UNIVERSIDADE FEDERAL DO PARANÁ

MICHELLY ALVES COUTINHO GEHLEN

MAPEAMENTO DE GENES *nif* PUBLICADOS NO NCBI USANDO CONCEITOS DE MINERAÇÃO DE DADOS E INTELIGÊNCIA ARTIFICIAL

CURITIBA
2011

MICHELLY ALVES COUTINHO GEHLEN

MAPEAMENTO DE GENES *nif* PUBLICADOS NO NCBI USANDO CONCEITOS DE MINERAÇÃO DE DADOS E INTELIGÊNCIA ARTIFICIAL

Dissertação apresentada ao Curso de Pós-Graduação em Bioinformática da Universidade Federal do Paraná, como requisito parcial para a obtenção do título de Mestre em Bioinformática.

Orientador : Roberto Tadeu Raittz, Dr.
Co-orientadora: Liu Un Rigo, Dra.

CURITIBA
2011

TERMO DE APROVAÇÃO

MICHELLY ALVES COUTINHO GEHLEN

MAPEAMENTO DE GENES *nif* PUBLICADOS NO NCBI USANDO CONCEITOS DE MINERAÇÃO DE DADOS E INTELIGÊNCIA ARTIFICIAL

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, Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientador : Prof. Dr. Roberto Tadeu Raittz

Co-orientador : Profª. Drª. Liu Un Rigo

Prof. Dr. João Artur de Souza
Universidade Federal de Santa Catarina

Profª. Drª. Maria Berenice Reynaud Steffens
Universidade Federal do Paraná

Prof. Dr. Roberto Tadeu Raittz
Universidade Federal do Paraná

Profª. Drª. Liu Un Rigo
Universidade Federal do Paraná

Curitiba, 17 de fevereiro de 2011.

Para Téco, Lucas, Tiago, Pai e Mãe.

Amo vocês.

AGRADECIMENTOS

À Deus, pela oportunidade de crescimento e aprimoramento, de vida e de fé.

A Mãezinha do Céu, por ficar sempre do meu lado, me carregando no colo, cada uma das muitas vezes em que eu caí.

Ao meu marido, Téco, pelo amor, demonstrado no silêncio do olhar e nos gestos concretos de apoio sempre presentes.

Aos meus Filhos, Lucas e Tiago, por existirem e serem a minha força e razão de tudo.

Aos meus Pais, Moacir e Soeli, pelo carinho e puxões de orelha.

À minha irmã Giselly, meu cunhado Fábio e meu afilhado Dudu pelo bom humor e descontração durante todo o decorrer do mestrado.

Às minhas Avós, Maria e Santina e ao meu tio Sérgio, sempre especial para mim, por todas as orações realizadas.

Aos meus orientadores, Roberto e Liu, pela amizade. Sabemos o quão difícil foi chegar até aqui, mas conseguimos. Juntos.

Ao programa de Pós Graduação em Bioinformática pela oportunidade.

As professoras Jeroniza e Berenice pelo apoio constante e paciência quando de meus desabafos.

Ao professor Fábio pelo auxílio sempre pontual, conselhos e atenção.

Ao professor Leonardo pelo brilhante tema proposto.

Ao professor Emanuel por suas colaborações e correções.

Ao professor Lucas pela sua amizade, conselhos, bom humor e apoio todos os dias no laboratório.

Aos meus “irmãos” Dieval e Leandro pela amizade, paciência, carinho, auxílio e apoio em todos os momentos, dentro e fora do ambiente do mestrado.

Ao Du pela amizade, paciência, otimismo, auxílio e bom humor.

Ao Luiz pela amizade e pelos “spams” que inúmeras vezes melhoraram os meus dias.

Ao Rodrigo por todos os conselhos e por ser minha voz cada uma das vezes que as palavras não saíram.

À Vanely pela amizade e por nossas longas conversas no laboratório.

À Suzana pelo apoio e orações, principalmente nos momentos em que eu mais precisei.

Aos colegas de mestrado Ademir, Daniela, Rosa, Terumi, Waldemar, Waldir, Lucas, Sérgio, Juliana e Danhylo pelas conversas, contribuições e pelo bom convívio sempre.

Ao Instituto Nacional em Ciência e Tecnologia em Fixação Biológica de Nitrogênio.

Ao National Center for Biotechnology Information.

Muito obrigada !!!!

“A inteligência é a insolência educada.”

Aristóteles

RESUMO

A Fixação Biológica de Nitrogênio é um importante processo biosustentável, aplicado na agricultura sob a forma de biofertilizantes. Na FBN ocorre a redução do dinitrigênio gasoso (N_2) à amônia (NH_4), mediante catalização realizada pelo Complexo da Nitrogenase. Este Complexo é codificado basicamente por um agrupamento (*cluster*) de genes conhecidos como *nif*. Esta pesquisa propõe, através de técnicas e aplicações em Bioinformática, o desenvolvimento de uma metodologia para mineração de dados relacionados a genes *nifHDKEN* publicamente disponibilizadas pelo Centro Nacional para Informações de Biotecnologia (NCBI), através do Banco de Dados GenBank®; a classificação automática das informações através da implementação e uso de uma rede neural artificial modelo FAN; o mapeamento relacional entre genes *nif* encontrados e seus respectivos organismos e o descritivo de um paralelo entre a literatura de referência e os resultados encontrados. A automatização dos processos foi realizada mediante a criação de programas (*scripts*) em linguagem de programação Python, versão, 2.6, suplantada pela biblioteca de algoritmos utilitários para Bioinformática, BioPython, versão, 1.5.2. O processo de coleta e mineração de dados baseou-se em resultados obtidos através da execução *online* da ferramenta NCBI BLASTP. Sobre os dados coletados foram aplicadas técnicas para extração de características para posterior classificação das informações via rede neural artificial. Para a maximização dos resultados referentes ao processo de aprendizagem da rede neural, aplicou-se ainda sobre os dados, a técnica de co-aprendizado supervisionado. Os dados classificados e mapeados foram submetidos a uma pós-análise, visando ancorar as informações adquiridas com a literatura referência da área. Até a data de 25/11/2010 foram encontrados e classificados 14988 registros referentes a anotações de sequências protéicas relacionadas a genes *nifHDKEN,* agrupados em 2125 organismos diferentes, considerando-se as estirpes dos mesmos. Sem considerar as estirpes dos organismos foram relacionados 646 organismos diferentes, contendo pelo menos um gene *nif* seqüenciado, anotado, depositado e disponibilizado no NCBI GenBank. Considerando-se os 1425 genomas completos já depositados no NCBI GenBank em 28/01/2011, 14,03% apresentam em sua anotação, pelo menos um gene *nif*.

Palavras-chave: Bioinformática, Fixação Biológica de Nitrogênio, Mineração de Dados, Redes Neurais Artificiais.

ABSTRACT

The Biological Nitrogen Fixation is an important bio-sustainable process, applied in agriculture in the form of bio-fertilizers. The BNF corrresponds to the reduction of dinitrigen gas (N2) to ammonia (NH4) by catalyzing performed by the Complex of Nitrogenase, encoded primarily by a cluster of genes known as *nif*. This research proposes by means of techniques and applications in bioinformatics, the development of a methodology for data mining-related genes *nifHDKEN* publicly available by GenBank ®; the automatic classification of information through the implementation and use of an artificial neural network model FAN; relational mapping between *nif* genes found and their organisms and the description of a parallel between the reference literature and the results. The automation process was accomplished by creating scripts in the Python programming language, version 2.6, supplanted by the library of algorithms for bioinformatics tools, BioPython, version 1.5.2. The collection process and data mining was based on results obtained by running the tool online NCBI BLASTP. About the data collected were applied techniques for extracting features for subsequent classification of information via artificial neural network. To maximize the results for the learning process of the neural network was applied also on the data, the technique of co-supervised learning. After data classification and mapping the data were subjected to a post-analysis in order to anchor the information with reference to the literature of the area. As of the date of 25.11.2010 were found and classified records pertaining to 14,988 notes from protein sequences related to genes *nifHDKEN*, grouped into 2125 different organisms, considering the strains of the same. Without considering the strains of organisms were listed 646 different organisms, containing at least one *nif* gene sequenced, annotated and deposited in NCBI GenBank and available. Considering the 1425 complete genomes already deposited in the NCBI GenBank on 28.01.2011, 14.03% are in your note, at least one gene nif.

Keywords: Bioinformatics, Biological Nitrogen Fixation, Data Mining, Artificial Neural Networks.

LISTA DE ILUSTRAÇÕES

FIGURA 1 - O USO DE COMPUTADORES PARA PROCESSAMENTO DA INFORMAÇÃO BIOLÓGICA 018
FIGURA 2 - DISPOSIÇÃO DOS GENES *nif* EM *Klebsiella pneumoniae* 023
FIGURA 3 - ESPÉCIES COM MAIOR NÚMERO DE SEQUÊNCIAS DEPOSITADAS NO GENBANK 028
FIGURA 4 - O PROCESSO DE KDD..... 035
FIGURA 5 - O NEURÔNIO BIOLÓGICO 044
FIGURA 6 - MODELO DE UM NEURÔNIO ARTIFICIAL 044
FIGURA 7 - EXEMPLO DE REDE NEURAL..... 045
FIGURA 8 - PARÁMETROS AJUSTADOS PARA EXECUÇÃO AUTOMÁTICA DA FERRAMENTA BLASTP 050
FIGURA 9 - SCRIPT PYTHON PARA CONTAGEM DE OCORRÊNCIAS DE AMINOÁCIDOS EM SEQUÊNCIAS PROTÉICAS CODIFICADAS POR GENES *nif* 056
FIGURA 10 - TRECHO DE UM ARQUIVO TEXTO SIMPLES RETORNADO PELA FERRAMENTA BLASTP 058
FIGURA 11 - EXEMPLO DA TÉCNICA APLICADA PARA A GERAÇÃO DAS MATRIZES DE CO-OCORRÊNCIA..... 061
FIGURA 12 - METODOLOGIA BASEADA EM RESULTADOS BLAST, PARA MINERAÇÃO DE DADOS DO NCBI GENBANK 065
FIGURA 13 - TRECHO DE CÓDIGO CORRESPONDENTE À GERAÇÃO DAS CARACTERÍSTICAS DESCRITAS NO GRUPO I – FÍSICO-QUÍMICAS 084
FIGURA 14 - TRECHO DE CÓDIGO CORRESPONDENTE À GERAÇÃO DA CARACTERÍSTICA RELACIONADA À EXISTÊNCIA DE DOMÍNIOS CONSERVADOS DA NITROGENASE..... 085
FIGURA 15 - TRECHO DE CÓDIGO CORRESPONDENTE À GERAÇÃO DAS CARACTERÍSTICAS DESCRITAS NO GRUPO II – INFERIDAS..... 086
FIGURA 16 - TRECHO DE CÓDIGO CORRESPONDENTE À GERAÇÃO DAS CARACTERÍSTICAS DESCRITAS NO GRUPO III – BLAST 087
FIGURA 17 - CÓDIGO FONTE CORRESPONDENTE À FUNÇÃO DE GERAÇÃO DAS CARACTERÍSTICAS DE ENERGIA E ENTROPIA, BASEADAS NOS DESCRITORES DE HARALICK (HARALICK, 1979) 088
FIGURA 18 - DISTRIBUIÇÃO DOS REGISTROS CLASSIFICADOS COMO GENES nif ENTRE OS GENES DE INTERESSE DE ESTUDO..... 091
FIGURA 19 - DISTRIBUIÇÃO DOS ORGANISMOS DETENTORES DE GENES nif ENCONTRADOS NO REINO ARCHAEA..... 093
FIGURA 20 - DISTRIBUIÇÃO DOS ORGANISMOS DETENTORES DE GENES nif ENCONTRADOS NO REINO BACTÉRIA 094
FIGURA 21 - DETALHAMENTO DO UNIVERSO COMPONENTE DO GRUPO DAS PROTEOBACTÉRIAS..... 095
FIGURA 22 - AFERIÇÃO DA RELAÇÃO TEMPO x RESULTADOS..... 097

QUADRO 1 - DIVISÃO DOS CONJUNTOS DE DADOS PARA UTILIZAÇÃO NO APRENDIZADO DA REDE FAN..... 089
QUADRO 2 - MEDIDAS DE ACERTIVIDADE DA REDE FAN ENTRE OS TRÊS CONJUNTOS DE DADOS 090
QUADRO 3 - RESULTADO QUANTITATIVO, REFERENTE Á CLASSIFICAÇÃO FINAL DOS DADOS, REALIZADA PELA REDE FAN 090
QUADRO 4 - DISPERSÃO DOS REGISTROS CLASSIFICADOS COMO GENES *nif* ENTRE O UNIVERSO DE GENES ESTUDADOS..... 090
QUADRO 5 - SUMARIZAÇÃO DOS DADOS APRESENTADOS NOS MAPAS DE RELACIONAMENTO ENTRE ORGANISMOS E GENES *nif* , PARA GENOMAS COMPLETOS E INCOMPLETOS..... 092
QUADRO 6 - DIVISÃO DOS CONJUNTOS DE DADOS PARA UTILIZAÇÃO NO PRIMEIRO CICLO DE APRENDIZADO DA REDE FAN 100
QUADRO 7 - DIVISÃO DOS CONJUNTOS DE DADOS PARA UTILIZAÇÃO NO SEGUNDO CICLO DE APRENDIZADO DA REDE FAN..... 101
QUADRO 8 - COMPARATIVO ENTRE A REDE NEURAL FAN E DIFERENTES LINHAS DE CORTE..... 105
QUADRO 9 - COMPARATIVO ENTRE AS REDES NEURAIS FAN E MLP..... 106
QUADRO 10 - SUMARIZAÇÃO DE RESULTADOS DA COMPARAÇÃO ENTRE OS REGISTROS ENCONTRADOS E O TRABALHO PUBLICADO POR YOUNG (YOUNG, 1992)..... 107

LISTA DE TABELAS

TABELA 1 - PRODUTOS DOS GENES *NIF* E SUAS FUNÇÕES (CONHECIDAS OU PROPOSTAS) NO PROCESSO DE FIXAÇÃO DE NITROGÊNIO 024
TABELA 2 - CARACTERÍSTICAS EXTRAÍDAS COM BASE NAS SEQUÊNCIAS DE AMINOÁCIDOS, DIVIDIDAS POR GRUPOS DE INTERESSE 052
TABELA 3 - GRUPOS E PESOS DE AMINOÁCIDOS 053
TABELA 4 - ASSINATURAS DE DOMÍNIOS CONSERVADOS UTILIZADAS 055
TABELA 5 - GENES *NIF* ENCONTRADOS EM PESQUISA MANUAL REALIZADA 066
TABELA 6 - MAPEAMENTO TAXONÔMICO DOS GENES *NIF* ENCONTRADOS EM PESQUISA MANUAL 067
TABELA 7 - SEQUÊNCIAS SELECIONADAS PARA A CONTINUIDADE DO PROCESSO DE MINERAÇÃO DE DADOS 068
TABELA 8 - DESCRIÇÃO DOS CAMPOS CONSTANTES NOS ARQUIVOS IDsValuesBLAST.txt 079
TABELA 9 - AFERIÇÃO DA RELAÇÃO TEMPO X RESULTADOS 096
TABELA 10 - AFERIÇÃO DA RELAÇÃO TEMPO X RESULTADOS – DETALHAMENTO POR GENES nif EM ESTUDO 098
TABELA 11 - RESULTADOS OBTIDOS DURANTE OS TESTES DE GRUPOS DE CARACTERÍSTICAS 099
TABELA 12 - TABELA DE PROGRESSÃO DO PERCENTUAL REAL DE ACERTOS DA REDE, BASEADOS NA TÉCNICA DE CO-APRENDIZADO 102

LISTA DE ABREVIATURAS E SIGLAS

DNA	-	Ácido desoxirribonucléico
NCBI	-	National Center for Biotechnology Information
RNA	-	Ácido ribonucléico
FeMoCo	-	Cofator Ferro Molibdênio
NLM	-	National Library of Medicine
NIH	-	US National Institutes of Health
EMBL	-	European Molecular Biology Laboratory Nucleotide Sequence Database
DDBJ	-	DNA Databank of Japan
WGS	-	Whole Genome Shotgun
BLAST	-	Basic Local Alignment Search Tool
KDD	-	Knowledge Discovery in Databases
HMM	-	Hidden Markov Models
Kg N/ha/ano	-	Kilograma de Nitrogênio por Hectare por Ano
FBN	-	Fixação Biológica de Nitrogênio
ATP	-	Adenosina Trifosfato
MoFe	-	Proteína Molibdênio-Ferro
Fe-S	-	Grupamento Ferro-Enxofre
NifB-co	-	Cofator relacionado à proteína NifB
MSP	-	Maximal Segment Pair
HSPs	-	High Scoring Pairs
PAM	-	Point Accepted Mutation
BLOSUM	-	BLOcks SUbstitution Matrix
SQL	-	Structured Query Language
MATLAB	-	Matrix Laboratory
IUPAC	-	International Union of Pure and Applied Chemistry
GRAVY	-	Grand Average of Hydropathy
ExPASy	-	Expert Protein Analysis System
SIB	-	Swiss Institute of Bioinformatics
CDD	-	Conserved Domains Database
FAN	-	Free Associative Neurons
FeMoCo	-	Cofator Ferro Molibdênio

LISTA DE SÍMBOLOS

N_2	-	Dinitrogênio Gasoso
ºC	-	Graus Celsius
$(NH_4)^+$	-	Íon amônico
$\alpha_2\beta_2$	-	Heterotetrâmero da dinitrogenase, onde α = proteína NifD e β = proteína NifK
γ_2	-	Homodímero da dinitrogenase reductase ou proteína NifH
$MoFe_7S_9$	-	Centro de metálico denominado cofator Ferro Molibdênio
4Fe-4S	-	Centro 4 Ferro-4 Enxofre
$\alpha_2\beta_2\delta_2$	-	Hexâmero composto pelo heterotetrâmero da dinitrogenase aliado a dinitrogenases alternativas codificadas pelos genes *vnfDKG* ou *anfDKG*, respectivamente.
®	-	Marca Registrada

SUMÁRIO

1. INTRODUÇÃO 017

1.1. BIOINFORMÁTICA ... 017
1.2. FIXAÇÃO BIOLÓGICA DE NITROGÊNIO ... 019
1.3. ORGANISMOS DIAZOTRÓFICOS ... 020
1.4. PERFIL GÊNICO DA FIXAÇÃO DE NITROGÊNIO ... 021
1.5. O CLUSTER *nif* ... 022
1.6. NCBI GENBANK ... 026
1.7. BLAST ... 028
1.8. ASSINATURAS DE DOMÍNIOS CONSERVADOS ... 031
1.9. MINERAÇÃO DE DADOS ... 033
1.10. EXTRAÇÃO DE CARACTERÍSTICAS ... 036
1.11. RECONHECIMENTO DE PADRÕES ... 039
1.12. REDES NEURAIS ARTIFICIAIS ... 042

2. JUSTIFICATIVA 047

2.1. OBJETIVOS ... 048
2.1.1. OBJETIVO GERAL ... 048
2.1.2. OBJETIVOS ESPECÍFICOS ... 048

3. METODOLOGIA 049

3.1 PYTHON ... 049
3.2. BIOPYTHON ... 049
3.3. BANCO DE DADOS NCBI GENBANK PROTEIN ... 050
3.4. BLASTP ... 050
3.5. MS EXCEL ... 051
3.6. MATLAB ... 051
3.7. CARACTERÍSTICAS CONTEMPLADAS ... 051
3.7.1. GRUPO I - FISICO-QUÍMICAS ... 052
3.7.1.1. Peso Molecular Normalizado ... 052
3.7.1.2. Ponto Isoelétrico ... 054
3.7.1.3. Percentual de Aromaticidade ... 054
3.7.1.4. Índice de Instabilidade ... 054
3.7.1.5. Índice de Hidropatia (*GRAVY INDEX*) ... 054
3.7.1.6. Assinaturas de Domínios Conservados ... 055
3.7.2. GRUPO II – INFERIDAS ... 055
3.7.3. GRUPO III – BLAST ... 057
3.7.4. GRUPO IV – MATRIZES DE CO-OCORRÊNCIA ... 059
3.8. PROCESSO DE CLASSIFICAÇÃO DOS DADOS ... 062
3.9. REDE NEURAL ARTIFICIAL MODELO FREE ASSOCIATIVE NEURONS (FAN) ... 062
3.10. SOFTWARE EASYFAN ... 063
3.11. PROCESSO DE CO-APRENDIZADO DA REDE NEURAL ARTIFICIAL ... 063

4. RESULTADOS E DISCUSSÃO 065

4.1. BUSCA MANUAL POR SEQUENCIAS DE GENES *nif* ... 066
4.2. SELEÇÃO DE DADOS, DIVISÃO E FORMAÇÃO DE GRUPOS ... 066
4.3. EXECUÇÃO AUTOMÁTICA DA FERRAMENTA BLASTP ... 077
4.4. INTERPRETAÇÃO E TABULAÇÃO AUTOMÁTICA DE DADOS ... 078
4.4.1. *PARSING* DE DADOS ... 079
4.4.2. GERAÇÃO DE TABELAS MESTRE DE DADOS ... 080
4.4.3. CONSTRUÇÃO DE TABELAS ÍNDICE ... 081
4.4.4. INCORPORAÇÃO DE DADOS ... 081
4.5. PROCESSO DE UNIÃO E LIMPEZA DE DADOS ... 082
4.5.1. UNIÃO DE DADOS ... 082

4.5.2. LIMPEZA DOS DADOS 083
4.6. EXTRAÇÃO DE CARACTERÍSTICAS 083
4.6.1. GRUPO I - FISICO-QUÍMICAS 084
4.6.1.1. Assinaturas de Domínios Conservados 084
4.6.2. GRUPO II – INFERIDAS 085
4.6.3. GRUPO III – BLAST 086
4.6.4. GRUPO IV – MATRIZES DE CO-OCORRÊNCIA 087
4.7. APRENDIZADO DA REDE NEURAL ARTIFICIAL 089
4.8. MAPEAMENTO ORGANISMOS X GENES *nifHDKEN* 091
4.9. TESTES DE REPETIBILIDADE DO PROCESSO DE MINERAÇÃO DE DADOS 096
4.10. AJUSTES NO PROCESSO DE SELEÇÃO DE CARACTERÍSTICAS 098
4.11. A IMPORTÂNCIA DO CO-APRENDIZADO EM REDES NEURAIS ARTIFICIAIS 100
4.12. POR QUE MINHA REDE AINDA APRESENTA ERROS? 103
4.13. REDE NEURAL *VERSUS* LINHA DE CORTE 104
4.14. POR QUE UTILIZAR A REDE FAN? 105
4.15. RESULTADOS X LITERATURA 106
4.16. APLICABILIDADE DO FERRAMENTAL DESENVOLVIDO 107

6. CONCLUSÕES 109

7. PERSPECTIVAS FUTURAS 111

REFERÊNCIAS 113

APÊNDICES 125

ANEXOS 199

1. INTRODUÇÃO

1.1. BIOINFORMÁTICA

Bioinformática é a ciência de gerenciar, analisar, extrair e interpretar informações a partir de sequências biológicas e moleculares. Tem sido uma área ativa de pesquisa desde o final de 1980. Depois da conclusão do projeto do genoma humano em abril de 2003, a área ganhou maior destaque, e, com a realização de mais projetos de sequenciamento de genomas, os dados disponíveis, como sequências de DNA, sequências de proteínas e estruturas de proteínas passaram a ter um crescimento exponencial. Diante desta enorme quantidade de dados, o biólogo não pode mais simplesmente fazer uso das técnicas tradicionais em biologia para a realização de análises. Ao invés disso, a fim de compreender o mistério da vida, o uso das tecnologias da informação passaram a ser essenciais (HSU, 2006; p. vi).

Para o National Center for Biotechnology Information (NCBI), a Bioinformática é o campo da ciência em que a biologia, ciência da computação e tecnologia da informação se fundem para formar uma única disciplina. O objetivo final do campo é o de permitir a descoberta de novos conhecimentos biológicos, bem como para criar uma perspectiva global de que princípios unificadores da biologia possam ser discernidos (NCBI, 2010).

Para Fox, a Bioinformática trata da conversão de observações biológicas em modelos computacionais processáveis (FOX, 2009), enquanto para Hughey e colaboradores, pode ser definida como a aplicação de técnicas computacionais para manipular dados biológicos (HUGHEY *et al,* 2001). Já Gibas define Bioinformática como sendo a Biologia Computacional, aplicando técnicas quantitativas e analíticas à modelação de sistemas biológicos (GIBAS *et al*, 2001).

A Bioinformática é interdisciplinar sendo que seus domínios permeiam vários campos do conhecimento como Biologia, Medicina, Matemática, Física, Ciência da Computação e Estatística (BAYAT, 2002).

O início da Bioinformática pode ser rastreado até Margaret Dayhoff em 1968 e sua coleção de sequências de proteínas conhecida como o Atlas de Sequências de Proteínas e Estruturas (DAYHOFF, 1968 *apud* FOX, 2009) e seu crescimento é paralelo ao desenvolvimento da tecnologia de sequenciamento de DNA. Da mesma

forma que o desenvolvimento do microscópio, em finais de 1600, revolucionou as ciências biológicas, a tecnologia de sequenciamento do DNA revolucionou o campo da Bioinformática (FOX, 2009).

Duas importantes atividades de larga escala que utilizam a Bioinformática são a genômica e a proteômica. A genômica se refere ao sequenciamento e análise do conjunto completo de sequências de DNA que codificam o material hereditário (como genes e transcritos de RNA) passado de geração para geração enquanto a proteômica trata da análise do conjunto completo de proteínas ou proteoma que os genes expressam. Somadas à genômica e a proteômica existem muitas outras áreas da biologia onde a Bioinformática é aplicada, como por exemplo a metabolômica e a transcriptômica (FOX 2009).

Segundo o NCBI (NCBI, 2010), a "revolução genômica", implicou na criação e manutenção de grandes bancos de dados para armazenar informações biológicas, tais como sequências de nucleotídeos e aminoácidos, envolvendo também o desenvolvimento de interfaces complexas, em que pesquisadores poderiam tanto ter acesso aos dados existentes, como também apresentar dados novos ou revisados. A Figura 1 apresenta um esquema gráfico abordando o uso de computadores para processamento da informação biológica. A riqueza de informações de sequenciamento do genoma tem exigido o projeto de software e o uso de computadores para processar essa informação (FOX, 2009).

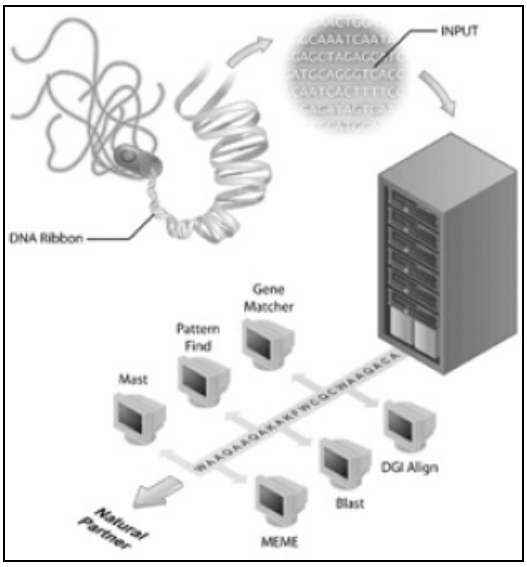


FIGURA 1 - O USO DE COMPUTADORES PARA PROCESSAMENTO DA INFORMAÇÃO BIOLÓGICA.
FONTE: Extraído de Fox, 2009 (FOX, 2009).

A justificativa para a aplicação de abordagens computacionais de modo a facilitar a compreensão dos vários processos biológicos inclui, além de uma perspectiva mais global em projeto experimental, a capacidade de capitalizar as tecnologias emergentes de mineração de dados - o processo pelo qual são geradas hipóteses testáveis sobre a função ou estrutura de um gene ou a proteína de interesse, identificação de seqüências semelhantes em organismos mais bem caracterizadas. Desta forma, a biologia no século 21 vem se transformando de uma ciência puramente baseada em laboratório em uma ciência da informação (NCBI, 2010).

O uso desenfreado de fertilizantes químicos nitrogenados utilizados nas lavouras resulta em grandes níveis de poluição ambiental. Desta maneira, o processo de FBN surge como alternativa ao uso desses fertilizantes químicos como uma forma renovável e não poluente de disponibilização de nitrogênio para agricultura (ZAHRAN, 1999 *apud* OSAKI, 2009).

1.2. FIXAÇÃO BIOLÓGICA DE NITROGÊNIO

O nitrogênio é um elemento químico estável, inerte, componente de moléculas de ácidos nucléicos, proteínas e polissacarídeos entre outras, e que compreende 78% da atmosfera da Terra (SUR *et al*, 2010). Apesar de sua abundância apenas uma pequena porcentagem é capaz de ser aproveitada pelo ciclo biogeoquímico, dada sua forma molecular (N_2) ser quimicamente muito estável (STEVENSON, 1972). Apenas 0,04% do nitrogênio disponível na biosfera ocorre em compostos que são acessíveis aos seres vivos (ROSSWALL, 1976). O nitrogênio disponível para uso dos seres vivos é também chamado de nitrogênio fixado (RAYMOND *et al*, 2004).

Fixação de nitrogênio é o processo químico ou biológico pelo qual o dinitrogênio gasoso (N_2) é reduzido à amônia (NH_4) (RAMOS, 2003).

A síntese química da amônia é utilizada para a produção de fertilizantes nitrogenados para a agricultura, sendo um processo de custo elevado devido à necessidade de condições especiais, como pressão de aproximadamente 200 atmosferas e temperaturas entre 400 e 600ºC (BOTHE et al., 1983). Apenas 25% do nitrogênio fixado na Terra decorre de processos industriais. A fixação biológica de nitrogênio contribui com 65% de todo o nitrogênio fixado, enquanto fenômenos

naturais, como a ação de descargas elétricas, radiação ultravioleta e a ação vulcânica, correspondem a 10% (HUNGRIA E CAMPO, 2005; MOREIRA E SIQUEIRA, 2006)

A Fixação Biológica de Nitrogênio foi inicialmente descrita em bactérias diazotróficas da rizosfera e do rizoplano de uma grande variedade de plantas não-leguminosas (DÖBEREINER, 1992) e, além de possuir importante papel na manutenção do ciclo do nitrogênio na biosfera, é também responsável pela disponibilização deste elemento para a utilização pelos seres vivos sendo catalisada pelo complexo enzimático da nitrogenase, encontrado apenas em alguns organismos procariontes, denominados diazotróficos (POSTGATE, 1982; SAIKIA E JAIN, 2007).

1.3. ORGANISMOS DIAZOTRÓFICOS

Somente algumas espécies de organismos procariontes, chamados de diazotrofos, encontrados nos domínios Archaea e Bactéria são capazes de fixar nitrogênio (POSTGATE, 1982).

As bactérias diazotróficas ocupam nichos distintos, podendo ser aeróbicos, anaeróbicos facultativos ou estritos, de vida livre, aquáticos ou do solo (YOUNG, 1992).

Sabe-se atualmente que algumas bactérias diazotróficas colonizam o interior da planta e são conhecidas como bactérias endofíticas fixadoras de nitrogênio (OLIVARES et al., 1996; REINHOLD-HUREK et al., 2000; URETA et al., 1995). Diversos estudos mostram que estes endofíticos colonizam seus hospedeiros em grande número causando aumentos na produção. Em culturas de arroz inundado, por exemplo, há uma contribuição da fixação de nitrogênio de 30 a 60 Kg N/ha /ano (SANTIAGO et al.,1986). Na cana-de-açúcar a contribuição pode chegar a 150 kg N/ha/ano (URQUIAGA et al., 1992). A possibilidade de se substituir o nitrogênio fertilizado com a fixação biológica de nitrogênio deve ser considerada pelo fato de ser econômica e ambientalmente vantajosa (REIS et al, 2000).

Os diazotrofos contribuem com aproximadamente 60% do nitrogênio fixado no planeta e são essenciais para a manutenção do ciclo do nitrogênio (BURNS e HARDY, 1975 *apud* KLASSEN, 2000).

A reação de incorporação do nitrogênio pelas bactérias diazotróficas é catalisada pelo complexo enzimático da nitrogenase e consiste na conversão do nitrogênio gasoso (N_2) na sua forma mais reduzida, como íons de amônio $(NH_4)^+$, que serão utilizados pelos seres vivos para a biossíntese de seus compostos nitrogenados (POSTGATE, 1982).

1.4. PERFIL GÊNICO DA FIXAÇÃO DE NITROGÊNIO

A enzima mais importante associada ao mecanismo de fixação de nitrogênio é a nitrogenase (SUR et al, 2010). É uma molécula biologicamente bastante complexa que reduz dinitrogênio gasoso à amônia sob temperatura e pressão adequadas (PETTERS E SZILAGYI, 2006). A nitrogenase convencional é uma mistura equilibrada das proteínas Ferro e Molibdênio-Ferro na proporção de 2:1 (SUR, 2010).

A nitrogenase em si é um complexo ATP-hidrolisante, redox-ativo composto por duas proteínas: o heterotetrâmero dinitrogenase $\alpha_2\beta_2$ (onde $\alpha =$ proteína *nifD* e $\beta =$ proteína *nifK*) correspondente à proteína Molibdênio-Ferro e o homodímero dinitrogenase reductase γ_2 (proteína *nifH*), correspondente à proteína Ferro. A subunidade α contém um sítio ativo para redução de dinitrogênio: tipicamente um centro de metálico $MoFe_7S_9$ (denominado FeMo-cofator) (RAYMOND et al, 2010).

Dois cofatores FeMoCo estão ligados às subunidades α da proteína MoFe. Além disso, os centros 4Fe-4S compreendem alguns outros grupos protéticos. Os centros P são ligados covalentemente aos resíduos de cisteína das proteínas Molibdênio-Ferro, formando uma ponte entre as subunidades α e β (RUBIO e LUDDEN, 2008).

Conforme descrito por Burris (BURRIS, 1991), a estequiometria da reação catalisada pelo complexo da nitrogenase é a seguinte:

$$N_2 + 8H^+ + 8e^- + 16MgATP \rightarrow 2\ NH^+_4 + H_2 + 16MgADP + 16P_i \quad (1)$$

O estabelecimento das estruturas tridimensionais das proteínas da nitrogenase e seus centros metálicos em 1992 mudou consideravelmente a pesquisa sobre fixação de nitrogênio (GEORGIADIS et al., 1992; KIM et al., 1993; EINSLE et al., 2002), pavimentando o caminho para a evolução de áreas como a bioquímica, espectroscopia e a biofísica, modelando e descrevendo quimicamente as características estruturais das proteínas da nitrogenase, seus mecanismos funcionais e capacidade de catalisar a fixação de nitrogênio em diferentes condições ambientais (SUR et al., 2010).

Várias técnicas como a mutagênese, o mapeamento de deleções, os vetores de clonagem, etc., têm facilitado a identificação e caracterização de genes associados à fixação biológica de nitrogênio. Sabe-se hoje que os genes *nod*, *nol*, *noe*, gln, *fix, fdx, res*, *vnf*, *anf* e *nif* são alguns genes encontrados em fixadores de nitrogênio (SUR et al., 2010).

1.5. O CLUSTER *nif*

As pesquisas genéticas envolvendo a fixação biológica de nitrogênio foram primeiramente iniciadas em *Klebsiella pneumoniae* e pela primeira vez uma detalhada organização de genes *nif* foi descrita neste organismo (ARNOLD et al., 1988 *apud* SUR, 2010).

Em *Klebsiella pneumoniae* foram identificados e sequenciados 21 genes *nif*, os quais se encontram numa região de aproximadamente 20 kb do genoma (Figura 2) (DIXON et al, 1986 *apud* PEDROSA et al, 2001).

Estudos confirmam que os genes *nifHDK* codificam a nitrogenase: o gene *nifH* codifica para as subunidades γ da proteína Fe e os genes *nifD* e *nifK* codificam as subunidades α e β da proteína MoFe respectivamente (ROBERTS et al., 1978 *apud* KLASSEN, 2000; SUNDARESAN e AUSUBEL, 1981 *apud* KLASSEN, 2000; IOANNIDIS e BUCK, 1987 *apud* KLASSEN, 2000; SUR, 2010).

Os genes *nifEN* codificam para as proteínas NifE e NifN, formando um tetrâmero $\alpha_2\beta_2$. Este tetrâmero está envolvido na síntese do FeMoCo (PAUSTIAN et al., 1989 *apud* KLASSEN, 2000, RAYMOND, 2004) provavelmente através da

formação de um molde para o cofator. Isto porque os genes *nifN* e *nifE* apresentam alta similaridade com os genes *nifK* e *nifD*, respectivamente (ROBERTS et al., 1978 *apud* KLASSEN, 2000; BRIGLE et al., 1987 *apud* KLASSEN, 2000; RAYMOND et al., 2004) acreditando-se inclusive que os genes *nifN* e *nifE* tenham sido originados a partir de uma antiga duplicação de um operon NifDK (FANI et al., 2000).

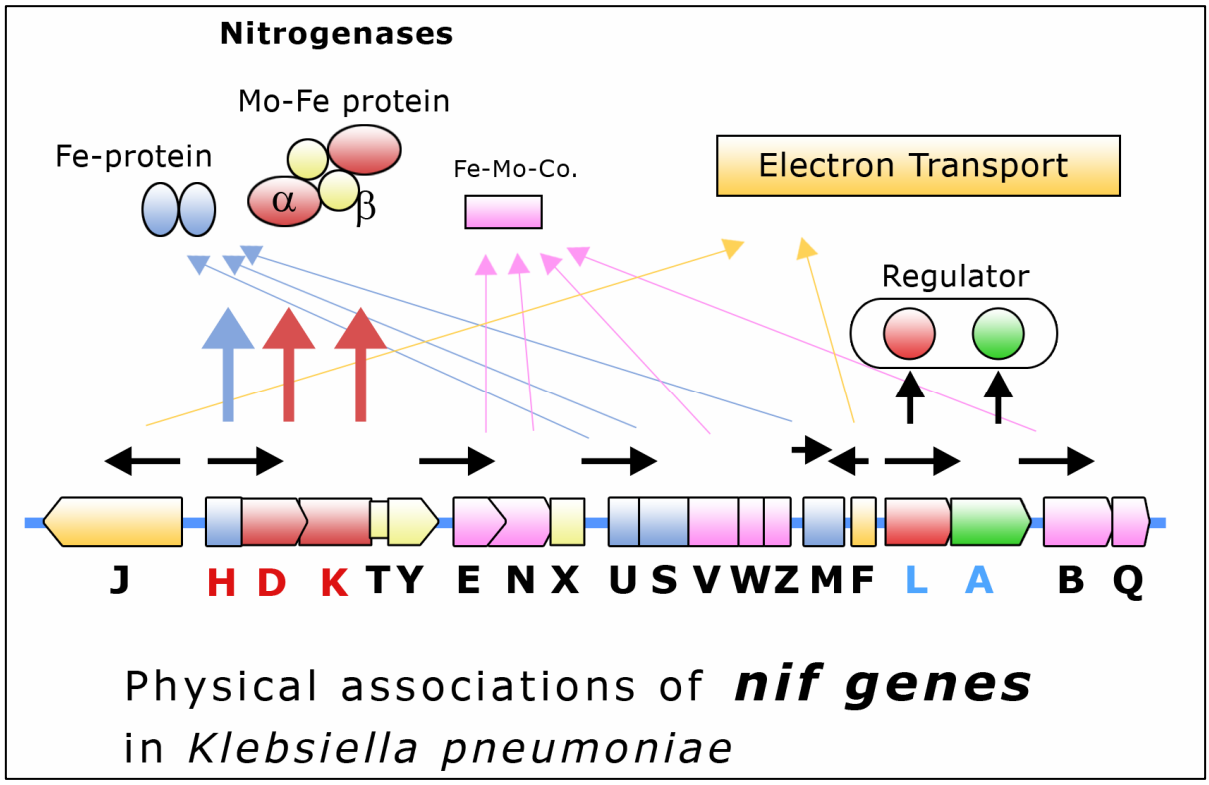


FIGURA 2 – DISPOSIÇÃO DOS GENES *nif* EM *Klebsiella pneumoniae*.
FONTE: Livre adaptação a partir de <http://www.asahi-net.or.jp/~it6i-wtnb/bnf.html>. Último acesso em 04/01/2011.

A Tabela 1 sumariza os produtos e funções catalogadas para os genes *nif* de acordo com Triplett (TRIPLETT, 2000):

TABELA 1 – PRODUTOS DOS GENES *nif* E SUAS FUNÇÕES (CONHECIDAS OU PROPOSTAS) NO PROCESSO DE FIXAÇÃO DE NITROGÊNIO.

Gene *nif*	**Identidade / Produto / Função**
nifH	Dinitrogenase reductase. Doador obrigatório de elétrons durante o *turnover* da nitrogenase. Também necessário para a biossíntese do FeMoCo e maturação da apodinitrogenase.
nifD	Subunidade α da dinitrogenase. Forma o tetrâmero $\alpha_2\beta_2$ juntamente com a subunidade β. Sítio de redução de substrato do FeMoCo se encontra inserido juntamente com a subunidade α da dinitrogenase.
nifK	β subunidade da dinitrogenase. Os centros P estão presentes na interface da subunidade β.
nifT	Desconhecido.
nifY	Em *K. pneumoniae* auxilia na inserção do FeMoCo dentro da apodinitrogense.
nifE	Forma o tetrâmetro $\alpha_2\beta_2$ juntamente com o gene *nifN*. Necessário para a síntese do FeMoCo.
nifN	Necessário para a síntese do FeMoCo.
nifX	Necessário para a síntese do FeMoCo.
nifU	Envolvido na síntese do FeMoCo. Função específica desconhecida.
nifS	Envolvido na mobilização de Fe para a síntese e reparo do cluster Fe-S.
nifV	Envolvido na mobilização de S para a síntese e reparo do cluster Fe-S.
nifW	Homocitrato sintase, envolvido na síntese do FeMoCo.
nifZ	Envolvido na estabilidade da dinitrogenase. Supõe-se que proteja a dinitrogenase da inativação de O_2.
nifM	Desconhecido.
nifF	Flavodoxina. Doador fisiológico de elétrons para a proteína NifH.
nifL	Elemento regulatório negativo.
nifA	Elemento regulatório positivo.
nifB	Necessário para a síntese do FeMoCo. O produto metabólico NifB-co é o doador específico de Fe e S para o FeMoCo.
nifQ	Envolvido na síntese do FeMoCo.
nifJ	Piruvato:flavodoxina (ferredoxina) oxidoredutase. Envolvida no transporte de elétrons para a nitrogenase.

FONTE: Traduzido a partir de Triplett, 2000 (TRIPLETT, 2000).

Numerosos estudos sugerem a existência de pelo menos 14 genes *nif* em comum encontrados entre os diazotrofos: genes *nifH, nifD, nifK, nifE, nifN, nifX, nifU, nifS, nifV, nifW, nifZ, nifQ, nifB* e *nifY*, cujos produtos são essenciais para a biossíntese da nitrogenase (ZHENG et al., 1998; MERRICK e EDWARDS, 1995; DIXON e KAHN, 2004; HU et al., 2007; RUBIO e LUDDEN, 2008).

A ativação da transcrição de genes *nif* ocorre em momentos de estresse de nitrogênio sendo dependente da proteína nitrogênio-sensível *nifA*. Se houver uma quantidade suficiente de oxigênio ou de nitrogênio reduzido presente, uma outra proteína é ativada: o NifL. A proteína NifL bloqueia a atividade do gene *nifA* resultando na inibição da nitrogenase em formação. A proteína NifL é regulada pelos produtos dos genes *glnD* e *glnB*. Os genes *nif* podem ser encontrados nos

cromossomos das bactérias, mas muitas das vezes eles são encontrados em plasmídeos juntamente com outros genes relacionados à fixação biológica de nitrogênio (tais como os genes *nod*) (PEDROSA et al, 2001).

Estudos conduzidos por Shultz e Kondorosi (SHULTZ e KONDOROSI, 1998) e por Perret e colaboradores (PERRET et al., 2000) verificaram que os genes *nod, nol,* e noe produzem sinais de nodulação. A interação dos diferentes genes *nod*, desencadeando a criação de nódulos nas raizes foram relatados por Yang e colaboradores (YANG et al., 1999), Long (LONG, 2001) e Geurts e Bisseling (GEURTS e BISSELING, 2002) entre outros.

Conforme Triplett (TRIPLETT, 2000), os genes *fdx* correspondem a ferredoxinas, servindo como doadores de elétrons para a nitrogenase em *R. capsulatus*. Em *Herbaspirillum seropedicae* os genes *fdxA* e *fdxN* são responsáveis pela regulação das ferredoxinas (SOUZA et al., 2010).

De acordo com Klassen (KLASSEN, 2000), os genes que codificam para as V e Fe nitrogenases são chamados de *vnf* e *anf*, respectivamente, e há significante similaridade entre estes e os genes relacionados às proteínas estruturais do complexo da nitrogenase. Robson e colaboradores, citado por Klassen (ROBSON et al., 1989 *apud* KLASSEN, 2000) complementa ainda que so genes *nif* codificam as proteínas da dinitrogenase formando um tetrâmero $\alpha_2\beta_2$ produtos dos genes *nifDK*, enquanto que as dinitrogenases alternativas contém uma terceira subunidade formando um hexâmero $\alpha_2\beta_2\delta_2$ e são codificados pelos genes *vnfDKG* ou *anfDKG*, respectivamente.

Os genes *fix* estão envolvidos com a fixação de nitrogênio em organismos simbióticos, não possuindo homólogos com os genes *nif* de *K. pneumoniae* (GUBLER e HENNECKE, 1996 *apud* KLASSEN, 2000).

1.6. NCBI GENBANK

O maior banco de dados da atualidade, em termos de informações sobre seqüências, é o NCBI GenBank (BENSON *et al*, 2002; BENSON *et al, 2003; BENSON et al,* 2007; BENSON *et al,* 2008).

O GenBank ® é um banco de dados compreensível e público que agrega informações sobre seqüências de DNA de diferentes organismos, obtidas principalmente através da submissão de dados individuais ou em lotes - gerados a partir de projetos de seqüenciamento em larga escala. Ele foi criado e é distribuído pelo National Center for Biotechnology Information (NCBI), uma divisão da National Library of Medicine (NLM), localizado no campus do US National Institutes of Health (NIH), em Bethesda, MD (BENSON *et al*, 2002; BENSON *et al, 2003; BENSON et al,* 2007; BENSON *et al,* 2008).

No tocante a colaboração mundial para a atualização dos dados depositados no GenBank, Benson e colaboradores (BENSON *et al,* 2008) explicam que, juntamente com o GenBank, o European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL), descrito por Stoesser (STOESSER *et al,* 2002) e Kulikova (*KULIKOVA et al,* 2007) *e* o DNA Databank of Japan (DDBJ), apresentado por Tateno (TATENO *et al*, 2002) e Sugawara (SUGAWARA *et al,* 2007) compreendem o International Nucleotide Sequence Database Collaboration (INSDC), responsável pela troca de informações diárias de forma a garantir a uniformidade e a compreensão dos dados colecionados.

Como descrito por Wheeler e colaboradores (WHEELER *et al, 2003;* WHEELER *et al, 2008)*, o NCBI disponibiliza gratuitamente os dados depositados através de arquivos em um formato texto padrão, específico do GenBank via FTP, mediante atualizações bimestrais completas e atualizações diárias além de uma vasta linha de ferramentas de busca, serviços e análise de dados. Porém mesmo dotado de boas ferramentas de análise e pesquisa, não permite o acesso inter-relacionado aos dados, dada a sua condição primária de armazenamento de informações, realizada via arquivos texto, conforme descrito por Benson e colaboradores (BENSON *et al,* 2008).

Benson e colaboradores (BENSON *et al*, 2008) relatam que desde sua implantação, o GenBank dobra seu tamanho a cada 18 meses. Suas divisões principais possuem mais de 80 bilhões de bases de nucleotídeos, referentes a mais

de 76 milhões de seqüências individuais, com 15 milhões de novas seqüências adicionadas apenas no ano de 2007. Contribuições do projeto Whole Genome Shotgun (WGS) suplementam os dados nas divisões tradicionais em aproximadamente 190 bilhões de bases. Genomas completos continuam a representar um segmento que cresce rapidamente, sendo 200 dos mais de 570 genomas microbianos completos existentes no GenBank, depositados no ano de 2007. Mais de 260000 espécies nomeadas estão representadas no GenBank e novas espécies estão sendo adicionadas à uma velocidade de mais de 1700 por mês.

Ainda segundo Benson e colaboradores (BENSON et al, 2008), cada entrada no GenBank inclui uma descrição concisa de dados sobre a(s) seqüência(s), nome científico e taxonomia do organismo de origem, referências bibliográficas e uma lista de características que engloba áreas de significância biológica tais como: regiões codificadoras e suas traduções em proteínas, unidades de transcrição, regiões repetitivas e sítios de mutação e modificação. Os arquivos dentro da distribuição do GenBank são particionados em “divisões” que rusticamente correspondem a grupos taxonômicos como as bactérias, vírus, primatas e roedores. A Figura 3 representa graficamente as espécies encontradas no GenBank com maior número de bases depositadas (em bilhões de bases).

Os dados relacionados às sequências depositadas no GenBank são acessíveis através do Entrez: um flexível sistema computacional para recuperação de dados, disponibilizado em: (www.ncbi.nlm.nih.gov/sites/gquery). O sistema engloba aproximadamente 35 bancos de dados integrando informações da maioria das bases de seqüências de DNA e proteínas juntamente com as de taxonomia, genomas, mapeamento, estruturas de proteínas e informações de domínios estruturais, e referencias bibliográficas biomédicas através do PubMed (BENSON et al, 2008).

Combinar novos dados com os de outros pesquisadores em todo o mundo em um banco de dados central fornece sólidos contextos biológicos estimulando novas descobertas (BENSON et al, 2008).

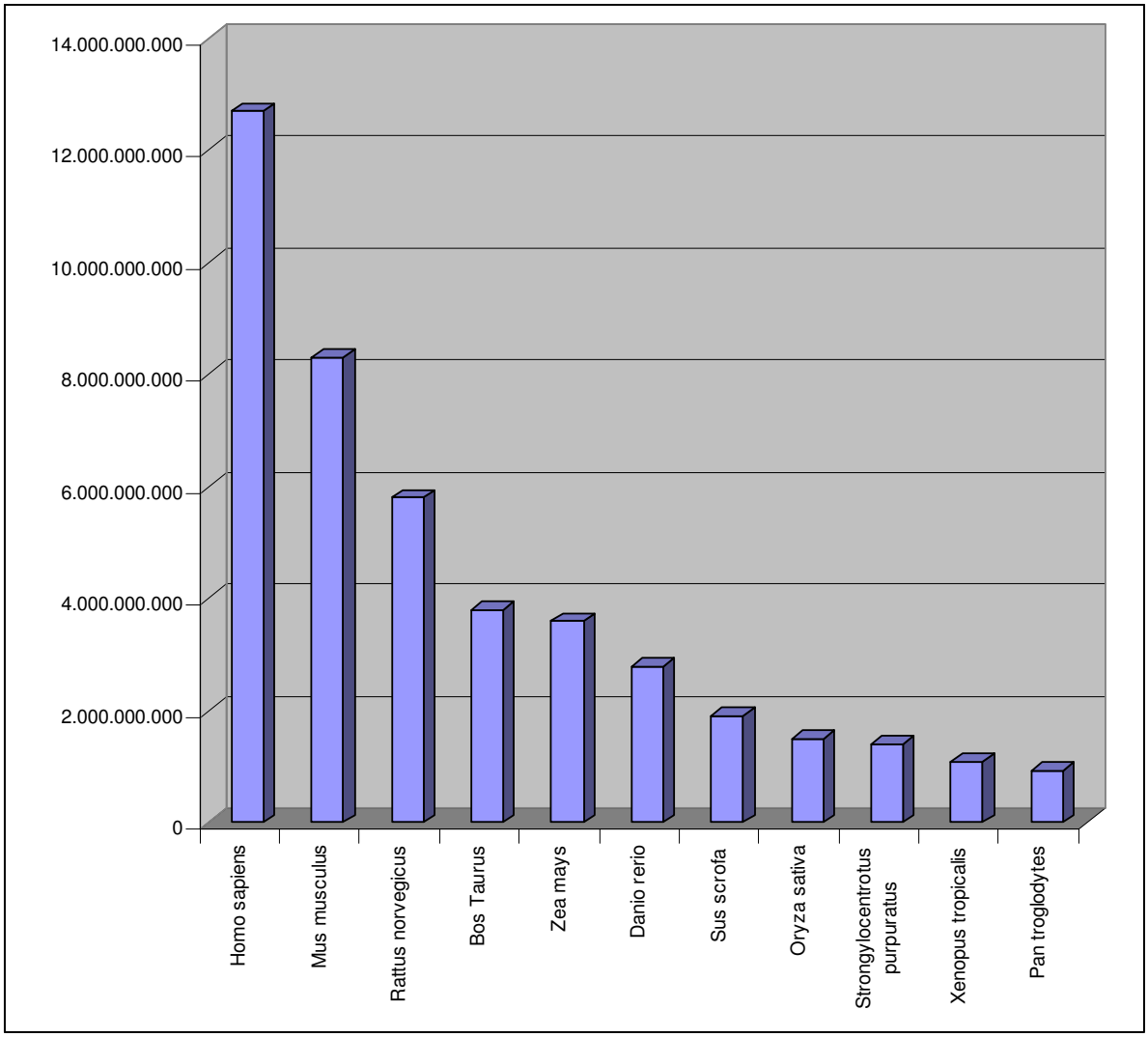


FIGURA 3 – ESPÉCIES COM MAIOR NÚMERO DE SEQUÊNCIAS DEPOSITADAS NO GENBANK. FONTE: Livre adaptação para Benson e colaboradores, 2008 (BENSON *et al.*, 2008).

1.7. BLAST

A determinação da similaridade de uma dada seqüência gênica com outras já existentes é uma poderosa ferramenta para descobrir sua função biológica. Assim como os antigos gregos utilizavam a anatomia comparada para compreender o corpo humano e lingüistas usavam a pedra de Rosetta para decifrar os hieróglifos egípcios, hoje podemos usar a análise comparativa da seqüência de entender genomas, RNA e proteínas (BEDELL et al., 2003).

BLAST, acrônimo para Basic Local Alignment Search Tool (ALTSCHUL et al., 1990; BEDELL et al., 2003; YE et al., 2006; JOHNSON et al., 2008), trata-se de

uma das mais utilizadas ferramentas de análise e pesquisa de sequências em bioinformática (JOHNSON et al., 2008).

Segundo Bedell e colaboradores (BEDELL et al., 2003), o programa BLAST implementa uma heurística para o alinhamento local de biossequências. Estas consistem em seqüências de nucleotídeos, formadas por repetições das letras A, T, C e G, e seqüências de aminoácidos, formadas por repetições das 20 letras respectivas a cada um dos 20 aminoácidos existentes.

Alinhamentos locais são usados para encontrar regiões nas quais duas seqüências possuam alto grau de similaridade, sendo diferentes do alinhamento global, que busca alinhar pares de sequências em toda a extensão. O BLAST compara seqüências de entrada contra todas as sequências encontradas em um banco de dados, realizando alinhamentos locais (ALTSCHUL et al., 1990).

Apesar de o adjetivo "Basic" em seu nome, o BLAST é um pacote de software sofisticado e importante na área de bioinformática por várias razões: em primeiro lugar, a análise de similaridade de uma seqüência é uma ferramenta de identificação. Em segundo lugar, o BLAST é rápido. O mundo das seqüências é grande e cresce exponencialmente, assim, velocidade é importante. Em terceiro lugar, o BLAST é confiável, tanto sob um ponto de vista rigorosamente estatístico quanto sob o ponto de vista de desenvolvimento de software. Em quarto lugar, é flexível e pode ser adaptado a vários cenários de análise de sequências. Finalmente, o BLAST se encontra enraizado na cultura da bioinformática na medida em que a palavra "blastar" é freqüentemente usada como verbo (BEDELL et al., 2003).

