C. fastigiata*
ESPECIE 5 *C. kempfii*
**GRUPO 5 (azul claro)**
ESPECIE 6 *C. lanuginosa*
ESPECIE 13 *C. scalpelliformis*
**GRUPO 6 (rosa)**
ESPECIE 7 *C. mexicana*
ESPECIE 10 *C. prolifera*
ESPECIE 12 *C. racemosa*
ESPECIE 15 *C. sertularioides*
**GRUPO 7 (amarelo)**
ESPECIE 1 *C. ashmeadii.*
ESPECIE 16 *C. taxifolia*

**Gráfico 5.6:** Representação Gráfica dos clusters formados após a sexta partição utilizando HSCM.

Segundo PILLMANN *et al* (1997), *C. scalpelliformes* é uma espécie separada das outras, ficando num clado isolado e terminal, em consonância com os nossos resultados. Da mesma que o comportamento de *C. brachypus*, ficando isolada (gráfico 5.7), demonstrando que a espécie mesmo semelhante a *C. prolifera*, apresenta-se distinta das demais espécies de *Caulerpa* em estudo, diante de estudos filogenéticos segundo WYNNE *et al* (2009), essa espécie foi proposta como um grupo filogeneticamente separado, o que vem corroborar os resultados encontrados em nossas análises.




Sétima Divisão: oito clusters.

**7ª PARTIÇÃO**
**GRUPO 1 (vermelho)**
ESPECIE 8 C. *microphysa*
ESPECIE 9 C. *murrayi*
ESPECIE 11 C. *pusilla*
ESPECIE 17 C. *verticillata*
ESPECIE 18 C. *webbiana*
**GRUPO 2 (preto)**
ESPECIE 2 C. ***brachypus***
**GRUPO 3 (azul)**
ESPECIE 4 C. *fastigiata*
ESPECIE 5 C. *kempfii*
**GRUPO 4 (verde)**
ESPECIE 3 C. *cupressoides*
ESPECIE 14 C. *serrulata*
**GRUPO 5 (azul claro)**
ESPECIE 1 C. *ashmeadii*
**GRUPO 6 (rosa)**
ESPECIE 7 C. *mexicana*
ESPECIE 10 C. *prolifera*
ESPECIE 12 C. *racemosa*
ESPECIE 15 C. *sertularioides*
**GRUPO 7 (amarelo)**
ESPECIE 6 C. *lanuginosa*
ESPECIE 16 C. *taxifolia*
**GRUPO 8 (marron)**
ESPECIE 13 *C.* ***scalpelliformis***

**Gráfico 5.7:** Representação Gráfica dos clusters formados após a sétima partição utilizando HSCM.

O sucesso alcançado pelo uso da nova metodologia se deve em grande parte a facilidade e rapidez da aplicação do método. Além disso, possibilita o incremento de novas representações através gráficos tridimensionais (3D), onde as interpretações serão bem mais fáceis de compreender, apresentando grande vantagem na visualização dos dados frente ao dendograma, podendo ser visto na sequencia de formação dos grupos não só as distâncias entre eles, mas também permite observar as espécies que se encontram na inserção entre grupos, ou seja, organismos que mantém alguma característica tanto de um como de outro grupo. Vale ressaltar que os resultados aqui apresentados utilizando uma metodologia alimentada com dados exclusivamente morfológicos demonstra que a análise da morfologia aliada a um método de partição pode ser uma alternativa para esclarecimento de novos sistemas de classificação.

# Capítulo 6

# Conclusão

> *"As nossas classificações voltarão a ser, tanto quanto puderem ser feitas, genealógicas; indicarão então o que se pode denominar o verdadeiro plano da criação.*
> *"Os órgãos rudimentares testemunharão infalivelmente a respeito da natureza de conformações há muito perdidas."*
>
> (Charles Darwin)

Os sistemas de classificação, na sua generalidade, são muito diversos, enfatizam pontos de vista pessoais dos seus atores, sendo por isso mesmo discutível e passiveis de correções e modificações, justamente porque tratam com organismos vivos sujeitos a contínuas alterações e influência do ambiente.

Sendo assim, a ferramenta de classificação mais adequada dependerá de quais características o taxonomista dispõe e do melhor tratamento, análise destas características. Assim, "cada caso é um caso" e o usuário sempre deve estudar cuidadosamente o seu problema, pois com o surgimento de novos métodos de agrupamento, tornou-se cada vez mais importante a escolha criteriosa do método com vistas à realmente solucionar, ou ao menos ajudar a resolver, a sua necessidade de

classificação. Sendo que um método de agrupamento que satisfaça os requisitos para um grupo de usuários pode não satisfazer os requisitos de outros. Pois o agrupamento está no olhar do especialista, assim realmente agrupamento de dados deve envolver as necessidades do usuário ou sua aplicação.

