UNIVERSIDADE FEDERAL DO PARANÁ

KELLY RAFAELA OTEMAIER

BioSOM: Metodologia para identificação de sinônimos de genes utilizando *Self-Organizing Maps*.

CURITIBA

2012

KELLY RAFAELA OTEMAIER

BioSOM: Metodologia para identificação de sinônimos de genes utilizando *Self-Organizing Maps*.

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de Mestre em Bioinformática.

Orientador:
Jeroniza Nunes Marchaukoski, Drª.
Coorientador:
Maria Berenice R. Steffens, Drª.

CURITIBA

2012

TERMO DE APROVAÇÃO

KELLY RAFAELA OTEMAIER

BIOSOM: Metodologia para identificação de sinônimos de genes utilizando Self-Organizing Maps

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Bioinformática, pelo Programa de Pós-graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientador: Profª Drª Jeroniza Nunes Marchaukoski

Coorientador: Profª Drª Maria Berenice Reynaud Steffens

Prof. Dr. Luiz Antonio Pereira Neves
Universidade Federal do Paraná - UFPR

Prof. Dr. Júlio Cesar Nievola
Pontifícia Universidade Católica do Paraná - PUCPR

Curitiba, 17 de fevereiro de 2012

À Claudinei Ribeiro,

que sempre me incentivou e deu apoio à minha

formação, dedico este trabalho.

AGRADECIMENTOS

À Deus, por estar sempre a meu lado.

À Drª Jeroniza Marchaukoski, obrigada pela confiança, apoio, incentivo e, sobretudo pela compreensão e oportunidade.

À Drª Maria Berenice, orientadora com uma paciência sem igual, obrigada pela confiança, incentivo e principalmente por ter me ensinado a ter sempre uma visão crítica sobre o meu trabalho.

Ao professor (e orientador) Dr. Roberto Raittz, pela maneira otimista de enfrentar os obstáculos, obrigada por sempre me motivar quando eu mais precisei e pelas ideias sem fim.

Ao professor Dr. Júlio Nievola que foi muito solícito em me atender e ao me dar dicas valiosas sobre Kohonen.

Ao professor Dr. Luiz A.P. Neves, que mesmo sem querer, me apresentou este PPG.

Ao professor Dr. Lucas Ferrari pela amizade, apoio, conversas divertidas e estar sempre presente no laboratório a disposição de quem precisa de ajuda.

A todos os professores do PPG em Bioinformática pelo aprendizado que cada deixou.

A Drª. Andreia Malucelli por me dar a oportunidade de conhecer e conviver em um ambiente de pesquisa. Um exemplo de docente.

Ao Dr. Orlando Alcântara Soares pelo incentivo de sempre, mesmo quando nem eu acreditava que era possível. Um dos professores mais admiráveis que já conheci.

Aos professores que me confiaram as carta de recomendação a este programa - Dr. Emerson Paraíso (a quem sou muito grata por ter me dado uma base sólida de conhecimento de programação), Dr. Mauro S. Pereira Fonseca e Dr. Ricardo Nabhen, Orlando e Andreia.

Às minhas afilhadas Letícia e Ana Carolina, pela ausência nestes dois anos, mas tenho a certeza que um dia elas terão a oportunidade de viver um mestrado e entenderão todas as minhas ausências.

Ao William, amigo e companheiro de bancada no laboratório, onde rimos, choramos, lamentamos, reclamamos e principalmente nos divertimos no dia a dia do mestrado.

Ao amigo Danhylo com quem iniciei este projeto e com quem aprendi muita coisa, o admiro por sua paciência e dedicação ao ensinar.

Aos meus colegas de laboratório em especial aos de 2009 (Rosa, Rodrigo e Guilherme), de 2010 (Sérgio, Lucas, Juliana, Vanely, Levston, Paula, Jessé) e aos de 2011 (Katia, Barbara, Rafael) pela excelente convivência que tivemos, risadas, medos, desperos e alegrias.

À Suzana pela amizade, paciência, alegria e disponibilidade de sempre.

À Lea, aos alunos de iniciação científica Gustavo, Alysson e principalmente ao Jeovane.

Ao programa de Pós Graduação em Bioinformática pela oportunidade

À Capes pelo auxílio financeiro.

Muito obrigada!!!

RESUMO

Genes e proteínas são de grande importância biológica para a compreensão de processos bioquímicos e requerem nomes consistentes. Existem diversas diretrizes para nomenclatura de genes, mas elas não são rigorosamente aplicadas à atribuição de nomes aos genes recém-identificados, gerando assim, inúmeras maneiras de nomear um mesmo gene. Este trabalho tem o objetivo de detectar e minimizar a redundância e a inconsistência de dados para colaborar com a identificação correta de genes. Para isso foram utilizadas técnicas de Inteligência Artificial para identificar os sinônimos realizando um estudo dirigido a dez experimentos distintos. Para selecionar os dados dos experimentos foi construído um banco de dados relacional para armazenar as informações constantes na base NR do NCBI e as informações identificadas neste estudo. Os dados do experimento foram minerados através das técnicas de mapas auto-organizáveis de Kohonen. A Rede SOM de Kohonen foi aplicada para exprimir as relações de similaridade entre os dados. Para identificação dos agrupamentos gerados pela rede SOM foi utilizada a técnica denominada Matriz-U. As informações resultantes deste trabalho permitem inferir os sinônimos dos genes, identificar prováveis nomes para genes nomeados como hipotéticos e apontar possíveis erros de anotação.

Palavras-chave: nomenclatura genes, agrupamento genes, Kohonen, Matriz-U, SOM.

ABSTRACT

Genes and proteins are of great biological importance for the understanding of bio-chemical processes and require consistent names. There are several guidelines for naming genes, but they are not strictly applied to the naming of the newly identified genes, thus generating many ways of naming the same gene. This work aims to minimize redundancy and inconsistency of data to work with the correct identification of genes. For that were used Artificial Intelligence techniques to identify synonyms conducting a study aimed at ten different experiments. To select the data for the experiments was built a relational database to store the information in the base of the NCBI NR and information identified in this study. The experimental data were mined through the techniques of self-organizing maps of Kohonen. The Kohonen SOM was applied to express the similarity relations between the data. To identify clusters generated by SOM was used a technique called U-Matrix. The information resulting from this work allow us to infer the synonyms of genes, identify potential names for appointment as hypothetical genes and to identify possible annotation errors.

Keywords: genes nomenclature, genes clustering, Kohonen, U-Matrix, SOM.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

API - Application Programming Interface
ASCII - American Standard Code for Information Interchange
BLAST - Basic Local Alignment Search Tool
BLASTP - Basic Local Alignment Search Tool Protein
BMU - Best Matching Unit
CGD - Saccharomyces Genome Database
CGSC - Coli Genetic Stock Center
DDBJ - DNA Database Bank of Japan
DER - Diagrama Entidade Relacionamento
EMBL - European Molecular Biology Laboratory
FTP - Protocolo de Transferência de Arquivo (File Transfer Protocol, em inglês)
GI - GeneInfo identifiers
GenBank - Banco de dados público do National Center for Biological Information, do Instituto de Saúde dos Estados Unidos da America
GPSDB - Gene and Protein Synonyms DataBase
HGNC - HUGO Gene Nomenclature Committee
HMM - Modelos ocultos de Markov
INREC - ÍNdice RECursivo
KDD - Knowledge-Discovery in Databases
MATLAB - MATrix LABoratory
MGD - Mouse Genome Database
NCBI - National Center for Biotechnology Information
NR - Non-Redundant (banco de dados de proteínas não-redundantes)
PIR-PSD - International Protein Sequence Database
POJO - Plain Old Java Object
PFam - Protein Family
RefSeq - Reference

RGD	- Rat Genome Database
SGD	- Saccharomyces Genome Database
SOM	- Self-organizing map
SQL	- Structured Query Language
UniProt	- Universal Protein Resource Sequence
XML	- Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

A preocupação com uma nomenclatura para genes existe desde que se iniciaram as anotações gênicas, ainda na época de Gregor Mendel na década de 1860.

Na década de 1960 começaram a ser estudados os primeiros problemas relacionados com a nomenclatura de gene, mesmo quando ainda eram poucos os genes humanos identificados (POVEY, *et al.*, 2001) (WAIN, *et al.*, 2002).

Vários esforços foram desenvolvidos para convencionar a nomenclatura de genes e proteínas: UniProt para a nomenclatura de proteínas (CONSORTIUM, 2009), HGNC(HUGO Gene Nomenclature Committee) comitê de nomenclatura para genes humanos (EYRE, *et al.* 2006), DEMEREC e colaboradores (1966) sugeriram uma nomenclatura para genes bacterianos. Ainda existem diversas outras bases de dados de organismo modelo que orientam sobre a nomenclatura desses genes (DWIGTH, *et al.*, 2004) (DRYSDALE, *et al.*, 2005) (BULT, *et al.,* 2008) (DWINELL, *et al.*, 2009), como por exemplo o Flybase, que orienta a nomenclatura para genes da *Drosophila.*

Apesar de existirem, essas diretrizes de nomenclatura não são rigorosamente aplicadas à atribuição de nomes aos genes recém-identificados, ou seja, todo pesquisador é livre para definir e atribuir o nome que achar mais adequado (FUNDEL e ZIMMER, 2006). Como resultado, pode haver inúmeras maneiras (nome completo, símbolo, sinônimo) de descrever o mesmo gene (PILLET, *et al.*, 2004).

Atualmente, é comum encontrar genes com a mesma função e nomes diferentes ou variações de um nome para o mesmo gene (TSURUOKA, *et al.*, 2007), fazendo assim, com que haja dúvidas e confusões em adotar um nome para anotar um novo gene. Uma maneira de minimizar o problema de nomenclatura de genes é a identificação dos sinônimos desses, ou seja, agrupar as diversas formas que um nome de gene está anotado.

A partir de 1995, com o surgimento de sequenciadores automáticos de DNA, houve um aumento exponencial no número de sequências gênicas, o que implicou no surgimento de grandes bases de dados públicas para armazenar informações biológicas (PROSDOCIMI, *et al.*, 2003). Também foi necessária a concepção de

ferramentas computacionais que auxiliassem no processo de análise dessa grande massa de dados.

Um número cada vez maior de genes é identificado a cada dia e são armazenados em diversos bancos de dados públicos, tornando ainda mais indispensável uma uniformização na nomenclatura gênica (SPLENDORE, 2005).

Com a identificação dos sinônimos dos genes haverá uma melhora na documentação das sequências gênicas depositadas nos bancos de dados públicos e isso facilitará os processos de análise de dados e anotações (principalmente as automáticas) de novas sequências gênicas.

1.1 **OBJETIVOS GERAIS**

O objetivo desse estudo é minimizar a redundância e a inconsistência de dados para colaborar com a identificação correta de genes. Para isso foi utilizada técnica de Inteligência Artificial, mais especificamente a técnica de Rede Neural Artificial denominada SOM (Self-Organizing Maps) para identificar os sinônimos dos nomes dos genes através de agrupamentos. Foram realizados dez experimentos distintos para verificar a eficiência do algoritmo SOM na identificação de sinônimos de genes.

1.1.1 Estratégias utilizadas

1. Analisar a estrutura do arquivo NR (Non-Redundant Data Base) disponibilizado pelo NCBI;

2. Criar um banco de dados relacional que integre as informações disponíveis no banco de dados NR do NCBI e as informações sobre os sinônimos dos genes adquiridas no decorrer da mineração de dados realizada neste trabalho;

3. Aplicar técnica de agrupamento para identificar os sinônimos dos nomes dos genes, baseado nos resultados obtidos com os alinhamentos de sequências de aminoácidos e na estrutura sintática dos nomes dos genes;

4. Desenvolver um script para interpretar automaticamente o formato do arquivo NR do NCBI para que as informações contidas nele possam ser armazenadas no banco de dados de forma automática;

5. Propor mecanismo de mineração de dados para a análise e interpretação dos dados;

6. Realizar experimentos com a Rede Som de Kohonen e aplicar uma técnica de visualização, para identificar e analisar os grupos obtidos de tal forma que se consiga inferir se os genes pertencentes ao mesmo grupo são sinônimos.

Estas seis estratégias são necessárias para que a metodologia possa ser desenvolvida e os experimentos possam ser realizados de forma automática.

1.2 JUSTIFICATIVA E RELEVÂNCIA DO TRABALHO

Genes e proteínas são importantes para compreensão da estrutura e funcionalidade genética dos organismos. A atribuição de nomes significativos e consistentes a eles é um requisito importante.

Wain *et al.* (2002) e Demerec *et al.* (1966) estabeleceram diretrizes de nomenclatura para nomear os genes recém-identificados, mas estas ainda são pouco aplicadas.

Atualmente há um aumento exponencial do sequenciamento de novos organismos e a tendência é aumentar ainda mais devido ao barateamento do processo e da importância de novas descobertas.

Considerando que:

1. Não há nenhum estudo de agrupamento de genes que incorpore todos os dados do Banco de Dados NR, que é uma base de dados importante e utilizada para pesquisa através do Blast;
2. Nenhum estudo analisa a presença de genes hipotéticos nos agrupamentos;
3. Nenhum estudo aponta divergências entre os nomes dos genes que podem indicar erros de anotação.

Baseado nos três itens descritos, considerando a importância do tema estudado e com o objetivo de colaborar com uma anotação automática mais consistente, é que a proposta da metodologia desenvolvida neste trabalho foi idealizada.

1.3 **ORGANIZAÇÃO DA DISSERTAÇÃO**

O presente trabalho está organizado em cinco seções da seguinte forma, seguindo as diretivas da Universidade Federal do Paraná:

Capítulo 1: introdução, objetivos e organização da dissertação;

Capítulo 2: levantamento do tema e revisão bibliográfica dos métodos para o desenvolvimento deste trabalho;

Capítulo 3: descrição da metodologia para desenvolvimento do trabalho;

Capítulo 4: resultados obtidos e discussões;

Capítulo 5: conclusões do estudo desenvolvido;

Capítulo 6: recomendações para trabalhos futuros.

2 REVISÃO DE LITERATURA

2.1 CONCEITOS BIOLÓGICOS

2.1.1 Nomenclatura de genes

Existem diversos comitês de nomenclatura de genes que têm a função de assegurar que cada gene tenha um nome e símbolo únicos que sejam usados de forma consistente na literatura científica.

Entre esses diversos comitês de nomenclatura podemos citar o CGSC (Coli Genetic Stock Center), HGNC (HUGO Gene Nomenclature Committee), FlyBase é o comitê de nomenclatura para genomas da *Drosophila* que mantém um banco de dados curado de genes e genomas (MCQUILTON, *et al.*, 2011) e o CGD (*Saccharomyces* Genome Database) é uma base de dados com informações genômicas e biológicas para os genes de *S. cerevisiae* mantido e atualizado por curadores (CHRISTIE, *et al.*, 2004).

O CGSC é um comitê de nomenclatura para os genes de *Eschirischia Coli*; um banco de dados é mantido por eles contendo informações sobre os nomes dos genes, sinônimos, propriedades, posição do mapa de genes, informações sobre o produto do gene, informações sobre mutações específicas e referências à literatura primária (BERLYN e LETOVSKY, 1992). Em 1966 DEMEREC e colaboradores sugeriram uma normatização para os genes bacterianos. A FIGURA 1 apresenta um esquema de como devem ser escritos estes nomes de acordo com a regra apresentada. O CGSC utiliza hoje (em fevereiro de 2012) esta normatização sugerida por DEMEREC *et al.* em 1966.

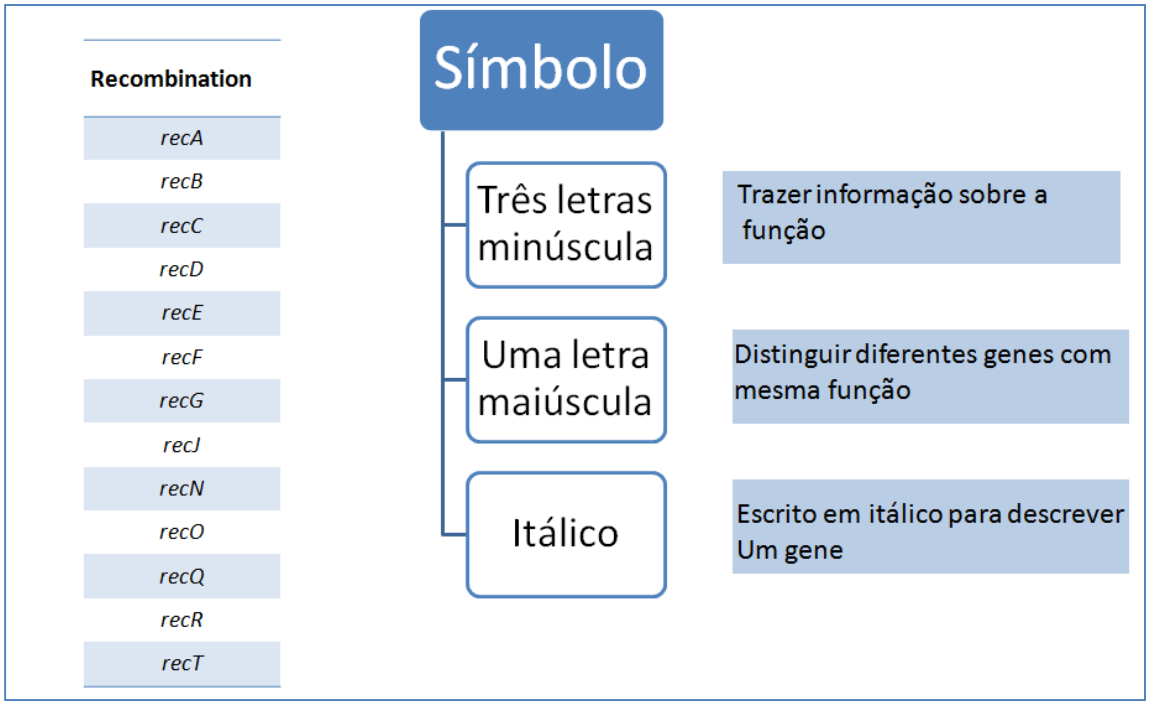


FIGURA 1 - COMPOSIÇÃO DA NOMENCLATURA DE GENE BACTERIANO SEGUNDO DEMEREC. FONTE: Adaptado de (DEMEREC, *et al*.., 1966)

O nome do gene deve ser composto de três letras minúsculas (que trazem informação sobre a função do gene), seguido de uma letra maiúscula (para distinguir diferentes genes com a mesma função) e deve ser escrito em itálico para definir que é um gene ou não itálico para definir que é uma proteína. Um bom exemplo da nomenclatura empregada corretamente, são os genes envolvidos na recombinação homóloga e reparo do DNA como, por exemplo, os genes *recA* e *recB*.

O HGNC que é um comitê de nomenclatura para genes humanos, atribui símbolos únicos para mais de 32.000 nomes de genes, dos quais mais de 19.000 são codificadores de proteínas. O genenames.org é um repositório *on-line* com curadoria de nomenclatura e contém links para informação genômica, proteômica e fenotípica, bem como páginas da família de cada gene (POVEY, *et al*., 2001). A FIGURA 2 mostra um esquema de como deve ser escrito o nome de um gene humano de acordo com o HGNC.

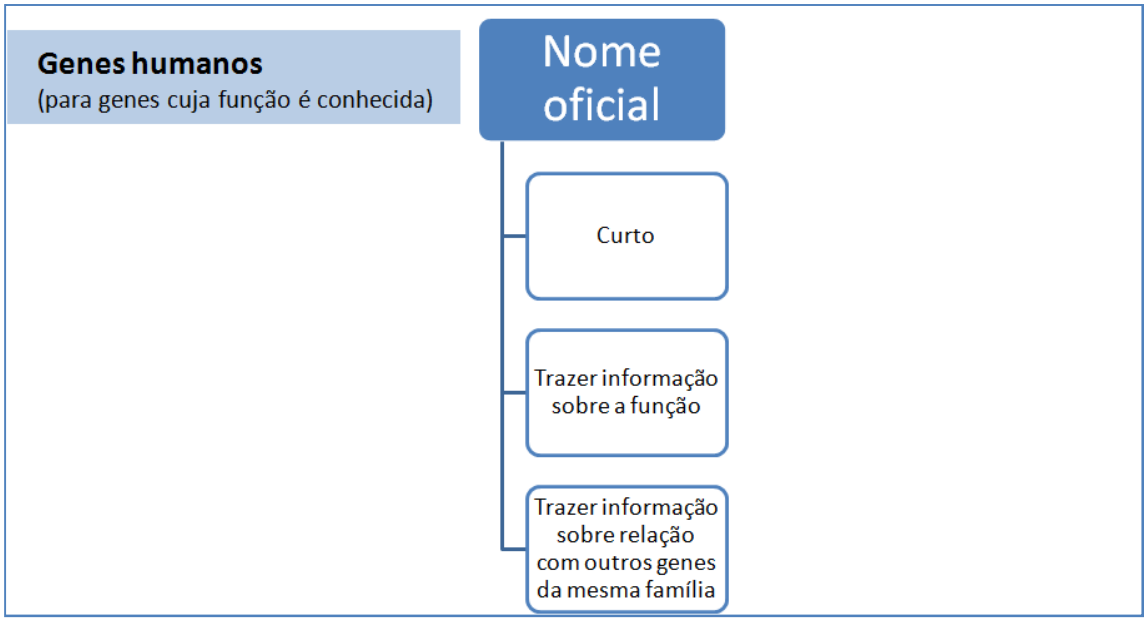


FIGURA 2 - COMPOSIÇÃO DA NOMENCLATURA PARA GENES HUMANOS SEGUNDO HGNC.
FONTE: Adaptado de (EYRE, *et al.*, 2006)

O nome do gene deve ser curto, deve trazer alguma informação sobre a função do gene e deve trazer alguma informação sobre a relação com outros genes da mesma família. Esta forma é bem semelhante à apresentada por DEMEREC e colaboradores.

2.1.2 Sinônimos de genes – Trabalhos correlatos

Uma maneira de identificar a diversidade de nomes que se referem ao mesmo gene é através da sinonímia. Além dos comitês de nomenclatura que asseguram um nome único e apontam seus sinônimos com acurácia há trabalhos que visam abranger um maior número de nomes de genes se dedicando apenas ao estudo dos sinônimos de genes.

Em 2004 PILLET e colaboradores construíram um BD de sinônimos denominado GPSDB[1] (Gene and Protein Synonyms DataBase). Este banco abrange 14 organismos modelo das seguintes bases de dados: LocusLink[2] e Swiss-Pro[3] para

[1] www.gpsdb.expasy.org

[2] www.ncbi.nlm.nih.gov/LocusLink

[3] www.expasy.org

multiespécies; GDB (PEARSON, 1991), HGNC[4] e OMIM[5] para genes humanos; MGD[6] para *Mouse*; RGD[7] e Ratmap[8] para *Rat*; Flybase[9] para Drosophila; SGD[10] para *Saccharomyces cerevisiae*; TAIR[11] para *Arabidopsis thaliana*; WormBase[12] para *Caenorhabditis elegans*; SubtiList (MOSZER, *et al.*, 2002) para *Bacillus subtilis*; e EcoGene[13] para *Escherichia coli*; para agrupar os sinônimos foram fundidas todas as entradas destes bancos de dados que eram relativas a um mesmo gene. Utilizaram o processo de normalização de nomes de genes e retirada de palavras irrelevantes através do uso de expressão regular, com foco em genes que codificavam alguma proteína e formaram assim, um banco de dados com 532.970 sinônimos que representam 319.386 proteínas.

Outro estudo desenvolvido foi o BioThesaurus (LIU, *et al.*, 2006), que é um sistema *web* que reúne informações de 13 fontes de dados distintas: UniProt, incluindo Swiss-Prot, TrEMBL e PIR-PSD; NCBI[14] incluindo Entrez Gene, RefSeq e GenPept; banco de dados de organismos modelo incluindo MGD, SGD, RGD, FlyBase e WormBase; outras bases de dados como o HGNC, EC enzyme nomenclature[15] e OMIM database of human genes and genetic disorders. O processo para a formação do banco de dados de sinônimos consiste em um filtro de nomes que foram considerados inconsistentes (por exemplo: *hypothetical protein*, *putative protein*, *novel protein*) por um curador. Uma normalização de nomes foi realizada a fim de agrupar nomes com variantes textuais, de pontuação ou variantes sintáticas. Seu banco de dados é composto por 7.144.420 de sinônimos que representam 2.869.972 de nomes de genes distintos.

---

[4] www.genenames.org

[5] www.ncbi.nlm.nih.gov/omim

[6] www.informatics.jax.org

[7] www.rgd.mcw.edu/wg/home

[8] www.ratmap.gen.gu.se

[9] www.flybase.org

[10] www.yeastgenome.org

[11] www.arabidopsis.org

[12] www.wormbase.org

[13] www.ecogene.org

[14] www.ncbi.nlm.nih.gov/

[15] www.chem.qmul.ac.uk/iubmb/enzyme

CD-HIT[16] é um software que gera agrupamentos levando em consideração os alinhamentos das sequências gênicas. Ele aponta uma sequência representativa do grupo, que é obtida através do alinhamento da maior sequência com a menor sequência contidas no mesmo grupo. Os dados de entrada do algoritmo devem estar no formato FASTA[17] e um valor de identidade deve ser informado, que irá servir como linha de corte para os agrupamentos (LI, JAROSZEWSKI e GODZIK, 2002) (LI, *et al.*, 2010).

2.1.3 Normalização dos nomes de genes

Um dos grandes obstáculos que impedem o uso efetivo de um dicionário de sinônimos é o problema de variação dos nomes. TSURUOKA e colaboradores (2007) descrevem alguns tipos de variações de termos que podem ocorrer nos nomes dos genes, como por exemplo, a ortografia e a morfologia (QUADRO 1).

TIPO DA VARIAÇÃO DO TERMO	EXEMPLO
Ortográficas	IL2 IL**-** 2
Morfológicas	GHF-1 **transcriptional** factor GHF-1 **transcription** factor
Romano/arábico	Synapsin **III** Synapsin **3**
Acrônimo/nome completo	**IL**-2 **interleukin**-2
Palavras extras	Zfp580 Zfp580 **protein**
Palavra entre parênteses	Ah receptor Ah **(dioxin)** receptor

QUADRO 1 - ALGUNS TIPOS DE VARIAÇÃO DE TERMOS QUE PODEM OCORRER NOS NOMES DOS GENES.

FONTE: Adaptado de (TSURUOKA, *et al.*, 2007)

[16] www.cd-hit.org/

[17] Formato onde uma sequência começa com uma descrição de uma única linha, seguida por linhas de dados de sequência gênica.

FANG e colaboradores (2006) sugerem que uma das maneiras de minimizar estes problemas é a normalização dos termos que minimiza as variações que podem ocorrer nos nomes dos genes, diminuindo a variabilidade dos termos e facilitando a construção de um dicionário de sinônimos (TSURUOKA, MCNAUGHT e ANANIADOU, 2008). O método de normalização proposto por eles consiste em retirar hífens e parênteses dos nomes dos genes, espaços em branco e remover todo tipo de pontuação.

