RAFAEL EDUARDO RUVIARO CHRIST

**Classificação de bactérias do gênero
Bradyrhizobium usando uma rede neural
ART2 com dados de eletroforese de genes
ribossomais**

Dissertação de Mestrado apresentada à Escola de Engenharia de São Carlos, da Universidade de São Paulo, como parte dos requisitos para obtenção do Título de Mestre em Engenharia Elétrica.

Área de Concentração: Processamento de Sinais e Instrumentação
Orientador: Prof. Dr. Carlos Dias Maciel

São Carlos
2007

Agradecimentos

A Deus.

Aos meus pais por todo esforço que realizaram para que eu chegasse até aqui.

A meu orientador Prof. Carlos Dias Maciel por ter me concedido a oportunidade de realizar este trabalho, pela orientação e pela amizade.

A minha Irmã Mara por ter me aguentar sempre.

A Patrícia por ser tão maravilhosa.

Aos docentes e funcionários do Departamento de Engenharia Elétrica da Escola de Engenharia de São Carlos, pela atenção recebida.

A todos os amigos pelo apoio.

Ao CPNq pelo suporte financeiro e bolsa de estudo e à Escola de Engenharia de São Carlos (EESC/USP) pela infra-estrutura e pós-graduação.

Resumo

Christ, Rafael Eduardo Ruviaro (2007). *Classificação de bactérias do gênero *Bradyrhizobium* usando uma rede neural ART2 com imagens de eletroforese de genes ribossomais*. Dissertação (Mestrado), Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2007.

Os microrganismos apresentam uma enorme diversidade genética, sendo componentes fundamentais de cadeias alimentares e ciclos biogeoquímicos. Estima-se que menos de 10% dos microrganismos existentes tenham sido caracterizados e descritos. A taxonomia de bactérias tropicais pertencentes ao gênero *Bradyrhizobium* é pouco conhecida. Essas bactérias fixadoras de nitrogênio são críticas para a produção com altos rendimentos em cultivos de soja no Brasil, com 25% da safra mundial. O objetivo deste trabalho é classificar uma coleção de estirpes brasileiras do gênero *Bradyrhizobium* utilizando uma rede neural ART2. São 119 estirpes com características fenóticas de *B. elkanii* e *B. japonicum*, isoladas de 33 espécies de leguminosas tropicais. Os dados consistem de imagens de gel de eletroforese obtidas a partir da análise RFLP-PCR de 3 regiões ribossomais: 16S, 23S e IGS. Foram utilizadas as enzimas de restrição: Cfo I, Dde I, Msp I, Hae III, Hha I e Hinf I. Este trabalho apresentou algoritmos que podem auxiliar no estudo taxonômico de estirpes de *Bradyrhizobium* a partir de imagens de gel de eletroforese. Os resultados são compatíveis com aqueles encontrados na literatura.

Palavras-Chave: Classificação, Rede ART2, *Bradyrhizobium*.

Abstract

Christ, Rafael Eduardo Ruviano (2007). *Classification of Bradyrhizobium strains using an ART2 neural network using electrophoresis in a gel images from ribosomal genes*. Dissertation (Master), Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2007.

The microorganisms show an enormous genetic diversity. They are rather important components of food chains and biogeochemical cycles. It is estimated that less than 10% of the living microorganisms have been studied. The taxonomy of tropical bacteria belonging to the genus *Bradyrhizobium* is still poorly known. Those N_2 fixing nitrogen bacteria are critical for high income soybean production in Brazil, responsible for 25% of the world-wide production. The main objective of this work is to classify a Brazilian collection of 111 *Bradyrhizobium* strains showing phenotypic characteristics of *B. elkanii* and *B. japonicum*, isolated from 33 tropical legume species. The data consist of electrophoresis in a gel images obtained from PCR-RFLP analysis of 3 ribosomal regions: 16S, 23S e IGS. They were digested with 3 restriction endonuclease: CfoI, MspI and DdeI. This work presented some algorithms which can help in taxonomic studies of *Bradyrhizobium* strains using electrophoresis in a gel images. The results are similar to that available in the literature.

Key-words: Classification, ART2 network, *Bradyrhizobium*.

Lista de Figuras

2.1	Bactérias do gênero <i>Bradyrhizobium</i> , quando associadas às raízes de soja, formam nódulos, onde ocorre a conversão do nitrogênio da atmosfera em compostos nitrogenados, que serão utilizados pela planta (Embrapa, 2007).	8
2.2	Exemplo de imagem produzida pela técnica de gel de eletroforese. Pode-se observar 7 canaletas com bandas em posições distintas. .	11
2.3	Exemplo de ensaio. Gel de eletroforese da região IGS rRNA de 24 estirpes digeridas com a enzima de restrição Msp I.	13
2.4	A estirpe 1 é descrita por um conjunto de 9 canaletas recortadas dos ensaios de gel de eletroforese. Cada canaleta corresponde a uma região do rRNA digerida por uma enzima de restrição. Para que as estirpes possam ser classificadas pela rede ART2, estas devem ser transformadas em um vetor.	14
2.5	A informação das canaletas está na posição de suas bandas. Bandas são regiões escuras que indicam a presença de material genético. Neste trabalho, cada canaleta é transformada em um eletroferograma, cujos picos representam as bandas.	15
2.6	Representação dos pontos gerados com o algoritmo DISTATIS para a base de dados Iris, a qual é descrita por 4 variáveis. MDS selecionou apenas as 3 variáveis mais relevantes. O espaço gerado $P_{150 \times 3}$ é uma aproximação do espaço original $P_{150 \times 4}$	18

2.7	Arquitetura da rede ART2 (modificado de Filho et al. (1997)). A camada F1 processa os dados na entrada, a camada F2 representa as classes e R é usado no teste de <i>reset</i>	21
2.8	O processamento na rede ART2 pode ser resumido nas seguintes etapas: Pré-processamento, Cálculo da similaridade, Ciclo de busca e Aprendizagem.	23
2.9	Exemplo de classificação com a rede ART2 em 3 dados tridimensionais A , B e C . A rede ART2 criou 2 classes, representadas espacialmente pelo cone 1 e 2, para classificar esses dados. Os dados A e B estão dentro da classe 1 e o dado C pertence a classe 2 (modificado de Whiteley et al. (1996)).	25
2.10	Visão dos subsistemas de atenção e orientação os quais controlam a dinâmica da informação na rede ART2 (modificado de Heins e Tauritz (1995)).	26
2.11	Passo 1 do ciclo de busca (modificado de (Heins e Tauritz, 1995)).	28
2.12	Passo 2 do ciclo de busca (modificado de (Heins e Tauritz, 1995)).	29
2.13	Passo 3 do ciclo de busca (modificado de (Heins e Tauritz, 1995)).	29
2.14	Passo 4 do ciclo de busca (modificado de (Heins e Tauritz, 1995)).	30
3.1	Base de dados Iris composta por três espécies de flores: (1) iris-setosa, (2) iris-versicolor e (3) iris-virginica.	34
3.2	Base de dados <i>Balance</i> . A combinação de quatro atributos leva a balança a três estados: esquerda, direita, ou equilibrada.	35
3.3	Geração das matrizes de distância a partir das estirpes. A primeira canaleta de cada estirpe é colocada na matriz $16S - CfoI$, a segunda na matriz $16S - DdeI$. Esse passo gera 9 matrizes de distância, uma para cada região do rRNA digerida por uma enzima específica.	41

3.4	As matrizes de distância são agrupadas em uma única C através de (3.2). A matriz C serve como entrada para o processamento com MDS. O algoritmo DISTATIS transforma C em uma matriz de pontos em que cada linha é um dado I representando uma estipe.	43
3.5	Valores de ρ produzidos com uma função logarítmica pelo algoritmo <i>rho_generator</i> com 27 valores no intervalo $0.94 \leq \rho < 1$.	45
3.6	Número de classes geradas pela rede ART2 em função de ρ usando os valores representados na Figura 3.5. O número de classes cresce mais rapidamente quando o valor de ρ se aproxima de 1. Este gráfico mostra os resultados da rede ART2.	45
3.7	Distâncias inter e intra-classe como função do parâmetro de vigilância (ρ). Cada mudança no gráfico mostra o aumento no número de classes. O ponto de intersecção entre as equações define o número ótimo de classes, neste caso são identificadas 13 classes com $\rho = 0.9873$.	46
3.8	Exemplo de grafo gerado pelo Algoritmo (1). A rede ART2 gerou 3 partições com 1, 2 e 3 classes respectivamente. O grafo mostra como os dados são redistribuídos pela rede quando se altera o número de classes.	48
4.1	Resultado dos classificadores para cada base de dados. Cada linha corresponde a uma base de dados.	50
4.2	Base Iris. Distâncias inter e intra-classe como função do parâmetro de vigilância (ρ). Cada mudança no gráfico mostra o aumento no número de classes. O ponto de intersecção entre as equações define o número ótimo de classes, neste caso 3 classes com $\rho = 0.998121$.	51
4.3	Base <i>Glass</i> . Distâncias inter e intra-classe como função de ρ . Cada mudança no gráfico mostra o aumento no número de classes. O ponto de intersecção entre as equações define o número ótimo de classes, neste caso 6 classes com $\rho = 0.999750$.	52

4.4	Base <i>Wine</i> . Distâncias inter e intra-classe como função de ρ . Cada mudança no gráfico mostra o aumento no número de classes. O ponto de intersecção entre as equações define o número ótimo de classes, neste caso 3 classes com $\rho = 0.999318$	52
4.5	Base <i>Zoo</i> . Distâncias inter e intra-classe como função de ρ . Cada mudança no gráfico mostra o aumento no número de classes. O ponto de intersecção entre as equações define o número ótimo de classes, neste caso 8 classes com $\rho = 0.966478$	53
4.6	Base <i>Balance</i> . Distâncias inter e intra-classe como função de ρ . Cada mudança no gráfico mostra o aumento no número de classes. O ponto de intersecção entre as equações define o número ótimo de classes, neste caso 13 classes com $\rho = 0.987316$	53
4.7	Base de estirpes do gênero <i>Bradyrhizobium</i> . Distâncias inter e intra-classe como função do parâmetro de vigilância (ρ). O ponto de intersecção entre as equações define o número ótimo de classes, neste caso 4 classes com $\rho = 0.994893$	54
4.8	Grafo produzido pelo algoritmo <i>DrawGraph</i> para a base de estirpes do gênero <i>Bradyrhizobium</i> . O grafo ilustra a dinâmica dos dados em 7 particionamentos distintos, iniciando com 1 classe até 10 classes.	56
4.9	Dendrograma gerado pelo método UPGMA para a base de estirpes do gênero <i>Bradyrhizobium</i> . Para facilitar a visualização, cortou-se o dendrograma em 10 classes. Seus nós representam a porcentagem de elementos.	56

Lista de Tabelas

3.1	Descrição dos 16 atributos usados para descrever 100 animais na base <i>Zoo</i> (Forsyth, 1990).	36
3.2	Descrição das bases usadas na validação da rede ART2.	36
3.3	Parâmetros da rede ART2.	37
3.4	Exemplo de cálculo do escore em um dos testes realizados com rede ART2 para a base Iris com $\rho = 0.98$	38
3.5	Relação entre enzimas de restrição usadas e respectivas regiões ribossomais a que se referem.	40
4.1	Resultado dos algoritmos para cada base.	49
4.2	Resultado do algoritmo <i>Distance</i> para as bases Iris, <i>Glass</i> , <i>Wine</i> , <i>Zoo</i> e <i>Balance</i>	54
4.3	Distribuição das estirpes de <i>Bradyrhizobium</i> , cortando o grafo em 4 classes. As estirpes de referência das espécies <i>B. japonicum</i> e <i>B. elkanii</i> estão em destaque.	57
4.4	Relação de estirpes de referência das espécies <i>B. japonicum</i> e <i>B. elkanii</i> usadas neste trabalho conforme Germano-Silva (2003). . .	58

Lista de Abreviaturas e Siglas

ART	<i>Adaptive Resonance Theory</i>
DNA	<i>Deoxyribonucleic Acid</i>
Embrapa	Empresa Brasileira de Pesquisa Agropecuária
FBN	Fixação Biológica de Nitrogênio
MCD	Memória de Curta Duração
MDS	<i>Multidimensional Scaling</i>
MLD	Memória de Longa Duração
PCA	<i>Principal Component Analysis</i>
PCR	<i>Polymerase Chain Reaction</i>
RBF	<i>Radial Basis Function</i>
RFLP	<i>Restriction Fragment Length Polymorphism</i>
RNA	<i>Ribonucleic Acid</i>
RNAs	Redes Neurais Artificiais
rRNA	<i>Ribosomal Ribonucleic Acid</i>
SOM	<i>Self-Organizing Map</i>
UPMGA	<i>Unweighted Pair Group Method with Arithmetic mean</i>
XML	<i>eXtensible Markup Language</i>

Lista de Símbolos

a, b	pesos fixos da camada F1
α	taxa de aprendizado
b_{iJ}	pesos <i>botton-up</i>
c	peso fixo usado no teste de <i>reset</i>
d	ativação do nó vencedor J na camada F2
e	parâmetro para prevenir a divisão por zero
i	índice do vetor
I	dado
J	índice do nó vencedor na camada F2
m	número de classes criadas pela rede ART2
n	dimensão do vetor
p	comprimento da canaleta medido em número de pixels
q	altura da canaleta medida em número de pixels
ρ	parâmetro de vigilância
t	eletroferograma
t_{Ji}	pesos <i>top-down</i>
θ	parâmetro de supressão de ruído

Conteúdo

Agradecimentos	iii
Resumo	v
Abstract	vii
Lista de Figuras	vii
Lista de Tabelas	xiii
Lista de Abreviaturas e Siglas	xv
Lista de Símbolos	xvii
1 Introdução	1
1.1 Objetivos Gerais e Específicos	5
2 Teoria	7
2.1 Rizóbios: Bactérias Fixadoras de Nitrogênio	7
2.2 Biologia Molecular	8
2.3 Classificação de Dados na Computação	12
2.4 Redes Neurais Artificiais	18
2.5 Rede Neural ART2	20
2.5.1 Visão Funcional	21
2.5.2 Pré-processamento	24
2.5.3 Cálculo da Similaridade	24
2.5.4 Ciclo de Busca	26

2.5.5	Aprendizagem	31
3	Materiais e Métodos	33
3.1	Bases de Dados de Validação	33
3.2	Validação do Algoritmo da Rede ART2 Implementado	37
3.3	Base de Dados de Estirpes Tropicais do Gênero <i>Bradyrhizobium</i> .	39
3.4	Regiões Ribossomais Analisadas e Enzimas de Restrição	39
3.5	Geração das Matrizes de Distância DP	40
3.5.1	Cálculo da Matriz de Distância Final C	42
3.6	Cálculo da Matriz de Pontos com MDS	42
3.7	Cálculo do Número Ótimo de Classes com o Algoritmo <i>Distance</i> .	44
3.8	Representação Gráfica da Dinâmica dos Dados da Base de Estirpes de <i>Bradyrhizobium</i>	46
4	Resultados e Discussão	49
4.1	Resultados da Rede ART2 Comparados com Outros Classificadores	49
4.2	Resultados das Distâncias Inter e Intra-classe	50
4.3	Resultados para a Base de Estirpes do gênero <i>Bradyrhizobium</i> . .	54
5	Conclusão	59
5.1	Próximos Passos	61
A	Algoritmo Detalhado da Rede ART2	63

Capítulo 1

Introdução

Os microrganismos apresentam uma enorme diversidade genética e desempenham funções cruciais na manutenção da biosfera, uma vez que são componentes fundamentais de cadeias alimentares e ciclos biogeoquímicos (Schimel, 1995; Meyers, 1996). Apesar de sua importância, estima-se que menos de 10% dos microrganismos existentes tenham sido caracterizados e descritos (Canhos e Manfio, 2000).

