Pontifícia Universidade Católica
do Rio de Janeiro

Felipe Raposo Passos de Mansoldo

Identificação e Rastreamento Epidemiológico de Bactérias: Desenvolvimento de Sistema Web e Avaliação de Métodos Inteligentes

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para obtenção do título de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica da PUC-Rio.

Orientadora: Profª. Marley Maria Bernardes Rebuzzi Vellasco

Rio de Janeiro
Abril de 2012

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Felipe Raposo Passos de Mansoldo

Identificação e Rastreamento Epidemiológico de Bactérias: Desenvolvimento de Sistema Web e Avaliação de Métodos Inteligentes

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica do Centro Técnico Científico da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Profa. Marley Maria Bernardes Rebuzzi Vellasco
Orientadora
Departamento de Engenharia Elétrica – PUC-RJ

Profa. Karla Tereza Figueiredo Leite
UEZO

Prof. André Vargas Abs da Cruz
DEE/PUC-Rio

Profa. Ana Luiza de Mattos Guaraldi
UERJ

Prof. José Eugênio Leal
Coordenador Setorial do Centro Técnico Científico – PUC-Rio

Rio de Janeiro, 20 de Abril de 2012

.

Felipe Raposo Passos de Mansoldo

Graduou-se em Engenharia de Controle e Automação pela Pontifícia Universidade Católica do Rio de Janeiro (PUC-RIO) em 2009.

Ficha Catalográfica

Mansoldo, Felipe Raposo Passos de

Identificação e rastreamento epidemiológico de bactérias : desenvolvimento de sistema web e avaliação de métodos inteligentes / Felipe Raposo Passos de Mansoldo ; orientadora: Marley Maria Bernardes Rebuzzi Vellasco. – 2012.
138 f. : il. (color.) ; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2012.
Inclui bibliografia

1. Engenharia elétrica – Teses. 2. Bioinformática. 3. Identificação de bactérias. 4. Classificação de bactérias. 5. Redes neurais artificiais. 6. Mapas auto-organizaveis. I. Vellasco, Marley. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. III. Título.

CDD: 621.3

Aos meus pais Cezar Augusto Mansoldo e Marise Raposo Passos de Mansoldo,
Ao meu irmão Marcelo Raposo Passos de Mansoldo.

Agradecimentos

Aos meus pais, pela educação, carinho e estímulo. Ao meu irmão pelo companheirismo.

À minha orientadora, Professora Marley, pela confiança, dedicação, apoio e compreensão nos momentos mais difíceis.

Aos professores que participaram da Comissão examinadora. À professora Ana Guaraldi, pela paciência e carinho nas explicações sobre seu trabalho.

À professora Maria Isabel Pais da Silva, e à amiga e principal incentivadora Dr. Sonia Letichevsky.

Aos fiéis amigos André Gosling, Felipe Vilaça, Lucas Dias, José Carlos de Carvalho Dias e Victor Passos Ferreira. Ao Yuri Vieira e Lincoln Sant'Anna pela ajuda e apoio no laboratório.

A todos os professores e funcionários da Pontifícia Universidade Católica de Rio de Janeiro. Á FAPERJ pelo apoio financeiro.

À J.S. Bach e Baden Powell pelos momentos de inspiração.

Resumo

Mansoldo, Felipe Raposo Passos de; Vellasco, Marley Maria Bernardes Rebuzzi. **Identificação e rastreamento epidemiológico de bactérias: desenvolvimento de sistema web e avaliação de métodos inteligentes.** Rio de Janeiro, 2012. 138p. Dissertação de Mestrado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

A maioria dos laboratórios não conta com um sistema informatizado para gestão dos procedimentos pertinentes a cada caso. A administração e controle das amostras é feito manualmente, através de diversas fichas que são preenchidas desde o colhimento do material biológico, no hospital, até a identificação final da bactéria no laboratório. Dessa forma, a organização das informações fica limitada, uma vez que, estando as informações escritas à mão e guardadas em livros, é quase impossível a extração de conhecimento útil que possa servir não só no apoio à decisão, como também, na formulação de simples estatísticas. Esta dissertação teve dois objetivos principais. O desenvolvimento de um sistema *Web*, intitulado BCIWeb (*Bacterial Classification and Identification for Web*), que fosse capaz de auxiliar na identificação bacteriológica e prover a tecnologia necessária para a administração e controle de amostras clínicas oriundas de hospitais. E a descoberta de conhecimento na base de dados do sistema, através da mineração de dados utilizando os métodos de Mapas Auto-Organizáveis (SOM: *Self-Organizing Maps*) e Redes *Multilayer Perceptrons* (MLP) para classificação e identificação de bactérias. A partir do desenvolvimento desta ferramenta amigável, no estudo de caso, os dados históricos do LDCIC (Laboratório de Difteria e Corinebactérias de Importância Clínica) do Departamento de Biologia da UERJ foram inseridos no sistema. Os métodos inteligentes propostos para classificação e identificação de bactérias foram analisados e apresentaram resultados promissores na área.

Palavras-chave:

Bioinformática; identificação de bactérias; classificação de bactérias; redes neurais artificiais; mapas auto-organizáveis.

Abstract

Mansoldo, Felipe Raposo Passos de; Vellasco, Marley Maria Bernardes Rebuzzi. (Advisor) **Identification and Epidemiological Surveillance of Bacteria: Web System Development and Evaluation of Intelligent Methods.** Rio de Janeiro, 2012. 138p. MSc Dissertation - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro..

Most laboratories do not have a computerized system for management procedures. The administration and control of the samples are made manualy through many forms of data sheets which are filled from the beginning, when the samples of biological materials are gathered at the hospital, up to the final identification at the laboratory. In this context, the organization of the information become very limited, while the information writting by hands and stored in books, its almost impossible to extract useful knowledge, which could help not only supporting decisions but also in the formulations of simples statistics. This thesis had two objectives. The development of a web system called BCIWeb (Bacterial Classifiation and Identification for Web) that could assist in bacterial identification and provide the technology necessary for the administration and control of clinical specimen coming from the hospitals and the discovery of knowledge in database system, through data mining methods using SOM (Self Organizing Maps) and Multilayer Perceptron Neural Networks (MLP) for classification and identificatin of bactéria. From the development of this friendly tool, in the case study, the historical data from LDCIC (Laboratório de Difteria e Corinebactérias de Importância Clínica) of UERJ Biology Department were entered into the system. The proposed intelligent methods for classification and identification of bacteria were analysed and showed promising results.

Keywords

Bioinformatics; identification of bacteria; bacterial classification; artificial neural networks; self-organizing map

Sumário

Lista de figuras

Lista de tabelas

Lista de siglas

AJAX *(Asynchronous Javascript and XML)*

BP (*Back-Propagation*)

BCIWeb (*Bacterial Classification and Identification for Web*)

BGPIs (Bastonetes Gram-positivos Irregulares)

CMS (*Content Management System*)

DCBD (Descoberta de Conhecimento em Bases de Dados)

DNA (*deoxyribonucleic acid;* em português: ADN, ácido desoxirribonucleico)

EQM (Erro Quadrático Médio)

E-R (Entidade-Relacionamento)

QE (*Quantization Error;* em português: Erro de Quantização)

TE (*Topographic Error;* em portguês: Erro Topográfico)

GenBank (*Genetic Sequence Database*)

GNU LGPL (*GNU Lesser General Public License*)

IA (Inteligência Artificial)

Identax (*IDENTAX Bacterial Identifier*)

JSON (*JavaScript Object Notation*)

KDD (*Knowledge Discovery in Databases*)

LDCIC (Laboratório de Difteria e Corinebactérias de Importância Clínica)

MLP (Multilayers Perceptron)

MSE (Mean Squared Error)

PCR (*Polymerase Chain Reaction*; em português: Reação em Cadeia da Polimerase)

PHP (Pré-Processador de Hipertexto)

PIBWin (*Probabilistic Identification of Bacteria for Windows*)

Retroprogação do Erro (*Backpropagation*)

RNA (Redes Neurais Artificiais)

SGBD (Sistema de Gerenciamento de Banco de Dados)

SNC (Sistema Nervoso Central)

SOM (Self-Organizing Maps)

SQL (*Structured Query Language*)

1Introdução

1.1. Motivação

Microrganismos do gênero Corynebacterium vêm sendo citados com crescente frequência como importantes patógenos de infecções. O aparecimento de amostras multiresistentes e o aumento do número de casos, por vezes culminando em óbito, têm contribuido para aumentar o interesse pelas corinebactérias (FUNKE, 1999) (JANDA,1998) (JANDA, 1999) (RIEGEL, 1996) (WILLIAMS, 1993). Recentes avanços na identificação de espécies deste gênero mostram que existe uma complexidade taxonômica considerável e que os métodos de identificação clássicos podem gerar resultados ambíguos (ADDERSON, 2008) (JANDA, 2002) (PASCUAL, 1995).

Com o avanço da tecnologia e a redução dos custos na área de biologia molecular, uma grande quantidade de informações tem sido armazenada em bancos de dados. No entanto, devido à complexidade desses dados, é necessária a integração de várias áreas do conhecimento para que se possa extrair conhecimento útil dessas bases de dados (BOGUSKI, 1998) (BOGUSKI, 1994).

As técnicas tradicionais de análise de dados não têm sido suficientes para a cobertura total do problema, seja pelo alto custo ou pelo tempo de execução. Deste modo, existe um aumento no interesse pelo desenvolvimento e aplicação de sistemas computacionais inteligentes (LEE, 2009) (JENA, 2009) (NARAYANAN, 2003).

Um método computacional inteligente de grande destaque são as redes neurais artificiais (RNA) (HAYKIN, 2008). Através deste método é possível simular o aprendizado, percepção, raciocínio e adaptação do cérebro humano, fazendo com que sejam muito utilizadas na construção de sistemas de suporte à decisão, classificação, previsão, entre outros, nas mais diversas áreas do conhecimento (HAYKIN, 2008) (RIPLEY, 2008) (KOHONEN, 1982). Estudos recentes demonstraram que RNAs são poderosas ferramentas na identificação de bactérias (AHMAD, 2008) (SAHIN, 2006) (MARIEY, 2001) (GARRIT, 2001).

O processo de identificação de bactérias deve ser rápido e preciso afim de que seja aplicado o tratamento adequado ao paciente. Porém, poucos laboratórios possuem recursos materiais e profissionais para análise fenotípica e/ou genotípica das amostras colhetadas (GUARALDI, 2010) (DAMASCO, 2005). Além disso, na maioria dos laboratórios o controle das amostras, as quais compreendem informações dos pacientes, históricos médico, resultados das provas bioquímicas e antibiograma, é realizado à mão, por fichas preenchidas em diversos setores do hospital e depois guardadas em livros.

A precariedade no controle e registro das informações médicas e biológicas aponta para a necessidade do estudo e desenvolvimento de sistemas e métodos inteligentes que sirvam de suporte para a área.

1.2. Objetivos

Deste modo, a principal meta deste trabalho foi desenvolver um sistema *web* que fosse capaz de dar suporte ao Laboratório de Difteria e Corinebactérias de Importância Clínica (LDCIC) no registro e controle das amostras que são analisadas, bem como prover métodos computacionais inteligentes que auxiliem na classificação e identificação dos micorganismos.

Dois objetivos principais podem então ser destacados: i) o desenvolvimento de um sistema genérico, em plataforma *Web*, capaz de auxiliar na identificação bacteriológica e prover a tecnologia necessária para a administração e controle de amostras clínicas oriundas de hospitais; ii) a descoberta de conhecimento na base de dados do sistema, através da mineração de dados utilizando métodos inteligentes. A seguir são detalhados os principais módulos desenvolvidos neste trabalho, de forma a alcançar os objetivos acima descritos:

- *Desenvolvimento de Banco de Dados*: Modelagem de uma base de dados para armazenamento e manipulação de grandes volumes de informações biológicas de forma rápida e eficiente.
- *Desenvolvimento de Modelos de Classificação*: Tradicionalmente a classificação de microrganismos é feita através de uma série de testes

que podem ser caros, demorados, ou produzirem resultados que dependam de interpretações subjetivas, como cores etc. Mapas Auto-Organizáveis (SOM: *Self-Organizing Maps*) foram aplicados no estudo dos atributos (testes biológicos) que são relevantes na formação de grupos de microrganismos. A investigação das relações entre os atributos através de métodos inteligentes é importante na definiçao do comportamento de cada grupo.

- *Desenvolvimento de Modelos de Identificação*: A identificação de espécies de bactérias através de marcadores fenotípicos é muito difícil devido à complexidade taxonômica e à variabilidade de resultados que podem existir para uma espécie. Diversas configurações de Redes *Multilayer Perceptrons* (MLP) foram treinadas com resultados históricos de testes biológicos, a fim de se obter a rede com melhor aprendizado e capacidade de generalização dos padrões biológicos de cada espécie.
- *Desenvolvimento do sistema web*: Criação de um sistema *online* de interface amigável integrado à base de dados que permita aos usuários registrar e controlar as amostras do laboratório. Para este sistema foram criados módulos personalizados como árvores de decisão, integração com *Google maps*, entre outros.

1.2.1. Organização da Dissertação

Esta dissertação está organizada em cinco capítulos adicionais, descritos a seguir.

O capítulo 2 descreve o processo de descoberta de conhecimento em bases de dados e apresenta uma breve introdução às técnicas de inteligência computacional, destacando redes neurais supervisionadas e mapas auto-organizáveis. Todos os conceitos básicos para o entendimento dessas técnicas são descritos neste capítulo.

No capítulo 3 é introduzido o conceito de bioinformática, seu histórico e principais bancos de dados mundiais. É feita, também, uma breve explicação

sobre bactérias e os métodos clássicos na sua identificação. Logo são apresentados mais detalhes sobre o gênero Corynebacterium pois este é o alvo do estudo de caso desta dissertação. Na seçao 3.4 é feita uma breve retrospectiva dos estudos e métodos relevantes na área biológica, computacional e de bioinformática. Os principais programas usados na identificação de bactérias são também descritos. Ao final é discutido o atual estado da arte do uso de redes neurais artificiais nos problemas taxonômicos.

No capítulo 4 o sistema BCIWeb é descrito detalhadamente. No início são enumeradas e explicadas todas as tecnologias usadas, seguido pela modelagem proposta para a base de dados levando-se em consideração os pré-requisitos do LDCIC. O capítulo termina apresentando o procedimento de uso do sistema e suas funcionalidades.

O capítulo 5 apresenta o estudo de caso do LDCIC da UERJ, onde se utiliza técnicas de inteligência computacional na base de dados do sistema BCIWeb.

No capítulo 6 estão descritas as conclusões e os possíveis trabalhos futuros.

2 Descoberta de Conhecimento em Bases de Dados

2.1. Introdução

Com a redução dos custos em dispositivos de armazenamento e o aumento da capacidade e velocidade dos sistemas de computação o interesse na descoberta de conhecimento das bases de dados aumentou. Nota-se que é crescente a quantidade de informações acumuladas. No entanto, essas informações armazenadas, na maioria das vezes, não significam de imediato algum conhecimento relevante.

O processo de DCBD (Descoberta de Conhecimento em Bases de Dados), ou KDD (*Knowledge Discovery in Databases*) é o processo de exploração da base de dados em busca da descoberta de novos conhecimentos. Trata-se de um processo não trivial de identificação válida, nova, potencialmente útil e compreensível de padrões de dados (FAYYAD,1996). Ou seja, não se trata de simplesmente encontrar padrões entre os dados, mas sim, a extração de conhecimento inteligível e utilizável para o apoio a decisões.

A não trivialidade se deve às condições em que se encontram as informações nas bases de dados. Pode ser que haja, em grande volume, ruídos, informações incompletas, faltosas, fora de escala e incorretas. Dada essa diversidade de fatores, é necessário uma etapa de pré-processamento e limpeza dos dados, afim de que se tenha dados úteis para o processo. Portanto, a DCBD é composta de várias etapas para que se tenha ao final um conhecimento válido, cada etapa está ilustrada na Figura 2.1-1.

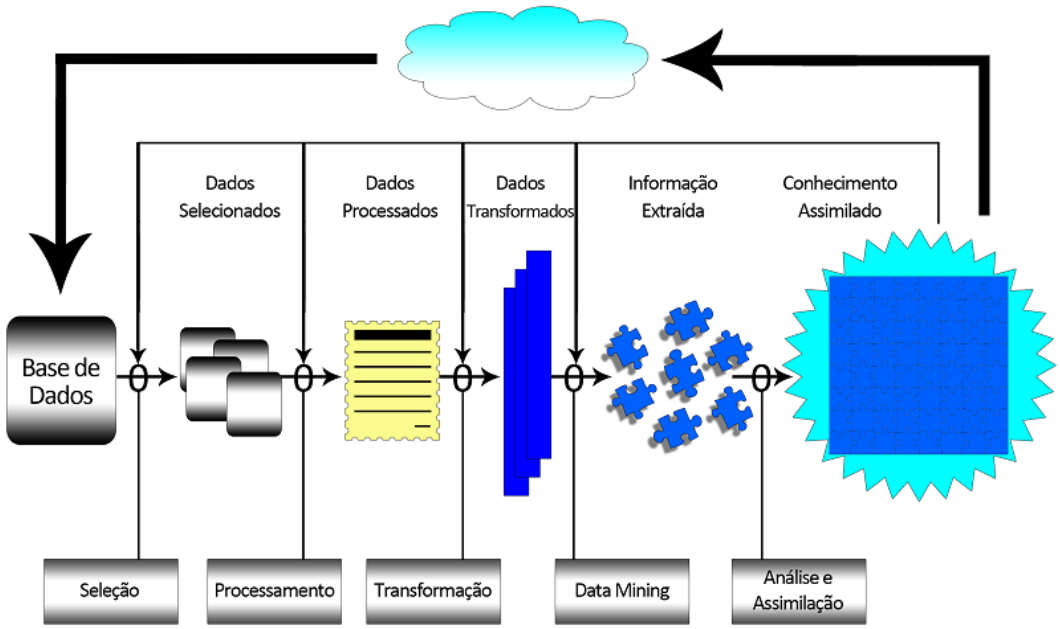


Figura 2.1-1 Etapas do processo de descoberta de conhecimento em bases de dados. Fonte: (Adaptação de GARCIA, 2003)

Conforme apresentado na Figura 2.1-1, o processo de DCBD é iterativo, envolvendo várias etapas que são executadas sequencialmente, sendo muitas vezes, necessário o retorno a etapas anteriores para a realização de ajustes, até que se tenha o resultado esperado (FAYYAD,1996) (CABENA,1997).

A seguir, são detalhadas cada uma das etapas do processo de Descoberta de Conhecimento em Bases de Dados, a saber: Determinação dos objetivos; Preparação dos dados; Mineração dos dados; Análise dos resultados; e Utilização do conhecimento.

2.1.1. Determinação dos objetivos

A determinação dos objetivos é a primeira e importante etapa do processo, onde são reconhecidos os problemas a se elucidar. Nesta etapa se tem uma avaliação das informações da base de dados e as possibilidades do que se pode obter, a fim de que seja criado um modelo que auxilie no suporte à decisão.

2.1.2. Preparação dos dados

Na etapa de preparação dos dados se cria uma representação dos mesmos que seja compatível com os objetivos da análise realizada e que se encaixe na resolução do problema em questão. Esta etapa consiste de alguns métodos de preparo dos dados como a consulta à base de dados, limpeza e tratamento de erros, determinação de valores fora dos domínios, inconsistências, transformação dos dados como normalizações etc. Pode-se organizar em três fases:

- *Seleção dos dados*: Os dados podem ser divididos em dois tipos principais: categóricos ou quantitativos. De acordo com os objetivos estabelecidos os dados são identificados, se necessário manipulados e seus domínios analisados.
    - Categóricos: são valores finitos e nomeiam um tipo de objeto, por exemplo, o sexo "masculino" ou "feminino".
    - Quantitativos: podem ser valores contínuos (números reais) ou discretos (números inteiros).
- *Pré-processamento*: O objetivo desta etapa é assegurar a qualidade dos dados envolvidos através da verificação de cada um deles. Devem ser avaliados dados com ruídos e valores desconhecidos.
- *Transformação*: Os dados devem ser manipulados de acordo com as exigências do formato de entrada do algoritmo de mineração de dados utilizado no processo. As mais comuns são:
    - Resumo: Quando um conjunto de atributos são unidos formando um só.
    - Discretização: Dados contínuos são transformados em discretos.
    - Normalização: É aplicada em valores contínuos afim de que fiquem restritos a um determinado intervalor de valores.

2.1.3. Mineração dos dados

A Mineração dos dados é a etapa fundamental do DCBD, cujo objetivo é extrair padrões dos conjuntos de dados que foram pré-processados, através do ajuste dos modelos de algoritmos escolhidos até que se tenha o resultado esperado, ou seja, alcançado o objetivo de buscar informação útil na base de

dados. Essa seleção de algoritmos é feita nos objetivos estabelecidos no início do processo.

Existem diversos algoritmos que podem ser utilizados nesta etapa do processo, tais como: Redes Neurais Artificiais, Lógica Fuzzy, Árvores de Decisão, Algoritmos Genéricos etc. Entre os diversos tipos de algoritmos e modelos, a escolha do mais apropriado depende do tipo de problema a ser resolvido e das características dos dados. Os problemas podem ser divididos basicamente em:

- Classificação
- Predição
- Estimação
- Agrupamento

Para uma melhor compreensão dos principais algoritmos para a solução desses problemas a seção 2.2 apresenta uma breve descrição de cada um deles.

2.1.4. Análise dos resultados

Esta etapa envolve a validação dos resultados obtidos, onde normalmente são feitas algumas consultas aos solicitantes do processo para se ter alguma medida de qualidade e segurança nos resultados. O processo como um todo envolve diversas consultas a etapa de análise, e então retornando à etapa de mineração de dados para realizar ajustes em busca de resultados satisfatórios.

2.1.5. Utilização do conhecimento

Esta etapa representa o fim do processo. As novas descobertas são apresentadas e o conhecimento válido adquirido deve ser incoporado em um modelo que sirva de suporte à decisão para os problemas especificados inicialmente.

2.2. Sistemas Inteligentes

2.2.1. Redes Neurais Artificiais

Redes Neurais são modelos computacionais não lineares, inspirados na estrutura e operação do cérebro humano, que procuram reproduzir características humanas, tais como*: aprendizado, associação, generalização e abstração* (CYBENKO, 1996) (REZENDE, 2003, p. 142). As Redes Neurais são compostas por diversos elementos processadores (neurônios artificiais), altamente interconectados, formando uma rede de nós exatamente como na estrutura do cérebro, que efetuam operações simples, transmitindo seus resultados aos processadores vizinhos (FEITOSA**,** 1999). A habilidade das Redes Neurais em realizar mapeamentos não-lineares entre suas entradas e saídas as tem tornado prósperas no reconhecimento de padrões e na modelagem de sistemas complexos (VALIATI, 2006).

Redes Neurais Artificiais simulam o "cérebro biológico" nos aspectos estruturais e comportamentais. Portanto, é capaz de aprender através dos erros e realizar descobertas, adquirindo conhecimento através da experiência. A sua capacidade de aprendizado e generalização faz com que o sistema forneça respostas coerentes para entradas até então desconhecidas, o que torna as RNAs uma poderosa e interessante ferramenta computacional na solução de problemas (BRAGA et al., 2000).

As pesquisas em RNAs (Redes Neurais Artificiais) e Inteligência Artificial (IA) tiveram início com McCulloch & Pitts (1943). Psiquiatra e neuroanatomista Warren McCulloch e o matemático William Pitts descreveram uma fórmula matemática que unifica os estudos de neurofisiologia e lógica matemática. Foram os responsáveis pelo primeiro modelo de um neurônio artificial, este modelo apresenta a capacidade de interação entre os neurônios.

Em 1949 o psicólogo Donadl Hebb, em seu livro The *Organization Of Behavior* (HEBB, 1949), descreve o processo de aprendizado humano, e propôs o uso de pesos como forma de representar o conhecimento em RNAs. Em 1958,

Frank Rosenblatt seguindo as regras propostas por Hebb, criou um modelo de neurônio artificial chamado de *perceptron*, surgiu então uma RNA de uma camada com capacidade de aprender e identificar padrões. Esse modelo básico de neurônio artificial é o mesmo usado ainda nos dias de hoje, conforme na Figura 2.2-1 Modelo do neurônio artificial com sua estrutura e conexões, pode ser visto o modelo de neurônio artificial com suas conexões em funções.

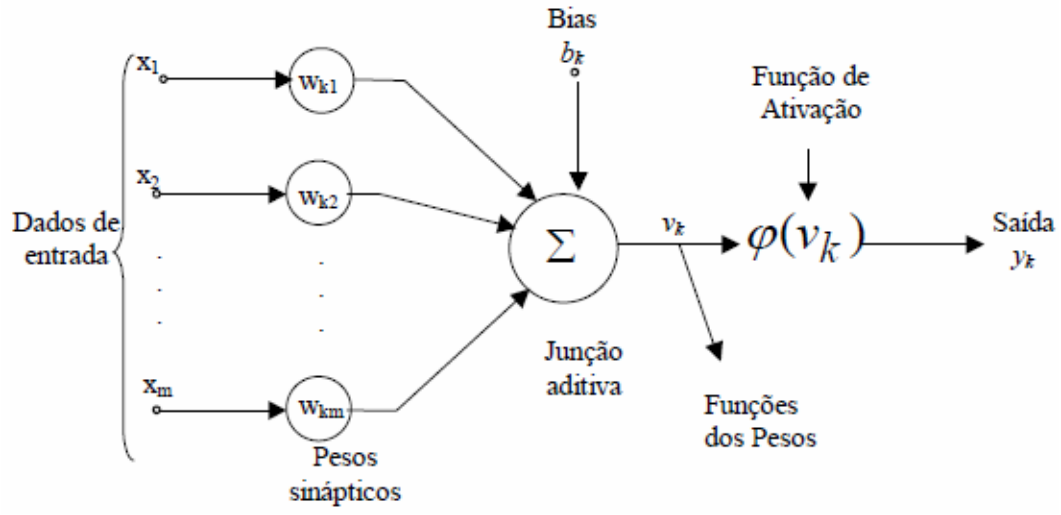


Figura 2.2-1 Modelo do neurônio artificial com sua estrutura e conexões

O neurônio apresentado na Figura 2.2-1, tipicamente denominado elemento processador, é inspirado no neurônio biológico, possuindo um conjunto de entradas x_m (dendritos) e uma saída y_k (axônio). As entradas são ponderadas por pesos sinápticos w_{km} (sinapses), que determinam o efeito da entrada x_m sobre o processador *k*. Estas entradas ponderadas são somadas, fornecendo o potencial interno do processador (net_k). A saída ou estado de ativação s_k do elemento processador *k* é finalmente calculada através de uma função de ativação *f*, tipicamente uma função sigmoidal. O estado de ativação pode então ser definido pela Equação 2.2-1:

$$y_k = \varphi \left(\sum_{m=1}^{N} x_m w_{km} + b_k \right) \qquad \text{Equação 2.2-1}$$

Onde: N é o número de entrada do neurônio k

b_k é o termo de polarização do neurônio k (bias)

φ é função de ativação e funciona restringindo a amplitude de saída de determinado neurônio e adicionando não-linearidade ao modelo.

2.2.1.1. Redes Multi-Layer Perceptron

A organização de *perceptrons* em mais de uma camada é chamada Multilayers Perceptron (MLP) e foi proposta por Rummelhart (Rummelhart, 1986). As RNAs do tipo MLP são formadas por conjuntos de neurônios dispostos em camada. Uma ou mais camadas intermediárias e uma camada de saída. Todos os neurônios de uma camada estão conectados a todos os neurônios da camada seguinte, desde a primeira camada intermediária até a camada de saída. Associado a cada conexão existe um peso sináptico, que modifica o sinal recebido através daquela conexão.

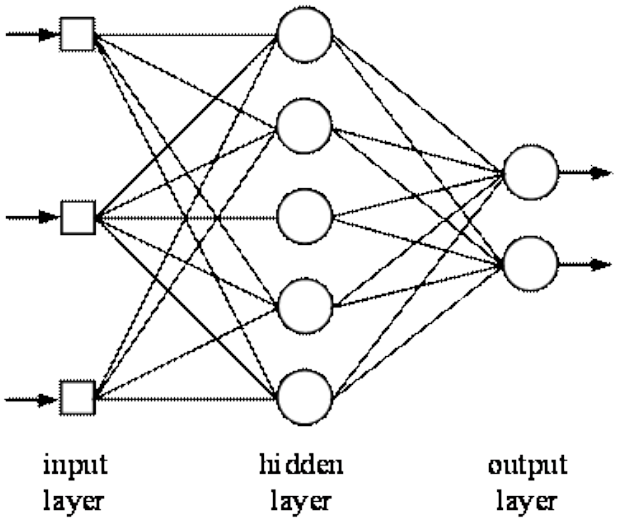


Figura 2.2-2 Multilayer perceptron com três neurônios na camada de entrada, cinco na camada intermediária e dois na camasa de saída

O processamento de uma Rede Neural pode ser dividido em duas fases: *Recuperação da Informação (recall)* e *Aprendizado* (*learning). A recuperação da Informação* é o processo de cálculo da saída da rede, dado um certo padrão de entrada. O *Aprendizado* é o processo de atualização dos pesos sinápticos para a aquisição do conhecimento. Existem três tipo básico de aprendizado: treinamento supervisionado, não-supervisionado e por reforço (FEITOSA**,** 1999).

Redes MLP são treinadas através de um processo supervisionado, o qual consiste, basicamente, na apresentação de um padrão de entrada e da saída desejada para aquele padrão. A saída produzida pela rede MLP em resposta àquele padrão é comparada com a saída desejada, ou seja, a saída correta. A diferença entre a resposta da rede e a esperada é calculada e, com base neste erro, os pesos sinápticos associados às conexões são ajustados, de forma a minimizar este erro. Este processo é repetido diversas vezes até que se chegue a um erro mínimo pré-estabelecido.

O algoritmo *Back-Propagation (BP)* (Haykins, 1994; Wassermann, 1993) define uma maneira sistemática de atualização dos pesos das diversas camadas da rede baseada na idéia que os erros dos neurônios das camadas escondidas são determinados pela retropropagação reversa dos erros dos neurônios da camada de saída.

Este tipo de treinamento supervisionado é baseado no método do gradiente decrescente (*gradient descent*), que busca minimizar o erro global da camada de saída. Deste modo, a atualização do peso (Δw_{km}) é proporcional ao negativo da derivada parcial do erro com relação ao próprio peso:

$$\Delta W_{km} = -\rho \frac{\partial \varepsilon}{\partial W_{km}} \qquad \text{Equação 2.2-2}$$

Onde ρ é a taxa de aprendizado;

ε é a função erro, definida como:

$$\varepsilon = \frac{1}{2} \sum_{m=1}^{N_0} t_k - s_k \ ^2 \qquad \text{Equação 2.2-3}$$

Onde N_0 é o número de processadores da camada de saída;

t_k é o valor esperado na saída do processador *k*;

s_k é o valor de saída do processador *k*.

O processo de atualização dos pesos sinápticos é repetido diversas vezes, até que, para todos os padrões de entrada a diferença entre a saída esperada e a da camada de saída tenha um erro menor do que o especificado.

Segundo Hornik (1989), o algoritmo *Back Propagation* é um aproximador universal, ou seja, é capaz de aprender qualquer mapeamento de entrada-saída. No entanto, pode apresentar alguns problemas básicos: a definição do tamanho da rede, o longo processo de treinamento, paralisia da rede e mínimo local.

2.2.1.2. O Cérebro, o computador e a inspiração biológica

Redes Neurais Artificiais são desenvolvidas como generalização de modelos matemáticos de neurônios biológicos resumindo-se em (FAUSETT, 1994, p.6):

- o processamento das informações ocorre por intermédio de elementos simples, chamados de neurônios;
- os sinais são passados entre os neurônios por meio de conexões;
- cada conexão tem um peso associado;
- para determinar o sinal de saída, cada neurônio aplica uma função de ativação à soma dos sinais de entrada ponderados pelos pesos sinápticos.

Dois pesquisadores realizaram estudos que foram fundamentais sobre o Sistema Nervoso Central (SNC) sua estrutura e funcionamento. Foram eles Santiago Ramón y Cajal (1852-1934) e Camillo Golgi (1843-1926), que receberam juntos em 1906 o Prêmio Nobel de Medicina.

Ramón y Cajal chegou à que conclusão que o sistema nervoso é composto por bilhões de neurônios distintos e que estas células se encontram polarizadas, sugeriu também que os neurônios comunicam-se através de ligações especializadas chamadas sinapses, que são as responsáveis pela memorização da informação (GOELZER, 2007).

A Figura 2.2-3 A estrutura do neurônio biológico e os locais onde ocorrem as sinpases. Fonte: (FAUSETT, 1994). ilustra as principais partes de um neurônio biológico.

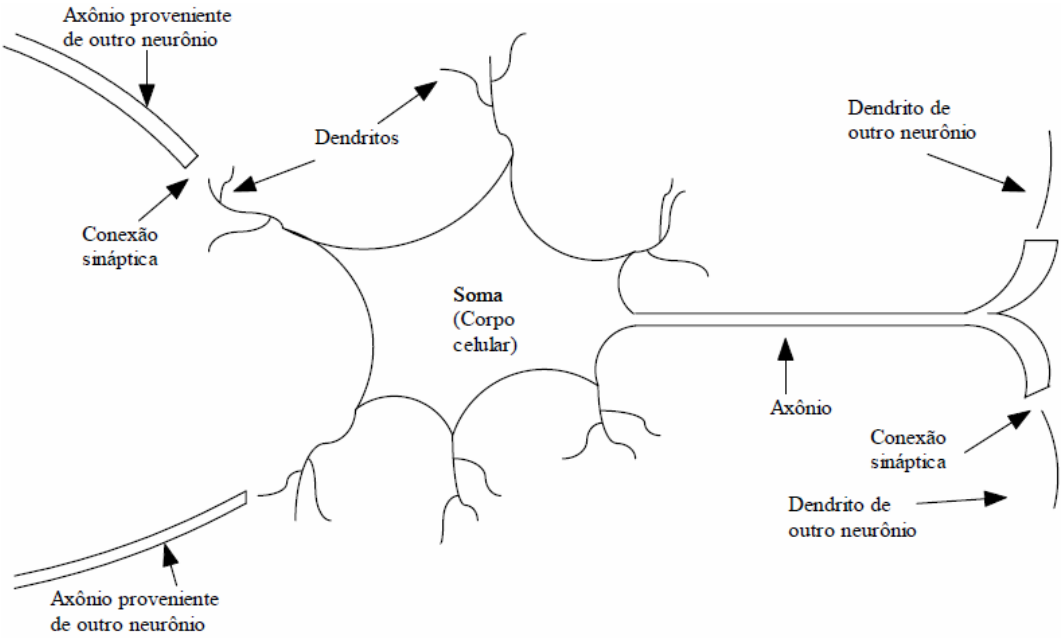


Figura 2.2-3 A estrutura do neurônio biológico e os locais onde ocorrem as sinpases. Fonte: (FAUSETT, 1994).

As principais partes do neurônio biológico são:

- corpo celular é a parte central do neurônio ou soma, é responsável pela geração dos impulsos nervosos;
- dentritos são ramificações que partem do soma e são responsáveis por receber os sinais externos;
- axônio é responsável pela propagação de impulsos nervosos que partem do corpo celular.

2.2.1.3. Vantagens e Desvantagens do seu uso

Qualquer modelo matemático está sujeito vantagens e limitações. Como vantagens das RNAs pode-se citar:

- *Acurácia*: Os resultados obtidos são geralmente superiores aos obtidos com técnicas estatísticas;
- *Auto-aprendizado*: Dispensa conhecimento de especialistas para tomar decisões, o aprendizado provém dos exemplos históricos;
- *Robustez*: As unidades de processamento (neurônios) operam em paralelo, portanto, a falha de algumas unidades não torna toda a rede inoperante;
- *Tolerância a ruídos*: As RNAs têm a capacidade de extrair a essência dos padrões de entrada, sendo imunes aos ruídos inerentes aos dados.

As desvantagens das RNAs:

- *Caixa-preta*: Após o treinamento de uma determinada rede é difícil extrair as regras que representam o conhecimento adquirido, ou seja, a justificativa de uma decisão tomada;
- *Volume de dados*: Para o correto aprendizado das RNAs, é necessário um grande volume de dados históricos com o menor número de dados faltosos possível.

2.2.2. Agrupamento por Mapas Auto-Organizáveis - SOM

Mapas Auto-Organizáveis ou Self-Organizing Maps (SOM), desenvolvido por (KOHONEN, 1982) são um tipo particular de RNA de treinando do tipo competitivo e não-supervisionado. O modelo, tem como dados de entrada vetores e como saída a organização dos vetores em agrupamentos (*clusters*) (KOHONEN, 1989). SOM realiza uma projeção não-linear de dados de alta dimensão em um mapa discreto usualmene de duas dimensões. Portanto, vetores de entrada de padrões similares serão mapeados em regiões espacialmente próximas no mapa de saída. Grupos que eram desconhecidos inicialmente, após o processo não-supervisionado de treinamento, se agrupam em regiões específicas do mapa. Desta forma, SOM é uma ferramenta para visualização e exploração de dados em alta dimensão. Um exemplo de arquitetura SOM está ilustrada na Figura 2.2-4.