2.1.4 Banco de dados e ferramentas de Bioinformática

Com o surgimento de sequenciadores automáticos além de grandes bases de dados públicas, foram necessárias ferramentas que auxiliassem as análises desse grande volume de dados (PROSDOCIMI, *et al.*, 2003).

A identificação de similaridade de sequências gênicas é uma tarefa das mais utilizadas em bioinformática (CAMACHO, *et al.*, 2009). A ferramenta Blast (ALTSCHUL, *et al.*, 1990) (ALTSCHUL e GISH, 1990) (JOHNSON, *et al.*, 2008) (YE, MCGINNIS e MADDEN, 2006) é utilizada para realizar alinhamentos entre sequências de nucleotídeos ou aminoácidos e fornece resultados estatísticos dos alinhamentos realizados.

Blast é uma das ferramentas mais utilizadas na área de bioinformática para busca de similaridade. Os alinhamentos são realizados entre uma sequência de consulta chamada de *query*, contra diversas sequências de um banco de dados que são denominadas sequências *subject*, fornecendo informações estatísticas sobre cada alinhamento realizado (CAMACHO, *et al.*, 2009).

Blast conta com diversos subprogramas, dentre eles o BlastP que compara uma sequência de aminoácido (*query*) contra um banco de sequências de proteínas (*subject*). A partir dos alinhamentos realizados podem ser identificadas semelhanças entre os genes.

Um dos bancos de dados mais utilizados para realizar alinhamentos de sequências é o NR (Non-Redundant Data Base), que é uma base de dados não-redundante de sequências protéicas, distribuído e mantido pelo NCBI (NCBI, 2011).

Ele incorpora sequências idênticas de diversos bancos de dados conforme mostra o QUADRO 2.

Banco de dados de origem	**Sintaxe do identificador**
GenBank	gb | accession | locus
EMBL Data Library	emb | accession | locus
DDBJ	dbj | accession | locus
NBRF PIR	pir || entry
Protein Research Foundation	prf || name
SWISS-PROT	sp | accession | entry name
Brookhaven Protein Data Bank	pdb | entry | chain
Patents	pat | country | number
GenInfo Backbone Id [18]	bbs | number
General database identifier [19]	gnl | database | identifier
NCBI Reference Sequence	ref | accession | locus
Local Sequence identifier	lcl | identifier

QUADRO 2 - SINTAXE PARA OS DIFERENTES BANCOS DE DADOS INCORPORADOS AO NR.
FONTE: http://www.ncbi.nlm.nih.gov

Neste quadro está disponível a sintaxe de cada identificador dos diferentes bancos de dados que são incorporados no NR.

A FIGURA 3 apesenta um trecho do arquivo NR disponibilizado pelo NCBI, onde é possível identificar as diferentes sintaxes que são agrupadas em um único número de GI [20] (*GeneInfo Identifiers*).

[18] Entradas antigas criadas manualmente a partir de *jornal-scan*.

[19] Geralmente utilizado para bancos de dados locais personalizados.

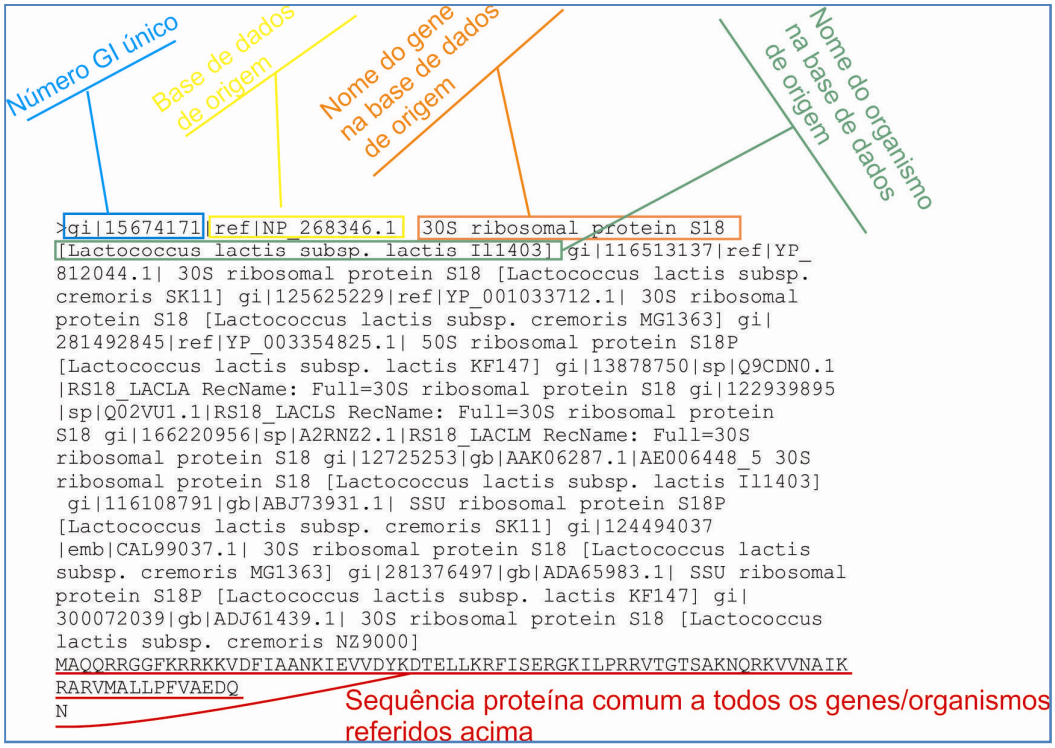


FIGURA 3 - TRECHO DO ARQUIVO NR REFERENTE AO GENE 30S RIBOSSOMAL PROTEIN S18, DIPONIBILIZADO PELO NCBI.

FONTE: Adaptado de (TAO, 2011)

Para que as sequências sejam incorporadas em um registro de GI único, elas devem ter o mesmo comprimento e cada aminoácido deve manter-se na mesma posição (TAO, 2011).

2.2 CONCEITOS COMPUTACIONAIS

2.2.1 Agrupamento de dados

Todo o processo de agrupamento de dados passa por diversas etapas que se inicia no pré-processamento dos dados até chegar à interpretação dos resultados obtidos com os agrupamentos (BATISTAKIS, HALKIDI e VAZIRGIANNIS, 2001). Os passos básicos para desenvolver um processo de agrupamento são apresentados na FIGURA 4.

[20] GI é uma numeração única que identifica a combinação de uma determinada sequência de aminoácido com o nome de um gene no NR.

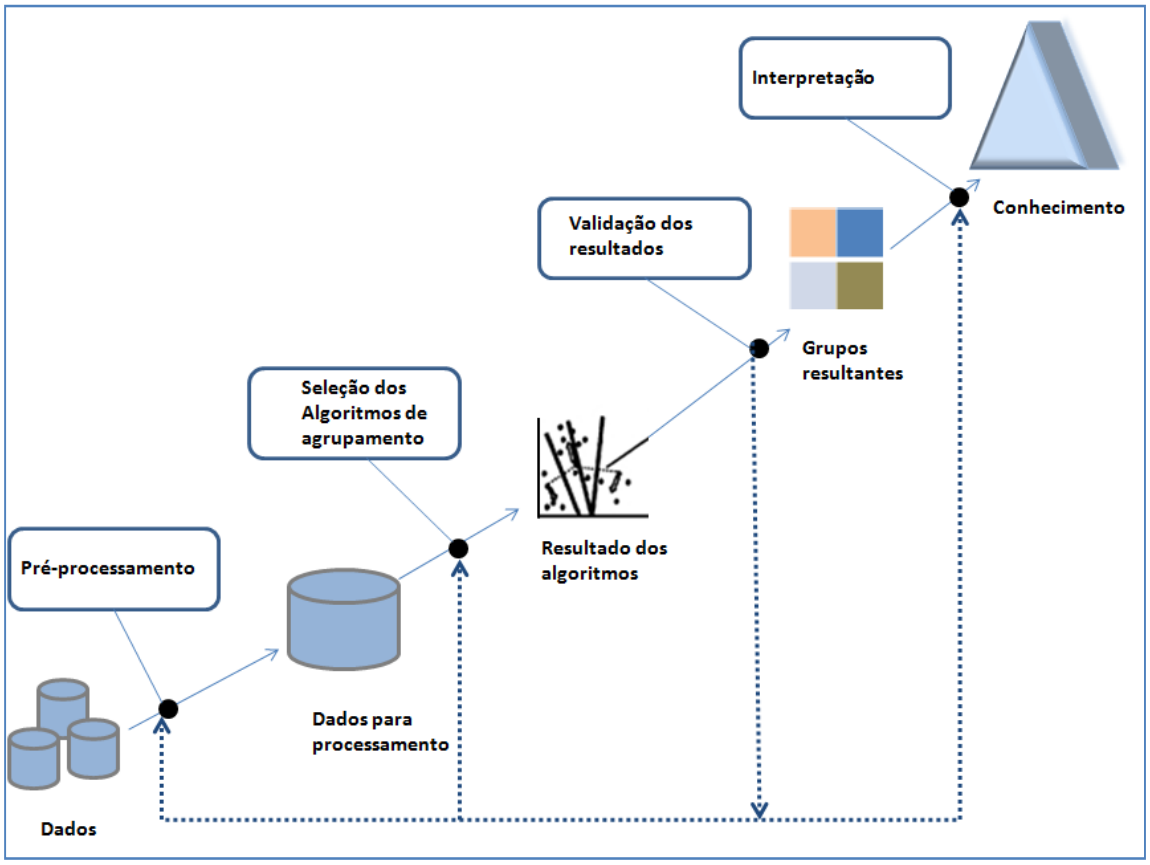


FIGURA 4 - ETAPAS DO PROCESSO DE AGRUPAMENTO DE DADOS.
FONTE: Adaptado de (BATISTAKIS, HALKIDI E VAZIRGIANNIS, 2001)

Segundo FAYYAD (1996) o processo de KDD (Knowledge-Discovery in Databases) em geral pode ser resumido da sequinte forma:

- **Pré-processamento**: O objetivo é selecionar adequadamente as características que descrevem um objeto e serão executadas de modo que codifiquem informações sobre uma tarefa de nosso interesse.
- **Seleção dos algoritmos de agrupamento**: Esta etapa refere-se à escolha de um algoritmo que resulte em um bom esquema de agrupamento.
- **Validação dos resultados**: os resultados dos algoritmos de agrupamento são verificados através de critérios e técnicas apropriadas. Podem-se utilizar três tipos de medidas para avaliar os resultados de um agrupamento: as internas, que visam avaliar os grupos entre si; as externas, que comparam o agrupamento obtido com um já conhecido; e as relativas, que comparam os

resultados de diversos algoritmos com o objetivo de selecionar aquele que obteve melhor desempenho (TAN, STEINBACH e KUMAR, 2006).

- **Interpretação**: existem casos em que os especialistas na área de aplicação adicionam outras evidências experimentais e análise aos resultados obtidos com os agrupamentos, a fim de chegar à conclusão correta.

2.2.2 Seleção de características

A extração de característica pode ser entendida como a extração de qualquer medida útil no processo de identificação de um padrão. Segundo SOUZA (1999) as características podem ser simbólicas, numéricas ou ambas, podendo ser variáveis contínuas ou discretas.

A extração de características é considerada uma importante etapa na fase do pré-processamento dos dados. Duas abordagens, segundo SEWEL (2007), podem ser utilizadas para a seleção de características:

- ***Forward Selection***: Inicia-se sem nenhuma variável e vai adicionando uma a uma e em cada etapa de adição, o erro é diminuído. Este processo deve ser repetido até que qualquer nova adição não diminua o erro de forma significativa.
- ***Backward Selection***: é o oposto do *Forward Selection*. Inicia-se o processo com todas as variáveis, devendo removê-las uma a uma. A cada remoção o erro é diminuído até ser estabilizado de forma que a qualquer nova remoção o erro aumente de forma significativa.

A literatura ainda apresenta diversas formas e discussões sobre a extração de característica:

- KIRA e RENDELL (1992) descreveram a seleção de características através de um algoritmo estatístico;
- Métodos flutuantes de pesquisa em seleção de características foram descritos por PUDIL, NOVOVICOVÁ e KITTLER (1994);
- Um método para seleção de subconjuntos de características baseado na Teoria da Informação foi o estudo de KOLLER e SAHAMI (1996);
- JOHN, KOHAVI e PFLEGER (1994) abordaram o problema das características irrelevantes na seleção das características.

2.2.3 Redes Neurais Artificiais (RNAs)

Redes Neurais Artificiais são sistemas computacionais que foram inspirados na estrutura, no método de processamento e na habilidade de aprendizado de um cérebro biológico (CYBENKO, 1996).

FAUSSET (1994) diz que uma rede neural artificial é um sistema de processamento de informações que têm características em comum com uma rede neural biológica, a FIGURA 5 representa um neurônio biológico segundo FAUSETT.

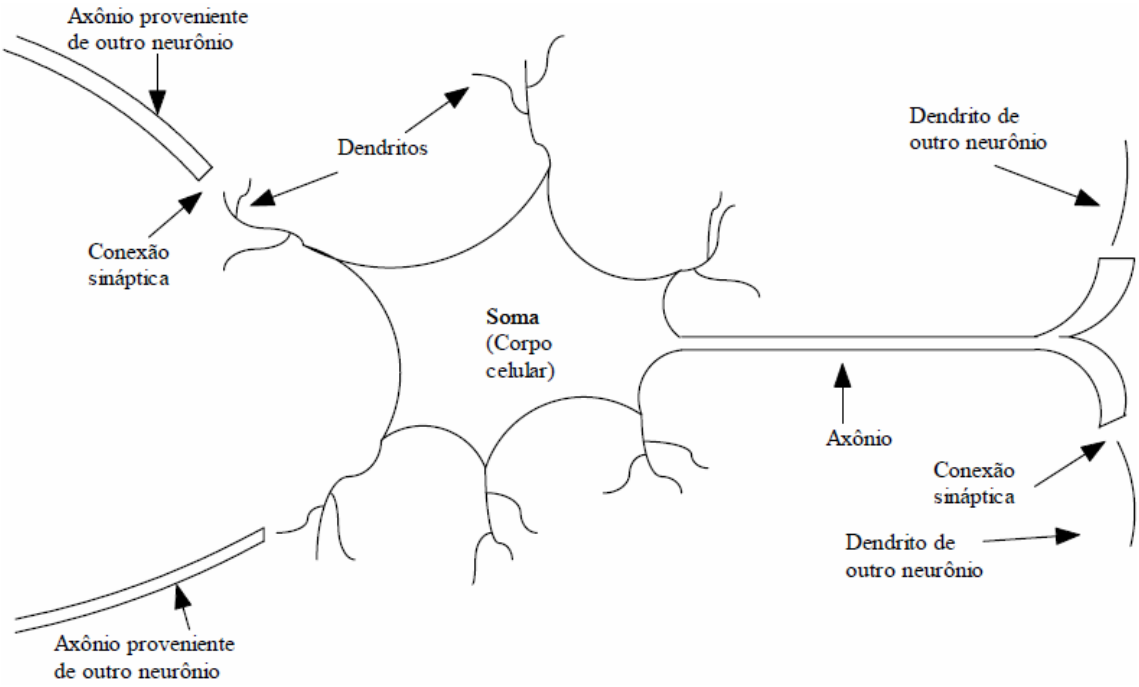


FIGURA 5 - O NEURÔNIO BIOLÓGICO.
FONTE: Adaptado de (FAUSETT, 1994)

O neurônio é uma unidade de processamento de informação fundamental para as operações em uma rede neural (HAYKIN, 1994). A FIGURA 6 mostra um modelo de neurônio artificial, baseado no neurônio biológico, que é à base das redes neurais artificiais.

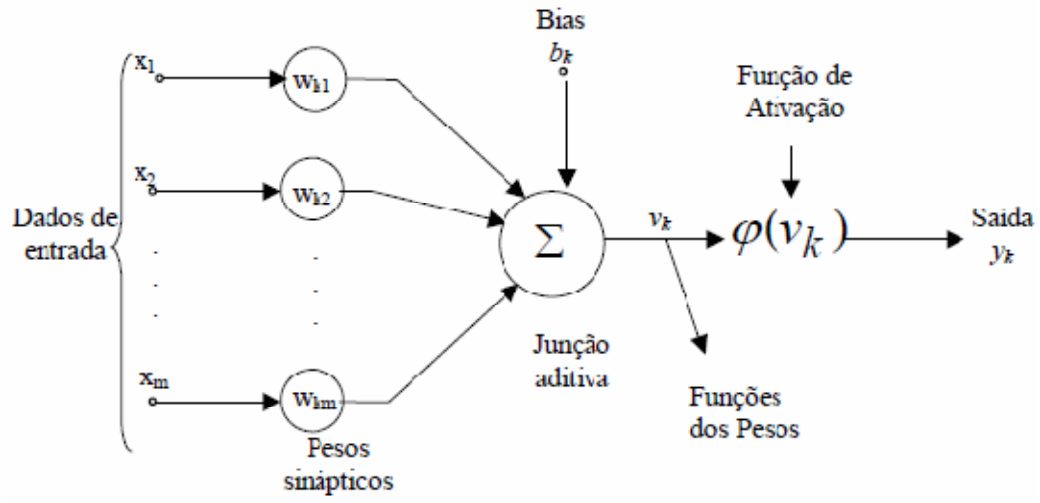


FIGURA 6 - MODELO DE UM NEURÔNIO ARTIFICIAL.
FONTE: Adaptado de (HAYKIN, 1994)

Do ponto de vista da neurociência o aprendizado ocorre através das alterações estruturais sinápticas entre os neurônios. Essas alterações podem ser realizadas de diversas formas, o que acaba gerando diversos tipos de aprendizagem (BASTOS, 2007).

Do ponto de vista da Inteligência Artificial, uma das características mais importante das redes neurais é a capacidade de aprendizado, onde a cada nova iteração, há uma melhora de desempenho no resultado final.

O QUADRO 3 apresenta formas de se agrupar as redes neurais de acordo com as características em função do paradigma (aprendizagem supervisionada, não supervisionada ou híbrida), regras de aprendizagem utilizadas, a arquitetura da rede, o algoritmo de aprendizagem e a tarefa para qual cada uma se propõem (NIEVOLA, 2004).

Este trabalho não discute de forma extensa cada uma delas, apenas apresenta uma visão geral dos pontos fundamentais de uma RNA.

Paradigma	Regra de Aprendizagem	Arquitetura	Algoritmo de Aprendizagem	Tarefa
Supervisionada	Correção do erro	Perceptron com uma camada	Algoritmos de aprendizagem do perceptron	Classificação de padrões
		Perceptron com várias camadas	Retro-propagação; Adaline e Madaline	Aproximação de funções, predição e controle
	Boltzmann	Recorrente	Algoritmo de aprendizagem de Boltzmann	Classificação de padrões
	Hebb	Multicamadas em avanço	Análise Discriminante linear	Análise de dados, classificação de padrões
	Competitiva	Competitiva	Quantização do vetor de aprendizagem	Categorização em classes internas, compressão de dados
		Rede ART	ARTMAP	Classificação de padrões, categorização em classes internas
Não supervisionado	Correção do erro	Multicamadas em avanço	Projeção de Sammon	Análise de dados
	Hebb	Em avanço ou competitiva	Análise da componente principal	Análise de dados, compressão de dados
		Rede Hopfield	Aprendizagem de memória associativa	Memória associativa
	Competitiva	Competitiva	Quantização de vetores	Categorização, compressão de dados
		SOM (Kohonen)	SOM (Kohonen)	Categorização, análise de dados

QUADRO 3 - OS ALGORITMOS DE APRENDIZAGEM MAIS CONHECIDOS. **(continua)**

Paradigma	Regra de Aprendizagem	Arquitetura	Algoritmo de Aprendizagem	Tarefa
Não Supervisionado	Competitiva	Rede ART	ART1, ART2	Categorização
Híbrido	Correção de erros e competitiva	Rede RBF	Algoritmo de aprendizagem RBF	Classificação de padrões, aproximação de funções, predição, controle

QUADRO 3 - OS ALGORITMOS DE APRENDIZAGEM MAIS CONHECIDOS.
FONTE: Adaptado de (JAIN, MAO e MOHIUDDIN, 1996)

Segundo HAYKIN (1994) define os paradigmas supervisionado, não supervisionado e híbrido da seguinte forma:

- No paradigma supervisionado é continuamente apresentada à rede conjuntos de padrões de entrada e seus respectivos padrões de saída. Durante esta etapa, a rede realiza uma adaptação dos pesos das conexões entre os dados de processamento, até que o erro entre os padrões de saída gerados pela rede atinja um valor mínimo satisfatório. É esperado que o sistema aprenda a correlacionar de forma correta uma entrada com sua classe correspondente;
- No paradigma não supervisionado não há a presença de um professor externo para supervisionar o processo de aprendizagem;
- No aprendizado híbrido, são utilizados de forma conjunta, o paradigma supervisionada e o não supervisionado, onde a primeira camada de conexões é treinada de forma não supervisionada e a segunda, de forma supervisionada.

2.2.4 O mapa auto-organizável de Kohonen

O objetivo principal de um SOM é mapear os dados de entrada de tamanho arbitrário em um mapa uni ou bidimensional, realizando esta transformação de forma

adaptativa e de uma maneira ordenada topologicamente, ou seja, que padrões de entrada próximos devem ativar unidades de saída próximas no mapa gerado.

Há também a possibilidade de um mapa ter mais de duas dimensionalidades, porém não são tão comuns (HAYKIN, 1994). A FIGURA 7 apresenta o modelo da rede de Kohonen que é de natureza não supervisionada e competitiva, onde os neurônios da camada de saída competem entre si para permanecerem ativos, mas apenas um neurônio fica ativo a cada vez.

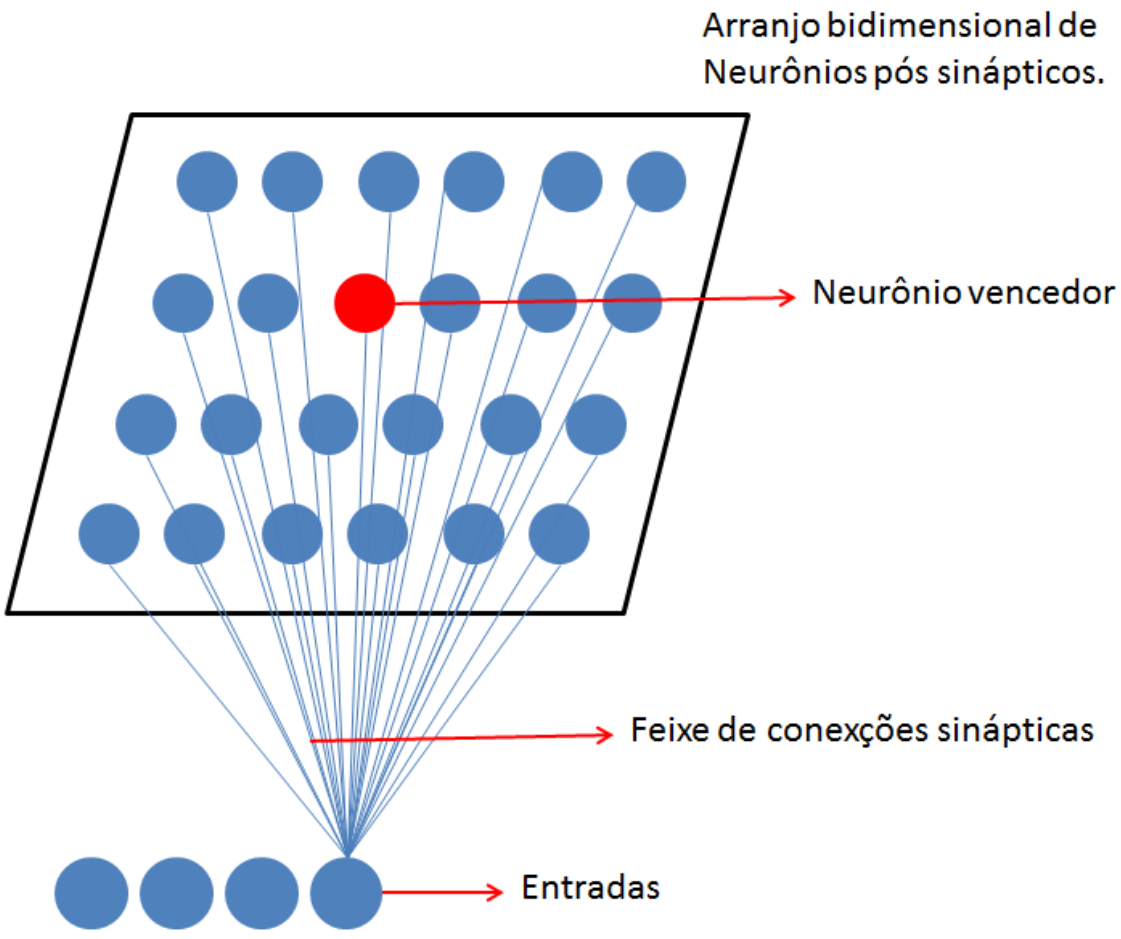


FIGURA 7 - MODELO DE KOHONEN
FONTE: Adaptado de (HAYKIN, 1994)

Este modelo de Kohonen apresenta um arranjo bidimensional de unidades de saída onde cada uma é conectada aos nós de entrada através de conexões

sinápticas, porém apenas um neurônio será ativado e considerado o vencedor. O neurônio vencedor é denominado BMU (*Best Matching Unit*).

Segundo HAYKIN (1994) há três processos essenciais para a formação de um mapa auto-organizável:

- Competição: os neurônios da grade calculam valores para cada padrão de entrada, de acordo com uma função discriminante, que fornece a base para a competição entre os neurônios. O neurônio que obtiver o maior valor da função é declarado o neurônio vencedor da competição.
- Cooperação: o neurônio vencedor determina a localização de uma vizinhança topológica de neurônios, fornecendo assim base para a cooperação entre os neurônios vizinhos.
- Adaptação sináptica: depois de determinado o neurônio vencedor e a vizinhança topológica ocorre o processo de adaptação dos pesos sinápticos, que consiste basicamente em fazer com que o neurônio vencedor e seus vizinhos se tornem mais especializados no reconhecimento do último padrão de entrada que foi apresentada a rede.

2.2.5 Interpretação do mapa produzido pelo SOM

Para interpretar o conteúdo de um mapa produzido pela rede SOM é necessário utilizar técnicas específicas de visualização que auxiliem nessa tarefa.

Existem diversos métodos de visualização, entre eles, a visualização do mapa como uma grade elástica (HAYKIN, 1994), o mapa contextual que utiliza arranjos bidimensionais e a Matriz-U (matriz de distância unificada). Este último foi utilizado neste trabalho e será detalhado a seguir.

Agrupamentos são úteis para identificar padrões nos dados (GUHA, RASTOGI e SHIM, 1998). A Matriz-U é uma forma de identificar agrupamentos no resultado gerado por um SOM.