A maioria dos estudos e esforços para uso sustentável da biodiversidade foca os macrorganismos. Estimativas indicam que microrganismos e invertebrados constituem quase 90% das espécies da Terra (Canhos e Manfio, 2000). Embora sejam menos estudados, são essenciais para a sobrevivência das formas de vida na Terra (Hammond, 1995).

Muitos avanços da biotecnologia moderna e agricultura são derivados de descobertas recentes nas áreas de genética, fisiologia e metabolismo de microrganismos. Sua diversidade genética e metabólica é explorada visando à obtenção de produtos biotecnológicos como: antibióticos, alimentos, processamento de alimentos, bebidas alcoólicas, ácidos orgânicos, álcoois, alimentos fermentados, tratamento de resíduos, e, na agricultura, na fertilização de solos e controle biológico de pragas e doenças (Canhos et al., 1998).

O conhecimento das funções dos diferentes componentes da biodiversidade de solos, de suas inter-relações e taxas de ocorrência é bastante limitada devido à escassez de informações taxonômicas disponíveis (Canhos et al., 1998). Estima-se que são descritas apenas 1% das bactérias de solos. Atualmente, são conhecidas

5.000 espécies de bactérias, sendo que a maioria não é de solos. Contudo, em 1 g de solo há aproximadamente 20 a 40 mil espécies de bactérias (Canhos e Manfio, 2000).

Através do emprego de técnicas moleculares, os microbiologistas estão apenas começando a desvendar a enorme diversidade microbiana dos solos (Canhos et al., 1998). Especificamente, estudos sobre a diversidade de organismos relacionados à fertilização biológica de solos estão em estágio mais avançado no Brasil (Embrapa, 2007). Isso decorre do anseio pela sustentabilidade na agricultura, com redução da dependência de insumos químicos. Os métodos de fertilização biológica são baseados no uso de inoculantes microbianos, principalmente bactérias fixadoras de nitrogênio.

A soja (*Glycine max* (L.) Merr.) tem grande importância econômica para o Brasil, sendo o segundo maior produtor com 25% da safra mundial (50 milhões de toneladas) (Embrapa, 2007). Sua cultura apresenta uma elevada demanda por nitrogênio devido aos altos teores de proteínas em seus grãos. Depois da água, o nitrogênio é o nutriente mais limitante para a produção agrícola.

Fertilizantes nitrogenados podem ser usados como fonte de nitrogênio. Contudo, têm baixo aproveitamento, cerca de 50%, custo elevado e risco de contaminação ambiental. Usá-los como única fonte de nitrogênio torna o plantio no Brasil inviável economicamente. Por outro lado, a Fixação Biológica de Nitrogênio (FBN) pode suprir totalmente essa demanda com um custo muito menor (Chueire et al., 2003). A FBN é um dos componentes críticos para a produção com altos rendimentos em cultivos de soja no Brasil.

A FBN é o sistema natural de transformação de N_2 atmosférico por bactérias que vivem no solo em formas assimiláveis pelas plantas (Milagre, 2003). Sua importância é comparável à fotossíntese, já que suprem os ecossistemas com novas quantidades de nitrogênio. Os rizóbios são bactérias capazes de fixar nitrogênio quando em simbiose com plantas leguminosas. Quando associados às raízes de soja, formam nódulos, onde ocorre a conversão do nitrogênio atmosférico em compostos nitrogenados, que serão utilizados pela planta.

Dentre os rizóbios, o gênero *Bradyrhizobium* é de grande importância na agricultura. Essas bactérias são capazes de fixar o nitrogênio atmosférico em simbiose com a soja. No Brasil, as espécies *B. japonicum* e *B. elkanii* têm sido amplamente utilizadas como inoculantes nessa leguminosa (Chueire et al., 2003).

A importância dessas bactérias é confirmada pelo crescente interesse da Embrapa. A Embrapa Soja é referência mundial no desenvolvimento de tecnologias para a cultura em regiões tropicais e possui várias linhas de pesquisa em Microbiologia do solo (Embrapa, 2007): seleção de estirpes de rizóbio mais eficientes e competitivas; caracterização fisiológica, bioquímica e genética de populações estabelecidas de rizóbio; obtenção de estirpes mutantes com maior tolerância ao *Al* (Alumínio) e *Mn* (Manganês) do solo; seleção de métodos de inoculação; seleção de genótipos de soja de alta eficiência de FBN. Todos esses estudos podem ser beneficiados com o aumento do conhecimento taxonômico dos rizóbios envolvidos na FBN da soja.

Há um grande número de fatores que afetam os níveis de rizóbios no solo e o desempenho da FBN (Embrapa, 2007): capacidade de fixar nitrogênio, tolerância à acidez do solo, à aplicação de pesticidas, ao tratamento de sementes e a secas prolongadas. Assim, é necessário um melhor conhecimento da classificação das estirpes brasileiras para obter alta eficiência na FBN.

A taxonomia das bactérias do gênero *Bradyrhizobium* é pouco refinada. Inicialmente, foram classificadas pela capacidade de nodular uma espécie específica de planta hospedeira (Jordan, 1938) *apud* (Milagre, 2003). Atualmente, há resultados conflitantes obtidos pelas análises fenóticas e genóticas (So et al., 1994; Rossuma e Schuurmans, 1995). A análise de genes ribossomais mostrou que várias espécies eram completamente diferentes, tendo mais parentescos com bactérias não fixadoras de nitrogênio (Garrity e Holt, 2001).

O uso de técnicas moleculares é uma abordagem nova e eficaz para a identificação de espécies bacterianas. A técnica de seqüenciamento de genes é muito eficiente, porém exige um investimento elevado (Chueire et al., 2003). Desse modo, relações genéticas entre as estirpes têm sido investigadas com outras téc-

nicas. A combinação das tecnologias de *Polymerase Chain Reaction* (PCR) (Saiki et al., 1988) com o método de *Restriction Fragment Length Polymorphism* (RFLP) (Laguerre et al., 1992) de regiões de genes cromossômicos conservadas permitem a identificação de uma ampla gama de bactérias (Milagre, 2003).

Dentre as regiões de genes cromossômicos, a região RNAr 16S é encontrada em todas as bactérias e tornou-se padrão universal para a sua identificação e classificação (Germano et al., 2006). Também é sugerido o uso das regiões com maior variabilidade como 23S e IGS (Milagre, 2003). A informação dessas regiões fornece conhecimento suficiente para identificar uma bactéria como membro de uma espécie ou grupo conhecido ou colocá-la em uma nova espécie.

O conhecimento da ecologia de estirpes de rizóbio presentes nos solos das regiões tropicais é ainda bastante restrito (Sá e Vargas, 1997). A caracterização genética desses microrganismos é de extrema importância para estudos taxonômicos e para conhecer melhor suas características fisiológicas, ecológicas e econômicas. Sendo assim, técnicas de biologia molecular e classificação poderão aumentar não só o conhecimento sobre a diversidade de rizóbios em solos tropicais, como também ajudar nos estudos para maximizar a eficiência da FBN (Hara et al., 2005).

O aumento no conhecimento da taxonomia de rizóbios nativos é essencial para propósitos de bioprospecção e biotecnologia. O conhecimento de biogeografia é importante para a definição de estratégias de busca e descoberta, ou seja, “onde procurar” por recursos biológicos potencialmente novos, e na definição de áreas de conservação de recursos biológicos e *pools* gênicos ricos em diversidade (Canhos e Manfio, 2000). Esse aumento do conhecimento taxonômico é impulsionado não somente por novas técnicas de biologia molecular, mas também pelo emprego de técnicas de computação, principalmente algoritmos de classificação e agrupamento.

Classificação pode ser definida como o particionamento de dados em classes onde elementos em uma classe são mais similares entre si do que elementos em outras classes (Halkidi et al., 2001). Técnicas de classificação estão ganhando importância nos últimos anos na análise exploratória de dados biológicos. Elas

são aplicadas principalmente na construção de taxonomias filogenéticas e avaliação da biodiversidade de grupos de organismos com base na análise de genes ribossomais (Talavera, 2007). Dentre as técnicas existentes, destacam-se as Redes Neurais Artificiais (RNAs) pois provêm uma abordagem robusta para problemas de classificação (Mitchell, 1997).

Adaptive Resonance Theory (ART) é uma teoria que descreve uma série de arquiteturas de RNAs para aprendizagem não-supervisionada propostas por Carpenter e Grossberg (Grossberg, 1976). As redes ART são usadas em classificação de dados. A rede original ART1 classifica dados binários com aprendizagem não-supervisionada (Carpenter e Grossberg, 2003). Já, a rede ART2 é sua extensão para dados analógicos (Carpenter e Grossberg, 1987b). Apesar da rede ART2 ser não supervisionada, é possível controlar o grau de similaridade entre dados colocados em uma mesma classe (Fausett, 1997). Assim, pode-se gerar classes com diferentes granularidades para uma mesma base. Esse aspecto é útil para estudar a taxonomia das estirpes em diferentes particionamentos.

1.1 Objetivos Gerais e Específicos

O objetivo deste trabalho é classificar uma coleção de estirpes brasileiras do gênero *Bradyrhizobium* utilizando uma rede neural ART2. São utilizadas 119 estirpes com características fenóticas de *B. elkanii* e *B. japonicum*, isoladas de 33 espécies de leguminosas tropicais (Embrapa, 2007). Os dados consistem de imagens de gel de eletroforese obtidas a partir da análise RFLP-PCR de 3 regiões ribossomais: 16S, 23S e IGS. Foram utilizadas as enzimas de restrição: Cfo I, Dde I, Msp I, Hae III, Hha I e Hinf I (Germano et al., 2006). Há também os seguintes objetivos específicos:

1. Criar uma representação gráfica que mostre a dinâmica dos dados da base de estirpes do gênero *Bradyrhizobium* nos diferentes números de classes geradas pela rede ART2;
2. Calcular automaticamente o número ótimo de classes baseado nas distâncias inter e intra-classe, algoritmo *Distance*;

3. Implementar um aplicativo para gerar automaticamente valores do parâmetro de vigilância, algoritmo *rhoGenerator*;
4. Validar o algoritmo implementado da rede ART2 e o algoritmo *Distance* em bases de dados cuja classificação é conhecida *a priori*;
5. Preocupação com procedimentos de projeto de *software* padronizados: (1) Criar documentação descrevendo o ambiente, funcionamento e aplicação dos algoritmos desenvolvidos; (2) Facilitar o compartilhamento de informações, criando arquivos padrões em formato XML para dados, parâmetros e resultados; (3) Conversão das bases de dados para um arquivo XML padrão.

Capítulo 2

Teoria

Taxonomia é a ciência de classificar organismos. A taxonomia existe em um estado dinâmico, onde os esquemas de classificação de microrganismos são modificados em função do acúmulo contínuo de novos dados. Inicialmente, eram usadas apenas características ecológicas, fisiológicas e morfológicas. Posteriormente, passou a respeitar também as relações evolutivas entre organismos (filogenia), principalmente baseadas na semelhança entre genomas.

A taxonomia molecular envolve o estudo do DNA (*Deoxyribonucleic Acid*) cromossômico e RNA (*Ribonucleic Acid*) ribossômico. Essas informações podem ser empregadas na classificação de bactérias em diversos níveis taxonômicos hierárquicos. Desde o estabelecimento de relações infra-específicas entre linhagens até relações entre espécies, gêneros e níveis taxonômicos supra genéricos (Williamson, 2006).

2.1 Rizóbios: Bactérias Fixadoras de Nitrogênio

Rizóbios são bactérias capazes de fixar nitrogênio atmosférico formando nódulos nas raízes das plantas. A Figura 2.1 mostra os nódulos nas raízes de uma leguminosa de soja. Atualmente, os rizóbios são distribuídos em 5 famílias e 6 gêneros (Garrity e Holt, 2001). Apenas os rizóbios de crescimento lento do gênero *Bradyrhizobium* como as espécies *Bradyrhizobium japonicum* e *Bradyrhizobium elkanii* nodulam soja. Essas espécies são divididas em estirpes, sendo que umas das características que as diferenciam é a quantidade de nitrogênio fixado. (Milagre, 2003).

Figura 2.1: Bactérias do gênero *Bradyrhizobium*, quando associadas às raízes de soja, formam nódulos, onde ocorre a conversão do nitrogênio da atmosfera em compostos nitrogenados, que serão utilizados pela planta (Embrapa, 2007).

Para classificar as diversas espécies de rizóbio, foram inicialmente empregados testes morfológicos, fisiológicos, bioquímicos e simbióticos como: taxa de crescimento, habilidade em utilizar fontes de carbono e de nodular leguminosas hospedeiras. Contudo, particularmente nas últimas duas décadas, as técnicas de biologia molecular ganharam mais espaço nos estudos taxonômicos dos rizóbios (Chueire et al., 2003).

2.2 Biologia Molecular

A Biologia Molecular é o estudo da biologia em nível molecular, com especial ênfase no estudo da estrutura e função do material genético (Williamson, 2006). Atualmente, novas técnicas de Biologia Molecular permitem uma caracterização mais precisa das estirpes: seqüenciamento (Butler, 2005), eletro-

forese (Westermeyer, 2005), hibridização de DNA (Glover, 2002), PCR (Saiki et al., 1988) e RFLP (Laguerre et al., 1992). Essas técnicas levaram ao rápido acúmulo de dados nas últimas décadas, aumentando o conhecimento sobre a taxonomia de bactérias. Isso é comprovado pelas consideráveis alterações nas posições taxonômicas no Manual de Sistemática Bacteriológica (Garrrity e Holt, 2001). Dentre essas técnicas de biologia molecular, destacam-se seqüenciamento, PCR e RFLP.

A análise de seqüenciamento é um processo que determina a ordem dos nucleotídeos em uma amostra de material genético (Butler, 2005). Nucleotídeos são os blocos que constituem os ácidos nucléicos. Há vários métodos de seqüenciamento disponíveis, mas todos com custo elevado (Chueire et al., 2003). Apesar da análise de seqüenciamento ser eficiente, existem outros métodos mais baratos para analisar genes ribossomais (Milagre, 2003).

O uso das técnicas PCR-RFLP associadas com genes ribossomais pode ser mais conveniente para estudos taxonômicos: apresentam baixo custo, boa concordância com resultados de seqüenciamento e alta reproduzibilidade (Germano et al., 2006). Assim, podem ser usadas para avaliar a diversidade e posição taxonômica de bactérias. Especificamente, podem ser úteis como passo inicial para investigar a diversidade dos rizóbios nos trópicos. Os poucos estudos sobre rizóbios, concentram-se em estirpes de regiões temperadas. Há poucos estudos sobre rizóbios de regiões tropicais, apesar das indicações de que essa região possui alta diversidade (Milagre, 2003). Deste modo, a análise dos padrões produzidos por RFLP-PCR pode ser crítica para a correta asserção de classes e espécies dessa região (Milagre, 2003).

PCR é a amplificação enzimática de uma seqüência específica de rRNA, visando à produção de cópias desta seqüência em um tubo de ensaio (Saiki et al., 1988). As análises genéticas de organismos sofreram um grande impulso com essa técnica. PCR possibilita a análise dos genes por meio de um método simples e rápido de amplificação de seqüências (Butler, 2005).

A obtenção de segmentos amplificados normalmente obedece às seguintes etapas (Butler, 2005): a) extração do rRNA que contém a região a ser amplificada; b) escolha do segmento a ser amplificado e obtenção de *primers* específicos para o reconhecimento desse segmento; c) amplificação que dará origem a várias cópias e d) leitura do produto amplificado após eletroforese e coloração.