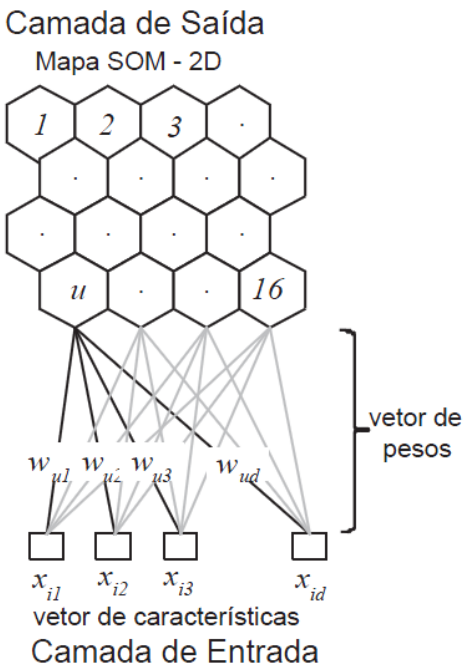


Figura 2.2-4 Arquitetura de uma rede SOM com o vetor de entrada abaixo e o mapa de saída acima. Configuração com 16 neurônios na camada de saída. Fonte: (SILVA, 2009)

Conforme Figura 2.2-4, cada unidade *u* (neurônio) é associada a um vetor de pesos $\mathbf{w}_u = (\mathbf{w}_{u1}, \mathbf{w}_{u2}, ..., \mathbf{w}_{ud})$ que tem dimensão *d*, como padrão de entrada, sendo $\mathbf{x}_{id}$ vetores de características de entrada . O SOM é definido por um conjunto de neurônios dispostos em um arranjo que define a vizinhança de cada neurônio, conforme Figura 2.2-5.

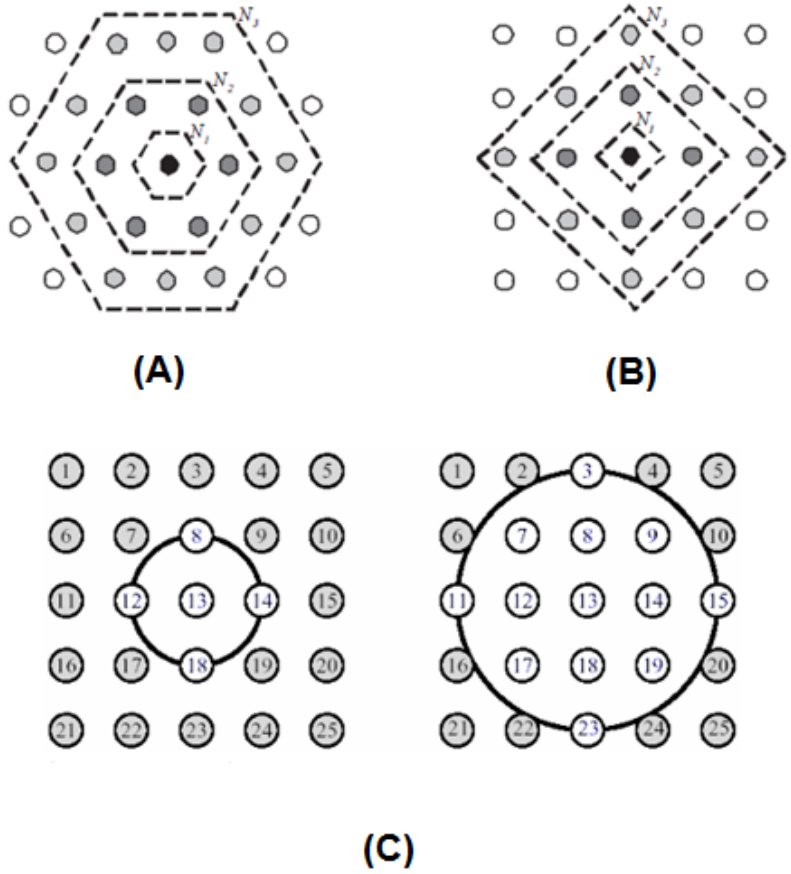


Figura 2.2-5 Exemplos de vizinhanças no mapa SOM. A vizinhança define a relação entre os neurônios, que pode ser (A) Hexagonal, (B) Retangular ou (C) Circular. Fonte: (Adaptação de MATHWORKS, 2005).

O processo de geração do mapa auto-organizável pode ser separado em três etapas (HAYKIN, 2001, p487):

- *Competição*: Para cada padrão de entrada, os neurônios da grade calculam seus respectivos valores de uma função discriminante.

Essa função discriminante fornece a base para a competição entre os neurônios. O neurônio com o maior valor da função discriminante é declarado vencedor da competição;

- *Cooperação*: O neurônio vencedor determina a localização espacial de uma vizinhança topolófica de neurônios excitados, fornecendo assim a base para a cooperação entre os neurônios vizinhos;
- *Adaptação Sináptica*: Este último mecanismo permite que os neurônios excitados aumentem seus valores individuais da função discriminante em relação ao padrão de entrada através de ajustes adequados aplicados a seus pesos sinápticos. Os ajustes feitos são tais que a resposta do neurônio vencedor à aplicação subsequente de um padrão de entrada similar é melhorada.

Os vetores de peso são iniciados aleatoriamente, com valores no domínio dos padrões de entrada ou por outros métodos mais sofisticados (KOHONEN, 2001), assim, segundo Haykin (2001), nenhuma organização prévia é imposta ao mapa de características. SOM são treinadas através de um processo não-supervisionado, que se inicia com o uso de algum algoritmo otimizador ou seleção aleatória de um vetor de características do conjunto de treinamento. A ordem de apresentação dos padrões de entrada para a rede interferem no resulado final do mapeamento, variações na sequência de apresentação podem gerar mapeamentos topologicamente incorretos (MIRANDA, 1998).

O neurônio vencedor (BMU de *best-matching unit*) denotado como *c*, é determinado através do menor valor da equação de distância entre seu vetor de pesos $\mathbf{w}_u$ e $\mathbf{x}_i$, conforme Equação 2.2-4.

$$c = argmin \quad x_i - w_u \qquad \text{Equação 2.2-4}$$

Portanto, a unidade *c* é a melhor representação de $\mathbf{x}_i$. Este calculo é feito para todos os padrões de entrada durante todo o processo de treinamento. As atualizações dos vetores de pesos são controladas pelo parâmetro de taxa de aprendizado α, que normalmente diminui de valor em função do tempo.

Com o objetivo de preservação das relações de similaridade entre os vetores de características e as unidade do mapa, o neurônio vencedor não é o único a ter seus pesos sinápticos atualizados, todos os pesos dos neurônios situados dentro de uma vizinhança pré-determinada são atualizados, conforme ilustrado na Figura 2.2-6. Portanto, o grau de adaptação do neurônio BMU e de seus vizinhos depende de uma função de vizinhança, denominada $\mathbf{h}_{ci}$, esta função deve reduzir o grau de vizinhança relativo ao BMU ao longo do treinamento para ocorrer a convergência do mapa. A função de vizinhança é, usualmente, uma função gaussiana conforme .

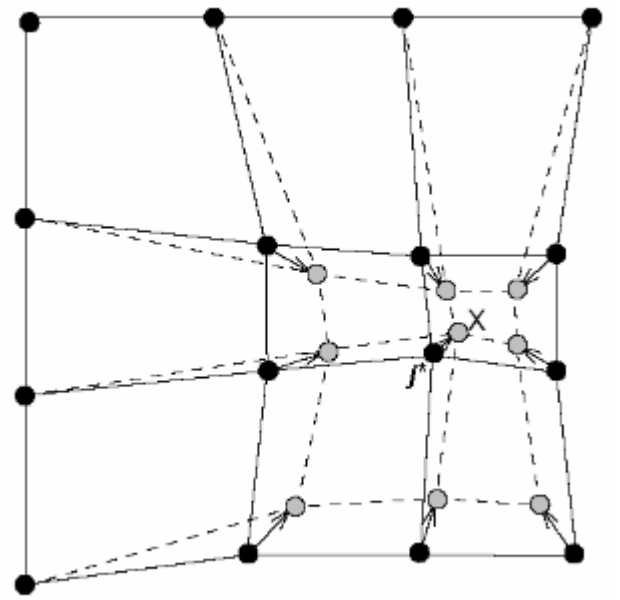


Figura 2.2-6 Mapa antes (neurônios em preto) e depois (neurônios em cinza) da atualização dos pesos do neurônio vencedor *j** e dos seus vizinhos em direção à entrada *X*. Fonte: (Adaptação de VESANTO, 2002).

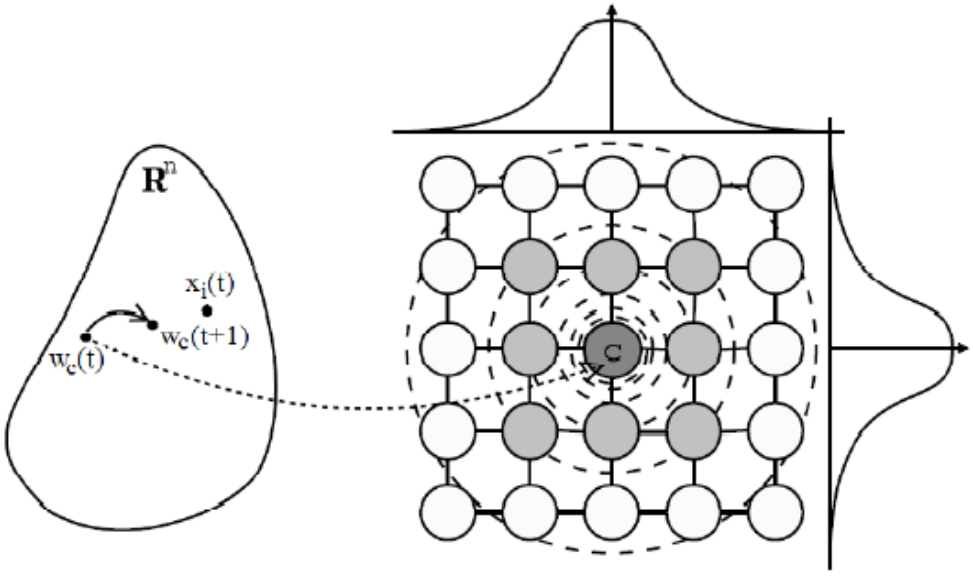


Figura 2.2-7 Atualização dos neurônios vizinhos ao BMU *c* conforme as funções gaussianas na horizontal e vertical. Fonte: (Adaptação de CHOW, 2006)

Um exemplo didático consiste no agrupamento de diferentes animais através de algumas características como tamanho, se é carnívoro ou não, duas ou quatro patos etc. Na Tabela 2.2-1 estão os vetores que foram usados na entrada do SOM de exemplo.

Tabela 2.2-1 Vetores de entrada para o SOM de exemplo (KOHONEN, 2001)

	Dove	Hen	Duck	Goose	Owl	Hawk	Eagle	Fox	Dog	Wolf	Cat	Tiger	Lion	Horse	Zebra	Cow
Small	1	1	1	1	1	1	0	0	0	0	1	0	0	0	0	0
Medium	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
Big	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
2 legs	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
4 legs	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
Hair	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
Hooves	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
Mane	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1	0
Feathers	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
Hunt	0	0	0	0	1	1	1	1	0	1	1	1	1	0	0	0
Run	0	0	0	0	0	0	0	0	1	1	0	1	1	1	1	0
Fly	1	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0
Swim	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0

O mapa de saída desta SOM está ilustrado na Figura 2.2-8 Mapa de saída do SOM de exemplo (KOHONEN,2001). Demonstrando o sucesso de agrupamento de uma RNA, em que o objetivo de mapear os conjuntos de entrada em um mapa topográfico foi alcançado. No mapa gerado (Figura 2.2-8) é possível notar a separação em três grandes grupos. No lado direito há o predomínio de animais

felinos de quatro patas, no lado esquerdo foram agrupadas as aves, é interessante notar a proximidade das aves de rapina com o grupo dos felinos. No canto superior ficaram agrupados os animais de quato patas herbívoros.

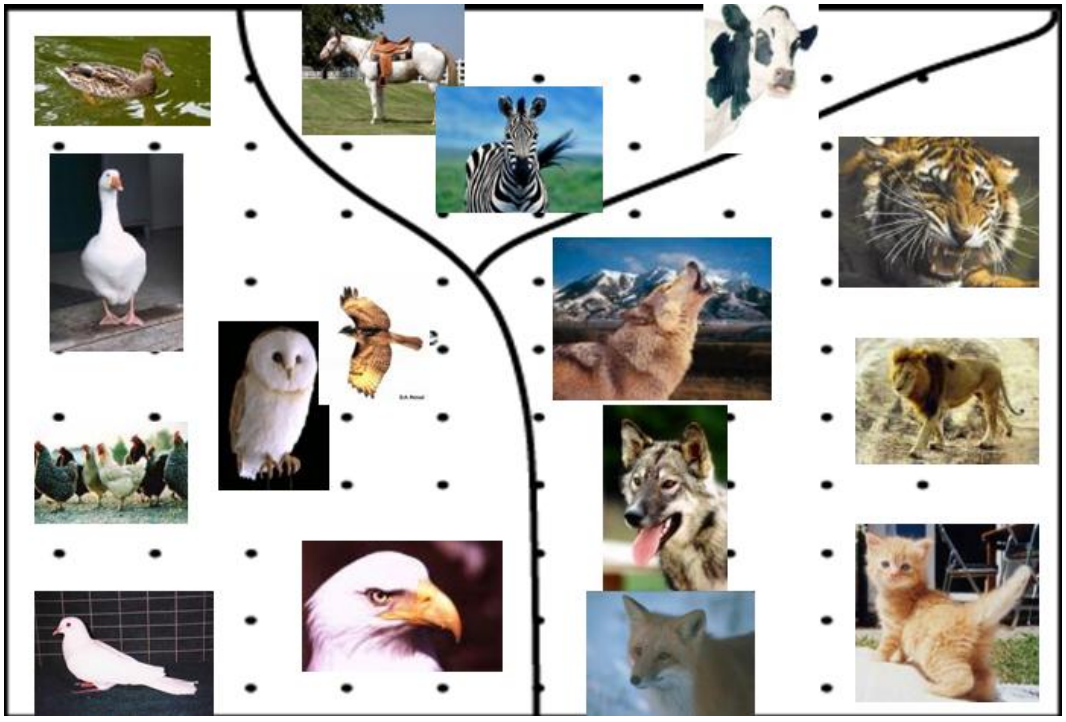


Figura 2.2-8 Mapa de saída do SOM de exemplo (KOHONEN,2001)

2.2.2.1. Conceito Biológico

Seu desenvolvimento teve base na auto-organização do cérebro humano, onde cada área do córtex é associada a várias funções. A Figura 2.2-9 Particionamento do cérebro humano - divisão do córtex cerebral em Lobos. ilustra o particionamento do cérebro humano.

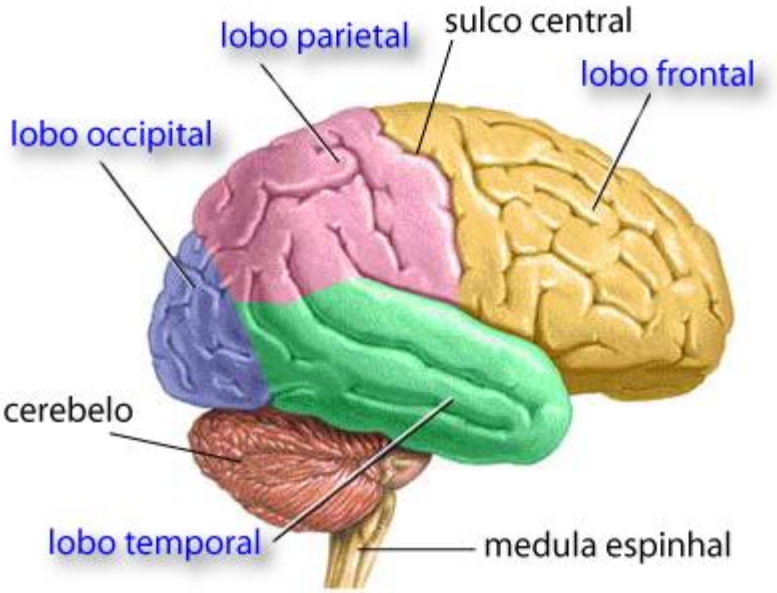


Figura 2.2-9 Particionamento do cérebro humano - divisão do córtex cerebral em Lobos.

Kohonen tinha como objetivo de suas pesquisas criar um modelo auto-organizado de informações que através de um aprendizado indutivo fosse capaz de simular o aprendizado e organização de informações no neocórtex cerebral (WANGENHEIM, 2006).

Este modelo auto-organizado deveria explicar matematicamente o comportamento biológico que ocorre no cérebro de que estímulos similares são aprendidos e agrupados em áreas próximas no cérebro, e que portanto, posteriormente podem ser categorizados e assimilados, temos como exemplos de partições e funções no cérebro humano, de acordo com a Figura 2.2-9 Particionamento do cérebro humano - divisão do córtex cerebral em Lobos.:

- *Lobo frontal*: Responsável pela elaboração do pensamento, planejamento e programação de necessidades;
- *Lobo Parietal*: Responsável pela sensação de dor, tato, gustação, temperatura, pressão;
- *Lobo temporal*: É relacionado primariamente com o sentido de audição, possibilitando o reconhecimento de tons específicos e intensidade do som;

- *Lobo Occipital*: Responsável pelo processamento da informação visual;
- *Lobo Límbico:* Está envolvido com aspectos do comportamento emocional, sexual e com o processamento da memória.

2.2.3. Classificação por Árvores de Decisão

Este método de aprendizagem automática se adapta a diferentes domínios de dados e é facilmente interpretável visualmente, devido a sua representação gráfica. É uma forma simples e eficaz de representar o conhecimento, basea-se na abordagem "dividir para conquistar" (QUINLAN, 1986), ou seja, um problema complexo é dividido em problemas mais simples.

Neste tipo de abordagem estratégica, o conjunto de dados de entrada é dividido sucessivamente em vários subconjuntos, até cada um destes pertencer à mesma classe, ou até, uma das classes ser maioria, não havendo portanto, necessidade de novas divisões. (GARCIA, 2003).

A Figura 2.2-10 apresenta uma árvore de decisão para a escolha de se jogar tênis baseada em variáveis do clima (ensolarado, nublado, chuvoso ou ventando). Para a criação desta árvore foram usados como exemplos de entrada os dados que constam na Figura 2.2-11.

Figura 2.2-10 Árvore de decisão na escolha de se jogar tenis baseado em variáveis do clima

Exemplos de Treino

Dia	Aspecto	Temp.	Humidade	Vento	Jogar Ténis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não

Figura 2.2-11 Dados de entrada para a criação da árvore de decisão

Os algoritmos baseados em árvores de decisão têm como vantagem sua simplicidade e eficiência computacional. A seguir serão apresentados dois dos principais algoritmos baseados em árvores de decisão.

2.2.3.1. Algoritmo ID3 e C4.5

O algoritmo ID3 foi desenvolvido por Ross Quinlan da Universidade de Sydney, Australia (QUINLAN, 1987). Sua metodologia de geração de árvore consiste em começar pelo atributo mais relevante e continuar pelos outros atributos segundo a avaliação da entropia de cada um. O algoritmo visa resolver problemas que tenham atributos categóricos e sem ruídos. Sendo assim, atributos ruidosos devem ser tratados previamente. A necessidade de se trabalhar com atributos do tipo contínuo fez com que sucessivas melhorias fossem desenvolvidas. Em 1993, Ross Quinlan apresenta seu trabalho intitulado "*C4.5: Programs for machine learning*" (QUINLAN, 1993). Trata-se do aprimoramento do algoritmo ID3, tornando possível trabalhar com atributos categóricos e contínuos.

Na Figura 2.2-12 está descrito o algoritmo C4.5. O funcionamento básico do algoritmo consiste na chamada da função de forma recursiva, passando como parâmetros de entrada a base de dados e o conjunto dos atributos desta base. A saída desta função é a árvore de decisão apontada pelo seu nó raiz em *D*.

```
   Entradas: uma base de dados T
             um conjunto de atributos A
   Saída:    uma (sub)árvore de decisão D

1  função C4.5( T, A, D )
2  inicio
3     cria um nó de decisão em D
4     se todas instâncias de T pertencem a mesma classe
5     então
6        atribui ao nó apontado por D uma única folha identificando a classe
7     senão
8        se A é um conjunto unitário
9        então
10          atribui ao nó apontado por D uma única folha identificando
               o valor mais comum do atributo preditor
11       senão
12          início
13             calcula o ganho de informação de cada um dos atributos de A
14             se um dos atributos de A possui ganho de informação médio
                  maior que os demais
15             então
16                início
17                   define at o atributo com maior ganho
18                   para cada valor v do atributo at faça
19                      início
20                         adiciona uma sub-árvore d ao nó apontado por D
21                         define Tv a base com instâncias de T onde at = v
22                         C4.5( Tv, A-{at}, d )
23                      fim.
24                fim.
25             senão
26                atribui ao nó apontado por D uma única folha identificando
                     o valor mais comum do atributo preditor
27          fim.
28 fim.
```

Figura 2.2-12 Função iterativa do algoritmo C4.5. Fonte: Lopes (2007).

Uma das dificuldades deste algoritmo é a escolha do atributo a ser utilizado em cada um dos nós de decisão, seja ele o nó raiz ou um dos demais nós de decisão. Essa escolha é feita nas linhas 13 e 14 do algoritmo (Figura 2.2-12). A primeira etapa (linha 13) consiste em considerar todos os testes que dividem a base em dois ou mais grupos. Esta etapa é feita observando para cada um dos atributos do conjunto *A* o ganho de informação em relação à classificação desejada. Os subconjuntos gerados são analisados através do cálculo da entropia de cada subconjunto de instâncias. Esta entropia é utilizada para calcular o ganho de informação que o atributo considerado obteve. A segunda etapa (linha 14) é um teste entre: (i) escolher o atributo com o maior ganho e chamar a função recursivamente a função C4.5 para criar uma sub-árvore (linhas 16 a 24); (ii) assumir não ser necessário criar um nó de decisão e apena adicionar uma folha com a classe mais frequente (linha 26) (LOPES. 2007).

2.2.3.1.1. Critério de Seleção de atributos

Durante a construção da árvore de decisão são utilizados critérios para a seleção dos atributos que contenham os melhores registros que se enquadram no

nó da árvore em questão. Para isso considera-se a capacidade informativa do atributo, determinada pelo cálculo da Entropia, definida na seção a seguir.

2.2.3.1.1.1. Entropia

Entropia é uma medida da aleatoridade de uma variável. A construção de uma árvore de decisão é guiada pelo objetivo de diminuir a entropia ou seja a aleatoridade, que é o que dificulta a previsão da variável objetivo. Seguindo esta heurística, o algoritmo busca o melhor atributo para classificar os registros, reduzindo sua aleatoridade, a fim de que os mesmos tenham consistência máxima, ou seja, registros da mesma classe.

A entropia usada para a construção de árvores de decisão tem sua origem na teorida da informação, com base nos trabalhos realizados por Claude Shannon e Warren Weaver (SHANNON, 1949). O cálculo da entropia é dado pela Equação 2.2-5 que determina o número de exemplos de *S* pertencentes à classe C_j, sendo os atributos com *m* possíveis valores:

$$Entropia\ S = \sum_{j=1}^{m} -p_j \log_2 p_j \qquad \text{Equação 2.2-5}$$

Onde: S é um conjunto de exemplos

m é o número de classes

p_j é a proporção de S que pertence à classe j, tendo:

$$p_j = \frac{S_j}{S} \qquad \text{Equação 2.2-6}$$

Onde: S_j é o número de exemplos classificados na j-ésima partição

S é o número total de exemplos do conjunto *S*

Uma classificação é considerada perfeita, se todos os membros de um conjunto *S* pertencem a uma mesma classe, portanto, a entropia é igual a *zero*. Se o conjunto S contiver números diferentes de exemplos e de suas classes, a entropia estará entre zero e um. Na Figura 2.2-13 Variação da entropia H(p) em função da propabilidade de p. está ilustrado o gráfico da entropia de S em função da probabilidade de S.

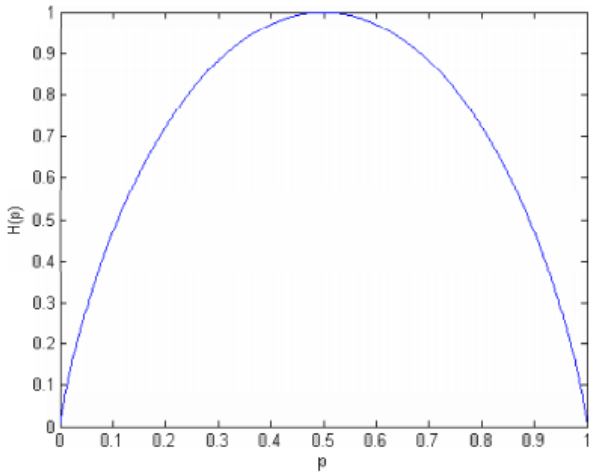


Figura 2.2-13 Variação da entropia H(p) em função da propabilidade de p.

2.2.3.1.1.2. Ganho de Informação médio

O ganho de informação é dado pela Equação 2.2-7 e mede a redução da entropia causada pela partição dos exemplos de acordo com os valores do atributo. Tem como base a medida da entropia.

$$Ganho\ S,A\ = Entropia\ S - \sum_{j=1}^{m} \frac{S_j}{S}\ Entropia(S_j) \qquad \text{Equação 2.2-7}$$

Onde *Ganho(S,A)* é o ganho do atributo *A* sobre o conjunto *S*

S_j é o subconjunto de *S* no qual o atributo A tem valor *j*

3 Bioinformática

3.1. Introdução

Bioinformática é a aplicação da ciência da computação e da tecnologia de informação na área de biologia e medicina. A bioinformática cuida do gerenciamento, análise, extração e interpretação de dados biológicos (BOGUSKI, 1998) (BOGUSKI, 1994).

Com o aumento da velocidade computacional e o avanço tecnológico no sequenciamento automatizado de DNA (*deoxyribonucleic acid;* em português: ADN, ácido desoxirribonucleico) houve um crescimento acelerado do volume de dados coletados. No entanto, por se tratar de um grande volume de dados em um espaço de tempo curto, a sua compreensão é lenta, sendo necessário portanto, equipamentos de armazenagem e métodos de análise eficientes para a nova área que surgiu (FERNANDES, 2009) (NAPOLI, 2003) (BOGUSKI, 1994).

No começo da década de 1990, o sequenciamento de DNA de diversas espécies, em larga escala, se acelerou. O projeto do genoma humano está incluso neste período e tinha como principal objetivo identificar e mapear os genes de todos os 23 pares de cromossomos humanos. Devido ao grande número de informações, surge a necessidade de se criar grandes banco de dados, desenvolver ferramentas para analisar esses dados e modelos de como transformar essas informações em conhecimento prático na biologia e medicina. Muitas destas base de dasos são públicas. A seguir apresenta-se, uma breve descrição das principais existentes:

- *GenBank (Genetic Sequence Database):* é um banco de dados que armazena sequências de DNA públicas, onde cada sequência é cadastrada com sua descrição, nome científico e taxonomia do organismo proveniente. Na Figura 3.1-1 se encontra o gráfico de crescimento desta base de dados até o ano de 2006. No release *187.0* a contagem total de pares de base foi de 135.117.731.375 (NCBI, 2012).

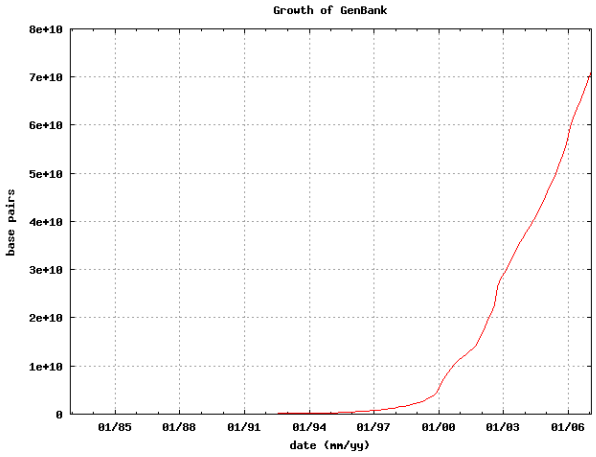


Figura 3.1-1 Crescimento da base de dados GenBank até o ano de 2006. Fonte: (GENBANK, 2012)

- *NCBI (National Center for Biotechnology Information):* Mantém atualizadas informações sobre genomas, possui o BLAST que é ferramentas para alinhamentos, *link* com o PubMed para citações e resumos de literatura da área biomédica.
- *EMBL (European Molecular Biology Laboratory):* Base com ênfase em biologia molecular, possibilitando o gerenciamento de múltiplos níveis da organização biológica, da molécula ao macro-organismo.
- *EBI (European Bioinformatics Institute):* Base de dados do centro de serviços e pesquisas na área de bioinformática, a qual incluci DNA, sequências de proteínas e estruturas macromoleculares.
- *PDB (Protein Data Bank):* a principal base de dados de proteínas. É capaz de gerenciar estruturas de moléculas e sequências de DNA em 3D.

Pode-se resumir os três principais objetivos da bioinformática em: organização do banco de dados; desenvolvimento de ferramentas que auxiliem na análise dos dados; e a criação de modelos que se tornem úteis nas áreas biomédicas, transformando portanto, os dados biológicos em conhecimento prático.

3.2. Identificação de Bactérias

3.2.1. Introdução

A identificação de bactérias por meio de classificação procura descrever a diversidades de espécies de bactérias agrupando e categorizando os organismos levando em consideração as suas similaridades. Bactérias podem ser agrupadas levando-se em consideração a estrutura celular, metabolismo, DNA, pigmentação etc. Porém, estes métodos são úteis na identificação e classificação de cepas de bactérias e não na separação por espécies.

A dificuldade em classificar espécies é decorrente da falta de estruturas distintas nas bactérias, ocasionada pela transferência horizontal de genes entre diferentes espécies (Boucher, 2003), que faz com que bactérias, inicialmente parecidas, se tornem muito distantes morfologicamente e metabolicamente (BACTERIA, 2012).

O isolamento e identificação da espécie de bactérias é crucial no tratamento do paciente, afinal, o tipo de tratamento é determinado pela espécie da bactéria que está causando a patogenia . O que torna esta identificação ainda mais difícil é o fato de diferentes espécies apresentarem morfologia e metabolismo idênticos.

O processo laboratorial de identificação de bactérias consiste em basicamente três etapas:

- *Coleta de amostras*: Varia conforme a fonte da amostra. No caso do Laboratório de Difteria e Corinebactérias de Importâncias Clínica (LDCIC) são colhidas amostras de fluidos orgânicos como sangue, urina etc.
- *Cultivo*: As amostras recolhidas são espalhadas em placas de meio de cultura e postas em incubação para que sejam formadas colônias de bactérias. Nesta etapa alguns métodos podem ser usados para se dar o primeiro passo na identificação da bactéria que está incubada, como a utilização de meios de cultura seletivos para determinados grupos metabólicos de bactérias e a observação das caracteríticas das colônias

como forma, textura, cor, reação de hemólise e se requer ou não oxigênio para crescimento.

- *Identificação*: Existem diversos métodos que podem ser usados na identificação de bactérias e normalmente são usados ao mesmo tempo. Nas seções a seguir serão explicados como funcionam as principais técnicas de identificação.

3.2.2. Identificação por coloração

A técnica de coloração ou técnica de Gram, foi desenvolvida pelo microbiologista Hans Christian Gram e divide as bactérias em dois grupos: Gram-positivas e Gram-negativas (BERGEY, 1994). Deve ser levada também em consideração a morfologia das bactérias (bacillus, coccus, spirillum etc.) e das colônicas. Algumas possibilidades de formação de colônias estão ilustradas na Figura 3.2-1. Na Figura 3.2-2 estão representadas as principais formas que as bactérias podem apresentar.

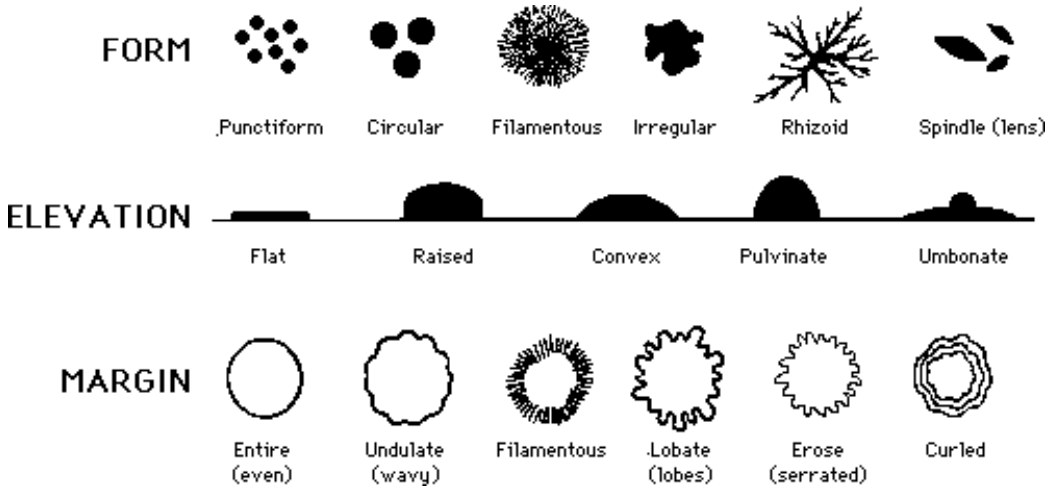


Figura 3.2-1 Descrição de algumas formas de colônias em placas. Fonte: (BACT, 2012).

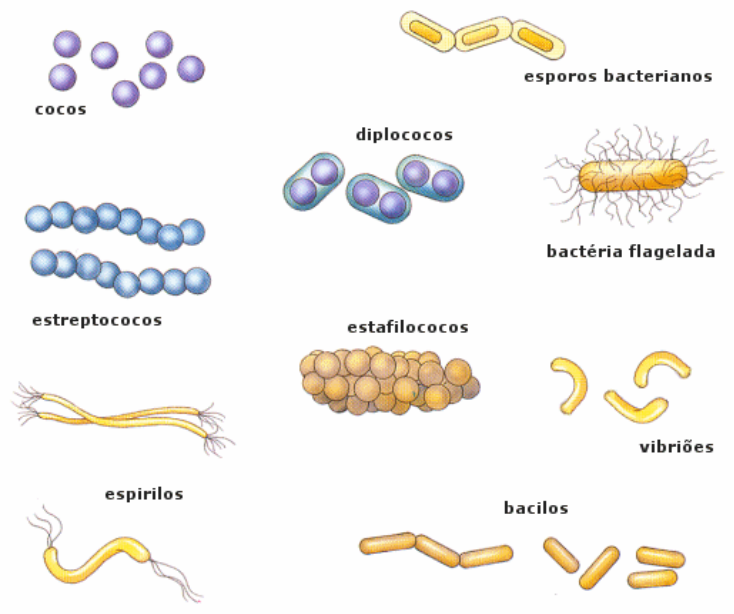


Figura 3.2-2 As principais formas que as bactérias podem apresentar. Fonte: (EARTHLIFE, 2012).

3.2.3. Identificação bioquímica

Consiste na avaliação da capacidade de determinadas substâncias serem metabolizadas pelas bactérias, que naturalmente realizam atividades bioquímicas através dos nutrientes do seu ambiente. Este método consiste em diversas provas como:

- Provas fermentativas: Glicose, lactose, sacarose, manose etc.
- Indol
- Motilidade
- Hidrólise do amido

Estas e outras provas são realizadas ao mesmo tempo e ao fim de cada reação o resultado é anotado para posterior comparação e identificação da espécie incubada. A Figura 3.2-3 e Figura 3.2-4 são dois exemplos de testes bioquímicos reais, respectivamente o teste de hidrólise de amido e indol.

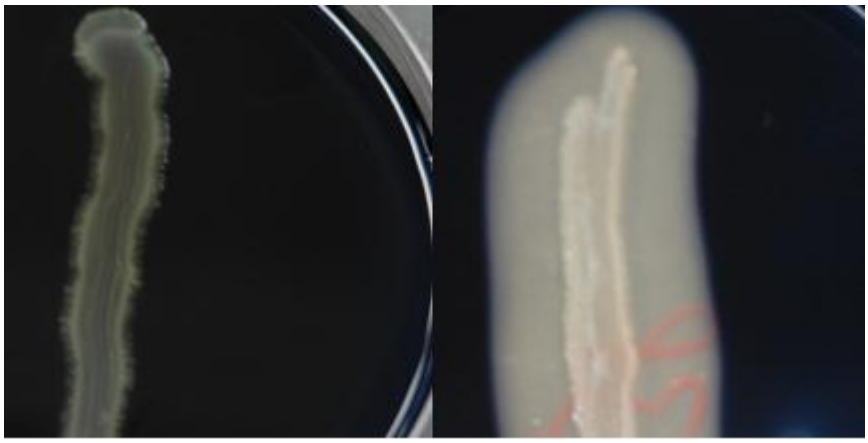


Figura 3.2-3 Teste de hidrólise do amido. Na direita um exemplo de microorganismo capaz de degradar o amido.