A Matriz-U armazena as distâncias entre os neurônios adjacentes (ULTSCH, 1993), o que possibilita a geração de um mapa de duas dimensões que pode ter colorações distintas de acordo com os valores das distâncias entre os neurônios vizinhos do mapa. Deste modo, é possível visualizar regiões de proximidade que

representam os grupos que foram formados, e também regiões com maior distância entre os neurônios, que representem as fronteiras entre esses grupos.

A FIGURA 8 mostra à esquerda a Matriz-U, onde podemos visualizar de forma nítida três agrupamentos e à direita visualizamos o mapa com seus respectivos rótulos, que facilita a identificação dos dados contidos em cada grupo.

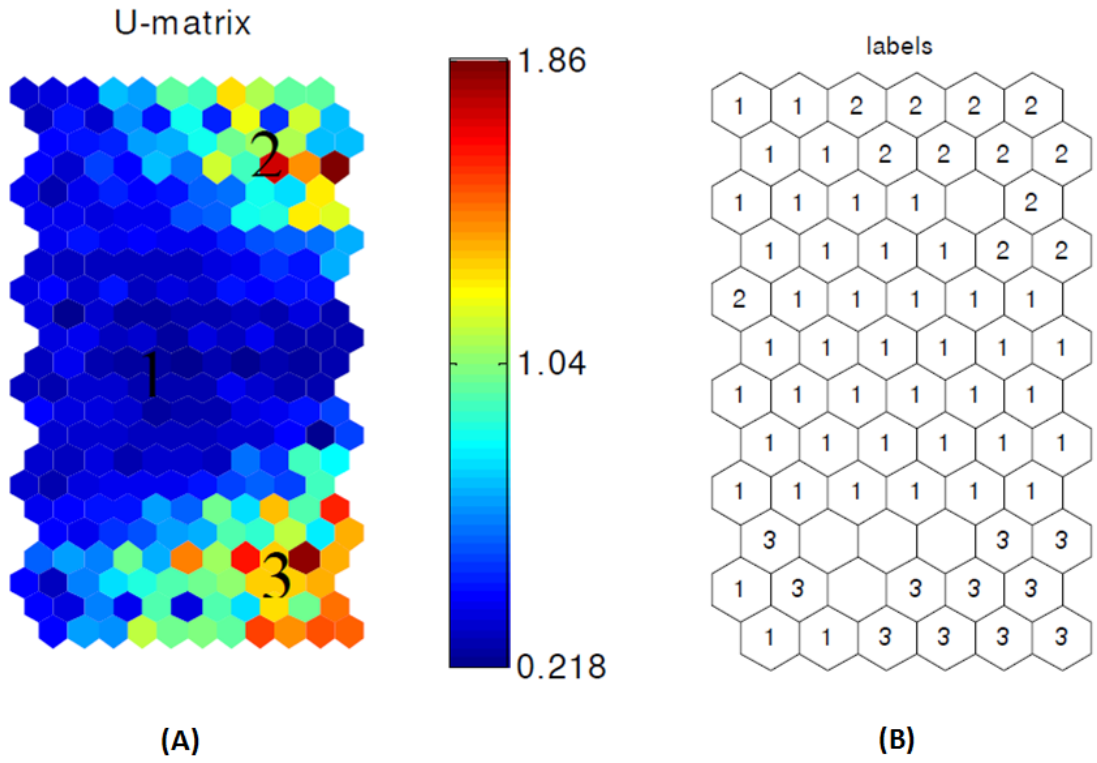


FIGURA 8 - EXEMPLO DE MATRIZ-U: (A) A MATRIZ-U, (B) MAPA COM RÓTULOS.
FONTE: (FARIA, *et al.*, 2010)

Outra forma de visualização que auxilia na análise do resultado é através dos *hits* de histograma (FIGURA 9), onde é possível identificar qual parte do mapa gerado apresentou maior quantidade de dados.

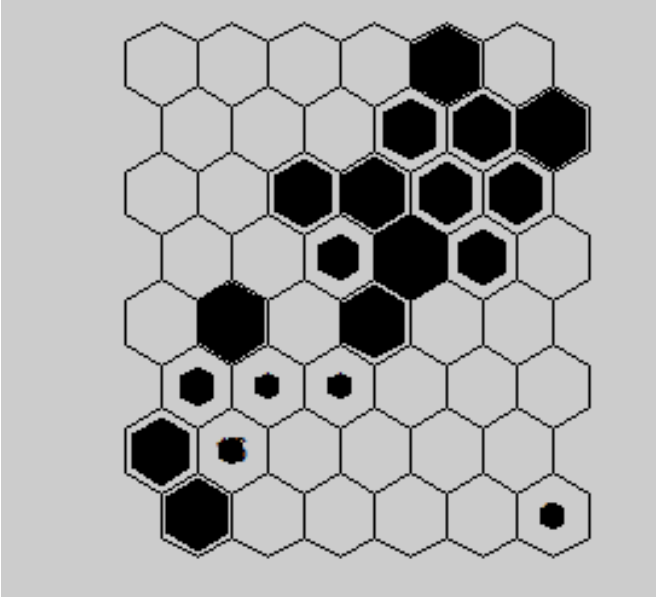


FIGURA 9 - *HITS* DE HISTOGRAMA
FONTE: A autora (2012)

Um BMU é um neurônio vencedor, escolhido através de uma medida de distância, que representa um determinado vetor de dados, deste modo a visualização dos hits de histograma apresenta quantos dados foram mapeados em cada unidade (neurônio) do mapa (cada unidade do mapa pode ter um ou vários dados).

2.2.6 Análise do mapa

KOHONEN (1995) diz que há um mapa ótimo para um dado conjunto de dados de entrada. Para avaliar a qualidade do mapa e do processo de aprendizagem duas métricas de erro podem ser utilizadas:

- O erro de quantização: é a média do erro que corresponde à diferença entre um vetor de entrada (X_j) e o vetor do neurônio vencedor ($p_{j(bmu)}$), avaliado para o mapa com os vetores de pesos treinados e dado pela seguinte equação (WU e TAKATSUKA, 2005): (VESANTO, *et al.*, 2000)

$$E_q = \frac{\sum_{j=1}^{n}(x_j - p_{j\,(bmu)})}{n} \quad (3.5.3.1)$$

É esperado que o melhor mapa tenha o menor valor de erro de quantização médio, pois este estaria mais bem ajustado aos vetores de entrada.

- O erro topográfico: visa avaliar quanto à estrutura da grade aproxima padrões próximos no espaço de entrada. Considerando que para cada padrão X_e temos o BMU como primeiro neurônio na ordem de competição da grade, sendo assim, o BMU2 corresponde ao segundo neurônio nessa escala. Deste modo, o erro topográfico corresponde ao percentual entre o neurônio vencedor (BMU) e o neurônio vizinho (BMU2) dado pela equação 3.5.3.2 (KOHONEN, 1990) (GUERRA, *et al.*, 2008):

$$E_t = \frac{\sum_{j=1}^{n} u(X_e)}{n} \quad (3.5.3.\ 2)$$

onde, $u(X_e)$ é igual a 1 se BMU e o BMU2 forem vizinhos, ou igual a 0 se BMU e BMU2 não forem vizinhos.

As métricas apresentadas são inversamente proporcionais ao número de neurônios, ou seja, os valores dos erros diminuem conforme há um aumento na quantidade de neurônios (VESANTO, SULKAVA e HOLLMEN, 2003).

Pölzlbauer, 2004 realizou uma discussão de outras medidas para verificação da qualidade do mapa de Kohonen como a medida de distorção, *trustworthiness* e produto topográfico. Este último pode ser utilizado para otimizar o tamanho do mapa para um determinado conjunto de dados, propriedade esta que não é trivial à maioria das medidas de qualidade.

Essas medidas discutidas por Pölzlbauer não estão disponíveis no pacote *SOM Toolbox*, e são de complexa implementação, e por este fato não serão utilizadas nesta dissertação.

3 MATERIAIS E MÉTODOS

Para o desenvolvimento deste trabalho foram utilizadas as abordagens: empírica e observacionais. Nas subseções 3.1, 3.2 e 3.3 serão apresentados os materiais utilizados para o desenvolvimento da pesquisa e nas subseções de 3.4 a 3.17 serão apresentadas todas as etapas da metodologia desenvolvida.

3.1 **Conjunto de dados**

O arquivo de dados do NR contendo nomes de genes e suas respectivas sequências de aminoácido foi extraído do sítio: **ftp://ftp.ncbi.nlm.nih.gov/blast/db/ em 03 de junho de 2011**. Com base neste arquivo, foi possível desenvolver um modelo de dados e implementar um Banco de Dados Relacional.

No QUADRO 4 estão relacionadas as dez sequências de aminoácidos com seus respectivos identificadores e nomes de genes que foram utilizadas para realizar os alinhamentos por meio da ferramenta BlastP. Estas sequências foram selecionadas de maneira aleatória.

Nome do Gene	Sequência de aminoácido	Identificador GI
Argininosuccinate lyase	MTENNEHLALWGGRFTSGPSPELARLSKSTQFDWRLADDDIAGSRAHARALGRAGLLTADELQR MEDALDTLQRHVDDGSFAPIEDDEDEATALERGLIDIAGDELGGKLRAGRSRNDQIACLIRMWLR RHSRVIAGLLLDLVNALIEQSEKAGRTVMPGRTHMQHAQPVLLAHQLMAHAWPLIRDVQRLIDWD KRINASPYGSGALAGNTLGLDPEAVARELGFSRVTDNSIDGTAARDLVAEFAFVAAMTGVDISRLS EEIIIWNTQEFAFVKLDDGYSTGSSIMPQKKNPDIAELARGKSGRLIGDLTGLLATLKGLPTAYARD LQEDKEAVFDQVDTLEVLLPAFTGMVRTMHFDGDRLEEEAPTGFALATDIAEWLVKNGVPFRHA HELSGACVKLAEGRGQELWDLTDNDFIETFAAFLPADKAPGVREVLSSHGSVDSRNGKGGTAYG RVREQIADAKAEVEELKLFPASTSDGSAYKAPGTF	23335287
ABC transporter ATP-binding protein	MAYTTFSQTKNDQLKEPMFFGQPVNVARYDQQKYDIFEKLIEKQLSFFWRPEEVDVSRDRIDYQA LPEHEKHIFISNLKYQTLLDSIQGRSPNVALLPLISIPELETWVETWAFSETIHSRSYTHIIRNIVNDPS VVFDDIVTNEQIQKRAEGISSYYDELIEMTSYWHLLGEGTHTVNGKTVTVSLRELKKKLYLCLMSV NALEAIRFYVSFACSFAFAERELMEGNAKIIRLIARDEALHLTGTQHMLNLLRSGADDPEMAEIAEE CKQECYDLFVQAAQQEKDWADYLFRDGSMIGLNKDILCQYVEYITNIRMQAVGLDLPFQTRSNPI PWINTWLVSDNVQVAPQEVEVSSYLVGQIDSEVDTDDLSNFQL	15802782
Resolvase	MTGQRIGYIRVSTFDQNPERQLEGVKVDRAFSDKASGKDVKRPQLEALISFARTGDTVVVHSMD RLARNLDDLRRIVQTLTQRGVHIEFVKEHLSFTGEDSPMANLMLSVMGAFAEFERALIRERQREGI ALAKQRGAYRGRKKSLSSERIAELRQRVEAGEQKTKLAREFGISRETLYQYLRTDQ	9507569
DNA-binding response regulator PhoB S4	MSRRILVVEDEAPIREMLCFVLEQKGYQAVEAEDYDSAMSKLAEPFPDLVLLDWMLPGGSGINLIK HMKREEMTRNIPVVMLTARGEEEDKVRGLEVGADDYITKPFSPKELVARLKAVIRRVTPTALEDVI DVQGLKLDPVSHRVTANDQPLDMGPTEFKMLHFFMTHQERVYSREQLLNNVWGTNVYVEDRTV DVHIRRLRKALEDAGHDKLIQTVRGAGYRFSTKA	15640738
Ribonucleotide-diphosphate reductase subunit beta	METLLSFEKVYKDYPSGPSIIHALKETNFEAKKGELIAIVGPSGSGKSTLLSLAGALLTPTGGTISIN GKSVGNLSSKEQTALRLEEIGFIFQAAHLVPYLHVKDQISFIGKMAGKSAAELEKDTASLLSQLGIS DRANFYPKDLSGGQKQRVAIARALINQPSVILADEPTASLDTERSREVVELIRNEVVQTSRTAIMVT HDER MLDLVNHVYRMEDGILTQES	16804410
fructose- specific phosphotransferase system protein FrvX	MNIELLQQLCEASAVSGDEQEVRDILINTLEPCVNEITFDGLGSFVARKGNKGPKVAVVGHMDEV GFMVTHIDESGFLRFTTIGGWWNQSMLNHRVTIRTHKGVKIPGVIGSVAPHALTEKQKQQPLSFD EMFIDIGANSREEVEKRGVEIGNFISPEANFACWGEDKVVGKALDNRIGCAMMAELLQTVNNPEIT LYGVGSVEEEVGLRGAQTSAEHIKPDVVIVLDTAVAGDVPGIDNIKYPLKLGQGPGLMLFDKRYFP NQKLVAALKSCAAHNDLPLQFSTMKTGATDGGRYNVMGGGRPVVALC LPTRYLHANSGMISKADYEALLTLIRGFLTTLTAEKVNAFSQFRQVD	16131738

QUADRO 4 - AS DEZ SEQUÊNCIAS DE AMINOÁCIDOS UTILIZADAS NESTE TRABALHO PARA GERAR DADOS PARA OS DEZ EXPERIMENTOS REALIZADOS. **(continua)**

Nome do Gene	Sequência de aminoácido	Identificador GI
pyridoxamine-phosphate oxidase protein	KLRFKDIGFILQASNLIPFLTVKQQLELVDKLMKNENNQLQESLFEDLGITHLKNKLPRDLSGGERQ RLAIARALYNDPAIVLADEPTASLDSEKAYEVVELLTKECKEKQKTVIMVTHDRRMIESCDKIFEIRD GVLKQQ	333905884
2-component transcriptional regulator	MSHKPAHLLLVDDDPGLLKLLGLRLTSEGYSVVTAESGAEGLRVLNREKVDLVISDLRMDEMDGM QLFAEIQKVQPGMPVIILTAHGSIPDAVAATQQGVFSFLTKPVDKDALYQAIDDALEQSAPATDER WREAIVTRSPLMLRLLEQARLVAQSDVSVLINGQSGTGKEIFAQAIHNASPRNSKPFIAINCGALPE QLLESELFGHARGAFTGAVSNREGLFQAAEGGTLFLDEIGDMPAPLQVKLLRVLQERKVRPLGSN RDIDINVRIISATHRDLPKAMARGEFREDLYYRLNVVSLKIPALAERTEDIPLLANHLLRQAAERHKP FVRAFSTDAMKRLMTASWPGNVRQLVNVIEQCVALTSSPVISDALVEQALEGENTALPTFVEARN QFELNYLRKLLQITKG NVTHAARMAGRNRTEFYKLLSRHELDANDFKE	15803079
Ferritin-like protein	MAYTTFSQTKNDQLKEPMFFGQPVNVARYDQQKYDIFEKLIEKQLSFFWRPEEVDVSRDRIDYQA LPEHEKHIFISNLKYQTLLDSIQGRSPNVALLPLISIPELETWVETWAFSETIHSRSYTHIIRNIVNDPS VVFDDIVTNEQIQKRAEGISSYYDELIEMTSYWHLLGEGTHTVNGKTVTVSLRELKKKLYLCLMSV NALEAIRFYVSFACSFAFAERELMEGNAKIIRLIARDEALHLTGTQHMLNLLRSGADDPEMAEIAEE CKQECYDLFVQAAQQEKDWADYLFRDGSMIGLNKDILCQYVEYITNIRMQAVGLDLPFQTRSNPI PWINTWLVSDNVQVAPQEVEVSSYLVGQIDSEVDTDDLSNFQL	15802782
response regulator receiver modulated metal dependent phosphohydrolase	MESMLDRPEQELVLVVDDTPDNLLLMRELLEEQYRVRTAGSGPAGLRAAVEEPRPDLILLDVNMP GMDGYEVCRRLKADPLTRDIPLMFLTARADRDDEQQGLALGAVDYLGKPVSPPIVLARVRTHLQL KANADFLRDKSEYLELEVRRRTRQLQQLQDAVIEALATLGDLRDNPRSRHLPRIERYVRLLAEHLA AQRAFADELTPEAVDLLSKSALLHDIGKVAVPDRVLLNPGQLDAADTALLQGHTRAGRDALASAE RRLGQPSGFLRFARQIAYSHHERWDGRGFPEGLAGERIPLAARIVALADRYDELTSRHAYRPPLA HAEAVLLIQAGAGSEFDPRLVEAFVAVADAFAEVARRYADSAEALDVEMQRLEQAVAESIELTAPP A	15599975

QUADRO 4 - AS DEZ SEQUÊNCIAS DE AMINOÁCIDOS UTILIZADAS NESTE TRABALHO PARA GERAR DADOS PARA OS DEZ EXPERIMENTOS REALIZADOS.

FONTE: A autora (2012)

3.2 **Softwares e linguagens de programação utilizadas**

- Matlab: é um *software* para o desenvolvimento de aplicativos de caráter técnico (MATHWORKS, 2008). Foi utilizado a versão 11.0 do software para execução dos scripts do Som *Toolbox* .
- Netbeans: é um ambiente de desenvolvimento (NETBEANS, 2011) para a escrita de *scripts*, a versão utilizada foi a 6.9.
- Excel: foi utilizado para auxiliar a análise dos dados dos agrupamentos, a versão utilizada foi Microsoft Office Excel 2007 (MICROSOFT., 2011)
- Blast: último acesso a ferramenta *on line* em 14/12/2011. Ferramenta utilizada para realizar alinhamentos. Foram utilizados os parâmetros padrões do programa:
    - Search Set Database: Non-redundant protein sequences (nr)
    - Program Selection Algorithm: blastp (protein BLAST)
    - Max target sequences: 100
    - Expect threshold: 10
    - Matrix: BLOSUM62
    - Gap Costs: Existence 11 Extension 1
    - Compositional adjustments: Conditional compositional score matrix adjustement
- PostgreSQL: é de Sistema Gerenciador de Banco de Dados *open-source* (POSTGRESQL, 2011) . Para este trabalho foi utilizada a versão 8.0.0.
- PgAdmin: *software* utilizado para escrever consultas SQL, apresenta uma interface gráfica para facilitar a administração do banco de dados (pgAdmin PostgreSQL tools, 2009). Foi utilizada a versão 1.14.1 para escrever consultas SQL.
- DbDesigner 4: é um sistema para modelagem de banco de dados relacional (ZINNER, 2003). Foi utilizado para modelar o banco de dados desenvolvido.
- Hibernate: versão utilizada foi a 3.6.9, empregado para mapear os atributos entre uma base de dados relacional e o modelo objeto de uma aplicação (HIBERNATE, 2011).
- PFam: é um banco de dados de famílias de proteínas, cada uma representada por alinhamentos de sequência múltipla e HMMs (modelos

ocultos de Markov) (SONNHAMMER, EDDY e DURBIN, 1997) (FINN, *et al.*, 2010). Também gera agrupamentos de famílias aparentadas, denominado como clãs. Um clã é uma coleção de PFam-A[21] relacionados por estrutura, sequência ou por modelos ocultos de Markov (PFAM, 2011). Neste trabalho foi utilizada a versão 26.0 para identificação das famílias dos genes.

- SQL: é uma das mais populares linguagens relacionais (SIMKOVICS, 1998). Foi utilizada para inserir dados no banco de dados bem como realizar consultas.
- JAVA: é uma linguagem de programação orientada a objetos (PROCESS, 2011), foi utilizada para escrever todos os scripts dos processos que antecedem a Rede SOM.

3.3 **Parâmetros do SOM**

A seguir são apresentados os parâmetros do SOM que foram utilizados neste trabalho:

- **Tamanho do mapa**: é calculado automaticamente na rotina de inicialização pelo Som Toolbox utilizando a seguinte heurística:
    - cinco vezes a raiz quadrada do tamanho da amostra - ou seja – no caso dos experimentos realizados neste estudo temos uma amostra igual a 100 que implicaria no seguinte cálculo: $5\ X\ \sqrt{100}$ = 50;
- **Dimensão do mapa**: é também utilizada uma heurística de modo que a razão entre os comprimentos dos dois lados do mapa seja calculada como a raiz quadrada da razão dos dois maiores autovalores do conjunto de dados. Os comprimentos dos lados são, então, ajustados de modo que seu produto seja próximo do número desejado de neurônios do mapa (VESANTO, *et al.*, 2000). Esta rotina do SOM *Toolbox* determinou um tamanho de mapa de dimensão 6×8 e, portanto, com 48 neurônios;
- **Algoritmo de treinamento**: em lote com duas fases (fase de ordenação e fase de ajuste fino ou convergência);

[21] PFam-A são entradas (sequências) de alta qualidade, de famílias manualmente curada.

- **Inicialização**: rotina Lininit (som lininit (base de dados)) que inicializa o mapa de forma linear;
- **Topologia**: em forma hexagonal;
- **Formato de vizinhança**: dado pela função gaussiana;
- **Taxa de aprendizagem**: o algoritmo em lote não a utiliza;
- **Número de Épocas**: 200;

3.4 **Fluxograma da Metodologia desenvolvida**

A metodologia desenvolvida para a execução deste trabalho está representada em um fluxograma na FIGURA 10, e cada um dos processos realizados será detalhado a seguir.

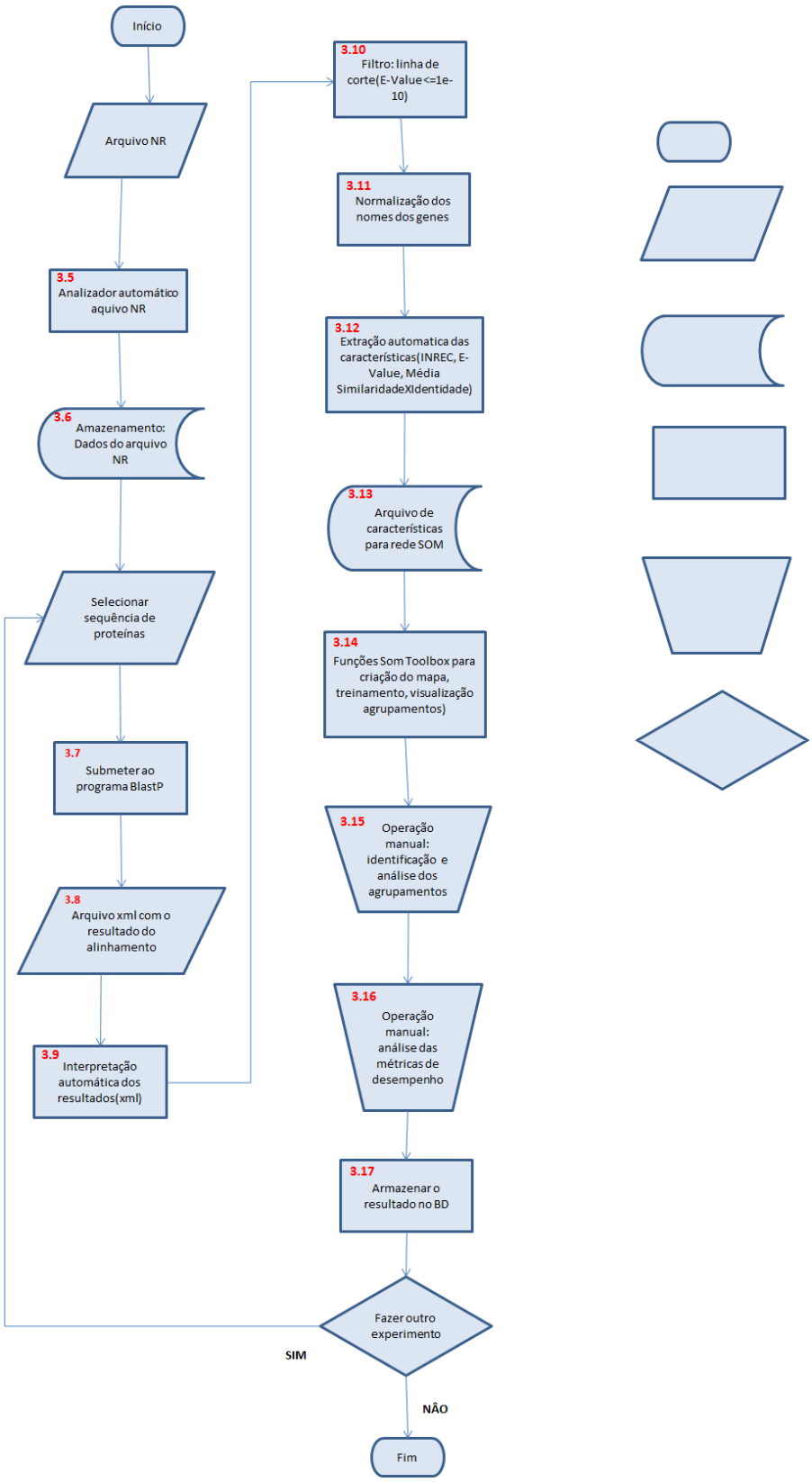


FIGURA 10 - FLUXOGRAMA DA METODOLOGIA DESENVOLVIDAPARA ESTE ESTUDO.

FONTE: A autora (2012)

3.5 **Analisador de arquivo NR**

Um analisador é um tipo de programa que tem o objetivo de quebrar um texto de entrada em partes menores e classificar essas partes segundo a gramática de uma linguagem (DELAMARO, 2004). Ainda segundo Delamaro, um analisador pode ser dividido em análise léxica, análise sintática e análise semântica. A análise léxica (AL) separa e identifica os símbolos do programa, a análise sintática (AS) verifica os símbolos contidos no programa e a ordem em que são encontrados, a análise semântica (ASem) verifica se os aspectos semânticos do programa estão corretos, ou seja, verifica se não existe incoerências em relação ao significado das construções feitas pelo programador.

Baseado nos fatores descritos anteriormente foi desenvolvido um *script* na linguagem de programação Java que executa a análise do arquivo de dados do NR. O *script* está descrito no Apêndice A.

A formatação do banco de dados NR é composta por um número de GI único que identifica a combinação de uma determinada sequência de aminoácido com o nome de um gene. (TAO, 2011).

3.6 **Banco de dados baseado no NR**

Foi realizada a implementação do modelo físico no sistema gerenciador de banco de dados PostgreSQL.

Foram utilizados *scripts* em SQL para criação das tabelas e seus relacionamentos. SQL é uma linguagem de pesquisa para banco de dados relacional. Diversas características dessa linguagem foram inspiradas na álgebra relacional (DATE e DARWEN, 1997)

Para armazenar no banco de dados as informações extraídas através do analisador de arquivo, foram desenvolvidos *scripts* em Java utilizando a Java Persistence API (PROCESS, 2011) que é um conjunto de interfaces que automatiza o processo de persistência de objetos em um banco de dados relacional. A biblioteca utilizada foi o Hibernate, que é uma coleção de projetos que permite ao

desenvolvedor utilizar modelos do tipo POJO [22] (*Plain Old Java Object*) e assim realizar o Mapeamento Objeto/Relacional (HIBERNATE, 2011). Esta ferramenta foi escolhida por agregar produtividade ao projeto.

