**UNIVERSIDADE FEDERAL DO PARANÁ**

**Julio Galvão Santana**

# SISTEMA COMPUTACIONAL BASEADO EM APRENDIZADO DE MÁQUINA PARA POSICIONAMENTO TAXONÔMICO DE BACTÉRIAS UTILIZANDO DADOS FENOTÍPICOS

**Curitiba**
**2013**

**Julio Galvão Santana**

# SISTEMA COMPUTACIONAL BASEADO EM APRENDIZADO DE MAQUINA PARA POSICIONAMENTO TAXONÔMICO DE BACTÉRIAS UTILIZANDO DADOS FENOTÍPICOS

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, área de concentração Bioinformática.

Orientador: Prof.ª Dr.ª Maria Berenice Steffens
Co-orientador: Prof. Dr. Roberto Tadeu Raittz
Colaboradores: Prof.ª Dr.ª Cláudia C. G. Martin Didonet
Prof. Dieval Guizelini (MsC Bioinformática)

Curitiba
2013

Santana, Julio Galvão
S232 Sistema computacional baseado em aprendizado de máquina para posicionamento taxonômico de bactérias utilizando dados fenotípicos / Julio Galvão Santana. - Curitiba, 2013.
108 f.: il., tabs, grafs.

Orientadora: Prof ª. Dra. Maria Berenice Steffens
Co-orientador: Prof º. Dr. Roberto Tadeu Raittz
Colaboradores: Prof.ª Dr.ª Cláudia C. G. Martin Didonet
Profº. Dieval Guizelini
Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Curso de Pós-Graduação em Bioinformática.
Inclui Bibliografia.

1. Bacteriologia - Classificação. 2 .Redes neurais (Computação). 3. Bioinformática. I. Steffens, Maria Berenice. II. Raitz, Roberto Tadeu. III. Didonet, Cláudia C. G. Martin. IV. Guizelini. V.Título. VI. Universidade Federal do Paraná.

CDD 589.9

À minha esposa Florida,
À meus pais e família.

## AGRADECIMENTOS

Agradeço a todas as pessoas que de alguma forma me ajudaram a concluir este trabalho.

Aos meus orientadores Prof.ª Dr.ª Maria Berenice Reynaud Steffens e Prof. Dr. Roberto Tadeu Raittz, que com muita dedicação, sabedoria e paciência conduziram este trabalho.

Ao Professor Dieval Guizelini que sempre apoiou e ajudou nos momentos mais difíceis.

A Dr.ª Cláudia C. G. Martin Didonet pelo apoio.

Ao programa de mestrado em Pós Graduação em Bioinformática pela oportunidade.

A todos os professores do programa de Pós Graduação em Bioinformática.

Os funcionários do programa de Pós Graduação em Bioinformática.

A todos os meus colegas de mestrado.

A minha querida irmã Aline pelo apoio e incentivo.

Aos meus queridos pais Dalton e Ilza pelo apoio e incentivo.

A minha amada esposa Florida por suportar pacientemente minha ausência e pelo o apoio nos momentos mais difíceis.

A toda minha família pelo apoio e incentivo.

E principalmente a Deus que sempre me abençoa e guia.

Meu Muito Obrigado.

**RESUMO**

As bactérias são organismos unicelulares que apresentam ampla diversidade morfológica, metabólica e ecológica. Estes microrganismos pertencem ao Domínio Bactéria que, atualmente, conta com 52 Filos. A taxonomia bacteriana inclui a descoberta, descrição e classificação de acordo com normas e princípios, o processo formal de atribuição de nome e a identificação, propriamente dita, de um organismo desconhecido. Historicamente, a identificação e classificação de bactérias tem se baseado principalmente na morfologia, composição do meio de cultivo, potencial de patogenicidade, fisiologia e bioquímica. Atualmente, são também utilizadas informações de ordem fenotípica, genotípica, ecológica e filogenética para produzir uma taxonomia multidimensional. A proposta deste trabalho foi auxiliar na aplicação dos métodos convencionais através da associação da abordagem computacional ao processo de identificação e classificação de bactérias. Foi aplicado o conceito de aprendizado de máquina no desenvolvimento uma ferramenta que permite realizar o posicionamento taxonômico de bactérias baseado em ensaios bioquímicos e fisiológicos. O sistema apresenta funcionalidades que permitem ao usuário cadastrar artigos científicos e espécies bacterianas; cadastrar diferentes categorias de testes e os respectivos resultados (características) disponíveis na literatura ou obtidos no laboratório; obter relatórios referentes aos resultados cadastrados e, finalmente, extrair características a serem utilizadas no treinamento da rede neural FAN (módulo integrado), para então obter o posicionamento taxonômico, em nível de gênero, de uma dada bactéria. Um protótipo foi construído com dados coletados de artigos que descrevem novas espécies de bactérias e o conjunto contem 228 espécies pertencentes a 10 gêneros. Em paralelo, foi estruturado um banco de dados para armazenamento e consulta dos artigos. O treinamento da rede foi validado pelo Cross-validation (leave one out) com uma taxa de acerto de 93%. Isto indica que é possível obter a classificação de bactérias utilizando somente resultados de ensaios bioquímicos e fisiológicos.

Palavras chaves: Taxonomia de bactérias, rede neural, bioinformática

**ABSTRACT**

Bacteria are unicellular organisms that display a wide morphological, metabolic and ecological diversity. These microorganisms belong to the domain Bacteria, which currently has 52 phyla. Bacterial taxonomy includes the discovery, description and classification according to rules and principles, the formal process of naming and identification, strictly speaking, an unknown organism. Historically, the identification and classification of bacteria has been mainly based on the morphology, composition of the culture medium, potential pathogenicity, physiology and biochemistry. Currently, phenotypic, genotypic, phylogenetic and ecological information is also used to produce a multidimensional taxonomy. The purpose of this study was to assist in the application of conventional methods by combining the computational approach to the identification and classification of bacteria process. The concept of machine learning as a tool which allows the taxonomic position of bacteria based on biochemical and physiological tests was applied in the development. The system displays features that allow the user to register scientific articles and bacterial species; to register different categories of tests and results (features) available in the literature or obtained in the laboratory; to obtain reports on the results registered and finally to extract features to be used in the FAN neural network training (integrated module), and then to obtain the taxonomic position of the genus of a certain bacterium. A prototype was built with data collected from articles describing new species of bacteria and the set contained 228 species belonging to 10 genera. In parallel, it was created a database for storage and retrieval of articles. Network training was validated by cross-validation (leave one out) with an accuracy rate of 93%. This indicates that it is possible to obtain the classification of bacteria using only results from biochemical and physiological tests.

Key words: Taxonomy of bacteria, neural network, bioinformatics

**LISTA DE FIGURAS**

## Lista de Gráficos

## LISTA DE QUADROS

## LISTA DE TABELAS

**SUMÁRIO**

## 1. INTRODUÇÃO

A taxonomia bacteriana inclui sistemática, nomenclatura e identificação de um organismo desconhecido. A sistemática, além de documentar, procura compreender a diversidade biológica através da classificação dos organismos. A identificação de bactérias envolve a caracterização de um dado gênero, uma dada espécie ou, ainda, uma dada estirpe, baseada na comparação dos dados referentes com dados de gêneros, espécies ou estirpes previamente classificados e nomeados. Um microrganismo recém-isolado só poderá ser identificado e colocado dentro de um determinado táxon se este táxon já existe. O conhecimento sobre as necessidades nutricionais das bactérias e as condições físicas necessárias para o seu crescimento ajuda a identificá-las e a agrupá-las em grupos taxonômicos distintos. Existem vários testes laboratoriais que podem determinar a atividade metabólica de um microrganismo e o registro detalhado das reações realizadas por uma espécie microbiana é bastante útil para se determinar a qual grupo taxonômico um dado isolado pertence. A metodologia convencional para o processo de isolamento de bactérias e sua caracterização consiste de observação de critérios morfológicos das colônias, testes nutricionais, bioquímicos e fisiológicos, testes de crescimento em meios seletivos, testes sorológicos e testes quimiotaxonômicos. Esta abordagem pode ser complementada pelos métodos moleculares de identificação e classificação de microrganismos, especialmente aqueles baseados na analise da seqüência gênica. A realização dos ensaios convencionais demanda uma infraestrutura básica, de laboratório e de pessoal, que já está acessível para os laboratórios de pesquisa de pequeno e médio porte. Entretanto, a aplicação de técnicas moleculares ainda demanda um elevado investimento de recursos e treinamento de pessoal especializado (VIDEIRA, ARAÚJO, BALDANI, 2007).

## 1.2 JUSTIFICATIVAS DO TRABALHO

Aprimorar a aplicação dos métodos convencionais com a associação de métodos computacionais ao processo de classificação de bactérias, ajudando com isso, suprir a carência existente na área de bactérias não clínicas.

## 1.3 OBJETIVOS

### 1.3.1 OBJETIVO GERAL

Aplicar o conceito de aprendizado de maquina no desenvolvimento de uma ferramenta computacional que permita realizar o posicionamento taxonômico de bactérias baseado em ensaios bioquímicos e fisiológicos e disponibilizar um banco de dados de acesso público com resultados destes testes.

### 1.3.2 OBJETIVOS ESPECÍFICOS

Construir uma ferramenta computacional que aplica técnicas de inteligência artificial para auxiliar no posicionamento taxonômico de bactérias baseado em análises bioquímicas e fisiológicas e paralelamente constituir um banco de dados para armazenamento dos resultados dos testes que descrevem espécies de bactérias.

## 2. REVISÃO BIBLIOGRÁFICA

### 2.1 CLASSIFICAÇÃO DOS SERES VIVOS

Desde os primórdios, a humanidade sentiu a necessidade de classificar tudo quanto existe no meio ambiente, sendo que os antigos gregos e romanos já nomeavam e classificavam os organismos que eram lhes eram úteis. A idéia que a natureza esta dividida em três grandes reinos, mineral, vegetal e animal, foram apresentada em 1675 pelo químico francês Nicholas Lemery (1645-1715) em sua obra *Cours de chymie*. No século XVIII, a classificação de Lemery foi popularizada pelo naturalista sueco Carl von Linné (1707-78), em sua obra *Systema Naturae* publicada em 1735, que estabelecia a classificação hierárquica das espécies e a nomenclatura científica binomial (SCHLEIFER, 2009). Nascia então a Taxonomia, ou seja, a ciência dedicada à descoberta, descrição e nomenclatura das espécies, bem como a organização destas em um sistema de classificação. Linné agrupou os seres vivos de acordo com as características morfológicas por eles partilhadas, mas, ao longo dos séculos, estes agrupamentos foram alterados múltiplas vezes para melhorar a consistência entre a classificação e o princípio darwiniano da ancestralidade comum (CAVALIER-SMITH, 1998, SCHLEIFER, 2009). Inicialmente, os seres vivos eram divididos nos Reinos Plantae e Animalia. No século XIX, a categoria Protista foi adicionada em 1865 por Ernest Haeckel com o objetivo de incluir algas, fungos, protozoários e bactérias. No século XX, em 1969, Robert Whittaker propõe um sistema de classificação composto por cinco reinos, com um reino independente para os fungos: Protista (protozoários e algas unicelulares), Monera (bactérias e cianobactérias), Fungi, Plantae e Animália. Em 1988, Lynn Margulis e Karalene Schwartz propõem um sistema de classificação baseado em dois Super-Reinos ou Domínios: Prokarya e Eukarya. Neste conceito, procariotos pertencem a um só reino Bactéria que se subdivide em dois sub-Reinos Archaeobacteria e Eubactéria e eucariotos pertencem a um só reino que subdivide em quatro sub-reinos: Protoctista, Animália, Fungi e Plantae. Em 1990, Carl Woese, Mark Wheellis e Otto Kandler propõem um sistema de classificação totalmente novo, baseado em comparações de seqüências nucleotídicas do RNA componente da subunidade menor do ribossomo (SSU rRNA). Estas moléculas passam então a ser consideradas cronômetros moleculares. Usando a seqüência de SSU rRNA 16S e 18S, de procariotos e eucariotos respectivamente, como um índice filogenético, Woese e colaboradores agruparam os cinco reinos, criados de acordo com a taxonomia proposta por Linné, em três grandes domínios: Archaea, Bactéria e Eucarya (WOESE; KANDLER & WHEELIS, 1990). Na figura 1 esta mostrada a árvore filogenética universal contendo os principais de procariotos (Archea e Bactéria) e eucariotos (Eucarya). Dentre os três domínios, o Bactéria é o que possui a maior quantidade de organismos.

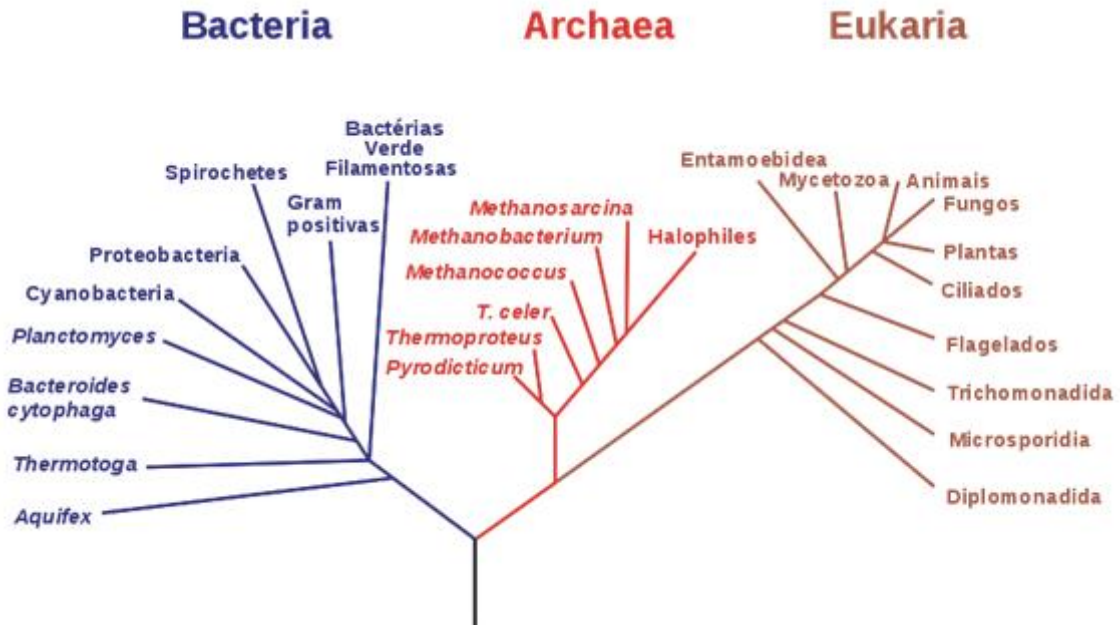



**Figura 1. Árvore filogenética universal determinada com base em comparações de seqüências de rRNA 16 e 18S. Fonte: Wheelis, Klander & Woese, 1992.**

CLASSIFICAÇÃO FILOGENÉTICA UNIVERSAL DOS SERES VIVOS PROPOSTA POR CARL WOESE E COLABORADORES (1990).

## 2.2 DOMÍNIO BACTÉRIA

Este domínio foi primeiramente proposto por Carl R. Woese (1987) e estava composto por 11 filos (Figura 2). No entanto, com passar dos anos e a incorporação da análise do gene 16S de rRNA de organismos independente de cultivo, o número passou para 36 (Figura 3) e a seguir para os atuais 52 filos (Figura 4) ((HUNGENHOLTZ, GOEBEL e PACE,1998; RAPPÉ & GIOVANONNI, 2003).

As bactérias são organismos unicelulares cujo material genético (DNA e plasmídeos) não está envolto por membrana nuclear e sim imerso no citosol. A célula bacteriana apresenta, normalmente, uma das três formas básicas: esféricas (cocos), cilíndricas (bacilos) ou curvadas (vibrios) ou espiriladas (espirilos). Podem conter flagelos que permitem mobilidade e/ou fimbrias (pelos) que estão envolvidas na reprodução sexual (HOGG, 2005).

Apresentam ampla diversidade metabólica e, em relação à fonte de carbono utilizada, podem ser subdivididos em: Autotróficos, que utilizam dióxido de carbono e Heterotróficos, que requerem um tipo ou mais de compostos orgânicos como fonte de carbono. No citoplasma de algumas bactérias podem ser encontradas estruturas denominadas corpos de inclusão que servem como reservatórios nutricionais, podendo conter compostos orgânicos como amido, glicogênio ou lipídeos.

Quanto ao habitat, por causa de sua capacidade de adaptação, sobrevivem em muitos ambientes que não sustentam outras formas de vida. Podem-se encontrar bactérias na atmosfera, oceanos, lagos e fontes termais ácidas, solo, corpo humano e de animais (pele, boca e intestino, por exemplo), associadas a plantas, vulcões, etc.

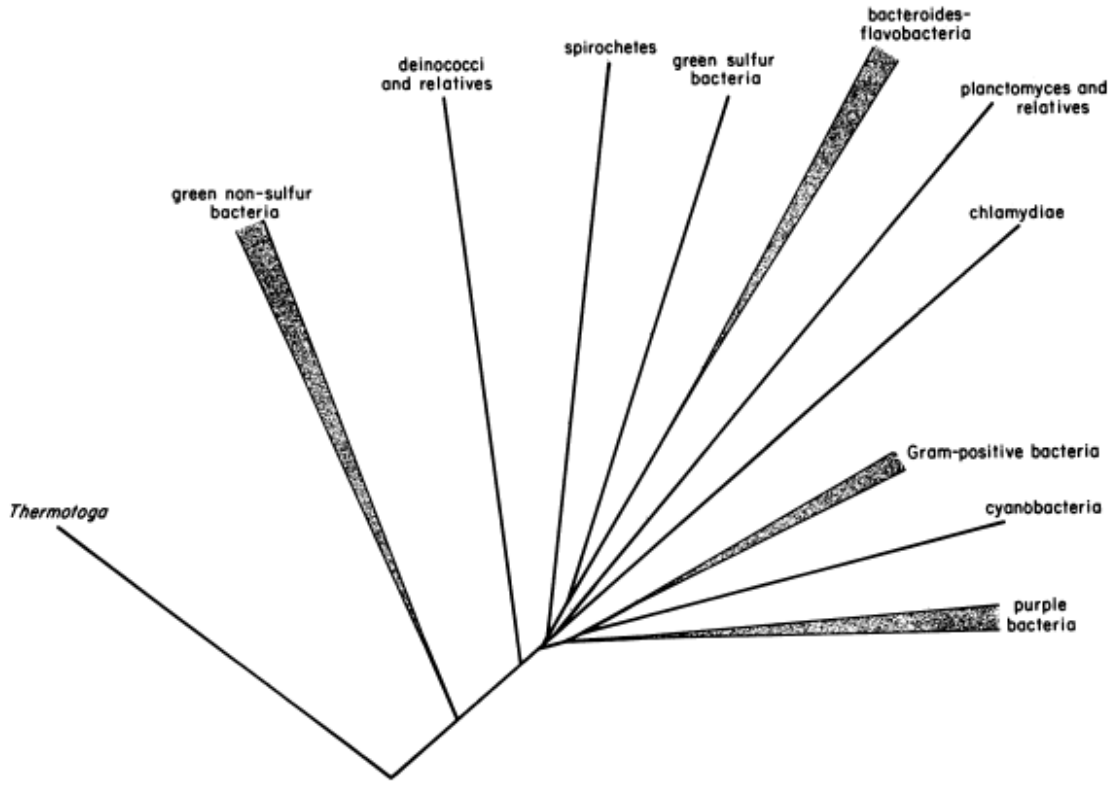



**Figura 2- ÁRVORE FILOGENÉTICA DO DOMÍNIO BACTERIA PROPOSTA POR CARL R. WOESE (1987)**

A árvore apresenta 11 filos bacterianos. O grupo das bactérias gram-positivas foi posteriormente dividido em Firmicutes e Actinobacteria. Árvore reproduzida a partir de Woese, 1987.

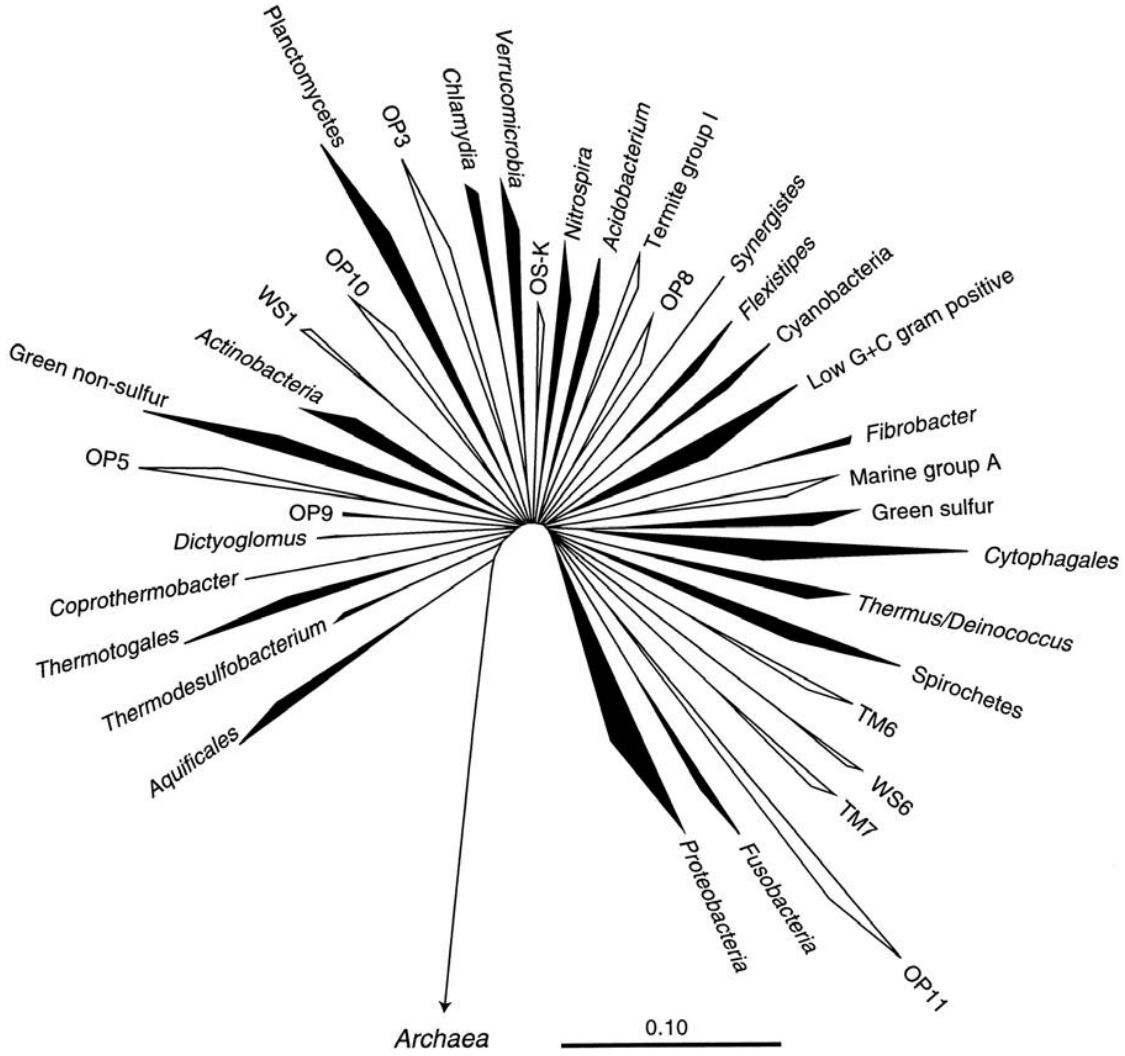



**Figura 3- ÁRVORE FILOGENÉTICA DO DOMÍNIO BACTERIA PROPOSTA POR HUNGENHOLTZ et al (1998b)**

A árvore apresenta 36 filos bacterianos. Ramos preenchidos indicam filos que apresentam representantes cultivados. Ramos não preenchidos indicam filos formados apenas por indivíduos não cultivados. A barra de escala representa 0,1 mudança nucleotídicas por posição. Árvore reproduzida a partir de HUNGENHOLTZ, GOEBEL e PACE, 1998.

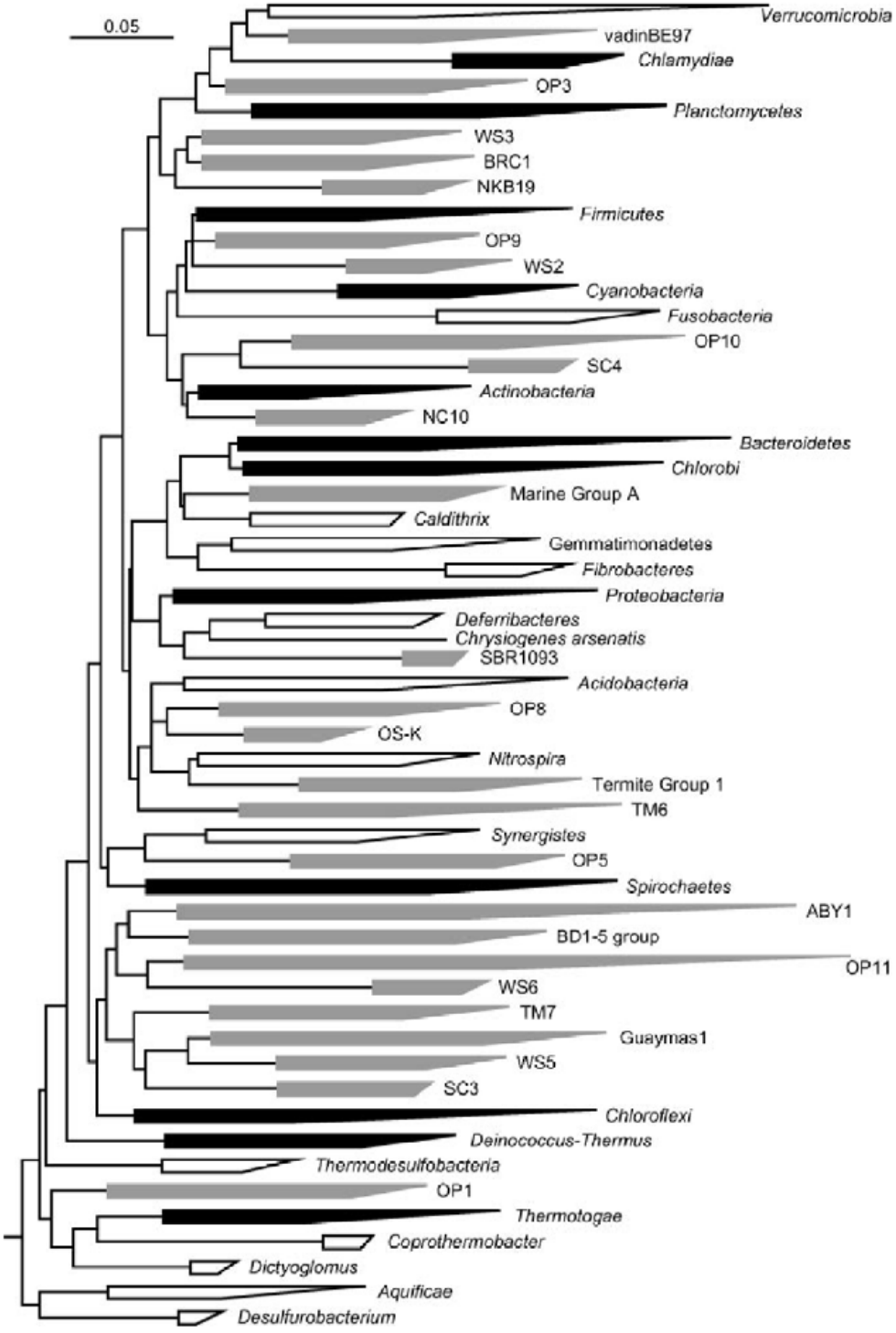



**Figura 4- ÁRVORE FILOGENÉTICA DO DOMÍNIO BACTERIA PROPOSTA POR RAAPÉ E GIOVANNONI (2003)**

A árvore apresenta 52 filos bacterianos. Setas preenchidas representam os 12 filos originais (bactérias gram negativas foram divididas em Firmicutes e Actinobacteria) descritos por Woese (WOESE, 1987), não preenchidas os filos que possuem representantes cultivados reconhecidos desde 1987 e em cinza estão os 26 filos candidatos que não possuem representantes cultiváveis conhecidos. A barra de escala representa 0,05 mudanças nucleotídicas por posição. Árvore reproduzida a partir de RAPPÉ e GIOVANNONI, 2003.

### 2.2.1 TAXONOMIA DE BACTÉRIAS

A taxonomia bacteriana inclui 1. Sistemática (descoberta, descrição e classificação de acordo com normas e princípios), 2. Nomenclatura (processo formal de atribuição de nome) e 3. Identificação de organismo desconhecido (OWEN, 2004). A identificação consiste em se determinar se um organismo pertence a uma das unidades definidas em 1 e 2. O objetivo atual da sistemática, além de documentar é compreender a diversidade biológica e por isto, deve reconstruir a história da diversidade bacteriana através de classificações naturais dos organismos. Existem duas abordagens básicas para a classificação, o Sistema Fenético (ou taxonomia numérica) e o Sistema Filogenético (LENGELER, DREWS, SCHLEGEL, 199). Na análise fenética os agrupamentos baseiam-se em padrões de semelhança e diferença, morfológicas e fisiológicas, entre organismos, baseados em características herdáveis. A organização do conhecimento sobre a diversidade dos organismos se baseia em um conjunto de métodos matemáticos uma vez que as características podem ser medidas, pesadas e numeradas (LENGELER, DREWS, SCHLEGEL, 1999). Na análise filogenética os agrupamentos baseiam-se no padrão da sua história evolutiva. Freqüentemente há descontinuidades, de modo que os padrões revelam agrupamentos com diferentes faixas de variação entre si e vários graus de diferença dentro do grupo. Os padrões filogenéticos mostram como os padrões fenéticos mudam com o tempo, formando uma árvore com diferentes ramificações. Historicamente, a classificação de bactérias tem se baseado principalmente na morfologia, composição do meio de cultivo, potencial de patogenicidade, fisiologia, bioquímica, taxonomia numérica e hibridização DNA-DNA (LENGELER, DREWS, SCHLEGEL, 199). Atualmente, na identificação e definição de novas espécies de bactérias é recomendado o uso da taxonomia polifásica que foi introduzida por COWELL (1970) e onde é utilizadas informações de ordem fenotípica, genotípica, ecológica e filogenética para produzir uma taxonomia multidimensional VANDAMME, P.; POT, B.; GILLIS, M.; De VOS, P.; KERSTERS, K.; SWINGS, J., 1996). Dentre estas abordagens, o seqüenciamento do gene *16S rDNA* é amplamente utilizado para determinar a posição filogenética dos procariotos.

A taxonomia polifásica trouxe a solução ao problema relativo à superficialidade e heterogeneidade dos grupos, dando origem a grupos taxonômicos mais robustos e homogêneos (COLWELL, 1970). Esta abordagem representou um grande avanço para a ciência, sendo que a mesma contribuiu fortemente para o trabalho de Carl R. Woese e colaboradores, trabalho este que agrupou os cinco reinos (Animália, Plantae, Fungi, Protista e Monera) propostos por Whittaker em 1969, em três grandes domínios (Archaea, Bactéria e Eucarya) (WOESE; KANDLER & WHEELIS, 1990).

No quadro 1 estão listadas as principais categorias e características aplicadas na taxonomia bacteriana.