O ferramental BLAST pode ser utilizado tanto através de uma interface web quanto via ferramenta *stand-alone* para comparar a consulta do usuário com um banco de dados de sequências (ALTSCHUL et al., 1990; ALTSCHUL et al., 1997). Diversas variantes do BLAST comparam todas as combinações de nucleótideos ou proteínas consultadas com respectivas bases de dados de nucleotídeos ou proteínas. BLAST trabalha uma heurística que considera pequenos pontos de partidas, (ou “sementes”) entre duas sequências, e tenta iniciar alinhamentos a partir destes "pontos quentes". Além dos alinhamentos realizados, BLAST fornece informações estatísticas como o valor “expect” e a taxa de falso-positivos (YE et al., 2006).

A suíte BLAST aproxima diretamente alinhamentos de sequências através de uma medida otimizada de similaridade local, a MSP (Maximal Segment Pair).

Resultados matemáticos sobre as propriedades estocásticas das pontuações MSP garantem a análise do desempenho deste método, bem como a significância estatística dos alinhamentos gerados (ALTSCHUL et al., 1990).

A idéia do algoritmo é permitir que buscas semi-ótimas em bancos de dados de sequências sejam realizadas com desempenho aceitável (ALTSCHUL et al., 1990), já que os algoritmos conhecidos para o cálculo de alinhamentos locais ótimos de biossequências apresentam desempenhos muito ruins (SETÚBAL e MEIDANIS, 1997). O BLAST não busca encontrar alinhamentos locais ótimos entre uma dada sequência de consulta e todas as sequências de um banco de dados, e sim HSPs (High Scoring Pairs), que são pares de segmentos de sequências de alta pontuação quando alinhados (ALTSCHUL et al., 1990).

O cálculo da pontuação difere para nucleotídeos e proteínas. Para determinar a pontuação de um par de segmentos de nucleotídeos, basta contar em quantas posições no alinhamento, as letras são iguais e em quantas posições estas são diferentes. Já no caso em que o par de segmentos é de aminoácidos, são usadas matrizes de pontuação, como a PAM e a BLOSUM, as quais associam pares de aminoácidos com a pontuação do alinhamento do par em questão, levando em conta semelhanças funcionais e evolutivas existentes entre os aminoácidos (ALTSCHUL et al., 2005; BEDELL et al., 2003).

O BLAST se baseia no fato de que um alinhamento local de alta pontuação entre duas seqüências possui diversos pequenos trechos de alinhamentos dos caracteres das mesmas. Assim, para encontrar os HSPs, um método de alinhamento por palavras é utilizado, através do qual palavras de um tamanho fixo extraídas da sequência de consulta são procuradas na sequência que está sendo alinhada (YE et al., 2006).

Conforme descrito por Altschul (ALTSCHUL et al., 1990), o algoritmo básico do BLAST possui três etapas, que são descritas a seguir :

1. **Construção da lista de palavras candidatas** – no caso dos nucleotídeos, são geradas todas as $L-w+1$ palavras de tamanho w presentes na sequência de consulta de tamanho L. No caso dos aminoácidos, são enumeradas todas as $20w$ palavras possíveis de um tamanho w fixo, e selecionadas as que possuam pontuação no mínimo igual a um limite T quando alinhadas, sem *gaps*, com alguma palavra (também de tamanho w) da seqüência de consulta.

2. **Determinação dos hits no banco de dados** – são encontradas todas as combinações exatas (hits) entre as palavras candidatas e as seqüências do banco de dados.

3. **Extensão dos hits** – cada equivalência exata encontrada no passo anterior é estendida até que a sua pontuação diminua um valor X abaixo da melhor pontuação encontrada para extensões menores.

O BLAST não garante que o alinhamento local ótimo será encontrado, mesmo sendo utilizadas técnicas estatísticas para apresentar medidas de qualidade dos resultados encontrados nas buscas (BEDELL et al., 2003).

Dentro do escopo deste trabalho, a ferramenta BLAST foi essencial para o processo de mineração de dados realizado.

1.8. ASSINATURAS DE DOMÍNIOS CONSERVADOS

O conceito de domínio foi proposto pela primeira vez em 1973 por Wetlaufer (WETLAUFER, 1973) depois de estudos cristalográficos de raios-x em lisozima de galinhas, desenvolvidos por Phillips (PHILLIPS, 1966), papaína, realizados por Drenth e colaboradores (DRENTH et al, 1968.) e por estudos de proteólise limitada de imunoglobulinas, desenvolvidos em 1973 por Porter e Edelman (PORTER, 1973; EDELMAN, 1973). Wetlaufer (WETLAUFER, 1973) definiu domínios como sendo unidades estáveis da estrutura da proteína que poderiam dobrar (encapsular-se) de forma autônoma. Complementando o conceito de domínio, em 1981, Richardson (RICHARDSON, 1981) os definiu como tendo uma estrutura compacta e em 1991, Bork (BORK, 1991) demonstrou que os domínios possuíam características de função e evolução.

De acordo o sítio WordLingo (WORLDLINGO, 2011), um domínio protéico é uma parte da sequência e estrutura da proteína que pode evoluir, funcionar e existir independentemente do resto da cadeia protéica. Cada domínio forma uma estrutura compacta tridimensional e muitas vezes pode ser independentemente estável e encapsulado. Muitas proteínas consistem de vários domínios estruturais. Um dado domínio pode aparecer em uma variedade de proteínas evolutivamente relacionadas. Domínios possuem tamanhos que variam entre cerca de 25 a 500 aminoácidos de comprimento, frequentemente podem formar unidades funcionais e,

pelo fato de serem auto-estáveis, podem ser "trocados" por engenharia genética entre uma proteína e outra para a geração de proteínas quimera.

Quando se analisa famílias de sequências de proteínas, nota-se que algumas regiões acabam sendo melhor conservadas do que outras, durante a evolução. Essas regiões são geralmente importantes para a função de uma proteína e/ou para a manutenção de sua estrutura tridimensional. Ao analisar as propriedades constantes e variáveis desses grupos de sequências similares, é possível derivar uma assinatura para uma família de proteínas ou domínio que distingua seus membros em relação à outras proteínas não relacionadas. Uma analogia pertinente é o uso de impressões digitais por parte da polícia para fins de identificação. Uma impressão digital é geralmente suficiente para identificar um determinado indivíduo. Da mesma forma, a assinatura da proteína pode ser usada para inserir uma proteína recentemente sequenciada a uma determinada família de proteínas e, portanto, formular hipóteses sobre sua função (SIGRIST, 2010).

Entre os bancos de dados públicos relacionados a domínos de proteínas, atualmente disponíveis podemos citar:

- **NCBI CDD**: do inglês Conserved Domains Database, consiste em uma coleção de modelos de múltiplos alinhamentos de seqüências, domínios conservados e proteínas completas. Estes estão disponíveis como matrizes de pontuação de posição específica (PSSMs: Position-Specific Score Matrices) para uma identificação mais rápida de domínios conservados em sequências de proteínas via RPS-BLAST. Seu conteúdo inclui os domínios curados do NCBI, que usam as informações da estrutura 3D de forma explícita para definir limites de domínio e proporcionar *insights* sobre as relações de sequência X estrutura X função, bem como modelos de domínio importados de várias bases de dados de origem externa como Pfam, SMART, COG e PRK TIGRFAM) (MARCHLER-BAUER et al., 2011).
- **ProDom**: Banco de dados compreensível de domínios de famílias protéicas gerados a partir de comparações globais de todas as sequências protéicas disponíveis (SERVANT et al., 2002; BRU et al., 2005).

- **Pfam**: Um dos mais importantes bancos públicos de informações sobre classificação de proteínas contando com 75% das proteínas conhecidas. Criado no Instituto Wellcome Trust Sanger em 1998, o projeto Pfam é conduzido pelo Dr. Alex Bateman (FINN et al, 2010).
- **ExPASy PROSITE:** O servidor de dados proteômicos ExPASy (acrônimo para **Ex**pert **P**rotein **A**nalysis **Sy**stem) do Swiss Institute of Bioinformatics (SIB) é dedicado à análise de sequências e estruturas protéicas (GASTEIGER, 2003). O PROSITE é um dos bancos de dados agregados ao ExPASy, correspondente a famílias de proteínas e domínios. Ele é baseado no fato de que, como existe um grande número de diferentes proteínas, a maioria delas pode ser agrupada, baseando-se nas similaridades em suas sequências, dentro de um limitado número de famílias. Domínios de proteínas ou proteínas pertencentes a uma determinada família, de modo geral, possuem os mesmos atributos funcionais e são derivadas de um ancestral comum. O PROSITE atualmente contém padrões e perfis específicos para mais de mil famílias de proteínas ou domínios. Cada uma destas assinaturas vem com documentação oferecendo informações básicas sobre a estrutura e função dessas proteínas (SIGRIST, 2010).

1.9. MINERAÇÃO DE DADOS

O progresso das tecnologias de informação tornou o armazenamento e distribuição de dados muito mais fácil nas últimas duas décadas. Enormes quantidades de dados têm sido acumuladas em um ritmo muito rápido. No entanto, os dados não são, por vezes, tão úteis e significativos, porque o que as pessoas desejam obter é o conhecimento - a informação “escondida” nos dados. Conhecimento ou informação pode ser visto como os padrões ou características dos dados, sendo muito mais valioso do que os próprios dados. Assim, um campo de novas tecnologias surgiu em meados de 1990 para lidar com a descoberta do conhecimento sobre os dados. Ele é chamado de descoberta de conhecimento em bases de dados, do inglês KDD (Knowledge Discovery in Databases), ou simplesmente de mineração de dados, do inglês *Data Mining* ou DM (CHEN et al, 1996; FAYYAD et al, 1996).

Dentre as técnicas englobadas pela tecnologia da informação, utilizadas em bioinformática, pode-se considerar a mineração de dados como sendo o núcleo, visando descobrir conhecimento a partir de grandes quantidades de dados (HSU, 2006).

Conforme descrito por Hsu (HSU, 2006) os dados se acumulam como uma grande montanha, porém, a maioria deles não é considerada útil. Assim como o mineiro deseja desenterrar metais e pedras preciosas da terra e da rocha, o profissional que realiza *data mining* revela conhecimento/informação ao processar grandes quantidades de dados.

Definições para mineração de dados são encontradas na literatura sob diferentes formas. Fayyad (FAYYAD, 1996, p. 20) descreve que: “A descoberta de conhecimento em bases de dados é o processo não trivial de identificação válida, nova, potencialmente útil e, finalmente, compreensível de padrões de dados".

Simoudis (SIMOUDIS, 1996, p. 26) relata que: "Mineração de Dados é o processo de extração de informações válidas, compreensíveis e úteis e previamente desconhecidas, a partir de grandes bases de dados e utilizá-las para a tomada de decisões cruciais”. Já para Dunham (DUNHAM, 2003, p. 3) "Minerar de dados é encontrar informações escondidas em um banco de dados".

Elmasri e Navathe (ELMASRI e NAVATHE, 2005, p. 264) e Hsu (HSU, 2006, p. 2) abordam a mineração de dados como sendo uma fase da KDD. Hsu (HSU, 2006, p. 2) descreve a mineração de dados como sendo o algoritmo ou método utilizado para encontrar informações a partir dos dados. Enquanto isso, KDD é todo o processo, incluindo a coleta de dados, o pré-processamento dos mesmos, a mineração de dados e a interpretação da informação. A Figura 4 demonstra graficamente o processo de KDD descrito por Hsu (HSU, 2006).

Para Elmasri e Navathe (ELMASRI e NAVATHE, 2005, p. 264) *data mining* se refere à mineração ou a descoberta de novas informações em função de padrões ou regras em grandes quantidades de dados e é uma parte integrante do processo de KDD, composto por seis fases:

- Seleção de dados;
- Limpeza;
- Enriquecimento;
- Transformação ou Codificação;

- *Data Mining*, e
- Apresentação de Resultados.

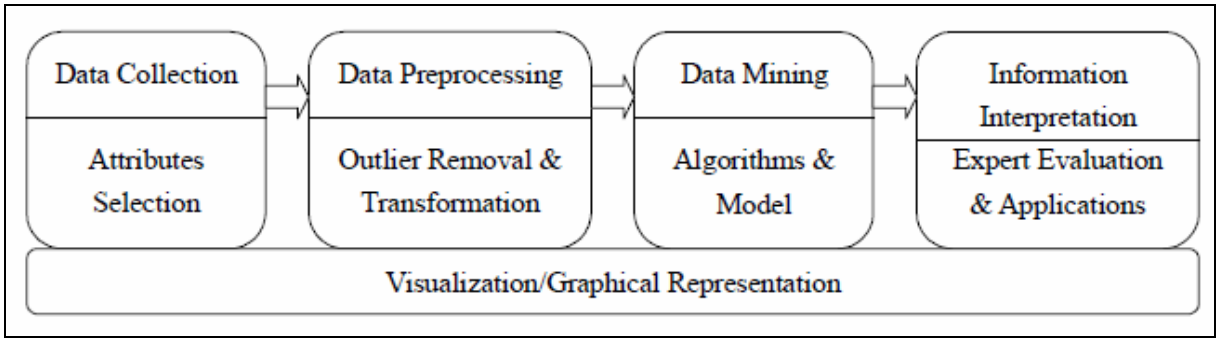


FIGURA 4 – O PROCESSO DE KDD.
FONTE: Extraído de Hsu (HSU, 2006).

Dentre as semelhanças encontradas nas definições extraídas a partir da literatura, o descobrimento de informações como sendo o objetivo da mineração de dados é fato. No entanto, Hsu (HSU, 2006) apresenta algumas premissas que a informação descoberta deve atender:

1. Inovação: O conhecimento ou informação deverá ser uma novidade. O senso comum ou fatos conhecidos não são o que se procura.
2. Corretude: Uma seleção inadequada ou uma má representação de dados irá levar à resultados incorretos. As informações extraídas precisam ser cuidadosamente verificadas por especialistas no domínio.
3. Significado: As informações extraídas deverão significar alguma coisa podendo ser facilmente compreendidas.
4. Aplicação: A informação extraída deve ser capaz de ser utilizada em um problema de domínio determinado.

Com o advento da biotecnologia de alto rendimento, o registro de dados biológicos como o DNA, RNA e proteínas entre outros são gerados mais rapidamente do que nunca. Enormes quantidades de dados estão sendo produzidos e recolhidos. A mineração de dados pode ajudar o biólogo na busca por novos conhecimentos a partir de pilhas de informações biológicas combinando bancos de dados tradicionais, estatísticas e tecnologias de aprendizado de máquina com um

mesmo objetivo. Entre as tecnologias de ponta aplicadas na mineração de dados podemos encontrar por exemplo, algoritmos de recuperação de informações e de aprendizagem de dados, armazenagem Bayesiana, modelos ocultos de Markov (do inglês Hidden Markov Models, HMM), redes neurais artificiais e algoritmos genéticos (HSU, 2006).

1.10. EXTRAÇÃO DE CARACTERÍSTICAS

Características podem ser entendidas como qualquer medição útil extraída no processo de identificação de um padrão. As características podem ser simbólicas, numéricas ou ambas, podendo ser variáveis contínuas ou discretas (SOUZA, 1999).

A extração de características é um processo comumente utilizado em aprendizagem de máquina, na qual um subconjunto das funcionalidades existentes a partir dos dados disponíveis é selecionado para aplicação de um algoritmo de aprendizagem. O melhor conjunto contém o menor número de dimensões que mais contribuem para a precisão; todo o restante deve ser descartado (SEWELL, 2007).

Segundo Sewell (SEWELL, 2007), a extração de características é uma fase importante do pré-processamento utilizado para o reconhecimento de padrões e pode ser realizada utilizando-se duas abordagens:

- *Forward Selection*: Inicia-se com nenhuma variável e adiciona-se uma a uma. Em cada etapa de adição, o erro é diminuido. O processo deve ser repetido até que qualquer nova adição não diminua o erro de forma significativa.
- *Backward Selection*: O início contempla todas as variáveis e as mesmas devem ser removidas uma a uma. A cada remoção o erro é diminuído até ser estabilizado de forma que nenhuma nova remoção o diminua de maneira significativa.

Kira e Rendell (KIRA e RENDELL, 1992) descreveram um algoritmo estatístico para seleção de características chamado RELIEF que usa aprendizagem baseada em exemplos para atribuir um valor relevante para cada característica.

John, Kohavi e Pfleger (JOHN, KOHAVI E PFLEGER, 1994) abordaram o problema das características irrelevantes na seleção de características

apresentando definições para a irrelevância e dois graus de pertinência (fracas e fortes). Também afirmam que as características selecionadas devem depender não só dos conceitos de característica e metas, mas também de um algoritmo de indução. Além disso, afirmaram que a abordagem do modelo de filtros para seleção de características pode ser substituído pelo modelo de padrões.

Pudil, Novovicová e Kittler (PUDIL, NOVOVICOVÁ e KITTLER, 1994) apresentaram Métodos “Flutuantes” de pesquisa em seleção de características. Esses são métodos de busca sequencial caracterizados por uma forma dinâmica de alteração de características incluídas ou eliminadas em cada etapa. Eles foram criados no intuito de apresentar resultados melhores do que as técnicas disponíveis até então e serem computacionalmente mais eficientes do que o método *Branch and Bound*. Conforme descrito por Bensana, Bel e Dubois (BENSANA, BEL e DUBOIS, 1988), o princípio do *Branch and Bound* (BB) é a enumeração de todas as soluções viáveis de um problema de otimização combinatória (problema de minimização) tal que propriedades ou atributos não compartilhados por qualquer solução ótima são detectados tão cedo quanto possível. Um atributo (ou ramo da árvore de enumeração) define um subconjunto do conjunto de todas as soluções viáveis do problema original onde cada elemento do subconjunto satisfaz este atributo. Para Ehlers e Van Rensburg (EHLERS e VAN RENSBURG, 1996), o método *Branch and Bound* é um algoritmo que busca por uma solução ótima através do exame de somente uma pequena parte do número total de possíveis soluções. Ele trabalha quebrando o espaço de soluções viáveis em subproblemas menores até que uma solução ótima seja alcançada. Para cada subproblema gerado o custo total ou lucro é calculado. Subproblemas com pior custo ou lucro são descartados até que não se possam criar mais subproblemas.

Koller e Sahami (KOLLER e SAHAMI, 1996) avaliaram um método para seleção de subconjuntos características com base na Teoria da Informação: eles apresentaram um modelo teoricamente justificado para a seleção da característica ideal baseada no uso de entropia cruzada para minimizar a quantidade de informações preditivas perdidas durante a eliminação de características.

Jain e Zongker (JAIN e ZONGKER, 1997) estudaram vários algoritmos de extração de características e descobriram que o algoritmo de seleção sequencial de Métodos “Flutuantes”, proposto por Pudil, Novovicová e Kittler (PUDIL, NOVOVICOVÁ E KITTLER, 1994), dominou os outros algoritmos testados.

Dash e Liu (DASH e LIU, 1997) fizeram um levantamento dos métodos de seleção de características para a classificação.

Em um estudo comparativo dos métodos de seleção de características na aprendizagem estatística de categorização de texto, Yang e Pedersen (YANG e PEDERSEN, 1997) avaliaram a freqüência documental, os ganhos de informação, a informação mútua, um teste qui-quadrado e uma medida de força-prazo, encontrando um ganho de informação mais eficaz.

Blum e Langley (BLUM e LANGLEY, 1997) centraram sua pesquisa em duas questões fundamentais: o problema da seleção de características relevantes e que o problema de selecionar exemplos relevantes.

Yang e Honavar (YANG e HONAVAR , 1998) utilizaram algoritmos genéticos para a seleção de um subconjunto de características.

Liu e Motoda (LIU e MOTODA, 1998) discorrem sobre a seleção de características uma visão geral dos métodos desenvolvidos desde os anos 1970 fornecendo um quadro geral a fim de examinar os métodos e categorizá-los.

Weston e colaboradores (WESTON et al., 2001) introduziram um método de seleção de características que se baseia em encontrar dados que minimizem os limites do erro. O método mostrou-se superior ao de alguns algoritmos padrão de seleção características nos conjuntos de dados testados.

Xing, Jordan e Karp (XING, JORDAN E KARP, 2001) aplicaram com sucesso os métodos de seleção de características (usando uma abordagem híbrida de filtros e padrões) para um problema de classificação em biologia molecular envolvendo apenas 72 pontos de dados em um espaço dimensional de 7130 pontos. Eles também investigaram métodos de regularização como uma alternativa para seleção de características demonstrando que tais métodos eram preferíveis para o problema que estavam abordando.

Forman (FORMAN, 2003) apresenta uma comparação empírica de doze métodos de seleção de características. Os resultados revelaram o surpreendente desempenho de uma então nova métrica de seleção de características, chamada BNS (Separação Bi-Normal, do inglês, Bi-Normal Separation).

Guyon e Elisseeff (GUYON e ELISSEEFF , 2003) fornecem uma introdução à variáveis e seleção de características. Eles recomendam o uso de um preditor linear de livre escolha e a seleção das variáveis de duas formas alternativas: (1) com um método de classificação variável utilizando um coeficiente de correlação de

informações ou mútuo; (2) com um método de seleção de características aninhadas realizando *forward selection*, *backward selection* ou atualizações multiplicativas.

Técnicas de seleção de características se tornaram uma necessidade aparente em muitas aplicações de bioinformática. Além do grande conjunto de técnicas já desenvolvidas para o aprendizado de máquina e campos de mineração de dados, aplicações específicas em bioinformática, levaram a uma variedade de técnicas recentemente propostas. Durante a última década, a motivação para a aplicação de técnicas de seleção de características em bioinformática deixou de ser um exemplo ilustrativo para se tornar um verdadeiro pré-requisito para a construção de modelos computacionais (SAEYS, 2007).

Como muitas das técnicas de reconhecimento de padrões não foram originalmente concebidas para lidar com grandes quantidades de características irrelevantes, seu uso combinado com técnicas de extração de características tornou-se uma necessidade na maior parte das aplicações (Guyon e Elisseeff, 2003; Liu e Motoda, 1998; Liu e Yu, 2005).

1.11. RECONHECIMENTO DE PADRÕES

Padrão é um conjunto de características que definem um objeto ou um grupo de objetos. É essencialmente um arranjo ou uma ordenação em que alguma organização de estrutura pode ser dita existir (PANDYA e MACY, 1995). Um padrão pode ser referenciado como uma quantidade ou descrição estrutural de um objeto ou algum outro item de interesse. Um padrão pode ser tão básico quanto um conjunto de medidas ou observações, geralmente sendo representado na forma de vetor ou matriz. O mundo pode ser visto como feito de padrões (SOUZA, 1999).

Existem muitas definições e muitas abordagens sobre Reconhecimento de Padrões (RP). Duda e Hart, (DUDA e HART, 1973), caracterizaram RP como sendo um campo que consiste no reconhecimento de regularidades significativas em meios ruidosos e complexos. Já Bezdek e Pal (BEZDEK E PAL,1992) definem RP como a busca por estruturas em dados.

Segundo Pao (PAO, 1989), o conhecimento de reconhecimento de padrões é importante devido às ocorrências na vida humana tomarem forma de padrões. A formação da linguagem, o modo de falar, o desenho das figuras, o entendimento das imagens, tudo envolve padrões. RP é uma tarefa complexa, onde o homem busca,

sempre, avaliar as situações em termos dos padrões das circunstâncias que as constituem e descobrir relações existentes no meio, para melhor entendê-lo e adaptar-se.

O conhecimento de RP é altamente especializado. As contribuições têm vindo de muitas áreas de pesquisas distintas, tais como: sistemas de processamento de sinais, inteligência artificial, modelagem conexionista, teoria de estimação/otimização, conjuntos difusos, modelagem estrutural, linguagem formal, etc) (KASABOV, 1996).

Segundo Souza (SOUZA, 1999), o termo reconhecimento de padrões envolve uma gama significativa de problemas de processamento de informação, com grande relevância prática, desde o reconhecimento de voz e de caracteres feitos manualmente, até à detecção de erros em equipamentos ou diagnósticos médicos. Não é possível abordar todas as áreas onde seria útil e, por vezes mesmo necessário, aplicar RP. Deve-se levar em conta que as pessoas aplicam seus próprios métodos de reconhecimento de padrões em praticamente todas as áreas da atividade humana. Na prática computacional, pode-se enumerar algumas áreas onde vem sendo usado RP (KASABOV, 1996):

- Análise, segmentação e pré-processamento de imagens;
- Reconhecimento de faces;
- Identificação de impressões digitais;
- Reconhecimento de caracteres;
- Análise de manuscritos;
- Visão computacional;
- Entendimento e reconhecimento de voz;
- Diagnóstico médico;
- Sinais biológicos.

Existem, hoje, muitas estratégias de RP desenvolvidas, que se baseiam em técnicas matemáticas, estatísticas e/ou incorporadas à Inteligência Artificial (Redes Neurais, Conjuntos Difusos, etc.). Cada uma tenta simular o RP de forma distinta (SOUZA, 1999).

Independente da técnica usada no RP, Bezdek e Pal (BEZDEK E PAL, 1992) conceitualizam o problema em três estágios ou espaços: o espaço do padrão, o espaço das características e o espaço da classificação. Conceitualmente, o problema de RP pode, então, ser descrito como uma transferência do espaço de padrões P (dimensão R), para o espaço de características F (dimensão N) e finalmente para o espaço de classificação C (dimensão K):

$$P \rightarrow F \rightarrow C \tag{2}$$

O problema em reconhecimento de padrões está na sua definição ou composição, já que definir um conjunto de características que o representa pode não ser uma tarefa trivial. A chave é escolher e extrair um conjunto finito de características que o represente totalmente e que possa ser passível de ser manuseado (BEZDEK E PAL, 1992).

A presença da complexidade nos padrões refere-se normalmente a uma representação em espaços de alta dimensão, levando ao problema da dimensionalidade. A redução de dimensionalidade torna-se uma parte do processo de reconhecimento de padrões que precisa ser melhor explorado. Trabalhos nesta linha mostram a preocupação e a necessidade de novas abordagens para o tratamento da alta dimensionalidade no processo de reconhecimento de padrões (CARREIRA –PERPIÑÁN, 1997; POSTON E MARCHETTE, 1998).

Conforme descrito por Souza (SOUZA, 1999) o processo de extração de características transforma o espaço dos dados (espaço do padrão) em um espaço de características que é de dimensão muito menor comparado com o espaço de dados original, ainda retendo a maioria do conteúdo de informação intrínseca dos dados.

O Reconhecimento de Padrões, assim como a Inteligência Artificial, teve seu início nos anos 50. O RP inicialmente utilizou-se de técnicas probabilísticas, mais precisamente, da Estimação e Teoria da Decisão para sua fundamentação (SOUZA, 1999). Assim, usou a linguagem da Probabilidade em sua origem, usando a abordagem Bayesiana (SCHALKOFF, 1992). Uma outra corrente apareceu depois, utilizando a abordagem estrutural, sendo usado o modelo sintático. Mais

recentemente, a Modelagem Neural e Difusa apareceram como vertentes para esta área (PAO, 1989; FU, 1994B ; BISHOP, 1995).

A teoria mais proeminente independente do domínio de aplicação é a teoria da classificação. Baseada na teoria de decisão estatística, fornece procedimentos matemáticos formais para a classificação de padrões uma vez que eles são representados abstratamente como vetores (DUDA E HART, 1973).

Nas últimas décadas e principalmente nos últimos anos, as Redes Neuronais Artificiais estão sendo cada vez mais usadas como classificadores e também estão sendo desenvolvidas novas redes neurais que simulam os métodos estatísticos (como a rede bayesiana e a probabilística) (PAO, 1989; BISHOP, 1995; PANDYA E MACY, 1995; OMIDVAR E DAYHOFF, 1998).

1.12. REDES NEURAIS ARTIFICIAIS

Redes neurais são uma técnica derivada da pesquisa em inteligência artificial que usa regressão generalizada e gera um método interativo para esse processo. Redes neurais usam a abordagem de curva-apropriada para inferir uma função a um conjunto de amostras. Essa técnica oferece uma “abordagem de aprendizado” e é direcionada por amostras de teste usadas para a inferência inicial e aprendizado. Com esse tipo de método de aprendizado, as respostas para as novas entradas podem ser interpoladas a partir dos exemplos conhecidos. Essa interpolação, entretanto, depende do modelo do mundo (representação interna do problema) desenvolvido pelo método de aprendizagem (ELMASRI e NAVATHE, 2005, p. 640).

Cybenko descreve as Redes Neurais Artificiais como sistemas computacionais inspirados na estrutura, método de processamento e habilidade de aprendizado de um “cérebro biológico” (CYBENKO, 1996).

Para Rezende (REZENDE, 2003, p. 142) as redes neurais artificiais são modelos matemáticos que se assemelham às estruturas neurais biológicas e que têm capacidade computacional adquirida por meio do aprendizado e generalização.

Neste sentido, Nievola afirma que as redes neurais artificiais consistem em uma abordagem de Inteligência Artificial, para soluções de problemas que têm por base o modelo conhecido e aceito de inteligência: o cérebro humano (NIEVOLA, 1998, p. 12).

Para Fausett (FAUSETT, 1994, p.6), uma rede neural artificial é um sistema de processamento de informações que apresenta certas características em comum com uma rede neural biológica. As redes neurais artificiais são desenvolvidas como generalização de modelos matemáticos de neurônios biológicos com base nas seguintes asserções:

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

A partir da observação do modelo natural, McCulloch e Pitts (MCCULLOC e PITTS, 1943) elaboraram um modelo matemático aproximado do comportamento do neurônio. Este modelo, o mais utilizado como fundamento básico para as redes neurais artificiais, sugere que todos os sinais fornecidos à rede atingem o núcleo por meio das entradas (dendritos), que têm a capacidade de atenuar parcial ou totalmente o sinal recebido, realizando a seguir uma reação bioquímica, a qual determina o estado ativo ou inativo da saída da célula nervosa. A Figura 5 exemplifica graficamente um modelo de neurônio biológico.

O neurônio é uma unidade de processamento de informação fundamental para as operações em uma rede neural (HAYKIN, 2001). A Figura 6 mostra o modelo de um neurônio que forma a base para o projeto de redes neurais artificiais.

Os neurônios artificiais, também denominados nodos ou nós de uma rede neural encontram-se organizados em uma ou mais camadas. Cada elemento de processamento é conectado a um ou vários outros por ligações ponderadas, que procuram simular sinapses biológicas (HAYKIN, 1999). Um exemplo esquemático de uma rede neural é ilustrado na Figura 7.

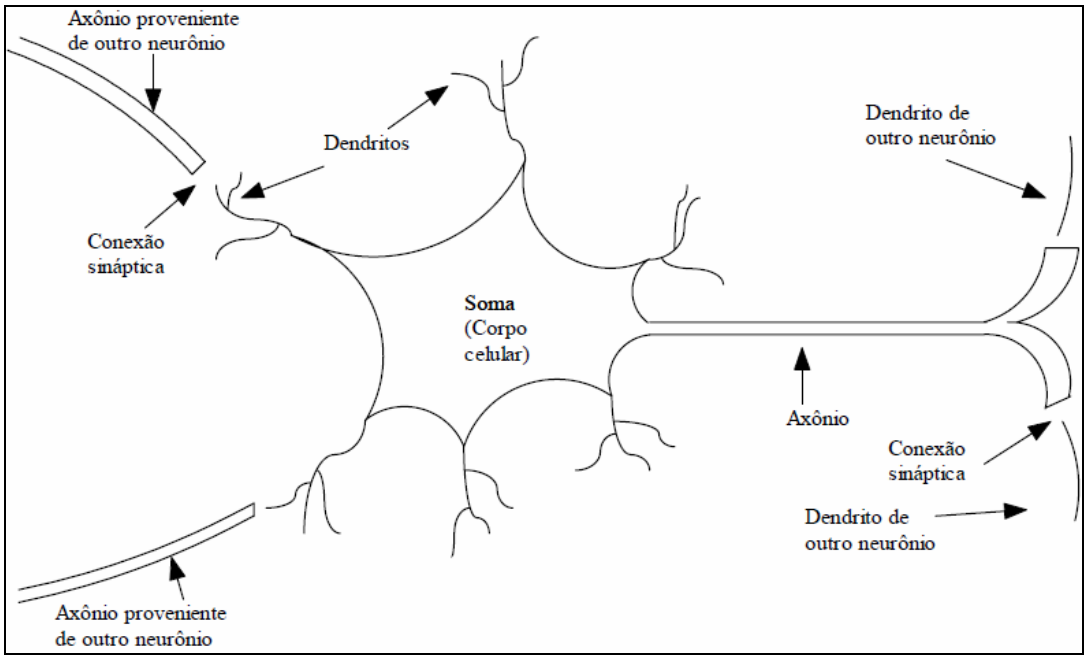


FIGURA 5 – O NEURÔNIO BIOLÓGICO.
FONTE: Adaptado de Fausett, 1994 (FAUSETT, 1994).

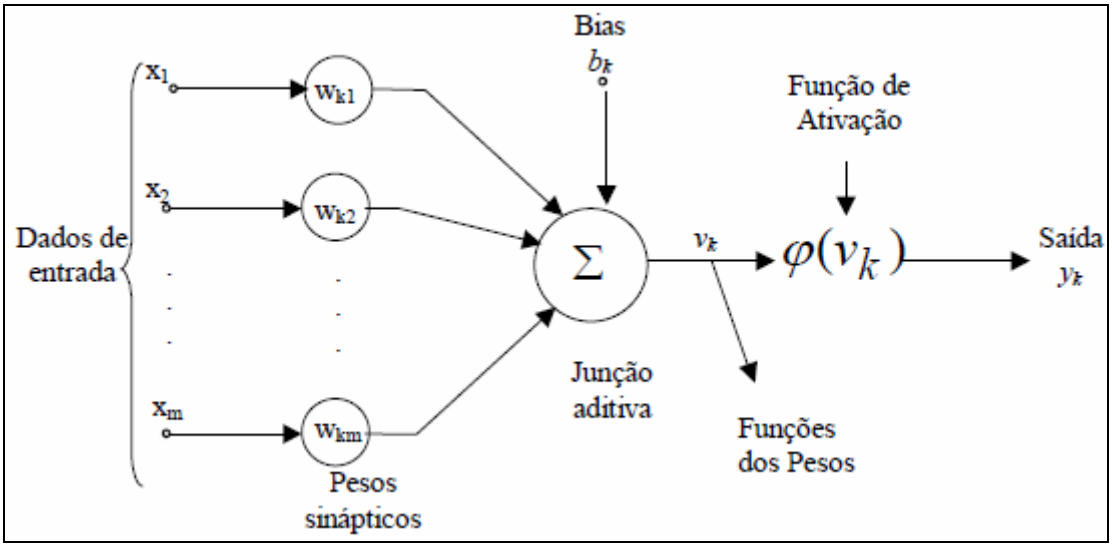


FIGURA 6 – MODELO DE UM NEURÔNIO ARTIFICIAL.
FONTE: Adaptado de Haykin, 2001 (HAYKIN, 2001).

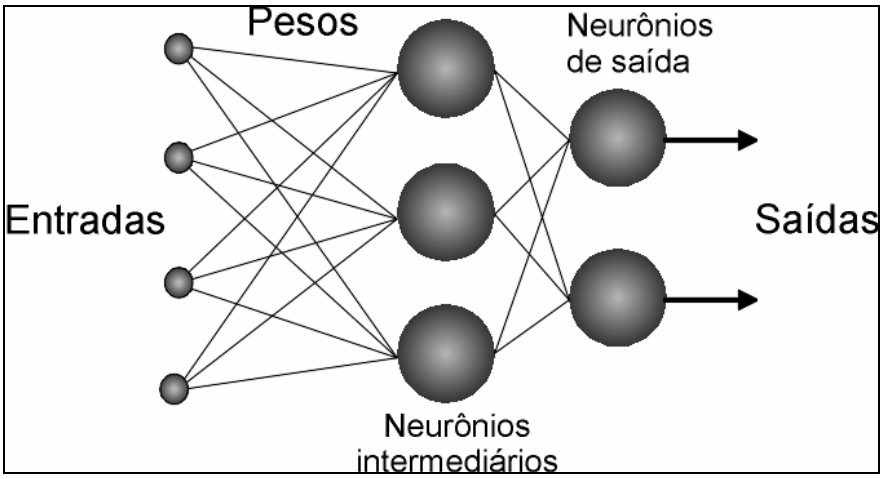


FIGURA 7 – EXEMPLO DE REDE NEURAL.
FONTE: <http://www.cerebromente.org.br/n05/tecnologia/rna.htm>. Último acesso em: 06/01/2011.

As redes neurais podem ser classificadas em duas categorias: redes supervisionadas e não supervisionadas. Os métodos adaptativos que tentam reduzir o erro de saída são os métodos de aprendizado supervisionado, enquanto aqueles que desenvolvem representações internas sem amostras de saída são chamados métodos não supervisionados de aprendizado (ELMASRI e NAVATHE, 2005, p. 640).

As redes neurais auto-adaptativas aprendem com base em informações sobre um problema específico. Elas trabalham bem nas tarefas de classificação e são úteis em mineração de dados (ELMASRI e NAVATHE, 2005, p. 640).

Por estes motivos, as redes neurais artificiais foram utilizadas neste trabalho não apenas como técnica de aprendizado de máquina, mas também como suporte às decisões aplicadas aos dados resultantes da estratégia de *data mining* desenvolvida.

2. JUSTIFICATIVA

Estudos sobre organismos fixadores de nitrogênio fornecem suporte ao desenvolvimento de tecnologias para aplicação de inoculantes bacterianos e biofertilizantes, que, utilizados na agricultura brasileira, levam à redução de custo de produção[1] e do impacto ambiental da atividade agrícola, aumento da competitividade dos produtos agropecuários brasileiros e ao aumento da renda do agricultor, trazendo um impacto positivo na sustentabilidade da agricultura brasileira e na formação de recursos humanos altamente qualificados em biologia molecular, genômica, proteômica, transcriptômica e bioinformática (CNPQ, 2011).

Em 1992, numa pesquisa abordando a classificação filogenética de organismos fixadores de nitrogênio, Young (YOUNG, 1992) elaborou manualmente uma lista contendo todos os conhecidos microorganismos fixadores de nitrogênio, descritos na literatura até aquele ano. Desde então o trabalho tem sido utilizado como referência para estudos posteriores sobre organismos fixadores de nitrogênio. Uma cópia da referida listagem pode ser encontrada no Anexo 1, deste trabalho.

Em 2004, Raymond e colaboradores (RAYMOND et al., 2004) apresentaram um estudo abordando as sequências anotadas de proteínas de todos os organismos com genomas completos e disponíveis no NCBI GenBank (101 na época) que correspondiam aos genes *nifHDKEN*, também visando a classificação filogenética dos mesmos. O download das sequências utilizadas na análise foi realizado manualmente, através do repositório de dados FTP disponibilizado pelo NCBI GenBank.

Considerando: (i) a importância do tema estudado, (ii) a lacuna temporal entre o primeiro e o segundo trabalhos (12 anos) bem como do segundo trabalho até os dias atuais (07 anos), (iii) o aumento exponencial do número de informações relacionadas a sequências genômicas, depositadas em bancos de dados públicos, por todo o mundo e (iv) a dificuldade do processo de se reunir, alocar, disponibilizar e atualizar informações consistentes sobre organismos fixadores de nitrogênio; é que a proposta tema deste trabalho foi idealizada.

[1] Estima-se que o uso destes organismos na agricultura brasileira possa resultar em uma economia anual da ordem de US$ 420 milhões em fertilizantes nitrogenados e, especificamente para o agricultor paranaense, de U$ 115 milhões de dólares (GENOPAR, 2011).

2.1. OBJETIVOS

Em face á justificativa deste trabalho, serão descritos a seguir os objetivos gerais e específicos que regem esta pesquisa.

2.1.1. OBJETIVO GERAL

Mapear automaticamente, através de mineração de dados, os genes *nifHDKEN* pertencentes a organismos fixadores de nitrogênio cujas sequências genômicas estão depositadas no NCBI GenBank.

2.1.2. OBJETIVOS ESPECÍFICOS

- Contemplar na pesquisa os genes *nif* de organismos com genomas completos e incompletos.
- Propor um mecanismo de mineração de dados que contemple a busca, analise, interpretação e classificação das informações relacionadas aos genes alvo da pesquisa.
- Criar rotinas computacionais que suplantem e automatizem o mecanismo de mineração, análise e classificação de dados.
- Criar, a partir dos resultados obtidos, um mapa de relacionamento entre os genes *nif* encontrados e os respectivos organismos que os possuem.
- Traçar um paralelo entre a literatura de referêncua e os dados mapeados visando a contribuição com futuros experimentos em genética, fisiologia e resultando num entendimento melhor dos mecanismos relacionados à fixação de nitrogênio.

3. METODOLOGIA

A metodologia descrita nas seções a seguir apresenta o ferramental necessário para a criação do mecanismo computacional para a mineração, classificação e mapeamento de dados relacionados a organismos fixadores de nitrogênio, de forma a cumprir os objetivos da pesquisa.

3.1 PYTHON

Toda a implementação computacional contida nesta pesquisa foi realizada em linguagem de programação Python 2.6.

Conforme descrito por Lutz e Ascher, o Python é uma linguagem de programação orientada a objetos, utilizada em uma variedade de domínios, tanto para programas independentes como para aplicações de script. Ela é gratuita, portável, poderosa e fácil de usar (LUTZ e ASCHER, 2007).

3.2. BIOPYTHON

Aliado a linguagem de programação Python foi utilizado o *toolkit* BioPython 1.5.2. para a codificação de rotinas e funções relacionadas a busca e interpretação de dados biológicos obtidos no GenBank.

O projeto BioPython é uma associação internacional de desenvolvedores de ferramentas computacionais para biologia molecular em Python. O site oficial, <http://www.biopython.org> é uma fonte gratuita de módulos, scripts e links para desenvolvedores de softwares Python para pesquisa em ciências biológicas (BIOPYTHON, 2011a).

O projeto BioPython, entre outras funcionalidades, converte e interpreta vários formatos de arquivos comumente utilizados em bioinformática em estruturas de dados Python. Entre os formatos suportados, encontram-se os arquivos de saída do BLAST, utilizados como base para a estratégia de mineração de dados descrita neste trabalho (BIOPYTHON, 2011a).

3.3. BANCO DE DADOS NCBI GENBANK PROTEIN

De acordo com a Biblioteca da Universidade de Cornell, o banco de dados Protein é um conjunto de seqüências de várias fontes, incluindo traduções de regiões codificantes anotadas no GenBank, RefSeq e TPA, bem como registros de SwissProt, PIR, PRF e APO. Seqüências protéicas são determinantes fundamentais da estrutura e função biológica de organismos (CORNELL UNIVERSITY LIBRARY, 2011).

O banco de dados Protein é utilizado como fonte de pesquisa dos dados utilizados neste trabalho.

3.4. BLASTP

Como base de busca para os dados utilizados nesta pesquisa, foi utilizada através do toolkit BioPython, a ferramenta BLASTP, versão 2.2.24.

Segundo o NCBI (NCBI, 2011) a versão padrão do BLAST para proteínas é usado tanto para identificar uma seqüência de aminoácidos em uma consulta quanto para encontrar seqüências similares nos bancos de dados de proteínas. Como outros programas BLAST, BLASTP é projetado para encontrar regiões locais de similaridade. Quando a similaridade se estender a toda a seqüência, o BLASTP apresentará igualmente um relatório de alinhamento global, que é o resultado preferido para fins de identificação de proteínas.

O algoritmo de execução automática da ferramenta BLASTP definiu como padrão os parâmetros apresentados na Figura 8:

Search Set Database ..... :	Non-redundant protein sequences (nr)
Program Selection Algorithm ..... :	blastp (protein-protein BLAST)
Max target sequences ..... :	20000
Expect threshold..... :	10
Matrix..... :	BLOSUM62
Gap Costs..... :	Existence 11 – Extension 1
Compositional adjustments..... :	Conditional compositional score matrix adjustement

FIGURA 8 – PARÁMETROS AJUSTADOS PARA EXECUÇÃO AUTOMÁTICA DA FERRAMENTA BLASTP.
FONTE: A autora (2011).

3.5. MS EXCEL

O Microsoft Office Excel trata-se de um software de planilha eletrônica de cálculo. Conforme descrito pela Microsoft, o software Excel possibilita a análise, o gerenciamento e o compartilhamento de informações auxiliando na tomada de decisões (MICROSOFT, 2011).

Nesta pesquisa foi utilizada a versão 2007 do Excel como suporte para a organização, ordenação e análise dos dados obtidos.

3.6. MATLAB

Conforme descrito pela MathWorks, o software MATLAB é um ambiente para o desenvolvimento de aplicativos de natureza técnica. É adequado àqueles que desejam implementar e testar soluções com facilidade e precisão possuindo facilidades de computação, visualização e programação. O nome MATLAB vem do inglês Matrix Laboratory. MATLAB foi originalmente desenvolvido para prover um acesso amigável ao tratamento de vetores e matrizes. Atualmente o MATLAB dispõe de uma biblioteca bastante abrangente de funções matemáticas, geração de gráficos e manipulação de dados que auxiliam muito o trabalho do programador. E ainda possui uma vasta coleção de bibliotecas, denominadas *toolboxes*, para áreas específicas como: bioinformática, estatística, processamento de imagens, processamento de sinais e redes neurais artificiais entre outras (MATHWORKS, 2011).

Para o desenvolvimento deste trabalho foi utilizada a versão 7.8.0.347(R2009a) 32-bit (win32) como auxílio nos testes de desempenho e comparação de redes neurais artificiais.

3.7. CARACTERÍSTICAS CONTEMPLADAS

As características, extraídas a partir das sequências de aminoácidos resultantes do processo de busca automatizado, através da ferramenta BLASTP e definidas ao longo de todo o processo de desenvolvimento serviram de base para o processo de classificação e aprendizado da Rede Neural Artificial.

No total foram extraídas 15 características, divididas em quatro grupos de interesse, apresentados na Tabela 2.

TABELA 2 – CARACTERÍSTICAS EXTRAÍDAS COM BASE NAS SEQUÊNCIAS DE AMINOÁCIDOS, DIVIDIDAS POR GRUPOS DE INTERESSE.

Grupo	Característica	
I - Físico – Químicas	01	Peso molecular normalizado
	02	Ponto isoelétrico
	03	Percentual de aromaticidade
	04	Índice de instabilidade
	05	Índice de Hidropatia (GRAVY)
	06	Domínios Conservados
II - Inferidas	07	Percentual de (F) Fenilalanina
	08	Percentual de (L) Leucina
	09	Percentual de (A) Alanina
	10	Percentual de (R) Arginina
	11	Percentual de (G) Glicina
III – BLAST	12	Tamanho da sequência anotada
	13	Média entre valores de Identidade X Similaridade retornados pelo BLAST, normalizados pelo tamanho da sequência Query de busca.
IV – Matrizes de Co-ocorrência	14	Entropia
	15	Energia

FONTE: A autora, (2011)

As próximas subseções apresentam as medidas utilizadas e os métodos e o ferramental necessário para que o processo de extração de características pudesse ser realizado.

3.7.1. GRUPO I - FISICO-QUÍMICAS

Através de um script em Python, a função **ProtParam.ProteinAnalysis()**, disponibilizada pela biblioteca BioPython, foi aplicada para a geração do grupo de características Físico-Químicas.

As próximas subseções descrevem com mais detalhes cada característica extraída através da função supracitada.

3.7.1.1. Peso Molecular Normalizado

Conforme a documentação disponibilizada pela equipe de desenvolvimento da biblioteca BioPython (BIOPYTHON, 2011b), o valor do peso molecular calculado pela função **ProtParam.ProteinAnalysis()** baseia-se na somatória das massas

moleculares de cada um dos aminoácidos componentes da sequência analisada, descontando-se o valor referente ao peso das moléculas de água resultantes da reação de formação da ligação peptídica. Os valores correspondentes às massas são medidos em Dalton (Da). A Tabela 3 apresenta os aminoácidos atendidos pela função **ProtParam.ProteinAnalysis()**, sua nomenclatura e pesos representados conforme o padrão definido pela União Internacional de Química Pura e Aplicada (International Union of Pure and Applied Chemistry, IUPAC).

TABELA 3 – GRUPOS E PESOS DE AMINOÁCIDOS.

Grupo	Descrição			Massa Molecular (Da)
2	Alanina	Ala	A	89.09
3	Asparagina ou Aspartato	Asx	B	132.61
2	Cisteina	Cys	C	121.16
2	Aspartato	Asp	D	133.10
2	Glutamato ou Ácido Glutâmico	Glu	E	147.13
1	Fenilalanina	Phe	F	165.19
2	Glicina	Gly	G	75.07
1	Histidina	His	H	155.16
1	Isoleucina	Ile	I	131.18
3	Leucina ou Isoleucina	Xle	J	131.18
1	Lisina	Lys	K	146.19
1	Leucina	Leu	L	131.18
1	Metionina	Met	M	149.21
2	Asparagina	Asn	N	132.12
2	Pirrolisina	Pyl	O	255.31
2	Prolina	Pro	P	115.13
2	Glutamina	Gln	Q	146.15
1	Arginina	Arg	R	174.20
2	Serina	Ser	S	105.09
1	Treonina	Thr	T	119.12
2	Selenocisteina	Sec	U	168.05
1	Valina	Val	V	117.15
1	Triptofano	Trp	W	204.23
3	Aminoácido não especificado ou desconhecido	Xaa	X	0.0
2	Tirosina	Tyr	Y	181.19
3	Glutamina ou Glutamato	Glx	Z	146.64

Legenda:
Grupo 1 – Correspondente aos aminoácidos essenciais.
Grupo 2 – Correspondente aos aminoácidos não-essenciais
Grupo 3 – Correspondente aos aminoácidos ambíguos.

FONTE: A autora com base em dados disponibilizados pela IUPAC (IUPAC, 2011)

De modo a normalizar os valores resultantes do cálculo aplicado às sequências em análise, o script tomou o cuidado de dividir valor total de massa

molecular encontrado para cada sequência, pelo seu respectivo tamanho (em quantidade de aminoácidos anotados).

3.7.1.2. Ponto Isoelétrico

De acordo com a documentação disponibilizada pela equipe de desenvolvimento da biblioteca BioPython (BIOPYTHON, 2011c), o cálculo do valor do ponto isoelétrico de uma proteína é calculado segundo os valores e metodologia descritos por Bjellqvist (BJELLQVIST, 1982).

3.7.1.3. Percentual de Aromaticidade

Em conformidade com a documentação BioPython disponibilizada (BIOPYTHON, 2011b) o percentual de aromaticidade de uma sequência protéica é calculado de acordo com a metodologia descrita em 1994, por Lobry e Gautier (LOBRY e GAUTIER, 1994) traduzida no simples cálculo da freqüência relativa dos aminoácidos aromáticos Fenilalanina, Triptofano e Tirosina.

3.7.1.4. Índice de Instabilidade

Conforme a documentação disponibilizada pela equipe de desenvolvimento da biblioteca BioPython (BIOPYTHON, 2011b), o cálculo do índice de instabilidade de uma proteína é realizado conforme a metodologia proposta em 1990 por Guruprasad (GURUPRASAD, 1990) que propõe que qualquer valor resultante, acima de 40 indica que a proteína é instável (ou seja, possui um tempo curto de meia-vida).

3.7.1.5. Índice de Hidropatia (*GRAVY INDEX*)

O índice GRAVY (do inglês Grand Average of Hydropathy), descrito por Kyte e Doolittle (KYTE e DOOLITTLE, 1982) indica a solubilidade de uma proteína: basicamente, um valor GRAVY positivo corresponde a uma proteína hidrofóbica, enquanto um valor GRAVY negativo corresponde a uma proteína hidrofílica. O valor GRAVY para um peptídeo ou proteína é calculado pela soma dos valores de

hidropatia de cada aminoácido, dividido pelo número de resíduos encontrados na sequência.

Os dados resultantes do cálculo foram armazenados e utilizados posteriormente, durante o processo de classificação de dados.

3.7.1.6. Assinaturas de Domínios Conservados

Em função da estrutura e documentação disponível, bem como do fato da representação gráfica dos domínios conservados de proteínas ser realizada através de expressões regulares, o banco de dados ExPASy PROSITE foi o escolhido dentre os disponíveis para ser utlizado nessa pesquisa.

As assinaturas de domínios conservados utilizadas nesta pesquisa se encontram listadas na Tabela 4.

TABELA 4 – ASSINATURAS DE DOMÍNIOS CONSERVADOS UTILIZADAS

Descrição ExPASy PROSITE	Sequência Consenso
NifH/frxC family signature 1	E-x-G-G-P-x(2)-[GA]-x-G-C-[AG]-G
NifH/frxC family signature 2	D - x - L - G - D - V - V - C - G - G - F - [A G S P] - x - P
Nitrogenases component 1 alpha and beta subunits signature 1	[L I V M F Y H] - [L I V M F S T] - H - [A G] - [A G S P] - [L I V M N Q A] - [A G] - C
Nitrogenases component 1 alpha and beta subunits signature 2	[S T A N Q] - [E T] - C - x (5) - G - D - [D N] - [L I V M T] - x - [S T A G R] - [L I V M F Y S T]
moaA / nifB / pqqE family signature	[L I V] - x (3) - C - [N D P] - [L I V M F] - [D N Q R S] - C - x - [F Y M] - C

FONTE: Obtido a partir de ExPASy Proteomics Server, 2011 (EXPASY PROTEOMICS SERVER, 2011a; EXPASY PROTEOMICS SERVER, 2011b; EXPASY PROTEOMICS SERVER, 2011c)

3.7.2. GRUPO II – INFERIDAS

A partir do processo de Classificação Prévia de Dados (descrito com detalhes em Resultados e Discussão), foram separados do montante de dados, com auxílio do software Excel, 1335 registros referentes à sequências protéicas codificadas pelos genes *nif*.

Através de um pequeno script desenvolvido em Python, apresentado na Figura 9, as sequências foram analisadas de forma que as ocorrências para cada um dos aminoácidos componentes das sequências foram somadas em variáveis totalizadoras e exibidas em tela.

```
import os
import sys
proteins = { "A": 0, "B": 0, "C": 0, "D": 0, "E": 0, "F": 0, "G": 0, "H": 0,\
             "I": 0, "J": 0, "K": 0, "L": 0, "M": 0, "N": 0, "O": 0, "P": 0,\
             "Q": 0, "R": 0, "S": 0, "T": 0, "U": 0, "V": 0, "W": 0, "X": 0,\
             "Y": 0, "Z": 0, "-": 0}
file   = open('aminoperc.txt', 'r')
data   = file.readlines()
for line in data:
    lAux = line.replace('\n', '')
    for letter in lAux:
        proteins[letter] = proteins[letter] + 1
file.close()
for item in proteins:
    print item, '\t', proteins[item]
trinca = []
tot    = len(data)
for line in data:
    lAux = line.replace('\n', '')
    ini  = 0
    while ini >= 0:
       atu = lAux.find("G", ini, len(lAux))
       if atu >= 0:
          trinca.append(lAux[(atu-1):(atu+2)])
          ini = atu + 1
       else:
          break
dic = {}
for l in trinca:
    if l in dic:
       dic[l] = dic[l] + 1
    else:
       dic[l] = 1
for l in dic:
    print l, "\t", dic[l]
file.close()
```

FIGURA 9 – SCRIPT PYTHON PARA CONTAGEM DE OCORRÊNCIAS DE AMINOÁCIDOS EM SEQUÊNCIAS PROTÉICAS CODIFICADAS POR GENES *nif*.
FONTE: A autora (2010).

Com base neste resultado, foi verificado que as proteínas codificadas pelos genes *nif* possuíam uma quantidade maior dos aminoácidos: Fenilalanina (F), Leucina (L), Alanina (A), Arginina (R) e Glicina (G), em relação a todos os outros aminoácidos analisados.

A partir daí inferiu-se que os valores percentuais correspondentes aos aminoácidos F, L, A, R e G apresentavam-se como características discriminatórias relevantes e poderiam ser utilizados para auxílio no processo de classificação de dados, detalhadamente descrito em Resultados e Discussão.