RFLP é uma técnica de fragmentação do rRNA através do uso de enzimas de restrição (Laguerre et al., 1992). É bastante utilizada em análise de genomas (Bhat et al., 1995). Enzimas de restrição são proteínas que cortam o rRNA em fragmentos menores. Uma enzima de restrição reconhece uma sequência de nucleotídeos específica (sítio de restrição), cortando o rRNA em cada ponto no qual essa sequência ocorre (Lewin, 2001). O número de fragmentos produzidos é igual ao número de sítios de restrição reconhecidos pela enzima utilizada. Quanto mais pares de base no sítio de restrição, menor é a probabilidade de se encontrar um local para cortar. As diferenças na sequência de rRNA dos indivíduos resulta em fragmentos de tamanhos distintos.

Eletroforese em gel é uma técnica separação de fragmentos moleculares em um determinado gel de acordo com seu tamanho (Westermeier, 2005). Fragmentos de moléculas de DNA ou RNA são colocados em pequenos poços de amostras. Os fragmentos migram com a aplicação de um campo elétrico formando canaletas, ver Figura 2.2. Esse movimento produz uma série de bandas. Cada banda corresponde a um conjunto de fragmentos de mesmo tamanho. Como os fragmentos com menor massa migram mais rapidamente, as bandas são separadas em ordem decrescente ao longo do gel (Lewin, 2001).

As diferenças entre indivíduos são notadas quando se visualiza diferentes tamanhos de fragmentos. Se os indivíduos diferem entre si em relação a posição dos sítios de restrição, são gerados fragmentos de tamanhos distintos. Então, as bandas serão observadas em posições distintas na eletroforese.

Figura 2.2: Exemplo de imagem produzida pela técnica de gel de eletroforese. Pode-se observar 7 canaletas com bandas em posições distintas.

O ribossomo é uma estrutura celular responsável pela síntese de proteínas. Os ribossomos bacterianos se dissociam em duas sub-unidades. A unidade maior contém o rRNA 23S e a menor, o rRNA 16S. A região localizada entre as regiões 16S e 23S é denominada região intergênica (IGS) (Lewin, 2001). Essas regiões podem revelar relações taxonômicas entre os diversos gêneros e espécies de rizóbios. O RNA ribossomal apresenta características bastante conservadas por ser crítico para a função celular. As análises dessas regiões possuem um aspecto altamente informativo e seus resultados são facilmente reproduzíveis (Thórsson e Jónsson, 2000).

A região rRNA 16S é importante para a taxonomia molecular de bactérias (Germano et al., 2006). Sequências parciais ou completas dessa região são usadas para estudar a taxonomia dos rizóbios (Chen et al., 2000; Vinuesa et al., 1998; Urtz e Elkan, 1996; Oyaizu e Gamou, 1992; Young et al., 1991). Contudo, há estudos mostrando que apesar do alto nível de diversidade morfológica, fisiológica e genética, a diversidade da região 16S rRNA é baixa nas estirpes de *Bradyrhizo-*

bium (Qian et al., 2003; Chen et al., 2000; Vinuesa et al., 1998). Assim, deve-se estudar também regiões com maior variabilidade.

A região 23S rRNA pode auxiliar para delinear relações próximas entre bactérias. Ela apresenta maior variabilidade e contém mais informações que a região 16S rRNA (Olsen e Woese, 1993). Essa região é útil para a especificação de muitos gêneros de bactérias (Ludwig e Schleifer, 1994), incluindo rizóbios (Germano et al., 2006). Também, análises de seqüências do espaço intergênico IGS são empregadas na especificação de bactérias (Germano et al., 2006).

Dado que a taxonomia dos rizóbios do gênero *Bradyrhizobium* é pouco refinada e existem resultados conflitantes das análises fenóticas e genóticas (Rossuma e Schuurmans, 1995). Deve-se usar também dados das análises das regiões 23S e IGS, que apresentam maior nível de divergência entre as espécies de *Bradyrhizobium* (Milagre, 2003).

2.3 Classificação de Dados na Computação

Classificação de Dados é um processo que se inicia pela aquisição de dados e termina com a interpretação dos resultados:

1. Aquisição de dados, os quais podem estar dispersos em vários sistemas, além de conter redundâncias e inconsistências;
2. Limpeza dos dados com remoção de ruídos, dados espúrios e erros;
3. Transformação dos dados em uma entrada compatível com o classificador utilizado;
4. Seleção das variáveis com base na quantidade de informação e nos objetivos do processo de classificação;
5. Aplicação do classificador na base formatada;
6. Interpretação dos resultados.

A Figura 2.3 corresponde a um ensaio gerado de acordo com Germano et al. (2006), incluindo as etapas de aquisição e limpeza dos dados. Um ensaio contém vários testes com uma enzima de restrição em uma região específica do rRNA para um conjunto de estirpes. Esse ensaio refere-se à região IGS rRNA digerida pela enzima de restrição Msp I das estirpes 51 a 75.

Figura 2.3: Exemplo de ensaio. Gel de eletroforese da região IGS rRNA de 24 estirpes digeridas com a enzima de restrição Msp I.

É importante entender a definição de estirpe adotada neste trabalho. Na aquisição dos dados 3 regiões ribossomais foram analisadas: 16S, 23S e IGS (Embrapa, 2007). Para cada região foram usadas 3 enzimas de restrição descritas em (Germano et al., 2006). Assim, cada estirpe é representada por 9 canaletas. A Figura 2.4 ilustra como os ensaios são usados para descrever as estirpes.

A informação na canaleta encontra-se na distribuição das bandas, regiões onde a cor é mais intensa devido à concentração de material genético. Um eletroferograma é usado para representar essa informação em cada canaleta (Talavera, 2007). Seus picos representam as bandas. Assim, o eletroferograma é um gráfico da distribuição de material genético em relação à posição na raia conforme mostrado na Figura 2.5.

Figura 2.4: A estirpe 1 é descrita por um conjunto de 9 canaletas recortadas dos ensaios de gel de eletroforese. Cada canaleta corresponde a uma região do rRNA digerida por uma enzima de restrição. Para que as estirpes possam ser classificadas pela rede ART2, estas devem ser transformadas em um vetor.

Figura 2.5: A informação das canaletas está na posição de suas bandas. Bandas são regiões escuras que indicam a presença de material genético. Neste trabalho, cada canaleta é transformada em um eletroferograma, cujos picos representam as bandas.

Os dados devem ser transformados antes da aplicação no classificador. A rede ART2 é um classificador de vetores. Assim, as informações usadas para descrever uma estirpe devem ser comprimidas em um vetor. Os passos necessários são: transformação dos dados em uma única matriz de distância e seleção das variáveis mais importantes.

As imagens de gel de eletroforese usadas para descrever as estirpes são redundantes. Como cada pixel é muito correlacionado com pixels vizinhos, os vetores que representam as bandas são altamente correlacionados. Além disso a dimensionalidade desses vetores é alta. As bandas foram cortadas com comprimento de 340 pixels, gerando vetores de 340 variáveis. Quanto maior a dimensionalidade dos vetores, mais parecidos eles se tornam. Assim, deve-se reduzir a redundância e a dimensionalidade dos dados, utilizando apenas a informação discriminatória, descartando aquela que é redundante.

Um dado I é um vetor de dimensionalidade n , onde n é o número de variáveis que descrevem esse dado. A seleção de variáveis desempenha uma tarefa essencial em classificação. Seu objetivo é selecionar as variáveis mais importantes, ou seja, mais relevantes e menos redundantes.

É importante que não haja redundância entre as variáveis, isto é, que sua covariância seja baixa. Quando essas variáveis contiverem somente as características mais salientes, o classificador terá melhor desempenho, será mais rápido e ocupará menos memória (Jain et al., 2000). A maioria dos classificadores, inclusive redes neurais, tem seu desempenho afetado pela dimensionalidade dos dados (Jain et al., 2000). Além disso, quanto maior for o número de variáveis, os vetores tornam-se mais similares. Isso prejudica a capacidade do classificador distinguir os dados.

Para selecionar apenas as variáveis mais importantes, é preciso determinar sua relevância segundo algum critério. Há técnicas como PCA (*Principal Component Analysis*) (Jain et al., 2000), redes neurais multicamada (Lowe e Webb, 1991), SOM (*Self-Organizing Map*) (Kohonen, 1982) e MDS (*Multidimensional Scaling*) (Kruskal e Wish, 1981). Neste trabalho foi usada o método MDS. Sua escolha é justificada por sua capacidade de encontrar uma configuração de pontos em um espaço de baixa dimensão que melhor representa um conjunto de dados multivariados. Também, pela disponibilidade do algoritmo de MDS DISTATIS (Abdi et al., 2005), o qual é estudado pelo grupo de pesquisa.

A correlação entre dois vetores pode ser usada como medida de similaridade. Esta reflete o grau no qual eles estão relacionados. Uma medida comum é a Correlação de Pearson (Hon, 2002). Seu coeficiente é definido como:

$$\phi_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (2.1)$$

no qual $\bar{x}$ e $\bar{y}$ são a média amostral dos vetores x e y respectivamente. Esse coeficiente pode ser transformado em uma medida de distância, Distância de Pearson, conforme (2.2).

$$d_p = 1 - \phi_{xy} \quad (2.2)$$

Uma matriz de distância é uma matriz triangular de valores que expressam a distância dado-dado de um conjunto de dados. A matriz de distância, dada em (2.3), define a distância de Pearson, calculada por (2.2) entre um conjunto de dados. Onde i e j são os índices dos dados no conjunto.

$$DP_{k_{ij}} = [d_p(p_i, p_j)] = \begin{vmatrix} 0 & & & & \\ d_{21} & 0 & & & \\ d_{31} & d_{32} & 0 & & \\ \dots & \dots & \dots & 0 & \\ d_{m1} & d_{m2} & \dots & \dots & 0 \end{vmatrix} \quad (2.3)$$

MDS é um método de ordenação multivariado que fornece uma representação espacial a partir da distância entre os dados (Kruskal e Wish, 1981). Dada uma matriz de distância, MDS gera um espaço de pontos que pode ser usado para reduzir a dimensionalidade dos dados. Nesse espaço, cada ponto representa um dado e as coordenadas são as variáveis.

Este trabalho empregou o algoritmo DISTATIS (Abdi et al., 2005). Ele transforma um conjunto de m dados com uma dimensionalidade d em um sub-espaço de dimensionalidade d' , com $d \leq d'$. Esse sub-espaço, apesar da menor dimensionalidade, preserva o máximo de informação do espaço original. Isso é possível pois MDS descarta a informação comum dos dados, preservando somente a informação discriminante.

Primeiramente, converte-se um conjunto de m dados de dimensionalidade d em uma matriz de distância dado-dado D_{mm} (Llerena et al., 2007). Então, DISTATIS transforma a matriz D_{mm} em um espaço de pontos. Esse espaço de pontos pode ser visto também como uma matriz dado-variável D_{mn} onde cada linha corresponde a um dado e as colunas às variáveis. Assim, m é o número de dados e n é número de variáveis que descrevem um dado. As variáveis da matriz P_{mn} estão ordenadas em ordem decrescente de seus auto-valores. Cada auto-valor indica a quantidade de informação que sua variável representa.

Escolhendo-se as d' primeiras variáveis da matriz P_{mn} gera-se uma nova matriz $P_{md'}$. A nova matriz $P_{md'}$ é uma aproximação de P_{mn} . Através dos autovalores retornados por DISTATIS é possível calcular a porcentagem de informação de $P_{md'}$ em relação à P_{mn} . A Figura 2.6 ilustra a aplicação de MDS na base de dados Iris (Fisher, 1936), na qual se reduziu a dimensionalidade dos dados de 4 para 3 variáveis.

Figura 2.6: Representação dos pontos gerados com o algoritmo DISTATIS para a base de dados Iris, a qual é descrita por 4 variáveis. MDS selecionou apenas as 3 variáveis mais relevantes. O espaço gerado $P_{150 \times 3}$ é uma aproximação do espaço original $P_{150 \times 4}$.

2.4 Redes Neurais Artificiais

RNAs são modelos computacionais que tentam imitar características de grandes redes biológicas altamente conectadas de elementos simples (Everitt et al., 2001). Esses elementos são aproximações de neurônios biológicos. Esses modelos usam alguns princípios organizacionais, tais como aprendizado, generalização, adaptabilidade, tolerância à falha e computação distribuída (Jain et al., 2000). A principal diferença entre RNAs e outras abordagens de classificação é sua habilidade de aprender relacionamentos complexos não lineares entre dados de entrada e saída.

As RNAs mais usadas para classificação são as redes de alimentação para frente (*feedforward*) como Perceptron Multicamada e Redes RBF (Fausett, 1997). Essas redes são organizadas em camadas, com conexões unidirecionais entre as

mesmas. Destaca-se também a rede SOM (Kohonen, 1982). Nesta, os dados são distribuídos em um mapa que considera a topologia das classes. Além disso, redes *Backpropagation* mostraram sucesso com problemas de áreas distintas do conhecimento: classificação de caracteres (LeCun et al., 1989), faces (Cottrell, 1990) e palavras faladas (Lang et al., 1990).

A popularidade do uso de RNAs para problemas de classificação é devido a sua baixa dependência de conhecimento de domínio específico (Mitchell, 1997). RNAs provêm uma abordagem robusta para problemas como aprender a interpretar dados reais complexos. Elas estão entre os métodos de aprendizagem mais eficientes conhecidos atualmente (Mitchell, 1997).

A aprendizagem envolve atualizações na arquitetura da rede e nos pesos das conexões objetivando executar eficientemente uma tarefa específica. Tipicamente, reúne-se um conjunto de dados para treinar a rede. Durante o treinamento, a informação é codificada pelo ajuste dos pesos (Valluru, 1995). O treinamento termina quando a resposta produzida pela rede é considerada apropriada para uma determinada aplicação. Então, os pesos não são mais atualizados.

Uma característica interessante da memória humana é sua habilidade para aprender novas coisas sem esquecer algo aprendido anteriormente. Por exemplo, há a habilidade de reconhecer rostos não vistos por algum tempo depois de ter aprendido novos rostos. É desejável importar essa capacidade para RNAs. A maioria das redes neurais tendem a apagar informações velhas quando se adiciona novas informações incrementalmente (Freeman e Skapura, 1991).

Espera-se que as RNAs classifiquem apropriadamente qualquer dado apresentado no futuro (Freeman e Skapura, 1991). Isso é aceitável dado um problema estável e com fronteiras bem definidas. Nessas condições, é possível definir um conjunto adequado de dados de treinamento para o problema. Infelizmente, em aplicações reais, o ambiente é instável e sem fronteiras (Freeman e Skapura, 1991). Assim, mudanças ao longo do tempo, no domínio do problema, reduzem gradualmente o desempenho da rede. A capacidade de um sistema de se adaptar a novos dados indefinidamente é chamada plasticidade (Carpenter e Grossberg, 1988).

Uma solução possível para manter a plasticidade é retreinar a rede somente com os novos dados ou juntamente com os dados iniciais (Braga et al., 2000). Essa abordagem tem limitações de custo computacional e indisponibilidade dos dados iniciais. Também, a rede pode apagar as informações aprendidas anteriormente causando instabilidade. Contudo, o aprendizado deve ser plástico e estável. Um algoritmo de aprendizado é estável quando permite que o conhecimento seja atualizado, sem perder aquele previamente armazenado (Senger, 2004). A maioria dos algoritmos de classificação são estáveis, mas incapazes de formar novas classes, ou incrementais, mas instáveis (Grossberg, 1976). Esse conflito é conhecido como dilema da plasticidade-estabilidade. Uma arquitetura de RNAs que resolve tal dilema são as redes ART.