**Quadro 1 - Categorias e características aplicadas na taxonomia bacteriana**

| Categorias | Características (exemplos) |
|---|---|
| Cultural | Morfologia da colônia<br>Cor da colônia<br>Corpos de frutificação<br>Micélio |
| Morfológica | Morfologia da célula<br>Tamanho da célula<br>Motilidade<br>Tipo de flagelo<br>Materiais de reserva<br>Coloração de Gram |
| Fisiológica | Faixa de temperatura<br>Faixa de pH<br>Tolerância a salinidade |
| Bioquímica | Utilização de fontes de carbono<br>Oxidação de carboidratos<br>Fermentação de carboidratos<br>Perfil enzimático |
| Testes inibitórios | Meios seletivos<br>Antibióticos<br>Corantes |
| Sorológica | Aglutinação<br>Imunodifusão |
| Quimiotaxonômica | Ácidos graxos<br>Lipídeos polares<br>Ácidos micólicos<br>Composição de lipopolissacarídeos<br>Aminoácidos de parede celular<br>Açúcares totais<br>Açúcares de parede celular<br>Pigmentos<br>Proteínas totais |
| Genotípica | Conteúdo de C+G<br>Polimorfismo de DNA randomicamente amplificado (RAPD)<br>Polimorfismo de tamanho de fragmentos de restrição (RFLP)<br>Eletroforese de campo pulsado de fragmentos de DNA (PFGE)<br>DNA sonda |
| Filogenética | Hibridização DNA:DNA<br>Hibridização DNA:rRNA<br>Seqüência do gene *16S rRNA*<br>Seqüência do gene *23S rRNA*<br>Seqüência da subunidade *β* da APT sintase<br>Seqüência da chaperona GroEL |

FONTE: Adaptado de Busse, Denner e Lubitz (1996)

### 2.2.2 DIVERSIDADE BACTERIANA E BACTÉRIAS DO SOLO

A diversidade microbiana considerando os parâmetros de diversidade de espécies e diversidade genética suplanta em algumas ordens de magnitude a diversidade existente em todos os demais grupos de seres vivos (MANFIO, 2000*). As bactérias são consideradas os microrganismos mais abundantes* e, segundo o Taxonomic Outline of Bacteria and Archaea (TOBA) Release 7.7 (http://www.taxonomicoutline.org/index.php/toba/index), o número de espécies descritas em 2007 já era superior a 7.000. Embora seja significativo, o número de espécies catalogadas não ultrapassa, possivelmente, 10 % de toda a biodiversidade de bactérias detectadas no meio ambiente. Isto se deve a necessidade do cultivo celular no processo de identificação dos microrganismos, o que na grande maioria das vezes não é viável devido às especificidades metabólicas de muitas espécies (DE LONG, PACE, 2001). Desta forma, o avanço do conhecimento da diversidade de microrganismos não-cultiváveis depende diretamente do desenvolvimento de técnicas que permitam a analise dessas comunidades microbianas de forma independente de cultivo.

Dentre os diferentes ambientes que podem ser ocupados por bactérias, o solo é um ambiente que se destaca. Juntamente com outros microrganismos que habitam o solo, as bactérias constituem uma interface biológica com os ambientes físicos e químicos da Terra, seja atuando diretamente em processos como a mineralização da matéria orgânica ou indiretamente, através de simbioses como na fixação de nitrogênio (O'DONNELL e GÖRRES, 1999). Essa comunidade apresenta propriedades características que dependem direta ou indiretamente dos aspectos climáticos, geográficos, geológicos, hidrológicos, florístico e faunístico, bem como de interferências antropogênicas locais (MOREIRA E SIQUEIRA, 2002).

## 2.2 BACTÉRIAS FIXADORAS DE NITROGÊNIO

Depois do carbono, o nitrogênio é o nutriente presente em maior abundância nos organismos, sendo que este composto está presente no material genético, polissacarídeos, proteínas, etc. (FRANCO & DÖBEREINER, 1994). Todavia, apesar de sua relevância para os seres vivos, o nitrogênio é abundante na natureza na forma de gás, sendo que nesta forma a existência da tripla ligação, torna essa molécula bastante estável e de difícil assimilação pelos organismos (SPRENT & SPRENT, 1990). Portanto, a presença de nitrogênio em formas possíveis de ser assimiladas por eucariotos, tornou-se um fator limitante para o crescimento vegetal, e com a necessidade do aumento da produção agrícola ocorreu também um aumento no uso de adubos químicos nitrogenados, o que além

de ser um fator encarecedor do produto agrícola, também passou a causar sérios problemas de contaminação do solo e da água (ROMERO et al., 1998). As principais conseqüências ambientais do elevado uso de fertilizantes nitrogenados incluem a eutrofização de rios e áreas costeiras, redução da biodiversidade do solo e águas, poluição de reservatórios de água subterrâneos com nitrito e nitrato, e produção $N_2O$, gás com efeito estufa 290 vezes superior ao do $CO_2$.

Uma alternativa ao uso de fertilizantes químicos é a exploração de um processo existente na natureza a milhares de anos, conhecido como Fixação Biológica de Nitrogênio (FBN). Esse processo consiste na conversão do dinitrigenio ($N_2$), presente na atmosfera, em amônia ($NH_3$), forma metabolicamente utilizável pela maior parte dos organismos (Postgate, 1998). O catalisador biológico deste processo é o complexo enzimático da nitrogenase (BURRIS, 1991), cuja reação estequiometricamente balanceada mostrada abaixo (SIMPSON & BURRIS, 1984):

$$N_2 + 8H+ + 8e- + 16Mg_2 - ATP \rightarrow 2NH_3 + H_2 + 16Mg.ADP + 16Pi$$

A fixação biológica de nitrogênio é realizada apenas por procariotos denominados diazotrofos, distribuídos nos domínios Bactéria e Archaea (YOUNG, 1992). As bactérias fixadoras de nitrogênio são comumente classificadas em três grupos (YOUNG, 1992):

**(i)** diazotrofos de vida livre, que fixam $N^2$ para seu próprio consumo

**(ii)** diazotrofos associativos, que colonizam plantas, porém não formam estruturas especializadas. Os organismos endofíticos facultativos podem colonizar tanto o exterior, quanto o interior de raízes. Os endofíticos obrigatórios colonizam apenas o interior de raízes

**(ii)** diazotrofos simbióticos, que estabelecem íntima relação com a planta hospedeira formando estruturas especializadas na fixação de nitrogênio denominadas nódulos.

No contexto de uma aplicação tecnológica, tomando-se como exemplo a produtividade média da soja brasileira de aproximadamente 2500 kg/ha (produção anual de 56 milhões de toneladas) que dependente exclusivamente da simbiose com a bactéria diazotrófica *Bradyrhizobium* sp., estima-se que a fixação biológica de nitrogênio foi responsável por uma

economia equivalente a pelo menos 6 bilhões de dólares americanos para o agricultor brasileiro em 2007/2008 ((BALDANI *et al*., 2002 ; INCT-Fixação Biológica de Nitrogênio).

## 2.3 MÉTODOS DE IDENTIFICAÇÃO E CARACTERIZAÇÃO MORFOFISIOLÓGICA DE BACTÉRIAS

A identificação de bactérias envolve a caracterização de um dado gênero, uma dada espécie ou, ainda, uma dada estirpe, baseada na comparação dos dados referentes com dados de gêneros, espécies ou estirpes previamente classificados e nomeados. Assim, a principio, um organismo recém isolado só poderá ser identificado e colocado dentro de um determinado táxon se este táxon já existe. Bactérias que não foram previamente isoladas não podem ser identificadas, devendo ser primeiramente reconhecidas como novas e então classificadas de acordo com a taxonomia existente (LENGELER, DREWS, SCHLEGEL, 1999).

O conhecimento sobre as necessidades nutricionais das bactérias e as condições físicas necessárias para o seu crescimento ajuda a identificá-las e a agrupá-las em grupos taxonômicos distintos. Alguns destes grupos são capazes de se desenvolverem utilizando compostos químicos simples enquanto outros requerem um sortimento elaborado de nutrientes. Condições físicas como temperatura, luminosidade e pressão osmótica também são importantes para sustentar a vida dos microrganismos. Estas características também podem ser uteis para a identificação e classificação (LENGELER, DREWS, SCHLEGEL, 1999).

As bactérias realizam uma grande variedade de reações químicas que resultam na conversão de nutrientes em macromoléculas complexas ou no catabolismo de macromoléculas em metabolitos mais simples. Existem vários testes laboratoriais que podem determinar a atividade metabólica de um microrganismo. Um registro detalhado das reações realizadas por uma espécie microbiana é útil e muitas vezes essencial para se determinar a qual grupo taxonômico um dado isolado pertence.

A metodologia convencional para o processo de isolamento de bactérias e sua caracterização consiste de observação de critérios morfológicos das colônias, testes nutricionais, bioquímicos e fisiológicos, testes de crescimento em meios seletivos, testes sorológicos e testes quimiotaxonômicos (vide quadro 1).

Atualmente esta metodologia complementa os métodos moleculares de identificação e classificação de microrganismos, especialmente aqueles baseados no estudo da seqüência do gene *16SrDNA*. Esta técnica se baseia na amplificação do gene *16SrDNA* por PCR e posterior caracterização por seqüenciamento (LENGELER, DREWS, SCHLEGEL, 1999).

Outras abordagens moleculares reconhecidas como genotípicas e filogenéticas estão listadas no quadro 1.

A seguir, estão brevemente descritas as principais metodologias convencionais utilizadas na caracterização de isolados bacterianos.

### 2.3.1 ANÁLISE MORFOLÓGICA

Descrita por BOONE & CASTENHOLZ (2001) como a etapa inicial do processo de identificação de uma bactéria esta analise verifica as características celulares como: flagelos, forma, dimensão, comportamento tintorial, estrutura, mobilidade, etc.

### 2.4.2 ANÁLISE BIOQUÍMICA E FISIOLÓGICA

As analises bioquímicas e fisiológicas verificam características como temperatura ideal de crescimento, crescimento na presença de vários substratos, metabolização de compostos variados, valores de pH ideais, coloração de Gram, atividade das enzimas Oxidase e Catalase, entre outras.

#### 2.4.2.1 COLORAÇÃO DE GRAM

A coloração de Gram, também chamada de coloração diferencial é uma técnica de preparação histológica que permite a visualização de bactérias ao microscópio ótico (VIDEIRA, ARAÚJO, BALDANI, 2007), ou seja, é uma técnica de coloração para diferenciação de microrganismos através das cores. É um dos testes bioquímicos mais empregados na caracterização bacteriológica, apresentando grande importância para a taxonomia bacteriana, uma vez que possibilita a separação da maioria das bactérias em dois grandes grupos: Gram positivos e Gram negativos. (CERQUEIRA, 2007; MAGNANI, 2005). Esta técnica se baseia na capacidade da parede bacteriana em reter o corante cristal violeta, após o tratamento com álcool (Figura 5) sendo que isso é possível graças às diferenças químicas existentes entre as paredes de bactérias Gram positivas e Gram negativas. As bactérias Gram positivas apresentam uma espessa camada de ácido teicóico e peptideoglicano que retém o corante, enquanto as Gram negativas apresentam uma

delgada camada de peptídeoglicano sobreposta por uma camada de lipopolissacarídeos, fosfolipídeos, lipoproteínas e proteínas, que não retém o corante (CERQUEIRA, 2007).

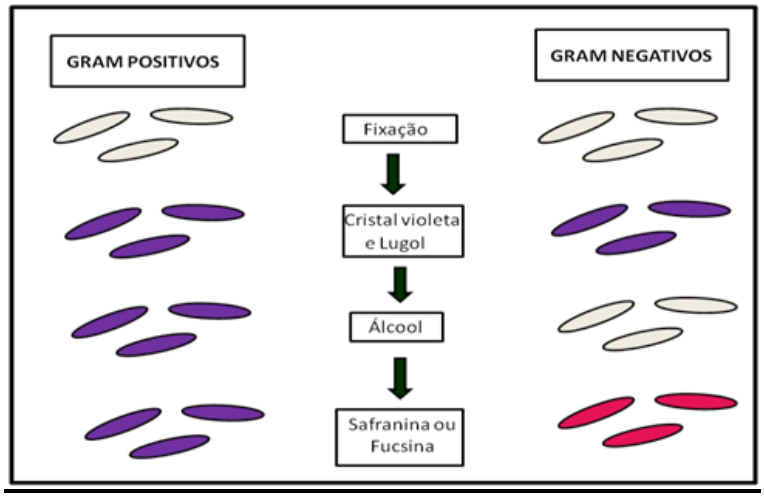



**Figura 5 Representação esquemática da técnica de coloração de Gram.**

Fonte – Adaptado de Videira, Araujo & Baldani, 2007.

### 2.4.2.2 ATIVIDADE DE OXIDASE

Padronizado por Kovacs em 1956 com a utilização do reagente tetrametil-p-fenilenodiamino (TMPD), o teste de atividade da enzima oxidase apresenta grande importância taxonômica, sendo que vários pesquisadores foram capazes de diferenciar espécies de bactérias da mesma família, utilizando o mesmo. Além disso, este teste também é bastante utilizado na caracterização de bactérias Gram negativas, apesar de algumas espécies de bastonetes Gram negativos apresentarem fraca positividade (JURTSHUK, JR, McQUITTY, 1976; TARRAND, GROSCHEL, 1982).

O teste baseia-se em verificar a atividade da enzima oxidase (enzima encontrada em algumas espécies de bactérias, e que tem como função transferir elétrons ao oxigênio) através da utilização do reagente TMPD, sendo que na presença da oxidase esse reagente é oxidado produzindo uma coloração arroxeada indicando a positividade do teste, como representado na figura 6 (VIDEIRA, ARAÚJO, BALDANI, 2007).

Alguns pesquisadores também vêm usando o teste de oxidase como uma ferramenta quantitativa, capaz de verificar o grau de atividade da oxidase, possibilitando também uma diferenciação bacteriana através de seu padrão metabólico (JURTSHUK, MILLIGAN, 1974; JURTSHUK, JR, McQUITTY, 1976).

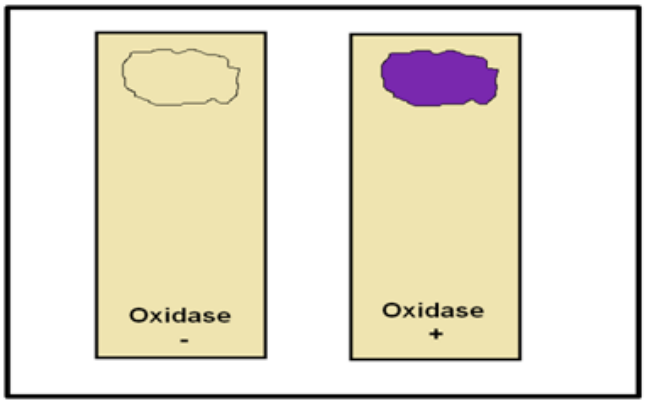



**Figura 6 - Representação do resultado do teste de oxidase.**
Fonte: Adaptado de Videira, Araujo & Baldani, 2007.

### 2.4.2.3 ATIVIDADE DE CATALASE

O teste da atividade da enzima catalase é largamente utilizado para a diferenciação de bactérias Gram positivas, sendo um teste simples, de baixo custo, e boa reprodutibilidade e rapidez nos resultados (CHESTER, 1979). Baseia-se em verificar a presença da enzima catalase através de sua capacidade de converter peróxido de hidrogênio ($H_2O_2$) em água ($H_2O$) e oxigênio molecular ($O_2$) (TAYLOR, ACHANZAR, 1972). Para a realização do teste utiliza-se uma gota de $H_2O_2$ à 3% (v/v) sobre uma gota de cultura líquida contendo a bactéria a ser testa ou uma gota de $H_2O_2$ à 3% (v/v) sobre uma cultura em placa de petri. O surgimento de bolhas indica a positividade do teste, sendo que estas são formadas em função do $O_2$ liberado durante a reação da catalase, como representado na figura 7 (VIDEIRA, ARAÚJO, BALDANI, 2007).

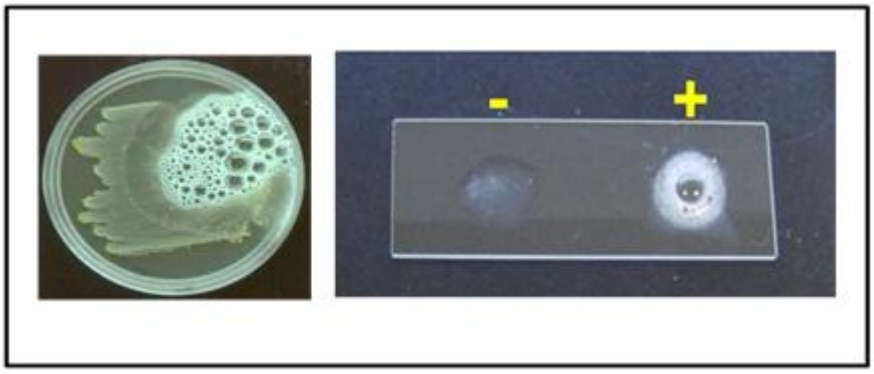



**Figura 7. REPRESENTAÇÃO DO RESULTADO DO TESTE DE CATALASE.**

Fonte: Adaptado de (Videira & Araujo & Baldani, 2007).

### 2.4.2.4 HIDROLISE DE GELATINA

Teste utilizado para classificar bacilos Gram positivos esporulados, bactérias fermentadoras e não fermentadoras.

Alguns tipos de bactérias possuem capacidade de produzir uma enzima proteolítica denominada gelatinase, que tem como função hidrolisar gelatina em componentes capazes de atravessar a membrana bacteriana e servir como nutrientes (VIDEIRA, ARAÚJO & BALDANI, 2007). Para a realização do teste de gelatinase as bactérias devem ser cultivadas em tubos de ensaio com meio contendo peptona, extrato de levedura e gelatina, e após incubação por 24 horas, esses tubos devem ser refrigerados por 2 horas. Após isso incubados novamente, sendo que esse procedimento deve ser repetido por cinco dias. A positividade do teste é dada pela liquefação do meio, como representado na figura 8 (YANO *et al*.,1991). Tubo A reação negativa (meio semi-sólido), tubos B e C reações positivas (meio líquido).

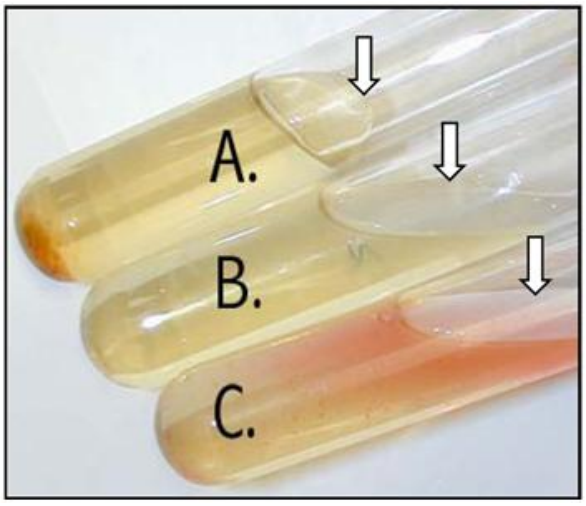



**Figura 8 - REPRESENTAÇÃO DO RESULTADO DE GELATINASE.**

Fonte: Imagem disponível em http://homepages.wmich.edu/~rossbach/bios312

### 2.4.2.5 TEMPERATURA ÓTIMA DE CRESCIMENTO

Cada tipo de bactéria possui uma temperatura ótima de crescimento, ou seja, uma temperatura onde o microrganismo melhor se desenvolve (VIDEIRA, ARAÚJO, BALDANI, 2007). Bactérias que se desenvolvem bem em temperaturas mais baixas são denominadas psicrófilas, as que se desenvolvem bem em temperaturas medianas são chamadas mesófilas, sendo que nesse grupo se incluí a maioria dos patógenos humanos, bactérias que se desenvolvem melhor em temperaturas mais elevadas são denominadas termófilas e os hipertermóficas são microrganismos que se desenvolvem em temperaturas de até 120ºC (VIDEIRA, ARAÚJO, BALDANI, 2007). Partindo do princípio que cada microrganismo possui uma temperatura ótima de crescimento, esse também é um critério de classificação bacteriana, sendo que para tal, a bactéria avaliada é incubada durante em diferentes temperaturas para avaliação de seu desenvolvimento (VIDEIRA, ARAÚJO, BALDANI, 2007). O tempo de incubação pode variar de acordo com o microrganismo.

### 2.4.2.6 pH ÓTIMO DE CRESCIMENTO

O pH do meio é importante para o crescimento das bactérias, uma vez que o potencial hidrogenioiônico influencia diretamente no metabolismo celular. Normalmente, o melhor desenvolvimento do microrganismo ocorre quando o pH do meio está entre o pH mínimo e o pH ideal, do que quando o pH do meio está entre o pH ideal e o pH máximo.
Existem três tipos de classificação para as bactérias em relação ao pH:

- acidófilos, crescem melhor em pH abaixo da neutralidade(1,0 – 5,5);
- neutrofilos, crescem melhor em pH neutro (5,5 – 8,0);
- alcalifilos, crescem melhor em pH alcalino (8,0 – 11,5)

Em meio de cultura, deve ser utilizado tampões para manter o pH em equilíbrio, mesmo após excreção de resíduos pelos organismos e considerando sempre o pH ótimo de crescimento.

Para a avaliação do crescimento devem ser utilizados frascos contendo meio de cultura mais indicado e condições ótimas de crescimento. Os frascos devem conter meio com diferentes pH's (por ex, 4.0; 5.0; 5.5; 6.0;6.5; 7.0; 7.5; 8.0; 9.0) e neles será inoculada a suspensão bacteriana. O tempo de incubação pode variar de acordo com o microrganismo (VIDEIRA, ARAUJO, BALDANI, 2007).

### *2.4.2.7 CRESCIMENTO NA PRESENÇA DE CLORETO DE SÓDIO (NaCl)*

Alguns organismos se desenvolvem em ambientes de altos teores de salinidade (NaCl) e para isto possuem o mecanismo conhecido como osmoadaptação que consiste em evitar a desidratação das células.

Para a avaliação deste processo as bactérias são cultivadas em meio liquido pelo período adequado, centrifugadas e as células ressuspensas em tampão fosfato 0,05M estéril. Em seguida, são inoculadas em placas de Petri contendo meio solido com diferentes concentrações de NaCl (por ex, 0,10,30,50,100 $g.L^{-1}$). A avaliação leva em conta a presença ou ausência e intensidade de crescimento no meio de cultura (VIDEIRA, ARAUJO, BALDANI, 2007).

### 2.4.2.8 HIDROLISE DE CASEÍNA

A caseína é uma das proteínas do leite que, devido à elevada massa molecular, é incapaz de penetrar na membrana celular dos microrganismos. A utilização da caseína pelos mesmos é possível pela sua degradação em oligopeptídios, dipeptídios e finalmente em aminoácidos, os quais depois são assimilados e catabolizados pelas células. A hidrólise da caseína é catalisada por enzimas proteolíticas (proteases) produzidas pelos microrganismos e o objetivo do teste é determinar a capacidade de um dado microrganismo excretar uma enzima (proteolítica extracelular) capaz de degradar a caseína.

No teste, o meio de cultivo é composto por agar nutritivo suplementado com leite e durante o período de incubação os microrganismos que secretam proteases exibem uma zona clara rodeando a zona de crescimento bacteriano. Isto caracteriza uma reação positiva, a perda da opacidade do meio é resultante de uma reação hidrolítica com formação de aminoácidos solúveis e não coloidais. A ausência do halo caracteriza uma reação negativa, ou seja, o meio que envolve o crescimento do microrganismo mantém-se opaco (LENGELER et. al., 1999).

### 2.4.2.9 CARACTERIZAÇÃO METABÓLICA - FERMENTAÇÃO DE FONTES DE CARBONO

Os microrganismos efetuam as suas variadas atividades bioquímicas utilizando nutrientes obtidos a partir do ambiente que os rodeia. É possível verificar algumas destas atividades através da observação da capacidade destes utilizarem enzimas para degradar carboidratos, por exemplo. Esta metabolização origina produtos finais como, por exemplo,

ácidos, gases ou outras moléculas orgânicas, cuja detecção pode ajudar na caracterização e identificação dos microrganismos.

A degradação fermentativa ocorre geralmente num meio líquido que contém o substrato especifico que determina a capacidade fermentativa. Após incubação, a libertação de compostos ácidos, resultantes da fermentação do carboidrato, provoca a redução do pH do meio. Isto é observado pela inclusão de um indicador de pH no meio de cultivo, o que leva à mudança da cor original do meio e que permite caracterizar o teste como uma reação positiva. As culturas que não são capazes de fermentar o carboidrato não conduzem à mudança de cor do meio nem apresentam produção de gás, isto caracteriza uma reação negativa (LENGELER et. al., 1999).

### 2.4.2.10 REDUÇÃO DE NITRATO

A redução dos nitratos por alguns microrganismos ocorre na ausência de oxigênio. Nestes microrganismos a respiração anaeróbia é um processo oxidativo, pois as células usam substâncias inorgânicas como os nitratos ($NO_3^-$) para fornecer oxigênio que subseqüentemente é utilizado durante a produção de energia. Com isto, os nitratos são reduzidos a nitritos ($NO_2^-$).

Para determinar a redução dos nitrato, inocula-se o microrganismo num meio de cultivo suplementado com 0,5% de nitrato de potássio ($KNO_3$) como fonte de nitrato. Após incubação, a cultura é examinada para a presença de íons nitrito no meio. A verificação da capacidade do microrganismo em reduzir o nitrato a nitrito é determinada pela adição de dois reagentes: ácido sulfanílico e $\alpha$-naftilamina. Os nitritos presentes no meio vão reagir com esses reagentes produzindo uma mudança de cor imediata para vermelho, caracterizando uma reação positiva. Entretanto, se a cultura não sofrer a alteração de cor existem duas possibilidades: o microrganismo possui enzimas que reduziram os nitratos a nitritos e estes foram transformados em amônia ou a nitrogênio molecular ou os nitratos não foram reduzidos pelo microrganismo. Para determinar se os nitratos foram ou não reduzidos a nitritos, adiciona-se uma pequena quantidade de zinco em pó à cultura incolor que já contém os reagentes. O zinco reduz os nitratos a nitritos, e o aparecimento de uma cor vermelha revelando que os nitratos não foram reduzidos a nitritos pelo microrganismo, caracterizando uma reação negativa. Por outro lado, se a adição de zinco não produzir uma mudança de cor indica que os nitratos já tinham sido reduzidos a nitritos e este a amônia ou a azoto e isto também caracteriza uma reação positiva (LENGELER et. al., 1999).

### 2.4.3 ANÁLISE MOLECULAR

É uma analise diretamente ligada ao DNA ou RNA que tem sido muito usada na taxonomia moderna pelo avanço tecnológico (VANDAMME *et al*, 1996). Com o advento das técnicas de reação em cadeia da polimerase (PCR) (SAIKI et al., 1988) e seqüenciamento de DNA (SANGER et al., 1977), os métodos moleculares, especialmente aqueles baseados no estudo da seqüência do 16S rDNA, tornaram-se muito úteis na descoberta de novos microrganismos. Estas técnicas se baseiam na amplificação do 16S rDNA por PCR e posterior caracterização por seqüenciamento. Outros métodos empregados consistem na Análise de Restrição do rDNA Amplificado (ARDRA), no Polimorfismo do Tamanho do Fragmento de Restrição Terminal (TRFLP), na Amplificação Aleatória de DNA Polimórfico (RAPD), na Análise do Espaço Ribossomal Intergênico (RISA), na Eletroforese em Gel com Gradiente Desnaturante (DGGE), na Eletroforese em Gel de Gradiente de Temperatura (TGGE) e no Polimorfismo Conformacional de Fita Simples (SSCP) (LENGELER, DREWS, SCHLEGEL, 1999).

## 2.5 SISTEMAS DE DETECÇÃO AUTOMÁTICA DE BACTÉRIAS

Devido às inúmeras atividades e aplicabilidades das bactérias, a identificação das mesmas se tornou muito útil tanto na saúde, quanto em estudos ecológicos e no mercado biotecnológico. Porém, como já citado, a identificação de uma dada espécie requer a aplicação de diversas analises e, além disso, diferentes espécies podem apresentar morfologia e metabolismo idênticos. Assim a correta identificação pode envolver a utilização de inúmeros testes químicos para observação de um conjunto de complexo de características também já citado acima. Uma maneira de facilitar a aplicação de um número grande de analises é através da utilização de sistemas automatizados que permitam a identificação bacteriana de forma mais rápida e eficaz. Atualmente existem no mercado vários sistemas de detecção automática e semi-automáticas de bactérias. Os mais conhecidos são Sistema Vitek (bioMérieux™), Sistema Biolog (Biolog™), Phoenix (Becton Dickinson Diagnostic Systems) e as características básicas de cada um estão descritas abaixo.

### 2.5.1 PHOENIX

O equipamento **BD Phoenix™** é utilizado na identificação rápida de bactérias clinicamente significantes e à realização de testes de sensibilidade a antimicrobianos. Para isto, o sistema Phoenix fornece resultados rápidos sobre a maioria das bactérias aeróbias e anaeróbias facultativas, Gram-positivas e Gram-negativas. Para a identificação são utilizados 45 cavidades contendo substratos bioquímicos desidratados e 2 cavidades para controle de fluorescência, dessa forma a identificação utiliza diversos testes bioquímicos convencionais, cromogênicos e fluorogênicos para identificar o organismo. O teste de sensibilidade contém até 84 cavidades com agentes antimicrobianos desidratados e uma cavidade para controle de crescimento. O sistema utiliza um indicador de redox colorimétrico otimizado para os testes de sensibilidade e diversos indicadores colorimétricos e fluorométricos para a identificação. O sistema Phoenix contém além do hardware, um software que apresenta uma base de dados, onde fica armazenado o perfil de inúmeras espécies bacterianas de interesse clínico, sendo assim, o software utiliza essa base de dados para identificar as espécies bacterianas e casos de resistência a antibióticos (PHOENIX).

### 2.5.2 VITEK

No Sistema Vitek os métodos clássicos de identificação foram miniaturizados e adaptados para sistemas de teste que empregam codificação numéricas computadorizadas utilizando uma base de dados gravada na memória do sistema, sendo que o resultado pode ser atingido após 2-6 horas de incubação. O sistema utiliza cromógeno ou substratos fluorogênicos nos testes químicos realizados por esse método de identificação. Este sistema pode ser acoplado a um microprocessador que lê e interpreta os testes enzimáticos, proporcionando assim uma maior padronização, precisão e reprodutibilidade e velocidade do que outros sistemas de identificação convencionais (M. A. PFALLER., et al, 1991). O crescimento no cartão de teste de poços, resulta em mudanças bioquímicas do substrato que pode ser interpretado por um leitor de placas especializado (WalkAway 40) para produzir um perfil bioquímico (chamado de Bionúmero). Este perfil é comparado com os perfis de microrganismos conhecidos cadastrados na base de dados para gerar sua identificação.

### 2.5.3 BIOLOG

O sistema Biolog é utilizado para verificar a capacidade de um microrganismo em consumir até 95 diferentes fontes de carbono, e pode ser utilizado para caracterizar tanto organismos Gram positivos, como organismos Gram negativos, uma vez que existem placas Biolog específicas para caracterização de bactérias pertencentes a cada um dos dois grupos (VIDEIRA, ARAÚJO & BALDANI, 2007; GUCKERT *et. al*., 1996).

Para a realização do teste se utiliza uma microplaca na qual existem 95 poços, sendo que cada um desses poços contêm uma fonte de carbono pré-seca diferente e o corante redox azul de tetrazólio (GUCKERT *et. al*., 1996). O princípio do teste baseia-se em adicionar bactérias crescidas em condições e meio específico e suspendidas em fluído inoculante que faz parte do Kit BIOLOG em cada um dos poços da microplaca e verificar seu padrão de utilização das diferentes fontes carbonos (GUCKERT *et. al*., 1996). Quando uma fonte de carbono é oxidada pelo microrganismo, o corante azul de tetrazólio é reduzido, passando de incolor para roxo, e esta mudança é percebida por um leitor de placas que fornece os resultados que são então comparados em um banco de dados, fornecendo a provável identidade da bactéria (VIDEIRA, ARAÚJO, BALDANI, 2007; GRAHAM, HAYNES, 2005; GUCKERT *et. al*., 1996). Para obter o resultado é utilizado o sistema MicroLog, que compara o padrão dos testes chamado de “impressão digital metabólica” com a sua base de dados.