3.7.3. GRUPO III – BLAST

A partir dos resultados obtidos com a execução automática da ferramenta BLASTP, descritos em detalhes em Resultados e Discussão, o dado referente ao tamanho da sequência foi julgado como relevante ao processo discriminatório de classificação de dados.

Além do tamanho das sequências, os valores referentes aos percentuais de Identidade e Similaridade retornados pela ferramenta BLAST (respectivamente os campos Identities e Positives) também foram julgados relevantes porém com necessidade de uma melhor adequação computacional para sua futura utilização como característica classificatória. Apenas como esclarecimento pontual, o campo de Identidade se refere ao total de aminoácidos idênticos que puderam ser alinhados entre as sequências objeto analisadas. Já o campo Similaridade refere-se à quantidade total de aminoácidos alinhados em conformidade com a matriz de substituição selecionada como parâmetro, antes da execução da ferramenta BLASTP, levando-se em consideração, neste caso as trocas de aminoácidos previstas pela matriz escolhida. No caso desta pesquisa, como já apresentado anteriormente, a matriz de substituição escolhida foi a BLOSUM62.

A primeira medida tomada com relação aos valores percentuais de Identidade e Similaridade encontrados em um alinhamento foi a normalização dos mesmos baseada no tamanho da sequência consulta (*query*).

Por padrão, o algoritmo BLAST retorna os valores dos percentuais de identidade e similaridade baseados no tamanho total do alinhamento realizado.

A Figura 10 apresenta um pequeno trecho de um arquivo tipo texto simples retornado pela execução automática da ferramenta BLASTP.

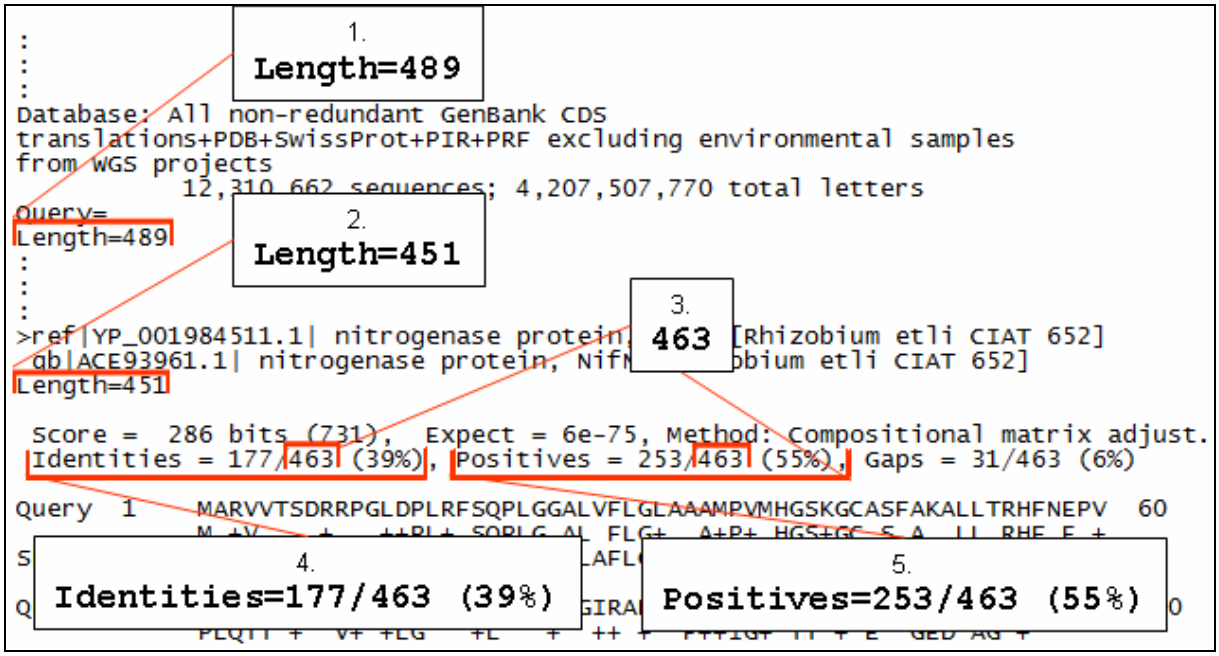


FIGURA 10 – TRECHO DE UM ARQUIVO TEXTO SIMPLES RETORNADO PELA FERRAMENTA BLASTP.
FONTE: A autora, (2011).

Na figura há cinco campos em destaque:

- Número 1: Correspondendo ao tamanho total da sequência consulta *query* passada como parâmetro para a pesquisa.
- Número 2: Indicando o tamanho total da sequência resultado (*subject* ou *hit*) retornada pela ferramenta BLASTP.
- Número 3: Apresentando o tamanho total do espaço do alinhamento.
- Número 4: Correspondendo ao valor encontrado e percentual correspondente à Identidade do alinhamento.
- Número 5: Indicando o valor encontrado e percentual correspondente à Similaridade do alinhamento.

Como dito, a ferramenta BLASTP retorna como resultado um valor percental para os campos Identidade e Similaridade baseando-se na divisão do valor de cada um destes campos (ou seja, 177 para o campo Número 4 e 253 para o campo Número 5), pelo tamanho total do alinhamento descrito no campo Número 3 (no caso 463), resultando nos valores de 39% para o campo Identidade e 55% para o campo Similaridade.

O grande problema deste resultado é o fato de que o tamanho do espaço total referente a cada um dos alinhamentos é extremamente variável, podendo inclusive ser maior que o tamanho total das sequências *query* e *subject*, não fornecendo, desta forma, uma medida normalizada, que possa ser utilizada como uma futura característica classificatória.

Mediante a normalização dos valores de Identidade e Similaridade pelo tamanho da sequência original de consulta (*query*), através de um script específico (detalhado em Resultados e Discussão), pode-se conseguir medidas com melhor padronização, expressando de maneira mais clara o valor real de cada campo em estudo.

Aplicando esta abordagema os dados relativos ao caso apresentado no exemplo da figura 10 ao invés de valores de 39% para Identidade e 55% para Similaridade, passa-se a ter 36% e 51% respectivamente.

Com os valores de Identidade e Similaridade normalizados pelo tamanho da sequência original de busca, foi gerada uma média aritmética simples entre eles e o valor resultante foi armazenado como característica discriminatória de dados.

3.7.4. GRUPO IV – MATRIZES DE CO-OCORRÊNCIA

Mediante a observação e análise constante dos registros obtidos dentro do processo de mineração de dados e visando a maximização dos resultados alcançados durante o processo de classificação informações através de Rede Neural Artificial, duas novas características ancoradas nos conceitos de visão computacional foram incorporadas aos grupos já existentes em caráter experimental:

- Energia da Sequência;
- Entropia da Sequência.

Em 1979, Haralick (HARALICK, 1979) propôs uma metodologia para descrição de texturas em imagens com base em estatísticas de segunda ordem, em que são definidas as características provenientes do cálculo de matrizes denominadas “matrizes de co-ocorrência”. As matrizes consistiam de uma contagem do número de combinações diferentes de níveis de cinza que poderiam ocorrer em uma imagem, em uma determinada direção. Para a obtenção de tais matrizes,

considerava-se a variação da distância e direção (d, θ) entre pixels vizinhos, sendo normalmente utilizados quatro diferentes direcionamentos: 0º, 45º, 90º e 135º.

As matrizes de co-ocorrência formam a base para elaboração de diversas medidas estatísticas conhecidas como descritores de Haralick (ALVES et al., 2006).

Foram utilizados nesta pesquisa, sob a forma de características experimentais, dois dos descritores apresentados por Haralick (HARALICK, 1979): Energia, ou, Segundo Momento Angular que avalia a uniformidade textural em uma imagem, e a Entropia, que mede a desordem em uma imagem, ou seja, seu grau de dispersão de níveis de cinza. Respectivamente suas fórmulas são:

$$Energia = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \left[P(i, j, \delta, \theta)\right]^2 \tag{3}$$

$$Entropia = -\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} P(i, j, d, \theta) \log_2 \left[P(i, j, \delta, \theta)\right] \tag{4}$$

Ambas as medidas foram contextualizadas, convergindo para o problema alvo deste trabalho, de modo que as matrizes de co-ocorrência refletissem a realidade encontrada nas sequências de proteínas expressas pelos genes *nif*, resultantes do processo de mineração de dados. Desta forma os descritores de Energia e Entropia passaram a medir, respectivamente, os graus de uniformidade e desordem de cada uma das sequências estudadas.

Foram geradas matrizes de co-ocorrência de 3X3 aminoácidos baseadas nas sequências estudadas.

Para cada registro processado, a sequência era lida no sentido da direita para a esquerda, amazenando-se os arranjos de 3 em 3 aminoácidos.

Com base nesta lista, as combinações em pares, relacionadas aos arranjos verificados eram analisadas uma a uma, e em caso de co-ocorrência, a contagem e o registro dos dados eram realizados, sendo estes armazenados em uma estrutura de dados Python denominada Dicionário[2] (dic).

[2] Coleção desordenada de dados onde os itens são armazenados e buscados a partir de uma chave única (LUTZ E ASCHER, 2007).

A Figura 11 apresenta um exemplo da aplicação da técnica em uma sequência fictícia.

1) Sequência exemplo: "**LMLALMAL**"

2) Lista de Arranjos possíveis: LML – MLA – LAL – ALM – LMA – MAL

3) Combinações possíveis entre os Arranjos encontrados:

ALM-ALM	LAL-ALM	LMA-ALM	LML-ALM	MAL-ALM	MAM-ALM
ALM-LAA	LAL-LAA	LMA-LAA	LML-LAA	MAL-LAA	MAM-LAA
ALM-LMA	LAL-LMA	LMA-LMA	LML-LMA	MAL-LMA	MAM-LMA
ALM-LML	LAL-LML	LMA-LML	LML-LML	MAL-LML	MAM-LML
ALM-MAL	LAL-MAL	LMA-MAL	LML-MAL	MAL-MAL	MAM-MAL
ALM-MAM	LAL-MAM	LMA-MAM	LML-MAM	MAL-MAM	MAM-MAM

4) Representação gráfica da matriz de co-ocorrência:

	LML	MLA	LAL	ALM	LMA	MAL
LML	0	0	0	1	0	0
MLA	0	0	0	0	1	0
LAL	0	0	0	0	0	1
ALM	1	0	0	0	0	0
LMA	0	1	0	0	0	0
MAL	0	0	1	0	0	0

5) $Energia = \sum_{i=0}^{n-1} \sum_{j=0}^{n-1} \left[P(i,j,\delta,\theta)\right]^2$ => **1.0**

6) $Entropia = -\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} P(i,j,d,\theta) \log_2 \left[P(i,j,\delta,\theta)\right]$ => **0.0**

FIGURA 11 – EXEMPLO DA TÉCNICA APLICADA PARA A GERAÇÃO DAS MATRIZES DE CO-OCORRÊNCIA.
FONTE: A Autora, (2011).

3.8. PROCESSO DE CLASSIFICAÇÃO DOS DADOS

Com base nas características extraídas e na análise subjetiva das informações, levando também em consideração o registro da anotação de cada uma das sequências resultantes do processo de *data mining*, os dados foram manualmente dispostos em duas classes distintas:

- **0** ou **"Não *nif*"** – Correspondendo a registros que não se tratavam de genes *nif*.
- **1** ou **"*nif*"** – Correspondendo a registros que se tratavam de genes *nif*.
-

Durante a primeira etapa do processo de classificação, apenas uma amostra significativa dos dados foi analisada.

Dos 14988 registros resultantes do processo de mineração de dados, 4500 foram classificados, correspondendo a cerca de 30%. Destes, 3165 (70,33%) correspondiam à classe 0 ou "Não *nif*" e 1335 (29,67%) pertenciam à classe 1 ou "*nif*".

Após esta primeira etapa de classificação e aprendizado da rede neural artificial, foi aplicado o processo de co-aprendizado da RNA descrito primeiramente por Cardoso e Cruz (CARDOSO E CRUZ, 2007) e mellhor detalhado adiante, de forma a maximizar os resultados obtidos durante o aprendizado da rede neural.

Concluído o processo de co-aprendizado, todos os 14988 registros puderam ser devidamente classificados.

3.9. REDE NEURAL ARTIFICIAL MODELO *FREE ASSOCIATIVE NEURONS* (FAN)

O modelo Free Associative Neurons (FAN) é composto por unidades independentes, com capacidade autônoma de processamento. O modelo se baseia em aprendizagem conexionista, modelagem difusa e representação de padrões. FAN ganha em termos de inexatidão por trabalhar com granularidade de informação sendo capaz de incluir métodos diferentes de associação de padrões para aumentar capacidades de aprendizagem (RAITTZ, 1998).

Segundo Raittz (RAITTZ, 1998), em FAN, cada padrão de entrada é expandido em uma vizinhança difusa em torno dele. Cada conjunto de vizinhança difusa é uma combinação de valores de características próximas às originais. A imprecisão mede o grau de similaridade entre o vizinho difuso e o padrão de entrada original. A aprendizagem ocorre com a projeção de todo o "bairro difuso" para o espaço FAN. Existe uma unidade FAN para cada classe no domínio do problema. Cada unidade é uma grade com todas as combinações de característica observadas em sua classe correspondente. Durante o treinamento, cada combinação de característica é representada por uma célula difusa, ponderada de acordo com sua frequência de ocorrência. O treinamento é baseado em reforço (se a classificação foi correta) ou em esquecimento (se houve um erro de classificação).

3.10. SOFTWARE EASYFAN

Para o processo de classificação dos dados foi escolhida a aplicação EasyFan[3] pelo fato de fornecer suporte ao aprendizado supervisionado e permitir a visualização gráfica das características e neurônios da rede treinada.

O processo de aprendizado da rede neural artificial FAN ocorreu de acordo com a especificação do software EasyFan, que necessita de uma amostra pré-classificada de dados, dividida em três conjuntos: Treinamento, Teste e Validação.

Devido a opção pelo uso da técnica de co-aprendizado (CARDOSO e CRUZ, 2007) o software foi utilizado em dois momentos: durante a classificação prévia de dados e após o processo de co-aprendizado e reclassificação.

Durante as duas etapas, a fase de Treinamento foi executada por 12 horas, na tentativa de maximizar os resultados de aprendizado.

3.11. PROCESSO DE CO-APRENDIZADO DA REDE NEURAL ARTIFICIAL

Após a análise da primeira classificação das informações realizada pelo software EasyFan, baseada numa amostra significativa de dados, deu-se início ao processo de co-aprendizado da rede neural.

[3] Implementação da Rede Neural FAN desenvolvida em linguagem Java, por Lenfers e colaboradores (LENFERS et al., 2006).

Segundo Cardoso e Cruz (CARDOSO e CRUZ, 2007) a classificação subjetiva dos dados, realizada após a extração de características pode apresentar incoerências, pois o padrão apresentado no início da classificação tende a sofrer sutis modificações até o término dessa, penalizando inclusive o aprendizado da rede. Assim sendo, é possível que melhorias significativas na classificação sejam implementadas através da análise e comparação dos resultados obtidos durante o uso da rede com a classificação subjetiva original.

Esta pesquisa partiu de uma classificação inicial correspondente a cerca de 30% dos dados totais para a classificação total dos dados obtidos (14988 registros) em apenas um ciclo de co-aprendizado. Houve a tentativa de repetição de mais um ciclo de co-aprendizado, porém sem melhora efetiva nos resultados obtidos com o uso da rede neural. Maiores informações a respeito deste processo serão apresentadas em Discussão.

4. RESULTADOS E DISCUSSÃO

O alcance das metas propostas nesta pesquisa somente foi possível graças ao desenvolvimento de uma metodologia de mineração de dados correspondentes a sequências depositadas no NCBI GenBank, sendo esse o primeiro resultado alcançado.

A metodologia desenvolvida baseia-se na aquisição, interpretação, classificação e organização de dados obtidos através da execução da ferramenta BLASTP, dentro da parametrização explanada na seção Metodologia.

A Figura 12 apresenta as etapas estipuladas e alcançadas durante o processo de pesquisa:

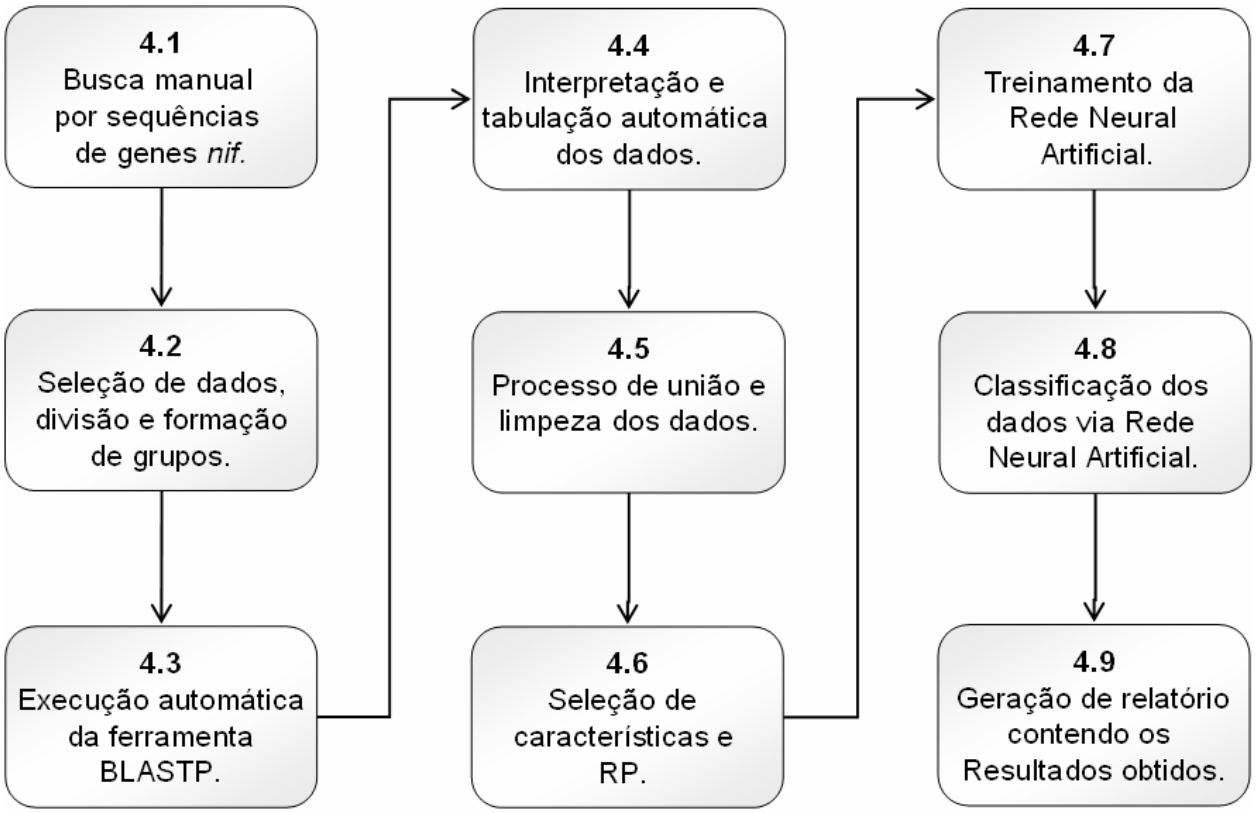


FIGURA 12 – METODOLOGIA BASEADA EM RESULTADOS BLAST, PARA MINERAÇÃO DE DADOS DO NCBI GENBANK.
FONTE: A autora (2010).

Os resultados obtidos em cada uma das etapas serão descritos em detalhes nas próximas seções.

4.1. BUSCA MANUAL POR SEQUENCIAS DE GENES *nif*

Foi realizada uma busca manual no banco de dados Protein, componente disponível no site do NCBI GenBank (http://www.ncbi.nlm.nih.gov), buscando pelas *strings:* "*nif*" e "gene".

Foram encontradas 168 sequências, divididas conforme exibido abaixo, na Tabela 5:

TABELA 5 - GENES *nif* ENCONTRADOS EM PESQUISA MANUAL REALIZADA.

Quantidade de Genes *nif* Encontrados em Pesquisa Manual																				
Gene nif	A	B	D	E	F	H	J	K	L	M	N	O	Q	S	T	U	V	W	X	Y	Z	Total
Qtde	9	9	9	9	9	9	9	9	9	7	9	6	5	9	5	9	9	8	7	5	8	168

FONTE: A autora, com base em pesquisa realizada no dia 23/03/2010 (2010).

4.2. SELEÇÃO DE DADOS, DIVISÃO E FORMAÇÃO DE GRUPOS

As ocorrências encontradas na busca manual foram agrupadas segundo sua taxonomia. Foram encontrados nove grupos taxonômicos distintos que agregavam os dados encontrados:

- Archaea
- Actinobacteria
- Chlorobium
- Firmicutes
- Cyanobacteria
- Alphaproteobacteria
- Beta proteobacteria
- Delta/Epsilonproteobacteria
- Gammaproteobacteria

A Tabela 6 mostra o resultado deste agrupamento trazendo o ID Genbank de cada uma das sequências.

TABELA 6 – MAPEAMENTO TAXONÔMICO DOS GENES *nif* ENCONTRADOS EM PESQUISA MANUAL.

Mapeamento de Genes *nif* X Taxonomia									
Gene *nif*	**Archaea**	**Actinobacteria**	**Chlorobium**	**Firmicutes**	**Cyanobacteria**	**Proteobacteria**			
						Alfa	**Beta**	**Delta/Epsilon**	**Gamma**
A	YP_304373.1	AAK29059.1	YP_001130856.1	ABO48894.1	ACB50848.1	AAG61008.1	AAA25027.1	YP_001380213.1	CAA61477.1
B	AAM30455.1	AAD17269.1	NP_662424.1	CAC27790.1	AAA93017.1	CAE30070.1	ACC76628.1	ACG74423.1	ACH83483.1
D	AAK33115.1	ABD13833.1	ABP37358.1	BAH24184.1	AAN63667.1	ACT67950.1	CAA90933.1	CAE10458.1	ACU32753.1
E	AAC45517.1	ABW16212.1	ACD89763.1	BAD95756.1	ABC99985.1	CAA40213.1	CAP64027.1	ABQ25405.1	ACH83473.1
F	NP_635211.1	YP_003153508.1	ACD89838.1	ABR49154.1	ABD00720.1	AAC05792.1	YP_933785.1	ABA88115.1	EFD86060.1
H	ACL15934.1	AAU00812.1	ACF13499.1	BAD95753.1	CAA24729.1	ACY38122.1	EFG73464.1	EET35459.1	ACO76403.1
J	ADG13989.1	EEZ60798.1	ACF45587.1	YP_001511796.1	EFA68677.1	ADE85473.1	YP_286040.1	ABS24454.1	CAA55302.1
K	AAK33116.1	AAD17261.1	YP_001130861.1	BAH23273.1	YP_324430.1	EFH04964.1	ACB33700.1	ACG74426.1	AAU90624.1
L	CAJ35150.1	ZP_06264124.1	ACD90870.1	EEU00690.1	ACC82181.1	CAM74810.1	AAM90968.1	NP_906990.1	ACU32739.1
M	CAJ52798.1	CBL03920.1		EFH06627.1		CAL74525.1	ACV37615.1	NP_953123.1	AAA25105.1
N	YP_686182.1	ABW16211.1	NP_662423.1	AAX73209.1	BAF47155.1	ACE93961.1	CAP64028.1	ADG91738.1	ACU32728.1
O			ACF44312.1		EAW33916.1	ACB94124.1	ADE11074.1	ABA87771.1	ACR13122.1
Q		ABK73085.1				AAG60755.1	AAZ46197.1	ABA87575.1	YP_112774.1
S	AAZ69707.1	AAD17272.1	ACF15138.1	CBL26306.1	AAA22006.1	CAA48487.1	ADE11045.1	ADH85543.1	CAA68020.1
T					EAZ92885.1	ACJ01036.1	ABE34510.1	YP_003654208.1	AAA64712.1
U	AAZ71345.1	AAD17270.1	ACE05136.1	EEG78998.1	ABA23845.1	AAA22184.1	ADE11044.1	ACS78205.1	ADC62337.1
V	ACX72384.1	CAJ65437.1	ABB28490.1	EEP52475.1	EDX84946.1	AAA26138.1	ACB33634.1	ABL01059.1	ABM61073.1
W	NP_377560.1	AAD17267.1		ZP_06409757.1	EAW46997.1	ACK52502.1	AAZ46144.1	YP_003654235.1	AAA64730.1
X		ABW16210.1		BAD95758.1	ACC79168.1	AAG27067.1	ABM37631.1	EET35455.1	ACH83471.1
Y					YP_001658823.1	AAB02346.1	AAG35590.1	NP_906473.1	EFD86070.1
Z		AAC82974.1	YP_001997174.1	YP_001422241.1	EAM50528.1	BAC47026.1	ABO58924.1	CAE10450.1	AAU90667.1

FONTE: A autora, com base em pesquisa realizada no dia 23/03/2010 (2010).

Devido ao fato de terem sido encontradas lacunas para genes *nif* em alguns grupos taxonômicos e na tentativa de se manter uma maior diversidade nos dados que seriam utilizados como fontes de busca, foram selecionadas apenas as sequências relacionadas aos genes *nifHDKEN.* Segundo Raymond (RAYMOND, 2004), estes genes são universais em organismos fixadores de nitrogênio. Além dos nove grupos taxonômicos apresentados acima, um décimo agrupamento contendo sequências randomicamente selecionadas dentre as encontradas foi montado de forma a maximizar os resultados obtidos com a execução da ferramenta BLASTP.

A Tabela 7 apresenta a relação Grupo Taxonômico X Organismos (Org.) X Sequências (Seq.) X Genes *nif* (*nif*) selecionados para a continuidade do processo de mineração de dados:

TABELA 7 – SEQUÊNCIAS SELECIONADAS PARA A CONTINUIDADE DO PROCESSO DE MINERAÇÃO DE DADOS. (**continua**)

Archaea		
nif	Org.	*Methanosphaerula palustris E1-9c*
H	Seq.	MKRQVAIYGKGGIGKSTTTQNTVAALAEQGKKIMVVGCDPKADSTRLLLHGLCQKTVLDTLRDE GDDIELDEILKPGYRDTLCVESGGPEPGVGCAGRGIITSINLLESLGAYTDDLDYVFYDVLGDV VCGGFAMPIREGKAEEIYIVASGELMALYAANNIAKGIHKYAVSGKVRLGGIICNSRQVDNEYP LLKAFAEELGSQLIYFVPRDNLVQRAEINKKTVIDFEPESNQAEEYRQLAKAIDENRMFVIPKP MTQDRLEELMMQHGFMDALV
D	Org.	*Methanosarcina mazei Go1*
	Seq.	MGAEINSEIGDKQQLVENMMKALPEKAARNRRKHIVVRDCSTEQHIEADDKVIPGILTNRGCAF AGTKGVVFGPIKDMVHIVHGPIGCAFFTWGTRRNFAKAEEGEDNYMNYCVCTDMKETDIVFGGE KKLKKAIDEVVKIFNPEAVTICATCPVGLIGDDIEAVSREAEKEHGIKVIPARCEGYRGVSQSA GHHIASNALMENLIGTEDIKNPTPFDINIFGEYNIGGDLWEIKPILEKIGYRIVSTFTGDGSFH KLAQAHRAKLSILLCHRSINYTNRMMEEKYGIPWLKVNYIGTKATEKSLKKMAEFFDDPEITRK TEEIIAEETAKYRDEIEKYRKKLQGKTAFIYAGGSRSHHYTNLFEELGMKVVVAGYQFAHRDDY EGRQIIPHMQEKALGSILEDIHYEKDENVKPAVSPERIEELKKKIGLMDYEGLFPEMKDGTIVI DDLNHHETEALLKTLKPDIFCSGIKDKYWAQKHGIPSRQIHSYDYSGRYTGFSGVVNFARDIDM AMHSPTWRFIRPPWKGEGEE
K	Org.	*Methanosarcina mazei Go1*
	Seq.	MLDYTPCEEITREAVTINPAKICQPIGAVYAALGVHNCMPHSHGSQGCLSYLRMCLSRHYRENP VATTSSFSEGTAVFGGAANLKEALGNLTAIYKPEVIAIHTTCVAETIGDDVGVIIEDVRAEELI DPSIKICAASTPSYVGTHITGYDNMVKSFVTIFARKTKPNGKLNLIPGFVEPGDIREIKRILSI MGIPNIVFPDTTDVFDVPLTGEPGMYPKGGTPIGDIEDSASSAGTIALCRMAGGSAAKILESKF KVPAKIGPTPIGIRNTDRFIMNAAKLANVAIPLELEDERGRLVDMMTDAHPHYHGKKVAIYGDP DIISGLTSLVMEMGMEPGVVLTGTQSSEFEKEVEGLTGSEYPEADIISGGDLFMLHQIIKRKPV DLLIGNTYGKFISRAEDVPLVRVGFPIMDRASLHYFPIMGYAGAARLVERIGNTLLDRKDRDAP DWLLETIQ

TABELA 7 – SEQUÊNCIAS SELECIONADAS PARA A CONTINUIDADE DO PROCESSO DE MINERAÇÃO DE DADOS. (**continuação**)

Archaea		
E	Org.	*Methanococcus maripaludis*
	Seq.	MVLNLDTENRKMQDGNNDDDFDLEVQIPNSIFEKLKSIEALKARESPMCVSGKDDSIPTCDQNS TPGMITQRSCVYGGARVVLMPITDAVHLVHGPIGCAACTWDIRGSKSTGDKLYKNGFSTDLQEK DIVFGGEKKLYESILEVNKLYHPGAIFVYSTCVVGLIGDDLKAVCRQAQEATGCRVIPVQSEGF KSFNKTAGHKLACDAMLDYVIGTEEPEEEHPYSINIIGEFNVAGDLWGIIPLYEKMGVKVHTAI TGDSTVAKVASAHRSKLNIVQCPKSSNYLAAQMDKKYGIPSFKVNFFGLDETTKSLRAVAEFFG DEEMIKRTEELIKSEIKNLRDEISEYQKDLSGRTVAIYSGAHKSWALVSAFGELDMEIIMSGTQ NGKPEDYQQIRDHVCEGTLIVDDASSMELVQLLKEYKPDILISGAKEKYLSLKSGIPHCDFNHD RITAFSGYQGFINFARVVHTAVMTPIWRLSRKMMI
N	Org.	*uncultured methanogenic archaeon RC-I*
	Seq.	MTDKVRMVNENQCQTCMPLGGVIAFKGVEGAMALVHGSQGCSTYMRLASVEHYNEPIDIASSSL NEKQTIHGGEGNLRKALDNVLRIYQPAVLGVVTTCLTETIGEDIQRMIKGYVDQRSIVDQRSVV ERGLKGVDIIPVNTPSYAGSQSEGFWAAVRAIVEYFARPAEKHGGINVIVPHISPADIREIKRI LDLAGIEYTLLPDYSMTLDRPYGGQYTRIAPGGTRTADIARMAGARATIQFGLTCPEELSPGRY LQQEYGVPLINLPLPIGLENTDRFVETLSQLTGMPLPETLVVERGWLLDAMADTHKYNAEGRPV IYGEPELVYAFASVCAENGAYPAVVATGTRSSRLTARLKLLLAEADRQCVLLEETDFTAIESAA VSTAANIAVGHSGGKFLTERRGIPVVRMGFPIHDRIGGQRILSAGYAGTLAFLDRFTNALLEQK YASYRQLKREQLLDPISIEGDQ
Actinobacteria		
H	Org.	*Frankia sp. ACN14a*
	Seq.	ILHSKAQTSVIQLAAEKGSVEDLELDEVLVEGQWGIKCVESGGPEPGVGCAGRGVITSITYLEE AGAYEDLDFVTYDVLGDVVCGGFAMPIRQGKAQEIYIVTSGEMMAMYAANNIARGILKYAHSGG VRLGGLICNSRKTDREDELIMELARRLNTQMIHFIPRYNVVQHAQLRRMTVIELHRKNSQADEY R
D	Org.	*Frankia sp. CcI3*
	Seq.	MTTTPAPIKAETEAMIAEVLSNYPAKTAKFRAKHLKANDPDGSKECEVKSNIKSRPGVMTIRGC AYAGSKGVVWGPVKDIVNISHGPVGCGQYSWATRRNYARGPWGVNNFAAMQFTTDFQEKDIVFG GDPKLETVCDEIVELFPLAKGISVQSECPIGLIGDDIEAVSRKASKKLELPVVPVRCEGFRGVS QSLGHHIANDAVRDHVLGTGGNTFKQTPYDVALIGDYNIGGDAWASRKILEEMGLRVIAQWSGD GTINEMASTHLSKLNLIHCYRSMNYICTTMEERYGTPWTEFNFFGPTKIVDSMRKIAAFFDDEI KQKTEAAIARYQVRFDEITNAFRPRLEGKRVMLAVGGLRPRHTIGAYEDLGMEVVGTGYEFAHK DDYTRTYPALAEGVVLYDDPTAFELEEFAKRLKPDLMGAGVKEKYVFHKMGVPFRQMHSWDYSG PYHGVDGFAVFARDMDIAINSPTWDLMETPWSKAGEVF
K	Org.	*Frankia sp. EuIK1*
	Seq.	MTTTPETGSSIPFRVLDHSELFKDVVYQKQFEGKAEFENGSSSAEVARVLEWTRGWEYREKNFA REALTVNPAKACQPLGAVLAGLGFQGTLPLVHGSQGCVAYFRSHFARHFKEPVPAASTSMTEDA AVFGGLNNLVEALENATRLYKPKMVAVSTTCMAEVIGEDLFAYIGAAKEKGAISTDYPVPYAHT PSFVGSHITGYDSMLKGILRNLTKSADRTEPAAGGKPRLNIIPGFETYTGNLREYRRVLELMGV DPLILGDHSDSLDSPADGEYDLYPGGTPLADAAKAKFSRATVLLQESATRKTTELIRDVWKQDT LVLETPIGPGGTDQFLTEISRLAGVEIPAELTVERGRLVDALTDSHAYLHGKRVAIAGDPDLVV ALTRFVLELGMIPVHVLSTNADTSFKARMEKVLSASKFGEAATVWPEKDLWHLRSLVFTEPVDL LIGSTYLKYIAREANVPLVRVGFPIFDRHHLHRFPIIGYTGRAALVTQLVNTVLDELETAADHS YDAVR
E	Org.	*Frankia sp. EAN1pec*
	Seq.	MAATDRATLFTEPACDHNREKSAKERKAGCPKPAPGGTSGGCTFDGAMITLVPIVDSAHVVHGP IACAGNSWDGRGSLSSGPDLYRRGFTSDVGEQDVIFGGEQRLFDTILEAVRRHHPPAVFVYSTC VTAMIGDDLDAVCSAAAEHTGVPVIPVHAPGFAGNKNLGNRLAGEALLEHVIGTVEPPDVTELD VNLVGEYNIAGELWDVLPVLAKMGIRVRACISGDARYADVAAAHRARATMVVCSRALLGLARGL EDRYGIPWFEGSFYGVRAMNDTLREFARLLGGAELARRAEEVIALEQTAVDLALEPYRERLAGR RAVLYTGGVKSWSIVSALQDLGIEVVANGITKSSDGDVEKIRELLGPDARIVSEGSPRELLRIA EETRADILVAGGRNQYTALKGRLPFLDINQERHIPYAGYRGAVELARRLDMALSNPVWEQVRAP APWDVEGVA

TABELA 7 – SEQUÊNCIAS SELECIONADAS PARA A CONTINUIDADE DO PROCESSO DE MINERAÇÃO DE DADOS. (**continuação**)

Actinobacteria		
N	Org.	*Frankia sp. EAN1pec*
	Seq.	MARVVTSDRRPGLDPLRFSQPLGGALVFLGLAAAMPVMHGSKGCASFAKALLTRHFNEPVPLQT TGVTEVSAVLGSGDDLVANLDGIRAKQNPRIIGLLTTGVTEVSGEDVAGQVRQYIAMMNHTTPE GAPLIVRVSTPDFAGGLSDGWSAALRSLVATVPFDHADSDEYPGTRSGFGAGTGSAPETVAVLV GPSLSAADLDELCALIRSFGMAPVLVPDLSGSLDGHLAPSWQPTTTGGTGLAQLRRLDEAGLII TAGATAAEAGVDLAARTAADLVQHDHLSGLAAVDSLVAELMTRSGRGPAPEVRRARARLADGLL DTHFVLGGARIALAMEPEALVAVGSLLHDVGAEIVAAVSPTDAPVLATAPWDEIVIGDLTDLEE RALEGGAELLIGSSHVRTVADRIGAAHLAVGFPIYDRLGSALRTTAGYGGSLRLLVDAANRLLD HHQADHQANHRADHRPGRHDVREHPLDSFDQLDVLCQESPC
Chlorobium		
H	Org.	*Chloroherpeton thalassium ATCC 35110*
	Seq.	MRKVAIYGKGGIGKSTTTQNTVAGLAEMGKKVMVVGCDPKADSTRLLLGGLIQKTVLDTLREEG EDVELDDIIKEGYSATRCVESGGPEPGVGCAGRGIITSVNLLEQLGAYDDEWNLDYVFYDVLGD VVCGGFAMPIRDGKAEEIYIVVSGEMMAMYAANNICKGILKYADAGGVRLGGLICNSRKVDNEQ EMIQELARQLGTQMIHFVPRDNMVQRAEINRKTVIDYDPTHSQADEYRTLAKKIDENEMFVIPK PLEIDALEKLLVDFGIAN
D	Org.	*Chlorobium phaeovibrioides DSM 265*
	Seq.	MESKTQFPDPSQVREELIEKYPAKVAKKRNKSIIINDPETVPEVQANVRTVPGIITQRGCSYAG CKGVVLGPTRDVVNITHGPIGCGFYSWLTRRNQTKPETDADANFITYCFSTDMQEENIVFGGEK KLKVAIQEAYDLFHPKAIAIFSTCPVGLIGDDVHAASREMKEKFGDCNVFGFSCEGYRGVSQSA GHHIANNGIFKHMVGRNNEIKEGKFKLNLLGEYNIGGDAFEIERIFNKCGLTINSSFSGNSTVG QLENAHMADLNAIMCHRSINYMGDMLETKYGIPWMKINFIGAESSAKSLRKIAEYFGDAELKAR VEEVIKEEMPKVKEVIDEILPRTSGKTAMLFVGGSRAHHYQDLFSELGMTTVAAGYEFAHRDDY EGREVLPNIKVDADSKNIEELKIEADPELYKPRKTEAELEELKAKGLEINSYEGMMKQMMKKSL VVDDISHYESEKLIEIYKPDIFCAGVKEKYVVQKMGVPLKQLHSYDYGGPYTGFEGAVNFYRDI DRMVNNPVWKLIKAPWQKTADSEKESLSASYVTQ
K	Org.	*Prosthecochloris vibrioformis DSM 265*
	Seq.	MLLRHTTKEVKVREGLTINPAKTCQPVGALYAALGIHGCLPHSHGSQGCCAYHRSMLTRHYKEP VMAATSSFTEGASVFGGQANLLASIDTIFSVYDPDIIAVHSTCLSETIGDDLQQITKKAKDDGK IPEGKHVIYASTPSFVGSHVTGYANMVVGMAEQFAESTGEKKEQVNLIAGWMEPSDMREIKKIS KELGARIVLFPDTSDVLDAPQTGKHEFYPKGGVTIPELKSTGDSLSSIALGCISAEPAAVALDK KFGVPFETVDMPIGISATDRFIMALSKAAGVSVPDEITSERGRLLDAMTDMDQYFYGKKVALFG DPDQLIPLTEFLLDLGMLPIHIVSGTPGKRFEKRMAEILERSPGANFKNGPGADTFLLHQWMKN EPVDLLIGNSYGKYIARDENIPFIRFGFPIVDRIGHSYFPSVGYSGGLRLVEKILSVFMDRQDR EAPEEKFELVM
E	Org.	*Chlorobium limicola DSM 245*
	Seq.	MKEEIGILEGRQGQVFEKKSGEAEQLDISCEKTSLSGSVSQRACVFCGSRVVLYPVADALHLVH GPIGCAAYTWDIRGAVSSGPELHRLSFSTDLGEMDVIYGGEKKLYLSLIELIDKYKPKAAFIYS TCIIGLIGDDIDAVCKKVSKETGIPVLPVHSEGFKGTKKDGYKAACTSLMKLVGTGSIEGISPY SINILGEFNLAGEAWIIREYYEKMGIEVVSTMTGDGRVDAVRRAHGATLNVVQCSGSMTTLAKE MEEKYGIPYMRVSYFGIEDMSKSLYDVAKHFSDRPDIMDAAKEIVSKEVAKLYPELQKFKKVLA GKKAAIYVGGAFKTFSLIKALRSIGMSVVLAGSQTGNKDDYERLREMCDEGTIIVDDSNPVELS KFVLEKEADLLIGGVKERPIAYKLGIGFCDHNHERKIPLAGFIGMYNFAKEVYQSVMSPVWQFA PRKGGKI
N	Org.	*Chlorobium tepidum TLS*
	Seq.	MTTNASAMTAKTATQNACKLCTPLGACLAFRGIESCVPFLHGSQGCATYIRRYLISHYKEPIDI ASSNFNEETAVFGGSHNLQLGLKNVTQQYKPQVIGIATTCLSETIGDDVPMILKDYKAIMNDPN LPTMIFASTPSYSGSHIDGFHTAVRSAVKTFAVGGAKKNLLNLFSGMISPADIRYLKEILKEFG MPFMLLPDYSQTLDGGPWGEYHRIPPGGTPTSAIADSGSAAASIEFGSTLEAKKSAAGYLEAEF GVPRHQLPLPIGIKATDRFFALLEELTEKPMPEKYEDERRRLVDAYADGHKYIFGRKAMLYGEE DLVISMAAFLREIGVVPVLCASGGKSGQMKQRMLELIPDMDEQGIEACEGVDFVDIEHEAERLK PDMLIGNSKGFTMSKKHELPLIRIGFPIHDRFGGQRLHHLGYRGTQELFDRIVNTVIEERQKSS PIGYTYM

TABELA 7 – SEQUÊNCIAS SELECIONADAS PARA A CONTINUIDADE DO PROCESSO DE MINERAÇÃO DE DADOS. (**continuação**)

Firmicutes		
H	Org.	*Heliobacterium chlorum*
	Seq.	MRQIAIYGKGGIGKSTTTQNTVSALAEMGKKIMIVGCDPKADSTRLILHSKAQATVMDLAREKG TVEDLELEDVLLTGFADIRCAESGGPEPGVGCAGRGVITAINFLEENGAYTPDLDYVFYDVLGD VVCGGFAMPIRENKAQEIYIVTSGEMMAMYAANNIARGILKYAATGKVRLGGLICNSRKTDKEY ELIAELASRLGTQMIHFLPRDNQVQRAELRRMTVIEYSPEHPQADEYRALAKKIDENKMLVVPT PLSMDELEDLLISYGILEDEETAMAKLG
D	Org.	*Paenibacillus abekawaensis*
	Seq.	MGLDIEANKKLVEELLEIYPKKARKDREKHYQINTEESKTCGTCGLKSNIKSRPGVMTPRGCSY AGSKGVVWGPIKDMVHISHGPIGCGQYSWGTRRNYANGTLGIDNFTAMQVTSDFQETDIVFGGD KKLEVIMREITEMFPLAKGISVQSECPVGLIGDDIEAVSKKMAKELDMPIVPVRCEGFRGVSQS LGHHIANDAIRDFVLGRAELAETGPYDVNIIGDYNIGGDAWASRILLEEMGLRVIAQWSGDGTL NEFQIAHKAKLNLIHCHRSMNYMVDHMEKAYGIPWMEYNFFGPSKTYESLRAIAALFDEKIQEN CEKMIAKYKPQMDAIINKYRPRLESKKVLLMIGGLRSRHTIGAYEDLGMEIVASGYEFAHKDDY ERTFPMMPEGAIIMDDPTAYELEELAQKLDVDLVASGVKEKYVYHKMGVPFRQMHSWDYSGPYH GFDGFKIFAKDMDMTVNNPVWNMISKKNKVQAEGASI
K	Org.	*Paenibacillus graminis*
	Seq.	MSKDRMEIPDYNSLFSEERYVQQRANKKQFEAPCSEQDTAEALAYSKSAEYMQKNFDRKAVVIN PHKACQPLGSVMAALGFEKTLPFVHGSQGCNSYFRSHLSRHFKEPTPAVSSSMTEDAAVFGGMN NLIDGLENSIALYKPEMIALCTTCMAEVIGDDLSSFIGNARIKGAITEDFPVAFCNTPSFVGSH ITGYDSMLKGILSYLYERSGMKASPGSGAESGEKLNVMLGFEPYTGNFAELRKILDAFDTKYTI LGDHSGNYDSPATGEYEYYYGGTKLADVPLAANALGTLSLQKYTLKKTQDFISSTWNQQLTAMS TPLGITGTDKLLDAISELTGLPIPASLMEERGRVVDALTDSHPYLHGKRVALVGDPDLLIGLIG FCLEIGMEPVHIVCSNGDVDFNEVKFKDEAEALLASSPYGSEAALYIGNDLWHMRSLLMNDPVD LAIGSSHLKFAAKDAGVPLVRVGFPILDRHHMHRYPIIGYQGTLNLLSLLVNTILDELDRNNSG FNYDLVR
E	Org.	*Heliobacterium chlorum*
	Seq.	MFVNLTKLDLKETCPLTDKGAPKLCMKALPGEGAERSCAFDGARVVLMPVTDAAHVVHGPIACA GNSWDNRGAKSSGPQLYRRGLTTDLMEMDVVYGGEKKLIQAIIDVARKHQPKAVFVYATCVSAL IGDDLETACKEAQKDVSVPVIPVNAPGFMGDKNIGNRVAGEVLFRQIIGTAEPETVHPFAINFI GEYNIAGDLWGIQPSLEAIGIKLQACISGDARFDDIRTAHRAKLNVLVCSKSLTNLVKQMEQKY GIPYIEGSFYGIHDTSATLRAIAKAFGDAGLIERTEAHILQKEAETRAVIEPLKAQLSGKRAVL FTGGVKTWSMVTTLRELGIDVLAGGTQNSTPDDFLRMKALMDPTAHIIEDTSAAGFLTLIKEKS PDLIVAGGKTKYLAHKTRTPFLDINHGRSLPYAGYAGMITFANAVARTVFSPVWEKMRVAFPPE VTVNE
N	Org.	*Paenibacillus massiliensis*
	Seq.	MPIRQISKPLSVNPLKVGQPLGGVMALQGVYRSMPLLHGAQGCSAFSKSLLTRHFREPIAVQTS ALQEMDVIFDADRNLEEALDSIWSKHHPDVIGVLSTALTEVAGVDFQSRVKTFKRERSMKDSLL FPVSLPDFQGSLETGYGSTVEAMIDAVIGHVGGRLPKKKRNGQVNLLPGSHLTAGDVMEIKDII ASFGLEVITLPDLSTSLSGHLLTGFSPLTRGGVPLDSMCQMLESSCTIVIGASMERSARRLENS AGIPYQTFAGLSGLAASDEFMCFLQHISRETTPVRFRWQRENLLDSMLDAHFYYMGASAVVALE PDHLLSTADWLGEMGVQLKRLIAPCRTSAILNSSLDIWIGDLDDAEDAAPSADLRISNSHGKKG AMRSDTAFVQAGFPVYDELGAHTSISVGYRGTMSLVNKVGNSLLHGRSRGG
Cyanobacteria		
H	Org.	*Nostoc sp. PCC 7120*
	Seq.	MTDENIRQIAFYGKGGIGKSTTSQNTLAAMAEMGQRIMIVGCDPKADSTRLMLHSKAQTTVLHL AAERGAVEDLELHEVMLTGFRGVKCVESGGPEPGVGCAGRGIITAINFLEENGAYQDLDFVSYD VLGDVVCGGFAMPIREGKAQEIYIVTSGEMMAMYAANNIARGILKYAHSGGVRLGGLICNSRKV DREDELIMNLAERLNTQMIHFVPRDNIVQHAELRRMTVNEYAPDSNQGQEYRALAKKINNDKLT IPTPMEMDELEALKIEYGLLDDDTKHSEIIGKPAEATNRSCRN

TABELA 7 – SEQUÊNCIAS SELECIONADAS PARA A CONTINUIDADE DO PROCESSO DE MINERAÇÃO DE DADOS. (**continuação**)

Cyanobacteria		
D	Org.	*Anabaena cylindrica PCC 7122*
	Seq.	MTPPDDKKIVEQRKELVKEVLKAYPEKAAKKREKHLNVFEEGKADCGVKSNIKSLPGVMTARGC AYAGSKGVVWGPIKDMIHISHGPVGCGYWSWSGRRNYYIGTTGIDTFGTMHFTSDFQERDIVFG GDKKLTKLIEELDVLFPLNRGVSIQSECPIGLIGDDIEAVAKKTSKTIGKPVIPVRCEGFRGVS QSLGHHIANDMIRDWVFPRADQGKKDGTLKFEGTPYDVAVIGDYNIGGDAWASRILLEEIGLRV VAQWSGDGTINEMLMTPNVKMNLIHCYRSMNYISRHMEEAYGIPWLEYNFFGPTKIAASLREIA SKFDAKIQANAEKVIAKYQPTMDAIVSKYRPRLEGKTVAMMVGGLRPRHVVPAFQDLGMKMIGT GYEFAHNDDYKRTTNYIENGTIVYDDVTAYEFEEFIKALKPDLVASGVKEKYVFQKMGLPFRQM HSWDYSGPYHGYDGFAIFARDMDLALNSPTWGLIGAPWNKKAQAKAKAKASV
K	Org.	*Anabaena variabilis ATCC 29413*
	Seq.	MPQNPERIVDHVDLFKQPEYTELFENKRKNFEGAHPPEEVERVSEWTKSWDYREKNFAREALTV NPAKGCQPVGAMFAALGFEGTLPFVQGSQGCVAYFRTHLSRHYKEPCSAVSSSMTEDAAVFGGL NNMVEGMQVAYQLYKPKMIAVCTTCMAEVIGDDLGAFITNSKNAGSIPQDFPVPFAHTPSFVGS HVTGYDNMMKGILSNLTEGKKKATSNGKINIIPGFDTYVGNNREVKRMLGVMGVDYTILSDSSD YFDSPNTGEYEMYPGGTKLEDAADSINAKATVALQAYTTPKTREYIKTQWKQETQVLRPFGVKG TDEFLTAISELTGKAIPEELEIERGRLVDAITDSYAWIHGKKFAIYGDPDLIISITSFLLEMGA EPVHILCNNGDDTFKKEMEAILAASPFGKEAKVWIQKDLWHFRSLLFTEPVDFFIGNSYGKYLW RDTKIPMVRIGYPLFDRHHLHRYSTLGYQGGLNILNWVVNTLLDEMDRSTNITGKTDISFDLIR
E	Org.	*Synechococcus sp. JA-3-3Ab*
	Seq.	MPLISRTKVAELLNETACEHNHKKDGKGKNKVCAEQAKPGGAQGGCAFDGASIALVPIADAAHL VHGPIACSGNSWGTRGSLSSGPQLYRVGFTTDLSENEIIFGGEEKLFQAILDLQAKYQPAAIFV YLTCVTALIGDDVEAVCQRASRQLGIPVIPVNAPGFVGSKNLGNRLAGEALLDYVIGTASPEYT TPRDINLIGEYNIAGELWDVLPLFDELGIRILAKITGDGRYREICTAHRAKLNVLICAKALLNV GRKMQERYGIPYIEASFYGLQEMNHCLRAIAQGLKDPELQERTEALIERETRATEAALAPYRQR LAGKRVVLYTGGVKSWSIISAARDLGMQVTATSTRKSTEEDKARIRELLGQDGILLDKGSPQEL LKVVERTKADLLIAGGRNQYTALKARIPFLDINQERHHAYAGYRGLVNLAREIDAALHSPIWEQ LRRPAPWETSAAKEVGGWPS
N	Org.	*Gloeothece sp. KO68DGA*
	Seq.	MTTILNPNKSLSVNPLKMSQPLGAALAFLGLKGMMPLFHGAQGCTAFAKVVLVRHFRESIPLST TAMTEVSTILGGEDHVEQAILTILEKYEPQIIGLLTTGLTETRGDDMKRILKTVREKHPELDSF PIVSVSTPDYKGSLQDGYTAAVESIVSADYSNVVDDEPPLVGMKPQVTVLCGSYLSPGDVEEIK AVIEGFGLTPVMIPDLSRSLDGHLEDNYHTVTTGGTSLTQLKQLNRSCLTLAIGESMRKAAAIL EKRFGTKYEIFPRLAGLDAVDDFLWKLSQILTARCDHHFPIVPNVPAHLQRQRRQLQDVILDTH FYFGGKKVSLGLEPDLLYQTAWLLTEMGANVQAAVTTTKSPALENLPIDNVTIGDLEDLEELAV GSDLIITNSHGTSLAKRLKAPLYRMGYPVFDQLGLGQRCLVGYRGTMKFLFDVGNILMEEEAKH SPVALH
Alphaproteobacteria		
H	Org.	*Bradyrhizobium japonicum*
	Seq.	TSQNTLAALAELGQKILIVGCDPKADSTRLILHAKAQDTILSLAANAGSVEDLEIEDVIKLGYK DIRCVESGGPEPGVGCAGRGVITSINFLEENGAYEGIDYVSYDVLGDVVCGGFAMPIRENKAQE IYIVMSGEMMAMYAANNISKGILKYANSGGVRLGGLICNERQTDKELELADALAKKLGSRLIYF VPRDNVVQHAELRRMTVLEYAPDSKQADHYRNLATKIHN
D	Org.	*Mesorhizobium amorphae*
	Seq.	MSLDYENDGALHAKLIEEVLSQYPDKAAKRRKKHLSVAKSEDEAGEEGEVLSECDVKSNIKSIP GVMTIRGCAYAGSKGVVWGPVKDMVHISHGPVGCGQYSWSQRRNYYVGTTGIDTFGTMQFTSDF QEKDIVFGGDKKLEQIIDEIEGLFPLNNGITVQSECPIGLIGDDTEAVSRNKAKEHGKTIVPVR CEGFRGVSQSLGHHIANDSIRDWVFEKKDVEFEAGPYDVNVIGDYNIGGDAWASRILLEEIGLR VVGNWSGDATLAEIERAPKAKLNLIHCYRSMNYICRHMEEKYGVAWMEYNFFGPSQIEASLRKI AKHFGPEIEEKTEKVIAKYRPLVDAVIAKYRPRLEGNTVMLYVGGLRPRHVITAYEDLGMVIVG TGYEFAHNDDYQRTGHYVKNGTLIYDDVTGYELEKFIEGVRPNLVGSGIKEKYPVQKMGIPFRQ MHSWDYSGPYHGYDGFAIFARDVDLAINNPVWGLYQAPWKTSPSRGRAMAAE

TABELA 7 – SEQUÊNCIAS SELECIONADAS PARA A CONTINUIDADE DO PROCESSO DE MINERAÇÃO DE DADOS. (**continuação**)