A ART é uma série de arquiteturas para aprendizagem não-supervisionada baseada em neurobiologia e comportamento humano (Grossberg, 1976). Seu objetivo é modelar uma arquitetura maciçamente paralela em uma rede do tipo auto-organizável. Redes auto-organizáveis são sistemas de aprendizado de máquina não-supervisionado, de forma a auto-organizar dados em classes, baseadas em alguma função de similaridade (Fausett, 1997).

As redes ART permitem que um exemplo de treinamento modifique uma classe somente se esta estiver suficientemente próxima do exemplo de treinamento, caso contrário uma nova classe é criada. Desta forma, resolvem o dilema da estabilidade-plasticidade, ou seja, como fazer um sistema se adaptar, evitando-se ao mesmo tempo que as entradas atuais destruam o treinamento passado.

2.5 Rede Neural ART2

A arquitetura da rede ART2, ver Figura 2.7, é similar àquela empregada em muitos sistemas auto-organizáveis. Há duas camadas de neurônios, camada de entrada (F1) e de saída (F2). Os neurônios são representados por nós.

Na camada F1, os nós W , X , V , U , Q e P representam o dado I em diferentes estágios do processamento da rede. Esses nós são vetores com a mesma dimensionalidade de I . O processamento dos dados ocorre com um nó para cada elemento

do vetor de entrada. As etapas são normalização, supressão de ruído e aumento de contraste. A camada F2 é competitiva onde cada nó Y corresponde a uma classe. Seu tamanho é decidido dinamicamente, assim que novas classes são necessárias. Finalmente, os nós R representam o vetor usado no teste de *reset*.

Figura 2.7: Arquitetura da rede ART2 (modificado de Filho et al. (1997)). A camada F1 processa os dados na entrada, a camada F2 representa as classes e R é usado no teste de *reset*.

2.5.1 Visão Funcional

A rede ART2 é um classificador de vetores. As etapas para um teste de aprendizagem são apresentadas na Figura 2.8. O teste de aprendizagem compreende o processamento de um dado. Seu funcionamento é apresentado de forma resumida a seguir:

1. Um dado I é apresentado à rede. Após normalização, supressão de ruído e aumento de contraste, as classes da camada F2 competem para classificá-lo. Aquela com maior ativação é definida como candidata;
2. Se a similaridade da classe candidata e o vetor é maior que um parâmetro definido pelo usuário (ρ), a aprendizagem é permitida. Os pesos da classe são atualizados para melhor refletir I ;

3. Se a classe candidata e I não são similares o suficiente, a classe candidata é inibida. As classes restantes competem novamente e a próxima melhor candidata é selecionada. Isso é repetido até que uma classe aceitável seja escolhida. Então, a aprendizagem é permitida;
4. Se nenhuma das classes é escolhida, cria-se uma nova classe em F2 para aprender I . Finalmente, todas as ativações são novamente habilitadas e um novo dado é apresentado.

O Algoritmo da Rede ART2 implementado neste trabalho, Algoritmo (2) está disponível de forma detalhada no Apêndice A.

Figura 2.8: O processamento na rede ART2 pode ser resumido nas seguintes etapas: Pré-processamento, Cálculo da similaridade, Ciclo de busca e Aprendizagem.

2.5.2 Pré-processamento

Os dados são inicialmente pré-processados através de uma normalização e da supressão de valores abaixo de um determinado limiar. Isso tem como efeito escalar os dados e eliminar pequenos ruídos.

O pré-processamento se inicia na camada F1 à partir do nó U que envia sinais para P e W . Cada nó U envia um sinal para W multiplicado pelo peso a . W soma os sinais provenientes de I e U . Os sinais que saem de W são normalizados antes de entrarem em X . A normalização ocorre quando um dado passa entre os seguintes nós: de W para X ; de V para U e de P para Q .

A eliminação de ruídos ajuda a rede na formação de classes estáveis (Filho et al., 1997). Assim, os sinais que saem de X passam por um filtro $f(X_i)$, dado por:

$$f(X_i) = \begin{cases} X_i & \text{se } X_i \leq \theta \\ 0 & \text{se } X_i > \theta \end{cases}$$

Simultaneamente, os sinais normalizados de Q recebem o mesmo tratamento:

$$f(Q_i) = \begin{cases} Q_i & \text{se } Q_i \leq \theta \\ 0 & \text{se } Q_i > \theta \end{cases}$$

Esses sinais que chegam a V são somados $V = bf(Q) + f(Q)$ e normalizados antes de serem enviados a U . Finalmente, são então enviados a P e R . Em P , são adicionados aos sinais dos nós ativos em F2. Em R , são usados no teste de *reset*.

2.5.3 Cálculo da Similaridade

É possível gerar classes com diferentes granularidades para uma mesma base com a rede ART2. Seu algoritmo permite que um dado modifique uma classe somente se a classe estiver suficientemente próxima do dado, caso contrário uma nova classe é criada. A similaridade entre a classe e um novo dado é controlada por um parâmetro de vigilância ρ especificado pelo usuário.

O parâmetro de vigilância é o principal parâmetro na rede ART2 pois determina o número final de classes. A similaridade, medida pela norma L_2 do vetor

R , é comparada ao parâmetro de vigilância definido pelo usuário. Se a similaridade excede o parâmetro de vigilância, a classe aprende a entrada. Como R é normalizado, ρ varia de 0 a 1. Quando $\rho = 1$, a entrada deve ser exatamente similar a classe. Assim, aumentando seu valor, a rede forma mais classes.

A similaridade entre um dado e o vetor de pesos *top-down* da classe candidata é baseada na direção dos vetores. Dois vetores são iguais se apontam para a mesma direção. Em vez de utilizar diretamente o ângulo entre os vetores, uma comparação alternativa é usada, (2.4). Ela contribui para se obter plasticidade com estabilidade (Carpenter e Grossberg, 1987b). Geometricamente, R corresponde a uma soma escalar com contribuições do vetor de entrada processado U e de P .

$$R_i = \frac{U_i + cP_i}{e + ||U|| + c||P||} \quad (2.4)$$

onde $||\bullet||$ denota a norma L_2 , U é a versão processada do dado atual e o vetor P representa a influência da Memória de Longa Duração (MLD).

Figura 2.9: Exemplo de classificação com a rede ART2 em 3 dados tridimensionais A , B e C . A rede ART2 criou 2 classes, representadas espacialmente pelo cone 1 e 2, para classificar esses dados. Os dados A e B estão dentro da classe 1 e o dado C pertence a classe 2 (modificado de Whiteley et al. (1996)).

A Figura 2.9 ilustra geometricamente a aplicação do cálculo da similaridade num espaço tridimensional. Nesse espaço, cada classe é um cone com seu centro definido pelo vetor de pesos *top-down*. O parâmetro de vigilância é inversamente proporcional ao raio do cone. Dado um valor de ρ , a rede ART2 distribui o menor número possível de cones no espaço para que todos dados fiquem dentro de alguma classe. Diminuindo-se ρ , aumenta-se o raio do cone e são necessários menos cones para classificar os mesmos dados.

2.5.4 Ciclo de Busca

O ciclo de busca é a etapa na qual a rede seleciona um nó na camada F2 para classificar uma entrada. Esses nós podem estar em 3 estados: ativo, inativo ou inibido. Um nó está ativo quando se torna um candidato a aprender uma entrada. Todos os nós que perdem essa competição ficam inativos. Se o nó candidato não combina satisfatoriamente com a entrada, ele torna-se inibido (Senger, 2004). Um nó inibido não pode competir novamente para classificar o dado atual.

Figura 2.10: Visão dos subsistemas de atenção e orientação os quais controlam a dinâmica da informação na rede ART2 (modificado de Heins e Tauritz (1995)).

A rede ART2 é auto-organizável, plástica e estável em resposta a dados arbitrários (Heins e Tauritz, 1995). Isso é possível pela dinâmica com a qual a rede processa um conjunto de dados. Essa dinâmica é controlada por dois subsistemas: atenção e orientação, representados por linhas pontilhadas na Figura 2.10.

Subsistema de Atenção

O subsistema de atenção tem a função de propor um nó para classificar uma dada entrada. É formado pela camada F1, F2 e mecanismo de controle de ganho. As unidades das camadas de entrada e saída são ligadas através de dois conjuntos de conexões direcionadas com pesos. Cada unidade na camada inferior envia seu sinal de ativação através dos pesos *bottom-up*. As unidades superiores respondem ao sinal através dos pesos *top-down*. Os pesos *bottom-up* e *top-down* são adaptativos e são chamados de Memória de Longa Duração (MLD) pois armazenam o conhecimento da rede. A estabilização da aprendizagem ocorre pela comparação entre a entrada modificada pelos pesos *bottom-up* e a expectativa da memória do nó ativo em F2 através dos pesos *top-down*.

Subsistema de Orientação

O subsistema de orientação ou mecanismo de *reset* funciona como um detector de novidade, controlando o grau de similaridade entre dados de uma mesma classe. Sua função é decidir se a rede deve classificar o dado na classe selecionada pelo sistema de atenção.

O mecanismo de controle de ganho controla o estado dos nós na camada F1. Um nó pode estar, "ligado" ou "desligado", nos diferentes estágios do processamento (Fausett, 1997).

Mecanismo de controle de ganho e regra $\frac{2}{3}$

As unidades de controle de ganho ajudam o mecanismo de *reset* a controlar a dinâmica da aprendizagem. Um nó está supraliminarmente ativado, suficientemente ativo para gerar uma saída, se recebe um sinal de pelo menos 2 de 3 possíveis fontes. Isso é chamado de regra $\frac{2}{3}$. Se um nó recebe um sinal de somente uma

fonte ele está subliminarmente ativado (Freeman e Skapura, 1991). O ciclo de busca, durante um teste de aprendizagem, é mostrado nos 4 passos abaixo :

1. Primeiramente, um dado I é apresentado aos nós em F1 com uma variável do vetor para cada nó. Isso gera a a Memória de Curta Duração (MCD) em X , ativa o controle de ganho G e o subsistema de orientação A . Note que G gera um sinal excitatório para cada nó na camada F1. Então, o aparecimento de X em F1 resulta em um sinal inibitório enviado à A . A rede é estruturada de forma que esse sinal inibitório cancela o efeito excitatório do sinal de I . Assim A continua inativo. O sinal inibitório também gera S que é transmitido para a camada F2 através dos pesos *bottom-up*. Finalmente, todos os nós F2 computam os valores em suas entradas criando Y . Em F2 só o nó com a maior entrada total é ativado. Todos os outros são desativados (fenômeno vencedor leva tudo).

Figura 2.11: Passo 1 do ciclo de busca (modificado de (Heins e Tauritz, 1995)).

2. O vetor Y em F2 gera o vetor descendente U . Este é transformado na expectativa V através dos pesos *top-down*. Também, Y inibe o controle de ganho de F1. O controle de ganho é configurado para cessar sua atividade quando receber algum sinal inibitório de F2. Há 3 possíveis fontes de sinal para F1, mas apenas duas são usadas ao mesmo tempo. Pela regra $\frac{2}{3}$, somente os nós F1 recebendo sinais de I e V continuam ativos. O vetor X^* que aparece em F1 é resultado da intersecção de I e V . X^* é muito

importante e representa a combinação da expectativa *top-down* da rede com o dado atual em sua entrada I . Finalmente, se V não combina com I , isso resulta em um decréscimo na inibição total de F1 para A e em uma nova saída S^* .

Figura 2.12: Passo 2 do ciclo de busca (modificado de (Heins e Tauritz, 1995)).

3. Se a diferença entre V e I é grande o suficiente, X^* será pequeno. Então, A não pode mais ser prevenido de gerar um sinal de *reset* para todos os nós em F2. Se estes estão ativos, tornam-se inibidos. O parâmetro de vigilância determina quanta diferença será tolerada. Então, já que Y não existe mais, a saída *top-down* e o sinal inibitório para o controle de ganho também desaparecem (Freeman e Skapura, 1991).

Figura 2.13: Passo 3 do ciclo de busca (modificado de (Heins e Tauritz, 1995)).

4. Depois que o nó F2 é inibido, sua expectativa *top-down* é eliminada e X pode ser reinstalado em F1 (Heins e Tauritz, 1995). Então, um novo ciclo inicia e X é novamente ativado. Finalmente, os nós participando no padrão original continuam inativos até que o controle de ganho de F2 é desligado pela remoção da entrada.

Figura 2.14: Passo 4 do ciclo de busca (modificado de (Heins e Tauritz, 1995)).

Esse ciclo de busca repete até:

1. Um nó F2 candidato é aceito, ou seja, sua expectativa *top-down* combina satisfatoriamente com a entrada;
2. Todos os nós F2 são inibidos. Um novo nó é criado para classificar I ;
3. Todos os nós F2 tornam-se inibidos e atingiu-se o limite de classes. A rede não pode classificar a entrada.

A MCD existe somente durante um ciclo de busca. A MLD é atualizada lentamente depois da apresentação dos dados várias vezes. Durante o ciclo de busca, não ocorre aprendizagem.

Após um ciclo de busca quando uma classe aceitável é selecionada, o subsistema de orientação não gerará um sinal de *reset* para a camada F2. O padrão de atividade em F1 leva a escolha do mesma classe em F2. Este gera a mesma expectativa em F1 novamente. Quando ocorre essa repetição a rede está em ressonância.

2.5.5 Aprendizagem

Aprendizagem é a processo de atualização do vetor de pesos da classe vencedora para melhor refletir o dado classificado e ocorre durante a ressonância. Nessa etapa, os pesos são atualizados de acordo com equações diferenciais dadas em (2.5) (Fausett, 1997). A aprendizagem está completa quando os pesos atingem seus valores assintóticos (Whiteley et al., 1996). Quando isso ocorre, a rede é dita estabilizada e o treinamento termina. A estabilidade significa que os dados não mudam mais de classe durante o treinamento.

$$\begin{aligned} t_{Ji} &= \alpha d U_i + 1 + \alpha d(d+1)t_{Ji} \\ b_{iJ} &= \alpha d U_i + 1 + \alpha d(d+1)b_{iJ} \end{aligned} \quad (2.5)$$

A aprendizagem em redes neurais é geralmente baseada em erro, como por exemplo o método de treinamento *Backpropagation*, aplicado em redes *Multi Layer Perceptron* (Mitchell, 1997). O objetivo é minimizar uma função de erro pela correção dos pesos da rede. Na rede ART2 a aprendizagem é incremental. Aprendizagem incremental significa que se pode alterar o conhecimento anterior quando o ambiente muda (Kondadadi e Kozma, 2002). Contudo, a atualização ocorre somente quando novos dados são suficientemente similares aos já classificados. Caso contrário, novas classes são criadas. Esse modelo de aprendizagem gera a estabilidade da rede. A memória só muda quando um dado é similar o suficiente com os dados classificados ou quando algo completamente novo ocorre. Isso torna a rede ART2 apropriada para problemas dinâmicos (Carpenter e Grossberg, 2003). Nesses problemas, o conhecimento está constantemente evoluindo assim como nos dados de estirpes de bactérias empregados neste trabalho.

Há dois modos de aprendizagem: rápida e lenta. A diferença básica entre esses métodos é a taxa de aprendizagem que ocorre a cada teste de aprendizagem (Senger, 2004).

Aprendizagem rápida significa que os pesos podem atingir equilíbrio durante um teste de aprendizagem (Pedini e Gaudiano, 1994). Os pesos são atualizados muitas vezes durante a ressonância. No modo lento, pesos não atingem o equilíbrio durante um teste de aprendizagem particular pois são modificados apenas uma vez. Os dados devem ser apresentados muitas vezes até os pesos atingirem seus valores assintóticos.