3.7 **BlastP**

O subprograma BlastP foi utilizado para realizar os alinhamentos das sequências de aminoácidos utilizadas nos experimentos.

Cada uma das dez sequências selecionadas anteriormente serviram de entrada (*query*) para o programa BlastP, gerando assim, dez conjuntos de dados distintos contendo 100 genes cada um.

3.8 **Arquivo XML**

Depois de concretizado cada alinhamento, o programa Blast disponibiliza para *download* um arquivo em formato XML que contém informações estatísticas expressas através dos alinhamentos. Sua finalidade principal é a facilitar o compartilhamento das informações através da *internet*.

Para cada experimento foi feito o *download* do arquivo XML correspondente ao alinhamento realizado e salvo em uma pasta temporária no computador para ser interpretado posteriormente.

3.9 **Interpretação automática do arquivo XML**

Foi criado um script em Java para interpretar automaticamente o arquivo XML e extrair as informações de interesse dos seguintes campos do arquivo:

- Hsp evalue: Campo que contém a informação do valor de E-Value de cada alinhamento;

[22] São objetos Java que seguem uma estrutura simplificada, porém rígida, onde a classe deve obrigatóriamente ser composta por: um construtor *default* sem argumentos e métodos que seguem o padrão de *getters* e *setters* para seus atributos.

- Hsp qseq: Campo que contém a sequência *query*;
- Hsp hseq: Campo que contém a sequência *subject;*
- Hsp hit-to: Campo que contém o tamanho da sequência *query;*
- Hsp positive: Campo que contém o valor da similaridade entre os alinhamentos;
- Hsp identity: Campo que contém o valor de identidade entre os alinhamentos.

O formato do arquivo é organizado de forma hierárquica, como mostra a FIGURA 11, o que facilita a identificação dos campos que são de interesse.

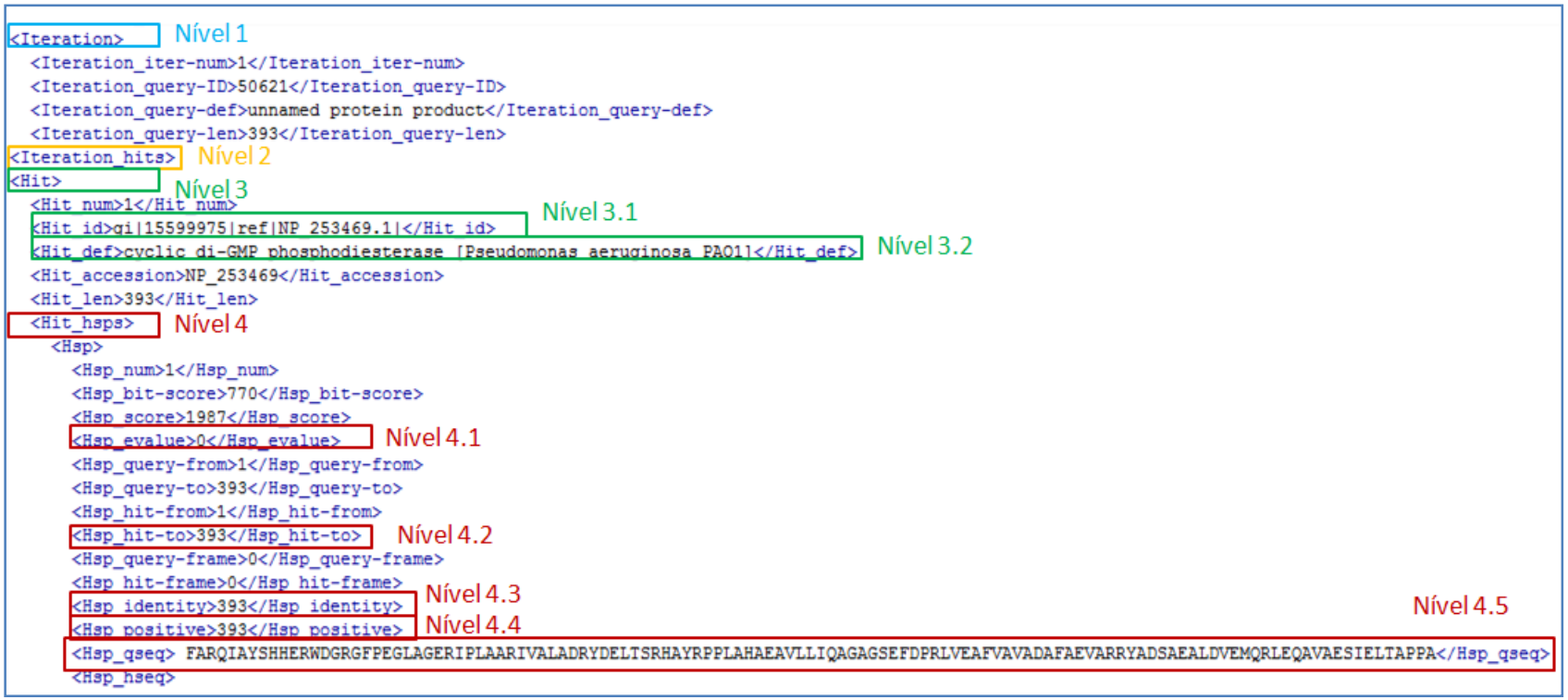


FIGURA 11 - TRECHO DE UM ARQUIVO XML.
FONTE: A autora (2012).

A figura mostra os níveis hierárquicos que foram utilizados neste trabalho até chegar aos campos de interesse. As informações dos níveis 3.1 e 3.2 foram utilizadas para rotulagem na Matriz-U e posteriormente esses dados foram armazenados no banco de dados. Os níveis 4.1 a 4.5 foram utilizados na fase de seleção de características.

3.10 **Linha de corte**

Após a extração das informações de interesse do arquivo XML, foi aplicada uma linha de corte de E-Value > $1e^{-10}$ nesses resultados. Este filtro foi necessário para que baixos índices de alinhamento (ou alinhamentos de pouca expressão) não interferissem no resultado final do agrupamento.

Este valor de corte do E-Value foi utilizado em diversos trabalhos de Bioinformática e Biologia Molecular (PAVY, *et al.*, 2005) (FRECH e CHEN, 2010) (YI e JUNG, 2011) (BELDA-FERRE, *et al.*, 2011).

No caso dos experimentos realizados neste estudo nenhum alinhamento ficou abaixo da linha de corte, mas essa medida pode ser útil em experimentos futuros realizados com um número maior de genes.

3.11 **Normalização dos nomes dos genes**

Foi utilizada a combinação de diversas técnicas de normalização de nomes de genes para minimizar a ambiguidade e pequenas variações de nomes que podem ocorrer.

Este processo foi realizado como um pré-processamento para calcular o valor do INREC (Índice Recursivo) do nome do gene. O INREC é utilizado para redução de dimensionalidade.

As técnicas utilizadas neste trabalho e um exemplo de como aplicá-las estão descritos no QUADRO 5.

TÉCNICA	EXEMPLO
Converter a letras maiúsculas em minúscula (COHEN, *et al.*, 2002)	**I**l2 → **i**l2
Retirar hífen (BRUIJN e MARTIN, 2003) (FANG, *et al.*, 2006)	il**-**2 → il2
Retirar espaços em branco extras (início, meio, fim) (FANG, *et al.*, 2006)	il 2 → il2
Remoção de parênteses (FANG, *et al.*, 2006)	ll **(**2**)** → il2
Converter Romano em Arábico (BRUIJN e MARTIN, 2003)	il**II** → il**2**

QUADRO 5 - TÉCNICAS UTILIZADAS PARA NORMALIZAÇÃO DOS NOMES DOS GENES E SEUS RESPECTIVOS EXEMPLOS DE UTILIZAÇÃO.

FONTE: A autora (2012)

3.12 **Extração das características**

Após a obtenção dos dados de interesse resultantes do alinhamento e da realização da normalização dos nomes dos genes, foram extraídas as características que serviram de base para o processamento da rede SOM.

As características contempladas nessa pesquisa estão descritas a seguir:

1. **INREC**: foi utilizado o INREC para redução de dimensionalidade do nome do gene em um valor numérico único que o representasse (SOUZA, 1999). Para cada letra do nome do gene foi atribuído um valor numérico de acordo com o padrão ASCII. O processo de atribuição dos valores ASCII para os nomes dos genes e o cálculo recursivo do INREC foram obtidos automaticamente através dos *scripts* desenvolvidos;
2. **Tamanho sequência**: a quantidade de aminoácidos presentes na sequência *subject* foi utilizada como característica.
3. **Similaridade X Identidade**: Para que os valores percentuais de Similaridade e Identidade ficassem com uma melhor padronização, foi realizada uma normalização desses valores pelo tamanho da sequência original de consulta (*query*) (COUTINHO, 2011), através de um script escrito em Java. A média simples entre os percentuais normalizados de Identidade e Similaridade foi utilizada como característica.

4. **E-Value**: Por indicar a probabilidade que o alinhamento entre duas sequências tenha acontecido ao acaso, (BEDELL, KORF e YANDELL, 2003), esta medida foi utilizada como característica.

3.13 **Arquivo de características**

Para submeter um arquivo contendo as características à rede SOM, foi gerado automaticamente um arquivo tabulado conforme especificação do Som *Toolbox* (VESANTO, *et al.*, 2000) e está esquematizado na FIGURA 12.

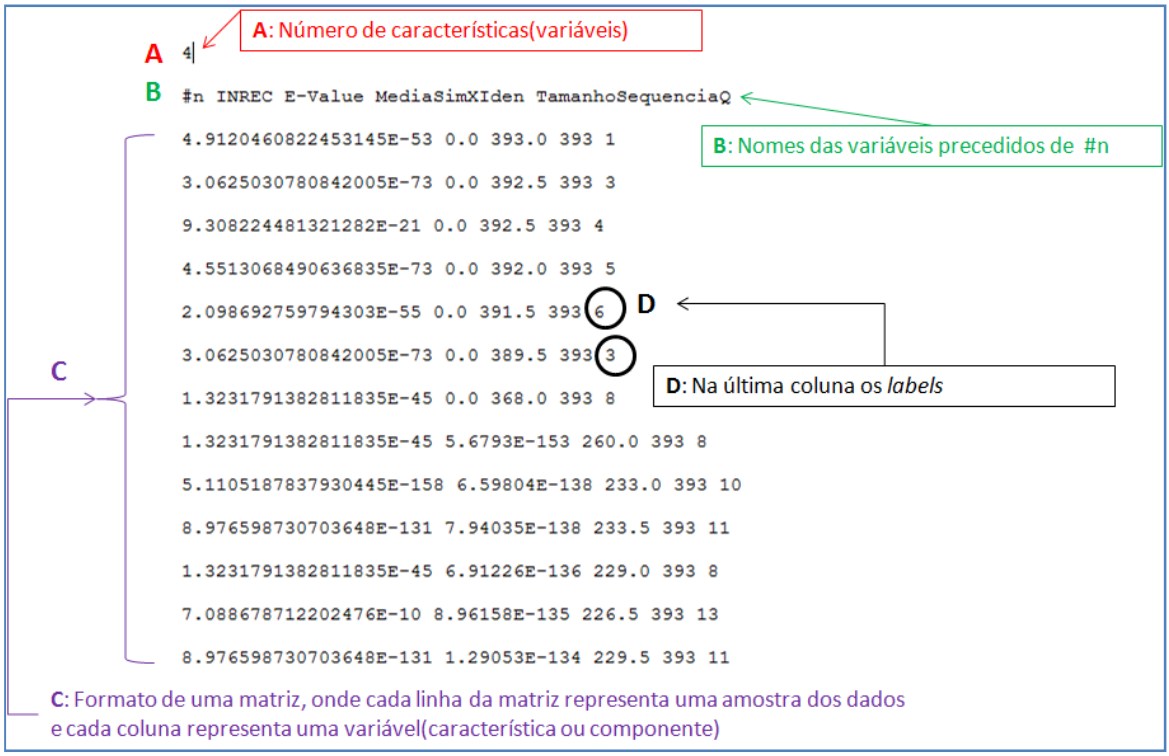


FIGURA 12 - TRECHO DE UM ARQUIVO FORMATADO CONFORME ESPECIFICAÇÃO DO SOM TOOLBOX CONTENDO AS CARACTERÍSTICAS QUE SERÃO SUBMETIDAS AO PROCESSO DE AGRUPAMENTO SOM.

FONTE: A autora (2012)

Na primeira linha contém o número de variáveis (características).

Na segunda linha deve conter os nomes das variáveis precedidos de **#n**

A partir da terceira linha o conteúdo deve estar no formato de uma matriz, onde cada linha representa uma amostra dos dados e cada coluna representa uma variável (característica ou componente).

Para cada amostra de dados (linha da matriz) é possível inserir rótulos que facilitam a análise dos dados após o treinamento (VESANTO, *et al.*, 2000). Os rótulos devem ser curtos para facilitar a identificação.

3.14 **Funções Som *Toolbox***

Foi utilizado o Matlab para execução do pacote *Som Toolbox* (VESANTO, *et al.*, 2000), desenvolvido na Universidade da Finlândia pela equipe de pesquisa em Mapas Auto-Organizáveis, dirigida por Teuvo Kohonen e disponível para download no sítio: http://www.cis.hut.fi/projects/somtoolbox/.

As funções utilizadas neste trabalho e suas respectivas finalidades estão apresentadas no **Erro! Fonte de referência não encontrada.**.

Função	Finalidade	Parâmetros
sd = som_read_data (**"Nome do arquivo"**)	Leitura do arquivo com os valores das características	**"Nome do arquivo"**: nome do arquivo que contém as variáveis. Exemplo: Experimento1.txt
som_normalize (**sd**, **'tipo da normalização'**, **1:X ou X**)	Normalizar os valores das características entre 0 e 1. Pode-se normalizar todas as características ou especificar quais serão normalizadas.	**sd:** é à base de dados em estudo **'tipo da normalização' utilizada**: • **'var' –** A variância é normalizada a um. **1:X ou X:** número da coluna da variável a ser normalizada, pode ser todas as variáveis da coluna 1 até a coluna final das variáveis (1:X) ou pode ser apenas uma determinada variável (X).

QUADRO 6 - FUNÇÕES DO SOM TOOLBOX. (**continua**)

Função	Finalidade	Parâmetros
sm = ***som_lininit*** *(sd ,**'tipo da topologia'**) OU*	Inicialização do mapa. ***som_lininit:*** inicialização linear	***sd****:* a base de dados em estudo **'tipo da topologia'**: • '**hexa':** topologia hexagonal
som_seqtrain*(sm,sd) OU* ***som_batchtrain****(sm,sd)*	Função de treinamento ***som_seqtrain:*** treinamento da rede SOM com o algoritmo de sequencial ***som_batchtrain:*** treina a rede SOM com algoritmo batch (lote)	**sm**: é o mapa criado e inicializado com uma das funções de inicialização **sd**: a base de dados em estudo
som_quality (*sm,sd*)	Métricas de desempenho: erro de quantização e erro topográfico	**sm**: é o mapa criado e inicializado com uma das funções de inicialização **sd**: a base de dados em estudo

QUADRO 6 - FUNÇÕES DO SOM TOOLBOX. **(continuação)**

Função	Finalidade	Parâmetros
som_autolabel (sm,sd,' mode')	Rotulagem do mapa	**sm:** é o mapa criado e inicializado com uma das funções de inicialização **sd**: a base de dados em estudo 'mode': algoritmo de rotulamento dos dados,: • '**freq'**: apenas uma instancia de cada rótulo é mantido e a frequência de cada rótulo é adicionada no final • '**vote'**: apenas o rótulo com a maioria dos casos é mantido

QUADRO 6 - FUNÇÕES DO SOM TOOLBOX. **(continuação)**

Função	Finalidade	Parâmetros
som_show(sm,'umat','all','comp',1:X,'empty','Labels','norm')	Visualização da Matriz-U	***sm**: é o mapa criado e inicializado com uma das funções de inicialização* ***'umat'**: visualização da matriz-U* ***'all':** visualização de todos os componentes* ***'comp':** adiciona o nome de cada componente* **1:X**: quais componentes devem ser mostrados, da coluna 1 até a coluna X(4 no caso deste trabalho)* ***'empty'**: faz um mapa vazio para ser preenchido posteriormente* ***Labels:** Adiciona os rótulos* ***'norm'**: Mostra os valores d dados na barra colorida*

QUADRO 6 - FUNÇÕES DO SOM TOOLBOX. **(continuação)**

Função	Finalidade	Parâmetros
som_show_add('label',sm,'subplot',X)	Adiciona rótulos no mapa formado	***'label':*** deve ser adicionado rótulo à Matriz-U ***sm****:* é o mapa criado e inicializado com uma das funções de inicialização ***'subplot':*** *tipo da plotagem que deve ser adicionado* ***X****:* número do mapa que deve ser adicionado os rótulos
som_show_add('hit', ***som_hits****(sm,sd))*	Adiciona hits o mapa formado	*'****hit'****:* deve ser adicionado hits de histograma à Matriz-U ***som_hits:*** adicionar o mapa criado à visualização

QUADRO 6 - FUNÇÕES DO SOM TOOLBOX.

FONTE: Adaptado de (VESANTO, *et al.*, 2000)

3.15 **Identificação e análise dos agrupamentos**

A identificação dos agrupamentos foi feita de forma manual, interpretando a Matriz-U apresentada na visualização. Para o suporte da organização e análises de verificação dos agrupamentos foi utilizada a planilha Excel.

O interesse foi em identificar o grupo onde se encontra a sequência *query*, o que indicaria que todos os genes que fazem parte desse grupo podem ser sinônimos.

Um esquema de um exemplo fictício é mostrado na FIGURA 13. O grupo que contém a sequência *query*, neste trabalho, será denominado neste trabalho de grupo de interesse.

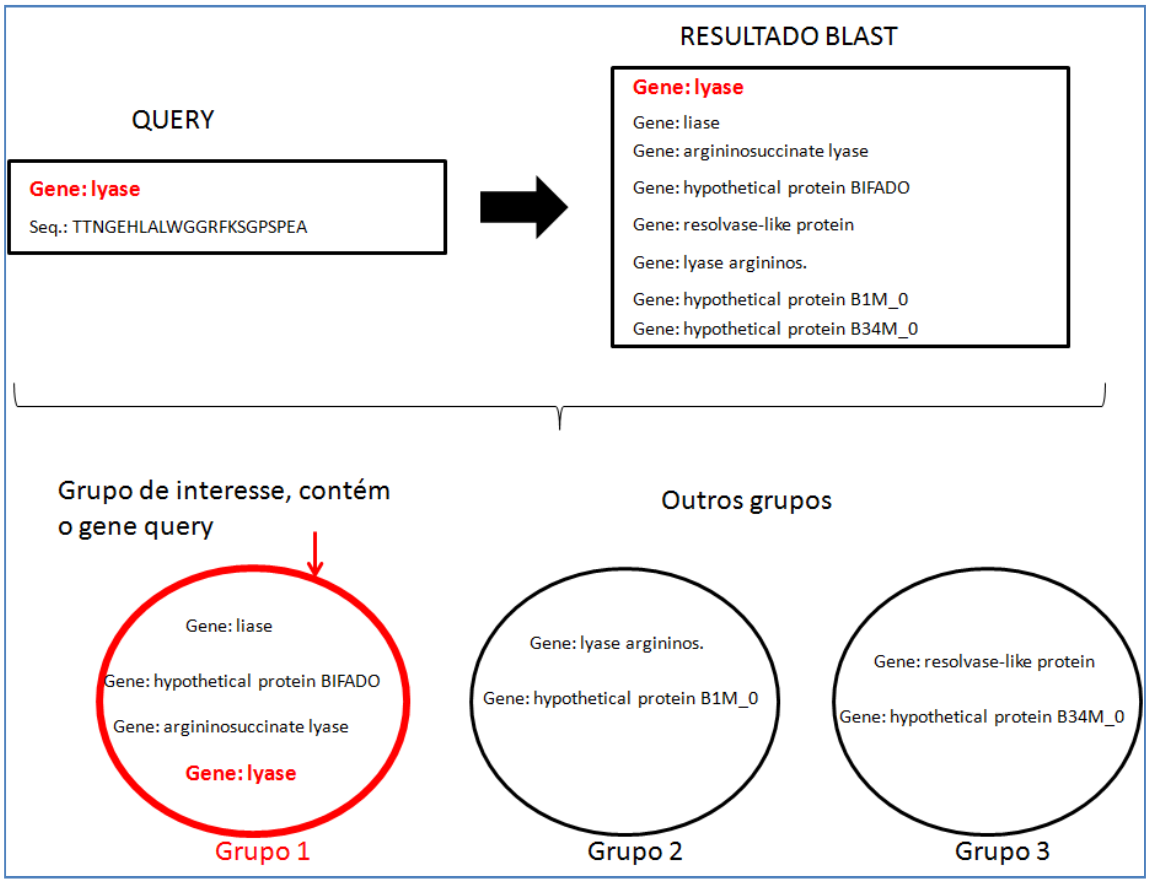


FIGURA 13 – ESQUEMA DE INTERPRETAÇÃO DO RESULTADO OBTIDO COM O AGRUPAMENTO.
FONTE: A autora (2012)

Neste exemplo fictício, o gene *query* é a *lyase*.

O resultado do Blast gerou oito nomes de genes. O resultado do agrupamento resultou em três grupos.

No Grupo 1 está a *lyase* (gene *query*) e outros três genes, dessa forma poderíamos inferir que estes outros três genes que estão no mesmo grupo da *lyase* são sinônimos dela.

Outra interpretação que pode ocorrer, além da identificação de possíveis sinônimos, é que há indícios que o gene hipotético (*hypothetical protein BIFADO*) pode ser uma *lyase* por pertencerem ao mesmo grupo.

3.15.1 Identificação dos agrupamento através da Matriz-U

Para identificar os grupos, foi aplicada a técnica denominada Matriz-U, onde a matriz apresentada é quase duas vezes maior que o mapa original, pois existem hexágonos adicionais entre todos os pares de unidades de mapa vizinhas, ou seja, considera cada célula como um hexágono central cercado por outros hexágonos, para delimitação dos agrupamentos.

As cores variam de acordo com uma escala de distâncias. No caso do Som *Toolbox* as cores variam do azul escuro ao vermelho, onde a cor azul escuro representa as células (neurônios) mais próximas, ou seja, os agrupamentos. As cores mais claras até o vermelho representam a separação dos agrupamentos (VESANTO et al., 2000a).

Na FIGURA 14 um exemplo da interpretação da U-Matrix. A rotulagem do mapa foi realizada através de uma numeração que cada gene recebeu. Um script em Java gerou uma legenda com os nomes para que facilitasse sua posterior identificação.

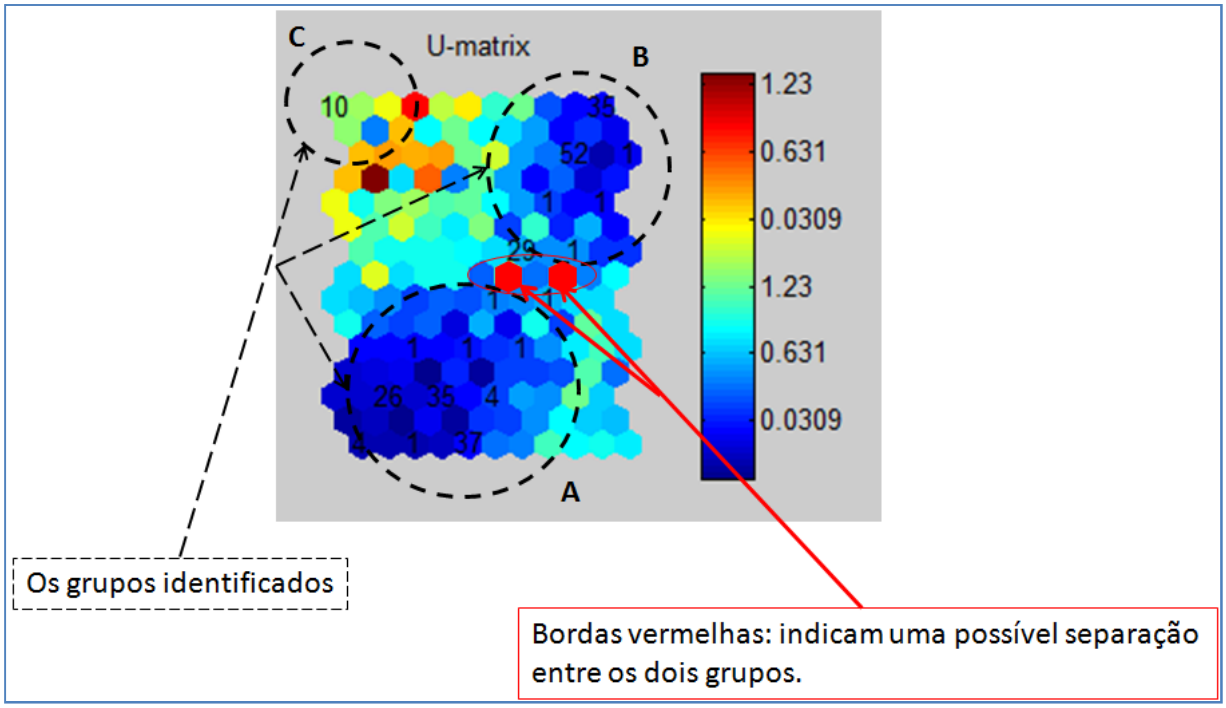


FIGURA 14 - INTERPRETAÇÃO DA MATRIZ-U COM RÓTULOS: DOIS GRANDES GRUPOS (A E B) E UM GRUPO MENOR(C). O QUE INDICA UMA SEPARAÇÃO DOS DOIS GRANDES (A E B) SÃO AS BORDAS DE COR VERMELHA (QUE FORAM EVIDENCIADAS NA IMAGEM).

FONTE: A autora (2012)

3.16 **Métricas de desempenho do mapa**

Para avaliar a qualidade do mapa gerado foram utilizadas duas medidas de desempenho (erro médio de quantização e erro topográfico) que já estão implementadas no *Som Toolbox* através da função [Qe,Te] = ***som_quality*** (sM,sD) como mostra a FIGURA 15.