Alphaproteobacteria		
K	Org.	*Methylosinus trichosporium OB3b*
	Seq.	MPQNAENVLDHFNLFRQPEYIELFERKKKEFENHVADSEVERVKDWAKTKEYQDKNFAREALTV NPAKACQPLGAVFASLGFESTIPFVHGSQGCVAYYRSHFSRHFKEPTSCVSSSMTEDAAVFGGL NNMIDGLANTYNMYKPKMISVSTTCMAEVIGDDLNAFIKTSKDKGSVPMEYDVPFAHTPAFVGS HVTGYDNVMKGIVDHFWNGKAGTAPKLERVPNEKINFIGGFDGYVVGNIREVKRMFDSMGVEYT ILGDPSDVWDTPTDGEFRMYDGGTTLEDAANAIHAKATFSMQHFSTEKTLPLIAASGQDTYSFH HPIGVRGTDEFLMKVSEVTGKPISKELTKERGRLVDAIADSSAHLHGKRYAIYGDPDLCYGLSA FLLELGAEPITVLASNGGKLWEEKMNALFASSPFGKNCKCYPGKDLWHMRSLLFTEPVDFLIGN TYGKYLERDTGTPLLRIGFPIFDRHHYHRYPVWGYQGGLNVLVWLLDKVFEETDRNTIVPAKSD YSFDIIR
E	Org.	*Bradyrhizobium japonicum*
	Seq.	MSSLAATVQDIFDEPGCAKNGSKSEAERRNGCTRQLQPGSAAGGCAFDGAKVALQPFNDVAHLV HGPIACEGNSWDNRAAASSGSDLWRTAFTTDLSETDIVFGGEKRLCKAIKEIIDKCDPPAIFVY QTCIPAMIGDDINAVCKAASRRFSKPVIPINSPGFAGSKNLGNKLAGEALLDYVIGTREPDYTT PYDINLIGEYNLSGELWQVKPVLDELGVRILCCISGDGKYREVASSHRARAAMLVCSKSMINVA RKMEQRYGIPFFEGSFYGIQDSSESLRQIARLLVERGAPADLLGRTEAVIAREEARAWAAIQPY KPRLEGKRALLMTGGVKSWSVVSALQEAGLELVGTSVKKSTMEDKERIKELMGQDAHMIDDMTA REMYKMLKDAEADIMLSGGKSQFVALKAAVPWVDINQERCHAYMGYAGIVKLVEEIDNSLSSPM WEQLRRPAPWEALAKAREQMQSMAAIAGDPVLAETARRARNICVCNRVDLGTIEDAISVHGLRS VAAVREHTNAAGGCCQGRIEDMLMSEPSDMRQAAE
N	Org.	*Rhizobium etli CIAT 652*
	Seq.	MVQVFPQTKSAAVNPLKSSQPLGAALAFLGVERAVPLFHGSQGCTSLALVLLVRHFKEAIPLQT TALDAVATILGGAGNLEEAILNLKRRAKPKLIGICTTALVETRGEDFAGDLANIKRDRADELAG TEVVLANTPDFEGAIEEGWAKAVISMVERITTPGEQSRHQKKIAILPGWHLTVADIELLRETVE SFGLKPVILPDISGSLDGTVPGDRWVPTTYGGTPVDEIQELGTAAQCIAIGEHMRRPAELLRTR TGVPYALFQSLTGLKRADRFVSFLSEISGAPVPANIRRRRAQLQDALLDGHFHFGGKKIAIAVE PDQLYQFATFFTGMGAEIAAAVTTTGTSKILAEMPAESLQVGDLGDLEELAVGADLLVTHSHGR QAAERLGIPLMRVGFPIYDRLGSQHKLTSLYQGTRDLIFDVSNIFQANQRAPCPGSLDPSRKGE LPR
Betaproteobacteria		
H	Org.	*Burkholderia sp. Ch1-1*
	Seq.	MSKLRQIAFYGKGGIGKSTTSQNTLAALTELGQKILIVGCDPKADSTRLILHAKAQDTILSLAA EAGSVEDLELEDVMKIGYRDIRCVESGGPEPGVGCAGRGVITSINFLEENGAYDGVDYVSYDVL GDVVCGGFAMPIRENKAQEIYIVMSGEMMAMYAANNISKGILKYANSGGVRLGGLVCNERQTDK ELELAEALAKMLGSRLIHFVPRDNIVQHAELRRMTVIEFAPESKQAEEYRQLATKVHNNAGNGT IPTPITMDQLEDLLMEHGIMKSIDESEVGKTAAELTA
D	Org.	*Herbaspirillum seropedicae*
	Seq.	MSLTVEETTARNTELINEVLKAYPDKTAKRRAKHLTTQEEGKSDCNVKSNIKSIPGVMTIPPCA YAGSKGVVWGPIKDMIHISHGPVGCGQYSWGSRRNYYIGKTGIDSFVTMQFTSDFQEKDIVFGG DKKLEKIVDEIQELFPLNKGISVQSECPIGLIGDDIEAVSKKKSKQYEGHTIVPVRCEGFRGVS QSLGHHVANDAIKEWVLDKMDPDKNQFVATPYDVAIIGDYNIGGDAWSSRILLEEIGLRVIAQW SGDGTLAEMENTPKAKLNVLHCYRSMNYISRHMEEKFGIPWVEYNFFGPSQIEASLRQIASHFD DKIKEGAERVIAKYKALTDAVIAKYRPRLEGKTVMLFVGGLRPRHVIDAYGDLGMKVVGTGYEF GHNDDYQRTTHYVEDGTLIYDDVTSYEFEKFVEKIEPDLVGSGIKEKYVFQKMGVPFRQMHSWD YSGPYHGYDRFAIFARDMDMAINSPVWGMAKAPWKA

TABELA 7 – SEQUÊNCIAS SELECIONADAS PARA A CONTINUIDADE DO PROCESSO DE MINERAÇÃO DE DADOS. (**continuação**)

Betaproteobacteria		
K	Org.	*Leptothrix cholodnii SP-6*
	Seq.	MPQNADKVIDHEMLFREPEYQELFASKKAEFEFNHSDTKVAEIRDWTKSAEYKDKNFAREALTV NPAKACQPLGAVFVANGFHKTLSFVHGSQGCVAYYRSHFSRHFKEPTSCVSSSMTEDAAVFGGL NNMIDGLANSYNLYKPDMIAVSTTCMAEVIGDDLNAFIKTSKEKGSVPAEFDVPFAHTPAFVGS HVTGYDNALLGILQHFWDGKAGTAPKLERVPDESINFIGGFDGFVVGNMLEIKRIFDLFGADYT VLCDPSETWNTPTDGQFRMYEGGTTKAQVERALNAKATIVFQEYSCEKTIKYLKEKGQEVVVLN SPIGVAGTDAFVMALSRLTGKPVPAVLEKERGQLVDAMADSQAHLHGKRYALYGDPDMMLGLTQ FLLELGAEPAHVLATNGSNEWAAKVQALLDASPYGAGCKVYPKRDLWHMRSLLFTEPVDFLIGN TYGKYLERDTGTPLIRMVFPIFDRHHYHRYPIWGYDGALRTLIMFLDEFFETLDANTIVPGKTD YSYDIIR
E	Org.	*Cupriavidus taiwanensis*
	Seq.	MSRKARAADLFDQPNCAKNRAKSEEERKLGCSRQLSPGAAAGGCAFDGAKIALQPVADVAHLVH GPIACEGNSWDNRQAASSGPTLYRTGFTTDINEFDVIYGGERRLFRSVREIIEKYNPPAVFVYQ TCVTALIGDDIDAVCKQASVEFSKPVIPVHAPGFAGSKNLGNKLGGEALLNYVIGTREPAYTTP TDINVIGEYNLSGELWQVKPLLDELGIRLLSCITGDGRYSDIASAHRARANMVVCSKSMVHVAT KMQERYGIPYLECSFYGIGDTSNALRQIASLLVQRGASGKLLDSTERLIQAQEERAWERIATYR ARLEGKRALLITGGVKSWSVVAALQEAGLEIIGTSVKKSTKEDKERIKGILGQDALMFDNMTAR EMYKLLHDAKADIMLSGGRSQFIALKALMPWLDVNQERHHPYAGYEGIVELIREIERTIYNPVW QQVRIPAP
N	Org.	*Cupriavidus taiwanensis*
	Seq.	MAIVIQSEKACTVNPLKTSQPLGASFAAMGLEACMPVLHGSQGCTSFALVLLTRHFKEAIPLQT TAMDEISAVLGGYDNVETALLNIRKREAPRIIVVCSTGLTETNGEDLDGHLRAIRKGNPELFNT EIVYVSTPDYVGAFEDGYSKAVVEIVMELVEPLPAIYRQISVLPGSHLSPRDVEELRVIVQSFG LDPIILPDISGSLDGHFDSEWRGVTRGGTTLEQIRAAGASSFTIGIGEQTRAGARALQEICGTP FEIFERLTGLEPNDRLLRLLAQLSGRAIPEKYRRQREQLLDAMLDSHFYTGGIKVAIGADPDML LSVGSLLHELGAELSVCVSTTPSPAHALLPASEVVLGDLEDMERAAGNCDVLITHSHGRQMAAR LGKPLMRVGFPVFDRVGNAHRCQIGYRGTMDLIFEIANLMLERIQFRAPKD
Delta / Epsilonproteobacteria		
H	Org.	*Geobacter sp. M18*
	Seq.	MRQIAIYGKGGIGKSTTTQNTVAGLASIGKKVMIVGCDPKADSTRLILHAKAQSTVMDLVRELG TVEDLELEDVMKVGYGEVKCVESGGPEPGVGCAGRGVITAINFLEENGAYTPDLDFVFYDVLGD VVCGGFAMPIREGKAEEIYIVCSGEMMAMYAANNIAKGILKYASSGKVRLAGLICNARKTDKEY ELIDALAKKLGTQMIHFVPRDNQVQRAELRRMTVIEYSPDHPQANEYRTLAQKIADNKMLVVPT PLEMEELEDLLMEFGIMEAEDESIVGVAEAAAV
D	Org.	*Wolinella succinogenes*
	Seq.	MRAEELKALQKEAIEEVLAAYPEKTAKNRSKHLGVGAPDDESQKTCGGVRSNKKSAPGVMTQRG CAYAGSKGVVWGPVKDMVHISHGPIGCGQYSRGGRRNYYIGTTGVDTFVTVNFSTDFQERDIVF GGDKKLQQAIDEINDLFPLNHGITVQSECPIGLIGDDIQAAARKKSAETGKTVVAVSCEGFRGV SQSLGHHIANDTIRDIMFPLSDEVNKDFVATPYDVAIIGDYNIGGDAWSSRILLEEMGLRVIAQ WSGDSTFKEITAGPKAKLNLLHCYRSMNYVARHMEKEYGIPWMEYNFFGPSKIEASLRAIAKHF GPEIEKKAEEVIAKYKAITEAVIAKYKPRLEGKTVMLYVGGLRSRHIIGAYEDLGMEVIGLGYE FAHDDDYQRTKEEVSGSVLVYDDVNEYELEKFVDKLRPDLVASGVKEKYVFQKMGLPFRQMHSW DYSGPYHGYDGFAIFARDIEMAVNSPVWAHNTAPWD
K	Org.	*Anaeromyxobacter sp. K*
	Seq.	MANNLGLTVKPVTPTTPEEEARVAAWIDTQEYREKNFARQALVVNPVHACQPLGAELAAHGFEG TLPFVHGSQGCASYYRSTLNRHFREPAPAVSDSMTEDGAVFGGQNNLHEGLENAVALYKPRMLA IFTSCMPEVIGDDLTAFIKNARQKGALPKDLPAPYANTPSFSGTHVTGYDAMLAAILQTLTEGK RVEGRCTGRLNLLTGFDANTGNYREYKRILAAFGIPATVLADISDTFDSPNDGTYRLYPGGTPL ADAADSINGKATLTVGPYSTAKTFGWIKEAYAGEHVSLPMPMGVARTDALMLELSRLFDRPVPE SIRAERGRAVDAMTDAQQYLHGKRFAVYGDPDQLLGYVAFLLEMGATPRHVLCSRGSKKLEKEL RAVLDASPYGKDGQVWMNRDLWHLRSLLVTDPVDAVIGDTHGKFAARDARVPLFRFGFPVFDRV NKHRSPIIGYQGAINMLTELCNKFLDLRDETCEERFFEMMR

TABELA 7 – SEQUÊNCIAS SELECIONADAS PARA A CONTINUIDADE DO PROCESSO DE MINERAÇÃO DE DADOS. (**continuação**)

Delta / Epsilonproteobacteria		
E	Org.	*Geobacter uraniireducens Rf4*
	Seq.	MAKPDYYDVTECETHDAGAPKFCKKSEPGEGTERSCAYDGARVVLMPITDVIHLVHGPIACAGN SWDNRGARSSDSQLYRRGFTTEMLENDVIFGGEKKLYKAILELAERYKPKAIFVYATCVTAMTG DDVEAVCTAAQEKVAMPIIPVNTPGFIGDKNIGNRLAGEVLFKYVIGTAEPEYTTDYDINLIGE YNIAGDLWGMLPLFDKLGIRVLSCFSGDAKFEDLRYAHRAKLNVIICSKSLTNLAKKMQKTYGM PYLEESFYGMTDVAKALRDIARELDNVSGGLEKRVMQERVERLIEEEEARCRELIAPYRARLEG KRAVLFTGGVKTWSMVNALAELGVEILAAGTQNSTLEDFYRMKALMHKDARIIDDTSTAGLLSV MYEKMPDLIVAGGKTKFLALKTKTPFLDINHGRSHPYAGYDGMVTFAKQLDLTVNNPIWPVLNA KAPWEKSDAELNADVALAAGHSTAHLNEDMKESRVKVPTKNATVNPQKNSPALGATLAYLGIDQ MLGLLHGAQGCSTFIRLQLSRHFKESIALNSTSMSEDTAIFGGWENLKKGIKRVIEKFGPQVVG VMTSGLTETMGDDVRSAIVHFRQENPEFAHVPVIHASTPDYCGSMQEGYAAAVEAIVATIPEGG EKIKGQVTILPGCHLTPADVEEVAEICEAFGLTPLVIPDISNALDGHIDETVSPLSVGGVTLDK VRLAGRSEATLYLGDSLAKAAEILKENFAIPCYGFTSITGLAETDSLMETLSAIAGRPVPEKLR RWRSRLMDAMVDCHYQFGLKRISLALEADLLKTMTLFLAGMGCRIQAAISATRVRGLDRLPTDN IFVGDLEDLENSAQGSDLLVANSNGRQAAARLGGIPLLRAGLPVFDRLGAHQKMYVGYRGTMNL VFETATIFQANAKEAQKLAHN
N	Org.	*Arcobacter nitrofigilis DSM 7299*
	Seq.	MESISAKPLQLNPIKLSQPMGAMLCFLGIKNCMPLMHGAQGCASFTKVFFTRHFNDPIAVQTTA VNDITAVIDGGDYAISESIKNITKKVQPDLVGLFTTGLTETKGDDIKGACLLVQDQQKMVYVNT PDFEGSIESGFAKSIEAIIDQLVASASEVDVNKAVIIPNVNLKPIEIEKIKDTIALFGYEVLSL PDLSDSLDGHLGLKQGALSSGGITVEDIEKLGTCSLAISIGSSVKKAGDKLKAKNENMKLLHFD SLGGLEDSDKFFKALCQIKNISTPHPSIVRWRKRLQDAMLDSHFAIGSASVVLALEPDQCISVA NTIIEAGANIKAIITTHKNDLLDNIECENILIGDFEDVEKYLKDSDVLISNFHGERYTMRHKKA LMLRGFPDFEGVGNQLKNDVLYEGSTYLLFELANLINHHNQGAFHGH
Gammaproteobacteria		
H	Org.	*Azotobacter vinelandii DJ*
	Seq.	MAMRQCAIYGKGGIGKSTTTQNLVAALAEMGKKVMIVGCDPKADSTRLILHSKAQNTIMEMAAE AGTVEDLELEDVLKAGYGGVKCVESGGPEPGVGCAGRGVITAINFLEEEGAYEDDLDFVFYDVL GDVVCGGFAMPIRENKAQEIYIVCSGEMMAMYAANNISKGIVKYANSGSVRLGGLICNSRNTDR EDELIIALANKLGTQMIHFVPRDNVVQRAEIRRMTVIEYDPKAKQADEYRALARKVVDNKLLVI PNPITMDELEELLMEFGIMEVEDESIVGKTAEEV
D	Org.	*Pantoea sp. At-9b*
	Seq.	MSNATADRNLEIIQEVLEIYPEKTRKERRKHMMVTDPEMESVGKCIISNRKSQPGVMTVRGCAY AGSKGVVFGPIKDMAHISHGPIGCGQYSRAGRRNYFTGISGVDSFVTLNFTSDFQERDIVFGGD KKLTKLIEEMEELFPLTKGISIQSECPVGLIGDDISAVAKASSTAINKPVVPVRCEGFRGVSQS LGHHIANDVIRDWVLDNREGQPFTTTPYDVAIIGDYNIGGDAWASRILLEEMGLRVVAQWSGDG TLVEMENTPFVKLNLVHCYRSMNYISRHMEEKHGIPWMEYNFFGPTKVAESLRKIADQFDDQIR ANAEAVIARYQAQNDAIIAKYRPRLEGRKVLLYMGGLRPRHLIGAYEDLGMEIIAAGYEFGHND DYDRTLPDLKEGTLLFDDASSYELEAFVKALKPDLIGSGIKEKYIFQKMGVPFRQMHSWDYSGP YHGYDGFAIFARDMDMTLNNPAWGQLTAPWLKSA
K	Org.	*Methylococcus capsulatus str. Bath*
	Seq.	MSQNAEKVLDHFNLFREPEYINLFESKRKEFEFMPPDQQVEEVREWAKTEEYKEKNFAREALVV NPAKACQPLGAVFAAVGFEGTIPFVHGSQGCVAYYRSHFSRHFKEPTSCVSSSMTEDAAVFGGL NNMIDGLANTYNMYKPKMIAVSTTCMAEVIGDDLNAFIKTSKEKGSIPADFDVPFAHTPAFVGS HITGYDNVMKGILQHFWDGKSGTVEPLVRQPNESINFLGGFDGYTVGNLREIKRIFNLFGIDYT IIGDNSDVWDTPTDGEFRMYDGGTTLEQAANALHAKATISMQEFCTEKTLPFIAEHGQEVVALN HPIGVKGTDRFLMEISRLTGKPIPVELEKERGRLVDAIADSTAHIHGQKFAIYGDPDLCYGLAE FLLELGAEPTHVLATNGGKAWETKMQALFDSSPFGKNCKVYPGRDLWHMRSLLFTEPVDFLIGN TYGKYLERDTGTPLIRIGFPIFDRHHKHRYPVWGYQGGLNVLVWILDRMFEAIDANTNIPAKTD YSFDIIR

TABELA 7 – SEQUÊNCIAS SELECIONADAS PARA A CONTINUIDADE DO PROCESSO DE MINERAÇÃO DE DADOS. (**continuação**)

Gammaproteobacteria		
E	Org.	*Acidithiobacillus ferrooxidans ATCC 53993*
	Seq.	MLQNKIQDVFNEPGCSKNQSKSDKERKKGCTKALQPGGAAGGCAFDGAKIALQPITDVAHLVHG PIACEGNSWDNRGSKSSGSQLYRTGFTTDINELDVVYGGEKHLFKSIKEVLDKYDPSAVFVYQT CVTAMIGDDIESVCKAASQKFAKPIIPVNAPGFVGAKNLGNKLAGEALLDYVIGTEEPEYSTPY DINIIGEYNLSGELWQVKPLLDHLGIRVTCCISGDAKYHDVAQSHRARANMMVCSKSMINIARK MEERYQIPFFEGSFYGISDTTESLREITRLLIQQGAPAELHDRTEALIAREEARAWQRIAEYTH RLRGKRVLLFTGGVKSWSVVSALQEGGMEVVGTSVKKSTREDKERIKEIMGQDAHMLDDLTPRE MYKMFQEARADVLLSGGRSQFAALKNKMPWVDINQERHQAYNGYEGMVNLVKQIDLALYNPMWA LLRKPAPWDMGEART
N	Org.	*Klebsiella variicola At-22*
	Seq.	MADIIRSEKPLAVSPIKTGQPLGAILASLGLAQAIPLVHGAQGCSAFAKVFFIQHFHDPVPLQS TAMDPTATIMGADGNIFTALDTLCQRHSPQAIVLLSTGLTEAQGSDIARVVRQFREAHPRHNGV AILTVNTPDFFGSMENGYSAVIESVIEQWVAPTPRPGQRPRRVNLLVSHLCSPGDIEWLGRCVE AFGLQPVILPDLSQSMDGHLGEGDFTPLTQGGASLRQIAQMGQSLGSFAIGVSLQRAASLLTQR SRGDVIALPHLMTLDHCDTFIHQLAKMSGRRVPAWIERQRGQLQDAMIDCHMWLQGQRMAMAAE GDLLAAWCDFARSQGMQPGPLVAPTSHPSLRQLPVEQVVPGDLEDLQQLLSHQPADLLVANSHA RDLAEQFALPLIRVGFPLFDRLGEFRRVRQGYAGMRDTLFELANLLRDRHHHTALYHSPLRQGA APQSASGDAYAAH
Random		
H	Org.	*Bradyrhizobium yuanmingense*
	Seq.	GIGKSTTSQNTLAALAEMGQKILIVGCDPKADSTRLILHAKAQDTILSLAASAGSVEDLELEDV MKVGYREIRCVESGGPEPGVGCAGRGVITSINFLEENGAYENIDYVSYDVLGDVVCGGFAMPIR ENKAQEIYIVMSGEMMAMYAANNISKGILKYANSGGVRLGGLICNERQTDKELELAEALAKKLG TQLIYFVPRDNVVQHAELRRMTVLEYAPDSKQADHYRNLATKVHNNGGKG
D	Org.	*Sinorhizobium medicae*
	Seq.	MSLDYENDDALHEKLIEEVLSHYPDKAAKRRKKHLSVAKNKQETAEEGQVVSECDVKSNIKSIP GVMTIRGCAYAGSKGVVWGPIKDMVHISHGPVGCGQYSWSQRRNYYVGTTGIDAFVTMQFTSDF QEKDIVFGGDKKLEKIIDEIEELFPLNNGVTVQSECPIGLIGDDIEAVSRKKAEEYNTTIVPVR CEGFRGVSQSLGHHIANDAIRDWVFDTTEVAYEAGRYDVNVIGDYNIGGDAWASRILLEEIGLH VVGNWSGDATLAEIERAPTAKLNLIHCYRSMNYICRHMEEKYGVPWMEYNFFGPSQIEASLRQI AKHFGPEIEERAERVIARYSALTYAVIDKYWPRLRGKRVMLYVGGLRPRHVITAYEDLGMEIVG TGYEFAHNDDYQRTGHYVKEGTLIYDDVTGYELEKFIERIRPDLVGSGIKEKYSVQKMGIPFRQ MHSWDYSGPYHGYDGFAIFARDMDLAVNNPVWELYDAPWQKAAMLAASGAAE
K	Org.	*Azospirillum brasilense*
	Seq.	MSMSHPVSQSADKVIDHFTLFRQPEYKELFERKKTEFEYGPTPTKEVARVSAWTKTEEYKEKNL PVEAVVINPTKACQPIGAMLAAQGFEGTLPFVHGSQGCVSYYRTHLTRHFKEPNSAVSSSMTED AAVFGGLNNMIDGLQRYALYKPKMIAVLTTCMAEVIGDDLSGFINNAKNKESVPADFPVPFAHT PAFVGSHIVGYDNMIKGVLTHFWGTSENFDTPKNETINLIPGFDGFAVGNNRELKRIAGLFGIQ MTILSDVSDNFDTPADGEYRMYDGGTPLEATKEAVHAKATISMQEYCTPQSLQFIKREGPAGRQ AYNYPMGVTGTDELLMKLAELSGKPSRGVKLERGRLVDAIADSHTHLHGKRFAVYGDPDFCLGM SKFLMELGAEPVHILSTSGSKKWEKQVQKVLDACRSASSGKAYGAKDLWHLRSLVFTDKVDYII GNSYGKYLERDTKIPLIRLTYPIFDRHHHHRYPTWGYQGALNVLVRILDRIFEDMDANTNIVGE TDYSFDLVR
E	Org.	*Bradyrhizobium japonicum*
	Seq.	MSSLAATVQDIFDEPGCAKNGSKSEAERRNGCTRQLQPGSAAGGCAFDGAKVALQPFTDVAHLV HGPIACEGNSWDNRGAASSGSDLWRTAFTTDLSETDIVFGGEKRLCKAIKEIIDKCDPPAIFVY QTCIPAMIGDDINAVCKAASRRFSKPVIPINSPGFAGSKNLGNKLAGEALLDYVIGTREPDYTT PYDINLIGEYNLSGELWQVKPVLDELGVRILCCISGDGKYREVASSHRARAAMLVCSKSMINVA RKMEQRYGIPFFEGSFYGIQDSSESLRQIARLLVERGAPADLLGRTEAVIAREEARAWAAIQPY KPRLEGKRALLMTGGVKSWSVVSALQEAGLELVGTSVKKSTMEDKERIKELMGQDAHMIDDMTA REMYKMLKDAEADIMLSGGKSQFVALKAAVPWVDINQERCHAYMGYAGIVKLVEEIDNSLSSPM WEQLRRPAPWEALAKAREQMQSMAAIAGDPVLAETARRARNICVCNRVDLGTIEDAISVHGLRS VAAVREHTNAAGGCCQGRIEDMLMSEPSDMRQAAE

TABELA 7 – SEQUÊNCIAS SELECIONADAS PARA A CONTINUIDADE DO PROCESSO DE MINERAÇÃO DE DADOS. (**conclusão**)

Random		
N	Org.	*Gluconacetobacter diazotrophicus PAl 5*
	Seq.	MATIVKPRKAASVNPAEIFDAAGRGAGLSGYRRRGAAVPWLAGLQSFALVLTVRHYKEAIPLQD HGDGRGRDILGAAGNLEEALLNLQRRMKPRFIGIASTALVETRGEDYAGDLKLILQRQPELADT RIVFASTPDYAGALEDGWAAAVSAIIESVVAPWSPTVTSFQQVNVLPGVHQTPADIEALRDLIE SFGLYPVILPDLSGSLDGHVAENWCPTTQGGARMEEVAQMARAVHTIAIGEHMRAPADLLGSVT GVPVTLFPTLTGLAANDRLMALLSRLSGRAVPGRYRRQRSQLLDAMLDGHFHFGGKRIAIAADP DLLYGLSAFFAGMGARIVAAVASTSNAPNLDSIPADSVIVGDLTDLEDAVHAAGGADLLVTHSH GRQSADRLGIPLMRVGFPIFDRLGTAHAQTIGYRGTRDLIFRVANLFLGQMHEHTPDDFGHVPS AHTIEEIVHDSASLAAH

FONTE: A autora (2010).

A continuidade do processo de mineração de dados foi efetuada mediante a execução automática da ferramenta BLASTP, explicada com detalhes na próxima seção.

4.3. EXECUÇÃO AUTOMÁTICA DA FERRAMENTA BLASTP

Com base na divisão taxonômica realizada sobre as sequências encontradas, foi criada uma estrutura física de diretórios que comportasse o armazenamento das informações relacionadas a cada um dos grupos estudados. Dessa forma, foram criados os seguintes diretórios:

- ..\BDActinobacteria
- ..\BDArchaea
- ..\BDChlorobi
- ..\BDCyanobacteria
- ..\BDFirmicutes
- ..\BDPAlpha
- ..\BDPBeta
- ..\BDPDeltaEpsilon
- ..\BDPGama
- ..\BDRandom

A execução da ferramenta BLASTP foi realizada de maneira automática, através de um script desenvolvido. O Apêndice 1 apresenta o código fonte do referido script.

O script foi replicado em cada um dos diretórios criados, sendo o trecho de código relacionado às sequências *querys* de busca alterado conforme a Tabela 7.

Como resultados da execução do script foram gerados cinco arquivos texto para cada diretório "taxonômico", relacionados a cada uma das sequências de genes *nif* procuradas (nifH.txt, nifD.txt, nifK.txt, nifE.txt e nifN.txt), contendo o retorno padrão de texto simples fornecido pela função NCBIWWW.qblast() disponibilizada pela biblioteca BioPython. No total foram gerados 50 arquivos de retorno.

O Anexo 2 apresenta um exemplo de arquivo texto simples, retornado pela função NCBIWWW.qblast(). O trecho do arquivo em questão apresenta alguns resultados da busca pela sequência relacionada ao gene *nifN* agregada ao grupo taxonômico Actinobacteria.

O processo de interpretação e tabulação automática dos dados resultantes da execução da ferramenta BLAST se encontra descrito na seção a seguir.

4.4. INTERPRETAÇÃO E TABULAÇÃO AUTOMÁTICA DE DADOS

Dando continuidade ao processo de mineração de dados a etapa de interpretação e tabulação automática de dados foi dividida em quatro procedimentos distintos:

- Análise e separação de informações (*parsing*);
- Geração de Tabelas Mestre de organismos encontrados;
- Construção de Tabela Índice, baseada no campo GenBank ID.
- Busca nos dados originais (através da Tabela Índice) e incorporação de informações relevantes à Tabela Mestre.

Cada procedimento foi executado de maneira automática por um script específico. As próximas subseções descrevem em detalhes, cada um destes processos.

Como resultado final após a execução das quatro etapas descritas, foram gerados dez arquivos texto (de nome IDsValuesBLAST.txt) um para cada grupo taxonômico estudado, contendo todas as informações contidas nos arquivos texto simples retornados pela execução da ferramenta BLASTP num formato tabulado, ou seja, com colunas correspondendo à campos e linhas correspondendo a registros.

A Tabela 8 apresenta uma descrição dos campos constantes no arquivo IDsValuesBLAST.txt:

TABELA 8 – DESCRIÇÃO DOS CAMPOS CONSTANTES NOS ARQUIVOS IDsValuesBLAST.txt

ID	Nome do Campo	Descrição
01	Organismo	Descrição do organismo.
02	*nif*	Gene *nif* ao qual o registro corresponde.
03	ID	ID Genbank, correspondende ao campo GI retornado pelo BLASTP.
04	Query Length	Tamanho da sequência Query enviada ao BLASTP.
05	Subject Length	Tamanho da sequência Subject retornada pelo BLASTP.
06	Identities	Valor do campo Identities.
07	% Identities	Valor percentual correspondente ao campo Identities.
08	Positives	Valor do campo Positives.
09	% Positives	Valor percentual correspondente ao campo Positives.
10	Gaps	Valor do campo Gaps.
11	% Gaps	Valor percentual correspondente ao campo Gaps.
12	Score	Valor do bit score atribuído ao alinhamento.
13	Expect	e-Value atribuído ao alinhamento.
14	Annotation	Anotação original do gene registrada no GenBank.
15	Query	Sequencia Query enviada.
16	IniQuery	Posição de início do alinhamento, correspondente a Query.
17	EndQuery	Posição final do alinhamento, correspondente a Query.
18	Sbjct	Sequência Subject encontrada.
19	IniSbjct	Posição de início do alinhamento, correspondente ao Subject.
20	EndSbjct	Posição final do alinhamento, correspondente ao Subject.

Fonte: A autora, (2010).

Estes arquivos correspondem a Tabelas Detalhe, ligadas a uma Tabela Mestre (descrita adiante) através de uma pseudo chave-primária composta dos campos Organismo e *nif.*

4.4.1. *PARSING* DE DADOS

O processo de interpretação e separação de dados a partir dos arquivos texto simples retornados pela execução da ferramenta BLASTP (nifH.txt, nifD.txt, nifK.txt, nifE.txt e nifN.txt) foi realizado utilizando-se o script **step02ParsingData.py**, apresentado no Apêndice 2.

O script foi replicado e executado em cada um dos diretórios referentes aos grupos taxonômicos em estudo sendo gerados, para cada diretório, cinco novos arquivos auxiliares, (ListnifH.txt, ListnifD.txt, ListnifK.txt, ListnifE.txt e ListnifN.txt), um para cada gene *nif* estudado, contendo a descrição dos organismos encontrados nos arquivos de retorno em texto simples do BLASTP. No total foram gerados mais 50 arquivos de retorno.

Com base nos resultados obtidos com a execução do script em questão, foi gerada uma Tabela Mestre de dados, para cada grupo taxonômico em estudo. Os detalhes deste procedimento são detalhados na subseção a seguir.

4.4.2. GERAÇÃO DE TABELAS MESTRE DE DADOS

O processo para a geração de uma Tabela Mestre (maybefix.txt), para cada grupo taxonômico estudado, contendo as ocorrências únicas dos organismos identificados e listados nos arquivos de retorno da função de *parsing* de dados (ListnifH.txt, ListnifD.txt, ListnifK.txt, ListnifE.txt e ListnifN.txt), foi realizado utilizando-se o script **step03MergeFiles.py**, apresentado no Apêndice 3.

Dentro dos processos executados pelo script, um passo importante é a busca pela taxonomia registrada no GenBank para cada um dos organismos constantes na Tabela Mestre. Este processo é realizado pela execução das funções **Entrez.esearch()** e **Entrez.efetch()**, disponibilizada pela biblioteca BioPython, passando como base de busca, o banco de dados “Taxonomy” e solicitando retorno em formato “XML”. De forma a agilizar os processos realizados dentro da estratégia de mineração de dados, considerando o acesso constante via web e suas limitações, um arquivo de log contendo o nome do organismo e sua respectiva “Taxonomia GenBank” é gravado localmente. A primeira consulta à taxonomia é realizada neste arquivo e, somente no caso da mesma não ser encontrada, a busca web é ativada.

Outro processo importante realizado pelo script **step03MergeFiles.py** é a verificação dos organismos que possuem ou não seu genoma completamente seqüenciado. Esta verificação é possível graças ao acesso local e busca através de instruções SQL (Linguagem de Consulta Estruturada, do inglês Structured Query Language)[4] ao banco de dados relacional de genomas completos, disponibilizado por Guizelini (GUIZELINI, 2010).

---

[4] Linguagem de pesquisa declarativa para bases de dados relacionais, desenvolvido originalmente no início dos anos 70 nos laboratórios da IBM dentro do projeto System R, com o objetivo de demonstrar a viabilidade da implementação do modelo relacional proposto por Codd (CODD et al., 1981). O nome original da linguagem era **SEQUEL**, acrônimo para **"Structured English Query Language"** (Linguagem de Consulta Estruturada em Inglês) (CHAMBERLIN, 1976).

Os organismos que possuem referência encontrada no banco de dados são marcacados com um atributo específico, dentro do arquivo de Tabela Mestre.

O script foi replicado e executado em cada um dos diretórios referentes aos grupos taxonômicos, sendo geradas 10 Tabelas Mestre.

A partir das Tabelas Mestre, foram gerados arquivos de índice de forma a melhorar o processo de busca por informações relevantes. A subseção a seguir explica detalhadamente como esta geração ocorreu.

4.4.3. CONSTRUÇÃO DE TABELAS ÍNDICE

A construção das tabelas de índice para os dados existentes nas Tabelas Mestre se deu pela execução do script, **step04FindIDs.py**, apresentado no Apêndice 4.

O script foi replicado e executado em cada um dos diretórios referentes aos grupos taxonômicos estudados, sendo geradas 10 tabelas de índices, relacionadas respectivamente às suas Tabelas Mestre originais.

Com base nas tabelas de índice geradas foi possível o resgate de informações relevantes, a partir dos arquivos texto simples originais, retornados pela execução da ferramenta BLASTP.

4.4.4. INCORPORAÇÃO DE DADOS

A partir da geração das tabelas de índices, novos dados puderam ser incorporados à Tabela Mestre de forma a serem utilizados posteriormente na pesquisa.

O script **step05AddValuesIDs.py**, apresentado no Apêndice 5, possibilitou a criação do arquivo IDsValuesBLAST.txt que continha, em formato tabulado todas as informações referentes aos arquivos texto simples de retorno do BLASTP, viabilizando sua manipulação computacional pelas demais ferramentas utilizadas na pesquisa, como planilhas eletrônicas e redes neurais artificiais.

Da mesma forma ocorrida nas etapas anteriores, o script foi replicado e executado em cada um dos diretórios referentes aos grupos taxonômicos estudados, sendo gerados no total 10 arquivos tabulados.

4.5. PROCESSO DE UNIÃO E LIMPEZA DE DADOS

Com os dados tabulados, armazenados em cada um dos diretórios correspondentes aos grupos taxonômicos estudados, foi possível o desenvolvimento de scripts que contemplassem as etapas de união e limpeza de dados, descritas com detalhes nas próximas sub-seções.

4.5.1. UNIÃO DE DADOS

O processo de união automática dos dados obtidos com a execução do BLASTP, armazenados de forma dispersa nos 10 diretórios correspondentes aos grupos taxonômicos estudados (arquivos IDsValuesBLAST.txt), foi possível graças ao desenvolvimento dos scripts **step06RescueData.py** e **step07NewDataUnion.py**, respectivamente apresentados nos Apêndices 6 e 7.

Um pré-processamento de limpeza é realizado pelo script **step06RescueData.py** de forma a eliminar as redundâncias de ID GenBank encontradas em cada um dos arquivos IDsValuesBLAST.txt. Estas redundâncias ocorrem devido à alta homologia encontrada entre os genes *nif*. Esta homologia é refletida nos resultados apresentados pela ferramenta BLASTP, pois, por exemplo, ao se realizar uma busca pelo gene *nifN*, vários resultados referentes ao gene *nifE* são apresentados, e vice-versa, resultando nas redundâncias encontradas nos arquivos IDsValuesBLAST.txt.

Um outro ponto interessante relacionado ao processamento realizado pelo script **step06RescueData.py** é o fato da eliminação dos registros relacionados a organismos não cultivados ou não identificados (portadores das *tags* 'uncultured' ou 'unidentified' em sua descrição, respectivamente), que estivessem marcados na Tabela Mestre como tendo seu genoma ainda incompleto. Tal cuidado é tomado devido ao enorme fluxo de dados resultante das pesquisas de metagenômica, depositados no NCBI GenBank.

Ao final do processo realizado pelo script **step06RescueData.py** foram criados um arquivo para cada diretório correspondente aos grupos taxonômicos estudados (BDActino.txt, BDArchaea.txt, BDChlorobi.txt, BDCyano.txt, BDFirmicutes.txt, BDPAlpha.txt, BDPBeta.txt, BDPDelta.txt, BDPGama.txt e BDRandom.txt), contendo as informações agora sem redundâncias. Estes arquivos

criados são utilizados como entrada para o script **step07NewDataUnion.py**, que irá realizar a união dos dados disponíveis.

Um novo diretório nomeado ..\BDUnion foi criado para armazenar os dados resultantes do proceso de união. Os scripts **step06RescueData.py** e **step07NewDataUnion.py** foram executados a partir deste diretório, resultando ao final dos processos, na criação do arquivo NewData.txt contendo os registros com ID GenBank único referentes a todos os grupos taxonômicos estudados.

4.5.2. LIMPEZA DOS DADOS

A conclusão do processo de limpeza dos dados ocorreu mediante a execução do script **step08CuringData.py**, apresentado no Apêndice 8, e consistiu da eliminação das redundâncias encontradas no arquivo NewData.txt, originadas no processo de união dos arquivos BDActino.txt, BDArchaea.txt, BDChlorobi.txt, BDCyano.txt, BDFirmicutes.txt, BDPAlpha.txt, BDPBeta.txt, BDPDelta.txt, BDPGama.txt e BDRandom.txt.

As redundâncias ocorreram devido ao fato de um mesmo ID GenBank ser trazido como resultado da execução da ferramenta BLASTP para mais de um grupo taxonômico diferente.

Ao final do processo uma nova tabela unificada, contendo todas as ocorrências únicas de organismos encontrados pela ferramenta BLAST em todos os grupos taxonômicos estudados é gerada com o nome de DataOk.txt. Esta tabela serviu de base para o processo de seleção de características e reconhecimento de padrões descrito na próxima seção.

4.6. EXTRAÇÃO DE CARACTERÍSTICAS

As características descritas na seção de Metodologia puderam ser extraídas a partir dos dados existentes na tabela DataOk.txt através de scripts desenvolvidos, detalhados nas subseções descritas a seguir.

A cada ciclo do processo de extração de características, os novos dados resultantes eram devidamente anexados ao seu registro de origem, sendo, porém gravados em uma nova tabela, chamada LastData.txt, de forma a manter a originalidade e integridade dos dados encontrados anteriormente.

Ao final do processo de extração de características, a tabela LastData.txt, contendo 14988 registros (com base em dados obtidos em pesquisa realizada no dia 25/11/2010), serviu de base para que a classificação prévia de dados (necessária para o processo de aprendizado da rede neural artificial) pudesse ser realizada.

4.6.1. GRUPO I - FISICO-QUÍMICAS

As cinco características pertencentes ao grupo I – Físico-Químicas foram extraídas com base na passagem da sequência protéica, de cada um dos registros arrolados na tabela DataOk.txt, como parâmetro para a função: **ProtParam.ProteinAnalysis()**, disponibilizada pela biblioteca de BioPython.

O trecho de código correspondente a esse processo pode ser visualizado na Figura 13.

```
:
:
## Chamada da função ProtParam.ProteinAnalysis
nif  = ProtParam.ProteinAnalysis(seq)       ## Análise da proteína

## Gravação da lista base para a posterior geração do arquivo LastData.txt
newList.append([…,                           \## Gravação dos demais dados
               str(molecular_weight(Seq)),   \## Campo Peso Molecular
               str(nif.isoelectric_point()),\## Campo Ponto Isoelétrico
               str(nif.aromaticity()),       \## Campo Percentual de Aromaticidade
               str(nif.instability_index()),\## Campo Índice de Instabilidade
               str(nif.gravy()),             \## Campo Índice GRAVY.
                …])
:
:
```

FIGURA 13 – TRECHO DE CÓDIGO CORRESPONDENTE À GERAÇÃO DAS CARACTERÍSTICAS DESCRITAS NO GRUPO I – FÍSICO QUÍMICAS.
FONTE: A Autora, (2010).

4.6.1.1. Assinaturas de Domínios Conservados

Com base nas assinaturas referentes aos domínios conservados de proteínas relacionadas à nitrogenase, catalogados e disponíveis no banco de dados ExPASy Prosite, obtidas conforme descrito em Metodologia, foi desenvolvido um procedimento automatizado para busca e verificação da existência das mesmas nas sequências de estudo.

O procedimento de busca utiliza-se de Expressões Regulares[5] para a realização da busca e determinação da existência ou não de domínios conservados relacionados à nitrogenase em cada uma das sequências analisadas.

A linguagem de programação Python disponibiliza o módulo **re** para processamento e interpretação de expressões regulares e a função **re.search()** para a realização de busca por determinadas expressões em uma sequência de caracteres qualquer. A Figura 14 apresenta o trecho de código que, ao ser executado, realizou este processo de busca.

```
## Lista das expressões regulares a serem verificadas em cada uma das
sequências de estudo.
expr = ['(E.GGP..[GA].GC[AG]G)',\
        '(D.LGDVVCGGF[AGSP].P)',\
        '([LIVMFW]..H.H[DN]D.G.[GAS].[GASLI])',\
        '([LIVMFYH][LIVMFST]H[AG][AGSP][LIVMNQA][AG]C)',\
        '([STANQ][ET]C.....GD[DN][LIVMT].[STAGR][LIVMFYST])',\
        '([LIV]...C[NDP][LIVMF][DNQRS]C.[FYM]C)']

## Laço para verificação e busca dos domínios conservados, nas sequências
de estudo.
for line in lines:
    ConsDom = 0
    for item in expr:
        result = re.search(item,line[24])
        if result <> None:
           ConsDom = 1
           break
    line.append(str(ConsDom))
```

FIGURA 14 – TRECHO DE CÓDIGO CORRESPONDENTE À GERAÇÃO DA CARACTERÍSTICA RELACIONADA À EXISTÊNCIA DE DOMÍNIOS CONSERVADOS DA NITROGENASE.
FONTE: A Autora, (2010).

4.6.2. GRUPO II - INFERIDAS

Os percentuais referentes aos aminoácidos Fenilalanina (F), Leucina (L), Alanina (A), Arginina (R) e Glicina (G), pertencentes ao grupo de características

[5] Expressão regular é um método formal de se especificar um padrão de texto. É uma composição de símbolos, caracteres com funções especiais, que, agrupados entre si e com caracteres literais, formam uma seqüência, uma expressão. Essa expressão é interpretada como uma regra, que indicará sucesso se uma entrada de dados qualquer “casar” com essa regra, ou seja, obedecer exatamente a todas as suas condições (JARGAS, 2001).

inferidas, conforme estudo apresentado na seção Metodologia, foram obtidos de forma automatizada.

A Figura 15 apresenta o trecho de código correspondente ao processo que totaliza a partir da sequência analisada, os aminioácidos de interesse e divide os mesmos pelo tamanho total da sequência, de modo a obter um determinado valor percentual.

```
:
:
:
print "Processing data..."
for line in lines:
    F = 0
    L = 0
    A = 0
    R = 0
    G = 0
    for letter in line[17]:
        if letter == "F":
           F = F + 1
        elif letter == "L":
           L = L + 1
        elif letter == "A":
           A = A + 1
        elif letter == "R":
           R = R + 1
        elif letter == "G":
           G = G + 1
    line.append(str(((float(F)/float(len(line[17])))*100.0)))
    line.append(str(((float(L)/float(len(line[17])))*100.0)))
    line.append(str(((float(A)/float(len(line[17])))*100.0)))
    line.append(str(((float(R)/float(len(line[17])))*100.0)))
    line.append(str(((float(G)/float(len(line[17])))*100.0)))
:
:
:
```

FIGURA 15 – TRECHO DE CÓDIGO CORRESPONDENTE À GERAÇÃO DAS CARACTERÍSTICAS DESCRITAS NO GRUPO II – INFERIDAS.
FONTE: A Autora, (2010).

4.6.3. GRUPO III - BLAST

As duas características relacionadas ao Grupo III – BLAST, (1) Tamanho total da sequência e (2) Média entre valores de Identidade X Similaridade foram obtidas automaticamente, via execução de script desenvolvido.

A Figura 16 apresenta o trecho do código relacionado a este processo:

```
:
:
for it in tSorted:
    ## Tamanho da Sequência Query
    q  = it[4]

    ## Valor Identidade trazido pelo BLAST
    idt= it[6].split('/')

    ## Valor Similaridade trazido pelo BLAST
    sim= it[8].split('/')

    ## Calculo da Média Simples
    med= ((((float(idt[0])/float(q))*100)+\
           ((float(sim[0])/float(q))*100))/2)

    ## Tamanho da Sequência Retorno
    s  = len(it[9])
:
:
```

FIGURA 16 – TRECHO DE CÓDIGO CORRESPONDENTE À GERAÇÃO DAS CARACTERÍSTICAS DESCRITAS NO GRUPO III – BLAST.
FONTE: A Autora, (2010).

4.6.4. GRUPO IV – MATRIZES DE CO-OCORRÊNCIA

Foi desenvolvida e implementada em Python uma técnica para a extração das medidas de Energia e Entropia baseadas no modelo de matrizes de co-ocorrência, utilizadas na análise de texturas em imagens, descrito por Haralick (HARALICK, 1979).

A Figura 17 apresenta o código fonte da função de geração das características de Energia e Entropia.

```
def features(seq,bases):
    dicl    = {}
    dicc    = {}
    dicx    = {}
    seqfind = ''
    atu     = seq
    for letter in seq:
        seqfind = seqfind+letter
        tam = len(atu)
        ini = 0
        while tam >= 0:
            trinca = atu[ini:ini+bases]
            ini = ini + bases
            tam = tam - bases
            if len(trinca) == bases :
               dicl[trinca] = 0
               dicc[trinca] = 0
        atu = seq.replace(seqfind,"")
    for l in dicl:
        for c in dicc:
            x       = l+c
            dicx[x] = 0
    for x in dicx:
        lAux = seq
        ini  = 0
        while ini >= 0:
              atu = lAux.find(x, ini, len(lAux))
              if atu >= 0:
                 dicx[x] = dicx[x]+1
                 ini = atu + 1
              else:
                 break

    sumall = 0
    for x in dicx:
        if dicx[x] > 0:
           sumall = sumall + dicx[x]

    entropy  = 0
    energy   = 0
    for x in dicx:
        if dicx[x] > 0:
           d = float(dicx[x])/float(sumall)
           entropy  = entropy + (d*log(d,2))
           energy   = energy + (d**2)
    entropy = entropy * -1
    return [entropy, energy]
```

FIGURA 17 – CÓDIGO FONTE CORRESPONDENTE À FUNÇÃO DE GERAÇÃO DAS CARACTERÍSTICAS DE ENERGIA E ENTROPIA, BASEADAS NOS DESCRITORES DE HARALICK (HARALICK, 1979).
FONTE: A Autora, (2010).

Ao final do processo, todas as co-ocorrências de 3X3 aminoácidos estavam devidamente totalizadas e armazenadas em uma estrutura de fácil acesso para que as fórmulas de Energia e Entropia pudessem ser aplicadas, fornecendo como resultados, as características desejadas para cada uma das sequências analisadas.

4.7. APRENDIZADO DA REDE NEURAL ARTIFICIAL

Com base nos resultados obtidos com o processo de co-aprendizado da rede neural, e, atendendo às especificações de software da EasyFan, com o auxílio do software Excel, os 14988 registros foram separados em três arquivos: Treinamento.txt, Teste.txt e Validação.txt. O Quadro 1 apresenta um melhor detalhamento sobre a divisão das informações, nos três arquivos de trabalho.

Conjuntos de Dados			
Total Geral de Registros	14988		
Total de Registros Previamente Classificados	14988 (100%)	**Registros por Classe**	
		Não *nifs*	***nifs***
		10385	4603
Total de Registros do Arquivo de Treinamento	4996 (1/3)	**Registros por Classe**	
		Não *nifs*	***nifs***
		3490	1506
Total de Registros do Arquivo de Teste	4996 (1/3)	**Registros por Classe**	
		Não *nifs*	***nifs***
		3450	1546
Total de Registros do Arquivo de Validação	4996 (1/3)	**Registros por Classe**	
		Não *nifs*	***nifs***
		3445	1551

QUADRO 1 – DIVISÃO DOS CONJUNTOS DE DADOS PARA UTILIZAÇÃO NO APRENDIZADO DA REDE FAN.
FONTE: A Autora, (2010).

Ainda com o auxílio do software Excel, os dados constantes em cada um dos arquivos foram “embaralhados” de modo a não apresentarem nenhum vício ou padrão que pudesse prejudicar o aprendizado da rede.

Submetidos ao processamento da EasyFan, o tempo total dispendido para o aprendizado dentro dos níveis desejados foi de 12 horas, com estabilidade nos parâmetros de médias encontrada em torno de 5 horas de treinamento.

Os resultados obtidos durante o processo de aprendizagem, considerando dados referentes à melhor média aritmética, se encontram apresentados no Quadro 2. Os valores se encontram separados, conforme cada uma das etapas de aprendizado da rede: Treinamento, Teste e Validação.

Percentual de Acertos da Rede FAN				
Medidas	**Treinamento**	**Teste**	**Validação**	**Média**
Acertividade	98,83	99,01	98,64	98,83

QUADRO 2 – MEDIDAS DE ACERTIVIDADE DA REDE FAN ENTRE OS TRÊS CONJUNTOS DE DADOS.
FONTE: A Autora, (2010).

Com a repercussão do percentual de acertos, superior a 98.5% em todos os conjuntos de dados, a rede gerada foi considerada satisfatória para a utilização dentro da pesquisa, como forma de classificação das informações obtidas no processo de mineração de dados. O arquivo **Rede.enn**, contendo os dados referentes ao aprendizado da rede foi gerado e gravado para futuras utilizações.

O Quadro 3 apresenta os resultados para a classificação realizada pela rede neural FAN, sobre os dados resultantes do processo de mineração de dados.

Classificação Final dos Dados		
Classe	**Quantidade**	**Percentual**
0 ou "Não *nif*"	10463	69,81%
1 ou "*nif*"	4525	30,19%
Total	**14988**	**100,00%**

QUADRO 3 – RESULTADO QUANTITATIVO, REFERENTE Á CLASSIFICAÇÃO FINAL DOS DADOS, REALIZADA PELA REDE FAN.
FONTE: A Autora, (2010).

Os 4525 registros, assinalados pela rede como genes *nif*, foram submetidos a uma análise manual realizada com base no retorno oferecido pela ferramenta BLASTP e o agrupamento dos mesmos entre os genes estudados (*nifHDKEN*) pode ser verificado no Quadro 4.

Registros Classificados como Genes *nif*	
Gene *nif*	**Quantidade**
D	749
E	313
H	2778
K	452
N	233
Total	**4525**

QUADRO 4 – DISPERSÃO DOS REGISTROS CLASSIFICADOS COMO GENES *nif* ENTRE O UNIVERSO DE GENES ESTUDADOS.
FONTE: A Autora, (2010).

A Figura 18 apresenta graficamente os dados arrolados no Quadro 4.

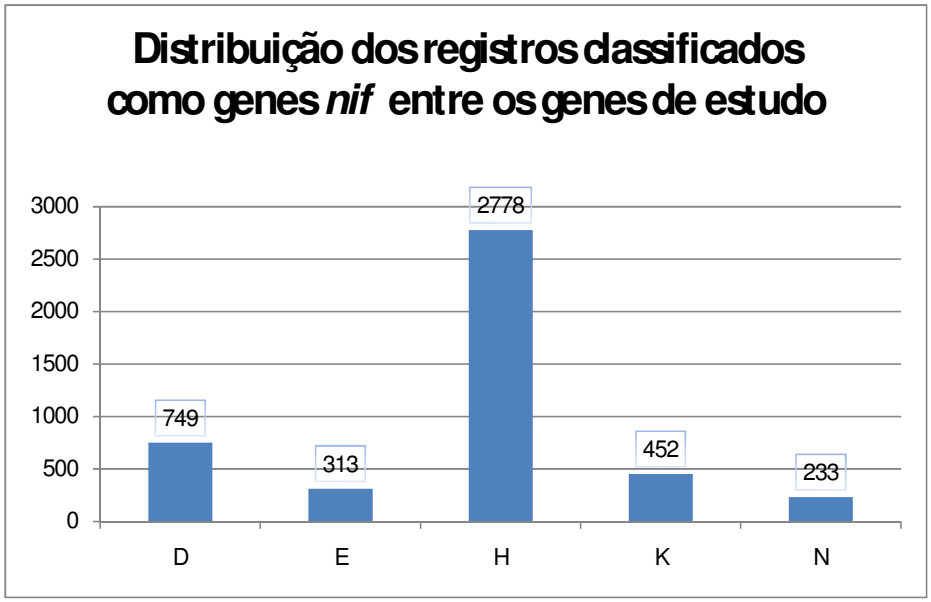


FIGURA 18 – DISTRIBUIÇÃO DOS REGISTROS CLASSIFICADOS COMO GENES *nif* ENTRE OS GENES DE INTERESSE DE ESTUDO.
FONTE: A Autora, (2010).

O gene *nifH* (2778 ocorrências) foi o gene de maior incidência apresentada entre os dados obtidos, seguido pelos genes *nifD* (749 ocorrências) e *nifK* (452 ocorrências) respectivamente.

Uma vez tendo-se a visão geral dos dados classificados foi possível realizar um mapeamento das informações obtidas, relacionando os genes *nif* encontrados com seus respecivos organismos. A próxima seção descreve em detalhes esse processo.

4.8. MAPEAMENTO ORGANISMOS X GENES *nifHDKEN*

Com a classificação dos genes *nif* realizada através da rede FAN, com média de 98,83% de certeza, partiu-se então para a criação de um mapa, agrupando os 4525 genes *nif* encontrados por seus respectivos organismos.

O resultado completo deste mapeamento pode ser verificado na seção de Apêndices deste trabalho, sendo que o Apêndice 9 apresenta o mapa relacionado aos organismos com genomas completos e o Apêndice 10 contempla os dados relacionados aos genomas incompletos. Para melhor organização das informações, os dados se encontram também agrupados por grupos taxonômicos.

Adicionalmente, a coluna **Young, 1992**, presente em ambos os mapas, indica quais organismos encontrados já faziam parte do estudo sobre organismos fixadores de nitrogênio, publicado por Young, em 1992 (YOUNG, 1992).

O Quadro 5 sumariza e compara os dados apresentados nos mapas gerados (de organismos com genomas completos e incompletos), realizando totalizações pelos grupos taxonômicos.