Neste contexto, este trabalho propõe uma ferramenta inovadora para realizar agrupamento de dados biológicos em especial das espécies de algas marinhas do Gênero *Caulerpa*, com o uso do algoritmo de Suavização Hiperbólica (HSCM) cujo maior diferencial em relação às abordagens originais é a similaridade dos agrupamentos obtidos com dados moleculares.

É importante reforçar a visão que, mesmo sendo uma área nova em taxonomia de algas marinhas, o algoritmo de suavização hiperbólica tem grande potencial e incontáveis aplicações.

Como perspectiva para futuros trabalhos, sugerimos sua utilização em outras áreas, em especial a de micro-organismos. Fungos e bactérias apresentam problemas de sistemática históricos, como reclassificações constantes, principalmente nos últimos trinta anos, com o advento da Biologia Molecular. Temos grandes expectativas que a utilização desse modelo possa solucionar várias questões que ainda hoje desafiam os pesquisadores dessas áreas.

O algoritmo proposto é funcional, robusto e os diversos usos se mostram promissores.

# Referências Bibliográficas

ACKERMAN, M. & S. BEN-DAVID, 2013. "A characterization of linkage-based hierarchical clustering." *Journal of Machine Learning Research.*

ALIGULIYEV, R. M., 2009. "Performance evaluation of density-based clustering methods." *Information Sciences* 179.20: 3583-3602.

ALLÈGRE, C. J. & SCHNEIDER, S. H., 1994. The evolution of the Earth. *Sci. Amer.*, 271:44 - 51.

AMORIM, D.S., 2002. *Fundamentos de sistemática filogenética.* Ribeirão Preto: Editora Holos.

BAGIROV, A. M. & J. YEARWOOD, 2006. "A new nonsmooth optimization algorithm for minimum sum-of-squares clustering problems." *European Journal of Operational Research* 170.2: 578-596.

BAGIROV, A. M., 2008. "Modified global k-means algorithm for minimum sum-of-squares clustering problems." *Pattern Recognition* 41.10: 3192-3199.

BARATA, D., 2008. *Taxonomia e filogenia do gênero Caulerpa J.V. Lamour (Bryopsidales, Chlorophyta) no Brasil*. Tese (Doutorado em Biologia Vegetal), Instituto de Botânica, São Paulo.

BELTON, G. S., et al. "Resolving phenotypic plasticity and species designation in the morphologically challenging *Caulerpa racemosa peltata* complex (Chlorophyta, Caulerpaceae)." *Journal of Phycology* 50.1 (2014): 32-54.

BHATTACHARYA, D. & MEDLIN, L., 1998. *Algal phylogeny and the origin of land plants.* Plant. Physiol., 116**:** 9 - 15.

BICUDO, C. E. M.; MENEZES, M., 2006. *Gênero de Algas de águas Continetais do Brasil*. Chave para identificação e descrição. Ed. Rima. 2.ed.

BOLD, H. C., 1972. *O Reino Vegetal.* São Paulo: Edgard Blucher LTDA.

BRAYNER, S.; SONIA, M. B. P. & MARIA, E. B. P., 2008. "Taxonomia e distribuição do gênero *Caulerpa* Lamouroux." *Acta Botanica Brasilica* 22.4: 914-928.

BULLERI, F., et al., 2010. "The seaweed Caulerpa racemosa on Mediterranean rocky reefs: from passenger to driver of ecological change." *Ecology* 91.8: 2205-2212.

CALIJURI, M. C.; ALVES, M. A. & SANTOS, A. C. A., 2006. *Cianobactérias e cianotoxinas em águas continentais.* São Carlos: Rima; 118 p.

COPPEJANS, E. & T. Beeckman, 1989. "*Caulerpa* section Sedoideae (Chlorophyta, Caulerpales) from the Kenyan coast." *Nova Hedwigia* 49.3-4: 381-393.

COPPEJANS, E., W. P. & REINE, V., 1992. "Seaweeds of the Snellius-II Expedition (E. Indonesia): the genus *Caulerpa* (Chlorophyta-Caulerpales)." *Buil. Séanc. Acad. r. Sei. Outre-Mer, nr. 37*: 667-712.