## 2.6 MINERAÇÃO DE DADOS

Mineração de dados (do inglês, data mining) é um processo que utiliza algoritmos para analisar grandes bases de dados de modo eficiente procurando extrair das mesmas conhecimento valioso. Uma das tarefas mais uteis da mineração de dados chama-se classificação. Seu objetivo é bastante simples: um programa de computador deve atribuir automaticamente uma classe para um objeto cuja classe seja desconhecida. A classificação consiste em associar objetos a um conjunto pré definido de classes de acordo com suas características (FAYYAD et al., 1996).

As aplicações da mineração de dados, na pratica incluem: aprovação de credito (classificar um cliente como *alto*, *médio* ou *baixo* risco para concessão de credito), filtro de spam (detectar se email é normal ou spam), detecção de fraudes (identificar se uma transação financeira é legal ou suspeita), medicina (auxiliar na definição do diagnostico), bioinformática (algoritmos de identificação da classe de proteínas).

O programa ou algoritmo criado para executar a tarefa de classificação é denominado classificador (GONÇALVES, 2013). Construir classificadores precisos e eficientes é um dos grandes desafios da mineração de dados e atualmente existem vários classificadores, como arvores de decisão, redes neurais, SVMs (Support Vector Machines), etc.

Alguns exemplos da utilização da mineração de dados:

1. Relação entre a compra de fraldas e cervejas na sexta-feira. Utilizando a técnica de mineração de dados a rede Wal-Mart de supermercados, descobriu que homens casados com idade entre 25 e 30 anos compravam fraldas e cerveja as sextas-feiras, no caminho do trabalho para casa. A rede então otimizou a posição das gôndolas nos pontos de vendas, colocando as estantes de fraldas ao lado das estantes de cervejas e com isto o consumo de ambos os produtos cresceu 30%.
2. Adequação do estoque de mercadorias nas redes de lojas de departamentos do Brasil. As grandes redes de lojas de departamentos que atuam no Brasil aplicou a estratégia da mineração de dados para realizar a adequação de seus estoques de mercadorias de acordo com o fluxo de vendas.com isso reduziram em media, de 51000 produtos para 14000 os produtos oferecidos em suas lojas. Foram encontradas anomalias tais como, *roupas de inverno e guarda chuvas encalhados no nordeste e eletrodomésticos 110v a venda em Santa Catarina, onde a corrente elétrica é 220v.*

### 2.6.1 EXTRAÇÃO DAS CARACTERÍSTICAS

Consiste em uma etapa essencial do processo de mineração de dados e pode ser definido como a captura das informações mais relevantes para fazer uma classificação de um dado fornecido (DEVIJVER, 1982). Envolve a simplificação do conjunto de dados obtido, de forma que seja possível descrevê-lo com mais precisão e menos dados.

Extração de características é um termo genérico para métodos de construção de combinações de valores para representar os dados com certa precisão (SEWELL, 2007).

Este processo é comumente utilizado em aprendizagem de máquina, onde é selecionado um subconjunto das funcionalidades existentes, a partir dos dados disponíveis, este então é utilizado na aplicação de um algoritmo de aprendizagem que validara o subconjunto. O melhor conjunto contém o menor número de dimensões que mais contribuem para a precisão; todo o restante deve ser descartado. Esta é uma fase importante do pré-processamento utilizado para o reconhecimento de padrões (SEWELL, 2007).

### 2.6.2 RECONHECIMENTO DE PADRÕES

Padrão é definido como um conjunto de características que descrevem um objeto ou um grupo de objetos (PANDYA, MACY, 1995). Um padrão pode ser desde um conjunto de medidas a um conjunto de observações, geralmente representado na forma de vetor. Tais características são semelhantes ente si (SOUZA, 1999).

A Inteligência Artificial utiliza-se do Reconhecimento de Padrões para analisar determinado conjunto de dados chamados de “conjunto de treinamento” e organizá-los de acordo com padrões. O reconhecimento de padrões visa classificar os dados baseados nas informações extraídas de padrões.

O reconhecimento de padrões é utilizado em varias áreas como:

- Processamento de sinais de voz
- Bioinformática
- Classificação de documentos
- Analise de imagem
- Reconhecimento Biométrico
- Automação industrial
- Mineração de dados
- Sensoriamento remoto
- Visão
- Geologia
- Identificação de assinaturas

Existem, hoje, muitas estratégias de reconhecimento de padrões, que se baseiam em técnicas matemáticas, estatísticas e/ou incorporadas à Inteligência Artificial (Redes Neurais, Conjuntos Difusos, etc.). (SOUZA, 1999).

## 2.7 WEKA

O software WEKA (*Waikato Environment for Knowledge Analysis*) foi desenvolvido na Universidade de Waikato, Nova Zelândia em 1993, para a mineração de dados. É um *software* livre (código aberto) desenvolvido na linguagem Java, dentro das especificações da GPL (*General Public License*). As suas características, bem como as técnicas nele implementadas são descritas de forma detalhada em Witten e Frank 2005.

Weka ao longo dos anos se consolidou como a ferramenta de mineração de dados mais utilizada em ambiente acadêmico. Seu ponto forte é a tarefa de classificação, mas

também é capaz de minerar regras de associação e clusters de dados. Pode ser utilizada no modo console ou através da interface gráfica Weka Explorer. Uma das suas características mais interessantes é o fato da ferramenta fornecer uma API bastante poderosa e flexível que permite a integração de suas classes a qualquer tipo de sistema Java (Weka API).

O sistema Weka possui vários algoritmos de classificação como: Naïve Bayes, árvores de decisão (ID3), redes neurais, k-Nearest Neighbor, Support Vector Machines (SVN), MLP, RBF, entre vários outros. Possui vários modos de exibição dos resultados, com geração de texto com os resultados da validação do algoritmo utilizado, neste texto também existe a matriz de confusão, onde é possível obter de forma rápido os acertos e erros obtidos (Witten & Frank, 2005).

O WEKA tem como objetivo agregar algoritmos provenientes de diferentes abordagens/paradigmas na subárea da inteligência artificial dedicada ao estudo da aprendizagem por parte de máquinas.

O formato ARFF é utilizado como padrão para estruturar as bases de dados manipuladas pelo sistema Weka.

### 2.7.1 FORMATO DO ARQUIVO ARFF

Este tipo de arquivo conter como primeiro campo o nome da base de conhecimento, logo após, os campos que representam os padrões (atributos), que podem ser de variados tipos. Em seguida, o campo com os nomes das classes. Este conjunto de campos compõe o cabeçalho (WITTEN & FRANK, 2005).

Após a apresentação do cabeçalho e exibido o conjunto de dados, conforme figura 9 abaixo:

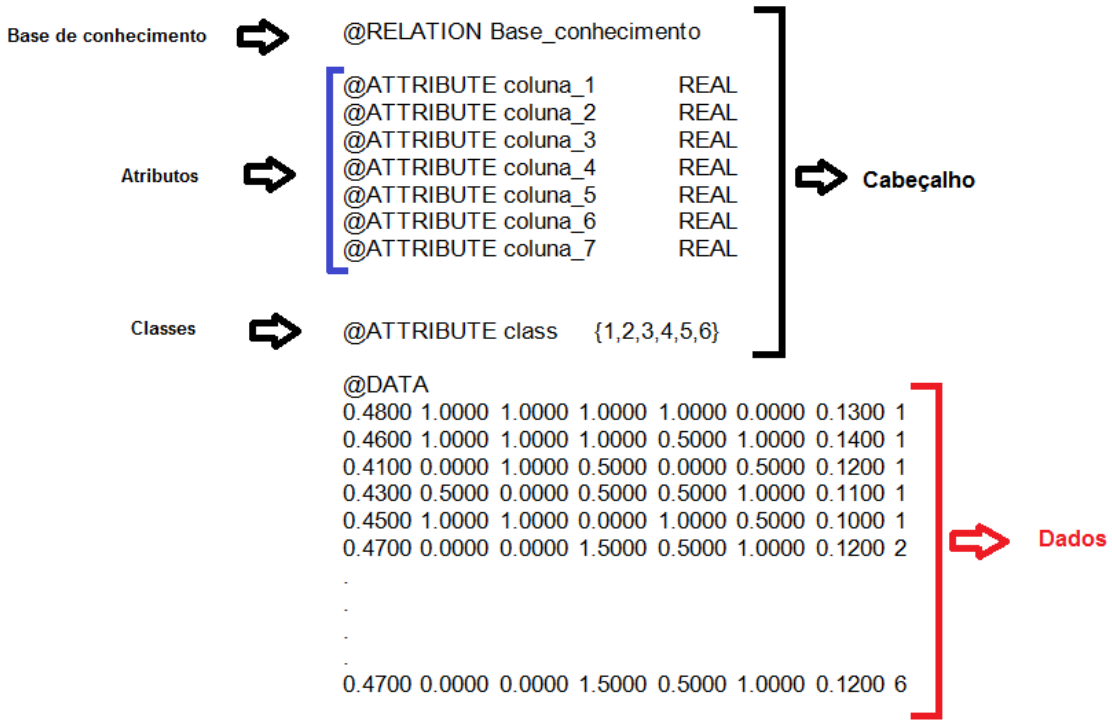



**Figura 9 – Arquivo .arff**

FONTE: Autor, 2013

## 2.8 REDES NEURAIS ARTIFICIAIS

Redes Neurais Artificiais são sistemas computacionais que foram inspirados na estrutura, no método de processamento e na habilidade de aprendizado de um cérebro biológico (CYBENKO, 1996). Baseiam-se em um modelo matemático que representa a estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência. Nas redes neurais artificiais, a idéia é realizar o processamento de informações tendo como princípio a organização de neurônios do cérebro. Como o cérebro humano é capaz de aprender e tomar decisões baseadas na aprendizagem, as redes neurais artificiais devem fazer o mesmo. Assim, uma rede neural pode ser interpretada como um esquema de processamento capaz de armazenar conhecimento baseado em aprendizagem (experiência) e disponibilizar este conhecimento para a aplicação em questão (ACHARYA et al., 2003). Uma grande rede neural artificial pode ter centenas ou milhares de unidades de processamento; já o cérebro de um mamífero pode ter muitos bilhões de neurônios (BRAGA, CARVALHO, LUDERMIR, 2000)

As figuras 10 abaixo representam o neurônio biológico e a figura 11 representa o modelo de um neurônio artificial

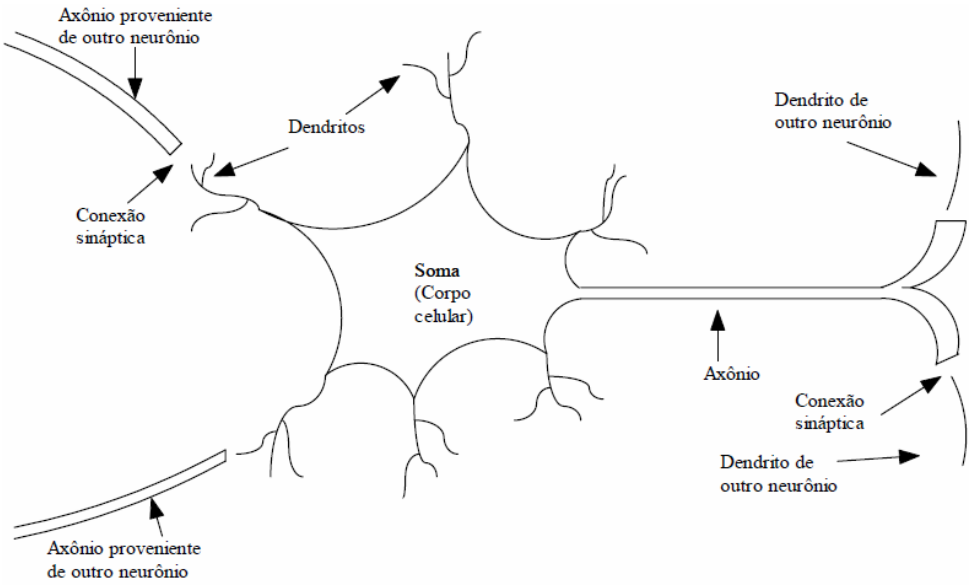



**Figura 10 - O NEURÔNIO BIOLÓGICO.**

FONTE: Adaptado de (FAUSETT, 1994)

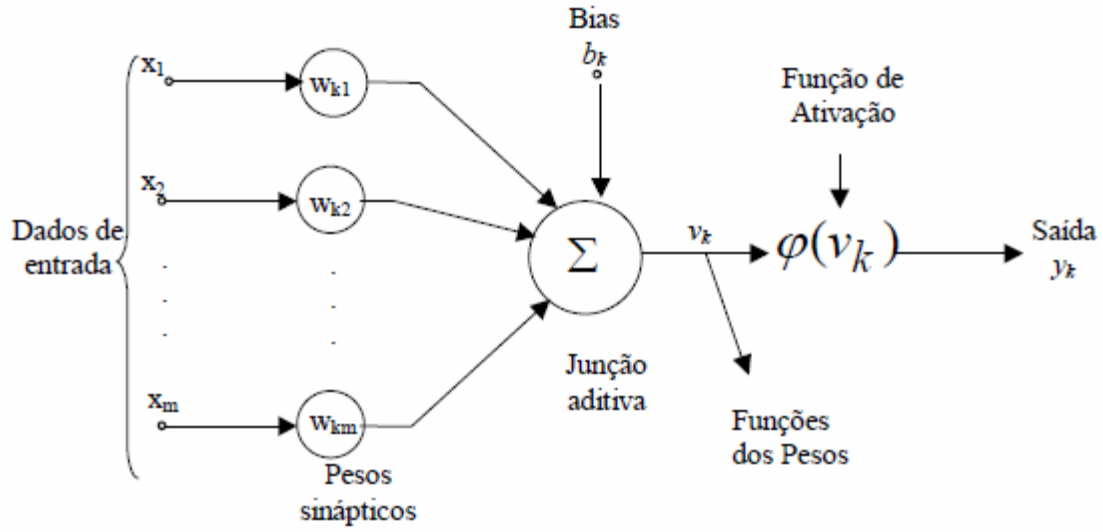



**Figura 11 – MODELO DE UM NEURÔNIO ARTIFICIAL.**

FONTE: Adaptado de Haykin, 2001 (HAYKIN, 2001).

Os elementos básicos de um neurônio artificial segundo MCCULLOCH e PITTS (1943) numa Rede Neural Artificial são: os pesos sinápticos, a função soma e a função de ativação, como mostra a Figura 11.

Basicamente, uma rede neural se assemelha ao cérebro em dois pontos: o conhecimento é obtido através de etapas de aprendizagem (HAYKIN, 2001) e pesos sinápticos são usados para armazenar o conhecimento. Uma sinapse é o nome dado à conexão existente entre neurônios. Nas conexões são atribuídos valores, que são chamados de pesos sinápticos. Isso deixa claro que as redes neurais artificiais têm em sua constituição uma série de neurônios artificiais (ou virtuais) que serão conectados entre si, formando uma rede de elementos de processamento (figura 12).

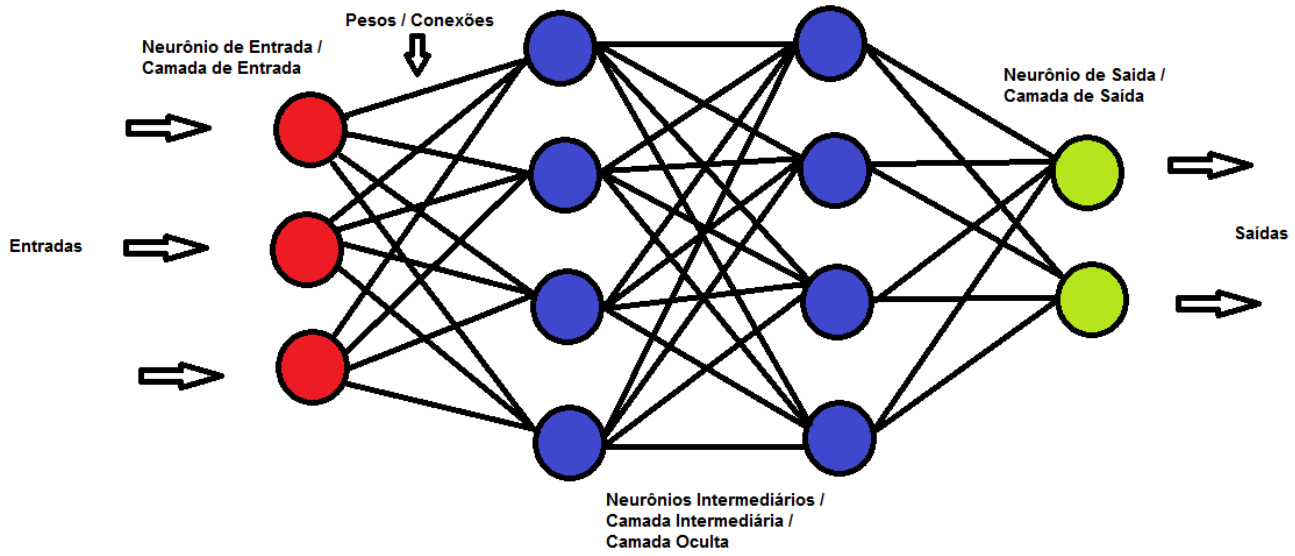



**Figura 12 – Esquema de Rede Neural**

FONTE: Autor, 2013

Com uma rede neural estabelecida, um conjunto de valores pode ser aplicado sobre um neurônio, sendo que este está conectado a outros pela rede. Estas entradas são multiplicadas no neurônio pelo valor do peso de sua sinapse (conexão), estes valores são somados e se o somatório ultrapassar o valor máximo estabelecido, um sinal é propagado pela saída (axônio) deste neurônio. Este processo é realizado com os demais neurônios da rede. Na pratica significa que os neurônios vão sofrer algum tipo de *ativação*, dependendo das entradas e dos pesos sinápticos.

O processo de aprendizagem das redes neurais é realizado quando ocorrem várias modificações significantes nas sinapses (pesos) dos neurônios. Essas alterações ocorrem de acordo com a ativação dos neurônios. Se determinadas conexões são mais usadas,

estas são reforçadas enquanto que as demais são enfraquecidas. Sempre que uma rede for ser utilizada para um fim, é necessário que ela seja treinada (ajuste dos pesos). Tipos de aprendizado nas redes neurais artificiais (ELMASRI e NAVATHE, 2005):

Supervisionado: a rede neural recebe um conjunto de entradas padronizados com os seus respectivos padrões de saída. Ocorrem os ajustes nos pesos sinápticos até que o erro entre os padrões de saída gerados pela rede tenha o valor desejado;

Não-supervisionado: a rede neural trabalha os dados de forma a determinar algumas propriedades dos conjuntos de dados. A partir destas propriedades é que o aprendizado é constituído;

A capacidade preditiva das redes neuronais não tem passado despercebida por nenhum ramo de atividade sendo utilizada nas mais variadas área como: telecomunicações, comércio, militar, turismo, robótica, visão, bioinformática, biologia, bolsa de valores, etc.

Abaixo, características de redes neurais utilizadas ou comentadas no presente trabalho. Todas elas utilizam o treinamento supervisionado.

### 2.8.1 REDE *FREE ASSOCIATIVE NEURONS (*FAN*)*

Free Associative Neurons (FAN) é um algoritmo que integra características de uma rede neural com técnicas de reconhecimento de padrões difusos (Fuzzy) e da lógica difusa (RAITTZ, 2002). FAN ganha em termos de inexatidão por trabalhar com granularidade de informação sendo capaz de incluir métodos diferentes de associação de padrões para aumentar capacidades de aprendizagem. Cada padrão de entrada é expandido em uma vizinhança difusa ao seu redor. Cada conjunto de vizinhança difusa é uma combinação de valores de características próximas às originais. A imprecisão mede o grau de similaridade entre o vizinho difuso e o padrão de entrada original (RAITTZ, 2002).

O processo de aprendizagem ocorre com a transformação dos dados difusos para o espaço FAN, é utilizado o reforço ou penalização. Graus de pertinência associam os padrões a cada neurônio representante de uma classe no domínio do problema. FAN associa características das Redes Neurais (aprendizado automático) e dos modelos difusos (representação da informação), ou seja, não necessita de configuração entre diferentes reconhecimentos de padrões (GUIZELINI et. al., 2011).

### 2.8.2 REDE *MULTILAYER PERCEPTRON* (MLP)

*Multilayer Perceptron* são redes que possuem uma ou mais camadas de neurônios entre as camadas de entrada e saída, a(s) chamada(s) camada(s) oculta(s) ou intermediária(s) (LIPPMANN, 1987). Este modelo difere do modelo original, com apenas um neurônio, o modelo Perceptron. Segundo CIBENKO, 1989 uma rede com uma camada intermediaria pode implementar qualquer função continua, e com duas camadas intermediarias é possível aproximar qualquer função matemática. Então, a vantagem da inserção de camadas intermediarias é aumentar o poder computacional do modelo.

No modelo MLP todos os neurônio são ligados aos neurônios da camada subseqüente, não havendo ligação com os neurônios laterais (mesma camada) e também não ocorre realimentação. O processo de aprendizagem é iterativo, conhecido como aprendizagem por experiência, aonde os ajustes dos pesos sinápticos são obtidos através dos padrões de treinamento, visando melhorar a taxa de acerto para a próxima iteração (HAYKIN, 1999). De acordo com BASHEER; HAJMEER, 2000 a configuração da rede não é determinada previamente, ou seja, a quantidade de camadas escondidas bem como o número de neurônios é determinada por tentativa e erro. São feitos vários testes e a partir da analise dos resultados obtidos, a melhor configuração é escolhida. Outra dificuldade é a determinação do número ideal de ciclos de treinamento da rede, que também é determinado por tentativa e erro (BASHEER; HAJMEER, 2000). Caso ocorra um número muito grande de ciclos de treinamento, a rede pode entrar em um processo chamado de "memorização" dos padrões (super-treinamento - do inglês *overtraining*), perdendo a capacidade de generalização. Ao contrário, se um número muito pequeno de ciclos for aplicado, a rede torna-se incapaz de representar os dados. O super ajuste do inglês *overfitting* é a conseqüência do *overtraining*.

### 2.8.3 REDE *RADIAL BASIS FUNCTIONS* (RBF)

Assim, como a rede MLP, a rede *Radial Basis* Functions é uma rede neural multicamadas (VON ZUBEN; ATTUX, 2008). A principal diferença é que a rede RBF representa a informação de forma localizada, facilitando a interpretação dos parâmetros de cada uma das funções componentes. Ela possui duas camadas de processamento: a primeira, a entrada é mapeada na camada intermediaria e na camada de saída é obtida uma combinação linear dos valores resultantes da camada intermediaria. A camada intermédia geralmente utiliza funções gaussiana (AGUIAR et. al., 2007).

Diferentemente das redes MLP as redes RBF trabalham o projeto de uma rede neural como um problema de ajuste de curvas (aproximação) em um espaço de alta dimensionalidade (HAYKIN, 1999). As redes RBF podem ser aplicadas principalmente em classificação de padrões, em que as saídas da rede são encaradas como estimadores estatísticos (GUPTA; JIN; HOMMA, 2003).

### 2.8.4 SUPPORT VECTOR MACHINES (SVM)

As Máquinas de Vetores de Suporte (SVMs, do Inglês Support Vector Machines) constituem uma técnica de aprendizado embasada pela teoria de aprendizado estatístico, desenvolvida por (VAPNIK, 1995; CHERVONENKIS, 1971). Essa teoria estabelece uma série de princípios que devem ser seguidos na obtenção de classificadores com boa generalização, definida como a sua capacidade de prever corretamente a classe de novos dados do mesmo domínio em que o aprendizado ocorreu. As técnicas de aprendizado de maquina empregam um princípio de inferência denominado indução, no qual se obtém conclusões genéricas a partir de um conjunto particular de exemplos. O objetivo é aprender a representar (ou agrupar) as entradas submetidas segundo uma medida de qualidade. Essas técnicas são utilizadas principalmente quando o objetivo for encontrar padrões ou tendências que auxiliem no entendimento dos dados (SOUTO, 2003). Basicamente, o SVM é um algoritmo linear que constrói hiperplanos, com o objetivo de encontrar hiperplanos ótimos, ou seja, hiperplanos que maximizem a margem de separação das classes, para separar os padrões de treinamento em diferentes classes (WAN & CAMPBELL, 2000).

As SVMs vêm recebendo crescente atenção da comunidade de Aprendizado de Máquina (MITCHELL, 1997), pois os resultados da aplicação dessa técnica são comparáveis e muitas vezes superiores aos obtidos por outros algoritmos de aprendizado, como as Redes Neurais Artificiais (RNAs). Exemplos de aplicações de sucesso podem ser encontrados em diversos domínios, como na categorização de textos, na análise de imagens e em Bioinformática.

### 2.8.5 ARVORE DE DECISÃO J48

As árvores de decisão classificam instâncias partindo da raiz da árvore para algum nodo folha que fornece a classe da instância. Cada nodo da árvore especifica o teste de algum atributo da instância, e cada arco alternativo que desce daquele nodo corresponde a um dos possíveis valores deste atributo. Uma instância é classificada começando no nodo raiz da árvore e testa o atributo relacionado a este nodo e segue o arco que corresponde ao

valor do atributo na instância em questão. Este processo é repetido então para a sub-árvore abaixo até chegar a um nodo folha.

O algoritmo J48 é a implementação em Java para o Weka da arvore de decisão C4.5 (QUINLAN, 1993) que, por sua vez, é uma significativa evolução do ID3. O algoritmo ID3 é baseado no conceito estatístico de entropia e no conceito de ganho. O algoritmo C4.5 lida tanto com atributos categóricos (ordinais ou não-ordinais) como com atributos contínuos. Para lidar com atributos contínuos, o algoritmo C4.5 define um limiar e então divide os exemplos de forma binária: aqueles cujo valor do atributo é maior que o limiar e aqueles cujo valor do atributo é menor ou igual ao limiar. Também permite que os valores desconhecidos para um determinado atributo sejam representados como '?', que são então tratados de forma especial. Esses valores não são utilizados nos cálculos de ganho e entropia (WITTEN & FRANK, 2005).

### 2.8.6 OVERFITTING

Overfitting (super ajuste) é um fenômeno que aparece como resultado de overtrainig (super treinamento), mas não só neste caso, pois, pode ocorrer quando muitos parâmetros são utilizados para determinar um conjunto de características (modelo). Sua principal conseqüência é a memorização dos padrões pela rede, e com isto a perda da capacidade de generalização.

Para detectar e evitar o overfitting o conjunto de dados deve ser dividido em dois subconjuntos um para treinamento e o outro para os testes, permitindo assim uma avaliação final e a obtenção de uma taxa real de acertos na classificação (REZENDE, 2005).

A seguir são comentados os métodos de divisão do conjunto de dados para a avaliação do modelo.

### 2.8.7 VALIDAÇÃO CRUZADA

A validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo, a partir de um conjunto de dados. Esta é a principal técnica e é amplamente empregada em problemas onde o objetivo da modelagem é a predição. Permite estimar o quão preciso é um modelo na prática, ou seja, o seu desempenho para um novo conjunto de dados.

A técnica consiste na divisão do conjunto de dados em subconjuntos mutuamente exclusivos, e posteriormente, na utilização de alguns destes subconjuntos para o treinamento e o restante dos subconjuntos para validação ou teste. O modelo é avaliado a partir dos resultados obtidos desta combinação.
Muitas são as maneiras de realizar a divisão dos dados, mas somente três são as mais utilizadas: o método *holdout*, o *k-fold* e o *leave-one-out* (KOHAVI, 1995).

#### 2.8.7.1 HOLDOUT

Este método é bem simples e consiste em dividir o conjunto total de dados em dois subconjuntos mutuamente exclusivos, um para treinamento e outro para teste. O conjunto de dados fornece dados para o treinamento da técnica utilizada e o conjunto de teste fornece dados novos, para testar a generalização do modelo. Geralmente uma medida muito utilizada é considerar 2/3 dos dados para treinamento e o 1/3 restante para teste (KOHAVI, 1995). Após a divisão dos conjuntos, a estimação do modelo é realizada (treinamento) e, posteriormente, os dados de teste são aplicados (validação) e o erro de predição calculado (THEODORIDIS & KOUTROUMBAS, 2003).

O resultado da avaliação pode depender, por exemplo, em que ponto terminou os dados de treinamento e começaram os dados de teste, ou seja, da quantidade de padrões existente em cada conjunto, pois pode ocorrer que no conjunto de treinamento não exista nenhum padrão representando classes do conjunto de teste. Outro fator que influencia na avaliação é a quantidade de padrões existentes de cada classe no conjunto de treinamento. Por exemplo, uma classe A com uma grande quantidade de padrões deverá influenciar mais o resultado final, ao contrário de uma classe B com poucos padrões do seu tipo. Neste caso a rede treinada terá uma melhor generalização para os dados da classe A do que para a classe B.

Esta abordagem é mais indicada quando existe uma farta quantidade de dados. Caso o conjunto total de dados seja pequeno, o erro calculado na predição pode sofrer muita variação.

### 2.8.7.2 K-FOLD

Este método consiste em dividir o conjunto total de dados em *k* subconjuntos mutuamente exclusivos e do mesmo tamanho e, a partir disto, um subconjunto é utilizado para teste e os *k-1* restantes são utilizados para estimação dos parâmetros. Calcula-se a acurácia do modelo. Este processo é realizado *k* vezes alternando de forma circular o subconjunto de teste. Por exemplo, se K = 4, a rede será treinada quatro vezes, na primeira vez o primeiro grupo será usado para teste e os outros três serão usados para treinamento. Na segunda vez, o segundo grupo será para teste e os outros três serão para treinamento, e assim sucessivamente. Uma demonstração gráfica está na figura 13 (DELEN, 2003; KOHAVI, 1995).

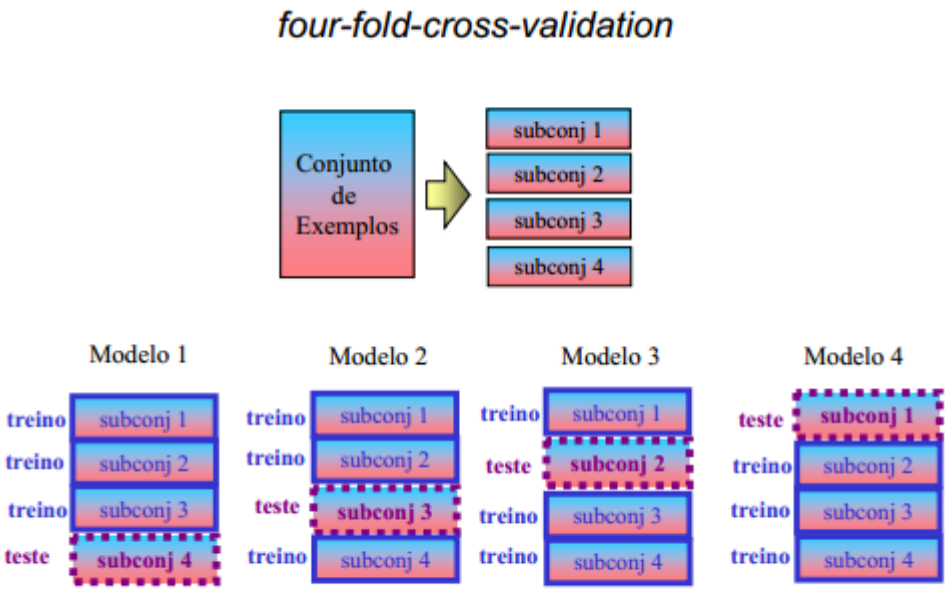



**Figura 13 –four-fold-Cross-validation.**

FONTE: Adaptado de http://www.inf.ufrgs.br/~alvares/CMP259DCBD/avaliacao.pdf

Ao final das *k* iterações calcula-se a acurácia sobre os erros encontrados, através da equação descrita anteriormente, obtendo assim uma medida mais confiável sobre a capacidade do modelo de representar o processo gerador dos dados

A vantagem de usar o método ao invés do método holdout é que nele o treinamento é feito com todos os dados, e por isso gera um resultado mais confiável, uma vez que no

método holdout os dados são divididos e essa divisão pode não gerar um resultado representativo dos padrões.