Figura 3.2-4 Teste Indol. O indol pode ser detectado pela formação de um anel rosa na parte superior do tubo.

A identificação de bactérias através de testes bioquímicos é uma prática rotineira e os resultados destes testes bioquímicos já são conhecidos e padronizados para várias espécies e seus grupos. Com o objetivo de otimizar o tempo e diminuir os custos, empresas criaram testes miniaturizados e compactos para que todas as provas sejam realizadas rapidamente. No entanto, ainda há a necessidade dos testes serem incubados por um período, geralmente, entre 24-48 horas. Um exemplo deste sistema é o API 20 fabricado pela empresa Biomerieux (BIOMERIEUX,2012). Na Figura 3.2-5 são ilustrados três testes bioquímicos usando o sistema API 20 para três amostras distintas.

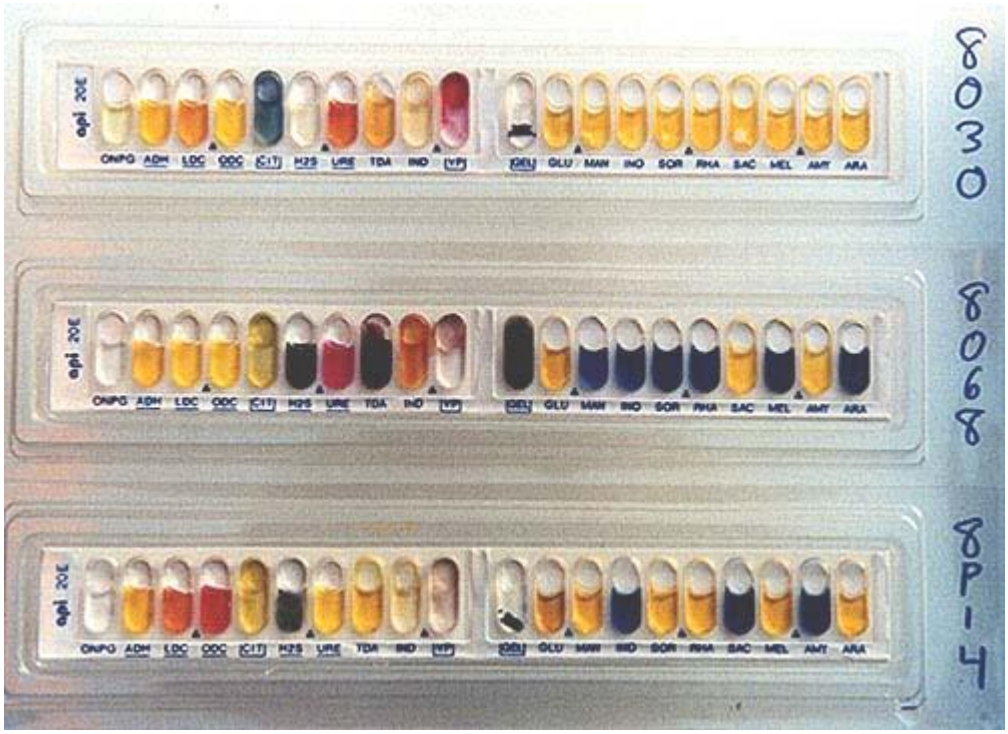


Figura 3.2-5 API 20. Testes para três amostras: 8030, 8068 e 8P14.

Conforme a Tabela 3.2-1, depois de realizado todos os testes bioquímicos e seus resultados anotados, é realizada a comparação entre os resultados das provas e as tabelas padronizadas, a espécie em questão é determinada de acordo com o maior índice de similaridade entre os testes.

Tabela 3.2-1 Resultados dos testes do API 20 para cada cultura e sua respectiva identificação segundo os testes bioquímicos listados nas colunas.

Número da cutura	ONPG	ADH	LDC	ODC	CIT	H2S	URE	TDA	IND	VP	GEL	GLU	MAN	INO	SOR	RHA	SAC	MEL	AMY	ARA	Identifiação
8030	+	–	+	–	+	–	+	–	–	+	–	+	+	+	+	+	+	+	+	+	*Klebsiella pneumoniae*
8068	–	–	–	–	–	+	+	+	+	–	+	+	–	–	–	–	+	–	+	–	*Proteus vulgaris*
8P14	–	–	+	+	–	+	–	–	–	–	–	+	+	–	+	+	–	+	–	+	*Salmonella* sp.

3.2.4. Identificação genotípica

Este tipo de identificação consiste em análises voltadas às moléculas de DNA ou RNA (KHAMIS, 2005). A melhor forma de identificar espécies é através da comparação genômica de sequências completas, no entanto, nos últimos anos poucos genomas completos de bactérias foram sequenciados, por ser uma tarefa complexa que envolve elevados gastos e o envolvimento de vários grupos de pesquisa. Por isso este tipo de análise é, ainda hoje, inviável de se tornar rotineiro nos laboratórios (EMBRAPA, 2012).

Como alternativa, surgiram técnicas menos complexas e despendiosas a nível molecular. Dentre elas estão metodologias baseadas em reação em cadeia da polimerase (PCR) e suas variações, comparação entre conteúdo de guanina mais citosina (%GC), hibridização DNA-DNA e RNAr 16S (BULL, 1992).

Atualmente o método "RNAr 16S" é o mais confiável no estudo da taxonomia bacteriana. Segundo PEIXOTO (2002), o RNAr 16S preenche todos os requisitos que definem um marcador molecular ideal, pois seu tamanho é grande o suficiente para permitir comparações significativas, possui regiões altamente conservadas, está presente e tem a mesma função em todas as espécies, não são afetadas por mudanças ambientais e apresenta um grande número de sequências disponíveis em bases de dados na internet.

3.3. Bactérias: O Gênero Corynebacterium

3.3.1. Introdução

Desde 1970, muitos pesquisadores se dedicam a estudar infecções graves causadas por BGPIs (Bastonetes Gram-positivos Irregulares) coriformes. Segundo Clarridge & Spiegel (1995), destacam-se como fatores de predisposição às infecções pelas corinebactérias os estados de malignidade, idade avançada, transplantes, AIDS, diabetes, neutropenia, hospitalização e/ou antibioticoterapia por período prolongado, além de procedimentos invasivos (cateterismo, implantes

e válvulas).

Em estudo realizado na Europa por Kraeva (2007), os exames bacteriológicos de 1589 pacientes demonstraram que 11% das infecções agudas do trato respiratório superior foram causados por *Corynebacterium* spp e que estas bactérias podem provocar infecções em várias localizações (bronquite, pielonefrite, uretrite, colpitis, dermatite, artrite, etc.).

O aparecimento de amostras multirresistentes aos antimicrobianos utilizados no tratamento e o aumento do número de casos de infecções de origens diversas, muitas vezes culminando em óbito tanto em países industrializados quanto em desenvolvimento, têm contribuído para aumentar o interesse pelas corinebactérias (GUARALDI, 2008).

3.3.2. Principais espécies de interesse médico

Até o final do ano de 2011, o gênero *Corynebacterium* é composto de 59 espécies, sendo 36 destas consideradas de relevância médica. Além das corinebactérias produtoras de toxina, que são patógenos obrigatórios para humanos e/ou animais (*C. diphtheriae*, *C. ulcerans* e *C. pseudotuberculosis*), diversas outras espécies podem fazer parte da microbiota[1] anfibiôntica normal[2] humana e/ou atuar como patógenos oportunistas. As infecções humanas causadas pelas corinebactérias podem levar ao óbito tanto pacientes imunocomprometidos[3] quanto imunocompetentes[4] (GUARALDI, 2008).

Devido à complexidade deste gênero, é frequência de reclassificação taxonômica de seus componentes, conforme aumenta o grau de conhecimento sobre as suas características fenotípicas e genotípicas.

Dentre as espécies de bactérias isoladas de seres humanos, a partir do ano de

[1] A palavra "microbiota" significa: conjunto dos microrganismos que habitam um ecossistema.
[2] O termo "microbiota anfibiôntica" é mais adequado que "microbiota normal" pois um microorganismo que não é patogênico em certas condições pode vir a ser.
[3] Pacientes que possuem o seu sistema imunológico (de defesa) fragilizado/deprimido. Em contato com um determinado antígeno, ele é incapaz de criar anticorpos.
[4] Pacientes que são capazes/competentes de produzir uma resposta imunológica a um determinado antígeno.

1990, podemos incluir: C. *auris* (FUNKE, 1995; BABAY, 2004), C. *argentoratense* (RIEGEL, 1995), C. *durum* (RIEGEL, 1997), C. *imitans* (FUNKE, 1997a), C. *mucifaciens* (FUNKE, 1997b), C. *lipophiloflavum* (FUNKE, 1997c), C. *coyleae* (FUNKE, 1997d; TAGUCHI, 2006; FERNANDEZ-NATAL, 2008), C. *riegelii* (FUNKE, 1998), C. *confusum* (FUNKE, 1998b), C. *falsenii* (SJODEN, 1998) C. *kroppenstedtii* (COLLINS, 1998), C. *thomssenii* (ZIMMERMANN, 1998) e C. *sundsvallense* (COLLINS, 1999).

3.3.3. Diagnóstico laboratorial

A identificação das bactérias em questão é fundamental para que não haja o estabelecimento ou o agravamento dos quadros infecciosos. Uma das dificuldades na identificação é o fato das bactérias corineformes representarem, aproximadamente, 60% das bactérias de tipo bastonetes Gram-positivos irregulares (BGPIs). Em função das frequentes alterações na taxonomia devido ao avanço no conhecimento genotípico e à descoberta de novas espécies do gênero *Corynebacterium,* a identificação dos BGPIs isolados de material clínico torna-se ainda mais difícil (CAMELLO, 2003).

Poucos laboratórios possuem recursos materiais e profissionais para análise fenotípica e/ou genotípica das amostras colhetadas. Além disso, muitas vezes é desconsiderada a relevância destes microrganismos, o que resulta em atraso no início do tratamento adequado e consequente agravamento do quadro clínico do paciente infectado.

Segundo Guaraldi (2010), é fundamental que os procedimentos laboratoriais sejam realizados de forma rápida e, ao mesmo tempo, precisa. Entretanto, a qualidade das investigações depende da disponibilidade de reagentes, da boa formação do pessoal do laboratório e dos recursos financeiros. A Figura 3.3-1 destaca, de forma generalizada, as etapas de isolamento e identificação de bactérias.

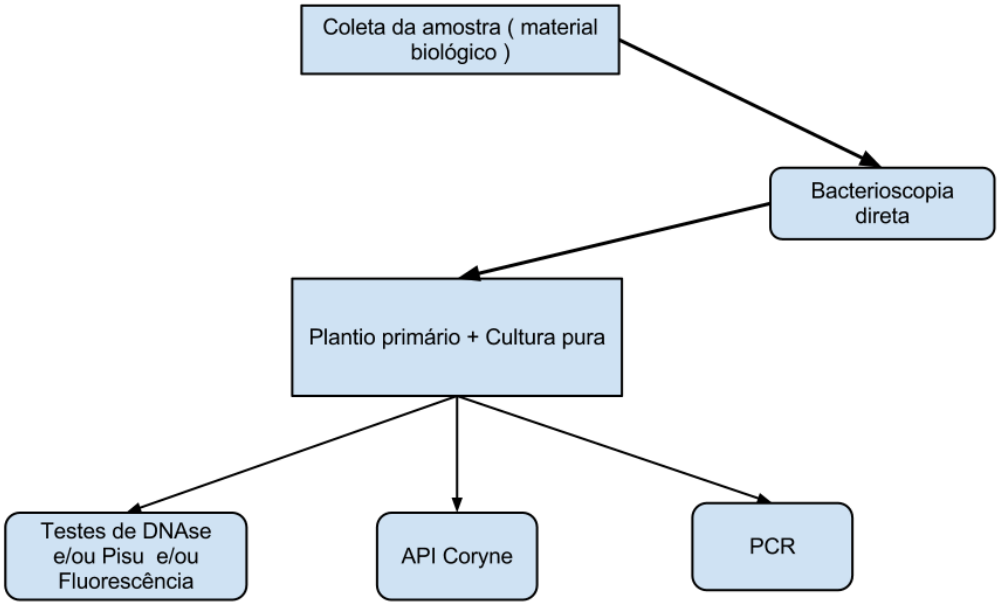


Figura 3.3-1 Principais etapas do diagnótico laboratorial.

O processo ilustrado na Figura 3.3-1 pode ser resumido em:

- *Coleta da amostra*: O material clínico, como sangue, secreções diversas, líquor etc. é colhido do paciente. O material é então encaminhado para o laboratório com as informações do paciente em questão;
- *Bacterioscopia direta*: Esta etapa serve como triagem, não tem o objetivo de realizar um diagnóstico, pois diversas características visuais através de métodos de coloração podem indicar se o microrganismo em questão é relevante ou não para análise;
- *Plantio primário + Cultura pura*: Após a constatação da relevância da amostra em questão, a mesma é direcionada para a cultura de condicionamento de preparo para os métodos de identificação;
- *Testes de DNAse e/ou Fluorescência, API Coryne, PCR*: Conforme visto nas seções anteriores, são algumas técnicas de identificação que podem ser aplicadas.

Todo processo deve ser meticulosamente registrado na ficha de controle. É feito o registro de dados clínicos e epidemiológicos necessários ao acompanhamento dos espécimes. Para a vigilância e monitoramento, o laboratório deve registrar detalhes como:

- *Detalhes do paciente*: Nome, idade, sexo, hospital de admissão e médico responsável pélo atendimento;
- *Detalhes do laboratório*: Sítio de coleto do espécime clínico, datas das coletas;
- *Detalhes clínicos*: Sintomas, data do início do quadro clínico.

Por fim são registrados os resultados dos testes para a identificação da bactéria isolada. No Anexo-A está digitalizada a ficha de dados, usada no Laboratório de Difteria e Corinebactérias de Importância Clínica (LDCIC), que engloba todas estas informações.

3.4. Estado da Arte na identificação de bactérias

No contexto biológico, a taxonomia dos microrganismos pode ser dividida em três conceitos básicos:

- *Classificação*: Divisão dos microrganismos em grupos. Possui tipicamente uma estrutura hierárquica, seguindo a ordem de classe, ordem, família, gênero e espécie.
- *Nomenclatura*: Nomeação do microrganismo. Seu nome científico é formado por duas partes, a primeira informa o gênero e a última a espécie.
- *Identificação*: Se trata da unidade básica da taxonomia, é o menor e mais definitivo nível de divisão.

Em inteligência comutacional, a diferenciação entre classificação e identificaçao também pode ser entendida como aprendizado não-supervisionado e supervisionado respectivamente (RIPLEY, 2008).

O conceito de espécie pode ser entendido como um grupo de microrganismos que compartilham diversas características e que são capazes de se reproduzir entre sí, porém muito se discute sobre uma denifição única deste conceito (VANDAMME, 1996). Deve ser detacado que não existe um método oficial de classificação e identificação de bactérias (GYLLENBERG, 2001) e segundo VANDAMME (1996) a classificação de bactérias sempre será instável.

Na identificação de bactérias, as primeiras técnicas basearam-se em algumas propriedades visuais e comportamentais, porém, estes métodos refletem a limitação tecnológica daquele período (BOONE, 2001). Tradicionalmente a caracterização de bactérias é feita através de provas bioquímicas, estes testes, detectam as respostas metabólicas dos microrganismos, são portanto, características fenotípicas que refletem a genotipia da bactéria que está sendo testada (GYLLENBERG, 2001). Nas análises de quais testes têm mais relevância na separação das espécies destacam-se Willcox (1973) e Gyllenberg (1963). Para ilustração, na Tabela 3.4-1, se encontram as respostas dos principais testes bioquímicos da espécie Corynebacterium accolens. Nesta tabela, estão listados os nomes dos testes bioquímicos e logo abaixo dos mesmos se encontram os respectivos resultados. Valores com o símbolo “-“ indicam que para aquele teste não houve reação, “+” indica que o teste reagiu positivamente e “V” significa que o resultado pode ser “+” ou “-“.

Tabela 3.4-1 Perfil bioquímico da espécie Corynebacterium accolens. Referências: Janda 1998; MacFaddin 1999; Murray 2007

Esculina	**Nitrato**	**Urease**	**Glicose**	**Maltose**	**Sacarose**	**Manose**	**Manitol**
-	+	V	+	-	V	V	V
Glicogenio	**PYZ**	**PYRA**	**PAL**	**Beta_GUR**	**Alpha_GLU**	**Beta_NAG**	**GEL**
+	+	ND	-	ND	ND	ND	-
Hemolise	**Agente_0129**	**CAMP**	**Ribose**	**Amido**	**BETA_GAL**	**DNAse**	**GLI_20°**
-	ND	-	-	ND	ND	-	ND
Frutose	**Xilose**	**Arabinose**	**Galactose**	**Trealose**	**Catalase**	**FLUOR**	**Lipofilia**
ND	-	-	ND	ND	+	ND	+
GLI_42°	**O_F**	**Tirosina**	**Mobilidade**	**AAR**			
ND	F	ND	-	-			

No fim da década de 1950 iniciou-se a identificação e classificação de microrganismos baseadas na taxonomia numérica (SNEATH, 1957a) (SNEATH, 1957b) (SNEATH, 1962), onde dados fenotípicos são analisados por coeficientes numéricos que expressam similaridade entre as linhagens. Nesta abordagem se utiliza um grande número de testes bioquímicos (100 a 200) e uma amostragem grande e diversificada de linhagens, tendo os seus resultados expressos em porcentagens (VANDAMME, 1996). Um método que pode ser destacado como exemplo é o de estatística empírica, onde os perfis bioquímicos das bactérias da tabela de referência são representados por elementos binários (Tabela 3.4-2).

Tabela 3.4-2 Perfil de testes bioquímicos fictício, exibindo seu resultado normal e abaixo a sua representação binária.

	Teste 1	**Teste 2**	**Teste 3**	**Teste 4**	**Teste 5**
Resultado normal	+	+	-	+	-
Resultado convertido	1	1	0	1	0

A

Tabela 3.4-3 apresenta quatro espécies da família *Enterobacteriaceae* e seus respectivos perfis bioquímicos convertidos para valores binários. Após a conversão de todos os registros da tabela de referência, é calculada a frequência dos elementos por espécie. Por exemplo a frequência do dígito "1" na sexta coluna de testes (destaque em amarelo) é calculada como $f_{61=}\frac{3}{4} = 75\%$. Com base nas frequências dos dígitos de cada coluna, monta-se uma outra tabela de frequências, que é normalizada segundo o critério de caso o valor de frequência seja maior que 50% e menor que 100% ela é transformada em "1", caso contrário, assume o valor "0".

Tabela 3.4-3 Quatro espécies da família *Enterobacteriaceae* e seus respectivos perfis bioquímicos codificados em binário. Fonte: Laboratório de Bacteriologia de Atlanta, EUA (GYLLENBERG, 2001).

Espécie	**Perfil bioquímico**
Budvicia aquatica	0100110000000011101001001010100000001000010100
Budvicia aquatica	0100100000010011101001001010100000011000010100
Budvicia aquatica	0100110000000011101001001010100000001100010100
Budvicia aquatica	0100110000000011101001001010100000001110010100

A partir da tabela de frequências normalizada o perfil bioquímico da bactéria sob análise é posto a prova, sendo comparado com cada registro, e a comparação que apresentar maior similaridade corresponde à melhor identificação hipotética.

Beers e Lockhard (1962) foram os primeiros a sugerir o uso de probabilidade no processo de identificação bacteriológica. Dybowski e Franklin (1968) detalharam o processo de probabilidade condicional na identificação de enterobactérias. Seguindo esta metodologia, Lapage (1970) identificou com sucesso de 70% a 80%, 279 cepas recém isoladas. A partir de então o computador

se mostrou indispensável na classificação e identificação de bactérias (BASCOMB, 1973) (GYLLENBERG, 1965) (LAPAGE, 1973) (SCHNIDER, 1979) (WILLCOX, 1973). O modelo básico de probabilidade condicional consiste no teorema de Bayes (BAYES, 1763), com algumas modificações para melhorar a eficiência do algoritmo (BRYANT, 2004). Para calcular a probabilidade de uma bacteria desconhecida pertencer a uma determinada espécie, aplica-se a Equação 3.4-1.

$$P\ t_i|\ R\ = \frac{P\ R|t_i\ \ P\ t_i}{j \in \{1,\ldots,n\}\ P\ R\ t_j\ \ P\ t_j} \qquad \text{Equação 3.4-1}$$

Onde P(*x*|*y*) é a probabilidade condicional do evento *x*, assumindo a ocorrência do evento y, P(*x*) é a probabilidade incondicional do evento x, e *j* percorre todos os *n* registros de espécies da tabela de referência. *R* representa o perfil da bactéria sob análise e t_i um registro específico da tabela. Este modelo é usado até os dias de hoje e exige algumas pré-configurações para seu perfeito funcionamento.

Ao final da década de 1960, houve uma revolução no processo laboratorial de identificaçao bioquímica. A indústria criou sistemas de testes compactos para realização dos testes, o que é muito vantajoso pois torna o processo padronizado, sequencial, mais barato devido à economia de insumos, redução da necessidade de espaço físico de armazenamento e incubação e em alguns casos, tornou a identificação mais rápida (JANDA, 2002). Dentre os testes criados, o *API 20E strip test* (bioMérieux-Vitek, Hazelwood, Mo.) se tornou o sistema commercial "padrão ouro" na maioria dos laboratórios de todo o mundo até os dias de hoje (ALMUZARA, 2006) (O'HARA, 1992). Trata-se de um *kit* com vinte provas bioquímicas diferentes, que gera como resultado um código de sete dígitos. A partir deste código pode-se pesquisar, em um livro fornecido pela empresa, a identificação da bactéria correspondente. Porém, as propriedades fenotípicas podem ser instáveis e apresentar variações de acordo com as condições ambientais (ROSSELLÓ-MORA, 2001). A partir de 1980 foram desenvolvidos algoritmos computacionais para identificação de grupos de bactérias, baseando-se

principalmente em informações morfológicas obtidas através dos métodos tradicionais ou dos resultados de sistemas comerciais (BOEUFGRAS, 1988) (BRYANT, 1986)(COX, 1990) (FRENEY, 1991) (MILLER, 1996)(RHODEN, 1993).

Segundo Janda (2002), para espécies que são encontradas com baixa frequência, os sistemas comerciais são falhos na sua identificação devido à insuficiência de registros fidedignos na sua base de dados, como nos casos das espécies *Pasteurella* (HAMILTON-MILLER, 1993) e *Haemophilus* (HAMILTON-MILLER, 1996).

Bryant (2004) anunciou o programa PIBWin (*Probabilistic Identification of Bacteria for Windows*), capaz de prover a identificação probabilística de uma bactéria isolada desconhecida baseada em uma matriz com testes bioquímicos de bactérias conhecidas. Como características gerais do programa pode-se destacar:

- A identificação de uma bactéria isolada desconhecida
- A possibilidade de se adicionar testes para desempate quando a identificação não é possivel
- Salvar e abrir resultados
- Suporte ao formato de arquivo Excel (2003)
- Programa específico para o sistema operacional Microsoft Windows

Flores (2009) avançou no segmento de programas para identificação de bactérias e publicou o *IDENTAX bacterial identifier* (Identax). Trata-se de um programa de código-fonte aberto para identificação de bactérias através de características fenotípicas. Seu processo de identificação é parecido com PIBWin pois faz uso de matrizes com informações de bactérias já conhecidas. Como características destacam-se:

- A identificação de uma bactérias isolada desconhecida
- Suporte aos formatos Excel e CSV
- Criação de uma árvore de decisão com a melhor sequência de testes para identificar o máximo de espécies possíveis
- Configuração dos parâmetros do algoritmo de análise

- Desenvolvido usando *Sun Microsystems Java* o que lhe garante portabilidade para a maioria dos sistemas operacionais.
- Distribuído sob licença *GNU Lesser General Public License* (GNU LGPL) (GNU, 2012)

A partir do ano de 1990 o uso de IA passou a ser largamente aceito em aplicações médicas. Este fenômeno pode ser percebido pelo aumento do número de aparelhos médicos disponíveis no mercado com algoritmos de IA embutidos. Nas revistas médicas houve também um grande aumento no número de publicações que usam IA (GANT, 2001).

Redes Neurais Artificiais (RNA) é a técnica de IA mais popular na medicina (STEIMANN, 2001), pois possui a habilidade de aprender com exemplos históricos de dados, realiza um mapeamento não linear entre as entradas e a saída, suporta dados imprecisos e ruidosos e tem a capacidade de generalizar o conhecimento. RNAs têm sido utilizadas em vários segmentios médicos como diagnósticos clínicos, análise de imagens radiológicas, interpretação de dados biológicos etc (PANDEY, 2009).

Stamey (1996) desenvolveu uma RNA chamada *ProstAsure Index* que possibilita a classificação de câncer na prostata como malígno ou benigno. Este modelo obteve uma precisão de acerto de 90%. O uso de RNAs com grau relevante de sucesso se extende a outros temas como Mama (DOWNS, 1996), Tireóide (KARAKITSOS, 1996), Infarto do Miocárdio (HEDEN, 1994), câncer no pulmão (ZHOU, 2002) e diabetes (PAGANO, 2004).

Alguns pesquisadores empregaram RNAs na identificação e classificação de bactérias. Ahmad (2008) aplicou RNAs na identificação de espécies da família *Peptococcaceae*, usando como dados de entrada as informações de testes bioquímicos do Manual Bergey de Sistemática Bacteriológica (GARRIT, 2001). A RNA funcionou com sucesso e apresentou uma taxa de erro inferior à 8%.

Sahin (2006) demonstrou que as RNAs têm grande potencial na solução de problemas taxonômicos. Vários outros trabalhos foram realizados no estudo das RNAs nos mais diversos grupos de bactérias (Duerden et al. 1989; Rataj et al.

1991; Kennedy et al. 1993; Goodacre et al. 1996, 1998; Giacomini et al. 1997, 2000; Mariey et al. 2001).

4 Sistema BCIWeb

Este capítulo apresenta o sistema computacional desenvolvido nessa dissertação, o qual tem como objetivo auxiliar laboratórios na administração dos dados gerados pelas amostras clínicas a serem analisadas. O sistema funciona em ambiente web e foi programado em PHP e MySQL. É descrita a modelagem do banco de dados e a metodologia de comparação das provas bioquímicas. Ao fim do capítulo, é detalhado o funcionamento do sistema.

4.1. Idealização do sistema BCIWeb

Na Figura 4.1-1 está ilustrado o diagrama em blocos das principais características que o sistema de auxílio aos laboratórios deve apresentar. Com base nestas necessidades foram escolhidas tecnologias especificas a fim de darem suporte a esta ferramenta, conforme será visto nas seções seguintes. A Figura 4.1-2 exibe as principais funcionalidades do sistema.

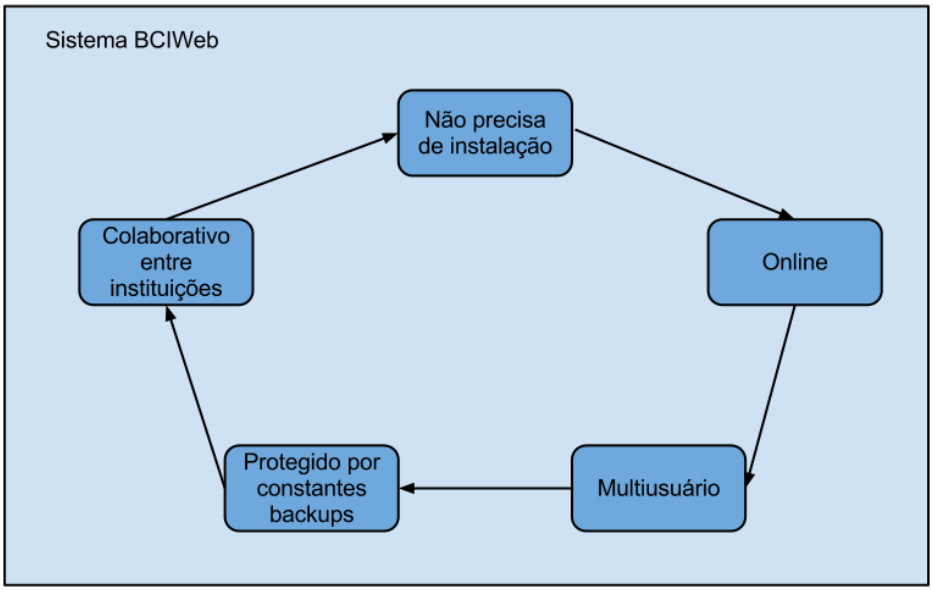


Figura 4.1-1 Diagrama em blocos das principais funcionalidades que devem existir na ferramenta.

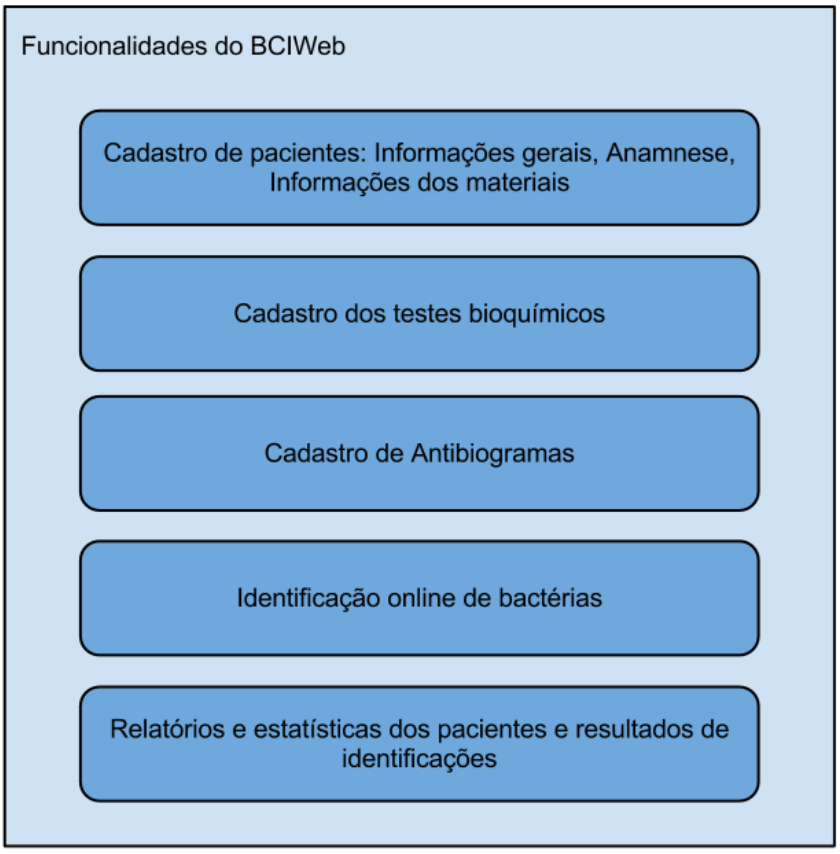


Figura 4.1-2 Diagrama com as principais funcionalidades do sistema BCIWeb.

4.2. Vantagens de um sistema WEB

Programas desenvolvidos em plataforma *web* funcionam através de navegadores como *Internet Explorer*, *Firefox*, *Chrome* etc. e ficam alojados em servidores, o que faz com que possam ser acessados ao mesmo tempo por diferentes usuários. Tipicamente apresentam conexão com bases de dados para registro e administração de informações.

Pode-se destacar algumas vantages de sistemas *web* como:

- Compatibilidade com qualquer sistema que tenha acesso à internet;
- Informações centralizadas e com *backups* constantes;
- Acesso rápido, fácil e atualização que independe do usuário, afinal todos os usuários acessam o mesmo sistema;

- Fácil acesso, uma vez que qualquer pessoa no mundo pode acessá-lo por ser *on-line*;
- Funciona 24 horas, 7 dias por semana.

4.3. Tecnologias Empregadas

4.3.1. Programação em PHP

PHP (Pré-Processador de Hipertexto), é uma linguagem de programação extremamante modularizada, o que a torna perfeita para o uso em servidores *web*. Foi desenvolvida por Rasmus Lerdorf em 1994 (WELLING, 2003) e suas funções, sintaxe e tipos de dados se assemelham com C e C++, podendo também, ser executada dentro de documentos HTML. Provém conexão com as principais base de dados como Oracle, MySQL, PostgreSQL, Firebird etc. A linguagem PHP tem código fonte aberto, pode ser utilizada de forma gratuita e é geralmente usada para criar páginas dinâmicas na *web.*

O PHP está, atualmente, estável na versão "5.3.9" (10 de Janeiro, 2012) com versões para Windows, Linux, FreeBSD, Mac OS, OS/2, AS/400, Novell Netware, RISC OS, AIX, IRIX e Solaris.

Estima-se que, atualmente, mais de 70% das páginas online estejam usando PHP. A base de dados do *WIKIPEDIA* funciona com PHP e MySQL, suportando mais de onze milhões (11.000.000) de artigos. O *Facebook*, que funciona com PHP, em janeiro de 2012, conta com mais de oitocentos milhões (800.000.000) de usuários ativos, com uma média de envio de fotos para seus servidores de duzentos e cinquenta milhões (250.000.000) de fotos por dia.

4.3.2. Banco de dados em MySQL

O MySQL é um sistema de gerenciamento de banco de dados (SGBD), que utiliza a linguagem SQL (*Structured Query Language*) como interface. O MySQL apresenta praticamente todas as funcionalidades das grandes bases de dados comerciais. Seu código-fonte é aberto e está disponível em mais de vinte plataformas.

O MySQL é também uma das bases de dados mais populares no mundo, com mais de 10 milhões de instalações, é usada em *sites* como Google, Facebook, Twitter, Wikipedia e Nokia.com

4.3.3. Sistema Gerenciador de Conteúdo Joomla

Um Sistema de Gerenciamento de Conteúdo para aplicações *web*, também chamados de CMS (*Content Management System*), tem como objetivo separar a gestão de conteúdo e o design gráfico das páginas *web*, organiza a estrutura do *website* visando uma facilitação na criação, distribuição e disponibilização das informações que devem ser publicadas. Todo gerenciamento, desde a instalação até a publicação, é feito *online* através do próprio navegador.

O Joomla é um CMS, criado em 2005, de código aberto e totalmente desenvolvido em PHP, HTML e MySQL. Atualmente está, estável na versão "1.7.3" (14 de Novembro, 2012). É dividido em duas partes: administrativa e exibição, as quais são detalhadas a segior:

- *Administrativa*: Conhecida também como *back-end*, é restrita para utilização dos administradores. Nesta área se tem todo o controle de configurações do sistema, criação, edição e gestão de conteúdo do site. Novos usuários com diferentes níveis de privilégio e acesso podem ser adicionados e editados.
- *Exibição*: Conhecida também como *front-end*, é a área onde os visitantes têm acesso e visualizam o conteúdo informativo publicado pelo sistema. Essa visualização e acesso a conteúdos pode ser controlada de acordo com cada usuário, localização de acesso, usuários registrados ou não etc.

O sistema tem suporte para o desenvolvimento de novos componentes que se integram ao Joomla a fim de que novas funcionalidades sejam desenvolvidas e aplicadas ao *site*. Para isso existe o *Joomla Framework*, ilustrado na Figura 4.3-1.

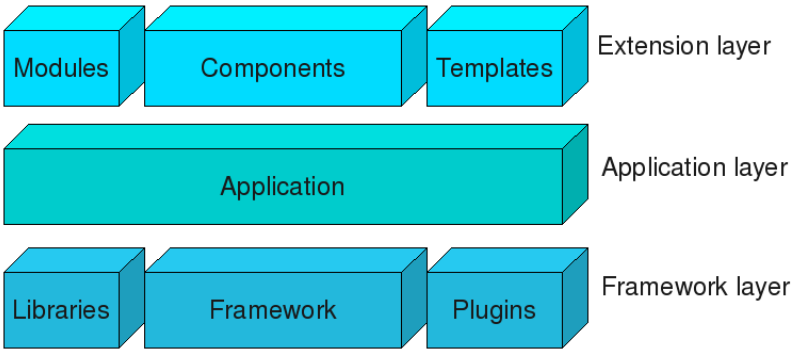


Figura 4.3-1 Joomla framework dividido em módulos. Fonte: (JOOMLA, 2012)

O *Framework* é dividido em três camadas: Extensão, Aplicação e Framework. O acesso à cada camada depende da complexidade e das funcionalidades do novo aplicativo que está sendo criado. Os aplicativos podem ser divididos em:

- *Modules:* É a camada mais básica de desenvolvimento, possibilita a interação do usuário com o sistema e seus componentes. Podem existir vários *modules* com as mais diversas funcionalidades na mesma página;
- *Components:* É a principal parte no desenvolvimento de uma aplicação *web*. Cuida das funcionalidades da aplicação e através dos *modules* interage com os usuários e exibe seus resultados;
- *Templates:* Possibilidade a criação e gerência de códigos HTML e CSS que permitem a personalização da estética das páginas;
- *Application:* Na camada aplicação se encontram as classes desenvolvidas para darem suporte ao *Framework*;
- *Framework:* Contém as classes do CMS;
- *Libraries:* Contém a biblioteca requerida pelo *Framework* e aceita adições vindas de outras aplicações;
- *Plugins:* Auxilia no acesso de outras aplicações ao *Framework*.

