Universidade Federal de Minas Gerais

Programa de Pós-Graduação em Engenharia Elétrica

Cristiano Luiz Silva Tavares

CLUSTER: um software para auxílio em estudos de dados biológicos

Belo Horizonte

Outubro de 2015

Cristiano Luiz Silva Tavares

CLUSTER: um software para auxílio em estudos de dados biológicos

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Dr. Adriano Vilela Barbosa

Belo Horizonte

Outubro de 2015

A Valentina, Melissa e Clarice.

Vocês que já são força, doçura e luz no meu viver.

AGRADECIMENTOS

Chegar para agradecer e louvar
O ventre que me gerou, o orixá que me tomou e a mão da doçura de Oxum que consagrou
Louvar a água de minha terra, o chão que me sustenta, o palco, o massapê, a beira do abismo, o punhal do susto de cada dia
Agradecer as nuvens que logo são chuva, que serenizam os sentidos e ensina a vida a reviver
Agradecer os amigos que fiz e que mantêm a coragem de gostar de mim, apesar de mim
Agradecer a alegria das crianças, as borboletas dos meus quintais; reais, ou não
Agradecer a cada folha, a toda raiz; as pedras majestosas e as pequeninas como eu, em Aruanda
Agradecer ao sol que raia o dia e a lua que, como o menino Deus, espraia luz e vira os meus sonhos de pernas pro ar
Agradecer as marés altas e também aquelas que levam para outros costados todos os males
Agradecer a tudo que canta livre no ar, dentro do mato, sobre o mar
As vozes que soam de cordas tênues e partem cristais
Agradecer aos senhores que acolhem e aplaudem esse milagre
Agradecer, ter o que agradecer, louvar e abraçar

Maria Bethânia, 2015

Especialmente,

Agradecer a confiança, ensinamentos e companheirismo do Prof. Adriano

Agradecer aos colegas biólogos colhidos nesta jornada; principalmente a grande amiga Carol

Agradecer aos colegas de mestrado que, juntos, construímos nossos diversos sonhos; principalmente Carlos, Diana, Douglas, Elder, Fredy, Lianny e Natália

Agradecer aos motoristas de viagem que trouxeram e levaram este sonho com a leveza de um bom sorriso todas as semanas

Agradecer os colegas do Instituto Federal do Espírito Santo

Agradecer aos velhos amigos que sempre me apoiaram

Agradecer a toda minha família, principalmente ao apoio incondicional dos meus irmãos: Rodrigo, Camila, Leandro e Cleidson

Agradecer a moça Sheila, companheira no início deste sonho e de diversos outros

Agradecer ao homem Sandro, que "boa aventurou-se" noites a dentro deste texto e do meu viver

Por fim, agradecer meus três maiores amores: Pai, Mãe e Vó. A vocês, só o mar

Ogunhê!

Agora que sinto amor
Tenho interesse no que cheira.
Nunca antes me interessou que uma flor tivesse cheiro.
Agora sinto o perfume das flores como se visse uma coisa nova.
Sei bem que elas cheiravam, como sei que existia.
São coisas que se sabem por fora.
Mas agora sei com a respiração da parte de trás da cabeça.
Hoje as flores sabem-me bem num paladar que se cheira.
Hoje às vezes acordo e cheiro antes de ver.

Alberto Caeiro

RESUMO

Com crescimento acelerado da quantidade de dados de origem biológica, surgem dois problemas: (i) o armazenamento e gestão de dados e (ii) a extração de informações a partir destes dados. O segundo problema é um dos principais desafios na biologia computacional, o que requer desenvolvimento de ferramentas e métodos capazes de transformar todos esses dados heterogêneos em conhecimento biológico. Parte deste conhecimento envolve determinar variações de expressões gênicas de dados biológicos. Descobrir o significado das expressões gênicas tem contribuído no desenvolvimento de técnicas na agricultura, na pecuária, no tratamento de doenças e em políticas de preservação de espécies de animais e plantas ameaçados de extinção. Desde modo, este trabalho propõe um software, intitulado Cluster, para auxiliar pesquisas em dados biológicos. Cluster atua diretamente na seleção de características, ou expressões gênicas, para a classificação de grupos de amostras. Cluster é capaz de otimizar a quantidade e a qualidade de características responsáveis para o agrupamento de indivíduos. A interface simples do software Cluster contribui de forma a facilitar sua configuração e apresentação de resultados claros. O software é testado em bases de dados com propriedades distintas. A especificidade, sensibilidade, eficiência e acurácia de classificação das amostras são métricas utilizadas para validar a seleção de características proposta em Cluster. Dentre os testes realizados destaca-se a determinação de alelos na distinção de tartarugas marinhas e seus híbridos, a determinação de características genômicas na distinção de tecidos gástricos cancerosos e a determinação de características morfológicas para a distinção de sementes de trigo.

Palavras-Chave: *Clustering*. Dados Biológicos. Otimização. Reconhecimento de Padrões. Seleção de características.

ABSTRACT

The ever increasing availability of biological data gives rise to two problems: (i) data storage and management and (ii) the extraction of useful information from these data. The latter problem is one of the main challenges in computational biology, and requires the development of tools and methods capable of transforming all these heterogeneous data into biological knowledge. Part of this knowledge involves determining variations in gene expression on biological data. Studies on biological data have contributed to the development of new techniques in agriculture, animal farming, in the treatment of diseases and in the development of policies for the preservation of endangered animal and plant species. Thus, this paper proposes a software, named Cluster, to assist research on genetic diversity. Cluster acts directly on the feature selection step of the classification problem. Cluster is able to optimize the quantity and quality of the features used to group individuals. The simple interface of the Cluster software helps its configuration and the presentation of clear results. The software is tested on databases with different properties. The specificity, sensitivity, efficiency and accuracy of the classification are metrics used to validate the feature selection mechanism proposed in Cluster. Tests performed on the software include: the determination of alleles for distinguishing sea turtles and their hybrids; the determination of genomic features for classification gastric cancer tissue and determination of morphological features for classification wheat seeds.

Keywords: Clustering. Biological Data. Optimization. Pattern Recognition. Feature Selection

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

AUC	*Area Under the ROC Curve*
B	Breast Cancer
C	Número de Características Selecionadas
Canad.	Canadense
Cc	*Carreta Carreta*
c_g	Centro de grupo
CGCA	*Complete Gradient Clustering Algorithm*
Cr	Coeficiente Relief
DNA	Ácido Desoxirribonucleico
Ei	*Eretmochelys imbricata*
EiCc	Híbrido entre as espécies *Eretmochelys imbricata* e *Carreta Carreta*
F	Expoente de fuzzificação
FNA	*Fine Needle Aspiration*
G	Número de grupos
G	Gastric Cancer
J	Medida de dissimilaridade dentro de grupos
K	Número de amostras vizinhas mais próximas utilizadas no ReliefF
KNN	*K-Nearest Neighbors*
Máx.	Máximo
MCMC	Monte Carlo via Cadeia de Markov
Mín.	Mínimo
P	Prostate Cancer
ROC	*Receiver Operating Characteristic*
S	Seeds
T	Turtles
TGD	Tumor Gástrico Difuso
TGI	Tumor Gástrico Intestinal
TN	Tecido normal

Sumário

Capítulo 1: Introdução

Com crescimento acelerado da quantidade de dados de origem biológica, surgem dois problemas: (i) o armazenamento e gestão de dados e (ii) a extração de informações a partir destes dados. O segundo problema é um dos principais desafios na biologia computacional, o que requer desenvolvimento de ferramentas e métodos capazes de transformar todos esses dados heterogêneos em conhecimento biológico (LARRAÑAGA, CALVO, *et al.*, 2006).

Grande parte destes dados biológicos tem origem em pesquisas sobre análise de diversidade genética. Segundo Primack (2014), diversidade genética é a variação da expressão gênica dentro de cada espécie, tanto entre populações geograficamente separadas como entre indivíduos de uma dada população. Ao analisar a diversidade genética de um grupo de indivíduos, pesquisadores buscam identificar características, ou expressões gênicas, que são capazes de diferenciar subgrupos de indivíduos ao longo do tempo.

O entendimento e reconhecimento da diversidade genética em dados biológicos tem permitido ao homem desenvolver e aprimorar técnicas empregadas em diversas áreas. Na agricultura e pecuária, por exemplo, o melhoramento genético tem ajudado a aumentar a qualidade e produção de carnes e vegetais (PIERCE, 2013). No campo da saúde, a genética está relacionada ao tratamento de diversas doenças, como o câncer (GRIFFITHS, WESSLER, *et al.*, 2008; ROBINSON, 2010). Bancos genéticos têm ajudado nos estudos de políticas para a preservação de espécies de animais e plantas ameaçadas de extinção (DINIZ e FERREIRA, 2000).

Dentre as técnicas utilizadas nas pesquisas sobre diversidade genética, destaca-se o *clustering*, conhecida como análise de grupos. Nesta análise, busca-se agrupar amostras ou indivíduos conforme suas características semelhantes (LARRAÑAGA, CALVO, *et al.*, 2006).

Ao agrupar amostras através de suas semelhanças, pesquisadores conseguem reconhecer certos padrões em expressões gênicas que auxiliam nos estudos genéticos. Percebe-se que a expressão gênica pode expressar-se em características morfológicas e em características moleculares. Por exemplo, enquanto é possível distinguir espécies de trigo avaliando as características físicas de suas sementes, a Síndrome de Down em um feto só é comprovada ao analisar características cromossômicas de suas células.

Contudo, encontrar as características, ou expressões gênicas, que melhor agrupam ou classificam amostras não é uma tarefa simples (GUYON e ELISSEEFF, 2006). Muitas vezes, é necessária a utilização de técnicas computacionais para a seleção destas características. Atualmente, vários softwares propõem métodos para o agrupamento de amostras, mas não lidam diretamente com a seleção das melhores características para este fim.

Sendo assim, esta dissertação propõe um software que auxilie pesquisadores, através de *clustering*, no estudo de dados biológicos. Este atua diretamente na otimização de seleção de características, designando aquelas que melhor separam, por análise de similaridade, as amostras nas classes em estudo. O software proposto é nomeado como Cluster.

O processo de seleção das características desenvolvido no software Cluster é baseado no método ReliefF (KONONENKO, 1994). As amostras são agrupadas seguindo o algoritmo *Fuzzy* C-Means (DUNN, 1974; BEZDEK, 1981). O software é composto ainda, por um algoritmo genético (HOLLAND, 1975) para a otimização. Desta forma, procura-se sempre obter as melhores características para a separação, ou classificação, das amostras.

Para melhor desempenho, o software Cluster trata de otimizar a seleção de características por dois meios: pela qualidade e pela quantidade de características selecionadas.

A qualidade das características é garantida ao se otimizar o número de amostras vizinhas mais próximas consideradas no cálculo do ReliefF. Já a quantidade é

otimizada diretamente, ao não selecionar todas as características de estudo para o agrupamento.

Junto a seleção de características, otimiza-se, ainda, o número de grupos utilizado do processo de *clustering*. Segundo Porras-Hurtado, Ruiz, *et al.* (2013) definir o número de grupos de uma base de dados não é uma tarefa simples para os pesquisadores.

Para demonstrar a eficiência da seleção de características proposta, o software Cluster é testado com diversas bases de dados. Seu desempenho é analisado pela acurácia, sensibilidade, especificidade e eficiência (FAWCETT, 2006) de classificação das amostras.

1.1 Objetivos do trabalho

1.1.1 Objetivo geral

Propor um software, para auxílio em análise de dados biológicos, que apresente bons resultados de desempenho ao atuar diretamente no modo de selecionar características para o processo de *clustering*.

1.1.2 Objetivos específicos

Dentre os objetivos específicos, acerca dos métodos trabalhados, destaca-se:

- Comprovar a influência de uma boa seleção de características para o processo de *clustering, e*;
- Propor um modo eficiente de se otimizar a seleção de características de uma base de dados em conjunto com número de grupos considerados no processo de *clustering*;

Perante a implementação computacional são destacados os seguintes objetivos:

- Criar um software que consiga bons resultados de acurácia na classificação de amostras, atuando na seleção de características para agrupamento destas;

- Criar uma ferramenta capaz de atuar em diversas bases de dados, principalmente, no que se refere a versatilidade dos tipos de dados que compõem estas bases e na aplicabilidade das mesmas, e;
- Apresentar um software com *layout* simples e de fácil operabilidade, com exposição de resultados claros e interativos para a análise de grupos.

1.2 Contribuições

Mais do que a proposta de um software para auxílio na análise de grupos de dados biológicos, esta dissertação busca apresentar técnicas otimizadas a serem utilizadas em diversas áreas de reconhecimento de padrões.

Reconhecimento de padrões é a ciência cujo o objetivo é a classificação de dados em um número de categorias ou classes a partir de suas características (THEODORIDIS e KOUTROUMBAS, 2009). Desta forma, as técnicas aqui apresentadas e otimizadas servem para diversos estudos, não sendo restrito a pesquisas de dados biológicos.

Sendo assim, de maneira específica, acredita-se que os pontos apresentados a seguir são contribuições desta dissertação.

Sobre os métodos trabalhados, são destacados:

- Otimização do método ReliefF pela alteração no número de amostras vizinhas mais próximas consideradas em seu cálculo;
- Utilização do método ReliefF otimizado com o algoritmo *Fuzzy* C-Means para geração de bons resultados de desempenho em classificação de dados, e;
- Criação de um algoritmo genético como ferramenta de otimização na determinação da quantidade de grupos presentes dentro de uma base de dados.

Sobre a implementação computacional é destacada:

- Apresentação de uma ferramenta simples de operação e exposição clara de seus resultados.

1.3 Estrutura do trabalho

Os assuntos discutidos neste trabalho estão organizados conforme mostrado a seguir.

- **Capítulo 2: Genética** – Apresenta revisão sobre genética, além de expor brevemente os softwares utilizados na atualidade em pesquisas sobre diversidade genética.
- **Capítulo 3: Metodologia** – Neste capítulo, os métodos e métricas empregados em toda a concepção do software Cluster são apresentados.
- **Capítulo 4: Apresentação do Software Cluster** – Exposição do *layout* final, estrutura, operações e forma de apresentação de resultados do software Cluster.
- **Capítulo 5: Testes e Discussão** – As bases de dados utilizadas na concepção do software são apresentadas e em seguida analisadas em uma série de testes. O desempenho do software Cluster é testado de acordo uma série de critérios.
- **Capítulo 6: Conclusões** – Neste capítulo, os objetivos propostos para esta dissertação são revisitados. São apresentadas perspectivas de trabalhos futuros e, por fim, são expostas publicações e premiação obtida durante o desenvolvimento deste trabalho.

Capítulo 2: Genética

Para melhor entendimento do software proposto é necessário compreender sua aplicabilidade no contexto atual. Para isto, dedica-se este capítulo. Brevemente, apresenta-se uma visão geral sobre genética, de forma a revelar a devida contribuição de softwares ligados à área. Cita-se, ainda, os principais softwares utilizados atualmente em pesquisas sobre diversidade genética, especialmente aqueles que utilizam algoritmos de agrupamento, assim como o software Cluster proposto nesta dissertação.

2.1 Visão geral sobre genética

Genética é o campo da ciência que examina como as características são passadas de uma geração para a seguinte (ROBINSON, 2010). A genética é fundamental para a vida de cada indivíduo pois ela influencia nossas características físicas, personalidade, inteligência e na susceptibilidade a inúmeras doenças (PIERCE, 2013). Genética se aplica a todo e qualquer ser vivo, não se restringindo a seres humanos.

A genética está intimamente ligada à evolução: a população evolui à medida que surgem novas variações de características e estas passam a ser herdadas, tornando-as mais comuns (GRIFFITHS, WESSLER, *et al.*, 2008; PIERCE, 2013). Novas populações, mais adaptadas ao meio, podem surgir deste processo.

O primeiro estudo detalhado sobre genética foi o trabalho de *Gregor Mendel*, um monge austríaco que viveu no século XIX (SNUSTAD e SIMMONS, 2013). Mendel

publicou o resultado de seus experimentos sobre cruzamentos controlados de ervilhas em 1865. Ele foi capaz de deduzir distintos fatores que levavam a informação sobre o desenvolvimento dos genitores à prole (GRIFFITHS, WESSLER, *et al.*, 2008).

Contemporâneos de Mendel, *Charles Darwin* e *Alfred Wallace* revolucionaram o pensamento científico ao darem credibilidade ao conceito de que a partir de um ancestral comum descenderam todos os seres vivos. Este pensamento fortaleceu e divulgou os estudos de Mendel (SNUSTAD e SIMMONS, 2013).

O século XIX é marcado com a primeira publicação de estudo ligado a área da genética, realizada por Mendel. Porém, a genética está presente junto a humanidade há muito mais tempo, de forma preciosa para nossa evolução. Talvez, o maior marco do trabalho da humanidade com a genética envolva a agricultura.

Ao aplicar os princípios genéticos para a domesticação de plantas e animais o homem inventou a agricultura. Hoje em dia, as principais culturas e animais usados na agricultura foram submetidos a extensas alterações genéticas, para aumentar significativamente os seus rendimentos e fornecer muitas características desejáveis. Dentre estas características destacam-se a resistência a doenças e pragas, qualidades nutricionais especiais, e características que facilitam a colheita (PIERCE, 2013).

Atualmente, o uso do conhecimento acerca da genética não está restrito a agricultura, sendo desenvolvido em diversas áreas. A indústria farmacêutica, por exemplo, é outra área em que a genética desempenha um papel importante. Medicamentos e aditivos alimentares são sintetizados por fungos e bactérias que têm sido manipulados geneticamente. A indústria da biotecnologia molecular emprega técnicas gênicas para o desenvolvimento e produção em massa de substâncias de valor comercial. O hormônio do crescimento, a insulina e o fator de coagulação são agora produzidos comercialmente por bactérias modificadas geneticamente (PIERCE, 2013).

Na área da medicina, a maior contribuição da genética envolve a pesquisa sobre a formação e destruição das diversas formas do câncer. O câncer é resultado da mutação genética de algumas células (GRIFFITHS, WESSLER, *et al.*, 2008; ROBINSON, 2010). Pesquisas sobre tumores de câncer contribuem essencialmente para o desenvolvimento de técnicas de tratamento dos enfermos. Esta contribuição é afirmada nos estudos sobre leucemia (GOLUB, SLONIM, *et al.*, 1999), câncer gástrico (HIPPO, TANIGUCHI, *et al.*, 2002), câncer de próstata (BEST, GILLESPIE, *et al.*, 2005), câncer de mama (CHARYTANOWICZ, NIEWCZAS, *et al.*, 2010), dentre outros.

O desenvolvimento nestas diversas áreas se deu, principalmente, após a apresentação da estrutura da molécula de DNA, ácido desoxirribonucleico, por Watson e Crick (1953a; 1953b). A partir daí a Genética Clássica, baseada nos experimentos de Mendel, deu lugar à Genética Molecular (SNUSTAD e SIMMONS, 2013). As pesquisas começaram a ser a nível molecular, através dos genes.

O gene corresponde a informação contida nas moléculas de DNA. São os genes que indicam como as células irão se reproduzir formando uma série de características nos indivíduos como: cor dos olhos, cor da pele e propensão a algumas doenças em seres humanos.

Sendo assim, com o estudo molecular, o desafio da genética passou a ser identificar e determinar o papel dos genes, ou seja, a entender como estes contribuem para constituição do indivíduo (GRIFFITHS, WESSLER, *et al.*, 2008). Esta não é uma tarefa simples. Para se ter uma ideia, entre os organismos celulares, o menor genoma conhecido é do *Mycoplasma genitalium* e possui 482 genes, enquanto um espermatozoide humano tem cerca de 20.500 genes (SNUSTAD e SIMMONS, 2013). A quantidade de genes e suas interações dificultam este estudo.

Mais que mapear o DNA de um indivíduo, ou seja, entender a influência dos genes na constituição dos indivíduos, é necessário compreender a aplicação e evolução dos genes dentro de uma população. Em outras palavras, é preciso entender que a constituição genética dos indivíduos de uma população varia. Surge, então, a Genética de Populações, onde os geneticistas buscam documentar a variabilidade genética e compreender seu significado perante grupos (SNUSTAD e SIMMONS, 2013). Neste tipo de estudo é possível identificar, por exemplo, evidências do surgimento de novas espécies.

O termo população é utilizado quando se tem um grupo de indivíduos em equilíbrio gênico segundo o princípio de Hardy-Weinberg (HARDY, 1908; WEINBERG, 1908). No equilíbrio gênico as frequências das expressões gênicas permanecem constantes com o passar de gerações, independente se um determinado gene é frequente ou raro neste grupo de indivíduos. Porém para se assumir tal equilíbrio, o grupo em análise deve ser grande, idealmente infinito. Este grupo, ainda, deve conter acasalamentos acontecendo de forma aleatória além de não ser afetado por mutações, migrações ou seleção natural (PIERCE, 2013). Sendo assim, entende-se que este grupo já não possui considerável interação gênica com outro grupo de indivíduos ao longo do tempo, formando assim as populações.

Nota-se que estudos de Genética de Populações exigem um nível de complexidade superior aos trabalhos desenvolvidos por Mendel. É necessária a aplicação de técnicas complexas para auxílio na determinação do papel dos genes. Softwares são empregados baseando-se em técnicas e análises distintas. Alguns destes softwares, voltados para o estudo sobre diversidade genética, são apresentados e discutidos a seguir. O software proposto neste trabalho utiliza dados moleculares e morfológicos para pesquisas de dados biológicos por meio da separação de grupos de indivíduos.