#### 2.8.7.3 LEAVE-ONE-OUT

O método *leave-one-out* é uma simplificação do *k-fold*, com *k* igual ao número total de dados *N (*KOHAVI, 1995). Onde os N padrões são divididos em dois conjuntos, o primeiro com somente um padrão e o segundo com todos os outros restantes (N-1). A rede é treinada com os N-1 padrões (segundo conjunto) e testada com o primeiro grupo que contem somente um elemento e o processo é refeito para todos os padrões do modelo. Nesta abordagem são realizados *N* cálculos de erro, um para cada dado.

Apesar de apresentar uma investigação completa sobre a variação do modelo em relação aos dados utilizados, este método possui um alto custo computacional, sendo indicado para situações onde existem poucos dados disponíveis.

### 2.8.8. BOOTSTRAP

O método bootstrap, introduzido por Efron (1979), é um método de reamostragem baseado na construção de sub-amostras a partir de uma amostra inicial. Consiste em retirar da amostra inicial (A) uma pseudo amostra com reposição, aonde cada elemento é retirado de forma aleatória. Esta amostra é chamada de A*, e o processo é repetido varias vezes, são feitos cálculos estatísticos para cada nova amostra gerada. (Silva Filho, 2000 ).

Segundo Breiman (1996), o ideal entretanto é utilizar replicas bootstrap de tamanho igual ao conjunto de dados original de aprendizagem, mesmo sendo de tamanho igual a amostra usara somente cerca de 63% dos exemplos. Para o conjunto de testes são utilizadas as instancias que não foram selecionadas no conjunto de treinamento (cerca de 27%).

Não existe um consenso em quantas replicas bootstrap utilizar, mas quanto maior for o número de classes, maior é a quantidade de necessária de replicadas bootstrap, pois tratando de redes neurais a convergência é mais lenta, nos testes feitos por Breiman(1996), o verificado é que a partir da vigésima quinta replica a media dos resultados sofre pouca alteração.

Observe-se que a reamostragem não adiciona nenhuma informação nova a amostra original. Pode parecer que o bootstrap crie dado a partir do nada, entretanto ele não utiliza as observações das reamostras como se elas fossem dados reais. O bootstrap não é um

substituto para o acréscimo de dados com vistas ao aumento da precisão, em vez disso, a idéia do bootstrap é a de se empregarem as medias das reamostras para se estimar como a media amostral de uma amostra de tamanho N, extraída dessa população, varia em decorrência da amostragem aleatória. Uma desvantagem é a falta de controle sobre a especialização produzida pela rede. (Silva Filho, 2000).

A figura 14 representa o funcionamento do bootstrap.

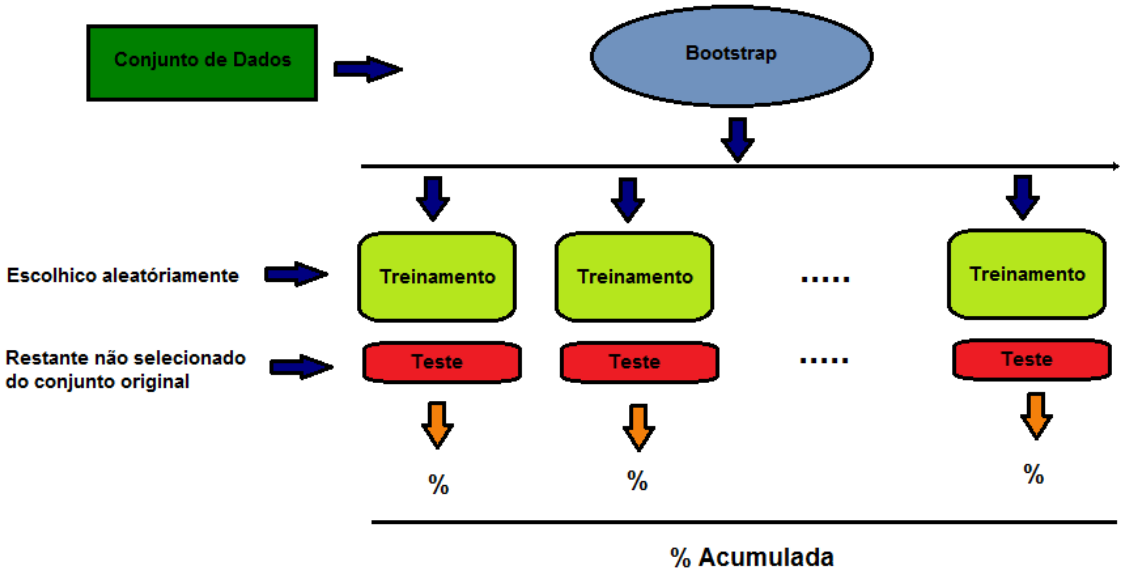



**Figura 14 – Bootstrap**

FONTE: Autor, 2013

O método Bootstrap também é conhecido como Bagging, que é o acrônimo de *"Bootstrap Aggreganting"*.

## 2.9 BANCO DE DADOS POSTGRESQL

PostgreSQL é um dos bancos de dados livre mais avançado do mundo e é utilizado por grandes empresas publicas brasileiras: Caixa Econômica Federal, Ministério da Saúde (Datasus), Serpro, Banco do Brasil, Celepar, Metrô de São Paulo, projeto SIVAM (Sistema de Vigilância da Amazônia), etc.

O pgAdmin é um software gráfico para administração do Sistema Gerenciador de Banco de Dados PostgreSQL (SGDB PostgreSQL), disponível para Windows e UNIX, que possui muitos recursos e onde é possível manipular todas as funcionalidades graficamente, permitindo ao usuário visualizar as consultas e históricos dos comandos efetuados, entre outros tantos recursos. Esta característica lhe confere segurança e facilidade de execução.

## 2.10 LINGUAGEM DE PROGRAMAÇÃO JAVA

Java é uma linguagem de programação muito utilizada no mundo e possui alguns diferenciais que a destacam (DEITEL & DEITEL, 2005):

- Orientação a objeto: Permite um maior reaproveitamento de código, possui componentes bem modularizados com funções bem definidas e com propósitos claros e delimitados, o que permite fazer um software com menos código e conseqüentemente com menor custo de manutenção;

- Portabilidade: Permite que software possa funcionar em vários sistemas operacionais (Independência de plataforma), pois o código escrito em Java é compilado em um "bytecode" que é executado por uma máquina virtual;

- Recursos de rede: Possui bibliotecas para todos os protocolos de rede;

- Segurança: Recursos de rede com criptografia e vários protocolos de validação de acesso.

### 2.10.1. NETBEANS

O NetBeans é um ambiente de desenvolvimento integrado (IDE) gratuito e de código aberto para desenvolvedores de software. Este IDE é executado em muitas plataformas, como Windows, Linux, Solaris e MacOS, e oferece ferramentas necessárias para criar aplicativos profissionais de desktop, Web e móveis. Em 1999 foi adquirido pela Sun Microsystems e transformado em código aberto, tornando-o uma plataforma OpenSource.

Desde então, a comunidade de desenvolvedores que utilizam esta plataforma contribuem para a ampliação do projeto original e por isso, tornou-se uma das IDEs mais populares.

# 3. MATERIAIS E MÉTODOS

## 3.1 CONSTRUÇÃO DA FERRAMENTA PARA POSICIONAMENTO TAXONÔMICO DE BACTÉRIAS

Para a construção do software que permite o posicionamento taxonômico de bactérias, primeiramente foram levantados os requisitos necessários para o seu funcionamento adequado. Estes requisitos foram denominados REQ1 a ReQ14 e estão listados no quadro 2.

**Quadro 2 – Requisitos do Sistema**

| Requisito | Descrição |
|---|---|
| REQ1 | Cadastro os artigos que contem a descrição das espécies, salvando o seu arquivo PDF |
| REQ2 | Cadastro dos gêneros |
| REQ3 | Cadastro das espécies |
| REQ4 | Cadastro das categorias |
| REQ5 | Cadastro das características (testes) |
| REQ6 | Cadastro dos resultados das características |
| REQ7 | Cadastro das categorias deve ser rápido e otimizado, não permitindo duplicidade |
| REQ8 | Treinamento da rede |
| REQ9 | Salvar a rede treinada |
| REQ10 | Posicionamento taxonômico em nível de gênero através da informação dos resultados das características utilizando técnicas de IA – Redes Neurais |
| REQ11 | Geração dos relatórios de características e resultados |
| REQ12 | Visualização dos PDFs dos artigos |
| REQ13 | Validação do modelo usando cross validation (leave-one-out) |
| REQ14 | Validação do modelo usando bootstrap |

FONTE: Autor, 2013

Com base na lista de requisitos foi criado o Diagrama de Casos de Uso que está apresentado na Figura 15.

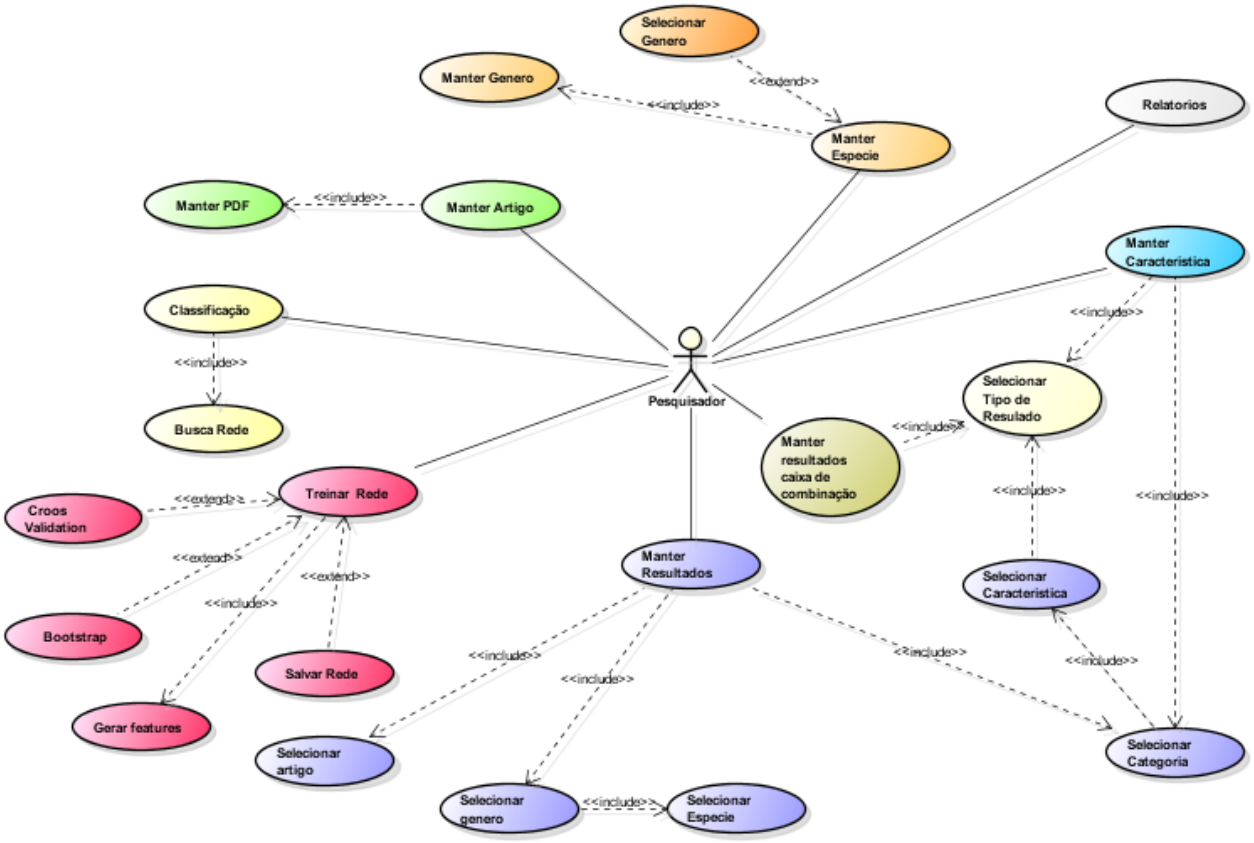



**Figura 15 – Diagrama de casos de uso**

FONTE: Autor, 2013

Neste diagrama estão representadas todas as funcionalidades do sistema, onde o usuário esta representado pelo pesquisador. As funcionalidades estão descritas abaixo:

1. Manter artigo: representa todas as opções da função Artigo (inclusão, alteração e exclusão) e sempre que esta função é chamada a função “Manter PDF", que é responsável por guardar os arquivos no formato PDF, também é utilizada.
2. Manter espécie: representa todas as opções da função Espécie (inclusão, alteração e exclusão) e sempre que esta função for chamada a função “Selecionar gênero” também é utilizada. Esta função é responsável pela busca de todos os gêneros cadastrados e caso o gênero não esteja cadastrado é possível utilizar a função “Manter gênero” que é responsável por todas as opções da função.

3. Relatórios: representa as funções Relatório. O usuário pode obter dois relatórios, um referente a todos os resultados relativos a uma dada característica e outro referente a todos os resultados de uma dada espécie.
4. Manter característica: representa todas as opções da função Características (inclusão, alteração e exclusão) e sempre que esta função é utilizada as funções Selecionar categoria e Selecionar tipo de resultado são utilizadas. Estas são responsáveis pela busca das categorias cadastradas e pela busca dos tipos de todos os tipos de resultados, respectivamente.
5. Manter resultados caixa de combinação: representa as opções da função (inclusão, alteração e exclusão) e sempre que esta função for utilizada a função Selecionar tipo de resultado será chamada.
6. Manter resultados: representa todas as opções da função (inclusão, alteração e exclusão) e sempre que esta função é utilizada são chamadas as funções Selecionar Artigo, Selecionar Espécie e Selecionar Característica. A função Selecionar Gênero é utilizado através da função Selecionar Espécie, e a função Selecionar Categoria é utiliza através da função Selecionar Característica, que por sua vez também utiliza a função Selecionar Tipo de Resultado.
7. Treinar rede: representa as opções de treinamento da rede e sempre que esta função é ativa a função Gerar Features também é utilizado. Nesta etapa também é possível utilizar as funções Salvar Rede, Cross validation e Bootsrap.
8. Classificação: representa a função de classificação e sempre que esta opção é utilizada é chamada a função Busca Rede, que retorna a rede.

A linguagem escolhida para a codificação foi Java e o ambiente IDE de desenvolvimento foi o Netbeans, ambos amplamente utilizados na comunidade acadêmica e reconhecidamente eficientes. Para a construção da base de dados foi escolhido o SGBD Postgresql.

O sistema foi projetado para utilizar a arquitetura Model-view-controller (MVC) que é considerado um padrão de projeto (do inglês, Design Pattern) e é atualmente muito utilizado. Este modelo isola a lógica da aplicação da interface do usuário, permitindo desenvolver, editar e testar separadamente cada parte. Para isto foram criados três pacotes que representam as camadas:

- **Controlador** (controller): É responsável por controlar todo o fluxo de informação que passa pelo sistema. Basicamente executa a regra de negócio (modelo) e repassa a informação para a visualização (visão).

- **Modelo** (model): É utilizado para manipular informações de forma mais detalhada, sendo recomendado que, sempre que possível, se utilize dos modelos para realizar consultas, cálculos e todas as regras de negócio do sistema. É o modelo que tem acesso a toda e qualquer informação sendo essa vinda de um banco de dados.

- **Visão** (view) : É responsável por tudo que o usuário visualiza.

A camada visão possui as classes Java responsáveis pela visualização e manipulação da interface do sistema. A camada modelo possui a classe de conexão com o banco de dados e as classes de manipulação dos dados (persistência). A camada controladora possui as classes que manipulam as classes do modelo (regra de negocio) e possui ligação com o pacote *FAN (*Figura 16).

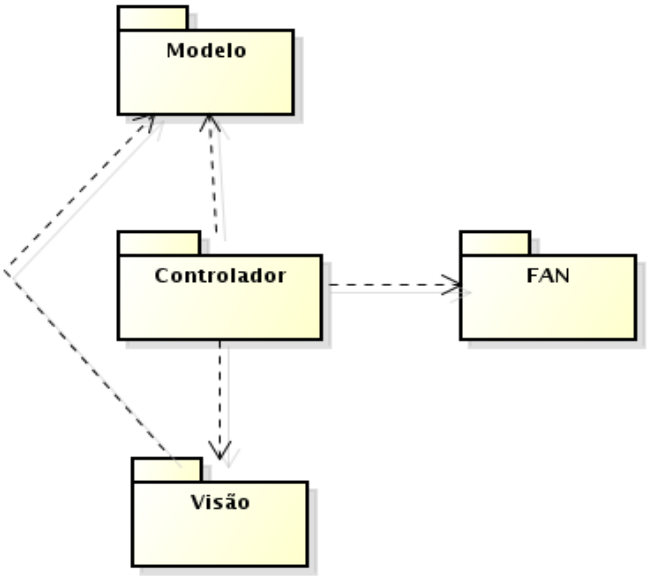



**Figura 16- Diagrama de Pacotes**

FONTE: Adaptado de Reenskaug, 1979

Definida a arquitetura do sistema, foi desenvolvido o diagrama de classes da camada modelo. Esta camada contém os objetos que devem ser persistidos no banco de dados (figura 17).

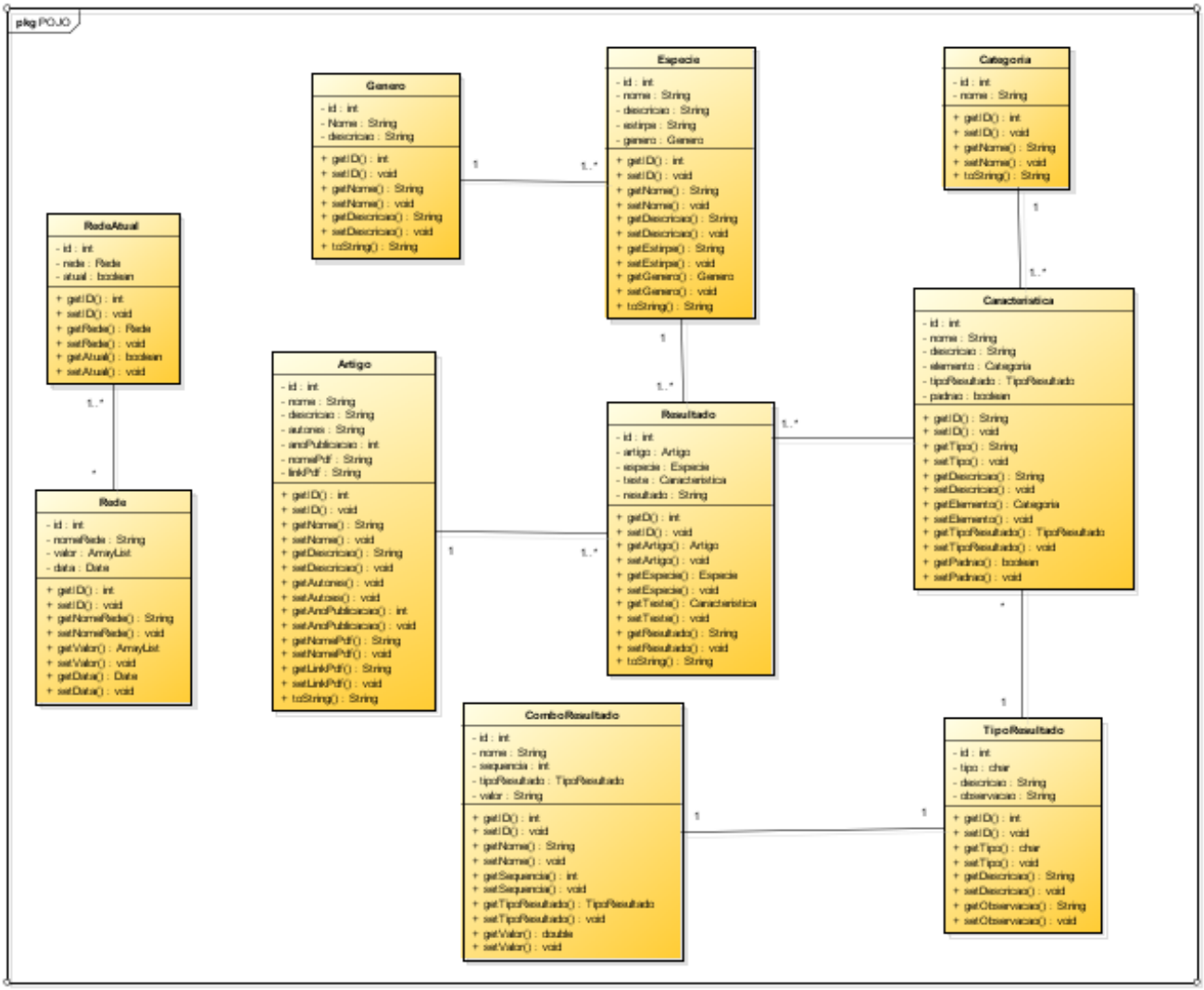



**Figura 17 – Diagrama de Classes**
FONTE: Autor, 2013

Diagrama de classes contendo as classes que serão persistidas na base de dados: Artigo, Resultados, Espécies, Gêneros, Categorias, Características, TipoResultados e ComboResultados

Com base no diagrama de classes da camada modelo foi desenvolvido o Diagrama Entidade-Relacionamento (DER) (figura 18), definindo-se assim quais eram os dados que deveriam ser armazenados na base de dados.

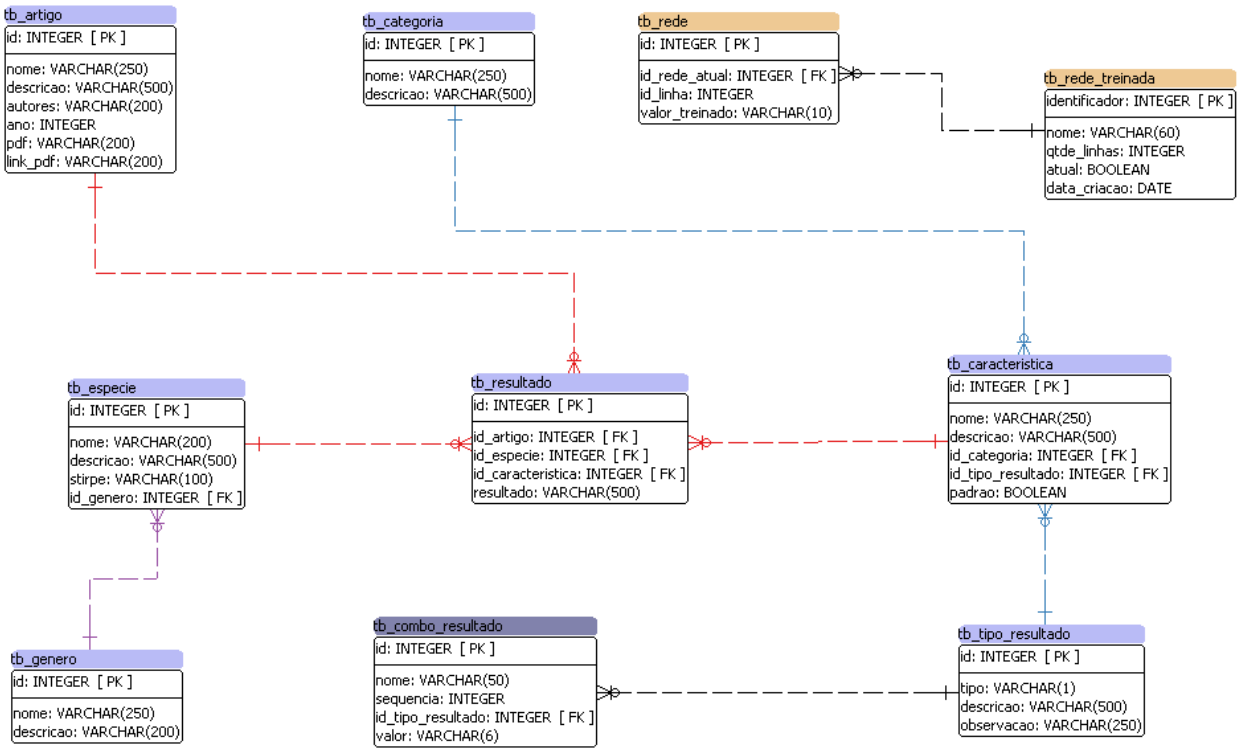



**Figura 18 – Base de dados**

FONTE: Autor, 2013

Diagrama de Entidade e Relacionamento, contendo as tabelas que guardaram os resultados dos testes que diferencias as espécies. A tabelas são: TB_genero, TB_Combo_Resultado, TB_Tipo_Resultado, TB_Caracteristica, TB_Resultado, TB_Especie, TB_Rede_Treinada, TB_Rede, TB_Categoria, TB_Artigo. As ligações entre as tabelas são destacadas.

Este Diagrama Entidade-Relacionamento está estruturado por um conjunto de dez tabelas cujos campos estão discriminados abaixo.

**Quadro 3 – Quadro da tabela artigo**

| **Tabela:** | TB_Artigo | |
|---|---|---|
| **Descrição:** | Tabela responsável por armazenar a descrição dos artigos | |
| **Campo** | **Domínio** | **Descrição** |
| Id | Numérico | Identificador do registro |
| Nome | Texto | Nome do artigo em pdf |
| descricao | Texto | Descrição do artigo, ou observação |
| autores | Texto | Nome dos autores do artigo |
| Ano | Numérico | Ano de publicação |
| link_pdf | Texto | Endereço do pdf |

FONTE: Autor, 2013

**Quadro 4 – Quadro da tabela Categoria**

| **Tabela:** | TB_Categoria | |
|---|---|---|
| **Descrição:** | Tabela responsável por armazenar as categorias que agruparão as características (testes) | |
| **Campo** | **Domínio** | **Descrição** |
| Id | Numérico | Identificador do registro |
| Nome | Texto | Nome da categoria |
| descricao | Texto | Descrição da categoria |

FONTE: Autor, 2013

**Quadro 5 – Quadro da tabela Característica**

| **Tabela:** | TB_Caracteristica | |
|---|---|---|
| **Descrição:** | Tabela responsável por armazenar a descrição das características (testes) | |
| **Campo** | **Domínio** | **Descrição** |
| Id | Numérico | Identificador do registro |
| Nome | Texto | Nome do artigo em pdf |
| descricao | Texto | Descrição do artigo, ou observação |
| id_categoria | Numérico | Identificador de registro da tabela TB_Categoria, indica a qual categoria a característica pertence |
| id_tipo_resultado | Numérico | Identificador de registro da tabela TB_Tipo_Resultado, indica o tipo de resultado da tabela |
| Padrao | Boolean | Indica se a característica será utilizada para geração (treinamento e classificação) da rede neural |

FONTE: Autor, 2013

**Quadro 6 – Quadro da tabela Tipo Resultado**

| **Tabela:** | TB_Tipo_Resultado | |
|---|---|---|
| **Descrição:** | Tabela responsável por armazenar os tipo de resultados | |
| **Campo** | **Domínio** | **Descrição** |
| id | Numérico | Identificador do registro |
| tipo | Texto | Nome da categoria |
| descricao | Texto | Descrição do tipo de resultado |
| observação | Texto | Observação ou informação que julgue necessário |

FONTE: Autor, 2013

**Quadro 7 – Quadro da tabela Combo Resultado**

| **Tabela:** | TB_Combo_Resultado | |
|---|---|---|
| **Descrição:** | Tabela responsável por armazenar as opções da tabela TB_Tipo_Resultado, esta tabela será controlada pela aplicação, pois somente tipos de resultados do tipo combo (caixa de seleção) serão armazenados | |
| **Campo** | **Domínio** | **Descrição** |
| Id | Numérico | Identificador do registro |
| Nome | Texto | Nome da categoria |
| sequencia | Numérico | Ordem de exibição |
| id_tipo_resultado | Numérico | Identificador do registro da TB_Tipo_Resultado |

FONTE: Autor, 2013

**Quadro 8 – Quadro da tabela Espécie**

| **Tabela:** | TB_Especie | |
|---|---|---|
| **Descrição:** | Tabela responsável por armazenar as espécies | |
| **Campo** | **Domínio** | **Descrição** |
| Id | Numérico | Identificador do registro |
| Nome | Texto | Nome da espécie |
| descricao | Texto | Descrição da espécies |
| Estipe | Texto | Estirpe da espécies |
| id_genero | Numérico | Identificador do registro da TB_Genero, indica a qual gênero a espécie pertence |

FONTE: Autor, 2013

**Quadro 9 – Quadro da tabela Gênero**

| **Tabela:** | TB_Genero | |
|---|---|---|
| **Descrição:** | Tabela responsável por armazenar os gêneros | |
| **Campo** | **Domínio** | **Descrição** |
| Id | Numérico | Identificador do registro |
| Nome | Texto | Nome do gênero |
| descricao | Texto | Descrição do gênero |

FONTE: Autor, 2013

**Quadro 10 – Quadro da tabela Resultado**

| **Tabela:** | TB_Resultado | |
|---|---|---|
| **Descrição:** | Tabela responsável por armazenar os resultados dos testes contidos nos artigos | |
| **Campo** | **Domínio** | **Descrição** |
| id | Numérico | Identificador do registro |
| id_artigo | Texto | Identificador da registro da TB_Artigo, indica a qual artigo pertence o resultado |
| id_especie | Texto | Identificador da registro da TB_Especie, indica a qual espécie pertence o resultado |
| id_caracteristica | Texto | Identificador da registro da TB_Artigo, indica a qual característica pertence o resultado |
| resultado | Texto | Resultado da característica |

FONTE: Autor, 2013

## 3.2 Funcionalidades da Ferramenta

A ferramenta para posicionamento taxonômico de bactérias apresenta as seguintes funcionalidades:

### 3.2.1 Cadastro dos Artigos

O cadastro dos artigos foi implementado para permitir a gravação de dados importantes sobre o artigo: título, ano de publicação e autores. Todo artigo cadastrado configura um registro único e também fica armazenado em formato PDF (*Portable Document Format),* disponível para consulta pelo usuário. Esta lista de artigos cadastrados pode ser consultada de forma rápida e organizada sendo possível, editar, excluir e visualizar o conteúdo de interesse. Também foi inserido um campo onde o usuário pode registrar observações.

### 3.2.2 Cadastro das Espécies

O cadastro das espécies foi idealizado para ser rápido e fácil de utilizar. Contem índice para consulta geral, onde os todos os cadastros podem ser filtrados pelo gênero bacteriano. Para o cadastro uma nova espécie basta informar o nome da espécie, a estirpe (se for conhecida) e o gênero a qual pertence. Se necessário é possível utilizar o campo

observações. Caso a espécie seja de um gênero ainda não cadastrado, ao lado da caixa de combinação existe a opção de cadastro de novo gênero, onde basta informar o nome e se necessário utilizar o campo de observações. Também é possível excluir o gênero, desde que não esteja vinculado a nenhuma espécie.

#### *3.2.2.1 Espécies de bactérias cadastradas*

Foram cadastradas 304 estirpes de 228 espécies de bactérias pertencentes 10 gêneros diferentes. Os microrganismos cadastrados estão listados no anexo 1.

Os dados referentes as bactérias cadastradas foram coletados do *International Journal of Systematic and Evolutionary Microbiology* (IJSEM), um periódico oficial para caracterizações taxonômicas, descrições de novos taxa e reclassificações de procariotos. O IJSEM é o periódico oficial de registro de nomes de bactérias do Comitê Internacional em Sistemática de Procariontes (ICSP) da União Internacional da Sociedade de Microbiologia (IUMS).