Grupos Taxonômicos			**Genomas**	
			Completos	**Incompletos**
Archaea	Euryarchaeota	environmental samples	1	0
		Halobacteria	1	0
		Methanobacteria	6	7
		Methanococci	12	6
		Methanomicrobia	14	3
		Methanopyri	1	0
	Total		***35***	***16***
Total			***35***	***16***
Bactéria	Actinobacteria		4	156
	Aquificae		2	0
	Bacteroidetes/Chlorobi group		12	6
	Chlamydiae/Verrucomicrobia group		2	3
	Chloroflexi		3	2
	Cyanobacteria		14	279
	Deferribacteres		1	0
	environmental samples		0	45
	Fibrobacteres/Acidobacteria group		1	0
	Firmicutes		34	131
	Fusobacteria		2	5
	Nitrospirae		1	2
	Proteobacteria	Alphaproteobacteria	37	740
		Betaproteobacteria	11	254
		delta/epsilon subdivisions	24	48
		Gammaproteobacteria	14	163
		unclassified Proteobacteria	1	4
	Spirochaetes		2	11
	Synergistetes		0	1
	unclassified Bactéria		0	56
Total			***165***	***1906***
Eukaryota	Amoebozoa	Polysphondylium pallidum PN500	0	1
	Bacillariophyta	Rhopalodia gibba	0	1
	Viridiplantae	Volvox carteri f. nagariensis	0	1
Total			***0***	***3***
TOTAL GERAL			**200**	**1925**

QUADRO 5 – SUMARIZAÇÃO DOS DADOS APRESENTADOS NOS MAPAS DE RELACIONAMENTO ENTRE ORGANISMOS E GENES *nif* , PARA GENOMAS COMPLETOS E INCOMPLETOS.
FONTE: A Autora, (2010).

De acordo com os resultados obtidos, os 4525 genes *nif* obtidos através dos processos de mineração e classificação de dados, se encontravam relacionados à 2125 organismos diferentes. Destes, 200 possuíam seu genoma completo já depositado no NCBI GenBank e 1925 ainda se encontram em fase de seqüenciamento.

Considerando-se os 1425 genomas completos já depositados no NCBI GenBank (dados obtidos em 28/01/2011, a partir do site do próprio NCBI), 14,03% apresentam em sua anotação, pelo menos um gene *nif*.

Sem se considerar estirpes dos organismos estudados, 646 espécies diferentes de organismos foram encontradas. Destas, 158 (24,46%) possuíam seu genoma completamente anotado, enquanto, 488 (75,54%) ainda permanecem com seu genoma incompleto.

A Figura 19 apresenta graficamente a distribuição dos organismos com genomas completos e incompletos, encontrados no Reino Archaea.

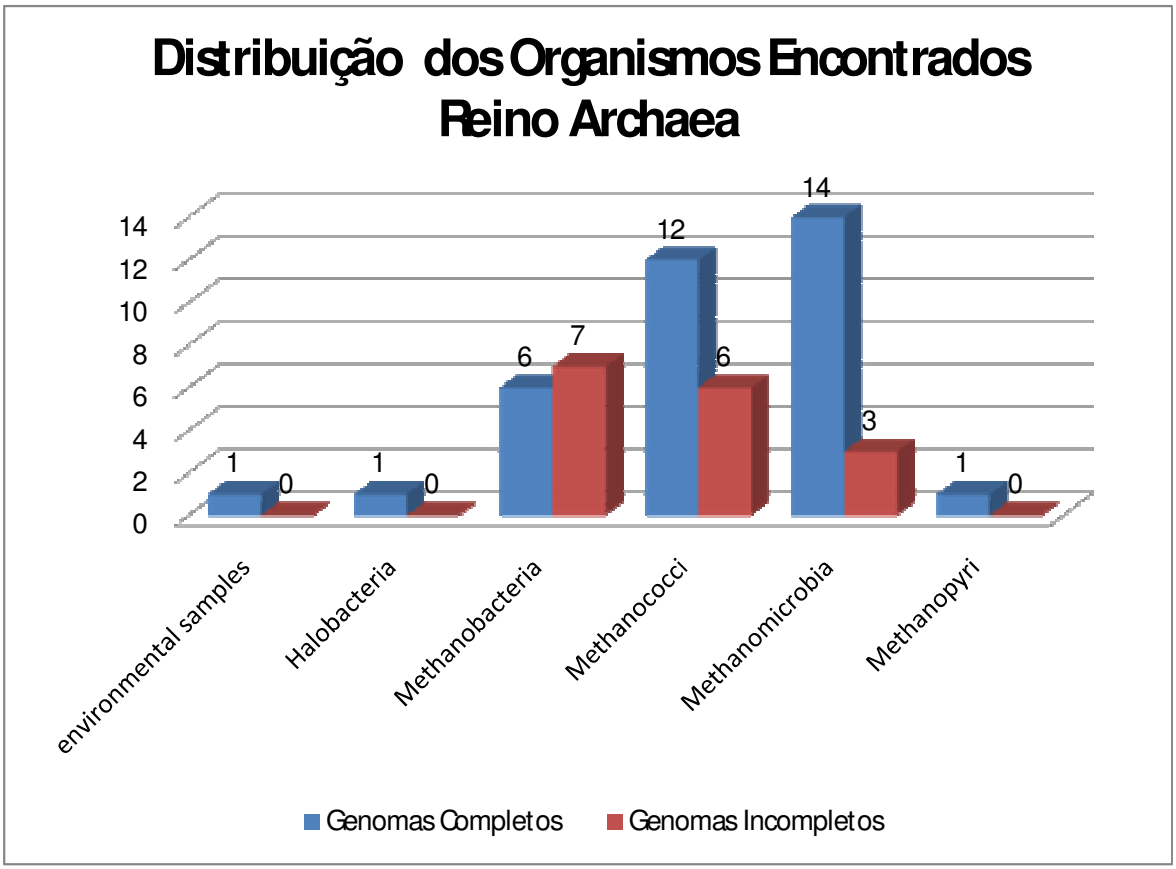


FIGURA 19 – DISTRIBUIÇÃO DOS ORGANISMOS DETENTORES DE GENES *nif* ENCONTRADOS NO REINO ARCHAEA.
FONTE: A Autora, (2011).

Conforme apresentado pelo gráfico, os grupos com maior expressão quantitativa encontrados foram: Methanobacteria, Methanococci e Methanomicrobia.

A Figura 20 apresenta graficamente a distribuição dos organismos com genomas completos e incompletos, encontrados no Reino Bactéria.

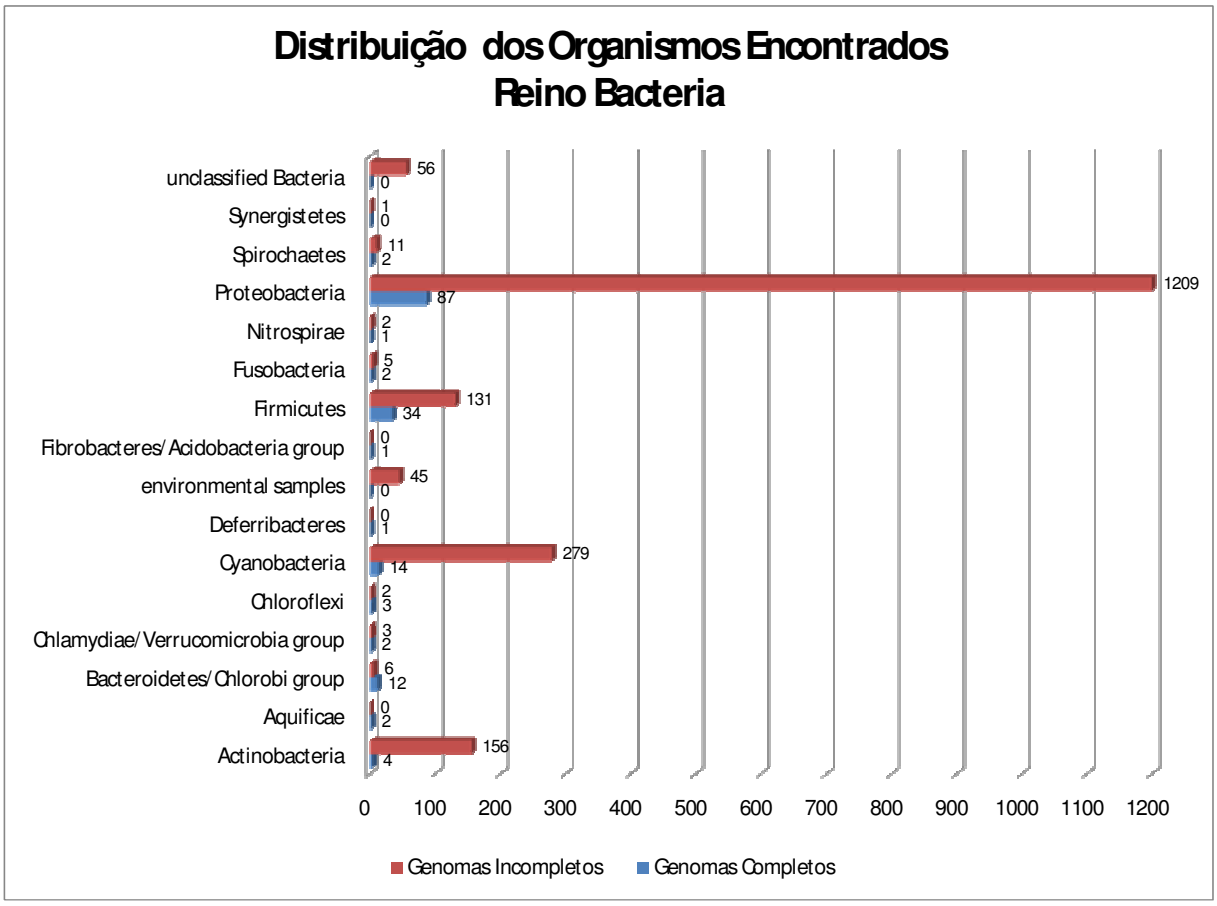


FIGURA 20 – DISTRIBUIÇÃO DOS ORGANISMOS DETENTORES DE GENES *nif* ENCONTRADOS NO REINO BACTÉRIA.
FONTE: A Autora, (2011).

Entre os grupos taxonômicos apresentados no gráfico destaca-se o grupo das Proteobactérias. A Figura 21 detalha graficamente o universo componente deste grupo.

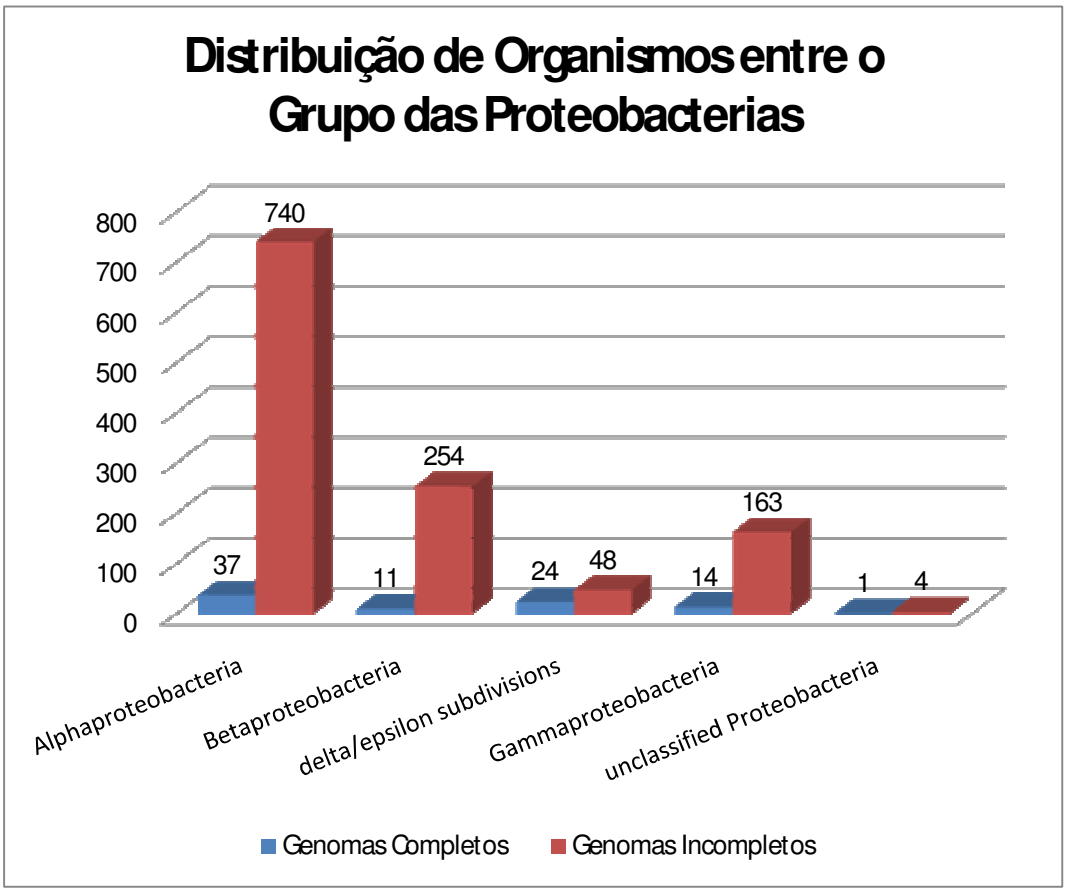


FIGURA 21 – DETALHAMENTO DO UNIVERSO COMPONENTE DO GRUPO DAS PROTEOBACTÉRIAS.
FONTE: A Autora, (2011).

Entre os 200 organismos com genomas completos, em 129 conseguiu-se detectar todos os cinco genes em estudo (*nifHDKEN*). Já entre os 1925 organismos com genomas incompletos, em apenas 68 foi possível a detecção de todos os genes estudados. A configuração estrutural da proteína nitrogenase, codificada pelos genes *nifHDK* foi encontrada em 48 organismos com genomas incompletos e em apenas 3 organismos com genomas completos. No restante dos organismos mapeados, as mais diversas combinações possíveis entre os cinco genes de estudo foi encontrada.

É importante frisar, que o fato de se encontrar a configuração total dos genes de estudo em um determinado organismo não atesta e nem muito menos garante que o mesmo se trate de um fixador de nitrogênio. Análises laboratoriais e estudos de inoculação, bem como constante consulta à literatura relacionada ao tema são indispensáveis para a confirmação das propriedades fixadoras de nitrogênio em um organismo.

4.9. TESTES DE REPETIBILIDADE DO PROCESSO DE MINERAÇÃO DE DADOS

Como o NCBI GenBank é atualizado diariamente, desenvolver um processo de mineração de dados passível de repetições e que permitisse atualizações constantes dos dados, sem perda de informações era essencial para esta pesquisa.

Ao longo do desenvolvimento do trabalho, inúmeras vezes o processo de mineração de dados foi testado, corrigido e aperfeiçoado. De forma a se manter um registro de métrica relacionando o tempo dispendido e o total de resultados obtidos, por três vezes os dados resultantes do processo de mineração de dados foram registrados (em 25/05/10, 25/08/10 e 25/11/10) também como forma de garantir a confiabilidade e a repetibilidade do processo de mineração de dados.

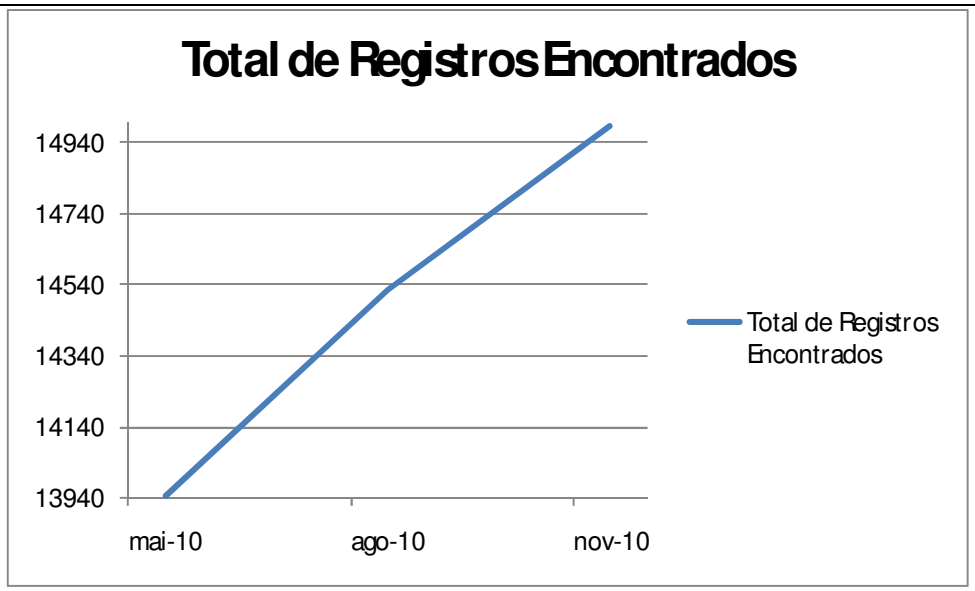
A Tabela 9 apresenta dados referentes às datas base para o registro da métrica adotada, ao número de informações relacionadas aos genes em estudo (*nifHDKEN*) encontradas a cada execução da técnica, (sempre crescentes) e ao tempo total gasto em cada ciclo completo do processo de mineração de dados (decrescente ao longo do tempo). O Apêndice 11 apresenta um quadro contendo as três aferições de tempo, detalhadas por procedimento (*step*) executado.

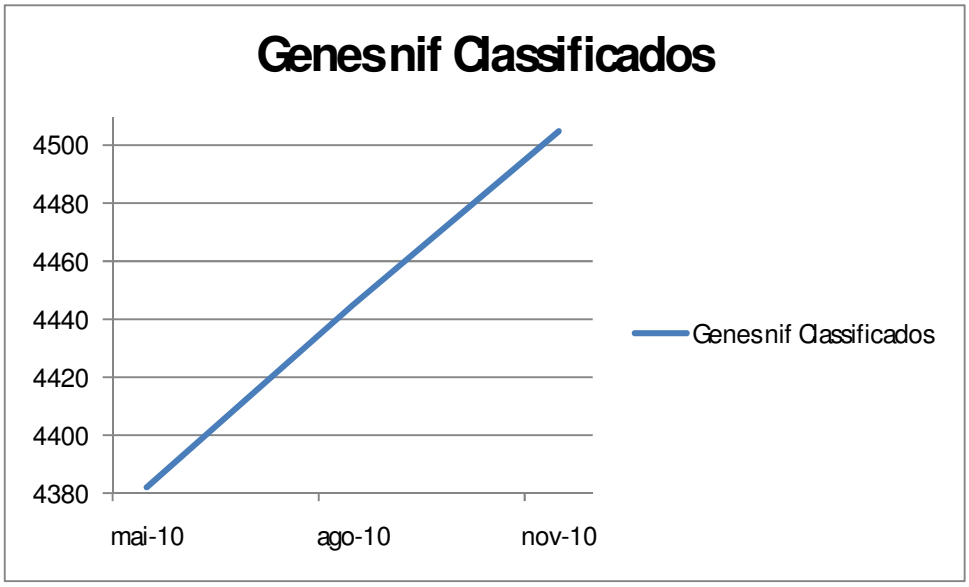
TABELA 9 – AFERIÇÃO DA RELAÇÃO TEMPO X RESULTADOS

Data de Execução	25/05/2010	25/08/2010	25/11/2010
Tempo (hh:mm:ss)	25:13:15	15:57:04	08:33:52
Total de Registros Encontrados	13949	14527	14988
Genes *nif Classificados*	4382	4445	4525

Fonte: A autora, (2010).

A Figura 22 apresenta graficamente as medidas da relação Tempo X Resultados, aferidas durante a pesquisa.





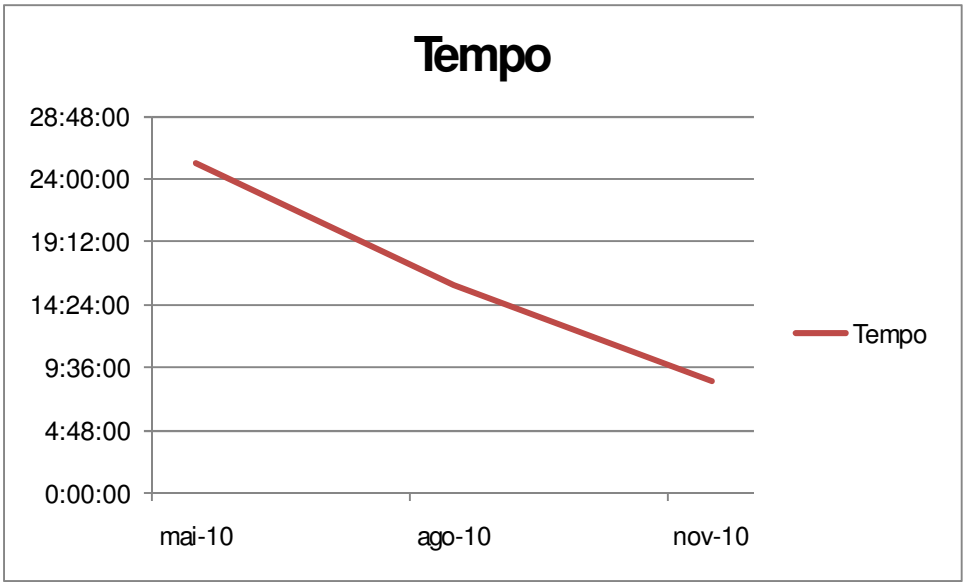


FIGURA 22-AFERIÇÃO DA RELAÇÃO TEMPO x RESULTADOS
Fonte: A autora, (2010).

A Tabela 10 apresenta a discriminação dos resultados obtidos na aferição realizada, por gene *nif* estudado.

TABELA 10 – AFERIÇÃO DA RELAÇÃO TEMPO X RESULTADOS – DETALHAMENTO POR GENES *nif* EM ESTUDO.

Datas	Tempo (hh:mm:ss)	Quantidades	Genes *nif*					Total
			D	E	H	K	N	
25/05/2010	25:13:15	Total de Registros Encontrados	923	806	10643	1084	493	13949
		Genes *nif Classificados*	661	301	2798	402	220	4382
25/08/2010	15:57:04	Total de Registros Encontrados	932	964	11045	1093	493	14527
		Genes *nif Classificados*	700	313	2785	425	222	4445
25/11/2010	08:33:52	Total de Registros Encontrados	938	1067	11378	1110	495	14988
		Genes *nif Classificados*	742	321	2764	455	223	4525

Fonte: A autora, (2010).

De acordo com uma comparação manual dos registros encontrados em cada uma das aferições, constatou-se a manutenção das informações de uma aferição para outra e a inclusão de novos dados a cada aferição, confirmando a credibilidade da metodologia criada para mineração e classificação de dados.

4.10. AJUSTES NO PROCESSO DE SELEÇÃO DE CARACTERÍSTICAS

Para se chegar a um conjunto consiso e eficiente de características, vários testes foram realizados.

O primeiro agrupamento de características extraídas a partir do dados foi o relacionado ao grupo I – Físico-Químicas, porém, os testes de treinamento da rede neural, apenas com esse grupo, apresentaram um resultado muito aquém do esperado, com valores de: 80,15% no Treinamento; 78,45% no Teste e 79,84% na Validação.

Em face a resultados tão inexpressivos, novos experimentos computacionais puderam extrair dos dados, as características relacionadas ao grupo II – Inferidas, porém com pouca melhora nos resultados referentes ao processo de aprendizagem da rede neural artificial. Os valores obtidos foram: 82,78% durante o Treinamento; 80,74% no Teste e 80,25% na Validação.

Como os resultados não melhoraram de acordo com o esperado, um terceiro grupo de características foi agregado (grupo III-BLAST) somando aos dados, valores normalizados, originados a partir da execução da ferramenta BLAST.

Com a inserção deste novo grupo, os valores relacionados ao grau de aprendizagem da rede neural elevaram-se: 90,52% na etapa de Treinamento; 91,15% no Teste e 91,37% na Validação.

No intuito de melhorar ainda mais os resultados, o grupo IV, relacionado às características experimentais foi incorporado ao conjunto de características. De fato, os valores relacionados ao aprendizado da rede melhoraram a contento, sendo então mantida a configuração relacionando os 4 grupos de características, sendo encontrado os valores: 98,83% no Treinamento; 98,64 no Teste e 99,01% na Validação.

Quinze testes completos, abrangendo as combinações possíveis entre os grupos de características encontradas foram realizados, na tentativa de se encontrar uma combinação de grupos que apresentasse um melhor resultado. Os resultados obtidos em cada uma das fases do aprendizado da rede neural se encontram descritos na Tabela 11.

TABELA 11 – RESULTADOS OBTIDOS DURANTE OS TESTES DE GRUPOS DE CARACTERÍSTICAS.

Situação	Treinamento	Teste	Validação	Média
Grupos I e II	82.78	80.74	80.25	81.26
Grupos I e III	80.29	78.93	79.99	80.25
Grupos I e IV	77.65	77.00	77.05	77.23
Grupos I, II e III	90.52	91.15	91.37	91.01
Grupos I, II e IV	79.75	78.75	79.75	79.42
Grupos I, II, III e IV	**98.83**	**98.64**	**99.01**	**98.83**
Grupos I, III e IV	89.97	89.15	88.48	89.20
Grupos II e III	71.65	71.50	71.65	71.60
Grupos II e IV	75.55	77.20	75.55	76.10
Grupos II, III e IV	90.15	91.00	90.47	90.54
Grupos III e IV	90.45	90.78	90.21	90.48
Somente Grupo I	80.15	78.45	79.84	79.48
Somente Grupo II	70.05	71.15	69.90	70.37
Somente Grupo III	75.65	75.66	75.67	75.66
Somente Grupo IV	64.75	66.25	66.75	65.92

FONTE: A Autora, (2011).

Como a combinação dos quatro grupos de características foi a melhor encontrada, a mesma foi mantida para utilização na pesquisa.

4.11. A IMPORTÂNCIA DO CO-APRENDIZADO EM REDES NEURAIS ARTIFICIAIS

Agregar à esta pesquisa a técnica de co-aprendizado descrita por Cardoso e Cruz (CARDOSO E CRUZ, 2007) foi relevante no sentido de minimizar os valores encontrados para erros de classificação apresentados pela rede neural artificial.

A técnica foi aplicada em três ciclos de análise do aprendizado da rede neural, sempre contemplando os 14988 registros encontrados no processo de mineração de dados.

Para o primeiro ciclo de aprendizagem da rede, os dados foram divididos em conjuntos de Treinamento, Teste e Validação, de acordo com as informações apresentadas no Quadro 6.

Conjuntos de Dados			
Total Geral de Registros	14988		
Total de Registros Previamente Classificados	1044 (6,96%)	**Registros por Classe**	
		Não *nifs*	***nifs***
		720	324
Total de Registros do Arquivo de Treinamento	348 (1/3)	**Registros por Classe**	
		Não *nifs*	***nifs***
		240	108
Total de Registros do Arquivo de Teste	348 (1/3)	**Registros por Classe**	
		Não *nifs*	***nifs***
		240	108
Total de Registros do Arquivo de Validação	348 (1/3)	**Registros por Classe**	
		Não *nifs*	***nifs***
		240	108

QUADRO 6 – DIVISÃO DOS CONJUNTOS DE DADOS PARA UTILIZAÇÃO NO PRIMEIRO CICLO DE APRENDIZADO DA REDE FAN.
FONTE: A Autora, (2010).

Com apenas 6,96% do total de registros previamente classificados, baseando-se nas características pré-definidas, a rede alcançou um percentual médio de acerto real de 94,53%. A veracidade da classificação foi constatada através de análise manual dos registros.

Os 5,47% de divergência, correspondente à 820 registros foram reavaliados, manualmente, um a um, dentro do processo de co-aprendizado. As sequências relacionadas a cada um destes registros foram submetidas novamente à ferramenta BLASTP.

Após a análise realizada, constatou-se uma taxa de erro real da rede de 80,76% em relação aos registros divergentes, caracterizando um valor muito alto.

Aplicando a técnica de co-aprendizado, os registros divergentes foram reclassificados com base nos resultados obtidos a partir do BLASTP.

A partir dos dados reclassificados, para o segundo ciclo de aprendizagem, houve uma nova divisão dos dados, em conjuntos de Treinamento, Teste e Validação, conforme demonstrado no Quadro 7.

Conjuntos de Dados			
Total Geral de Registros	14988		
		Registros por Classe	
Total de Registros Previamente Classificados	4500 (30,02%)	**Não *nifs***	***nifs***
		3165	1335
		Registros por Classe	
Total de Registros do Arquivo de Treinamento	1500 (1/3)	**Não *nifs***	***nifs***
		1034	466
		Registros por Classe	
Total de Registros do Arquivo de Teste	1500 (1/3)	**Não *nifs***	***nifs***
		1074	426
		Registros por Classe	
Total de Registros do Arquivo de Validação	1500 (1/3)	**Não *nifs***	***nifs***
		1057	443

QUADRO 7 – DIVISÃO DOS CONJUNTOS DE DADOS PARA UTILIZAÇÃO NO SEGUNDO CICLO DE APRENDIZADO DA REDE FAN.
FONTE: A Autora, (2010).

Após o primeiro ciclo de co-aprendizado, já foi possível a classificação correta de um número maior de dados, sendo que, cerca de 30% pôde ser utilizado para a divisão entre os grupos de Treinamento, Teste e Validação.

Depois da submissão dos dados à um novo processo de aprendizagem, o valor do percentual médio de acerto real da rede subiu para 97,62%.

Novamente, os registros divergentes, correspondentes a 2,38% (ou 357 registros) foram reavaliados manualmente e reclassificados conforme retorno fornecido pela ferramenta BLASTP. Neste ciclo, com base na análise realizada, foi constatado uma taxa de erro real da rede de 54,34% para os registros divergentes. O resultado foi bastante animador, pois em apenas um ciclo de co-aprendizado, conseguiu-se baixar a taxa de erros nas divergências em 25%.

Em face aos bons resultados obtidos com o processo de co-aprendizado, os dados reclassificados foram novamente divididos entre os conjuntos de Treinamento, Teste e Validação.

Conforme o Quadro 1, nesta fase do co-aprendizado, todos os 14988 registros foram utilizados para o aprendizado da rede, sendo divididos em 3 partes de igual proporção.

Após o aprendizado, constatou-se apenas 176 registros com valores divergentes da pré classificação realizada durante o co-aprendizado, perfazendo um percentual de 1,17%. Porém, entre os valores divergentes, após um novo ciclo de análise dos resultados foi verificado que a rede errou 100% dos casos divergentes, ou seja, atingiu seu valor máximo de aprendizado.

Este conjunto de 176 registros que a rede não conseguiu classificar de maneira correta, mesmo após um último ciclo de co-aprendizado era composto principalmente de anotações sequências protéicas com tamanho menor que 50 aminoácidos e sequências anotadas como hipotéticas.

Com a obtenção de um alto valor médio de acerto, de 98,83% e uma estabilização no valor de erros obtidos nas divergências da rede, considerou-se como atingido o objetivo de alcance do melhor valor de classificação possível para esta rede neural, baseado nas características analisadas. A Tabela 12 apresenta de forma resumida os resultados alcançados no processo de co-aprendizado da rede.

TABELA 12 – TABELA DE PROGRESSÃO DO PERCENTUAL REAL DE ACERTOS DA REDE, BASEADOS NA TÉCNICA DE CO-APRENDIZADO.

Ciclo	**% Divergência**		**% Real de Acerto**
01	5,47		94,53
	% Erro	**% Acerto**	
	80,76%	19,24	
02	2,38		97,62
	% Erro	**% Acerto**	
	54,34%	45,66	
03	1,17		98,83
	% Erro	**% Acerto**	
	100%		

FONTE: A Autora, (2011).

Para se atingir este objetivo, o uso da técnica de co-aprendizado da rede foi essencial, pois possibilitou a redução da taxa de erro real da classificação em 78,61% (saindo de 5,47% para 1,17%).

4.12. POR QUE MINHA REDE AINDA APRESENTA ERROS?

Mesmo com cuidados tomados como testes para maximização de grupos de características e co-aprendizado, a rede neural treinada ainda não conseguiu atingir 100% de acerto.

Após análise realizada sobre o fato, foram verificadas as seguintes potenciais situações que interferiram na qualidade do conjunto de treinamento e consequentemente nos resultados do processo de aprendizado como um todo:

- Nem todos os dados de sequências protéicas, disponibilizados pelo NCBI GenBank correspondem a informações curadas experimentalmente em laboratório. Mesmo embora originadas de montagens genômicas consistentes, nem todas as sequências são analisadas laboratorialmente de forma a garantir sua expressão e função gênica.
- Como tem sido amplamente apresentado na literatura, em trabalhos como de Poptsova e Gogarten (POPTSOVA E GOGARTEN, 2010) e Schnoes e colaboradores (SCHNOES et al., 2009), problemas de anotação nos genomas fechados, pode-se considerar como possibilidade, a eventual existência de inconsistências nas anotações utilizadas.
- É da natureza das redes neurais artificiais, como mecanismos de classificação tolerantes a erros, a busca por situações próximas a um patamar ótimo, porém a totalidade de acertos não faz parte de suas metas.

Em face as situações apresentadas, concluiu-se que o valor de 98,83% como média de aprendizado pode ser caracterizado como muito bom, validando assim a pesquisa realizada.

4.13. REDE NEURAL *VERSUS* LINHA DE CORTE

De forma a assegurar a potencialidade de uso de uma rede neural artificial, em detrimento de outras técnicas de classificação, como as linhas de corte, foi realizado um teste comparativo entre estas duas técnicas, baseado nos resultados finais obtidos no processo de mineração de dados.

O teste avaliou a acurácia da classificação da rede neural artificial em comparação com outras três estratégias diferentes de linhas de corte (todas baseadas em campos específicos, retornados pela ferramenta BLAST), a saber:

- **e-Value > 10E-06**: Classificação de um registro como "1 – Gene *nif*", baseada no valor do campo e-Value maior que 10E-06.
- **0.30 Id X 0.50 Sm (QUERY)**: Classificação de um registro como "1 – Gene *nif*", baseada nos valores dos campos Identidade >= 30% e Similaridade >= 50% a partir do tamanho total da sequência Query de busca.
- **0.30 Id X 0.50 Sm (BLAST)**: Classificação de um registro como "1 – Gene *nif*", baseada nos valores dos campos Identidade >= 30% e Similaridade >= 50% a partir do espaço total do alinhamento retornado pela ferramenta BLAST.

Conforme visto em trabalhos como o de Shoemaker e colaboradores, em 2010 (SHOEMAKER et al., 2010), Ye e Godzik, de 2004 (YE E GODZIK, 2004) e McClure, Vasi e Fitch, em 1994 (MCCLURE, VASI E FITCH, 1994), entre outros, os valores para classificação de homologia em sequências protéicas relacionados a Identidade >= 30% e Similaridade >= 50% são largamente utilizados, sendo este o motivo por terem sido escolhidos para uso nesta pesquisa.

Os 14988 registros, já devidamente classificados de maneira correta, em sua totalidade foram submetidos as 3 linhas de corte propostas e o Quadro 8 apresenta um comparativo com os resultados alcançados, contemplando, além de cada uma das situações em análise, a quantidade de registros correspondentes a genes *nif* e a genes não-*nif*. São apresentados os valores totais e os percentuais de acertos e erros em cada uma das situações analisadas.

Avaliação	***nifs***		**Não *nifs***		**Acertos**	**Erros**	**% de Acertos**	**% de Erros**
Situação real	4603		10385					
Rede	4525		10463		14812	176	98,83	1,17
	nifs Reais	*Não nifs Reais*	*nifs Reais*	*Não nifs Reais*				
	4476	49	127	10336				
e-Value>10E-06	5308		9680		14075	913	93,91	80,79
	nifs Reais	*Não nifs Reais*	*nifs Reais*	*Não nifs Reais*				
	4499	809	104	9576				
0.30 Id X 0.50 Sm (QUERY)	4398		10590		13943	1045	93,03	83,21
	nifs Reais	*Não nifs Reais*	*nifs Reais*	*Não nifs Reais*				
	3978	420	625	9965				
0.30 Id X 0.50 Sm (BLAST)	3760		11228		13373	1615	89,22	89,15
	nifs Reais	*Não nifs Reais*	*nifs Reais*	*Não nifs Reais*				
	3374	386	1229	9999				

QUADRO 8 – COMPARATIVO ENTRE A REDE NEURAL FAN E DIFERENTES LINHAS DE CORTE.
FONTE: A Autora, (2010).

Conforme apresentado pelo Quadro 6, a classificação via rede neural artificial FAN obteve um percentual médio de acertos 6,85% maior que as demais linhas de corte e uma média de erros 84,38% menor, provando assim a eficiência da utilização da rede FAN para o problema classificatório em questão.

4.14. POR QUE UTILIZAR A REDE FAN?

A decisão pelo uso da rede neural artificial FAN se deu pelos resultados obtidos a partir da comparação da mesma com uma implementação nativa da rede neural modelo MLP, no software MatLab.

Na melhor das situações encontradas, a rede FAN atingiu um resultado médio 3,83% melhor que a implementação MLP do MatLab. O Quadro 9 apresenta em detalhes todos os valores obtidos nos testes realizados nas redes neurais FAN e MLP, destacando o melhor resultado médio encontrado pelas duas redes testadas. Os mesmos conjuntos de dados para treinamento, teste e validação, em cada um dos casos analisados, foi utilizado para a comparação das redes.

Situação	Rede FAN				Rede MLP			
	Treinamento	Teste	Validação	Média	Treinamento	Teste	Validação	Média
Grupos I e II	82.78	80.74	80.25	81.26	78.64	82.35	77.73	79.58
Grupos I e III	80.29	78.93	79.99	80.25	78.28	76.96	81.19	78.81
Grupos I e IV	77.65	77	77.05	77.23	73.77	78.16	73.20	75.04
Grupos I, II e III	90.52	91.15	91.37	91.01	92.33	88.87	91.83	91.01
Grupos I, II e IV	79.75	78.75	79.75	79.42	80.95	79.93	80.15	80.34
Grupos I, II, III e IV	**98.83**	**98.64**	**99.01**	**98.83**	**93.89**	**93.71**	**97.52**	**95.04**
Grupos I, III e IV	89.97	89.15	88.48	89.2	87.14	89.60	88.92	88.55
Grupos II e III	71.65	71.5	71.65	71.6	69.40	72.93	69.86	70.73
Grupos II e IV	75.55	77.2	75.55	76.1	71.77	78.36	71.77	73.97
Grupos II, III e IV	90.15	91	90.47	90.54	91.95	88.14	92.28	90.79
Grupos III e IV	90.45	90.78	90.21	90.48	88.19	86.24	90.66	88.36
Somente Grupo I	80.15	78.45	79.84	79.48	81.75	79.63	77.84	79.74
Somente Grupo II	70.05	71.15	69.9	70.37	66.55	72.57	70.95	70.02
Somente Grupo III	75.65	75.66	75.67	75.66	77.16	73.77	71.89	74.27
Somente Grupo IV	64.75	66.25	66.75	65.92	61.51	67.58	65.08	64.72

QUADRO 9 – COMPARATIVO ENTRE AS REDES NEURAIS FAN E MLP.
FONTE: A Autora, (2010).

Para o uso da rede MLP embarcada na *toolbox* de Redes Neurais do MatLab, foram utilizadas as seguintes funções: mlp_tr(), para treinamento da rede MLP e mlp_ts(), para teste e validação da rede MLP treinada.

As funções foram desenvolvidas e gratuitamente distribuídas pelo professor Roberto Raittz, orientador desta pesquisa. Seus códigos fontes se encontram disponíveis no Apêndice 12 deste trabalho.

Em face ao resultado encontrado e considerando o potencial gráfico e a facilidade de acesso disponibilizada pela implementação EasyFan, a rede FAN foi escolhida para compor a etapa classificatória de dados, desta pesquisa.

4.15. RESULTADOS X LITERATURA

Os 2125 organismos mapeados foram confrontados com o trabalho referência publicado por Young, em 1992 (YOUNG, 1992).

Se considerarmos as estirpes dos organismos, 762 dos 2125 (35,85%) se encontravam descritos como fixadores biológicos de nitrogênio no trabalho de Young: 77 com genoma completo e 685 com genoma incompleto.

Levando em consideração apenas o gênero e a espécie dos organismos, foram mapeados 646 diferentes organismos. Destes, 205 foram anteriormente

descritos por Young, e, com base neste último número, verificou-se que 51 organismos possuíam seu genoma completo, depositado no NCBI GenBank, enquanto 154 ainda se encontravam com o genoma incompleto. O Quadro 10 sumariza os dados encontrados.

Situações Analisadas	Considerando Estirpes			Sem Considerar as Estirpes		
	Completos	Incompletos	Total	Completos	Incompletos	Total
A) Resultados Mapeados	200 (9.41%)	1925 (90.59%)	2125	158 (24.46%)	488 (75.54%)	646
B) Constantes em Young, 1992	70 (9.19%)	685 (89.91%)	762	51 (24.88%)	154 (75.12%)	205

QUADRO 10 – SUMARIZAÇÃO DE RESULTADOS DA COMPARAÇÃO ENTRE OS REGISTROS ENCONTRADOS E O TRABALHO PUBLICADO POR YOUNG (YOUNG, 1992).
FONTE: A Autora, (2011).

Traçando um pararelo percentual entre os resultados encontrados nesta pesquisa e o trabalho original publicado por Young, já levando-se em consideração os agrupamentos realizados por este último para as Cianobactérias, cerca de 64% dos organismos descritos na pesquisa realizada por Young em 1992 possuem pelo menos uma sequência protéica, relacionada a um gene *nif*, depositada do NCBI GenBank.

É importante notar que cerca de apenas 30% dos resultados obtidos nesta pesquisa, correspondem a 64% dos organismos descritos como fixadores biológicos de nitrogênio, por Young. Ou seja, aproximadamente 70% dos organismos listados, embora necessitem de maiores confirmações, tanto na literatura, quanto em experimentos laboratoriais podem se tratar de novos potenciais fixadores de nitrogênio. Por outro lado, também é interessante observar que 36% dos organismos descritos por Young há mais de uma década, até o ano de 2010 ainda não possuíam sequer uma sequência relacionada a genes *nif* depositada em um banco de dados público.

4.16. APLICABILIDADE DO FERRAMENTAL DESENVOLVIDO

Todo o ferramental descrito neste trabalho, bem como o passo a passo de sua execução para usos futuros serão disponibilizados em um espaço apropriado, dentro do sítio do Programa de Pós-Graduação em Bioinformática (http:\\www.bioinfo.ufpr.br).

6. CONCLUSÕES

- Até a data de 25/11/2010 foram encontrados, através da metodologia desenvolvida para a mineração de dados, 14988 registros referentes a anotações de sequências protéicas relacionadas a genes *nifHDKEN.*

- Estes puderam ser agrupados em 2125 organismos diferentes, considerando-se as estirpes dos mesmos, sendo que 200 (9,42%) possuíam genomas completos anotados no NCBI GenBank e 1925 (90,58%) se encontravam com seus genomas ainda incompletos.

- O grupo taxonômico com maior concentração de organismos encontrados foi o das Proteobactérias, com 1296 (60,98%) organismos diferentes: 87 com genomas completos e 1209 com genomas incompletos.

- Sem considerar as estirpes dos organismos mapeados, puderam ser relacionados 646 organismos diferentes, divididos em 158 (24,46%) organismos com genoma completamente seqüenciado e anotado e 488 (75,54%) organismos com genomas incompletos.

- Considerando-se os 1425 genomas completos já depositados no NCBI GenBank (dados obtidos em 28/01/2011, a partir do site do próprio NCBI), **14,03**% apresentam em sua anotação, pelo menos um gene *nif*.

- A utilização de redes neurais artificiais foi comprovadamente testada como melhor metodologia para classificação dos dados obtidos, em detrimento de outras estratégias de linhas de corte.

- O uso da técnica de co-aprendizado para maximização dos resultados do aprendizado da rede neural artificial provou-se eficiente, proporcionando uma redução real do erro apresentado pela rede em 21,38%.

- A partir do paralelo traçado entre os dados encontrados e a literatura de referência pôde-se concluir que há muito campo ainda a ser pesquisado em relação aos organismos fixadores de nitrogênio e o trabalho conjunto da Bioinformática e das pesquisas bioquímicas experimentais em laboratório pode acelerar este processo.

7. PERSPECTIVAS

A disponibilização das informações encontradas no processo de mineração e classificação de dados, em forma de um sítio de acesso público, com um banco de dados automaticamente atualizável seria a grande meta deste trabalho, podendo contribuir significativamente com à comunidade de pesquisa em fixação biológica de nitrogênio considerando a especificidade das informações que seriam reunidas em um único local, dotado de consultas e relacionamentos direcionados, todos focando os organismos, autores e publicações e processos envolvidos com fixação biológica de nitrogênio.

O aumento do escopo de busca do processo de mineração de dados, de modo a atender todos os genes relacionados ao processo de fixação biológica de nitrogênio, bem como publicações e autores correlatos, de forma a se conseguir informações mais detalhadas sobre os organismos fixadores de nitrogênio seria de grande valia e auxílio nas pesquisas. Saber de uma maneira fácil e simplificada, quais grupos taxonômicos são capazes de fixar nitrogênio, ou em quais os ambientes já foi detectada a presença de fixadores de N_2, ou ainda se os organismos fixadores de nitrogênio podem ocorrer em ambientes aeróbios, anaeróbios, extremos (em termos de temperatura, pH, salinidade, etc.) ou oligotróficos, e se existe relacão entre estes ambientes e a ocorrencia dos organismos fixadores de N_2, além de corroborar com as pesquisas em andamento sobre fixação biológica de nitrogênio, instigaria novos *insights* aos pesquisadores.

A consolidação das informações disponibilizadas, ou ainda a confirmação de indícios de novos possíveis organismos fixadores de nitrogênio através de literatura especializada e experimentos laboratoriais poderia contribuir de maneira incisiva nas pesquisas.

REFERÊNCIAS

1. ALTSCHUL, S.F.; GISH, W.; MILLER, W.; MYERS, E.W.; LIPMAN, D.J.; Basic local alignment search tool. J. Mol. Biol., 215:403–410, 1990.

2. ALTSCHUL, S.F.; MADDEN, T.L.; SCHAFFER, A.A.; ZHANG, J.; ZHANG, Z.; MILLER, W.; LIPMAN, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25:3389–3402, 1997.

3. ALTSCHUL, S.F.; WOOTTON, J.C.; GERTZ, E.M; AGARWALA, R.; MORGULIS, A.; SCHAFFER, A.A.; YU, Y. Protein database searches using compositionally adjusted substitution matrices", FEBS J. 272:5101-5109, 2005.

4. ALVES, W.A.L.; ARAÚJO, S.A.; LIBRANTZ, A.F.H. Reconhecimento de Padrões de Texturas em Imagens Digitais Usando uma Rede Neural Artificial Híbrida. Exacta, v. 4, n.2, p. 325-332, 2006.

5. ARNOLD, W.; RUMP, A.; KLIPP, W.; PRIEFER, U. B.; PUHLER, A. Nucleotide sequence of a 24206 base pair DNA fragment carrying the entire nitrogen fixation cluster of Klebsiella pneumoniae. J. Mol. Biol., Londres, Vol.230, p. 714-738, 1988.

6. BAYAT, A. Science, medicine, and the future Bioinformatics. BMJ, v. 324 p. 1018–22, 2002

7. BEDELL, J.; KORF, I.; YANDELL, M. BLAST. O'Reilly, p.360., 2003.

8. BENSANA, E.; BEL, G.; DUBOIS, D. Opal: A multi-Knowledge Based System for Industrial Job-Shop Scheduling. International Journal of Production Research, vol 26, Pp 795 –819, 1988.

9. BENSON, D. A.; KARSCH-MIZRACHI, I.; LIPMAN, D. J.; OSTELL, J.; RAPP, B. A.; WHEELER, D. L. GenBank. Nucleic Acids Research, Vol. 36(Database issue), p. D25–D30. 2008.

10. BENSON, D. A.; KARSCH-MIZRACHI, I.; LIPMAN, D. J.; OSTELL, J.; RAPP, B. A.; WHEELER, D. L. GenBank. Nucleic Acids Research, Vol. 35(Database issue), p. 21–25. 2007.

11. BENSON, D. A.; KARSCH-MIZRACHI, I.; LIPMAN, D. J.; OSTELL, J.; DAVID L.; WHEELER, D. L. Genbank. Nucleic Acids Research, Vol. 31-1, p. 23–27. 2003.

12. BENSON, D. A.; KARSCH-MIZRACHI, I.; LIPMAN, D. J.; OSTELL, J.; RAPP, B. A.; WHEELER, D. L. GenBank. Nucleic Acids Research, Vol. 30, p. 17–20. 2002.

13. BEZDEK, J. C.; PAL, S. K. (Editores) Fuzzy Models Pattern Recognition: Methods That Search for Structures in Data, IEEE, 1992.

14. BIOPYTHON. BioPython. Disponível em: <http://pyscience-brasil.wikidot.com/module:biopython>. Ultimo acesso: 11/01/2011a.

15. BIOPYTHON. Source Code for Module Bio.SeqUtils.ProtParam. Disponível em: <http://biopython.org/DIST/docs/api/Bio.SeqUtils.ProtParam-pysrc.html#ProteinAnalysis>. Último acesso: 11/01/2011b.

16. BIOPYTHON. Source Code for Module Bio.SeqUtils.IsoelectricPoint. Disponível em: <http://biopython.org/DIST/docs/api/Bio.SeqUtils.IsoelectricPoint-pysrc.html#IsoelectricPoint.pi>. Último acesso: 11/01/2011c.

17. BISHOP, C. M. Neural Networks for Pattern Recognition. Oxford University Press, 1995.

18. BJELLQVIST, B.; EK, K.; RIGHETTI, P.G.; GIANAZZA, E.; GÖRG, A.; WESTERMEIER, R.; POSTEL, W. Isoelectric focusing in immobilized pH gradients: principle, methodology and some applications. J Biochem Biophys Methods. Sep;6(4):317-39, 1982.

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

20. BORK, P. Shuffled domains in extracellular proteins. FEBS Lett, 286:47-54, 1991.

21. BOTHE, H.; BARBOSA, G.; DÖBEREINER, J. Nitrogen fixation and nitrate respiration by *Azospirillum brasilense.* Z Naturforsch, 38c : 571-577, 1983.

22. BRIGLE, K.E.; WEISS, M.C.; NEWTON, W.E.; DEAN, D.R. Products of the iron-molybdenum cofator –specific biosynthetic genes, nifE and nifN, are structurally homologous to the products of the Nitrogenase molybdenum-irom protein genes, nifD and nifK. J. Bacteriol., Washington, Vol.169, p.1547-1553, 1987.

23. BRU, C.; COURCELLE, E.; CARRÉRE, S.; BEAUSSE, Y.; DALMAR, S.; KAHN, D. The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Research, Vol. 33, D212–D215 (Database issue), 2005.

24. BURNS, R.C.; HARDY, K.W. Nitrogen Fixation in Bacteria and Higher Plants, Berlin:Springer-Verlag, 1975.

25. BURRIS, R.H. Nitrogenases. The Journal of Biological Chemistry, v.266, n.15 p. 9339-9342, 1991.

26. CARDOSO, R.L.A.; CRUZ, L.M. Análise *in silico* de proteínas transportadoras presentes no genoma de *Herbaspirillum seropedicae*. Monografia de Conclusão de Curso, UFPR, Curitiba, 2007.

27. CARREIRA-PERPIÑÁN, M. Á. A Review of Dimension Reduction Techniques, Technical Report CS-96-09, 1997.

28. CHAMBERLIN, D.D; ASTRAHAN, M.M.; ESWARAN, K.P; GRIFFITHS, P.P.; LORIE, R.A.; MEHL, J. W.; REISNER, P.; WADE, B.W. SEQUEL 2: a unified approach to data definition, manipulation, and control. IBM J. Res. Dev. 20, 6, 560-575, 1976.

29. CHEN, M-S.; HAN, J.; YU, P.S. Data mining: An overview from a database perspective. IEEE Trans. on Knowledge and Data Eng., 8(6), 866-883, 1996.

30. CNPQ. Instituto Nacional de Ciência e Tecnologia de Fixação Biológica de Nitrogênio. Disponível em: <http://www.cnpq.br/programas/inct/_apresentacao/inct_fixacao_biol_nitro.html>. Último acesso em: 17/01/2011.

31. CORNELL UNIVERSITY LIBRARY. NCBI Entrez Protein. Disponível em: <http://vivo.cornell.edu/individual/vivo/individual5746>. Último acesso em: 11/01/2011.

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

33. DASH, M.; LIU, H. Feature selection for classification. Intelligent Data Analysis, 1(1–4), 131–156, 1997.

34. DAYHOFF, M.O. Computer analysis of protein evolution. Sci Am., Jul, 221(1):86-95, 1969.

35. DIXON, R.A.; AUSTIN, S.; BUCK, M.; DRUMMOND, M.; HILL, S.; HOLTEL, A.; MACFARLANE, S.; MERRICK, M.; MINCHIN, S. Genetics and regulation of *nif* and related genes in *Klebsiella pneumoniae*. Philosophical Transactionns of the Royal Society, London, Série B, v.317, p.147, 1987.

36. DIXON, R.; KAHN, D. Genetic regulation of biological nitrogen fixation. Nat. Rev. Microbiol., 2:621-631, 2004.

37. DÖBEREINER, J. Recent changes in concepts of plant bacteria interactions: Endophytic N2 fixing bactéria. Ciência e Cultura, São Paulo, v.44, n.5, p.310-313, 1992

38. DRENTH, J.; JANSONIUS, J. N.; KOEKOEK, R.; SWEN, H. M.; WOLTHERS, B. G. Structure of papain. Nature, 218:929-932, 1968.

39. DUDA, O.; HART, P. E. Pattern classification and scene analysis. John Wiley & Sons, Inc., 1973.

40. DUNHAM, M. H. Data mining: Introductory and advanced topics. Upper Saddle River, NJ: Prentice Hall.1.6, 2003.

41. EDELMAN, G. M. Antibody structure and molecular immunology. Science, 180:830-840, 1973.

42. EHLERS, E. M.; VAN RENSBURG, E. An Object Oriented Manufacturing Scheduling Approach. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, vol 26 n 1, Pp 17 – 26, 1996.

43. ELMASRI, E.; NAVATHE, S. Sistemas de Bancos de Dados. Addison Wesley, Pearson, São Paulo, 4ª ed., 2005.

44. EXPASY PROTEOMICS SERVER. NifH/frxC family signatures and profile. Disponível em: <http://expasy.org/prosite/PDOC00580>. Último acesso em 12/11/2011. 2011a.

45. EXPASY PROTEOMICS SERVER. Nitrogenases component 1 alpha and beta subunits signatures. Disponível em: <http://expasy.org/prosite/PDOC00085>. Último acesso em 12/11/2011. 2011b.

46. EXPASY PROTEOMICS SERVER. moaA / nifB / pqqE family signature. Disponível em: <http://expasy.org/prosite/PDOC01009>. Último acesso em 12/11/2011. 2011c.

47. FANI, R.; GALLO, R.; LIO, P. Molecular evolution of nitrogen fixation: The evolutionary history of the nifD, nifK, nifE and nifN genes. J. Mol. Evol., 51:1-11, 2000.

48. FAUSETT, L. Fundamentals of Neural Networks Prentice Hal , Englewood, New Jersey. 1994, 461p.

49. FAYYAD, U. M. Data mining and knowledge discovery: Making sense out of data. IEEE Expert, 11(5), 20-25, 1996.

50. FAYYAD, U.M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (Ed.). Advances in knowledge discovery and data mining. Menlo Park, Cambridge, MA: AAAI/MIT Press, 1996.

51. FINN, R.D.; MISTRY, J.; TATE, J.; COGGILL, P.; HEGER, A.; POLLINGTON, J.E.; GAVIN, O.L.; GUNESEKARAN, P.; CERIC, G.; FORSLUND, K.; HOLM, L.; SONNHAMMER, E.L.; EDDY, S.R.; BATEMAN, A. The Pfam protein families database. Nucleic Acids Research. Database Issue 38:D211-222, 2010.