Aprendizagem Rápida

1. Vetores de pesos: são atualizados muitas vezes durante a ressonância. Poucos ciclos (testes de aprendizagem) são necessários para a rede estabilizar. Também, os pesos finais assemelham-se o último vetor classificado na classe;
2. Dados: mais adequada para dados binários com ou sem ruído. A rede é mais sensível à ordem de apresentação dos vetores (Jianguo e Jingdong, 2002);
3. Estabilidade: ocorre quando os dados são consecutivamente classificados nas mesmas classes.

Aprendizagem Lenta

1. Vetores de pesos: são atualizados 1 vez durante a ressonância. Um número maior de ciclos é necessário para a rede estabilizar;
2. Dados: mais adequada para dados contínuos e menos sensível a ordem de apresentação;
3. Estabilidade: ocorre quando os pesos atingem seus valores assintóticos.

Capítulo 3

Materiais e Métodos

Todos os programas desenvolvidos neste trabalho foram escritos em C++ com o auxílio de *shell script* para automatizar os testes. As bases de dados, parâmetros e resultados foram salvos em arquivos XML. Arquivos XML permitem a separação do conteúdo da formatação em arquivos. Isso aumenta a simplicidade e legibilidade, tanto para humanos quanto para computadores. O ambiente de programação usado foi o Eclipse.

Eclipse é um IDE (*Integrated Development Environment*) focado no desenvolvimento de ferramentas e aplicações de *software*. Este IDE foi escolhido pois reúne um conjunto de ferramentas de apoio ao desenvolvimento com o objetivo de agilizar este processo. Especificamente, utilizou-se o *plugin Eclipse C/C++ Development Tooling* que permite programação em C/C++. O compilador usado foi o gcc 3.4.6. O computador é um AMD64 3000+ com 1Gb de RAM rodando Linux Gentoo 3.4.6-r2.

3.1 Bases de Dados de Validação

Neste trabalho foi desenvolvida uma rede ART2 para classificar estirpes de *Bradyrhizobium*. O algoritmo ART2, disponível na literatura, foi implementado usando o paradigma de orientação a objetos, através da linguagem C++.

O objetivo da validação é comparar o desempenho da rede ART2 com classificadores que usam diferentes medidas de similaridade em 3 bases de áreas distintas do conhecimento. As bases são conhecidas na literatura e largamente utilizadas: Iris (Fisher, 1936), Wine (Forina, 1991) e *Balance* (Siegler, 1976). Os classificadores são: *Twostep*, *K-means* (Kanungo et al., 2002), *Positive Approach* (Spiegler,

2006), *Ward Methods* (Ward, 1963), Kohonen (Kohonen, 1982), *Furthest Neighbor* (Mitchell, 1997), *Within-groups Linkage* (Spiegler, 2006), *Centroid Clustering* (Sokal e Michener, 1958), *Between-groups Linkage* (Sokal e Michener, 1958), *Median Clustering* (Spiegler, 2006) e *Nearest Neighbor* (Mitchell, 1997).

As bases de dados empregadas na validação estão disponíveis em Newman et al. (2007). Elas são mantidas pela Universidade da Califórnia em Irvine (UCI). Esse repositório contém mais de cem bases usadas especificamente em aprendizagem de máquina.

Iris

Iris é provavelmente a base de dados mais usada em classificação, sendo regularmente citada (Duda e Hart, 1973; Spiegler, 2006). Ela foi produzida por (Fisher, 1936) e inclui 3 espécies de plantas. Cada espécie refere-se a um tipo de flor: iris-setosa, iris-versicolor e iris-virginica, ver Figura 3.1.

Figura 3.1: Base de dados Iris composta por três espécies de flores: (1) iris-setosa, (2) iris-versicolor e (3) iris-virginica.

Wine

Os dados da base *Wine* foram coletados na Itália pelo *Institute of Pharmaceutical and Food Analysis and Technologies* (Forina, 1991). Esses dados são resultados de análises químicas de 13 substâncias presentes em vinhos. Os vinhos analisados são de uma mesma região, mas de 3 vinícolas distintas.

Balance

A base sintética *Balance* foi gerada para modelar experimentos psicológicos por Siegler (1976) e possui 4 atributos: 2 distâncias e 2 forças, ver Figura 3.2. A combinação desses atributos leva a balança a 3 estados possíveis:

1. Se força-esquerda * distância > força-esquerda * distância, então esquerda;
2. Se força-esquerda * distância < força-esquerda * distância, então direita;
3. Se força-esquerda * distância = força-esquerda * distância, então equilíbrio.

Figura 3.2: Base de dados *Balance*. A combinação de quatro atributos leva a balança a três estados: esquerda, direita, ou equilibrada.

Zoo

A base *Zoo*, criada por Forsyth (1990), descreve 100 animais através de 16 atributos, Tabela 3.1. Os animais estão distribuídos em 8 classes: *amphibians*, *birds*, *arthropods*, *crustaceans*, *fish*, *insect*, *mammals* e *reptiles*. Sua escolha é justificada, pois foi usada para testar uma rede ART2 modificada em Qian et al. (2005).

Tabela 3.1: Descrição dos 16 atributos usados para descrever 100 animais na base *Zoo* (Forsyth, 1990).

Número	Nome	Tipo
1	<i>hair</i>	booleano
2	<i>feathers</i>	booleano
3	<i>eggs</i>	booleano
4	<i>milk</i>	booleano
5	<i>airborne</i>	booleano
6	<i>aquatic</i>	booleano
7	<i>predator</i>	booleano
8	<i>toothed</i>	booleano
9	<i>backbone</i>	booleano
10	<i>breathes</i>	booleano
11	<i>venomous</i>	booleano
12	<i>fins</i>	booleano
13	<i>legs</i>	numérico
14	<i>tail</i>	booleano
15	<i>domestic</i>	booleano
16	<i>catsize</i>	booleano

Glass

A base *Glass* descreve 214 vidros usados em diferentes aplicações como janela de automóveis, edifícios ou lâmpadas. O estudo da classificação de vidros foi motivada pela sua contribuição em investigações criminais (German, 1987). Os dados são descritos por 9 elementos químicos e separados em 6 classes (originalmente eram 7, mas a classe *non-float processed vehicle windows* contém 0 dados e foi descartada).

As bases usadas nesta etapa são resumidas na Tabela 3.2.

Tabela 3.2: Descrição das bases usadas na validação da rede ART2.

Nome da base	Número de classes	Número de variáveis	Número de dados (por classe)
<i>Iris</i>	3	4	150 (50) (50) (50)
<i>Wine</i>	3	13	178 (59) (71) (48)
<i>Balance</i>	3	4	625 (288) (49) (288)
<i>Zoo</i>	7	16	100 (40) (20) (13) (10) (8) (5) (4)
<i>Glass</i>	6	9	214 (76) (70) (29) (17) (13) (9)

3.2 Validação do Algoritmo da Rede ART2 Implementado

Como as bases empregadas têm atributos contínuos, a validação foi feita treinando a rede ART2 em aprendizagem lenta. Todos os parâmetros escolhidos estão relacionados na Tabela 3.3, com exceção de ρ . Varia-se ρ até se obter o mesmo número de classes da classificação original. A etapa de validação para cada base segue os seguintes passos:

1. Carregar os parâmetros e um novo ρ ;
2. Treinar a rede até estabilizar;
3. Se o número de classes for igual ao original, salvar os resultados em uma tabela;
4. Converter a tabela em um escore;
5. Salvar o melhor escore.

Tabela 3.3: Parâmetros da rede ART2.

Parâmetro	Descrição	Valor
a, b	pesos fixos da camada F1	$a = 10, b = 10$
c	peso fixo usado no teste de reset	$c = 0.1$
d	ativação do nó vencedor J na camada F2	$d = 0,9$
e	parâmetro para prevenir a divisão por zero	$e = 10^{-9}$
θ	parâmetro de supressão de ruído	$\theta = \frac{1}{\sqrt{n}}$
α	taxa de aprendizado	$\alpha = 0,6$
b_{iJ}	pesos <i>bottom-up</i> iniciais	$b_{iJ} = \frac{1}{(1-d)\sqrt{n}}$
t_{Ji}	pesos <i>top-down</i> iniciais	$t_{Ji} = 0$
ρ	parâmetro de vigilância	$\rho = 0,9$

A rede ART2 é pouco sensível aos valores de a e b . Foram utilizados valores sugeridos na literatura, $a = 10$ e $b = 10$. Valores de c e d devem satisfazer a seguinte inequação: $\frac{cd}{1-d} \leq 1$. Para maximizar a abrangência de ρ , deve-se

Tabela 3.4: Exemplo de cálculo do escore em um dos testes realizados com rede ART2 para a base Iris com $\rho = 0.98$.

classes	Classe 1	Classe 2	Classe 3	Total
Iris-setosa	50	0	0	50
Iris-versicolor	0	3	47	50
Iris-virginica	1	44	5	50
Total	50	47	52	150

escolher d próximo de 1. Assim, após testes, definiu-se $d = 0,9$ e, por consequência, $c = 0,1$. O parâmetro e deve receber um valor muito pequeno para não influenciar nos cálculos, usou-se $e = 10^{-9}$. Escolheu-se uma taxa de aprendizado $\alpha = 0,6$. Baixos valores de *alpha* garantem que os pesos vão atingir seus valores assintóticos. Os pesos *bottom-up* iniciais foram escolhidos de acordo com a equação $b_{iJ} = \frac{1}{(1-d)\sqrt{n}}$ para prevenir que um novo neurônio vencedor seja selecionado durante a ressonância. Os pesos *top-down* iniciais devem ser zerados para garantir que não ocorrerá um sinal de *reset* na primeira vez que uma classe receber um dado.

O escore é a porcentagem de acerto do algoritmo em relação à classificação original, Tabela 3.4. Cada classe resultante é comparada com sua classe original correspondente, aquela com a qual ela tem mais elementos em comum:

1. Iris-setosa corresponde a Classe 1, 50 dados de 50, 100% de acerto;
2. Iris-versicolor corresponde a Classe 3, 44 dados de 50, 88% de acerto;
3. Iris-virginica corresponde a Classe 2, 47 dados de 50, 94% de acerto.

Para o teste com a base Iris mostrado na Tabela 3.4, o escore é calculado como a média dos acertos em cada classe, $(100\% + 94\% + 88\%)/3 = 94\%$.

3.3 Base de Dados de Estirpes Tropicais do Gênero *Bradyrhizobium*

Os dados utilizados neste trabalho foram fornecidos pelo laboratório de Biotecnologia do Solo da Embrapa, Centro Nacional de Pesquisa de Soja (Embrapa, 2007), Londrina, Paraná. São 119 estirpes brasileiras de *Bradyrhizobium* que apresentam características fenóticas de *B. elkanii* e *B. japonicum*. As estirpes foram isoladas de 33 espécies de leguminosas tropicais. Essa base foi usada em vários trabalhos recentes (Germano-Silva, 2003; Milagre, 2003; Sawada et al., 2003; Germano et al., 2006).

3.4 Regiões Ribossomais Analisadas e Enzimas de Restrição

As estirpes de espécies do gênero *Bradyrhizobium* foram submetidas às etapas de crescimento, extração e amplificação do rRNA através de PCR. O rRNA é então cortado com enzimas de restrição com a técnica RFLP. Cada enzima corta o rRNA somente onde encontra uma sequência específica. As diferenças na sequência de rRNA dos indivíduos resulta em fragmentos de tamanhos distintos, que são separados através de eletroforese em gel.

Cada estirpe é descrita por 9 canaletas de gel de eletroforese. Analisaram-se 3 regiões ribossomais: 16S, 23S e IGS. Para cada região foram usadas 3 enzimas de restrição descritas em (Germano et al., 2006). Portanto, são 9 enzimas de restrição para cada estirpe conforme relacionado na Tabela 3.5.

A extração das informações foi feita de acordo com Talavera (2007). O objetivo do processamento das imagens de gel de eletroforese é detectar as bandas nas canaletas do gel. Esta etapa transforma canaletas em eletroferogramas t . Resumidamente, cada imagem do gel de eletroforese tem suas canaletas recortadas gerando imagens individuais. Cada canaleta corresponde a uma matriz de tons de cinza $R = [r_{pq}]$, em que p e q são o comprimento e altura da canaleta em número de pixels. Todas as canaletas têm o mesmo comprimento, $p = 340 \text{ pixels}$. Já, a altura varia entre $18 \leq q \leq 25 \text{ pixels}$. Como há ruído de fundo, deformações

Tabela 3.5: Relação entre enzimas de restrição usadas e respectivas regiões ribossomais a que se referem.

Número	Enzima de Restrição	Região Ribossomal
1	Cfo I	16S
2	Dde I	16S
3	Msp I	16S
4	Hae III	23S
5	Hha I	23S
6	Hinf I	23S
7	Dde I	IGS
8	Hae III	IGS
9	Msp I	IGS

nas bandas e variabilidade de iluminação, a transformação das imagens em eletroferogramas busca minimizar esses inconvenientes (Talavera, 2007). Assim, um filtro FIR passa-baixas é usado para reduzir ruídos de fundo. No último passo, as matrizes R são comprimidas em eletroferogramas t (Talavera, 2007).

Os passos anteriores foram executados para as 119 estirpes. São 9 canaletas de gel de eletroforese para cada estirpe, gerando um total de 1152 eletroferogramas t de 340 variáveis. O objetivo agora é transformar as imagens de gel de eletroforese em uma entrada para a rede ART2. Cada estirpe é representada por 9 eletroferogramas. Essa informação deve ser colocada em um vetor.

3.5 Geração das Matrizes de Distância DP

Esta etapa recebe como entrada as 119 estirpes. O resultado são 9 matrizes de distância. Todos os eletroferogramas que representam uma tupla específica, por exemplo (16S,Cfo I) foram agrupados em uma matriz de vetores T , como pode ser visto na Figura 3.3. São 9 matrizes, uma para cada tupla, $T_{ij} = [t_i]$, conforme Tabela 3.5. Onde i é o número de estirpes e j é o comprimento do vetor t .

A matriz de distância, (3.1), define a distância de Pearson, (2.2), entre todos os eletroferogramas t da matriz T , em que k é número da matriz, i e j são os índices dos eletroferogramas.

$$DP_{k_{ij}} = [dp(t_i, t_j)] \quad (3.1)$$

Figura 3.3: Geração das matrizes de distância a partir das estirpes. A primeira canaleta de cada estirpe é colocada na matriz 16S – *CfoI*, a segunda na matriz 16S – *DdeI*. Esse passo gera 9 matrizes de distância, uma para cada região do rRNA digerida por uma enzima específica.

3.5.1 Cálculo da Matriz de Distância Final C

Esta etapa recebe como entrada 9 matrizes de distância e retorna 1 matriz de distância, Figura 3.4. O objetivo é condensar as k matrizes de distância (3.1) em uma única matriz C . As matrizes têm suas distâncias somadas de acordo com (3.2).

$$C_{ij} = \sqrt{\sum_{k=1}^n (DP_{k_{ij}})^2} \quad (3.2)$$

3.6 Cálculo da Matriz de Pontos com MDS

Esta é a última etapa na transformação da base de estirpes de *Bradyrhizobium* em uma entrada para a rede ART2, a qual recebe como entrada uma matriz de distância e retorna um espaço de pontos (Llerena et al., 2007). O objetivo é reduzir a dimensionalidade dos eletroferogramas.

A idéia básica é mapear um conjunto de dados d -dimensional em um novo espaço d' dimensional. A matriz de distância C é transformada em um espaço de pontos com o algoritmo DISTATIS (Abdi et al., 2005). Esse algoritmo inicia com a matriz de distâncias dos eletroferogramas e executa sucessivas manipulações até encontrar um conjunto de pontos em algum espaço euclidiano k -dimensional. Esse espaço representa os eletroferogramas e suas coordenadas geram as distâncias contidas na matriz de distâncias.

