

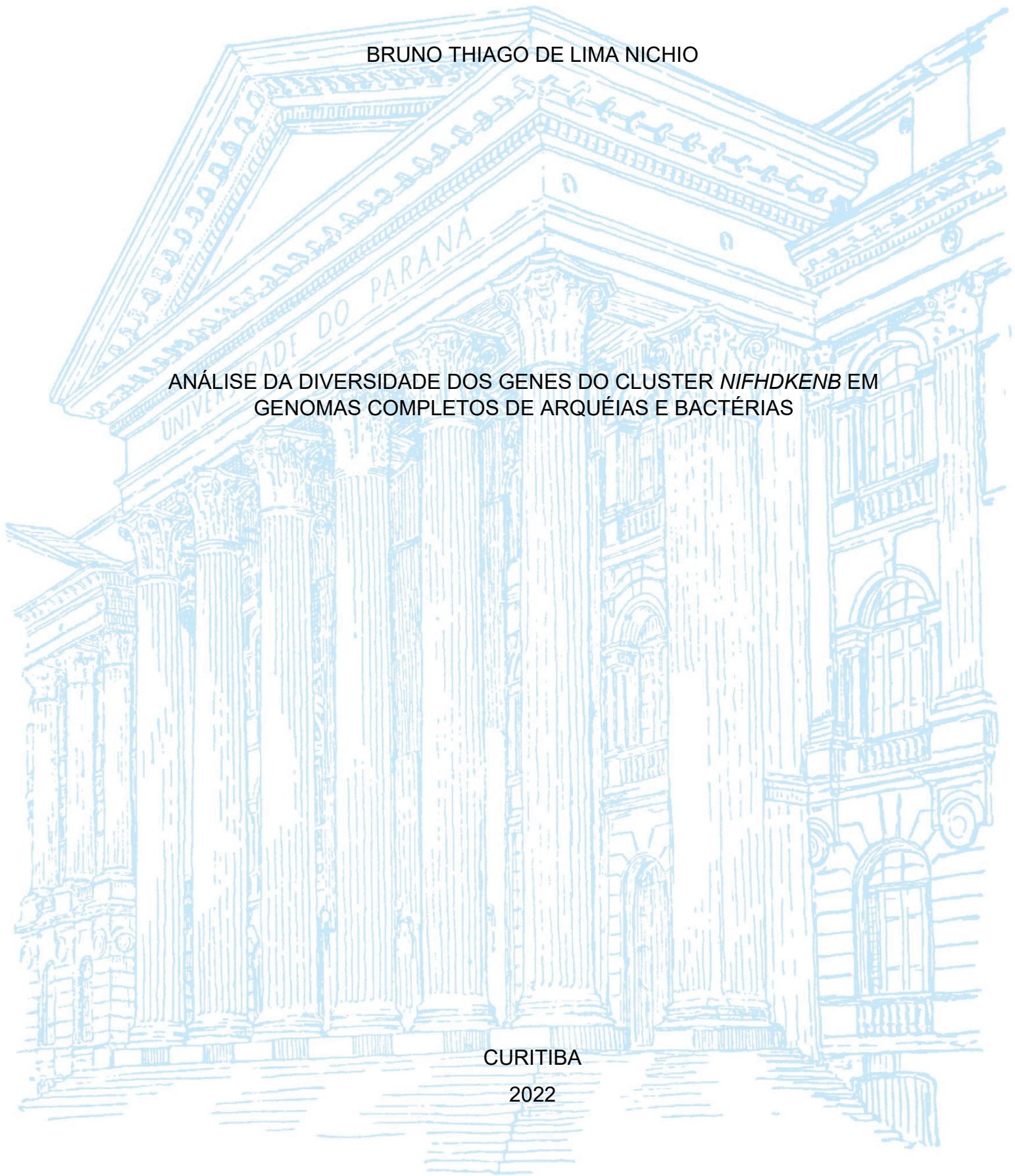
UNIVERSIDADE FEDERAL DO PARANÁ

BRUNO THIAGO DE LIMA NICHIO

ANÁLISE DA DIVERSIDADE DOS GENES DO CLUSTER *NIFHDKENB* EM  
GENOMAS COMPLETOS DE ARQUÉIAS E BACTÉRIAS

CURITIBA

2022



BRUNO THIAGO DE LIMA NICHIO

ANÁLISE DA DIVERSIDADE DOS GENES DO CLUSTER *NIFHDKENB* EM  
GENOMAS COMPLETOS DE ARQUÉIAS E BACTÉRIAS

Tese de doutorado apresentada ao Programa de Pós-Graduação em Ciências (Bioquímica), Setor de Ciências Biológicas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de doutor em Ciências (Bioquímica).

Orientador: prof. Dr. Fábio de Oliveira Pedrosa  
Coorientador: prof. Dr. Roberto Tadeu Raitz

CURITIBA

2022

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)  
UNIVERSIDADE FEDERAL DO PARANÁ  
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIAS BIOLÓGICAS

Nichio, Bruno Thiago de Lima.

Análise da diversidade dos genes do cluster *Nifhdkenb* em genomas completos de arqueias e bactérias. / Bruno Thiago de Lima Nichio. – Curitiba, 2022.

1 recurso on-line : PDF.

Orientador: Fábio de Oliveira Pedrosa.

Coorientador: Roberto Tadeu Raittz.

Tese (Doutorado) – Universidade Federal do Paraná, Setor de Ciências Biológicas. Programa de Pós-Graduação em Ciências – Bioquímica.

1. Bactérias Diazotróficas. 2. Nitrogenase. 3. Inteligência artificial. 4. Archaea. I. Título. II. Pedrosa, Fábio de Oliveira. III. Raittz, Roberto Tadeu. IV. Universidade Federal do Paraná. Setor de Ciências Biológicas. Programa de Pós-Graduação em Ciências - Bioquímica.



MINISTÉRIO DA EDUCAÇÃO  
SETOR DE CIÊNCIAS BIOLÓGICAS  
UNIVERSIDADE FEDERAL DO PARANÁ  
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO  
PROGRAMA DE PÓS-GRADUAÇÃO CIÊNCIAS  
(BIOQUÍMICA) - 40001016003P2

## TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação CIÊNCIAS (BIOQUÍMICA) da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **BRUNO THIAGO DE LIMA NICHIO** intitulada: **ANÁLISE DA DIVERSIDADE DOS GENES DO CLUSTER NIFHDKENB EM GENOMAS COMPLETOS DE ARQUÉIAS E BACTÉRIAS**, sob orientação do Prof. Dr. FABIO DE OLIVEIRA PEDROSA, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 29 de Março de 2022.

Assinatura Eletrônica

12/05/2022 18:44:00.0

FABIO DE OLIVEIRA PEDROSA  
Presidente da Banca Examinadora

Assinatura Eletrônica

19/04/2022 13:51:05.0

ANA CLAUDIA BONATTO

Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica

19/04/2022 15:19:53.0

EMANUEL MALTEMPI DE SOUZA

Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica

25/04/2022 11:42:01.0

HELISSON FAORO

Avaliador Externo (INSTITUTO CARLOS CHAGAS)

Assinatura Eletrônica

20/04/2022 11:48:45.0

DANILO SIPOLI SANCHES

Avaliador Externo (UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ)

Assinatura Eletrônica

26/04/2022 19:27:24.0

DIEVAL GUIZELINI

Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica

19/04/2022 11:07:28.0

LEONARDO MAGALHÃES CRUZ

Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

SETOR DE CIÊNCIAS BIOLÓGICAS - CENTRO POLITÉCNICO - CURITIBA - Paraná - Brasil

CEP 81531-980 - Tel: (41) 3361-1672 - E-mail: pgbioq@ufpr.br

Documento assinado eletronicamente de acordo com o disposto na legislação federal Decreto 8539 de 08 de outubro de 2015.

Gerado e autenticado pelo SIGA-UFPR, com a seguinte identificação única: 175784

Para autenticar este documento/assinatura, acesse <https://www.prppg.ufpr.br/siga/visitante/autenticacaoassinaturas.jsp> e insira o código 175784

*A Deus, Inteligência Suprema, que nos ensina que a Fé contempla, muitas vezes, a uma lógica ainda desconhecida pela racionalidade humana.*

## AGRADECIMENTOS

Agradeço aos meus orientadores, prof. Dr. Fábio de Oliveira Pedrosa e prof. Dr. Roberto Tadeu Raittz, os quais me acolheram com seus generosos braços, através da oportunidade, da experiência e do valioso companheirismo e são, portanto, contribuidores inestimáveis para essa realização.

Agradeço à banca examinadora composta pelos professores Dr<sup>a</sup> Ana Cláudia Bonatto, Dr. Danilo Sipoli Sanches, Dr. Dieval Guizelini, Dr. Emanuel Maltempi de Souza e Dr. Leonardo Magalhães Cruz pelas valiosas contribuições que nos agregou ainda mais qualidade e zelo a este trabalho.

Agradeço aos colegas do programa de Pós-Graduação em Bioinformática. Felizmente, a lista é gigante de egressos, mestrandos e doutorandos que me ajudaram bastante, em especial ao grupo de Inteligência Artificial Aplicada à Bioinformática, pois ensinaram-me a ser um melhor programador e melhor ser humano.

Agradeço aos colegas do Programa de Pós-Graduação em Ciências (Bioquímica). Em especial a Dr<sup>a</sup> Roxana Beatriz Chaves, minha colega, amiga e irmã. E aos amigos do Núcleo de Fixação Nitrogênio (NFN), que me deram suporte em tempo integral e que trilharam comigo por este caminho.

Agradeço a todos os demais envolvidos direta ou indiretamente nos setores administrativos, técnicos, docentes e colaboradores dos PPGs em Ciências- Bioquímica e em Bioinformática da UFPR que prestaram valiosos serviços e para que fosse possível chegarmos à conclusão deste trabalho.

Agradeço à prof<sup>a</sup> Dr<sup>a</sup> Maria Berenice Reynauld Steffens, pelo brilhante exemplo de profissional e de mãe para com os seus alunos, amigos e familiares. A professora deixa um valioso legado para nós.

Agradeço ao meu amigo e irmão M.Sc Alexandre Quadros Lejambre, que sempre me incentivou nos trabalhos, meu exemplo de dedicação e esforço. Deixa-me em profunda gratidão.

Agradeço à família, em especial à minha mãe Maria Sílvia de Lima, por seu amor incondicional, e aos meus irmãos Alexandro de Lima Nichio e Cláudio Nichio Júnior, por tantos anos compartilhados com amor, luta e valiosas experiências que nos uniram e as quais me renovaram as forças para não desistir mesmo diante do distanciamento físico.

Agradeço à todas as agências financiadoras, em especial, a CAPES e Fundação Araucária, apoios fundamentais na viabilidade deste trabalho desde o seu projeto.

*“You can’t go back and make a new start, but you  
can start right now and make a brand new ending.”  
— James R. Sherman, Rejection*

*“Não são as respostas que movem o mundo.  
São as perguntas!”  
- Albert Einstein*



## RESUMO

A Fixação Biológica do Nitrogênio (FBN) ocorre com a ação do Complexo Enzimático (CEN) da Nitrogenase molibdênio dependente (Mo-nitrogenase) codificado pelos genes *nif*. Estes genes se distribuem em grande diversidade em espécies diazotróficas de *Archaea* e *Bacteria*. Evidências de novas espécies fixadoras acarretam volume crescente de informações disponíveis as quais precisam ser investigadas, ao passo que, métodos e tecnologias em bioinformática possibilitam diferentes abordagens na manipulação desse volume de informações e na análise dos genes *nif*. Propusemos, o estudo do conjunto mínimo (*core*) da Mo-nitrogenase que compreende as sequências NifH, NifD, NifK, NifE NifN e NifB em espécies diazotróficas utilizando técnicas de inteligência artificial (IA). Para isso, avaliamos se a herança da Mo-nitrogenase nos fixadores possui relação com aquisição, subtração ou substituição de *nif* acessórios e os impactos dessas relações para a manutenção do CEN. Exploramos como principais estratégias de IA: representação vetorial de sequências biológicas por SWeeP, Redes Neurais Artificiais (RNA), Análise dos Componentes Principais (PCA), Clusterização Hierárquica em Componentes Principais (HCPC), construções de super-árvores e agrupamento de sequências homólogas por RAFTS<sup>3</sup>G. As proteínas codificadas pelos elementos *nif* inseridos em *clusters* (ou núcleos) foram identificadas através de sintenia genômica. Identificamos 662 espécies diazotróficas das quais 52 espécies são potenciais fixadores desconhecidos pela literatura. Além disso, devido à projeção vetorial dos genomas dos diazotrofos, foi possível classificar todos os organismos em 10 grupos pela integração dos dados com o método de HCPC. Observamos significativa relação desses grupos no tocante à aquisição, perda e/ou transferência de genes *nif*, sugerindo a necessidade da subclassificação dos Clusters da Mo-nitrogenase proposta no trabalho de Raymond e colaboradores (2004). Essas evidências propõem ainda que, a Transferência Horizontal de Genes (THG) foi crucial na adaptação da FBN. Por fim, este trabalho demonstra o potencial da integração do método SWeeP e dos avanços nas técnicas de aprendizado de máquina à resolução de problemas biológicos complexos.

**Palavras-chave:** Diazotrofos. Nitrogenase. Inteligência Artificial. SWeeP.

## ABSTRACT

Biological Nitrogen Fixation (BNF) occurs with the action of the Nitrogenase Enzymatic Complex (NEC) of nitrogenase-dependent molybdenum (Mo-nitrogenase) encoded by the *nif* genes. These genes distribute in great diversity in diazotrophic species of Archaea and Bacteria genomes. Evidence of new diazotrophic species results in an increasing volume of available information that needs investigation. Methods and technologies in bioinformatics enable different approaches in the manipulation of large information and the *nif* genes analyses. We proposed the study of the minimum set (core) of Mo-nitrogenase comprising the sequences NifH, NifD, NifK, NifE, NifN, and NifB in diazotrophic species using techniques in Artificial Intelligence (AI). For this, we evaluated whether the inheritance of Mo-nitrogenase in fasteners is related to acquisition, subtraction, or replacement of accessory *nif* and the impacts of these relationships on the maintenance of NEC. We explore as the main AI strategies: vector representation of biological sequences by SWeeP, Artificial Neuronal Networks (ANN), Principal Component Analysis (PCA), Hierarchical Clusterization of Principal Components (HCPC), super-tree constructions, and grouping of homologous sequences by RAFTS<sup>3</sup>G. We identified the *nif* elements inserted in clusters through genomic synthesis. We obtained 662 diazotrophic genomes which 52 species are potential diazotrophs unknown in the literature. In addition, due to the vector projection of diazotroph genomes, it was possible to classify all organisms into ten groups with the HCPC method. We observed a significant relationship between these groups regarding the acquisition, loss, and transfer of *nif* genes, suggesting the need for the subclassification of Monitrogenase Clusters proposed in Raymond et al. (2004). Those shreds of evidence also propose that Horizontal Gene Transfer (HGT) was crucial in BNF evolution. Finally, this work demonstrates the potential of the SWeeP method for machine learning techniques advances to solve complex biological problems.