52. FORMAN, G. An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 3, 1289–1305, 2003.

53. FOX, J. What is Bioinformatics? The Science Creative Quarterly, Issue Four, 2009. Extraído de : http://www.scq.ubc.ca/what-is-bioinformatics/. Último acesso em: 16/12/2010.

54. FU, L. M. Neural Networks in Computer Intelligence. 2º Ed. New York: McGraw-Hill, Inc., 1994b.

55. GASTEIGER, E.; GATTIKER, A.; HOOGLAND, C.; IVANYI, I.; APPEL, R.D.; BAIROCH, A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. 31:3784-3788, 2003.

56. GENOPAR. Histórico. Disponível em: <http://nfn.genopar.org/nfn/historico.html>. Último acesso: 17/01/2011.

57. GEURTS, R.; BISSELING, T. Rhizobium nod factor perception and signaling. Plant Cell, 14:239-249, 2002.

58. GIBAS, C.; JAMBECK, P. Developing Bioinformatics Computer Skills, First Edition, April 2001, O'Reilly & Associates, Inc.

59. GUBLER, M.; HENNECKE, H. fixA B and C are essential for symbiotic and free living microaerobic nitrogen fixation. FEBS Lett, 200: 186-192, 1996.

60. GUIZELINI, D. Banco de Dados Biológico no Modelo Relacional para Mineração de Dados em Genomas Completos de Procariotos Disponibilizados pelo NCBI GenBank. Universidade Federal do Paraná. Dissertação de mestrado. 2010.

61. GURUPRASAD K.; REDDY, B.V.B.; PANDIT, M.W. Correlation between stability of a protein and its dipeptide composition: A novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Engineering, 4:155-161, 1990.

62. GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157–1182, 2003.

63. HARALICK, R.M. Statistical and structural approaches to texture. Proc IEEE. 67:786-804, 1979.

64. HAYKIN, S. Neural Networks – A Compreensive Foundation. Prentice-Hall, New Jersey, 2nd edition, 1999.

65. HAYKIN, S. Redes Neurais Princípios e Prática. Tradução de: Paulo Martins Engel. Porto Alegre: Bookman, 2001.

66. HSU, H. Advanced Data Mining Technologies in Bioinformatics. Idea Group Publishing, 329P, 2006.

67. HU, Y.; FAY, A.W.; LEE, C.C.; RIBBE, M.W. P-cluster maturation on Nitrogenase MoFe protein. Proc. Natl. Acad. Aci. USA, 104: 10424-10429, 2007.

68. HUGHEY, R.; KARPLUS, K. Bioinformatics: a new field in engineering education. J. Engineering Educ., p.101–104, 2003.

69. HUNGRIA, M.; CAMPO, R.J. Fixação biológica do nitrogênio em sistemas agrícolas. In: Congresso Brasileiro de Ciência do Solo, 30, 2005, Recife. Anais Recife: SBS, 2005.

70. IOANNIDIS, I.; BUCK, M. Nucleotide sequence of the Klebsiella pneumoniae nifD gene and predicted amino acid sequence of the subunit of Nitrogenase MoFe protein. Biochem. J., London, Vol.247, p.287-291, 1987.

71. IUPAC. Nomenclature and symbolism for amino acids and peptides (Recommendations 1983) Pure and Applied Chemistry 56 (5), 595 – 624, 1984. Disponível em: <http://www.iupac.org/objID/Article/pac5605x0595>. Último acesso: 15/01/2011.

72. JAIN, A.; ZONGKER, D. Feature selection: Evaluation, application, and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(2), 153–158, 1997.

73. JARGAS, A.M. Expressões Regulares: Guia de Consulta Rápida. Disponível em: <http://guia-er.sourceforge.net/index.html>. Último acesso: 17/01/2011.

74. JOHN, G. H.; KOHAVI, R.; PFLEGER, K. Irrelevant features and the subset selection problem. In: COHEN, W.W.; HIRSH, H. eds. Machine Learning: Proceedings of the Eleventh International Conference. San Francisco, CA: Morgan Kaufmann Publishers, pp. 121–129, 1994.

75. JOHNSON, M.; ZARETSKAYA, I.; RAYTSELIS, Y.; MEREZHUK, Y.; MCGINNIS, S.; MADDEN, T. L. NCBI BLAST: a better web interface. Nucleic Acids Res., 36(Web Server issue): W5–W9, 2008.

76. KASABOV, K. N. Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering. USA, Massachusetts Institute of Technology Press, 1996.

77. KIRA, K.; RENDELL, L. A. A practical approach to feature selection. In: Derek H. SLEEMAN and Peter EDWARDS, eds. ML92: Proceedings of the Ninth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 249–256, 1992.

78. KLASSEN, G. Análise genética e funcional dos genes *nifENXorf1orf2*, *nifQmodABCfixXC* de *Herbaspirillum seropedicae*. Universidade Federal do Paraná, Tese de Doutorado, 2000.

79. KOLLER, D.; SAHAMI, M. Toward optimal feature selection. In: Proceedings of the Thirteenth International Conference on Machine Learning. Morgan Kaufmann, pp. 284–292, 1996.

80. KULIKOVA, T.; AKHTAR, R.; ALDEBERT, P.; ALTHORPE, N.; ANDERSSON, M.; BALDWIN, A.; BATES, K.; BHATTACHARYYA, S.; BOWER, L. *et al*. EMBL Nucleotide Sequence Database in 2006. Nucleic Acids Research, Vol. 35(Database issue), p. 16–20. 2007.

81. KYTE, J.; DOOLITTLE, R.F. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105-132, 1982.

82. LENFERS, F. P. ; IGNACIO, F. A. ; KUSTER, C. V. ; GARRET, L. F. V. ; ZOTTO, S. GUIZELINI, D. ; RAITTZ, R. T. EasyFan. Trabalho de Conclusão de Curso. Universidade Federal do Paraná. 2006.

83. LIU, H.;MOTODA, H. Feature Selection for Knowledge Discovery and Data Mining. The Kluwer International Series in Engineering and Computer Science. Kluwer Academic Publishers, 1998.

84. LOBRY, J.R.; GAUTIER, C. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acids Research, 22:3174-3180, 1994.

85. LONG, S.R. Gene and signals in the Rhizobium-legume symbiosis. Plant Physiol, 125:69:72, 2001.

86. LUTZ, M.; ASCHER, D. Aprendendo Python. Tradução TORTELLO, J. 2 Ed. Porto Alegre: Bookman, 568p. 2007.

87. MARCHLER-BAUER, A.; LU, S.; ANDERSON, J.B.; CHITSAZ, F.; DERBYSHIRE, M.K.; DEWEESE-SCOTT, C.; FONG, J.H.; GEER, L.Y.; GEER, R.C.; GONZALES, N.R.; GWADZ, M.; HURWITZ, D.I.; JACKSON, J.D.; KE, Z.; LANCZYCKI, C.J.; LU, F.; MARCHLER, G.H.; MULLOKANDOV, M.; OMELCHENKO, M.V.; ROBERTSON, C.L.; SONG, J.S.; THANKI, N.; YAMASHITA, R.A.; ZHANG, D.; ZHANG, N.; ZHENG, C.; BRYANT, S.H. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. 2011 Jan; 39(Database issue):D225-D229. Epub 2010 Nov 24.

88. MATHWORKS. MATLAB - The Language Of Technical Computing. Disponível em: <http://www.mathworks.com/products/matlab/>. Último acesso em: 11/01/2011.

89. MCCLURE, M.A.; VASI, T.K.; FITCH, W.M. Comparative Analysis of Multiple Protein-Sequence Alignment Methods. Mol. Biol. Evol. 11(4):57 l-592, 1994.

90. MCCULLOCH, W.S.; PITTS, W. A Logical Calculus of Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 5 p 115-133, 1943.

91. MERRICK, M.J.; EDWARDS, R.A. Nitrogen control in bacteria. Microbiol. Rev., 59: 604-622, 1995.

92. MICROSOFT. Recursos e benefícios do Excel. Disponível em: <http://office.microsoft.com/pt-br/excel/recursos-e-beneficios-do-excel-2010-HA101806958.aspx>. Último acesso em: 11/01/2011.

93. MOREIRA, F.M.S.; SIQUEIRA, J.O. Microbiologia e bioquímica do solo. Lavras: Universidade Federal de Lavras, 2006. 729p.

94. NCBI. A science primer - just the facts: a basic introduction to the science underlying NCBI Resources. Disponível em: <http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html>, Último acesso: 12/12/2010.

95. NIEVOLA, J. C. Uso de Redes Neurais Artificiais do Tipo RTRL e MLP FIR para Previsão de Séries Temporais. Tese para professor titular da PUCPR, 1998.

96. OLIVARES, F. L.; BALDANI, V. L. D.; REIS, V. M.; BALDANI, J. I.; DÖBEREINER, J. Occurrence of endophytic diazotrophic Herbaspirillum spp in roots, stems and leaves predominantly of gramineae. Biology and Fertility of Soils, Berlin, v.21, n.3, p.197-200,1996.

97. OMIDVAR, O.; DAYHOFF, J. (Ed.) Neural Networks and Pattern Recognition, USA: Academic Press, 1998.

98. OSAKI, J. H. Caracterizacao Funcional Da Proteina NtrX de Herbaspirillum seropedicae. Universidade Federal do Paraná, Dissertação de Mestrado, 2003.

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

100. PAO, Y. Adaptive Pattern recognition and neural networks. Addison-Wesley, 1989.

101. PAUSTIAN, T.D.; SHAH, V.K.; ROBERTS, G.P. Purification and characterization of nifN and nifE gene products from Azotobacter vinelandii. Proc Natl. Acad. Sci. USA, Washington, Vol.86, p.6082-6086, 1989.

102. PEDROSA, F.O.; BENELLI, E.M.; YATES, M.G.; WASSEM, R.; MONTEIRO, R.A.; KLASSEN, G.; STEFFENS, M.B.R.; SOUZA, E.M.; CHUBATSU, L.S.; RIGO, L.U. Recent developments in the structural organization and regulation of nitrogen fixation genes in Herbaspirillum seropedicae. Journal of Biotechnology 91 (2-3):189-195, 2001.

103. PERRET, X.; STACHELIN, C.; BROUGHTON, W.J. Molecular basis of symbiotic promiscuity. Mol. Biol. Rev., 64: 180-201, 2000.

104. PETTERS, J. W.; SZILAGYI, R. K. Exploring new frontiers of nitrogenase structure and mechanism. Curr. Opin. Chem. Biol., 10: 101-108, 2006.

105. PHILLIPS, D. C. The three-dimensional structure of an enzyme molecule. Sci Am, 215:78-90, 1966.

106. POPTSOVA, M. S.; GOGARTEN, J. P. Using comparative genome analysis to identifyproblems in annotated microbial genomes. Microbiology. 156, 1909–1917, 2010.

107. PORTER, R. R. Structural studies of immunoglobulins. Science, 180:713-716, 1973.

108. POSTGATE, J. R. The fundamentals of nitrogen fixation. Cambridge University Press, p. 201-241, 1982.

109. POSTON, W. L.; MARCHETTE, D. J. Recursive dimensionality reduction using Fisher's linear discriminant. Pattern Recognition, v. 31, Issue: 7, p. 881-888, Julho 1998.

110. PUDIL, P.; NOVOVICOVÁ, J.; KITTLER, J. Floating search methods in feature selection. Pattern Recognition Letters, 15(11), 1119–1125, 1994.

111. RAITZ, R.T.; SOUZA, J.A.; DANDOLINI, G.A.; PACHECO, R.C.S.; MARTINS, A.; GAUTHIER, F.A.; BARCIA, R.M. Learning by means of free associative neurons. Fuzzy Information Processing Society, 1997. NAFIPS '97., 1997 Annual Meeting of the North American , pp.434-439, 21-24, 1997;

112. RAMOS, J. R. L. S. Análises Moleculares Comparativas de Estirpes de *Herbaspirillum* por PFGE, RAPD, RFLP e Sequenciamento do gene que codifica o 16SrRNA. Universidade Federal do Paraná, Tese de Doutorado, 2003.

113. RAYMOND, J.; SIEFERT, J.L.; STAPLES, C.R.; BLANKENSHIP, R.E. The natural history of nitrogen fixation. Mol. Biol. Evol., 21:541-554, 2004.

114. REINHOLD-HUREK, B.; HUREK, T. Reassessment of the taxonomic structure of the diazotrophic genus Azoarcus sensu lato and description of three new genera and new species, Azovibrio restrictus gen. nov., sp. nov., Azospira oryzae gen. nov., sp. nov. and Azonexus fungiphilus gen. nov., sp. nov. Int. J. Syst. Evol. Microbiol., 50, 649-659, 2000.

115. REIS, V. M. Biological Dinitrogen Fixation in Gramineae and Palm Trees. Critical Reviews in Plant Science, v. 19, n.3, p. 227-247, 2000.

116. REZENDE, S. O. Sistemas Inteligentes Fundamentos e Aplicações. São Paulo: Editora Manole Ltda. 2003, 525 p.

117. RICHARDSON, J. S. The anatomy and taxonomy of protein structure. Adv Protein Chem, 34:167-339, 1981.

118. ROBERTS, G.P.; MacNEIL, T.; MacNEIL, D.; BRILL, W.J. Regulation and characterization of protein products coded by the nif (nitrogen fixation) genes of Klebsiella pneumoniae. J. Bacteriol., Washington, Vol.136:1, p. 267-279, 1978.

119. ROBSON, R.L.; WOODLEY, P.R.; PAU, R.N.; EADY, R.R. Structural genes for the vanadium Nitrogenase from Azotobacter chroococcum. EMBO J., Oxford, Vol.8:4, p. 1217-1224, 1989.

120. ROSSWALL, T. The internal nitrogen cycle between microorganisms, vegetation and soil. Ecol. Bull., 22: 157-167, 1976.

121. RUBIO, L. M.; LUDDEN, P. W. Biosyntesis of the iron molybdenum cofator of nitrogenase. Ann. Rev. Microbiol., 62: 93-111, 2008.

122. SAIKIA, S.P.; JAIN, V. Biological nitrogen fixation with non-legumes: An achievable target or a dogma? Curr. Sci., 92: 317-322, 2007.

123. SANTIAGO, V. T.; BRAVO, M.; DAEZ, G., VENTURA, V.; WATANABE, I.; APP, A. Effect on n-fertilizers, straw and dry follow on the nitrogen balance of a flodded soil planted with rice. Plant and Soil, 93:405-11, 1986.

124. SCHALKOFF, R. J. Pattern Recognition: Statistical, Structural and Neural Approaches. John Wiley & Sons, Inc., 1992.

125. SERVANT, F.; BRU, C.; CARRÉRE, S.; COURCELLE, E.; GOUZY, J.; PEYRUC, D.; KAHN, D. ProDom: Automated clustering of homologous domains. Briefings in Bioinformatics. Vol 3, no 3:246-251, 2002.

126. SETÚBAL, J.; MEIDANIS J. Introduction to Computational Molecular Biology, PWS Publishing Company, 1997.

127. SEWELL, M. Feature Selection. 2007. Disponível em <http://machine-learning.martinsewell.com/feature-selection/>. Último acesso: 05/01/2011.

128. SHULTZ, M.; KONDOROSI,A. Regulation of symbiotic root nodule development. Ann. Rev. Genet., 32:33-57, 1998.

129. SIGRIST, C.J.A.; CERUTTI, L.; DE CASTRO, E.; LANGENDIJK-GENEVAUX. P.S.; BULLIARD, V.; BAIROCH, A.; HULO, N. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. 38(Database issue)161-6 , 2010.

130. SIMOUDIS, E. Reality check for data mining. IEEE Expert, 11(5), 26-33, 1996.

131. SOUZA, A.L.F.; INVITTI, A.L.; REGO, F.G.M; MONTEIRO,R.A.; KLASSEN, G.;SOUZA, E. M.; CHUBATSU, L.S.; PEDROSA, F.O.; RIGO, L.U. The involvement of the nif-associated ferredoxin-like genes fdxA and fdxN of Herbaspirillum seropedicae in nitrogen fixation, The Journal of Microbiology, volume 48:1, 77-83, 2010.

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

133. STEVENSON, F. J. Nitrogen: Element and Geochemistry. In : The Encyclopedia of Geochemistry and Environmental Sciences, Fairbridge, R.W. (Ed.). Van Nostrand Reinhold publisher, New York, London, pp: 795-801, 1972.

134. STOESSER, G.; BAKER, W.; VAN DEN BROEK, A.; CAMON, E.; GARCIA-PASTOR, M.; KANZ, C.; KULIKOVA, T.; LEINONEN, R.; LIN, Q.; LOMBARD, V.; LOPEZ, R.; REDASCHI, N.; STOEHR, P.; TULI, M. A.; VAUGHAN, R. The EMBL nucleotide sequence database. Nucleic Acids Research, Vol. 30, p. 21–26. 2002.

135. SAEYS, Y.; INZA, I.; LARRAÑAGA, P. A review of feature selection techniques in bioinformatics. Bioinformatics, Vol. 23 no. 19, pages 2507–2517, 2007.

136. SCHNOES, A.M.; BROWN, S.D.; DODEVSKI, I.; BABBITT, P.C. Annotation Error in Public Databases: Misannotation of Molecular Function in Enzyme Superfamilies. PLoS Comput Biol 5(12).1-13, 2009.

137. SHOEMAKER, B.A.; ZHANG, D.; THANGUDU, RR; TYAGI, M.; FONG, J.H.; MARCHLER-BAUER, A.; BRYANT, S.H.; MADEJ, T.; PANCHENKO, AR. Inferred Biomolecular Interaction Server--a web server to analyze and predict protein interacting partners and binding sites. Nucleic Acids Res.38(D)518-24, 2010.

138. SUGAWARA, H.; ABE, T.; GOJOBORI, T.; TATENO, Y. DDBJ working on evaluation and classification of bacterial genes in INSDC. Nucleic Acids Research, Vol. 35(Database issue), p. 13–15. 2007.

139. SUNDARESAN, V.; AUSUBEL, F.M. Nucleotide sequence of the gene coding for the Nitrogenase iron protein from Klebsiella pneumoniae. J. Biol. Chem., Baltimore, v. 256:6, p. 2808-2812, 1981.

140. SUR, S.; BOTHRA, A.K. ; SEN, A. Symbiotie Nitrogen Fixation – A Bioinformatics Perspective. Biotechnology, 9(3):257-273, 2010.

141. TATENO, Y.; IMANISHI, T.; MIYAZAKI, S.; FUKAMI-KOBAYASHI, K.; SAITOU, N.; SUGAWARA, H. E GOJOBORI, T. DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Research, Vol. 30, p. 27–30. 2002.

142. TRIPLETT, E.W. Prokaryotic nitrogen fixation – a model system for the analysis of a biological process. Norfolk, UK: Horizon Scientific Press, 800pp., 2000.

143. URETA, A; ALVAREZ, B.RÁMON, A.; VERA, M. A.; MARTINÉZ-DRETS, G. Identification of Acetobacter diazotrophicus, Herbaspirillum seropedicae and Herbaspirillum rubrisubalbicans using biochemichal and genetic criteria. Plant and Soil., v. 127, p. 271-277, 1995.

144. WESTON, J. et al. Feature selection for SVMs. In: Todd K. LEEN, Thomas G. DIETTERICH, and Volker TRESP, eds. Advances in Neural Information Processing Systems 13. Cambride, MA: The MIT Press, pp. 668– 674, 2001.

145. WETLAUFER, D. B. Nucleation, rapid folding, and globular intrachain regions in proteins. Proc Natl Acad Sci U S A, 70:697-701, 1973.

146. WHEELER, D. L.; BARRETT, T.; BENSON, D. A.; BRYANT, S. H.; CANESE, K.; CHETVERNIN, V.; CHURCH, D.M.; DICUCCIO, M.; EDGAR, R. *et al*. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, Vol. 36, p. D25 – D30. 2008.

147. WHEELER, D.L.; CHURCH, D. M.; FEDERHEN, S.; LASH, A. E.; MADDEN, T. L.; PONTIUS, J. U.; SCHULER, G. D.; SCHRIML, L. M.; SEQUEIRA, E.; TATUSOVA, T. A.; WAGNER, L. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research, Vol. 31, p. 28–33. 2003.

148. WORLDLINGO. Protein domain. Disponível em: <http://www.worldlingo.com/ma/enwiki/en/Protein_domain>. Último acesso: 12/01/2011.

149. XING, E. P.; JORDAN, M. I.; KARP, R. M. Feature selection for high-dimensional genomic microarray data. In: ICML ’01: Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann, pp. 601–608, 2001.

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

151. YE, Y.; GODZIK, A.. Comparative analysis of protein domain organization. Genome Research, 14(3), 343-53, 2004.

152. YANG, G.P.; DEBELLÉ, F.; SAVAGNAC, A.; FERRO, M.; SCHILTZ, O. et al. Structure of the mesishizobium huakuii and Rhizobium galegae Nod factors: A cluster of phylogenetically related legumes are nodulated by rhizobia producing Nod factors with alpha, beta-unsaturated N-acyl substitutions. Mol. Microbiol., 34: 227-237, 1999.

153. YANG, J.; HONAVAR, V. Feature subset selection using a genetic algorithm. IEEE Intelligent Systems, 13(2), 44–49, 1998.

154. YANG, Y; PEDERSEN, J. O. A comparative study of feature selection in text categorization. In: ICML ’97: Proceedings of the Fourteenth International Conference on Machine Learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 412–420, 1997.

155. YOUNG, J.P.W. Phylogenetic classification of nitrogen-fixing organisms, Em: STACEY, G.; BURRIS, H.; EVANS, H. J. (ed), Biological nitrogen fixation. Chapman & Hall, New York, N.Y., 1992. 943p.

156. ZHENG, L.; WHITE, R.; CASH, V.; JACK, R.; DEAN, D. Assembly of iron-sulfur clusters. J. Biol. Chem. Baltimore, Vol. 273, p. 13264-13272, 1998.

APÊNDICES

APÊNDICE 1 – SCRIPT **step01DoBLAST.py** PARA EXECUÇÃO AUTOMÁTICA DA FERRAMENTA BLASTP.

```
# -*- coding: cp1252 -*-
import os.path
from Bio.Blast import NCBIWWW
from datetime import datetime

lSequences = [## Trecho de código alterado conforme grupo taxonômico, de acordo com a Tabela
7##]

lGenes = [ 'nifH', 'nifD', 'nifK', 'nifE', 'nifN' ]
nCount = 0
today  = datetime.today()
print "Step 01"
print "BLASTP process for all nif genes. Start at:", today
for sequence in lSequences :
    print "Processing BLASTP of :", lGenes[nCount], "gene"
    FileName = lGenes[nCount] + '.txt'
    if os.path.isfile(FileName):
       print "Ok!"
    else:
       result_handle = NCBIWWW.qblast( "blastp", "nr" , sequence, None   ,\
                                       None   , None , None    ,'(none)',\
                                       10.0   , None , None    , None   ,\
                                       20000  , None , None    , None   ,\
                                       None   , None , None    , None   ,\
                                       None   , None , None    , None   ,\
                                       None   , None , None    , None   ,\
                                       None   , None , None    , 20000  ,\
                                       None   , 20000, None    , None   ,\
                                       None   , None , None    , 'Text' ,\
                                       None   , None , None)
       print "Writing results in te file :", FileName
       lines     = result_handle.readlines()
       save_file = open(FileName, "w")
       for line in lines :
           text = line.split('\n')
           save_file.write(text[0])
           save_file.write('\n')
       save_file.close()
    nCount = nCount + 1
today  = datetime.today()
print "Step 01 - BLASTP process for all nif genes. Finish at:", today
```

APÊNDICE 2 – SCRIPT **step02ParsingData.py** PARA EXECUÇÃO DO PROCESSO DE *PARSING* DE DADOS.

```
# -*- coding: cp1252 -*-
import os
from Bio import SeqIO
from datetime import datetime

today = datetime.today()
print "Step 02"
print "Parsing data... Start at:", today
lGenes = [ 'nifH', 'nifD', 'nifK', 'nifE', 'nifN' ]
for gene in lGenes :
    FileBLAST = open(gene + '.txt', "r")
    print "Processing file :", gene + '.txt'
    lines       = FileBLAST.readlines()
    lWrite      = 0
    area        = ''
    ListResults = []
    dic         = {}
    for line in lines:
        if lWrite > 0 :
           if line.startswith('  Database: All non-redundant') > 0:
              break
           if line.startswith('>') :
              lArea = 1
           elif line.startswith('Length=') :
              lArea = 0
              ListResults.append(area.replace('\n', ''))
              area  = ''
           if lArea > 0 :
              area = area + line
        else :
           if line.startswith('ALIGNMENTS') > 0:
              lWrite = 1
    for line in ListResults :
        newline = line.split('|')
        for phrase in newline :
            lName = 0
            name  = ''
            for letter in phrase :
                if letter == '[' :
                   lName = 1
                elif letter == ']' :
                   break
                if lName > 0 :
                   if letter <> '[' :
                      name = name + letter
            if len(name) > 0 :
               dic[name.strip()] = ''
    Results = []
    for item in dic :
        Results.append(item)
    Results.sort()
    FileResult = open( 'List' + gene + '.txt', 'w')
    for item in Results :
        FileResult.write(item)
        FileResult.write('\n')
    FileBLAST.close()
    FileResult.close()
today  = datetime.today()
print "Step 02 - Parsing data... Finish at:", today
```

APÊNDICE 3 – SCRIPT **step03MergeFiles.py** DE TABELAS MESTRE DE DADOS.

```
# -*- coding: cp1252 -*-
import os
import mysql.connector

from Bio      import SeqIO
from datetime import datetime
from Bio      import Entrez

Entrez.email = "A.N.Other@example.com"     # Always tell NCBI who you are

hoje         = datetime.today()
print "Step 03 - Begin at:", hoje
dFechados    = []
db              =  mysql.connector.Connect(host='localhost',  database='genbank',  user='root',
password='')
cursor      = db.cursor()
stmt_select = "SELECT DISTINCT source FROM genbankseg"
cursor.execute(stmt_select)
rows        = cursor.fetchall()
for row in rows:
    dFechados.append(row[0])
dFechados.sort()
db.close()

fileH     = open("ListnifH.txt" , "r")
fileD     = open("ListnifD.txt" , "r")
fileK     = open("ListnifK.txt" , "r")
fileE     = open("ListnifE.txt" , "r")
fileN     = open("ListnifN.txt" , "r")
maybefix  = open("maybefix.txt" , "w")
taxonomy  = open("C:\\Python26\\nifBank\\taxonomy.txt" , "r")
dadosH    = fileH.readlines()
dadosD    = fileD.readlines()
dadosK    = fileK.readlines()
dadosE    = fileE.readlines()
dadosN    = fileN.readlines()
fileH.close()
fileD.close()
fileK.close()
fileE.close()
fileN.close()
merge = {}

for item in dadosH:
    if len(item) > 0:
       data = item.replace('\n','')
       merge[data] = ''
for item in dadosD:
    if len(item) > 0:
       data = item.replace('\n','')
       merge[data] = ''
for item in dadosK:
    if len(item) > 0:
       data = item.replace('\n','')
       merge[data] = ''
for item in dadosE:
    if len(item) > 0:
       data = item.replace('\n','')
       merge[data] = ''
for item in dadosN:
    if len(item) > 0:
       data = item.replace('\n','')
       merge[data] = ''
for item in merge:
    merge[item] = []
dtaxonomy = taxonomy.readlines()
dicTax    = {}
for lines in dtaxonomy:
    data = lines.split('\t')
    aux  = data[1].replace('\n', '')
    dicTax[data[0]] = aux
taxonomy.close()
```

```
dmaybefix = {}
dAux      = {}
dMatriz   = {}
cont      = 0
remain    = len(merge)

for item in merge:
    if item in dFechados :
       merge[item].append('X')
    else :
       merge[item].append(' ')
    taxon   = ''
    if len(item) > 0 :
       if item in dicTax:
          merge[item].append(dicTax[item])
       else:
          handle  = Entrez.esearch(db="Taxonomy", term=item)
          record  = Entrez.read(handle)
          if len(record["IdList"]) > 0 :
             if record["IdList"][0] > 0 :
                handle  = Entrez.efetch(db="Taxonomy",\
                                        id=record["IdList"][0],\
                                        retmode="xml")
                records = Entrez.read(handle)
                taxon   = records[0]["Lineage"]
          merge[item].append(taxon)
          dicTax[item] = taxon
    remain = remain - 1
    if (remain%50) == 0 :
       print "Remain :", remain

for key in merge :
    maybefix.write(key)
    maybefix.write('\t')
    maybefix.write(merge[key][0])
    maybefix.write('\t')
    maybefix.write(merge[key][1])
    maybefix.write('\n')

maybefix.close()

taxonomy  = open("C:\\Python26\\nifBank\\taxonomy.txt" , "w")

for key in dicTax :
    taxonomy.write(key)
    taxonomy.write('\t')
    taxonomy.write(dicTax[key])
    taxonomy.write('\n')
taxonomy.close()

hoje  = datetime.today()
print "Step 03 - Finish at:", hoje
```

APÊNDICE 4 – SCRIPT **step04FindIDs.py** PARA GERAÇÃO DAS TABELAS DE ÍNDICES.

```
# -*- coding: cp1252 -*-
import os
import mysql.connector
from Bio import SeqIO
from datetime import datetime
from Bio import Entrez

today  = datetime.today()
print "Step 04"
print "Begin at:", today

fMaybeFix = open("maybefix.txt" , "r")
dMaybeFix = fMaybeFix.readlines()

print "Preparing BLAST results..."

## Bloco X
print "Processing file : nifH.txt"
nifH   = open("nifH.txt", "r")
lines  = nifH.readlines()
lWrite = 0
area   = ''
ListH  = []
for line in lines:
    if lWrite > 0 :
       if line.startswith('  Database: All non-redundant') > 0:
          break
       if line.startswith('>') :
          lArea = 1
       elif line.startswith('Length=') :
          lArea = 0
          ListH.append(area.replace('\n', ''))
          area  = ''
       if lArea > 0 :
          area = area + line
    else :
       if line.startswith('ALIGNMENTS') > 0:
          lWrite = 1
nifH.close()
## Final do Bloco X

## Repete o Bloco X para cada um dos demais genes nif em estudo: DKEN.

print "Looking for IDs ..."

count  = 0
OrgsID = {}
SnifH  = {}
SnifD  = {}
SnifK  = {}
SnifE  = {}
SnifN  = {}
for item in dMaybeFix :
    if count >= 2 :
       data = item.split('\t')
       if data[1] <> '@' :
           organism = data[0]
           print organism

           ## Bloco Y
           nifH  = {}
           for line in ListH :
               if "[" + organism + "]" in line :
                  lineInv = line[::-1]
                  orgInv  = organism[::-1]
                  start   = lineInv.find("]" + orgInv + "[")
                  string  = lineInv[((start + len("]" + \
                                      orgInv + "["))+1): len(lineInv)]
                  while True :
                     sinInv  = ''
```

```
                    codInv  = ''
                    lEndSin = 0
                    cPipe   = 0
                    cLetter = 0
                    for letter in string :
                        if letter == "|" :
                           cPipe = cPipe + 1
                           lEndSin = 1
                        if lEndSin <= 0 :
                           sinInv = sinInv + letter
                        else :
                           if cLetter <= 3 :
                              if letter <> '>' :
                                 codInv = codInv + letter
                           else :
                              break
                           if cPipe == 2 :
                              cLetter = cLetter + 1
                    nifH[codInv[::-1].strip()]  = sinInv[::-1].strip()
                    SinAux = sinInv[::-1].strip()
                    SnifH[SinAux.upper()] = SinAux
                    if string.find("]" + orgInv + "[") > 0 :
                       start   = string.find("]" + orgInv + "[")
                       sAux    = string[((start + len("]" +\
                                         orgInv + "["))+1): len(string)]
                       string  = sAux
                    else :
                        break
          ## Final do Bloco Y

          ## Repete o Bloco y para os demais genes nif: DKEN

          OrgsID[organism] = [data[1], nifD, nifE, nifH, nifK, nifN]
    count = count + 1
fMaybeFix.close()
fileID = open("IDsFind.txt" , "w")
lGenes = [ "nifD", "nifE", "nifH", "nifK", "nifN" ]
print "Writing IDs..."

for orgs in OrgsID :
    count = 1
    while count < 6 :
       if len(OrgsID[orgs][count]) > 0 :
          for item in OrgsID[orgs][count] :
              fileID.write( orgs )
              fileID.write('\t')
              fileID.write(OrgsID[orgs][0])
              fileID.write('\t')
              fileID.write(lGenes[(count-1)])
              fileID.write('\t')
              fileID.write(item)
              fileID.write('\t')
              fileID.write(OrgsID[orgs][count][item])
              fileID.write('\n')
       else :
              fileID.write( orgs )
              fileID.write('\t')
              fileID.write(OrgsID[orgs][0])
              fileID.write('\t')
              fileID.write(lGenes[(count-1)])
              fileID.write('\t')
              fileID.write('')
              fileID.write('\t')
              fileID.write('')
              fileID.write('\n')
       count = count + 1
fileID.close()
today  = datetime.today()
print "Step 04 - Parsing data... Finish at:", today
```

APÊNDICE 5 – CÓDIGO FONTE REFERENTE AO SCRIPT **step05AddValuesIDs.py** PARA INCORPORAÇÃO DE INFORMAÇÕES ÀS TABELAS MESTRE.

```
# -*- coding: cp1252 -*-
import os
import mysql.connector
from Bio import SeqIO
from datetime import datetime
from Bio import Entrez

today  = datetime.today()
print "Step 05"
print "Begin at:", today

fIDs = open("IDsFind.txt", "r")
dIDs = fIDs.readlines()

print "Preparing BLAST results..."

files = [ [ "nifD.txt", "nifD" ],\
          [ "nifE.txt", "nifE" ],\
          [ "nifH.txt", "nifH" ],\
          [ "nifK.txt", "nifK" ],\
          [ "nifN.txt", "nifN" ] ]

bigList = []
dicLen  = {}
for File in files:
    print "Processing file: ", File[0]
    file   = open(File[0], 'r')
    lines  = file.readlines()
    lWrite = 0
    area   = ''
    lArea  = 0
    Len    = 0
    for line in lines:
        if lWrite > 0 :
           if line.startswith('  Database: All non-redundant') > 0:
              space  = area.split('\n')
              Query  = ''
              Sbjct  = ''
              header = ''
              Field  = ''
              for ln in space:
                  if ln.startswith(' Score =') :
                     break
                  header = header + ln
              lBegin = 0
              lField = 0
              for ln in space:
                  if lBegin > 0 :
                     if ln.startswith('Query') :
                        lField = 0
                        Query  = Query + ln
                     if ln.startswith('Sbjct') :
                        Sbjct = Sbjct + ln
                     if ln.startswith(' Score ='):
                        lBegin = 0
                        fLength     = header.split("Length=")
                        fIdentities = Field.split("Identities =")
                        fPositives  = Field.split("Positives =")
                        fScore      = Field.split("Score =")
                        fExpect     = Field.split("Expect =")
                        fGaps       = Field.split("Gaps =")
                        Length      = ''
                        Identities  = ''
                        Positives   = ''
                        Gaps        = ''
                        Score       = ''
                        Expect      = ''
```

```
                    for letter in fLength[1] :
                        if letter == 'S' :
                           break
                        else :
                           Length = Length + letter
                    for letter in fScore[1] :
                        if letter == ',' :
                           break
                        else :
                           Score = Score + letter
                    for letter in fExpect[1] :
                        if letter == ',' :
                           break
                        else :
                           Expect = Expect + letter
                    for letter in fIdentities[1] :
                        if letter == '(' :
                           break
                        else :
                           Identities = Identities + letter
                    for letter in fPositives[1]:
                        if letter == '(' :
                           break
                        else :
                           Positives = Positives + letter
                    for letter in fGaps[1]:
                        if letter == '(' :
                           break
                        else :
                           Gaps = Gaps + letter
                    bigList.append([File[1], header, Length.strip(), Score.strip(),\
                                    Expect.strip(), Identities.strip(), Positives.strip(),\
                                    Gaps.strip(), Query, Sbjct])
                    Query  = ''
                    Sbjct  = ''
                if ln.startswith(' Identities ='):
                   lBegin = 1
                if ln.startswith(' Score =') :
                   lField = 1
                   Field  = ''
                if lField > 0 :
                   Field = Field + ln
            fLength     = header.split("Length=")
            fIdentities = Field.split("Identities =")
            fPositives  = Field.split("Positives =")
            fScore      = Field.split("Score =")
            fExpect     = Field.split("Expect =")
            fGaps       = Field.split("Gaps =")
            Length      = ''
            Identities  = ''
            Positives   = ''
            Gaps        = ''
            Score       = ''
            Expect      = ''
            for letter in fLength[1] :
                if letter == 'S' :
                   break
                else :
                   Length = Length + letter
            for letter in fScore[1] :
                if letter == ',' :
                   break
                else :
                   Score = Score + letter
            for letter in fExpect[1] :
                if letter == ',' :
                   break
                else :
                   Expect = Expect + letter
            for letter in fIdentities[1] :
                if letter == '(' :
                   break
                else :
                   Identities = Identities + letter
            for letter in fPositives[1]:
                if letter == '(' :
```

```
                    break
                else :
                    Positives = Positives + letter
            for letter in fGaps[1]:
                if letter == '(' :
                    break
                else :
                    Gaps = Gaps + letter
            bigList.append([File[1], header, Length.strip(), Score.strip(), Expect.strip(),
Identities.strip(), Positives.strip(), Gaps.strip(), Query, Sbjct])
            break
         if lArea > 0 :
            if line.startswith('>'):
               space  = area.split('\n')
               Query  = ''
               Sbjct  = ''
               header = ''
               Field  = ''
               for ln in space:
                   if ln.startswith(' Score =') :
                      break
                   header = header + ln
               lBegin = 0
               lField = 0
               for ln in space:
                   if lBegin > 0 :
                      if ln.startswith('Query') :
                         score = Field
                         lField = 0
                         Query  = Query + ln
                      if ln.startswith('Sbjct') :
                         Sbjct = Sbjct + ln
                      if ln.startswith(' Score ='):
                         lBegin = 0
                         fLength     = header.split("Length=")
                         fIdentities = Field.split("Identities =")
                         fPositives  = Field.split("Positives =")
                         fScore      = Field.split("Score =")
                         fExpect     = Field.split("Expect =")
                         fGaps       = Field.split("Gaps =")
                         Length      = ''
                         Identities  = ''
                         Positives   = ''
                         Gaps        = ''
                         Score       = ''
                         Expect      = ''
                         for letter in fLength[1] :
                             if letter == 'S' :
                                break
                             else :
                                Length = Length + letter
                         for letter in fScore[1] :
                             if letter == ',' :
                                break
                             else :
                                Score = Score + letter
                         for letter in fExpect[1] :
                             if letter == ',' :
                                break
                             else :
                                Expect = Expect + letter
                         for letter in fIdentities[1] :
                             if letter == '(' :
                                break
                             else :
                                Identities = Identities + letter
                         for letter in fPositives[1]:
                             if letter == '(' :
                                break
                             else :
                                Positives = Positives + letter
                         for letter in fGaps[1]:
                             if letter == '(' :
                                break
                             else :
                                Gaps = Gaps + letter
```

```
                            bigList.append([File[1], header, Length.strip(), Score.strip(),\
                                        Expect.strip(), Identities.strip(), Positives.strip(),\
                                        Gaps.strip(), Query, Sbjct])
                            Query = ''
                            Sbjct = ''
                        if ln.startswith(' Score ='):
                            lField = 1
                            Field  = ''
                        if lField > 0 :
                            Field = Field + ln
                        if ln.startswith(' Identities ='):
                            lBegin = 1
                    fLength     = header.split("Length=")
                    fIdentities = Field.split("Identities =")
                    fPositives  = Field.split("Positives =")
                    fScore      = Field.split("Score =")
                    fExpect     = Field.split("Expect =")
                    fGaps       = Field.split("Gaps =")
                    Length      = ''
                    Identities  = ''
                    Positives   = ''
                    Gaps        = ''
                    Score       = ''
                    Expect      = ''
                    for letter in fLength[1] :
                        if letter == 'S' :
                            break
                        else :
                            Length = Length + letter
                    for letter in fScore[1] :
                        if letter == ',' :
                            break
                        else :
                            Score = Score + letter
                    for letter in fExpect[1] :
                        if letter == ',' :
                            break
                        else :
                            Expect = Expect + letter
                    for letter in fIdentities[1] :
                        if letter == '(' :
                            break
                        else :
                            Identities = Identities + letter
                    for letter in fPositives[1]:
                        if letter == '(' :
                            break
                        else :
                            Positives = Positives + letter
                    for letter in fGaps[1]:
                        if letter == '(' :
                            break
                        else :
                            Gaps = Gaps + letter
                    bigList.append([File[1], header, Length.strip(), Score.strip(),\
                                    Expect.strip(), Identities.strip(), Positives.strip(),\
                                    Gaps.strip(), Query, Sbjct])
                else :
                    area = area + line
                if line.startswith('>') :
                    lArea = 1
                    area  = ''
                    area  = area + line
            else :
                if line.startswith('ALIGNMENTS') > 0:
                    lWrite = 1
                elif line.startswith('Length=') > 0 and Len <= 0:
                    length = line.split('Length=')
                    Len    = length[1].replace('\n','')
                    dicLen[File[1]] = Len
        file.close()

print "Looking for IDs..."

nCount = 0
dic    = {}
```

```
for item1 in dIDs :
    lAux     = item1.replace('\n', '')
    fAux     = lAux.split('\t')
    data     = fAux[0:len(fAux)]
    organism = data[0]
    if len(data[3]) > 0 :
       for item2 in bigList :
           if ((data[2] == item2[0]) and\
               (data[0] in item2[1]) and\
               (data[3] in item2[1]) and\
               (data[4] in item2[1])) :
              querySpace = item2[8].split("Query")
              sbjctSpace = item2[9].split("Sbjct")
              iniQ  = ''
              endQ  = ''
              query = ''
              lIni  = 1
              for line in querySpace:
                  if len(line) > 0:
                     for letter in line :
                         if letter in "0123456789":
                            if lIni > 0 :
                               iniQ = iniQ + letter
                            endQ = endQ + letter
                         elif letter <> ' ':
                            if lIni > 0 :
                               lIni = 0
                            query = query + letter
                            endQ = ''
              iniS  = ''
              endS  = ''
              sbjct = ''
              lIni  = 1
              for line in sbjctSpace:
                  if len(line) > 0:
                     for letter in line :
                         if letter in "0123456789":
                            if lIni > 0 :
                               iniS = iniS + letter
                            endS = endS + letter
                         elif letter <> ' ':
                            if lIni > 0 :
                               lIni = 0
                            sbjct = sbjct + letter
                            endS = ''
              dic[nCount] = [data[0], data[1], data[2], data[3], dicLen[data[2]], item2[2],\
                             item2[3], item2[4], item2[5], item2[6], item2[7], data[4],\
                             query, iniQ, endQ, sbjct, iniS, endS]
              nCount = nCount + 1
    else :
       dic[nCount] = [data[0], data[1], data[2], data[3], dicLen[data[2]], '', '', '', '',\
                      '', '', data[4], '', '', '', '', '', '']
       nCount = nCount + 1
fIDs.close()

file = open("IDsValuesBLAST.txt", "w" )
for item in dic:
    file.write(dic[item][0])
    file.write('\t')
    file.write(dic[item][1])
    file.write('\t')
    file.write(dic[item][2])
    file.write('\t')
    file.write(dic[item][3])
    file.write('\t')
    file.write(dic[item][4])
    file.write('\t')
    file.write(dic[item][5])
    file.write('\t')
    file.write(dic[item][8])
    file.write('\t')
    op1  = ''
    op2  = ''
    perc = ''
    lDiv = 0
```

```
    for letter in dic[item][8] :
        if letter == '/' :
           lDiv = 1
        if lDiv > 0 :
           if letter <> '/' :
              op2 = op2 + letter
        else :
           op1 = op1 + letter
    if len(op1) > 0 :
       perc = str((int(op1)*100)/int(op2))
    file.write(perc)
    file.write('\t')
    file.write(dic[item][9])
    file.write('\t')
    op1  = ''
    op2  = ''
    perc = ''
    lDiv = 0
    for letter in dic[item][9] :
        if letter == '/' :
           lDiv = 1
        if lDiv > 0 :
           if letter <> '/' :
              op2 = op2 + letter
        else :
           op1 = op1 + letter
    if len(op1) > 0 :
       perc = str((int(op1)*100)/int(op2))
    file.write(perc)
    file.write('\t')
    file.write(dic[item][10])
    file.write('\t')
    op1  = ''
    op2  = ''
    perc = ''
    lDiv = 0
    for letter in dic[item][10] :
        if letter == '/' :
           lDiv = 1
        if lDiv > 0 :
           if letter <> '/' :
              op2 = op2 + letter
        else :
           op1 = op1 + letter
    if len(op1) > 0 :
       perc = str((int(op1)*100)/int(op2))
    file.write(perc)
    file.write('\t')
    file.write(dic[item][6])
    file.write('\t')
    file.write(dic[item][7])
    file.write('\t')
    file.write(dic[item][11])
    file.write('\t')
    file.write(dic[item][12])
    file.write('\t')
    file.write(dic[item][13])
    file.write('\t')
    file.write(dic[item][14])
    file.write('\t')
    file.write(dic[item][15])
    file.write('\t')
    file.write(dic[item][16])
    file.write('\t')
    file.write(dic[item][17])
    file.write('\n')
file.close()
today  = datetime.today()
print "Step 05 - Parsing data... Finish at:", today
```

APÊNDICE 6 – CÓDIGO FONTE REFERENTE AO SCRIPT **step06RescueData.py** PARA RECUPERAÇÃO DE DADOS DISPERSOS NOS DIRETÓRIOS REFERENTES AOS GRUPOS TAXONÔMICOS ESTUDADOS.

```
# -*- coding: cp1252 -*-
import os
from Bio import SeqIO
from datetime import datetime
from Bio import Entrez
from operator import itemgetter

today  = datetime.today()
print "Step 6"
print "Begin at:", today

files=[['C:\\Python26\\..\\BDActinobacteria\\IDsValuesBLAST.txt', \
        'C:\\Python26\\..\\BDActinobacteria\\maybefix.txt',\
        'BDActino.txt'],\
       ['C:\\Python26\\..\\BDArchaea\\IDsValuesBLAST.txt',\
        'C:\\Python26\\..\\BDArchaea\\maybefix.txt',\
        'BDArchaea.txt'],\
       ['C:\\Python26\\..\\BDChlorobi\\IDsValuesBLAST.txt',\
        'C:\\Python26\\..\\BDChlorobi\\maybefix.txt',\
        'BDChlorobi.txt'],\
       ['C:\\Python26\\..\\BDCyanobacteria\\IDsValuesBLAST.txt',\
        'C:\\Python26\\..\\BDCyanobacteria\\maybefix.txt',\
        'BDCyano.txt'],\
       ['C:\\Python26\\..\\BDFirmicutes\\IDsValuesBLAST.txt',\
        'C:\\Python26\\..\\BDFirmicutes\\maybefix.txt',\
        'BDFirmicutes.txt'],\
       ['C:\\Python26\\..\\BDPAlpha\\IDsValuesBLAST.txt',\
        'C:\\Python26\\..\\BDPAlpha\\maybefix.txt',\
        'BDPAlpha.txt'],\
       ['C:\\Python26\\..\\BDPBeta\\IDsValuesBLAST.txt',\
        'C:\\Python26\\..\\BDPBeta\\maybefix.txt',\
        'BDPBeta.txt'],\
       ['C:\\Python26\\..\\BDPDeltaEpsilon\\IDsValuesBLAST.txt',\
        'C:\\Python26\\..\\BDPDeltaEpsilon\\maybefix.txt',\
        'BDPDelta.txt'],\
       ['C:\\Python26\\..\\BDPGama\\IDsValuesBLAST.txt',\
        'C:\\Python26\\..\\BDPGama\\maybefix.txt',\
        'BDPGama.txt'],\
       ['C:\\Python26\\..\\BDRandom\\IDsValuesBLAST.txt',\
        'C:\\Python26\\..\\BDRandom\\maybefix.txt',\
        'BDRandom.txt']]
ftaxon = open('C:\\Python26\\Novos Testes\\taxonomy.txt', "r")
ltaxon = ftaxon.readlines()
dTaxon = {}
for lines in ltaxon:
    data = lines.split('\t')
    aux  = data[1].replace('\n', '')
    dTaxon[data[0]] = aux
ftaxon.close()
for f in files:
    fIDs  = open(f[0], "r")
    dIDs  = fIDs.readlines()
    lIDs  = []
    count = 0
    for line in dIDs :
        if count > 0 :
           lAux   = line.replace('\n', '')
           fAux   = lAux.split('\t')
           fields = fAux[0:len(fAux)]
           fields.append('')
           ftuple = tuple(fields)
           lIDs.append(ftuple)
        count = 1
    fIDs.close()
    tSorted = sorted(lIDs,\
        key = itemgetter(0,2,5,6,3,13,7,8,9,10,11,12,15,16,17,18,19,20))
    tAux    = sorted(lIDs,\
        key = itemgetter(0,2,5,6,3,13,7,8,9,10,11,12,15,16,17,18,19,20))
```

```
    lSorted = []
    for line in tSorted :
        lSorted.append(list(line))
    print "Processing data..."
    print "Looking for duplicated lines in :", f[0]
    print "Len :", len(lSorted)
    f00    = lSorted[0][0]
    f01    = lSorted[0][1]
    f02    = lSorted[0][2]
    f03    = lSorted[0][3]
    f04    = lSorted[0][4]
    f05    = lSorted[0][5]
    f06    = lSorted[0][6]
    f07    = lSorted[0][7]
    f08    = lSorted[0][8]
    f09    = lSorted[0][9]
    f10    = lSorted[0][10]
    f11    = lSorted[0][11]
    f12    = lSorted[0][12]
    f13    = lSorted[0][13]
    f14    = lSorted[0][14]
    f15    = lSorted[0][15]
    f16    = lSorted[0][16]
    f17    = lSorted[0][17]
    f18    = lSorted[0][18]
    f19    = lSorted[0][19]
    f20    = lSorted[0][20]
    dSame  = {}
    lFirst = 0
    for line1 in lSorted:
        if lFirst > 0 :
           if len(line1[3]) > 0 :
              if ((line1[0]  == f00) and\
                  (line1[2]  == f02) and\
                  (line1[5]  == f05) and\
                  (line1[6]  == f06) and\
                  (line1[7]  == f07) and\
                  (line1[8]  == f08) and\
                  (line1[9]  == f09) and\
                  (line1[10] == f10) and\
                  (line1[11] == f11) and\
                  (line1[12] == f12) and\
                  (line1[15] == f15) and\
                  (line1[16] == f16) and\
                  (line1[17] == f17) and\
                  (line1[18] == f18) and\
                  (line1[19] == f19) and\
                  (line1[20] == f20) ):
                 line1[21] = 'X'
              f00 = line1[0]
              f01 = line1[1]
              f02 = line1[2]
              f03 = line1[3]
              f04 = line1[4]
              f05 = line1[5]
              f06 = line1[6]
              f07 = line1[7]
              f08 = line1[8]
              f09 = line1[9]
              f10 = line1[10]
              f11 = line1[11]
              f12 = line1[12]
              f13 = line1[13]
              f14 = line1[14]
              f15 = line1[15]
              f16 = line1[16]
              f17 = line1[17]
              f18 = line1[18]
              f19 = line1[19]
              f20 = line1[20]
        else:
           lFirst = 1

    print "Writing data..."
    file   = open(f[2], "w")
    nCount = 0
```

```
    for line in lSorted:
        if len(line[3]) > 0 :
            if line[21] <> 'X' :
               if ((line[0].startswith('uncultured') or\
                    line[0].startswith('unidentified')) and line[1] <> 'X') :
                  a = 1
               else :
                  file.write(line[0])
                  file.write('\t')
                  file.write(line[1])
                  file.write('\t')
                  file.write(line[2])
                  file.write('\t')
                  file.write(line[3])
                  file.write('\t')
                  file.write(line[4])
                  file.write('\t')
                  file.write(line[5])
                  file.write('\t')
                  file.write(line[6])
                  file.write('\t')
                  file.write(line[7])
                  file.write('\t')
                  file.write(line[8])
                  file.write('\t')
                  file.write(line[9])
                  file.write('\t')
                  file.write(line[10])
                  file.write('\t')
                  file.write(line[11])
                  file.write('\t')
                  file.write(line[12])
                  file.write('\t')
                  file.write(line[13])
                  file.write('\t')
                  file.write(line[14])
                  file.write('\t')
                  file.write(line[15])
                  file.write('\t')
                  file.write(line[16])
                  file.write('\t')
                  file.write(line[17])
                  file.write('\t')
                  file.write(line[18])
                  file.write('\t')
                  file.write(line[19])
                  file.write('\t')
                  file.write(line[20])
                  file.write('\t')
                  file.write(dTaxon[line[0]])
                  file.write('\n')
                  nCount = nCount + 1
    file.close()
    ftaxon.close()
    print "Final Len :", nCount
today  = datetime.today()
print "Step 6 - End at:", today
```

APÊNDICE 7 – CÓDIGO FONTE REFERENTE AO SCRIPT **step07NewDataUnion.py** PARA UNIÃO DOS DADOS DISPERSOS NOS DIRETÓRIOS REFERENTES AOS GRUPOS TAXONÔMICOS ESTUDADOS.

```
import os
import mysql.connector
from Bio import SeqIO
from datetime import datetime
from Bio import Entrez

Entrez.email = "A.N.Other@example.com"

today  = datetime.today()
print "Step 7 - Begin at:", today

files = [ [ "BDActino.txt"    ],\
          [ "BDArchaea.txt"   ],\
          [ "BDChlorobi.txt"  ],\
          [ "BDCyano.txt"     ],\
          [ "BDFirmicutes.txt"],\
          [ "BDPAlpha.txt"    ],\
          [ "BDPBeta.txt"     ],\
          [ "BDPDelta.txt"    ],\
          [ "BDPGama.txt"     ],\
          [ "BDRandom.txt"    ] ]

ListGer = []
for Files in files :
    File  = open(Files[0], "r")
    dFile = File.readlines()
    valid = 0
    blank = 0
    for line in dFile :
        lAux   = line.replace('\n', '')
        fAux   = lAux.split('\t')
        fields = fAux[0:len(fAux)]
        ListGer.append(fields)
    File.close()
    print "File :", Files[0],"\t", "Len:", len(dFile)

dic = {}
for line in ListGer:
    chave  = ''
    nCount = 0
    for item in line:
        if nCount <= 21 :
           chave  = chave + item + '\t'
           nCount = nCount + 1
    dic[chave] = ''

print "Total Len:", len(dic)

file = open("NewData.txt", "w")
for item in dic:
    fields = item.split('\t')
    file.write(fields[0])
    file.write('\t')
    file.write(fields[1])
    file.write('\t')
    file.write(fields[2])
    file.write('\t')
    file.write(fields[3])
    file.write('\t')
    file.write(fields[4])
    file.write('\t')
    file.write(fields[5])
    file.write('\t')
    file.write(fields[6])
    file.write('\t')
    file.write(fields[7])
    file.write('\t')
```

```
    file.write(fields[8])
    file.write('\t')
    file.write(fields[9])
    file.write('\t')
    file.write(fields[10])
    file.write('\t')
    file.write(fields[11])
    file.write('\t')
    file.write(fields[12])
    file.write('\t')
    file.write(fields[13])
    file.write('\t')
    file.write(fields[14])
    file.write('\t')
    file.write(fields[15])
    file.write('\t')
    file.write(fields[16])
    file.write('\t')
    file.write(fields[17])
    file.write('\t')
    file.write(fields[18])
    file.write('\t')
    file.write(fields[19])
    file.write('\t')
    file.write(fields[20])
    file.write('\t')
    file.write(fields[21])
    file.write('\n')
file.close()

today  = datetime.today()
print "Step 7 - End at:", today
```

APÊNDICE 8 – CÓDIGO FONTE REFERENTE AO SCRIPT **step08CuringData.py** PARA LIMPEZA DOS DADOS DISPERSOS NOS DIRETÓRIOS REFERENTES AOS GRUPOS TAXONÔMICOS ESTUDADOS.

```
# -*- coding: cp1252 -*-
import os
from Bio import SeqIO
from datetime import datetime
from Bio import Entrez
from operator import itemgetter
today  = datetime.today()
print "Step 8 - Begin at:", today
print "Reading data..."
fIDs  = open("NewData.txt", "r")
dIDs  = fIDs.readlines()
lIDs  = []
for line in dIDs :
    lAux   = line.replace('\n', '')
    fAux   = lAux.split('\t')
    fields = fAux[0:len(fAux)]
    fields.append('')
    ftuple = tuple(fields)
    lIDs.append(ftuple)
fIDs.close()
print "Sorting data..."
tSorted = sorted(lIDs, key = itemgetter(0,2,5,6,3,13,7,8,9,10,11,12,15,16,17,18,19,20))
tAux    = sorted(lIDs, key = itemgetter(0,2,5,6,3,13,7,8,9,10,11,12,15,16,17,18,19,20))
lSorted = []
for line in tSorted :
    lSorted.append(list(line))
print "Processing data..."
print "Looking for duplicated lines..."
print "Len :", len(lSorted)
f00    = lSorted[0][0]
f02    = lSorted[0][2]
f03    = lSorted[0][3]
f05    = lSorted[0][5]
f06    = lSorted[0][6]
f07    = lSorted[0][7]
f08    = lSorted[0][8]
f09    = lSorted[0][9]
f10    = lSorted[0][10]
f11    = lSorted[0][11]
f12    = lSorted[0][12]
f13    = lSorted[0][13]
f15    = lSorted[0][15]
f16    = lSorted[0][16]
f17    = lSorted[0][17]
f18    = lSorted[0][18]
f19    = lSorted[0][19]
f20    = lSorted[0][20]
lFirst = 0
for line1 in lSorted:
    if lFirst > 0 :
       if ((line1[0]  == f00) and\
           (line1[2]  == f02) and\
           (line1[3]  == f03) and\
           (line1[5]  == f05) and\
           (line1[6]  == f06) and\
           (line1[7]  == f07) and\
           (line1[8]  == f08) and\
           (line1[9]  == f09) and\
           (line1[10] == f10) and\
           (line1[11] == f11) and\
           (line1[12] == f12) and\
           (line1[15] == f15) and\
           (line1[16] == f16) and\
           (line1[17] == f17) and\
           (line1[18] == f18) and\
           (line1[19] == f19) and\
           (line1[20] == f20)):
```

```
          line1[22] = 'X'
       f00    = line1[0]
       f02    = line1[2]
       f03    = line1[3]
       f05    = line1[5]
       f06    = line1[6]
       f07    = line1[7]
       f08    = line1[8]
       f09    = line1[9]
       f10    = line1[10]
       f11    = line1[11]
       f12    = line1[12]
       f13    = line1[13]
       f15    = line1[15]
       f16    = line1[16]
       f17    = line1[17]
       f18    = line1[18]
       f19    = line1[19]
       f20    = line1[20]
    else:
      lFirst = 1
print "Writing data..."
lReSorted = []
for line in lSorted :
    lReSorted.append(tuple(line))
tSorted = sorted(lReSorted, key = itemgetter(1,3,6,7,4,14,8,9,10,11,12,13,16,17,18,19,20,21))
file    = open("DataOK.txt", "w")
for item in tSorted:
    if item[22] <> "X":
        file.write(item[0])
        file.write("\t")
        file.write(item[1])
        file.write("\t")
        file.write(item[2])
        file.write("\t")
        file.write(item[3])
        file.write("\t")
        file.write(item[4])
        file.write("\t")
        file.write(item[5])
        file.write("\t")
        file.write(item[6])
        file.write("\t")
        file.write(item[7])
        file.write("\t")
        file.write(item[8])
        file.write("\t")
        file.write(item[9])
        file.write("\t")
        file.write(item[10])
        file.write("\t")
        file.write(item[11])
        file.write("\t")
        file.write(item[12])
        file.write("\t")
        file.write(item[13])
        file.write("\t")
        file.write(item[14])
        file.write("\t")
        file.write(item[15])
        file.write("\t")
        file.write(item[16])
        file.write("\t")
        file.write(item[17])
        file.write("\t")
        file.write(item[18])
        file.write("\t")
        file.write(item[19])
        file.write("\t")
        file.write(item[20])
        file.write("\t")
        file.write(item[21])
        file.write("\n")
file.close()
file2.close()
print "Step 8 - End at:", today
```

APÊNDICE 9 – MAPA RELACIONAL DE ORGANISMOS X GENES *nif* ENCONTRADOS. DADOS REFERENTES A GENOMAS COMPLETOS.