Os resultados dos variados testes de classificação das espécies utilizadas, foram obtidas neste periódico e estavam descritas em diversos artigos. Foram escolhidos, preferencialmente, gêneros de bactérias que contem espécies diazotróficas, ou seja, bactérias capazes de realizar a fixação biológica de nitrogênio. A maioria dos artigos utilizados para a extração das características e dos respectivos resultados (testes) que levaram à classificação da bactéria, apresenta os dados na forma de tabela (Figura 19).

| Characteristic | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| Biotin requirement | − | + | − | + | − | − | + | − | − |
| Growth in 3 % NaCl | + | − | − | − | V | − | + | + | − |
| Carbon source: | | | | | | | | | |
| *N*-Acetylglucosamine | + | + | + | ND | − | V | ND | + | − |
| L-Arabinose | + | + | + | + | V | + | V | + | ND |
| D-Cellobiose | + | − | − | − | − | + | ND | + | − |
| D-Fructose | + | + | + | + | + | + | + | V | + |
| L-Fucose | + | V (+)* | − | ND | − | + | ND | + | ND |
| D-Galactose | + | + | + | + | V | + | − | + | V |
| Gentiobiose | + | − | + | ND | − | + | ND | + | − |
| D-Gluconate | + | +(−)* | − | ND | + | − | ND | − | V |
| D-Glucose | + | + | + | + | V | + | − | + | V |
| Glycerol | + | + | + | ND | + | − | + | − | + |
| *myo*-Inositol | + | V (+)* | − | − | − | + | ND | − | − |
| Lactose | + | − | − | − | − | V | ND | + | − |
| Maltose | + | − | − | − | − | + | ND | + | − |
| D-Mannitol | + | + | + | − | − | − | + | − | + |
| D-Mannose | + | V (+)* | − | ND | − | + | + | + | − |
| L-Rhamnose | + | − | − | + | − | V | ND | + | − |
| D-Ribose | ND | + | + | + | − | + | + | V | − |
| D-Sorbitol | + | + | + | − | − | − | − | − | + |
| Sucrose | + | − | ND | − | − | + | − | + | − |
| D-Trehalose | + | − | − | ND | − | + | ND | + | − |
| DNA G+C content (mol%) | 68·7 | 69–70 | 70 | 66·8 | 69–71 | 66–68 | 68–70 | 64–67 | 70·7 |

**Figura 19 – Exemplo de tabela consultada no artigo referente à descrição da bactéria *Azospirillum melinis*, e que contem as informações referentes às características utilizadas para a sua classificação taxonômica. Onde + significa resultado positivo para o teste e – significa resultado negativo para o teste, ND significa não declarado (não conhecido), V significa variado, V(+) significa variado com maior tendência para ser positivo.**

FONTE: Adaptado de Guixiang et. al. 2006

### 3.2.3 Cadastro dos Tipos de Resultados

Esta tabela foi preenchida direto na base de dados, pois uma vez definidos os tipos de resultados não será permita a sua alteração. Os tipos de resultados podem ser de variadas formas conforme quadro 5:

**Quadro 11 – Tipos de resultados**

| Tipo de Resultado | Domínio | Descrição |
|---|---|---|
| Numérico | Numérico | Resultados números |
| Alfanumérico | Alfanumérico | Resultados alfanuméricos |
| Caixa de combinação | Texto | Utiliza uma caixa de combinação para exibir os possíveis resultados |
| Temperatura | Numérico | Utiliza tela exclusiva para exibição dos resultados |
| pH | Numérico | Utiliza tela exclusiva para exibição dos resultados |
| Resistência a antibióticos | Numérico | Utiliza tela exclusiva para exibição dos resultados |
| Crescimento em NaCl | Numérico | Utiliza tela exclusiva para exibição dos resultados |

FONTE: Autor, 2013

Este quadro contém todos os possíveis tipos de resultados que podem ser utilizadas.

### 3.2.4 Cadastro de Resultados das Caixas de Combinação

O cadastro dos resultados das caixas de combinação foi projetado para ser de fácil utilização, onde é possível cadastrar, alterar ou excluir um resultado. As caixas de combinação foram preenchidas direto na base, pois uma vez definidas não será permitido a sua alteração.

### 3.2.5 Cadastro das Categorias

Esta tabela foi preenchida direto na base de dados, pois uma vez que as categorias foram definidas não será permitida a sua alteração. As categorias estão listadas no quadro 6.

**Quadro 12 – Categorias cadastradas. Aquelas que foram efetivamente utilizadas estão destacadas em negrito.**

| Categorias | |
|---|---|
| Redução de Acetileno | **Hidrolise de** |
| **Produção de acido a partir de** | **Redução de nitrato** |
| Resistência a Ampicilina | Fonte de nitrogênio |
| Crescimento anaeróbico | pH ótimo de crescimento |
| Assimilação de | **Temperatura ótima de crescimento (ºC)** |
| **Fonte de Carbono** | **Atividade de Oxidase** |
| **Atividade de Catalase** | Produção de |
| **Coloração de Gram** | Atividade de Uréase |
| Crescimento na presença de NaCl | |

FONTE: Autor, 2013

Este quadro contém todas as categorias que o sistema pode utilizar, as em destaque são que foram efetivamente utilizadas.

### 3.2.6 Cadastro de Características

As características referem-se ao testes bioquímicos e fisiológicos registrados nos artigos científicos e cujos resultados são utilizados na classificação das bactérias. O cadastro das características foi projetado para ser ágil e fácil de usar. Contem um índice com uma consulta geral onde são exibidas todas as características cadastradas, que podem ser filtradas por Categoria. Para o cadastro de uma característica, basta selecionar a qual categoria a característica pertence e o tipo de resultado, e então informar o nome da característica. Se necessário, pode ser utilizado o campo de observações para o registro de informações importantes.

### 3.2.7 Cadastro dos Resultados das Características.

Este cadastro é o mais importante e o que exige maior rapidez de execução e facilidade de uso, pois a sua usabilidade deve ser a melhor possível. Para isto a tela foi projetada de modo a permitir que o cadastro dos Resultados das Características demande o menor esforço possível.

Para um novo cadastro, o usuário deve seguir os seguintes passos:

1. Na caixa de combinação, deve ser selecionado o artigo a que estes resultados pertencem. Se existirem registros previamente lançados, estes serão exibidos em uma tabela abaixo da caixa de combinação.

2. Na caixa de combinação dentro do painel *Espécie* deve ser selecionado o gênero e na caixa de combinação logo abaixo serão apresentadas todas as espécies referentes ao gênero selecionado. Após ser selecionada a espécie que se deseja utilizar, deve-se clicar no botão Adicionar. A espécie será adicionada na tabela de resultados logo abaixo.

3. Na caixa de combinação dentro do painel *Característica* deve ser selecionada a categoria de interesse e na caixa de combinação logo abaixo serão apresentadas todas as características referentes. Após ser selecionada a característica que se deseja utilizar deve-se clicar no botão Adicionar.

O diagrama abaixo ilustra o fluxo para o cadastro dos resultados que caracteriza uma espécie, estes resultados são obtidos dos artigos do período *International Journal of Systematic and Evolutionary Microbiology* (IJSEM).

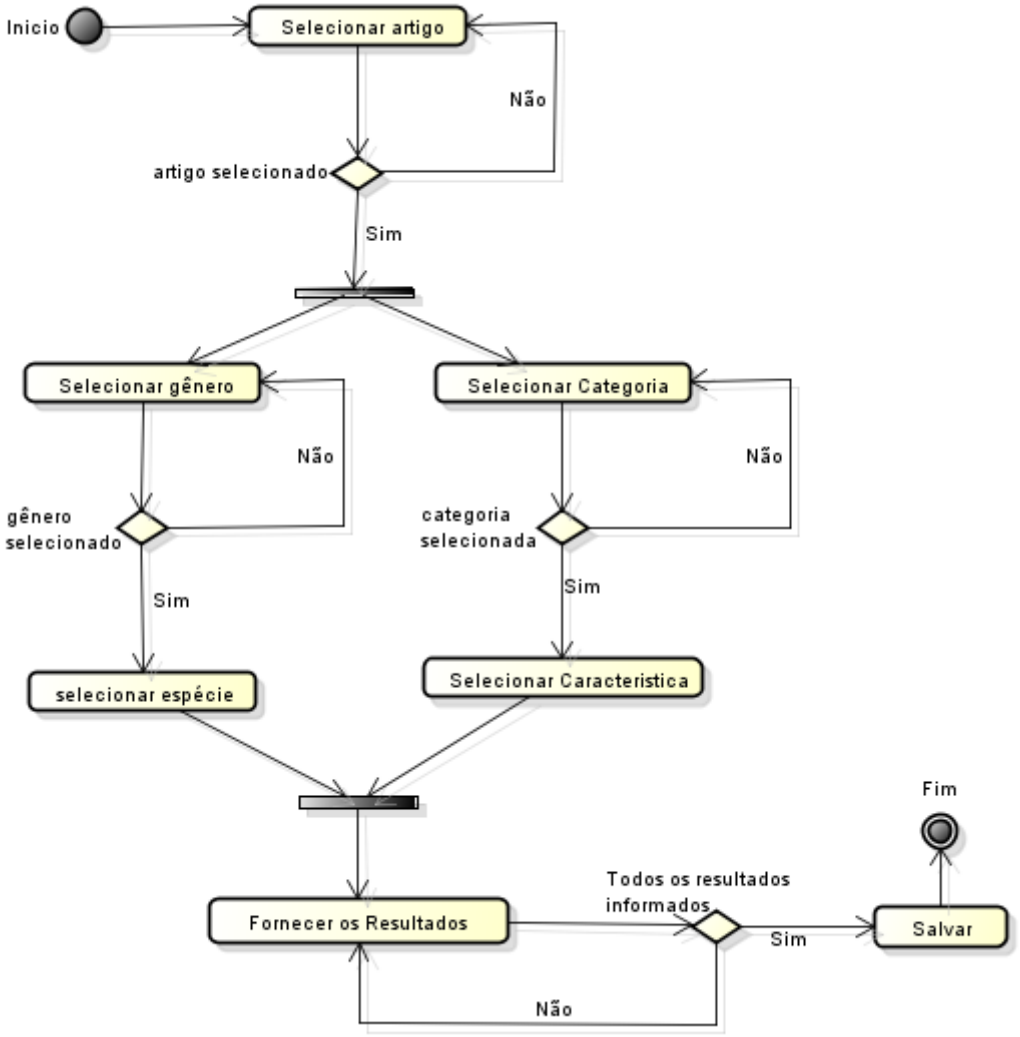



Diagrama de atividade representado o fluxo para cadastrar os resultados das características que identificam uma espécie.

FONTE: Autor, 2013

Após a inclusão das espécies e características, a tabela está pronta para receber os resultados, onde a primeira coluna representa a espécie, e as demais representam as características. Para introduzir os resultados basta clicar na coluna referente a espécie e característica desejadas. Dependendo do tipo de resultado da característica será habilitado um modo diferente de fornecer os resultados. Se for do tipo *Caixa de combinação*, será habilitada uma caixa de combinação, onde deve ser selecionado o resultado (Quadro 7).

**Quadro 13 – Possíveis resultados caixa de combinação**

| Resultado |
|---|
| Positivo (+) |
| Fracamente Positivo (>+) |
| Indefinido (-/+) |
| Fracamente Negativo (> -) |
| Negativo (-) |

Resultados possíveis: positivo, fracamente positivo (quando o resultado é mais para positivo que indefinido), indefinido, fracamente negativo (quando o resultado é mais para negativo do que indefinido.

FONTE: Autor, 2013

Se for do tipo Temperatura, será habilitada uma janela pop-up (figura 20) com opções seleção de 10 a 60 graus Celsius. Para se selecionar um intervalo de temperatura, por exemplo, de 30 a 37ºC, basta clicar em 30 e 37 e o restante será preenchido automaticamente. Estará então registrado que a temperatura de ideal de crescimento da bactéria abrange a faixa de 30 a 37 graus Celsius.

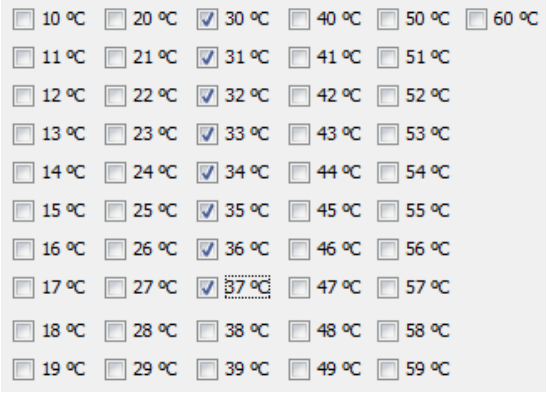



**Figura 20 – Temperatura de Crescimento**

FONTE: Autor, 2013

Se a característica for do tipo pH, será habilitada uma janela pop-up (figura 21) com opção de seleção de valores de pH de 0 ate 14, separados em intervalos de 0,5. Para se

selecionar um intervalo de pH, por exemplo do pH 6 ate o pH 8, basta clicar em 6 e 8 e o intervalo será preenchido automaticamente.

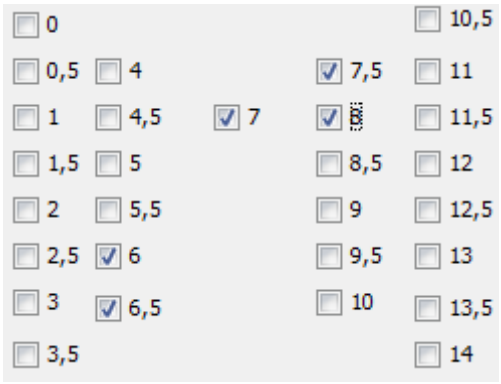



**Figura 21 – Faixa de pH**

FONTE: Autor, 2013

Se a característica for do tipo Crescimento em NaCl, será habilitada uma janela pop-up (figura 22) para seleção dos percentuais de NaCl dentro do intervalo de 0,1% a 5%. Em cada nível será indicado se a bactéria cresce ou não em meio contendo o percentual de NaCl. Se for indicado que a bactéria cresce na presença de NaCl 2%, o sistema preencherá automaticamente a opção *Sim* nos percentuais inferiores. Da mesma forma, se for indicado que a bactéria não cresce na presença de NaCl 2%, o sistema preencherá automaticamente a opção *Não* nos percentuais superiores. Quando não se dispõe do resultado, seleciona-se a opção *ND.*

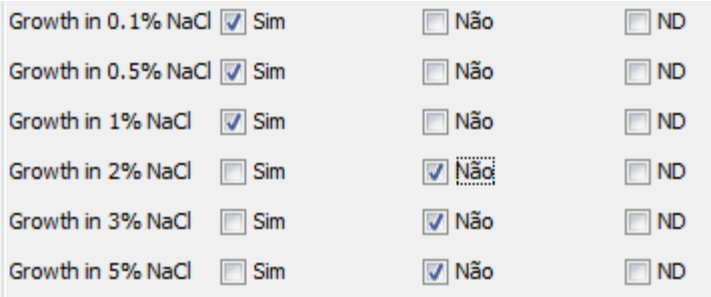



**Figura 22 – janela pop-up para a característica Crescimento em Cloreto de Sódio (NaCl)**

FONTE: Autor, 2013

Se a característica for do tipo Numérico ou Alfanumérico, o resultado deverá ser informado diretamente na célula da tabela.

Se a característica for do tipo Resistência a Antibiótico, será habilitada uma janela pop-up (figura 23) para a seleção das concentrações do antibiótico dentro do intervalo 50 a 150 µg/mL. Se for indicado que a bactéria cresce na presença de 100 µg/mL de um dado antibiótico, o sistema preenchera automaticamente a opção Sim para a concentração inferior. Da mesma forma, se for indicado que a bactéria não cresce na presença de 100 µg/mL, o sistema preenchera automaticamente a opção Não para a concentração superior. Quando não se dispõe do resultado, seleciona-se a opção ND.

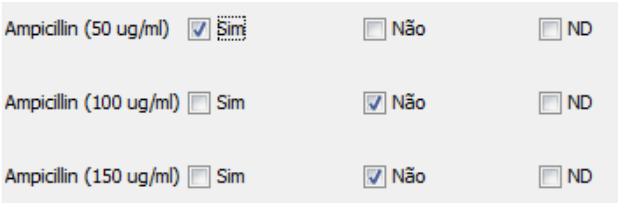



**Figura 23 – Janela pop-up para a característica Resistência a antibiótico Ampicilina**
FONTE: Autor, 2013

Após todos os resultados terem sido inseridos a tabela já pode ser salva, porém, ressalta-se que não é permitido salvar sem que todos os resultados sejam informados.

É possível visualizar o artigo em arquivo formato PDF que originou os resultados da tabela, para isto basta clicar com o botão direito e solicitar o PDF. Também é possível excluir todos os resultados de uma espécie ou todos os resultados de uma característica.

### 3.2.8 Relatórios

O usuário poderá consultar um relatório contendo todos os resultados cadastrados para uma determinada característica, para isto deverá selecionar a categoria na caixa de combinação e logo abaixo na caixa de características selecionar a opção desejada. É feita uma busca na base de dados retornando todos os resultados para característica selecionada e exibindo os artigos aos quais pertencem.

O usuário também poderá consultar o relatório contendo todos os resultados de uma determinada espécie, para isto deverá selecionar o gênero e na caixa de combinação, logo abaixo, a espécie desejada. Será feita uma busca na base de dados retornando todas as características da espécie e exibindo a quais artigos pertencem.

### 3.2.9 Cadastro dos Resultados das Características

Num primeiro momento foram cadastrados todos os resultados referentes aos testes bioquímicos e fisiológicos que correspondem aos Resultados das Características (contidos nos artigos em PDF), o que resultou em um total de mais de 14.000 registros. Entretanto durante a fase de treinamento foi necessário selecionar um conjunto mínimo de características que permitisse a viabilidade de uso da ferramenta.

A definição do conjunto mínimo de características baseou-se em dois fatores: 1. Testes que possam ser realizados por laboratórios com estrutura mínima de pesquisa e sem a demanda de equipamentos de alto custo e 2. A freqüência de utilização de um dado teste em relação aos artigos consultados. Com estas duas premissas, os testes relativos a categoria filogenética, como seqüenciamento do gene *16SrRNA*, e relativos a categoria genotípica, como porcentagem de C/G, não foram considerados.

Foi definido um conjunto preliminar com 40 características que mais possuíam resultados. Com a definição das características preliminares, surgiu à necessidade de completar os valores ausentes (atributos não determinados) de um dado padrão para o qual o resultado não estava disponível no artigo de referencia. Na literatura existem variadas abordagens, conhecido pelo termo em inglês “missing values”, em que se utilizam valores como a media, maior freqüência, ou constante global, para o preenchimento do valor desconhecido (MACHADO FILHO, 2006).

Neste trabalho foram aplicadas cinco estratégias, conforme é descrito abaixo:

**Valor Central**: Consiste em obter o valor central da característica (agrupando por gênero), se não se dispor do valor real por causa da ausência de dados o valor central será obtido pelo total dos dados. A formula utilizada é o ((máximo – mínimo)/2), onde o máximo representa o maior valor da característica e o mínimo representa o menor valor respectivamente.

**Mediana**: Consiste em obter o valor mediano da característica (agrupado por gênero), se não se dispor do valor real por causa da ausência de dados a mediana será obtida pelo total dos dados. A fórmula utilizada foi mediana.

**Media:** Consiste em obter o valor da media da característica (teste), levando em consideração os dados agrupados do gênero, se não se dispor do valor real por causa da ausência de dados a media será obtida pelo total dos dados**.** A fórmula utilizada foi a media.

**Moda (maior freqüência):** Consiste em obter a moda da característica (teste), levando em consideração os dados agrupados do gênero, se não se dispor do valor real por causa da ausência de dados a moda será obtida pelo total dos dados. A fórmula utilizada é a moda.

**Valor Fora (outlier):** Para as testes do tipo *caixa de combinação,* foi arbitrado o valor 2 (dois) que esta fora do intervalo real dos resultados, pois os valores dos resultados variam de 0 (zero) a 1 (um). Para os testes de temperatura foi arbitrado setenta, pois os valores variam de dez a sessenta.

Após o preenchimento dos valores ausentes, conforme a abordagem escolhida, os valores foram normalizados entre zero é um, utilizando a formula representada na (figura 24):

$$\left(z_i\right) = \frac{z_i - z_{min}}{z_{max} - z_{min}}$$

**Figura 24 – Formula de Normalização**

Onde:

Z: representa o número a ser normalizado

i: representa o índice

max: maior valor

min: menor valor

O valor original será subtraído do menor valor da característica, o resultado deverá ser dividido pela subtração do maior com o menor valor da característica, isto deve ser feito para todos os valores.

Após a geração dos arquivos completos os mesmos foram utilizadas para treinamento da rede MLP da biblioteca desenvolvida pelo Dr. Roberto Tadeu Raittz para Matlab (comunicação pessoal), e da rede FAN da ferramenta EasyFan (EASYFAN, 2006). Os modelos foram recriados varias vezes para todas as abordagens, inserindo ou retirando características e validados nas redes FAN do EasyFan e MLP do Matlab. Através desta abordagem foram selecionadas as 8 categorias subdivididas em 36 características que apresentaram o melhor resultado de classificação (Quadro 8).

**Quadro 14 – Categorias e características selecionadas para o treinamento**

| **Categoria** | **Característica (teste)** | |
|---|---|---|
| Produção de acido a partir de | D-Fructose | D-Mannitol |
| | D-Glucose | D-Mannose |
| | D-Glycerol | D-Xylose |
| | D-Maltose | Inulin |
| Fonte de Carbono | D-Arabinose | D-Sucrose |
| | D-Arabitol | D-Trehalose |
| | D-Fructose | D-Xylose |
| | D-Galactose | Glycerol |
| | D-Gluconate | L-Arabinose |
| | D-Glucose | L-Fucose |
| | D-Maltose | L-Rhamnose |
| | D-Mannitol | Lactose |
| | D-Mannose | N-Acetyl-D-glucosamine |
| | D-Ribose | Sodium citrate |
| | D-Sorbitol | |
| Atividade de Catalase | Catalase | |
| Coloração de Gram | Gram | |
| Hidrolise de | Caseina | |
| | Gelatina | |
| Redução de Nitrato | Redução de Nitrato | |
| Temperatura ótima de crescimento (ºC) | Temperatura ótima de crescimento (ºC) | |
| Atividade de Oxidase | Atividade de Oxidase | |

FONTE: Autor, 2013

Para as Categorias Produção de acido a partir de, Fonte de Carbono, Atividade de Catalase, Coloração de Gram, Hidrolise de, Redução de Nitrato e Atividade de Oxidase, todos do tipos de resultado caixa de combinação, foram considerados os valores default, conforme quadro 7. Para a categoria Temperatura ótima de crescimento (ºC), foi utilizada a media dos resultados quando existia mais de uma temperatura, ou a própria temperatura quando o valor era único.

Após a análise do treinamento e validação da rede MLP, com diversos parâmetros de entrada (quantidade de camadas e neurônios), em comparação aos resultados obtidos com a rede FAN, foi verificado que esta ultima sempre forneceu as melhores taxas percentuais de acerto. Com base nestas observações a rede FAN foi a escolhida para ser incorporada a ferramenta como um módulo. Este módulo foi originalmente desenvolvido por Dieval Guizelini (MsC em Bioinformática), para a ferramenta SIBILA (comunicação pessoal). Para as validações do modelo foram implementados o método Bootstrap (2.8.8) e Cross validation - leave-one-out (2.8.7.3).

## 4. Resultados e Discussão

Neste trabalho foi desenvolvida uma ferramenta que aplica técnicas de inteligência artificial para o posicionamento taxonômico de bactérias baseada em análises fisiológicas e bioquímicas. A ferramenta foi estruturada com o objetivo de se cumprir os critérios de desempenho e usabilidade, bem como todos os requisitos listados no quadro 2. A funcionalidade que demandou mais tempo e analise para ser concluída foi a de cadastro dos Resultados das Características, que correspondem aos testes bioquímicos e fisiológicos registrados nos diferentes artigos, devido a grande quantidade de dados utilizados. Por este motivo, logo na primeira versão foi notável a necessidade da usabilidade desta função, visto que na versão final, este processo ficou em media quatro vezes mais rápido que as primeiras versões.

### 4.1 Funcionalidades disponíveis na ferramenta

A primeira funcionalidade introduzida foi o Cadastro de Artigos que está apresentado na figura 25. Esta função foi projetada para permitir o rápida cadastro e fácil visualização dos artigos, pois é possível ver o nome do artigo e ano de publicação (figura 26). Para visualizar o artigo, basta o usuário clicar com mouse sobre a linha correspondente e solicitar a exibição do artigo no formado PDF. Estão registrados 73 artigos referentes à descrição de 228 espécies de bactérias pertencentes 10 gêneros diferentes (Quadro 4). Se o usuário desejar cadastrar um novo artigo deverá utilizar a janela de Cadastro de Novos Artigos (figura 25) onde preencherá os campos *nome do artigo, ano de publicação, nome dos autores*. Se achar necessário, poderá utilizar o campo *descrição* para registrar anotações que julgue importantes. É obrigatório ter o PDF do artigo no formato PDF, este arquivo será armazenado no sistema.

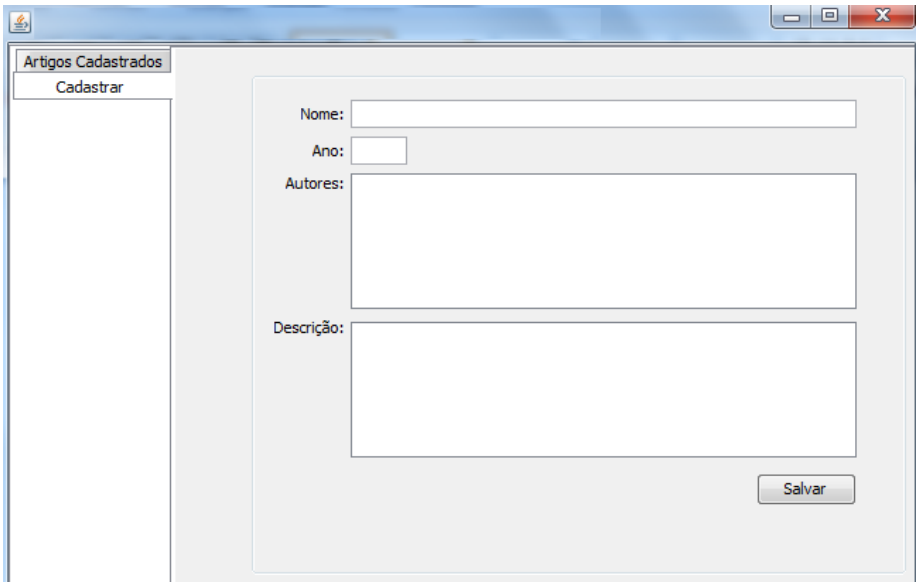



**Figura 25- Captura de janela Cadastro de Novos Artigos**

FONTE: Autor, 2013

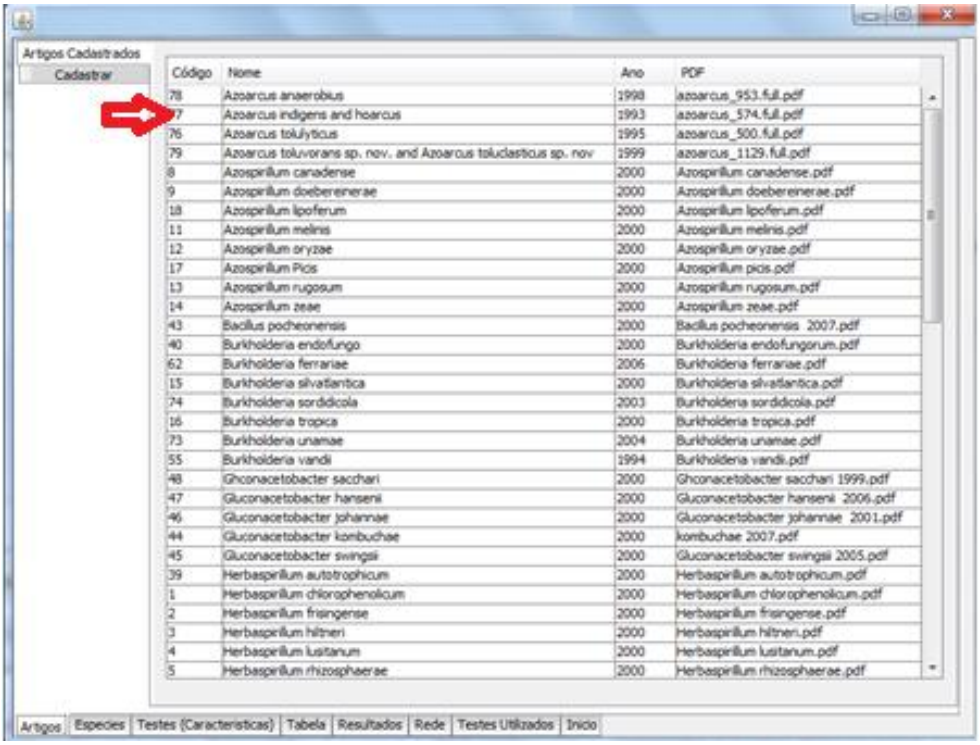



**Figura 26 – Captura da janela Consulta de Artigos**

FONTE: Autor, 2013

A próxima funcionalidade adicionada foi Cadastro de Espécies, que esta apresentada na figura 27. Esta é uma funcionalidade de consulta que permite a visualização de todos os

registros já cadastrados. São exibidos os campos *nome da espécie, estirpe* (se houver) e o *gênero* a qual pertence. Inicialmente são apresentadas todas as espécies, mas é possível filtrar uma espécie especifica através da opção *caixa de combinação*, no inicio da janela. Também é possível a visualização do artigo em formato PDF. Se o usuário desejar incluir uma nova espécie pertencente a um gênero já cadastrado devera utilizar a janela Cadastro de Nova Espécie (figura 28), onde preenchera os campos *nome da espécie* e *estirpe* (se houver) e, na caixa de combinação, selecionará o *gênero*. Se achar necessário, poderá utilizar o campo *descrição* para registrar anotações que julgue importantes. Caso o gênero não esteja cadastrado, o usuário utilizará a opção Novo Gênero, que abre uma janela pop-up, para incluir o novo registro.