4.3.4. Google Chart Tools

O Google Chart Tools (CHART, 2011) permite gerar dinamicamente visualizações gráficas interativas e imagens, fornece suas classes em *Javascript* e cria as imagens através de HTML5/SVG o que lhe confere compatibilidade entre a maioria dos navegadores. Na Figura 4.3-2 um *screenshot* da galeria de estilos que a ferramenta pode exibir.

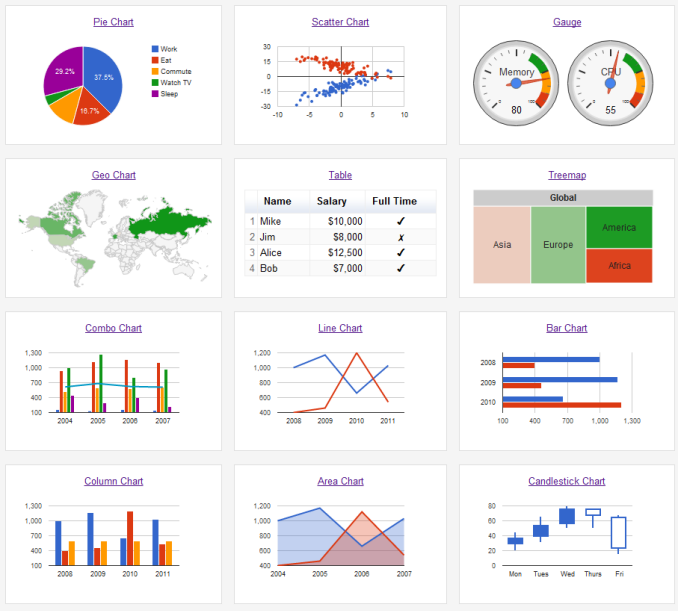


Figura 4.3-2 Google Chart e alguns exemplos de estilos que podem ser usados na visualização dos dados. Fonte: (GCHART, 2012).

4.3.5. JavaScript InfoVis Toolkit

O *JavaScript InfoVis Toolkit* (INFOVIS, 2011) é uma biblioteca em *Javascript* que cria visualizações interativas. As informações que vão compor os grafos devem ser apresentadas no formato JSON (*JavaScript Object Notation*) e carregadas via AJAX *(Asynchronous Javascript and XML)*. A biblioteca pode criar tipos avançados de grafos como TreeMaps, SpaceTrees, Hyperbolic Trees e Radial Trees, incluindo animações das mesmas. Algumas das possibilidades de visualização estão ilustradas na Figura 4.3-3.

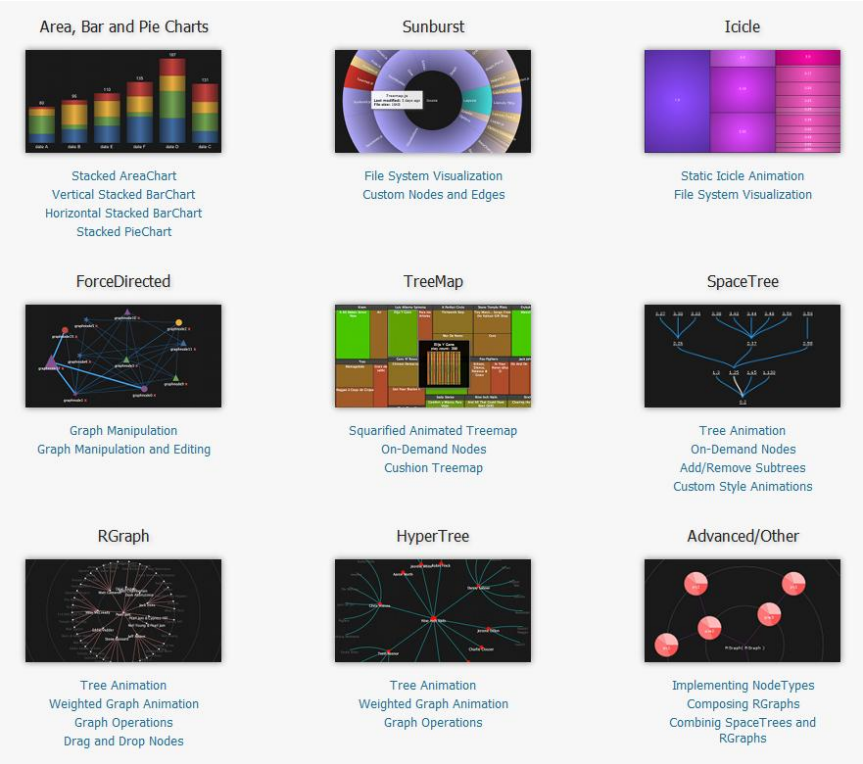


Figura 4.3-3 Alguns exemplos de visualizações que podem ser usadas no Infovis.
Fonte: http://thejit.org/demos/

4.3.6. Componente LeavesPHP

O componente LeavesPHP foi criado exclusivamente para o projeto desta dissertação. Surgiu da necessidade de se apresentar, em uma plataforma *web*, uma árvore de decisão (Seção 2.2.3) interativa, onde o usuário pudesse navegar por entre os ramos da árvore.

Foi criada uma função em PHP baseada no algoritmo C4.5 (Seção 2.2.3.1), na qual sua resposta é convertida para o formato JSON para que seja compatível com o InfoVis (Seção 4.3.5), deste modo é possível exibir a árvore que foi gerada *online*.

4.4. Arquitetura do Sistema

O Sistema BCIWeb (*Bacterial Classification and Identification for Web*) é composto pelo *Joomla CMS* conectado a um banco de dados relacional *MySQL*. O banco de dados foi criado de acordo com as necessidades deste trabalho, conforme será explicado na seção 4.5. O sistema tem como principais funcionalidades:

- Identificação *on-line* de bactérias através de testes bioquímicos
- Gerenciamento de pacientes, diagnósticos e prontuários
- Árvore de apoio à decisão
- Estatísticas
- Importação e exportação de resultados
- Compartilhamento colaborativo entre Instituições

Pelo *back-end* do *site* o administrador define os privilégios de acesso para cada grupo e gerencia os usuários cadastrados.

4.5. Modelagem do Banco de Dados

4.5.1. Levantamento de requisitos

Através de entrevitas com especialistas do Laboratório de Difteria e Corinebactérias de Importância Clínica (LDCIC) e de acordo com os objetivos estabelecidos no início do processo, define-se os atributos e relacionamentos dos campos que vão compor as tabelas da base de dados.

De acordo com as necessidades do sistema o levantamento de requisitos foi dividido em três etapas:

- Prontuários
- Tabela bioquímica padrão
- Registro de usuários

No processo de modelagem do banco de dados cada etada é minunciosamente discutida com os especialistas para que seja criado no banco de dados um modelo fiel do que já existe, ou seja, se trata da “virtualização” do trabalho existente. Na seção 4.5.2 será discutido cada etapa com mais detalhes e seu respectivo processo.

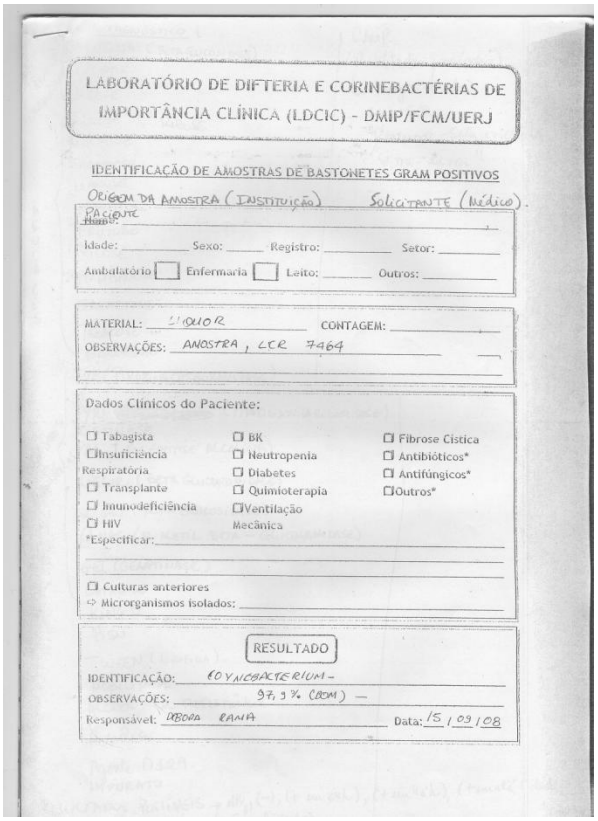
4.5.2. Criação da base de dados

As figuras desta seção seguem um modelo modelo E-R (Entidade-Relacionamento), onde são exibidas as entidades, seus atributos, e os relacionamentos entre as entidades, portanto, se trata de um modelo conceitual de alto nível com as principais características das tabelas (ELMASRI et al, 2005).

Nas seções a seguir são apresentados os processos de modelagem das principais etapas deste projeto.

4.5.2.1. Prontuários

As Figura 4.5-1 e Figura 4.5-2 apresentam digitalizações dos prontuários usados no LDCIC. A utilização do prontuário e o processo de identificação bacteriológica “manual” está detalhado na seção 3.3.3.



LABORATÓRIO DE DIFTERIA E CORINEBACTÉRIAS DE IMPORTÂNCIA CLÍNICA (LDCIC) - DMIP/FCM/UERJ

IDENTIFICAÇÃO DE AMOSTRAS DE BASTONETES GRAM POSITIVOS

ORIGEM DA AMOSTRA (INSTITUIÇÃO) SOLICITANTE (MÉDICO)

PACIENTE
Nome:
Idade: Sexo: Registro: Setor:
Ambulatório ☐ Enfermaria ☐ Leito: Outros:

MATERIAL: LÍQUOR CONTAGEM:
OBSERVAÇÕES: AMOSTRA, LCR 7464

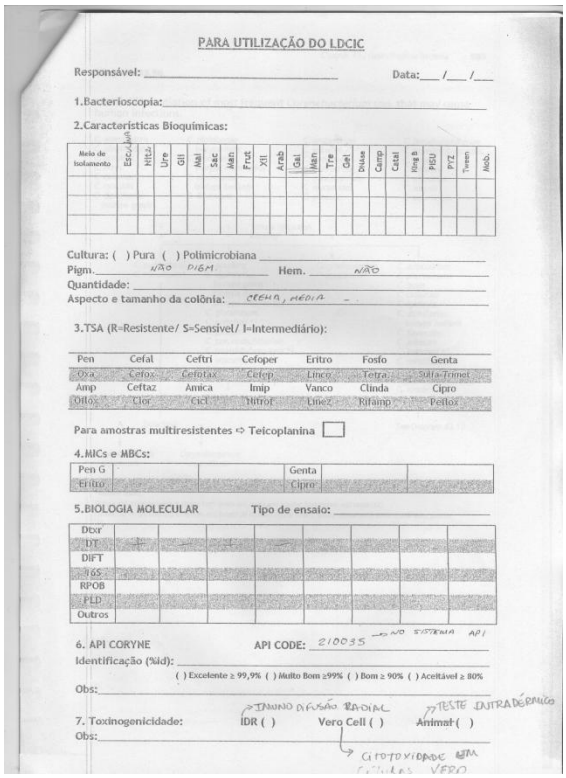
Dados Clínicos do Paciente:

☐ Tabagista
☐ Insuficiência Respiratória
☐ Transplante
☐ Imunodeficiência
☐ HIV
☐ BK
☐ Neutropenia
☐ Diabetes
☐ Quimioterapia
☐ Ventilação Mecânica
☐ Fibrose Cística
☐ Antibióticos*
☐ Antifúngicos*
☐ Outros*
*Especificar:

☐ Culturas anteriores
⇨ Microrganismos isolados:

RESULTADO
IDENTIFICAÇÃO: CORYNEBACTERIUM -
OBSERVAÇÕES: 97,1% (BOM) -
Responsável: DEBORA RANIA Data: 15/09/08

Figura 4.5-1 Digitalização da primeira folha do prontuário usada no laboratório da UERJ



PARA UTILIZAÇÃO DO LDCIC

Responsável: Data: ___/___/___

1.Bacterioscopia:

2.Características Bioquímicas:

Meio de Isolamento	Esc. Alcal	Hidr.	Ure	Gli	Mal	Sac	Man	Fruc	Xil	Arab	Gal	Inos	Tre	Cel	Dulse	Camp	Catal	Hog B	PIGU	PYZ	Tween	Mob.

Cultura: ( ) Pura ( ) Polimicrobiana
Pigm. NÃO DIGM. Hem. NÃO
Quantidade:
Aspecto e tamanho da colônia: CREMA, MEDIA

3.TSA (R=Resistente/ S=Sensível/ I=Intermediário):

Pen	Cefal	Ceftri	Cefoper	Eritro	Fosfo	Genta
Oxa	Cefox	Cefotax	Cefep	Linco	Tetra	Sulfa-Trimet
Amp	Ceftaz	Amica	Imip	Vanco	Clinda	Cipro
Oflox	Clor	Cicl	Nitrof	Linez	Rifamp	Peflox

Para amostras multiresistentes ⇨ Teicoplanina ☐

4.MICs e MBCs:

Pen G			Genta		
Eritro			Cipro		

5.BIOLOGIA MOLECULAR Tipo de ensaio:

Dcor									
DT									
DIFT									
16S									
RPOB									
FLD									
Outros									

6. API CORYNE API CODE: 210035 ⇨ NO SISTEMA API
Identificação (%id):
( ) Excelente ≥ 99,9% ( ) Muito Bom ≥99% ( ) Bom ≥ 90% ( ) Aceitável ≥ 80%
Obs:

7. Toxinogenicidade: IDR ( ) Vero Cell ( ) Animal ( )
Obs:

Figura 4.5-2 Digitalização da segunda folha do prontuário usada no laboratório da UERJ

A modelagem no banco de dados destes prontuários foi dividida em cinco tabelas levando em consideração a organização proposta pelos especialistas, a saber:

- Pacientes: Informaçõs gerais
- Pacientes: Anamnese
- Pacientes: Informações dos materiais
- Testes bioquímicos
- Antibiograma

As tabelas de Testes bioquímicos e Antibiograma serão abordadas na seção a seguir. A Figura 4.5-3 apresenta o modelo E-R das três primeiras tabelas do prontuário.

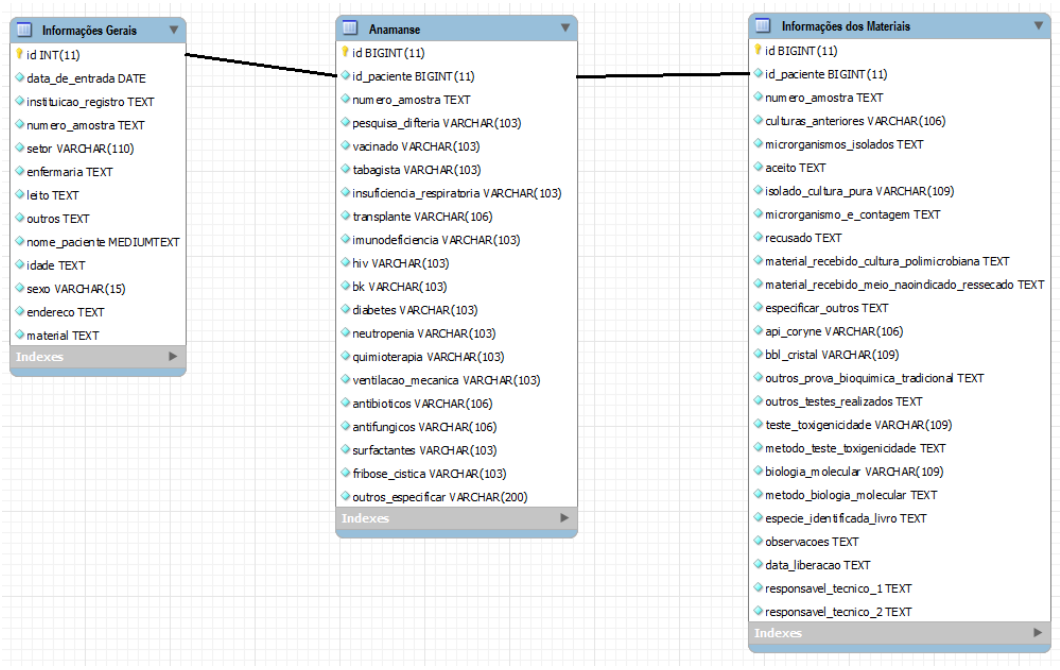


Figura 4.5-3 Diagrama E-R da primeira parte do prontuário

4.5.2.2. Testes bioquímicos e Antibiograma

Na Figura 4.5-4 está ilustrada a modelagem das tabelas de testes bioquímicos e antibiogramas.

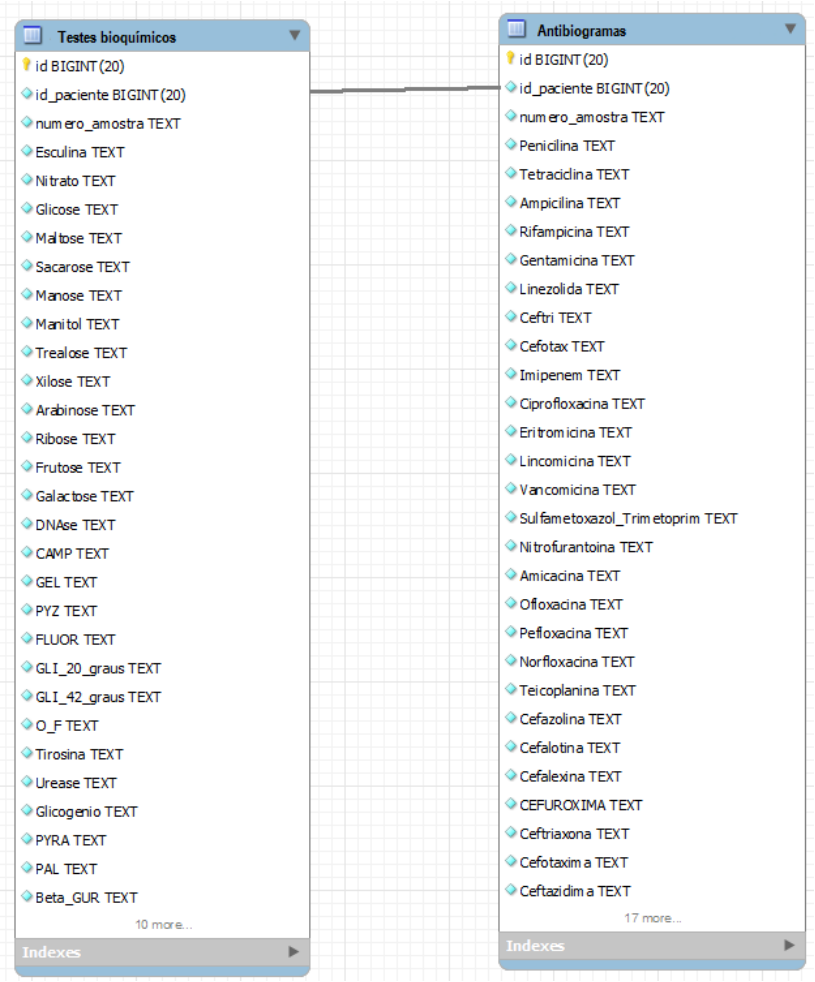


Figura 4.5-4 Diagrama E-R das tabelas de testes bioquímicos e antibiogramas

4.6. Desenvolvimento do Componente LeavesPHP

Conforme visto na seção 4.3.6, o algoritmo C4.5 foi implementado em PHP e em conjunto com o InfoVis (Seção 4.3.5) é capaz de gerar uma árvore de decisão *online* interativa.

A Tabela 4.6-1 fornecida por Lopes (2007) serviu de modelo para teste e validação do componente criado. Esta base de dados possui quatorze registros e quatro atributos, sendo um deles o atributo preditor.

Tabela 4.6-1 Dados médicos na predição de medicação.

Instância	Frequência Cardíaca	Frequência Respiratória	Perda do Apetite	Medicação (atributo preditor)
1	normocardico	taquipneia	nao	sim
2	normocardico	taquipneia	sim	sim
3	taquicardico	taquipneia	nao	sim
4	bradicardico	taquipneia	nao	nao
5	bradicardico	taquipneia	nao	sim
6	bradicardico	eupneia	sim	nao
7	taquicardico	eupneia	sim	sim
8	normocardico	taquipneia	nao	nao
9	normocardico	eupneia	nao	nao
10	bradicardico	eupneia	nao	nao
11	normocardico	eupneia	sim	nao
12	taquicardico	taquipneia	sim	nao
13	taquicardico	eupneia	nao	sim
14	bradicardico	taquipneia	sim	nao

O código e procedimento de entrada dos dados na função estão descritos no Anexo-B. A árvore criada pelo algoritmo C4.5 para os dados da Tabela 4.6-1 Tabela 4.6-1está representada na Figura 4.6-1 (LOPES, 2007), enquanto que na Figura 4.6-2 Árvore de decisão gerada pelo componente LeavesPHP encontra-se a árvore gerada pelo componente LeavesPHP para os mesmos dados. Nota-se que os mesmos resultados foram atingidos demonstrando a perfeita implementação do algoritmo.

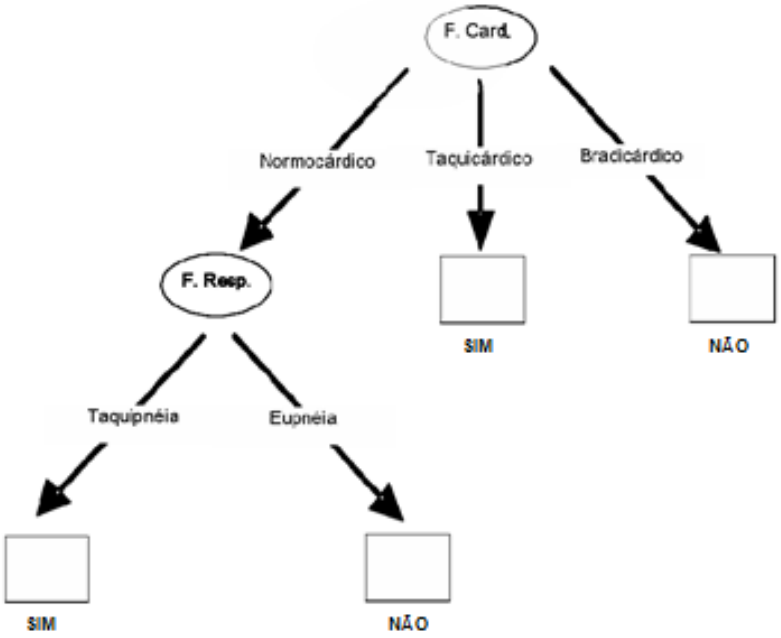


Figura 4.6-1 Árvore de decisão calculada pelo algoritmo C4.5. Fonte: Lopes (2007)

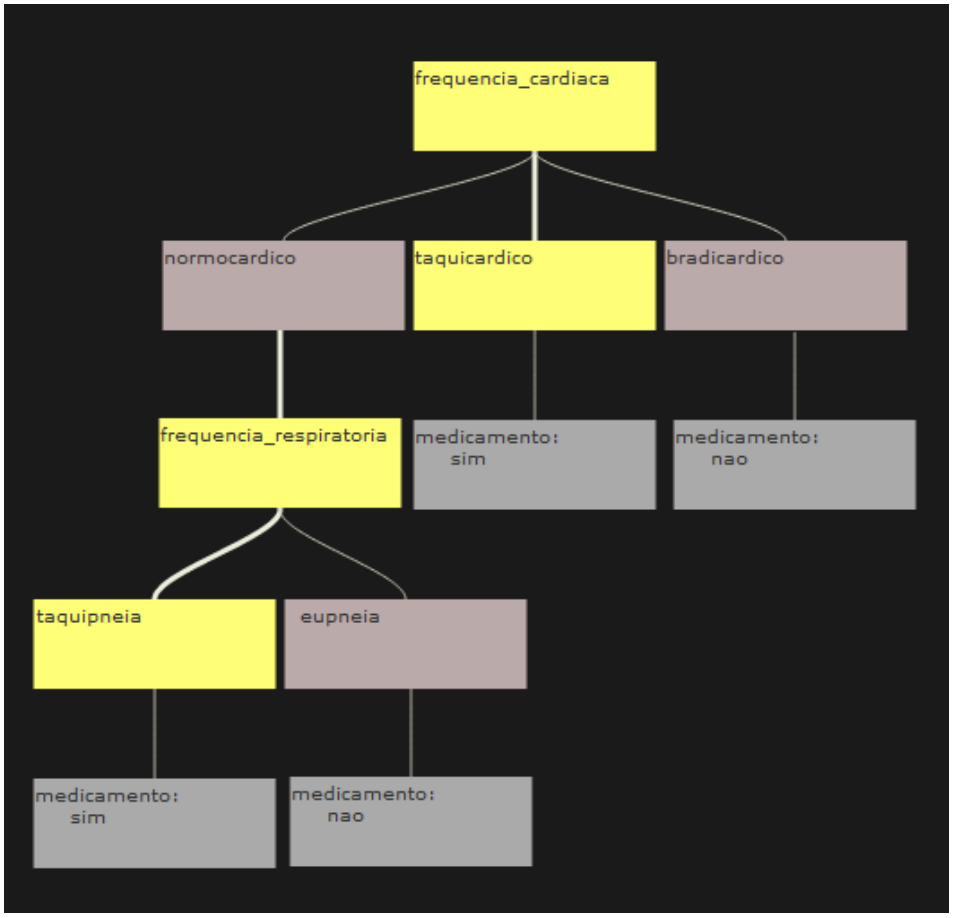


Figura 4.6-2 Árvore de decisão gerada pelo componente LeavesPHP

4.7. Método de comparação dos testes bioquímicos

Na seção 3.4 foi revisado o histórico de diversas metodologias de identificação e classificação de bactérias. O algoritmo desenvolvido para este sistema usa testes bioquímicos descritos em artigos científicos e capítulos de livros especializados por diferentes grupos de pesquisadores (ALMUZARA, 2006) (CAMELLO, 2003) (FUNKE, 2007) (THOMSON, 2007). Para cada espécie existe uma configuração padrão de resultados.

Porém, a simples comparação dos resultados dos testes bioquímicos da amostra em análise com a tabela de referência não garante a identificação precisa do microrganismo. No auxilio deste problema foram criados fluxogramas com os principais testes que influenciam a diferenciação das espécies do gênero *Corynebacterium.* A Figura 4.7-1 e Figura 4.7-2 apresentam digitalizações de dois fluxogramas.

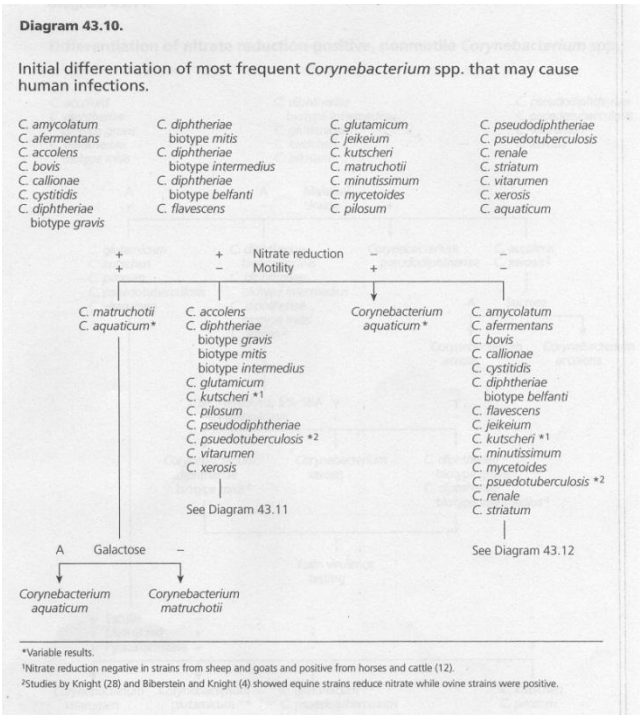


Figura 4.7-1 Digitalização do diagrama dos testes iniciais que auxiliam na diferenciação entre espécies do gênero *Corynebacterium.* Fonte: LDCIC.

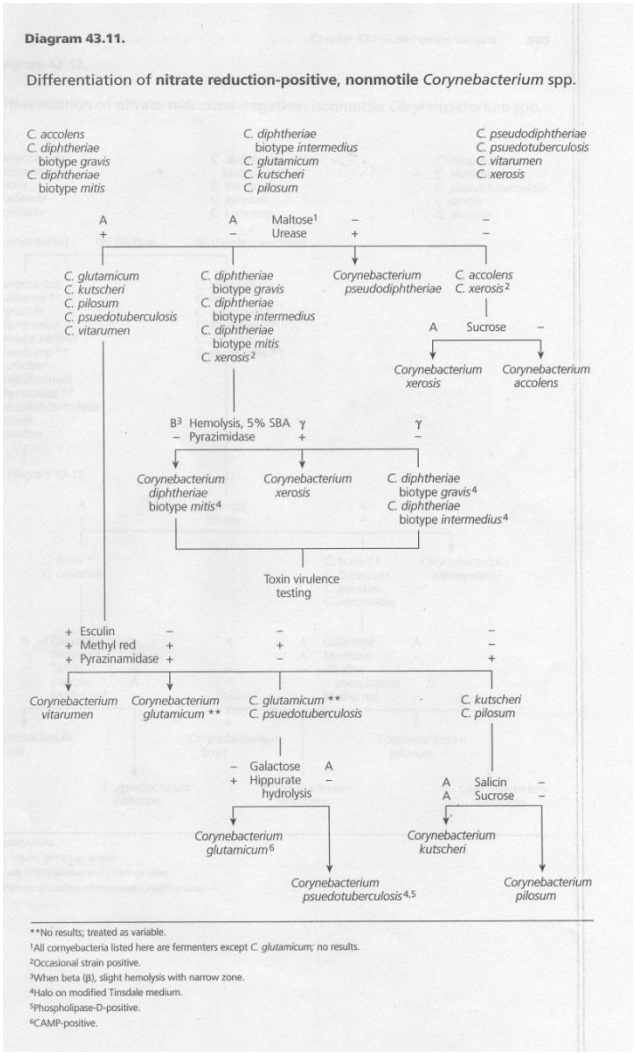


Figura 4.7-2 Digitalização do diagrama da continuação de testes que auxiliam na diferenciação entre espécies do gênero Corynebacterium. Fonte: LDCIC.

Portanto, no desenvolvimento e modelagem do algoritmo de comparação dos testes, os fluxogramas também foram levados em consideração.

4.8. Apresentação do Sistema

Nesta seção são apresentadas algumas telas e funcionalidades desenvolvidas no sistema.

4.8.1. Tela de login

A tela de *login* solitica ao usuário a entrada do seu nome de usuário e senha para efetuar o acesso ao sistema. A Figura 4.8-1 apresenta a página referente à esta tela. Grande parte das operações só são possíveis de serem realizadas por um usuário autenticado. Quem define os privilégios de cada usuário é o administrador do sistema, é possível que seja criado diversos níveis de controle de acesso por usuário.

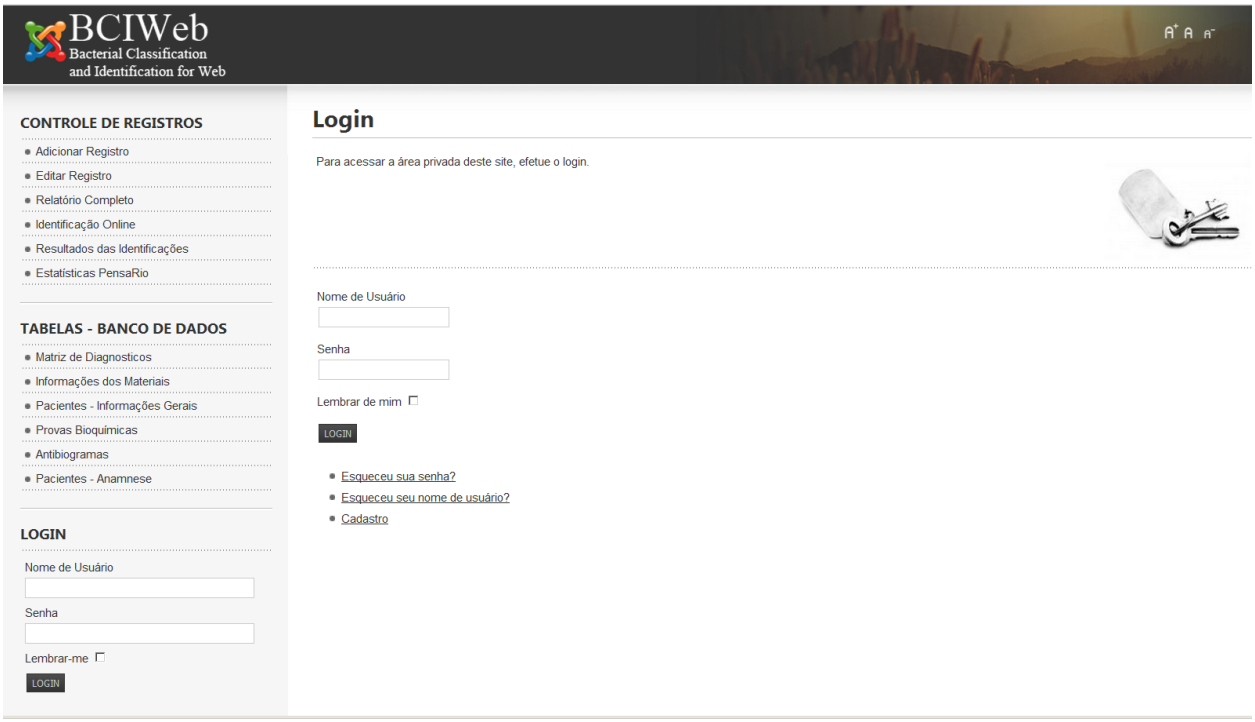


Figura 4.8-1 Tela de login do sistema. Através desta tela também é possível fazer o castro de usuários ou recuperar senhas.

4.8.2. Menu principal

Através do menu principal é possível ter acesso à todas as páginas do sistema. O menu sempre estará localizado ao lado esquerdo da tela conforme em destaque vermelho na Figura 4.8-2.

Figura 4.8-2 Menu principal em destaque vermelho.

4.8.3. Tela de cadastro de amostra

O processo de cadastro de amostra é dividido em cinco etapas. A tela inicial (Figura 4.8-3) solicita ao usuário a inserção das informações básicas do paciente, após o preenchimento dos campos o usuário prossegue com o cadastro clicando no botão “Continuar”.

Figura 4.8-3 Tela da primeira etapa no processo de adição de registro.

A segunda etapa (Figura 4.8-4) é relativa à informações de anamnese do paciente. Na terceira etapa (Figura 4.8-5) o usuário deve fornecer os dados sobre culturas anteriores, microrganismos isolados e informações gerais dos materiais.