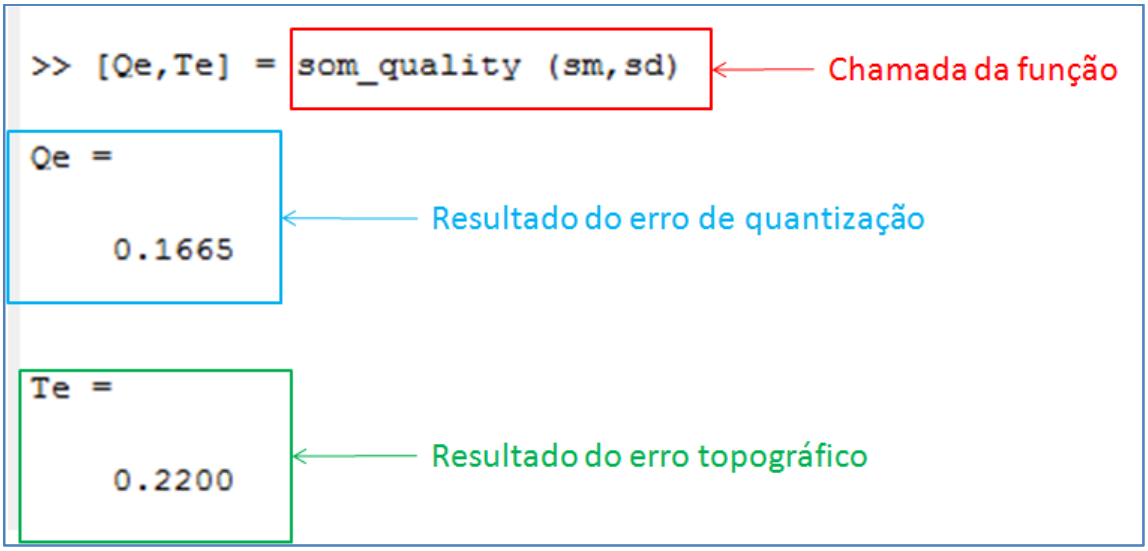


FIGURA 15 - CAPTURA DE TELA DOS ERROS DE QUANTIZAÇÃO E TOPOGRÁFICO.
FONTE: A autora (2012)

Os parâmetros para utilização dessa função:

- Qe: representa o resultado do erro de quantização
- Te: representa o resultado do erro topográfico
- sM: representa o mapa criado treinado, é um parâmetro obrigatório
- sD: representa a base de dados utilizada, é um parâmetro obrigatório

3.17 **Armazenagem dos dados**

Os dados resultantes da interpretação dos agrupamentos foram armazenados no Bando de Dados de acordo com o esquema apresentado na FIGURA 16.

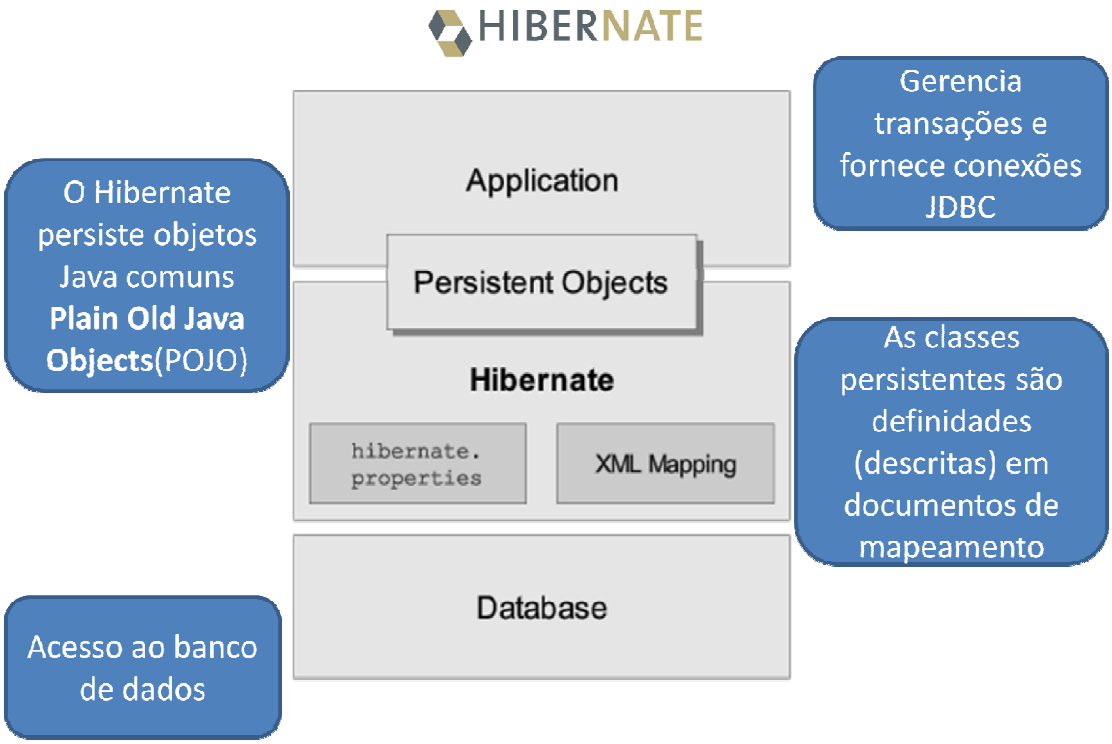


FIGURA 16 - ESQUEMA PARA ARMAZENAR AS INFORMAÇÕES NO BANCO DE DADOS. FONTE: (HIBERNATE, 2011)

Cada camada presente no esquema apresentado na FIGURA 16 tem uma função, são elas:

- a camada de aplicação gerencia as transações e fornece as conexões JDBC;
- a camada de persistência de objetos interage entre a camada de aplicação e a camada do Hibernate, ela é responsável por persistir os *scripts* que contém as classes POJOs;
- a camada do Hibernate é onde as classes persistentes são definidas em um documento de mapeamento denominado web.xml;
- e, por fim há o acesso ao banco de dados onde as transações são efetivadas.

O diagrama de entidade e relacionamento desenvolvido será detalhado em Resultados e Discussão.

4 RESULTADOS E DISCUSSÃO

Foram realizados diversos experimentos variando a topologia, funções e forma de treinamento até a obtenção de mapas que apresentassem o menor valor de erro topográfico e erro de quantização. Os dez experimentos realizados utilizaram os mesmos parâmetros já descritos anteriormente em Materiais e Métodos.

4.1 O Banco de Dados

A implementação do banco de dados possibilitou a armazenagem das informações identificadas com os agrupamentos. Foi realizada uma modelagem do banco de dados que comportasse tanto os dados do NR como as novas informações obtidas com os agrupamentos.

- O modelo de dados: Um DER (diagrama de entidade e relacionamento) foi construído para atender as necessidades descritas anteriormente. A FIGURA 17 demonstra o DER do banco de dados deste trabalho.

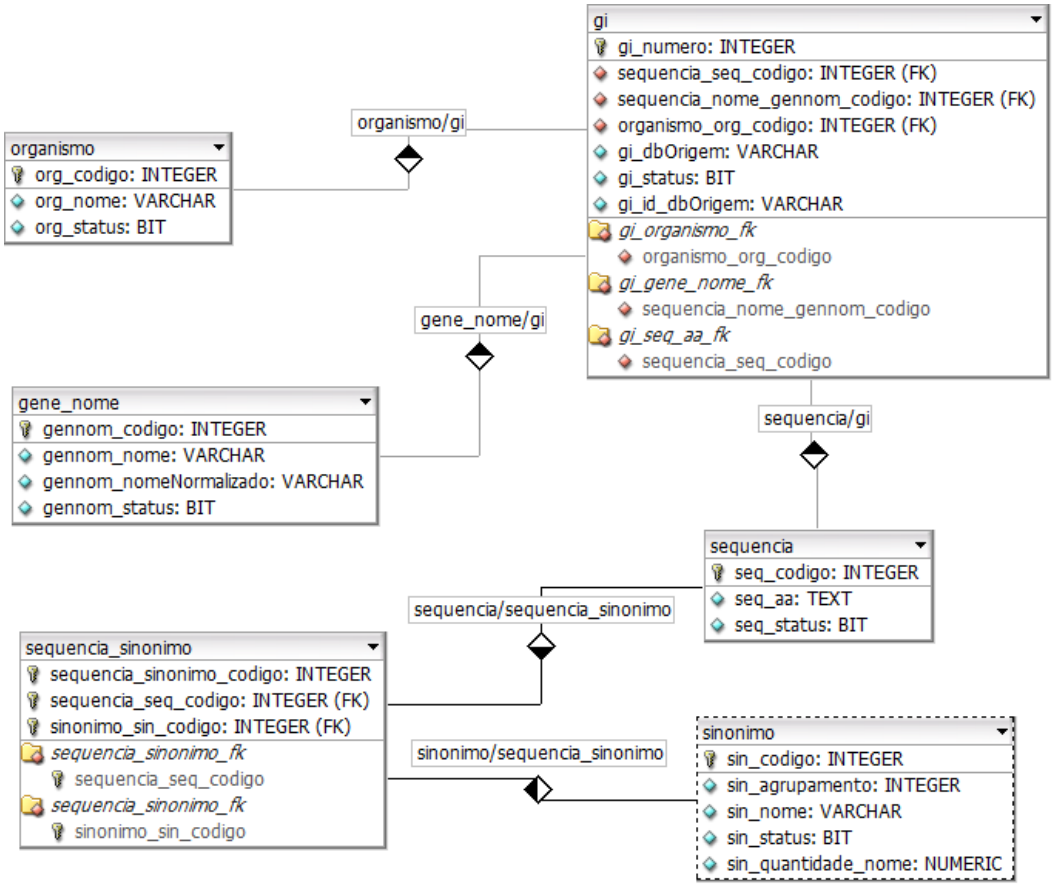


FIGURA 17- DIAGRAMA DE ENTIDADE E RELACIONAMENTO DESENVOLVIDO PARA ESTE TRABALHO.

FONTE: A AUTORA.

A criação deste banco de dados facilitou a seleção de sequências de aminoácidos utilizadas para os experimentos deste trabalho e permitiu armazenar as informações identificadas nos agrupamentos. No Apêndice B está disponível o dicionário de dados deste modelo.

4.2 **Experimentos realizados**

Foram realizados dez experimentos onde cada experimento continha um conjunto de dados de 100 genes (esta quantidade de genes foi arbitrariamente selecionada nos parâmetros do BlastP), gerados pela ferramenta BlastP. Esses

experimentos foram realizados para analisar e validar os grupos encontrados, a FIGURA 18 exibe de forma resumida os resultados obtidos nos experimentos realizados.

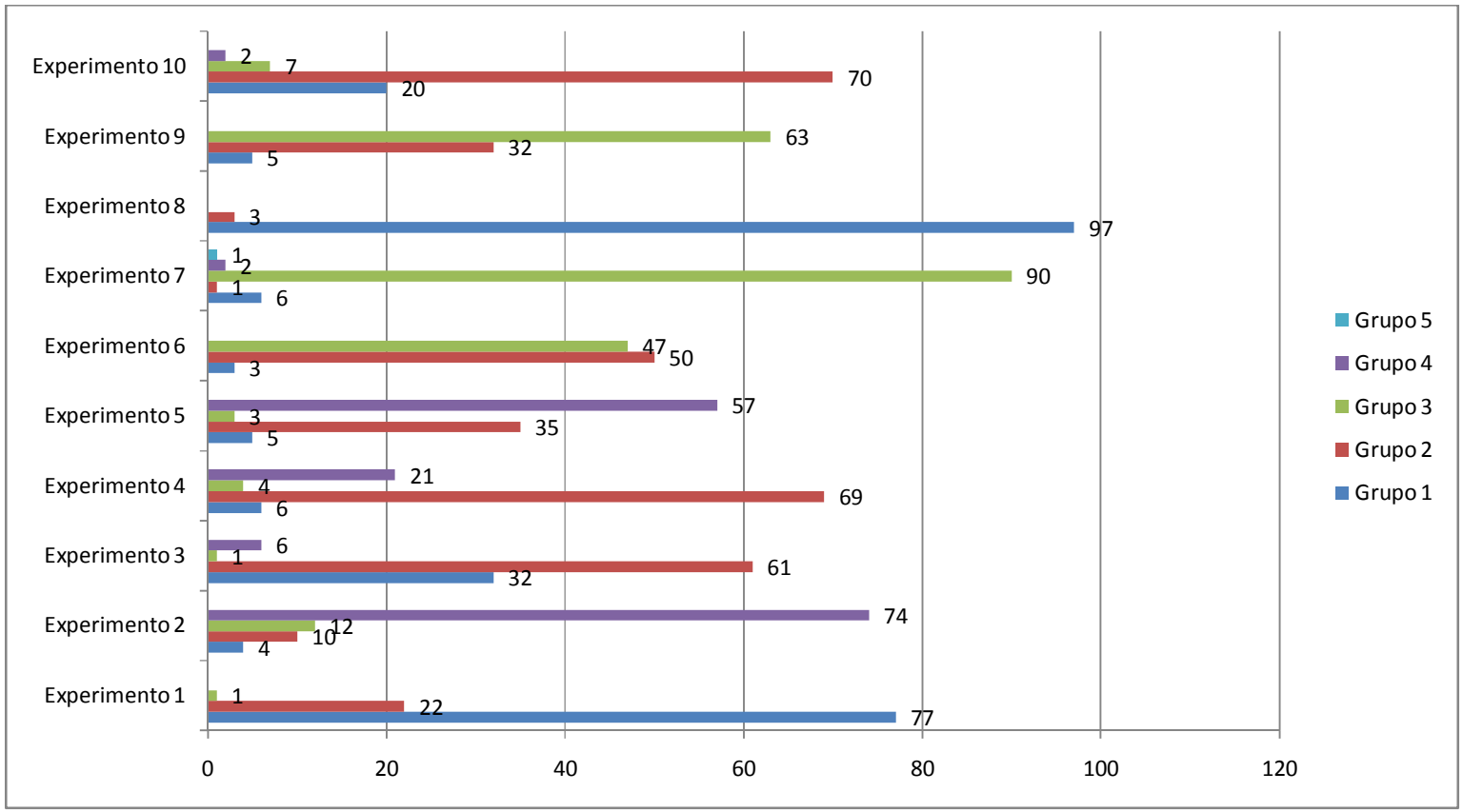


FIGURA 18 - RESUMO DOS RESULTADOS DOS DEZ EXPERIMENTOS REALIZADOS.

FONTE: A autora (2012)

No gráfico apresentado as barras indicam os grupos e cada barra apresenta a quantidade de genes pertencente ao grupo.

Nos agrupamentos podem ocorrer diversas ocorrências: nomes de genes pertencentes à mesma família, nomes de genes pertencentes à família diferentes, nome de genes hipotéticos, nome de genes genéricos (como por exemplo, somente nome da família nome de um domínio, até mesmo nenhum nome), quando encontrado um gene pertencente a uma família diferente em um grupo que predominada uma determinada família, este será investigado a fim de identificar o motivo.

O tamanho dos mapas formados em todos os experimentos é o mesmo, pois há o mesmo número de dados (100) em cada um deles. O mapa original é formado por uma matriz de 6X8 totalizando 48 neurônios e o mapa formado para visualização da Matriz-U é uma matriz de 11X15. O mapa formado para visualização da matriz-U é quase o dobro do tamanho do mapa original, isso ocorre porque são adicionados hexágonos coloridos entre os neurônios adjacentes que representam a distância entre esses neurônios.

Na FIGURA 19 há um exemplo representativo dos mapas, onde se pode observar a diferença de tamanho entre eles.

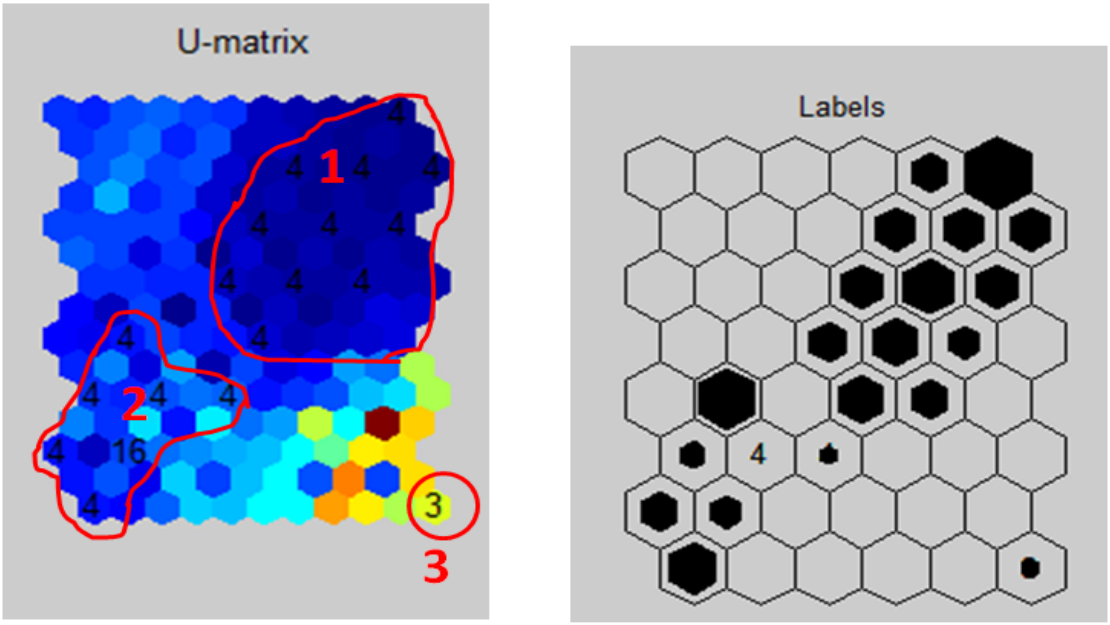


FIGURA 19 – EXEMPLO DOS MAPAS FORMADOS. (A) MAPA DA MATRIZ-U. MAPA ORIGINAL. (B)

FONTE: A AUTORA (2012)

O mapa (A) representa a Matriz-U de tamanho 11X15 e o mapa (B) representa o mapa original de tamanho 6X8.

4.2.1 Experimento 1

Por questões de legibilidade serão detalhados dois dos dez experimentos realizados, os demais estão disponíveis no Apêndice C. Estes dois experimentos foram escolhidos por apresentarem todas as características abordadas nos demais experimentos.

O gene *query* desse experimento é o *Argininosuccinate lyase*. É apresentada através da FIGURA 20 a visualização da Matriz-U referente ao Experimento 1.

FIGURA 20 - MATRIZ-U DO EXPERIMENTO 3.2.1. (A) CIRCULADO EM VERMELHO OS TRÊS GRUPOS IDENTIFICADOS. (B) MAPA COM A FREQUÊNCIA DOS GENES EM CADA UNIDADE.
FONTE: A AUTORA (2012).

A Matriz-U resultante mostra a existência de três grupos. O grupo 1 apresenta 77 genes, o grupo 2 apresenta 22 genes enquanto que o grupo 3 apresenta apenas um gene. A TABELA 1 apresenta detalhes das informações apresentadas e identificadas graficamente.

TABELA 1 - GENES IDENTIFICADOS EM CADA UM DOS AGRUPAMENTOS DO EXPERIMENTO 1.

Grupo	Nome do gene	Quantidade de genes	Família
GRUPO 1	*Argininosuccinate lyase*	72	Lyase
	hypothetical protein BIFANG	1	-
	hypothetical protein BIFDEN_01897	1	-
	RecName: Full=Argininosuccinate lyase	1	Lyase
	argh	1	Lyase
	putative argininosucci-nate lyase	1	Lyase
Grupo 2	*Argininosuccinate lyase*	20	Lyase
	putative argininosuccinate lyase	1	Lyase
	hypothetical protein BIFANG_03230	1	-
Grupo 3	*hypothetical protein BIFADO_01950*	1	-

FONTE: A autora (2012)

No grupo 1 foram identificados dois genes hipotéticos, um gene *putative Argininosuccinate lyase*, uma sigla (argh), um gene com o nome *RecName: Full=Argininosuccinate lyase* e 72 genes referente a *Argininosuccinate lyase.*

Com base nesse resultado há uma forte indicação de que os genes descritos como hipotéticos no grupo 1 sejam uma *Argininosuccinate lyase,* porém, não se pode afirmar pois apenas procedimentos realizados em laboratório é que confirmariam essa hipótese.

Entretanto é possível afirmar que a sigla *argh* presente neste grupo representa uma *Argininosuccinate lyase.*

O gene *RecName: Full=Argininosuccinate lyase* ilustra que ainda poderia ser realizada uma validação retirando palavras que não fazem parte do nome do gene (neste caso específico "*RecName: Full="* poderia ser retirado do nome do gene).

Todas essas informações identificadas com os agrupamentos foram armazenadas no banco de dados como esquematizado na FIGURA 21.

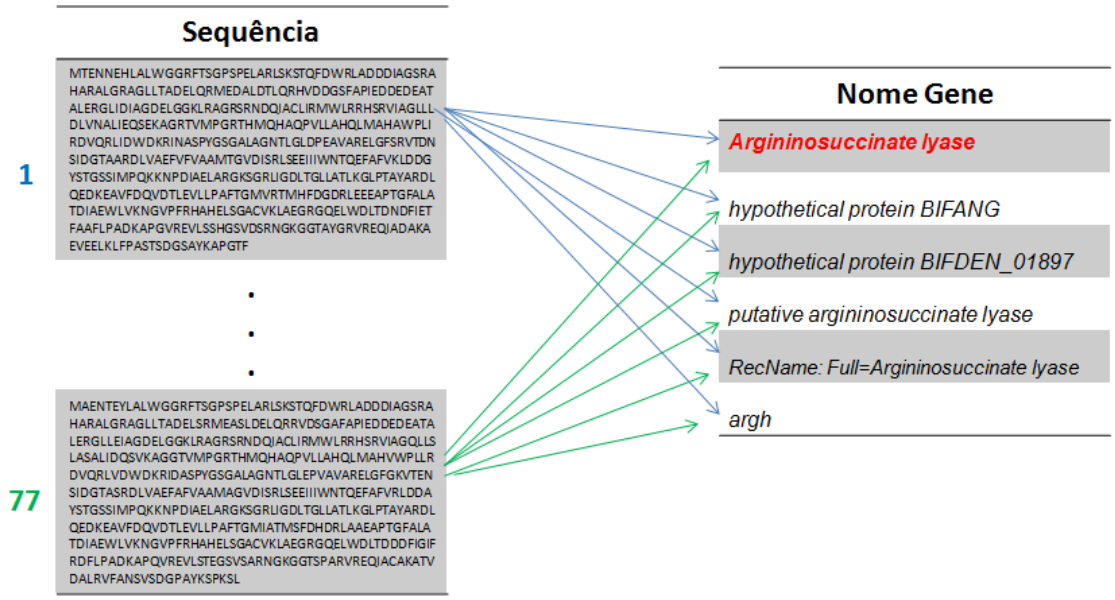


FIGURA 21 - ESQUEMA REPRESENTANDO A FORMA EM QUE AS INFORMAÇÕES SÃO ARMAZENADAS NO BANCO DE DADOS.

FONTE: A autora (2012)

Cada uma das 77 sequências contidas no grupo 1 tem seis nomes de genes relacionados à elas. Em letras vermelhas é apontado o nome do gene que aparece

com maior frequência (*Argininosuccinate lyase*), que pode fornecer um subsídio para a escolha do nome de um novo gene a ser anotado.

4.2.2 Experimento 2

Neste experimento dos 100 genes que foram obtidos com o resultado da ferramenta BlastP, 26 deles tinham nomes distintos. Como pode ser observado na TABELA 2 é possível visualizar o nome original do gene e o respectivo rótulo (que consiste em um número aleatório produzido por um script automático escrito em Java levando em consideração a normalização realizada no nome do gene) utilizado para auxiliar na análise do resultado dos agrupamentos obtidos.

TABELA 2 - RÓTULOS UTILIZADOS PARA AUXILIAR A ANÁLISE E A QUANTIDADE DE NOMES DE GENES QUE OS REPRESENTA.

Nome original do gene (sem normalização)	Rótulo (considerando a normalização)	Quantidade
ABC superfamily ATP binding cassette transporter, ABC protein	13	05
ABC transporter	65	06
ABC transporter ATP-binding protein	3	40
ABC transporter, ATP-binding protein	3	14
ABC- transporter ATP binding protein	3	1
ABC transporter family protein	15	01
ABC transporter related protein	77	04
ABC transporter-like ATP-binding protein	45	01
ABC-type antimicrobial peptide transport system, ATPase component	40	02

TABELA 2 - RÓTULOS UTILIZADOS PARA AUXILIAR A ANÁLISE E A QUANTIDADE DE NOMES DE GENES QUE OS REPRESENTA.
(continua)

Nome original do gene (sem normalização)	Rótulo (considerando a normalização)	Quantidade
hypothetical protein CAT7_09005	86	01
hypothetical protein HMPREF0428_00848	22	01
hypothetical protein HMPREF0433_01076	28	01
hypothetical protein LfarK3_01742	21	01
hypothetical protein lin2471	4	01
hypothetical protein lin2725	23	01
hypothetical protein lmo2372	1	01
hypothetical protein lmo2580	33	01

TABELA 2 - RÓTULOS UTILIZADOS PARA AUXILIAR A ANÁLISE E A QUANTIDADE DE NOMES DE GENES QUE OS REPRESENTA. **(continuação)**

Nome original do gene (sem normalização)	Rótulo (considerando a normalização)	Quantidade
hypothetical protein LmonF_01221	14	01
hypothetical protein LmonocytFSL_07750	11	01
hypothetical protein LMRG_02687	29	01
lipoprotein releasing system, ATP-binding protein	27	01
lipoprotein-releasing system ATP-binding protein LolD	8	05
macrolide ABC transporter ATP-binding protein/permease	35	01
macrolide export ATP-binding/permease protein MacB	16	02
Phosphonate-transporting ATPase	87	01
putative ABC transporter ATP-binding protein	57	02

TABELA 2 - RÓTULOS UTILIZADOS PARA AUXILIAR A ANÁLISE E A QUANTIDADE DE NOMES DE GENES QUE OS REPRESENTA.
(conclusão)

Nome original do gene (sem normalização)	**Rótulo (considerando a normalização)**	**Quantidade**
putative hemin import ATP-binding protein HrtA	5	02
pyridoxamine 5'-phosphate oxidase protein	79	01
TOTAL	**-**	**100**

FONTE: A autora (2012)

Nesta tabela é possível observar como a normalização dos nomes dos genes foi útil. Escrito em letras vermelhas os genes com os mesmos nomes, diferenciados apenas por uma vírgula e um hífen (grifados em azul), mas que foram agrupados em um único nome de gene (*ABC transporter ATP-binding protein*), tendo assim o mesmo rótulo para os três nomes de genes.

Após os dados serem submetido ao SOM treinado, foi possível visualizar a Matriz-U correspondente, onde se pode observar a presença de quatro grupos.

Este resultado pode ser observado na FIGURA 22 que mostra também uma sobreposição entre a Matriz-u e a matriz que contém os rótulos (C), o que facilita ainda mais a interpretação dos agrupamentos, além dos hits de histograma (D) onde podemos visualizar a quantidade de genes que ativaram determinados neurônios.