**Keywords:** Diazotrophic. Nitrogenase. AI. SWeeP.

## LISTA DE FIGURAS

FIGURA 1 - REAÇÃO GLOBAL DA FBN CATABOLIZADA PELO COMPLEXO DA NITROGENASE .....	24
FIGURA 2 - ESTRUTURA MOLECULAR DA MO-NITROGENASE EM <i>AZOTOBACTER VINELANDII</i> .....	26
FIGURA 3 - EVOLUÇÃO TAXONÔMICA DE ARQUÉIAS E BACTÉRIAS.....	30
FIGURA 4 – DISTRIBUIÇÃO DE NIF EM ORGANISMOS ANAERÓBIOS E AERÓBIOS/FACULTATIVOS.....	31
FIGURA 5 - MÉTODO DE DIVISÃO DO CONJUNTO EM TREINAMENTO/TESTE.....	34
FIGURA 6 – CATEGORIAS DE APRENDIZADO DE MÁQUINA .....	36
FIGURA 7 – ESTRUTURA DE UMA REDE NEURAL ARTIFICIAL TÍPICA .....	37
FIGURA 8 - FLUXOGRAMA DE SWEEP NA OBTENÇÃO DE VETOR DE SEQUÊNCIAS.....	40
FIGURA 9 – MAPA DE FATORES DE UMA PCA EM 1, 2 E 3 DIMENSÕES.....	42
FIGURA 10 – CLUSTERIZAÇÃO HIERÁRQUICA NO MAPA DE FATORES .....	44
FIGURA 11 – ESTRUTURA DE UMA ÁRVORE FILOGENÉTICA ENRAIZADA.....	45
FIGURA 12 – BLOCOS DE SINTENIA E AS REGIÕES DE QUEBRA. ....	48
FIGURA 13 - FLUXOGRAMA METODOLÓGICO .....	51
FIGURA 14 – ETAPAS DE DESENVOLVIMENTO DO MODELO DA RNA SWEEP. ....	53
FIGURA 15 - FLUXOGRAMA PARA CONSTRUÇÃO DE FILOGENIA SWEEP.....	57
FIGURA 16 - INTERFACE WEB DE SYNTTAX.....	58
FIGURA 17 - ORGANISMOS IDENTIFICADOS POR DIAZOINSPECTOR E VALIDADOS.....	60
FIGURA 18 - DEFINIÇÃO DO MODELO PARA NIFFINDER.....	65
FIGURA 19 - MAPA DE CALOR ORDENADO PARA NIF <i>CORE</i> .....	68
FIGURA 20 – MAPA DE FATORES DOS 10 GRUPOS POR HCPC. ....	69
FIGURA 21- SUPER-ÁRVORE NIF <i>CORE</i> <i>VERSUS</i> PROTEOMA DE <i>ARCHAEA</i> E <i>BACTERIA</i> .....	71
FIGURA 22 – PROTEOBACTÉRIAS GAMA RELACIONADAS NAS ÁRVORES <i>CORE</i> E PROTEOMA.....	72
FIGURA 23 – MAPAS DE CALOR NIF <i>CORE</i> E PROTEOMA .....	73
FIGURA 24 – DISTRIBUIÇÃO DOS GRUPOS NA FILOGENIA NIF <i>CORE</i> .....	74

FIGURA 25 - SINTENIA GENÔMICA PARA G1 .....	75
FIGURA 26- SINTENIA GENÔMICA PARA G2 .....	75
FIGURA 27 – SINTENIA GENÔMICA PARA G3.....	76
FIGURA 28 - SINTENIA GENÔMICA PARA G4 .....	78
FIGURA 29 - SINTENIA GENÔMICA PARA G5 .....	79
FIGURA 30 - SINTENIA GENÔMICA PARA G6 .....	80
FIGURA 31 - SINTENIA GENÔMICA PARA G7 .....	81
FIGURA 32 - SINTENIA GENÔMICA PARA G8 .....	82
FIGURA 33 SINTENIA GENÔMICA PARA G9.....	83
FIGURA 34 – ÁRVORE <i>CORE</i> E NIF IDENTIFICADAS .....	86
FIGURA 35 – FILOGENIA <i>CORE</i> DAS CIANOBACTÉRIAS (G6).....	89
FIGURA 36 – NIF <i>CORE</i> E PROTEOMA DOS GÊNEROS <i>AZOARCUS</i> E <i>AZOSPIRA</i> . .....	92

## LISTA DE QUADROS

QUADRO 1 - GENES <i>NIF</i> E FUNÇÕES DESCRITAS PARA <i>AZOTOBACTER VINELANDII</i> .....	27
QUADRO 2 - ORGANISMOS SEM VALIDAÇÃO EM LITERATURA COM <i>NIF CORE</i> COMPLETO .....	61
QUADRO 3 – PREDIÇÃO DE <i>NIFPRED</i> E <i>NIFFINDER</i> PARA <i>APHANIZOMENON FLOS-AQUAE DEX188</i> .....	66
QUADRO 4 – FILOS CARACTERIZADOS PARA OS 10 GRUPOS OBTIDOS POR HCPC .....	69

## LISTA DE TABELAS

TABELA 1 - PROTEÍNAS COM OUTRAS DEFINIÇÕES IDENTIFICADAS EM CLUSTERS NIF HOMÓLOGOS.....	63
TABELA 2 – VALORES ESTATÍSTICOS PARA K GRUPOS INDICADOS.....	67
TABELA 3 – NIF IDENTIFICADAS PARA <i>AZOARCUS</i> E <i>AZOSPIRAE</i> .....	93

## LISTA DE ABREVIATURAS E SIGLAS

ADP	Adenosina Di-Fostato
ATP	Adenosina Tri-Fosfato
BCOM	<i>Binary Co-Occurrence Matrix</i>
BLAST	<i>Basic Local Alignment Search Tool</i>
CEN	Complexo Enzimático da Nitrogenase
DBSCAN	<i>Density-Based Spatial Clustering of Applications with Noise</i>
FBN	Fixação Biológica do Nitrogênio
FTP	<i>File Transfer Protocol</i>
GAF	<i>cGMP-specific phosphodiesterases, Adenylyl cyclases and FhIA</i>
GEO	Grande Evento da Oxigenação
HCPC	<i>Hierarchical Clustering on Principal Components</i>
HDV	<i>Higher-Dimensional Vector</i>
HIF	<i>Hormogonium-Inducing Factor</i>
HTH	<i>Helix turn Helix</i>
IA	Inteligência Artificial
IS	<i>Insertion Sequence</i>
e.g.	<i>exempli gratia</i> (por exemplo)
i.e.	<i>id est</i> (ou seja)
KNN	<i>K-nearest neighbors</i>
MCA	<i>Multiple Correspondence Analysis</i>
MFA	<i>Multiple Factor Analysis</i>
MLP	<i>MultiLayer-Perceptron</i>
Mo-nitrogenase	Nitrogenase Molibdênio dependente
NCBI	<i>National Center for Biotechnology Information</i>
NOR	<i>Nitric Oxide Reductase</i>
NJ	<i>Neighbor Joining</i>
Obs.	Observação
PAS	<i>Per-Arnt-Sim</i>
PCA	<i>Principal Component Analysis</i>
PVC	<i>Planctomycetes, Verrucomicrobia e Chlamydiae</i>
RAFTS3	<i>Rapid Alignment-Free Tool for Sequence Similarity Search</i>
RAFTS <sup>3</sup> G	<i>Rapid Alignment-Free Tool for Sequence Similarity Search to Groups</i>

RefSeq	<i>NCBI Reference Sequence Database</i>
RFC	Radiação dos Filos Candidatos
RNA	Redes Neurais Artificiais
SAM	<i>S-adenosyl-L-methionine</i>
SIB	<i>Swiss Institute of Bioinformatics</i>
SVM	<i>Support-Vector Machine</i>
SWeeP	<i>Spaced Words Projection</i>
UPGMA	<i>Un-weighted Pair Group Method with Arithmetic mean</i>
THG	Transferência Horizontal de Genes
XRE	<i>Xenobiotic Response Element</i>



## LISTA DE SÍMBOLOS

$\alpha$	Alfa
$\beta$	Beta
$\gamma$	Gama
$\delta$	Delta
$\varepsilon$	Épsilon
$k$ -mer	<i>Substrings</i> de comprimento $k$
®	Marca registrada
Cys	Cisteína
Fe	Ferro
FeMo	Ferro-Molibdênio
FeMo-co	Cofator Ferro-Molibdênio
S	Enxofre
H <sub>2</sub>	Diidrogênio (Hidrogênio)
$\log_2(n)$	Logaritmo de 2 na base $n$ (logaritmo binário)
NH <sub>3</sub>	Hidróxido de Amônia (Amônia)
NO <sub>3</sub>	Nitrato
Mo	Molibdênio
O <sub>2</sub>	Dióxido (Oxigênio)
V	Vanádio

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO</b>	<b>20</b>
<b>2</b>	<b>OBJETIVOS</b>	<b>23</b>
2.1	GERAL	23
2.2	ESPECÍFICOS	23
<b>3</b>	<b>REVISÃO TEÓRICA-EMPÍRICA</b>	<b>24</b>
3.1	ASPECTOS MOLECULARES E EVOLUTIVOS DA FBN	24
3.1.1	A BIOLOGIA MOLECULAR DA FBN	24
3.1.2	THG E A DIVISÃO DAS MO-NITROGENASES	28
3.1.3	NIF EM FIXADORES ANAERÓBIOS E AERÓBIOS	30
3.2	TÉCNICAS E MÉTODOS DE IA APLICADA À BIOINFORMÁTICA	32
3.2.1	APRENDIZADO DE MÁQUINA	32
3.2.2	APRENDIZADO SUPERVISIONADO E NÃO SUPERVISIONADO	34
3.2.3	A REDES NEURAIS ARTIFICIAIS	37
3.2.4	A REDUÇÃO DE DIMENSIONALIDADE	37
3.2.5	<i>SPACED WORDS PROJECTION</i> (SWEEP)	38
3.2.6	ANÁLISE DE COMPONENTES PRINCIPAIS (PCA)	40
3.2.7	AGRUPAMENTO HIERÁRQUICO EM PCA (HCPC)	42
3.2.8	SISTEMÁTICA FILOGENÉTICA	44
3.2.9	RAFTS <sup>3</sup> G: CLUSTERIZAÇÃO DE HOMÓLOGOS	46
3.2.10	ANÁLISE DE SINTENIA GENÔMICA	47
<b>4</b>	<b>MATERIAIS E MÉTODOS</b>	<b>50</b>
4.1	DEFININDO CONJUNTO PROTEOMA E NIF <i>CORE</i>	52
4.2	PROTEOMAS BACTERIANOS OBTIDOS PELO NCBI	52
4.3	PROTEOMAS REPRESENTADOS EM BASE VETORIAL	52
4.4	DIAZOTROFOS OBTIDOS PELA LITERATURA	52
4.5	MINERAÇÃO DE FIXADORES CANDIDATOS	53
4.6	PROTEÍNAS NIF HOMÓLOGAS IDENTIFICADAS POR RAFTS <sup>3</sup> G (I)	54
4.7	PREDIÇÃO DE NIF EM DIAZOTROFOS CANDIDATOS	54

4.8	VALIDAÇÃO DE DIAZOTROFOS POR REVISÃO SISTEMÁTICA .....	54
4.9	NIF HOMÓLOGAS IDENTIFICADAS POR RAFTS <sup>3</sup> G (II) .....	55
4.10	PCA SWEEP DO CONJUNTO NIF <i>CORE</i> .....	55
4.11	APLICAÇÃO DO MÉTODO HCPC EM PCA .....	55
4.12	MAPAS DE CALOR.....	56
4.13	FILOGENIA SWEEP .....	56
4.14	SINTENIA POR SYNTTAX .....	57
<b>5</b>	<b>RESULTADOS</b> .....	<b>60</b>
5.1	CURADORIA DOS ORGANISMOS DIAZOTRÓFICOS .....	60
5.2	NIF IDENTIFICADAS POR HOMOLOGIA .....	62
5.3	EXTRAÇÃO DE NIF <i>CORE</i> , PROTEOMA E <i>CLUSTER</i> NIF .....	64
5.3.1	NIFFINDER.....	64
5.4	GRUPOS OBTIDOS POR HCPC.....	66
5.5	MAPAS DE FATORES OBTIDOS POR PCA.....	68
5.6	FILOGENIA SWEEP .....	70
5.7	SINTENIA DOS NÚCLEOS NIF EM GRUPOS DIAZOTRÓFICOS .....	74
<b>6</b>	<b>DISCUSSÃO</b> .....	<b>84</b>
6.1	PREDIÇÃO DE ORGANISMOS FIXADORES .....	84
6.2	ANÁLISE DOS GRUPOS.....	85
6.3	THG: CASOS <i>AZOARCUS</i> E <i>AZOSPIRA</i> .....	91
<b>7</b>	<b>CONCLUSÕES</b> .....	<b>94</b>
	<b>REFERÊNCIAS</b> .....	<b>95</b>
	<b>APÊNDICE</b> .....	<b>107</b>
	<b>ANEXO</b> .....	<b>200</b>

## 1 INTRODUÇÃO

A Fixação Biológica do Nitrogênio (FBN) é um processo bioquímico crucial para o ciclo do nitrogênio, caracterizado pela redução do nitrogênio atmosférico ( $N_2$ ) em amônia ( $NH_3$ ). A FBN caracteriza-se pela quebra da tripla ligação da molécula  $N_2$  através da nitrogenase, sendo a principal nitrogenase envolvida, a nitrogenase molibdênio dependente (Mo-nitrogenase). Proteínas que compõem a biossíntese, relacionadas à maturação da Mo-nitrogenase e ao carregamento de íons viabilizam a bioquímica do processo, constituindo assim, o Complexo Enzimático da Nitrogenase (CEN) (BURÉN *et al.*, 2020). O CEN é codificado pelos genes *nif* em organismos específicos de *Archaea* e *Bacteria*, outrossim, é reportado significativa variedade dos genes *nif* requeridos para a exequidade da FBN o que tornam as estruturas codificadas pelos elementos *nif* alvo de investigação. Além disso, a capacidade dos diazotrofos usarem  $N_2$  como única fonte de nitrogênio, estarem inseridos em diversas condições ambientais e alguns deles estabelecerem interações com organismos hospedeiros, possibilitam-lhes diversas vantagens evolutivas e, desse modo, o mecanismo da FBN possui diversas alterações, resultante das sucessivas transferências horizontais (ou laterais) e verticais ao longo de bilhões de anos (BOYD, Eric S. *et al.*, 2015; ESTEVES-FERREIRA *et al.*, 2017; INOUE *et al.*, 2015). A FBN é considerada tão primitiva que acompanha o ciclo da vida terrestre, mais antiga que a fotossíntese - e também possui relações diretas e indiretas com a origem desse outro processo (BOYD, Eric S.; PETERS, 2013; MUS *et al.*, 2019). Exames filogenéticos e estruturais de múltiplas proteínas Nif sugerem que a Mo-nitrogenase emergiu de um ancestral anaeróbio termofílico e hidrogenotrófico de arqueia há cerca de 3,2 bilhões de anos. Em seguida, deu origem às suas duas outras formas alternativas, a Vanádio (V) e a Ferro (Fe) nitrogenase dependente, propondo que estas surgiram mais tarde em resposta à limitação local do elemento Molibdênio (MUS *et al.*, 2019).

Identificar e selecionar novos organismos fixadores têm sido objetivo de vários trabalhos e, a despeito desse tema, os métodos *in silico* de classificação de organismos fixadores são distintos. Algumas metodologias, limitam-se em um único gene, usualmente o gene *nifH* que codifica para a nitrogenase redutase NifH. Esta abordagem, inclusive, é amplamente utilizada como marcador molecular filogenético, além de marcador para inferência de novos organismos diazotróficos (BAHULIKAR *et al.*, 2014; CHEUNG *et al.*, 2015; COELHO *et al.*, 2009; ROMÁN-PONCE *et al.*, 2016).

Entretanto, é reportado que cópias de *nifH* estão presentes em organismos não fixadores, culminando em um “viés” na predição de potenciais diazotrofos (WATANABE; HORIIKE, 2021; ZANI *et al.*, 2000). A esse respeito, vários trabalhos buscam pelo melhor método *in silico* para a predição de novas espécies diazotróficas. Alguns postulam que os componentes estruturais do arcabouço da Mo-nitrogenase (NifH, NifD e NifK) são estruturas suficientes à predição (BOYD, *et al.*, 2015). Para outros estudos, esse conjunto é insuficiente e necessita englobar os elementos que compõem também a biossíntese dos grupos prostéticos, resultando em um conjunto mínimo composto por NifH, NifD, NifK, NifE, NifN (BOYD, E. S. *et al.*, 2011; PAULITSCH *et al.*, 2020). Em 2012, Dos Santos e colaboradores propuseram o “conjunto mínimo” (ou *core*) da Mo-nitrogenase compreendido pelas proteínas NifH, NifD, NifK, NifE, NifN e NifB, presentes nos genomas diazotróficos e imprescindíveis para a continuidade da FBN (DOS SANTOS *et al.*, 2012; LI *et al.*, 2014). Aliado a essa conjectura, a semelhança funcional e filogenética de NifE com NifD e NifK com NifN sugere a origem de *nifEN* pela duplicação de um ancestral dos genes *nifDK*. Ao passo que os genes estruturais da nitrogenase (*nifHDK*) e das proteínas de montagem do cofator Mo-Fe (*nifENB*) possuem distribuição filogenética similar, indicando que todos os seis genes co-evoluíram e se mantêm juntos (KOIRALA; BRÖZEL, 2021). Portanto, diante dos trabalhos supramencionados, verificamos a relevância da seleção dos dados *core* da Mo-nitrogenase, bem como dos produtos de genes adicionais envolvidos na biossíntese de grupos prostéticos, e da aplicação de métodos comparativos, pois maiores percepções podem ser obtidas a partir deles (BOYD, Eric S.; PETERS, 2013). Ademais, os impactos envolvidos na transferência da Mo-nitrogenase e suas implicações na construção do *cluster* de genes *nif* (ou do núcleo<sup>1</sup> *nif*) em genomas de *Archaea* e *Bacteria* continuam indefinidos.

Nesse cenário em paralelo ao desenvolvimento da Inteligência Artificial (IA) aplicada à bioinformática, esta última torna-se importante aliada no emprego de estratégias e de métodos computacionais, tanto à manipulação como à extração de conhecimento da FBN, retirando, a partir dos dados coletados, informações implícitas que colaboram significativamente em ultrapassar a fronteira do conhecimento. Dentre

---

<sup>1</sup> O termo núcleo adotado para este trabalho também se refere ao agrupamento de elementos conservados e homólogos, os quais podem ser co-expressos e estão envolvidos numa mesma via metabólica. No entanto, quando descrevemos padrões de núcleos *nif*, damos maior ênfase aos *nif* identificados.

muitos exemplos, podemos citar que a bioinformática é capaz de viabilizar a análise de grande volume de informações de espécies diazotróficas, em reduzir a complexidade dos dados biológicos disponíveis, em utilizar métricas comparativas em diversos dados de genes *nif*, em estabelecer grupos de organismos que descrevam padrões semelhantes. (GAUTHIER *et al.*, 2019; SCHMIDT; FORNE; IMHOF, 2014; SISÓ; DOLORS, 2020). No entanto, para que todo esse trabalho seja possível, é crucial a integração de grande volume de dados com técnicas de redução de complexidade informacional, sem que haja perda do sentido biológico.