2.2 Softwares aplicados a diversidade genética

O software Cluster, desenvolvido neste trabalho, busca auxiliar pesquisadores no estudo de dados biológicos. Este ajuda a descrever padrões importantes dentro da Genética de Populações através da análise de grupos. Diversos outros softwares são encontrados na literatura para o estudo de diversidade genética e se diferem por diversos motivos, como: a técnica utilizada, o tipo de dado utilizado e o tipo de resultado apresentado. Diversos volumes do periódico *Molecular Ecology Resources* (ISSN 1755-098X) apresentam revisões e novos softwares aplicados à diversidade genética. Dentre os mais utilizados podemos destacar alguns, incluídos na Tabela 2.1.

Tabela 2.1: Softwares aplicados ao estudo de diversidade genética
(Fonte: ZOLET, SEGATTO, *et al. 2013 - adaptada*)

Software	Tipos de dados	Referência
ARLEQUIN	DNA, SNP, Microssatélite, MULT, FREQ	(EXCOFFIER e LISCHER, 2010)
BAPS	MULT	(CORANDER, WALDMANN, *et al.*, 2004)
FSTAT	Microssatélite, MULT	(HERED, 1995)
GENEPOP	Microssatélite, MULT	(RAYMOND e ROUSSET, 1995) (ROUSSET, 2008)
MSA	Microssatélite, MULT	(DIERINGER e SCHOLÖTTERER, 2003)
STRUCTURE	SNP, Microssatélite, MULT	(PRITCHARD, STEPHENS e DONNELLY, 2000) (FALUSH, STEPHENS e PRITCHARD, 2003)

Sendo: DNA, dados de sequência; FREQ, dados de frequência; MULT, marcadores multialélicos; SNP, polimorfismo de único nucleotídeo.

O software ARLEQUIN é um dos mais utilizados no estudo de diversidade genética devido à versatilidade de suportar diversos tipos de dados. O software MSA é considerado ponto inicial de muitas análises por ser capaz de converter arquivos em

vários formatos a serem utilizados em diversos softwares como o GENEPOP, STRUCTURE e ARLEQUIN (ZOLET, SEGATTO, *et al.*, 2013). O software GENEPOP efetua testes de equilíbrio de Hardy–Weinberg para a diferenciação de populações e para o desequilíbrio genotípico entre pares de *loci*; entre outras funções (ROUSSET, 2008). O software FSTAT é especialista em calcular índices da Estatística-F (*F-Statistics),* difundidos entre os biólogos como medidas para avaliação da diferenciação genética entre populações (HERED, 1995). Os softwares STRUCTURE e BAPS se diferenciam dos demais por utilizar inferência Bayesiana em seus algoritmos de análise de grupos.

Destaca-se um software brasileiro dentro do cenário de programas que auxiliam pesquisas com dados genéticos: o GENES (CRUZ, 2013). Este software foi desenvolvido na Universidade Federal de Viçosa e auxilia estudos genéticos aplicados ao melhoramento vegetal e animal. O mesmo é composto por sete módulos: Estatística Experimental, Biometria, Análise Multivariada, Diversidade Genética, Simulação, Matrizes e Integração. Cada módulo possui diversas funções relativas à sua área. Dentro do módulo Diversidade Genética é encontrada semelhanças ao software Cluster, como a análise de grupos e otimização; porém outros métodos e técnicas são aplicadas.

Dentro os softwares citados acima, aquele que serviu como base para o software proposto é o STRUCTURE, por isto destina-se uma seção para descrição mais detalhada do mesmo.

2.2.1 Software STRUCTURE

O STRUCTURE é um software gratuito para análise genética de populações desenvolvido no *Pritchard lab*, pertencente ao *Department of Genetics and Biology*, *Stanford University*. Seu algoritmo básico foi descrito por Pritchard, Stephens e Donnelly (2000) e possui atualizações e extensões do método publicado por Falush, Stephens e Pritchard (2003; 2007) e Hubisz, Falush, Stephens e Pritchard (2009). O software é um dos mais utilizados por geneticistas para avaliar o nível de diversidade genética (EARL e VONHOLDT, 2012).

O STRUCTURE utiliza um algoritmo iterativo de inferência Bayesiana para agrupar amostras que compartilham padrões similares dentro de suas variáveis genéticas. O fato do software ter como princípio inferência Bayesiana, faz com que

seja necessária informação prévia da classificação do conjunto de amostras para auxiliar na obtenção de melhores resultados (PORRAS-HURTADO, RUIZ, *et al.*, 2013). O domínio de uma população será definido pelo pesquisador, podendo ser um fenótipo, comportamento dos indivíduos amostrados, ou até mesmo características linguísticas e culturais ao se pensar em populações formadas por seres humanos.

A estimação de valores para o agrupamento formado pelo STRUCTURE segue o método de Monte Carlo via Cadeias de Markov – MCMC (GAMERMAN e LOPES, 2006) . O algoritmo de Monte Carlo via Cadeias de Markov distribui, inicialmente, os indivíduos de forma aleatória em um número pré-determinado de grupos. A cada iteração, calculam-se as estimativas de frequência das amostras para cada grupo, de modo a convergir para a separação das amostras nos grupos. Muitas vezes, o número de iterações pode chegar a de 100.000 (PORRAS-HURTADO, RUIZ, *et al.*, 2013). Para aumentar a confiabilidade dos resultados, normalmente, aumenta-se o número de iterações, com o cuidado de não exagerar, pois valores demasiadamente grandes são inúteis pela involução do algoritmo. Considera-se como involução a falta de progresso do algoritmo. Evanno, Regnaut e Goudet (2005) concluíram em testes que 10.000 iterações era suficiente para seu trabalho, ou seja, não é possível generalizar um bom número de iterações. O número de iterações varia conforme a base de dados utilizada.

Porras-Hurtado, Ruiz, *et al*. (2013) lembram que embora o número de grupos, no caso populações, seja um parâmetro pré-selecionado pelo usuário do software, este valor não é facilmente definido. Evanno, Regnaut e Goudet (2005) salientaram que pouco se sabe sobre a capacidade do STRUCTURE para detectar o número real de populações que compõe uma base de dados. A preocupação em determinar o número de populações acompanha o desenvolvimento do STRUCTURE desde sua proposta inicial por Pritchard, Stephens e Donnelly (2000), com maior foco e discussão por Hubisz, Falush, *et al.* (2009).

Dentre os modos desenvolvidos para determinação do número de populações, destaca-se o estudo por simulações de Evanno, Regnaut e Goudet (2005) e o software CLUMPP (JAKOBSSON e ROSENBERG, 2007). O método desenvolvido por Evanno, Regnaut e Goudet (2005) foi implementado por Earl e vonHoldt (2012) e está disponível *on line* para uso gratuito, conhecido como STRUCTURE HARVESTER. No entanto, para se valer deste método é necessário fazer diversas simulações no software STRUCTURE com variação do número de populações. Os arquivos gerados após as simulações são entregues ao STRUCTURE HARVESTER que avalia o real

número de populações. O software CLUMPP também depende de diversas simulações com variação do número de populações dentro do STRUCTURE. Esta dependência torna o processo complicado e demorado, especialmente quando se trata com busca em mais de 10 possíveis números de populações (PORRAS-HURTADO, RUIZ, *et al.*, 2013).

2.3 Marcadores Genéticos

Os tipos de dados trabalhados pelos softwares citados acima são na verdade marcadores genéticos. Um marcador genético é qualquer característica hereditária que pode ser empregada para avaliar diferenças genéticas entre indivíduos. Estes marcadores podem ser divididos em dois grupos básicos, marcadores morfológicos e marcadores moleculares (BERED, NETO e DE CARVALHO, 1997).

Os marcadores morfológicos são aqueles provenientes de observações de fenótipos de fácil identificação visual (FERREIRA e GRATTAPAGLIA, 1998). Estes são utilizados desde as pesquisas de Mendel. Ele utilizou marcadores como a altura da planta, a cor da flor e a textura da semente em seu estudo sobre de ervilhas (SNUSTAD e SIMMONS, 2013).

Os marcadores morfológicos são susceptíveis a mudanças perante ao meio em que as amostras vivem. Sendo assim, para originar bons resultados, os estudos com marcadores morfológicos, muitas vezes, são restritos a espécies controladas em laboratório (ZOLET, SEGATTO, *et al.*, 2013). Ferreira e Grattapaglia (1998) salientam que só ocasionalmente encontram-se marcadores morfológicos de importância econômica para programas de melhoramento genético.

Desde modo, a maior parte dos softwares atuais que tratam de diversidade genética utilizam bases de dados compostas por marcadores moleculares. Segundo Milach (1998), marcadores moleculares são sequências de DNA herdadas geneticamente que diferenciam dois ou mais indivíduos. Estas sequências são capazes de detectar, diretamente do DNA, diversas formas de especialidades desenvolvidas pelos indivíduos.

Os marcadores moleculares são empregados na identificação de clones, linhagens, híbridos, culturas, paternidade, estimativas de diversidade, fluxo gênico, taxa de cruzamento, parentesco e na construção de mapas genéticos, além de outras

diversas finalidades (BUSO, CIAMPI, *et al.*, 2003). Dentre os marcadores mais utilizados em estudos de diversidade genética, destacam-se:

- RFLP – *Restriction Fragment Length Polymorphism* (GRODZICKER, WILLIANS, *et al.*, 1974)
- ALFP – *Amplified Fragment Length Polymorphism* (ZABEAU e VOS, 1993)
- RAPD – *Random Amplified Polymorphic DNA* (WILLIAMS, KUBELIK, *et al.*, 1990; WELSH e MCCLELLAND, 1990)
- Microssatélites SSR – *Simple Sequence Repeats* (LITT e LUTY, 1989)

Entre as vantagens dos marcadores moleculares pode-se citar a não influencia do ambiente, o que propicia maior credibilidade as pesquisas. Alguns marcadores morfológicos são dependentes de amostras de indivíduos inteiros e adultos, enquanto marcadores moleculares são capazes de caracterizar o genótipo de um indivíduo a partir apenas de amostras de células ou de tecidos (FERREIRA e GRATTAPAGLIA, 1998). Porém, trabalhar com marcadores moleculares necessita-se de técnicas e equipamentos mais complexos.

Capítulo 3: Metodologia

Neste capítulo são discutidas as técnicas e algoritmos utilizados na construção do software Cluster. Inicialmente, apresenta-se uma visão geral da estrutura do software. Segue-se discussão sobre técnicas de imputação de dados para base de dados incompletas. Posteriormente descreve-se técnicas para seleção de características mais relevantes ao agrupamento de amostras. Discute-se, em seguida, algoritmos para *clustering* e apresenta-se o algoritmo genético proposto para otimização. Por fim, expõe-se os parâmetros de desempenho considerados neste trabalho.

3.1 Visão geral

A estrutura do software Cluster pode, resumidamente, ser representada no diagrama em blocos da Figura 3.1.

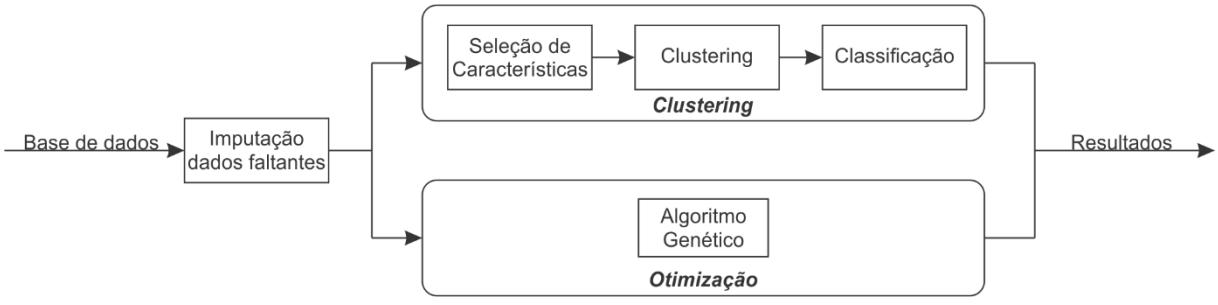


Figura 3.1: Diagrama em Blocos

O primeiro passo desempenhado pelo software Cluster ao receber uma base dados é a imputação de dados faltantes. Isto se faz necessário pois rotineiramente encontrar-se base de dados com amostras incompletas em estudos biológicos.

Em seguida, tem-se as duas operações básicas do software Cluster: Clustering e Otimização.

A Operação de Clustering é dividida em três partes:

- Seleção de características, executada pelo método ReliefF (KONONENKO, 1994);
- *Clustering*, propriamente dito, executado pelo algoritmo *Fuzzy* C-Means (DUNN, 1974; BEZDEK, 1981) e;
- Classificação, desempenhada por um classificador com o intuito de validar por meio da acurácia de classificação, a boa seleção de características.

A Operação de Otimização é desempenhada por meio de um algoritmo genético (HOLLAND, 1975). Este foi desenvolvido para otimizar a Operação de Clustering desempenhada pelo próprio software Cluster.

O modo de apresentação de resultados varia conforme a operação desempenhada pelo software. Em comum as duas operações, tem-se a acurácia de classificação como resultado. Junto a acurácia, o software calcula a sensibilidade, especificidade e eficiência para a Operação de Clustering. Busca-se, assim, melhor validação das caraterísticas selecionadas.

Nas seções a seguir, cada um dos blocos da Figura 3.1 é discutido assim como a técnica ou método escolhido para tal.

Para melhor entendimento deste capítulo, as bases de dados trabalhadas são consideradas como uma matriz $X \in \mathbb{R}^{M \times N}$ onde M é o número de amostras e N é o número de características de cada amostra. As amostras $\{x_m\}$ são agrupadas em classificadas G grupos a fim de compor as classes do problema. Chama-se de dado, um elemento desta matriz.

A Figura 3.2 representa uma base de dados genérica.

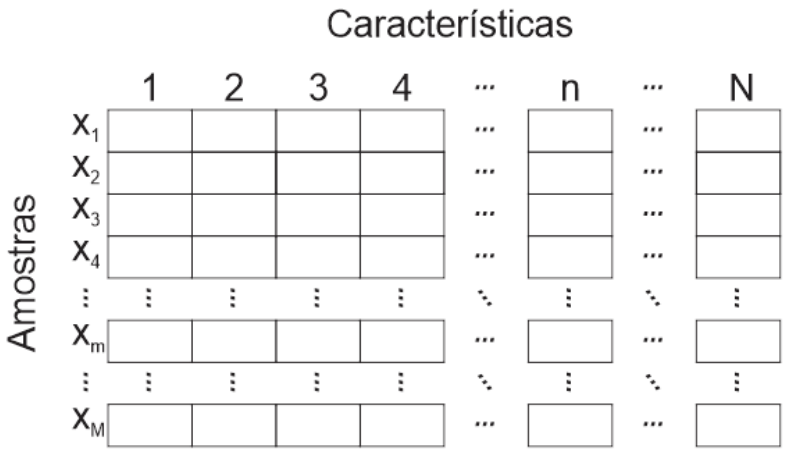


Figura 3.2: Modelo de base de dados

3.2 Imputação de dados faltantes

Um problema rotineiro ao se trabalhar com o reconhecimento de padrões em bases de dados com grande volume de dados é encontrar bases compostas por amostras incompletas (RUBIN, 1976; RUBIN, 1987; RUBIN, 1996; ACID, CAMPOS e HUETE, 2001; CHIU, CHAN, *et al.*, 2013). Na área da genética, devido às dificuldades de aquisição de material na natureza e métodos laboratoriais complexos para genotipagem, a criação de base de dados muitas vezes é condicionada a amostras com dados faltantes, ou incompletas. No contexto da matriz X, significa que uma amostra x_m pode não conter todas as N características do problema.

Uma solução simples para resolver este problema é repetir os experimentos. Entretanto, é fácil perceber que esta solução é cara e ineficiente. Quando se trabalha com material genético obtido de animais silvestres, por exemplo, esta solução é praticamente inviável. Outra solução é o descarte das amostras com dados faltantes. Esta solução pode implicar na perda de parte importante da informação contida na base de dados. Assim, as técnicas para imputação de dados faltantes foram desenvolvidas para imputar de forma precisa os dados faltantes (CHIU, CHAN, *et al.*, 2013).

A determinação da técnica para imputação de dados faltantes é uma questão delicada, pois a escolha de um método inapropriado pode resultar em conclusões errôneas (CHIU, CHAN, *et al.*, 2013). Basicamente, existem duas modalidades de técnicas para a imputação de dados faltantes; a imputação simples e a imputação múltipla.

As técnicas de imputação simples foram as primeiras técnicas estatísticas para se completar bases de dados incompletas (RUBIN, 1987). Estas técnicas são relativamente simples pois consistem, essencialmente, em completar os dados pela média, pela mediana, por interpolação, por regressão linear ou até mesmo por dados aleatórios dentro de um intervalo proposto (ACID, CAMPOS e HUETE, 2001). Estas técnicas são caracterizadas por preencher o dado faltante uma única vez, não modificando a base de dados novamente. A base de dados passa a ser analisada como se não houvesse dados faltantes (MCKNIGHT, MCKNIGHT, *et al.*, 2007). Por este motivo, estas técnicas são também conhecidas como técnicas de imputação única.

As primeiras técnicas de imputação múltipla foram propostas por Rubin (1987) e se diferem da imputação simples, ou imputação única, por substituir cada valor faltante por dois ou mais valores distintos. Desta forma, várias bases de dados são geradas para que possam ser analisadas do modo convencional. Existem diversos modos de ser realizar a imputação múltipla. Basicamente, observam-se os dados não faltantes e tenta-se completar a base preservando a conformidade da distribuição dos dados (MCKNIGHT, MCKNIGHT, *et al.*, 2007). Comumente, a distribuição utilizada é a normal (ALLISON, 2001), mesmo quando alguns dados não possuem tal distribuição (SCHAFER, 1997).

Métodos para selecionar a melhor técnica de imputação de dados é um estudo amplo e complexo presente em larga literatura (RUBIN, 1996; CHIU, CHAN, *et al.*, 2013). A escolha da técnica de imputação não é o foco deste trabalho. Devido às características das bases de dados utilizadas nesta dissertação, descritas na seção 4.2, optou-se por utilizar uma técnica de imputação simples. Os dados faltantes são completados pela média dos dados não faltantes de amostras da mesma classe da amostra incompleta. Vale lembrar que a classificação prévia das amostras é fornecida pelo pesquisador. Apesar de simples, esta técnica originou bons resultados sem prejudicar o desempenho do software Cluster. Obteve-se boa eficiência na classificação das amostras, não sendo necessário recorrer à utilização de uma técnica de imputação múltipla.

3.3 Seleção de características

A seleção de características é parte primordial deste trabalho. Muitas bases de dados possuem centenas ou até milhares de características a serem estudadas. Destas características, apenas algumas podem ser necessárias para distinguir as classes em estudo, as demais podem, inclusive, atrapalhar a pesquisa. Sendo assim, a redução de dimensionalidade do problema é uma das principais questões associadas ao reconhecimento de padrões (THEODORIDIS e KOUTROUMBAS, 2009).

Se selecionadas características com pouco poder de discriminação das classes, por consequência o classificador poderá ter mau desempenho. Por outro lado, ao selecionar características ricas em informação, o projeto do classificador será simplificado (THEODORIDIS e KOUTROUMBAS, 2009).

Reduzir o número de características trabalhadas traz outra vantagem: a diminuição da complexidade computacional. Grupos compostos por características com correlação mútua alta tendem a aumentar a complexidade computacional sem ganhos efetivos para o classificador, por não ajudar a distinguir as classes (THEODORIDIS e KOUTROUMBAS, 2009). Com menos características sendo utilizadas, menor será o recurso necessário para armazenamento, além de implicar em tempos menores de execução dos algoritmos (GUYON e ELISSEEFF, 2003).

Existem duas formas distintas para se analisar uma base de dados de modo a efetuar a seleção das características. Estas formas são conhecidas como análise univariada e análise multivariada. A diferença entre as duas análises é que a análise multivariada leva em consideração que as características podem estar relacionadas entre si, de forma que uma interfere em outra no processo de classificação. Já a análise univariada aborda individualmente cada uma das características, sem levar em consideração qualquer relação entre elas. A análise multivariada costuma produzir melhores resultados, visto que se assemelha mais da realidade, porém possui maior custo computacional (GUYON e ELISSEEFF, 2003; THEODORIDIS e KOUTROUMBAS, 2009).

3.3.1 Relief e ReliefF

Um método difundido para a seleção de características é o Relief (KIRA e RENDELL, 1992). A ideia principal deste algoritmo é a estimação de quanto uma

determinada característica ajuda na distinção de classes levando em consideração amostras vizinhas (KONONENKO, 1994). Este método possui análise multivariada e é baseado no algoritmo KNN (*K-Nearest Neighbors)* – K-Vizinhos Mais Próximos (GUYON e ELISSEEFF, 2006).

O algoritmo do método Relief calcula, dada uma base de dados formada por M amostras de N características, um coeficiente Cr para cada uma das características. Quanto maior o valor do coeficiente Cr, maior é o poder de discriminação das classes pela característica em análise.