DAWES, C. J. *Marine botany*. John Wiley & Sons, 1998.

DING, W. et al., 2009. "Discovery of feature-based hot spots using supervised clustering." *Computers & Geosciences* 35.7: 1508-1516.

DUVE, C., 1996. The birth of complex cells. *Sci. Amer*. 274: 50 - 57.

ESTER, M. et al., 1998. *Clustering for mining in large spatial databases. KI* 12.1: 18-24.

EVERITT BS, LANDAU S. & LEESE, M., 2001. *Cluster Analysis*, 4th edn. Arnold, London.

FASULO, D., 1999. "An analysis of recent work on clustering algorithms. "*Department of Computer Science & Engineering, University of Washington*.

FISHER, R. A., 1936. "The use of multiple measurements in taxonomic problems." *Annals of eugenics* 7.2: 179-188.

GARAI, G. & CHAUDHURI, B. B., 2004. "A novel genetic algorithm for automatic clustering." *Pattern Recognition Letters* 25.2: 173-187.

GHOSH, P. et al., 2004. "In vitro anti-herpetic activity of sulfated polysaccharide fractions from Caulerpa racemosa" *Phytochemistry* 65.23: 3151-3157.

GORDON, A.D., 1998. How many clusters? An Investigation of five procedures for detecting nested cluster structure. In: *Data Science, Classification, and Related Methods*, edited by C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. Bock, and Y. Baba. Tokyo: Springer-Verlag.

GRAHAM, L. E. & WILCOX, L. W., 2009. *Algae*. Rio de Janeiro: Prentice-Hall do Brasil.

GUIRY, M. D. & GUIRY, G. M. 2013. *AlgaeBase. World-wide electronic publication, National University of Ireland*, Galway. *Available at:*http://www.algaebase.org *(accessed September 4, 2013).*

HAN, J., KAMBER, M., 2001. *Data Mining*: Concepts and Techiniques. 1.ed. New York: Morgan Kaufmann.

HAN, T. & B. RUNNEGAR, 1992. "Megascopic eukaryotic algae from the 2.1-billion-year-old Negaunee Iron-Formation, Michigan." *Science* 257.5067: 232-235.

HICKMAN JR., C. P.; ROBERTS, L. S.; LARSON, A., 2004. *Princípios Integrados de Zoologia*.11. ed., Ed. Guanabara Kogan: Rio de Janeiro.

HILL, D. et al., 1998. "An algorithmic model for invasive species: Application to Caulerpa taxifolia (Vahl) C. Agardh development in the North-Western Mediterranean Sea." *Ecological modelling* 109.3: 251-266.

JAIN, A. K. & R. C. DUBES, 1988. *Algorithms for clustering data*. Prentice-Hall, Inc..

JAIN, A. K., 2010. "Data clustering: 50 years beyond K-means." *Pattern Recognition Letters* 31.8: 651-666.

JAIN, A. K.; NARASIMHA, M. M. & PATRICK, J. F., 1999. "Data clustering: a review." *ACM computing surveys (CSUR)* 31.3: 264-323.

JAJUGA, K.; SOKOLOWSKI, A. & BOCK, H. H., 2002. *Classification, clustering, and data analysis: Recent advances and applications (studies in classification, data analysis, and knowledge organization).*

JOUSSON, O. et al., 1998. "Molecular evidence for the aquarium origin of the green alga *Caulerpa taxifolia* introduced to the Mediterranean Sea." *Marine Ecology Progress Series* 172.0: 275-280.

JUDD, W. S.; CAMPBEL, C. S.; KELLONGG, E. A.; STEENS P. F.; DONOGUE, M. J., 2009. *Sistemática Vegetal: um enfoque filogenético.* 3.ed. Porto Alegre: Artmed. 612p.

KAANDORP, J. A. & J. E. KÜBLER, 2001. *The algorithmic beauty of seaweeds, sponges and corals*. Springer.

KAPRAUN, D. F., 2005. "Nuclear DNA content estimates in multicellular green, red and brown algae: phylogenetic considerations." *Annals of Botany* 95.1: 7-44.

KARABOGA, D. & C. OZTURK., 2011. "A novel clustering approach: Artificial Bee Colony (ABC) algorithm." *Applied Soft Computing* 11.1: 652-657.

KASTING, J. F., 1993. Earth's early atmosphere. *Science*, 259: 920 - 926.