Portanto, trazemos o método desenvolvido com estratégias de Inteligência Artificial (IA), *Spaced Words Projection* (SWeeP), o qual, além de representar grande número de sequências biológicas, os vetores gerados por SWeeP podem ser utilizados em várias técnicas de mineração de dados, a exemplo, no treinamento de modelos de aprendizado de máquina (SCHREINER, 2021), na construção de super árvores, como demonstrado na construção de filogenias de 13.941 sequências aminoácidas do gene *neuB*, em 8.426 Proteomas mitocondriais disponíveis no RefSeq, para 10.324 Proteomas bacterianos e no conjunto de 1.000.588 Proteomas virais de SarsCov-2 (DE PIERRI *et al.*, 2020; PERICO *et al.*, 2021; RAITTZ *et al.*, 2021; VIEIRA; RAITTZ; FAORO, 2021). Além dos exemplos supracitados, a utilização de SWeeP tem demonstrado resultados satisfatórios na análise multivariada, por exemplo, na Análise dos Componentes Principais (PCA) (SCHREINER, 2021).

## 2 OBJETIVOS

### 2.1 GERAL

Construir e a interpretar as possíveis relações entre grupos de organismos diazotróficos sob a análise das estruturas *core* codificadas pelos genes *nif* em *Archaea* e *Bacteria* obtidos de genomas completos disponíveis na base do NCBI através do uso e desenvolvimento de estratégias de Inteligência Artificial.

### 2.2 ESPECÍFICOS

Os objetivos específicos dessa tese são:

- 1) Construir uma base de representação vetorial dos Proteomas<sup>2</sup> disponíveis na base de dados do *National Center for Biotechnology Information* (NCBI);
- 2) Propor e consolidar base de dados de diazotrofos identificados através de estratégia de mineração de dados;
- 3) implementar a Análise dos Componentes Principais (PCA) no conjunto dos vetores obtidos dos proteomas dos fixadores identificados,
- 4) Avaliar o agrupamento obtido pelo método *Hierarchical Clustering on Principal Components* (HCPC);
- 5) Identificar os padrões compartilhados pelos organismos de cada grupo ou correlacionar os grupos com as funções de fixação;
- 6) Identificar as alterações estruturais (inserções, exclusões e/ou substituições) no *cluster nif* através da análise de sintenia;
- 7) Avaliar a distribuição e os agrupamentos dos organismos fixadores de nitrogênio através da maior composição comum dos genes associados a FBN;
- 8) Correlacionar "os padrões" e os "não padrões" dos agrupamentos de organismos fixadores de nitrogênio com a taxonomia correspondente e,

---

<sup>2</sup> Quando nos referimos ao termo Proteoma (com P em maiúsculo), entende-se como o conjunto global de proteínas anotadas de genes ou regiões codificantes (CDS) obtido a partir do dado genômico de um organismo.

por fim,

- 9) Correlacionar as evidências observadas com os potenciais eventos evolutivos envolvidos entre as espécies.

### 3 REVISÃO TEÓRICA-EMPÍRICA

A revisão sistemática de literatura foi dividida em duas partes: 3.1 aspectos moleculares e evolutivos da FBN e 3.2 técnicas e métodos de Inteligência Artificial (IA) aplicada à bioinformática.

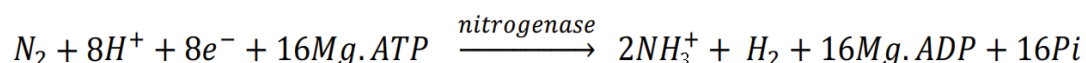
#### 3.1 ASPECTOS MOLECULARES E EVOLUTIVOS DA FBN

Neste capítulo elencamos a biologia molecular da Fixação Biológica do Nitrogênio (FBN), Transferência Horizontal de Genes (THG) e a divisão das Mo-nitrogenases e *nif* em fixadores anaeróbicos e aeróbicos.

##### 3.1.1 A BIOLOGIA MOLECULAR DA FBN

A FBN é o principal processo biológico pelo qual plantas tornam-se capazes de assimilar nitrogênio através da interação com bactérias, denominadas de diazotróficas. Esses micro-organismos reduzem o N<sub>2</sub> atmosférico a amônia (NH<sub>3</sub>) catalisado pelo complexo da nitrogenase que é sensível à certos níveis de oxigênio e possui alto custo energético (DOBEREINER, 1997; HOFFMAN *et al.*, 2014). Na equação estequiométrica da FBN, a partir de uma molécula de N<sub>2</sub> e à custa de 16 moléculas de ATP magnésio dependente e de um suprimento de elétrons e de prótons (íons de hidrogênio) duas moléculas de NH<sub>3</sub> e uma de H<sub>2</sub> são produzidos (POSTGATE, 1983).

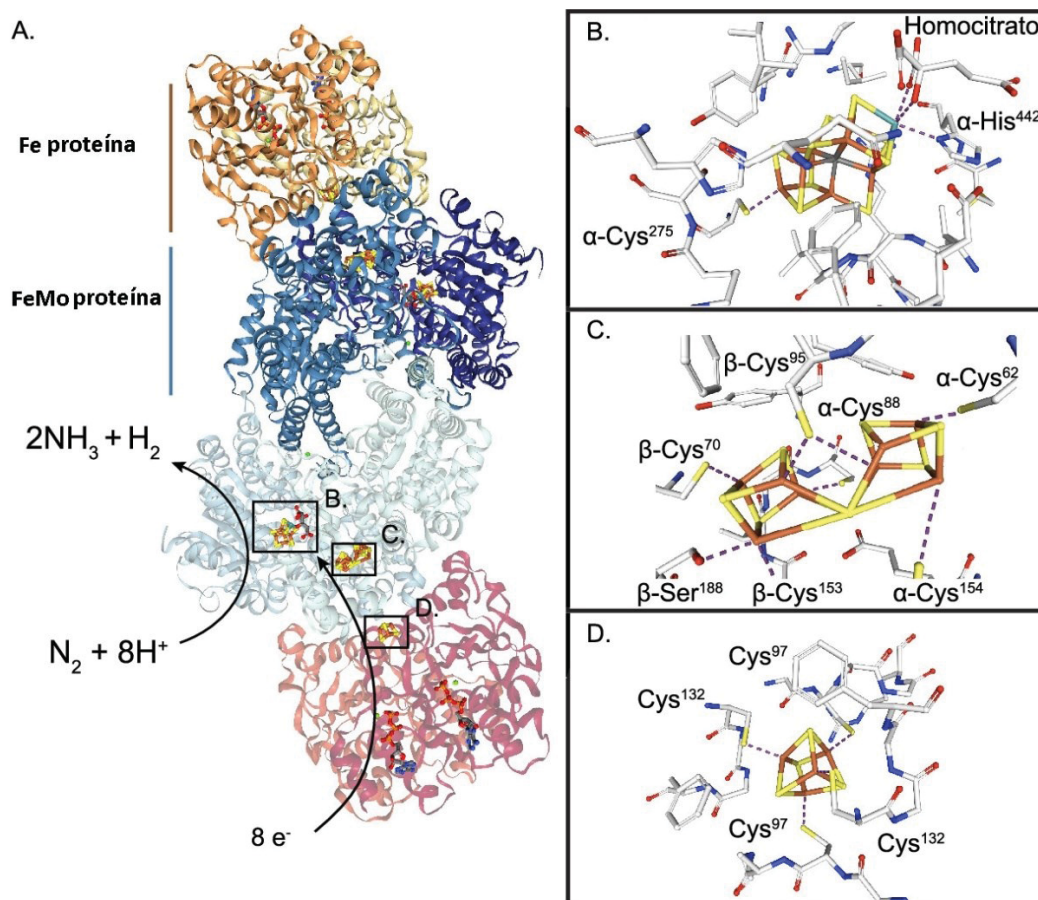
FIGURA 1 - REAÇÃO GLOBAL DA FBN CATABOLIZADA PELO COMPLEXO DA NITROGENASE





FONTE: Postgate (1983).

A principal via envolvida na FBN é atribuída ao complexo enzimático da nitrogenase molibdênio dependente (Mo-nitrogenase). O modelo mais estudado desse complexo é o encontrado no diazotrofo *Azotobacter vinelandii*, o qual é constituído por um tetrâmero de proteína Ferro-Molibdênio (FeMo) central com composição de subunidade  $\alpha_2\beta_2$  e um homodímero de proteína Ferro (Fe) ligado a cada metade  $\alpha\beta$  (FIGURA 2-A). São nessas estruturas também que encontramos os principais sítios enzimáticos: o cofator Fe-Mo (FeMo-co) (FIGURA 2-B), cluster P (FIGURA 2-C) e o cluster [4Fe-4S] (FIGURA 2-D). FeMo-co é composto por um cluster inorgânico [7Fe-9S-C-Mo] e uma molécula de R-homocitrato exclusivamente ligada ao átomo de Mo [7Fe-9S-C-Mo-R-homocitrato] (EINSLE *et al.*, 2002; KIRN; REES, 1992; SPATZAL *et al.*, 2011). Os clusters P são grupos [8Fe-7S] localizados na interface das subunidades  $\alpha$  e  $\beta$  dentro de cada metade  $\alpha\beta$  da proteína FeMo. (PETERS *et al.*, 1997; SPATZAL *et al.*, 2011) O cluster [4Fe-4S] ocupa uma posição de ponte na interface de ambas as subunidades e é coordenado simetricamente por dois resíduos Cys de cada um (e.g.: Cys97 e Cys132 em *A. vinelandii*).

FIGURA 2 - ESTRUTURA MOLECULAR DA MO-NITROGENASE EM *AZOTOBACTER VINELANDII*.

FONTE: Adaptado de Búren et al (2020).

LEGENDA: Estrutura da Mo-nitrogenase descrito para *A. vinelandii*. Ferro-Molibdênio (FeMo) central com composição de subunidade  $\alpha_2\beta_2$  e um homodímero de proteína Ferro (Fe) ligado a cada metade  $\alpha\beta$  (FIGURA 2-A). São nessas estruturas também que encontramos os principais sítios enzimáticos: o cofator Fe-Mo (FeMo-co) (FIGURA 2-B), cluster P (FIGURA 2-C) e o cluster [4Fe-4S] (FIGURA 2-D)

Cada uma das estruturas do complexo tem em sua composição proteínas codificadas pelos genes *nif* e a participação de genes denominados de acessórios. Os papéis propostos para os produtos descritos em *A. vinelandii* são: três genes estruturais da nitrogenase para a proteína FeMo são *nifD* para a subunidade  $\alpha$  e *nifK* para a subunidade  $\beta$  e o gene *nifH* para a proteína Fe. Os genes que codificam proteínas envolvidas na biossíntese de cofator *nifB* que catalisa a formação de NifB-co, o primeiro passo na biossíntese de FeMo-co juntamente com as proteínas codificadas por *nifE* e *nifN* e genes *nifO* e *nifQ* envolvidos no processamento de Mo para nitrogenase. Genes relacionados à maturação da proteína Fe pelo gene *nifM* e da proteína FeMo por *nifW* e *nifZ*. Proteínas para a biossíntese de clusters de [Fe-S] são codificados por *nifU*, *nifS*, *iscA<sup>nif</sup>* e *nifP*). Uma R-homocitrato sintase é codificada por *nifV*, *nifF* para proteínas envolvidas na transferência de elétrons para nitrogenase.

O controle da expressão de genes *nif* formado por *nifA* (ativador) e por *nifL* (anti-ativador) em resposta a sinais ambientais, como taxas de assimilação de nitrogênio, presença de O<sub>2</sub>, estado redox da célula, e níveis de ATP (BURÉN *et al.*, 2020; MARTINEZ-ARGUDO *et al.*, 2004; MUS *et al.*, 2018). A TABELA 1 resume os fenótipos Nif, obtidos por análise genética. No caso de *A. vinelandii*, outros genes descritos são acionados durante o processo da FBN: genes *naf* que codificam para proteínas acessórias da nitrogenase, genes *rnf* para flavodoxinas que compõem o sistema transferência de elétrons (e<sup>-</sup>), genes *mos* (ou *mod*) para proteínas transportadoras de molibdênio e genes *fix* para proteínas de fenótipo da fixação de nitrogênio em simbiose (HAMILTON *et al.*, 2011; JACOBSON *et al.*, 1989).

QUADRO 1 - GENES *NIF* E FUNÇÕES DESCRITAS PARA *AZOTOBACTER VINELANDII*.

<b>Gene</b>	<b>Produto(s)</b>	<b>Função(ões)</b>
<i>nifH</i>	Nitrogenase redutase, Fe proteína, NifH;	Doador de elétron (e <sup>-</sup> ) obrigatório para proteína FeMo. Também necessária para a biossíntese de FeMo cofator (FeMo-co) e maturação apo-MoFe.
<i>nifD</i>	Nitrogenase oxido-redutase subunidade alfa, NifD	Proteína MoFe subunidade α. Subunidade de ligação com FeMo-co. Forma o tetrâmero α <sub>2</sub> β <sub>2</sub> com NifK.
<i>nifK</i>	Nitrogenase oxido-redutase subunidade beta, NifK;	Proteína MoFe subunidade β. Forma o tetrâmero α <sub>2</sub> β <sub>2</sub> com NifD.
<i>nifT</i>	Proteína NifT, FixU;	função desconhecida.
<i>nifY</i>	Proteína NifY;	Interage com a proteína apo-MoFe antes da inserção do FeMo-co.
<i>nifE</i>	Nitrogenase oxido-redutase subunidade NifE, Nitrogenase componente I;	Andaime para maturação do cofator NifB em FeMo-co. Forma o tetrâmero α <sub>2</sub> β <sub>2</sub> com NifN.
<i>nifN</i>	Nitrogenase oxido-redutase subunidade NifN, Nitrogenase componente II;	Andaime para maturação do cofator NifB em FeMo-co. Forma o tetrâmero α <sub>2</sub> β <sub>2</sub> com NifE.
<i>nifX</i>	Proteína NifX;	Liga ao cofator NifB e interage com NifEN. Transportador proposto de precursores FeMo-co.
<i>nifU</i>	Proteína de andaime ( <i>scaffold</i> ) NifU;	Fornece precursores de clusters [Fe-S] para proteínas e cofatores da Nitrogenase.

<i>nifS</i>	Cisteína dessulfurase, proteína NifS;	Mobiliza S para síntese do cluster [Fe-S] em NifU.
<i>nifV</i>	Homocitrato sintase;	Envolvido na síntese de FeMo-co.
<i>nifP</i>	Serina O-acetiltransferase;	Participa na formação de cisteína na biossíntese de [Fe-S]
<i>nifW</i>	Proteína estabilizadora NifW;	Associa-se à proteína apo-MoFe precedendo à maturação do cluster P.
<i>nifZ</i>	Proteína NifZ;	Envolvido na maturação do cluster P em combinação com a Fe proteína.
<i>nifM</i>	Peptidil prolil cis-trans isomerase, rotamase;	Essencial para maturação da Fe proteína. Propõe que atua sobre o resíduo Pro <sup>258</sup> de NifH.
<i>nifF</i>	NAD(P)H dependente oxido-redutase;	Doador fisiológico de e <sup>-</sup> para Fe proteína.
<i>nifL</i>	Proteína histidina quinase; NifL	Proteína sensor. Anti-ativador de NifA.
<i>nifA</i>	Regulador transcricional família Fis $\sigma$ 54-dependente;	Forma o sistema de dois componentes com NifL.
<i>nifB</i>	Fe-S óxido-redutase;	Essencial para síntese de FeMo-co. Catalisa NifB-co a primeira etapa comprometida na síntese de FeMo-co.
<i>nifO</i>	Arsenato redutase ( <i>ArsC-like</i> ), proteína NifO;	Necessário para a ativação simultânea de nitrogenase e nitrato redutase.
<i>nifQ</i>	Proteína de processamento NifQ;	Processa Mo antes da sua incorporação ao FeMo-co.

FONTE: O Autor (2022).