O cálculo do coeficiente Cr para cada característica é dado pelo somatório das distâncias entre cada uma das amostras $\{x_m\}$ e K amostras vizinhas de classes diferentes $\{x_{P_k(m)}\}$ dividido pelo somatório das distâncias entre cada uma das amostras $\{x_m\}$ e K amostras vizinhas de mesma classe $\{x_{Q_k(m)}\}$. São consideradas sempre as amostras vizinhas mais próximas. Para melhor entendimento, apresenta-se a seguir a equação do cálculo do coeficiente Cr extraído de GUYON e ELISSEEFF (2006):

$$Cr_{(n)} = \frac{\sum_{m=1}^{M}\sum_{k=1}^{K}\left|x_{m,n} - x_{P_k(m),n}\right|}{\sum_{m=1}^{M}\sum_{k=1}^{K}\left|x_{m,n} - x_{Q_k(m),n}\right|}, \qquad para\ n = 1,2,3,\dots,N\ . \tag{3.1}$$

Pela equação, percebe-se melhor que, para se conseguir maiores coeficientes Cr, a característica em estudo deverá conseguir aumentar a distância entre amostras de classes distintas e diminuir a distância entre amostras de mesma classe.

Diversas variações do método Relief original são encontradas na literatura. Muitos buscam a otimização na seleção de características. Dentre os métodos propostos destacam-se: ReliefF (KONONENKO, 1994), RRelieff (ROBNIK-SIKONJA e KONONEKO, 1997) e IRelief (SUN e JIAN, 2006; SUN, 2007). Neste trabalho, utiliza-se o ReliefF.

O Relief é um método limitado a trabalhar com problemas de apenas duas classes e apresenta dificuldades em lidar com dados incompletos. Em contrapartida, o ReliefF não possui esta limitação além de ser mais robusto e lidar melhor com dados incompletos ou ruidosos em comparação com o simples Relief (KONONENKO, 1994). Enquanto no numerador do cálculo do coeficiente Cr pelo Relief calcula a distância para K amostras vizinhas da classe distinta da amostra em questão; o ReliefF calcula

a distância de K amostras vizinhas para cada classe distinta da amostra em questão. Lembrando que para característica analisada varre-se todas as amostras.

Para a maioria dos casos, o número de amostras vizinhas, K, utilizado para o cálculo do ReliefF pode seguramente ser ajustado com o valor 10 (ROBNIK-SIKONJA e KONONENKO, 2003). Em busca de melhores resultados, no software Cluster existe a opção de buscar o melhor valor para K utilizado pelo ReliefF. Detalhes da otimização são discutido nas próximas seções.

3.4 Clustering

Clustering é uma técnica que consiste em particionar um conjunto de amostras em subconjuntos de acordo com as diferenças entre elas. Pode se dizer também que é o processo de agrupamento de amostras semelhantes (LARRAÑAGA, CALVO, *et al.*, 2006).

Algoritmos de *clustering*, ou algoritmos de agrupamento, podem ser estruturados de diversos modos para produzir saídas distintas. De acordo com a saída produzida, os algoritmos de clustering podem ser classificados como *Hard* ou *Fuzzy*. A saída de um algoritmo *Hard* é uma matriz com cada amostra pertencendo apenas a um grupo, dentre todos os grupos formados. Já a saída de um algoritmo *Fuzzy* é uma matriz de pertinência de cada amostra para cada um dos grupos formados (JAIN, MURTY e FLYNN, 1999).

Levando em consideração que este trabalho envolve dados biológicos, onde muitas vezes uma amostra pode pertencer a mais de uma classe, decidiu-se trabalhar com algoritmos *Fuzzy* para o processo de *clustering*. Segundo Jain, Murty e Flynn (1999), o algoritmo de agrupamento *Fuzzy* mais popular é algoritmo nebuloso *Fuzzy* C-Means (DUNN, 1974; BEZDEK, 1981). Por este motivo, o trabalho presente lida com este algoritmo. O mesmo é descrito a seguir.

De forma genérica, um algoritmo de agrupamento *Fuzzy* aplicado a uma base de dados com M amostras $X = \{x_1, x_2, \dots, x_{\mathrm{m}}, \dots, x_M\}$ produz uma matriz de pertinência $U_{G\times M}$, onde G é o número de grupos. Cada elemento u_{gm} da matriz U representa o grau de pertinência da m-ésima amostra ao g-ésimo grupo *Fuzzy*.

Geralmente, os algoritmos de agrupamento *Fuzzy* procuram minimizar a dissimilaridade das amostras dentro dos grupos através da função:

$$J = \sum_{m=1}^{M} \sum_{g=1}^{G} u_{gm}^{F} D_{gm} , \tag{3.2}$$

onde: J é a medida de dissimilaridade dentro dos grupos;

$F \in (1, \infty)$ é um expoente de *fuzzificação* que determina a influência das pertinências no agrupamento, e;

D_{gm} é a distância entre a m-ésima amostra e o centro do g-ésimo grupo.

Entretanto, a otimização de J possui duas restrições, expostas nas equações a seguir:

$$\sum_{g=1}^{G} u_{gm} = 1, \qquad 1 \leq m \leq M . \tag{3.3}$$

$$0 < \sum_{m=1}^{M} u_{gm} < 1, \qquad 1 \leq g \leq G . \tag{3.4}$$

A equação (3.3) indica que a soma dos graus de pertinências associados a uma amostra deve ser igual a 1. Ressalta-se que $u_{gm} \in [0,1]$. Já a equação (3.4), em associação a equação (3.3), adverte que cada grupo deve conter pelo menos uma amostra com grau de pertinência maior que zero e não conter todas as amostras com grau de pertinência unitário.

Por se tratarem de métodos iterativos, os algoritmos *Fuzzy* atualizam a matriz de pertinência U a cada iteração. O método mais popular de se atualizar esta matriz segue a fórmula apresentada por (BABUSKA, 2000), adaptada aqui para melhor entendimento:

$$u_{gm} = \frac{1}{\sum_{l=1}^{G} \left(\frac{D_{gm}}{D_{lm}} \right)^{\frac{2}{F-1}}}, \qquad 1 \leq g \leq G, \ 1 \leq m \leq M . \tag{3.5}$$

De forma geral, a distância D_{gm} é calculada pela seguinte equação:

$$D_{gm} = \left\| x_m - c_g \right\|_A^2 = \left(x_m - c_g \right)^T A \left(x_m - c_g \right), \quad (3.6)$$

ou seja, D_{gm} é o quadrado de uma norma calculada a partir de um produto interno com a matriz A. O termo A é uma matriz definida positiva e quadrada $[M \times M]$ que define o formato dos grupos as serem formados (VEDRAMIN, 2012). Sendo assim, diferentes modelos de algoritmos Fuzzy podem ser formados pela variação na forma de se medir a distância D_{gm} com alteração da matriz A.

3.4.1 Algoritmo Fuzzy C-Means

O algoritmo *Fuzzy* C-Means foi apresentado inicialmente por DUNN (1974) e posteriormente aperfeiçoado por BEZDEK (1981). Este algoritmo é "utilizado para determinar agrupamentos e seus centros segundo a norma Euclidiana existente entre uma amostra e os centros dos agrupamentos. O raciocínio para entender a relação entre uma amostra e o agrupamento é o seguinte: quanto mais próximo do centro de um agrupamento a amostra estiver, maior será seu grau de pertinência a esse agrupamento" (GUIERA, CENTENO, *et al.*, 2005).

Para se conseguir a distância Euclidiana a partir da equação (3.6), base do algoritmo *Fuzzy* C-Means, basta adotar a matriz A como uma matriz identidade $I_{M \times M}$. O fato de utilizar a distância Euclidiana faz com que este algoritmo apresente grupos hiperesféricos (VEDRAMIN, 2012).

O algoritmo atualiza, a cada iteração, os centros de seus grupos pela equação:

$$c_g = \frac{\sum_{m=1}^{M} u_{gm}^F \, x_m}{\sum_{m=1}^{M} u_{gm}^F}, \qquad 1 \leq g \leq G\,. \quad (3.7)$$

O algoritmo *Fuzzy* C-Means pode ser interpretado da seguinte forma:

Algoritmo 3.1: *Fuzzy* C-means

Requer: Base de dados $X = \{x_1, x_2, \dots, x_M\}$, número de grupos $G \in \{2, 3 \dots, M-1\}$, expoente de *fuzzificação* $F \in (1, \infty)$ e critério de convergência

1: Crie G centros aleatórios
2: it ← 0
3: Repita
4: Calcule o centro de cada grupo através da equação (3.7)
5: Calcule as distâncias entre as amostras e os centros de cada grupo através da equação (3.6)
6: Atualize a matriz de partição através da equação (3.5)
7: it ← it +1
8: Até atingir o critério de convergência
9: Retorna $U = [u_{gm}]_{G \times M}$ e $C = [c_g]_{G \times N}$

Para este trabalho, utilizou-se dois critérios de convergência: o atingir do número máximo de iterações igual a 200 e o valor absoluto da modificação dos centros menor que $1x10^{-10}$, caracterizando involução do algoritmo. Qualquer um dos critérios caracteriza convergência. O expoente de *fuzzificação* F escolhido neste trabalho foi igual a 2, dentro do intervalo indicado por Bezdek, Ehrlich e Full (1984) para geração de bons resultados. Como o número de grupos G é uma grandeza de estudo importante para pesquisadores na área da genética, muitas vezes representando o número de populações, esta variável pode ser facilmente modificada pelo usuário do software Cluster. Este valor pode, ainda, ser otimizado pelo próprio software Cluster através de um algoritmo genético discutido na seção 3.5 desta dissertação.

3.4.2 Outros algoritmos Fuzzy

Conforme já mencionado, o fato do algoritmo *Fuzzy* C-Means utilizar a distância Euclidiana faz com que este algoritmo apresente grupos hiperesféricos (VEDRAMIN, 2012), ou seja, hiperesferas são formadas em tornos dos centros calculados. Isto faz com que a pertinência de cada amostra para cada um dos grupos seja proporcional à distância Euclidiana entre a amostra e os centros dos grupos.

Por apresentar grupos com centros iniciais aleatórios, ou seja, que variam conforme a cada execução, o algoritmo *Fuzzy* C-Means apresenta variação em sua saída. Em consequência, existe variação na acurácia de classificação das amostras a

cada execução. Acurácia é a taxa de acertos de classificação calculada pela quantidade de amostras classificadas corretamente pelo total do número de amostras da base de dados em análise.

Para garantir melhores resultados de acurácia, o algoritmo *Fuzzy* C-Means deve ser executado diversas vezes. É considerado somente o melhor resultado obtido, visto que uma vez encontrado os melhores centros dos grupos, estes serão utilizados pelo classificador para validação das características selecionadas.

Ressalta-se que outros dois algoritmos *Fuzzy* foram testados: Gustafson-Kessel (GUSTAFSON e KESSEL, 1978) e Gath-Geva (GATH e GEVA, 1989). Ainda que possuem centros iniciais aleatórios, ambos diferem do algoritmo *Fuzzy* C-Means pelo modo de se calcular a distância entre os centros dos grupos e as amostras. Eles adicionam matrizes de covariância no cálculo de distâncias produzindo assim regiões hiperelípticas. Os testes mostraram que estes dois algoritmos conseguiram acabar com a variação de resultados apresentada pelo *Fuzzy* C-Means, porém os resultados de acurácia dos classificadores foram reduzidos. Por exemplo, para a base de dados Turtles, descrita na seção 5.1.1, estes algoritmos produziram uma acurácia de 97,34% enquanto o algoritmo *Fuzzy* C-Means apresentou acurácia de 99,47%. Esta característica, menor acurácia de classificação, torna desinteressante trabalhar com esses algoritmos. Logo, decidiu-se continuar a utilizar apenas o algoritmo *Fuzzy* C-Means, principalmente, pela sua robustez ao apresentar bons resultados com diversas bases de dados.

3.4.3 Classificador

Conforme explicado, o algoritmo *Fuzzy* C-Means retorna apenas as coordenadas dos centros dos grupos formados e a matriz de pertinência das amostras para cada um destes grupos. Muitas vezes, o número de grupos não representa exatamente o número de classes do problema. Sendo assim, é necessário determinar a qual classe cada grupo de amostras pertence. Para isto, o software Cluster necessita de uma classificação prévia, já fornecida pelo pesquisador.

Para determinar a classe de cada grupo, cada amostra é designada para o grupo a que possui maior pertinência. Dentro deste grupo, a classe que possuir maior número de indivíduos determinará a classificação do mesmo. É levada em consideração a classificação prévia, ou real, já fornecida pelo pesquisador.

Ressalta-se que a finalidade do software proposto é melhorar parâmetros de *clustering*, sendo a classificação apenas ponto de análise de desempenho. Selecionar as melhores características para distinção de classes é o objetivo central deste trabalho.

3.5 Algoritmo genético

Os algoritmos primeiros genéticos foram propostos por John Holland (HOLLAND, 1975) na década de 70 e aperfeiçoado por Holland e seus alunos na Universidade de Michigan na década subsequente (MITCHELL, 2002). Estes algoritmos são técnicas que simulam o processo de evolução natural. Denominadas gerações, suas iterações geram através de processos de seleção, elitismo e operações de cruzamento e mutação, novos indivíduos mais adaptados ao meio (SRINIVAS e PATNAIK, 1994). Considera-se adaptado ao meio, indivíduos que produzem otimização da função objetivo em questão. Indivíduos são formados por variáveis do problema que influenciam a função objetivo.

Em outras palavras, dadas variáveis aleatórias iniciais, estas vão sendo combinadas de forma a produzir a otimização da função objetivo. As combinações são baseadas em fenômenos evolucionários da natureza (LINDEN, 2008).

Segundo Goldberg (1989), um dos alunos de Holland, os algoritmos genéticos são distintos dos algoritmos tradicionais de otimização em quatro aspectos:

- Algoritmos genéticos trabalham com parâmetros codificados;
- Algoritmos genéticos realizam a otimização a partir de uma população de pontos, e não de um ponto único;
- Algoritmos genéticos utilizam o cálculo da função objetivo para evolução de otimização, não necessitando do uso de derivadas ou outros conhecimentos auxiliares, e;
- Algoritmos genéticos utilizam regras probabilísticas e não regras determinísticas.

O algoritmo genético presente neste trabalho visa encontrar o número de características selecionadas, o número de grupos formados e o número de amostras vizinhas consideras pelo ReliefF para otimização da acurácia do classificador (função objetivo). O usuário é capaz de selecionar quais destas variáveis formarão o espaço de busca para otimização, conforme sua pesquisa.

Para melhor exemplificar, o algoritmo genético implementado neste trabalho é apresentado na Figura 3.3 e analisado em seguida.

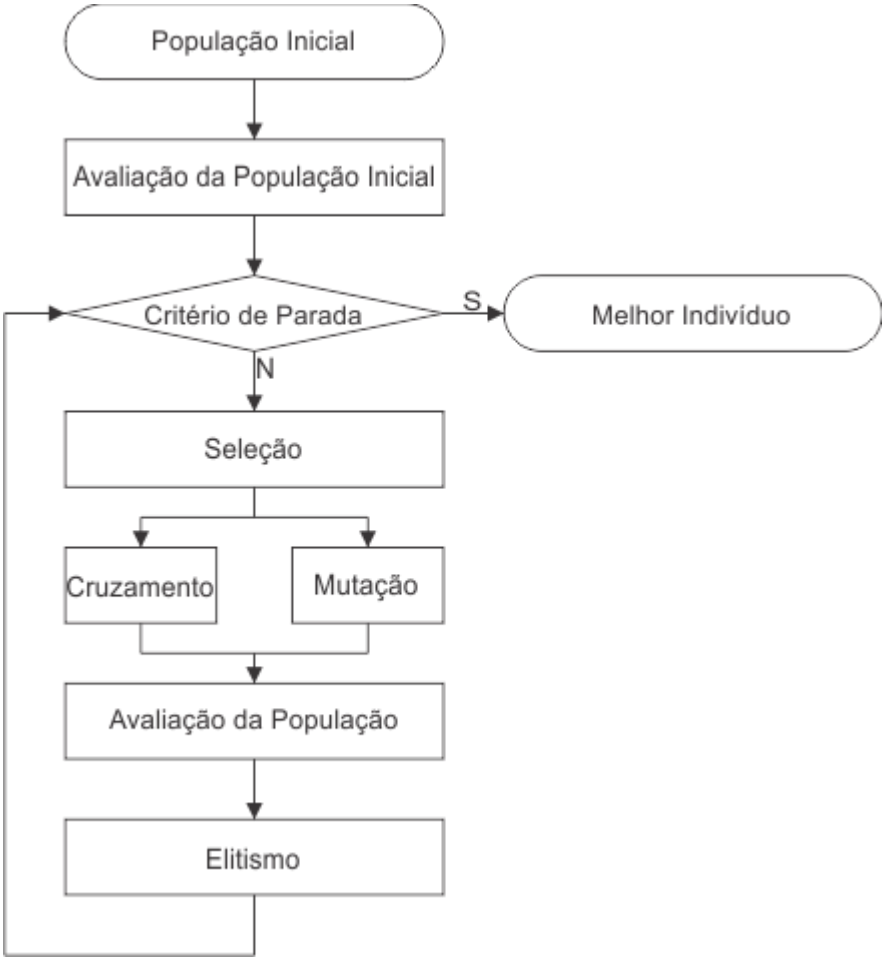


Figura 3.3: Algoritmo Genético – Fluxograma

Uma população inicial aleatória é gerada com base nas variáveis a serem trabalhadas: o número de características selecionadas, C; o número de grupos formados, G; e a quantidade de amostras vizinhas consideradas pelo ReliefF, K; conforme decisão do pesquisador. Estas variáveis formam o que se denomina neste trabalho por espaço de busca. Avalia-se cada um dos indivíduos desta população, constituídos pelas variáveis do espaço de busca. A avaliação é feita utilizando a seleção de características pela técnica do ReliefF aliada a execução do algoritmo *Fuzzy* C-Means. Devido à aleatoriedade de resultados do algoritmo *Fuzzy* C-Means este é executado 10 vezes para cada indivíduo, sendo considerado apenas o melhor resultado destas 10 execuções. O intuito desta avaliação é encontrar aquele indivíduo

que produz melhor acurácia de classificação pela combinação das variáveis do espaço de busca.

Após a avaliação da população inicial entra-se no loop de gerações. Este é encerrado quando atingido um dos três critérios de parada. O primeiro deles é estipulado pelo número de gerações, ou iterações, fornecido pelo pesquisador. Em concomitância, tem-se um segundo critério de parada; o atingir 100% de acurácia pelo classificador. Isto caracteriza involução do algoritmo, logo desnecessário continuar as iterações. O pesquisador pode, ainda, optar que o algoritmo genético encerre sua atividade ao não conseguir melhoria de resultados após seguidamente alcançar 30% do número total de gerações estipuladas. Este é, também, um modo que caracteriza involução do algoritmo, sendo o terceiro critério de parada. Assim, consegue-se economizar esforço computacional e tempo para entrega de resultados.

O processo de evolução, dentro do loop de gerações, começa com a seleção de indivíduos pelo método torneio. No método torneio, dentro de um par de indivíduos selecionados aleatoriamente, apenas aquele que produziu maior acurácia segue para a etapa de cruzamento ou mutação. Todos os indivíduos passam por este processo de seleção.

A cada geração a população é dividida aleatoriamente de forma que 85% dos indivíduos sofram a operação de cruzamento e 15% sofram a operação de mutação. Na operação de cruzamento, o indivíduo selecionado tem parte de suas variáveis trocadas por variáveis de outros indivíduos selecionados aleatoriamente. Na operação de mutação seleciona-se três indivíduos distintos e cria-se um único indivíduo a partir deles. Este indivíduo é formado pelo resultado da soma das variáveis de um dos três indivíduos com o produto do fator de mutação pelo valor absoluto da diferença das variáveis dos outros dois indivíduos. O fator de mutação utilizado foi de 0.85.

Para melhor entendimento, a fórmula da operação de mutação para gerar o novo indivíduo i_n é exemplificada abaixo, onde i_a, i_b e i_c representam os três indivíduos distintos e fm o fator de mutação. Lembra-se que os indivíduos são vetores das variáveis que formam o espaço de busca para a otimização do problema.

$$i_n = i_a + fm \,.\, |i_b - \, i_c| \,. \tag{3.7}$$

Criada esta nova população após as operações de cruzamento e mutação, a mesma é avaliada conforme feito com a população inicial. Compara-se a população

formada com a população da iteração anterior, ou com a população inicial no caso da primeira iteração. Desta comparação, permanece na nova população apenas os indivíduos que produziram melhores acurácias. Este processo denomina-se elitismo.

Repete-se o processo de seleção, cruzamento e mutação, avalição e elitismo de modo a evoluir o algoritmo de otimização até que se atinja um dos critérios de parada. Atingido um destes, informa-se ao pesquisador a melhor acurácia obtida e o melhor indivíduo. Será o melhor indivíduo, a melhor combinação entre o número de características selecionadas, o número de grupos formados e o número de amostras vizinhas consideradas pelo ReliefF que produziu melhor acurácia na classificação.

3.6 Parâmetros de desempenho

Conforme dito anteriormente, o software Cluster utiliza a acurácia do classificador como função objetivo para otimização. Porém, este não é o único parâmetro de desempenho para avaliar o software Cluster. Ressalta-se, mais uma vez, que o objetivo do software não é apresentar o melhor classificador, mas sim, um selecionador de características ótimo para distinção das classes em estudo.

Todos os parâmetros de desempenho utilizados nesta dissertação podem ser obtidos a partir da análise da matriz de confusão. Uma matriz de confusão mostra o número de classificações corretas atribuídas por um classificador comparando às classificações reais. Portanto, esta é uma matriz quadrada $Q \times Q$, onde Q é a quantidade de classes do problema em questão (PROVOST e KOHAVI, 1998).

Uma matriz de confusão genérica pode ser observada na Tabela 3.1.