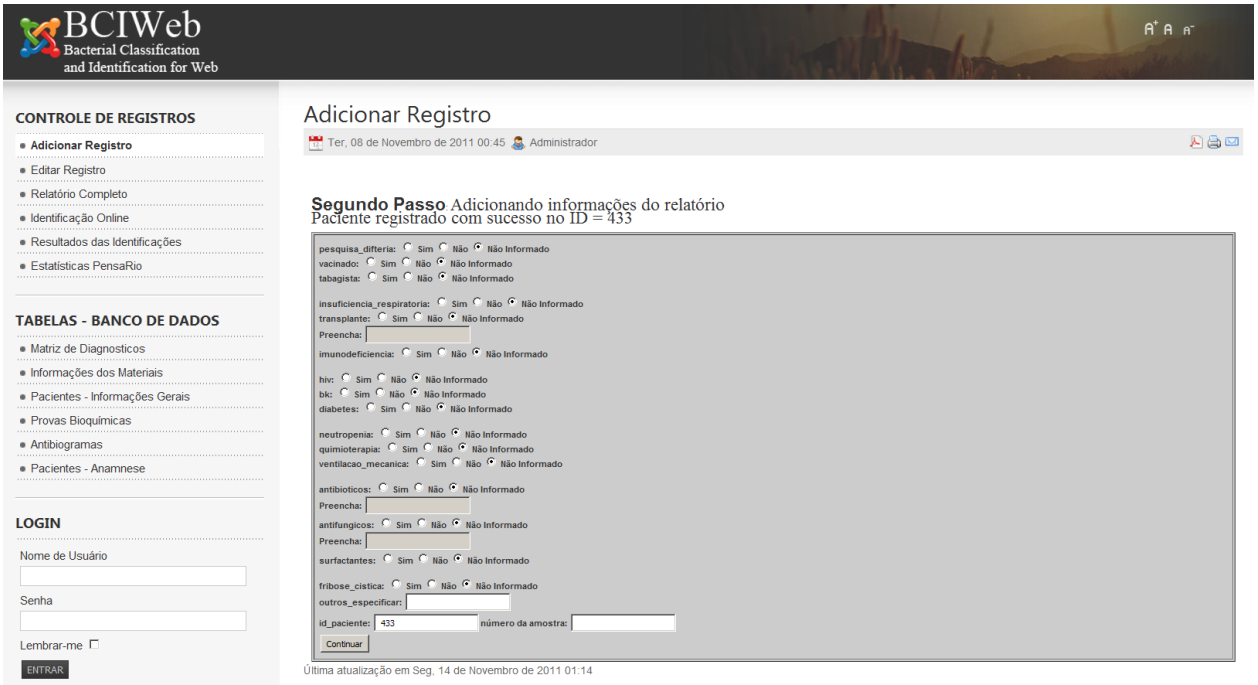


Figura 4.8-4 Tela da segunda etapa no processo de adição de registro.

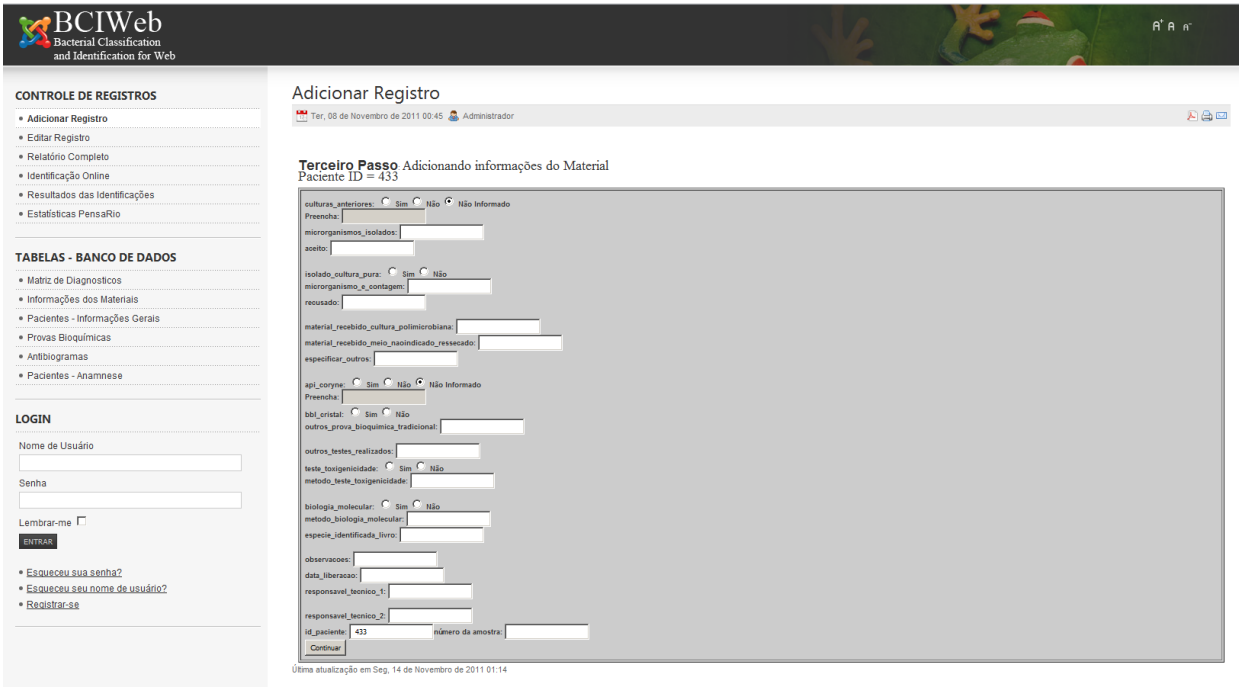


Figura 4.8-5 Tela da terceira etapa no processo de adição de registro.

Os resultados das provas bioquímicas da amostra que está sendo cadastrada devem ser inseridos na quarta etapa do processo (Figura 4.8-6). Nesta página todos os possíveis resultados dos campos são previamente conhecidos, por exemplo, para o teste de “Catalase” o usuário só pode escolher quatro opções de resultado: “não informado”, “+”, “-“ e “(+)”.

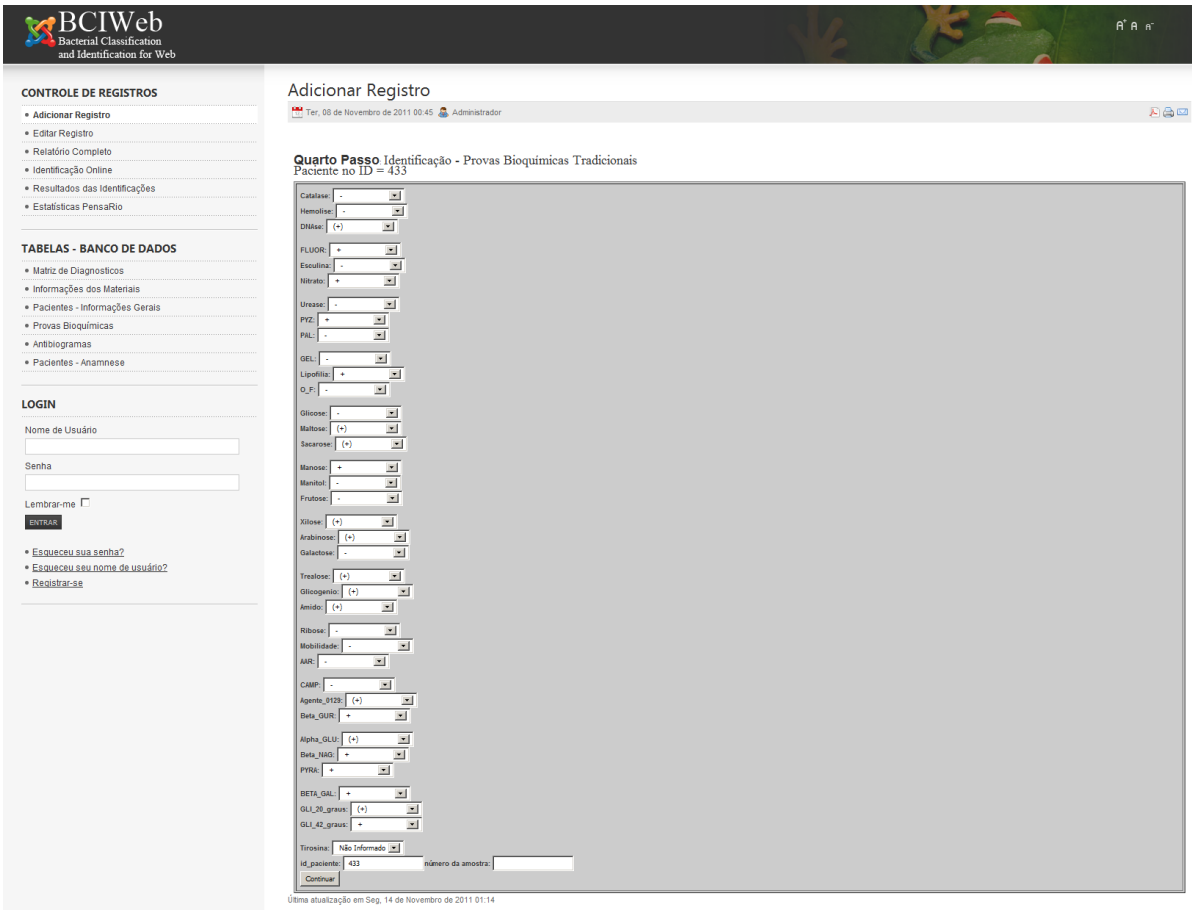


Figura 4.8-6 Tela da quarta etapa no processo de adição de registro.

Na seguinte e última etapa (Figura 4.8-7), o usuário deve fornecer os resultados do teste de antibiograma da amostra.

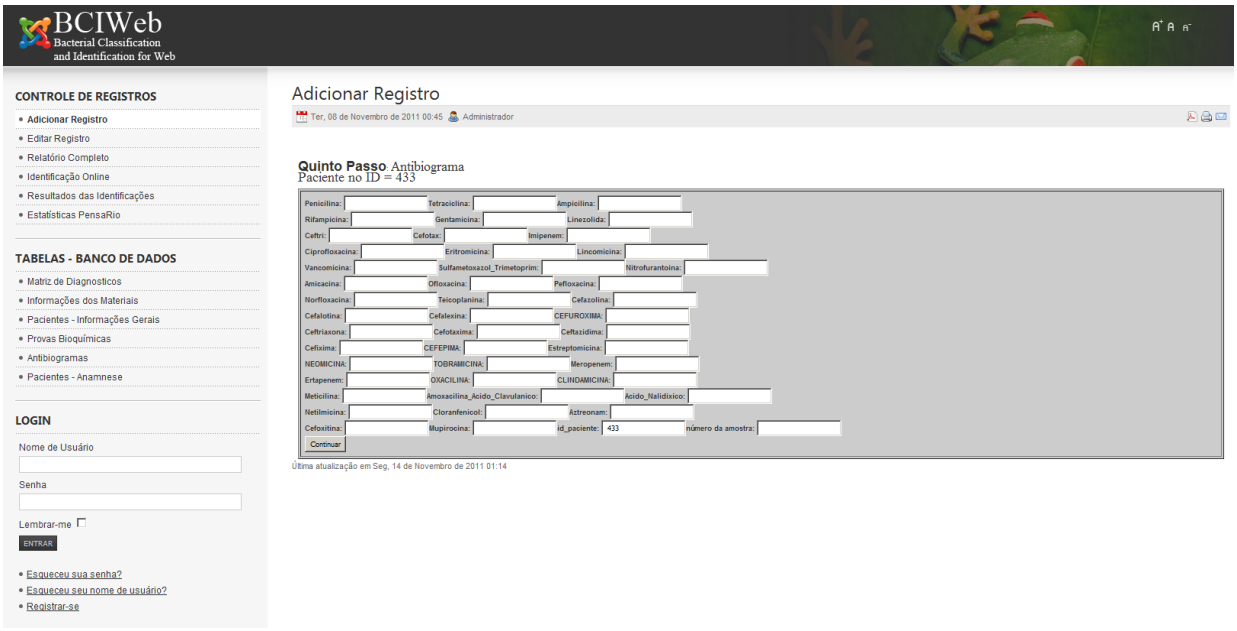


Figura 4.8-7 Tela da quinta etapa no processo de adição de registro.

Ao clicar em "Continuar" o cadastro é finalizado e o usuário é redirecionado automaticamente para a página de identificação online, onde os resultados dos testes bioquímicos recém inseridos são colocados à prova (Figura 4.8-8).

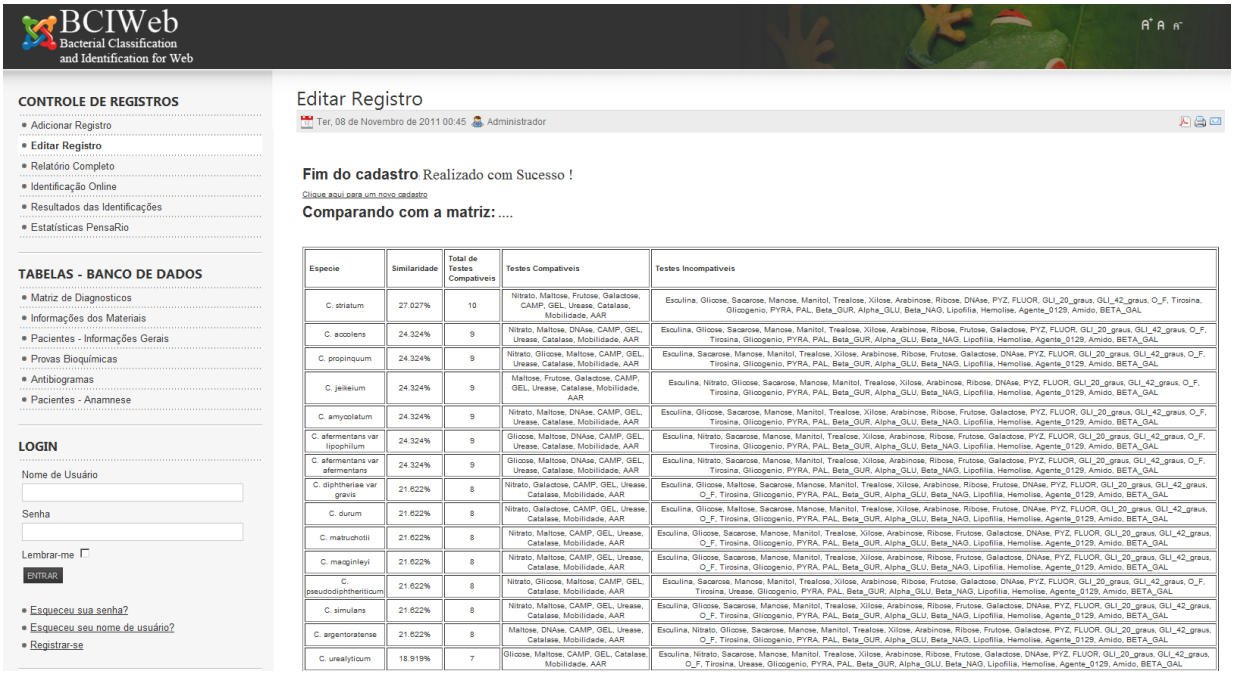


Figura 4.8-8 Tela da última etapa no processo de adição de registro. Ao final é realizada a identificação do registro recém adicionado.

4.8.4. Tela de edição de amostra

O acesso à página de edição de amostras é realizado através do menu principal na lateral esquerda. Inicialmente na tela de edição de amostras (Figura 4.8-9), o usuário deve escolher a amostra de interesse que deseja editar. Feita a escolha na caixa de seleção e clicando em "Continuar" o sistema o redirecionará à tela com as informações relativas à primeira etada do cadastro da amostra (Figura 4.8-10), a partir dai pode-se realizar alterações em todos as etapas da amostra escolhida.

Figura 4.8-9 Tela de edição de registros. Nesta primeira tela o usuário deve selecionar o registro que deseja editar.

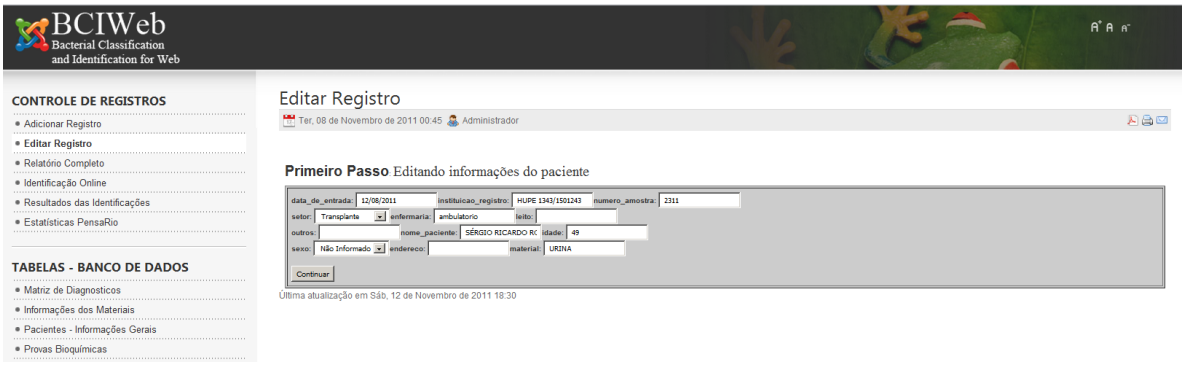


Figura 4.8-10 Segunda tela de edição de amostra. Nesta etapa o registro de interesse já foi selecionado e suas informações são exibidas para alterações.

4.8.5. Tela de identificação online de bactéras

Após toda adição de amostra, a identificação da mesma é automaticamente realizada. O resultado, bem como algumas informações relevantes ao processo de

identificação, são registradas junto ao registro da amostra.

Na tela de identificação *online* (Figura 4.8-11) é possível conferir rapidamente o resultado de uma sequência de testes bioquímicos avulso. Por ser uma página de consulta, nenhum resultado de identificação é registrado na base de dados.

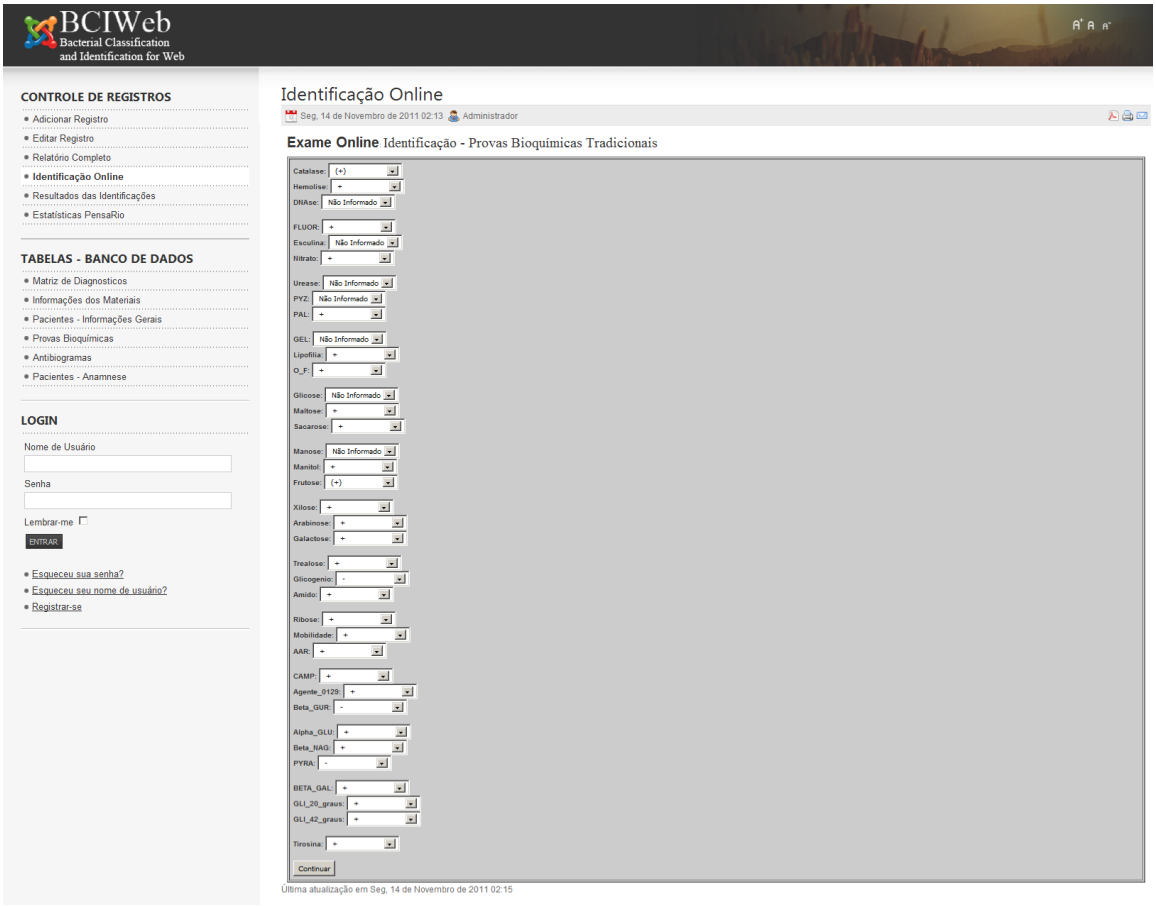


Figura 4.8-11 Tela de identificação on-line de bactérias. São apresentadas para preenchimento todas as provas bioquímicas cadastradas.

4.8.6. Tela de relatório

Na tela inicial (Figura 4.8-12) o usuário deve selecionar quais campos devem constar no relatório que será gerado, é possível escolher e unificar campos das mais diversas tabelas da base de dados.

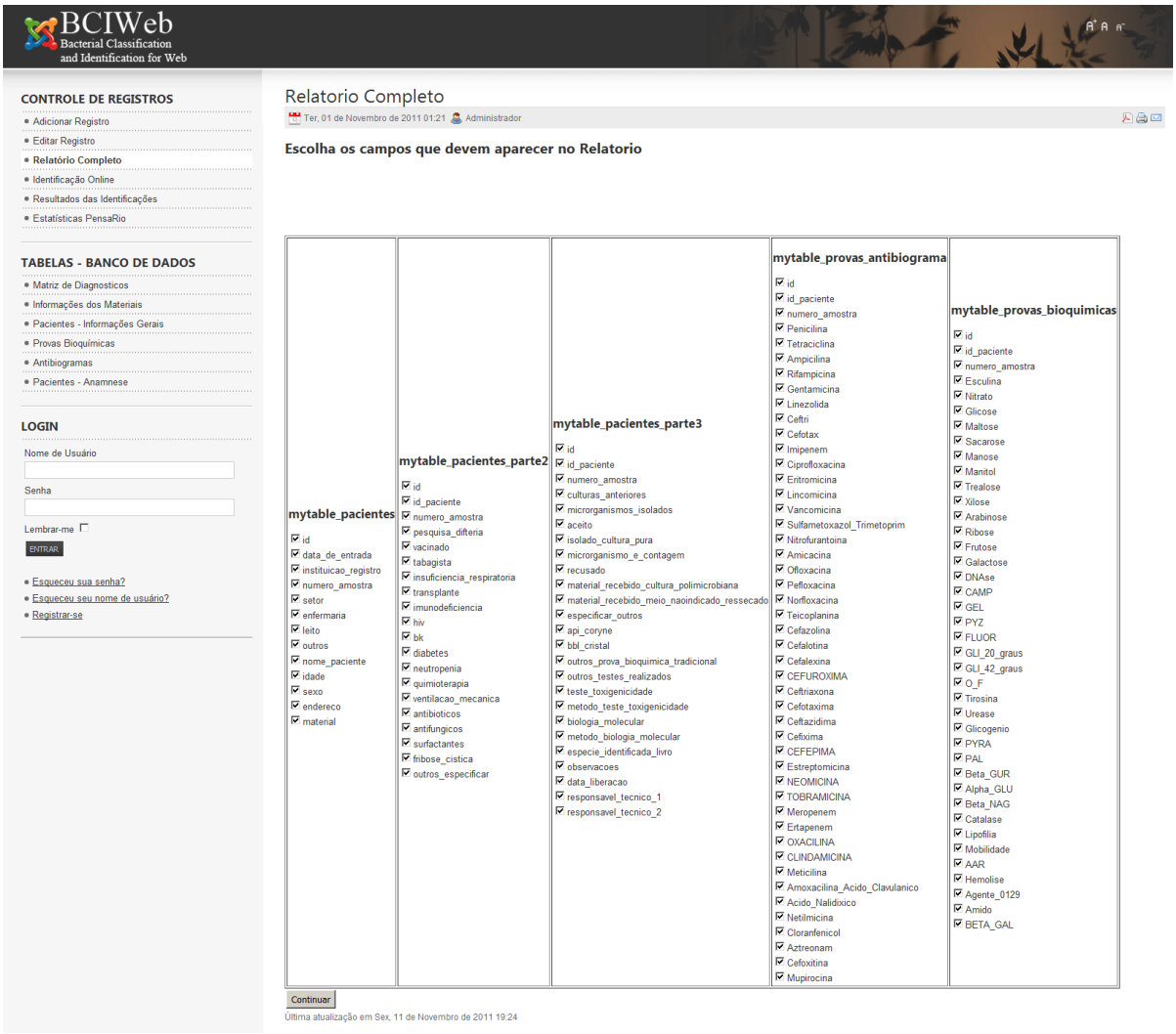


Figura 4.8-12 Tela de configuração para exibição do relatório completo. É possível escolher qualquer campo de qualquer uma tabelas para exibição em conjunto.

4.8.7. Tela de estatísticas

As estatísticas da base de dados do sistema são atualizadas a cada exibição de página. Na tela (Figura 4.8-16) é possível conferir diversos tipos de informações sobre as amostras, pacientes, identificações, incidências por datas e setores etc.

Todos os gráficos são interativos, permitindo ao usuário navegador por entre os dados e organiza-los da melhor forma que lhe convir como em ordem decrescente ou crescente. Nesta página destacam-se o gráfico de distribuição de

bactérias (Figura 4.8-13), gráfico dos números de ocorrências de bactérias distribuídas nos meses dos anos (Figura 4.8-14) e dois mapas (Figura 4.8-15) que exibem círculos, de variados tamanhos, nos locais onde o número de ocorrências se tornam relevantes, este número é configurável e o tamanho do círculo é proporcional ao número de casos no local.

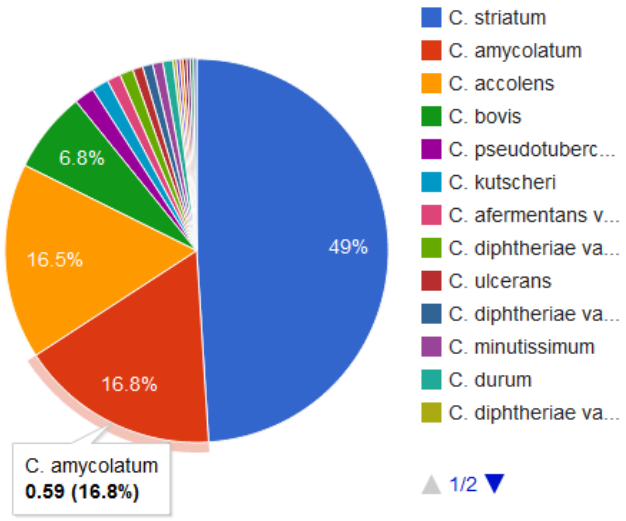


Figura 4.8-13 Gráfico de distribuição de bactérias da base de dados, localizado na página de estatísticas.

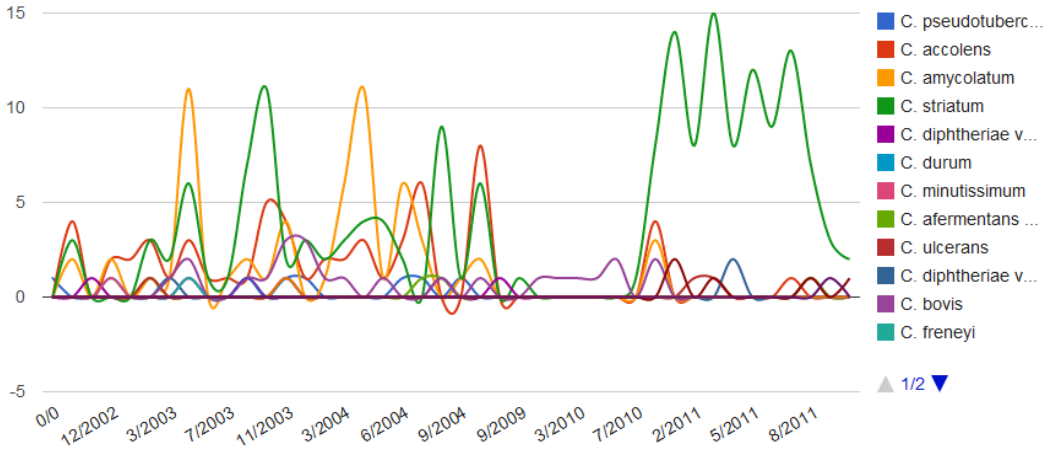


Figura 4.8-14 Gráfico dos números de ocorrências de bactérias distribuídas nos meses dos anos, localizado na página de estatísticas.

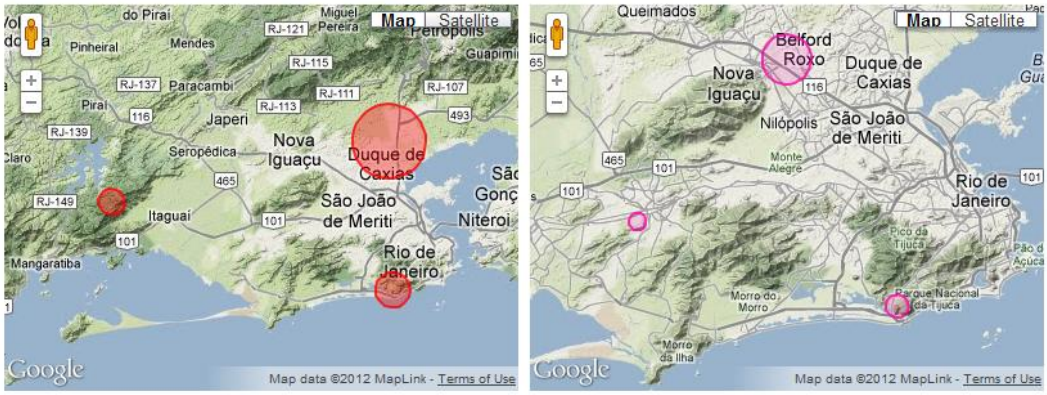


Figura 4.8-15 Mapas com raios de ocorrência. (A) Mapa com os focos de incidência relativos às residencias dos pacientes, (B) Mapa com os focos de incidência relativos aos locais de trabalho dos pacientes.

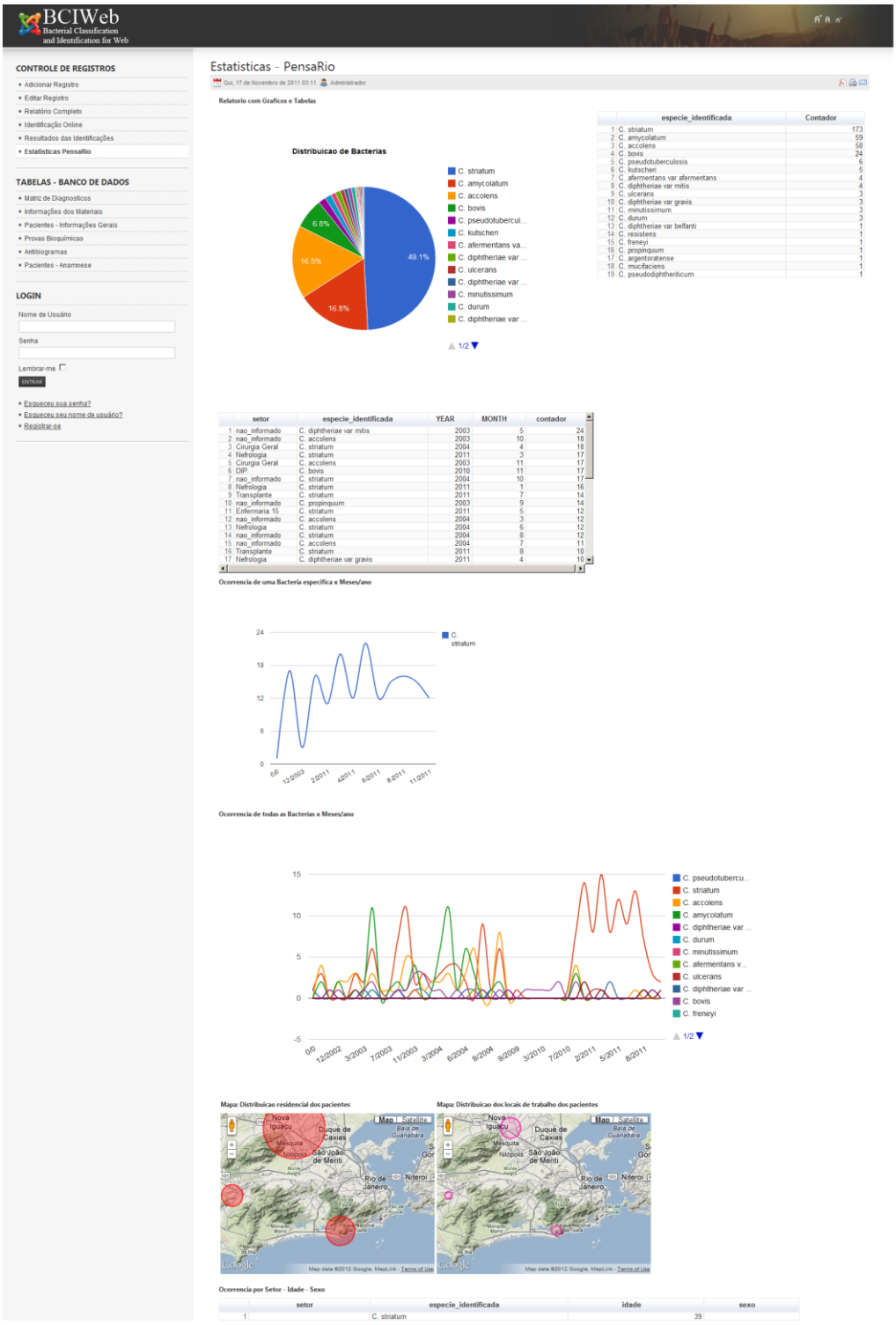


Figura 4.8-16 Tela completa onde se exibe os gráficos informativos e as estatísticas do sistema.

4.8.8. Tela de Árvore de Decisão

A página de árvore de decisão utiliza o componente *LeavesPHP*, apresentado na seção 4.3.6, para criar a árvore de decisão da base de dados de referência do sistema. Desse modo será criada uma árvore com os testes bioquímicos da tabela de referência, que tem como resultado o número de identificação das espécies (ANEXO C).

Na Figura 4.8-17 está ilustrado um exemplo de navegação no componente, onde o usuário selecionou os testes bioquímicos (apresentados em blocos amarelos):

- Manose: +
- Arabinose: -
- Ribose: +

Para esta configuração de testes bioquímicos, o resultado é a bactéria com número de identificação igual a doze (C. coyleae). Na Figura 4.8-18 é apresentado outro exemplo, onde foi feita a seleção dos testes bioquímicos:

- Manose: -
- Frutose: +
- Galactose: +

E portanto, obteve como resultado a bactéria com número de identificação igual a dez (C. bovis).

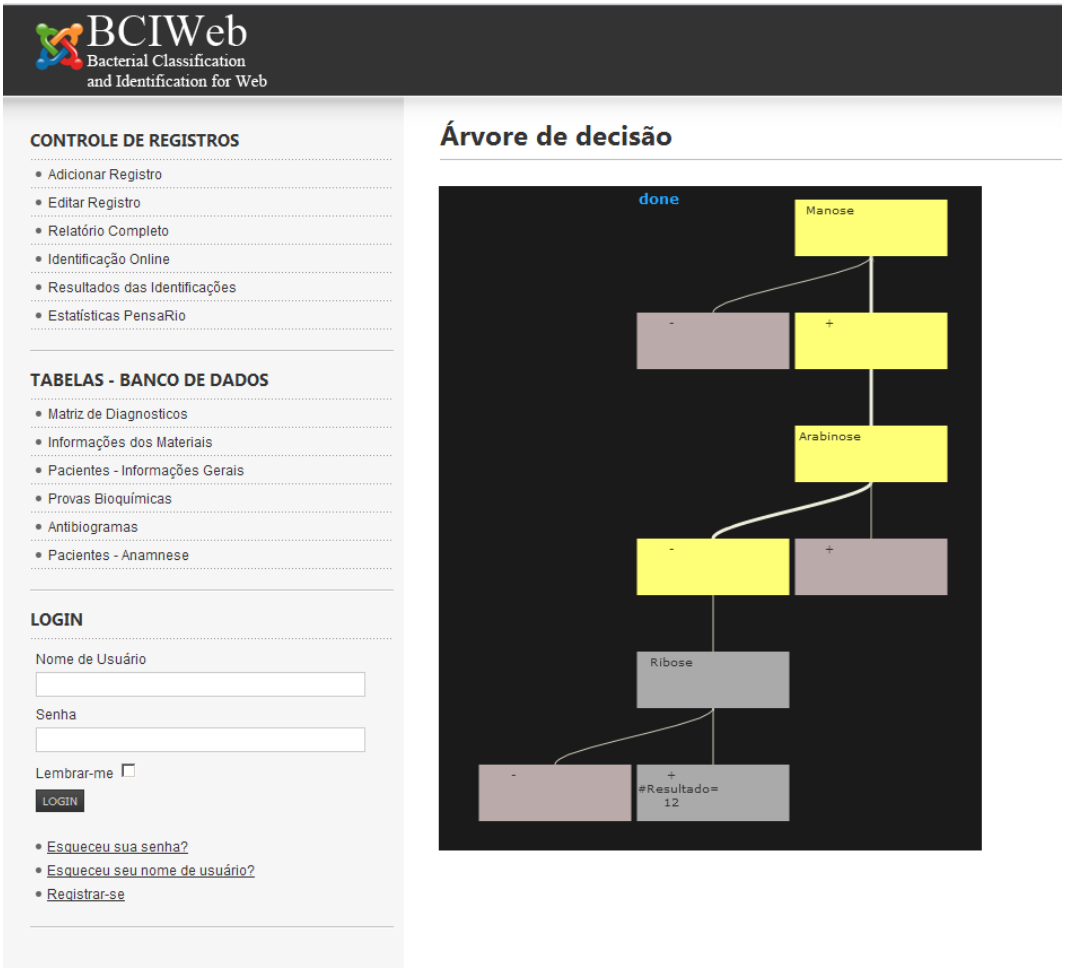


Figura 4.8-17 Página da árvore de decisão da base de dados de referência.