### 3.1.2 THG E A DIVISÃO DAS MO-NITROGENASES

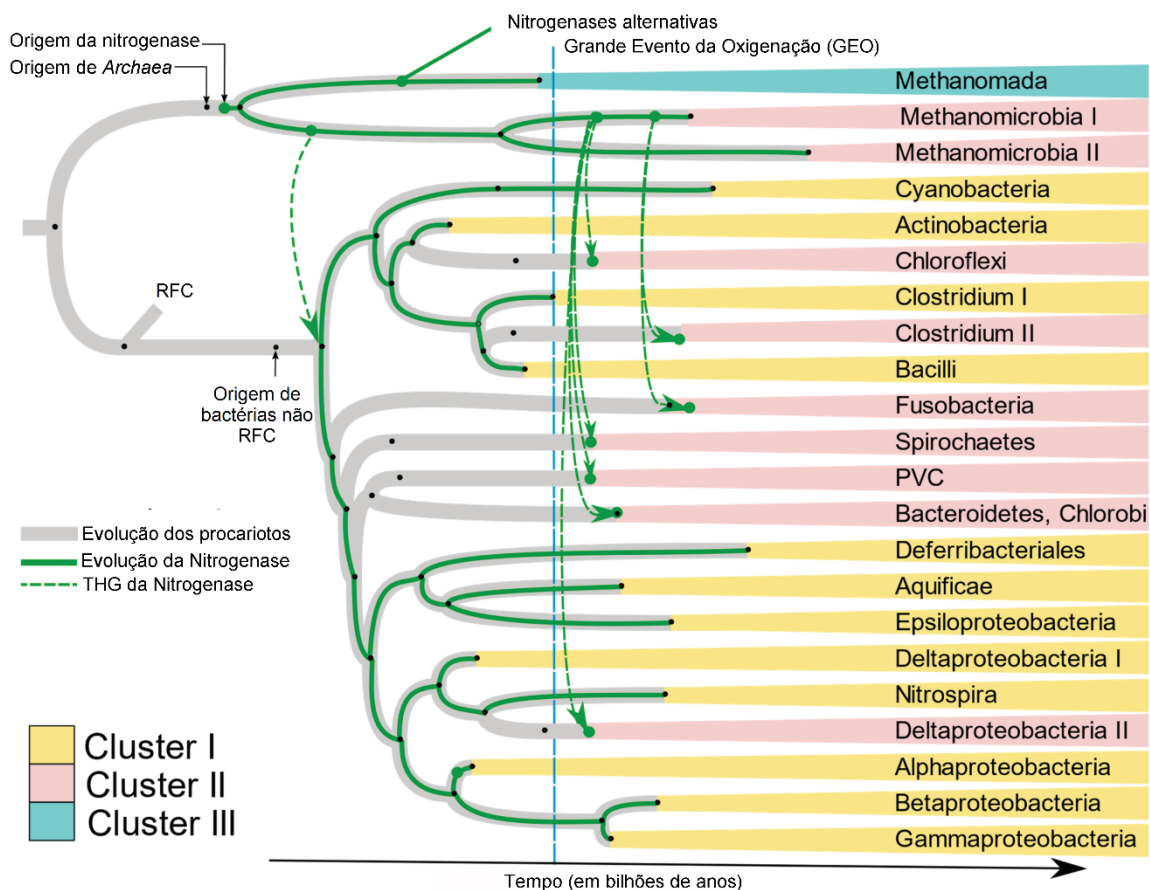
A hipótese mais aceita é de que a Mo-nitrogenase originou nas arqueias metanogênicas, com transferência dos genes *nif* para bactérias por Transferência Horizontal de Genes (THG) (KOIRALA; BRÖZEL, 2021; RAYMOND *et al.*, 2004). Ausência de genes *nif* no grupo Radiação dos Filos Candidatos (RFC), *Eukarya* e *Archaea* não metanogênica, e ramificação do *nif* bacteriano de *Archaea* sugere fortemente que a nitrogenase se originou primeiro em uma arqueia metanogênica primitiva. Estima-se que bactérias antigas da classe *Methanomicrobia* transferiram horizontalmente o primeiro *nif* em bactérias, entre 3,5 e 3,18 bilhões de anos, e, posteriormente, para todas as bactérias subsequentes por transferência vertical. A

distribuição da nitrogenase em vários filos sugere que provavelmente o Grande Evento da Oxigenação (GEO) do planeta em 2,3 bilhões de anos deve ter tornado muitos ambientes bacterianos aeróbios, criando nichos anaeróbios isolados. Este fato teria trazido diversos organismos anaeróbios em estreita proximidade nos nichos anaeróbios disponíveis. A co-ocorrência de táxons não relacionados nas proximidades aumentaria a troca de genes, favorecendo THG entre espécies anaeróbias o que tem suporte devido a semelhança dos genes *nif* de anaeróbios como *Bacteroidetes*, *Chlorobi*, *Spirochaetes*, *Deltaproteobacteria*, *Clostridium* e do grupo PVC (*Planctomycetes*, *Verrucomicrobia* e *Chlamydiae*) para *Metanomicrobia* (FIGURA 3). A nitrogenase prosseguiu das bactérias anaeróbicas incluindo filos de Proteobacterias e Chlorobi, finalmente chegando a linhagens com a capacidade de integrar O<sub>2</sub> em seu metabolismo de energia, há cerca de 800 *Mega-annum* (Ma) (BOYD, Eric S. *et al.*, 2015; LYONS; REINHARD; PLANAVSKY, 2014).

Raymond e colaboradores, em 2004, elaboraram um critério de classificação das nitrogenases as quais se dividem em 5 *clusters*, sendo que as Mo-nitrogenases estão subdivididas em dois deles: o *Cluster I* - representado pelas Mo-nitrogenases em espécies aeróbicas e anaeróbicas facultativas, são as nitrogenases mais estudadas e relacionam-se com os maiores *operons nif* conhecidos; e o *Cluster II* - representado pelas nitrogenases inseridas em espécies fixadoras anaeróbicas, bem caracterizadas e similares em estrutura das nitrogenases do *Cluster I* (RAYMOND *et al.*, 2004). O padrão de ramificação no Cluster II, sugere que dois eventos de THG distintos devem ter ocorrido para atingir a distribuição atual dos genes *nif* entre as bactérias anaeróbicas, onde *Clostridium* (II) e *Fusobacteria* obtiveram os genes muito mais tarde do que outros anaeróbios (FIGURA 3). O *cluster I*, a maioria pertence a  $\alpha$ -*proteobacteria*, *Cyanobacteria* e algumas espécies de *Firmicutes*. Embora a maioria das  $\alpha$ -*proteobacteria* tenha perdido genes *nif*, eles são preservados nas espécies que evoluíram para estar em simbiose com as plantas (e.g.: *Rhizobiales*) e espécies de vida livre que ocorrem principalmente em habitats com deficiência de nitrogênio, como a água (e.g.: *Rodospirillales* e *Rhodobacterales*). A ocorrência de *Aquificae* e *Nitrospirae* na raiz do *Cluster I* e agrupamento de *Cyanobacteria*, *Actinobacteria* e *firmicutes* aeróbicos na árvore *nif* sugere fortemente que a nitrogenase nesses filos evoluíram por THG de bactérias antigas não CPR. Em síntese, a THG tem impacto na distribuição da nitrogenase nos diazotrofos, portanto, vários táxons do mesmo filo podem ter nitrogenases distintas, com base em sua co-ocorrência fisiológica e ser

ecologicamente dirigida com táxons não relacionados. Portanto, conhecer quem são esses organismos e avaliar características na engenharia molecular dos genes *nif* pode auxiliar no entendimento do porquê de tantas espécies de mesmo táxon se distanciarem no contexto da FBN.

FIGURA 3 - EVOLUÇÃO TAXONÔMICA DE ARQUÉIAS E BACTÉRIAS.



FONTE: Adaptado de Koirala e Brözel (2021).

LEGENDA: Evolução taxonômica bacteriana com a direção (setas intermitentes em verde) das transferências da Nitrogenase entre arqueias e bactérias. Três das cinco classificações da nitrogenase segundo Raymond e colaboradores são reportadas: Cluster I em espécies aeróbicas e facultativas, Cluster II em espécies anaeróbicas e Cluster III.

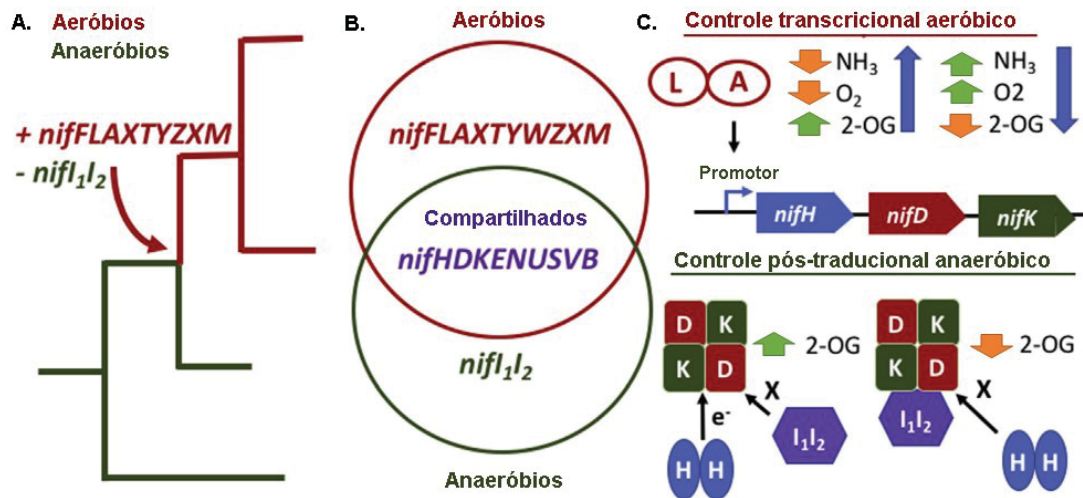
### 3.1.3 NIF EM FIXADORES ANAERÓBIOS E AERÓBIOS

Independente de quando as bactérias aeróbicas adquiriram a nitrogenase, o recrutamento e perda substanciais de genes *nif* acompanhou esse evento (representado pelos sinais “+” para recrutamento de genes e “-” para perda de genes

*nif*, FIGURA 4A), bem como uma reconfiguração dos mecanismos de transporte de elétrons para o CEN (BOYD, Eric S. *et al.*, 2015; POUDEL *et al.*, 2018). A codificação do complexo em linhagens anaeróbicas de evolução precoce (a exemplo, metanogênicos e clostrídios) tendem a abrigar núcleos *nif* menos complexos na maioria das vezes, constituído por 6 a 11 genes, os quais abrangem o conjunto *core nifHDKENB* e, para alguns casos, com o acréscimo de *nifV*, *nifU* e *nifS* e dos reguladores P-II *like* Nifl<sub>1</sub>, Nifl<sub>2</sub>. Em linhagens aeróbicas (a exemplo, a proteobactéria gama *Azotobacter vinelandii*) é aumentado substancialmente para quase 20 genes *nif*, os quais acrescentam-se ainda os genes *nifF*, *nifL*, *nifA*, *nifT*, *nifY*, *nifW*, *nifZ*, *nifX* e *nifM* (FIGURA 4B) (HAMILTON *et al.*, 2011; MUS *et al.*, 2019). Esse recrutamento de genes é necessário por codificarem produtos com papéis relacionados ao melhoramento e à maturação dos aglomerados de metais FeMo-co ou da proteína FeMo. Enquanto o aumento observado no número de genes *nif* e sua distribuição filogenética estão fortemente correlacionados com a adaptação para integrar O<sub>2</sub> em seus metabolismos de energia, o aumento não está correlacionado com qualquer um dos mecanismos de proteção O<sub>2</sub> conhecidos (BOYD, Eric S. *et al.*, 2015). Em vez disso, a maioria dos genes *nif* recrutados ou perdidos durante a transição da FBN de origens metabólicas anaeróbicas para aeróbicas parecem estar associadas à biossíntese de aglomerados de metais e à regulação da nitrogenase. Consistente com esta hipótese, a transição dos diazotrofos de ambientes anaeróbios para aeróbios está associada a uma mudança de regulação dos genes *nif* de pós-traducional dependente de Nifl<sub>1</sub> e Nifl<sub>2</sub> em anaeróbios para regulação transcricional dependente do produto NifA em aeróbios obrigatórios e anaeróbios facultativos (FIGURA 4C). Nesse contexto, tanto o aumento no número de genes *nif* quanto a mudança em seu modo de regulação durante a transição do metabolismo anaeróbio para aeróbio, pode se relacionar com a maior eficiência da síntese de proteínas Nif para atender a valores de FBN mais elevados. Dado que o N<sub>2</sub> fixado muitas vezes limita a produtividade do ecossistema, trabalhos indicam uma grande mudança na composição e na função dos *regulons nif*, enfatizando ainda mais a integração dinâmica entre a evolução dos ciclos biogeoquímicos de oxigênio, nitrogênio e carbono na Terra (MUS *et al.*, 2019; ZERKLE; MIKHAIL, 2017) .

FIGURA 4 – DISTRIBUIÇÃO DE NIF EM ORGANISMOS ANAERÓBIOS E

## AERÓBIOS/FACULTATIVOS.



FONTE: Adaptado de Mus (2019).

LEGENDA: (A.) distribuição filogenética de espécies aeróbicas (em vermelho) e anaeróbicas (em verde). (B.) Genes *nif* identificados em aeróbicos, anaeróbicos e compartilhados entre ambos (em azul). (C.) controle regulatório da nitrogenase, sendo em organismos aeróbicos mediado a nível transcricional por *nifLA* e, em anaeróbicos, a nível pós-traducional pelo parálogos *nifI<sub>1</sub>I<sub>2</sub>*.

### 3.2 TÉCNICAS E MÉTODOS DE IA APLICADA À BIOINFORMÁTICA

Neste capítulo elencamos as estratégias de Inteligência Artificial (IA) em Bioinformática na manipulação e análise de dados biológicos. Mais especificamente, abordaremos sobre Aprendizado de Máquina, Aprendizado Supervisionado e Não supervisionado, Redes Neurais Artificiais (RNAs), Redução de Dimensionalidade, *Spaced Words Projection* (SWeeP), Análise dos Componentes Principais (PCA), Agrupamento Hierárquico em Componentes Principais (HCPC), sistemática filogenética, *Rapid Alignment-Free Tool Sequence Similarity Search to Groups* (RAFTS<sup>3</sup>G): clusterização de homólogos e análise de sintonia genômica.

#### 3.2.1 APRENDIZADO DE MÁQUINA

Aprendizado de Máquina (*Machine Learning*) é um processo automático no qual possibilita aos computadores a habilidade de aprender sem serem explicitamente programados com base na experiência (SIMON, 2015). Basicamente, regras lógicas



são estabelecidas e visam melhorar o desempenho de uma tarefa ou, dependendo da aplicação, tomar decisão mais apropriada para determinado contexto. Os algoritmos de aprendizado de máquina implementam conjuntos de regras, baseados em métodos estatísticos, por exemplo, *Support-Vector Machine* (SVM), *K-nearest neighbors* (KNN), *k-means*, entre outros, os quais extraem padrões de grandes quantidades de dados (SARKER, 2021). Esses padrões reconhecidos permitem a criação de modelos de forma automática, os quais poderão ser usados no entendimento dos dados em si e até extrapolados no entendimento de dados externos (GHAHRAMANI, 2015). De modo geral, diz-se que a interferência humana no processo de aprendizado é mínima.

Além da escolha do modelo de aprendizado de máquina, é crucial a amostragem e o ajuste dos dados. Modelos que se ajustam excessivamente aos dados, tornam-se ineficientes para inferir dados novos os quais não se integraram ao conjunto utilizado no processo de aprendizado. Este fenômeno é chamado de *overfitting* e ocorre quando o modelo "decora" o padrão dos dados durante o processo de aprendizagem e não consegue extrapolar esse padrão para novos dados (JAMES, 2013). O fenômeno contrário também pode acontecer, assim sendo, quando o modelo de aprendizado não consegue encontrar as relações entre os dados durante a aprendizagem, caracteriza-se por *underfitting*.

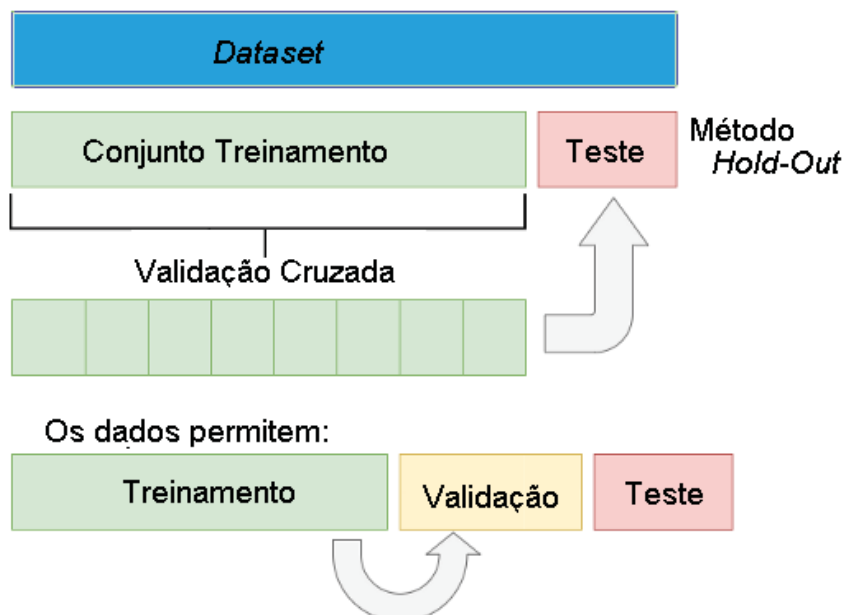
Para evitar tais fenômenos, é essencial que o conjunto total de dados disponível seja dividido em três subconjuntos para o ajuste do modelo:

- **Conjunto de treino:** como padrões apresentados para treinar o modelo, ou seja, definir estimativas dos parâmetros para o ajustamento do modelo de aprendizado.
- **Conjunto de validação:** para avaliar a performance de modelos com diferentes parâmetros através das conexões ponderadas.
- **Conjunto de teste:** para avaliar a performance final do modelo já ajustado.