Tabela 3.1: Matriz de confusão genérica

		Classificação atribuída (Software Cluster)			
		Classe 1	***Classe 2***	**...**	***Classe Q***
Classificação Real	***Classe 1***	$h_{1,1}$	$h_{1,2}$	...	$h_{1,Q}$
	Classe 2	$h_{2,1}$	$h_{2,2}$	...	$h_{2,Q}$
	...	...	...	...	...
	Classe Q	$h_{Q,1}$	$h_{Q,2}$	...	$h_{Q,Q}$

Nesta tabela, $h_{1,2}$ representa o número de amostras pertencentes a classe 1, segundo classificação prévia e real, atribuídas como da classe 2, segundo o software Cluster; e assim sucessivamente.

A acurácia é, simplesmente, a taxa de acertos do classificador, ou seja, a quantidade de amostras que o classificador conseguiu atribuir corretamente. Utilizando a nomenclatura da Tabela 3.1 pode-se calcular a acurácia do seguinte modo:

$$\text{Acurácia} = \frac{\sum_{i=1}^{Q} h_{i,i}}{\sum_{i=1}^{Q}\sum_{j=1}^{Q} h_{i,j}}. \tag{3.8}$$

Nesta dissertação, escolheu-se trabalhar com a acurácia pela facilidade de seu cálculo e, principalmente, por este ser um parâmetro de desempenho que avalia unicamente todas as classes envolvidas na base em estudo. Entretanto, o uso da acurácia pode mascarar maus resultados de classificação quando se trabalha com bases de dados desbalanceadas (SUN, KAMEL, *et al.*, 2007; CASTRO e BRAGA, 2011). Base de dados desbalanceadas são aquelas que possuem o número desigual de amostras por classes (CASTRO e BRAGA, 2011). Essas são comuns em estudos na área da genética.

Devido a esta particularidade, outros parâmetros de desempenho também podem ser analisados através do software Cluster. Nos testes desta dissertação, são analisadas a sensibilidade, a especificidade e a eficiência do classificador (FAWCETT, 2006). Diferentemente da acurácia, estes parâmetros avaliam separadamente cada uma das classes em estudo.

Dada uma classe em estudo, a sensibilidade é a capacidade do classificador em incluir, ou classificar, corretamente as amostras que pertencem a esta classe. Já a especificidade é a capacidade de excluir corretamente as amostras que não pertencem a esta classe. As equações para cálculo destes parâmetros são mostradas a seguir:

$$\text{Sensibilidade}_q = \frac{\mathrm{h}_{q,q}}{\sum_{j=1}^{Q} h_{q,j}}, \qquad 1 \le q \le Q; \tag{3.9}$$

$$\text{Especificidade}_q = \frac{\sum_{i=1}^{Q}\sum_{j=1}^{Q} h_{i,j} - \sum_{j=1}^{Q} h_{q,j} - \sum_{i=1}^{Q} h_{i,q} + \mathrm{h}_{q,q}}{\sum_{i=1}^{Q}\sum_{j=1}^{Q} h_{i,j} - \sum_{j=1}^{Q} h_{q,j}}, 1 \le q \le Q. \tag{3.10}$$

A eficiência é a média destes dois parâmetros de desempenho e pode ser expressa da seguinte forma:

$$\text{Eficiência}_q = \frac{\text{Sensibildiade}_q + \text{Especificidade}_q}{2}, \qquad 1 \leq q \leq Q. \tag{3.11}$$

Diversos outros parâmetros podem ser retirados de uma matriz de confusão, como o coeficiente de correlação de Matthews (BRAGA, 2000), *F-measure* e *G-mean* lembrados por Sun, Kamel, *et al.* (2007). Outra forma de avaliar o desempenho de um classificador é por meio da curva ROC, Receiver Operating Characteristic (BRAGA, 2000; FAWCETT, 2006; SUN, KAMEL, *et al.*, 2007). As curvas ROC são baseadas na sensibilidade e especificidade (BRAGA, 2000). A área abaixo da curva ROC (*Area Under the ROC Curve*, ou AUC) é outro parâmetro de desempenho de classificadores largamente utilizado na literatura (CASTRO e BRAGA, 2011). Porém, a análise de desempenho do classificador não é o objetivo deste trabalho.

Capítulo 4: Apresentação do Software Cluster

Este capítulo é dedicado à apresentação da funcionalidade do software Cluster. Inicialmente é exposta e descrita sua interface gráfica. O software desempenha duas funções básicas: Clustering e Otimização. Estas operações são discutidas de acordo com o modo de configuração e modo de apresentação de seus resultados. Toda a interface do Software Cluster foi criada com o intuito de facilitar a interação do pesquisador com a ferramenta proposta.

4.1 Descrição do software Cluster

O software proposto nesta dissertação, nomeado como Cluster, foi construído utilizando outros dois softwares: Matlab R2013a versão 8.1.0.604 e Microsoft Visual Studio 2010 versão 10.0.30319.1 RTMRel. O Matlab é responsável pelas funções que exigem cálculos mais complexos utilizados por Cluster, como; o método ReliefF, a etapa de *clustering* pelo algoritmo *Fuzzy* C-Means e as operações de mutação e cruzamento do algoritmo genético. Já o Visual Studio, configurado através da linguagem de programação Basic, é responsável pela interface gráfica e cálculos com menor esforço computacional, como; a estrutura para o algoritmo genético.

Conforme dito anteriormente, um dos objetivos do software Cluster é apresentar um *layout* amigável, de fácil entendimento para pesquisadores não familiarizados a trabalhar com matrizes, gráficos e índices matemáticos. Desta forma, a preocupação com o design e funcionalidade foi presente durante toda a sua concepção. Procurou-se desempenhar todo o trabalho em uma única tela com requisição de poucas variáveis de configuração. Sua interface ocupa uma tela com dimensões fixadas em 1200 x 635 pixels.

A interface gráfica do software Cluster é apresentado na Figura 4.1. Todas as figuras expostas neste capítulo foram adquiras a partir do software Cluster e referem-se ao estudo de uma base de dados sobre tartarugas marinhas híbridas da costa brasileira.

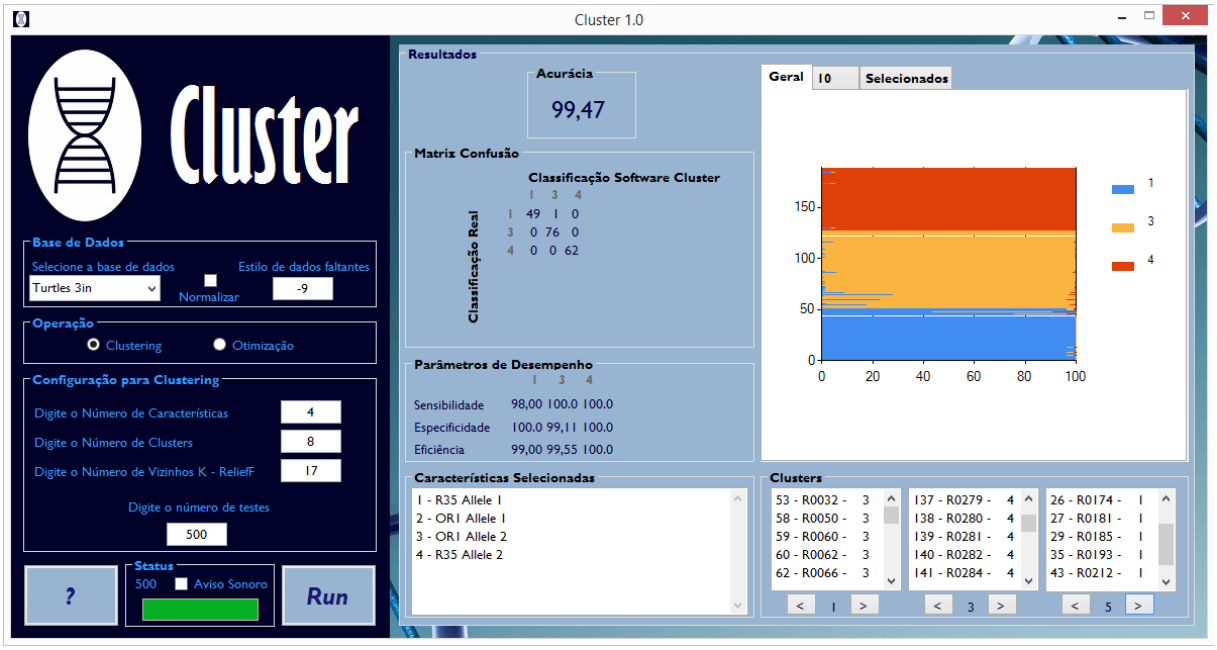


Figura 4.1: Software Cluster – Tela principal

O software Cluster foi estruturado de forma que ao lado esquerdo da tela, em azul escuro, se tenha o menu de configuração. Para configuração do software temos os quadros para: Seleção da base de dados a ser utilizada, Seleção da operação, Configuração da operação, e Verificação sobre *status*, ou progresso, da operação do software.

Os resultados das operações são mostrados sempre ao lado direito da tela, em azul claro. Estes estão divididos nos seguintes quadros: Acurácia, Matriz de confusão, Parâmetros de desempenho, Características selecionadas, Clusters e Gráficos.

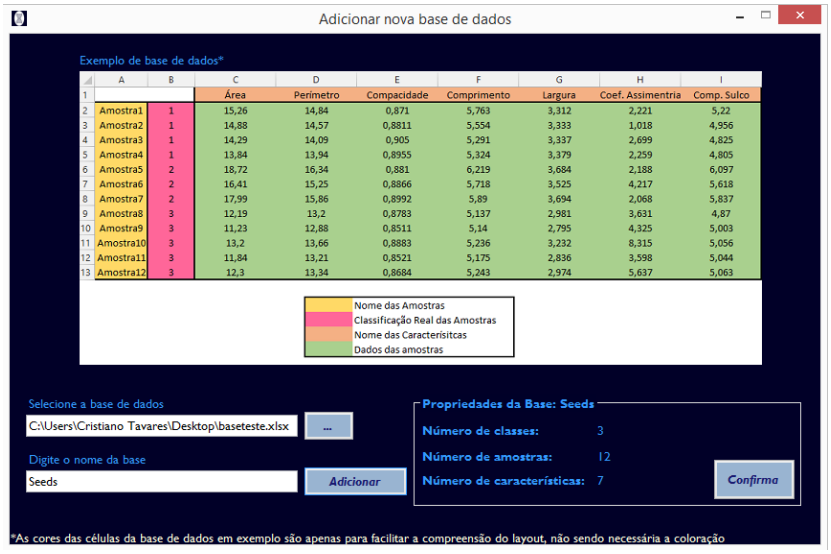
Cada um dos quadros da interface do software Cluster, para configuração e apresentação de resultados, é discutido a seguir.

4.2 Base de Dados

O software Cluster possui 5 bases de dados já ofertadas. Estas bases são distintas entre si pelo número de características, número de classes, número de amostras e tipo de dados que as compõem. Algumas destas bases são largamente utilizadas na literatura, enquanto outras proveem de publicações recentes. Todas elas serviram para a concepção e testes de validação do software. Detalhes destas bases de dados são apresentados no próximo capítulo, dedicado a tratar dos testes com o software Cluster.

O pesquisador, usuário do software Cluster, pode escolher trabalhar com qualquer uma destas bases de dados ou adicionar sua base de pesquisa ao mesmo. Ao escolher adicionar uma base de dados, uma tela secundária irá auxiliar o pesquisador a configurar sua base, para que esta possa ser utilizada no software Cluster. Isto se torna importante visto que cada software de diversidade genética possui uma forma diferente de configurar estas bases. É possível incluir novas bases de dados armazenadas em arquivos do software Microsoft Excel (extensões .xls e .xlsx) e do software Matlab (extensão .mat).

A tela para adicionar uma nova base de dados é exposta na Figura 4.2. Nela pode-se perceber o layout de configuração dos dados.



	A	B	C	D	E	F	G	H	I
1			Área	Perimetro	Compacidade	Comprimento	Largura	Coef. Assimentria	Comp. Sulco
2	Amostra1	1	15,26	14,84	0,871	5,763	3,312	2,221	5,22
3	Amostra2	1	14,88	14,57	0,8811	5,554	3,333	1,018	4,956
4	Amostra3	1	14,29	14,09	0,905	5,291	3,337	2,699	4,825
5	Amostra4	1	13,84	13,94	0,8955	5,324	3,379	2,259	4,805
6	Amostra5	2	18,72	16,34	0,881	6,219	3,684	2,188	6,097
7	Amostra6	2	16,41	15,25	0,8866	5,718	3,525	4,217	5,618
8	Amostra7	2	17,99	15,86	0,8992	5,89	3,694	2,068	5,837
9	Amostra8	3	12,19	13,2	0,8783	5,137	2,981	3,631	4,87
10	Amostra9	3	11,23	12,88	0,8511	5,14	2,795	4,325	5,003
11	Amostra10	3	13,2	13,66	0,8883	5,236	3,232	8,315	5,056
12	Amostra11	3	11,84	13,21	0,8521	5,175	2,836	3,598	5,044
13	Amostra12	3	12,3	13,34	0,8684	5,243	2,974	5,637	5,063

Figura 4.2: Tela para adicionar nova base de dados

O software Cluster suporta dados do tipo microssatélite, marcadores multialélicos ou qualquer tipo de dado que possua um valor numérico associado a uma das características em estudo. Esta propriedade faz com que o software não fique restrito a trabalhar somente com características genômicas, mas também atuando com características morfológicas.

Ainda sobre a configuração da base de dados, o pesquisador deve informar o estilo de dado faltante, para que o software Cluster faça a imputação destes. No exemplo da Figura 4.1 todos os dados que apresentam o número '-9' será imputado conforme explicado na seção 3.2 desta dissertação.

Algumas bases de dados, especialmente aquelas que tratam de dados morfológicos ou mistos, podem ter seus elementos em diversas grandezas ou unidades. Desta forma, o software Cluster possui a opção de normalizar os dados entre 0 e 1. Procura-se assim, dar a todas as características peso igual no processo de *clustering*, evitando a dependência na escolha da unidade de medida. Como esta equalização de dependência não é requerida por todas as bases de dados e estudos, o processo de normalização é uma escolha do pesquisador. Basta o mesmo clicar na caixa apropriada.

4.3 Operações

Conforme já dito, o software Cluster desempenha duas operações básicas: Clustering e Otimização. De forma resumida, o software Cluster pode ser representado pelo diagrama em blocos na Figura 4.3. Nele apresenta-se as entradas e saídas de acordo com cada uma das operações desempenhadas.

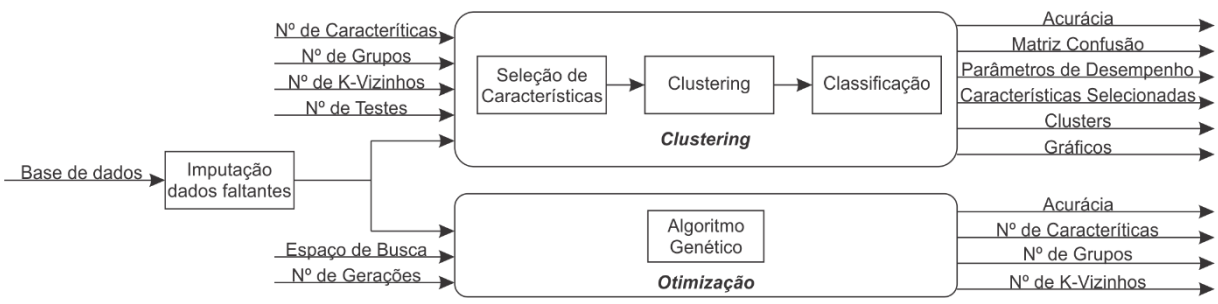
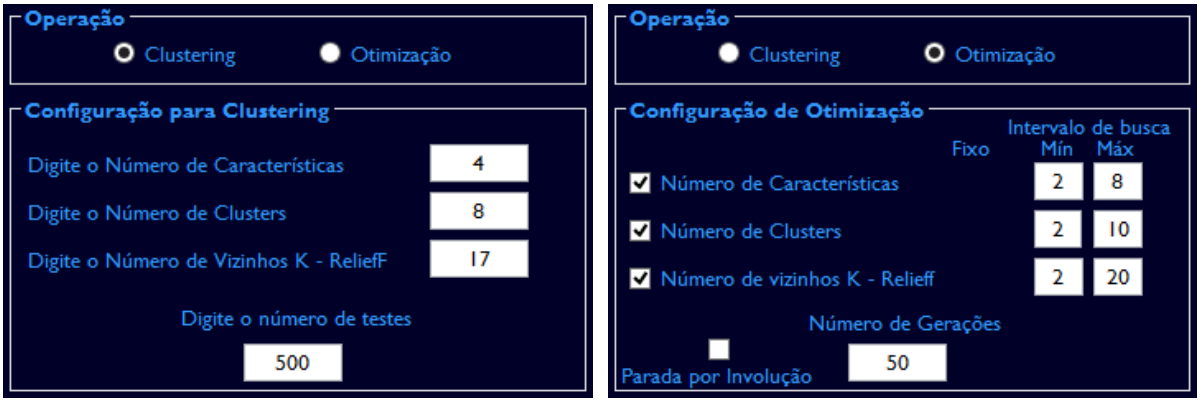


Figura 4.3: Diagrama em blocos – Entradas e saídas

Nota-se que as operações de Clustering e Otimização acontecem de modos distintas, sendo escolha do usuário a seleção de qual das duas operações irá trabalhar no software Cluster, uma por vez. Salienta-se que os parâmetros entregues como resultado pela Operação de Otimização são, em parte, as variáveis utilizadas para configuração da Operação de Clustering. Portanto, o usuário poderá utilizar as operações em sequência para obtenção de melhores resultados.

A seleção entre as operações de Clustering ou Otimização é algo trivial e pode se dar apenas por um clique. As variáveis de configurações são modificadas na tela conforme a operação selecionada pelo usuário. A Figura 4.4 mostra, em dois quadros, a diferença na tela do software Cluster conforme a operação selecionada.



(a) (b)

Figura 4.4: Configurações para cada uma das operações do software Cluster
(a) Operação de Clustering | (b) Operação de Otimização

A seguir são discutidas as configurações, funções e resultados apresentados por cada uma das operações do software Cluster.

4.3.1 Operação de Clustering

A Operação de Clustering é dividida em três etapas:

- Seleção de características, executada pelo método ReliefF;
- *Clustering*, propriamente dito, executado pelo algoritmo *Fuzzy* C-Means e
- Classificação.

As variáveis de configuração para a Operação de Clustering são: o número de características selecionadas, C; o número de grupos formados, G; e o número de

amostras vizinhas consideradas pelo ReliefF, K. Conforme mencionado acima, os valores destas variáveis para se obter o valor otimizado da acurácia de classificação podem ser conseguidos a partir da Operação de Otimização. Na Operação de Clustering, deve-se informar, ainda, a quantidade de testes a serem realizados. A quantidade de testes corresponde a quantidade de execuções do algoritmo *Fuzzy* C-Means.

Parte da contribuição deste trabalho é dada pela forma de apresentar os resultados do Clustering. O resultado de uma Operação de Clustering é exibido na Figura 4.5

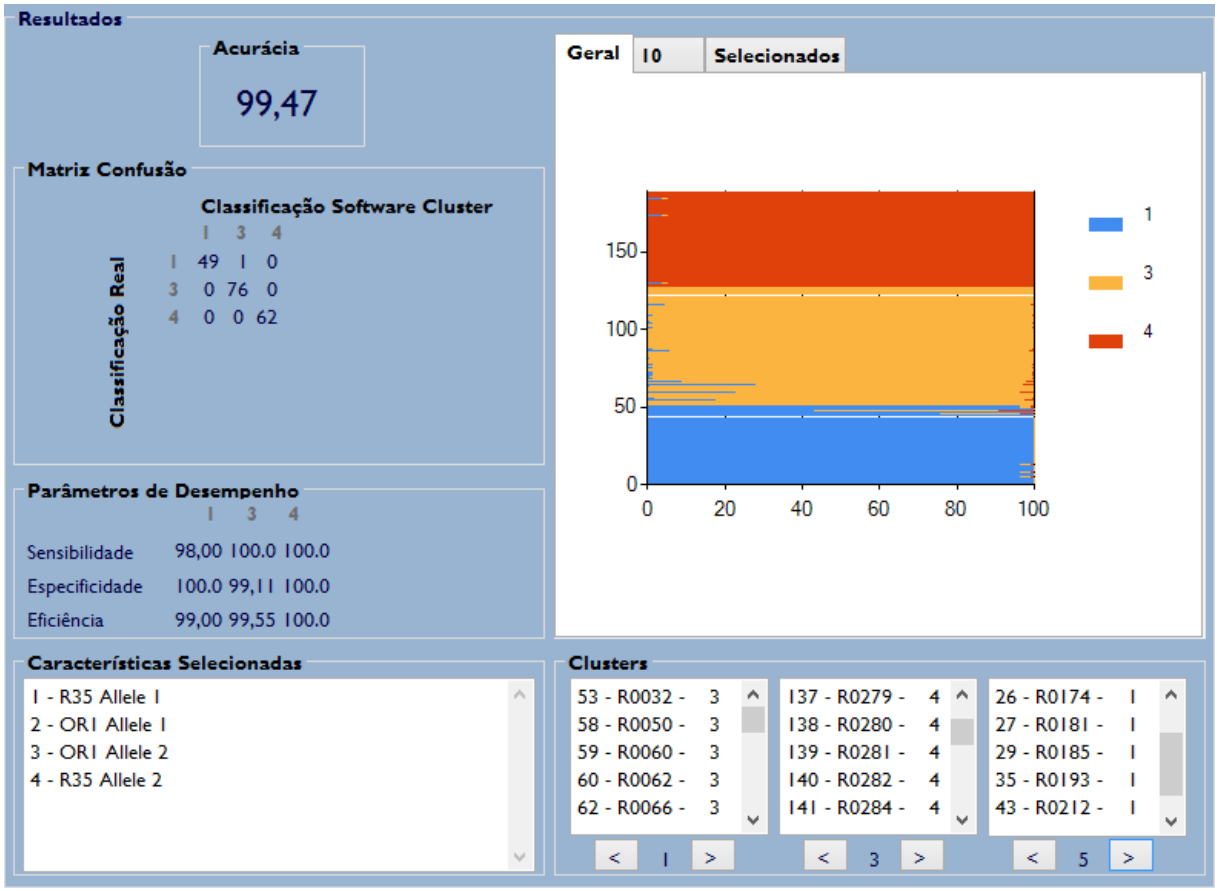


Figura 4.5: Resultado da Operação de Clustering

Na Figura 4.5 tem-se vários quadros a serem analisados. A seguir é apresentada breve explanação sobre cada um destes quadros.