Taxonomia			Organismo	Genes *nif*					YOUNG, 1992
				H	D	K	E	N	
Archaea	Euryarchaeota	environmental samples	uncultured methanogenic archaeon RC-I	1	1	1	1	1	
Archaea	Euryarchaeota	Halobacteria	Natronomonas pharaonis DSM 2160	1	0	0	0	0	
Archaea	Euryarchaeota	Methanobacteria	Methanothermobacter marburgensis str. Marburg	1	0	1	1	1	
Archaea	Euryarchaeota	Methanobacteria	Methanothermobacter thermautotrophicus str. Delta H	1	0	1	1	0	X
Archaea	Euryarchaeota	Methanobacteria	Methanobrevibacter ruminantium M1	1	0	0	0	0	
Archaea	Euryarchaeota	Methanobacteria	Methanobrevibacter smithii ATCC 35061	1	0	0	0	0	
Archaea	Euryarchaeota	Methanobacteria	Methanosphaera stadtmanae DSM 3091	1	0	0	0	0	
Archaea	Euryarchaeota	Methanobacteria	Methanothermus fervidus DSM 2088	1	0	0	0	0	X
Archaea	Euryarchaeota	Methanococci	Methanocaldococcus infernus ME	1	1	1	1	0	
Archaea	Euryarchaeota	Methanococci	Methanocaldococcus sp. FS406-22	1	1	1	1	0	
Archaea	Euryarchaeota	Methanococci	Methanocaldococcus vulcanius M7	1	1	1	1	0	
Archaea	Euryarchaeota	Methanococci	Methanococcus maripaludis C5	1	1	1	1	0	X
Archaea	Euryarchaeota	Methanococci	Methanococcus maripaludis C6	1	1	1	1	0	X
Archaea	Euryarchaeota	Methanococci	Methanococcus maripaludis C7	1	1	1	1	0	X
Archaea	Euryarchaeota	Methanococci	Methanococcus maripaludis S2	1	1	1	1	0	X
Archaea	Euryarchaeota	Methanococci	Methanococcus vannielii SB	1	1	1	1	0	X
Archaea	Euryarchaeota	Methanococci	Methanococcus aeolicus Nankai-3	1	1	1	0	0	X
Archaea	Euryarchaeota	Methanococci	Methanocaldococcus fervens AG86	1	0	0	0	0	
Archaea	Euryarchaeota	Methanococci	Methanocaldococcus jannaschii DSM 2661	1	0	0	0	0	
Archaea	Euryarchaeota	Methanococci	Methanococcus voltae A3	1	0	0	0	0	X
Archaea	Euryarchaeota	Methanomicrobia	Candidatus Methanoregula boonei 6A8	1	1	1	1	1	
Archaea	Euryarchaeota	Methanomicrobia	Methanoplanus petrolearius DSM 11571	1	1	1	1	1	X
Archaea	Euryarchaeota	Methanomicrobia	Methanosarcina acetivorans C2A	1	1	1	1	1	
Archaea	Euryarchaeota	Methanomicrobia	Methanosarcina barkeri str. Fusaro	1	1	1	1	1	X
Archaea	Euryarchaeota	Methanomicrobia	Methanosarcina mazei Go1	1	1	1	1	1	
Archaea	Euryarchaeota	Methanomicrobia	Methanosphaerula palustris E1-9c	1	1	1	1	1	
Archaea	Euryarchaeota	Methanomicrobia	Methanocella paludicola SANAE	1	0	0	0	0	
Archaea	Euryarchaeota	Methanomicrobia	Methanococcoides burtonii DSM 6242	1	0	0	0	0	
Archaea	Euryarchaeota	Methanomicrobia	Methanocorpusculum labreanum Z	1	0	0	0	0	
Archaea	Euryarchaeota	Methanomicrobia	Methanoculleus marisnigri JR1	1	0	0	0	0	
Archaea	Euryarchaeota	Methanomicrobia	Methanohalobium evestigatum Z-7303	1	0	0	0	0	
Archaea	Euryarchaeota	Methanomicrobia	Methanohalophilus mahii DSM 5219	1	0	0	0	0	
Archaea	Euryarchaeota	Methanomicrobia	Methanosaeta thermophila PT	1	0	0	0	0	
Archaea	Euryarchaeota	Methanomicrobia	Methanospirillum hungatei JF-1	1	0	0	0	0	
Archaea	Euryarchaeota	Methanopyri	Methanopyrus kandleri AV19	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Frankia alni ACN14a	1	1	1	1	1	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. CcI3	1	1	1	1	1	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. EAN1pec	1	1	1	1	1	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. EuI1c(2009)	1	0	0	0	0	X
Bacteria	Aquificae	Aquificae	Thermocrinis albus DSM 14484	1	1	0	1	1	
Bacteria	Aquificae	Aquificae	Hydrogenobacter thermophilus TK-6	1	0	0	1	1	
Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Candidatus Azobacteroides pseudotrichonymphae genomovar. CFP2	1	1	1	1	1	

Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Chlorobaculum parvum NCIB 8327	1	1	1	1	1	
Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Chlorobium chlorochromatii CaD3	1	1	1	1	1	
Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Chlorobium limicola DSM 245	1	1	1	1	1	X
Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Chlorobium luteolum DSM 273	1	1	1	1	1	X
Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Chlorobium phaeobacteroides BS1	1	1	1	1	1	X
Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Chlorobium phaeovibrioides DSM 265	1	1	1	1	1	
Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Chlorobium tepidum TLS	1	1	1	1	1	
Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Chloroherpeton thalassium ATCC 35110	1	1	1	1	1	X
Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Pelodictyon phaeoclathratiforme BU-1	1	1	1	1	1	
Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Prosthecochloris aestuarii DSM 271	1	1	1	1	1	X
Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Chlorobium phaeobacteroides DSM 266	1	1	1	0	1	X
Bacteria	Chlamydiae/Verrucomicrobia group	Chlamydiae/Verrucomicrobia group	Coraliomargarita akajimensis DSM 45221	1	1	1	1	1	
Bacteria	Chlamydiae/Verrucomicrobia group	Chlamydiae/Verrucomicrobia group	Methylacidiphilum infernorum V4	1	1	1	1	1	
Bacteria	Chloroflexi	Chloroflexi	Dehalococcoides ethenogenes 195	1	1	1	1	1	
Bacteria	Chloroflexi	Chloroflexi	Roseiflexus castenholzii DSM 13941	1	1	1	0	0	
Bacteria	Chloroflexi	Chloroflexi	Roseiflexus sp. RS-1	1	1	1	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena variabilis ATCC 29413	1	1	1	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	cyanobacterium UCYN-A	1	1	1	1	1	
Bacteria	Cyanobacteria	Cyanobacteria	Cyanothece sp. ATCC 51142	1	1	1	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	Cyanothece sp. PCC 7424	1	1	1	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	Cyanothece sp. PCC 7425	1	1	1	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	Cyanothece sp. PCC 7822	1	1	1	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	Cyanothece sp. PCC 8801	1	1	1	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	Cyanothece sp. PCC 8802	1	1	1	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	'Nostoc azollae' 0708	1	1	1	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc sp. PCC 7120	1	1	1	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	Synechococcus sp. JA-2-3B'a(2-13)	1	1	1	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	Synechococcus sp. JA-3-3Ab	1	1	1	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	Trichodesmium erythraeum IMS101	1	1	1	1	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc punctiforme PCC 73102	1	0	1	1	1	X
Bacteria	Deferribacteres	Deferribacteres	Denitrovibrio acetiphilus DSM 12809	1	1	1	1	1	
Bacteria	Fibrobacteres/Acidobacteria group	Fibrobacteres/Acidobacteria group	Fibrobacter succinogenes subsp. succinogenes S85	1	0	1	1	0	
Bacteria	Firmicutes	Firmicutes	Alkaliphilus metalliredigens QYMF	1	1	1	1	1	
Bacteria	Firmicutes	Firmicutes	Clostridium acetobutylicum ATCC 824	1	1	1	1	1	X
Bacteria	Firmicutes	Firmicutes	Clostridium beijerinckii NCIMB 8052	1	1	1	1	1	X
Bacteria	Firmicutes	Firmicutes	Clostridium cellulovorans 743B	1	1	1	1	1	
Bacteria	Firmicutes	Firmicutes	Clostridium kluyveri DSM 555	1	1	1	1	1	X
Bacteria	Firmicutes	Firmicutes	Clostridium kluyveri NBRC 12016	1	1	1	1	1	X
Bacteria	Firmicutes	Firmicutes	Clostridium ljungdahlii DSM 13528	1	1	1	1	1	
Bacteria	Firmicutes	Firmicutes	Desulfitobacterium hafniense DCB-2	1	1	1	1	1	
Bacteria	Firmicutes	Firmicutes	Desulfitobacterium hafniense Y51	1	1	1	1	1	
Bacteria	Firmicutes	Firmicutes	Desulfotomaculum acetoxidans DSM 771	1	1	1	1	1	
Bacteria	Firmicutes	Firmicutes	Desulfotomaculum reducens MI-1	1	1	1	1	1	
Bacteria	Firmicutes	Firmicutes	Heliobacterium modesticaldum Ice1	1	1	1	1	1	
Bacteria	Firmicutes	Firmicutes	Thermincola sp. JR	1	1	1	1	1	
Bacteria	Firmicutes	Firmicutes	Thermoanaerobacterium thermosaccharolyticum DSM 571	1	1	1	1	1	
Bacteria	Firmicutes	Firmicutes	Candidatus Desulforudis audaxviator MP104C	1	1	1	0	0	
Bacteria	Firmicutes	Firmicutes	Caldicellulosiruptor kronotskyensis 2002	1	1	0	1	0	
Bacteria	Firmicutes	Firmicutes	Caldicellulosiruptor saccharolyticus DSM 8903	1	0	1	1	0	

Bacteria	Firmicutes	Firmicutes	Clostridium thermocellum ATCC 27405	1	0	1	1	0	
Bacteria	Firmicutes	Firmicutes	Caldicellulosiruptor hydrothermalis 108	1	0	0	1	0	
Bacteria	Firmicutes	Firmicutes	Caldicellulosiruptor kristjanssonii 177R1B	1	0	0	1	0	
Bacteria	Firmicutes	Firmicutes	Alkaliphilus oremlandii OhILAs	1	0	0	0	1	
Bacteria	Firmicutes	Firmicutes	Clostridium botulinum A str. ATCC 3502	1	0	0	0	1	
Bacteria	Firmicutes	Firmicutes	Clostridiales genomosp. BVAB3 str. UPII9-5	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium botulinum A str. ATCC 19397	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium botulinum A str. Hall	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium botulinum A2 str. Kyoto	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium botulinum A3 str. Loch Maree	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium botulinum B1 str. Okra	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium botulinum Ba4 str. 657	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium botulinum F str. Langeland	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium cellulolyticum H10	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Eubacterium limosum KIST612	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Syntrophomonas wolfei subsp. wolfei str. Goettingen	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Moorella thermoacetica ATCC 39073	0	1	0	0	0	
Bacteria	Fusobacteria	Fusobacteria	Ilyobacter polytropus DSM 2926	1	1	1	1	1	
Bacteria	Fusobacteria	Fusobacteria	Fusobacterium nucleatum subsp. nucleatum ATCC 25586	1	0	0	0	0	
Bacteria	Nitrospirae	Nitrospirae	Thermodesulfovibrio yellowstonii DSM 11347	1	1	1	1	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Azorhizobium caulinodans ORS 571	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. B510	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Beijerinckia indica subsp. indica ATCC 9039	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium japonicum USDA 110	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. BTAi1	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS278	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Gluconacetobacter diazotrophicus PAl 5	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Magnetospirillum magneticum AMB-1	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium loti MAFF303099	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Methylobacterium nodulans ORS 2060	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylobacterium sp. 4-46	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylocella silvestris BL2	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium etli CFN 42	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium etli CIAT 652	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium leguminosarum bv. trifolii WSM1325	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium leguminosarum bv. trifolii WSM2304	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium leguminosarum bv. viciae 3841	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. NGR234	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacter capsulatus SB 1003	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacter sphaeroides 2.4.1	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacter sphaeroides ATCC 17025	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacter sphaeroides KD131	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodomicrobium vannielii ATCC 17100	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodopseudomonas palustris BisA53	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodopseudomonas palustris BisB18	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodopseudomonas palustris BisB5	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodopseudomonas palustris CGA009	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodopseudomonas palustris HaA2	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodopseudomonas palustris TIE-1	1	1	1	1	1	X

Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillum rubrum ATCC 11170	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium medicae WSM419	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium meliloti 1021	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Xanthobacter autotrophicus Py2	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Zymomonas mobilis subsp. mobilis NCIB 11163	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Zymomonas mobilis subsp. mobilis ZM4	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacter sphaeroides ATCC 17029	1	1	1	1	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillum centenum SW	1	1	1	1	0	
Bacteria	Proteobacteria	Betaproteobacteria	Azoarcus sp. BH72	1	1	1	1	1	X
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia phymatum STM815	1	1	1	1	1	
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia vietnamiensis G4	1	1	1	1	1	
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia xenovorans LB400	1	1	1	1	1	
Bacteria	Proteobacteria	Betaproteobacteria	Candidatus Accumulibacter phosphatis clade IIA str. UW-1	1	1	1	1	1	
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus taiwanensis	1	1	1	1	1	
Bacteria	Proteobacteria	Betaproteobacteria	Dechloromonas aromatica RCB	1	1	1	1	1	
Bacteria	Proteobacteria	Betaproteobacteria	Herbaspirillum seropedicae	1	1	1	1	1	X
Bacteria	Proteobacteria	Betaproteobacteria	Leptothrix cholodnii SP-6	1	1	1	1	1	
Bacteria	Proteobacteria	Betaproteobacteria	Polaromonas naphthalenivorans CJ2	1	1	1	1	1	
Bacteria	Proteobacteria	Betaproteobacteria	Sideroxydans lithotrophicus ES-1	1	1	1	1	1	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfarculus baarsii DSM 2075	1	1	1	1	1	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfatibacillum alkenivorans AK-01	1	1	1	1	1	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfobacterium autotrophicum HRM2	1	1	1	1	1	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfomicrobium baculatum DSM 4028	1	1	1	1	1	X
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfovibrio magneticus RS-1	1	1	1	1	1	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfovibrio salexigens DSM 2638	1	1	1	1	1	X
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfovibrio vulgaris DP4	1	1	1	1	1	X
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfovibrio vulgaris str. Hildenborough	1	1	1	1	1	X
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfovibrio vulgaris str. 'Miyazaki F'	1	1	1	1	1	X
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfurivibrio alkaliphilus AHT2	1	1	1	1	1	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Pelobacter carbinolicus DSM 2380	1	1	1	1	1	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Syntrophobacter fumaroxidans MPOB	1	1	1	1	1	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Wolinella succinogenes	1	1	1	1	1	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Wolinella succinogenes DSM 1740	1	1	1	1	1	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Anaeromyxobacter sp. Fw109-5	1	1	1	1	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Anaeromyxobacter sp. K	1	1	1	1	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geobacter bemidjiensis Bem	1	1	1	1	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geobacter lovleyi SZ	1	1	1	1	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geobacter metallireducens GS-15	1	1	1	1	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geobacter sp. FRC-32	1	1	1	1	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geobacter sp. M21	1	1	1	1	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geobacter sulfurreducens PCA	1	1	1	1	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geobacter uraniireducens Rf4	1	1	1	1	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Pelobacter propionicus DSM 2379	1	1	1	1	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Acidithiobacillus ferrooxidans ATCC 23270	1	1	1	1	1	X
Bacteria	Proteobacteria	Gammaproteobacteria	Acidithiobacillus ferrooxidans ATCC 53993	1	1	1	1	1	X
Bacteria	Proteobacteria	Gammaproteobacteria	Allochromatium vinosum DSM 180	1	1	1	1	1	X
Bacteria	Proteobacteria	Gammaproteobacteria	Azotobacter vinelandii DJ	1	1	1	1	1	X
Bacteria	Proteobacteria	Gammaproteobacteria	Dickeya dadantii 3937	1	1	1	1	1	
Bacteria	Proteobacteria	Gammaproteobacteria	Dickeya dadantii Ech703	1	1	1	1	1	

Bacteria	Proteobacteria	Gammaproteobacteria	Halorhodospira halophila SL1	1	1	1	1	1	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella pneumoniae 342	1	1	1	1	1	X
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella variicola At-22	1	1	1	1	1	
Bacteria	Proteobacteria	Gammaproteobacteria	Methylococcus capsulatus str. Bath	1	1	1	1	1	X
Bacteria	Proteobacteria	Gammaproteobacteria	Pectobacterium atrosepticum SCRI1043	1	1	1	1	1	
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonas stutzeri A1501	1	1	1	1	1	X
Bacteria	Proteobacteria	Gammaproteobacteria	Teredinibacter turnerae T7901	1	1	1	1	1	
Bacteria	Proteobacteria	Gammaproteobacteria	Tolumonas auensis DSM 9187	1	1	1	1	1	
Bacteria	Proteobacteria	unclassified Proteobacteria	Magnetococcus sp. MC-1	1	1	1	1	1	
Bacteria	Spirochaetes	Spirochaetes	Spirochaeta smaragdinae DSM 11293	1	1	1	1	1	
Bacteria	Spirochaetes	Spirochaetes	Spirochaeta thermophila DSM 6192	1	1	1	1	1	

APÊNDICE 10 – MAPA RELACIONAL DE ORGANISMOS X GENES *nif* ENCONTRADOS. DADOS REFERENTES A GENOMAS INCOMPLETOS.

Taxonomia			Organismo	Genes *nif*					YOUNG, 1992
				H	D	K	E	N	
Archaea	Euryarchaeota	Methanobacteria	Methanothermobacter thermautotrophicus	1	0	1	1	0	X
Archaea	Euryarchaeota	Methanobacteria	Methanobacterium ivanovii	1	0	0	0	0	
Archaea	Euryarchaeota	Methanobacteria	Methanobrevibacter arboriphilus	1	0	0	0	0	
Archaea	Euryarchaeota	Methanobacteria	Methanobrevibacter ruminantium	1	0	0	0	0	
Archaea	Euryarchaeota	Methanobacteria	Methanobrevibacter smithii	1	0	0	0	0	
Archaea	Euryarchaeota	Methanobacteria	Methanobrevibacter smithii DSM 2374	1	0	0	0	0	
Archaea	Euryarchaeota	Methanobacteria	Methanobrevibacter smithii DSM 2375	1	0	0	0	0	
Archaea	Euryarchaeota	Methanococci	Methanococcus maripaludis	1	1	1	1	0	X
Archaea	Euryarchaeota	Methanococci	Methanothermococcus okinawensis IH1	1	1	1	0	0	
Archaea	Euryarchaeota	Methanococci	Methanothermococcus thermolithotrophicus	1	1	0	0	0	X
Archaea	Euryarchaeota	Methanococci	hyperthermophilic methanogen FS406-22	1	0	0	0	0	
Archaea	Euryarchaeota	Methanococci	Methanococcus vannielii	1	0	0	0	0	X
Archaea	Euryarchaeota	Methanococci	Methanococcus voltae	1	0	0	0	0	X
Archaea	Euryarchaeota	Methanomicrobia	Candidatus Methanosphaerula palustris E1-9c	1	1	1	1	1	
Archaea	Euryarchaeota	Methanomicrobia	Methanosarcina barkeri	1	1	1	1	0	X
Archaea	Euryarchaeota	Methanomicrobia	Methanosarcina lacustris	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. ACN14a	1	1	1	1	1	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. EuIK1	1	1	1	1	1	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. EUN1f	1	1	1	1	1	X
Bacteria	Actinobacteria	Actinobacteria	Frankia symbiont of Datisca glomerata	1	1	1	1	1	X
Bacteria	Actinobacteria	Actinobacteria	Frankia alni	1	1	1	1	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp.	1	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mrp182	1	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Slackia exigua ATCC 700122	1	1	1	0	0	
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. ARgP5ag	1	1	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	actinomycete 7501	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	actinomycete L3	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Agromyces sp. ORS 1437	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Arthrobacter sp. 18/1	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Arthrobacter sp. 18/4	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. 32-72	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. 32-75	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. 32-85	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. 55005	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. ACN1ag(2009)	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Ag45/Mut15	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Ag8c	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AgB16	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AgB20	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AgB32	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AgGS'84/18	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AgGS'84/44	1	0	0	0	0	X

Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AgKG'84/4	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AgKG'84/5	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AgN2Cl1	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AgN2Cl2	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AgP1R1	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AgP1R2	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AgP1R3	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AgP1R4	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AgPm24	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Ai14a	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Ai7a	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AiBp5	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AiPa1	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AiPs1	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AiPs3	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. AiPs4	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Airl1	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. An2.1	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. An2.2	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Arl3	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Arl4	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Arl5	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Avcl1	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Avsl4	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. BMG5.10	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. BMG5.11	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. BMG5.12	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. BMG5.2	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. BMG5.3	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. BMG5.4	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. BMG5.5	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. BMG5.6	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Cc1.17	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. CeF	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. CeSI5	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Cg70.1	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. CjI-82	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. CpI1(2009)	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. D11	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. DSM 43829	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Ea1.12	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. G2	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Hrl1	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. IPN Ce16	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mgl5	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mpl1	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. ORS020608	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. R43(2009)	1	0	0	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. R96	1	0	0	0	0	X

Bacteria	Actinobacteria	Actinobacteria	Gordonibacter pamelaeae 7-10-1-b	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Microbacterium sp. ORS 1417	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Microbacterium sp. ORS 1418	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Microbacterium sp. ORS 1472	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Microbacterium sp. Zapt11	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Micrococcus sp. Y70	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Micromonospora lupini	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Micromonospora saelicesensis	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Micromonospora sp. MIA32	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Micromonospora sp. MIA38	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Micromonospora sp. MIA57	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Micromonospora sp. SB1-39	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Micromonospora sp. SB1-46	1	0	0	0	0	
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. 202	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. D1	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. D19	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. D2	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. D21	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. D3	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. D6	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. EaI1	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. EaI2	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. EaI3	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. EaI4	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. EaI5	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. EaI6	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. EaI7	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. H1	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. H11	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. H3	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRbt16	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRbt19	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRbt20	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRcf21	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRcf23	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRhbt17	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRhbt29	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRhbt30	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRhbt34	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRhdm1	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRhj6	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRjn1	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRlcF	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRnad12	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRnad14	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRnd4	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRwc2	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. HRwh6	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. K1	0	1	1	0	0	X

Bacteria	Actinobacteria	Actinobacteria	Frankia sp. K8	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mbj1	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mbj2	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mbj3	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mbj4-L	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mbj4-S	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mbm	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mdk1	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mdk2-L	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mdk2-S	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mdk3-L	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mdk3-S	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mdz-L	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mdz-S	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mmy1	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mmy2	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Msm-L	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Msm-S	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mwd	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mwy-L	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Mwy-S	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Myl-L	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Myl-S	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. N7	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. P7	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Te	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. X11	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Y8	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Z5	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Z7	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. Z8	0	1	1	0	0	X
Bacteria	Actinobacteria	Actinobacteria	Frankia sp. M2	0	1	0	0	0	X
Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Chlorobium ferrooxidans DSM 13031	1	1	1	1	1	
Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Prosthecochloris vibrioformis DSM 265	1	1	1	1	1	X
Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Paludibacter propionicigenes WB4	1	1	1	1	0	
Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Chlorobaculum macestae	1	0	0	0	0	
Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Chlorobaculum tepidum	1	0	0	0	0	
Bacteria	Bacteroidetes/Chlorobi group	Bacteroidetes/Chlorobi group	Prevotella bryantii B14	1	0	0	0	0	
Bacteria	Chlamydiae/Verrucomicrobia group	Chlamydiae/Verrucomicrobia group	Methylacidiphilum fumariolicum	1	1	1	1	1	
Bacteria	Chlamydiae/Verrucomicrobia group	Chlamydiae/Verrucomicrobia group	Opitutaceae bacterium TAV2	1	1	1	1	1	
Bacteria	Chlamydiae/Verrucomicrobia group	Chlamydiae/Verrucomicrobia group	Verrucomicrobiae bacterium DG1235	1	1	1	1	1	
Bacteria	Chloroflexi	Chloroflexi	Oscillochloris trichoides DG6	1	1	1	0	0	
Bacteria	Chloroflexi	Chloroflexi	Oscillochloris trichoides	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Crocosphaera watsonii WH 8501	1	1	1	1	1	
Bacteria	Cyanobacteria	Cyanobacteria	cyanobacterium endosymbiont of Rhopalodia gibba	1	1	1	1	1	
Bacteria	Cyanobacteria	Cyanobacteria	Cyanothece sp. CCY0110	1	1	1	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii CS-505	1	1	1	1	1	
Bacteria	Cyanobacteria	Cyanobacteria	Gloeothece sp. KO68DGA	1	1	1	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	Lyngbya sp. PCC 8106	1	1	1	1	1	X

Bacteria	Cyanobacteria	Cyanobacteria	Microcoleus chthonoplastes PCC 7420	1	1	1	1	1	
Bacteria	Cyanobacteria	Cyanobacteria	Nodularia spumigena CCY9414	1	1	1	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	Lyngbya majuscula CCAP 1446/4	1	1	1	1	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Synechococcus sp. PCC 7335	1	1	1	1	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena cylindrica PCC 7122	1	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena siamensis TISTR 8012	1	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena sp. CA	1	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Chlorogloeopsis fritschii PCC 6912	1	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Fischerella ambigua UTEX 1903	1	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Fischerella sp. UTEX 'LB 1931'	1	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Microcoleus chthonoplastes CCY9605	1	1	1	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Microcoleus chthonoplastes CCY9606	1	1	1	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Microcoleus chthonoplastes CCY9607	1	1	1	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Microcoleus chthonoplastes CCY9608	1	1	1	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc sp. 1189P	1	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc sp. 1190P	1	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Oscillatoria sp. PCC 6506	1	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Scytonema hofmanni PCC 7110	1	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Synechococcus sp.	1	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena azollae	1	1	0	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena sp. L-31	1	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Chroococcidiopsis thermalis PCC 7203	1	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Gloeothece sp. PCC 6909	1	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Leptolyngbya boryana IAM M-101	1	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Leptolyngbya sp. PCC 7104	1	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Leptolyngbya sp. PCC 7375	1	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Lyngbya aestuarii PCC 7419	1	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus	1	1	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Microcoleus chthonoplastes CCY0002	1	1	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Microcoleus chthonoplastes CCY9602	1	1	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Microcoleus chthonoplastes CCY9603	1	1	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Microcoleus chthonoplastes CCY9604	1	1	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc commune	1	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc sp. PCC 6720	1	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Oscillatoria sancta PCC 7515	1	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Pseudanabaena sp. PCC 7403	1	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Symploca atlantica PCC 8002	1	1	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Trichodesmium thiebautii	1	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Xenococcus sp. PCC 7305	1	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Chroococcales cyanobacterium LEGE 060123	1	0	1	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Tolypothrix sp. PCC 7101	1	0	1	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Scytonema hofmanni UTEX 2349	1	0	0	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena ambigua LH-M	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena aphanizomenoides M17	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena aphanizomenoides NRE2	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena azotica FACHB-118	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena cylindrica	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena cylindrica UTAD_A212	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena oscillarioides	1	0	0	0	0	X

Bacteria	Cyanobacteria	Cyanobacteria	Anabaena sp. A2	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena sp. BC1	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena sp. CH1	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena sp. I1	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena sp. LG1	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena sp. LG2	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena sp. PA1	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena sp. PCC 9109	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena sp. PO1	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena sphaerica UTEX 'B 1616'	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena sphaerica var. tenuis PMC188.03	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena sphaerica var. tenuis PMC229.04	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena sphaerica var. tenuis PMC246.05	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena sphaerica var. tenuis PMC266.06	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaenopsis sp. NRE1	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Aphanizomenon aphanizomenoides UADFA13	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Aphanizomenon aphanizomenoides UADFA3	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Aphanizomenon aphanizomenoides UADFA5	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Aphanizomenon aphanizomenoides UADFA6	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Aphanizomenon aphanizomenoides UADFA7	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Aphanizomenon aphanizomenoides UADFA8	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Aphanizomenon gracile UADFA10	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Aphanizomenon gracile UADFA11	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Aphanizomenon gracile UADFA12	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Aphanizomenon gracile UADFA16	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Aphanizomenon gracile UADFA2	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Aphanizomenon issatschenkoi UADFA1	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Aphanizomenon sp. KAC15	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Bahamian heterocystous cyanobacterium C1C5	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Calothrix sp.	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Calothrix sp. MCC-3A	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Chroococcidiopsis sp. 'Bad Sachsa'	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Chroococcidiopsis sp. MMG-5	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Chroococcidiopsis sp. MMG-6	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	cyanobacterium KM9	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cyanothece sp. SKTU126	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Cyanothece sp. TW3	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Cyanothece sp. WH 8902	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Cyanothece sp. WH 8904	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis africana PMC115.02	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis curvispora PMC144.02	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis curvispora PMC262.06	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii FLcultureE-3	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii FLcultureL-5	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii LD-D	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii LD-E	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii LD-F	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii LD-G	1	0	0	0	0	

Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii LD-I	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii LG-L	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii NJ3	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii PMC114.02	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii PMC117.02	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii PMC118.02	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii PMC145.02	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii PMC98.14	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii PMC99.06	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii PMC99.08	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermopsis raciborskii PMC99.12	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Dolichospermum flos-aquae PMC207.03	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Dolichospermum flos-aquae PMC208.03	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Dolichospermum planctonicum PMC196.03	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Dolichospermum planctonicum PMC200.03	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Dolichospermum planctonicum PMC230.04	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	filamentous thermophilic cyanobacterium tBTRCCn 101	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	filamentous thermophilic cyanobacterium tBTRCCn 24	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	filamentous thermophilic cyanobacterium tBTRCCn 301	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Gloeocapsa sp. KO20B5	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Gloeocapsa sp. KO30D1	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Gloeocapsa sp. KO38CU6	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Gloeothece sp. KO11DG	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Gloeothece sp. SK40	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Halothece sp. MPI 96P605	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	heterocystous cyanobacterium LD-B	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Katagnymene spiralis	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Leptolyngbya boryana IU 594	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Leptolyngbya nodulosa UTEX 2910	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Leptolyngbya sp. MMG-1	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Leptolyngbya sp. PCC 7410	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Lyngbya lagerheimii UTEX 1930	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Lyngbya sp. CCAP 1446/10	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Lyngbya wollei	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	marine stromatolite cyanobacterium HBC2C2	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	marine stromatolite cyanobacterium HBC5C3	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	marine stromatolite cyanobacterium HBC6C3	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5186	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5192	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5193	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5198	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5201	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5202	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5203	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5204	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5205	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5207	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5208	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5267	1	0	0	0	0	

Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5268	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5272	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5273	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5281	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5282	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5318	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5319	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5320	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5321	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5323	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5324	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5326	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5327	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5328	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5329	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5331	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5332	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5335	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5337	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5379	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus CCMEE 5380	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W1110	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W112	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W114	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W118	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W119	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W156	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W157	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W213	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W226	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W241	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W246	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W247	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W249	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W3110	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W312	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W313	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W321	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W327	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W334	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W344	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W411	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W4110	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W412	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W422	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W435	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W4510	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W457	1	0	0	0	0
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W511	1	0	0	0	0

Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W512	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W515	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W519	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W522	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Mastigocladus laminosus W523	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Microcoleus sp. PCC 8701	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Myxosarcina sp.	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nodularia harveyana CCAP 1452/1	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nodularia sp. KAC 13	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nodularia sphaerocarpa Up16a	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nodularia sphaerocarpa Up16f	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nodularia spumigena AV1	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nodularia spumigena FL2f	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc commune CP	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc commune SC	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc muscorum	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc muscorum 'LB UTEX 1933'	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc muscorum UTAD_N213	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc sp.	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc sp. 'J. Gallon'	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc sp. MCT-1	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc sp. MFG-1	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc sp. N2	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Oscillatoria sp. CCAP 1459/26	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Oscillatoria sp. MMG-2	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Oscillatoriales cyanobacterium JSC-1	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Phormidium sp. AD1	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Richelia sp. SC01	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Scytonema sp.	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Scytonema sp. DC-A	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Scytonema sp. FGP-7A	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Scytonema sp. NC-4B	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Scytonema sp. NCC-4B	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Spirirestis rafaelensis SRS70	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Stanieria cyanosphaera PCC 7437	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Synechocystis sp. WH 001	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Synechocystis sp. WH 002	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Synechocystis sp. WH 003	1	0	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Tolypothrix distorta var. symplocoides UTEX 'B 424'	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Tolypothrix sp. JCT-1	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Tolypothrix sp. LQ-10	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Tolypothrix sp. PCC 7601	1	0	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena sp. PCC 7108	0	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena variabilis	0	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Calothrix desertica PCC 7102	0	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Calothrix sp. PCC 7507	0	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermum stagnale PCC 7417	0	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Fischerella muscicola PCC 7414	0	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Fischerella sp. PCC 7603	0	1	1	0	0	X

Bacteria	Cyanobacteria	Cyanobacteria	Nodularia sphaerocarpa PCC 7804	0	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc sp. PCC 7423	0	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Scytonema sp. PCC 7814	0	1	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Leptolyngbya boryana PCC 6306	0	1	0	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	Chlorogloeopsis fritschii PCC 6718	0	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Chroococcidiopsis sp. PCC 6712	0	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Cylindrospermum majus PCC 7604	0	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Fischerella muscicola SAG 1427-1	0	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Gloeothece membranacea PCC 6501	0	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Leptolyngbya sp. PCC 7004	0	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Leptolyngbya sp. PCC 73110	0	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Lyngbya sp. CCY 9616	0	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nodularia spumigena	0	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc insulare SAG 54.79	0	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc sp. CCMP2511	0	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Pleurocapsa sp. PCC 7327	0	1	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Pleurocapsa sp. PCC 7516	0	1	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Pseudanabaena sp. CCY9509	0	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Pseudanabaena sp. PCC 6802	0	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Pseudanabaena sp. PCC 7409	0	1	0	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Stanieria sp. PCC 7301	0	1	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Tolypothrix tenuis	0	1	0	0	0	
Bacteria	Cyanobacteria	Cyanobacteria	Lyngbya aestuarii	0	0	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Nodularia spumigena PCC 73104	0	0	1	0	0	X
Bacteria	Cyanobacteria	Cyanobacteria	Anabaena affinis UTEX 2649	0	0	0	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	Fischerella muscicola UTEX 1829	0	0	0	1	1	X
Bacteria	Cyanobacteria	Cyanobacteria	Microcoleus sp.	0	0	0	1	1	
Bacteria	Cyanobacteria	Cyanobacteria	Nostoc commune UTEX 584	0	0	0	1	1	X
Bacteria	environmental samples	environmental samples	enrichment culture bacterium 01	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	enrichment culture bacterium 02	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	enrichment culture bacterium 37	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	marine stromatolite eubacterium HB(0697) A06	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	marine stromatolite eubacterium HB(0697) A100	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	marine stromatolite eubacterium HB(0697) A101A	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	marine stromatolite eubacterium HB(0697) C06A	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	marine stromatolite eubacterium HB(0697) C102	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	marine stromatolite eubacterium HB(0697) C104	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	marine stromatolite eubacterium HB(0898) A101	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	marine stromatolite eubacterium HB(0898) A69	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	marine stromatolite eubacterium HB(0898) C03	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	marine stromatolite eubacterium HB(0898) C06	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	marine stromatolite eubacterium HB(0898) Z02	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	marine stromatolite eubacterium HB(0898) Z05	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	marine stromatolite eubacterium HB(0898) Z07	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	marine stromatolite eubacterium HB(0898) Z12	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	marine stromatolite eubacterium HB(0898) Z14	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone AO1102	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone AO1104	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone AO1113	1	0	0	0	0	

Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone BH1132	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone BT1101	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone BT1118	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone HT1103	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone HT1150	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone HT1169	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone HT1177	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone HT1192	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone HT1193	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone HT1195	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone HT1196	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone HT1197	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone HT1198	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone HT1199	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone HT1200	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone HT1201	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone HT1202	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone HT1203	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone HT1204	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone HT1205	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone PO3120	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone PO3133	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone PO3135	1	0	0	0	0	
Bacteria	environmental samples	environmental samples	Unidentified marine eubacterium clone PO3137	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Acetivibrio cellulolyticus CD2	1	1	1	1	1	
Bacteria	Firmicutes	Firmicutes	Clostridium beijerinckii	1	1	1	1	1	X
Bacteria	Firmicutes	Firmicutes	Clostridium butyricum E4 str. BoNT E BL5262	1	1	1	1	1	X
Bacteria	Firmicutes	Firmicutes	Clostridium lentocellum DSM 5427	1	1	1	1	1	
Bacteria	Firmicutes	Firmicutes	Clostridium papyrosolvens DSM 2782	1	1	1	1	1	
Bacteria	Firmicutes	Firmicutes	Clostridium pasteurianum	1	1	1	1	1	X
Bacteria	Firmicutes	Firmicutes	Dethiobacter alkaliphilus AHT 1	1	1	1	1	1	
Bacteria	Firmicutes	Firmicutes	Ethanoligenens harbinense YUAN-3	1	1	1	1	1	
Bacteria	Firmicutes	Firmicutes	Heliobacterium chlorum	1	1	1	1	1	X
Bacteria	Firmicutes	Firmicutes	Paenibacillus massiliensis	1	1	1	1	1	
Bacteria	Firmicutes	Firmicutes	Thermincola potens JR	1	1	1	1	1	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sabinae	1	1	1	0	1	
Bacteria	Firmicutes	Firmicutes	Paenibacillus abekawaensis	1	1	1	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus durus	1	1	1	0	0	X
Bacteria	Firmicutes	Firmicutes	Paenibacillus fujiensis	1	1	1	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus graminis	1	1	1	0	0	
Bacteria	Firmicutes	Firmicutes	Syntrophothermus lipocalidus DSM 12680	1	1	1	0	0	
Bacteria	Firmicutes	Firmicutes	Heliobacterium gestii	1	1	0	1	0	
Bacteria	Firmicutes	Firmicutes	Heliobacillus mobilis	1	1	0	0	0	X
Bacteria	Firmicutes	Firmicutes	Heliobacterium modesticaldum	1	1	0	0	0	
Bacteria	Firmicutes	Firmicutes	Heliorestis baculata	1	1	0	0	0	
Bacteria	Firmicutes	Firmicutes	Heliorestis daurensis	1	1	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium thermocellum DSM 2360	1	0	1	1	0	
Bacteria	Firmicutes	Firmicutes	Clostridium thermocellum JW20	1	0	1	1	0	
Bacteria	Firmicutes	Firmicutes	Ruminococcus albus 7	1	0	1	1	0	

Bacteria	Firmicutes	Firmicutes	Ruminococcus albus 8	1	0	1	1	0	
Bacteria	Firmicutes	Firmicutes	Ruminococcus flavefaciens FD-1	1	0	1	0	0	
Bacteria	Firmicutes	Firmicutes	Caldicellulosiruptor lactoaceticus 6A	1	0	0	1	0	
Bacteria	Firmicutes	Firmicutes	Acetobacterium woodii	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Anaerobacillus alkalilacustre	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Bacillus alkalidiazotrophicus	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Bacillus arseniciselenatis	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Bacillus cereus	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Bacillus macyae	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Bacillus megaterium	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Bacillus sp. BT97	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Bacillus sp. c6	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Bacillus sp. CCBAU 15518	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Bacillus sp. CCBAU 15524	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Bacillus sp. M4	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Bacillus sp. w5	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Blautia hydrogenotrophica DSM 10507	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Bryantella formatexigens DSM 14469	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Butyrivibrio fibrisolvens 16/4	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridiales bacterium 1_7_47_FAA	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium bartlettii DSM 16795	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium bolteae ATCC BAA-613	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium botulinum Bf	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium botulinum F str. 230613	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium botulinum NCTC 2916	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium carboxidivorans P7	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium cellobioparum	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium difficile NAP07	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium difficile NAP08	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium difficile QCD-23m63	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium sp. B901-1b	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium sp. Kas104-4	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium sp. Kas106-4	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium sp. Kas107-1	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium sp. Kas107-2	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium sp. MK11	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium sp. MK12	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium sp. MK31	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium sp. MK8	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium sp. Sukashi-1	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium sporogenes ATCC 15579	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Coprococcus catus GD/7	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Coprococcus comes ATCC 27758	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Desulfitobacterium dehalogenans	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Desulfosporosinus orientis	1	0	0	0	0	X
Bacteria	Firmicutes	Firmicutes	Desulfotomaculum nigrificans DSM 574	1	0	0	0	0	X
Bacteria	Firmicutes	Firmicutes	Dialister invisus DSM 15470	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Dorea formicigenerans ATCC 27755	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Dorea longicatena DSM 13814	1	0	0	0	0	

Bacteria	Firmicutes	Firmicutes	Eubacterium hallii DSM 3353	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Megasphaera genomosp. type_1 str. 28L	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Megasphaera micronuciformis F0359	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Mitsuokella multacida DSM 20544	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Natronobacillus azotifigens	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus borealis	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus brasilensis	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus forsythiae	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus macerans	1	0	0	0	0	X
Bacteria	Firmicutes	Firmicutes	Paenibacillus odorifer	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus polymyxa	1	0	0	0	0	X
Bacteria	Firmicutes	Firmicutes	Paenibacillus riograndensis	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. AS2(2010)	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. Bb24	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. Bb54	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. Bd43	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. Be17	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. Bs57	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. BY55	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. By56	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. DF4MM9	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. g2	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. GA12	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. GA5	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. Gc58	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. GJ10	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. GJ11	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. GJ24	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. GJ46	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. GJ52	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. GJ9	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. Hp7	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. JT91	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. Nz28	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. Nz29	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. Nz30	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. S27	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. Sl51	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. Ss35	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. Sx52	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus sp. X19-5	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus stellifer	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus wynnii	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Paenibacillus zanthoxyli	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Ruminococcus obeum A2-162	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Ruminococcus obeum ATCC 29174	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Ruminococcus sp. SR1/5	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Ruminococcus torques L2-14	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Selenomonas flueggei ATCC 43531	1	0	0	0	0	

Bacteria	Firmicutes	Firmicutes	Selenomonas noxia ATCC 43541	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Selenomonas sp. oral taxon 149 str. 67H29BP	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Selenomonas sputigena ATCC 35185	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Turicibacter sanguinis PC909	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Turicibacter sp. PC909	1	0	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium butyricum 5521	0	1	1	1	1	X
Bacteria	Firmicutes	Firmicutes	Clostridium hungatei	0	1	0	0	0	
Bacteria	Firmicutes	Firmicutes	Clostridium sp. IBUN 22A	0	0	0	0	1	
Bacteria	Fusobacteria	Fusobacteria	Fusobacterium nucleatum subsp. nucleatum ATCC 23726	1	0	0	0	0	
Bacteria	Fusobacteria	Fusobacteria	Fusobacterium sp. 3_1_27	1	0	0	0	0	
Bacteria	Fusobacteria	Fusobacteria	Fusobacterium sp. 3_1_33	1	0	0	0	0	
Bacteria	Fusobacteria	Fusobacteria	Fusobacterium ulcerans ATCC 49185	1	0	0	0	0	
Bacteria	Fusobacteria	Fusobacteria	Fusobacterium varium ATCC 27725	1	0	0	0	0	
Bacteria	Nitrospirae	Nitrospirae	Leptospirillum ferrodiazotrophum	1	1	1	1	1	
Bacteria	Nitrospirae	Nitrospirae	Leptospirillum ferrooxidans	1	1	1	1	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum brasilense	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium japonicum	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Magnetospirillum gryphiswaldense MSR-1	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Magnetospirillum magnetotacticum MS-1	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium loti	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium opportunistum WSM2075	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylosinus trichosporium OB3b	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium etli 8C-3	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium etli CIAT 894	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium etli GR56	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium etli Kim 5	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium leguminosarum bv. trifolii	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacter capsulatus	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacter sp. SW2	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodopseudomonas palustris DX-1	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium meliloti AK83	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium meliloti BL225C	1	1	1	1	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Zymomonas mobilis subsp. mobilis ATCC 10988	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Zymomonas mobilis subsp. mobilis NCIMB 11163	1	1	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Gluconacetobacter diazotrophicus	1	1	1	1	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Gluconacetobacter johannae	1	1	1	1	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium amorphae	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium caraganae	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium chacoense	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium ciceri	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium huakuii	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium mediterraneum	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium plurifarium	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium septentrionale	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 11185	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 11196	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 11206	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 11214	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 11217	1	1	1	0	0	

Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 11226	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 11231	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 11257	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium temperatum	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium tianshanense	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp.	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. ANU289	1	1	1	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Citreicella sp. SE45	1	1	0	1	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium meliloti	1	1	0	0	1	X
Bacteria	Proteobacteria	Alphaproteobacteria	Beijerinckia derxii subsp. venezuelae	1	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Beijerinckia indica subsp. lacticogenes	1	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Beijerinckia mobilis	1	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium canariense	1	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium elkanii	1	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium liaoningense	1	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium yuanmingense	1	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Gluconacetobacter azotocaptans	1	1	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylocapsa acidiphila B2	1	1	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylocella palustris	1	1	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylocella tundrae	1	1	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylocystis echinoides	1	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Methylocystis sp. H9a	1	1	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylosinus sporium	1	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Methylosinus trichosporium	1	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales bacterium lut6	1	1	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium etli	1	1	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium medicae	1	1	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium etli Brasil 5	1	0	1	1	1	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodospirillum rubrum	1	0	1	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Agrobacterium rhizogenes	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Agrobacterium tumefaciens	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Aminobacter sp. BA135	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Amorphomonas oryzae	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Aurantimonas sp. CCNWGS0021-2	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Azorhizobium caulinodans	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Azorhizobium doebereinerae	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Azorhizobium sp. KNUC169	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum amazonense	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum canadense	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum doebereinerae	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum halopraeferens	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum irakense	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum lipoferum	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum melinis	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum oryzae	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum picis	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum rugosum	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. BV-s	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. CC-Nfb-7	1	0	0	0	0	

Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. ptl-3	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSA10	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSA11	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSA12	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSA13	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSA14w	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSA15c	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSA16	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSA18	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSA19	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSA20c	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSA2s	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSA36t	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSA6c	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSH10	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSH100	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSH19	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSH2	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSH21	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSH5	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSH58	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSH6	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSH64	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSH78	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSH81w	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. TSH96	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum zeae	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Beijerinckia derxii subsp. derxii	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Beijerinckia indica subsp. indica	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Beijerinckiaceae bacterium AR4	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Beijerinckiaceae bacterium BW863	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Bosea thiooxidans	1	0	0	0	0		
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium canariense bv. genistearum	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium denitrificans	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium elkanii USDA 76	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium genosp. alpha bv. genistearum	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium iriomotense	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium japonicum bv. genistearum	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium japonicum bv. glycinearum	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium jicamae	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium liaoningense bv. glycinearum	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium pachyrhizi	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 108a1	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 114d	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 123b	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 125e	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 15LBIV	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 16LBIV	1	0	0	0	0	X	
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 2	1	0	0	0	0	X	

Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 21LBIV	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 22LBI	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 29d	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 3LBC	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 3LBIV	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 48a	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 49b	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 55a	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 60	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 70b2	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 9LBC	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. AT1	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. BTAi 1	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 051011	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 051107	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 05588	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 23005	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 23011	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 23086	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 23155	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 23160	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 23236	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 25360	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 25374	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 33011	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 33012-2	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 33037	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 33039	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 33062	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 33067	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 33128	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 35186-1	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 35272	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 43058	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 43135	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 45066	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 45281	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 45384	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 45389	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 51012	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 51159	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 51172	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 51377	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 61069	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 61071	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 61073	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 61434	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 65066	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 65718	1	0	0	0	0	X

Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 65788	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CCBAU 65884	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CTAW11	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CTAW22	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CTAW38	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. CTAW71	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. DASA 68056	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. EPxn1	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. GM	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. IRBG 2	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. IRBG 228	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. IRBG 230	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. KO11	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. KO13	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. KO14	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. KO17	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. KO19A	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. KO20A	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. KO23	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. KO3A	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. KO3B	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. KO3C	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. KO3D	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. KO3E	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. KO3F	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. KO3G	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. KO3H	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. KO5	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. LMTR 14	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. LMTR 19	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. LMTR 21	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. LMTR 3	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. MAFF 210190	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. MAFF 210197	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. MAFF 210318	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. NZP2309	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS277	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS279	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS285	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS287	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS294	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS297	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS298	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS300	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS301	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS303	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS304	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS305	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS307	1	0	0	0	0	X

Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS309	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS310	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS318	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS324	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS327	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS328	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS330	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS331	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS336	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS344	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS354	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS356	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS358	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS361	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS372	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS375	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS377	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS391	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS393	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ORS400	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. RRD24	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. RRM3	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. RRM8	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. RRSSET16	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. RS13	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. RSA104	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. RSA3	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. RSB2	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. RSB6	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. RSS137	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. RST88bis	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. RST89	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. S4 0	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. SR106	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. SR66	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. SR69	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. SR78	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. TSA1	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. TSA15y	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. TSA26	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. TSA27b	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. TSA27s	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. TSA43	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. TSA44	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Devosia neptuniae	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Ensifer mexicanus	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Magnetospirillum sp. J10	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	marine proteobacterium 'Bird Shoal 1-2'	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	marine proteobacterium 'Bird Shoal 8-1'	1	0	0	0	0	

Bacteria	Proteobacteria	Alphaproteobacteria	marine proteobacterium 'Sippewissett 2-21'	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	marine proteobacterium 'Tomales Bay lg wh'	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium albiziae	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium alhagi	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium australicum WSM2073	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium ciceri biovar biserrulae	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium gobiense	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium robiniae	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. 102-Evora	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. 128a	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. 64b-Beja	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. 78-Elvas	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. 85-Elvas	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. 90-Evora	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. 94-Evora	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. 98-Evora	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. AC100c	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. AC21c2	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. AC39e1	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. AC98c	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. AC98e	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. ACMP18	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. BA134	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. BA151	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. BD56	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. BD68	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. BR3804	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 01550	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 11166	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 11270	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 11300	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 25282	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 25300	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 45265	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 45272	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 65321	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 65323	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 65324	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 65327	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 65336	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCBAU 73184	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCNWGS0010-1	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCNWGS0011	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCNWGS0015-1	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCNWXJ16-1	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCNWXJ32-3	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CCNWXJ40-4	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CIAM0210	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CMSS27	1	0	0	0	0

Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CPS13	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. CV-18-Elvas	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. DWO366	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. GGS20	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. GMSS1	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. GMSS22	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. HAMBI1487	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. INPA78B	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. LMG 11892	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. LMG 4270	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. M72P	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. ML105	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. ML108	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. ML128	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. N105	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. N141	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. N2076	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. N3	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. N33	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. N362	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. N441	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. ORS 3324	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. ORS 3355	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. ORS 3356	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. ORS 3365	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. ORS 3369	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. ORS 3404	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. rob10	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. rob11	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. rob12	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. rob13	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. rob6	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. S1274	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. S1302	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. S673	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. S786	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. S789	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. S792	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. S828	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. S948	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. S956	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. SCAU11	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. SCAU2	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. SCAU5	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. SCAU7	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. SCAU9	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. U510	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. USDA 3350	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. USDA 4207	1	0	0	0	0

Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. USDA 4214	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. USDA 4236	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. USDA 4255	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. USDA 4297	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. USDA 4318	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. USDA 4322	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. WSM1283	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. WSM1284	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. WSM1497	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. WSM2074	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. WSM3862	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. WSM3865	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. WSM3866	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. WSM3869	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. WSM3871	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. WSM3872	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. WSM3875	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. WSM3877	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. WSM3883	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. xhj20	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. xhj23	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. xhj26	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. xhj7	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium tarimense	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	methanotroph E10	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylobacterium nodulans	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylocapsa acidiphila	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylocapsa sp. KYG	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylocella silvestris	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylocystis methanolicus	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylocystis minimus	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylocystis parvus	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Methylocystis rosea	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylocystis sp. LW2	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylocystis sp. LW5	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylocystis sp. m261	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylocystis sp. SB2	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylosinus sp. LW3	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylosinus sp. LW4	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylosinus sp. LW8	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Methylosinus sp. PW1	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Novosphingobium nitrogenifigens	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Ochrobactrum cytisi	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Ochrobactrum lupini	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Phaeospirillum fulvum	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Phyllobacterium sp. ORS 1402	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Phyllobacterium sp. ORS 1403	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Pleomorphomonas oryzae	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales bacterium ltg2	1	0	0	0	0	

Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium alkalisoli	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium daejeonense	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium etli bv. mimosae	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium etli bv. phaseoli	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium etli IE4771	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium galegae	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium gallicum	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium gallicum bv. gallicum	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium gallicum bv. phaseoli	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium giardinii	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium hainanense	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium huautlense	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium leguminosarum	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium leguminosarum bv. viciae	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium leguminosarum bv. viciae USDA 2370	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium loessense	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium lusitanum	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium mongolense	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium multihospitium	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium phaseoli	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. 525W	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. BR6001	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 23084	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 33220a	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 65118	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 65255	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 65647	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 65813	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83266	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83268	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83309	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83452	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83457	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83475	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83476	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83477	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83480	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83482	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83485	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83489	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83490	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83491	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83514	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83515	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83517	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83526	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCBAU 83530	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCNWGS0022	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCNWGS0187	1	0	0	0	0	

Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCNWQTX14	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CCNWSX0011-1	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. CFN ESH34	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. DASA 68006	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. KNUC172	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. NCHA22	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. NGR181	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. RPJ16	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. RPJ3	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. RPJ5	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. RPJ6	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. RPP14	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. RPP20	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. SA3	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. SL-1	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. TJ167	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. TJ171	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. TJ172	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. TJ173	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. USDA 1920	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium sp. W3	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium tibeticum	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium tropici	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium yanglingense	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacter azotoformans	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacter blasticus	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacter sp. AP-10	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacter sphaeroides	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodoblastus acidophilus	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodopseudomonas faecalis	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodopseudomonas lichen	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodopseudomonas oryzae	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodopseudomonas palustris	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodopseudomonas sp. 99D	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodopseudomonas sp. HMD88	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodopseudomonas sp. HMD89	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodovulum sp. CP-10	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodovulum strictum	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodovulum sulfidophilum	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Shinella kummerowiae	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium americanum	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium arboris	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium chiapanecum	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium fredii	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium fredii bv. mediterranense	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium kostiense	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium kummerowiae	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium meliloti bv. mediterranense	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium saheli	1	0	0	0	0	

Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium saheli bv. sesbaniae	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. BR4007	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. BR827	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. C2	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. C4	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. C5	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. C9	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 05600	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 05606	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 05617	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 05631	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 05638	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 05640	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 05646	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 05672	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 05684	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 31015	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 33036a	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 53044B	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 65732	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 83317	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 83643	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 83647	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 83666	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 83742	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 83751	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCBAU 83765	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. CCNWGS114	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. DASA 68012	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. DWO607	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. gx-153	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. HAMBI1394	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. HAMBI1478	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. HAMBI1480	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. HAMBI1499	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. ITTG S8	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. LILM2009	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. M6	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. Rch-9813	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. Rch-9868	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. S002	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. S005	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. S007	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium sp. TJ170	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium terangae	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sinorhizobium xinjiangense	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas azotifigens	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12190	1	0	0	0	0
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12195	1	0	0	0	0

Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12200	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12245	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12246	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12247	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12248	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12249	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12250	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12252	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12253	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12254	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12255	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12256	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12257	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12258	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12259	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12260	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12261	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12262	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12263	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. BR12264	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. KNUC167	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sphingomonas sp. Y49	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Starkeya sp. ORS 1474	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Sulfitobacter sp. EE-36	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Telmatospirillum siberiense	1	0	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Xanthobacter autotrophicus	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Xanthobacter flavus	1	0	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Azospirillum sp. LOD4	0	1	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 1808N	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 5028A	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 5029F	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 5057B	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 5111P	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 5329H	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 5493M	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 5563D	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 5680G	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 5792C	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. 5915J	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. A110	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. A113	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. A12	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. A120	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. A13	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. A14	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. A210	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. A211	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. A214	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. A218	0	1	0	0	0	X

Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. A23	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. A26	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. A28	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. A29	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Aec12.12	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Aec12.3	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Aei-1	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. aeky10	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. aeky3	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Ai1a-2	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Ai4.2	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Albi6	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ApB16	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ApCV2	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ApE4.8	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ApT12	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. ApT2	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. AU5sfb	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. bfs1b	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. BtLT4	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. BtRI3	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. C11	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. C111	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. C13-2660	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. C14-2660	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. C15-2660	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. C16-2660	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. C17-2660	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. C18	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. C20-2420	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. C21	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. C21-2420	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Cf1	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Cj3.3	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Cp5.3	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Crj-1	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Da3.1	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. DcE2.8	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. DesT1	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. DesT10	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Dm4.3	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Dp6.5	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Dr12.2	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Dr3b-11	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. DrLT5	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Ec3.3	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. H11	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. H13	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. H15	0	1	0	0	0	X

Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. H16	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. H31	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. H32	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. H33	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Ih3.3	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. jwc91-2	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. jws91-2	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Kus-2	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Kus-4	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. La5-8	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. LcCT6	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. LcRI3	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Leb-12	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Leb-16	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Lh10	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Lop10.4	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Lop14.22	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Lop2.4b	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Lppb2	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Lpsp.1b	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Ma9.4	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Mm1.3	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. onl92.10	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. onl92.6	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Pe1.3	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Pe4	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Pe5.2b	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. phym.6a	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Pp2-4	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Pp3a.1	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Ppar1-21	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Ppar1-31	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Ppau3-41	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Pter17	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Pter29	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Pter37	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Pter4	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Pter7	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Rp2.1	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Soph313_CPI-0246	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Tep5	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. th.b2	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. tpar1.1	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. tpma1.5	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Tv2a-2	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. U11	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. U115	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. U118	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. U12	0	1	0	0	0	X

Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. U13	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. U21	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. U214	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. U23	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. U29	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. UU22sfb	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Vgn-2	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Vgn-5	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Vp13.3	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Wall12	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Wall28	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Bradyrhizobium sp. Wall9	0	1	0	0	0	X
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. Lc3-2	0	1	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. Lo5-9	0	1	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. Lo7-12	0	1	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Mesorhizobium sp. Lo9-4	0	1	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobiales bacterium lut5	0	1	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Roseomonas genomospecies 6	0	1	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Roseomonas gilardii	0	1	0	0	0	
Bacteria	Proteobacteria	Alphaproteobacteria	Rhizobium leguminosarum bv. trifolii TA1	0	0	0	1	1	X
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. CCGE1002	1	1	1	1	1	
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. Ch1-1	1	1	1	1	1	
Bacteria	Proteobacteria	Betaproteobacteria	Herbaspirillum seropedicae SmR1	1	1	1	1	1	X
Bacteria	Proteobacteria	Betaproteobacteria	Delftia tsuruhatensis	1	1	1	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Alcaligenes faecalis	1	1	0	0	0	X
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia tropica	1	1	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia tuberum	1	1	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Herbaspirillum sp. B501	1	1	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Azoarcus sp.	1	0	1	0	0	X
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia vietnamiensis	1	0	1	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Dechloromonas sp. SIUL	1	0	1	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Azoarcus communis	1	0	0	0	0	X
Bacteria	Proteobacteria	Betaproteobacteria	Azoarcus indigens	1	0	0	0	0	X
Bacteria	Proteobacteria	Betaproteobacteria	Azoarcus sp. DQS-4	1	0	0	0	0	X
Bacteria	Proteobacteria	Betaproteobacteria	Azoarcus tolulyticus	1	0	0	0	0	X
Bacteria	Proteobacteria	Betaproteobacteria	Azohydromonas australica	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Azohydromonas lata	1	0	0	0	0	X
Bacteria	Proteobacteria	Betaproteobacteria	Azonexus caeni	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Azonexus fungiphilus	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Azospira oryzae	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Azospira restricta	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Azovibrio restrictus	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	beta proteobacterium d8-1	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	beta proteobacterium d8-2	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	beta proteobacterium IMCC1716	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia brasilensis	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia caryophylli	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia cepacia	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia ferrariae	1	0	0	0	0	

Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia fungorum	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia kururiensis	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia mimosarum	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia nodosa	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia phymatum	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia phytofirmans	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia silvatlantica	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia silvatlantica SRMrh-20	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. Br3407	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. Br3469	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. Cs130	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. DUS751	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. gx97	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY155	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY156	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY157	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY158	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY159	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY160	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY161	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY162	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY164	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY166	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY167	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY168	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY169	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY170	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY171	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY251	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY255	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY256	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY257	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY258	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY259	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY262	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY264	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY266	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY268	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY270	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY271	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY272	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY276	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY277	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY278	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY279	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY280	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY281	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY282	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY283	1	0	0	0	0

Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY284	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY285	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY286	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY288	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY294	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY296	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY298	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY300	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY301	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY303	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY306	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY307	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY309	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY310	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY312	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY314	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY315	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY317	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY321	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY323	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY324	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY325	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY326	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY345	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY346	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY349	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY350	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY351	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY359	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY366	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY368	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY380	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY381	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY382	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY384	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY385	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY386	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY388	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY389	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY395	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY396	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY401	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY403	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY406	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY407	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY409	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY410	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY411	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY413	1	0	0	0	0

Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY414	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY417	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY418	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY419	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY420	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY421	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY422	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY425	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY430	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY431	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY432	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY434	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY439	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY441	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY449	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY452	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY453	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY454	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY461	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY465	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY467	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY468	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY471	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY472	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY476	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY478	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY479	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY480	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY481	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY484	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY485	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY487	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY489	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY490	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY492	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY504	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY578	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY581	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY582	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY585	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY587	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY595	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY604	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY605	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY622	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY623	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY629	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY634	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY636	1	0	0	0	0

Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. JPY637	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. KJ006	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. Ms116	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. NGR190	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. PTK47	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. SWF66044	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. TJ182	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. TS2	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. WSM3930	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. WSM3937	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia unamae	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia unamae MTI-641	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia xenovorans	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. amp18	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. amp34	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. amp45	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. amp6	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. cmp2	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. cmp29	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. cmp36	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. cmp52	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. cmp57	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. cmp60	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. ip2.14	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. ip2.35	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. ip2.89	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. MApud10.1	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. mapud10.5	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. mapud3.2	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. MApud3.4	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. mapud4.3	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. MApud8.1	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. mip15	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. mip2	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. mpp1.1	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. mpp1.13	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. mpp1.16	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. mpp1.6	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. mpp2.1	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. mpp2.10	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. mpp2.18	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. mpp2.26	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. mym10	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. myp8	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. pp2.22	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. pp2.3	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. pp2.43	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. pp2.49	1	0	0	0	0
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. pp2.75	1	0	0	0	0

Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. SWF66166	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. SWF66194	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. SWF66294	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. SWF66316	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. SWF66322	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. SWF67044	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. Tpig.6a	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. tpud17.3	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. tpud23.3	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. Tpud27.2	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. tpud28.1	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Cupriavidus sp. tpud31.3	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Derxia gummosa	1	0	0	0	0	X
Bacteria	Proteobacteria	Betaproteobacteria	Herbaspirillum rubrisubalbicans	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Herbaspirillum sp. KP1-50	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Herbaspirillum sp. KP1-77	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Ideonella dechloratans	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Ideonella sp. 1a22	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Ideonella sp. Long 7	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Pelomonas saccharophila	1	0	0	0	0	X
Bacteria	Proteobacteria	Betaproteobacteria	Pseudacidovorax intermedius	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Pseudacidovorax sp. ptl-2	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Ralstonia sp. ISSDS-784	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Rhodoferax antarcticus	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Roseateles depolymerans	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Rubrivivax gelatinosus	1	0	0	0	0	X
Bacteria	Proteobacteria	Betaproteobacteria	Zoogloea oryzae	1	0	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. mpig8.9	0	1	0	0	0	
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderia sp. mpud5.2	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Arcobacter nitrofigilis DSM 7299	1	1	1	1	1	X
Bacteria	Proteobacteria	delta/epsilon subdivisions	delta proteobacterium MLMS-1	1	1	1	1	1	
Bacteria	Proteobacteria	delta/epsilon subdivisions	delta proteobacterium NaphS2	1	1	1	1	1	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfonatronospira thiodismutans ASO3-1	1	1	1	1	1	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfovibrio aespoeensis Aspo-2	1	1	1	1	1	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfovibrio fructosovorans JJ	1	1	1	1	1	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfovibrio sp. FW1012B	1	1	1	1	1	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfovibrio vulgaris RCH1	1	1	1	1	1	X
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfovibrio vulgaris subsp. vulgaris DP4	1	1	1	1	1	X
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfuromonas acetoxidans DSM 684	1	1	1	1	1	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geobacter sp. M18	1	1	1	1	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geobacter sulfurreducens KN400	1	1	1	1	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfobacter curvatus	1	0	0	0	0	X
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfobacter latus	1	0	0	0	0	X
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfomicrobium baculatum	1	0	0	0	0	X
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfomonile tiedjei	1	0	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfonema limicola str. Jadebusen	1	0	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfovibrio africanus	1	0	0	0	0	X
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfovibrio dechloracetivorans	1	0	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfovibrio gigas	1	0	0	0	0	X

Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfovibrio salexigens	1	0	0	0	0	X
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfovibrio vulgaris	1	0	0	0	0	X
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geoalkalibacter ferrihydriticus	1	0	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Sulfurospirillum multivorans	1	0	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfuromonas acetexigens	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfuromonas michiganensis	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfuromonas palmitatis	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfuromonas thiophila	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfuromusa bakii	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfuromusa kysingii	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfuromusa sp. S1	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Desulfuromusa succinoxidans	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geobacter bremensis	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geobacter chapelleii	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geobacter grbiciae	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geobacter humireducens	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geobacter hydrogenophilus	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geobacter pelophilus	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geobacter psychrophilus	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geobacter thiogenes	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geopsychrobacter electrodiphilus	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Geothermobacter ehrlichii	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Malonomonas rubra	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Pelobacter acetylenicus	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Pelobacter acidigallici	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Pelobacter masseliensis	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Pelobacter propionicus	0	1	0	0	0	
Bacteria	Proteobacteria	delta/epsilon subdivisions	Pelobacter venetianus	0	1	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Azotobacter vinelandii	1	1	1	1	1	X
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella pneumoniae	1	1	1	1	1	X
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. 1_1_55	1	1	1	1	1	
Bacteria	Proteobacteria	Gammaproteobacteria	Methylobacter tundripaludum SV96	1	1	1	1	1	
Bacteria	Proteobacteria	Gammaproteobacteria	Pantoea sp. At-9b	1	1	1	1	1	
Bacteria	Proteobacteria	Gammaproteobacteria	Acidithiobacillus ferrooxidans	1	1	1	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Azotobacter chroococcum	1	1	1	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Halorhodospira halophila	1	1	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Methylobacter luteus	1	1	0	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Methylococcus capsulatus	1	1	0	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonas azotifigens	1	1	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonas stutzeri	1	0	1	1	1	X
Bacteria	Proteobacteria	Gammaproteobacteria	Pantoea agglomerans	1	0	0	1	1	X
Bacteria	Proteobacteria	Gammaproteobacteria	Acidithiobacillus ferrivorans	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Aeromonas sp. gx-126	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Aeromonas sp. IPPW-29	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Aeromonas sp. IPPW-33a	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Alkalilimnicola halodurans	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Allochromatium minutissimum	1	0	0	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Azomonas agilis	1	0	0	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Azomonas macrocytogenes	1	0	0	0	0	X

Bacteria	Proteobacteria	Gammaproteobacteria	Brenneria salicis	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Celerinatantimonas diazotrophica	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Ectothiorhodospira haloalkaliphila ATCC 51935	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Ectothiorhodospira mobilis	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Ectothiorhodospira shaposhnikovii	1	0	0	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Ectothiorhodospira shaposhnikovii DSM 2111	1	0	0	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Ectothiorhodospira sp. B7-7	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Ectothiorhodospira variabilis	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacter arachidis	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacter cloacae	1	0	0	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacter sp. gx-25	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacter sp. MTP_050512_17	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacter sp. Y11	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacter sp. Y4	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacter sp. Y61	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacter sp. Y79	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Erwinia sp. gx104	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	gamma Proteobacterium BAL281	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	gamma Proteobacterium BAL286	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Halomonas maura	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Halorhodospira abdelmalekii	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Halorhodospira halochloris	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella oxytoca	1	0	0	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp.	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. AF-4C	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. AL050511_01	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. AL060224_03	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. AL060225_04	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. BM92	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. BN-4A	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. CRLI0713	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. CRLI0715	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. CRLI0718a	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. CRLI0728	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. CRLI0730	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. CRLS0617	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. CRLS064	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. CRLS069a	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. CRPV0610	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. CRPV0611a	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. ECI-10A	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. gx-3	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. JT42	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. MNFG_801b	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. MNFG_802	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. MNS_801a	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. MNS_801b	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. MNW_801b	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. NTI31	1	0	0	0	0	

Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. P0638	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. P0640	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. P0643	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. P0646	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. TT001	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. Y57	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella sp. Y83	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Klebsiella variicola	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Kluyvera ascorbata	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	mangrove root bacterium HS001	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Marichromatium purpuratum	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Bird Shoal 12-2'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Bird Shoal 1-31'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Bird Shoal 1-9'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Bird Shoal 4-6'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Bird Shoal 6-2'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Bird Shoal sa2'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Bird Shoal sa3'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Bird Shoal sa4'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Bird Shoal sw4'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Sippewissett 2'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Sippewissett 2-32'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Sippewissett 24'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Tomales Bay 13-32r'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Tomales Bay 14-24'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Tomales Bay 2-31'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Tomales Bay 5-24'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Tomales Bay med white'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Tomales Bay mg10'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Tomales Bay mg15'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Tomales Bay wc1-2 sm white'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Tomales Bay wc1-28'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Tomales Bay wc2-3 lg wh'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Tomales Bay wc2-3 sm wh'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Tomales Bay wh'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	marine proteobacterium 'Tomales Bay wh clear'	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Methylobacter bovis	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Methylobacter chroococcum	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Methylobacter marinus	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Methylobacter vinelandii	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Methylocaldum szegediense	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Methylocaldum szegediense O-12	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Methylomonas methanica	1	0	0	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Methylomonas rubra	1	0	0	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Methylomonas sp. LW13	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Methylomonas sp. LW15	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Methylosoma difficile	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Methylovulum miyakonense	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	nitrogen-fixing bacterium HS002	1	0	0	0	0	

Bacteria	Proteobacteria	Gammaproteobacteria	Pantoea sp. A0305	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pantoea sp. A0310	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pantoea sp. CRLI0712a	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pantoea sp. CRLI0721a	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pantoea sp. CRLI0724b	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pantoea sp. CRLI0725a	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pantoea sp. CRLS0621	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pantoea sp. CRLS062b	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pantoea sp. EC080527_02	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pantoea sp. gx-125	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pantoea sp. P0352	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pantoea sp. P0356	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pantoea sp. P0359	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonas balearica	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonas putida	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonas sp. AF-4B	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonas sp. DC	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonas sp. gx-127	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonas sp. IPPW-2	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonas sp. IPPW-3	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonas sp. K1-2004	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonas stutzeri A15	1	0	0	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Pseudomonas stutzeri CMT.9.A	1	0	0	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Raoultella ornithinolytica	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Raoultella terrigena	1	0	0	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Serratia marcescens	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Stenotrophomonas maltophilia	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Stenotrophomonas sp. gx-44	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Thermochromatium tepidum	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Thioalkalispira microaerophila	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Thiocapsa bogorovii	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Thiocapsa roseopersicina	1	0	0	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Thiorhodospira sibirica ATCC 700588	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Vibrio cincinnatiensis	1	0	0	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Vibrio diazotrophicus	1	0	0	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Vibrio mangrovi	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Vibrio natriegens	1	0	0	0	0	X
Bacteria	Proteobacteria	Gammaproteobacteria	Vibrio parahaemolyticus	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Vibrio porteresiae	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Vibrio sp. MSSRF39	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Vibrio sp. MSSRF60	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Xanthomonas translucens	1	0	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Azorhizophilus paspali	0	1	0	0	0	
Bacteria	Proteobacteria	Gammaproteobacteria	Azotobacter salinestris	0	0	0	1	0	X
Bacteria	Proteobacteria	unclassified Proteobacteria	marine proteobacterium L2-3	1	0	0	0	0	
Bacteria	Proteobacteria	unclassified Proteobacteria	marine proteobacterium WC1-2(LW)	1	0	0	0	0	
Bacteria	Proteobacteria	unclassified Proteobacteria	marine proteobacterium WC1-2(sm)	1	0	0	0	0	
Bacteria	Proteobacteria	unclassified Proteobacteria	bacterium DU1	0	1	0	0	0	
Bacteria	Spirochaetes	Spirochaetes	Spirochaeta aurantia	1	0	0	0	0	

Bacteria	Spirochaetes	Spirochaetes	Spirochaeta stenostrepta	1	0	0	0	0
Bacteria	Spirochaetes	Spirochaetes	Spirochaeta zuelzerae	1	0	0	0	0
Bacteria	Spirochaetes	Spirochaetes	Treponema azotonutricium	1	0	0	0	0
Bacteria	Spirochaetes	Spirochaetes	Treponema bryantii	1	0	0	0	0
Bacteria	Spirochaetes	Spirochaetes	Treponema denticola	1	0	0	0	0
Bacteria	Spirochaetes	Spirochaetes	Treponema pectinovorum	1	0	0	0	0
Bacteria	Spirochaetes	Spirochaetes	Treponema primitia ZAS-1	1	0	0	0	0
Bacteria	Spirochaetes	Spirochaetes	Treponema primitia ZAS-2	1	0	0	0	0
Bacteria	Spirochaetes	Spirochaetes	Treponema sp. AC3	1	0	0	0	0
Bacteria	Spirochaetes	Spirochaetes	Treponema sp. Ru1	1	0	0	0	0
Bacteria	Synergistetes	Synergistetes	Pyramidobacter piscolens W5455	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium S5	1	1	1	1	1
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen fixing bacterium ANFK33	1	0	0	1	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium CZ152S	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium DZY-HS14	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium DZY-N56	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium G25-1	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium G25-1-2	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium G33-1	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium GG165E	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium GG41E	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium GG42E	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium GG49E	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium GG50E	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium GG53E	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium HX148S	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium LA11E	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium LA4E	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium LZ83E	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium LZ84E	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium NN143E	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium NN144E	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium NN145S	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium NN208E	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium PG132S	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium Q25-2	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium QZ25S	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium QZ33S	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium QZ80E	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium SS82E	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium Y41	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium YL34S	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	cyanobacteria-associated bacterium #2	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	cyanobacteria-associated bacterium #3	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	lactate SRB-Enrichment culture clone HBLac1	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen fixing bacterium NKNRT11	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen fixing bacterium NKNRT16	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen fixing bacterium NKNRT2-2	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen fixing bacterium NKNRT23	1	0	0	0	0

Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen fixing bacterium NKNRT26	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen fixing bacterium NKNRT27	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen fixing bacterium NKNRT2-7	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen fixing bacterium NKNRT2-8	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen fixing bacterium NKNRT6	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen-fixing bacterium JC110	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen-fixing bacterium KNUC115	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen-fixing bacterium SC16	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen-fixing bacterium SG21	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen-fixing bacterium TS210	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	temperate forest soil bacterium AC06	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	temperate forest soil bacterium YC07	1	0	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	bacterium TP1	0	1	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen-fixing bacterium SM1	0	1	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen-fixing bacterium SM2	0	1	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen-fixing bacterium WA1	0	1	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen-fixing bacterium WB2	0	1	0	0	0
Bacteria	unclassified Bacteria	unclassified Bacteria	nitrogen-fixing bacterium AN1	0	0	0	1	0
Eukaryota	Amoebozoa	Amoebozoa	Polysphondylium pallidum PN500	1	0	0	0	0
Eukaryota	Bacillariophyta	Bacillariophyta	Rhopalodia gibba	0	1	0	0	0
Eukaryota	Viridiplantae	Viridiplantae	Volvox carteri f. nagariensis	1	0	0	0	0

APÊNDICE 11 – QUADRO CONTENDO AS MEDIDAS DE TOMADAS DE TEMPO PARA CADA UM DOS PROCESSOS REALIZADOS NAS AFERIÇÕES REALIZADAS

	Data	Step01	Step02	Step03	Step04	Step05	Step06	Step07	Step08	Step09	Total
Actinobacteria	25/05/2010	00:15:07	00:01:00	00:02:00	00:04:00	00:12:00					00:34:07
	25/08/2010	00:10:00	00:00:30	00:01:25	00:02:34	00:08:04					00:22:33
	25/11/2010	00:06:56	00:00:02	00:00:17	00:00:47	00:05:34					00:13:36
Archaea	25/05/2010	00:14:36	00:01:00	00:04:00	00:04:00	00:15:00					00:38:36
	25/08/2010	00:08:45	00:00:46	00:01:56	00:02:34	00:11:24					00:25:25
	25/11/2010	00:05:29	00:00:03	00:00:54	00:01:47	00:09:58					00:18:11
Chlorobium	25/05/2010	00:13:07	00:01:00	00:01:00	00:02:00	00:17:00					00:34:07
	25/08/2010	00:10:42	00:00:35	00:00:24	00:02:00	00:12:52					00:26:33
	25/11/2010	00:09:59	00:00:03	00:00:04	00:01:40	00:10:07					00:21:53
Cyanobacteria	25/05/2010	00:16:58	00:01:00	00:01:00	00:03:00	00:14:00					00:35:58
	25/08/2010	00:10:21	00:00:36	00:00:21	00:02:45	00:10:47					00:24:50
	25/11/2010	00:06:01	00:00:03	00:00:01	00:01:32	00:08:41					00:16:18
Firmicutes	25/05/2010	00:16:54	00:01:00	00:01:00	00:05:00	00:15:00					00:38:54
	25/08/2010	00:12:28	00:00:25	00:00:32	00:02:58	00:09:57					00:26:20
	25/11/2010	00:10:58	00:00:03	00:00:03	00:01:30	00:08:59					00:21:33
Alphaproteobacteria	25/05/2010	00:17:56	00:01:00	00:01:00	00:03:00	00:15:00					00:37:56
	25/08/2010	00:08:23	00:00:41	00:00:29	00:01:56	00:09:49					00:21:18
	25/11/2010	00:03:49	00:00:02	00:00:01	00:00:52	00:05:38					00:10:22
Betaproteobacteria	25/05/2010	00:19:14	00:01:00	00:01:00	00:05:00	00:21:00					00:47:14
	25/08/2010	00:15:36	00:00:25	00:00:12	00:03:25	00:16:01					00:35:39
	25/11/2010	00:10:34	00:00:03	00:00:01	00:01:35	00:08:43					00:20:56
DeltaEpsilon	25/05/2010	00:14:02	00:01:00	00:01:00	00:05:00	00:12:00					00:33:02
	25/08/2010	00:10:47	00:00:23	00:00:42	00:02:15	00:09:14					00:23:21
	25/11/2010	00:07:52	00:00:03	00:00:03	00:01:33	00:08:49					00:18:20
Gamaproteobacteria	25/05/2010	00:16:23	00:01:00	00:01:00	00:08:00	00:18:00					00:44:23
	25/08/2010	00:12:45	00:00:23	00:00:55	00:05:09	00:14:28					00:33:40
	25/11/2010	00:08:10	00:00:03	00:00:55	00:01:28	00:10:02					00:20:38
Random	25/05/2010	00:15:26	00:01:00	00:01:00	00:04:00	00:15:00					00:36:26
	25/08/2010	00:12:47	00:00:27	00:00:15	00:01:12	00:10:23					00:25:04
	25/11/2010	00:07:27	00:00:04	00:00:01	00:00:49	00:05:32					00:13:53
Union	25/05/2010						00:03:08	00:06:54	00:05:45	18:36:45	18:52:32
	25/08/2010						00:01:14	00:04:27	00:02:04	11:24:36	11:32:21
	25/11/2010						00:00:30	00:00:10	00:00:02	05:37:30	05:38:12
Total	25/05/2010										25:13:15
	25/08/2010										15:57:04
	25/11/2010										8:33:52

APÊNDICE 12 – SCRIPTS EM MATLAB PARA TREINAMENTO E TESTE DE REDE MLP.

```
function mret = mlp_tr(w,epc)
m = length(w(1,:));
P = double(w(:,1:m-1))';
T = double(w(:,m)');
rand('seed', 672880951);
net = newff(P,T,epc);          % create neural network
net.trainParam.epochs = 1500;
%net.trainParam.show = 'off';
[net,tr] = train(net,P,T); % trains
mret = net;



function mret = mlp_ts(w,net)
m = length(w(1,:));
n  = length(w(:,1));
P = double(w(:,1:m-1))';
T = double(w(:,m)');
nc = max(T);
testInputs = P;
testTargets = T;
out0 = sim(net,testInputs);
out = round(out0);
xout = out;% Get response of trained networ
if min(T) == 1
    T = T-1;
    out = out - 1;
    nc = nc-1;
end
out(find(out>nc)) = nc;
out(find(out<0)) = 0;
conf = zeros(nc+1,nc+1);
for i=1:n
    conf(T(i)+1,out(i)+1) = conf(T(i)+1,out(i)+1)+1;
end
mret.conf = conf;
r = [];
for i=1:(nc+1)
    u = conf(i,i)/sum(conf(:,i));
    r = [r u];
end
certos = 100*(sum(out==T))/n;
mret.certos = certos;
mret.prop = r;
mret.Result = xout;
mret.Res0 = out0;
```

ANEXOS

ANEXO 1 – LISTAGEM ORIGINAL, FORMULADA POR YOUNG (YOUNG, 1992) CONTENDO TODOS OS CONHECIDOS FIXADORES DE NITROGÊNIO DESCRITOS NA LITERATURA . FONTE: YOUNG, (1992).

Table 1.1 The Known Nitrogen-Fixing Organisms Arranged in Phylogenetic Groups

Genus	Species	Literature References[a] Name	Phylogeny	Fixation
Green sulfur bacteria				
Chlorobium	*limicola* (also *limicola* f. sp. *thiosufatophilum*)	120	139	120
	phaeobacteroides	120	129	94
	vibrioforme	120	139	120
Chloroherpeton	*thalassium*	120	139	120
Pelodictyon	*luteolum*	120	120	50,95
Prosthecochloris	*aestuarii*	120	43	122
Firmibacteria				
Bacillus	*azotofixans*	111	111	111
	macerans	114	114	114
	polymyxa	114	114	114
Clostridium	*aceticum*	114	139	107
	acetobutylicum	114		107
	arcticum	114		114
	beijerinckii	114		107
	butylicum	114		107
	butyricum	114	139	114
	felsineum	114		107
	formiaceticum	114		114
	kluyveri	114		107
	"lactoacetophilum"[b]			107
	"madisoni"			107
	pasteurianum	114	139	114
	"pectinovorum"			107
	"saccharobutyricum"			95
	"tetanomorphum"	114		107
	tyrobutyricum	114		95
Desulfotomaculum	*orientis*	114	28	93
	nigrificans	114	28	84
	ruminis	114	28	93
Thallobacteria				
Arthrobacter	sp. *"fluorescens"*		117	18,19
Frankia	spp.	138	48	138
Streptomyces	spp.		117	32,63
Propionibacterium	*freudenreichii*	114	48	8
	jensenii (*peterssonii* is now *jensenii*) (*shermanii* is now *freudenreichii*)	114		8
Heliobacteria				
"Heliobacillus	*mobilis"*	11	11	11
Heliobacterium	*chlorum*	120	139	120
"Heliospirillum	*gestii"*	11		50
Cyanobacteria		120	139,46	120,104[c]
Cyanothece group	(Section I)			
Gloeocapsa group	(Section I)			
Gloeothece	(Section I)			

Table 1.1 Continued

Genus	Species	Literature References[a]		
		Name	Phylogeny	Fixation
Synechococcus group	(Section I)			
Synechocystis group	(Section I)			
Chroococcidiopsis	(Section II)			
Dermocarpa	(Section II)			
Myxosarcina	(Section II)			
Pleurocapsa group	(Section II)			
Xenococcus	(Section II)			
Lyngbya	(Section III)			
Oscillatoria	(Section III)			
Pseudanabaena	(Section III)			
Spirulina	(Section III)			
(*Trichodesmium* included in *Oscillatoria*)				
Anabaena	(Section IV)			
Aphanizomenon	(Section IV)			
Calothrix	(Section IV)			
Cylindrospermum	(Section IV)			
Nodularia	(Section IV)			
Nostoc	(Section IV)			
Scytonema	(Section IV)			
Chlorogloeopsis	(Section V)			
Fischerella	(Section V)			
Geitleria	(Section V)			
Stigonema	(Section V)			
Prochloron	*didemni*	120	127	120
Campylobacter				
Campylobacter	*nitrofigilis*	79	124	79
Proteobacteria: alpha subdivision				
Acetobacter	*diazotrophicus*	45	45	45
Agrobacterium	*tumefaciens*	67,27	134,140	59,60
Ancylobacter	*aquaticus*	101	17	77
[*Aquaspirillum*][c]	*fasciculus*[d]	67		67
	itersonii	67	139	67
	magnetotacticum[d]	120		120
	peregrinum[d]	67		67
Azorhizobium	*caulinodans*	34	34	34
Azospirillum	*amazonense*	38	38	38
	brasilense	67	140	67
	halopraeferens	103	103	103
	lipoferum	67	103	67
Beijerinckia	*derxii*	67		67
	fluminensis	67		67
	indica	67	34	67
	mobilis	67		67
Bradyrhizobium	*japonicum*	58	51,55	67
	sp. (other hosts)	67	51,55	67
"[*Chromobacterium*]	*folium*"	44	44	14
"*Methylocystis*	*echinoides*"[d]	120		120
	"*parvus*"		126	106,83

Table 1.1 Continued

Genus	Species	Literature References[a]		
		Name	Phylogeny	Fixation
"Methylosinus	*sporium"*			83
	"trichosporium"	67	126	83
(Microcyclus aquaticus and *M. eburneus* are now *Ancylobacter aquaticus)*				
Mycoplana	*bullata*	114	27	90a
	dimorpha	114	27	90a
"Photorhizobium	*thompsonum"*	35a,37	144a	35a,37
"[Pseudomonas]	*azotocolligans"*	30	30	30
[Pseudomonas]	*diazotrophicus*	131	136	131
[Pseudomonas]	*paucimobilis*	30	30	9
"Renobacter	*vacuolatum"*	72	17	77
Rhizobium	*fredii*	109	55,133	109
	galegae	74	55	74
	leguminosarum	67	140	67
	loti	56	55	56
	meliloti	67	51	67
Rhodobacter	*adriaticus*	53,120		120
	capsulatus	53,120	140	76
	sphaeroides	53,120	140	76
	sulfidophilus	53,120		76
	veldkampii	120		120
Rhodomicrobium	*vannielii*	120	140	76
Rhodopila	*globiformis*	53	140	76
Rhodopseudomonas	*acidophila*	120	140	76
	blastica	120	120	76
	marina	120,78		78
	palustris	120	140	76
	rutila	120	120	120
	sulfoviridis	120	120	76
	viridis	120	140	76
Rhodospirillum	*fulvum*	53	126	76
	molischianum	53	140	76
	photometricum	53	140	76
	rubrum	53	140	76
(Sinorhizobium see *Rhizobium)*		23		
Xanthobacter	*agilis*	57,5		57
	autotrophicus	67		67
	flavus	67	34	67
Proteobacteria: beta subdivision				
Alcaligenes	*faecalis*	67	67	21
	latus	67	67	16
	paradoxus	67	67	65
"Azoarcus"	spp.	103a	103a	103a
Derxia	*gummosa*	67	29	67
Herbaspirillum	*seropedicae*	7	103	7
"Lignobacter"	sp.		27	25
[Pseudomonas]	*saccharophila*	67	67,30	10

Table 1.1 Continued

Genus	Species	Literature References[a]		
		Name	Phylogeny	Fixation
Rhodocyclus	*gelatinosus*	53,120	142	76
	tenuis	53,120	142	76
Thiobacillus	*ferrooxidans*	120	71	75,100
Proteobacteria: gamma subdivision				
Amoebobacter	*roseus*	120		95
Azomonas	*agilis*	67	26	67
	insignis	67	26	67
	macrocytogenes	67	26	67
Azotobacter	*armenaicus*	67	26	67
	beijerinckii	67	26	67
	chroococcum	67	26	67
	nigricans	67	26	67
	paspali	67	26	67
	vinelandii	67	26	67
Beggiatoa	*alba*	120	119	120
Chromatium	*gracile*	120		120
	minus	120		120
	minutissimum	120		120
	vinosum	120	141	120
	violascens	120		120
	warmingii	120	141	120
	weissei	120	141	120
Citrobacter	*freundii*	67	2	67
Ectothiorhodospira	*shaposhnikovii*	120	141	120
	vacuolata	120		120
Enterobacter	*aerogenes*	67	67	67
	agglomerans (Pantoea?)	67,42	67,42	99
	cloacae	67	67	67
Erwinia	*herbicola (Pantoea?)*	67,42	67,42	89
(Escherichia intermedia is now *Citrobacter freundii)*				
Klebsiella	*pneumoniae*	67	67	
	oxytoca	67	67	62,67
	planticola	67	67	
	terrigena	67	67	
Lamprobacter	*modestohalophilus*	120		120
"Methylobacter	*capsulatus" Y*[d]	67	17	86[e]
Methylococcus	*capsulatus*	67	126	106,83
	luteus	67		106
	thermophilus	67		106
	ucrainicus	67		106
Methylomonas	*methanica*	67	126	106
	"rubra"			106
(Pantoea see *Erwinia* and *Enterobacter)*		42		
Pseudomonas	*stutzeri*	67	67	9
Thiocapsa	*pfennigii*	120	141	120
	roseopersicina	120	141	120

Table 1.1 Continued

Genus	Species	Literature References[a]		
		Name	Phylogeny	Fixation
Thiocystis	*violacea*	120	139	120
Vibrio	*cincinnatiensis*	4		128
	diazotrophicus	67		67
	natriegens	67	67	135
	pelagius	67		128
Proteobacteria: delta subdivision				
Desulfobacter	*curvatus*	137,6	28	137
	hydrogenophilus	137,6	28	137
	latus	137,6	28	137
Desulfovibrio	*africanus*	67		84,98
	baculatus	67		84,98
	desulfuricans	67	28	84,98
	gigas	67	28	98
	salexigens	67	28	98
	thermophilus	67		84,98
	vulgaris	67	28	98
Archaeobacteria				
Halobacterium	*halobium*	120	139	92[e]
Methanobacterium	*formicicum*	120		92[e]
	ivanovii	5		92,115
	thermautotrophicum	120		92[e]
Methanococcus	*aeolicus*	120		120
	maripaludis	120	139,120	120
	thermolithotrophicus	120	139,120	13,120,115
	vannielii	120		92[e]
	voltae	120	139,120	116
Methanolobus	*tindarius*	120		64
Methanoplanus	sp.	120		92[e]
Methanosarcina	*barkeri*	120	139,120	80,120
Methanothermus	*"facilis"*			92[e]
	fervidus	120		92[e]
Unknown phylogenetic position				
Propionispira	*arboris*	108,3		108

[a]Three types of references are given: for identification and naming, for phylogenetic position (preferably based on 16S rRNA), and for evidence that the organism fixes nitrogen.

[b]" " indicates that the name has not been validly published; [ ] indicates that the organism is not related to the type species of the genus to which it is assigned.

[c]cyanobacteria in Sections IV and V have heterocysts and are therefore assumed to fix nitrogen, but this has not always been examined experimentally.

[d]The phylogenetic position of this organism is uncertain; it is listed, for convenience, with supposed relatives that belong in this group.

[e]DNA from this species hybridizes with *nif* gene probes, but there is as yet no direct evidence for nitrogen fixation.

ANEXO 2 – EXEMPLO DE ARQUIVO TEXTO SIMPLES RETORNADO PELA FUNÇÃO NCBIWWW.qblast(). TRECHO EXTRAÍDO DO ARQUIVO nifN.txt.

Diretório : ..\BDActinobacteria

Arquivo : nifN.txt

Tamanho : 1349 KB

```
<p><!--
QBlastInfoBegin
            Status=READY
QBlastInfoEnd
--><p>
<PRE>
BLASTP 2.2.24+
Reference: Stephen F. Altschul, Thomas L. Madden, Alejandro
A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and
David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new
generation of protein database search programs", Nucleic
Acids Res. 25:3389-3402.


Reference for compositional score matrix adjustment: Stephen
F. Altschul, John C. Wootton, E. Michael Gertz, Richa
Agarwala, Aleksandr Morgulis, Alejandro A. Schaffer, and
Yi-Kuo Yu (2005) "Protein database searches using
compositionally adjusted substitution matrices", FEBS J.
272:5101-5109.

RID: EMTCBBS501N

Database: All non-redundant GenBank CDS
translations+PDB+SwissProt+PIR+PRF excluding environmental samples
from WGS projects
           12,310,662 sequences; 4,207,507,770 total letters
Query=
Length=489
                                                                    Score     E
Sequences producing significant alignments:                        (Bits)  Value

ref|ZP_06475517.1|  nitrogenase molybdenum-iron cofactor biosy...   448    8e-124
gb|AAD17263.1|  NifN [Frankia sp. EuIK1]                            422    4e-116
ref|YP_483559.1|  nitrogenase molybdenum-iron cofactor biosynt...   371    1e-100


ALIGNMENTS
>ref|ZP_06475517.1| nitrogenase molybdenum-iron cofactor biosynthesis protein NifN
[Frankia symbiont of Datisca glomerata]
 gb|EFD27772.1| nitrogenase molybdenum-iron cofactor biosynthesis protein NifN
[Frankia symbiont of Datisca glomerata]
Length=463

 Score =  448 bits (1152),  Expect = 8e-124, Method: Compositional matrix adjust.
 Identities = 259/468 (56%), Positives = 322/468 (69%), Gaps = 32/468 (6%)

Query  1    MARVVTSDRRPGLDPLRFSQPLGGALVFLGLAAAMPVMHGSKGCASFAKALLTRHFNEPV  60
            MARV T   R   DPL+ SQPLGGAL FLGLA  +P++HG++GC++FAKALLTRHF EP+
Sbjct  1    MARVTTGGHRAAFDPLKHSQPLGGALAFLGLARCIPMLHGAQGCSAFAKALLTRHFREPI  60

Query  61   PLQTTGVTEVSAVLGSGDDLVANLDGIRAKQNPRIIGLLTTGVTEVSGEDVAGQVRQYIA  120
            PLQ++ + +VSAVLGS + L+  LD I  +Q P IIG+LTTG+TEV+ ED+ G     +
Sbjct  61   PLQSSAINDVSAVLGSSESLLGALDTINERQRPDIIGVLTTGLTEVTEEDMFG----VLG  116

Query  121  MMNHTTPEGAPLIVRVSTPDFAGGLSDGWSAALRSLVATVPFDHADSDEYPGTRSGFGAG  180
               +      PL+V VSTPDF GGLSDGWSAALR +VA VP
Sbjct  117  ASKYAPGSRGPLVVAVSTPDFHGGLSDGWSAALRGIVAAVP------------------A  158

Query  181  TG-SAPETVAVLVGPSLSAADLDELCALIRSFGMAPVLVPDLSGSLDGHLAPSWQPTTTG  239
            TG + P   AVLVGP+L+A D+DE+  L+R+FG+ PV+VPDLSGSLDGHLA +W P TTG
Sbjct  159  TGVTVPGRAAVLVGPTLTAVDIDEIADLVRAFGLDPVVVPDLSGSLDGHLAANWSPVTTG  218
```

```
Query  240  GTGLAQLRRLDEAGLIITAGATAAEAGVDLAARTAADLVQHDHLSGLAAVDSLVAELMTR  299
            GT L  L  L    +++  G+ AA A  +LA +  A L+ HDHLSGLA  D LV +L+
Sbjct  219  GTTLDALLGLAGCEVVVAVGSAAATAAGELAVKAGAPLIAHDHLSGLAVTDLLVTDLIRH  278

Query  300  SGRGPAPEVRRARARLADGLLDTHFVLGGARIALAMEPEALVAVGSLLHDVGAEIVAAVS  359
            +G      VRR R RLADGL+D+HFVLGGA++ALA+EP+ LVAVGSL +DVGAEIV AVS
Sbjct  279  TGATAPEAVRRWRRRLADGLMDSHFVLGGAKVALALEPDQLVAVGSLFYDVGAEIVTAVS  338

Query  360  PTDAPVLATAPWDEIVIGDLTDLEERALEGGAELLIGSSHVRTVADRIGAAHLAVGFPIY  419
            PT+A VL  APW+E+V+GD  DL ERA   GAEL++ SSH    A   GAAHL +GFP+Y
Sbjct  339  PTNADVLRQAPWEEVVVGDFVDLAERAAAAGAELVVASSHGADAAAEAGAAHLTMGFPVY  398

Query  420  DRLGSALRTTAGYGGSLRLLVDAANRLLDHHQADHQANHRADHRPGRH  467
            DR+G+ LR T+GY GSL LLVDAANRLL+         HR  HRP RH
Sbjct  399  DRMGAQLRATSGYRGSLHLLVDAANRLLE---------HRESHRPSRH  437


>gb|AAD17263.1| NifN [Frankia sp. EuIK1]
Length=445

 Score =  422 bits (1086),  Expect = 4e-116, Method: Compositional matrix adjust.
 Identities = 289/494 (59%), Positives = 312/494 (64%), Gaps = 54/494 (10%)

Query  1    MARVVTSDRRPGLDPLRFSQPLGGALVFLGLAAAMPVMHGSKGCASFAKALLTRHFNEPV  60
            MARVVTSD RPGLDPLRFS  LGG LVFL LA    VMHGS+GCAS AKALLTRHFNEPV
Sbjct  1    MARVVTSDTRPGLDPLRFSLALGGRLVFLRLAQRDAVMHGSQGCASLAKALLTRHFNEPV  60

Query  61   PLQTTGVTEVSAVLGSGDDLVANLDGIRAKQNPRIIGLLTTGVTEVSGEDVAGQVRQYIA  120
            PLQTT    V AVLGSGDD        R ++ P            V  +D   Q
Sbjct  61   PLQTTCHRGV-AVLGSGDD--------RGRRRP---------GPPVHRDDEPHQ------  96

Query  121  MMNHTTPEGAPLIVRVSTPDFAGGLSDGWSAALRSLVATVPFDHADSDEYP-GTRSGFGA  179
                  PEGAPLIVRVSTP FAGGLSDGWSAA     ATVPFDHADSDEYP   +   G+
Sbjct  97   ------PEGAPLIVRVSTPPFAGGLSDGWSAACARWSATVPFDHADSDEYPVRLQDRLGS  150

Query  180  GTGSAPETVAVLVGPSLSAADLDELCALIRSFGMAPVLVPDLSGSLDGHLAPSWQPTTTG  239
            G   +P + A      LSAADLDELCALIRSFGMAPVLVPDLSGS      P   PTTT
Sbjct  151  GPRRSPCSSA-----RLSAADLDELCALIRSFGMAPVLVPDLSGSSTAP-GPVLAPTTTV  204

Query  240  GTGLAQLRRLDEAGLIITAGATAAEAGVDLAARTAADLVQHDHLSGLAAVDSLVAELMTR  299
              G ++   LDEAGLIITA       G     R           S LAA+D LV EL+
Sbjct  205  APG-SRSCALDEAGLIITARRPPRRPGSTWP-RAPPPTSSSTTTSRLAAMDRLVTELIDG  262

Query  300  SGRGPAPEVRRARARLADGLLDTHFVLGGARIALAMEPEALVAVGSLLHDVGAEIVAAVS  359
                 APEVR   ARLA GLLDTHFVLGGAR+ALAMEPEALVAVGSLLHDVG      VS
Sbjct  263  LLGRAAPEVRPP-ARLAHGLLDTHFVLGGARVALAMEPEALVAVGSLLHDVGGGDRRPVS  321

Query  360  PTDAPVLATAPWDEIVIGDLTDLEERALEGGAELLIGSSHVR--TVADRIGAAHLAVGFP  417
            PTDAPVLATAPWDEIV   LT+LE+       ELLIGSSHVR  T + R    H   G
Sbjct  322  PTDAPVLATAPWDEIVDRHLTELEDL-----PELLIGSSHVRGATASRR----HCRSGSD  372

Query  418  IYDRLGSALRTTAGYGGSLRLLVDAANRLLD--HHQADHQANHRADHRPGRHDVREHPLD  475
            +  R  +ALRTT GYG  LRLLVDAANRLLD  HH +DH+ +  +  R G  D  + PL
Sbjct  373  LR-RAPAALRTTGGYGAGLRLLVDAANRLLDHQHHHSDHRPDLPSGSRAGHEDAAQPPLA  431

Query  476  SFDQLDVLCQESPC  489
             FDQLDVLCQESPC
Sbjct  432  PFDQLDVLCQESPC  445


>ref|YP_483559.1| nitrogenase molybdenum-iron cofactor biosynthesis protein NifN
[Frankia sp. CcI3]
 gb|ABD13830.1| nitrogenase molybdenum-iron cofactor biosynthesis protein NifN
[Frankia sp. CcI3]
Length=575

 Score =  371 bits (953),  Expect = 1e-100, Method: Compositional matrix adjust.
 Identities = 213/313 (69%), Positives = 245/313 (79%), Gaps = 8/313 (2%)

Query  185  PETVAVLVGPSLSAADLDELCALIRSFGMAPVLVPDLSGSLDGHLAPSWQPTTTGGTGLA  244
            P  +AVLVGPSL+AADLDEL  LIR+FG+ PVLVPDLSGS+DGHLAP+WQPTTTGGTGLA
Sbjct  263  PRQLAVLVGPSLTAADLDELGELIRAFGLDPVLVPDLSGSVDGHLAPAWQPTTTGGTGLA  322

Query  245  QLRRLDEAGLIITAGATAAEAGVDLAARTAADLVQHDHLSGLAAVDSLVAELMTRSGRGP  304
            +LR L  +  ++ AGATAA AG  LAART A +++H HLSGL  +D+LV EL+  +G
Sbjct  323  RLRALGRSRAVLVAGATAAAAGDLLAARTGARILRHRHLSGLTEMDTLVTELIELTGASA  382
```

```
Query  305  APEVRRARARLADGLLDTHFVLGGARIALAMEPEALVAVGSLLHDVGAEIVAAVSPTDAP  364
               VR+ARARLADGLLDTHFVLGGAR+ALAMEPE LVAVGSLLHDVGAE+VAAVSPT AP
Sbjct  383  PARVRQARARLADGLLDTHFVLGGARVALAMEPETLVAVGSLLHDVGAEVVAAVSPTAAP  442

Query  365  VLATAPWDEIVIGDLTDLEERALEGGAELLIGSSHVRTVADRIGAAHLAVGFPIYDRLGS  424
            VLA APWDE+V+GDLTDL ERA  GGA L++GSSH R VADRIGAAHL VGFPI+DRLG+
Sbjct  443  VLADAPWDEVVVGDLTDLAERARAGGAHLVLGSSHAREVADRIGAAHLPVGFPIFDRLGA  502

Query  425  ALRTTAGYGGSLRLLVDAANRLLDHHQADHQANHR-----ADHRPG-RHDVREHP--LDS  476
            AL  TAGY GSLRLL+DAANRLLDH     ++  R     AD +P  R ++   P
Sbjct  503  ALAGTAGYAGSLRLLIDAANRLLDHEHVHRRSGRRFSVSGADGQPAHRTEIPTQPGSAGE  562

Query  477  FDQLDVLCQESPC  489
             DQLD L QESPC
Sbjct  563  SDQLDNLFQESPC  575


 Score =  235 bits (599),  Expect = 1e-59, Method: Compositional matrix adjust.
 Identities = 125/204 (62%), Positives = 151/204 (75%), Gaps = 19/204 (9%)

Query  1    MARVVTSDRRPGLDPLRFSQPLGGALVFLGLAAAMPVMHGSKGCASFAKALLTRHFNEPV  60
            MA +VT +R+  +DPL+ SQPLGGALVFLGLA ++P+MHG++GCASFAKALLTRHFNEP+
Sbjct  1    MAEIVTGERQATIDPLKHSQPLGGALVFLGLAGSLPIMHGAQGCASFAKALLTRHFNEPI  60

Query  61   PLQTTGVTEVSAVLGSGDDLVANLDGIRAKQNPRIIGLLTTGVTEVSGEDVAGQVRQYIA  120
            PLQTT +TEV+AVLGSG+ +V  LD IR KQ P IIGLLTTGVTEVSGEDV GQ+R+Y++
Sbjct  61   PLQTTAITEVTAVLGSGEAMVETLDAIRKKQRPEIIGLLTTGVTEVSGEDVGGQLRKYLS  120

Query  121  MMNHTTPEG------------APLIVRVSTPDFAGGLSDGWSAALRSLV-ATVPFDHADS  167
              N T   G             APLIV VSTPDF GGLSDGWSAAL +LV A VP +H+++
Sbjct  121  AWNETDSSGPEAGANGSAPGRAPLIVGVSTPDFIGGLSDGWSAALEALVRAVVPAEHSET  180

Query  168  D----EYPGTRSGFGAGTGSAPET  187
            +     Y   R+ F A   SAP T
Sbjct  181  EPSEARYFTGRTAFIA--ASAPVT  202


  Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF
excluding environmental samples from WGS projects
    Posted date:  Nov 18, 2010  5:42 PM
  Number of letters in database: -87,459,522
  Number of sequences in database:  12,310,662

Lambda     K      H
   0.318    0.135    0.395
Gapped
Lambda     K      H
   0.267   0.0410    0.140
Matrix: BLOSUM62
Gap Penalties: Existence: 11, Extension: 1
Number of Sequences: 12310662
Number of Hits to DB: 209585443
Number of extensions: 9431180
Number of successful extensions: 27634
Number of sequences better than 100: 89
Number of HSP's better than 100 without gapping: 0
Number of HSP's gapped: 27563
Number of HSP's successfully gapped: 93
Length of query: 489
Length of database: 4207507770
Length adjustment: 142
Effective length of query: 347
Effective length of database: 2459393766
Effective search space: 853409636802
Effective search space used: 853409636802
T: 11
A: 40
X1: 16 (7.3 bits)
X2: 38 (14.6 bits)
X3: 64 (24.7 bits)
S1: 41 (20.4 bits)
S2: 74 (33.1 bits)
```