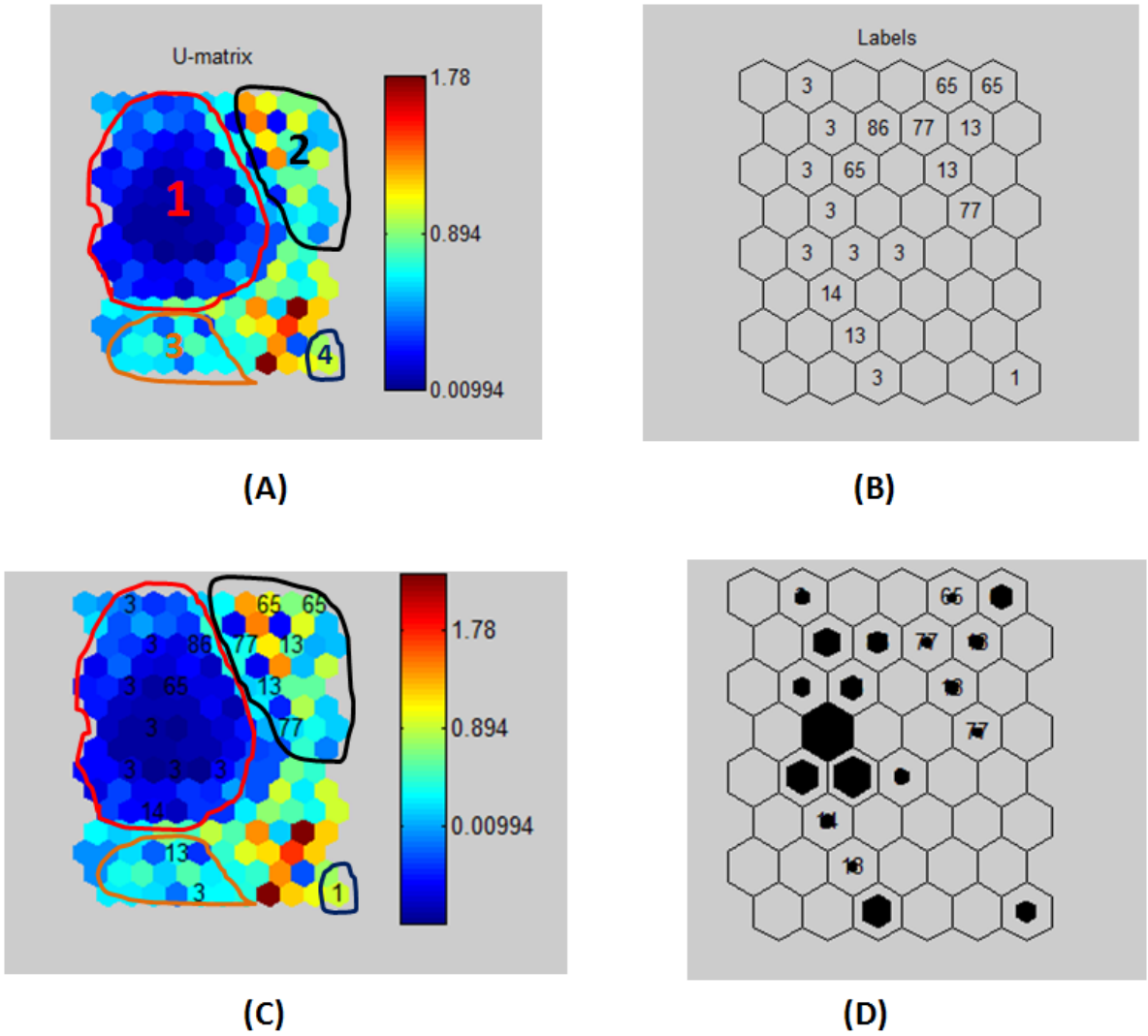


FIGURA 22 - VISUALIZAÇÃO DO RESULTADO OBTIDO COM O EXPERIMENTO 3.2.2. (A) MAPA COM A MATRIZ-U COM OS GRUPOS IDENTIFICADOS MARCADOS DE 1-4, (B) MAPA COM LABELS, (C) RÓTULOS SOBREPOSTOS NA MATRIZ-U E (D) MAPAS COM HITS DE HISTOGRAMA.

FONTE: A AUTORA (2012)

Para visualizar quais genes foram ativados em cada unidade, é utilizada a função *som_autolabel* (**sM**, **sD**, **'freq'**) do Som Toolbox, onde: **sM** é o mapa criado, inicializado e treinado, **sD** é a base de dados e o terceiro parâmetro '**freq'** representa a frequência em que cada dado (gene) ativou determinado neurônio.

Na FIGURA 23 é possível observar uma ampliação da captura de tela de uma unidade do mapa desse experimento, os demais serão omitidos por questões de legibilidade, porém os dados encontrados em cada unidade estão disponíveis na

TABELA 3.

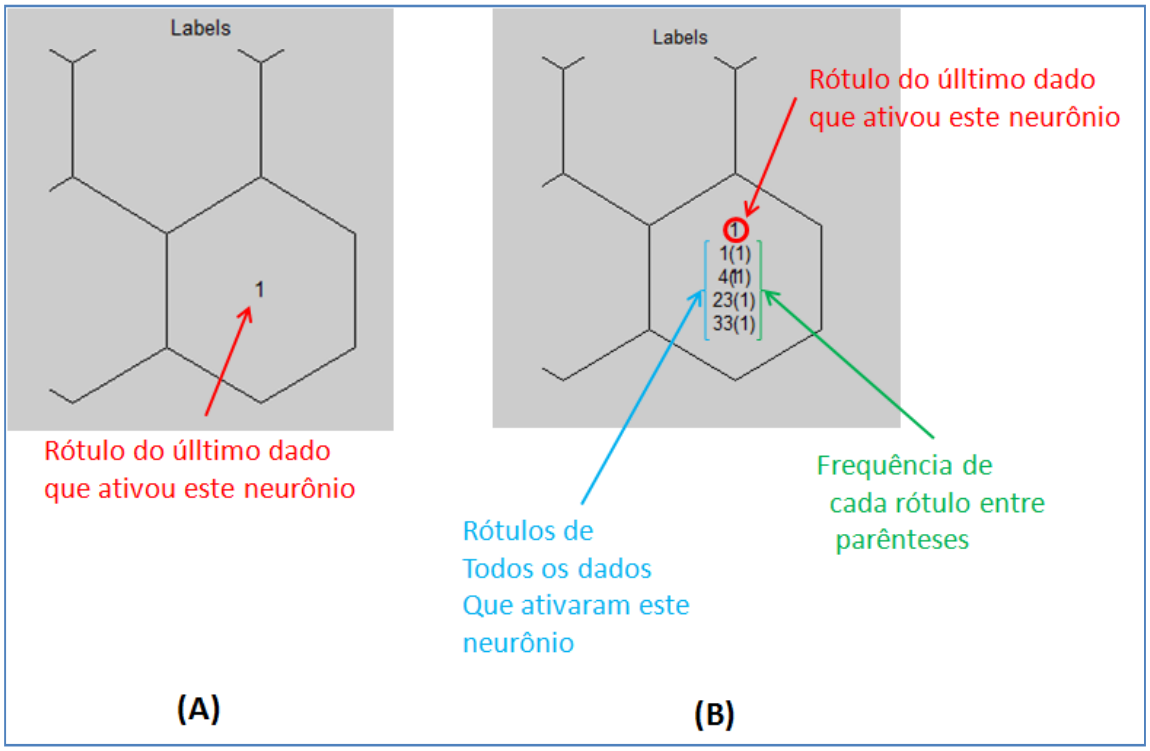


FIGURA 23 - FREQUÊNCIA DOS DADOS EM UM NEURÔNIO. (A) VISUALIZAÇÃO SOMENTE COM RÓTULO. (B) VISUALIZAÇÃO UTILIZANDO A FUNÇÃO DE FREQUÊNCIA.

FONTE: A autora (2012)

Nessa unidade do mapa a FIGURA 23.A apresenta apenas o rótulo do último dado que ativou este neurônio (rótulo: 1), na FIGURA 23.B com a ampliação é possível verificar todas as entradas que ativaram este neurônio (rótulos: 1, 4, 23, 33) e suas respectivas frequências.

No grupo 1 foram identificados quatro genes todos hipotéticos e que pertencem ao mesmo neurônio, a distância entre o neurônio em que eles estão contemplados e os vizinhos mais próximos é grande o que indica que são genes diferentes (não tem nenhuma relação de vizinhança).

No grupo 2 foram identificados dez genes (sendo um hipotético) e no grupo 3 13 genes em sua maioria pertencente à família *ABC Transporter*.

No grupo 4 foram identificados 74 genes, este é o grupo que contém a maioria dos genes incluindo o gene *query ABC-transporter ATP binding protein*.

A

TABELA 3 mostra quais genes foram identificados em cada grupo e suas respectivas famílias de acordo com o *software* PFAM.

TABELA 3 - GENES IDENTIFICADOS EM CADA UM DOS AGRUPAMENTOS DO EXPERIMENTO 2.

Grupo	Nome do gene	Quantidade de genes	Família
Grupo 1	*hypothetical protein lmo2580*	01	-
	hypothetical protein lin2725	01	-
	hypothetical protein lin2471	01	-
	hypothetical protein lmo2372	01	-
Grupo 2	*ABC transporter, ATP-binding protein*	04	ABC_tran (PF00005)
	putative hemin import ATP-binding protein HrtA	02	ABC_tran (PF00005)
	lipoprotein-releasing system ATP-binding protein LolD	02	ABC_tran (PF00005)
	hypothetical protein LmonocytFSL_07750	01	-
	ABC superfamily ATP binding cassette transporter, ABC protein	01	ABC_tran (PF00005)

TABELA 3 - GENES IDENTIFICADOS EM CADA UM DOS AGRUPAMENTOS DO EXPERIMENTO 2. **(continua)**

Grupo	Nome do gene	Quantidade de genes	Família
Grupo 3	*ABC transporter ATP-binding protein*	02	ABC_tran (PF00005)
	lipoprotein-releasing system ATP-binding protein LolD	01	ABC_tran (PF00005)
	ABC superfamily ATP binding cassette transporter, ABC protein	04	ABC_tran (PF00005)
	ABC transporter protein	03	ABC_tran (PF00005)
	ABC transporter related protein	03	ABC_tran (PF00005)
Grupo 4	***ABC-transporter ATP binding protein***	**49**	**ABC_tran (PF00005)**
	pyridoxamine 5'-phosphate oxidase protein	01	?
	lipoprotein-releasing system ATP-binding protein LolD	02	ABC_tran (PF00005)
	hypothetical protein LmonF_01221	01	-
	ABC transporter family protein	01	ABC_tran (PF00005)

TABELA 3 - GENES IDENTIFICADOS EM CADA UM DOS AGRUPAMENTOS DO EXPERIMENTO 2. **(continuação)**

Grupo	Nome do gene	Quantidade de genes	Família
	macrolide export ATP-binding/permease protein MacB	02	ABC_tran (PF00005)
	macrolide export ATP-binding/permease protein MacB	01	ABC_tran (PF00005)
	hypothetical protein LfarK3_01742	01	-
	hypothetical protein HMPREF0428_00848	01	-
	lipoprotein releasing system, ATP-binding protein	01	ABC_tran (PF00005)
	hypothetical protein HMPREF0433_01076	01	-
	hypothetical protein LMRG_02687	01	ABC_tran (PF00005)
	macrolide ABC transporter ATP-binding protein/permease	01	ABC_tran (PF00005)
	ABC-type antimicrobial peptide transport system, ATPase component	02	ABC_tran (PF00005)

TABELA 3 - GENES IDENTIFICADOS EM CADA UM DOS AGRUPAMENTOS DO EXPERIMENTO 2. **(conclusão)**

Grupo	Nome do gene	Quantidade de genes	Família
	ABC transporter-like ATP-binding protein	01	ABC_tran (PF00005)
	putative ABC transporter ATP-binding protein	02	ABC_tran (PF00005)
	ABC transporter	03	ABC_tran (PF00005)
	ABC transporter related protein	01	ABC_tran (PF00005)
	hypothetical protein CAT7_09005	01	-
	Phosphonate-transporting ATPase	01	ABC_tran (PF00005)
TOTAL	-	100	-

FONTE: A autora (2012)

No grupo quatro foi encontrado o gene *pyridoxamine 5'-phosphate oxidase protein* que não pertence à família ABC_tran (PF00005) que é a família da maioria dos genes que estão neste grupo.

Foi então realizada uma pesquisa no PFAM utilizando apenas o nome do gene *pyridoxamine 5'-phosphate oxidase protein* e o resultado obtido foi que este gene pertence a família Pyridox_oxidase (PF01243).

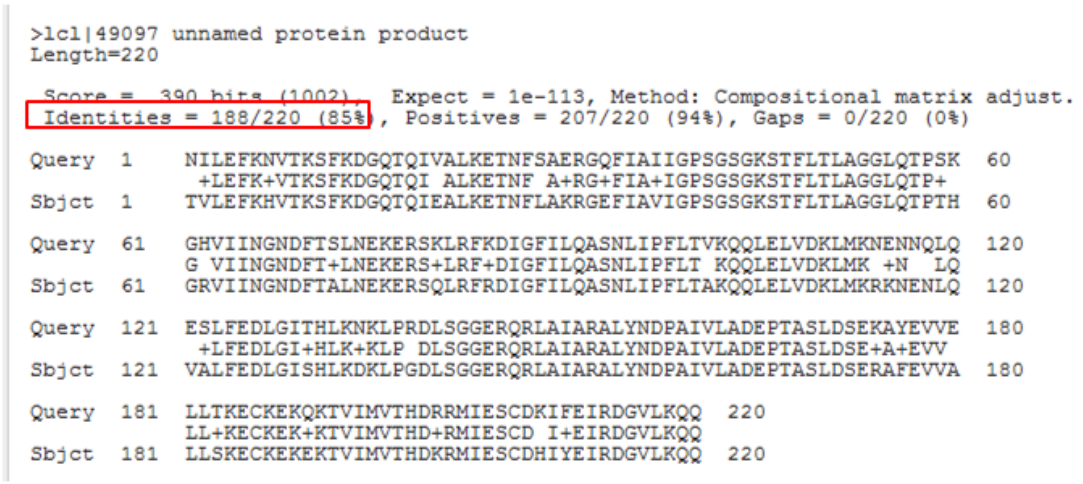
Depois de encontrada a família do gene *pyridoxamine 5'-phosphate oxidase protein* foi realizada uma nova consulta para identificar em qual agrupamento de

famílias esta estava presente. O resultado obtido através do PFAM a família Pyridox_oxidase (PF01243) encontra-se no agrupamento denominado FMN-binding (CL0336).

Neste agrupamento ainda são encontradas mais sete famílias(DUF385, DUF447, Flavin_Reduc, FMN_bind_2, Pyrid_oxidase_2, Pyridox_ox_2 e Pyridox_oxase_2) e nenhuma delas se refere a família ABC_tran (PF00005) o que descartou qualquer possibilidade do gene *pyridoxamine 5'-phosphate oxidase protein* ter alguma ligação com a família ABC_tran (PF00005).

Com o intuito de investigar o motivo de este gene estar em um grupo onde a maioria dos genes pertencem à família ABC Transporter, foi realizado um alinhamento entre a sequência de aminoácido referente ao gene *pyridoxamine 5'-phosphate oxidase protein* contra todas as outras sequências pertencentes ao mesmo grupo, a fim de verificar qual o percentual de similaridade dos alinhamentos.

A FIGURA 24 apresenta o melhor alinhamento obtido entre a sequência de aminoácido (*query*) referente ao gene *pyridoxamine 5'-phosphate oxidase protein* e a sequência referente ao gene *ABC transporter ATP-binding protein (subject)* que corresponde a 85% de identidade, ou seja, este resultado é mais um indício que o gene em questão deveria ser anotado com um *Transportador ABC* e não como um gene *pyridoxamine 5'-phosphate oxidase protein.*



```
>lcl|49097 unnamed protein product
Length=220

 Score =  390 bits (1002),  Expect = 1e-113, Method: Compositional matrix adjust.
 Identities = 188/220 (85%), Positives = 207/220 (94%), Gaps = 0/220 (0%)

Query  1    NILEFKNVTKSFKDGQTQIVALKETNFSAERGQFIAIIGPSGSGKSTFLTLAGGLQTPSK  60
            +LEFK+VTKSFKDGQTQI ALKETNF A+RG+FIA+IGPSGSGKSTFLTLAGGLQTP+
Sbjct  1    TVLEFKHVTKSFKDGQTQIEALKETNFLAKRGEFIAVIGPSGSGKSTFLTLAGGLQTPTH  60

Query  61   GHVIINGNDFTSLNEKERSKLRFKDIGFILQASNLIPFLTVKQQLELVDKLMKNENNQLQ  120
            G VIINGNDFT+LNEKERS+LRF+DIGFILQASNLIPFLT KQQLELVDKLMK +N  LQ
Sbjct  61   GRVIINGNDFTALNEKERSQLRFRDIGFILQASNLIPFLTAKQQLELVDKLMKRKNENLQ  120

Query  121  ESLFEDLGITHLKNKLPRDLSGGERQRLAIARALYNDPAIVLADEPTASLDSEKAYEVVE  180
            +LFEDLGI+HLK+KLP DLSGGERQRLAIARALYNDPAIVLADEPTASLDSE+A+EVV
Sbjct  121  VALFEDLGISHLKDKLPGDLSGGERQRLAIARALYNDPAIVLADEPTASLDSERAFEVVA  180

Query  181  LLTKECKEKQKTVIMVTHDRRMIESCDKIFEIRDGVLKQQ  220
            LL+KECKEK+KTVIMVTHD+RMIESCD I+EIRDGVLKQQ
Sbjct  181  LLSKECKEKEKTVIMVTHDKRMIESCDHIYEIRDGVLKQQ  220
```

FIGURA 24 - ALINHAMENTO REALIZADO ENTRE OS GENES*: PYRIDOXAMINE 5'-PHOSPHATE OXIDASE PROTEIN* (*QUERY*) X *ABC TRANSPORTER ATP-BINDING PROTEIN* (*SUBJECT*)

FONTE: A autora (2012)

4.3 **Erro topográfico e erro de quantização**

Os valores apresentados na TABELA 4 são referentes a cada experimento realizado e são obtidos depois de treinada a rede através da função do Som Toolbox: som_quality (sm,sd). Os parâmetros utilizados em cada experimento foram os mesmos conforme descrito em Materiais e Métodos.

TABELA 4 - VALORES DO ERRO TOPOGRÁFICO E ERRO DE QUANTIZAÇÃO OBTIDOS EM CADA UM DOS 10 EXPERIMENTOS REALIZADOS.

Experimento	Erro topográfico	Erro de quantização
1	0,1470	0,2500
2	0,1000	0,1617
3	0,3392	0,0700
4	0,2663	0,0900
5	0,3145	0,0500
6	0,2550	0,1500
7	0,3613	0,1100
8	0,1122	0,3800
9	0,1575	0,2000
10	0,3792	0,1400

FONTE: A autora (2012)

É possível observar que quando o erro de quantização diminui, há um aumento do erro topográfico. Segundo KOHONEN (1990) o melhor mapa é o que tem o menor valor de erro de quantização, pois este estaria mais bem ajustado aos vetores de entrada.

Seria possível diminuir ainda mais os erros de quantização dos experimentos: 1, 8 e 9; mas para isso seria necessário modificar os parâmetros a cada novo experimento realizado.

O objetivo foi deixar um grupo de parâmetros padrões (apresentado na seção 4.3) onde a maioria dos casos apresentados obtivesse um erro de quantização menor que 0,2 (valor arbitrário - inferido pela média da variação dos erros de quantização obtidos nos experimentos, que foi no máximo de 0,49).

4.4 **Comparativo deste trabalho com os trabalhos correlatos**

Os trabalhos descritos na seção 2.1.2 foram comparados com este trabalho e os elementos de comparação incluem as seguintes características:

a) Alinhamento de sequência: se considera o alinhamento entre sequências como uma característica para auxílio na identificação de sinônimo;
b) Forma sintática do nome: se considera a estrutura sintática do nome do gene para auxílio na identificação de sinônimo;
c) Conexão com banco de dados: permite realizar consultas através de um banco de dados previamente populado com os sinônimos;
d) Bases de dados utilizadas: quais bases de dados com informações gênicas são utilizadas como base para o processo de agrupamento;
e) Inclui genes hipotéticos: inclui genes descritos como hipotéticos nos agrupamentos;
f) Formato arquivo para agrupamento: qual formato de arquivo é utilizado para realizar agrupamentos;
g) Pesquisa através da *internet*: se é possível realizar a pesquisa de sinônimos através da internet;
h) Identifica erro de anotação: Capacidade de identificar possíveis erros de anotação;

No QUADRO 7 é possível observar as diferenças entre os principais trabalhos de relacionados a sinonímia de genes.

	Este trabalho	**Biothesaurus**	**GPSDB**	**CD-HIT versão 4.5.4**
(a) Alinhamento sequência	sim	não	não	sim
(b) Forma sintática do nome	sim	sim	sim	não
(c) Conexão com banco de dados	sim	sim	sim	não
(d) Bases de dados utilizadas	A	B	C	D
(e) Inclui genes hipotéticos	sim	não	não	-
(f) Formato arquivo para agrupamento	Arquivo de texto (.txt)	Agrupamento pré processados através do nome do gene	Agrupamento pré processados através do nome do gene	Arquivo Fasta
(g) Pesquisa através *internet*	não	sim	sim	sim
(h) Identifica erro de anotação	sim	não	não	não

QUADRO 7 - COMPARATIVO ENTRE OS ESTUDOS SOBRE SINÔNIMOS DE GENES.
FONTE: A autora (2012)

No item (**d**) referente às bases de dados utilizadas para agrupamentos, há uma diversidade de conjuntos referentes a elas e estão descritos a seguir:

A – GenBank, EMBL Data Library, DDBJ, NBRF PIR, Protein Research Foundation, SWISS-PROT, Brookhaven Protein Data Bank, Patents, NCBI Reference Sequence;

B – UniProt, Swiss-Prot, TrEMBL, PIR-PSD, Entrez Gene, RefSeq e GenPept, MGD, SGD, RGD, FlyBase e WormBase, HUGO, EC enzyme nomenclature e OMIM;

C – LocusLink, Swiss-Prot, GDB, HUGO, OMIM, MGD, RGD, Ratmap; Flybase; SGD, TAIR, WormBase, SubtiList e EcoGene

D – NCBI NR, Swissprot e PDBO. O usuário também pode fornecer uma base de dados específica.

5 CONCLUSÃO

Foi analisada a estrutura do arquivo NR (Non-Redundant Data Base) disponibilizado pelo NCBI. Com a identificação dos padrões existentes neste arquivo, foi possível desenvolver um script em Java para interpretar automaticamente as informações contidas nele.

Um banco de dados relacional foi criado para integrar as informações disponíveis no banco de dados NR do NCBI com as informações identificadas sobre os sinônimos dos genes, adquiridas no decorrer da mineração de dados realizada neste trabalho.

Foram aplicadas técnicas de KDD para identificar os sinônimos dos nomes dos genes, baseado nos resultados obtidos através dos alinhamentos de sequências de aminoácidos e na estrutura sintática dos nomes dos genes. Estas técnicas foram fundamentais para compor a metodologia desenvolvida.

Nos experimentos realizados com a Rede de Kohonen foi possível identificar em um dos grupos no experimento 2, um gene que não pertencia à mesma família dos demais, o que pode indicar um provável erro de anotação. Com essa indicação, poderia ser realizados procedimentos laboratoriais que confirmassem a hipótese levantada.

A metodologia desenvolvida é aplicável para descrever genes hipotéticos, *putative* e outros sem uma função descrita ou conhecida, podendo indicar uma possível função a estes genes após o agrupamento.

Outras vantagens observadas com a utilização do algoritmo SOM: rapidez com que gera os mapas, não havendo a obrigação de realizar um grande número de iterações para se obter um bom resultado; A forma de visualização do mapa do através da Matriz-U permite identificar a quantidade de grupos formados por meio de uma matriz de cores. Deste modo, esta técnica de visualização é muito interessante, pois a princípio ela não necessita que seja informado o número de grupos a serem formados.

A metodologia descrita pode ser utilizada com qualquer sequência de aminoácido que gere um grupo de dados através de alinhamentos, onde possa ser submetidos à rede SOM.

6 TRABALHOS FUTUROS

Este trabalho é apenas o início do estudo desse assunto tão vasto. Para melhorar ainda mais a metodologia, poderia ser aplicadas em conjunto técnicas de validação biológica como, por exemplo, as ontologias de funções, de processos biológicos e de componente celular que são fornecidas através do consórcio *Gene Ontology* (ASHBURNER, *et al.*, 2000), o que enriqueceria os agrupamentos com coerências biológicas.

Para a metodologia ser aplicada em uma grande quantidade de dados como, por exemplo, todo o Banco de Dados NR, seria necessário estudar ou desenvolver uma forma de automatizar o processo de identificação dos agrupamentos, que neste trabalho foi feito de forma manual.

No provável erro de anotação identificado em um dos experimentos, poderia ser realizado um procedimento laboratorial para confirmar a hipótese levantada neste estudo, com isso, haveria mais um subsídio de confiabilidade aos resultados apresentados.

7 REFERÊNCIAS

ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal of Molecular Biology**, v. 215(3), p. 403–410, 1990.

ALTSCHUL, S. F.; GISH, W. Local alignment statistics. **Journal of Molecular Biology**, v. 215, p. 403-410, 1990.

ASHBURNER, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. **Nat. Genet.**, 25(1), May 2000. 25-29.

BASTOS, E. N. F. **Uma Rede Neural Auto-Organizável Construtiva para Aprendizado Perpétuo de Padrões Espaço-Temporais**. UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL, PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO. Porto Alegre, p. 137. 2007.

BATISTAKIS, Y.; HALKIDI, M.; VAZIRGIANNIS, M. On clustering validation techniques. **J. Intell. Inf. Syst.**, 2001. 107–145.

BEDELL, J.; KORF, I.; YANDELL, M. **BLAST**. [S.l.]: O'Reilly Media, 2003.

BELDA-FERRE, P. et al. Mining Virulence Genes Using Metagenomics. **PLoS One**, v. 6(10), p. e24975, 2011.

BERLYN, M. B.; LETOVSKY, S. Genome-related datasets within the E.coli Genetic Stock Center database. **Nucl. Acids Res.**, n. 20(23), 1992. 6143-6151.

BLUM, A. L.; LANGLEY, P. Selection of relevant features and examples in machine learning. **Artificial Intelligence**, n. 97(1–2), 1997. 245–271.

BRUIJN, B. D.; MARTIN, J. **Finding gene function using LitMiner**. [S.l.]: [s.n.]. 2003. p. 486-494.

BULT, C. J. et al. The Mouse Genome Database (MGD): mouse biology and model systems. **Nucleic Acids Res.**, v. 36, p. D724–D728, 2008.

CAMACHO, C. et al. BLAST+: architecture and applications. **BMC Bioinformatics**, v. 10, n. 421, 2009.

CHRISTIE, K. R. et al. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. **Nucl. Acids Res.**, 32(suppl 1), 2004. D311-D314.

COHEN, K. B. et al. Contrast And Variability In Gene Names. **Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain**, p. 14–20, 2002.

CONSORTIUM, T. U. The Universal Protein Resource (UniProt) 2009. **Nucleic Acids Res.**, v. 37, p. D169–D174, 2009.