4.3.1.1 Acurácia e Matriz de Confusão

A acurácia obtida pela classificação após a etapa do algoritmo *Fuzzy* C-Means é apresentada no canto superior esquerdo da tela de resultados.

Para auxiliar o pesquisador, a matriz de confusão é mostrada logo abaixo. No exemplo mostrado, a análise da matriz de confusão possibilitou, em primeira percepção, a descoberta de um *outlier* dentro da base de dados.

4.3.1.2 Parâmetros de Desempenho

Uma constante em trabalhos que envolvam dados biológicos é se trabalhar com bases desbalanceadas, ou seja, com base de dados em que a quantidade de amostras é desigual entre as classes. Sabendo que a acurácia é uma medida que pode mascarar resultados ruins quando se trabalha com base de dados desbalanceadas (SUN, KAMEL, *et al.*, 2007; CASTRO e BRAGA, 2011) e que a matriz de confusão não é a melhor forma para se perceber os resultados de desempenho do classificador, o software Cluster apresenta outro quadro de parâmetros de desempenho. Este quadro é composto pela sensibilidade, especificidade e eficiência do algoritmo no processo de classificação de cada uma das classes.

4.3.1.3 Características Selecionadas

Uma das diferenças do software Cluster, perante diversos outros softwares que lidam com dados biológicos, é o modo de selecionar as melhores características para a distinção das classes em estudo. Estas características são mostradas no ranque do quadro ‘Características Selecionadas’.

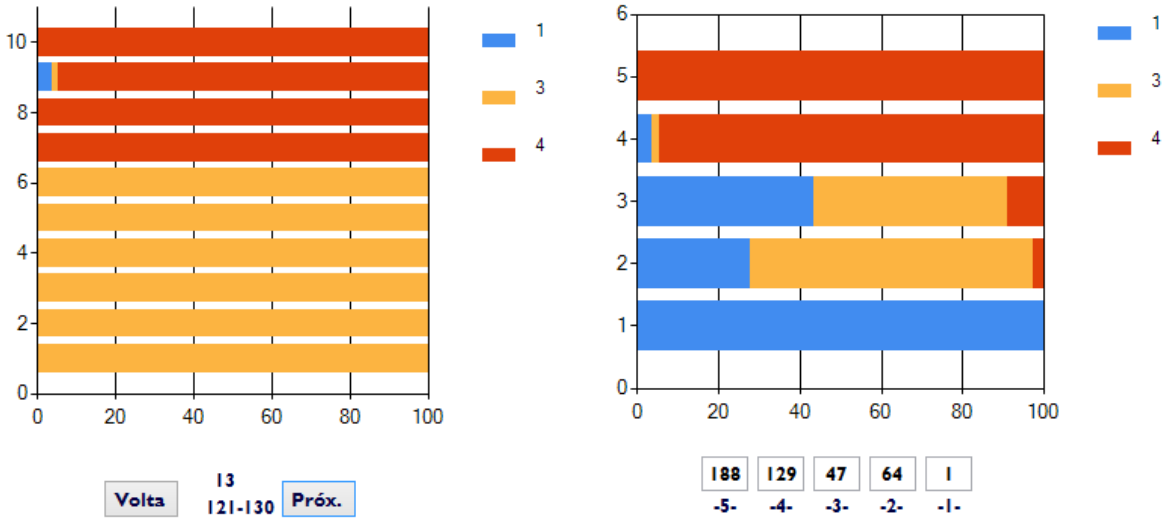
4.3.1.4 Clusters

Para auxílio no estudo de análise de grupos, cada grupo formado pelo software Cluster pode ser observado por três caixas no canto direito inferior da tela, Figura 4.5. É possível navegar entre os diversos grupos formados a partir das teclas ‘<’ e ‘>’. Mostra-se em cada caixa: o número da amostra, o nome da amostra na base de dados e a classe que a amostra foi classificada pelo software Cluster.

4.3.1.5 Gráficos

Por fim, em consonância com os outros softwares que tratam de dados biológicos, é apresentada, graficamente, a pertinência, ou contribuição, de cada classe para a constituição das amostras, segundo o software Cluster. Para ajudar na melhor visualização, o software Cluster apresenta a possibilidade de verificar com maiores detalhes as amostras. Estas são apresentadas em grupos de 10 amostras ou, ainda, por 5 amostras selecionadas pelo próprio pesquisador. Estas opções visam auxiliar a comparação entre a formação das amostras. Um exemplo do gráfico geral, com todas as amostras, pode ser observado no lado direito na Figura 4.5. Com maior detalhe, um exemplo, com um grupo de 10 amostras pode ser observado na Figura 4.6(a), e um grupo com 5 amostras selecionadas na Figura 4.6(b).

Todas a imagens foram adquiridas de uma única simulação. As 10 amostras mostradas na Figura 4.6(a), são as amostras numeradas na base de dados entre 121 a 130, inclusive. As 5 amostras selecionadas para comparação na Figura 4.6(b) são as numeradas como 188, 129, 47, 64 e 1 na base de dados. Cada amostra é indicada por uma barra horizontal onde as cores indicam a pertinência destas amostras a cada uma das classes. Nesta simulação as classes foram nomeadas como: 1, 3 e 4.



(a) (b)

Figura 4.6: Gráficos de pertinência das amostras

(a) Grupo de 10 amostras | (b) Grupo de 5 amostras selecionadas

4.3.2 Operação de Otimização

A Operação de Otimização é, simplesmente, a execução do algoritmo genético proposto nesta dissertação. Como variáveis de configuração, o pesquisador deverá informar o número de gerações a ser evoluídas pelo algoritmo genético, se este deverá ser encerrado pela involução do mesmo e o espaço de busca para otimização. O espaço de busca é compreendido com o intervalo sobre cada uma das variáveis de trabalho para otimização. O pesquisador é capaz de escolher quais variáveis formarão o espaço de busca. Conforme já discutido, está disponível a ele como variáveis de trabalho: o número de características selecionadas, o número de grupos formados e da quantidade de amostras vizinhas consideradas pelo ReliefF. Se o usuário desejar não trabalhar com uma destas variáveis, basta digitar um valor constante. Como *default*, o software sugere:

- Para o número de características selecionadas, C, o valor de um décimo do total de características da base de dados estudo;
- Para o número de grupos, G, o valor do número de classes da base de dados em estudo, e;
- Para o número de amostras vizinhas consideradas pelo Relief, K, o valor 10.

Ressalta-se que a variável K, pouco conhecida por pesquisadores na área de diversidade genética, pode ter o valor 10 adotado seguramente para diversas bases de dados, segundo Robnik-Sikonja e Kononenko (2003).

O número de gerações implica diretamente no tempo de processamento da Operação de Otimização. Para se evitar cálculos desnecessários ou de pouca valia, adotou-se critérios de parada por involução. Estes critérios foram debatidos no Capítulo 4. O pesquisador deverá selecionar, caso deseje, que o algoritmo interrompa seu trabalho por não conseguir melhoria de resultados.

No caso mostrado na Figura 4.4 (b), a Operação de Otimização foi configurada para um espaço de busca formada pelas três variáveis de trabalho. À frente destas variáveis, tem-se os valores mínimos e máximos a serem considerados para a Operação de Otimização. O número de gerações considerado foi de 50 e não foi desejada a interrupção do algoritmo genético pela involução na melhoria de resultados. O resultado desta simulação é apresentado na Figura 4.7.

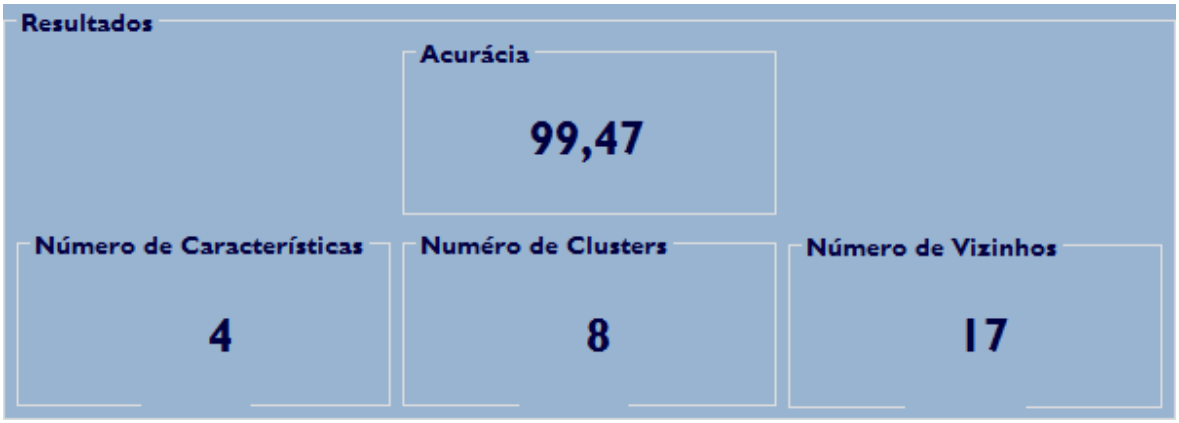


Figura 4.7: Resultado da Operação de Otimização

4.4 Status

O quadro de Status apresentado no menu de configurações, observado na Figura 4.1, informa o pesquisador sobre o desenvolvimento da operação executada pelo software Cluster. Neste quadro encontra-se uma barra de progresso que é preenchida conforme evolução do número de iterações ou gerações, dada a operação desempenhada pelo software. Este progresso também pode ser acompanhado numericamente.

É disponibilizado ao usuário a opção de um aviso sonoro ao fim de uma operação. Este aviso ajuda ao se trabalhar com bases de dados com grandes números de características aliadas a um número elevado de iterações ou gerações. Nestes casos, as operações podem levar um tempo considerável de execução, chegando a algumas horas.

4.5 Botões

Conforme já lembrado, uma das preocupações na concepção do software Cluster é sua fácil utilização. Desta forma, o mesmo possui apenas dois botões na parte de configuração. O primeiro deles, representado por '?', abre uma segunda tela com informações sobre o software Cluster. A tela de informações é apresentada na Figura 4.8. O outro botão, '*Run*', é responsável por iniciar a operação do software Cluster.

Figura 4.8: Software Cluster – Tela de informações

Capítulo 5:

Testes e Discussão

Após apresentação e entendimento das funcionalidades do software Cluster, agora são expostos os diversos testes a que este foi submetido. Os testes serviram para planejamento e adequações do software. Os parâmetros de desempenho analisados nos testes foram discutidos na seção de Metodologia (Capítulo 3). Os parâmetros serviram para confirmar a eficiência de otimização proposta na etapa de seleção de características.

Antes de apresentar qualquer teste, são, brevemente, expostas as bases de dados utilizadas neste capítulo. Todas as bases de dados aqui citadas estão inclusas no software Cluster.

5.1 Base de dados

Durante o processo de desenvolvimento do software Cluster, buscou-se fazer testes com base de dados de propriedades diversas. Esta diversidade se dá pelo número de classes, número de amostras, número de características e o tipo de dados analisados por cada base. Isto serviu para verificar a robustez do software Cluster, não sendo este um software dedicado apenas a uma base de dados.

A Tabela 5.1 mostra um resumo das bases de dados utilizadas durante os testes.

Tabela 5.1: Base de dados

Base de dados	Especificidades	Tipo de dados	Referência
Turtles	3 Classes 188 Amostras 18 Características	Marcadores multialélicos e Microssatélites	(VILAÇA, VARGAS, *et al.*, 2012)
Breast Cancer	2 Classes 569 amostras 30 características	Características Morfológicas	(LICHMAN, 1995)
Gastric Cancer	3 Classes 30 amostras 7130 características	Sondas Gênicas	(HIPPO, TANIGUCHI, *et al.*, 2002)
Prostate Cancer	2 Classes 20 amostras 22283 características	Sondas Gênicas	(BEST, GILLESPIE, *et al.*, 2005)
Seeds	3 Classes 210 amostras 7 características	Características Morfológicas	(LICHMAN, 2012)

Para melhor entendimento da abrangência do software Cluster, a seguir, as bases de dados utilizadas nos testes são brevemente apresentadas.

5.1.1 Turtles

A base de dados nomeada como Turtles, neste trabalho, refere-se a parte do estudo de tartarugas marinhas híbridas desenvolvido por Vilaça, Vargas, *et al.* (2012). Esta base é fruto do primeiro estudo de sequência nucleica da população de tartarugas marinhas do Brasil (VILAÇA e SANTOS, 2013). O DNA mitocondrial de 387 amostras de quatro espécies de tartarugas marinhas da costa brasileira foi sequenciado.

Das 387 amostras, 66 indivíduos foram identificados como híbridos; com a morfologia de uma espécie e o DNA mitocondrial de outra. Dos indivíduos híbridos, sua maioria (50 amostras) pertencem ao cruzamento genético de animais da espécie *Caretta caretta* com animais da espécie *Eretmochelys imbricata*. O número maior de híbridos entre estas espécies faz com que este trabalho foque na busca somente de características para diferenciar amostras destas duas espécies e seus híbridos. Assim, a base de dados original foi reduzida à base Turtles. Esta é formada por 188 amostras, sendo 50 pertencentes a indivíduos híbridos, 76 pertencentes a *Eretmochelys imbricata* e 62 pertencentes a *Caretta caretta*. Os pontos de amostragem, ou colheita de material, são expostos na Figura 5.1.

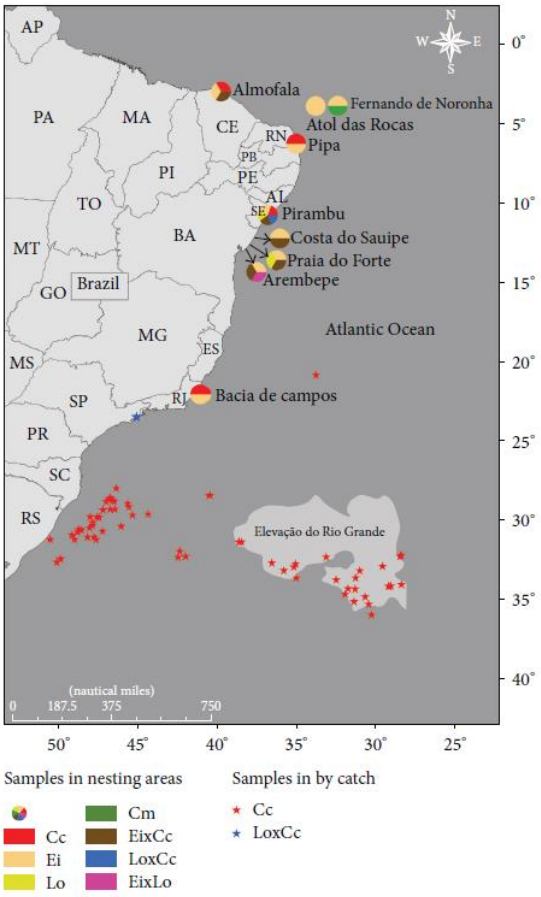


Figura 5.1: Mapa com os locais de amostragem ao longo da costa brasileira.

Círculos não se referem à proporções amostral, mas representam espécies ou híbridos encontrados em cada área. Cc – *Caretta caretta* | Ei – *Eretmochelys imbricata* | Lo – *Lepidochelys olivacea* | Cm – *Chelonia mydas*.

(VILAÇA e SANTOS, 2013)

Cada amostra contém 9 pares de alelos, sendo estas as características de estudo. Estes alelos proveem de cinco marcadores nucleares que foram sequenciados, e quatro microssatélites que foram genotipados em amostras que já possuem um *locus* mitocondrial conhecido (VILAÇA e SANTOS, 2013).

Os cinco marcadores nucleares sequenciados pertencem a quatro éxons: *Brain-Derived Neurotrophic Factor* (BDNF), Oocyte Maturation Factor (CMOS), e *Recombination Activatinggenes* (RAG1 e RAG2)); e a um íntron: (RNA *Fingerprint Protein* 35 Gene (R35)) (VILAÇA e SANTOS, 2013).

Os quatro microssatélites autossomos utilizados pertencem aos *loci* OR1 e OR3 (AGGARWAL, VELAVAN, *et al.*, 2004) e Cc1G02 e Cc1G03 (SHAMBLIN, FAIRCLOTH, *et al.*, 2007; VILAÇA e SANTOS, 2013).

Segundo Ministério do Meio Ambiente Brasileiro, todas as espécies de tartarugas marinhas encontradas no Brasil estão ameaçadas de extinção (BRASIL,

2003). Sabe-se que a hibridação natural tem desempenhado um papel importante na evolução de muitas plantas e animais. Entretanto, a hibridação não natural tem contribuído para a extinção de muitas espécies por meios diretos e indiretos (ALLENDORF, LEARY, *et al.*, 2001). Logo, o estudo desta particularidade, o hibridismo, visa contribuir com políticas específicas para a manutenção e crescimento da população de tartarugas marinhas na costa brasileira.

5.1.2 Breast Cancer

A base de dados Breast Cancer refere-se a um estudo para a classificação no diagnóstico de tumores de câncer de mama em benigno ou maligno feito pela University of Wisconsin (LICHMAN, 1995). O material utilizado para criação da base de dados oriunda de amostras extraídas de tumores de mama, por meio de uma biópsia por aspiração com agulha fina (*Fine needle aspiration, FNA*). Este procedimento de extração é considerado mais simples e seguro do que a biopsia tradicional, realizada por um procedimento cirúrgico invasivo (STREET, WOLDBERG e MANGASARIAN, 1993).

Imagens microscópicas do material colhido na biópsia por aspiração com agulha fina foi digitalizado por Street, Woldber e Mangasarian (1993). Estes, por meio de processamento digital de imagens, extraíram as informações morfológicas das células, formando, assim, a base de dados Breast Cancer. O exemplo de uma imagem digitalizada pode ser observado na Figura 5.2.

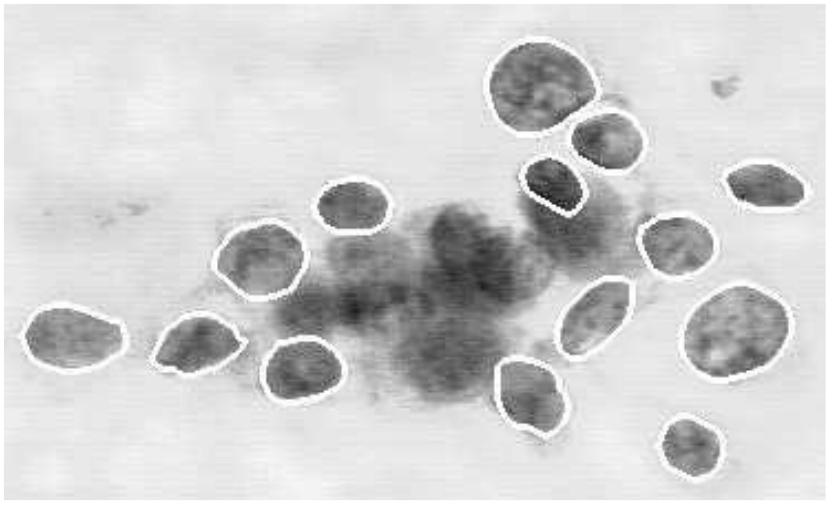


Figura 5.2: Imagem digitalizada de uma amostra da base de dados Breast Cancer (STREET, WOLDBERG e MANGASARIAN, 1993)

Na Figura 5.2, as células analisadas para a formação da base de dados estão contornadas em branco. Foram analisadas 10 informações morfológicas destas células, expressas em valores reais:

- Raio (média das distâncias do centro para pontos no perímetro);
- Textura (desvio padrão dos valores em escala de cinza);
- Perímetro;
- Área;
- Suavidade (variação local nos comprimentos de raio);
- Compacidade;
- Concavidade;
- Número de porções côncavas no perímetro;
- Simetria e
- Dimensão fractal;

Como em uma imagem amostrada possui diversas células, os valores considerados para a formação da base de dados são: os valores médios, os valores de desvio padrão e os valores maiores, ou piores, de cada uma das informações citadas acima. Deste modo, a base de dados possui 30 características para cada amostra, ou cada imagem.

Breast Cancer é constituído por 569 amostras, sendo 357 de tumores benignos e 212 de tumores malignos. A distribuição das amostras entre classes representa 62,74% e 37,26% o que caracteriza esta base como desbalanceada.

5.1.3 Gastric Cancer

Também em torno de estudos sobre o câncer, a base de dados Gastric Cancer lida com a classificação de tecidos de câncer gástrico. Busca-se diferenciar amostras consideradas de tecidos normais, ou não cancerosos, de amostras de tecidos cancerosos gástricos difusos e amostras de tecidos cancerosos gástricos intestinais, sendo estas as classes em estudo.