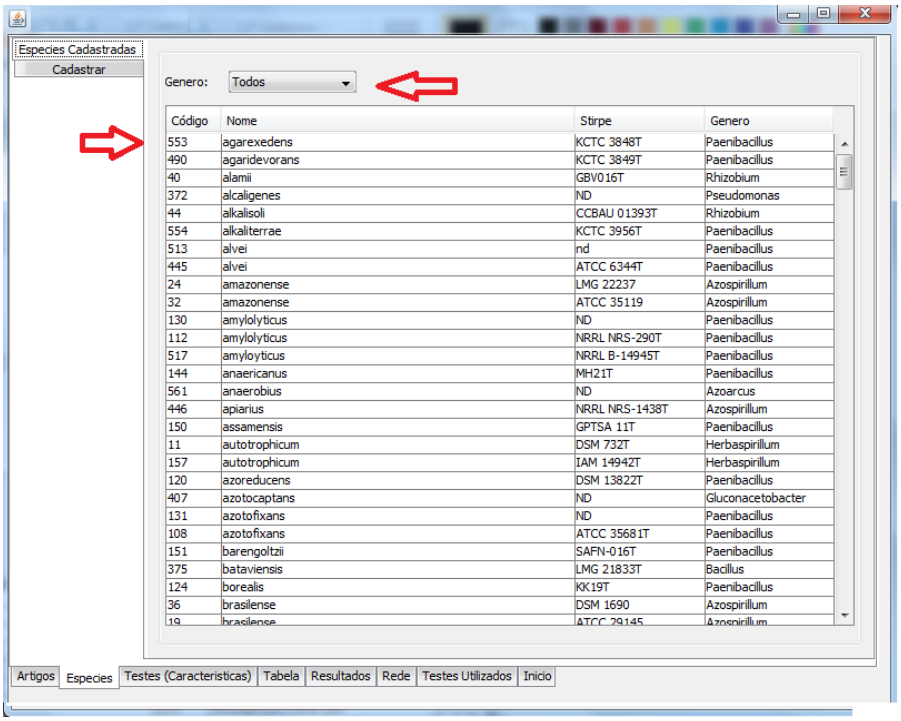



**Figura 27 – Captura da janela Consulta das Espécies Cadastradas**

FONTE: Autor, 2013

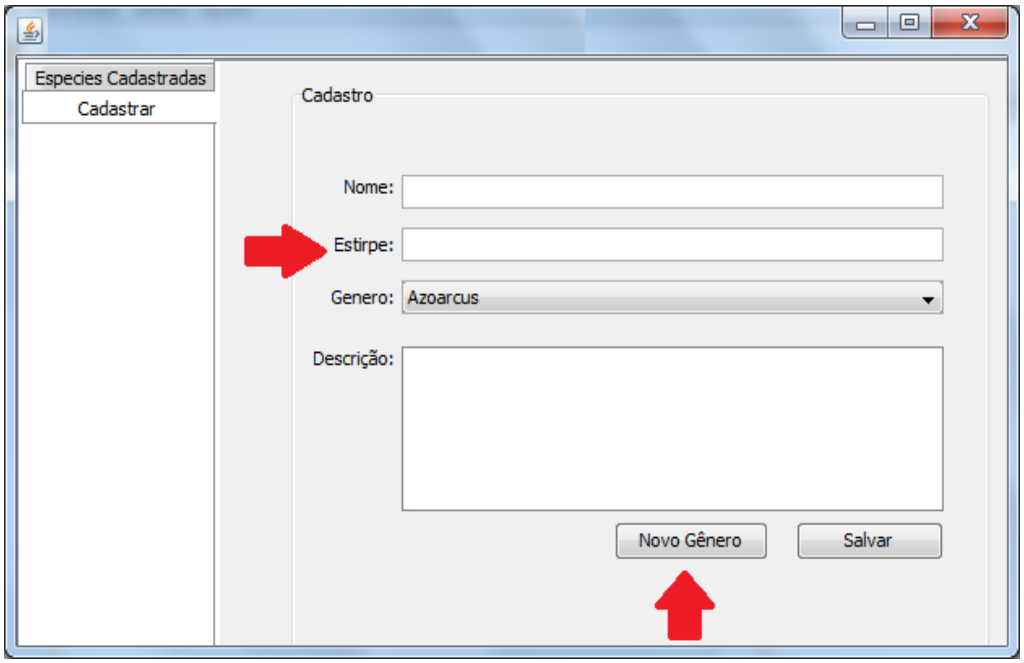



**Figura 28 – Captura da janela Cadastro de Nova Espécie**

FONTE: Autor, 2013

A funcionalidade seguinte foi Cadastro dos Resultados das Caixas de Combinação que esta apresentada na figura 29. Os resultados são correspondentes ao tipo de resultado Caixa de Combinação (quadro 5). O usuário deve selecionar Tipo de Resultado na caixa de combinação e na tabela de visualização são exibidos os registros referentes. Para alterar o registro, basta clicar na linha desejada para que os campos da tabela sejam carregados nos campos do painel *Alteração* e com isto é possível alterar ou excluir o registro. Caso o usuário deseje cadastrar um novo registro basta preencher todos os campos do painel *Alterar* e clicar na opção Salvar.

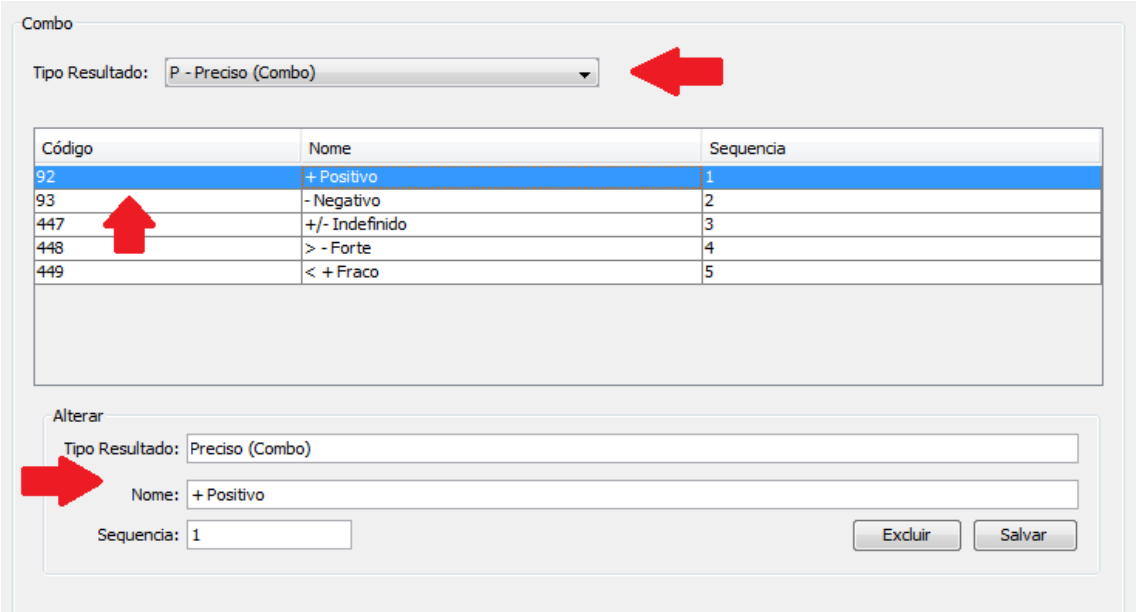



**Figura 29 – Captura da janela de Cadastro dos Resultados das Caixas de Combinação**

É possível escolher a seqüência de exibição dos dados na caixa de combinação, para isto, escolher a posição de exibição no campo seqüência.

FONTE: Autor, 2013

A funcionalidade seguinte foi Cadastro de Características que está apresentada na figura 30. Esta é uma funcionalidade de consulta que permite a visualização de todos os registros já cadastrados. São exibidos os campos *nome da característica, categoria* e o *tipo de resultado* a qual pertence. Inicialmente são apresentadas todas as categorias, mas é possível filtrar uma categoria especifica através da opção *caixa de combinação*, no inicio da janela. Se o usuário desejar incluir uma nova característica deverá utilizar a janela Cadastro de Nova Característica (figura 31), onde selecionará a *categoria* e o *tipo de resultado.* Também deverá preencher o campo *nome*. Se achar necessário, poderá utilizar o campo *descrição* para registrar anotações que julgue importantes.

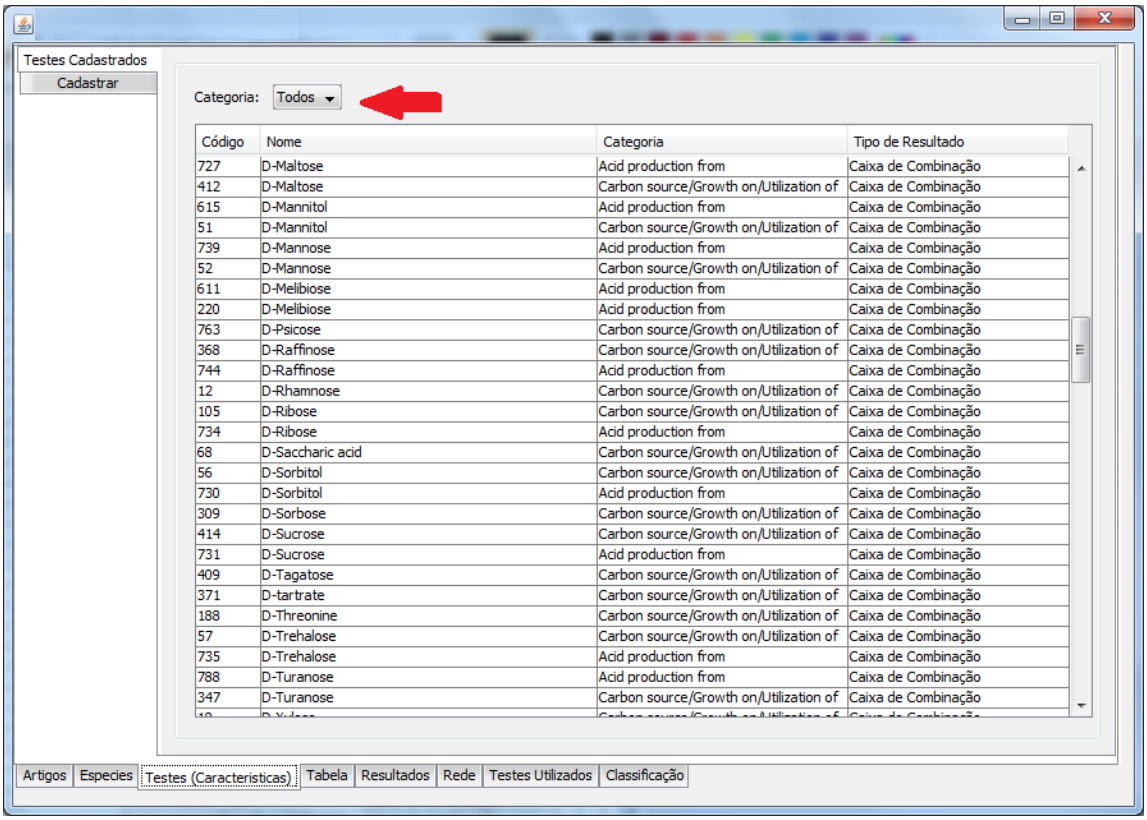



**Figura 30 – Captura da janela Consulta de Testes Cadastrados**

FONTE: Autor, 2013

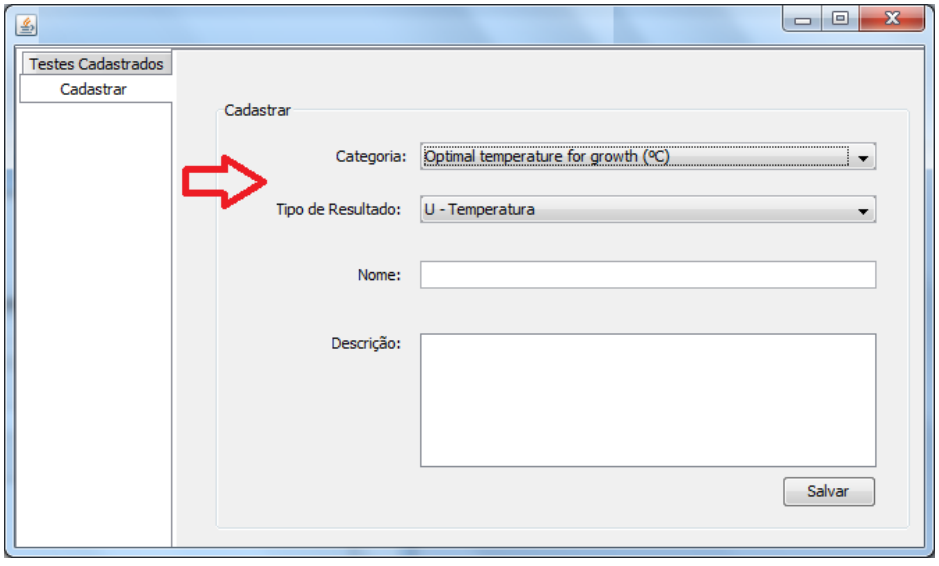



**Figura 31 – captura da janela Cadastro de Nova Característica**

FONTE: Autor, 2013

A funcionalidade seguinte foi Resultados das Características que está apresentada na figura 32. Ao ser selecionado o artigo de interesse, os registros referentes são exibidos na tabela de visualização. Se o usuário desejar incluir uma nova espécie deverá selecionar, no painel *Espécie*, o Gênero, que filtrara a caixa de combinação Espécie com base na opção escolhida. A adição da nova espécie ocorrerá ao clicar na opção Adicionar, e a espécie selecionada aparecerá na tabela de visualização. Se o usuário desejar incluir uma nova característica deverá selecionar a categoria no painel *Característica*, que filtrará a caixa de combinação características com base na opção escolhida. A adição da nova característica ocorrerá ao clicar na opção Adicionar, e a mesma aparecerá na tabela de visualização.

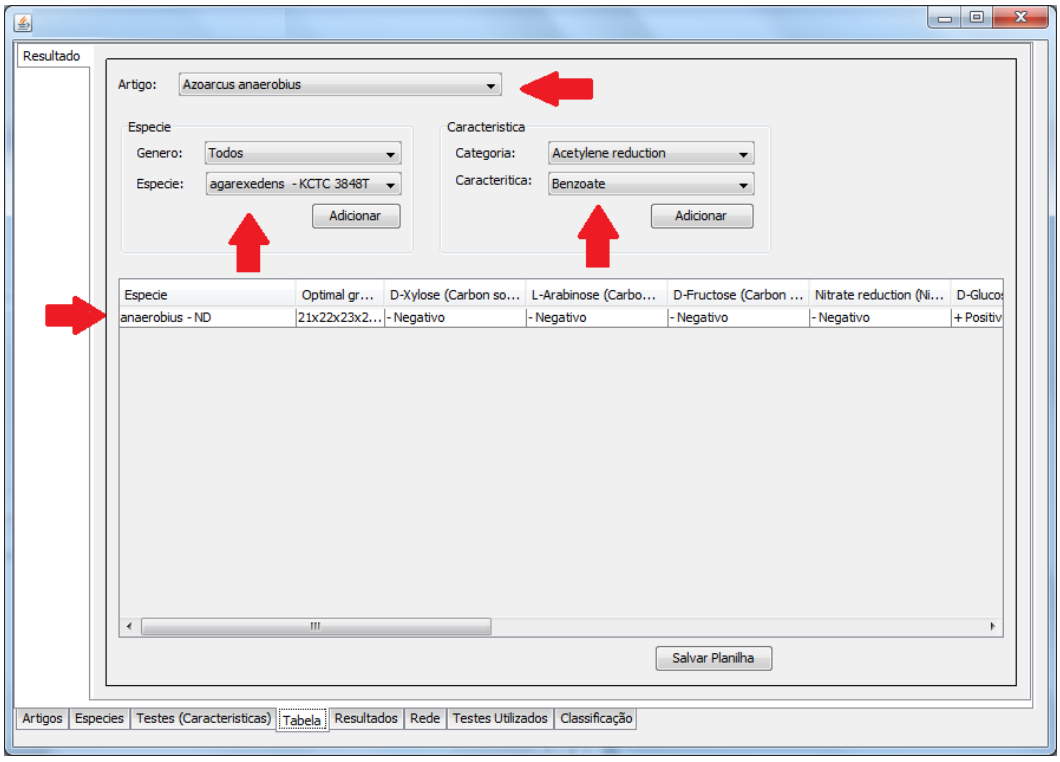



**Figura 32 – Captura da janela Resultados das Características Cadastradas**

FONTE: Autor, 2013

A forma de inclusão do dado referente ao Resultado da Característica depende do tipo de resultado que ativará dinamicamente a janela correta de lançamento do mesmo. Se o tipo de resultado for Temperatura, uma janela própria para o lançamento do resultado será carregada (figura 33) e o usuário deverá selecionar a temperatura adequada.

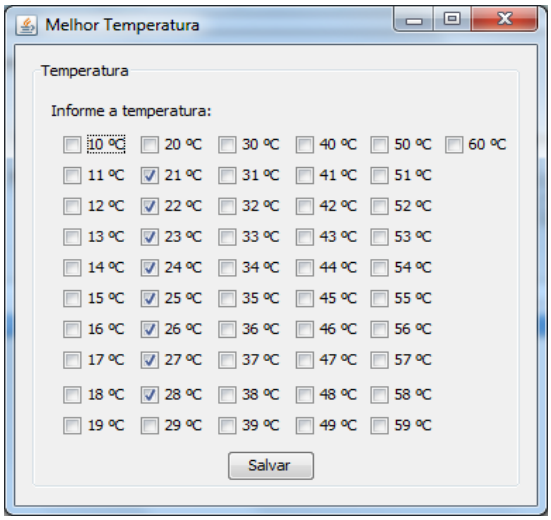



**Figura 33 – Captura da janela pop-up para a categoria Temperatura**

FONTE: Autor, 2013

Se o tipo de resultado for Caixa de Combinação, será carregada na célula da tabela referente ao resultado uma caixa de combinação (figura 34) e o usuário deverá selecionar a opção mais adequada. Caso seja selecionado um teste cujo não existe tela de lançamento (pop-up) é carregado uma caixa de combinação, possibilitando a simples seleção do resultado.

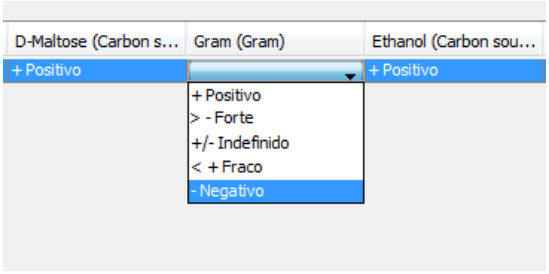



**Figura 34 – captura da janela Caixa de combinação**

FONTE: Autor, 2013

A funcionalidade Resultados das Características apresenta ainda as opções: excluir uma espécie (*Excluir linha*), excluir uma característica (*Excluir coluna*) e visualizar o artigo que originou os resultados (*Exibir PDF*) (figura 35).

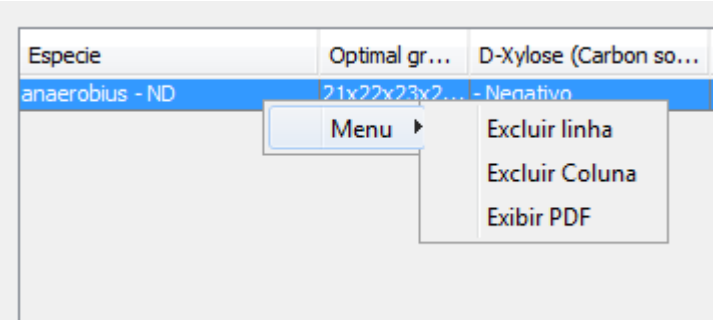



**Figura 35 – Captura da janela Opções na funcionalidade Resultados das Características Cadastradas**

FONTE: Autor, 2013

A funcionalidade seguinte foi *Extração de Características* (figura 36A) e permite extrair as características a serem utilizadas no treinamento da rede. Isto pode ser realizado utilizando a função Gerar Features. A extração das características seguiu o modelo já explicado (4.2.9)**,** em que cinco abordagens foram utilizadas (Valor central, Mediana, Média, Moda e Valor fora). Após geração das características (features) é possível salva-las através da opção “Salvar Features” (figura 36B). O arquivo gerado é salvo no diretório “rede” do sistema com a extensão “.dat”. Foram gerados cinco arquivos dat, um para cada metodologia, e estes arquivos foram utilizados para as validações na plataforma Weka (WITTEN & FRANK, 2005).

O treinamento (figura 36C) utiliza a rede neural FAN (2.8.1) e pode ser iniciado através da opção Treinar Rede que carrega a janela (figura 37) e utiliza o conjunto de features salvo na etapa anterior. A funcionalidade permite a configuração de parâmetros de treinamento da rede, como a escolha da quantidade de épocas e de quantas em quantas épocas o conjunto de treinamento será embaralhado (periodicidade). Durante o treinamento será exibida a taxa percentual de acerto, da melhor rede, no campo *Melhor media*. Além disso, é exibida a media atual de acerto no campo *Media atual*, bem como a época de treinamento no campo *Época*. Após o termino do treinamento a rede pode ser salva através da opção “Salvar Rede” (figura 37).

Se o usuário desejar obter o posicionamento taxonômico de uma dada bactéria deverá utilizar a janela Classificação (figura 38). Com a rede treinada e salva o sistema esta apto a classificar um conjunto de valores obtidos nos experimentos bioquímicos e fisiológicos (padrão desconhecido). O usuário deverá preencher os valores das características e clicar na opção Classificação e o resultado do posicionamento taxonômico no nível de gênero será exibido no campo *Gênero* (figura 38).

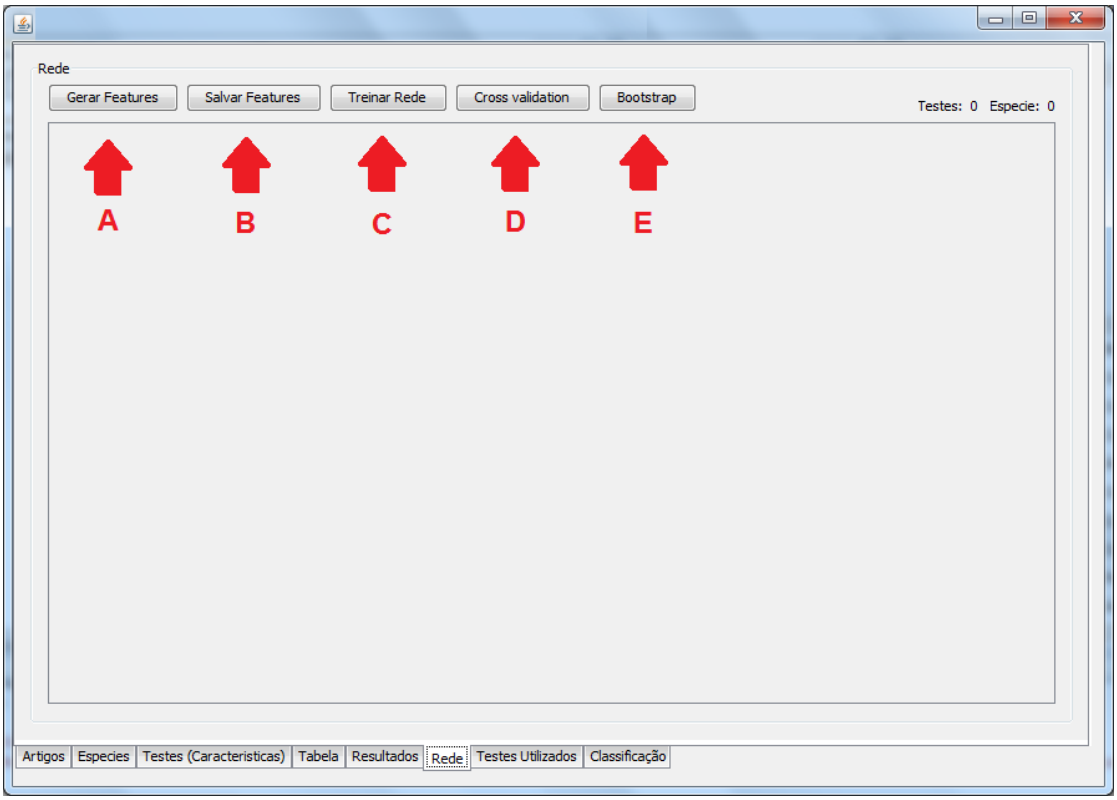



**Figura 36 – Captura da janela Treinamento do Modelo**

FONTE: Autor, 2013

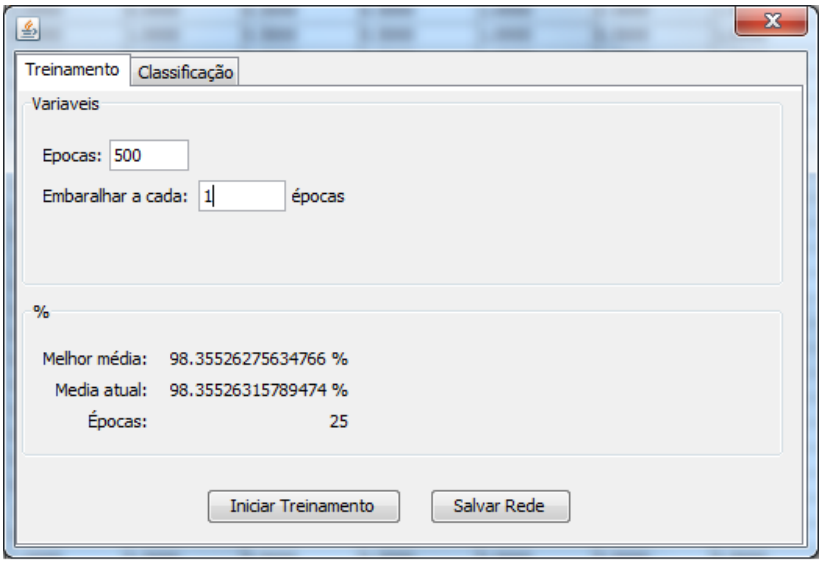



**Figura 37 – Captura da janela Treinamento**

FONTE: Autor, 2013

A quantidade de épocas pode ser definida (quantidade de vezes que a rede neural artificial repete o processo de aprendizagem) bem como a freqüência que o conjunto de treinamento é embaralhado.

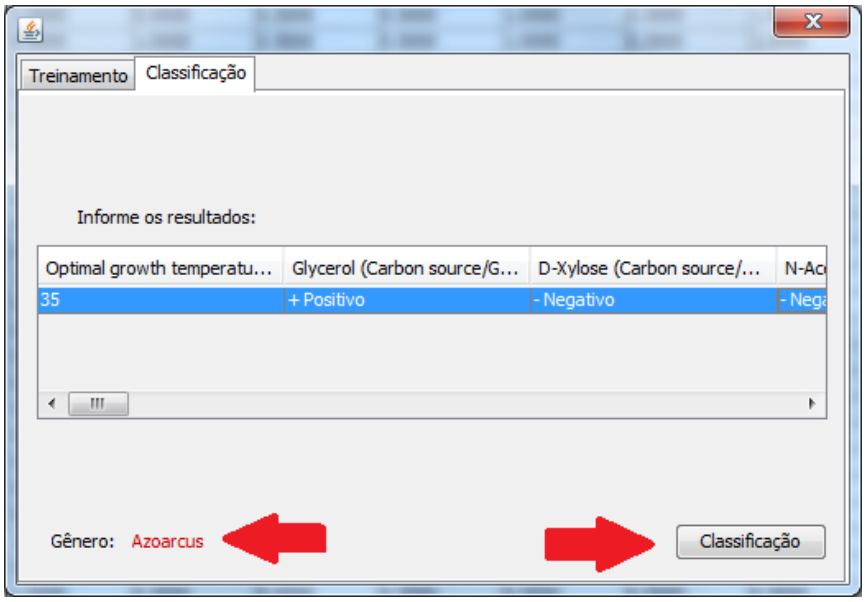



**Figura 38 – Captura da janela de Classificação**

FONTE: Autor, 2013

As opções Cross-validation (figura 36D) e Bootstrap (figura 36E) direcionam para a validação dos modelos. Os resultados são exibidos no campo *porcentagem de acerto*. Os dois métodos permitem a configuração dos parâmetros de treinamento da rede, como a escolha da quantidade de épocas de treinamento e a periodicidade em que o conjunto de treinamento será embaralhado (figuras 39 e 40). O método Bootstrap também permite a configuração da quantidade de copias bootstrap a ser utilizada (figura 40).

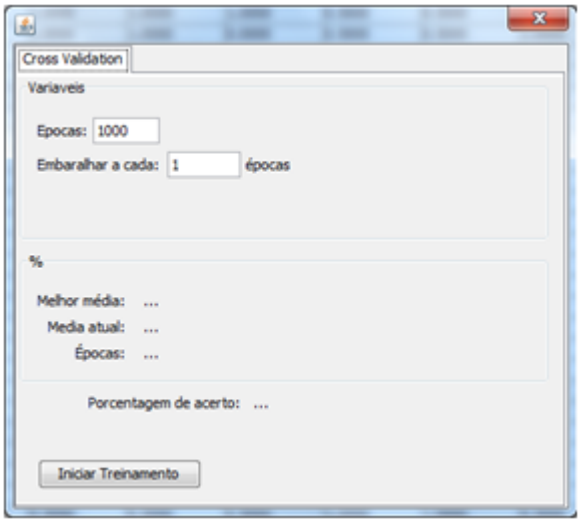



**Figura 39 – Captura da janela Cross Validation (leave-one-out)**

FONTE: Autor, 2013

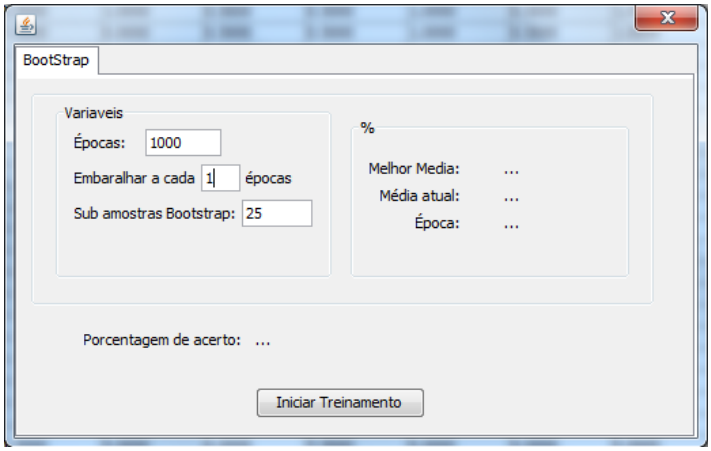
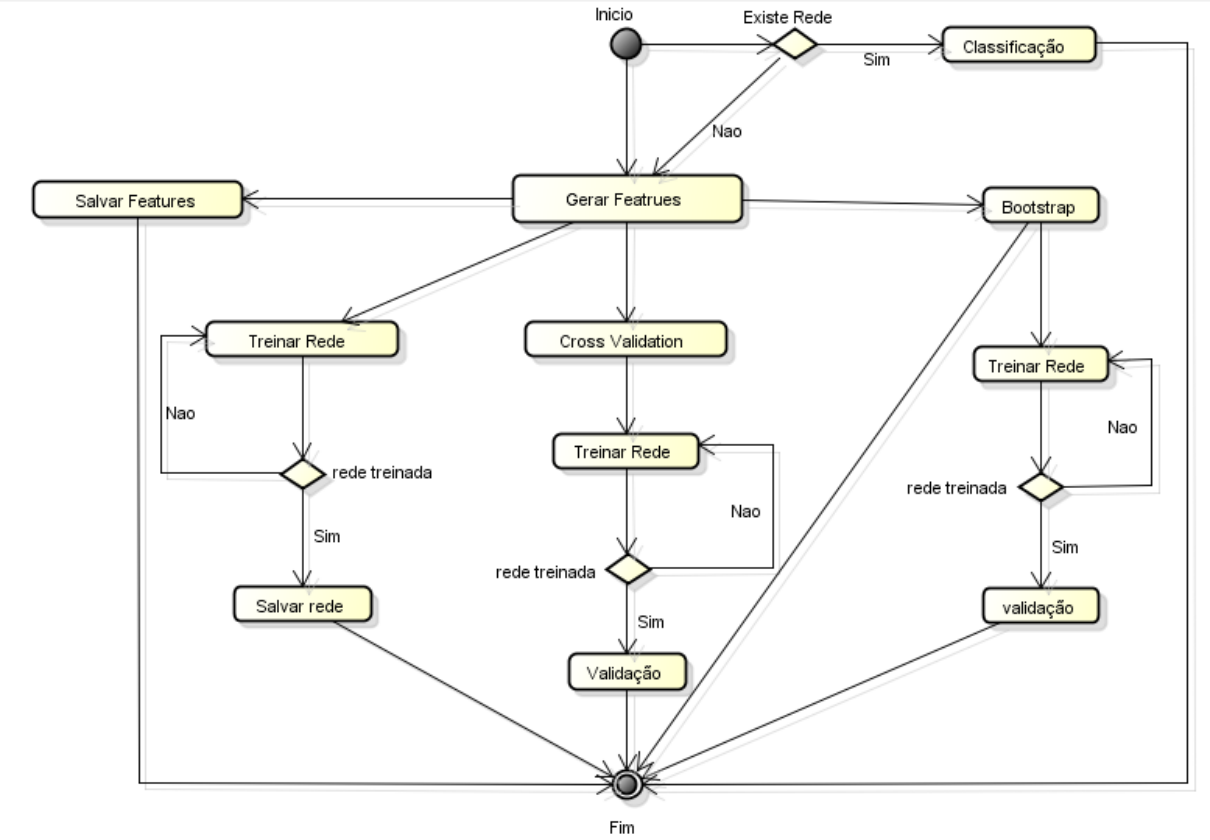



**Figura 40 – Captura da janela Bootstrap**

FONTE: Autor, 2013




Fonte: O autor, 2013

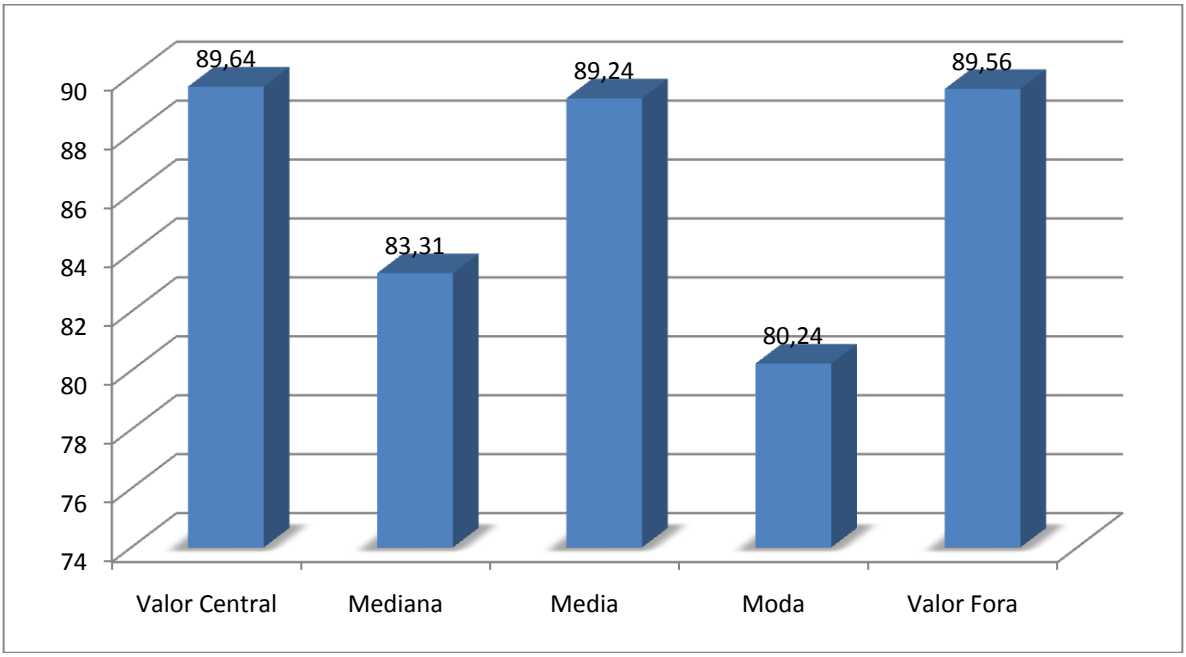
O diagrama de atividades representa os passos necessários para gerar e salvar uma rede neural artificial, bem como utilizar os métodos de validação Bootstrap e Croos Validation. Também representa o fluxo para utilizar a classificação quando já existe uma rede treinada salva.