COUTINHO, M. A. **Mapeamento de genes nif publicados no NCBI usando conceitos de mineraçã de dados e inteligência artificial**. Universidade Federal do Paraná. [S.l.]. 2011.

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

DATE, C. J.; DARWEN, H. **A Guide to the SQL Standard:** A user's guide to the standard database language SQL. [S.l.]: Addison-Wesley, 1997.

DELAMARO, M. **Como construir um compilador utilizando ferramentas Java**. [S.l.]: Novatec, 2004.

DEMEREC, M. et al. A proposal for a uniform nomenclature in bacterial genetics. **Genetics**, v. 54, p. 61–76, 1966.

DRYSDALE, R. A.; CROSBY, M. A. FlyBase: genes and gene models. **Nucleic Acids Res**, v. 33, p. D390–D395, 2005.

DWIGHT, S. S. et al. Saccharomyces genome database: Underlying principles and organisation. **Brief Bioinform.**, v. 5, p. 9–22, 2004.

DWINELL, M. R. et al. The Rat Genome Database 2009: variation, ontologies and pathways. **Nucleic Acids Res.**, v. 37, p. D744–D749, 2009.

EYRE, T. A. et al. The HUGO Gene Nomenclature Database, 2006 updates. **Nucleic Acids Res.**, v. 34, p. D319–D321, 2006.

FANG, H. R. et al. **Human Gene Name Normalization using Text Matching with Automatically Extracted Synonym Dictionaries**. [S.l.]: Association for Computational Linguistics. 2006. p. 41-48.

FARIA, E. L. D. et al. **Introdução ao Toolbox de Redes Neurais de Kohonen**. Centro Brasileiro de Pesquisas Físicas;Departamento de Física - Universidade Federal do Espírito Santo. [S.l.]. 2010. (CBPF-NT-001/10).

FAUSETT, L. **Fundamentals of Neural Networks Prentice Hal**. [S.l.]: Englewood, 1994.

FAYYAD, M. U. et al. **Advances in Knowledge Discovery and Data Mining**. [S.l.]: [s.n.], 1996.

FINN, R. D. et al. The Pfam protein families database. **PFAM**, 2010. Disponivel em: <http://pfam.sanger.ac.uk/>. Acesso em: 28 dezembro 2011.

FRECH, C.; CHEN, N. Genome-Wide Comparative Gene Family Classification. **PLoS One**, v. 5(10), p. e13409, 2010.

FUNDEL, K.; ZIMMER, R. Gene and protein nomenclature in public databases. **BMC Bioinformatics**, v. 7, p. 372–384, 2006.

GUERRA, A. M. et al. Assessment of self-organizing map artificial neural networks for the classification of sediment quality. **Environment International**, 2008. 1-9.

GUHA, S.; RASTOGI, R.; SHIM, K. **CURE:** An Efficient Clustering Algorithm for Large Databases. [S.l.]: [s.n.]. 1998.

HANISCH, D. et al. ProMiner: rule-based protein and gene entity recognition. **BMC Bioinformatics**, p. 6(Suppl 1):S14, 2005.

HAYKIN, S. **Neural Networks:** A Comprehensive Foundation. [S.l.]: New York : Macmillan College, 1994.

HIBERNATE. Relational persistence for java and.net. **J. Enterprise**, 2011. Disponivel em: <http://www.hibernate.org/>. Acesso em: 21 Dezembro 2011.

JAIN, A. K.; MAO, J.; AND MOHIUDDIN, K. M. Artificial Neural Networks: A Tutorial. **IEEE Computer**, March 1996. 31-44.

JOHN, G. H.; KOHAVI, R.; PFLEGER, K. **Irrelevant features and the subset selection problem**. Machine Learning: Proceedings of the Eleventh International Conference. San Francisco, CA: Morgan Kaufmann Publishers. 1994. p. 121-129.

JOHNSON, M. et al. NCBI BLAST: a better web interface. **Nucleic Acids Res.**, v. 36(Web Server issue), p. W5-W9, 2008.

KIRA, K.; RENDELL, L. A. **A practical approach to feature selection**. Proceedings of the Ninth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 1992. p. 249–256.

KOHONEN, T. The Self-Organizing Maps. **Proceedings of the IEEE**, v. 78(9), p. 1464-1480, 1990.

KOHONEN, T. et al. **SOM_PAK: the self-organizing map program package.** Helsinki University of Technology, Laboratory of Computer and Information Science. Espoo, Finland. 1995.

KOLLER, D.; SAHAMI, M. **Toward optimal feature selection**. In: Proceedings of the Thirteenth International Conference on Machine Learning. [S.l.]: Morgan Kaufmann. 1996. p. 284–292.

LI, W. et al. CD-HIT Suite: a web server for clustering and comparing biological sequences. **Bioinformatics**, n. 26(5), 1 March 2010. 680–682.

LI, W.; JAROSZEWSKI, L.; GODZIK, A. Tolerating some redundancy significantly speeds up clustering of large protein databases. **Bioinformatics**, n. 18(1), Janeiro 2002. 77-82.

LIU, H. et al. BioThesaurus: a web-based thesaurus of protein and gene names. **Bioinformatics** , v. 22(1), p. 103-105, 2006.

MATHWORKS. The Language Of Technical Computing. **MATLAB**, 2008. Disponivel em: <http://www.mathworks.com/>. Acesso em: 18 janeiro 2012.

MCQUILTON, P. et al. FlyBase 101 – the basics of navigating FlyBase. **Nucleic Acids Research** , 29 Novembro 2011.

MICROSOFT. Excel 2007 help and how-to. **MICROSOFT.**, 2011. Disponivel em: <http://office.microsoft.com/en-us/excelhelp/CL010072903.aspx>. Acesso em: 27 dezembro 2011.

MOSZER, I. et al. SubtiList: the reference database for the Bacillus subtilis genome. **Nucleic Acids Res**, 30(1), 1 janeiro 2002. 62–65.

NCBI. The blast databases. **NCBI**, 2011. Disponivel em: <ftp://ftp.ncbi.nih.gov/blast/db>. Acesso em: 21 Dezembro 2011.

NETBEANS. Code assistance in the netbeans ide java editor: A reference guide. **Netbeans**, 2011. Disponivel em: <http://netbeans.org/kb/docs/java/editor-codereference.html>. Acesso em: 12 Dezembro 2011.

NIEVOLA, J. C. Redes Neurais Artificiais. In: _____ **Sociedade Brasileira de Computação. (Org.) Inteligência Artificial.** [S.l.]: Porto Alegre: Editora da Sociedade Brasileira de Computação - ESBC, v. 1, 2004. Cap. 1, p. 1-50.

PAVY, N. et al. Generation, annotation, analysis and database integration of 16,500 white spruce EST clusters. **BMC Genomics**, v. 6, 2005.

PEARSON, P. L. The genome data base (GDB)—a human gene mapping repository. **Nucleic Acids Res.**, n. 19(Suppl), 25 abril 1991. 2237–2239.

PFAM. The Pfam database. **PFAM**, 2011. Disponivel em: <http://pfam.sbc.su.se/>. Acesso em: 19 janeiro 2012.

PGADMIN PostgreSQL tools. **pgAdmin**, 2009. Disponivel em: <http://www.pgadmin.org/>. Acesso em: 21 janeiro 2012.

PILLET, V. et al. GPSDB: a new database for synonyms expansion of gene and protein names. **Bioinformatics**, v. 21(8), p. 1743-1744, 2004.

PÖLZLBAUER, G. **Survey and comparison of quality measures for self-organizing maps**. Proceedings of the 5th Workshop on Data Analysis (WDA'04). [S.l.]: [s.n.]. 2004. p. 67-82.

POSTGRESQL. Postegresql. **Postegresql.**, 2011. Disponivel em: <http://www.postgresql.org/>. Acesso em: 04 Janeiro 2012.

POVEY, S. et al. The HUGO Gene Nomenclature Committee (HGNC). **Hum. Genet.**, 109(6), 2001. 678-80.

PROCESS, J. C. JSR-000317 JavaTM Persistence 2.0. **JAVA**, 2011. Disponivel em: <http://jcp.org/aboutJava/communityprocess/final/jsr317/index.html>. Acesso em: 28 Dezembro 2011.

PROSDOCIMI, F. et al. Bioinformática: Manual do Usuário. Um guia amplo e básico sobre diversos aspectos desta nova ciência. **Revista Biotecnologia Ciência e Desenvolvimento**, n. 29, p. 12-25, Janeiro 2003.

PUDIL, P.; NOVOVICOVÁ, J.; KITTLER, J. Floating search methods in feature selection. **Pattern Recognition Letters**, n. 15(12), 1994. 1119–1125.

SEWELL, M. Feature Selection, 2007. Disponivel em:<. Acesso em: 17 janeiro 2012.

SIMKOVICS, S. **Enhancement of the ANSI SQL Implementation of PostgreSQL**. Department of Information Systems, Vienna University of Technology. [S.l.]. 1998.

SONNHAMMER, E. L. L.; EDDY, S. R.; DURBIN, R. Pfam: a comprehensive database of protein families based on seed alignments. **Proteins**, n. 28(3), 1997. 405-20.

SOUZA, J. A. D. **RECONHECIMENTO DE PADRÕES USANDO INDEXAÇÃO RECURSIVA**. UNIVERSIDADE FEDERAL DE SANTA CATARINA. [S.l.]. 1999.

SPLENDORE, A. Para que existem as regras de nomenclatura genética? **Rev. bras. hematol. hemoter**, n. 27(2), 2005. 148-152.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. **Introduction to Data Mining**. [S.l.]: Addison-Wesley, 2006.

TAO, T. Ftp downloadable blast databases from ncbi. **NCBI**, 2011. Disponivel em: <http://www.ncbi.nlm.nih.gov/staff/tao/URLAPI/blastdb.html6>. Acesso em: 23 Dezembro 2011.

TSURUOKA, Y. et al. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. **Bioinformatics**, v. 23(20), p. 2768-2774, 2007.

TSURUOKA, Y.; MCNAUGHT, J.; ANANIADOU, S. Normalizing biomedical terms by minimizing ambiguity and variability. **BMC Bioinformatics**, v. 9(Suppl 3), p. S2, 2008.

ULTSCH, A. Self-Organizing Neural Networks for Visualization and Classification. **In Information and Classification (1993)**, p. 307-313, 1993.

VESANTO, J. et al. **SOM Toolbox for Matlab 5**. Helsinki University of Technology. [S.l.]. 2000.

VESANTO, J. et al. **SOM Toolbox for matlab 5**. Helsinki University of Technology. Finland. 2000.

VESANTO, J.; SULKAVA, M.; HOLLMEN, J. **On the decomposition of the self-organizing map distortion measure**. Proceedings of the Workshop on Self-Organizing Maps (WSOM'03). Hibikino, Kitakyushu, Japão: [s.n.]. 2003. p. 11-17.

WAIN, H. et al. Guidelines for human gene nomenclature. **Genomics**, 79(4), 2002. 464-70.

WU, Y.; TAKATSUKA, M. The Geodesic Self-Organizing Map and its error analysis. **Proceedings of the Twenty-eighth Australasian conference on Computer Science**, Darlinghurst, Australia, 38, 2005. 343-351.

YE, J.; MCGINNIS, S.; MADDEN, T. L. BLAST: improvements for better sequence analysis. **Nucleic Acids Res.**, v. 34(Web Server issue), p. W6-W9, 2006.

YI, G.; JUNG, J. Algorithm for large-scale clustering across multiple genomes. **Biomedical Informatics**, v. 7(5), p. 251–256, 2011.

ZINNER, M. G. DbDesigner. **DbDesigner**, 2003. Disponivel em: <http://fabforce.net/dbdesigner4/>. Acesso em: 21 janeiro 2012.

8 APÊNDICES

Apêndice A – Script para interpretação arquivo NR

FIGURA A. 1 - MÉTODO EM JAVA REFERENTE À LEITURA E INTERPRETAÇÃO DO ARQUIVO NR UTILIZANDO EXPRESSÕES REGULARES.

```
public static void lerArquivoNr() throws Exception {
    ArrayList<SeqI> arquivo;
    SeqI itemArquivo;
    String[] idGI;
    String[] stringNomesGene;
    Sequencia seqObj = new Sequencia();
    Organismo orgObj = new Organismo();
    GI giObj = new GI();
    GeneNome geneNomeObj = new GeneNome();

    // leitura d arquivo NR, a partir de um local especificado
    arquivo = Fasta.readMany("c:/BlastNr/nr2.txt", 10240);

    // pega cada sequencia de todo o texto que se referente a ela
    for (int k = 0; k < arquivo.size(); k++) {
        itemArquivo = arquivo.get(k);

        //adiciona a sequencia
        seqObj.setSequencia(itemArquivo.subseq(itemArquivo.bounds().plus()).toString());

        idGI = itemArquivo.id().replace("|", ",,").split(",,");
        giObj.setGi_codigo(Integer.parseInt(idGI[1]));

        //get nome do gene e nome do [ organismo ]
        stringNomesGene = itemArquivo.description().split(" ");
        giObj.setGi_dbOrigem(idGI[2] + " " + idGI[3]);


        //Expressões regulares para distiguir as tabulações do arquivo NR
        stringNomesGene[0] = stringNomesGene[0].replace("[", ";;");
        stringNomesGene[0] = stringNomesGene[0].replace("]", ",,");
        stringNomesGene[0] = stringNomesGene[0].replace("|", "@@");

        Pattern pOrganismo = Pattern.compile(";;(.*?),,");
        Matcher mOrganismo = pOrganismo.matcher(stringNomesGene[0]);


        Pattern pNomeGene = Pattern.compile("(.*?);;");
        Matcher mNomeGene = pNomeGene.matcher(stringNomesGene[0]);

        if (mNomeGene.find()) {

            geneNomeObj.setGennom_nome(mNomeGene.group(1));

        }

        if (mOrganismo.find()) {

            orgObj.setNome(mNomeGene.group(1));
        }

        giObj.setGi_id_dbOrigem(idGI[1]);


        // Armazenagem no BD
        AcessoDB acesso = new AcessoDB();
        acesso.inserirGene(geneNomeObj);
        acesso.inserirOrganismo(orgObj);

    }
```

Apêndice B – Dicionário de dados

QUADRO B. 1 - DICIONÁRIO DE DADOS DA ENTIDADE ORGANISMO.

Entidade: Organismo				
Atributo	**Classe**	**Domínio**	**Tamanho**	**Descrição**
org_codigo	Determinante	Numérico	-	Chave primária da entidade Organismo
org_nome	Simples	Texto	300	Nome do organismo
org_status	Simples	Texto	1	Status ativo ou inativo (A ou I)

QUADRO B. 2 - DICIONÁRIO DE DADOS DA ENTIDADE SEQUENCIA.

Entidade: Sequencia				
Atributo	**Classe**	**Domínio**	**Tamanho**	**Descrição**
seq_codigo	Determinante	Numérico		Chave primária da entidade Sequencia
seq_aa	Simples	Texto	5000	Sequência de aminoácido
seq_status	Simples	Texto	1	Status ativo ou inativo (A ou I)

QUADRO B. 3 - DICIONÁRIO DE DADOS DA ENTIDADE GENE_NOME.

Entidade: Gene_Nome				
Atributo	**Classe**	**Domínio**	**Tamanho**	**Descrição**
gennom_codigo	Determinante	Numérico	-	Chave primária da entidade Gene_Nome
gennom_nome	Simples	Texto	300	Nome do gene
gennom _nomeNormalizado	Simples	Texto	300	Nome do gene após normalização
gennom _status	Simples	Texto	1	Status ativo ou inativo (A ou I)

QUADRO B. 4 - DICIONÁRIO DE DADOS DA ENTIDADE GI.

Entidade: GI				
Atributo	**Classe**	**Domínio**	**Tamanho**	**Descrição**

gi_numero	Determinante	Numérico		Chave primária da entidade GI – é o número apresentado no arquivo NR
sequencia_seq_codigo	Determinante	Numérico	-	Chave estrangeira da entidade Sequencia
sequencia_nome_gennom_codigo	Determinante	Numérico	-	Chave estrangeira da entidade Gene_Nome
organismo_org_codigo	Determinante	Numérico	-	Chave estrangeira da entidade Organismo
Gi_dbOrigem	Simples	Texto	100	Banco de dados de origem do nome do gene
gi_status	Simples	Texto	1	Status ativo ou inativo (A ou I)

QUADRO B. 5 - DICIONÁRIO DE DADOS DA ENTIDADE SINÔNIMO.

Entidade: Sinonimo				
Atributo	**Classe**	**Domínio**	**Tamanho**	**Descrição**
sin_codigo	Determinante	Numérico	-	Chave primária da entidade Sinonimo
gene_nome_gennom_codigo_	Determinante	Numérico	-	Chave estrangeira da entidade Nome_Gene

sin_agrupamento	Determinante	Numérico	-	Número identificador do agrupamento
sin_nome	Simples	Texto	300	Nome de maior frequência que irá representar o grupo
sin_status	Simples	Texto	1	Status ativo ou inativo (A ou I)

Apêndice C – Experimentos realizados

Experimento 3

Neste experimento foi observado a formação de 4 grupos conforme mostra a Matriz-U referente a este processo (Figura C.1).

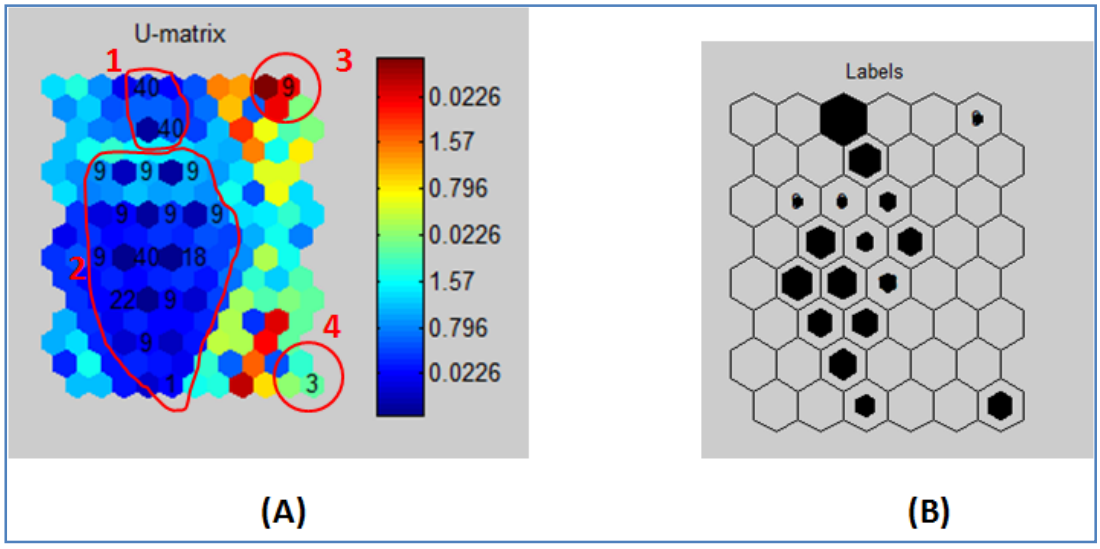


FIGURA C. 1 - CAPTURA DE TELA DA VISUALIZAÇÃO DA MATRIZ-U DO EXPERIMENTO 3. (A) MATRIZ-U COM OS RESPECTIVOS LABELS. (B) HITS DE HISTOGRAMA.

Na Tabela C.1 **Erro! Fonte de referência não encontrada.**podemos verificar quais genes pertencem a cada um dos grupos identificados, o grupo um tem 32 genes, o grupo dois 61 genes, o grupo três 01 gene e o grupo quatro 06 genes.

TABELA C. 1 - GENES IDENTIFICADOS NO EXPERIMENTO 3.

Grupo	**Nome do gene**	**Quantidade de genes**
Grupo 1	transposon Tn21 resolvase	02
	Tn3 family resolvase	01
	putative resolvase	01
	transposon Tn21 resolvase	06
	resolvase-like protein	01
	Resolvase domain protein	03
	site-specific recombinase	01
	resolvase domain-containing protein	06
	transposon resolvase	01
	site-specific recombinase, DNA invertase Pin	02

	hypothetical protein p49879_2p17	01
	hypothetical protein plpl0043	01
	TnpR-like resolvase	01
	recombinase	01
	hypothetical protein pOU7519_72	01
	transposase	01
	hypothetical protein ABTW07_3879	01
	helix-turn-helix protein	01
Grupo 2	resolvase for Tn21	01
	truncated resolvase	01
	transposon Tn21 resolvase	03
	Tn3 family resolvase	01
	putative resolvase	01
	resolvase	**20**
	hypothetical protein B1M_04234	01
	hypothetical protein METUNv1_03110	01

	resolvase-like protein	04
	Resolvase domain protein	02
	site-specific recombinase	02
	resolvase family recombinase	01
	inversion of adjacent DNA	01
	putative resolvase HDEF_p0007	01
	Tn1721 resolvase	01
	TnAtcArs resolvase	01
	Tn4653 resolvase	01
	TnpR resolvase	01
	resolvase family recombinase	02
	putative Tn4653A-like resolvase protein	01
	resolvase domain-containing protein	04
	Tn501 transposition resolvase	01
	resolvase/integrase TnpR	01

	Tn4656/Tn4658 resolvase	01
	TnpR recombinase	03
	Transposon Tn501 resolvase	01
	transposition resolvase	01
	hypothetical protein SMAC_09673	01
	site specific resolvase	01
Grupo 3	resolvase	01
Grupo 4	TnpR	06

Experimento 4

Neste experimento foi identificado a presença de 4 grupos conforme mostra a Matriz-U referente a este processo FIGURA C. 2

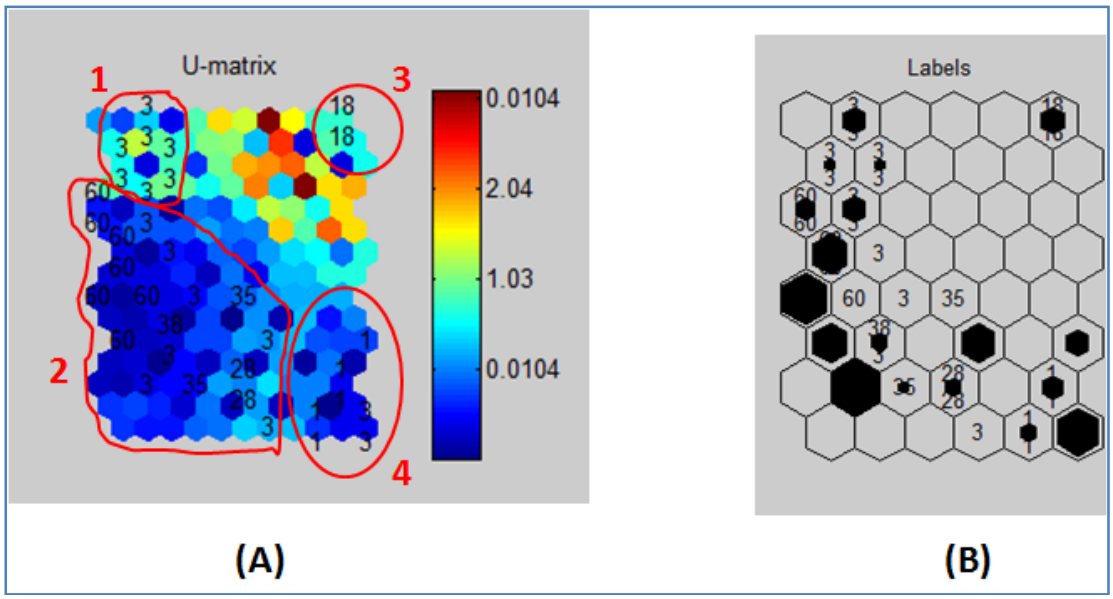


FIGURA C. 2 - CAPTURA DE TELA DO EXPERIMENTO 4 DA VISUALIZAÇÃO DA MATRIZ-U. (A) MATRIZ-U COM OS RESPECTIVOS LABELS. (B) HITS DE HISTOGRAMA.

Na Tabela C.2 podemos verificar quais genes pertencem a cada um dos grupos identificados, o grupo foi identificado 06 genes, no grupo dois 69 genes, no grupo três 04 genes e no grupo quatro 21 genes.

TABELA C. 2 - GENES IDENTIFICADOS NO EXPERIMENTO 4

Grupo	**Nome do gene**	**Quantidade de genes**
Grupo 1	phosphate regulon transcriptional regulatory protein PhoB	04
	two component transcriptional regulator PhoB, winged helix family	01
	transcriptional regulator PhoB	01

Grupo 2	phosphate regulon transcriptional regulatory protein PhoB	23
	putative DNA-binding response regulator PhoB	04
	DNA-binding response regulator in two-component regulatory system with PhoR	01
	response regulator homolog PhoB (phosphate regulon transcriptional regulatory protein)	01
	two component transcriptional regulator	01
	two component transcriptional regulator PhoB, winged helix family	05
	transcriptional regulatory protein	01
	phosphate regulon transcriptional regulatory protein	05
	two-component regulatory system response regulator	01
	phosphate regulon	03

	response regulator	
	phosphate regulon transcriptional regulatory protein PhoB (SphR)	01
	winged helix family two component transcriptional regulator	01
	phosphate regulon transcriptional regulator PhoB	01
	transcriptional regulator PhoB	18
	DNA-binding response regulator in two-component regulatory system with PhoR (or CreC)	01
	two-component system response regulator PhoB	01
	phosphate regulon two-component system, response regulator	01
Grupo 3	PhoB	04
Grupo 4	**DNA-binding response regulator PhoB**	**09**

	phosphate regulon transcriptional regulatory protein PhoB	08
	response regulator homolog PhoB	01
	recombination associated protein	01
	response regulator	02

Experimento 5

Neste experimento foi identificada a presença de quatro grupos conforme mostra a Matriz-U referente a este processo (**Erro! Fonte de referência não encontrada.**).