Diferente das amostras da base Breast Cancer que possuíam características morfológicas a serem analisadas, a base Gastric Cancer possui uma longa cadeia de dados formada por sondas gênicas. Foi feita a análise da expressão gênica de cerca

de 6800 genes totalizando uma base de dados com 7130 características (HIPPO, TANIGUCHI, *et al.*, 2002).

A distribuição das amostras pelas classes pode ser observada na Tabela 5.2.

Tabela 5.2: Base de Dados Prostate Cancer - Distribuição de amostras

Classe	Número de amostras
Tecido Normal	8 (26,67%)
Tumor Gástrico Difuso	5 (16,66%)
Tumor Gástrico Intestinal	17 (56,67%)

A distribuição desigual das amostras nas classes caracteriza esta base de dados também como uma base desbalanceada.

Segundo Hippo, Taniguchi, *et al.* (2002), os resultados conseguidos a partir desta base de dados fornecem não só uma nova base molecular para entender as propriedades biológicas do câncer gástrico. Estes resultados fornecem, também, recursos úteis para o desenvolvimento futuro de terapias e marcadores para diagnóstico de câncer gástrico.

5.1.4 Prostate Cancer

O câncer de próstata é um tipo de câncer caracterizado pela alta reincidência mesmo após ablação do tumor (BEST, GILLESPIE, *et al.*, 2005). Na reincidência, o novo tumor pode ser classificado como um tumor andrógeno dependente ou assumir a forma de um tumor andrógeno independente. A base de dados Prostate Cancer lida justamente com a procura de características que diferenciem estas duas formas do novo tumor.

Segundo Best, Gillespie *et al.* (2005), a identificação destas características para diferenciar as duas formas de andrógenos pode conter potencial para o desenvolvimento de terapias para tratamento de enfermos. Andrógenos independentes são altamente agressivos (BEST, GILLESPIE, *et al.*, 2005) e deixam de responder ao tratamento dado ao primeiro tumor (FELDMAN e FELDMAN, 2001).

A base de dados Prostate Cancer possui 22283 características provindas de análises gênicas de 20 amostras de tumores de câncer de próstata. Desta vez, tem-se 10 amostras de tumores considerados andrógenos dependentes e 10 amostras de

tumores considerados andrógenos independentes, caracterizando esta base como uma base balanceada.

5.1.5 Seeds

A base de dados intitulada Seeds provém do estudo de grãos de trigo. Foram obtidas imagens por meio de raios X de 210 amostras pertencentes a três variedades de trigo: Kama, Rosa e Canadense. Selecionou-se aleatoriamente 70 amostras de cada variedade. Os grãos foram colhidos de campos experimentais cultivados no Institute of Agrophysics of the Polish Academy of Sciences, em Lublin (CHARYTANOWICZ, NIEWCZAS, *et al.*, 2010).

O uso de raios X, para o estudo, torna o processo não-destrutivo e consideravelmente mais barato do que se utilizadas outras técnicas de imagem mais sofisticadas; como a microscopia eletrônica de varredura ou tecnologias a laser (LICHMAN, 2012). Porém, o uso de raios X deixa o processo de classificação dependente de um bom processo de análise de grupos.

Um exemplo de imagem utilizada para a formação da base de dados Seeds pode ser observado na Figura 5.3.

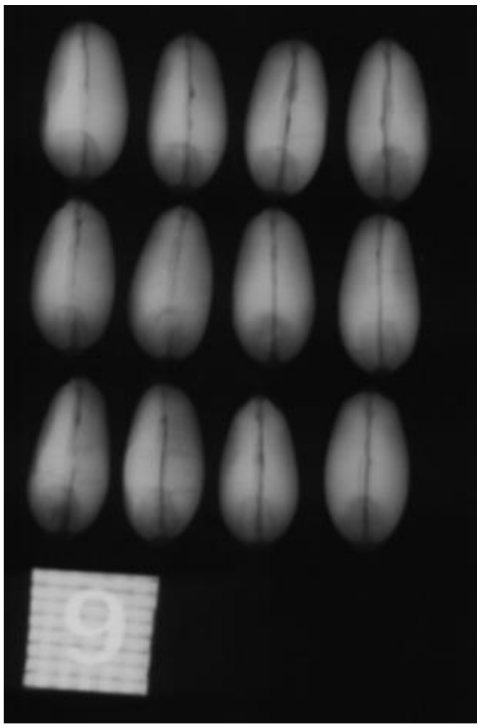


Figura 5.3: Fotografia de grãos de trigo por meio de raios X (18x13cm) (CHARYTANOWICZ, NIEWCZAS, *et al.*, 2010)

Para construir os dados, sete parâmetros geométricos dos grãos de trigo foram medidos:

- Área;
- Perímetro;
- Compacidade;
- Comprimento;
- Largura;
- Coeficiente de assimetria e
- Comprimento do sulco.

Todos esses parâmetros são valores reais e contínuos (CHARYTANOWICZ, NIEWCZAS, *et al.*, 2010).

5.2 Testes

A seguir são, propriamente, apresentados os testes e suas análises. Primeiro são testados, isoladamente, a otimização do número de características selecionadas, C; e a otimização do número de amostras vizinhas mais próximas analisadas pelo ReliefF, K. Com o parâmetro de desempenho utilizado para otimização; acurácia, não justifica um teste isolado da otimização do número de grupos G, pois quanto maior o número de grupos considerados melhor será o resultado deste parâmetro. Ressalta-se que um número elevado de grupos caracteriza um não agrupamento, visto que diversos grupos poderão conter apenas uma amostra. Isto desconfigura uma operação de *clustering*. Válido lembrar que a proposta deste trabalho é de otimizar os resultados atuando não apenas no número de grupos. Busca-se a otimização através do número de grupos junto a melhorias na seleção de características.

Após estes testes isolados, realizou-se testes em base de dados com elevado número de características. Atuou-se em otimizar o número de grupos G e o número de amostras vizinhas K, tendo o número características selecionadas C reduzido e fixo. Este teste está em consonância com estes realizados em outra universidade.

Como a proposta do software Cluster é atuar nestas três variáveis: número de características selecionadas, C; número de grupos, G; e número de amostras vizinhas mais próximas consideradas no ReliefF, K; estudos completos com algumas bases de dados foram realizados.

Por fim, é apresentado um breve teste comparativo entre o resultado obtido pelo software Cluster e o resultado obtido pelo software STRUCTURE no agrupamento das amostras da base Turtles.

Para facilitar a busca e entendimento da variedade de testes realizados, a Tabela 5.3 apresenta um sumário destes. A tabela é um resumo das avaliações executadas nesta dissertação com a indicação da variável otimizada e da base de dados utilizada em cada um dos testes realizados.

Tabela 5.3: Sumário de testes

Teste	Variável Otimizada			Base de dados					Pág.
	C	G	K	*T*	*B*	*G*	*P*	*S*	
Teste sobre otimização do número de características selecionadas	X			X	X				61
Teste sobre otimização do número de amostras vizinhas mais próximas			X	X	X			X	64
Teste sobre base de dados com elevado número de características		X	X			X	X		67
Teste de otimização geral: Turtles	X	X	X	X					70
Teste de otimização geral: Breast Cancer	X	X	X		X				72
Teste de otimização geral: Seeds	X	X	X					X	72
Teste comparativo com o software STRUCTURE	X	X	X	X					75

C – Número de características selecionadas | G – Número de grupos |
K – Número de amostras vizinhas mais próximas consideradas no ReliefF
T – Turtles | B – Breast Cancer | G – Gastric Cancer | P – Prostate Cancer | S – Seeds

Para todos os testes de otimização, os parâmetros de desempenho são comparados em simulações sem e com otimização. Para melhor efeito de comparação, tendo como referência os testes sem otimização, as células dos testes com otimização são realçadas do seguinte modo:

- Células Verdes, indicam melhoria no parâmetro de desempenho;
- Células Amarelas, indicam não alteração do parâmetro de desempenho, e;
- Células Vermelhas, indicam piora no parâmetro de desempenho.

5.2.1 Teste sobre otimização do número de características selecionadas

Um dos objetivos principais deste trabalho é propor uma maneira eficaz de se selecionar o melhor número de características para a classificação de amostras em

uma base de dados. O processo de seleção de características e uma boa classificação estão fortemente relacionados. Um seletor de característica ideal deve produzir uma representação que torna o trabalho do classificador trivial (DUDA, HART e STORK, 2000), sendo que o uso demasiado de características pode prejudicar o processo de classificação (THEODORIDIS e KOUTROUMBAS, 2009).

Para comprovar o ganho de otimização ao escolher o número de características a serem analisadas, o software Cluster foi testado com duas bases de dados: Breast Cancer e Turtles. As bases de dados tiveram os resultados da Operação de Clustering analisados sem e com otimização.

Como este teste envolve o estudo somente do número de características selecionadas as demais variáveis de otimização foram fixadas para as duas análises: sem e com otimização. O número de grupos considerado foi igual ao número de classes da base de dados em teste e o número de amostras vizinhas considerado foi igual a 10, conforme sugere Robnik-Sikonja e Kononenko (2003).

Para a análise sem otimização, utilizou-se o número total de características de cada base de dados em teste. Já para a análise com otimização, o espaço de busca do algoritmo genético foi formado somente pelo número de características, compreendido com valor mínimo igual a uma característica e valor máximo igual ao número total de características da base em teste. A configuração e resultado da Operação de Otimização é apresentado na Tabela 5.4.

Tabela 5.4: Teste sobre otimização do número de características selecionadas – configuração

	Número de características selecionadas			**Número de grupos**	**Número de amostras vizinhas**
	Mín.	***Máx.***	***Ótimo***		
Breast Cancer	1	30	10	2	10
Turtles	1	18	3	3	10

Mín. – Mínimo | Máx. – Máximo

Conforme dito, os testes sem e com otimização do número de características foram realizados por meio de Operações de Clustering no software Cluster. Os resultados dos parâmetros de desempenho são mostrados nas tabelas a seguir. Cada Operação de Clustering ocorreu por meio de 1000 testes, ou seja, 1000 execuções do

algoritmo *Fuzzy* C-Means. Válido ressaltar que a função objetivo otimizada é acurácia de classificação.

Tabela 5.5: Teste sobre otimização do número de características selecionadas – Base: Breast Cancer

	Sem otimização		Com otimização	
	Classe		Classe	
	Maligno	***Benigno***	***Maligno***	***Benigno***
Sensibilidade	61,32%	99,72%	72,64%	99,72%
Especificidade	99,72%	61,32%	99,72%	72,64%
Eficiência	80,52%	80,52%	86,18%	86,18%
Acurácia	85,41%		89,63%	

Tabela 5.6: Teste sobre otimização do número de características selecionadas – Base: Turtles

	Sem otimização			Com otimização		
	Classe			Classe		
	EiCc	***Ei***	***Cc***	***EiCc***	***Ei***	***Cc***
Sensibilidade	74,00%	100,00%	75,81%	80,00%	97,37%	100,00%
Especificidade	90,58%	87,50%	99,21%	98,55%	100,00%	92,06%
Eficiência	82,29%	82,29%	87,51%	89,28%	98,68%	96,03%
Acurácia	85,11%			93,62%		

EiCc – Híbrido | Ei – Eretmochelys imbricata | Cc – Caretta caretta

A Tabela 5.5 apresenta o resultado dos parâmetros de desempenho para a base Breast Cancer. Observa-se que nenhum dos parâmetros de desempenho analisados apresentaram piora comprovando a eficiência da otimização do número de características selecionadas neste teste. Ressalta-se que apenas um terço das características totais da base de dados foram selecionadas. Isto é mostrado na Tabela 5.4, onde o número ótimo de características é igual a 10, de um total de 30.

Ao analisar a Tabela 5.6, com os resultados do teste sobre otimização do número de características selecionadas para base Turtles, depara-se com a piora de dois parâmetros de desempenho. Apesar desta piora obteve-se melhoria na acurácia do classificador, função objetivo. Destacam-se as eficiências de classificação das

amostras onde para todas conseguiu-se atingir valores superiores a 89,00%. Nota-se, ainda, que os demais sete critérios de desempenho analisados, excluído a função objetivo, obtiveram melhoria com a introdução da otimização. Destaque para a sensibilidade em classificar animais *Caretta caretta* e a especificidade em classificar animais *Eretmochelys imbricata* que atingiram 100%. Desta vez, apenas um sexto das características totais da base de dados foi utilizado.

5.2.2 Teste sobre otimização do número de amostras vizinhas mais próximas

Conforme já mencionado anteriormente, Robnik-Sikonja e Kononenko (2003) sugerem que o número de amostras vizinhas mais próximas, utilizado para o cálculo do ReliefF, pode, seguramente, ser ajustado com o valor 10, para a maioria das bases de dados. Porém, é possível através do software Cluster otimizar este valor. Em termos práticos, otimizar o valor de K representa conseguir melhoria na qualidade das características selecionadas.

As bases de dados utilizadas para este teste foram: Seeds, Turtles e Breast Cancer.

O teste para validar a real otimização do número de amostras vizinhas mais próximas consistiu em fixar os valores do número de características selecionadas e o número de grupos utilizados na Operação de Clustering e modificar somente o valor de K. Isto ocorrerá nos testes para a obtenção de resultados sem e com otimização.

Como as bases de dados possuem especificidades em suas dimensões, considerou-se, para este teste, o número de características selecionadas igual a um terço do número das características totais da base em teste e o número de grupos utilizados igual a um décimo do número total de amostras da base em teste. Estes valores foram escolhidos apenas para validar a otimização do número de amostras vizinhas consideradas pelo ReliefF separadamente das outras variáveis. A observação destes valores pode ser realizada na Tabela 5.7.

Tabela 5.7: Teste sobre otimização do número de amostras vizinhas mais próximas – configuração

	Número de Características Selecionadas	Número de Grupos	Número de amostras vizinhas		
			Mín.	***Máx.***	***Ótimo***
Seeds	2	21	2	21	2
Turtles	6	19	2	19	17
Breast Cancer	10	57	2	57	9

Mín. – Mínimo | Máx. – Máximo

Para uma análise sem otimização, o valor de K considerado foi igual a 10. Já para uma análise com otimização utilizou-se o valor entregue pelo software Cluster após a Operação de Otimização. A Operação de Otimização foi configurada com os valores fixos do número de características selecionadas e número de grupos. Desta forma o espaço de busca é formado apenas pelo número de amostras vizinhas utilizado para o cálculo do ReliefF. O intervalo de busca foi estipulado com valor mínimo igual a 2 e valor máximo igual um décimo do número de amostras da base em teste. Foram executadas 50 gerações do algoritmo de otimização. As variáveis de trabalho podem ser observadas na Tabela 5.7.

Os resultados deste teste foram analisados a partir de 100 execuções da Operação de Clustering do software Cluster para cada uma das bases.

Os resultados comparativos entre as análises utilizando ou não a otimização do número de amostras vizinhas são apresentados nas Tabelas 5.8, 5.9 e 5.10. Válido lembrar que na análise sem otimização K = 10.

Tabela 5.8: Teste sobre otimização dos K-Vizinhos mais próximos – Base: Seeds

	Sem otimização			**Com otimização**		
	Classe			**Classe**		
	Kama	***Rosa***	***Canad.***	***Kama***	***Rosa***	***Canad.***
Sensibilidade	74,29%	97,14%	88,57%	92,86%	97,14%	97,14%
Especificidade	93,57%	97,86%	88,57%	98,57%	99,29%	95,71%
Eficiência	83,93%	97,50%	88,57%	95,71%	98,21%	96,43%
Acurácia	86,67%			95,71%		

Canad. – Canadense

Tabela 5.9: Teste sobre otimização dos K-Vizinhos mais próximos – Base: Turtles

	Sem otimização			Com otimização		
	Classe			Classe		
	EiCc	***Ei***	***Cc***	***EiCc***	***Ei***	***Cc***
Sensibilidade	92,00%	100,00%	100,00%	98,00%	100,00%	100,00%
Especificidade	100,00%	99,11%	97,62%	100,00%	99,11%	100,00%
Eficiência	96,00%	99,55%	98,81%	99,00%	99,55%	100,00%
Acurácia	97,87%			99,47%		

EiCc – Híbrido | Ei – Eretmochelys imbricata | Cc – Caretta caretta

Tabela 5.10: Teste sobre otimização dos K-Vizinhos mais próximos – Base: Breast Cancer

	Sem otimização		Com otimização	
	Classe		Classe	
	Maligno	***Benigno***	***Maligno***	***Benigno***
Sensibilidade	91,98%	97,48%	91,98%	97,48%
Especificidade	97,48%	91,98%	97,48%	91,98%
Eficiência	94,73%	94,73%	94,73%	94,73%
Acurácia	95,43%		95,43%	

Ao observar as tabelas, percebe-se a melhoria de resultados a partir da otimização do número de amostras vizinhas analisadas pelo ReliefF, K. Nenhum dos parâmetros de desempenho em nenhum dos testes apresentou piora.

Para base de dados Seeds, apenas um parâmetro de desempenho permaneceu como no teste que utiliza o número de amostras vizinhas proposto por Robnik-Sikonja e Kononenko (2003). Todos os demais foram melhorados. Observa-se, ainda, que o número de K foi diminuído para apenas duas amostras. Isto implica em menor esforço computacional.

Ao se trabalhar com a base Turtles percebe-se, também, claramente, a melhoria nos parâmetros de desempenho. Houve uma melhoria em quatro dos seis parâmetros possíveis, visto que três deles já possuíam valor de 100% e a acurácia é a função objetivo de otimização. Foi possível atingir 100% de eficiência na classificação de uma das classes. A acurácia esteve muito próxima ao valor máximo, tendo apenas uma

amostra classificada erroneamente. Para se obter estes resultados o número de K foi elevado para 17 amostras vizinhas consideradas.

Na tabela de resultados sobre a base Breast Cancer, não consegue-se ver melhorias nos parâmetros de desempenho, porém o número de K foi reduzido para 9 amostras vizinhas. A redução, ainda que pequena de K, implica em uma operação com menor esforço computacional. Desta forma conclui-se que é possível obter o mesmo resultado utilizando menor custo computacional.

5.2.3 Teste sobre base de dados com elevado número de características

Conforme já comentado, a maior contribuição de estudos genéticos na medicina envolve pesquisas acerca das diversas formas do câncer (GRIFFITHS, WESSLER, *et al.*, 2008; ROBINSON, 2010). Por este motivo, parte das bases de dados que ajudaram na concepção do software Cluster envolvem testes sobre esta enfermidade.

Geralmente, estas bases são compostas por longas cadeias de características, conhecidas como sondas gênicas. Cada amostra chega a ter milhares de características. Descobrir quais sondas gênicas que de fato contribuem para a distinção de um tipo de tumor é algo precioso pois pode influenciar diretamente no tratamento a que o paciente é submetido (GOLUB, SLONIM, *et al.*, 1999; HIPPO, TANIGUCHI, *et al.*, 2002; BEST, GILLESPIE, *et al.*, 2005).

Para este teste, sobre base de dados com elevado número de características, considerou-se duas bases de dados que envolvem a análise gênica de tecidos cancerosos: Gastric Cancer e Prostate Cancer.

Conforme os outros testes com o software Cluster, utilizou-se a comparação de resultados da Operação de Clustering sem e com otimização das variáveis de configuração. Cada teste contou com 1000 execuções da Operação de Clustering.

Desta vez, considerou-se o número de características selecionadas para as duas bases fixo em 8 características, para as duas análises: sem e com otimização. Este valor segue o modelo de testes realizado pelo Bioinformatics Laboratory da University of Ljubljana, para estas mesmas bases de dados, com o software Orange, desenvolvido por eles.

Para a análise sem otimização, o número de grupos considerados foi igual ao número de classes da base de dados em estudo e o número de amostras consideradas pelo ReliefF igual a 10. Para a análise com otimização, o espaço de busca foi formado

por estas duas variáveis: o número de grupos e o valor de K. Ambas variáveis tiveram seus valores mínimos estipulados no número de classes da base em estudo e seus valores máximos igual a 10. As variáveis de configuração para Operação de Otimização e o resultado desta operação é exposto na Tabela 5.11.

Tabela 5.11: Teste sobre base de dados com elevado número de características – configuração

	Número de Características Selecionadas	Número de Grupos			Número de amostras vizinhas		
		Mín.	***Máx.***	***Ótimo***	***Mín.***	***Máx.***	***Ótimo***
Gastric Cancer	8	3	10	10	3	10	5
Prostate Cancer	8	2	10	7	2	10	8

Mín. – Mínimo | Máx. – Máximo

Na Tabela 5.12 é apresentado os resultados de acurácias obtidas pelo software Cluster considerando testes sem e com otimização. Observa-se melhoria na acurácia em ambas as bases de dados quando se utiliza a otimização. As características selecionadas foram distintas entre as duas análises.

Tabela 5.12: Teste sobre base de dados com elevado número de características – Acurácia

	Grupo	
	Sem otimização	***Com otimização***
Gastric Cancer	80,00%	93,33%
Prostate Cancer	90,00%	95,00%

O maior interesse do software Cluster não é atingir a melhor acurácia. Esta é apenas a função objetivo de otimização, servindo como guia na seleção de características. Desta forma, a seguir, apresenta-se os resultados dos parâmetros de desempenho nos testes sobre base de dados com elevado número de características. Na Tabela 5.13 tem-se resultados para a base Gastric Cancer e na Tabela 5.14 resultados para a base Prostate Cancer.

Ao analisar ambas as tabelas, percebe-se a melhoria em todos os parâmetros de desempenho possíveis.