O método *Hold-Out* é o mais usual na amostragem dos dados e caracterizado pela divisão do conjunto de dados (*Dataset*) em um único conjunto de treinamento/teste. Uma proporção comum é utilizar 80% dos dados para treinamento e os 20% restantes para teste. A validação cruzada é outra opção para amostragens que consiste em dividir aleatoriamente em grupos  $k$  um conjunto em que, um dos grupos é usado como conjunto de teste e os demais como conjunto de treinamento. O modelo é treinado no conjunto de treinamento e pontuado no conjunto de teste, em

seguida, o processo é repetido até que cada grupo único seja usado como conjunto de teste (KUMAR, 2021). Na FIGURA 6 encontra-se ilustrado o método de divisão dos conjuntos.

FIGURA 5 - MÉTODO DE DIVISÃO DO CONJUNTO EM TREINAMENTO/TESTE



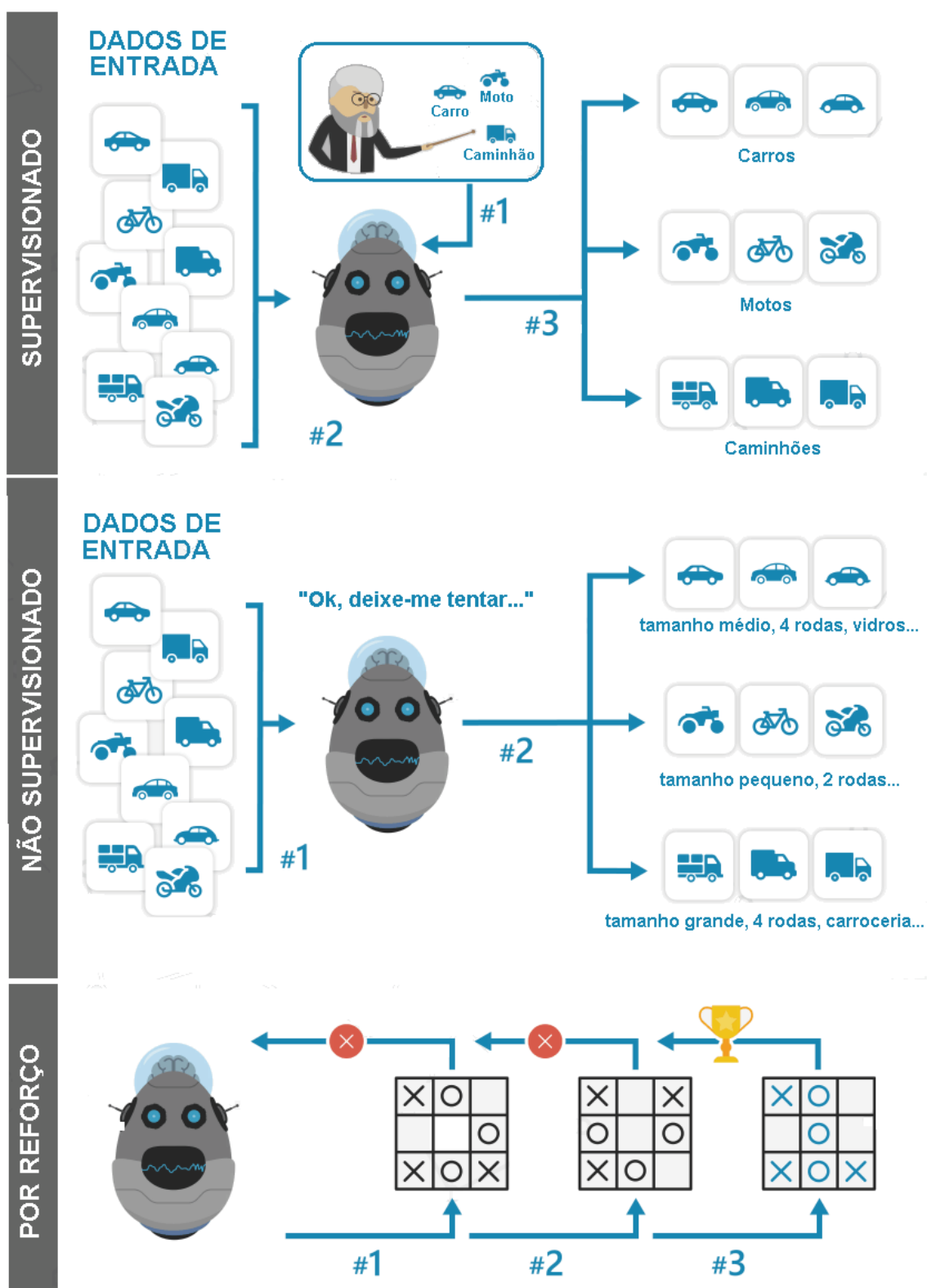
FONTE: O Autor (2022).

### 3.2.2 APRENDIZADO SUPERVISIONADO E NÃO SUPERVISIONADO

Os métodos de aprendizado podem ser divididos em três categorias principais (RUSSELL; NORVIG, 2013): aprendizado supervisionado, não supervisionado e por reforço (FIGURA 6). No aprendizado supervisionado são realizadas previsões mapeando os dados previamente rotulados. De regra geral, o algoritmo realiza ciclos e infere a classificação dos dados comparando-os com a classificação real. Os erros são usados para controlar o processo de aprendizagem, repetindo os ciclos até que o erro seja minimizado. No aprendizado não-supervisionado envolve a análise dos dados não rotulados e o padrão para a classificação é aprendido pelo algoritmo durante o processo de aprendizado. Várias hipóteses de classificação são testadas pelo algoritmo e a mais otimizada é utilizada. Por fim, no aprendizado por reforço, o agente interage com o ambiente através de observações, ações e recompensas. Ou seja, no lugar de usar dados para o aprendizado os quais mostram a classificação real correta, os dados são usados apenas como uma indicação sobre se uma ação é

correta ou não. Neste tipo de aprendizagem, a recompensa é relacionada ao conjunto das ações tomadas e não a ações individualizadas. Neste processo, o agente aprende a atingir uma meta em um ambiente incerto e potencialmente complexo (CHAKRABORTY; CHOUDHURY, 2017; JORDAN; MITCHELL, 2015; MNIH *et al.*, 2015).

FIGURA 6 – CATEGORIAS DE APRENDIZADO DE MÁQUINA



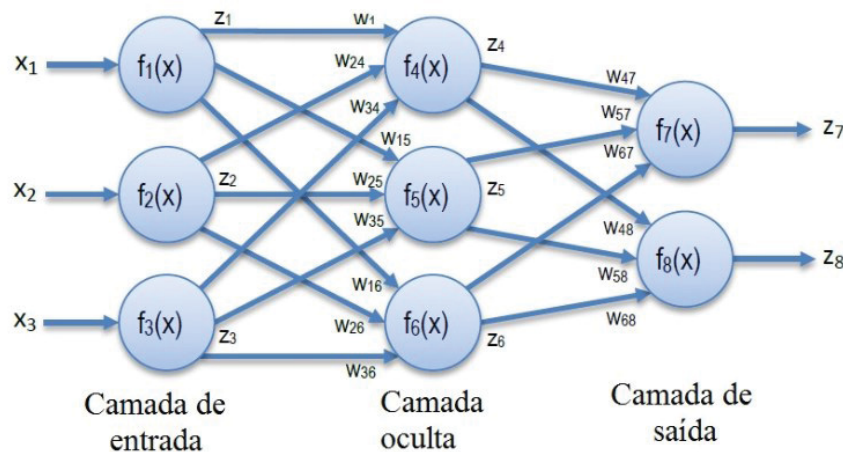
FONTE: *Busines Process Incubator* (2019).

LEGENDA: Aprendizado supervisionado em que o computador aprende a partir de dados previamente rotulados. Aprendizado não-supervisionado em que o computador agrupa os dados de acordo com o padrão encontrado durante o treinamento. Aprendizado por reforço em que o computador aprende através de recompensas e punições.

### 3.2.3 AS REDES NEURAIS ARTIFICIAIS

As Redes Neurais (ou neuronais) Artificiais (RNAs), são modelos de aprendizado de máquina influenciados pelas estruturas das redes neurais biológicas. Tipicamente, as redes são organizadas na forma de três camadas (entrada, intermediária ou oculta e saída) e compostas por unidades de processamento (neurônios) e conexões entre as unidades (sinapses) (GARCÍA KERDAN; MORILLÓN GÁLVEZ, 2020) (FIGURA 7). Os neurônios da camada de entrada recebem os padrões apresentados a ele, os neurônios da camada intermediária (ou oculta) são responsáveis pelo processamento e aprendizado das informações (conexões ponderadas) e os neurônios da camada de saída apresentam o resultado do processamento (VALAFAR, 2003). Durante o processo de aprendizado, a rede ajusta a intensidade de suas sinapses (peso) de acordo com o algoritmo de aprendizado e dos dados observados. Este procedimento é conhecido como Regra Delta (HU; HWANG, 2002).

FIGURA 7 – ESTRUTURA DE UMA REDE NEURAL ARTIFICIAL TÍPICA



FONTE: Adaptado de Ashwood (2013).

LEGENDA: Esquema de Rede Neural Artificial com neurônios distribuídos em 3:3:2 (neurônios da camada de entrada, oculta e da saída, respectivamente). Os *inputs* são representados pelas três entradas  $X_1$ ,  $X_2$  e  $X_3$ .  $f(x)$  representa a função de ativação para cada nó. A saída de cada nó é representada por  $Z$  e os pesos por  $W$ .

### 3.2.4 A REDUÇÃO DE DIMENSIONALIDADE

A redução de dimensionalidade tem por objetivo representar dados observados de alta dimensão em outra dimensão reduzida, porém, significativa dos dados originais, visando à dimensionalidade intrínseca correspondente ao número mínimo de parâmetros necessários para representar as propriedades dos dados observados (VAN DER MAATEN; POSTMA; VAN DEN HERIK, 2009). A aplicação de algoritmos, principalmente os mais extensos e complexos, é facilitada com esta compactação dos dados (ASGARI; MOFRAD, 2015). As técnicas mais usuais para a redução de dimensionalidade são:

- **Amostragem de dados:** indivíduos são amostrados da população, criando uma base de dados menor, mas representativa da população;
- **Seleção de atributos (*feature selection*):** alguns atributos são utilizados no processo analítico;
- **Redução com rotação de eixo:** a correlação dos dados é aproveitada para representá-los em um número menor de dimensões;
- **Redução com transformação de tipo:** a portabilidade do tipo do dado ocorre pela conversão em tipos de menor complexidade.

No presente trabalho, utilizamos técnicas de redução com rotação de eixo (e.g.: SWeeP e a análise de componentes principais (PCA), redução com transformação de tipo (e.g.: SWeeP) e amostragem de dados (e.g.: RAFTS<sup>3</sup>G).

### 3.2.5 SPACED WORDS PROJECTION (SWEEP)

O método SWeeP (*Spaced Words Projection*) explora a abordagem livre de alinhamento para representar sequências biológicas, convertendo-as na forma de vetores numéricos, o que reduz significativamente o volume de dados ao permitir representar sequências de um ou de vários organismos inteiros (DE PIERRI *et al.*, 2020). Na prática, SWeeP representa quaisquer informações contidas em um arquivo em formato FASTA ou multiFASTA (e.g.: o conjunto de caracteres correspondentes às sequências aminoacídicas) as quais são concatenadas e representadas em um vetor de dimensão (Vetor SWeeP) que mantém praticamente todas as informações de comparação. Concatenar as sequências consiste em flanquear, por um delimitador especial (e.g.: o delimitador “\*\*\*\*\*”) e o qual não será lido como caractere, cada

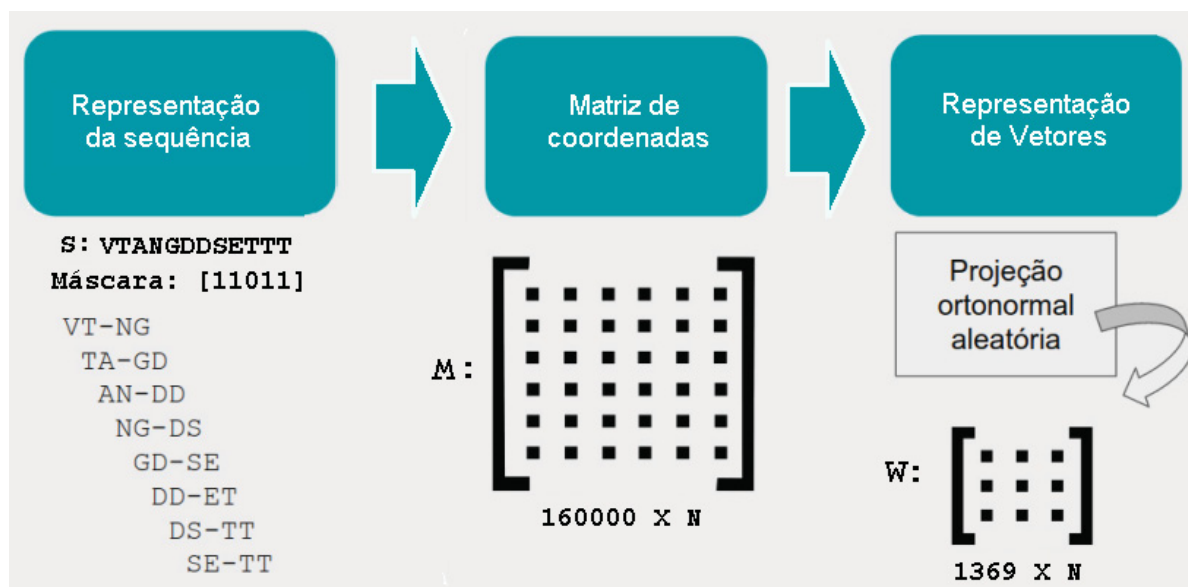
sequência contida no FASTA, construindo, assim, uma sequência unificada - o que definimos por proteoma<sup>3</sup> para o conjunto das proteínas unificadas. Cada proteoma é então representado por uma matriz bidimensional  $M$ , de palavras espaçadas de toda a sequência concatenada usando uma máscara pré-definida. A matriz  $M$  pode ser construída em dois formatos: 1) reversível, baseada na média geométrica dos números primos, ou 2) binária, que é econômica e de rápido processamento. A forma binária foi usada para os estudos descritos neste trabalho. A matriz  $M$ , quando linearizada em colunas, é um vetor que reflete dados de sequência altamente representativos, que designamos como um vetor de dimensão superior (HDV, do inglês *higher-dimensional vector*). O modelo SWeeP projeta esses HDVs em uma base quase ortonormal, de acordo com o lema de *Johnson-Lindstrauss* (JOHNSON; LINDENSTRAUSS; SCHECHTMAN, 1986), permitindo um desempenho eficiente com grandes conjuntos de dados. O método pode ser dividido em três processos principais (FIGURA 8):

- 1) **Representação da sequência:** a partir de janelas móveis, SWeeP utiliza as palavras espaçadas (análogo ao  $k$ -mer) sob uma configuração de máscara binária que representa a informação da sequência (S), por exemplo máscara "101", em que "1" é a leitura da sequência que será representada em uma matriz de coordenadas e "0" representa os "saltos" na leitura da sequência - permite tratar situações de mutações e inclusões de caracteres na sequência;
- 2) **Matriz de Coordenadas:** A matriz de coordenadas (M), que possui tamanho 160000 (400 x 400), é convertida em uma base quase-ortonormal orientada aleatoriamente, o qual possui número de coordenadas suficientes para manter a comparabilidade das sequências (vetor de dimensionalidade);
- 3) **Representação de vetores:** O vetor de dimensionalidade é projetado em outro vetor (W) que mantém a maioria das informações de comparação, ideais para várias tarefas de visualização. Essa etapa é pré-definida pelo usuário.

---

<sup>3</sup> No uso de SWeeP, o termo proteoma também é associado às sequências de proteínas unificadas para cada organismo. Por exemplo, o conjunto Nif *core* pode ser visto, neste caso, como proteoma do Nif, pois as sequências NifH, NifD, NifK, NifE, NifN e NifB são concatenadas e unificadas.

FIGURA 8 - FLUXOGRAMA DE SWEEP NA OBTENÇÃO DE VETOR DE SEQUÊNCIAS.



FONTE: Adaptado de Kulik (2020).

LEGENDA: Uma sequência qualquer de proteína (S) é lida por uma janela móvel de máscara binária, neste exemplo [11011]. Os caracteres de S lidos são representados em uma matriz  $M$  que é projetada com a base randômica quase-ortonormal gerada com 160.000 coordenadas (compatível com a projeção da máscara) em outro vetor ( $W$ ) de tamanho predefinido.