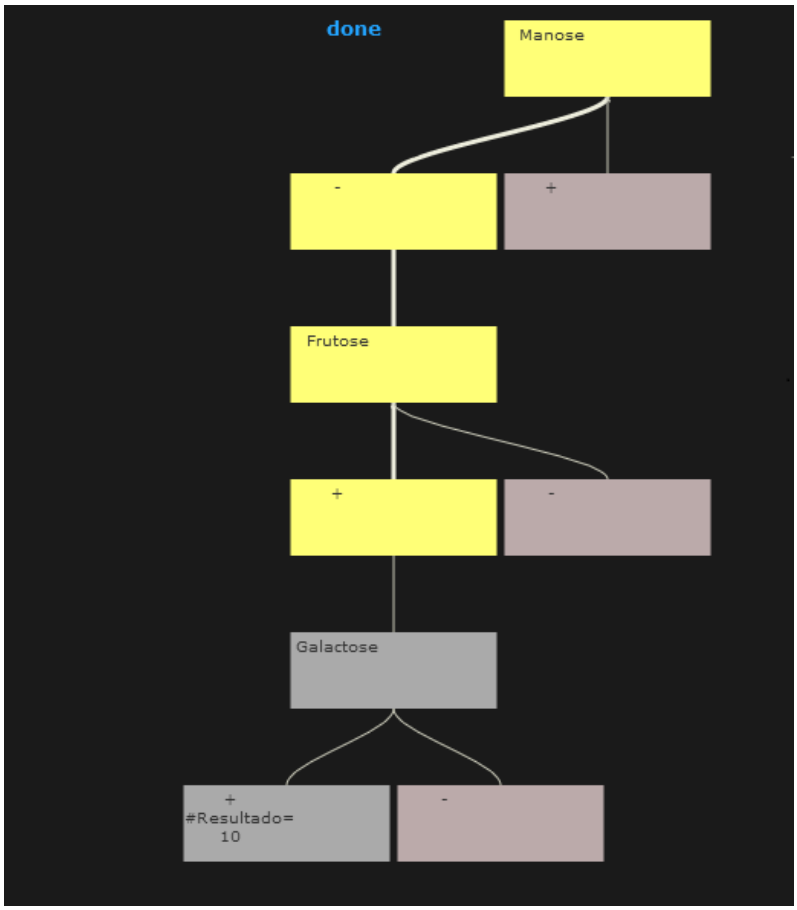


Figura 4.8-18 Recorte da árvore de decisão onde a configuração de testes bioquímicos leva ao resultado de identificação numérica igual a dez (C. bovis).

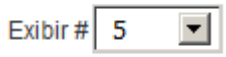
4.8.9. Páginas de administração das tabelas da base de dados

O acesso às páginas de administração de todas as tabelas da base de dados do sistema é feito através do menu principal. Conforme visto na seção 4.8.3, o cadastro das amostras é realizado em cinco etapas, cada etapa está relacionada a uma tabela da base de dados, o acesso a elas é nomeado como:

- Pacientes - Informações gerais
- Pacientes – Anamnese
- Informações dos materiais
- Provas bioquímicas
- Antibiogramas

As tabelas da base de dados não se restrigem apenas ao cadastro de amostras. Ao fim do cadastro de uma amostra seu registro de provas bioquímicas é posto à prova na identificação online e este resultado é inserido na tabela "Resultados de identificações". A matriz de referência para a identificação, explicada na seção 4.7, se encontra na tabela "Matriz de diagnósticos".

Todas as páginas de administração seguem um *layout* básico (Figura 4.8-19), suas principais funcionalidades são:

-  Adicionar registro
-  Editar registro
-  Excluir registro
- Importar ou exportar toda a tabela ou registros selecionados
- 

Caixa de seleção de quantos registros devem ser exibidos por página

A última funcionalidade de destaque é a possibilidade de se pesquisar termos de interesse em qualquer campo da tabela (Figura 4.8-20) , sendo uma busca *online* com tecnologia AJAX *(Asynchronous Javascript and XML).*

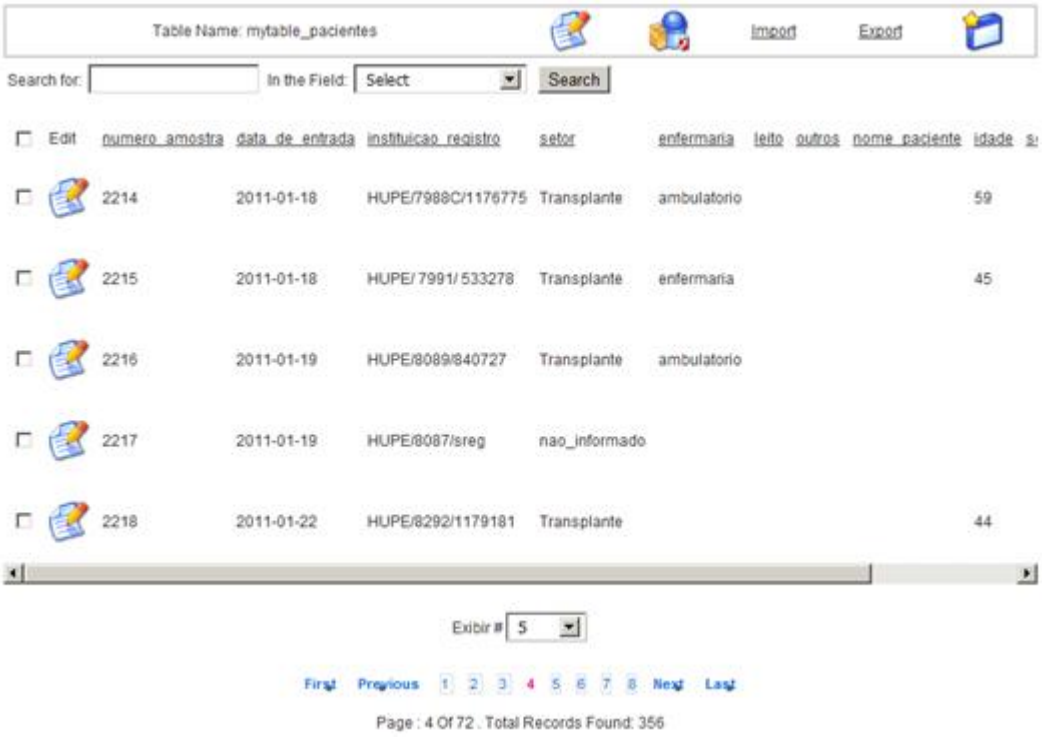


Figura 4.8-19 Página de administração da tabela de informações gerais dos pacientes.

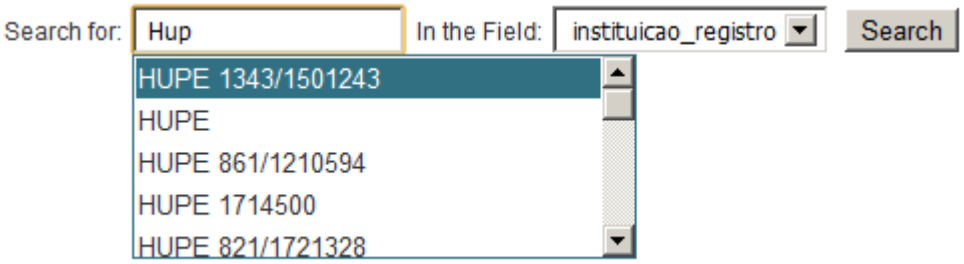


Figura 4.8-20 Destaque de uma parte da página de administração de tabelas mostrando a funcionalidade de busca pelo termo "Hup" no campo "instituição de registro" da tabela de informações gerais dos pacientes.

5 Estudo de Caso UERJ

Este capítulo busca aplicar alguns dos métodos de mineração de dados apresentados nos capítulos anteriores, com ênfase em RNAs (Seção 2.2.1) e SOM (Seção 2.2.2). Neste estudo de caso são analisados dois conjuntos de dados sendo um público e outro proveniente do LDCIC.

É importante destacar que, no contexto desta dissertação, a classificação de bactérias têm um viés de agrupamento, pois, no treinamento da rede não foram usadas informações sobre os grupos. Após o treino, as informações sobre os grupos foram aplicadas nos mapas para análise da relevância do método não-supervisionado no agrupamento de espécies e em como o método pode ajudar na proposta de novas separações de espécies.

Conforme visto na Capítulo 2, a descoberta do conhecimento em bases de dados envolve uma sequência de tarefas, sendo que seu nível de complexidade aumenta conforme as perguntas que devem ser respondidas e a dimensionalidade dos dados, o que é o caso dos conjuntos de dados dessa dissertação. Por isso, para que se chegue a resultados relevantes e interpretáveis é fundamental a preparação dos dados de entrada.

O *software* MATLAB foi utilizado para integração, processamento dos dados e visualização dos gráficos. Através do SOM *Toolbox* (VESANTO, 2000), em conjunto com MATLAB, foram gerados os mapas auto-organizáveis e feitas as análises exploratórias dos mapas.

5.1. Conjuntos de dados

Os conjuntos de dados que foram utilizados nesta dissertação são complementares. Os dados públicos foram obtidos de Koneman (2001) e Murray (2007) e os provenientes do LDCIC se encontram na base de dados do sistema BCIWeb na tabela de testes bioquímicos.

A complementaridade dos dois conjuntos, reside no fato de ambos serem resultados de testes bioquímicos. Os conjuntos são compostos por espécies de microrganismos e seus respectivos resultados para uma bateria de provas

bioquímicas. Conforme visto na seção 4.7, a etapa inicial na identificação bacteriológica é a comparação dos resultados da amostra de interesse com uma tabela de referência, neste caso, a tabela de referência é o conjunto de dados público provenientes de Koneman (2001) e Murray (2007). Portanto, esta tabela, é um referenicial seguro, afinal, são testes bioquímicos relativos à espécies que foram comprovadas genotipicamente.

O conjunto de dados do LDCIC, é formado inicialmente apenas pelos testes bioquímicos das amostras que chegam até o laboratório, não é sabido até então qual a espécie do microrganismo. Após o processo de identificaçao da bactéria (Seção 3.2) é chegada a uma conclusão, então seu registro é editado e atualizado com o seu resultado.

Os dois conjuntos possuem um total de 37 atributos (Tabela 5.1-1). O conjunto público possui 51 registros únicos, são portanto, 51 espécies de microrganismos pertencentes ao gênero Corynebacterium, no caso do conjunto do LDCIC, são 354 amostras de espécies diversas.

Tabela 5.1-1 Atributos dos conjuntos públicos e do LDCIC.

Identificação	**Atributo**	**Identificação**	**Atributo**
1	Esculina	19	GLI_20_graus
2	Nitrato	20	GLI_42_graus
3	Glicose	21	O_F
4	Maltose	22	Tirosina
5	Sacarose	23	Urease
6	Manose	24	Glicogenio
7	Manitol	25	PYRA
8	Trealose	26	PAL
9	Xilose	27	Beta_GUR
10	Arabinose	28	Alpha_GLU
11	Ribose	29	Beta_NAG
12	Frutose	30	Catalase
13	Galactose	31	Lipofilia
14	DNAse	32	Mobilidade
15	CAMP	33	AAR
16	GEL	34	Hemolise
17	PYZ	35	Agente_0129
18	FLUOR	36	Amido
		37	BETA_GAL

5.2. Pré-processamento dos Dados

Todos os 37 atributos dos conjuntos são discretos. Na Tabela 5.2-1 estão listados os atributos e seus possíveis valores. O formato dos dados na entrada dos modelos apresentados precisam ser numéricos, portanto, os valores dos atributos foram convertidos (Tabela 5.2-2) para que fossem compatíveis com o formato de entrada.

Tabela 5.2-1 Atributos e seus possíveis valores

Identificação	Atributo	Valores	Identificação	Atributo	Valores
1	Esculina	+ (+) - V	19	GLI_20_graus	+ (+) - V
2	Nitrato	+ (+) - V	20	GLI_42_graus	+ (+) - V
3	Glicose	+ (+) - V	21	O_F	O F
4	Maltose	+ (+) - V	22	Tirosina	+ (+) - V
5	Sacarose	+ (+) - V	23	Urease	+ (+) - V
6	Manose	+ (+) - V	24	Glicogenio	+ (+) - V
7	Manitol	+ (+) - V	25	PYRA	+ (+) - V
8	Trealose	+ (+) - V	26	PAL	+ (+) - V
9	Xilose	+ (+) - V	27	Beta_GUR	+ (+) - V
10	Arabinose	+ (+) - V	28	Alpha_GLU	+ (+) - V
11	Ribose	+ (+) - V	29	Beta_NAG	+ (+) - V
12	Frutose	+ (+) - V	30	Catalase	+ (+) - V
13	Galactose	+ (+) - V	31	Lipofilia	+ (+) - V
14	DNAse	+ (+) - V	32	Mobilidade	+ (+) - V
15	CAMP	+ (+) - V REV	33	AAR	+ (+) - V
16	GEL	+ (+) - V	34	Hemolise	+ (+) - V
17	PYZ	+ (+) - V	35	Agente_0129	+ (+) - V S R
18	FLUOR	+ (+) - V	36	Amido	+ (+) - V
			37	BETA_GAL	+ (+) - V

Tabela 5.2-2 Conversão numérica dos possíveis valores dos atributos.

Valor	Valor convertido
+	0.9
-	0.01
(+)	0.80
V	0.5
REV	2
O	3
F	4
S	5
R	6

Tabela 5.2-3 Quantidade de registros inválidos para cada atributo que foi removido.

Atributo	Registros nulos
Glicogenio	100
PYZ	72
PYRA	100
PAL	100
Beta_GUR	100
Alpha_GLU	100
Beta_NAG	100
FLUOR	72
Lipofilia	95
Hemolise	99
Agente_0129	100
Amido	100
BETA_GAL	100
GLI_20_graus	100
GLI_42_graus	100
O_F	100
Tirosina	100

A próxima etapa consiste na validação dos atributos, onde é avaliada sua integridade, média e desvio padrão dos valores. No conjunto público, devido à natureza bem comportada dos dados não houve necessidade de remoção ou alteração de qualquer atributo, em contrapartidada, para o conjunto de dados do LDCIC houve a necessidade de remoção de 17 atributos (Tabela 5.2-3), pois não apresentavam valores válidos. Este conjunto de dados teve sua quantidade de registros reduzida para 100 amostras, pois o restante dos registros não apresentavam a identificação bacteriológica das respectivas amostras.

Neste estudo de caso, pouco se sabe sobre a importância de cada atributo para seu respectivo conjunto. Por isso, para evitar que um atributo qualquer tenha maior relevância no agrupamento e que um atributo importante não tenha seu peso minimizado, é necessário aplicar o processo de normalização dos dados. Este processo consiste na aplicação de uma transformação linear em todos os valores de cada atributo. Com a ferramenta SOM *Toolbox* foram estudados dois tipos de normalização, sendo eles:

- *Método "var":* A variância do atributo é normalizada para um.
- *Método "range":* Os valores são normalizados entre zero e um.

Para os dois métodos foram realizados diversos testes, afim de se verificar, qual deles é o mais eficaz para estes conjuntos de dados. Segundo Zuchini (2003), no caso de SOM, uma boa heurística na seleção da melhor configuração de mapa, é selecionar os três mapas com menor erro topográfico TE (*Topographic Error*), dentre eles, o que apresentar valor intermediário de erro de quantização QE (Quantization Error) é o mapa eleito. Valores com QE ou TE iguais a zero devem ser desconsiderados devido à possibilidade de sobre-ajuste ou sub-ajuste.

Para o conjunto de dados públicos, os testes estão distribuídos nas Tabela 5.2-4 e Tabela 5.2-5Tabela 5.2-5, onde são apresentados os resultados do método "var" e "range" respectivamente. O método "range", apresentou em todas as configurações, um melhor desempenho.

Tabela 5.2-4 Resultados de QE e TE para diversas configurações de mapas, suas dimensões e especificações das fases de treinamento. Usando conjunto de dados público completo. Método de normalização "var".

Configurações						
Dimensões	Fase 1		Fase 2		QE	TE
	Épocas	Raio	Épocas	Raio		
7 x 5	7	1 -> 1	28	1 -> 1	3.987412	0.019608
10 x 10	20	2 -> 1	79	1 -> 1	2.975499	0.019608
14 x 10	28	2 -> 1	110	1 -> 1	2.554777	0.000000
15 x 15	45	2 -> 1	177	1 -> 1	1.578495	0.000000
20 x 20	79	3 -> 1	314	1 -> 1	0.467534	0.000000
25 x 25	123	4 -> 1	491	1 -> 1	0.082950	0.000000
27 x 27	143	4 -> 1	572	1 -> 1	0.040710	0.039216
28 x 28	154	4 -> 1	615	1 -> 1	0.028325	0.039216
29 x 29	165	4 -> 1	660	1 -> 1	0.015197	0.019608
30 x 30	177	4 -> 1	706	1 -> 1	0.005955	0.000000
32 x 32	201	4 -> 1	804	1 -> 1	0.009209	0.039216
32 x 30	189	4 -> 1	753	1 -> 1	0.007520	0.019608
34 x 34	227	5 -> 1.25	907	1.25 -> 1	0.000688	0.000688
34 x 30	200	5 -> 1.25	800	1.25 -> 1	0.005478	0.058824
36 x 36	255	5 -> 1.25	1017	1.25 -> 1	0.000312	0.039216
40 x 40	314	5 -> 1.25	1255	1.25 -> 1	0.000011	0.019608

Tabela 5.2-5 Resultados de QE e TE para diversas configurações de mapas, suas dimensões e especificações das fases de treinamento. Usando conjunto de dados público completo. Método de normalização "range".

Configurações						
Dimensões	Fase 1		Fase 2		QE	TE
	Épocas	Raio	Épocas	Raio		
7 x 5	7	1 -> 1	28	1 -> 1	1.520560	0.000000
10 x 10	20	2 -> 1	79	1 -> 1	1.110101	0.000000
14 x 10	28	2 -> 1	110	1 -> 1	0.890238	0.000000
15 x 15	45	2 -> 1	177	1 -> 1	0.579766	0.000000
16 x 16	51	2 -> 1	201	1 -> 1	0.511469	0.000000
17 x 17	57	3 -> 1	227	1 -> 1	0.370159	0.000000
18 x 18	64	3 -> 1	255	1 -> 1	0.336315	0.000000
19 x 19	71	3 -> 1	284	1 -> 1	0.263784	0.000000
20 x 20	79	3 -> 1	314	1 -> 1	0.220976	0.058824
21 x 21	87	3 -> 1	346	1 -> 1	0.139667	0.000000
22 x 22	95	3 -> 1	380	1 -> 1	0.121313	0.039216
23 x 23	104	3 -> 1	415	1 -> 1	0.071229	0.039216

24 x 24	113	3 -> 1	452	1 -> 1	0.048371	0.000000
25 x 25	123	4 -> 1	491	1 -> 1	0.044840	0.000000
27 x 27	143	4 -> 1	572	1 -> 1	0.027883	0.019608
28 x 28	154	4 -> 1	615	1 -> 1	0.008402	0.039216
29 x 29	165	4 -> 1	660	1 -> 1	0.008748	0.019608
30 x 30	177	4 -> 1	706	1 -> 1	0.002742	0.039216
32 x 32	201	4 -> 1	804	1 -> 1	0.002765	0.058824
32 x 30	189	4 -> 1	753	1 -> 1	0.001906	0.039216
34 x 34	227	5 -> 1.25	907	1.25 -> 1	0.000255	0.000000
36 x 36	255	5 -> 1.25	1017	1.25 -> 1	0.000219	0.078431
40 x 40	314	5 -> 1.25	1255	1.25 -> 1	0.000012	0.000000

5.3. Experimentos

5.3.1. Mapas Auto-Organizáveis

Conforme visto na seção 4.7, as espécies do gênero Corynebacterium podem ser separadas em grupos (Tabela 5.3-2) que compartilham as mesmas repostas à determinados testes bioquímicos. O conjunto de dados públicos possui 51 registros de espécies únicas e sua separação por grupos está representada na Tabela 5.3-1, o código de identificação de cada microrganismo se encontra listado na Tabela 5.3-3.

Tabela 5.3-1 Relação das espécies do conjunto público e seus respectivos grupos.

Nome	grupo	Nome	grupo
C. accolens	grupo_a	C. diphtheriae var gravis	grupo_e
C. pilosum	grupo_a	C. diphtheriae var intermedius	grupo_e
C. pseudodiphtheriticum	grupo_a	C. diphtheriae var mitis	grupo_e
C. xerosis	grupo_a	C. durum	grupo_e
C.afermentans var afermentans	grupo_b	C. falsenii	grupo_e
C. bovis	grupo_b	C. freneyi	grupo_e
C. jeikeium	grupo_b	C. glucuronolyticum	grupo_e
C. amycolatum	grupo_c	C. imitans	grupo_e
C. flavescens	grupo_c	C. kroppenstedtii	grupo_e
C. minutissimum	grupo_c	C. lipophiloflavum	grupo_e
C. striatum	grupo_c	C. macginleyi	grupo_e
C. cystitidis	grupo_d	C. matruchotii	grupo_e
C. kutscheri	grupo_d	C. mucifaciens	grupo_e
C. pseudotuberculosis	grupo_d	C. propinquum	grupo_e
C. renale	grupo_d	C. resistens	grupo_e
C. afermentans var lipophilum	grupo_e	C. riegelii	grupo_e
C. appendicis	grupo_e	C. seminale	grupo_e
C. argentoratense	grupo_e	C. simulans	grupo_e
C. atypicum	grupo_e	C. singulare	grupo_e
C. aurimucosum	grupo_e	C. sundsvallense	grupo_e
C. auris	grupo_e	C. thomssenii	grupo_e
C. confusum	grupo_e	C. tuberculostearicum	grupo_e
C. coyleae	grupo_e	C. tuscaniae	grupo_e
C. diphtheriae var belfanti	grupo_e	C. ulcerans	grupo_e
C. urealyticum	grupo_e	CDC group F-1	grupo_e
		CDC group G	grupo_e

Tabela 5.3-2 Os cinco grupos de divisão das bactérias do gênero Corynebacterium e suas descrições.

Grupo	Descrição
grupo a	Nitrato - positivo, imóvel
grupo b	Nitrato - negativo, imóvel
grupo c	Nitrato - negativo, imóvel, fermentador de glicose (ou glicose - positivo)
grupo d	Nitrato - negativo, imóvel, fermentador de glicose (ou glicose - positivo) , urease - positivo, fosfatase alcalina - negativo
grupo e	não definido

Tabela 5.3-3 Relação das espécies do conjunto público e seus respectivos números de identificação.

Nome	codigo	Nome	codigo
C. accolens	1	C. diphtheriae var gravis	15
C. pilosum	32	C. diphtheriae var intermedius	16
C. pseudodiphtheriticum	34	C. diphtheriae var mitis	17
C. xerosis	49	C. durum	18
C. afermentans var afermentans	2	C. falsenii	19
C. bovis	10	C. freneyi	21
C. jeikeium	24	C. glucuronolyticum	22
C. amycolatum	4	C. imitans	23
C. flavescens	20	C. kroppenstedtii	25
C. minutissimum	30	C. lipophiloflavum	27
C. striatum	42	C. macginleyi	28
C. cystitidis	13	C. matruchotii	29
C. kutscheri	26	C. mucifaciens	31
C. pseudotuberculosis	35	C. propinquum	33
C. renale	36	C. resistens	37
C. afermentans var lipophilum	3	C. riegelii	38
C. appendicis	5	C. seminale	39
C. argentoratense	6	C. simulans	40
C. atypicum	7	C. singulare	41
C. aurimucosum	8	C. sundsvallense	43

C. auris	9	C. thomssenii	44
C. confusum	11	C. tuberculostearicum	45
C. coyleae	12	C. tuscaniae	46
C. diphtheriae var belfanti	14	C. ulcerans	47
CDC group F-1	50	C. urealyticum	48
		CDC group G	51

Na seção 5.2 foram testadas diversas configurações de mapas a fim de se chegar ao modelo que melhor se adapte ao conjunto de dados públicos, deste modo, o mapa escolhido é o arranjo plano de 29 x 29 neurônios com vizinhança hexagonal e treinado pelo algoritmo “batch”.

Um tipo de mapa utilizado para identificar agrupamentos graficamente é chamado matriz-U. A matriz-U apresenta, através das cores nas unidades, as distâncias entre os agrupamentos, é portanto, um método de visualização de um SOM treinado, que permite a detecção visual das relações topológicas dos neurônios. Usa-se a mesma forma de cálculo utilizada durante o treinamento para determinar a distância entre os vetores de peso de neurônios adjacentes.

Utilizando a ferramenta SOM *Toolbox* na plataforma MATLAB, o mapa foi gerado e treinado com o conjunto completo de dados públicos. Inicialmente deve-se checar se a quantidade de neurônios é suficiente para representação dos dados do conjunto de treinamento. Esta análise (Figura 5.3-1), leva em consideração se é homogênea a frequência de resposta dos neurônios (círculos pretos) no mapa, portanto, de acordo com a Figura 5.3-1, a configuração do mapa é condizente com a quantidade de dados do conjunto.

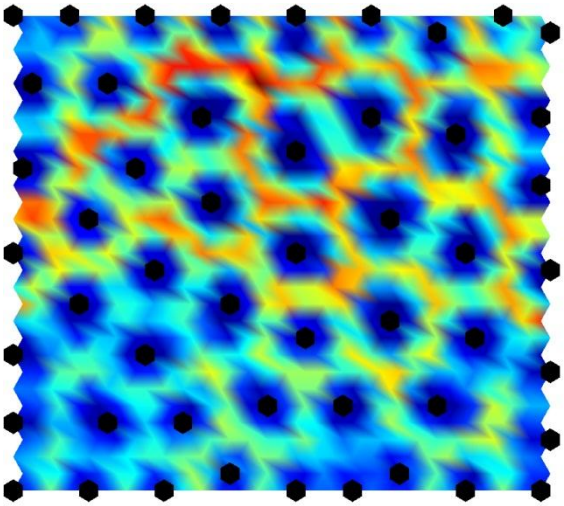


Figura 5.3-1 SOM - Matriz-U do conjunto de dados público com pontos escuros que indicam a frequência de resposta dos neurônios. completo.

Na Figura 5.3-2 são exibidas diversas visualizações da matriz-U. Em (B) é exibida na forma clássica, sendo (D) sua exibição com cores interpoladas, o que garante uma melhor percepção na formação e transição de grupos. É importante destacar que, por padrão, a cor azul denota a formação de agrupamentos, enquanto que a cor vermelha, significa a separação dos mesmos. Em todos os gráficos é claramento visível a formação de diversas ilhas de agrupamento, ficando ainda mais evidente, na visualização (C). Em (A) a matriz-U é colorida em tons de cinza e, para cada um dos 51 grupos, está destacado o rótulo de identificação da espécie que domina aquela região. É possível notar também que na parte inferior esquerda os grupamentos possuem transição mais suave entre sí, enquanto que na parte superior direita, as separações são fortemente marcadas pela coloração avermelhada, ou seja, são mais bem separados.

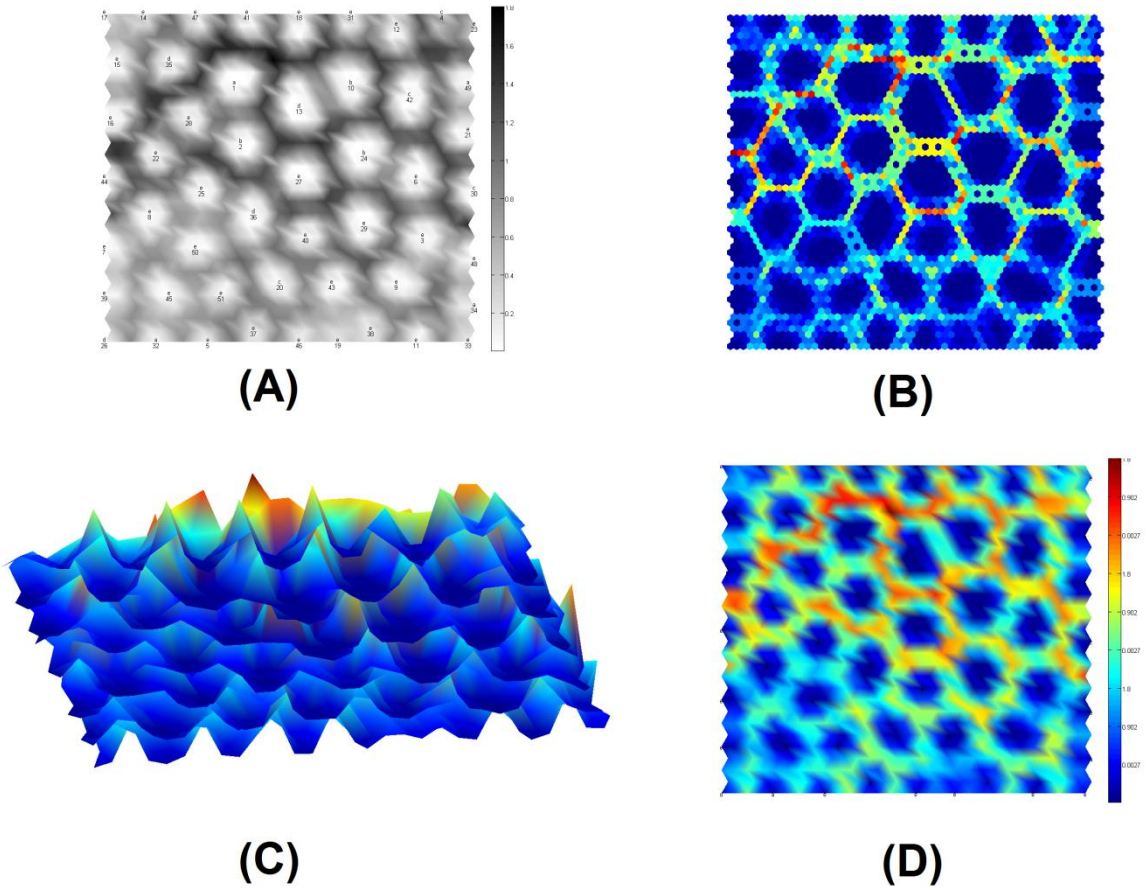


Figura 5.3-2 Diversas visualizações da matriz-U. Em (A) a matriz-U em escala cinza com os rótulos das espécies dominantes. A matriz-U original (B) e com interpolação de cores em (D). (C) Apresenta a matriz de distâncias em três dimensões.

O conjunto de dados públicos possui 5 classes (Tabela 5.3-2), que divide as 51 espécies em grupos que compartilham semelhanças em determinadas provas bioquímicas. Na Figura 5.3-3 são exibidas quatro tipos de visualizações de grupamentos, sendo (C) o único com informação prévia do número de grupos existentes. Em (A), a coloração dos grupos retrata o nível de similaridade entre eles, sendo que em (B), as cores se mantém, porém, são exibidas em um mapa de distâncias. A Figura 5.3-3-C tem seu grupamento formado através do método hierárquico, disponível no SOM *Toolbox*, que realiza esta formação baseada em um número definido de classes. Para os grupamentos formados em (D), cada neurônio recebeu o rótulo da classe que mais se assemelha ao seu vetor de pesos.

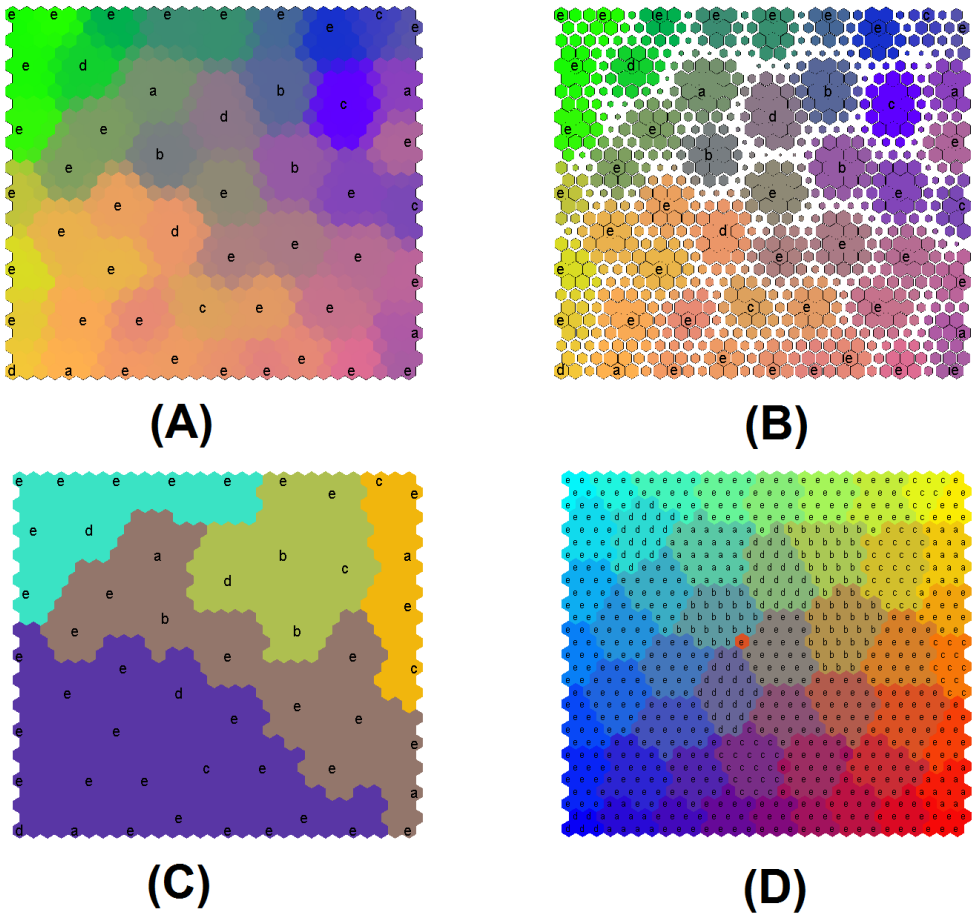


Figura 5.3-3 Sugestões de agrupamentos. Em (A), a coloração é definida de acordo com a similaridade entre os neurônios, para evidenciar a formação de agrupamentos, em (B) preservando a coloração, adiciona-se a matriz de distâncias. (C) apresenta cinco agrupamentos. Na representação em (D), cada neurônio recebe o rótulo do grupo que tem mais similaridade.

Devido à remoção dos 17 atributos (Tabela 5.2-3) do conjunto do LDCIC, os testes de agrupamento para este conjunto podem ser prejudicados. Desta forma é preciso replicar esta remoção no conjunto de dados público e treinar novamente o mapa (Tabela 5.2-5) afim de que se tenha um resultado condizente com a realidade dos testes que são feitos no laboratório.

Com a redução da dimensionalidade do conjunto de treinamento em 17 atributos, o tamanho do mapa com melhor desempenho é também reduzido. Conforme a Tabela 5.3-4, o mapa escolhido possui um arranjo plano de 25 x 25 neurônios com vizinhança hexagonal e treinado pelo algoritmo “batch”.

Tabela 5.3-4 Resultados de QE e TE para diversas configurações de mapas, suas dimensões e especificações das fases de treinamento. Usando conjunto de dados público incompleto (os atributos da Tabela 5.2-3 foram removidos). Método de normalização "range".