Tabela 5.13: Teste sobre base de dados com elevado número de características – Base: Gastric Cancer

	Sem otimização			Com otimização		
	Classe			Classe		
	TN	***TGD***	***TGI***	***TN***	***TGD***	***TGI***
Sensibilidade	100,00%	00,00%	94,12%	100,00%	60,00%	100,00%
Especificidade	90,91%	100,00%	69,23%	100,00%	100,00%	84,62%
Eficiência	95,45%	50,00%	81,67%	100,00%	80,00%	92,31%
Acurácia	80,00%			93,33%		

TN – Tecido Normal | TGD – Tumor Gástrico Difuso | TGI – Tumor Gástrico Intestinal

Tabela 5.14: Teste sobre base de dados com elevado número de características – Base: Prostate Cancer

	Sem otimização		Com otimização	
	Classe		Classe	
	TAD	***TAI***	***TAD***	***TAI***
Sensibilidade	80,00%	100,00%	90,00%	100,00%
Especificidade	100,00%	80,00%	100,00%	90,00%
Eficiência	90,00%	90,00%	95,00%	95,00%
Acurácia	90,00%		95,00%	

TAD – Tumor Andrógeno Dependente | TAI – Tumor Andrógeno independente

Especialmente, no teste com a base Gastric Cancer, vemos que, ao considerar a não otimização, o classificador possuiu sensibilidade igual a 00,00% para a classe de tumores gástricos difusos. Isto significa que nenhuma amostra foi classificada pertencente a esta classe. Vale lembrar que esta classe possui menor representatividade de amostras dentro da base de dados. Já no teste com otimização destaca-se que mais da metade dos nove parâmetros de desempenho analisados atingiram o valor de 100%. Em todas as classes, a eficiência de classificação foi superior a 92,30%. Importante ressaltar que a eficiência de 100% na classificação da classe de tecidos normais indica que as oito características selecionadas pelo software Cluster conseguem distinguir perfeitamente tecidos cancerosos de tecidos não cancerosos.

O resultado com a base Prostate Cancer mostrou que somente uma amostra foi erroneamente classificada ao se utilizar apenas oito características das 22283 que formam a base de dados. Isto representa uma redução de 99,96% no número de características analisadas.

Estes resultados de desempenho só foram possíveis devido a otimização no número de grupos junto ao número de amostras vizinhas mais próximas consideradas pelo ReliefF.

5.2.4 Teste de otimização geral

Na maioria dos testes já apresentados foram testadas as otimizações isoladas, ou seja, considerando um parâmetro por vez. Porém, em pesquisas de com dados biológicos o pesquisador costuma começar seus estudos com várias incógnitas sobre a parametrização a ser utilizada para minerar seus dados. O objetivo da Operação de Otimização do software Cluster é justamente auxiliá-lo nesta busca inicial. Por isto, se torna necessário testar a Operação de Otimização com as três variáveis de trabalho que formam o espaço de busca. A seguir são descritos os resultados obtidos com algumas das bases de dados utilizadas na concepção do software Cluster.

5.2.4.1 Turtles

A base de dados Turtles, que trata de tartarugas marinhas híbridas que nidificam no Brasil, foi a primeira base a ser trabalhada na concepção do software Cluster. Resultados expressivos, expostos a seguir, foram encontrados ao utilizar o software para análise de grupos desta base. Primeiramente, aplicou-se a Operação de Otimização e os resultados obtidos configuraram a Operação de Clustering do software Cluster.

A Operação de Otimização mostrou que com quatro características selecionadas, oito grupos e dezessete amostras consideradas pelo ReliefF é possível atingir uma acurácia de 99,47%. As quatro características selecionadas, nomeadas conforme a base de dados original, foram: R35 Allele 1, OR1 Allele 1, OR1 Allele 2 e R35 Allele 2.

Estes valores aplicados na Operação de Clustering originou os resultados expostos nas próximas tabelas.

Tabela 5.15: Teste sobre base de dados Turtles – Parâmetros de desempenho

	Classe		
	EiCc	***Ei***	***Cc***
Sensibilidade	98,00%	100,00%	100,00%
Especificidade	100,00%	99,11%	100,00%
Eficiência	99,00%	99,55%	100,00%
Acurácia	99,47%		

EiCc – Híbrido | Ei – Eretmochelys imbricata | Cc – Caretta caretta

Tabela 5.16: Teste sobre base de dados Turtles – Matriz de Confusão

		Atribuída (Cluster)		
		EiCc	***Ei***	***Cc***
Real	***EiCc***	49	1	0
	Ei	0	76	0
	Cc	0	0	62

EiCc – Híbrido | Ei – Eretmochelys imbricata | Cc – Caretta caretta

Na Tabela 5.15 destaca-se a eficiência de 100,00% para a classificação de indivíduos da espécie *Caretta caretta* além de sensibilidade perfeita para indivíduos da espécie *Eretmochelys imbricata* e especificidade perfeita para indivíduos híbridos. Todos os parâmetros analisados estão acima ou igual a 98,00%.

Na Tabela 5.16 observa-se que apenas um indivíduo foi erroneamente classificado. Com auxílio dos gráficos e opções de navegação entres os grupos fornecidos pelo software Cluster, facilmente, encontrou-se o número e nome deste indivíduo na base de dados original. Ao investigar este indivíduo, verificou-se que o mesmo se trata de uma amostra composta por material genético de três espécies distintas: *Eretmochelys imbricata*, *Caretta caretta* e *Chelonia mydas*. Este híbrido está erroneamente presente na base de dados Turtles, pertencendo a uma classe não considerada nos estudos. Sendo assim, considerado um *outlier.*

Desta forma, pode-se concluir, por meio da utilização do software Cluster, que os pares de alelos R35 e OR1, assim nomeados na base de dados original, formam uma boa combinação genética para a distinção de tartarugas marinhas *Eretmochelys imbricata*, *Caretta caretta* e animais híbridos formados por estas duas espécies, que nidificam do Brasil.

Todas as figuras apresentadas no Capítulo 4, Apresentação do Software Cluster, correspondem a este estudo por agora descrito.

5.2.4.2 Breast Cancer

Breast Cancer é uma das bases de dados mais acessadas no site UC Irvine Machine Learning Repository. Consta-se mais de 340 mil acessos desde 2007 (LICHMAN, 2013).

Os primeiros pesquisadores a utilizar esta base, Street, Woldberg e Mangasarian (1993), conseguiram selecionar 3 características e atingir a acurácia de 97% após validação cruzada com 10 partes. As três características selecionadas foram: pior textura, pior suavidade e maior número de porções côncavas no perímetro.

O melhor resultado de acurácia obtido pelo software Cluster foi de 98,07%. Porém, para este resultado, a Operação de Otimização designou o uso de 8 características selecionadas, 291 grupos e 168 amostras consideradas pelo ReliefF. O número alto de grupos considerados significa que muitos grupos são formados por somente uma amostra. Isto informa que, apesar de boa acurácia no total de classificação, a Operação de Otimização não se mostrou eficiente. Grupos com apenas uma amostra indicam que a solução proposta é individual a esta base de dados, sendo improvável uma boa generalização para o problema do câncer de mama.

Este teste, de otimização geral de Breast Cancer, mostrou que o software Cluster, apesar de ser eficiente em diversas bases de dados, possui dificuldade em trabalhar com bases de tal especificidade: elevado número de amostras separadas em poucas classes através de poucas características selecionadas.

5.2.4.3 Seeds

A base de dados Seeds, sobre a classificação de amostras de grãos de trigo por meio de imagens de raios x, foi debatida por Charytanowicz, Niewczas e *et al.* (2010). Estes utilizaram o Complete Gradient Clustering Algorithm, CGCA (KULCZYCKI e CHARYTANOWICZ, 2010), como método para análise de grupos da base Seeds. CGCA não exige um valor fixo quanto ao número de grupos, o que, permite-o determinar o valor real presentes na base em análise (KULCZYCKI e

CHARYTANOWICZ, 2010). Para isto, CGCA cria seus grupos baseado em estimativas de densidades de probabilidades (CHARYTANOWICZ, NIEWCZAS, *et al.*, 2010).

Os resultados de classificação, da base Seeds, obtidos por (CHARYTANOWICZ, NIEWCZAS, *et al.*, 2010) são mostrados na Tabela 5.17.

Tabela 5.17: Teste sobre base de dados Seeds – Matriz de Confusão (CGCA)

		Atribuída (Cluster)		
		Rosa	***Kama***	***Canad.***
Real	***Rosa***	67	3	0
	Kama	2	59	9
	Canad.	0	3	67

Canad. – Canadense

O software Cluster foi empregado para análise de grupos da base Seeds. Ao utilizar a Operação de Otimização conseguiu-se a acurácia de classificação de 95,71% com apenas 2 características selecionadas, 19 grupos e 2 amostras vizinhas analisadas pelo ReliefF. As duas características selecionadas foram: comprimento do grão e comprimento do sulco do grão. Os resultados da Operação de Clustering configurada com estes valores são expostos na Tabela 5.18.

Tabela 5.18: Teste sobre base de dados Seeds – Matriz de Confusão (Software Cluster)

		Atribuída (Cluster)		
		Rosa	***Kama***	***Canad.***
Real	***Rosa***	65	1	4
	Kama	0	68	2
	Canad.	2	0	68

Canad. – Canadense

Para melhor comparação, na Tabela 5.19, tem-se os parâmetros de desempenho dos dois métodos; CGCA e Cluster. Os resultados obtidos de CGCA serviram de parâmetro para os resultados obtidos pelo software Cluster. Células pintadas em verde indicam melhoria e as células pintadas em vermelhas piora de resultados.

Tabela 5.19: Teste sobre base de dados Seeds – Parâmetros de desempenho

	CGCA			Software Cluster		
	Classe			*Classe*		
	Rosa	***Kama***	***Canad.***	***Rosa***	***Kama***	***Canad.***
Sensibilidade	95,71%	84,29%	95,71%	92,86%	97,14%	97,14%
Especificidade	98,57%	95,71%	93,57%	98,57%	99,29%	95,71%
Eficiência	97,14%	90,00%	94,64%	95,71%	98,21%	96,43%
Acurácia	91,90%			95,71%		

Canad. – Canadense

Ao observar Tabela 5.19 percebe-se piora na eficiência de classificação para grãos da variedade Rosa. Porém, houve melhorias na eficiência de classificação para grãos das variedades Kama e Canadense. Dos nove parâmetros de desempenho analisados, seis obtiveram melhorias. A acurácia total que atingiu 95,71%.

Charytanowicz, Niewczas, *et al.* (2010) compararam ainda, os resultados de acurácias obtidas por CGCA com o algoritmo K-Means (KANUNGO, MOUNT, *et al.*, 2002). Vale lembrar que o K-Means é base do algoritmo selecionador de características ReliefF e do algoritmo de *clustering Fuzzy* C-Means, ambos utilizados no software Cluster. Os resultados do número de amostras classificadas corretamente para cada uma das classes obtidos pelos três algoritmos são mostrados na Tabela 5.20.

Tabela 5.20: Teste sobre base de dados Seeds – Comparação de acurácias

	K-Means	CGCA	Software Cluster
Rosa	66 (94,29%)	67 (95,71%)	65 (92,86%)
Kama	65 (92,86%)	59 (84,29%)	68 (97,14%)
Canadense	62 (88,57%)	67 (95,71%)	68 (97,14%)
Total	193 (91,94%)	193 (91,94%)	201 (95,71%)

Pela Tabela 5.20 vê-se que o CGCA apresentou piora expressiva na classificação de amostras da variedade Kama se comparado com o K-Means. Ambos obtiveram a mesma acurácia total. Como dito acima, o software Cluster apresentou dificuldade para separação das amostras da variedade Rosa, porém melhor acurácia total.

Ressalta-se, que, em testes com o software Cluster, a acurácia total de 96,19% foi atingida. Para isto, a Operação de Otimização indicou o uso de 3 características selecionadas, 53 grupos e 54 amostras vizinhas analisadas pelo ReliefF. Estes resultados não foram analisados e comparados com o CGCA pois julgou-se o valor de 53 grupos um valor elevado, preferindo assim, restringi-lo em no máximo um décimo do total das amostras da base Seeds.

5.2.5 Teste comparativo com o software STRUCTURE

Para validar os resultados obtidos com o Software Cluster, utilizou-se o software STRUCTURE. Comparou-se os resultados obtidos pelos dois softwares perante a análise de grupos com a base de dados Turtles.

O software STRUCTURE foi configurado para trabalhar com 1.000.000 de repetições de Monte Carlo via Cadeias de Markov (MCMC) com um período de *burn-in* igual a 100.000. Assumiu-se frequências alélicas não correlacionadas e misturadas. Variou-se o número de grupos, ou populações testadas, entre 3 e 10 com 5 simulações independentes para cada valor. Esta configuração segue a mesma executada por Vilaça, Vargas, *et al.* (2012).

Para se descobrir o número de grupos na base de dados, utilizou-se o método de Evanno (EVANNO, REGNAUT e GOUDET, 2005) por meio da ferramenta on-line STRUCTURE HARVESTER (EARL e VONHOLDT, 2012).

O parâmetro Delta K, calculado no método de Evanno, é exposto na Figura 5.4. O número de grupos simulado que gera maior valor de Delta K é aquele considerado por Evanno, Regnaut e Goudet (2005) como melhor representação do número de populações presentes na base em estudo.

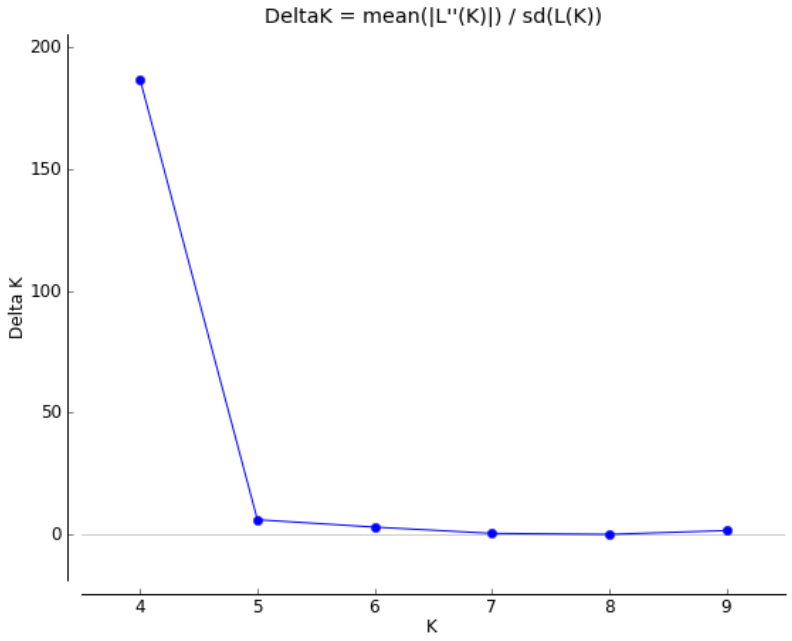


Figura 5.4: Delta K calculado por Structure Harvester – Base: Turtles

Neste caso, pela Figura 5.4, chega-se à conclusão que, de acordo com o método de Evanno, a base de dados Turtles possui 4 populações.

O software Cluster, em estudos com a mesma base de dados, apontou 8 grupos para melhor obtenção dos parâmetros de desempenho observados nesta dissertação. Entretanto, ao forçar o software Cluster a assumir o número de grupos igual a 4, conforme indicado pelo STRUCTURE HARVESTER, conseguiu-se a mesma classificação das amostras quando se trabalhou com 8 grupos. Desta vez, ouve apenas a necessidade de aumentar o número de amostras analisadas pelo ReliefF. As características selecionadas foram: R35 Allele 1, R35 Allele 2, RAG2 Allele 2 e OR1 Allele 1.

A Tabela 5.21 mostra a matriz de confusão obtida pela classificação das amostras a partir da matriz de pertinência fornecida pelo STRUCTURE considerando uma simulação com 4 grupos. O classificador foi o mesmo utilizado nos testes do software Cluster.

Ao comparar esta tabela com a Tabela 5.16, que apresenta os resultados obtidos pelo software Cluster para a base Turtles, percebe-se que STRUCTURE apresentou maior dificuldade em agrupar as amostras.

Tabela 5.21: Resultado do Software STRUCTURE – Base: Turtles | Matriz de Confusão

		Reais		
		EiCc	***Ei***	***Cc***
Previsão	***EiCc***	47	2	1
	Ei	1	75	0
	Cc	0	0	62

EiCc – Híbrido | Ei – Eretmochelys imbricata | Cc – Caretta caretta

Para melhor comparação, tem-se, na Tabela 5.22, a análise dos parâmetros de desempenho considerados nesta dissertação. A mesma proposta de cores é utilizada nesta comparação de resultados, sendo o resultado do software STRUCTURE utilizado como referência.

Tabela 5.22: Teste comparativo com o Software STRUCTURE – Base: Turtles | Parâmetros de desempenho

	Software STRUCTURE			**Software Cluster**		
	Classe			**Classe**		
	EiCc	***Ei***	***Cc***	***EiCc***	***Ei***	***Cc***
Sensibilidade	94,00%	98,68%	100,00%	100,00%	98,70%	100,00%
Especificidade	99,28%	98,21%	99,21%	99,28%	100,00%	100,00%
Eficiência	96,64%	98,45%	99,60%	99,64%	99,35%	100,00%
Acurácia	97,87%			99,47%		

EiCc – Híbrido | Ei – Eretmochelys imbricata | Cc – Caretta caretta

Percebe-se que o software Cluster obteve melhoria em todos os parâmetros de desempenho possíveis. Isto confirma a boa seleção de características desempenhada por Cluster.

Ressalta-se que o software STRUCTURE tem seus agrupamentos formados de maneira distinta da proposta do software Cluster. Por empregar métricas especificas de estudos genéticos, os grupos formados pelo STRUCTURE podem ser consideradas populações gênicas. Além disto, o software Cluster lida apenas com a similaridade entre as amostras da base de dados presente, não sendo possível fazer uma análise entre as gerações dos indivíduos.

5.3 Conclusão sobre os testes

Depois desta batelada de testes pode-se chegar a algumas conclusões sobre o desempenho do software Cluster.

No primeiro teste à que o software foi submetido, viu-se que o número de características selecionadas interfere diretamente no desempenho do classificador. Ao otimizar este valor conseguiu-se melhoria de desempenho na maioria dos parâmetros observados.

No teste de otimização do número de amostras vizinhas consideradas no cálculo do ReliefF, K, comprovou-se que a sugestão proposta por Robnik-Sikonja e Kononenko (2003) não atende todas a bases de estudo. Ao otimizar este valor, obteve-se melhoria nos parâmetros de desempenho. Na base em que estes parâmetros não foram melhorados, conseguiu-se diminuir o valor de K, o que significa a possibilidade de menor esforço computacional para a obtenção de mesmos resultados de desempenho.

No teste sobre base de dados com elevado número de características, trabalhou-se com as bases de dados acerca do câncer, formadas por dados gênicos. Ao reduzir o número de características selecionadas e fixar este valor, conseguiu-se melhorias de resultados ao otimizar as outras duas variáveis de trabalho. Destaca-se o encontrar um grupo de oito características genômicas capaz de separar perfeitamente as amostras de tecidos não cancerosos de tecidos gástricos cancerosos.

Ao testar a otimização de todas as variáveis de estudo com a base Turtles, descobriu-se 2 pares de alelos para distinguir indivíduos da espécie *Caretta Caretta*, de indivíduos espécie *Eretmochelys imbricata* e indivíduos híbridos destas duas espécies. Análise rápida mostrou a presença de um *outlier*.

Teste com a base de dados Breast Cancer mostrou que o software Cluster possui dificuldades em trabalhar com base de dados com elevado número amostras a serem separadas em poucas classes através de poucas características selecionadas. Criou-se diversos grupos unitários, desconfigurando uma operação de *clustering*.

O teste de otimização com a base Seeds mostrou a melhoria na acurácia de classificação dos grãos de trigos se comparados com o algoritmo K-Means e com o algoritmo CGCA.

No teste comparativo com o software STRUCTURE observou-se que o número de grupos indicado pelos softwares foi distinto. Através da ferramenta STRUCTURE

HARVESTER obteve-se o número de 4 populações enquanto Cluster indicou 8 grupos. Entretanto, ao utilizar a indicação do STRUCTURE HARVESTER, Cluster não teve seu desempenho alterado, continuando a conseguir melhores parâmetros de desempenho de classificação se comparação ao software STRUCTURE.

Capítulo 6: Conclusão

Este trabalho trata da proposta de um software para auxiliar pesquisadores em estudos com dados biológicos. Nomeado como Cluster, o software baseado em técnica de agrupamento, trabalha diretamente na seleção das características que melhor distinguem as amostras em classes. Como método para ordenar as melhores características utilizou-se o ReliefF. O algoritmo *Fuzzy* C-Means foi escolhido para a etapa de *clustering.* Aliou-se ainda o desenvolvimento de um algoritmo genético para otimizar a seleção de características em conjunto com o número de grupos utilizados durante *clustering.*

O resultado deste trabalho, o software Cluster, foi envolvido em uma interface de fácil operabilidade. Para comprovar o real funcionamento das técnicas envolvidas, Cluster foi testado com diversas bases de dados, contemplando as principais áreas de estudo da biologia.

Os objetivos específicos deste trabalho foram cumpridos e podem ser observados como seguem:

- A influência de uma boa seleção de característica para o processo de *clustering* foi comprovada nos testes realizados com o software Cluster. Isoladamente a quantidade das características selecionadas foi testada na seção 5.2.1 e a qualidade das características selecionadas na seção 5.2.2 desta dissertação. Os demais testes apresentados nas seções 5.2.3 e 5.2.4 mesclam a otimização na seleção de características com o número de grupos considerados na etapa de *clustering*.