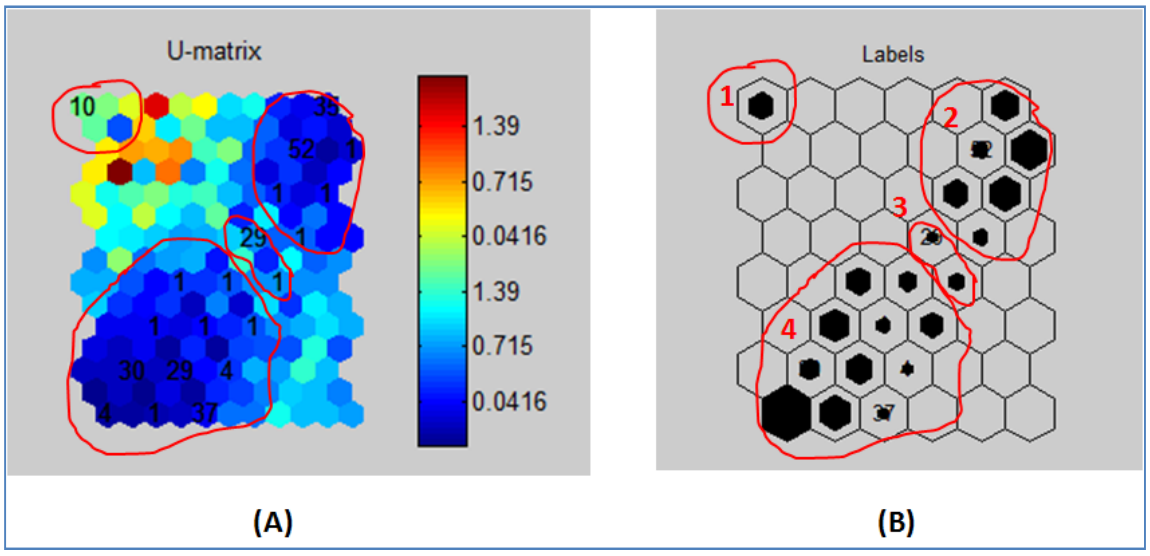


FIGURA C. 3 - CAPTURA DE TELA DO EXPERIMENTO 5 DA VISUALIZAÇÃO DA MATRIZ-U. (A) MATRIZ-U COM OS RESPECTIVOS LABELS. (B) HITS DE HISTOGRAMA.

Na Tabela C.3 podemos verificar quais genes pertencem a cada um dos grupos identificados, o grupo um tem 05 genes, o grupo dois 35 genes, o grupo três 03 genes e no grupo quatro 57 genes.

TABELA C. 3 - GENES IDENTIFICADOS NO EXPERIMENTO 5

Grupo	**Nome do gene**	**Quantidade de genes**
Grupo 1	ribonucleotide reductase	05
Grupo 2	Chain A, Crystal Structure Of Manganese Substituted R2-D84e (D84e Mutant Of The R2 Subunit Of E. Coli Ribonucleotide Reductase)	01
	Chain A, Ribonucleoside-Diphosphate Reductase 1 Beta Chain	01
	Chain A, Ribonucleotide Reductase R2 Subunit From E. Coli	01
	Chain A, Y122f Mutant Of Ribonucleotide Reductase From Escherichia Coli	01
	putative ribonucleoside diphosphate reductase 1, beta subunit, ferritin-like protein	01
	Ribonucleoside-diphosphate reductase	05

	ribonucleoside-diphosphate reductase 1 beta chain	01
	ribonucleoside-diphosphate reductase 1 subunit beta	02
	ribonucleoside-diphosphate reductase, beta subunit	03
	ribonucleotide reductase of class Ia (aerobic), subunit beta	01
	ribonucleotide reductase, beta subunit	01
	ribonucleotide-diphosphate reductase subunit beta	16
	Hypothetical protein	01
Grupo 3	ribonucleoside-diphosphate reductase 1 subunit beta	01
	ribonucleotide-diphosphate reductase subunit beta	02
Grupo 4	**ribonucleotide-diphosphate reductase subunit beta**	**21**
	ribonucleoside-diphosphate reductase, beta subunit	17
	ribonucleoside-diphosphate reductase 1, beta subunit, B2	01
	Chain A, Ribonucleoside-Diphosphate	01

	Reductase 1 Beta Chain	
	Chain A, Crystal Structure Of The Y122h Mutant Of Ribonucleotide Reductase R2 Protein From E. Coli	01
	Chain A, Ribonucleotide Reductase R2 Soaked With Ferrous Ion At Neutral Ph	01
	Chain A, Ribonucleotide Reductase Y122no2y Modified R2 Subunit Of E. Coli	01
	ribonucleoside-diphosphate reductase 1 subunit beta	03
	Chain B, Ribonucleoside-Diphosphate Reductase 1 Beta Chain Mutant E238a	01
	Chain A, Dithionite Reduced E. Coli Ribonucleotide Reductase R2 Subunit, D84e Mutant	01
	Chain A, Azide Complex Of The Diferrous F208a Mutant R2 Subunit Of Ribonucleotide Reductase	01
	Chain A, Substitution Of Manganese For Iron In Ribonucleotide Reductase From Escherichia Coli.	01

	Spectroscopic And Crystallographic Characterization	
	Hypothetical protein	01
	Chain A, Ribonucleotide Reductase R2-D84eW48F MUTANT SOAKED WITH Ferrous Ions At Neutral Ph	01
	Chain A, Autocatalytic Generation Of Dopa In The Engineered Protein R2 F208y From Escherichia Coli Ribonucleotide Reductase And Crystal Structure Of The Dopa-208 Protein	01
	Ribonucleotide reductase of class 1a beta subunit	02
	Ribonucleoside-diphosphate reductase	01
	ribonucleoside-diphosphate reductase class Ia beta subunit	01

Experimento 6

Neste experimento foi identificado a presença de três grupos conforme mostra a Matriz-U referente a este processo (figura C.4).

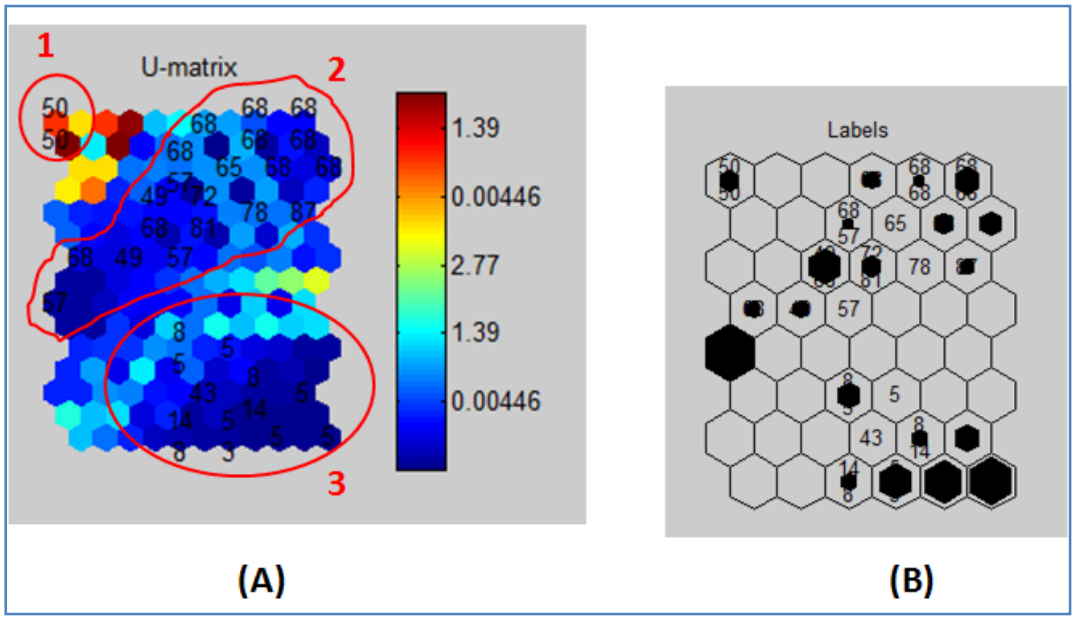


FIGURA C. 4 - CAPTURA DE TELA DO EXPERIMENTO 6 DA VISUALIZAÇÃO DA MATRIZ-U. (A) MATRIZ-U COM OS RESPECTIVOS LABELS. (B) HITS DE HISTOGRAMA.

Na Tabela C.4 podemos verificar quais genes pertencem a cada um dos grupos identificados, o grupo um tem 03 genes, o grupo dois 50 genes e o grupo três 47 genes.

TABELA C. 4 - GENES IDENTIFICADOS NO EXPERIMENTO 6

Grupo	**Nome do gene**	**Quantidade de genes**
Grupo 1	YsdC	03
Grupo 2	M42 glutamyl aminopeptidase	03
	hypothetical protein SDB_00048	01
	Endo-1,4-beta-glucanase	06

	M42 family peptidase	03
	hypothetical protein GK2713	01
	peptidase, M42 family	05
	cellulase	06
	glutamyl-aminopeptidase	02
	deblocking aminopeptidase, M42	13
	hypothetical protein BSNT_04206	01
	putative endo-1,4-beta-glucanase	02
	glucanase/ deblocking aminopeptidase	02
	cellulase M related protein	01
	COG1363: Cellulase M and related proteins	01
	deblocking aminopeptidase	01
	hypothetical protein GYO_3133	01
	M42 family deblocking aminopeptidase	01

Grupo 3	predicted peptidase	01
	M42 glutamyl aminopeptidase	15
	putative fructose-specific phosphotransferase system protein FrvX	10
	frv operon protein frvX	02
	aminopeptidase	05
	putative aminopeptidase ysdC	01
	hypothetical protein ECoL_03844	01
	fructose-specific phosphotransferase system protein FrvX	**07**
	frv operon protein	01
	conserved hypothetical protein	01
	putative peptidase	01
	hypothetical protein ENTCAN_09569	01
	deblocking aminopeptidase, M42	01

Experimento 7

Neste experimento foi identificado a presença de cinco grupos conforme mostra a Matriz-U referente a este processo (figura C.5).

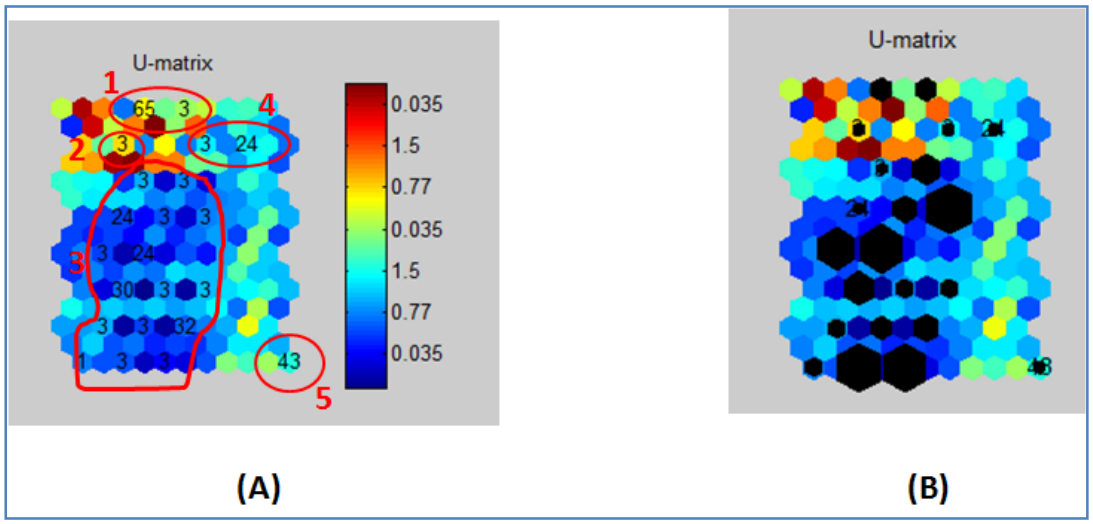


FIGURA C. 5- CAPTURA DE TELA DO EXPERIMENTO 7 DA VISUALIZAÇÃO DA MATRIZ-U. (A) MATRIZ-U COM OS RESPECTIVOS LABELS. (B) HITS DE HISTOGRAMA.

Na Tabela C.5 podemos verificar quais genes pertencem a cada um dos grupos identificados, o grupo um tem 06 genes, o grupo dois 01 genes, o grupo três 90 genes, grupo quatro 02 genes e grupo cinco 01 gene.

TABELA C. 5 - GENES IDENTIFICADOS NO EXPERIMENTO 7

Grupo	**Nome do gene**	**Quantidade de genes**
Grupo 1	ABC transporter ATP-binding protein	03
	ABC superfamily ATP binding cassette	01

	transporter, ABC protein	
	peptide ABC transporter ATPase	01
	ABC-type antimicrobial peptide transport system, ATPase component	01
Grupo 2	ABC transporter ATP-binding protein	01
Grupo 3	**pyridoxamine 5'-phosphate oxidase protein**	**01**
	ABC transporter ATP-binding protein	43
	hypothetical protein HMPREF0433_01076	01
	hypothetical protein HMPREF0428_00848	01
	lipoprotein releasing system, ATP-binding protein	03
	putative lipoprotein ABC transporter, ATP-binding protein	01
	ABC superfamily ATP binding cassette transporter, ABC protein	12
	peptide ABC transporter ATP-	03

	binding protein	
	ABC transporter	09
	lipoprotein-releasing system ATP-binding protein LolD	03
	multidrug resistance ABC superfamily ATP binding cassette transporter, ABC protein	01
	peptide ABC transporter ATPase	02
	putative ABC exporter, ATP-binding subunit	01
	hypothetical protein STRINF_01601	01
	putative ABC transporter ATP-binding protein	03
	ABC-type antimicrobial peptide transport system, ATPase component	01
	antimicrobial peptide ABC transporter ATP-binding protein	01
	ABC transporter ATPase	01
	antimicrobial peptide ABC transporter	01

	ATPase	
	ABC-type antimicrobial peptide transport system protein	01
Grupo 4	ABC transporter ATP-binding protein	01
	ABC superfamily ATP binding cassette transporter, ABC protein	01
Grupo 5	hypothetical protein llmg_2548	01

Experimento 8

Neste experimento foi identificado a presença de 2 grupos conforme mostra a Matriz-U referente a este processo (figura C.6).

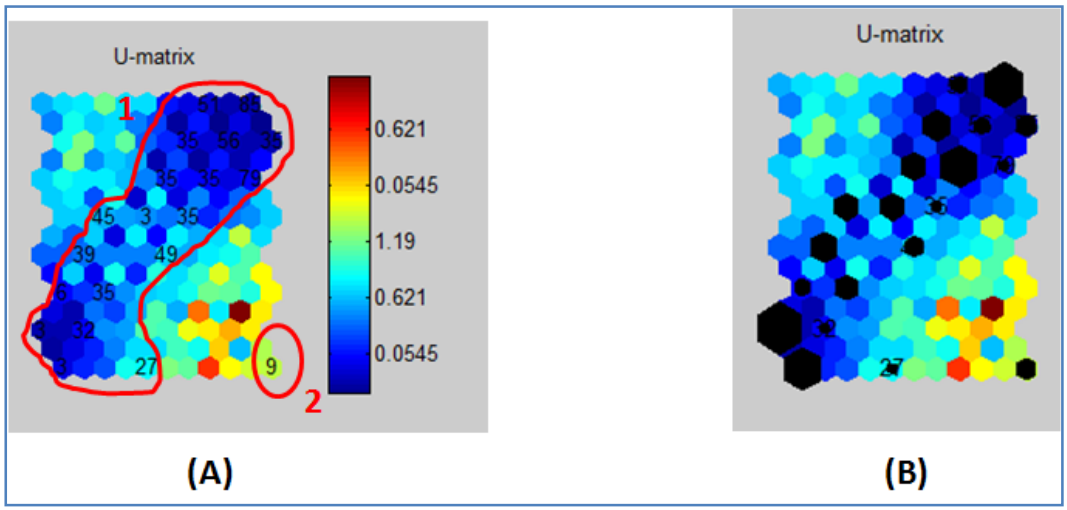


FIGURA C. 6- CAPTURA DE TELA DO EXPERIMENTO 8 DA VISUALIZAÇÃO DA MATRIZ-U. (A) MATRIZ-U COM OS RESPECTIVOS LABELS. (B) HITS DE HISTOGRAMA.

Na Tabela C.6 podemos verificar quais genes pertencem a cada um dos grupos identificados, o grupo um tem 97 genes e no grupo dois 03 genes.

TABELA C. 6 - GENES IDENTIFICADOS NO EXPERIMENTO 8

Grupo	**Nome do gene**	**Quantidade de genes**
Grupo 1	**2-component transcriptional regulator**	**04**
	putative 2-component transcriptional regulator	08
	transcriptional regulatory protein zraR	02
	sigma-54 dependent response regulator	05
	two-component response regulator	01
	DNA-binding response regulator in two-component system	04
	conserved hypothetical protein	02

	hypothetical protein ECNA114_2627	01
	sigma-54 interaction domain-containing protein	03
	putative DNA-binding response regulator in two-component system	02
	putative C4-dicarboxylate transport transcriptional regulatory protein DctD	03
	hypothetical protein ECP_2556	01
	hypothetical protein c3077	01
	hypothetical protein UTI89_C2873	01
	nitrogen regulator I homolog	01
	Two-component system response regulator	12
	C4-dicarboxylate transport transcriptional regulatory protein DctD	04
	hypothetical protein	01

	CKO_00233	
	transcriptional regulator	03
	sigma-54 dependent transcriptional regulator/response regulator	01
	putative transcriptional regulator	01
	hypothetical protein SARI_00314	01
	two component, sigma54 specific, Fis family transcriptional regulator	03
	putative two component, sigma54 specific, transcriptional regulator	01
	Response regulator containing CheY-like receiver, AAA-type ATPase, and DNA-binding domains	01
	two-component system, NtrC family, response regulator YfhA	01

	DNA-binding response regulator HsfA	01
	response regulator GlrR	03
	sigma-54 dependent DNA-binding response regulator	01
	putative two component, sigma54 specific, transcriptional regulator, Fis family	05
	hypothetical protein ESA_00708	01
	hypothetical protein plu3311	01
	two-component transcriptional regulator	01
	putative Fis family two component sigma-54 specific transcriptional regulator	01
	transcriptional regulatory protein	01
	two component, sigma54 specific, transcriptional regulator	01

	response regulator in two-component regulatory system (EBP family)	02
	two component sigma-54 specific Fis family transcriptional regulator	01
	hypothetical protein yruck0001_6810	01
	two component YfhA	01
	hypothetical protein yaldo0001_25090	01
	hypothetical protein yfred0001_36530	01
	hypothetical protein yrohd0001_20800	01
	hypothetical protein yberc0001_30440	01
	2-component transcriptional regulator yfhA	01
	hypothetical protein ymoll0001_29720	01
	two-component system, NtrC family, response regulator	01
	DNA-binding reponse regulator	01

	HsfA	
Grupo 2	yfhA protein	03

Experimento 9

Neste experimento foi identificado a presença de três grupos conforme mostra a Matriz-U referente a este processo (figura C.7).

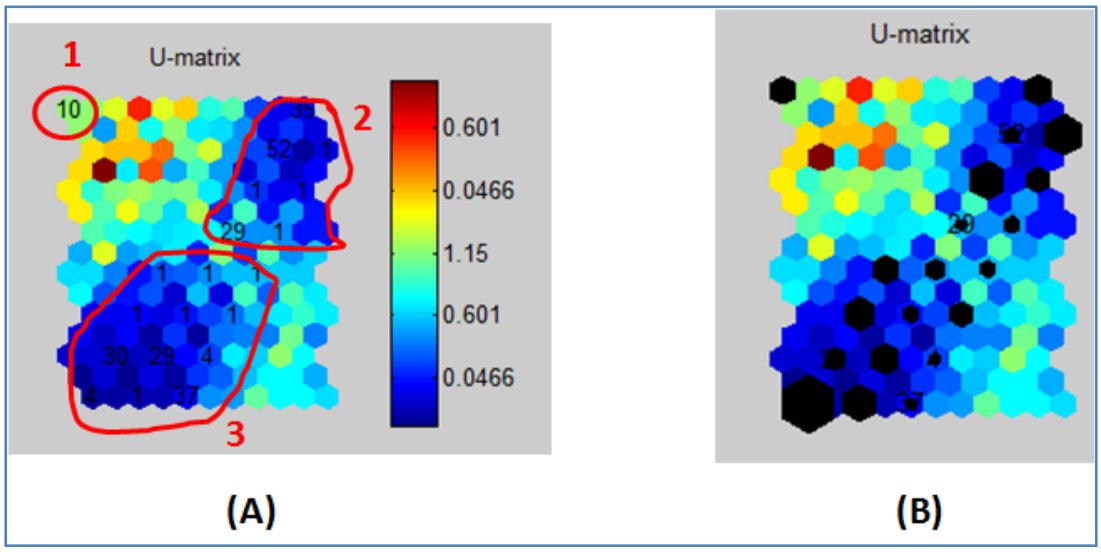


FIGURA C. 7 - CAPTURA DE TELA DO EXPERIMENTO 9 DA VISUALIZAÇÃO DA MATRIZ-U. (A) MATRIZ-U COM OS RESPECTIVOS LABELS. (B) HITS DE HISTOGRAMA.

Na Tabela C.7 podemos verificar quais genes pertencem a cada um dos grupos identificados, o grupo um tem 05 genes, o grupo dois 32 genes e o grupo três 63 genes.

TABELA C. 7 - GENES IDENTIFICADOS NO EXPERIMENTO 9

Grupo	**Nome do gene**	**Quantidade de genes**

Grupo 1	ribonucleotide reductase	05
Grupo 2	ribonucleotide-diphosphate reductase subunit beta	14
	ribonucleoside-diphosphate reductase, beta subunit	02
	Chain A, Ribonucleoside-Diphosphate Reductase 1 Beta Chain	01
	ribonucleoside-diphosphate reductase 1 subunit beta	03
	Hypothetical protein	01
	Ribonucleoside-diphosphate reductase	04
	ribonucleotide reductase of class Ia (aerobic) subunit beta	01
	Chain A, Ribonucleotide Reductase R2 Subunit From E. Coli	01
	Chain A, Y122f	01

	Mutant Of Ribonucleotide Reductase From Escherichia Coli	
	ribonucleoside-diphosphate reductase 1 beta chain	01
	ribonucleotide reductase, beta subunit 01	01
	Chain A, Crystal Structure Of Manganese Substituted R2-D84e (D84e Mutant Of The R2 Subunit Of E. Coli Ribonucleotide Reductase)	01
	putative ribonucleoside diphosphate reductase 1, beta subunit, ferritin-like protein	01
Grupo 3	ribonucleotide-diphosphate reductase subunit beta	26
	ribonucleoside-diphosphate reductase, beta subunit	17
	ferritin-like protein	**01**

	Chain A, Ribonucleoside-Diphosphate Reductase 1 Beta Chain	01
	Chain A, Crystal Structure Of The Y122h Mutant Of Ribonucleotide Reductase R2 Protein From E. Coli	01
	Chain A, Ribonucleotide Reductase R2 Soaked With Ferrous Ion At Neutral Ph	01
	Chain A, Ribonucleotide Reductase Y122no2y Modified R2 Subunit Of E. Coli	01
	ribonucleoside-diphosphate reductase 1 subunit beta	03
	Chain B, Ribonucleoside-Diphosphate Reductase 1 Beta Chain Mutant E238a	01
	Chain A, Dithionite Reduced E. Coli Ribonucleotide	01

	Reductase R2 Subunit, D84e Mutant	
	Chain A, Azide Complex Of The Diferrous F208a Mutant R2 Subunit Of Ribonucleotide Reductase	01
	Chain A, Substitution Of Manganese For Iron In Ribonucleotide Reductase From Escherichia Coli. Spectroscopic And Crystallographic Characterization	01
	protein	01
	Chain A, Ribonucleotide Reductase R2-D84eW48F MUTANT SOAKED WITH Ferrous Ions At Neutral Ph	01
	Chain A, Autocatalytic Generation Of Dopa In The Engineered Protein R2 F208y From Escherichia Coli Ribonucleotide Reductase And Crystal Structure Of The Dopa-208	01

	Protein	
	Ribonucleotide reductase of class 1a beta subunit	01
	Ribonucleotide reductase of class la (aerobic), beta subunit	02
	ribonucleoside-diphosphate reductase	01
	ribonucleoside-diphosphate reductase class la beta subunit	01

Experimento 10

Neste experimento foi identificado a presença de 4 grupos conforme mostra a Matriz-U referente a este processo (figura C.8).

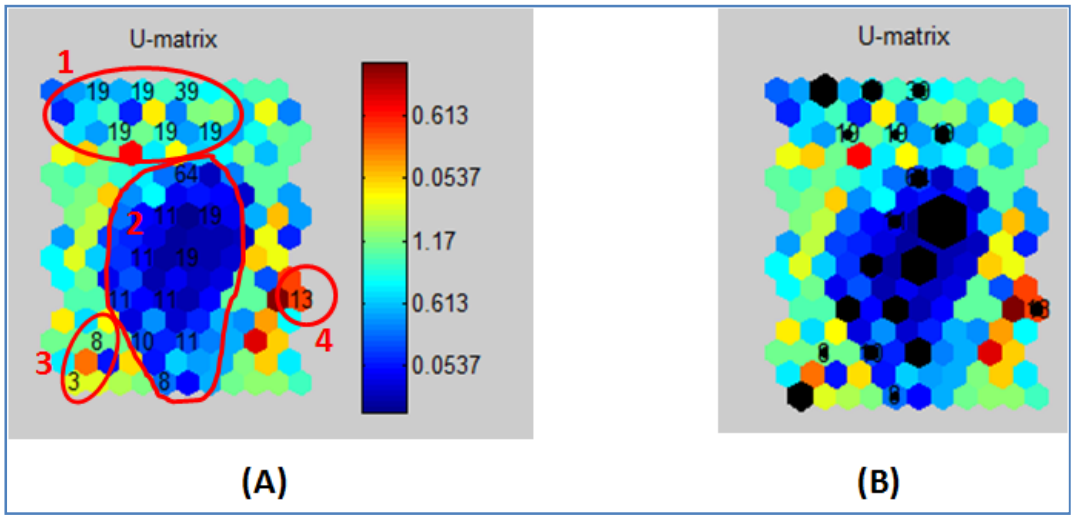


FIGURA C. 8 - CAPTURA DE TELA DO EXPERIMENTO 10 DA VISUALIZAÇÃO DA MATRIZ-U. (A) MATRIZ-U COM OS RESPECTIVOS LABELS. (B) HITS DE HISTOGRAMA.

Na Tabela C.8 podemos verificar quais genes pertencem a cada um dos grupos identificados, o grupo um tem 20 genes, o grupo dois 71 genes, o grupo três 07 genes e o grupo quatro 02 genes.

TABELA C. 8 - GENES IDENTIFICADOS NO EXPERIMENTO 10

Grupo	**Nome do gene**	**Quantidade de genes**
Grupo 1	Two-component response regulator	01
	response regulator	14
	two-component system response regulator	01
	response regulator receiver modulated metal dependent phosphohydrolase	**04**
Grupo 2	two-component response regulator	02
	response regulator receiver protein	07
	COG3437: Response regulator containing a CheY-like receiver domain and an HD-GYP domain	01

	response regulator receiver modulated metal dependent phosphohydrolase	24
	Response regulator receiver:Metal-dependent phosphohydrolase, HD subdomain	03
	response regulator	15
	regulatory components of sensory transduction system	03
	two component system, transcriptional regulatory protein	02
	pole remodelling regulatory diguanylate cyclase	07
	response regulator rpfG	03
	metal-dependent phosphohydrolase HD sub domain-containing protein	01
	signal transduction histidine-protein kinase BarA	02
	Response regulator containing a CheY-like receiver domain and an HD-GYP	01

	domain	
Grupo 3	cyclic di-GMP phosphodiesterase	01
	putative two-component response regulator	02
	hypothetical protein HMPREF1030_01984	01
	probable two-component response regulator	01
	two-component response regulator	01
	response regulator receiver	01
Grupo 4	hypothetical protein Csp_C22930	01
	HD domain protein	01