Configurações						
Dimensões	Fase 1		Fase 2		QE	TE
	Épocas	Raio	Épocas	Raio		
7 x 5	7	1 -> 1	28	1 -> 1	1.025767	0.000000
10 x 10	20	2 -> 1	79	1 -> 1	0.721076	0.000000
15 x 15	45	2 -> 1	177	1 -> 1	0.374350	0.000000
17 x 17	57	3 -> 1	227	1 -> 1	0.219531	0.000000
19 x 19	71	3 -> 1	284	1 -> 1	0.123850	0.000000
20 x 20	79	3 -> 1	314	1 -> 1	0.086404	0.000000
22 x 22	95	3 -> 1	380	1 -> 1	0.050983	0.000000
23 x 23	104	3 -> 1	415	1 -> 1	0.032814	0.000000
24 x 24	113	3 -> 1	452	1 -> 1	0.029544	0.058824
25 x 25	123	4 -> 1	491	1 -> 1	0.017157	0.019608
26 x 26	133	4 -> 1	531	1 -> 1	0.010742	0.039216
27 x 27	143	4 -> 1	572	1 -> 1	0.006448	0.000000
28 x 28	154	4 -> 1	615	1 -> 1	0.006137	0.098039
29 x 29	165	4 -> 1	660	1 -> 1	0.002266	0.078431
30 x 30	177	4 -> 1	706	1 -> 1	0.000640	0.058824
32 x 32	201	4 -> 1	804	1 -> 1	0.000771	0.019608
34 x 34	227	5 -> 1.25	907	1.25 -> 1	0.000068	0.000000

O conjunto do LDCIC serviu como entrada de teste no mapa SOM que foi gerado. Na Figura 5.3-4 Matriz-U do SOM com arranjo plano de 25 x 25 neurônios com vizinhança hexagonal. Os rotulos representam os grupos dos conjuntos de treino e teste, nas cores azul e vermelho respectivamente.Figura 5.3-4 está representada a matriz-U e em destaque estão os neurônios vencedores de cada grupo, onde apresentam os nomes dos seus respectivos grupos nas cores azul (conjunto de treino) e vermelho (conjunto de teste). Neste teste o mapa apresentou um erro topográfico de 0.26 e um erro de quantização de 1.67. Porém, é possível constatar que não houve uma boa generalização dos padrões, afinal, poucos grupos dos testes foram bem agrupados.

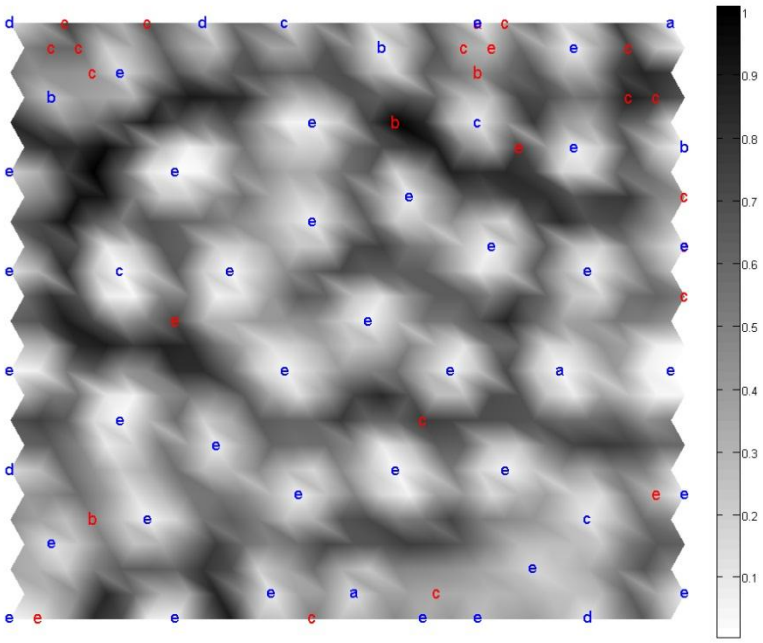


Figura 5.3-4 Matriz-U do SOM com arranjo plano de 25 x 25 neurônios com vizinhança hexagonal. Os rotulos representam os grupos dos conjuntos de treino e teste, nas cores azul e vermelho respectivamente.

Visando uma melhor generalização, este mapa foi treinado novamente com o conjunto de dados públicos alterado. Quando um registro apresenta o valor “V” significa que é um valor variável, podendo ser tanto “+” quanto “-“, dessa forma, o conjunto de dados foi reformulado com todas as possíveis combinações para cada caso. Na Figura 5.3-5, são exibidos diversos tipos de visualizações da matriz-U, sendo possível notar que houve uma suavização na separação dos grupos, exceto no canto inferior esquerdo, onde quatro agrupamentos estão fortemente demarcados.

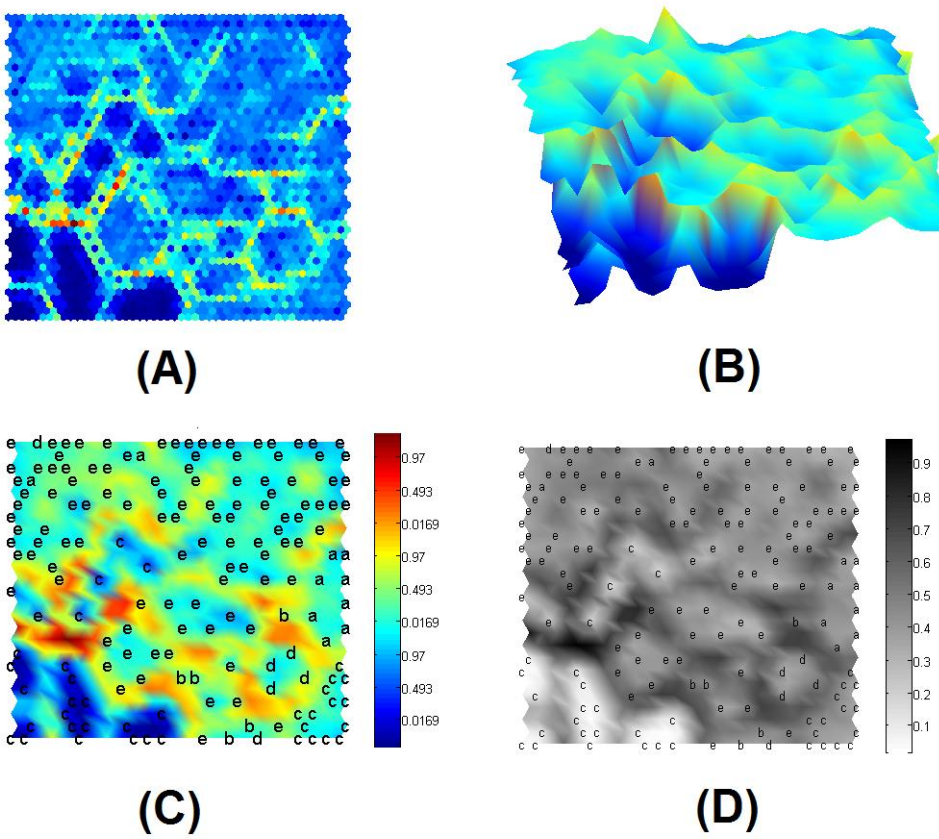


Figura 5.3-5 Diversas visualizações da matriz-U. A matriz-U original (A) e com interpolação de cores (C) com os rótulos dos grupos dominantes. (B) Apresenta a matriz de distâncias em três dimensões. Em (D) a matriz-U em escala cinza com os rótulos dos grupos dominantes.

Através da análise das Figura 5.3-6-A e Figura 5.3-6-B, pode-se observar a formação de agrupapamentos de espécies bem definidos para quatro dos cinco grupos propostos (Tabela 5.3-2). No canto inferior esquerdo (Figura 5.3-6-C) há o predomínio do grupo "grupo_c", especificamente da espécie com identificação 42 (*C. striatum*), ainda nesta área, observa-se a formação de quatro agrupamentos, sugerindo que apesar da espécie apresentar diversos valores variáveis foi possível agrupar todas suas nuâncias em uma região específica. Enquanto que no canto inferior direito, nota-se que o agrupamento do grupo "grupo_c" (dominado pela espécie *C. amycolatum*) está isolado através de vales dos agrupamentos dos grupos "grupo_a", "grupos_b" e "grupo_d".

Interessante notar no agrupamento na parte central do mapa (grupo "grupo_e"), o domínio de quatro sub-espécies *diphtheriae* na mesma região (*C.*

diphtheriae var belfanti , *C. diphtheriae var gravis* , *C. diphtheriae var intermedius* , *C. diphtheriae var mitis*). Este agrupamento ilustra o bom treinamento do mapa, pois a diferença entre elas consiste em algumas poucas provas bioquímicas.

As espécies pertencentes ao grupo "grupo_e" são as que não tiveram suas provas bioquímicas relacionadas a um grupo previamente estabelecido. É possível notar que as espécies deste grupo estão espalhadas por todo o mapa, algumas isoladas e outras formando diversos agrupamentos, o que permite que sejam identificadas e que tenham seus atributos correlacionados.

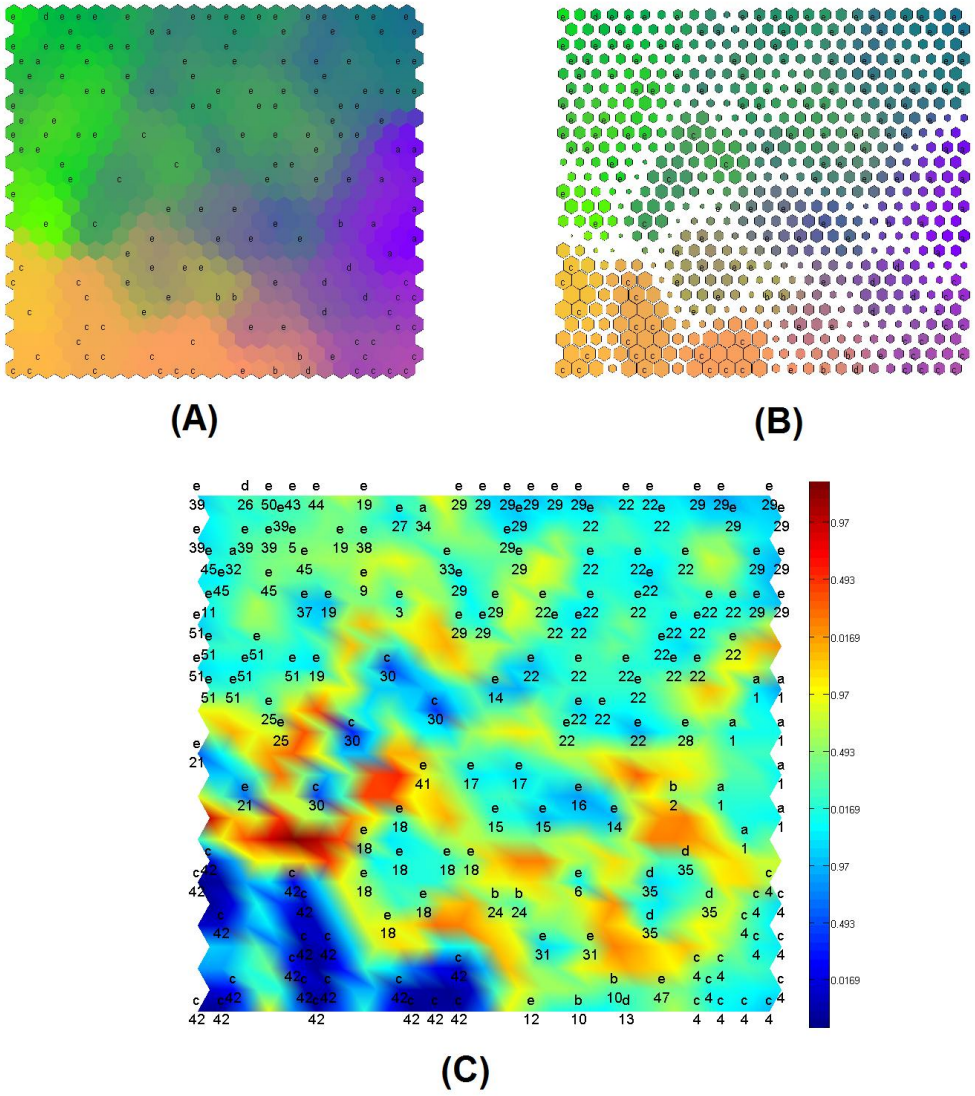


Figura 5.3-6 Sugestões de agrupamentos. Em (A), a coloração é definida de acordo com a similaridade entre os neurônios, para evidenciar a formação de agrupamentos, em (B) preservando a coloração, adiciona-se a matriz de distâncias. (C) apresenta a matriz-U com os neurônios vencedores de cada espécie e suas respectivas classes.

Com o intuito de testar o mapa SOM recém criado, novamente, o conjunto de dados do LDCIC foi apresentado. Na Figura 5.3-7-A está representada a matriz-U, onde os neurônios vencedores de cada grupo apresentam os nomes dos seus respectivos grupos nas cores azul (conjunto de treino) e vermelho (conjunto de teste). Sendo Figura 5.3-7-B a ampliação da área selecionada da Figura 5.3-7-A, é possível notar uma grande melhora na adaptação do conjunto de testes no mapa, portanto, pode-se afirmar que houve uma boa generalização dos padrões de treinamento. Neste teste o mapa apresentou um erro topográfico de 0.07 e um erro de quantização de 1.69.

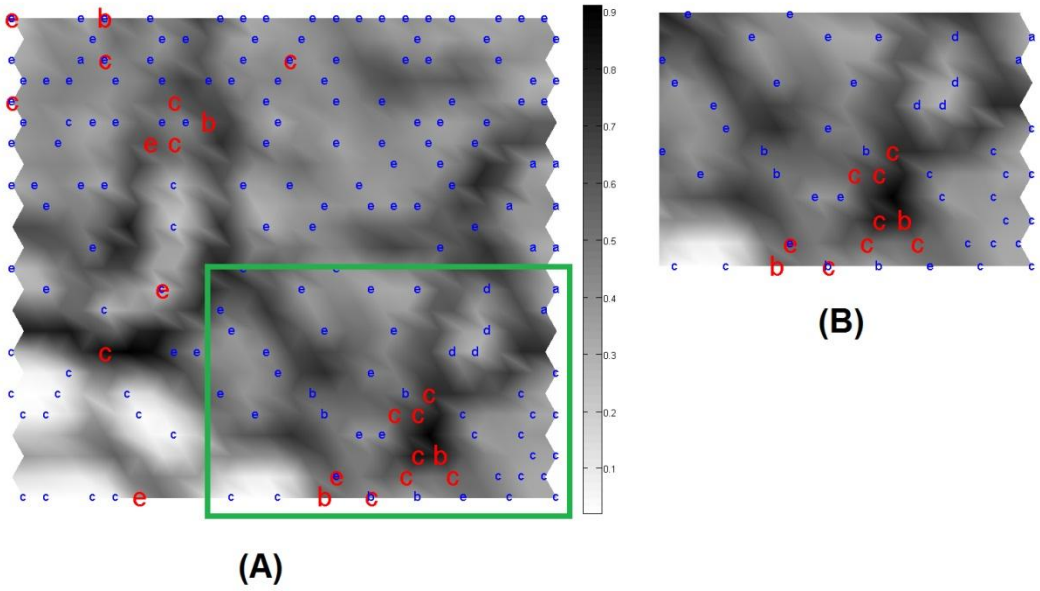


Figura 5.3-7 (A) Matriz-U do SOM com arranjo plano de 25 x 25 neurônios com vizinhança hexagonal. A área selecionada em está ampliada em (B). Os rotulos representam os grupos dos conjuntos de treino e teste, nas cores azul e vermelho respectivamente.

Nas Figura 5.3-8-A,B,C,D são apresentadas quatro matrizes-U relativas aos atributos Esculina, Nitrato, Urease e Glicose, neste tipo de representação por atributo único, pode-se observar as tendências de agrupamento particular à cada um e a sua contribuição para o resultado geral. Analisando as matrizes-U individuais de cada atributo pode-se entender melhor as correlações entre os mesmos, uma correção positiva existe quando as figuras são semelhantes,

enquanto que figuras com padrões de cores invertidos indicam correlação negativa. Através do estudo das correlações entre os atributos pode-se eliminar atributos redundantes e identificar atributos de grande relevância no resultado final. Na Figura 5.3-8-E a matriz-U é formada, apenas, pelos quatro atributos apresentados e na Figura 5.3-8-F apresenta-se a matriz-U composta por todos os vinte atributos.

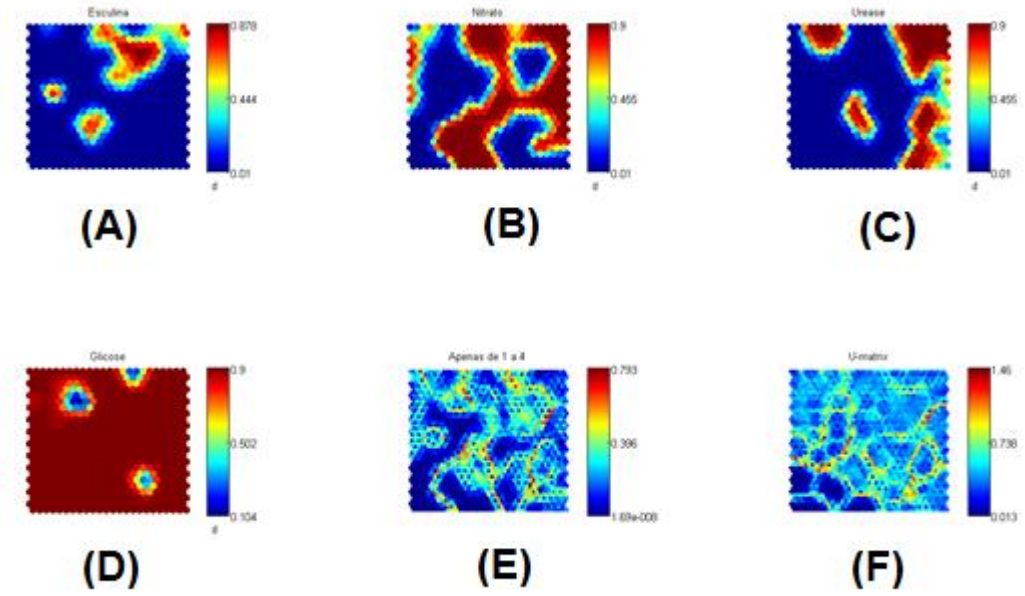


Figura 5.3-8 As figuras A, B, C e D são as matrizes-U relativas aos atributos Esculina, Nitrato, Urease e Glicose respectivamente. (E) representa a matriz-U composta exclusivamente pelos quatro atributos A,B,C e D. Em (F) é apresentada a matriz-U completa, ou seja, composta pela composição de todos os vinte atributos.

A Figura 5.3-9 apresenta as matrizes-U relativas a cada um dos vinte atributos usados nos experimentos. Analisando visualmente cada matriz pode-se chegar a diversas conclusões, o atributo Catalase e sua homogeneidade de cor indica que não há alterações no valor deste atributo para todos os registros. Existe uma correlação positiva entre Mobilidade e AAR, o que indica que um destes atributos pode ser removido, do ponto de vista prático pode-se optar por remover o teste bioquímico que seja mais trabalhoso, caro ou demorado.

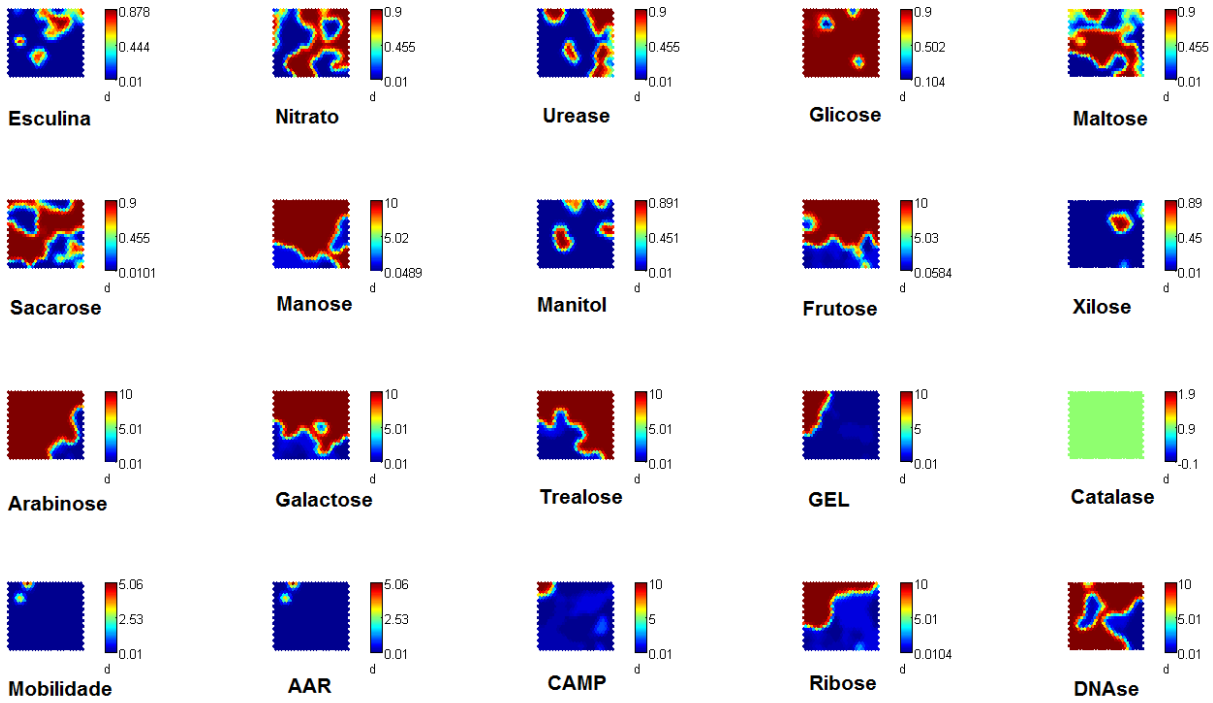


Figura 5.3-9 As matrizes-U relativas a cada um dos vinte atributos.

5.3.2. Redes Multilayer Perceptron

Na continuação dos experimentos, para avaliação da metodologia proposta, foram realizados testes com RNAs MLP, utilizando como conjuntos de entrada os mesmos dados utilizados nos mapas auto-organizáveis (Tabela 5.2-2), o conjunto de saída de treino, validação e testes e sua codificação está apresentado na Tabela 5.3-3. Diversas configurações foram testadas para as redes, sendo a métrica de avaliação de saída das redes apresentada na Equação 5.3-1. O MSE (Mean Squared Error) ou EQM (Erro Quadrático Médio) é determinado somando-se os erros da previsão ao quadrado e dividindo pelo número de épocas percorrido.

$$MSE = \frac{\sum_{t=1}^{N} A_t - P_t \ ^2}{N} \qquad \text{Equação 5.3-1}$$

Onde: A_t é o valor real na época t

P_t é o valor previsto na época t

N é o número total de épocas

A configuração básica de todas as redes testadas consiste na múltipla camada de neurônios *Perceptrons* (MLP, *Multilayer Perceptrons*), foram testadas diversas configurações variando de 20 até 80 neurônios para cada camada escondida. O algoritmo de aprendizado supervisionado utilizado é o de retroprogação do erro (*backpropagation*), que é essencialmente, baseado no gradiente da função do erro. Entre diversos tipos de algoritmos de aprendizado, dois foram utilizados e comparados nesta seção, sendo eles o *Resilient backpropagation* (RIEDMILLER, 1993) e *Scaled conjugate gradient backpropagation* (MOLLER, 1993). As funções de ativação dos neurônios escolhidas são respectivamente “tansig” (Figura 5.3-10) e “tansig”.

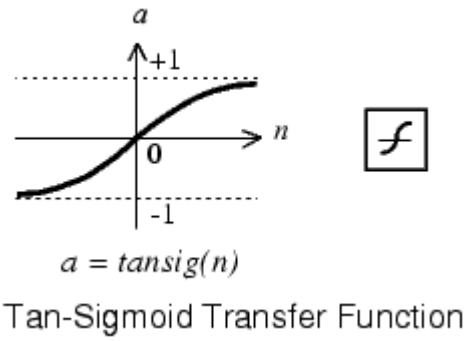


Figura 5.3-10 Gráfico de comportamente da função 'tansig'. Fonte: MATLAB.

Os conjuntos de treino, validação e teste foram separados, respectivamente, em 70%, 15% e 15%, de um total geral de 391 registros. A seguir são apresentados em tabelas, todos os testes realizados neste experimento, cada configuração foi testada 100 vezes, portanto sua apresentação estará na forma da média de todos estes experimentos. As tabelasTabela 5.3-5 Resultados de diversas configurações de topologia de rede, apresentando a média dos MSE do treino, validaçao e teste, assim como a média das porcentagens de acertos no treino, validação e teste. Todas as redes usaram o *Resilient backpropagation* como algoritmo de aprendizado. Tabela 5.3-5, Tabela 5.3-6 e Tabela 5.3-7 estão divididas em 7 colunas, sendo cada uma delas comentada abaixo.

- *Topologia da rede*: Número de neurônios das camadas escondidas da rede neural, com variação de 20 até 80.
- *MSE Treino*: Resultado da média do erro quadrático médio (MSE) do conjunto separado para o treino da RNA.

- *MSE Validação*: Resultado da média do erro quadrático médio (MSE) do conjunto de validação, calculado depois do treinamento da rede.
- *MSE Teste*: Resultado da média do erro quadrático médio (MSE) do conjunto de teste, é importante ressaltar que este conjunto de dados é inédito para a rede. O erro é calculado após o treino da mesma.
- *Acertos Treino*: Média da porcentagem de itens que foram corretamente previstos pela RNA durante a fase de treinamento.
- *Acertos Validação*: Média da porcentagem de itens que foram corretamente previstos pela RNA durante a fase de validação.
- *Acertos Teste*: Média da porcentagem de itens que foram corretamente previstos pela RNA durante a fase de testes.

Tabela 5.3-5 Resultados de diversas configurações de topologia de rede, apresentando a média dos MSE do treino, validaçao e teste, assim como a média das porcentagens de acertos no treino, validação e teste. Todas as redes usaram o *Resilient backpropagation* como algoritmo de aprendizado.

Rede neural - *Resilient backpropagation*							
Topologia da rede		MSE Treino	MSE Validação	MSE Teste	Acertos Treino	Acertos Validação	Acertos Teste
10	10	0,00971	0,01156	0,01192	61,3	54,2	54,2
20	20	0,00716	0,01003	0,011	73,7	61,4	59,4
20	30	0,00563	0,00968	0,00992	80,3	65,3	64,8
20	40	0,00572	0,00939	0,00995	80,3	65,7	64,6
20	50	0,00552	0,00958	0,00974	81,1	66,6	65,5
20	60	0,00651	0,0105	0,01095	80,6	65,9	64,2
20	70	0,00874	0,01276	0,01297	77,3	62,8	62,3
30	20	0,0067	0,0094	0,01011	75,3	64,2	62,4
30	30	0,00591	0,00929	0,00951	79,4	66	66
30	35	0,00578	0,00918	0,00985	80	66,6	64,6
30	40	0,00561	0,00918	0,00993	80,5	66,9	64,3
50	50	0,00521	0,00955	0,00993	81,8	65,8	63,9
60	60	0,00535	0,00966	0,0101	81,4	64,7	63,7
70	70	0,00537	0,00926	0,01	81,2	66,6	63,6
80	80	0,01047	0,0143	0,01476	76,4	61,8	59,8

Tabela 5.3-6 Resultados de diversas configurações de topologia de rede, apresentando a média dos MSE do treino, validaçao e teste, assim como a média das porcentagens de acertos no treino, validação e teste. Todas as redes usaram o *Scaled conjugate gradient backpropagation* como algoritmo de aprendizado.

Rede neural - *Scaled conjugate gradient backpropagation*							
Topologia da rede		MSE Treino	MSE Validação	MSE Teste	Acertos Treino	Acertos Validação	Acertos Teste
10	10	0,02541	0,0249	0,02654	20,2	17,7	17,7
20	20	0,01401	0,01518	0,01517	35,4	30,3	31,5
20	30	0,01519	0,01589	0,01616	32	27,9	28,7
20	50	0,01558	0,01586	0,0166	31,5	31,4	28,5
50	50	0,0127	0,01363	0,01418	41	39,1	36,9
60	60	0,01317	0,01417	0,01481	37,8	34,6	32,1
70	70	0,02038	0,02074	0,02211	36,4	35,3	32,6
80	80	0,01353	0,01456	0,01496	37,6	33,9	33,2

Analisando os resultados exibidos nas Tabela 5.3-5 e Tabela 5.3-6, é fácil notar a superioridade do modelo que faz uso do algoritmo de treinamento *Resilient backpropagation.* Especialmente a topologia que apresenta 30x35 (Tabela 5.3-7) neurônios nas camadas ocultas, apresentada na Tabela 5.3-7, esta foi a que apresentou o maior número de acertos (82%) no conjunto de dados de testes e o menor MSE de validação.

Tabela 5.3-7 Resultados da configuração de topologia da rede com melhor desempenho, apresentando a média dos MSE do treino, validaçao e teste, assim como a porcentagem de acertos no treino, validação e teste. Rede treinada usando o Resilient backpropagation como algoritmo de aprendizado.

Rede neural de melhor desempenho							
Topologia da rede		MSE Treino	MSE Validação	MSE Teste	Acertos Treino	Acertos Validação	Acertos Teste
30	35	0,0042	0,0113	0,0055	85,7	56,4	82

5.4. Discussão dos Resultados

Apesar das dificuldades encontradas no pré-processamentos dos dados e no desconhecimento da importância de cada atributo, os mapas auto-organizáveis e as redes MLP apresentaram surpreendentes resultados na classificação e identificação de bactérias.

É importante ressaltar que os grupos de bactérias estabelecidos na Tabela 5.3-2 e usados nos experimentos, seguem o algoritmo proposto na seção 4.7. Estes grupos são definidos através de um diagrama de testes bioquímicos e as espécies que compartilham uma sequência de resultados em comum são agrupadas. A rotulação em grupos pre-estabelecidos auxilia a análise e o estudo na classificação de bactérias, porém, não se pode ficar restrido à elas, afinal, conforme visto na seção 3.3, o gênero *Corynebacterium* é complexo e está em constante alteração devido à novas descobertas e avanços na análise genotípica. Sendo assim os mapas auto-organizáveis se mostram uma poderosa ferramenta no estudo da relevância de cada prova bioquímica nos grupos.

Ao contário dos mapas SOM, as RNA do tipo MLP possuem um índice mais prático e confiável na avaliação da topologia e configuração das redes que estão sendo testadas. Uma RNA com um bom desempenho, consiste e uma rede de menor tamanho possível e que tenha poder de generalização, ou seja, é capaz de responder corretamente a padrões jamais vistos. Portanto, é fundamental a validação através do MSE, na fase de treinamento, pois assim evita-se o super-treinamento e consequentemente a incapacidade de generalização.

6 Conclusões e Trabalhos Futuros

6.1. Conclusões

A pesquisa desenvolvida ao longo deste trabalho buscou oferecer contribuições na área da bioinformática, especificamente na classificação e identificação de bactérias. As informações biológicas usadas nos experimentos são provenientes do BCIWeb. Este sistema foi criado devido à necessidade de informatização dos processos de registro e controle das amostras no Laboratório de Difteria e Corinebactérias de Importância Clínica (LDCIC).

Esta dissertação, demonstrou no estudo de caso que redes neurais artificiais MLP e mapas auto-organizáveis possuem grande aplicabilidade no objetivo proposto. É importante destacar que todos o treinos, validações e testes foram realizados com conjuntos de dados reais, o que sugere a sua viabilidade no uso prático.

Os mapas SOM permitem uma boa análise gráfica do conjunto de dados, sendo possível estudar minunciosamente as relações entre as características pertinentes a cada grupo, tornando possível um melhor entendimento da importância de cada prova bioquímica na separação de espécies. Este método foi colocado à prova para o conjunto de dados do LDCIC, e após a sequência de tratamento de dados, chegou-se a uma boa generalização classificatória. As RNA MLP apresentaram um grande potencial no aprendizado dos padrões de informações biológicas, onde a melhor configuração de rede chegou à uma taxa de acerto de 82% na identificação de espécies.

O sistema desenvolvido foi muito importante na otimização do processo laboratorial e, através dele, com as informações digitalizadas, foi possível realizar todos os experimentos deste trabalho. Suas possibilidades de uso não ficam restritas apenas ao gênero Corynebacterium e as provas bioquímicas aqui usadas, afinal, seu desenvolvimento modular permite que sejam criados diversos tipos de páginas e novos campos nas bases de dados, o que possibilita seu uso nas mais diversas áreas.

6.2. Trabalhos Futuros

Diversas atividades, relacionadas ou não com esta linha de pesquisa, podem ser citadas como sugestão de trabalhos futuros:

- A criação de um modelo *web* de rede neural artificial já treinada para a melhor configuração do gênero de interesse, possibilitando a identificação *online* de bactérias através de RNA;
- As informações armazenadas no sistema BCIWeb não consistem apenas nos resultados de provas bioquímicas, estas amostras estão relacionadas com diversas outras tabelas que envolvem informações do paciente, do material colhido etc. Dessa forma um estudo mais abrangente das informações seria interessante;
- Estudo dos grupos sugeridos no mapa SOM e como eles podem ajudar na formação de novos e otimizados diagramas de testes bioquímicos na identificação de bactérias;
- O estudo de outros modelos inteligentes neste tema como Neuro-Fuzzy, *Support vector machine* etc;
- Complementação da base de dados, devido a muitos atributos com valores faltosos.
- O sistema BCIWeb é totalmente modular, deste modo, é fácil sua adaptação para as mais diversas áreas como oncologia, imunologia, genética etc.

7. Referências Bibliográficas

ADDERSON EE, BOUDREAUX JW, CUMMINGS JR, POUNDS S, WILSON DA, PROCOP GW, HAYDEN RT. **Identification of clinical coryneform bacterial isolates: comparison of biochemical methods and sequence analysis of 16S rRNA and rpoB genes**. Journal of Clinical Microbiology, 46(3):921-927, 2008.

AHMAD, N. ABDULLAH, S.R.S. ANUAR, N. HUSIN, H. (2008) **Bacteria identification using artificial neural network: a case study of Peptococcaceae Family identification**. Electronic Design, 2008. ICED 2008. International Conference on, 2008.

ALMUZARA, M. N., de Meer, C., Rodríguez, C. R., Famiglietti, A. M. R. & Vay, C. A. (2006). **Evaluación del sistema API Coryne, versión 2.0, para la identificación de bacilos gram-positivos difteroides de importancia clínica**. Rev Argent Microbiol 38, 197–201.

Babay HA, Kambal AM. **Isolation of coryneform bacteria from blood cultures of patients at a University Hospital in Saudi Arabia**. Saudi Med J 2004; 25:1073-1079.

BACTERIA, WIKIPEDIA. Disponível em: <http://en.wikipedia.org/wiki/Bacteria#Classification_and_identification>. Acesso em: 06 de janeiro de 2012.

BACT. Disponível em: < http://inst.bact.wisc.edu/inst/index.php?module=book&func=displayarticle&art_id=119>. Acesso em: 06 de janeiro de 2012.

BASCOMB, S., S. P. LAPAGE, M. A. CURTIS, AND W. R. WILLCOX. 1973. **Identification of bacteria by computer: identification of reference strains**. J. Gen. Microbiol. 77:291-315.

BAYES, THOMAS, AND PRICE, RICHARD (1763). **"An Essay towards solving a Problem in the Doctrine of Chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S.".** Philosophical Transactions of the Royal Society of London 53 (0): 370–418.

BEERS, R. J. & LOCKHART, W. R. (1962). **Experimental methods in computer taxonomy.** Journal of General Microbiology 28, 633-640.

BERGEY, DAVID H.; JOHN G. HOLT; NOEL R. KRIEG; PETER H.A. SNEATH (1994). **Bergey's Manual of Determinative Bacteriology** (9th ed.). Lippincott Williams & Wilkins. ISBN 0-683-00603-7.

BIOMERIEUX. Disponível em: <http://www.biomerieux-diagnostics.com>. Acesso em: 05 de janeiro de 2012.

BOEUFGRAS, J. M., J. L. BLAZER, F. ALLARD, AND I. DIAZ. 1988**. A new computer program for routine interpretation of API identification systems**, p. 125-137. In J. Schindler and M. Chyle (ed.), Selected papers, 2nd Conference on Taxonomy and Automatic Identification of Bacteria. Czechoslovak Society for Microbiology of the Czechoslovak Academy of Sciences, Prague, Czech Republic.

BOGUSKI, M. S. "**Bioinformatics.**" Curr Opin Genet Dev 4(3): 383-8. (1994).

BOGUSKI, M. S. **Bioinformatics - a new era. Trends in Biochemical Sciences, Supplement**, Trends Guide to Bioinformatics, 1–3 (1998)

BOONE, D.R. & CASTENHOLZ, R.W. 2001. **Bergey´s Manual of Systematic Bacteriology** 2nd ed., Volume One, Springer-Verlag, USA.

BOUCHER Y, DOUADY CJ, PAPKE RT, WALSH DA, BOUDREAU ME, NESBO CL, CASE RJ, DOOLITTLE WF (2003). "**Lateral gene transfer and the origins of prokaryotic groups"**. Annu Rev Genet 37: 283–328. doi:10.1146/annurev.genet.37.050503.084247. PMID 14616063.

BRAGA, A. P.; CARVALHO, A. P. L. F.; LUDERMIR, T. B. **"Redes Neurais Artificiais Teoria e Aplicações"** . Livros Técnicos e Científicos Editora, Rio de Janeiro, 2000.

BRYANT TN. **PIBWin - software for probabilistic identification. Journal of Applied Microbiololgy**. 2004;97(6):1326-7.

BRYANT, T. N., J. V. LEE, AND P. A. WEST. 1986. **Numerical classification of species of Vibrio and related genera.** J. Appl. Bacteriol. 61:437-467.