O conjunto de pontos é salvo em uma matriz P . A Figura 3.4 mostra que cada linha é um dado I representando uma estirpe. Assim a linha 1 corresponde à primeira estirpe. Cada dado tem 119 variáveis que definem um ponto no espaço. Suas variáveis estão ordenadas em ordem de importância. A escolha do número ótimo de variáveis é feita observando os autovalores retornados pelo algoritmo DISTATIS, os quais indicam a quantidade de informação representada por cada variável. A idéia é escolher um número de variáveis, cuja soma cumulada dos autovalores seja maior que uma determinada porcentagem total de autovalores. Na base de dados de estirpes de *Bradyrhizobium* essa porcentagem foi escolhida em 75%, correspondendo às primeiras 15 variáveis da matriz de pontos.

Figura 3.4: As matrizes de distância são agrupadas em uma única C através de (3.2). A matriz C serve como entrada para o processamento com MDS. O algoritmo DISTATIS transforma C em uma matriz de pontos em que cada linha é um dado I representando uma estirpe.

3.7 Cálculo do Número Ótimo de Classes com o Algoritmo *Distance*

O objetivo de classificação é particionar os dados em classes onde dados em uma mesma classe sejam mais similares que dados em classes distintas (Guha et al., 1998). Assim, deve-se minimizar a distância entre os dados em uma mesma classe (distância intra-classe) e maximizar a distância entre dados em classes distintas (distância inter-classe).

O algoritmo *Distance* determina o número ótimo de classes em uma base através da Distância de Pearson. A distância intra-classe, (3.3), calcula a distância média entre os dados p_{ij} em uma classe i e o centróide da classe c_i .

$$intra = \frac{\sum_{i=1}^m \sum_{j=1}^n d_p(p_{ij}, c_i)}{m} \quad (3.3)$$

onde m é o número de classes criadas pela rede ART2 e n é o número de dados em cada classe. A distância inter-classe, (3.4), define a distância média entre os centróides das classes criadas pela rede ART2.

$$inter = \frac{\sum_{i=1}^m \sum_{j=1}^m d_p(c_i, c_j)}{m} \quad (3.4)$$

Distance usa a rede ART2 para gerar particionamentos com diferentes números de classes. Primeiramente é necessário um conjunto de valores de ρ . Para tanto, implementou-se o algoritmo *RhoGenerator*, o qual gera valores de ρ em escala logarítmica, ver Figura 3.5. Isso é feito já que a rede ART2 é mais sensível aos valores finais de ρ (Fausett, 1997) onde o número de classes muda rapidamente, Figura 3.6. Apesar do parâmetro de vigilância variar entre $0 \leq \rho \leq 1$, valores práticos estão entre $0.7 \leq \rho < 1$ (Fausett, 1997).

Figura 3.5: Valores de ρ produzidos com uma função logarítmica pelo algoritmo *rho_generator* com 27 valores no intervalo $0.94 \leq \rho < 1$.

Figura 3.6: Número de classes geradas pela rede ART2 em função de ρ usando os valores representados na Figura 3.5. O número de classes cresce mais rapidamente quando o valor de ρ se aproxima de 1. Este gráfico mostra os resultados da rede ART2.

A cada execução da rede, um novo ρ é usado e as distâncias são calculadas. Trata-se de um problema de otimização multi-objetivo. Procura-se maximizar a distância inter-classe, dado que a distância intra-classe é mínima. Assim, como as distâncias inter e intra classes dadas por (3.3) e (3.4) estão na mesma escala, é possível encontrar um ponto no plano R2 onde as duas equações são iguais. Este ponto define o número ótimo de classes, ver Figura 3.7.

Figura 3.7: Distâncias inter e intra-classe como função do parâmetro de vigilância (ρ). Cada mudança no gráfico mostra o aumento no número de classes. O ponto de intersecção entre as equações define o número ótimo de classes, neste caso são identificadas 13 classes com $\rho = 0.9873$.

3.8 Representação Gráfica da Dinâmica dos Dados da Base de Estirpes de *Bradyrhizobium*

O objetivo desta etapa é gerar uma representação gráfica da dinâmica dos dados da base de estirpes do gênero *Bradyrhizobium* quando esta é dividida em diferentes partições pela rede ART2. Quando se aumenta a similaridade necessária para classificar dados em uma mesma classe, a rede ART2 cria partições mais

granulares. Então, a distribuição dos dados através das classes, em comparação com a partição anterior, é alterada. A informação sobre a distribuição dos dados nas classes em diferentes particionamentos é relevante.

O Algoritmo *DrawGraph* (1) gera automaticamente um grafo para ilustrar como as classes estão relacionadas quando seus dados são redistribuídos devido à mudança no número de classes. Neste grafo, cada vértice corresponde a uma classe e as relações entre as classes são representadas por arestas direcionadas (setas). Para desenhar o grafo é utilizado o aplicativo Graphviz (Ellson et al., 2007). Graphviz é um aplicativo gráfico *open-source*, no qual grafos podem ser especificados em uma linguagem de descrição nomeada DOT (Ellson et al., 2007).

Algoritmo 1 *DrawGraph*

```

 $E \leftarrow \{base\ de\ dados\}$ 
 $R \leftarrow \{valores\ de\ \rho\}$ 
 $K \leftarrow nmero\ de\ passos$ 

%criar particoes
for  $k = 0$  to  $K$  do
     $k = k + 1$ 
     $\rho \leftarrow R[i]$ 
     $P(k) \leftarrow ART2(E, \rho)$ 
end for

%criar regras
for  $k = 0$  to  $K - 1$  do
     $L(k) \leftarrow create\_rules(P(k), P(k + 1))$ 
end for

%converter regras para linguagem dot
for  $k = 0$  to  $K - 1$  do
     $script[k] \leftarrow convert\_rule(L(K))$ 
end for

%criar arquivo dot
 $dot\_file \leftarrow create\_dot\_file(script)$ 

%desenhar grafo
 $graph \leftarrow graphviz(dot\_file)$ 

```

A rede ART2 é empregada para gerar k partições, com um número crescente de classes. A rede é executada para um conjunto de parâmetros de vigilância $R = \{r_1, r_2, \dots, r_k\}$ produzindo um conjunto de partições $P_k = \{c_{k1}, c_{k1}, \dots, c_{km}\}$, onde $P_{k+1} > P_k$. Classes em partições consecutivas são consideradas relacionadas quando compartilham dados. Se c_k está relacionada com $c_{(k+1)}$, então uma regra de associação $c_k \rightarrow c_{(k+1)}$ é criada. Todas as classes geradas em um passo k são comparadas com aquelas do passo $k + 1$, gerando um conjunto de regras $L_k = \{(c_{k1} \rightarrow c_{(k+1)1}), (c_{k1} \rightarrow c_{(k+1)2}), \dots, (c_{kx} \rightarrow c_{(k+1)y})\}$. Este passo é repetido para todas as partições consecutivas, gerando $k - 1$ conjuntos de regras. No grafo, essas regras representam vértices direcionados. Finalmente, as regras são usadas para gerar um *script* na linguagem DOT e o aplicativo Graphviz é executado para desenhar o grafo.

O exemplo na Figura 3.8 ilustra como as classes estão relacionadas em diferentes partições. Neste caso, o Algoritmo (1) foi executado para a base $E = \{A, B, C, D, E\}$, com os seguintes valores de ρ $R = \{0.8, 0.9, 0.95\}$ em $K = 3$ passos. Para estes valores de ρ , a rede ART2 produziu as seguintes partições $P_1 = \{c_{11}\}$, $P_2 = \{c_{21}, c_{22}\}$ e $P_3 = \{c_{31}, c_{32}, c_{33}\}$. Analisando essas partições, gerou-se as regras $L_1 = \{(c_{11} \rightarrow c_{21}), (c_{11} \rightarrow c_{22})\}$ e $L_2 = \{(c_{21} \rightarrow c_{31}), (c_{21} \rightarrow c_{32}), (c_{22} \rightarrow c_{32}), (c_{22} \rightarrow c_{33})\}$.

Figura 3.8: Exemplo de grafo gerado pelo Algoritmo (1). A rede ART2 gerou 3 partições com 1, 2 e 3 classes respectivamente. O grafo mostra como os dados são redistribuídos pela rede quando se altera o número de classes.

Capítulo 4

Resultados e Discussão

4.1 Resultados da Rede ART2 Comparados com Outros Classificadores

O objetivo desta etapa é apresentar os resultados da rede ART2 contra outros classificadores usados na literatura. Foram empregadas 3 bases de áreas distintas, nas quais o número original de classes é conhecido: *Wine*, *Balance* e *Iris*.

A Tabela 4.1 mostra os resultados obtidos com os classificadores para cada uma das bases testadas. Cada coluna apresenta os escores para uma base. As linhas mostram os diferentes classificadores.

Tabela 4.1: Resultado dos algoritmos para cada base.

Algoritmo	<i>Wine</i> (%)	<i>Balance</i> (%)	<i>Iris</i> (%)	Escore médio (%) dos classificadores
<i>Twostep</i>	95	50	86	77
<i>K-means</i>	95	53	80	76
<i>Positive Approach</i>	90	35	97	74
ART2	68	49	97	71
Kohonen	76	45	86	69
<i>Wards Methods</i>	69	47	89	68
<i>Furthest Neighbor</i>	65	47	84	65
<i>Within-groups Linkage</i>	56	47	84	62
<i>Centroid Clustering</i>	56	49	91	62
<i>Between-groups Linkage</i>	56	47	75	59
<i>Median Clustering</i>	49	49	75	58
<i>Nearest Neighbor</i>	36	34	68	46
Mínimo (aleatório)	33	33	33	33
Escore médio das bases (%)	67	46	84	

A Figura 4.1 apresenta o escore dos classificadores para cada base de dados, com uma linha por base. Os escores das bases ficaram em 3 níveis bem definidos. A base Iris recebeu escores altos, *Wine* escores intermediários e *Balance* escores baixos. Além disso, os classificadores obtiveram escores distintos. Isso era esperado pois estes classificadores usam diferentes medidas de similaridades.

Figura 4.1: Resultado dos classificadores para cada base de dados. Cada linha corresponde a uma base de dados.

4.2 Resultados das Distâncias Inter e Intra-classe

Os resultados do algoritmo *Distance* para calcular automaticamente o número de classes são mostrados a seguir. Foram usadas as bases: Iris, *Glass*, *Wine* e *Zoo*.

As Figuras 4.2 a 4.5 apresentam o resultado das distâncias inter (3.4) e intra-classe (3.3) para as bases escolhidas, mostrando a influência do parâmetro de vigilância no desempenho da rede ART2. Cada Figura começa com valores de ρ que produzem uma classe. Aumentando-se ρ , a rede cria mais classes. Quando uma nova classe é criada, os valores das distâncias mudam.

Especificamente, pode-se observar a Figura 4.2 da base Iris. Essa base tem uma partição natural de 3 classes. Inicialmente há uma classe, a distância intra-classe é máxima e a distância inter-classe mínima. Incrementando o total de classes, a distância intra diminui e a inter aumenta. Isso mostra que o algoritmo está se aproximando da partição natural para esta base. As distâncias são iguais na terceira mudança no gráfico, ou seja, em 3 classes.

Figura 4.2: Base Iris. Distâncias inter e intra-classe como função do parâmetro de vigilância (ρ). Cada mudança no gráfico mostra o aumento no número de classes. O ponto de intersecção entre as equações define o número ótimo de classes, neste caso 3 classes com $\rho = 0.998121$.

Figura 4.3: Base *Glass*. Distâncias inter e intra-classe como função de ρ . Cada mudança no gráfico mostra o aumento no número de classes. O ponto de intersecção entre as equações define o número ótimo de classes, neste caso 6 classes com $\rho = 0.999750$.

Figura 4.4: Base *Wine*. Distâncias inter e intra-classe como função de ρ . Cada mudança no gráfico mostra o aumento no número de classes. O ponto de intersecção entre as equações define o número ótimo de classes, neste caso 3 classes com $\rho = 0.999318$.

Figura 4.5: Base *Zoo*. Distâncias inter e intra-classe como função de ρ . Cada mudança no gráfico mostra o aumento no número de classes. O ponto de intersecção entre as equações define o número ótimo de classes, neste caso 8 classes com $\rho = 0.966478$.

Figura 4.6: Base *Balance*. Distâncias inter e intra-classe como função de ρ . Cada mudança no gráfico mostra o aumento no número de classes. O ponto de intersecção entre as equações define o número ótimo de classes, neste caso 13 classes com $\rho = 0.987316$.

A Tabela 4.2 compara o número de classes gerado pelo algoritmo *Distance* e aquele proposto pelos autores. Também apresenta o valor de ρ usado na rede ART2. Os resultados são consistentes, o algoritmo *Distance* acertou o total de classes para todas as bases, exceto na base *Balance*.

Tabela 4.2: Resultado do algoritmo *Distance* para as bases *Iris*, *Glass*, *Wine*, *Zoo* e *Balance*.

Base	Parâmetro de vigilância	Número de classes <i>Distance</i>	Número de classes original
<i>Iris</i>	0.998121	3	3
<i>Glass</i>	0.999750	6	6
<i>Wine</i>	0.999318	3	3
<i>Zoo</i>	0.966478	8	8
<i>Balance</i>	0.987316	13	3

4.3 Resultados para a Base de Estirpes do gênero *Bradyrhizobium*

Figura 4.7: Base de estirpes do gênero *Bradyrhizobium*. Distâncias inter e intra-classe como função do parâmetro de vigilância (ρ). O ponto de intersecção entre as equações define o número ótimo de classes, neste caso 4 classes com $\rho = 0.994893$.

A Figura 4.7 ilustra o resultado do algoritmo *Distance* para a base de estirpes do gênero *Bradyrhizobium*. Nesta base, não há informação *a priori* sobre o número ótimo de classes. O algoritmo *Distance* sugeriu que a base possui uma partição natural de 4 classes.

A Figura 4.8 ilustra o resultado do algoritmo *DrawGraph* para a base de estirpes de *Bradyrhizobium* em 7 particionamentos distintos. Em cada nível do grafo, pode-se observar os valores de ρ usados e o corresponde particionamento produzido. Os dados foram particionados em até 10 classes.

Para realizar uma comparação com outra representação gráfica, usou-se um dendrograma gerado pelo método UPGMA (Unweighted Pair Group Method with Arithmetic mean) (Sneath e Sokal, 1973), ver a Figura 4.9. Este dendrograma foi cortado em 10 classes e seus nós representam a porcentagem de elementos como no grafo ilustrado na Figura 4.8. É possível notar uma boa semelhança entre as duas estruturas. A criação das classes apresentou um padrão semelhante, conforme pode-se observar no detalhe vermelho.

O algoritmo *Distance* sugeriu que a base de estirpes do gênero *Bradyrhizobium* possui 4 classes. Assim, o grafo gerado na Figura 4.8 foi cortado em 4 classes. A distribuição das estirpes nas classes é mostrada na Tabela 4.3.

O resultado da classificação foi avaliado utilizando-se estirpes de referência. Essas estirpes são selecionadas por especialistas como referência para uma determinada espécie. Também, dentre as estirpes de referência, é escolhida aquela mais representativa da espécie. A Tabela 4.4 apresenta as estirpes selecionadas como referência das espécies *B. japonicum* e *B. elkanii* usadas neste trabalho (Germano-Silva, 2003).

Figura 4.8: Grafo produzido pelo algoritmo *DrawGraph* para a base de estirpes do gênero *Bradyrhizobium*. O grafo ilustra a dinâmica dos dados em 7 particionamentos distintos, iniciando com 1 classe até 10 classes.