- A otimização na seleção de características em conjunto ao número de grupos foi uma proposta válida comprovada nos testes sobre base de dados com elevado número de características e testes de otimização geral, tratados nas seções 5.2.3 e 5.2.4, respectivamente, desta dissertação.
- O bom desempenho software Cluster pode ser observado nos resultados de acurácia apresentados nas tabelas de testes do capítulo 5, principalmente na comparação de resultados entre operações sem e com otimização.
- A versatilidade do software Cluster foi comprovada ao conseguir bons resultados em diversas bases de dados que contemplam as principais áreas de pesquisa da biologia. As bases de dados usadas como teste foram apresentadas na seção 5.1. Estas possuíam dados morfológicos e gênicos, para atender pesquisas: (i) na área da agricultura, (ii) na área da saúde e (iii) em preservação de animais em extinção.
- O layout simples e de fácil operabilidade foi apresentado no Capítulo 4. Nos testes realizados o layout facilitou a análise dos parâmetros de desempenho. Por exemplo, nos estudos com a base de dados Turtles, apresentado na seção 5.2.4.1, identificou-se um *outlier* de maneira fácil.

De forma geral pode-se concluir que o software Cluster cumpre os objetivos que foi designado. Cluster é uma proposta para auxílio em análises de dados biológicos atuando diretamente na seleção otimizada de características, ou expressões gênicas, para a separação de amostras em grupos de interesses.

6.1 Trabalhos futuros

Porras-Hurtado, Ruiz, *et al.* (2013) lembram que embora existam muitas opções de softwares disponíveis para a análise de dados sobre genética de populações e inferência de ancestralidade, não há nenhum programa aplicável a todas as situações ou tipos de dados.

Testes mostraram que o software Cluster cumpriu seu propósito com base na análise dos parâmetros de desempenho escolhidos. Entretanto, entende-se que esta é apenas a primeira versão do mesmo, cabendo melhoramentos, principalmente, no que envolve métricas relacionadas a formação de populações.

Dentre as melhorias previstas, inclui-se:

- A inclusão de cálculos de índices sobre diversidade genética como a Estatística-F (WRIGHT, 1951), Distância Gênica entre populações (NEI, 1972), Estatística-G (NEI, 1977; NEI, 1987), AMOVA (EXCOFFIER, SMOUSE e QUATTRO, 1992), dentre outros.
- Consideração do equilíbrio de Hardy–Weinberg (HARDY, 1908; WEINBERG, 1908) na formação de grupos em bases compostas por dados gênicos, de forma que os grupos representem populações.
- A determinação do número de grupos, além da otimização proposta, deve-se dar por outros métodos a serem implementados. Como sugestão o método da silhueta (KAUFMAN e ROUSSEEUW, 2005) e o método de Evanno (EVANNO, REGNAUT e GOUDET, 2005), sendo este específico para base com dados gênicos.
- A Operação de Clustering deverá contar com a possibilidade de parada por estabilização de resultado, de modo a independer do pesquisador para informar o número de testes a serem executados. Esta mudança facilita a operabilidade do software ao diminuir mais uma variável de configuração.
- Aprimorar o algoritmo de *clustering* de forma a diminuir a dependência dos bons resultados aos seus centros iniciais aleatórios do algoritmo *Fuzzy* C-Means. Esta mudança tende a diminuir o número de testes da Operação de Clustering.

O relacionamento com pesquisadores na área de genética é peça fundamental para a realização de adequações e implementações de novas operações no software Cluster. Buscar outras técnicas e métodos de seleção de características pode ser um aliado na obtenção de melhores resultados pelo software Cluster.

Atualmente a análise desempenhada pelo software Cluster é por similaridade das amostras. Através da similaridade não é possível retirar conclusões sobre a evolução dos grupos formados pois não se considera gerações distintas.

6.2 Publicações e premiação

O desenvolvimento desta dissertação gerou trabalhos apresentados em congressos nacionais. São eles:

- Resumo intitulado "*Estudo de técnicas de clustering aplicadas a distinção de espécies de tartarugas marinhas da costa brasileira"* apresentado no 5º Congresso Brasileiro de Biologia Marinha, Ipojuca 2015. Este trabalho obteve o 3º lugar no prêmio da Associação Brasileira de Biologia Marinha para estudantes de pós-graduação.
- Artigo intitulado *"Otimização da performance de um classificador por modificação no processo de seleção de características"* apresentado no 12º Congresso Brasileiro de Inteligência Computacional, Curitiba 2015.

Referências

ACID, S.; CAMPOS, L. M. D.; HUETE, J. F. Estimating probability values from an incomplete dataset. **International Journal of Approximate Reasoning**, v. 27, p. 183-204, 2001.

AGGARWAL, R. K. et al. Development and characterization of novel microsatellite markers from olive ridley sea turlte (Lepidochelys olivacea). **Molecular Ecology Notes**, v. 4, n. 1, p. 77-79, 2004.

ALLENDORF, F. W. et al. The problems with hybrids: setting conservation guidelines. **Trends in Ecology & Evolution**, v. 16, n. 11, p. 613-622, Novembro 2001.

ALLISON, P. D. Missing Data. **Sage University Papers Series on Quantitative Applications in the Social Sciences**, Thousand Oaks, 2001.

BABUSKA, R. Fuzzy Clustering Algorithms with Applications to Rule Extraction. In: SZCZEPANIAK, P. S.; LISBOA, P. J. G.; KACPRZYK, J. **Fuzzy Systems in Medicine**. New York: Physica-Verlag HD, v. 41, 2000. p. 139-173.

BERED, F.; NETO, J. F. B.; DE CARVALHO, F. I. F. Marcadores Moleculares e sua Aplicação no Melhoramento Genético de Plantas. **Ciência Rural**, Santa Maria, v. 27, n. 3, p. 513-520, 1997.

BEST, C. J. M. et al. Molecular Alterations in Primary Prostate Cancer after Androgen Ablation Therapy. **Clinical Cancer Research**, v. 11, n. 19, p. 6823-6834, Outubro 2005.

BEZDEK, J. C. **Pattern Recognition with Fuzzy Objective Function Algoritms**. New York: Plenum Press, 1981.

BEZDEK, J. C.; EHRLICH, R.; FULL, W. FCM: The Fuzzy C-Means Clustering Algorithm. **Computers & Geosciences**, v. 10, n. 2-3, p. 191-203, 1984.

BRAGA, A. C. D. S. **ROC: aspectos funcionais e aplicações.** Universidade do Minho. Braga. 2000.

BRASIL. **Instrução Normativa Nº 003, DE 26 de maio 2003**. Ministério do Meio Ambiete. Brasília, p. 20. 2003.

BUSO, G. S. C. et al. Marcadores microssatélites em espécies vegetais. **Biotecnologia Ciência e Desenvolvimento**, Brasília, v. 30, p. 46-50, Janeiro/Junho 2003.

CASTRO, C. L. D.; BRAGA, A. P. Aprendizado Supervisionado com Conjunto de Dados Desbalanceados. **Revista Controle & Automação**, Campinas, v. 22, n. 5, p. 441-466, Setembro/Outubro 2011.

CHARYTANOWICZ, M. et al. Complete Gradient Clustering Algorithm for Features Analysis of X-Ray Images. **Information Technologies in Biomedicine**, Berlin, v. 69, p. 15-24, 2010.

CHIU, C.-C. et al. **Missing value imputation for microarray data:** a comprehensive comparison study and a web tool. 24th International Conference on Genome Informatics. Singapore: BMC Systems Biology. 2013.

CORANDER, J. et al. BAPS 2: enhanced possibilities for the analysis of genetic population structure. **Bioinformatics**, v. 20, n. 15, p. 2363-2369, Abril 2004.

CRUZ, C. D. GENES - a software package for analysis in experimental statistics and quantitative genetics. **Acta Scientiarum**, v. 35, p. 271-276, Julho-Setembro 2013.

DIERINGER, D.; SCHOLÖTTERER, C. Microsatellite Analyser (MSA): a platform independent analysis tool for large microsatellite data sets. **Molecular Ecology Notes**, v. 3, p. 167-169, Março 2003.

DINIZ, M. F.; FERREIRA, L. T. Bancos Genéticos de Plantas, Animais e Microrganismos. **Biotecnologia Ciência & Desenvolvimento**, v. 13, p. 34-38, Março/Abril 2000.

DUDA, R. O.; HART, P. E.; STORK, D. G. **Pattern Classification**. 2ª. ed. New York: Willey, 2000.

DUNN, J. C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. **Journal of Cybernetics**, v. 3, n. 3, p. 32-57, 1974.

EARL, D. A.; VONHOLDT, B. M. STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. **Conservation Genetics Resources**, v. 4, n. 2, p. 359-361, Junho 2012.

EVANNO, G.; REGNAUT, S.; GOUDET, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. **Molecular Ecology**, v. 14, n. 8, p. 2611-2620, Julho 2005.

EXCOFFIER, L.; LISCHER, H. E. L. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windons. **Molecular Ecology Resources**, v. 10, n. 3, p. 564-567, Maio 2010.

EXCOFFIER, L.; SMOUSE, P. E.; QUATTRO, J. M. Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application to Human

Mitochondrial DNA Restriction Data. **Genetics**, Bethesda, v. 131, n. 1, p. 479-491, Junho 1992.

FALUSH, D.; STEPHENS, M.; PRITCHARD, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. **Genetics**, v. 164, n. 4, p. 1567-1587, Agosto 2003.

FALUSH, D.; STEPHENS, M.; PRITCHARD, J. K. Inference of population structure using multilocus genotype data: dominant markers and null alleles. **Molecular Ecology Notes**, v. 7, n. 4, p. 574-578, Julho 2007.

FAWCETT, T. An introduction to ROC analysis. **Pattern Recognition Letters**, v. 27, p. 861-874, 2006.

FELDMAN, B. J.; FELDMAN, D. The development of androgen-independent prostate cancer. **Nature Reviews Cancer**, London, v. 1, p. 34-45, Outubro 2001.

FERREIRA, M. E.; GRATTAPAGLIA, D. **Introdução ao uso de Marcadores Moleculares em Análise Genética**. 3ª. ed. Brasília: EMBRAPA-CENARGEN, 1998.

GAMERMAN, D.; LOPES, H. **Markov Chain Monte Carlo:** Stochastic Simulation for Bayesian Inference. 2ª. ed. Boca Raton: Chapman & Hall, v. 68, 2006.

GATH, I.; GEVA, A. B. Unsupervised optimal fuzzy clustering. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 11, n. 7, p. 773-780, Julho 1989.

GOLDBERG, D. E. **Genetic Algorithms in Search, Optimization, and Machine Learning**. 1ª. ed. Boston: Addison-Wesley Publishing Company, 1989.

GOLUB, T. R. et al. Molecular Classifcation of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. **Science**, v. 286, p. 5313-537, Outubro 1999.

GRIFFITHS, A. J. F. et al. **Introdução à genética**. 9ª. ed. Rio de Janeiro: Guanabara Koogan, 2008.

GRODZICKER, T. et al. Physical Mapping of Temperature-sensitive Mutations of Adenoviruses. **Cold Spring Harbor Symposia on Quantitative Biology**, v. 39, p. 439-446, 1974.

GUIERA, A. J. A. et al. Segmentação por agrupamento fuzzy c-means em imagens LiDAR aplicados na identificação de linhas de transmissão de energia elétrica. **Espaço Energia**, v. 3, p. 24-31, Outubro 2005.

GUSTAFSON, D. E.; KESSEL, W. C. **Fuzzy clustering with a fuzzy covariance matrix**. 17th Symposium on Adaptive. [S.l.]: IEEE Conference. 1978. p. 761-766.

GUYON, I.; ELISSEEFF, A. An Introduction to Variable and Feature Selection. **Journal of Machine Learning Research**, v. 3, p. 1157-1182, 1 Março 2003.

GUYON, I.; ELISSEEFF, A. **An Introduction to Feature Extraction**. 1ª. ed. New York: Springer, 2006.

HARDY, G. H. Mendelian Proportions in a Mixed Population. **Science**, Cambridge, v. 28, n. 706, p. 49-50, Julho 1908.

HERED, J. FSTAT (version 1.2): a computer program to calculate F-statistics. **The Journal of Heredity**, v. 86, n. 6, p. 485-486, 1995.

HIPPO, Y. et al. Global Gene Expression Analysis of Gastric Cancer by Oligonucleotide Microarrays. **Cancer Research**, v. 62, p. 233-240, Janeiro 2002.

HOLLAND, J. H. **Adaptation in natural and artificial systems:** An introductory analysis with applications to biology, control, and artificial intelligence. Oxford: University Michigan Press, 1975.

HUBISZ, M. J. et al. Inferring weak population structure with the assistance of sample group information. **Molecular Ecology Resources**, v. 9, n. 5, p. 1322-1332, Setembro 2009.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM computing surveys**, New York, v. 31, n. 3, p. 264-323, Setembro 1999.

JAKOBSSON, M.; ROSENBERG, N. A. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. **Bioinformatics**, v. 23, n. 14, p. 1801-1806, Julho 2007.

KANUNGO, T. et al. An efficient k-means clustering algorithm: analysis and implementation. **Pattern Analysis and Machine Intelligence**, v. 24, n. 7, p. 881-892 , Julho 2002.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding Groups in Data:** An Introduction to Cluster Analysis. 1ª. ed. Hoboken: John Wiley & Sons, 2005.

KIRA, K.; RENDELL, L. A. **A practical approach to feature selection**. International Conference on Machine Learning. Aberdeen: Morgan Kaufmann. 1992. p. 249-256.

KONONENKO, I. **Estimating Attributes:** Analysis and Extensions of RELIEF. Machine Learning: ECML-94. [S.l.]: Springer Verlag. 1994. p. 171-182.

KULCZYCKI, P.; CHARYTANOWICZ, M. A Complete Gradient Clustering Algorithm Formed with Kernel Estimators. **International Journal of Applied Mathematics and Computer Science**, v. 20, n. 1, p. 123-134, 2010.

LARRAÑAGA, P. et al. Machine learning in bioinformatics. **Briefings in Bioinformatics**, v. 7, n. 1, p. 86-112, Fevereiro 2006.

LICHMAN, M. Breast Cancer Wisconsin (Diagnostic) Data Set. **UCI Machine Learning Repository**, 1995. Disponivel em: <https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)>. Acesso em: Junho 2015.

LICHMAN, M. seeds Data Set. **UCI Machine Learning Repository**, 2012. Disponivel em: <https://archive.ics.uci.edu/ml/datasets/seeds>. Acesso em: Junho 2015.

LICHMAN, M. Welcome to the UC Irvine Machine Learning Repository! **UCI Machine Learning Repository**, 2013. Disponivel em: <https://archive.ics.uci.edu/ml/index.html>. Acesso em: Junho 2015.

LINDEN, R. **Algoritmos Genéticos**. 2ª. ed. Rio de Janeiro: Brasport, 2008.

LITT, M.; LUTY, J. A. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. **The American Journal of Human Genetics**, v. 44, n. 3, p. 397-401, Março 1989.

MCKNIGHT, P. E. et al. **Missing Data:** A Gentle Introduction. 1ª. ed. New York: The Guilford Press, 2007.

MILACH, S. C. K. **Marcadores moleculares em plantas**. Porto Alegre: [s.n.], 1998.

MITCHELL, M. **An introduction to genetic algorithms**. Cambridge: MIT Press, 2002.

NEI, M. Genetic Distance between Populations. **The American Naturalist**, v. 106, n. 949, p. 283-292, Maio/Junho 1972.

NEI, M. F-statistics and analysis of gene diversity in subdivided populations. **Annals of Human Genetics**, v. 41, n. 2, p. 225-233, Outubro 1977.

NEI, M. **Molecular evolutionary genetics**. 1ª. ed. New York: Columbia University Press, 1987.

PIERCE, B. A. **Genetics:** A Conceptual Approach. 5ª. ed. New York: W. H. Freeman, 2013.

PORRAS-HURTADO, L. et al. An overview of STRUCTURE: applications, parameter settings, and supporting software. **Frontiers in Genetics**, v. 4, Maio 2013.

PRIMACK, R. B. **Essentials of Conservation Biology**. 6ª. ed. Sunderland: Sinauer Associates, Inc., 2014.

PRITCHARD, J. K.; STEPHENS, M.; DONNELLY, P. Inference of population structure using multilocus genotype data. **Genetics**, v. 155, p. 945-959, Junho 2000.

PROVOST, F.; KOHAVI, R. Glossary of Terms. **Machine Learning**, v. 30, n. 2-3, p. 271-274, Fevereiro 1998.

RAYMOND, M.; ROUSSET, F. GENEPOP, version 1.2: population genetics software for exact tests and ecumenicism. **Journal of Heredity**, v. 86, p. 248-249, 1995.

ROBINSON, T. R. **Genetics for Dummies**. 2ª. ed. Indianapolis: Wiley Publishing, 2010.

ROBNIK-SIKONJA, M.; KONONEKO, I. **An adaptation of Relief for attribute estimation in regression**. 14th International Conference on Machine Learning. Nashville: The International Machine Learning Society. 1997. p. 296-304.

ROBNIK-SIKONJA, M.; KONONENKO, I. Theoretical and Empirical Analysis of ReliefF and RReliefF. **Machine Learning Journal**, v. 53, n. 1-2, p. 23-69, Outubro 2003.

ROUSSET, F. Genepop'007: a complete re-implementation of the Genepop software for Windows and Linux. **Molecular Ecology Resources**, v. 8, n. 1, p. 103-106, Janeiro 2008.

RUBIN, D. B. Inference and Missing Data. **Biometrika**, v. 63, n. 3, p. 581-592, Dezembro 1976.

RUBIN, D. B. **Multiple imputation for nonresponse in surveys**. 1. ed. Hoboken: John Wiley & Sons, Inc, 1987.

RUBIN, D. B. Multiple Imputation After 18+ Years. **Journal of the American Statistical Association**, v. 91, p. 473-489, Junho 1996.

SCHAFER, J. L. **Analysis of Incomplete Multivariate Data**. 1ª. ed. Boca Raton: Chapman & Hall, 1997.

SHAMBLIN, B. M. et al. Tetranucleotide microsatellites from the loggerhead sea turtle (Caretta caretta). **Molecular Ecology Notes**, v. 7, n. 5, p. 784-787, Setembro 2007.

SNUSTAD, D. P.; SIMMONS, M. J. **Fundamentos de genética**. 6ª. ed. Rio de Janeiro: Guanabara Koogan, 2013.

SRINIVAS, M.; PATNAIK, L. M. Genetic algorithms: a survey. **Computer**, v. 27, n. 6, p. 17-26, Junho 1994.

STREET, W. N.; WOLDBERG, W. H.; MANGASARIAN, O. L. Nuclear Feature Extraction For Breast Tumor Diagnosis. **International Symposium on Electronic Imaging: Science and Technology**, San Jose, v. 1905, p. 861-870, 1993.

SUN, Y. Iterative RELIEF for Feature Weighting: Algorithms, Theories, and Applications. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 26, n. 6, p. 1035-1051, Junho 2007.

SUN, Y. et al. Cost-sensitive boosting for classification of imbalanced data. **Pattern Recognition**, v. 40, n. 12, p. 3358-3378, Dezembro 2007.

SUN, Y.; JIAN, L. **Iterative RELIEF for Feature Weighting**. 23ª International Conference on Machine Learning. Pittsburgh: [s.n.]. 2006. p. 913-920.

THEODORIDIS, S.; KOUTROUMBAS, K. **Pattern Recognition**. 4ª. ed. San Diego: Academic Press, 2009.

VEDRAMIN, L. **Estudo e desenvolvimento de algoritmos para agrupamento fuzzy de dados em cenários centralizados e distribuídos**. Universidade de São Paulo. São Carlos, p. 160. 2012.

VILAÇA, S. T. et al. Nuclear markers reveal a complex introgression pattern among marine turtle species on the Brazilian coast. **Molecular Ecology**, v. 21, n. 17, p. 4300-4312, Setembro 2012.

VILAÇA, S. T.; SANTOS, F. R. D. Molecular Data for the Sea Turtle Population in Brazil. **Dataset Papers in Science**, v. 2013, Junho 2013.

WATSON, J. D.; CRICK, F. H. C. A Structure for Deoxyribose Nucleic Acid. **Nature**, v. 4356, n. 171, p. 737-738, Abril 1953a.

WATSON, J. D.; CRICK, F. H. C. Genetical Implications of the structure of Deoxyribonucleic Acid. **Nature**, v. 4361, n. 171, p. 964-967, Maio 1953b.

WEINBERG, W. Über den Nachweis der Vererbung beim Menschen. **Jahreshefte Verein für vaterländische Naturkunde in Württemberg**, v. 64, p. 369-382, 1908.

WELSH, J.; MCCLELLAND, M. Fingerprinting genomes using PCR with arbitrary primers. **Nucleic Acids Res**, v. 18, n. 24, p. 7213-7218, Dezembro 1990.

WILLIAMS, J. G. et al. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. **Nucleic Acids Res.**, v. 18, n. 22, p. 6531-6535, Novembro 1990.

WRIGHT, S. The general structure of populations. **Annals of Eugenics**, v. 15, n. 4, p. 323-354, Março 1951.

ZABEAU, M.; VOS, P. **Selective restriction fragment amplification:** a general method for DNA fingerprinting. EP0534858A1, 31 Março 1993.

ZOLET, A. C. T. et al. **Guia Prático pra Estudos Filigeográficos**. 1ª. ed. Ribeirão Preto: SBG, 2013.