BULL, A.T. , GOODFELLOW, M. & SLATER, J.H. 1992**. Biodiversity as a source of innovation in biotechnology.** Annual Review of Microbiology 46:219-252.

CABENA, P. et al. **Discovering Data Mining from concept to implementation**. New Jersey: Prentice-Hall, 1997, 195p.

CAMELLO TCF, MATTOS-GUARALDI AL, FORMIGA LCD, MARQUES EA 2003. **Nondiphtherial Corynebacterium species isolated from clinical specimens of patients in a University hospital**, Rio de Janeiro, Brazil. Brazilian J Microbiol 34: 39-44.

CAMELLO, THEREZA CRISTINA FERREIRA. **Corinebacterioses humanas em países emergentes - Corynebacterium pseudodiphtheriticum - patógenos relevantes.**. 2008. Tese (Doutorado em Ciências Médicas) - Universidade do Estado do Rio de Janeiro, . Orientador: Ana Luiza de Mattos Guaraldi.

CHART, GOOGLE. Disponível em: <http://code.google.com/apis/chart>. Acesso em: 10 de dezembro de 2011.

CHOW, T., RAHMAN, M., WU, S. (2006). **Contentbased image retrieval by using tree-structured features and multi-layer self-organizing map, Pattern Analysis and Applications**, Vol. 9, No 1, Springer-Verlag, London, pp. 1-20.

CLARRIDGE, J.E., AND C.A. SPIEGEL. 1995. **Corynebacterium and miscellaneous irregular grampositive rods, Erysipelothrix, and Gardnerella**. In: Murray, P. R., E. J. Baron, M. A. Pfaller, F. C. Tenover, and R. H. Yolken (ed.). Manual of clinical microbiology, 6th ed. American Society for Microbiology, Washington, D.C.

COLLINS, M.D., BERNARD, K.A., HUTSON, R.A., SJODEN, B., NYBERG, A., AND FALSEN, E. "**Corynebacterium sundsvallense sp. nov., from human clinical specimens.**" Int. J. Syst. Bacteriol. (1999) 49:361-366.

COLLINS, M.D., FALSEN, E., AKERVALL. E., SJODEN, B., AND ALAVAEZ, A. "**Corynebacterium kroppenstedtii sp. nov., a novel corynebacterium that does not contain mycolic acids**." Int. J. Syst. Bacteriol. (1998) 48:1449-1454.

COX, R. P., AND J. K. THOMSEN. 1990. **Computer-aided identification of lactic acid bacteria using the API CHL system**. Lett. Appl. Microbiol.

DAMASCO PV, PIMENTA FP, FILARDY AA, BRITO SM, ANDRADE AF, LOPES GS, HIRATA R JR, MATTOS-GUARALDI AL 2005. **Prevalence of IgG diphtheria antitoxin in blood donors in Rio de Janeiro**. Epidemiol Infect 133: 911-914.

DOWNS J, HARRISON RF, KENNEDY RL, CROSS SS. **Application of the fuzzy ARTMAP neural network model to medical pattern classification tasks**. Artif Intell Med 1996; 8: 403–28.

DUERDEN B.I., ELEY A., GOODWIN L., MAGEE J.T., HINDMARCH J.M., BENNET K.W.: **A comparison of Bacteroides ureolyticus isolates from different clinical sources.** J.Med.Microbiol. 29, 63–73 (1989).

DYBOWSKI, W. & FRANKLIN, D. A. (1968). **Conditional probability and the identification of bacteria : a pilot study**. Journal of General Microbiology 54, 21 5-229.

EMBRAPA. Disponível em: <http://www.cnpab.embrapa.br/publicacoes/download/doc234.pdf>. Acesso em: 05 de janeiro de 2012.

EARTHLIFE. Disponível em: < http://earthlife.wikispaces.com/space.discussion.Kingdom+Monera>. Acesso em: 05 de janeiro de 2012

FEITOSA, RAUL Q., VELLASCO, MARLEY M.B.R., OLIVEIRA, DANILO T., ANDRADE, DIOGO V., MAFFRA, SÉRGIO A. R. S. **Classificação de Expressões Faciais Utilizando Redes Neurais Back Propagation e RBF**,

Workshop de Computação (WORKCOMP´99), pp. 69-76, ITA, São José dos Campos, SP, 5 e 6 de outubro de 1999.

FERNADEZ-NATAL MI, SAEZ-NIETO JA, FERANDEZ-ROBLAS R, ASENCIO M, VALDEZATE S, LAPEÑA S, RODRIGUEZ-POLLAN RH, GUERRA JM, BLANCO J, CACHON F, SORIANO F. **The isolation of Corynebacterium coyleae from clinical samples: clinical and microbiological data**. Eur J Clin Microbiol Infect Dis 2008; 27(3):177-184.

FERNANDES, J. M. C. **GeneBrowser – Sistema de Recuperação de Dados Biológicos**. 118f. Dissertação (Mestrado) - Universidade de Aveiro. Engenharia Electrónica e Telecomunicações. Aveiro, PT, 2009.

FLORES, O. , L.A. BELANCHE, AND A.R. BLANCH. 2009. **New multiplatform computer program for numerical identification of microorganisms**. J Clin Microbiol 47:4133-4135.

FRENEY, J., M. T. DUPERRON, C. COURTIER, W. HANSEN, F. ALLARD, J. M. BOEUFGRAS, D. MONGET, AND J. FLEURETTE. 1991. **Evaluation of API Coryne in comparison with conventional methods for identifying coryneform bacteria**. J. Clin. Microbiol. 29:38-41.

FUNKE, G., EFSTRATIOU, A,. KUKLINSKA, D., HUTSON, R.A, DE ZOYSA, A., ENGLER, K.H., AND COLLINS, M.D. "**Corynebacterium imitans sp. nov. isolated from patients with suspected diphtheria**." J. Clin. Microbiol. (1997) 35:1978-1983.

FUNKE, G., HUTSON, R.A., HILLERINGMANN, M., HEIZMANN, W.R., AND COLLINS, M.D. "**Corynebacterium lipophiloflavum sp. nov. isolated from a patient with bacterial vaginosis**." FEMS Microbiol. Lett. (1997) 150:219-224.

FUNKE, G., LAWSON, P.A., AND COLLINS, M.D. "**Corynebacterium mucifaciens sp. nov., an unusual species from human clinical material.**" Int. J. Syst. Bacteriol. (1997) 47:952-957.

FUNKE, G., LAWSON, P.A., AND COLLINS, M.D. "**Corynebacterium mucifaciens sp. nov., an unusual species from human clinical material.**" Int. J. Syst. Bacteriol. (1997) 47:952-957.

FUNKE, G., LAWSON, P.A., AND COLLINS, M.D. "**Corynebacterium riegelii sp. nov., an unusual species isolated from female patients with urinary tract infections.**" J. Clin. Microbiol. (1998) 36:624-627.

FUNKE, G., OSORIO, C.R., FREI, R., RIEGEL, P., AND COLLINS, M.D. "**Corynebacterium confusum sp. nov., isolated from human clinical specimens**." Int. J. Syst. Bacteriol. (1998) 48:1291-1296.

FUNKE, G., RAMOS, C.P., AND COLLINS, M.D. "**Corynebacterium coyleae sp. nov., isolated from human clinical specimens**." Int. J. Syst. Bacteriol. (1997) 47:92-96.

FUNKE, G.; BERNARD, K.A. Coryneform Gram-Positive Rods. In Murray, P.R.; Baron E.J.; Pfaller, M.A. Tenover, C.; Yolken, R.H. (eds). **Manual of Clinical Microbiology**, 9th ed. ASM Press. Washington, D.C, 2007, p: 485-514.

GANT, V., RODWAY, S., & WYATT, J. (2001). **Artificial neural networks: Practical considerations for clinical applications.** In V. Gant, & R. Dybowski (Eds.), Clinical applications of artificial neural networks (pp. 329–356). Cambridge: Cambridge University Press.

GARCIA, S. C. O **Uso de Árvores de Decisão na Descoberta de Conhecimento na Área da Saúde**, 2003. 88f. Dissertação (Mestrado) - Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR-RS, 2003.

GARRIT, GEORGE M.; DE VOS, PAUL; KRIEG, NOEL R. **Bergey's manual of systematic bacteriology.** 2nd ed. New York : Springer, 2001-2011.

GENBANK, WIKIPEDIA. Disponível em: < http://en.wikipedia.org/wiki/File:Growth_of_genbank.png >. Acesso em: 06 de janeiro de 2012.

GIACOMINI M., RUGGIERO C., BERTONE S., CALEGARI L.: **Artificial neural network identification of heterotrophic marine bacteria based on their fatty-acid composition**. IEEE T.Bio-Med.Eng. 44, 1185–1191 (1997).

GIACOMINI M., RUGGIERO C., CALEGARI L., BERTONE S.: **Artificial neural network based identification of environmental bacteria by gas-chromatographic and electrophoretic data**. J.Microbiol.Meth. 43, 45–54 (2000).

GCHART. Disponível em: <http://code.google.com/apis/chart/interactive/docs/gallery.html>. Acesso em: 05 de janeiro de 2012.

GNU. Disponível em: < http://www.gnu.org/licenses/gpl-howto.pt-br.html >. Acesso em: 05 de janeiro de 2012.

GOELZER, M. P. **Análise de Técnicas de Mineração de Dados Aplicada na Bioinformática na Descoberta de Drogas**. 2007. 121f. Monografia - Universidade de Passo Fundo. Intituto de Ciências Exatas e Geociências, Graduação em Ciência da Computação. Passo Fundo, BR-RS, 2007

GOODACRE R., TIMMINS E.M., BURTON R., KADERBHAI N., WOODWARD A.M., KELL D.B., ROONEY P.J.: **Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks**. Microbiology 144, 1157–1170 (1998).

GOODACRE R., TIMMINS E.M., ROONEY P.J., ROWLAND J.J., KELL D.B.: **Rapid identification of Streptococcus and Enterococcus species using diffuse reflectance-absorbance Fourier transform infrared spectroscopy and artificial neural networks**. FEMS Microbiol.Lett. 140, 233–239 (1996).

GUARALDI, A. L. ; LEVY, C. E. . **Bacilos Gram Positivos. Manual de Identificação de Bactérias de Importância Médica**. Brasília: ANVISA, 2008, v. , p. 1-13.

GUARALDI, A. L.. **Manual de Diagnóstico Laboratorial**. Rio de Janeiro: UERJ - Laboratório de Difteria e Corinebactérias de Importância Clínica, 2010.

GYLLENBERG, H. G. 1965. **A model for computer identification of microorganisms**. J. Gen. Microbiol. 39:401-405.

GYLLENBERG, M. & KOSKI, T. 2001. **Probabilistic models for bacterial taxonomy.** International Statistical Review 69: 249-276

HAMILTON-MILLER, J. M. T. 1993. **A possible pitfall in the identification of Pasteurella spp. with the API system**. J. Med. Microbiol. 39:78-79.

HAMILTON-MILLER, J. M. T., and S. Shah. 1996. **Anomalous but helpful findings from the BBL Crystal ID kit with Haemophilus spp**. Lett. Appl. Microbiol. 23:47-48.

HAYKIN, S. **Neural Networks and Learning Machines**. Prentice Hall, 3rd edition, (2008).

HAYKIN, S., **Neural Networks: A Comprehensive Foundation**, Macmillan College Publishing Company, Inc., 1994.

Heden B, Edenbrandt L, Haisty Jr WK, Pahlm O. **Artificial neural networks for the electrocardiographic diagnosis of healed myocardial infarction**. Am J Cardiol 1994; 74: 5–8.

HORNIK, K., STINCHCOMBE, M, WHITE, H., **Multilayer Feedforward Networks are Universal Approximators**, Neural Networks, Vol. 2, pp. 359-366, 1989.

INFOVIS. Disponível em: <http://thejit.org/>. Acesso em: 05 de novembro de 2011.

JANDA, J. M., AND S. L. ABBOTT. 2002. **Bacterial identification for publication: when is enough enough?** J. Clin. Microbiol. 40:1887-1891.

JANDA, W.M. **Corynebacterium species and the coryneform bacteria Part I: New and emerging species in the genus Corynebacterium**. Clin. Microbiol. Newslett., 20: 41-52, 1998.

JANDA, W.M. **The corynebacteria revisited: new species, identification kits, and antimicrobial susceptibility testing**. Clin. Microbiol. Newslett., 21: 175-182, 1999.

JENA RK, AQEL MM, SRIVASTAVA, MAHANTI PK. **Soft Computing Methodologies in Bioinformatics**. Europian Journal of Scientific Research, 2009. 26(2): p. 189-203.

JOOMLA, Disponível em: <http://docs.joomla.org/Framework/>, Acesso em: 10 de janeiro de 2012.

KARAKITSOS P, COCHAND-PRIOLLET B, GUILLAUSSEAU PJ, POULIAKIS A. **Potential of the back propagation neural network in the morphologic examination of thyroid lesions**. Anal Quant Cytol Histol 1996; 18: 495–500.

KENNEDY M.J., THAKUR M.S.: **The use of neural networks to aid in microorganism identification: a case study of Haemophilus species identification**. Antonie van Leeuwenhoek 63, 35–38 (1993).

KHAMIS, A.; RAOULT, D.; La SCOLA, B. **Comparison between rpoB and 16S rRNA gene sequencing for molecular identification of 168 clinical isolates of Corynebacterium.** Journal of Clinical Microbiology, v. 43, n. 4, p. 1934–1936, Apr. 2005.

KOHONEN T. (1984). **Self-Organization and Associative Memory**. Springer, Berlin.

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

KOHONEN, T. (2001). **Self-Organizing Maps. Third, extended edition**. Springer.

KONEMAN, E.W.; ALLEN, S.D.; JANDA, W.M.; SCHRECKENBERGER, P.C.; WINN Jr., W.C. **Diagnóstico Microbiológico**. 5.ed., Rio de Janeiro: MEDSI, 2001.

KONONENKO I., BRATKO I., KUKAR M., **Application of machine learning to medical diagnosis**. In R.S.Michalski, I.Bratko, and M.Kubat (eds.): Machine Learning, Data Mining and Knowledge Discovery: Methods and Applications, John Wiley & Sons, 1998.

KRAEVA LA, MANINA Z, TSENEVA GI, RADCHENKO AG. **Etiologic role of *Corynebacterium* non diphtheriae in patients with different pathology**. Zh.Mikrobiol.Epidemiol.Immunobiol. 2007; 3-7.

LAPAGE, S. P., BASCOMB, S., WILLCOX, W. R. & CURTIS, M. A . (1970). **Computer identification of bacteria**. In Automation, Mechanization and Data Handling in Microbiology (Society for Applied Bac- teriology Technical Series no. 4), pp. 1-22. Edited by A. Baillie & R. J . Gilbert. London: Academic Press.

LAPAGE, S. P., S. BASCOMB, W. R. WILLCOX, AND M. A. CURTIS. 1973. **Identification of bacteria by computer: general aspects and perspectives**. J. Gen. Microbiol. 77:273-290.

LEE J. LANCASHIRE, CHRISTOPHE LEMETRE, AND GRAHAM R. BALL. **An introduction to artificial neural networks in bioinformatics—application to complex microarray and mass spectrometry datasets in cancer studies**. Brief Bioinform (2009) 10(3): 315-329 first published online March 23, 2009 doi:10.1093/bib/bbp012

LOPES, L. **Aprendizagem de máquina baseada na combinação de classificadores em bases de dados da área de saúde**, 2007. 117f. Dissertação (Mestrado) – Pontifícia Universidade Católica do Paraná – Centro de Ciências Biológicas e da Saúde. Programa de Pós-Graduação em Tecnologia em Saúde. Curitiba, BR-PR, 2007.

MARIEY L., SIGNOLLE J.P., AMIEL C., TRAVERT J.: **Discrimination, classification, identification of microorganisms using FTIR spectroscopy and chemometrics**. Vib.Spectrosc. 26, 151–159 (2001).

MATHWORKS. **Neural network toolbox. User's guide - Version 4. The MathWorks**, Inc. 2005.

MICHIE D., SPIEGELHALTER D.J., TAYLOR C.C (eds.) **Machine learning, neural and statistical classi¯cation**, Ellis Horwood, 1994.

MILLER, J. M., AND P. ALACHI. 1996. **Evaluation of new computer-enhanced identification program for microorganisms: adaptation of BioBase for identification of members of the family Enterobacteriaceae**. J. Clin. Microbiol. 34:179-181.

MIRANDA, J.M. **Redes neurais de Kohonen como modelos da topografia do sistema visual**. Dissertação de mestrado. COPPE/UFRJ, Rio de Janeiro, RJ, Brasil, 1998.

MITCHELL T., **Machine Learning**, MCGraw Hill, 1997.

MOLLER, **Neural Networks**, Vol. 6, 1993, pp. 525–533

MURRAY, PATRICK R; BARON, ELLEN JO. . **Manual of clinical microbiology**. 9th. ed. Washington: American Society for Microbiology, 2007 2v. ISBN 9781555813710

MYSQL, WIKIPEDIA. Disponível em: < http://en.wikipedia.org/wiki/MySQL>. Acesso em: 10 de dezembro de 2011.

NAPOLI, W. F. M. **Introdução a Bioinformática**, 2003. 163f. Monografia - Universidade Federal de Goiás. Graduação em Engenharia de Computação. Goiânia, BR-GO, 2003.

NARAYANAN A., E. KEEDWELL, AND B. OLSSON (2003), "**Artificial Intelligence Techniques for Bioinformatics**", Applied Bioinformatics, Vol.1, No. 4, pp. 191-222.

NCBI. Disponível em: <ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt>. Acesso em: 06 de janeiro de 2012.

O'HARA, C. M., D. L. RHODEN, AND J. M. MILLER. 1992. **Reevaluation of the API 20E identification system versus conventional biochemicals for identification of members of the family Enterobacteriaceae: a new look at an old product**. J. Clin. Microbiol. 30:123-125.

PAGANO N., M.BUSCEMA, E.GROSSI, M.INTRALIGI, G.MASSINI, P.SALACONE, ET AL., **Artificial neural networks for the prediction of diabetes mellitus occurrence in patients affected by chronic pancreatitis**, Journal of Pancreas 5(Suppl.5) (2004) 405–453.

PANDEY B, Mishra RB, "**Knowledge and intelligent computing system in medicine**", Computers in Biology and Medicine, vol. 39, (2009), pp. 215-230.

PASCUAL C, LAWSON PA, FARROW JA, GIMENEZ MN, COLLINS MD. **Phylogenetic analysis of the genus Corynebacterium based on 16S rRNA gene sequences**. Int J Syst Bacteriol. 1995 Oct;45(4):724-8. PubMed PMID: 7547291.

PEIXOTO, R. S. ; COUTINHO, H. L. ; RUMJANEK, N. G. ; MACRAE, A. ; ROSADO, A. S.. **Use of rpoB and 16S rRNA genes to analyse bacterial diversity from a tropical soil using PCR/DGGE**. Letters in Applied Microbiology, Inglaterra, v. 35, p. 316-320, 2002.

PHP, WIKIPEDIA. Disponível em: < http://en.wikipedia.org/wiki/PHP >. Acesso em: 02 de dezembro de 2011.

RATAJ T., SCHINDLER J.: **Identification of bacteria by a multilayer neural network**. Binary 3, 159–164 (1991).

RHODEN, D. L., G. A. HANCOCK, AND J. M. MILLER. 1993. **Numerical approach to reference identification of Staphylococcus, Stomatococcus, and Micrococcus spp**. J. Clin. Microbiol. 31:490-493.

RIEDMILLER, **Proceedings of the IEEE International Conference on Neural Networks (ICNN)**, San Francisco, 1993, pp. 586-591

RIEGEL, P., HELLER, R., PREVOST, G., JEHL, F., AND MONTEIL, H. "**Corynebacterium durum sp. nov., from human clinical specimens**." Int. J. Syst. Bacteriol. (1997) 47:1107-1111.

RIEGEL, P., RUIMY, R., DE BRIEL, D., PREVOST, G., JEHL, F., BIMET, F., CHRISTEN, R., MONTEIL, H. "**Corynebacterium argentoratense sp. nov., from the human throat**." Int. J. Syst. Bacteriol. (1995) 45:533-537.

RIEGEL, P.; RUIMY, R.; CHRISTEN, R.; MONTEIL, H. **Species identities and antimicrobial susceptibilities of Corynebacteria isolated from various clinical sources**. Eur. J. Clin. Microbiol. Infect. Dis., 15: 657-662, 1996.

RIPLEY, B. D. **Pattern Recognition and Neural Networks**. Cambridge University Press, (2008).

ROSSELLÓ-MORA, R., AND R. AMANN. 2001. **The species concept for prokaryotes**. FEMS Microbiol. Lett. 25:39-67.

RUMMELHART RUMELHART, DAVID E., GEOFFREY E. HINTON, AND R. J. WILLIAMS. "**Learning Internal Representations by Error Propagation**". David E. RUMELHART, JAMES L. MCCLELLAND, AND THE PDP RESEARCH GROUP. (editors), Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations. MIT Press, 1986.

SAHIN, N., AYDIN, S., 2006. **Identification of oxalotrophic bacteria by neural network analysis of numerical phenetic data**. Folia Microbiol (Praha). 51(2), 87-91.

SCHNIDER, J. 1979. **Computer-aided numerical identification of gram negative fermentative rods on a desk-top computer**. J. Appl. Bacteriol. 47:45-51.

SEWELL DL, COYLE MB, FUNKE G. **Prosthetic valve endocarditis caused by Corynebacterium afermentans subsp. lipophilum (CDC coryneform group ANF-1)**. J Clin Microbiol 1995; 33:759-61.

SHANNON, CLAUDE E. & WEAVER, WARREN (1949): **The Mathematical Theory of Communication**. The University of Illinois Press, Urbana (Ill.). ISBN 0-252-72548-4.

SHAVLIK J.W., DIETTERICH T.G. (eds.) **Readings in machine learning**, Morgan Kaufmann Publ., 1990. Spiegelhalter D.J., Philip Dawid A., Lauritzen S.L. and Cowell R.G., Bayesian analysis in expert systems, Statistical Science, 8(3):219-283, 1993.

SILVA, LEANDRO AUGUSTO DA. **Categorização de imagens médicas baseada em transformadas wavelet e mapas auto-organizáveis**. 2009. 127f. Tese (Doutorado) – Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Sistemas Eletrônicos. São Paulo, BR-SP, 2009.

SJODEN, B., FUNKE, G., IZQUIERDO, A., AKERVALL, E., AND COLLINS, M.D. "**Description of some coryneform bacteria isolated from human clinical specimens as Corynebacterium falsenii sp. nov**." Int. J. Syst. Bacteriol. (1998) 48:69-74.

SNEATH, P. H. A. (1964). **New approaches t o bacterial taxonomy: The use of computers.** Ann. Rev. Microbiolog. 18, 335.

SNEATH, P. H. A. 1957. **Some thoughts on bacterial classification**. J. Gen. Microbiol. 17:184-200.

SNEATH, P. H. A. 1957. **The application of computers to taxonomy**. J. Gen. Microbiol. 17:201-226.

SNEATH, P.H. & R. R. SOKAL. 1962. **Numerical taxonomy**. Nature 193:855-860.

STAMEY TA, BARNHILL SD, ZANG Z. **Effectiveness of ProstAsureTM in detecting prostate cancer (PCa) and benign prostatic hyperplasia (BPH) in men age 50 and older**. J Urol 1996; 155: 436A.

STEIMANN F. **On the use and usefulness of fuzzy sets in medical AI**. Artif Intell Med 2001; 21: 131–7.

TAGUCHI M, NISHIKAWA S, MATSUOKA H, NARITA R, ABE S, FUKUDA K, MIYAMOTO H, TANIGUCHI H, OTSUKI M.. **Pancreatic abscess caused by Corynebacterium coyleae mimicking malignant neoplasm**. Pancreas 2006; 33:425-429.

THOMSON RB JR, MILLER M 2007. **Specimen collection, transport and processing: bacteriology.** In Manual of Clinical Microbiology. Edited by P. R. Murray, E. J. Baron, J. H. Jorgensen, M. L. Landry & M. A. Pfaller. Washington, D.C.: ASM Press. p. 286-330.

VALIATI, J. F. **Redes Neurais Aplicadas ao Reconhecimento de Regiões Promotoras na Família Mycoplasmataceae.** 2006. 150f. Tese (Doutorado) - Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR-RS, 2006.

VANDAMME, P., B. POT, M. GILLIS, P. DE VOS, K. KERSTERS & J. SWINGS. 1996. **Polyphasic taxonomy, a consensus approach to bacterial systematics.** Microbiol Rev 60:407-438.

VESANTO, J. **Data exploration process based on the Self-Organizing Map. Dissertation for the degree of Doctor of Technology**. Helsinki University of Technology. Espoo, Finland. 2002.

VESANTO, J.; HIMBERG, J.; ALHONIEMI, E.; PARHANKANGAS, J. **SOM toolbox for Matlab 5**. Helsink University of Technology, 2000

WASSERMAN, P.D., **Advanced Methods in Neural Computing**, Van Nostrand Reinhold, 1993.

WELLING, L. & THOMSON, L. (2003). **PHP and MySQL Web development** (2ª ed). Sams Publishing

WILLCOX, W. R., S. P. LAPAGE, S. BASCOMB, AND M. A. CURTIS. 1973. **Identification of bacteria by computer: theory and programming**. J. Gen. Microbiol. 77:317-330.

WILLIAMS, D.Y.; SELEPAK, S.T.; GILL, V.J. **Identification of clinical isolates of nondiphtherial Corynebacterium species and their antibiotic susceptibility patterns**. Diag. Microbiol. Infect. Dis., 17: 23-28, 1993.

Z.-H.ZHOU,Y.JIANG,Y.-B.YANG,S.-F.CHEN,**Lung cancer cell identification based on artificial neural network ensemble,** Artificial Intelligence in Medicine 24 (2002) 23 -36.

ZIMMERMANN, O., SPROER, C., KROPPENSTEDT, R.M., FUCHS, E., KOCHEL, H.G., AND FUNKE, G. "**Corynebacterium thomssenii sp. nov., a Corynebacterium with N-acetyl-beta-glucosaminidase activity from human clinical specimens**." Int. J. Syst. Bacteriol. (1998) 48:489-494.

8 Anexos

ANEXO A - PLANILHA EXAME LABORATORIAL DE CORINEBACTERIOSES

LABORATÓRIO DE DIFTERIA E CORINEBACTERIOSES
DE IMPORTÂNCIA CLÍNICA – (LDCIC)
DMI/FCM/UERJ

LDCIC PLANILHA EXAME LABORATORIAL DE CORINEBACTERIOSES

DATA DE ENTRADA:____/____/____. **Nº DA AMOSTRA:**

Responsável pelo envio:_____ E-mail e/ou telefone:_____
Responsável pelo recebimento:_____ _____

Conformidade da ☐ amostra clínica e/ou ☐ cultura bacteriana: ☐ ACEITO ☐ ACEITO COM RESTRIÇÕES ☐ RECUSADO
Observação:_____

DADOS DO PACIENTE

Nome:_____ Iniciais:_____ Idade:_____ Sexo: ☐ Fem ☐ Masc
Registro de origem:_____ Bairro:_____ Município:_____
Ocupação:_____ Instituição de origem:_____ Data de Internação:____/____/____
☐ Processo infeccioso comunitário
☐ Processo infeccioso hospitalar
Setor: ☐ Amb ☐ Enf ☐ CTI Leito:_____ Outros:_____ Sítio de isolamento:_____

Imunização prévia contra difteria: ☐ SIM ☐ NÃO ☐ Completa ☐ Incompleta Número de doses:_____ **☐ Desconhecido**

MATERIAL BIOLÓGICO

☐ Swab ☐ Nasofaringe ☐ Orofaringe ☐ Ouvido ☐ Lesão cutânea
Localização:_____
☐ Aspiração traqueal ☐ Lavado bronco-alveolar ☐ Escovado brônquico ☐ Hemocultura: ☐ Periférica ☐ Cateter ☐ Urina: ☐ Jato médio
☐ Cateter de alívio ☐ Cateter vesical de demora ☐ Punção suprapúbica ☐ Saco coletor ☐ Abscesso Nível da coleta: ☐ Superficial ☐ Profunda
☐ Líquido: ☐ Cefalorraquidiano ☐ Pericárdico ☐ Pleural ☐ Peritoneal ☐ Sinovial
☐ Prótese ☐ Ponta de cateter Outros:_____

DADOS EPIDEMIOLÓGICOS

☐ Internação hospitalar recente - Motivo:_____ Setor:_____
☐ Viagens recentes Local:_____ ☐ Reside em área rural ☐ Acidente recente Tipo:_____
☐ Contato com animais Quais?_____

SINAIS E SINTOMAS

Início:____/____/____
Pseudomembrana: ☐ NÃO ☐ SIM **Localização:** ☐ Nasofaringe ☐ Tonsilas ☐ Faringe ☐ Laringe ☐ Palato ☐ Pele
☐ Edema ganglionar ☐ Edema de pescoço ☐ Prostração ☐ Miocardite ☐ Febre ☐ Palidez ☐ Artrite ☐ Endocardite ☐ Lesão cutânea
☐ Outros:_____ ☐ Complicações:_____

COMORBIDADES

☐ Transplantado ☐ HIV ☐ Hepatite ☐ Imunodeficiência ☐ Diabético ☐ Fibrose Cística ☐ Quimioterapia ☐ Insuficiência Respiratória
☐ Ventilação Mecânica ☐ Traqueostomia ☐ Cateter venoso profundo ☐ Cateter vesical ☐ Cateter de diálise ☐ Esplenectomia ☐ BK
☐ Anemia Tipo:_____ ☐ Outras:_____
Uso de Antibióticos: ☐ NÃO ☐ SIM Uso de Antifúngicos: ☐ NÃO ☐ SIM Uso de Surfactantes: ☐ NÃO ☐ SIM
Data:____/____/____ Especificar:_____ Data:____/____/____ Especificar:_____
Data:____/____/____ Especificar:_____ Data:____/____/____ Especificar:_____
Motivo:_____

DADOS COMPLEMENTARES

Eritrograma		**Leucograma**		**Uréia:**_____ mg/dL	8 – 20
Hemácias:_____ milhões/mm³	4,5 - 5,9	Leucócitos:_____ mm³	4.300 – 10.000	**Creatinina:**_____ mg/dL	0,8 – 1,2
Hemoglobina:_____ g/dL	13,5 – 17,5	Basófilos:_____ mm³	0 – 100	**AST:**_____ UI/L	12 – 31
Hematócrito:_____ %	41,0 – 52,0	Eosinófilos:_____ mm³	45 – 500	**ALT:**_____ UI/L	8 – 50
VCM:_____ fL	80 – 100	Neutrófilos:_____ mm³	1.500 – 7.800	**Gama GT:**_____ UI/L	6 – 38

DADOS RELEVANTES DE IMAGENS

_____
_____
_____
_____

DADOS DA AMOSTRA

Situação: □ CULTURA PURA □ CULTURA MISTA

Microrganismos associados e contagem: _____

EXAME MICROSCÓPICO

□ Direto □ A fresco □ Campo escuro

□ Gram □ Albert-Layboum □ Kinyoun

Resultado: _____

ANÁLISE GENOTÍPICA

□ mPCR

□ PFGE

□ RAPD

□ SDS-PAGE

ANÁLISE FENOTÍPICA

PROVAS BIOQUÍMICAS CONVENCIONAIS																SISTEMAS SEMI - AUTOMATIZADOS		
DNase	Uréia	NO_3	Gli	Mal	Sac	Tre	Rib	Mn	Mt	Ara	Xil	Fru	Gal	CAMP	O129	API Coryne Código /%ID	BBL CRYSTAL	OUTROS
Observações adicionais: _____ _____																		

ANTIBIOGRAMA: □ Disco de difusão □ E-teste

Antibiótico	Pen	Amp	Tet	Rif	Gen	Lzd	Ceftri	Cefotax	Imip	Cipro	Eri	Clin	Van	Trim-Sulfa	Nitro
Halos/ Mm															
Perfil: S/I/R															
MICS															
Outros antimicrobianos testados: _____ _____															
Observações adicionais: _____ _____															

RESULTADO FINAL

□ Pesquisa **NEGATIVA** para *C. diphtheriae*, *C. ulcerans* e *C. pseudotuberculosis*
□ Pesquisa **POSITIVA** para *C. diphtheriae* □ Atoxinogênica □ Toxinogênica
□ Pesquisa **POSITIVA** para *C. ulcerans* □ Atoxinogênica □ Toxinogênica
□ Pesquisa **POSITIVA** para *C. pseudotuberculosis* □ Atoxinogênica □ Toxinogênica

Observações: _____
_____
_____
_____
_____

Laudo enviado para: _____ Forma de envio: □ Eletrônica □ Correspondência

Endereço: _____
_____

DATA DE SAÍDA: ___/___/_____ **RESPONSÁVEL:** _____

Anexo B - Código e procedimento para a execução do componente LeavesPHP

```php
<?php

$atributos = array("frequencia_cardiaca", "frequencia_respiratoria", "perda_apetite", "medicacao");


$tabela = array(
array("frequencia_cardiaca"=>"normocardico", "frequencia_respiratoria"=>"taquipneia", "perda_apetite"=>"nao", "medicacao"=>"sim"),
array("frequencia_cardiaca"=>"normocardico", "frequencia_respiratoria"=>"taquipneia", "perda_apetite"=>"sim", "medicacao"=>"sim"),
array("frequencia_cardiaca"=>"taquicardico", "frequencia_respiratoria"=>"taquipneia", "perda_apetite"=>"nao", "medicacao"=>"sim"),
array("frequencia_cardiaca"=>"bradicardico", "frequencia_respiratoria"=>"taquipneia", "perda_apetite"=>"nao", "medicacao"=>"nao"),
array("frequencia_cardiaca"=>"bradicardico", "frequencia_respiratoria"=>"taquipneia", "perda_apetite"=>"nao", "medicacao"=>"sim"),
array("frequencia_cardiaca"=>"bradicardico", "frequencia_respiratoria"=>"eupneia", "perda_apetite"=>"sim", "medicacao"=>"nao"),
array("frequencia_cardiaca"=>"taquicardico", "frequencia_respiratoria"=>"eupneia", "perda_apetite"=>"sim", "medicacao"=>"sim"),
array("frequencia_cardiaca"=>"normocardico", "frequencia_respiratoria"=>"taquipneia", "perda_apetite"=>"nao", "medicacao"=>"nao"),
array("frequencia_cardiaca"=>"normocardico", "frequencia_respiratoria"=>"eupneia", "perda_apetite"=>"nao", "medicacao"=>"nao"),
array("frequencia_cardiaca"=>"bradicardico", "frequencia_respiratoria"=>"eupneia", "perda_apetite"=>"nao", "medicacao"=>"nao"),
array("frequencia_cardiaca"=>"normocardico", "frequencia_respiratoria"=>"eupneia", "perda_apetite"=>"sim", "medicacao"=>"nao"),
array("frequencia_cardiaca"=>"taquicardico", "frequencia_respiratoria"=>"taquipneia", "perda_apetite"=>"sim", "medicacao"=>"nao"),
array("frequencia_cardiaca"=>"taquicardico", "frequencia_respiratoria"=>"eupneia", "perda_apetite"=>"nao", "medicacao"=>"sim"),
array("frequencia_cardiaca"=>"bradicardico", "frequencia_respiratoria"=>"taquipneia", "perda_apetite"=>"sim", "medicacao"=>"nao"),
);

$sub_arvore = array();
$backup = "";

$json_saida = arvore($tabela, $atributos, $sub_arvore,$backup);

?>
```

Anexo C – Tabela de microrganismos com seus respectivos nomes e números de identificação

Identificação	Microrganismo
1	C. accolens
2	C. afermentans var afermentans
3	C. afermentans var lipophilum
4	C. amycolatum
5	C. appendicis
6	C. argentoratense
7	C. atypicum
8	C. aurimucosum
9	C. auris
10	C. bovis
11	C. confusum
12	C. coyleae
13	C. cystitidis
14	C. diphtheriae var belfanti
15	C. diphtheriae var gravis
16	C. diphtheriae var intermedius
17	C. diphtheriae var mitis
18	C. durum
19	C. falsenii
20	C. flavescens
21	C. freneyi
22	C. glucuronolyticum
23	C. imitans
24	C. jeikeium
25	C. kroppenstedtii

26	C. kutscheri
27	C. lipophiloflavum
28	C. macginleyi
29	C. matruchotii
30	C. minutissimum
31	C. mucifaciens
32	C. pilosum
33	C. propinquum
34	C. pseudodiphtheriticum
35	C. pseudotuberculosis
36	C. renale
37	C. resistens
38	C. riegelii
39	C. seminale
40	C. simulans
41	C. singulare
42	C. striatum
43	C. sundsvallense
44	C. thomssenii
45	C. tuberculostearicum
46	C. tuscaniae
47	C. ulcerans
48	C. urealyticum
49	C. xerosis
50	CDC group F-1
51	CDC group G