Figura 4.9: Dendrograma gerado pelo método UPGMA para a base de estirpes do gênero *Bradyrhizobium*. Para facilitar a visualização, cortou-se o dendrograma em 10 classes. Seus nós representam a porcentagem de elementos.

Tabela 4.3: Distribuição das estirpes de *Bradyrhizobium*, cortando o grafo em 4 classes. As estirpes de referência das espécies *B. japonicum* e *B. elkanii* estão em destaque.

Classes	A	B	C	D
Número de				
estirpes	53 (44,5%)	43 (36,1%)	13 (10,9%)	10 (8,4%)
Estirpes	SEMIA 501	SEMIA 535	SEMIA 6169	SEMIA 587
	SEMIA 509	PRY 40	SEMIA 6175	SEMIA 590
	SEMIA 511	PRY 42	SEMIA 6192	SEMIA 598
	SEMIA 515	PRY 49	SEMIA 6208	SEMIA 656
	SEMIA 518	PRY 52	SEMIA 6319	SEMIA 658
	SEMIA 528	R17	SEMIA 6382	SEMIA 662
	SEMIA 538	R35	SEMIA 6387	SEMIA 695
	SEMIA 542	R45	SEMIA 6420	SEMIA 696
	SEMIA 543	AM-Cp 17	SEMIA 6424	SEMIA 5003
	SEMIA 549	AM-P2 Lima	SEMIA 6425	SEMIA 5006
	SEMIA 560	AM-P5 Abac	SEMIA 5011	
	SEMIA 562	AM-2-855	SEMIA 5012	
	SEMIA 565	AM-01-517	SEMIA 5016	
	SEMIA 566	SEMIA 5019		
	SEMIA 567	SEMIA 5021		
	SEMIA 568	SEMIA 5024		
	SEMIA 569	SEMIA 5026		
	SEMIA 571	SEMIA 5027		
	SEMIA 576	SEMIA 5029		
	SEMIA 577	SEMIA 5030		
	SEMIA 580	SEMIA 5036		
	SEMIA 581	SEMIA 6093		
	SEMIA 583	SEMIA 6099		
	SEMIA 586	SEMIA 6101		
	SEMIA 5000	SEMIA 6144		
	SEMIA 5005	SEMIA 6145		
	SEMIA 5022	SEMIA 6146		
	SEMIA 5023	SEMIA 6148		
	SEMIA 5038	SEMIA 6149		
	SEMIA 5042	SEMIA 6150		
	SEMIA 5045	SEMIA 6152		
	SEMIA 5048	SEMIA 6154		
	SEMIA 5055	SEMIA 6156		
	SEMIA 5056	SEMIA 6157		
	SEMIA 5059	SEMIA 6158		
	SEMIA 5060	SEMIA 6159		
	SEMIA 5062	SEMIA 6160		
	SEMIA 5063	SEMIA 6161		
	SEMIA 5068	SEMIA 6163		
	SEMIA 5069	SEMIA 6164		
	SEMIA 5070	SEMIA 6165		
	SEMIA 5075	USDA 76		
	SEMIA 5079	SEMIA 6167		
	SEMIA 5080			
	SEMIA 5082			
	SEMIA 5084			
	SEMIA 5085			
	SEMIA 6002			
	SEMIA 6028			
	SEMIA 6053			
	SEMIA 6155			
	SEMIA 6166			
	BTail			

Tabela 4.4: Relação de estirpes de referência das espécies *B. japonicum* e *B. elkanii* usadas neste trabalho conforme Germano-Silva (2003).

Número	Estirpe	Espécie de Referência
1	SEMIA 566	<i>B. japonicum</i>
2	SEMIA 586	<i>B. japonicum</i>
3	SEMIA 5056	<i>B. japonicum</i> (estirpe escolhida como representativa da espécie)
4	SEMIA 5079	<i>B. japonicum</i>
5	SEMIA 5080	<i>B. japonicum</i>
6	SEMIA 5085	<i>B. japonicum</i>
7	USDA 76	<i>B. elkanii</i> (estirpe escolhida como representativa da espécie)
8	SEMIA 587	<i>B. elkanii</i>
9	SEMIA 5019	<i>B. elkanii</i>

Assumindo que as estirpes de referência mostradas na Tabela 4.4 são boas representantes de suas espécies, tem-se que:

1. Classe *A*: As estirpes desta classe devem ser da espécie *B. japonicum*. Esta classe contém todas as estirpes referência da espécie *B. japonicum*, inclusive aquela escolhida como representativa da espécie, SEMIA 5056. Também, como todas as referências foram agrupadas na mesma classe, confirma-se que elas são boas representantes desta espécie;
2. Classe *B*: Possui 2 referências de *B. elkanii*, sendo a estirpe USDA 76 a representativa desta espécie; Provavelmente, as estirpes classificadas aqui são da espécie *B. elkanii*;
3. Classe *C*: Não reuniu nenhuma espécie de referência. É possível que as estirpes desta classe façam parte de uma nova subespécie, no mesmo nível evolutivo de *B. elkanii* e *B. japonicum*;
4. Classe *D*: Contém uma referência de *B. elkanii*. Esta classe pode ser uma subespécie de *B. elkanii*.

Capítulo 5

Conclusão

Este trabalho utilizou uma rede neural ART2 para classificação. Primeiramente, a rede ART2 foi comparada com 11 classificadores clássicos em 3 bases de dados de áreas distintas do conhecimento. Os escores das bases ficaram em 3 níveis bem definidos, ver Figura 4.1. A base *Iris* recebeu escores mais altos, média de 86%, possivelmente por possuir 2 classes linearmente separadas. A base *Wine* recebeu escores intermediários, média de 67%. Finalmente, a base *Balance* apresentou escores baixos para todos os classificadores. Esta base apresentou um escore médio de 46%, relativamente próximo do mínimo de 33%, e baixo em relação às bases *Iris* e *Wine*. Como descrito anteriormente, a base *Balance* foi gerada pela combinação de atributos. Todavia, os classificadores usados empregam métricas que não usam combinação dos atributos para classificação dos dados.

Observando a Tabela 4.1 na vertical, nota-se que alguns algoritmos mostraram grande diferença de desempenho em uma mesma base. Por exemplo, na base *Wine*, tem-se *K-means* com 95% e *Nearest Neighbor* com apenas 36%. Essa variância também foi observada em menor escala nas outras bases. Isso ocorreu pois, geralmente, classificadores com métricas e funcionamento distintos produzem resultados diferentes em uma mesma base. Quando se aplica um determinado classificador nos dados, este vai impor sua “métrica” nesses dados. Para ilustrar essa diferença, pode-se comparar os algoritmos de *K-means* e *Nearest Neighbor*.

O algoritmo *K-means* empregado neste trabalho usa a distância euclidiana média como medida de dissimilaridade entre os dados. *K-means* tenta distribuir as classes, buscando minimizar a distância dos dados ao centro de suas classes (Kanungo et al., 2002). Já, no classificador *Nearest Neighbor*, a distância entre as

classes é calculada como a distância euclidiana mínima entre seus dados. Também, vale a comparação espacial. Num espaço tridimensional, para *K-means* as classes seriam esferas e para a rede ART2, cones.

Pode-se concluir que a rede ART2 obteve um bom desempenho, ficando entre os 4 maiores melhores classificadores. A rede recebeu o melhor escore para a base *Iris*, escore intermediário para a base *Wine* e baixo para a base *Balance*. Apesar da rede ter apresentado desempenho inferior aos classificadores *Twostep*, *K-means* e *Positive Approach*. Sua escolha implica em várias vantagens: (1) Não é necessário informar *a priori* o número de classes; (2) A rede ART2 pode ser aplicada em problemas dinâmicos, atualizando o conhecimento obtido de maneira automática. Isso ocorre pois ela é capaz de aprender novos dados e ainda reter o conhecimento daqueles já classificados; (3) Seu algoritmo é plástico e estável. A rede é capaz de classificar novos dados sem ter que ser treinada com todos os dados novamente; (4) A rede ART2 é tolerante a ruídos nos dados. Todavia, o resultado mais relevante é o fato da rede ART2 ter apresentado um comportamento compatível com os outros classificadores nas diferentes bases.

Os diferentes escores para as bases mostram a necessidade de considerar as características dos dados quando se escolhe um classificador. Também, a metodologia usada para calcular os escores pode ser questionada. Assume-se que quanto mais similares os testes são da classificação original, melhor os resultados (Spiegler, 2006). Contudo, os testes com a base *Balance* mostram que os resultados originais nem sempre podem ser obtidos com os classificadores existentes, o que não reduz a qualidade dos algoritmos utilizados.

O algoritmo *Distance* utiliza uma rede ART2 para descobrir o número de classes em uma base considerando simultaneamente as distâncias inter e intra-classe. Os resultados mostraram que ele é capaz encontrar com sucesso o particionamento natural das bases testadas: *Iris*, *Glass*, *Wine*, *Zoo* e *Balance*. O algoritmo *Distance* errou o número de classes apenas na base *Balance*. O cálculo das distâncias inter e intra classe é baseado na classificação produzida pela rede ART2. Nesta base, a rede obteve seu pior desempenho, 49%. É possível que o desempe-

nho ruim na classificação tenha prejudicado o algoritmo *Distance*. Finalmente, o algoritmo *Distance* automatiza os testes com a rede ART2, uma vez que reduz a interferência do usuário na escolha de valores de ρ .

Finalmente, a taxonomia de bactérias tropicais pertencentes ao gênero *Bradyrhizobium* é pouco conhecida. Este trabalho apresentou algoritmos que podem auxiliar no estudo taxonômico de estirpes de *Bradyrhizobium* a partir de imagens de gel de eletroforese. Implementou-se o algoritmo *DrawGraph* para construir uma representação gráfica da dinâmica dos dados desta base nos diferentes números de classes geradas pela rede ART2. A estrutura gerada mostrou-se compatível com o dendrograma produzido por UPGMA. O grafo ilustrado na Figura 4.8 apresentou um padrão semelhante ao do dendrograma da Figura 4.9. Também, o algoritmo *Distance* sugeriu que a base de estirpes do gênero *Bradyrhizobium* possui uma partição natural de 4 classes. Esse resultado é compatível com o trabalho de Milagre (2003), o qual também sugere 4 classes.

5.1 Próximos Passos

Como sugestão para trabalhos futuros, poderia-se utilizar o método de análise de estabilidade de agrupamentos (Ben-Hur e Guyon, 2003) para reamostrar os dados e avaliar o número de grupos estáveis. Isso possibilitaria uma abordagem mais robusta ao processo de determinação do número ótimo de classes. O cálculo do número ótimo de classes pelo algoritmo *Distance* é prejudicado quando a rede ART2 classifica os dados de maneira errônea. O método de análise de estabilidade de agrupamentos pode indicar a qualidade da classificação produzida pela rede ART2. Essa informação serviria para aumentar a confiabilidade no resultado produzido pelo algoritmo *Distance*. Além disso, algumas imagens de gel de eletroforese usadas apresentavam baixa qualidade. Assim, sugere-se a implementação de algum método para avaliar a qualidade das imagens. Imagens com qualidade inferior poderiam receber pesos diferentes em relação a imagens de boa qualidade.

Apêndice A

Algoritmo Detalhado da Rede ART2

Algoritmo 2 Rede ART2

inicie parâmetros $a, b, \theta, c, d, e, \alpha, \rho$

repeat

for para cada padrão de entrada faça **do**

 atualize as ativações da camada F1:

$$U_i = 0, W_i = I_i, P_i = 0$$

$$X_i = \frac{I_i}{e + \|I\|}, Q_i = 0, V_i = f(X_i)$$

 atualize as ativações da camada F1 novamente:

$$U_i = \frac{V_i}{e + \|V\|}, W_i = I_i + \alpha U_i$$

$$P_i = U_i, X_i = \frac{W_i}{e + \|W\|}$$

$$Q_i = \frac{P_i}{e + \|P\|}, V_i = f(X_i) + bf(Q_i)$$

 calcule os sinais para as unidades da camada F2:

$$Y_i = \sum_j P_i b_{ij} P_j$$

repeat

 encontre um neurônio candidato J em que $Y_J \geq Y_j$, para aprender a entrada corrente:

 teste a condição de reset:

$$U_i = \frac{V_i}{e + \|V\|}, P_i = U_i + dt_{Ji}$$

$$R_i = \frac{U_i + cP_i}{e + \|U\| + c\|P\|}$$

if $\|R\| < (\rho - e)$ **then**

$Y_J = -1$ (neurônio candidato J é inibido, *reset* é verdadeiro)

else

reset é falso

$$W_i = I_i + \alpha U_i, X_i = \frac{W_i}{e + \|W\|}$$

$$Q_i = \frac{P_i}{e + \|P\|}, V_i = f(X_i) + bf(Q_i)$$

end if

until *reset* é verdadeiro

repeat

 aprender o padrão, modificando os pesos do neurônio vencedor:

$$t_{Ji} = \alpha d U_i + 1 + \alpha d(d+1)t_{Ji}$$

$$b_{iJ} = \alpha d U_i + 1 + \alpha d(d+1)b_{iJ}$$

 atualize as ativações da camada F1:

$$U_i = \frac{V_i}{e + \|V\|}, W_i = I_i + \alpha U_i$$

$$P_i = U_i + dt_{Ji}, X_i = \frac{W_i}{e + \|W\|}$$

$$Q_i = \frac{P_i}{e + \|P\|}, V_i = f(X_i) + bf(Q_i)$$

until condição final para atualização dos pesos seja alcançada

end for

until condição de parada for falsa

Bibliografia

- Abdi, H., Valentin, D., O'Toole, A. J. e Edelman, B. (2005). DISTATIS: The analysis of multiple distance matrices, *IEEE Computer Society: International Conference on Computer Vision and Pattern Recognition* **3**(1): 42–50.
- Ben-Hur, A. e Guyon, I. (2003). Detecting Stable Clusters Using Principal Component Analysis, *Proceedings of Methods in Molecular Biology*, Humana Press, pp. 159–182.
- Bhat, K. V., Jarret, R. L. e RANA, R. S. (1995). DNA profiling of banana and plantain cultivars using random amplified polymorphic DNA (RAPD) and restriction fragment length polymorphism (RFLP) markers, *Electrophoresis* **16**(1): 1736–1745.
- Braga, A. P., Ludermir, T. B. e Carvalho, A. P. L. (2000). *Redes Neurais Artificiais: Teoria e Aplicações*, LTC.
- Butler, J. M. (2005). *Forensic DNA Typing, Second Edition: Biology, Technology, and Genetics of STR Markers*, second ed, Academic Press.
- Canhos, V. P., Coutinho, H. L. C. e Vazoller, R. (1998). Estratégia Nacional de Diversidade Biológica. Grupo de Trabalho Temático: Microrganismos e Biodiversidade de Solos, *Relatório técnico*.
- Canhos, V. P. e Manfio, G. P. (2000). Recursos microbiológicos para biotecnologia, *Relatório técnico*, Ministério da Ciência e Tecnologia (MCT).
- Carpenter, G. A. e Grossberg, S. (1987a). A massively parallel architecture for a self-organizing neural pattern recognition machine, *Computer Vision, Graphics, and Image Processing* **37**(1): 54–115.