## 4.2 Validações do modelo e seleção da estratégia de preenchimento de atributos não determinados

Para melhorar a interpretação dos resultados obtidos durante a validação do processo de treinamento e em decorrência do reduzido conjunto de padrões disponível para cada gênero bacteriano, comparou-se os resultados obtidos com o aprendizado supervisionado do conjunto total (onde todos os padrões são utilizados para o treinamento e validação) com os resultados obtidos da subdivisão do conjunto total para teste e validação. Além disso, pode ser escolhida a melhor estratégia de preenchimento de atributos não determinados (valores ausentes). Foram utilizados os métodos Cross validation – leave one out e Bootstrap, ambos já implementados na ferramenta (figura 39 e 40).

No método Bootstrap foram realizados testes com os parâmetros: 500 épocas, conjunto de treinamento embaralhado a cada época e 25 cópias bootstrap. Isto foi replicado para as cinco estratégias: Valor Central, Mediana, Media Moda e Valor Fora (4.2.9). Os resultados são exibidos na tabela 1 e o melhor resultado foi obtido com a estratégia *valor central*, com 89,64% de acerto.

**Gráfico 1 – Seleção da estratégia de preenchimento de atributos não determinados pelo método Bootstrap (25 cópias).**

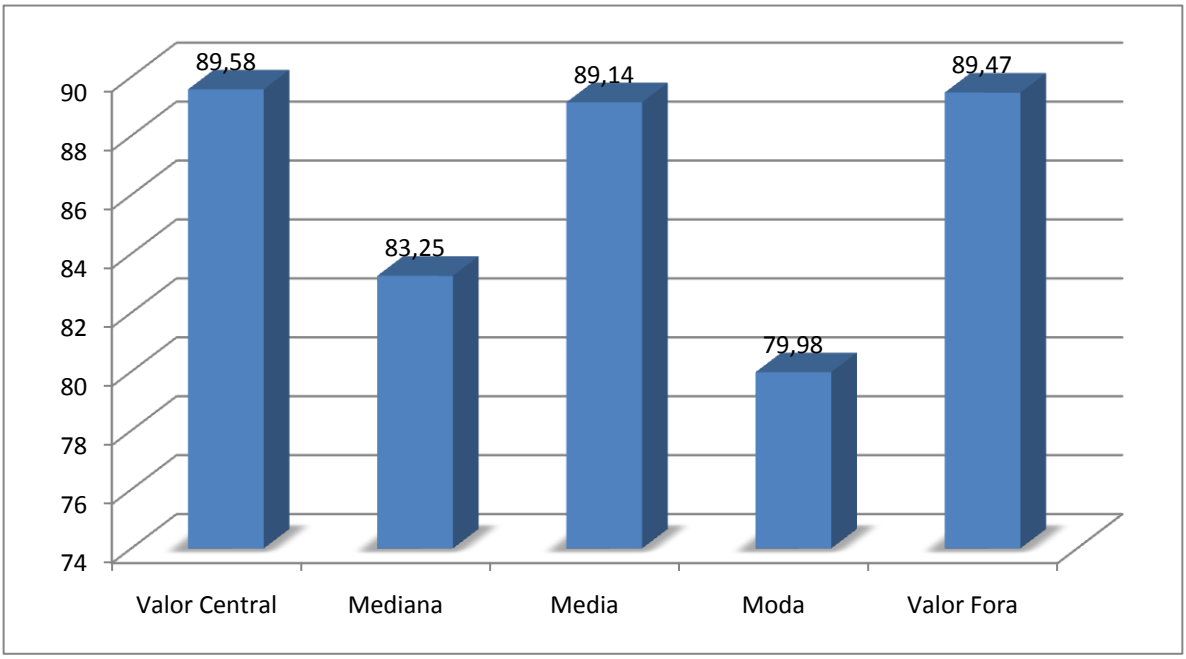



FONTE: Autor, 2013

Os mesmo testes foram realizados com o parâmetro número de copias aumentado para 50. Os resultados são mostrados na tabela 2 e o melhor resultado também foi obtido com a estratégia *valor central*, com 89,58% de acerto. Comparando-se os valores das duas

tabelas conclui-se que mesmo com o dobro de replicas a taxa de acerto sofreu pouca alteração.

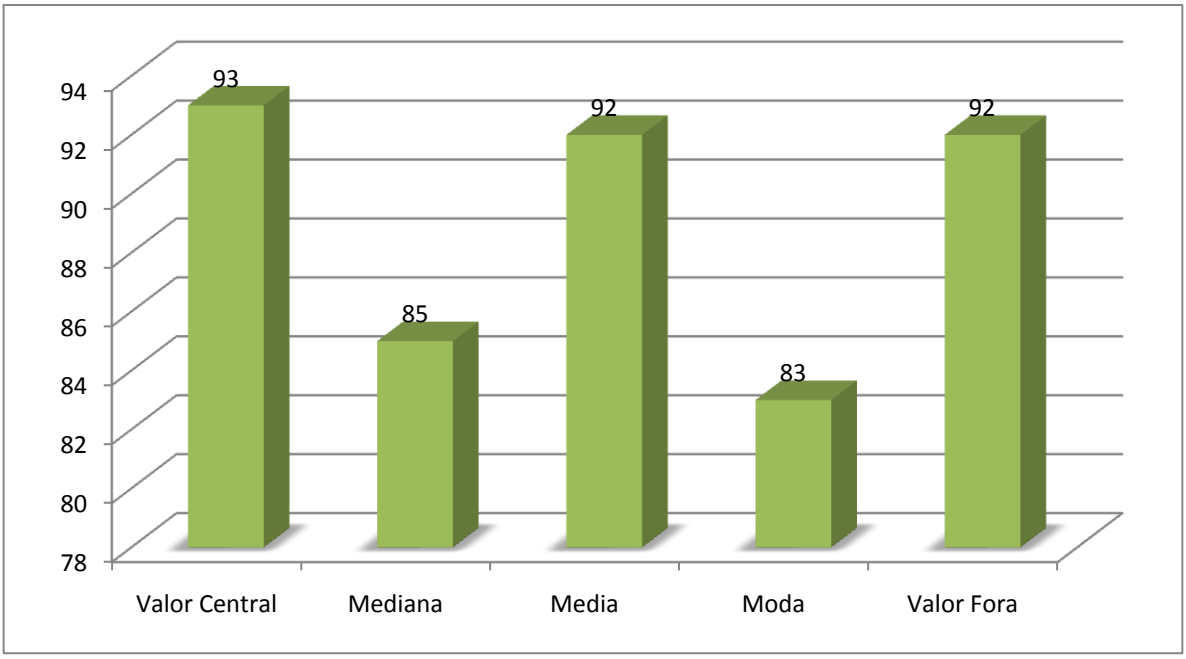
**Gráfico 2 – Seleção da estratégia de preenchimento de atributos não determinados pelo método Bootstrap (50 cópias).**




FONTE: Autor, 2013

Utilizando método Cross Validation em seu modelo extremo, o leave one out, foram feitos testes utilizando os parâmetros: 500 épocas e o conjunto de treinamento embaralhado a cada época. Isto foi replicado para as cinco estratégias: Valor Central, Mediana, Media Moda e Valor Fora (4.2.9). Os resultados são exibidos na tabela 3, e o melhor resultado foi obtido com a estratégia *valor central*, com 93% de acerto.

**Gráfico 3 – Seleção da estratégia de preenchimento de atributos não determinados pelo método Cross Validation – leave one out**




FONTE: Autor, 2013

Os percentuais de acerto do método Bootstrap foram inferiores ao método Cross Validation porque este método não possui controle sobre a especialização da rede treinada (2.8.8), pois pode acontecer que uma classe inteira seja colocada no conjunto de testes não tendo, com isto, nenhuma representante no conjunto de treinamento (2.8.8). Este fato pode gerar erro na classificação e, conseqüentemente diminuir a taxa porcentual de acerto.

Baseado nesta avaliação pode-se concluir que o modelo utilizando a estratégia valor central é a melhor opção entre as testadas para o preenchimento de atributos não determinados (valores ausentes) e que a taxa percentual de desempenho da ferramenta (89 a 93%) pode ser considerada aceitável.

## 4.3 Comparações do desempenho de diferentes algoritmos em relação aos modelos

Durante o desenvolvimento da ferramenta foram testadas as redes neurais artificiais MLP e FAN. A primeira por ser o algoritmo referencia e mais utilizado na maioria dos problemas de classificação da área de reconhecimento de padrões. E a segunda por ter sido desenvolvida por membro do grupo de pesquisa e por apresentar características consideradas relevantes ao domínio do problema, entre outras, o fato da independência entre os atributos.

Verificou-se que durante os testes, a rede FAN obteve o melhor desempenho, apresentando sempre as melhores taxa percentuais de acerto (superior a 90%). Por este motivo ela foi incorporada a ferramenta de forma definitiva.

Para se confirmar que a rede FAN foi à melhor escolha, foram comparados os resultados obtidos com outros algoritmos. Para isto foi utilizada a opção Salvar Features (janela Treinamento do Modelo - Figura 36B), que gera um arquivo dat com o conjunto de treinamento. Este conjunto foi transformado no formato arff para teste na plataforma Weka (2.7.1). Os algoritmos selecionados foram FAN, MLP, J48 (ID3), SVM e RBF e os testes de validação foram feitos com a configuração default de cada algoritmo. O módulo FAN denominado de *FANClassifier* (baseado em FAN e desenvolvido e integrado no WEKA por Dieval Guizelini - MsC em Bioinformática) é o mesmo utilizado pelo sistema.

Na plataforma Weka é possível realizar o treinamento da rede neural bem como a validação do modelo de diferentes modos. Os testes foram feitos de duas abordagens:

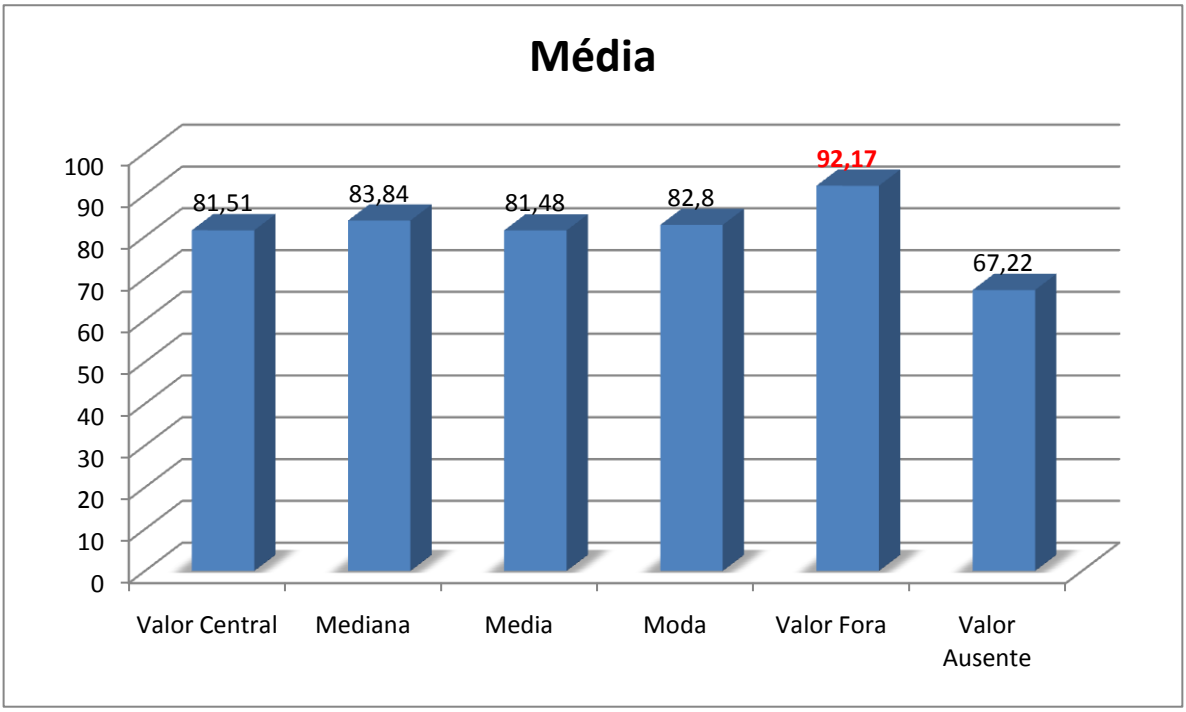
A: Validando o modelo com o próprio conjunto de treinamento;

B: Utilizado a função de Cross Validation do próprio Weka. Neste caso foi escolhida a opção 3 folds, uma vez que a classe com menor quantidade de padrões possui três padrões. Assim, pode-se garantir que cada subconjunto gerado tenha no mínimo um representante.

A árvore de decisão J48 é uma implementação na linguagem Java do algoritmo C4.5 para a plataforma Weka e consiste em uma melhoria do algoritmo ID3 (2.8.5). Uma das melhorias apresentada é o tratamento dos atributos não determinados (valores ausentes) e, para isto, no conjunto de treinamento deve ser inserido o símbolo “?” como resultado das características ausentes. Com base nesta capacidade, foi inserida mais uma estratégia de avaliação, o *Valor ausente*. O resultado da aplicação desta estratégia foi obtido somente para a árvore de decisão J48 (tabela 7). Foram obtidas as taxas de acerto de 71,28% e 63,16% para as abordagens A e B respectivamente. Estes valores indicam que desconsiderar os valores ausentes como é feito no algoritmo J48 não produz bons resultados, pois os percentuais de acerto são inferiores quando comparados às outras estratégias (tabela 4).

Foram comparados os resultados obtidos com a estratégia de treinamento utilizando Cross Validation 3-folds da plataforma WEKA. Na tabela 4 estão apresentadas as medias dos desempenhos obtidos, e pode-se verificar que a estratégia “valor fora” foi à melhor opção para o presente problema. Obtendo uma media de acerto de 92,17%.

**Gráfico 4 – Media das metodologias**

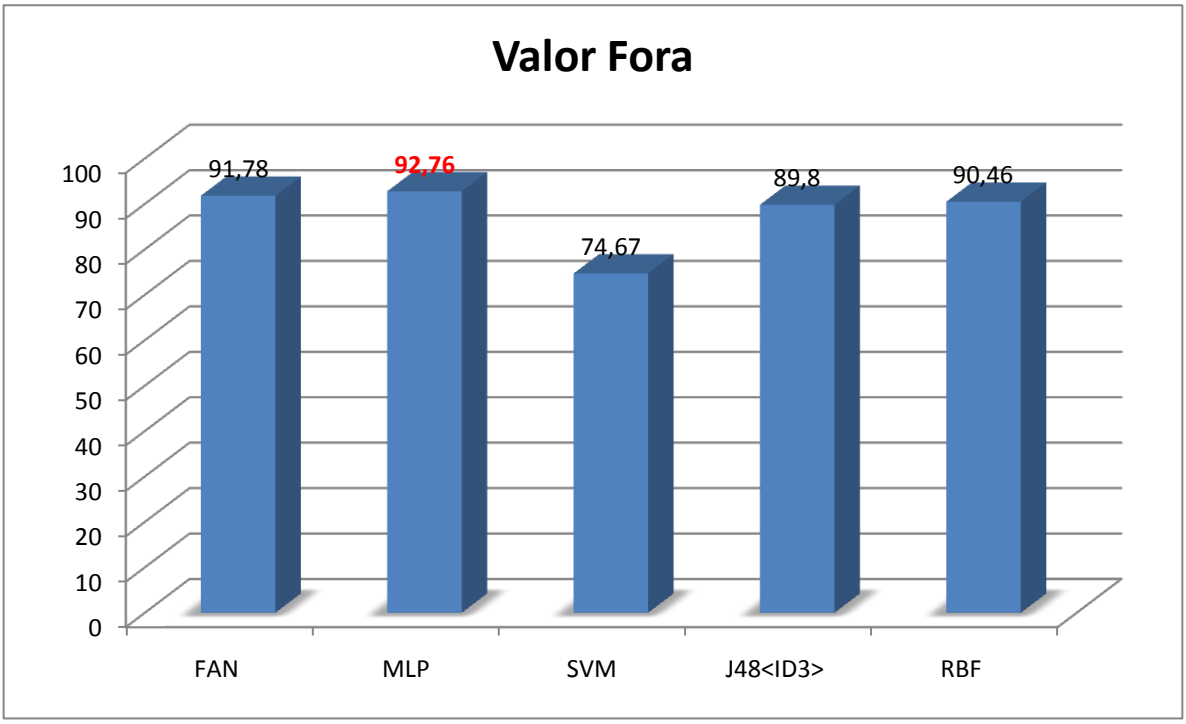



FONTE: Autor, 2013

Medias obtida para todas as estratégias de preenchimento dos valores ausentes

Comparando a estratégia de tratamento dos valores, entre os diferentes algoritmos, obtemos o MLP com a melhor taxa de acerto, com 92,76% (tabela 5).

**Gráfico 5– Acertos da estratégia Valor fora**

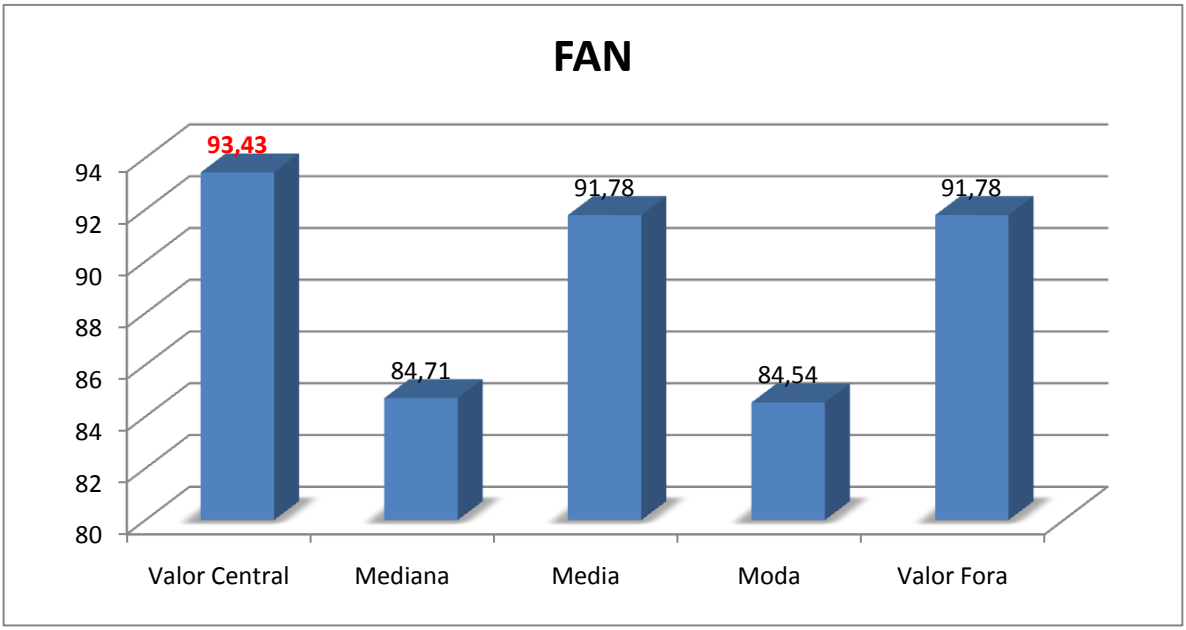



FONTE: Autor, 2013

Valor obtido para a estratégia de preenchimento Valor Fora para todos os algoritmos.

Porém, analisando a melhor taxa de acerto entre todas as estratégias (tabela 6) o algoritmo FAN com a estratégia valor central forneceu o melhor resultado, sendo superior ao melhor resultado obtido no algoritmo MLP (tabela 4).

**Gráfico 6 – Resultados de todas as estratégias de preenchimento de valores ausentes para a rede FAN validada pelo método Cross Validation 3-folds.**




FONTE: Autor, 2013

Valor obtido pelo o algoritmo FAN para as cinco estratégias de preenchimento dos valores ausentes.

A estratégia valor central preenche os valores ausentes, com um valor padrão que não influencia o resultado, este valor é obtido de forma a permanecer exatamente entre o máximo e o mínimo. Por exemplo, ao considerar os valores *SIM* ou *NÃO* como elementos possíveis do conjunto de resultados, para a ausência de um resultado é incluído um valor intermediário, aqui traduzido como TALVEZ, ou seja, não tendendo para nenhum lado.

Considerando-se apenas as cinco estratégias iniciais, a validação baseada em Cross Validation 3-folds forneceu valores que variaram de 74,67 ate 93,43% (tabela 7), sendo que o maior valor corresponde à estratégia valor central e rede neural FAN. Este resultado confirma que a escolha da estratégia de preenchimento de valores ausentes valor central e da rede FAN foi à correta (5.2).

**Tabela 1 – Resultados obtidos da comparação entre os algoritmos FAN, MLP, SVM, RBF e J48 na plataforma WEKA.**

| | Valor Central | Mediana | Media | Moda | Valor Fora | Valor Ausente |
|---|---|---|---|---|---|---|
| FAN (treino) | 100,00 | 98,14 | 100,0 | 97,37 | 100,00 | ... |
| FAN (cross validation 3-folds) | **93,43** | 84,71 | 91,78 | 84,54 | 91,78 | ... |
| MLP (treino) | 95,72 | 99,07 | 93,75 | 98,03 | 98,68 | ... |
| MLP(cross validation 3-folds) | 73,03 | 81,01 | 73,03 | 80,59 | **92,76** | ... |
| SVM(treino) | 55,92 | 74,90 | 52,30 | 72,37 | 87,83 | ... |
| SVM(cross validation 3-folds) | 44,74 | 61,46 | 41,45 | 60,53 | **74,67** | ... |
| J48<ID3> (treino) | 92,43 | 92,21 | 94,41 | 90,79 | 97,04 | 71,28 |
| J48<ID3> (cross validation 3-folds) | 77,30 | 80,76 | 83,22 | 76,97 | **89,80** | 63,16 |
| RBF (treino) | 94,41 | 88,09 | 96,38 | 88,49 | 98,68 | ... |
| RBF (cross validation 3-folds) | 88,16 | 78,03 | 88,49 | 78,29 | **90,46** | ... |

Tabela com todos os resultados obtidos pela ferramenta WEKA, para as cinco estratégias de preenchimento para os dados ausentes. Para cada estratégia é exibido as porcentagens de acerto usando os mesmos dados para treino e testes e Cross Validation (3-folds).

FONTE: Autor, 2013

### 4.4 Correções de erros da classificação prévia com base no resultado do classificador

Durante o processo de treinamento, a análise da matriz de confusão produzida pelo WEKA permitiu identificar inconsistência entre a indicação do modelo e a classificação inicialmente atribuída a um conjunto de padrões. Foi observado que algumas espécies sempre eram incorretamente posicionadas e por isto foi necessário fazer a revisão do cadastro. Foi verificado que os resultados das características estavam corretamente incluídos, mas algumas espécies de bactérias estavam vinculadas a gêneros incorretos. Após a correção, as mesmas foram corretamente posicionadas.

### 4.5 Interpretações da distribuição da taxa de erro entre as classes

A existência de erros no processo de classificação pode ser verificada no relatório gerado pelo WEKA (figura 41A) e que contem os resultados da validação, na coluna *TP Rate (taxa de verdadeiro positivo).* O número é proporcional a taxa de acerto e se a classe não apresentar nenhum erro de classificação este número será 1.000, caso contrario, se a classe for classificada erroneamente este valor será 0.000.

Para o gênero *Klebsiella* (*classe* nove - figura 41A) pode ser observado o valor 0,333, que representa a proporção de acertos para classe (verdadeiros positivos). No caso, a classe possui três padrões e somente um padrão esta sendo corretamente classificado, gerando assim o valor 0.333. O valor restante 0.666 esta sendo classificado erroneamente e este valor representa os outros dois padrões. Isto também pode ser confirmado na matriz de confusão gerada no relatório (figura 41B), onde é possível verificar que a classe nove, representada pela letra i, possui somente um registro classificado corretamente.

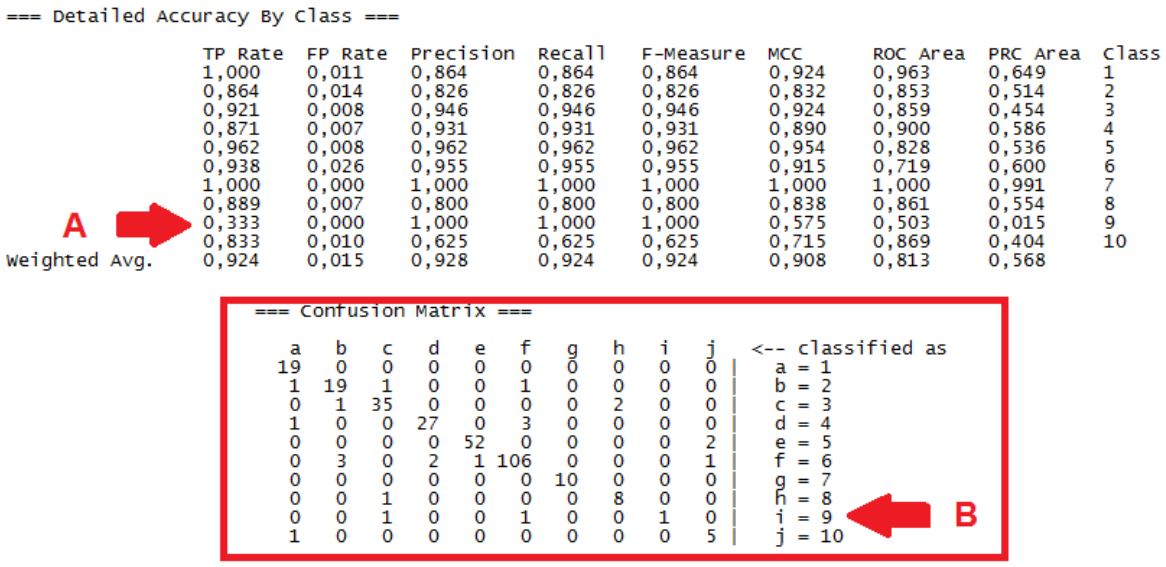



**Figura 41 – Relatório gerado pela plataforma WEKA para a rede FAN. A.coluna TP Rate e B. matriz de confusão. Gêneros de bactérias: *1 Herbaspirillum, 2 Azospirillum, 3 Burkholderia, 4 Gluconacetobacter. 5 Rhizobium, 6 Paenibacillus, 7 Bacillus 8 Pseudomonas 9 Klebsiella, 10 Azoarcus.***

FONTE: Adaptado da plataforma WEKA

## 5. Conclusões

- Um protótipo para o posicionamento taxonômico utilizando redes neurais artificiais foi construído, utilizando dados coletados de artigos que descrevem espécies de bactérias. O conjunto cadastrado contém 228 espécies pertencentes a 10 gêneros.
- Em paralelo, foi estruturado um banco de dados para armazenamento dos artigos consultados.
- A melhor estratégia para o preenchimento de dados ausentes entre as estratégias Valor Central, Mediana, Média, Moda (maior freqüência) e Valor Fora (outilier), para a utilização da ferramenta foi o Valor Central;
- A comparação entre as redes MLP, J48<ID3>, RBF, SVM e FAN, mostrou que a melhor rede neural para a utilização da ferramenta é a *FAN;*
- Gêneros que possuem poucas espécies não apresentam bons resultados na classificação;
- Foi possível realizar o posicionamento taxonômico de bactérias, em nível de gênero, utilizando somente os resultados de testes bioquímicos e fisiológicos e com a utilização de redes neurais, o que contribui com a comunidade cientifica.

## 6. Perspectivas

- Atualizar o banco de dados, cadastrando mais gêneros de bactérias e as respectivas espécies;
- Utilizar outras fontes de resultados para complementar os resultados ausentes, para isto utilizar os periódicos que forneçam artigos com os resultados dos testes que diferenciem uma espécie de outra.
- Aprimorar a seleção do conjunto mínimo de características, necessário para o treinamento da rede neural, assim permitindo obter os mesmos resultados com menos características;
- Aprimorar a ferramenta para atingir o nível taxonômico de espécie, assim sendo possível obter um resultado mais refinado.
- Disponibilizar a ferramenta para a plataforma WEB, tornando seu uso mais fácil e universal.

## 7. Referencias bibliográficas


1. ACHARYA, U. R. et al. Classifiation of heart rate data using aritificial neural network and fuzzy equivalence relation. Pattern Recognition, v. 36, p. 61-68, 2003.

2. BALDANI, J. I.; REIS, V. R. S.; TEIXEIRA, K. R. S.; BALDANI, V. L. D. Potencial biotecnológico de bactérias Diazotroficas associativas e endofíticas. In: SERAFINI, L. A.; BARROS, N. M.; AZEVEDO, J. L. (org) Biotecnologia: avanços na agricultura e na agroindústria. EDUCS, Caxias do Sul, 2002, 433p.

3. BALDANI, J.I.; CARUSO, L.; BALDANI, V.L.D.; GOI, S.R.; DÖBEREINER, J. Recent advances in BNF with non-legume plants. Soil Biology and Biochemistry, Oxford, v.29, n.5/6, p.911-922, 1997.

4. BALDANI, V.L.D. Efeito da inoculação de Herbaspirillum ssp. no processo de colonização e infecção de plantas de arroz e ocorrência e caracterização parcial de uma nova bactéria diazotrófica. Itaguaí: Universidade Federal Rural do Rio de Janeiro, 1996. 234p. Tese de Doutorado.