### 3.2.6 ANÁLISE DE COMPONENTES PRINCIPAIS (PCA)

A Análise dos Componentes Principais (PCA, abreviação de *Principal Component Analysis*) é um método exploratório, pois auxilia na elaboração de hipóteses gerais a partir dos dados coletados e é capaz de separar a informação importante da redundante e aleatória. Outrossim, é uma abordagem clássica de redução de dimensão dos dados que constrói combinações lineares, denominadas de Componentes Principais (PCs), de modo que seja facilitada a aplicação de algoritmos, sobretudo, os mais complexos (AGGARWAL, 2015). Os PCs são ortogonais entre si e podem explicar efetivamente os padrões principais das variações presentes nos dados originais (JOLLIFFE; CADIMA, 2016). Em PCA, os valores exatos de cada dado obtido não indicam necessariamente a natureza do valor (valores positivos ou negativos) e sim o quanto duas grandezas podem contribuir para o terceiro componente (vetor) gerado. A técnica constrói uma representação de baixa dimensão dos dados através de combinações lineares que descreve a maior variância possível das variáveis originais (VAN DER MAATEN; POSTMA; VAN DEN HERIK, 2009),

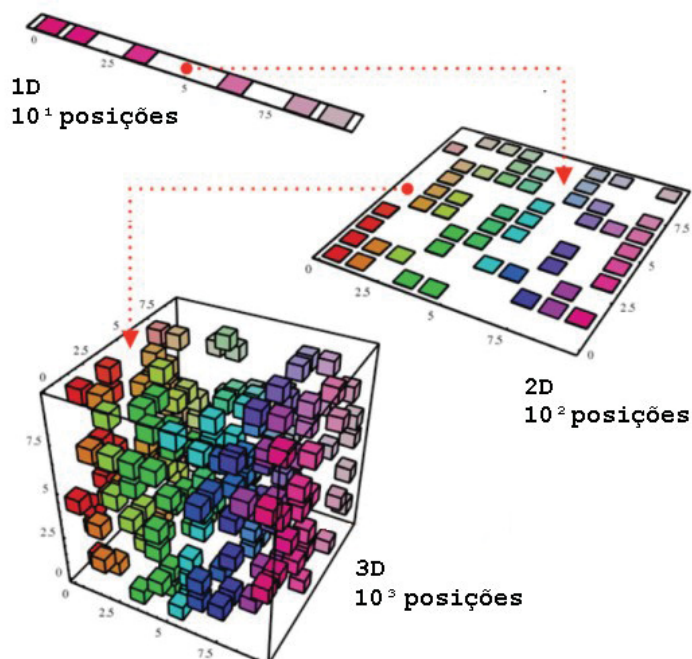


portanto, valores próximos a zero indicam que os componentes gerados não possibilitaram uma boa diferenciação dos dados (SAS, 2004). Os passos para calcular as componentes principais são:

- obter os dados ou as  $M$  amostras de vetores de dimensão  $n$ ;
- calcular a média ou o vetor médio destes dados;
- subtrair a média de todos os itens de dados;
- calcular a matriz de co-variância usando todas as subtrações (ela é o resultado da média do produto de cada subtração por ela mesma e terá dimensão  $n \times n$ .);
- calcular os auto valores e autovetores da matriz de covariância.
- Arranjar a matriz da Transformada de *Hotelling* cujas linhas são formadas a partir dos autovetores da matriz de covariância arranjados de modo que a primeira linha, o elemento (0,0), seja o auto vetor correspondente ao maior auto valor, e assim sucessivamente até que a última linha corresponda ao menor auto valor (HOTELLING, 1933).

Os PCs gerados irão englobar grande parte das variações observadas dentro das variáveis originais e podem ser utilizados com critérios de variabilidade em análises subsequentes (FIGURA 9) (JOLLIFFE; CADIMA, 2016; TODOROV; FOURNIER; GERBER, 2018). Estudos sugerem que os PCs podem corresponder a causas latentes, entretanto, devido a nossa literatura escassa, a seleção de variáveis e de métodos híbridos permitem desfrutar de melhor interpretabilidade dos dados, a exemplo do Agrupamento Hierárquico em Componentes Principais (HCPC, sigla para *Hierarchical Clustering on Principal Components*) (S; Y, 2011).

FIGURA 9 – MAPA DE FATORES DE UMA PCA EM 1, 2 E 3 DIMENSÕES.



FONTE: Adaptado de *Free Code Camp* (2019).

LEGENDA: Representação de uma distribuição de dados em uma PCA (mapa de fatores) em 1D, 2D e 3D.

### 3.2.7 AGRUPAMENTO HIERÁRQUICO EM PCA (HCPC)

A primeira vantagem agrupamento hierárquico em componentes principais (HCPC) é que ele envolve a aplicação de técnicas de agrupamento específicos aos resultados da análise de componentes principais, o que leva a uma melhor solução de agrupamento sobre o uso da análise fatorial somente (HUSSON; JOSSE; PAGÈS, 2010). Uma segunda vantagem é a utilização de um algoritmo misto para o processo de agrupamento – uma combinação do método de classificação de Ward com o algoritmo *K-means* – que melhora a robustez dos resultados finais (ARGÜELLES; BENAVIDES; FERNÁNDEZ, 2014).

O HCPC (FIGURA 10) pode ser dividido em quatro etapas (KASSAMBARA, 2017), a saber:

- 1) **Cálculo de Componentes Principais:** PCA, MCA ou MFA dependendo dos tipos de variáveis no conjunto de dados e da estrutura do conjunto de dados. Nesta etapa, você pode escolher o número de dimensões a serem retidas na saída;
- 2) **Cálculo do agrupamento hierárquico:** o agrupamento hierárquico é

realizado usando o critério de Ward nos componentes principais selecionados. O critério Ward é usado no agrupamento hierárquico porque é baseado na variância multidimensional como análise de componentes principais.

3) **Escolha do número de clusters com base na árvore hierárquica:** um particionamento inicial é realizado cortando a árvore hierárquica.

4) **Execução do agrupamento por *K-means*:** para melhorar a partição inicial obtida do agrupamento hierárquico. A solução final de particionamento, obtida após consolidação com *k-means*, pode ser (ligeiramente) diferente daquela obtida com o agrupamento hierárquico.

*K-means*, por sua vez, utiliza um método simples e eficiente baseado no cálculo de distância. de forma iterativa e atribui os pontos de dados ao grupo que representa a menor distância, ou seja, ao grupo de dados que seja mais similar (SYAKUR *et al.*, 2018). O processo executado pelo *k-means* é composto por quatro etapas (PAUL, 2018):

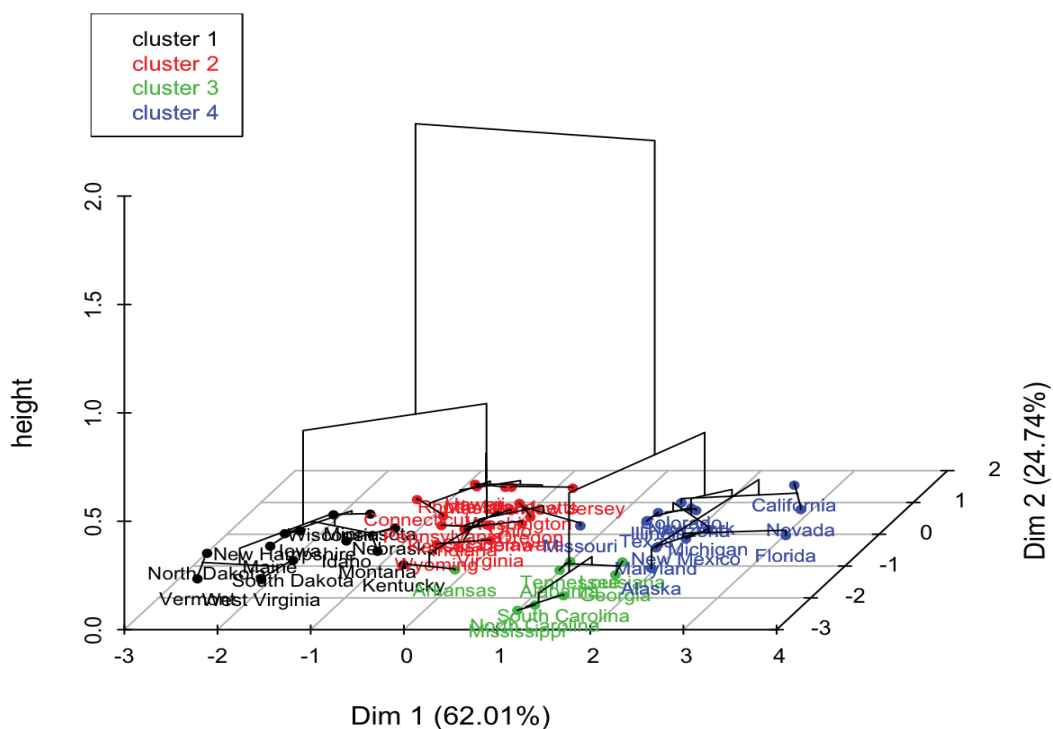
1) **Inicialização:** o algoritmo gera de forma aleatória  $k$  centroides, onde o número de centroides é representado pelos pontos centrais do *cluster* (grupo),

2) **atribuição ao *cluster*:** em que é calculado a distância entre todos os dados e cada um dos centroides (O cálculo de distância é utilizado através da distância Euclidiana). Finalizando esta etapa com os dados divididos conforme o número de centroides estipulado pelo argumento  $k$ ;

3) **Movimentação de centróides:** uma vez que os pontos de dados foram atribuídos aos *clusters* conforme sua distância, o próximo passo é recalcular o valor dos centróides. Nesta etapa é calculada a média dos valores dos pontos de dados de cada cluster e o valor médio será o novo centróide. O termo movimentação se refere a alteração da localização do centróide em um plano se pensarmos em um gráfico;

4) **Otimização do *K-means*:** na fase final da execução do *K-means* as fases Atribuição ao Cluster e Movimentação de Centroides são repetidas até o cluster se tornar estático ou algum critério de parada tenha sido atingido. O cluster se torna estático quando nenhum dos pontos de dados alteram de cluster. Um critério de parada pode ser o número de iterações máximas que o algoritmo irá fazer durante a fase de otimização.

FIGURA 10 – CLUSTERIZAÇÃO HIERÁRQUICA NO MAPA DE FATORES



FONTE: STHDA (2017).

LEGENDA: obtenção de 4 *clusters* por HCPC a partir de um mapa de fatores (PCA) em um estudo com dados de estados americanos. Neste exemplo é exibido o percentual (%) explicado por cada componente.

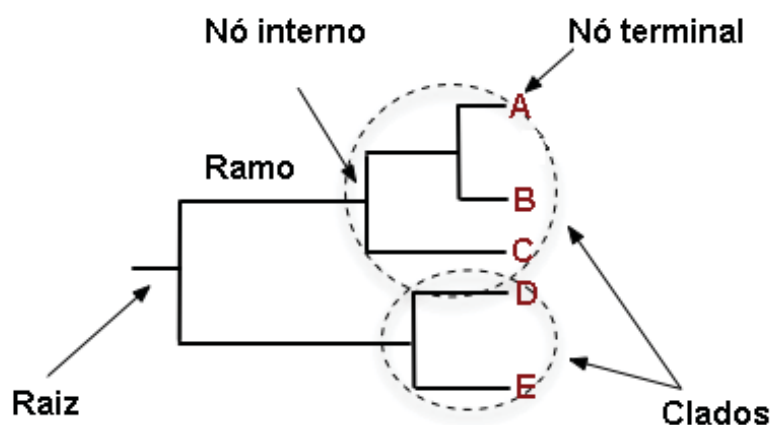
### 3.2.8 SISTEMÁTICA FILOGENÉTICA

A sistemática é o campo de pesquisa da biologia que fornece nomes científicos para organismos, descreve-os, classifica-os, investiga suas histórias evolutivas, considerando suas adaptações ambientais e preserva coleções de sua biodiversidade (RICHARD FORTEY, 2016; SCHUH; BROWER, 2009). A sistemática filogenética, ou filogenia, é o ramo da sistemática que atribui metodologia de classificação para reconstruir a história evolutiva e estudar os padrões (ou “pistas”) de relacionamento entre os organismos (PROCTER *et al.*, 2010). Este conhecimento pode ser empregado nas áreas de bioinformática, de ciências forenses e de conservação (EMERY, 2019).

As árvores filogenéticas representam as relações evolutivas e os nós terminais representam as unidades taxonômicas dos organismos. Cada nó interno é uma unidade taxonômica hipotética, o qual simboliza um antepassado comum aos

ramos que descendem deste nó. A raiz da árvore retrata o ancestral comum de todas as unidades taxonômicas representadas. Quando a direção do processo evolutivo e o ancestral comum a todas as unidades taxonômicas não são indicados, a árvore é denominada por não-enraizada. Em algumas árvores, o tamanho de ramo pode representar a quantidade de mudanças ou o tempo ocorrido entre os nós (LEMEY; SALEMI; VANDAMME, 2009). Na FIGURA 11 ilustramos um esquema de árvore enraizada e sua composição.

FIGURA 11 – ESTRUTURA DE UMA ÁRVORE FILOGENÉTICA ENRAIZADA



FONTE: O Autor (2022).

LEGENDA: Na árvore enraizada o ancestral comum é apresentado. A, B, C, D e E são as unidades taxonômicas dos organismos vivos. Os nós internos representam as unidades taxonômicas hipotéticas. Os organismos ABC e DE se organizam em clados distintos.

A construção da filogenia por marcadores moleculares é composta por duas etapas principais: o alinhamento de sequências e o processamento da árvore a partir do alinhamento realizado (WARNOW, 2018). Na construção de filogenias extensas utiliza-se o método de super-árvore, em que árvores de grupos menores (sub-táxons) são combinadas para evitar o processamento simultâneo de uma quantidade grande de dados. Isso se deve, pois o alinhamento de sequências e os métodos de construção de árvores de maior acurácia requerem grande capacidade computacional. Portanto, a filogenia completa obtida pelo método de super-árvore reflete a relação entre as espécies dos táxons menores e não do táxon como um todo (HAESELER, 2012). O método de super-árvore consiste em três etapas (HALGASWATHTHA *et al.*, 2012; WARNOW, 2018): 1) o alinhamento múltiplo de

sequências entre as espécies compreendidas nos sub-táxons, 2) a construção da topologia das árvores dos sub-táxons e, por fim, 3) a construção da filogenia do táxon completo através da combinação das árvores geradas para os sub-táxons.

Além da filogenia em si, o processamento é imprescindível para o ajuste da topologia das árvores e super-árvores. Os principais são (HAUBOLD, 2014; LEMEY; SALEMI; VANDAMME, 2009):

- **Métodos de parcimônia:** identificam a topologia a qual requer a menor taxa de mudanças de estado para produzir as características das unidades taxonômicas observadas, (e.g.: máxima parcimônia). (SOBER, 2004);

- **Métodos de probabilidade:** através de modelos evolutivos, encontram a topologia com a maior probabilidade de gerar as características das unidades taxonômicas observadas (e.g.: máxima verossimilhança e inferência Bayesiana) (RANNALA; YANG, 1996);

- **Métodos de distância:** ajustam a topologia a uma matriz de distância estimada através das verdadeiras distâncias genéticas (e.g.: UPGMA, sigla para *Unweighted Pair Group Method with Arithmetic mean* e NJ, abreviação para *Neighbor-joining*) (SAITOU; NEI, 1987; WEISS; GÖKER, 2011).

### 3.2.9 RAFTS<sup>3</sup>G: CLUSTERIZAÇÃO DE HOMÓLOGOS

A clusterização (ou agrupamento) é uma forma de aprendizado do tipo não supervisionado, ou seja, que não trabalha com dados rotulados (ANDREOPOULOS *et al.*, 2009). RAFT<sup>3</sup>G (DE LIMA NICHIO *et al.*, 2019) é uma técnica de agrupamento que visa a redução da dimensionalidade em grandes conjuntos com o mínimo de degradação da informação biológica. Na prática, a partir de um arquivo de entrada FASTA contendo as sequências proteicas não agrupadas, RAFTS<sup>3</sup>G explora a base de dados formatada segundo o mecanismo de RAFTS3 para a busca de sequências com similaridade (VIALLE *et al.*, 2016). Para selecionar sequências para o mesmo agrupamento (ou *cluster*), as sequências de consulta são avaliadas em pares. A primeira sequência, a maior, torna-se a primeira representativa do *cluster* e as demais similares a ela são inseridas ao *cluster* se respeitarem a condição do critério de seleção ou valor de corte (*self-score*) escolhido pelo usuário. O usuário pode escolher:

- 1) valor de *self-score* obtido pelo alinhamento da sequência consulta até um número

limitado de  $n$  candidatos ordenados da matriz binária de Co-ocorrência (BCOM) ou 2) fazendo busca binária alinhando  $n$  candidatos para encontrar o ponto de corte em que todas as sequências de ordem inferior deverão ser tão semelhantes ou maiores que a sequência neste ponto. As sequências de ordem superior são, provavelmente, menos semelhantes do que as estipuladas pelo critério de corte e são deixadas fora do agrupamento. O principal ganho da abordagem da busca binária é permitir a construção de um agrupamento em menos etapas, uma vez que encontra a maioria das sequências relacionadas com uma consulta em uma única pesquisa, alinhando apenas um número relativamente pequeno de candidatos ( $O(\log_2(n))$ ). Em ambos os casos, temos uma lista de sequências a agrupar que devem ser pelo menos tão semelhantes à consulta quanto a medida definida no corte.