- Carpenter, G. A. e Grossberg, S. (1987b). ART2: self-organization of stable category recognition codes for analog input patterns, *Applied Optics* **26**(23): 4919–4930.
- Carpenter, G. A. e Grossberg, S. (1988). The ART of adaptive pattern recognition by a self-organizing neural network, *Computer* **21**(3): 77–88.
- Carpenter, G. A. e Grossberg, S. (2003). Adaptive Resonance Theory, in M. A. Arbib (ed.), *The Handbook of Brain Theory and Neural Networks*, MIT Press, Cambridge, pp. 87–90.
- Carpenter, G. A., Grossberg, S. e Reynolds, J. (1991a). ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network, *Neural Networks* **4**(5): 565–588.
- Carpenter, G. A., Grossberg, S. e Reynolds, J. (1992). Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps, *IEEE Transactions on Neural Networks* **3**(1): 698–713.
- Carpenter, G. A., Grossberg, S. e Rosen, D. B. (1991). Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system, *Neural Networks* **4**(1): 759 – 771.
- Carpenter, G. A., Grossberg, S. e Rosen, D. B. (1991b). ART2-A: An adaptive resonance algorithm for rapid category learning and recognition, *Neural Networks* **4**(4): 493–504.
- Chen, L. S., Pedrosa, F. O. e Hungria, M. (2000). Genetic characterization of soybean rhizobia in paraguay, *Applied and Environment Microbiology* **66**(1): 5099–5103.
- Chueire, L. M. O., Bangel, E. V., Mostasso, F. L., Campo, R. J., Pedrosa, F. O. e Hungria, M. (2003). Classificação Taxonômica das estirpes de rizóbio reco-

- mendadas para as culturas da soja e do feijoeiro baseada no sequenciamento do gene 16S rRNA, *Revista Brasileira de Ciência do Solo* **27**(5): 833–839.
- Chueire, M. L. (2000). Identificação das estirpes de *Bradyrhizobium* e *Rhizobium* utilizadas em inoculantes comerciais para as culturas da soja e do feijoeiro pela técnica de PCR com primers aleatórios ou específicos, *Agricultura Tropical* **4**(1): 80–95.
- Cottrell, G. W. (1990). Extracting features from faces using compression networks: Face, identity, emotion and gender recognition using holons, in T. D. (ed.), *Connection Models: Proceedings of the 1990 Summer School*, Morgan Kaufmann, San Mateo, C.A., pp. 459–473.
- Dasarathy, B. V. (1980). Nosing Around the Neighborhood: A New System Structure and Classification Rule for Recognition in Partially Exposed Environments, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2**(1): 67–71.
- Duda, R. O. e Hart, P. E. (1973). *Pattern Classification and Scene Analysis*, John Wiley and Sons.
- Duda, R. O., Hart, P. E. e Stork, D. G. (2001). *Pattern Classification*, second ed, John Wiley, New York.
- Ellson, J., Gansner, E., Koren, Y., Koutsofios, E. e Mocenigo, J. (2007). Graphviz - Graph Visualization Software. <http://www.graphviz.org/>.
- Embrapa, S. (2007). A Soja. Disponível em:<<http://www.cnpso.embrapa.br/>>. Acesso em: 10 jan. 2007.
- Everitt, B. S., Landau, S. e Leese, M. (2001). *Cluster Analysis*, fourth ed, Oxford University Press Inc., New York.
- Faria, C. A. (2006). *Aplicação do Escalonamento Multidimensional ao Julgamento e Classificação de Conceitos Emocionais*, Tese de Mestrado, Universidade Federal de Uberlândia.

- Fausett, L. V. (1997). *Fundamentals of Neural Networks Architectures, Algorithms, and Applications*, first ed, Prentice Hall International Inc, New Jersey.
- Fayyad, M. U., Piatesky-Shapiro, G., Smuth, P. e Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*, AAAI Press.
- Federici, C. T., Roose, M. L. e Scora, R. W. (2000). RFLP analysis of the origin of citrus bergamia, citrus jambhiri, and citrus limonia. *Acta Horticulturae, Proceedings of the First International Citrus Biotechnology Symposium* **535**(1): 55–62.
- Filho, A. M. V., Carraro, J. L., Bastos, L. C., Pezzi, S. e Romão, W. (1997). Uma visão geral sobre rede neural artificial com arquitetura ART2, *Revista Tecnológica, Maringá* **6**(1): 59–71.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, *Annual Eugenics* **7**(2): 179–188.
- Forina, M. (1991). PARVUS - An Extendible package for data exploration, classification and correlation, *Relatório técnico*, Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy.
- Forsyth, R. S. (1990). Zoo (animals) database.
- Freeman, J. A. e Skapura, D. M. (1991). *Neural Networks - Algorithms, Applications and Programming Techniques*, Addison Wesley.
- Garrity, G. e Holt, J. G. (2001). Rhizobiaceae Conn, in G. M. Garrity, D. R. Boone e C. R. W. (eds), *Bergey's Manual of Systematic Bacteriology*, Vol. 1, Springer, pp. 119–154.
- German, B. (1987). Glass Identification Database. Disponível em: <<http://mlearn.ics.uci.edu/databases/glass/>>. Acesso em: 10 jan. 2007.

-
- Germano, M. G., Menna, P., Mostasso, F. L. e Hungria, M. (2006). RFLP analysis of the rRNA operon of a Brazilian collection of bradyrhizobial strains from 33 legume species, *International Journal of Systematic and Evolutionary Microbiology* **56**(1): 217–229.
- Germano-Silva, M. G. (2003). *Avaliação da Diversidade Genética de Bradyrhizobium pela Análise de Genes Ribossomais*, Tese de Doutorado, Universidade Estadual de Londrina.
- Glover, D. M. (2002). *DNA Cloning: A Practical Approach Volume 1: Core Techniques*, second ed, Oxford University Press, USA.
- Green, P. E. e Robison, P. J. (1967). Cluster analysis in test market selection, *Management Science* **13**(8): 387–400.
- Grossberg, S. (1976). Adaptive pattern classification and universal recoding, i: Parallel development and coding of neural feature detectors and ii: Feedback, expectation, olfaction, and illusions, *Biological Cybernetics* **23**(3): 121–134 and 187–202.
- Grossberg, S. (1998). The link between brain, learning, attention, and consciousness, *Relatório técnico*, Boston University, Boston, MA, USA.
- Guha, S., Rastogi, R. e Shim, K. (1998). CURE: An Efficient Clustering Algorithm for Large Databases, *Proceedings of the ACM SIGMOD Conference*, ACM Press, pp. 73–84.
- Halkidi, M., Batistakis, Y. e Vazirgiannis, M. (2001). On clustering validation techniques, *Journal of Intelligent Information Systems* **17**(2/3): 107–145.
- Hammond, P. M. (1995). The current magnitude of biodiversity, in V. H. Heywood (ed.), *Global biodiversity assessment*, Cambridge University Press, pp. 113–128.
- Hara, F. A. S., Oliveira, L. A., Hungria, M. e Ferreira, M. C. (2005). Caracterização genética de rizóbios tolerantes à acidez e ao alumínio, nativos de solos

da Amazônia, como subsídio para estudos de ecologia da fixação biológica do N₂, *VIII Congresso de Ecologia do Brasil* .

Heins, L. G. e Tauritz, D. R. (1995). Adaptive Resonance Theory (ART): An Introduction, *Relatório técnico*, Leiden University.

Hon, M. J. L. (2002). The C Clustering Library for cDNA microarray data. Disponível em:<<http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster>>. Acesso em: 07 out. 2006.

Jain, A. K., Duin, R. P. W. e Mao, J. (2000). Statistical pattern recognition: A review, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(1): 4–37.

Jianguo, L. e Jingdong, W. (2002). Fuzzy ART based Texture Classification, *Relatório técnico*, Department of Automation, TsingHua University, Beijing, China.

Jordan, D. C. (1938). Rhizobiaceae conn, in K. N. R. e H. J. G. (eds), *Bergey's manual of systematic bacteriology*, Vol. 1, The Williams and Wilkins, pp. 235–244.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C., Silverman, R. e Wu, A. Y. (2002). An Efficient k-means clustering algorithm: Analysis and Implementation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(7): 881–892.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps, *Biological Cybernetics* **43**(1): 59–69.

Kondadadi, R. e Kozma, R. (2002). A modified fuzzy ART for soft document clustering, *Proceedings of the 2002 International Joint Conference on Neural Networks*, Honolulu, HI, USA, pp. 2545–2549.

Kruskal, J. B. e Wish, M. (1981). *Multidimensional Scaling*, Sage Publications.

-
- Laguerre, G., Mazurier, S. I. e Amarger, N. (1992). Plasmid profiles and restriction fragment length polymorphism of *Rhizobium leguminosarum* bv. *viciae* in field populations, *FEMS Microbiology Ecology* **101**(1): 17–26.
- Laguerre, G., R., A. M. e F., R. (1994). Rapid identification of rhizobia by restriction fragment length polymorphism analysis of PCR-amplified 16S genes, *Applied and Environment Microbiology* **60**(1): 56–63.
- Lang, K. J., Waibel, A. H. e Hilton, G. E. (1990). A time-delay neural network architecture for isolated word recognition, *Neural Networks* **3**(1): 23–43.
- LeCun, Y., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. e Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition, *Neural Computation* **1**(4): 541–551.
- Lewin, B. (2001). *Genes VII*, Artmed Editora, Porto Alegre.
- Llerena, S. E., Maciel, C. D. e Boaventura, M. (2007). Biological data representation by dendograms and multidimensional scalling, *Proceedings of the 6th Brazilian Conference on Dynamics, Control and Their Applications* .
- Lowe, D. e Webb, A. R. (1991). Optimized Feature Extraction and the Bayes Decision in Feed-Forward Classifier Networks, *IEEE Transactions Pattern Analysis and Machine Intelligence* **13**(4): 355–264.
- Ludwig, W. e Schleifer, K. H. (1994). Bacterial phylogeny based on 16S and 23S rRNA sequence analysis, *FEMS Microbiology* **15**(1): 155–173.
- Meyers, N. (1996). Environmental services of biodiversity, *Proceedings of the National Academy of Sciences* **93**(7): 2764–2769.
- Milagre, S. T. (2003). *Análise de Estabilidade de Cluster em uma Coleção Brasileira de Bactérias Diazotróficas do Gênero Bradyrhizobium*, Tese de Mestrado, Universidade Estadual de Londrina.
- Mitchell, T. M. (1997). *Machine Learning*, first ed, MIT Press and The McGraw-Hill Companies, Inc., Singapore.

- Newman, D. J., Hettich, S., Blake, C. L. e Merz, C. J. (2007). UCI Repository of machine learning databases. Disponível em:<<http://www.ics.uci.edu/~mllearn/MLRepository.html>>. Acesso em: 22 ago. 2006.
- Olsen, G. J. e Woese, C. R. (1993). Ribosomal RNA: a key to phylogeny, *The Federation of American Societies for Experimental Biology Journal* **7**(1): 113–123.
- Oyaizu, H. e Gamou, N. T. (1992). Molecular methods of analysing bacterial diversity: the case of rhizobia, *Biodiversity and Conservation* pp. 237–249.
- Pedini, D. e Gaudiano, P. (1994). A collection of ART-family graphical simulators, *Relatório técnico*, Boston University, Dept. of Cognitive and Neural Systems, 111 Cummington Street, Boston, MA 02215.
- Qian, J., Kwon, S. W. e Parker, M. A. (2003). rRNA and nifD phylogeny of Bradyrhizobium from sites across the Pacific Basin, *FEMS Microbiology Letters* **219**(1): 159–165.
- Qian, X., Wang, Z. e Wang, Y. (2005). A method of data clustering based on improved algorithm of ART2, *Machine Learning and Cybernetics* **4**(1): 2021–2026.
- Rossuma, D. V. e Schuurmans, F. P. (1995). Genetic and phenetic analyses of Bradyrhizobium strains nodulating peanut (Arachis hypogaea L.) roots, *Applied and Environment Microbiology* **61**(4): 1599–1609.
- Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B. e Erlich, H. A. (1988). Primer-Directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase, *Science* **239**(4839): 487–491.
- Sawada, H., Kuykendall, L. D. e Young, J. M. (2003). Changing concepts in the systematics of bacterial nitrogen-fixing legume symbiotics, *Journal of General and Applied Microbiology* **49**(3): 155–179.

-
- Schimel, J. (1995). Ecosystem consequences of microbial diversity and community structure, in F. S. Chapin e C. Körner (eds), *Arctic and Alpine Biodiversity: Patterns, Causes, and Ecosystem Consequences*, Springer, pp. 239–254.
- Senger, L. J. (2004). *Escalonamento de processos: uma abordagem dinâmica e incremental para a exploração de características de aplicações paralelas*, Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação - ICMC-USP.
- Siegler, R. S. (1976). Three aspects of cognitive development, *Cognitive Psychology* **8**(1): 481–520.
- Sneath, P. H. e Sokal, R. R. (1973). *Numerical taxonomy: The principles and practice of numerical classification*, Prentice Hall International Inc, San Francisco.
- So, R. B., Ladha, J. K. e Young, J. P. (1994). Photosynthetic symbionts of *Aeschynomene* spp. form a cluster with bradyrhizobia on the basis of fatty acid and rRNA analyses, *International Journal of Systematic Bacteriology* **44**(1): 392–403.
- Sokal, R. R. e Michener, C. D. (1958). A statistical method for evaluating systematic relationships, *Univ. Kansas Sci. Bull.* **38**(1): 1409–1438.
- Spiegler, I. (2006). Representation and Interpretation of Clustering Techniques, *Relatório técnico*, Tel Aviv University.
- Sá, N. M. H. e Vargas, M. A. T. (1997). Fixação biológica do nitrogênio por leguminosas forrageiras, *Biologia do solos de Cerrados*, pp. 127 – 152.
- Talavera, E. V. (2007). *Métodos Bayesianos Aplicados em Taxonomia Molecular*, Tese de Mestrado, Escola de Engenharia de São Carlos.
- Thórsson, A. T. e Jónsson, K. A. (2000). Genotyping Icelandic isolates of rhizobia based on rDNA-RFLP, *Icelandic Agricultural Science* **13**(1): 17–25.

- Urtz, B. E. e Elkan, G. H. (1996). Genetic diversity among Bradyrhizobium isolates that effectively nodulate peanut (*Arachis hypogaea*), *Canadian Journal of Microbiology* **42**(1): 1121–1130.
- Valluru, B. R. (1995). *C++ Neural Networks and Fuzzy Logic*, IDG Books Worldwide, Inc.
- Vinuesa, P., de Bruijn, J. L. W. e Werner, D. (1998). Genotypic characterization of Bradyrhizobium strains nodulating endemic woody legumes of the Canary Islands by PCR-restriction fragment length polymorphism analysis of genes encoding 16S rRNA (16S rDNA) and 16S-23S rDNA intergenic spacers, repetitive extragenic palindromic PCR genomic fingerprinting, and partial 16S rDNA sequencing, *Applied and Environment Microbiology* **64**(1): 2096–2104.
- Ward, J. H. (1963). Hierarchical grouping to optimize an objective function., *Journal of the American Statistical Association* **58**: 236–244.
- Westermeier, R. (2005). *Electrophoresis in Practice: A Guide to Methods and Applications of DNA and Protein Separations*, fourth ed, Wiley-VCH.
- Whiteley, J. R., Davis, J. F., Mehrotra, A. e Ahalt, S. C. (1996). Observations and Problems Applying ART2 for Dynamic Sensor Pattern Interpretation, *IEEE Transactions on Systems, Man and Cybernetics-Part A: Systems and Humans* **26**(4): 423–437.
- Williamson, D. I. (2006). Hybridization in the evolution of animal form and life-cycle, *Zoological Journal of the Linnean Society* **148**(4): 585–602.
- Young, J. P. W., Downer, H. L. e Eardly, B. D. (1991). Phylogeny of the phototropic rhizobium strain BTAi1 by polymerase chain reaction-based sequencing of a 16S rRNA gene segment, *Journal of Bacteriology* **173**(1): 2271–2277.