KAUFMAN, L., & ROUSSEEUW, P. J., 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley & Sons, Inc. (March)

KAZI, M A.; REDDY, C. R. K. & BHAVANATH, J., 2013. "Molecular Phylogeny and Barcoding of *Caulerpa* (Bryopsidales) Based on the tufA, rbcL, 18S rDNA and ITS rDNA Genes." *PloS one* 8.12: e82438.

KITCHING, Ian J. et al., 1998. Cladistics: the theory and practice of parsimony analysis. *Systematics Association Publication*, n. 11.

KRISHNER, B.P., 1994. The Erth's Elements. *Sci. Amer*. 271: 37-43

LAM, D. W. & F. W. ZECHMAN, 2006. "Phylogenetic analyses of the Bryopsidales (ULVOPHYCEAE, CHLOROPHYTA) Based on Rubisco large subunit gene sequences." *Journal of Phycology* 42.3: 669-678.

LASZLO, M. & S. MUKHERJEE, 2007. "A genetic algorithm that exchanges neighboring centers for means clustering." *Pattern Recognition Letters* 28.16: 2359-2366.

LAVESSON, N., 2006. *Evaluation and Analysis of Supervised Learning Algorithms and Classifiers*. Blekinge Institute of Technology.

LEE, R. E., 2008. *Phycology*. Cambridge University Press.

LEGENDRE, P. & D. J. ROGERS, 1972. "Characters and clustering in taxonomy: a synthesis of two taximetric procedures." *Taxon*: 567-606.

LEGENDRE, P. & LEGENDRE, F. J. L., 2012. *Numerical ecology*. Vol. 20. Elsevier.

MACQUEEN, J., 1967. "Some methods for classification and analysis of multivariate observations." *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. No. 14.

MADL, P. & M. YIP., 2003. "Literature review of *Caulerpa taxifolia*."*HTTP: http://www. sbg. ac. at/ipk/avstudio/pierofun/ct/caulerpa.* Accessed 12 February 2004.

MARGULIS, L. & SCHWARTZ, K. V., 2001. *Cinco Reinos: Um Guia Ilustrado dos Filos da Vida na Terra*. Editora Guanabara Koogan S. A. 497 p.

MEUSNIER, I. et al., 2004. "Analysis of rDNA ITS1 indels in *Caulerpa taxifolia* (Chlorophyta) supports a derived, incipient species status for the invasive strain." *European Journal of Phycology* 39.1: 83-92.

OLIVEIRA, M.C. & MILSTEIN, D., 2010. Taxonomia molecular. In: PEDRINI, A de G. (Og.), *Macroalgas: uma introdução à taxonomia*, Technical Books Editora, Rio de Janeiro: 71-82

OLSEN, J. L. et al., 1998. "Mediterranean *Caulerpa taxifolia* and *C. mexicana* (Chlorophyta) are not conspecific." *Journal of Phycology* 34.5: 850-856.

PAPAVERO, N., 1994. *Fundamentos práticos de taxonomia zoológica*. Unesp.

PARK, H. & C. JUN, 2009. "A simple and fast algorithm for K-medoids clustering." *Expert Systems with Applications* 36.2: 3336-3341.

PARRA, O. O. & BICUDO, C. E. M., 1995. *Introducción a la biologia y sistemática de lãs algas de águas continentales*. Santiago, Chile: Gráfica Andes.

PEDRINI, A. G.., 2010. *Macroalgas: uma introdução à taxonomia*. Rio de Janeiro: Technical Books.

PILLMANN, A. et al., 1997. "Inter-and intraspecific genetic variation in *Caulerpa* (Chlorophyta) based on nuclear rDNA ITS sequences." *European Journal of Phycology* 32.4: 379-386.

REVIERS, B., 2003. *Biologie et phylogénie des algues*. Paris: Belin,. Tome 2. 255 p.

REVIERS, B., 2006. *Biologia e filogenia das algas*. Artmed. Porto Alegre. 280p.

RODRIGUES, J. A. G. & FARIAS, W. R. L., 2005. Extração, fracionamento, purificação e atividade anticoagulante dos polissacarídeos sulfatados da alga marinha verde *Caulerpa racemosa* (Caulerpales, Chlorophyta). In: *Congresso Brasileiro de Engenharia de Pesca* (pp. 1693-1701). Fortaleza: CD-Room do CONBEP, 14.

RODRIGUES, J. A. G., et al., 2010. "Polissacarídeos sulfatados isolados das clorofíceas *Caulerpa racemosa* e *Caulerpa cupressoides*–extração, fracionamento e atividade anticoagulante-doi: 10.4025/actascibiolsci. v32i2. 5923." *Acta Scientiarum. Biological Sciences* 32.2: 113-120.