5. BALDANI, V.L.D., Baldani, J.I., OLIVARES, F.L., DÖBEREINER, J. 1992. Identification and ecology of Herbaspirillum seropedicae and the closely related Pseudomoas rubusubalbicas. Symbiosis 13: 65-73.

6. BALDANI, V.L.D.; ALVAREZ, M.A. de B.; BALDANI, J.I.; DÖBEREINER, J. Establishment of inoculated Azospirillum spp. in the rhizosphere and in roots of field grown wheat and sorghum. Plant and Soil, Dordrecht, v.90, n.1, p.35-46, 1986.

7. BASHEER, I. A.; HAJMEER, M. Artificial neural networks: fundamentals, computing, design, and application. Journal of Microbiological Methods, v. 43, p. 3-31, 2000.

8. Boone D. R.; CASTENHOLZ, R. W. Bergey's manual of systematic bacteriology. 2. Ed. New York: SpringVerlag, 2001. V.

9. BRAGA, A., CARVALHO, A., LUDERMIR, T. Redes Neurais Artificiais: Teoria e Aplicações, Livro Técnico e Científico, Rio de Janeiro, 2000.

10. BREIMAN, L & Spector, Submodel selection and evaluation in regression the x random case Inttrnational Statistical Review 60(3), 291-319 , 1992

11. BREIMAN, L. Bagging predictors Machine Learning, Kluwer Academic Publishers,Volume 24, 123-140, 1996.

12. CANHOS, V.P.; MANFIO, G.P.; VAZOLLER, R.F.; PELLIZARI, V.H. Diversidade no domínio bactéria. In: CANHOS, V.P.; VAZOLLER, R.F. Biodiversidade do Estado de São Paulo, Brasil: síntese do conhecimento ao final do século XX. São Paulo, FAPESP, p. 1-13. 1997.

13. CARL R. WOESE , MICROBIOLOGICAL REVIEWS, June 1987, p. 221-271 Vol. 51, No. 2 Bacterial Evolution,

14. CARL R. WOESE, OTTO KANDLER, MARK L. WHEELIS Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya

15. CAVALCANTE, V. A.; DÖBEREINER, J. A new acid-tolerant nitrogen-fixing bacterium associated with sugarcane. Plant and Soil, n. 108, p. 23 – 31, 1988.

16. CAVALIER-SMITH, T. A revised six-kingdom system of life. Biol. Rev. v. 73, p.203-66, 1998.

17. CERQUEIRA, A. Apostila de Aulas Práticas - Disciplina de Bacteriologia. Departamento de Microbiologia e Parasitologia - Instituto Biomédico. Universidade Federal Fluminense.

18. CHARLES E. Stager; Automated Systems for Identification of Microorganisms; 1992

19. CHESTER, B. Semiquantitative Catalase Test as an Aid in Identification of Oxidative and Nonsaccharolytic Gram-Negative Bacteria. Journal Of Clinical Microbiology, v. 10, nº 4, p. 58-61. 1979.

20. Colwell, R. R. & D. J. Grimes. Nonculturable microorganisms. American Society for Microbiology, Washington. 2000.

21. Colwell, R.R. Polyphasic taxonomy of bacteria. In Culture Collections of Microorganisms, pp. 421-436. H. Iizuka & T. Hasegawa. (eds.) Tokyo, University of Tokyo Press, 1970.

22. CYBENKO, G. Neural Networks in Computational Science and Engineering. IEEE Computacional Science and Engineering, 3(1):36-43, 1996.

23. DE LONG E.F., Pace N.R. Environmental diversity of bacteria and archaea. Syst Biol. v. 50:470-8, 2001

24. DEITEL, HARVEY M.; DEITEL, PAUL J. Java: Como Programar. Prentice-Hall, 2005.

25. DELEN, D. et al. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine.

26. Devijver P. and K. J., Pattern Recognition: A Statistical Approach. Londres: Prentice-Hall, 1982.

27. DÖBEREINER, J. Biological nitrogen fixation in the tropics: social and economic contributions. In: INTERNATIONAL SYMPOSIUM ON SUSTAINABLE AGRICULTURE FOR THE TROPICS – THE ROLE OF BIOLOGICAL NITROGENFIXATION, Angra dos Reis, 1995. Abstracts... Angra dos Reis: The National Centre for Agrobiology Research (Embrapa-CNPAB), 1995. p.3-4.

28. DÖBEREINER, J.; BALDANI, J.I. Bases científicas para uma agricultura biológica. Ciência e Cultura, São Paulo, v.34, n.7, p.869-881, 1982.

29. DREYFUS, B.; GARCIA, J.L.; GILLIS, M. Characterization of Azorhizobium caulinodans gen. nov., sp. nov., a stem-nodulating nitrogen-fixing bacterium isolated from Sesbani rostrata. International Journal of Systematic Bacteriology, Baltimore, v.38, n.1, p.89-98, 1988.

30. EASYFAN. Kuster, C. V.; Ignacio, F. A.; Lenfers, F. P.; Garrett, L. F. V.; Zotto, S. EasyFan. 2006. Trabalho de Conclusão de Curso. (Graduação em Tecnólogo em Informática) - Universidade Federal do Paraná. Curitiba. Disponivel em HTTP://easyfan.souceforge.net/

31. EFRON B., “Bootstrap Methods: Another Look at the Jackknife”, Annals of Statistics, Vol. 7,1979, pp. 1-26

32. FAYYAD, U. M.; PIATESKY-SHAPIRO, G.; SMYTH, P. From Data Mining to Knowledge Discovery: An Overview. In: Advances in Knowledge Discovery and Data Mining, AAAI Press, 1996.

33. FRANCO, A.A.; DÖBEREINER, J. A biologia do solo e a sustentabilidade dos solos tropicais. Summa Phytopathológica, São Paulo, v.20, n.1, p.68-74, 1994.

34. Gonçalves ; E. C. Gonçalves, "Mineração de dados na pratica com Weka API", sql magazine,v 107, 2013

35. GRAHAM, M.H; HAYNES, R.J. Catabolic diversity of soil microbial communities under sugarcane and other land uses estimated by Biolog and substrate-induced respiration methods. Applied Soil Ecology. v. 29, nº 2, p. 155-164. 2005.

36. GRISI, T. C. S. L. Diversidade de Bactéria e Archaea do solo do Cariri paraibano e prospecção de celulases e xilanases em clones metagenômicos e isolados bacterianos. João Pessoa, 2011. Tese (Doutorado em Biotecnologia em recursos naturais) – Programa de Pós-Graduação da Rede Nordeste de Biotecnologia – RENORBIO. Universidade Federal da Paraíba.

37. GUCKERT, J.B.;CARRB, G.J.; JOHNSONB, T.D.; HAMM, B.G.; DAVIDSONA, D.H.; KUMAGAI, Y. Community analysis by Biolog: curve integration for statistical analysis of activated sludge microbial habitats. Journal of Microbiological Methods. v. 27, p. 183-197. 1996.

38. GUPTA, M. M.; JIN, L.; HOMMA, N. Static and Dynamic Neural Networks: From Fundamentals to Advanced Theory. [S.l.]: Wiley-IEEE Press, 2003.

39. GUIXIANG P.,HUARONG W., GUOXIA Z., Wei H., Yang L., En T. W., ZHIYUAN T.; Azospirillum melinis sp. nov., a group of diazotrophs isolated from tropical molasses grass, International Journal of Systematic and Evolutionary Microbiology (2006), 56, 1263–1271

40. GUIZELINI, D., Pedrosa, F. O., MARCHAUKOSKI, J. N. , FERREIRA, L. M. , STEFFENS, M. B. R., Gehlen, M. A. C. , RAITTZ, R. T., GENEBINGO: IDENTIFICAÇÃO DE GENES UTILIZANDO REDE NEURAL ARTIFICIAL, 10th Brazilian Congress on Computational Intelligence (CBIC’2011), November 8 to 11, 2011, Fortaleza, Ceará Brazil, 2011

41. HAYKIN, S. Redes Neurais, Principios e pratica. 2. ed. [S.l.]: Bookman, 1999.

42. HAYKIN, S. Redes neurais: princípios e prática. 2.ed. Porto Alegre, Bookman, 2001.

43. HOGG, S. Essential Microbiology 2005 John Wiley & Sons Ltd, West Sussex England 468 pp

44. HUGENHOLTZ, P.; GOEBEL, B. M.; PACE, N. R. Impact of cultureindependent studies on the emerging phylogenetic view of bacterial diversity. J.Bacteriol., v. 180, n 18, p. 4765-4774, 1998b.

45. IJSEM. International Journal of Systematic and Evolutionary Microbiology. <http://ijs.sgmjournals.org/>, Ultimo acesso 04/05/2013

46. Aguiar, H., Junior, O. Caldeira, A. M., Machado, M. A. S., Souza, R. C., Tanscheit R., Inteligência Computacional Aplicada à Administração, Economia e engenharia em Matlab. Ed. Thomson, p 370, 2007.

47. JURTSHUK, P. JR.; McQUITTY, D.N. Quantitation of the Tetramethyl-p-Phenylenediamine Oxidase Reaction in Neisseria Species. Applied and Environmental Microbiology. v. 31, n. 5, p. 668-679. 1976.

48. KOHAVI, R. A study a cross validation a bootstrap for accuracy estimation and a model selection. In: International Joint Conference on Artificial Intelligence (IJCAI). [S.l.: s.n.], 1995.

49. Lengeler, J.W., Drews, G., Schlegel H.G. Biology of Prokariotes. New York, Blackwell Sciences, 921p. 1999.

50. LIPPMANN, R. An introduction to computing with neural nets. ASSP Magazine, IEEE, v. 4, n. 2, p. 4{22, 1987}.

51. M. A. PFALLER,Comparison of the autoScan-w/a rapid bacterial system and the Vitek dor identificationof gram-negativo bacill. 1991.

52. M. C. P. Souto, A. C. Lorena, A. C. B. Delbem, and A. C. P. L. F. Carvalho. Técnicas de Aprendizado de Máquina para problemas de Biologia Molecular, pages 103–152. Minicursos de Inteligência Artificial, Jornada de Atualização Científica em Inteligência Artificial, XXIII Congresso da Sociedade Brasileira de Computação, 2003.

53. Machado Filho O. M. AMBIENTE DE MINERAÇÃO DE DADOS UTILIZANDO REDES NEURAIS OTIMIZADAS POR ALGORITMOS GENÉTICOS E TÉCNICA DE VISUALIZAÇÃO, 2006

54. MAGNANI, G. S.; Diversidade de bactérias endofíticas em cana-de-açúcar. Curitiba, 2005. Dissertação (Mestrado em Ciências – Bioquímica). Departamento de Bioquímica e Biologia Molecular – Setor de Ciências Biológicas. Universidade Federal do Paraná.

55. MARK L. WHEELIS, OTTO KANDLER, CARL R. WOESE On the nature of global classification Proc. Nati. Acad. Sci. USA Vol. 89, pp. 2930-2934, April 1992

56. MCCULLOCH, W. S.; PITTS, W. H. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, v. 5, p. 115-133, 1943.

57. Moreira, F. M. S., Siqueira, J. O. Microbiologia e Bioquímica do Solo. Lavras, MG, Editora UFLA, 2006, 729 p.

58. O'DONNELL, A. G.; CORRES. H 16s rDNA methods in soil microbiology. Currente Opinion In Biotechnology, v. 10, p 225-229, 1999.

59. OKON, Y.; LABANDERA-GONZÁLEZ, C.A. Agronomic applications of Azospirillum: an evaluation of 20 years worldwide field inoculation. Soil Biology and Biochemistry, v. 26, n.12, p. 1591-1601, 1994.

60. PANDYA, A.; MACY, R. B. Pattern Recognition with Neural Networks in C++. CRC Press, 1995.

61. POSTGATE, J. R. Nitrogen fixation. Cambridge, Cambridge Univ. Press. 112p. 1998.

62. PHOENIX, Disponível em <http://www.bd.com/scripts/brasil/productsdrilldown.asp?CatID=115&SubID=308&siteID=10056&d=brasil&s=brasil&sTitle=&metaTitle=Microbiologia&dc=brasil&dcTitle=BD+-+Brasil>, Acessado em 02/02/2013

63. Proc. Nati. Acad. Sci. USA Vol. 87, pp. 4576-4579, June 1990

64. QUINLAN, J. R. C4.5: Programs for machine learning. San Mateo, CA: Morgan Kaufmann Publishers, 1993.

65. Raittz, R. T. Fan 2002: Um modelo neuro-fuzzy para reconhecimento de padroes. 2002. Tese (Doutorado em Engenharia de Produção), Universidade Federal de Santa Catarina, Florianópolis.

66. RAPPÉ, M. S.; GIOVANONNI, S. J. The uncultured microbial majority. Annu. Rev. Microbiol., v. 67, p. 369-394, 2003.

67. REZENDE, S. O., Sistemas Inteligentes: Fundamentos e aplicações, Ed. Manole, p. 535, 2005.

68. ROMERO, E. M.; PALACIOS, R.; MORA, J. Cepas mejoradas de Rhizobium. Investigación y Ciencia,n. 8, p. 14 – 19, 1998.

69. SAIKI, R. K; GELFAND, D. H.; STOFFEL, S. SHARF, S. J.; HIGUCHI, R.; HORN, G. T.; MULLIS, K.; ERLICH, H. A. primer-directed enzymatic amplification of DNA witha thermostable DNA polymerase. Science, v. 239, p487-,1988.

70. SANGER, F. NICKLEN, S., COULSON, A. R. DNA sequencing with chain terminating inhibitors. Proc. Natl. Acad. Sci,. V. 74. 5463-5467, 1977.

71. Schleifer, K.H. Classification of Bacteria and Archaea: Past, present and future. System. Appl. Microbiol. V. 32, p. 533-42, 2009.

72. Sellenriek, Patricia; Comparison of MicroScan Walk-way®, Phoenix™ and VITEK-TWO® Microbiology Systems Used in the Identification and Susceptibility Testing of Bacteria

73. SEWELL, M. Feature Selection. 2007. Disponível em http://machine-learning.martinsewell.com/feature-selection/feature-selection.pdf. Ultimo acesso 02/05/2013.

74. Silva Filho , A. S., Inferência em Amostras Pequenas: Métodos Bootstrap.

75. SOUZA, J. A. Reconhecimento de padrões usando indexação recursiva.Tese de Doutorado, Universidade Federal de Santa Catarina, 1999.

76. SPRENT, J.I.; SPRENT, P. Nitrogen fixing organisms. London: Chapman and Hall, 2ed., 1990. 256p.

77. T. Mitchell. Machine Learning. McGraw Hill, 1997.

78. TARRAND, J. J.; GROSCHEL, D. H.; Rapid, Modified Oxidase Test for Oxidase-Variable Bacterial Isolates. Journal of Clinical Microbiology, v. 16, nº 4, p. 772-774. 1982.

79. TAYLOR, W. I.; ACHANZAR, D. Catalase Test as an Aid to the Identification of Enterobacteriaceae. Applied Microbiology, v. 24, nº 1, p. 58-61. 1972.

80. THEODORIDIS, S. e KOUTROUMBAS, K. Pattern Recognition. Elsevier, second edition, 2003.

81. V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications, 16(2):283–305, 1971.

82. Vandamme, P.; POT, B; GILLS, M.; DEVOS, P; KERSTERS, K; SWINGS, J. Polyphansic taxonomy, a consensus approach to bacterial systematics. Microbiology Reviews, Washington, v 60, n2, p. 407-437, 1996.

83. VAPNIK, V. N. The nature of Statistical learning theory. Springer-Verlag, New York, 1995.

84. VIDEIRA, S. S.; ARAÚJO, J. L. S.; BALDANI, V. L. D. Metodologia para Isolamento e Posicionamento Taxonômico de Bactéria Diazotróficas Oriundas de Plantas Não-Leguminosas. Seropédica: Embrapa Agrobiologia, (Documentos/Embrapa Agrobiologia ISSN 1577-8498, 234), p. 74, 2007.

85. VON ZUBEN, F.; ATTUX, R. R. Redes Neurais com Funcao de Base Radial. 2008. Disponivel em: <ftp://ftp.dca.fee.unicamp.br/pub/docs/vonzuben/ia353 1s07/topico9 07.pdf>.

86. WHITMAN, W. B.; COLEMAN, D. C.; WIEBE, W. Prokaryotes: the unseen majority. Proc. Natl. Acad. Sci. USA, v. 95, n. 12, p. 6578-6583, 1998.

87. Witten, I.H. and Frank, E. Data Mining: Practical machine learning tools and techniques. 2nd edition Morgan Kaufmann, San Francisco, 2005.

88. WOESE, C. R. Bacterial evolution. Microbial Rev., v. 51, n. 2, p. 221-271, 1987.

89. WOESE, C. R.; GUTELL, R.; GUPTA, R.; NOLLER, H. F. Detailed Analysis of the Higher-Order Structure of 16S-Like Ribosomal Ribonucleic Acids. Microbial Rev, v. 47, n°4, p. 621-669, 1983.

90. WOESE, C. R.; KANDLER, O.; WHEELIS, M. L. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. USA, v. 87, p. 4576-4579, 1990.

91. WAN, V. e CAMPBELL, W. Support vector machines for speaker verification and identification, IEEE Proceeding, 2000.

92. YABUUCHI, E.; KOSAKO, Y.; OYAIZU, H.; YANO, I.; HOTTA, H.; HASHIMOTO, Y.; EZAKI, T.; ARAKAWA, M. Proposal of Burkholderia gen. nov. and transfer of seven species of the genus Pseudomonas homology group II to the new genus, with the type species Burkholderia cepacia (Palleroni and Holmes, 1981) comb. nov. Microbiology and Immunology, Tokyo, v.36, p.1251-1275, 1992.

93. YANO, D.M.Y.; ATTILI, D.S.; GATTI, M.S.V.; EGUCHI, S.Y.; OLIVEIRA, U.M. Técnicas de Microbiologia em controle de qualidade. Campinas: Fundação Tropical de Pesquisa e Tecnologia “André Tosello”, 1991.

94. YOUNG, J.P.W. Phylogenetic Classification of Nitrogen-Fixing Organisms. In:Biological Nitrogen Fixation. Ed. G. STACEY, R.M. BURRIS, H.S. EVANS. London Chapman & Hall, p. 43-86, 1992.

# 8. Anexos

## Anexo 1 – Gêneros e espécies de bactérias cadastradas.

| *Azoarcus* | | | |
|---|---|---|---|
| *Espécie* | Estirpe | *Espécie* | Estirpe |
| *A.anaerobius* | | *A.toluclasticus* | |
| *A.communis* | | *A.tolulyticus* | |
| *A.indigens* | | *A.toluvorans* | |
| *Azospirillum* | | | |
| *Espécie* | Estirpe | *Espécie* | Estirpe |
| *A.amazonense* | ATCC 35119 | *A.irakense* | |
| *A.amazonense* | LMG 22237 | *A.largimobile* | ACM 2041T |
| *A.brasilense* | ATCC 29145 | *A.largimobile* | |
| *A.brasilense* | DSM 1690 | *A.lipoferum* | ATCC 29707T |
| *A.canadense* | DS2 | *A.melinis* | TMCY 0552 |
| *A.canadense* | LMG 23617 | *A.oryzae* | COC8T |
| *A.dobereinerae* | | *A.oryzae* | IAM 15130 |
| *A.doebereinerae* | DSM 13131T | *A.picis* | IMMIB TAR-3T |
| *A.halopraeferens* | DSM 3675T | *A.rugosum* | IMMIB AFH-6T |
| *A.halopraeferens* | | *A.zeae* | LMG 23989T |
| *A.irakense* | CIP 103311 | *A.zeae* | N7T and N6 |
| *Bacillus* | | | |
| *Espécie* | Estirpe | *Espécie* | Estirpe |
| *B.bataviensis* | LMG 21833T | *B.niacini* | DSM 2923T |
| *B.drentensis* | LMG 21831T | *B.novalis* | LMG 21837T |
| *B.foraminis* | LMG 23174T | *B.pocheonensis* | Gsoil 420T |
| *B.fumarioli* | LMG 17489T | *B.soli* | LMG 21838T |
| *B.jeotgali* | JCM 10885T | *B.vireti* | LMG 21834T |
| *Burkholderia* | | | |
| *Espécie* | Estirpe | *Espécie* | Estirpe |
| *B.caribensis* | MWAP64T | *B.sacchari* | |
| *B.caribensis* | KCTC 2964T | *B.sacchari* | LMG 19450T |
| *B.caryophylli* | KCTC 2965T | *B.silvatlantica* | |
| *B.cepacia* | ALQ 8281 | *B.solanacearum* | ATCC 1 1696T |
| *B.cepacia* | ATCC 6344T | *B.sordidicola* | KCTC 12081 |
| *B.cepacia* | ATCC 25416T | *B.sordidicola* | KCTC 12082 |
| *B.cepacia* | KCTC 2966T | *B.thailandensis* | DSM 13276T |
| *B.cepacia* | LMG 1222T | *B.tropica* | |
| *B.ferrariae* | feGl01T | *B.unamae* | |
| *B.fungorum* | kCTC 12917 | *B.unamae* | |
| *B.gladioli* | ATCC 1024gT | *B.vandii* | CY-0619 |
| *B.gladioli* | ATCC 19302 | *B.vandii* | CY-0627 |

| | | | |
|---|---|---|---|
| *B.glathei* | KCTC 2968T | *B.vandii* | D-2251 |
| *B.kururiensis* | KP23T | *B.vandii* | VA-1316 |
| *B.mimosarum* | | *B.vandii* | VU-0563 |
| **Continuação** | | | |
| *B.phenazinium* | KCTC 2971T | *B.vietnamiensis* | |
| *B.pickettii* | JCM 5969T | *B.vietnamiensis* | KCTC 2974T |
| *B.rhizoxinica* | HKI 454T | *B.vietnamiensis* | TVV75T |
| *B.sacchari* | IPT 101 | *B.xenovorans* | LMG 21463T |

| *Gluconacetobacter* | | | |
|---|---|---|---|
| *Espécie* | Estirpe | *Espécie* | Estirpe |
| *G.azotocaptans* | | *G.oboediens* | LMG 1688 |
| *G.diazotrophicus* | | *G.oboediens* | LMG 1689 |
| *G.entanii* | | *G.oboediens* | LMG 18849T |
| *G.entanii* | LTH 4560T | *G.oboediens* | NBRC 14822 |
| *G.europaeus* | NBRC 3261 | *G.rhaeticus* | LMG 22126T |
| *G.europeaus* | | *G.rhaeticus* | |
| *G.hansenii* | NBRC 14815 | *G.saccharivorans* | LMG 1582T |
| *G.hansenii* | NBRC 14816 | *G.saccharivorans* | LMG 1584 |
| *G.hansenii* | NBRC 14817 | *G.swingsii* | |
| *G.hansenii* | NBRC 14820T | *G.swingsii* | LMG 22125T |
| *G.intermedius* | LMG 18909T | *G.xylinus* | ACM19 |
| *G.johannae* | | *G.xylinus* | |
| *G.kombuchae* | RG3T | *G.xylinus* | JCM 10150 |
| *G.liquefuciens* | LMG 1381T | *G.xylinus* | JCM 7644T |
| *G.nataicola* | LMG 1536 | *G.xylinus* | JCM 9730 |
| *G.oboediens* | | | |

| *Herbaspirillum* | | | |
|---|---|---|---|
| *Espécie* | Estirpe | *Espécie* | Estirpe |
| *H.autotrophicum* | DSM 732T | *H.lusitanum* | LMG 21760 |
| *H.autotrophicum* | IAM 14942T | *H.lusitanum* | LMG 21710T |
| *H.chlorophenolicum* | CPW301T | *H.lusitanum* | P6-12T |
| *H.chlorophenolicum* | IAM 15024T | *H.putei* | IAM 15032 |
| *H.frisingense* | IAM 14974 | *H.rhizosphaerae* | UMS-37T |
| *H.frisingense* | GSF30T | *H.rubrisubalbicans* | DSM 9440T |
| *H.hiltneri* | N3T59 | *H.rubrisubalbicans* | IAM 14976 |
| *H.Hiltneri* | | *H.seropedicae* | DSM 6445T |
| *H.huttiense* | DSM 10281 | *H.seropedicae* | IAM 14977 |
| *H.Huttiensis* | IAM 14941T | | |

| *Klebsiella* | | | |
|---|---|---|---|
| *Espécie* | Estirpe | *Espécie* | Estirpe |
| *K.singaporensis* | lx3 | *K.terrigena* | |
| *K.trevisanii* | | | |

| *Paenibacillus* |
|---|

| *Espécie* | Estirpe | *Espécie* | Estirpe |
|---|---|---|---|
| *P.agarexedens* | KCTC 3848T | *P.montaniterrae* | MXC2-2T |
| *P.agaridevorans* | KCTC 3849T | *P.naphthalenovorans* | |
| **Continuação** | | | |
| *P.alkaliterrae* | KCTC 3956T | *P.odorifer* | TOD45T |
| *P.alvei* | | *P.pabuli* | |
| *P.alvei* | ATCC 6344T | *P.pabuli* | CIP 103119T |
| *P.amylolyticus* | | *P.pabuli* | NRRL NRS-924T |
| *P.amylolyticus* | NRRL NRS-290T | *P.pasadenensis* | SAFN-007T |
| *P.amyloyticus* | NRRL B-14945T | *P.pasadenensis* | SAFN-016T |
| *P.anaericanus* | MH21T | *P.pasadenensis* | SAFN-125 |
| *P.assamensis* | GPTSA 11T | *P.peoriae* | |
| *P.azoreducens* | DSM 13822T | *P.peoriae* | |
| *P.azotofixans* | | *P.peoriae* | IFO 15541T |
| *P.azotofixans* | ATCC 35681T | *P.peoriae* | LMG 14832T |
| *P.barengoltzii* | SAFN-016T | *P.phyllosphaerae* | |
| *P.borealis* | KK19T | *P.phyllosphaerae* | CCM 7310T |
| *P.brasilensis* | DSM 14914T | *P.polymyxa* | ATCC 842T |
| *P.brasilensis* | PB172T | *P.polymyxa* | |
| *P.campinasensis* | KCTC 0364BPT | *P.polymyxa* | CIP66.22T |
| *P.chibensis* | HSCC | *P.polymyxa* | DSM 36T |
| *P.chibensis* | NRRL B-142T | *P.polymyxa* | NRRL B-4317T |
| *P.chinjuensis* | WN9T | *P.provencensis* | 4401170T |
| *P.curdlanolyticus* | | *P.pueri* | b 13i |
| *P.curdlanolyticus* | CCM 4536T | *P.pueri* | b09i |
| *P.dendritiformis* | T168 | *P.riograndensis* | SBR5T |
| *P.dendritiformis* | T168T | *P.sabinae* | G18-7 |
| *P.durus* | | *P.sabinae* | JD2 |
| *P.favisporus* | GMP01T | *P.sabinae* | T2712 |
| *P.fonticola* | ZLT | *P.sabinae* | T49 |
| *P.forsythiae* | DSM 17842T | *P.sabinae* | T67 |
| *P.ginsengihumi* | DCY16T | *P.sabinae* | DSM 17841T |
| *P.glucanolyticus* | | *P.sanguinis* | 2301083T |
| *P.glycanilyticus* | JCM 11221T | *P.septentrionalis* | X13-1T |
| *P.glycanilyticus* | KCTC 3808T | *P.sepulcri* | CCM 7311T |
| *P.graminis* | RSA19T | *P.siamensis* | S5-3T |
| *P.illinoisensis* | CIP105253T | *P.sonchi* | X19-5T |
| *P.illinoisensis* | | *P.stellifer* | DSM 14472T |
| *P.illinoisensis* | NRRL NRS-1356 | *P.terrae* | |
| *P.jamilae* | | *P.terrae* | AM141T |
| *P.jamilae* | B.3455 | *P.terrae* | MH72 |
| *P.kobensis* | | *P.thailandensis* | MX2-3T |
| *P.kobensis* | CCM 4537T | *P.thailandensis* | S3-4A |

| | | | |
|---|---|---|---|
| *P.kobensis* | IFO 15729T | *P.thiaminolyticus* | JCM 8360T |
| *P.koreensis* | KCTC 2393T | *P.timonensis* | 2301032T |
| *P.kribbensis* | | *P.timonensis* | CCUG 48216T |
| **Continuação** | | | |
| *P.kribbensis* | AM49T | *P.tundrae* | |
| *P.larvae* | | *P.turicensis* | MOL722T |
| *P.lautus* | | *P.urinalis* | 5402403T |
| *P.lautus* | NRRL NRS-666T | *P.urinalis* | |
| *P.macerans* | | *P.validus* | DSM 3037T |
| *P.macerans* | ATCC 8244T | *P.validus* | |
| *P.macerans* | CIP 66.19T | *P.woosongensis* | YB-45T |
| *P.macerans* | NRRL B-172T | *P.wynnii* | LMG 22176T |
| *P.macquariensis* | ATCC 23464 | *P.xylanexedens* | |
| *P.macquariensis* | | *P.xylanilyticus* | CIP 109086T |
| *P.massiliensis* | 2301065T | *P.zanthoxyli* | |
| *P.massiliensis* | CIP 107939T | *P.zanthoxyli* | DSM 18202T |
| *P.mendelii* | CCM 4839T | | |

| *Pseudomonas* | | | |
|---|---|---|---|
| *Espécie* | Estirpe | *Espécie* | Estirpe |
| *P.alcaligenes* | | *P.koreensis* | |
| *P.citronellolis* | | *P.nitroreducens* | |
| *P.glumae* | KCTC 2969T | *P.pavonaceae* | |
| *P.jessenii* | | *P.umsongensis* | |
| *P.jinjuensis* | | | |

| *Rhizobium* | | | |
|---|---|---|---|
| *Espécie* | Estirpe | *Espécie* | Estirpe |
| *R.alkalisoli* | CCBAU 01393T | *R.loessense* | |
| *R.cellulosilyticum* | ALA10B2T | *R.loessense* | CCBAU 7190BT |
| *R.cellulosilyticum* | ALA38.2 | *R.loti* | NZP 2213T |
| *R.cellulosilyticus* | LMG 23642T | *R.lusitanum* | P1-7T |
| *R.ciceri* | IC-60 | *R.mesosinicum* | CCBAU 25010T |
| *R.ciceri* | UPM-Ca7 | *R.mesosinicum* | CCBAU 25217 |
| *R.daejeonense* | L22 | *R.mesosinicum* | CCBAU 41044 |
| *R.daejeonense* | CCBAU 10050T | *R.miluonense* | CCBAU 41251T |
| *R.etli* | | *R.mongolense* | USDA 1844T |
| *R.etli* | CFN 454 | *R.mulithospitium* | CCBAU 83401T |
| *R.etli* | CFN 42T | *R.oryzae* | Alt 505T, Alt 501 |
| *R.fabae* | | *R.phaseoli* | ATCC 14482T |
| *R.galegae* | | *R.pisi* | DSM 30132T |
| *R.galegae* | ATCC 43677T | *R.radiobacter* | DSM 30148T |
| *R.galegae* | USDA 4128T | *R.rhizogenes* | LMG 150T |
| *R.galegae* | HAMBI 540T | *R.rubi* | IFO 13261T |
| *R.gallicum* | USDA 2918T | *R.sullae* | USDA 4950T |

| | | | |
|---|---|---|---|
| *R.gallicum* | R602spT | *R.sullae* | IS123T |
| *R.giardinii* | H152T | *R.tibeticum* | CCBAU 85039T |
| *R.hainanense* | 166T | *R.tropici* | CIAT 889 |
| **Continuação** | | | |
| *R.hainanense* | | *R.tropici* | |
| *R.huautlense* | huautlense | *R.tropici* | CFN 899 |
| *R.huautlense* | SO2T | *R.tropici* | CFN 299T |
| *R.larrymoorei* | AF3-10T | *R.undicola* | LMG11875T |
| *R.leguminosarum* | | *R.vitis* | NCPPB 3554T |
| *R.leguminosarum* | ATCC 14480 | *R.yanglingense* | CCBAU 71623T |
| *R.leguminosarum* | USDA 2048 | | |

FONTE: Autor, 2013