Após a etapa de corte, o algoritmo revisa os grupos montados com base nas sequências para agrupar:

a) se a consulta encontrou sequências já agrupadas, todos os grupos encontrados são reunidos em um único e todas as demais sequências a agrupar são adicionadas neste grupo;

b) se nenhuma das sequências a agrupar for membro de um grupo criado anteriormente, um novo grupo será construído e essas sequências serão adicionadas a ele.

Enquanto houver sequências a serem analisadas, essas etapas serão repetidas para cada uma delas.

Devido a sua estrutura, RAFTS<sup>3</sup>G possui notável desempenho em valores de *self-score* 0.4 e 0.5 (F1-score de 0.91) e uma estratégia aliada a identificação de sequências homólogas, mesmo remotamente conservadas, em grande conjunto de dados.

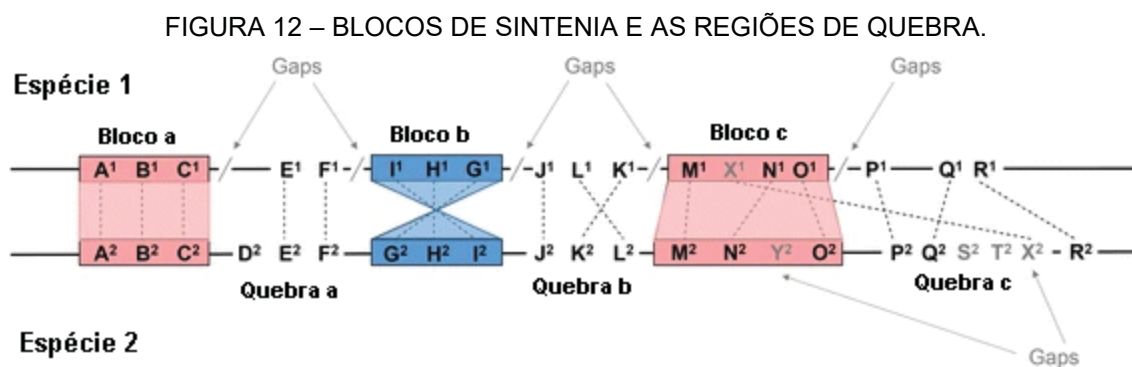
### 3.2.10 ANÁLISE DE SINTENIA GENÔMICA

O bloco de sintenia (que etimologicamente significa “na mesma fita”) é uma coleção de genes contíguos localizados no mesmo cromossomo. Esses blocos foram evolutivamente preservados, principalmente, por rearranjos do genoma e, assim, blocos de sintenia de duas espécies semelhantes podem estar preservados, porém invertido no cromossomo de um em relação ao outro. Ovcharenko e colaboradores (2005) definem como ‘qualquer’ bloco de sequência conservada, independente de

englobar múltiplos genes, uma área contendo genes únicos ou áreas desprovidas de genes conhecidos a serem considerados como bloco de sintenia, desde que haja conservação a nível de sequência. Portanto, pode-se referir também à presença de genes sintênicos em grupos ancestralmente relacionados, descrevendo assim o estado de sintenia conservada (OVCHARENKO *et al.*, 2005; STEIN, 2013).

Em genômica comparativa, sintenia refere-se à conservação de blocos ordenados de genes dentro de dois cromossomos, os quais estão sendo comparados entre si (PASSARGE; HORSTHEMKE; FARBER, 1999). Tais comparações comumente são feitas utilizando um software de alinhamento, tal como o *Basic Local Alignment Search Tool* (BLAST). O passo seguinte, frequentemente, consiste em utilizar alguma ferramenta de visualização gráfica dos resultados dos alinhamentos obtidos, de modo a facilitar a identificação das sintenias (COIMBRA, 2010).

Conforme ilustra a FIGURA 12 definimos para este trabalho a estrutura dos possíveis blocos de sintenia (LIU; HUNT; TSAI, 2018). Neste exemplo, os blocos estão sob o critério de pelo menos três genes ortólogos (denominados por âncoras) e podem ser divididos em: ortólogos com a mesma ordem entre as espécies (FIGURA 12, Bloco a), sintenia com ordem inversa (FIGURA 12, Bloco b) ou permitindo algumas lacunas (FIGURA 12, Bloco c). Em contraste, as situações de causa da quebra da sintenia podem ser: pela falta de ortólogos (Quebra a), pela ordem de genes co-arranjados (FIGURA 12, Quebra b) ou pelas lacunas (FIGURA 12, Quebra c).



FONTE: Adaptado de Liu, Hunt e Tsai (2018).

LEGENDA: Genes localizados em cromossomos de duas espécies são indicados em letras. Cada gene está associado a um número que representa a espécie a que pertence (espécie 1 ou 2). Os genes ortólogos são conectados por linhas tracejadas e os genes sem uma relação ortóloga são tratados como lacunas em programas de sintenia. Um bloco de sintenia pode ser ortólogos com a mesma ordem (Bloco a), com ordem inversa (Bloco b) ou permitindo lacunas (Bloco c). Em contraste, os casos com causa de quebra de sintenia podem ser devido à falta de ortólogos (Quebra a), ordem de genes co-



arranjados (Quebra b) ou lacunas (Quebra c)

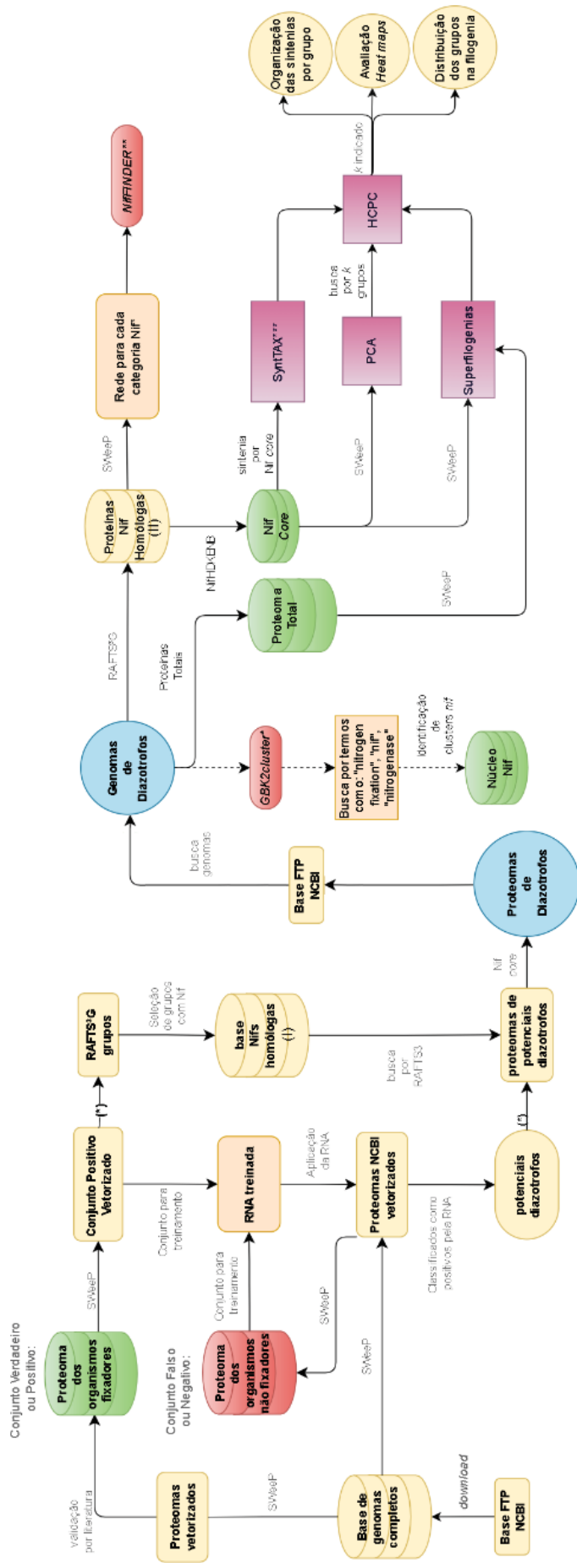
A análise sintênica, portanto, torna-se um método inestimável para estabelecer a ortologia das regiões genômicas em diferentes espécies bem como para inferir relações funcionais entre os genes (OBERTO, 2013).

## 4 MATERIAIS E MÉTODOS

Essa seção foi estruturada na mesma ordem em que as etapas do trabalho foram desenvolvidas. Estão descritos: 4.1) Definindo conjunto Proteoma e Nif core; 4.2) Proteomas bacterianos obtidos pelo NCBI; 4.3) Proteomas do NCBI representados em base vetorial; 4.4) Diazotrofos obtidos pela literatura; 4.5) Mineração de fixadores candidatos; 4.6) Proteínas Nif homólogas identificadas por RAFTS<sup>3</sup>G (I); 4.7) Predição de Nif em diazotrofos candidatos; 4.8) Validação de diazotrofos por revisão sistemática; 4.9) Nif homólogas identificadas por RAFTS<sup>3</sup>G (II); 4.10) PCA SWeeP do conjunto Nif core; 4.11) Aplicação do método HCPC em PCA; 4.12) Mapas de calor; 3.13) Filogenia SWeeP e 4.14) Sintenia por SyntTax.

Na FIGURA 13 encontra-se o fluxograma de toda a seção metodológica.

FIGURA 13 - FLUXOGRAMA METODOLÓGICO



FONTE: O Autor (2022).

LEGENDA: Base de genomas completos de bactérias obtido do NCBI e manipulação por SWveeP. Extração do conjunto treinamento (positivo e negativo) para classificação da RNA. Busca de Nif por homologia nos Proteomas dos organismos candidatos. (\*) representa a etapa de reversibilidade do vetor SWveeP para o retorno da informação em FASTA. A estratégia de busca por termos para extração de *clusters nif* (GBK2cluster\*) foi produzida, porém não utilizada neste trabalho. Com a obtenção dos conjuntos das sequências Nif homólogas foi possível a criação do preditor (NifFINDER\*\*). Com os dados Nif core, utilizamos a PCA e HCPC visando a criação de grupos e procuramos, com auxílio da sintonia (SyNTAX\*\*), por padrões moleculares similares nos grupos.

#### 4.1 DEFININDO CONJUNTO PROTEOMA E NIF CORE

Elencamos as principais abordagens de conjuntos de dados, para serem analisadas e correlacionadas nesse estudo através de estratégias da integração de SWeeP com outros métodos (e.g.: PCA, HCPC e Filogenia).

**Conjunto Proteoma:** foram obtidos os Proteomas a partir de genomas completos procarióticos.

**Conjunto Nif core:** a obtenção das 6 sequências que integram o Nif core: NifH, NifD, NifK, NifE, NifN e NifB. Incluímos proteínas bifuncionais relacionadas a fusão de genes, a exemplo, as proteínas NifEN e NifNB. Em casos de proteínas repetidas ou duplicadas (e.g.: NifH), desconsideramos para esta seleção, as sequências sem relação de homologia ou sintênica com as demais sequências core supracitadas.

#### 4.2 PROTEOMAS BACTERIANOS OBTIDOS PELO NCBI

O acesso do acervo dos proteomas da base de organismos bacterianos do NCBI deu-se conforme material descrito e publicado no trabalho de De Pierri et al (2020). No total foram utilizados 10.324 genomas completos extraídos da base do NCBI (versão de 30 agosto de 2018).

#### 4.3 PROTEOMAS REPRESENTADOS EM BASE VETORIAL

Extraímos todos os Proteomas para cada um dos 10.324 genomas completos baixados na seção 4.2. Para a manipulação dos dados, utilizamos SWeeP (disponível em: <https://sourceforge.net/projects/spacedwordsprojection/>) com os seguintes parâmetros: máscara “11011” para palavras espaçadas e projeção de 1369 coordenadas. Foi utilizado o servidor cedido pelo PPG em Bioinformática da UFPR com configurações descritas no APÊNDICE, TABELA 1.

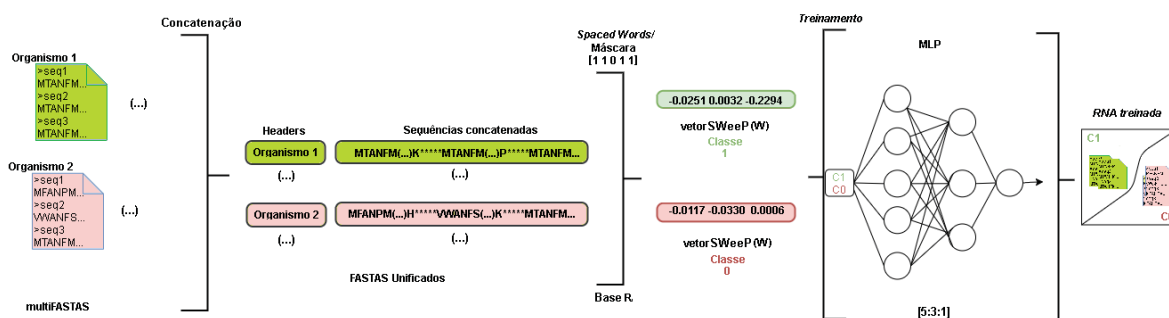
#### 4.4 DIAZOTROFOS OBTIDOS PELA LITERATURA

A coleta de procariotos diazotróficos teve como objetivo a aquisição de uma lista de organismos validados através da identificação do conjunto *core* da Monitrogenase ou ensaio bioquímico (e.g.: teste-padrão de redução de acetileno). Como referencial, utilizamos as literaturas de Dos Santos e colaboradores (2012) e o catálogo descrito em Chapman & Hall (2005). Dessa literatura, foram extraídos os dados genômicos disponíveis no NCBI. Extraímos os Proteomas de cada genoma e os vetorizamos, utilizando as mesmas configurações descritas na seção 4.3 para SWeeP.

#### 4.5 MINERAÇÃO DE FIXADORES CANDIDATOS

O objetivo desta etapa foi o de criar estratégia de mineração de dados na busca por potenciais genomas de diazotrofos. Consistiu em convertermos os proteomas de fixadores obtidos da literatura em vetores SWeeP, conforme descrito na seção 4.4, os quais utilizamos para comporem o conjunto Classe 1, ou verdadeiro, no treinamento e classificação da rede neural artificial. Selecionamos a classe de rede neuronal *MultiLayer-Perceptron* (MLP), configurado em 5:3:1 (valores para a camada de entrada, camadas ocultas e a camada de saída, respectivamente). O conjunto Classe 0 compôs-se de 5000 proteomas de espécies não fixadoras (teste do tipo não independente, treinamento realizado de modo a depurar). As etapas da criação da RNA SWeeP podem ser ilustradas na FIGURA 14. Desenvolvemos a RNA em ambiente MATLAB® licenciada para a versão 2019b (<https://www.mathworks.com/products/matlab.html>) com uso de *scripts* padrões. Utilizamos o servidor do PPG em Bioinformática da UFPR (APÊNDICE, TABELA 1).

FIGURA 14 – ETAPAS DE DESENVOLVIMENTO DO MODELO DA RNA SWEEP.



FONTE: O Autor (2022).

LEGENDA: As sequências contidas em multiFASTAs foram concatenadas e delimitadas por um caractere especial (\*\*\*\*\*) em uma única sequência (Proteoma) para dados de Fixadores (FASTAS em verde) e de não fixadores (FASTAS em vermelho). Foram obtidos os Vetores SWeeP para Fixadores, os quais compuseram o *input* da classe 1 e os vetores dos não fixadores o *input* da classe 0 para treinamento da RNA por MLP [5:3:1].

#### 4.6 PROTEÍNAS NIF HOMÓLOGAS IDENTIFICADAS POR RAFTS<sup>3</sup>G (I)

A ferramenta RAFTS<sup>3</sup>G (DE LIMA NICHIO *et al.*, 2019) (disponível em: <https://sourceforge.net/projects/rafts-g/>) foi adotada para este trabalho para a obtenção de agrupamentos de proteínas Nif homólogas, partindo da identificação dos 22 classes de Nif conhecidas, a saber: NifH, NifD, NifK, NifE, NifN, NifB, NfX NifU, NifS, NifV, NifP, NifW, NifZ, NifM, NifT, NifF, NifJ, NifO, NifA NifL, Nifl<sub>1</sub> e Nifl<sub>2</sub>. A partir dos Proteomas de diazotrofos obtidos na seção 4.4 foram gerados agrupamentos homólogos e extraídos aqueles contendo cada uma das 22 classes Nif. As informações dos grupos geraram a primeira base de Nif (I). O parâmetro de RAFTS<sup>3</sup>G foi de 0.4 de similaridade entre as sequências com busca binária.