SANTOS, CHARLES MORPHY DIAS DOS. "Os dinossauros de Hennig: sobre a importância do monofiletismo para a sistemática biológica." *Scientiae Studia*6.2 (2008): 179-200.

SCHOPF, J. W., 1993. Microfossils of the early Archaean Apex Chert: new evidence of the antiquity of life. *Science*, 260: 640 - 646.

SCHUH, RANDALL T. *Biological systematics: principles and applications*. Cornell University Press, 2000.

SOGIN, M. L.; GUNDERSON, J. H.; Elwood, H. J.; Alonso, R. A. & Peattie, D. A., 1989.Phylogenetic significance of the Kingdom concept: an unusual eukaryotic 16S-like ribosomal RNA from *Giardia lamblia*. *Science*, 243: 75 - 77.

SOUZA, MARTA MARIA CAETANO DE. "Avaliação dos efeitos renais e vasculares das lectinas das algas *Caulerpa cupressoides* e *Pterocladiela capillacea*." (2013).

TAYLOR, W. R., 1960. "Marine algae of the eastern tropical and subtropical coasts of the Americas.".

TEIXEIRA, VALÉRIA L., ALPHONSE KELECOM, OTTO R. GOTTLIEB. "Produtos naturais de algas marinhas." *Quim Nova* 14 (1991): 83-90.

TEIXEIRA, V. L., 2010. Taxonomia química. In: PEDRINI, A de G. (Og.), *Macroalgas: uma introdução à taxonomia, Technical Books Editora, Rio de Janeiro*: 83-97.

TORRANO-SILVA; BEATRIZ, N.; CARLOS, E. A. & EURICO, C. O., 2013. "Algas de aquarios ornamentales en Brasil: previsión de las introducciones." *Latin american journal of aquatic research* 41.2: 344-350.

TRONO, J. & GAVINO, C., 1999. "Diversity of the seaweed flora of the Philippines and its utilization." *Hydrobiologia* 398: 1-6.

WEBER-VAN, B. A., 1898. *Monographie des Caulerpes*.

WYNNE, M. J.; HEROEN, V. & DROR, L. A., 2009. "The recognition of *Caulerpa integerrima* (Zanardini) comb. et stat. nov. (Bryopsidales, Chlorophyta) from the Red Sea." *Phycologia* 48.4: 291-301.

UKABI, S., et al. "Surveying *Caulerpa* (Chlorophyta) species along the shores of the eastern Mediterranean." *Mediterranean Marine Science* 13.1 (2012): 5-11.

VALENTIM DA SILVA, EVANDRO, SOUZA, ANTÔNIA DE . "Estudo toxicológico, atividade antioxidante e antitumoral da macroalga marinha *Caulerpa taxifolia* (Vahl) Agardh (1817) CAULERPACEAE." (2011).

VERLAQUE, M., DURAND, C., HUISMAN, J. M., BOUDOURESQUE, C.-F. & LE PARCO, Y. 2003. On the identity and origin of the Mediterranean invasive *Caulerpa racemosa* (Caulerpales, Chlorophyta). *Eur. J. Phycol.* **38**:325–39.

VAN REINE, W.F. PRUD'HOMME, E. VERHEIJ, and E. COPPEJANS. "Species and ecads of Caulerpa (Ulvophyceae, Chlorophyta) in Malesia (South-East Asia): taxonomy, biogeography and biodiversity." *Netherland Journal of Aquatic Ecology* 30.2-3 (1996): 83-98.

XAVIER, A. E. & V. L. XAVIER, 2011. "Solving the minimum sum-of-squares clustering problem by hyperbolic smoothing and partition into boundary and gravitational regions." *Pattern Recognition* 44.1: 70-77.

XAVIER, A. E., 2010. "The hyperbolic smoothing clustering method." *Pattern Recognition* 43.3: 731-737.

XU, R. & D. WUNS, 2005. "Survey of clustering algorithms." *Neural Networks, IEEE Transactions on* 16.3: 645-678.

YEH, W. & G., CHEN, 2004. "Nuclear rDNA and internal transcribed spacer sequences clarify *Caulerpa racemosa* vars. from other *Caulerpa* species." *Aquatic botany* 80.3: 193-207.

ZADEGAN, R. S. M.; MEHDI, M. & FARAHNAZ, S., 2013. "Ranked medoids: A fast and accurate rank-based partitioning algorithm for clustering large datasets." *Knowledge-Based Systems* 39: 133-143.