#### 4.7 PREDIÇÃO DE NIF EM DIAZOTROFOS CANDIDATOS

Após obtermos a base de proteínas Nif (I), conforme descrito em 4.6 criamos, utilizamos a ferramenta de busca RAFTS3 (de modo análogo ao realizado com a ferramenta BLAST) para busca de sequências Nif nos organismos candidatos selecionados pela rede. Na prática, a base Nif serviu como estrutura de consulta (*subject*) para ferramenta RAFTS3 (VIALLE *et al.*, 2016). O objetivo da busca de proteínas Nif nos organismos candidatos foi para identificar as 6 estruturas Nif *core* da Mo-nitrogenase como um segundo critério na seleção de novos organismos fixadores. Utilizamos a ferramenta de busca RAFTS3 (disponível em: <https://sourceforge.net/projects/rafts3>) nos fixadores candidatos trazidos pela etapa 4.5 desta seção.

#### 4.8 VALIDAÇÃO DE DIAZOTROFOS POR REVISÃO SISTEMÁTICA

Após a obtenção e a confirmação das sequências Nif *core* nos proteomas dos organismos, realizamos uma varredura sistemática por literatura contendo estudo relativo aos organismos fixadores, validados entre os períodos de 2012 a 2022. Utilizamos três principais fontes de busca, sendo as bases do *Expasy* (SIB *institute*) (disponível em: <https://www.expasy.org/>), *PubMed*® (NCBI) (disponível em: <https://pubmed.ncbi.nlm.nih.gov/>), e *Scholar* (Google) (disponível em: <https://scholar.google.com.br/>).

#### 4.9 NIF HOMÓLOGAS IDENTIFICADAS POR RAFTS<sup>3</sup>G (II)

Com a obtenção de novos proteomas de fixadores, foi realizada nova estratégia clusterização das sequências para identificação de grupos homólogos de Nif (II). O parâmetro de RAFTS<sup>3</sup>G foi de 0.5 de similaridade entre as sequências com busca binária.

#### 4.10 PCA SWEEP DO CONJUNTO NIF CORE

A versão escrita em linguagem R de SWEEP, rSWEEP (disponível em <http://bioconductor.riken.jp/packages/3.11/bioc/html/rSWEEP.html>), foi utilizada para a integração dos vetores com as funções em ambiente R (versão 4.1.2 ) utilizadas para a construção de PCA do conjunto Nif *core* (disponível em: <https://www.r-project.org/>). Visualizamos a PCA em 2D construída através do pacote *Factoshiny* (disponível em: <http://factominer.free.fr/graphs/factoshiny.html>) e em 3D função plotPCA (APÊNDICE, FIGURA 4).

#### 4.11 APLICAÇÃO DO MÉTODO HCPC EM PCA

Assim como a PCA, a qual nos permite desenhar um panorama da distribuição da Mo-nitrogenase nos diazotrofos, o método HCPC permite ir além ao permitir identificarmos padrões ou agrupamentos de organismos os quais se relacionam e podem nos trazer informações implícitas e inerente aos dados Nif *core*. Obtemos

estimativas para o número de grupos ( $k$ ) pela metodologia de HCPC e comparamos também com estatística *Gap* (TIBSHIRANI; WALTHER; HASTIE, 2001), método Cotovelo (THORNDIKE, 1953) e coeficiente de silhueta (ROUSSEEUW, 1987). Foi estipulado o valor máximo de agrupamentos em 20 *clusters* devido a convergência dos dados em valores superiores a 20. Todas as metodologias foram obtidas em ambiente R, sendo que para HCPC utilizamos a função *consensusclusterplus* (disponível pelo bioconductor em: <https://bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html>) e para as metodologias estatística *gap*, método cotovelo e coeficiente da silhueta usamos a função *fviz\_nbclust* (disponível através do pacote *factoextra* em: <https://rpkgs.datanovia.com/factoextra/index.html>).

#### 4.12 MAPAS DE CALOR

A utilização de matrizes de distâncias dividiu-se em dois momentos:

1) na criação de mapa de calor ordenado obtido do resultado da matriz de dissimilaridade ordenada para a verificação da consistência das espécies agrupadas em auxílio a metodologia da seleção de  $k$  (3.10), com o objetivo de identificar grupos de organismos mais coesos e, a partir deles e dos  $k$  obtidos, elencar o valor de  $k$  mais ponderado na divisão de grupos. Para a criação da matriz, utilizamos a função *coldiss* (BOCARD; GILLET; LEGENDRE, 2018) em ambiente R (APÊNDICE, FIGURA 7).

2) para a criação de mapa de calor obtida da matriz de distância euclidiana ordenada para os dados dos grupos Nif e do Proteoma. O objetivo foi para a verificação das informações presentes do Nif *core* possuíam algum grau de divergência evolutiva em relação às informações do Proteoma. Neste caso, utilizamos *scripts* padrões em ambiente MATLAB® e foi construído dois mapas de calores: a) a distância euclidiana entre as informações de Nif *core* versus a distância do Proteoma e b) a distância do Proteoma versus a distância do Nif *core* dos grupos obtidos.

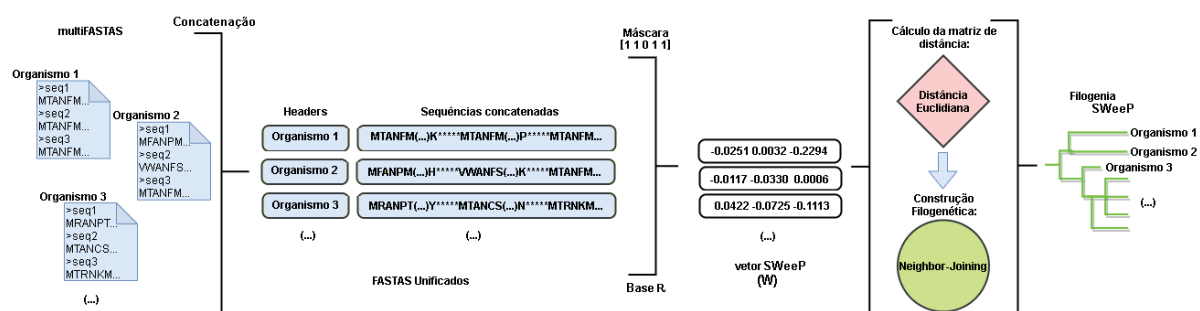
#### 4.13 FILOGENIA SWEEP

Construímos os vetores SWEEP para todos os proteomas de diazotrofos



obtidos em duas abordagens principais: 1) com dados do Proteoma e 2) com dados do Nif core. Os parâmetros usados para SWeeP foram: palavras espaçadas (*Spaced Words*) de máscara “11011” em matriz (W) de 1369 coordenadas. Para cada uma das duas abordagens, foi calculada a distância euclidiana entre os vetores dos organismos. Utilizamos o método NJ para a construção filogenética (FIGURA 15). A confiabilidade dos ramos foi avaliada através do método *bootstrap*, função disponível pelo pacote *Analyses of Phylogenetics and Evolution* (APE) versão 5.6-1 (disponível em <https://cran.r-project.org/web/packages/ape/index.html>) com amostragem de 1000 réplicas. A visualização dos dendrogramas foi realizada pelo Dendroscope versão 3.8.1 (disponível em: <https://software-ab.informatik.uni-tuebingen.de/download/dendroscope3/welcome.html>). Para comparar a topologia entre a árvore do Nif core e a árvore do proteoma utilizamos o tanglegram, algoritmo disponível pelo Dendroscope (<https://software-ab.informatik.uni-tuebingen.de/download/dendroscope/manual.pdf>).

FIGURA 15 - FLUXOGRAMA PARA CONSTRUÇÃO DE FILOGENIA SWEEP



FONTE: O Autor (2022).

LEGENDA: Dados do proteoma e do conjunto Nif core foram obtidos a partir de arquivos multiFASTAs para cada organismo. As sequências de cada proteoma foram unificadas na etapa de concatenação. Os proteomas foram representados por vetores SWeeP os quais utilizamos para o cálculo de distância (euclidiana) e construção filogenética pelo método NJ com valor de 1000 para bootstrap.

#### 4.14 SINTENIA POR SYNTAX

Utilizamos a ferramenta web SyntTax (FIGURA 16) para busca de sintenia procariótica (disponível em: <https://archaea.i2bc.paris-saclay.fr/SyntTax/>). A vantagem é que SyntTax possui base relacional própria de genomas completos

disponíveis e atualizada com a base do NCBI. A busca é feita a partir das sequências em proteínas por blocos de sintenia. O uso de SyntTax é limitado em até 120 genomas por análise e a varredura dos blocos de sintenia é exclusiva aos dados cromossômicos das espécies, portanto, ineficaz se existe sintenia no plasmídeo de algum organismo da base, por exemplo.

Buscamos por blocos sintênicos nas espécies disponíveis utilizando as proteínas Nif core (NifH, NifD, NifK, NifE, NifN e NifB) extraídos de *Klebsiella variicola strain 342* (ARNOLD *et al.*, 1988) para as bactérias e em *Methanosarcina mazei strain Go1* (EHLERS *et al.*, 2002) para as espécies de arqueias. Foi selecionado o score mínimo de 10% do algoritmo BLAST normalizado.

FIGURA 16 - INTERFACE WEB DE SYNTTAX

**A.** **SyntTax**  
Prokaryotic Synteny & Taxonomy Explorer

Prokaryotic taxonomy

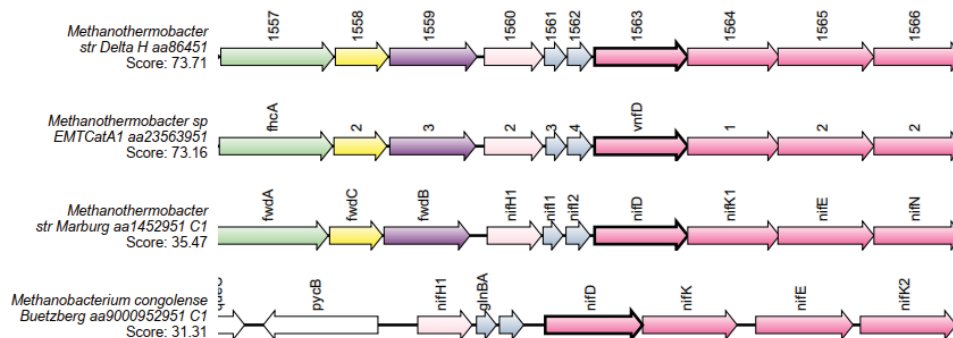
Protein sequence (Raw or Fasta)

Selected chromosomes : 63 (max 120)

Search params Minimal score

Submit

**B.** **SyntTax Report** <http://archaea.u-psud.fr/SyntTax>  
Genomic contexts



FONTE: O Autor (2022).

LEGENDA: (A.) janela principal de consulta sintênica e (B.) aba de resultados de SynTax (*SyntTax Report*). Resumo da sintenia obtida para o conjunto Nif *core* de proteínas identificadas em *Methanothermobacter* spp.

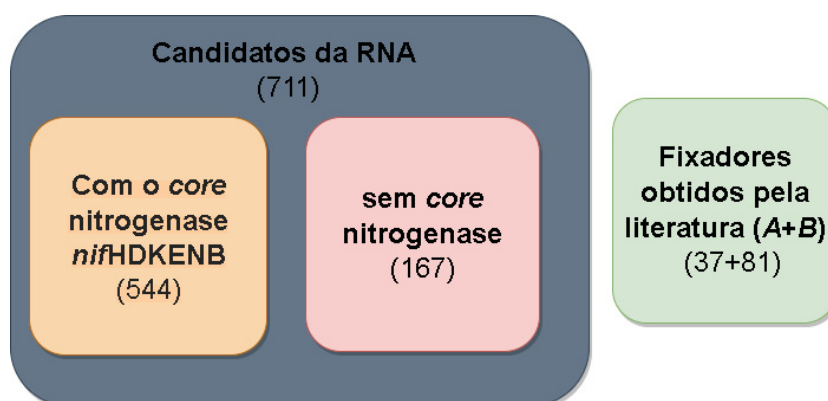
## 5 RESULTADOS

### 5.1 CURADORIA DOS ORGANISMOS DIAZOTRÓFICOS

A partir da estratégia de revisão sistemática da literatura, obtemos 118 genomas completos de espécies diazotróficas, dos quais 81 compuseram de espécies bacterianas e 37 de arqueias. A partir destes genomas foi obtido os dados de Proteomas dos quais extraímos, com RAFTS<sup>3</sup>G a 0.5 de similaridade, 336 agrupamentos relacionados às 22 classes de proteínas Nif (contendo o total de 2156 sequências Nif) que constituíram, assim, a base de proteínas Nif (I) (APÊNDICE, FIGURA 2).

A partir da criação da rede neural, desenvolvemos a estratégia para a prospecção de potenciais diazotrofos em dados de proteomas bacterianos contidos na base de dados do NCBI. A rede trouxe 711 organismos potenciais fixadores, dos quais foram identificados em 544 as proteínas do Nif *core* (NifH, NifD, NifK, NifE, NifN e NifB) e em 167 por não os conterem integralmente (FIGURA 17). Em APÊNDICE, TABELA 3, estão descritas as informações das 711 espécies com código de acesso ao dado genômico pelo NCBI e as sequências do Nif *core* identificadas em cada uma delas.

FIGURA 17 - ORGANISMOS IDENTIFICADOS POR DIAZOINSPECTOR E VALIDADOS.



FONTE: O Autor (2022).

LEGENDA: **A+B** – *Archaea* e *Bacteria*.

Dos 167 organismos sem *nif core*, em 87 deles (cerca de 52%) não identificamos nenhuma sequência do conjunto, a exemplo: *Asaia bogorensis* NBRC 16594 (ASM154799v1), *Rhizobium* sp Y9 (ASM281403v1), *Synechococcus* sp PCC

6312 (ASM31668v1) e *Agrobacterium tumefaciens* (ASM97156v1). Em 12 organismos (aproximadamente 7%) obtemos alguma similaridade de sequências com NifD (e.g.: *Moorella thermoacetica* (ASM126743v1), NifN (e.g.: *Aminobacter* sp MSH1) e em 33 com NifH (acerca de 20%), tais como *Clostridium botulinum* CDC1436 (ASM81793v1), *Fibrobacter succinogenes* subsp *succinogenes* S85 (ASM2466v1), *Fusobacterium varium* (ASM235645v1), *Desulfotobacterium dehalogenans* ATCC 51507 (ASM24315v3), *Megasphaera elsdenii* 1414 (ASM130471v1) entre outros. Para outros organismos desse conjunto sem o *nif* core, identificamos o conjunto parcialmente completo, dos quais podemos citar: *Thermacetogenium phaeum* DSM 12270 (ASM30593v1) no qual verificamos as proteínas NifH, NifD, NifK, NifE e NifB, *Clostridium drakei* (ASM309617v1), *Roseiflexus castenholzii* DSM 13941 (ASM1780v1), *Roseiflexus* sp RS1 (ASM1666v1), *Moorella thermoacetica* (ASM187460v1), *Caldicellulosiruptor kronotskyensis* 2002 (ASM16677v1), *Candidatus Desulforudis audaxviator* MP104C (ASM1842v1) e *Syntrophothermus lipocalidus* DSM 12680 (ASM9240v1) nos quais detectamos as estruturas NifH, NifD, NifK e NifB.

Por fim, catalogamos 662 organismos contendo as estruturas do Nif core completas. Destes, foi impossível estabelecer validação, por literatura, para 52 organismos (QUADRO 2). No APÊNDICE, QUADRO 1 consta a relação de todos os diazotrofos identificados neste estudo, com o número de acesso ao genoma e respectiva referência bibliográfica que traga indício na validação do organismo identificado como diazotrófico.

QUADRO 2 - ORGANISMOS SEM VALIDAÇÃO EM LITERATURA COM NIF CORE COMPLETO

	<b>Organismo</b>	<b>Filo</b>	<b>Código de acesso NCBI</b>
1	<i>Carboxydocella thermautotrophica</i> 019	<i>Firmicutes</i>	ASM304720v1
2	<i>Carboxydocella thermautotrophica</i> 41	<i>Firmicutes</i>	ASM305449v1
3	<i>Clostridium lentocellum</i> DSM 5427	<i>Firmicutes</i>	ASM17883v2
4	<i>Clostridium</i> sp BNL1100	<i>Firmicutes</i>	ASM24487v1
5	<i>Defluviicoccus</i> sp. isolate SSA4	$\alpha$ - <i>proteobacteria</i>	ASM1341470v1
6	<i>Defluviicoccus vanus</i> strain Ben 114	$\alpha$ - <i>proteobacteria</i>	ASM1467269v1
7	<i>Dehalogenimonas</i> sp WBC-2	<i>Chloroflexi</i>	ASM100526v1
8	<i>Desulfosarcina alkanivorans</i> PL12	$\delta$ - <i>proteobacteria</i>	ASM968894v1
9	<i>Desulfosarcina ovata</i> subsp <i>ovata</i> oXyS1	$\delta$ - <i>proteobacteria</i>	ASM968900v1
10	<i>Desulfurispirillum indicum</i> S5	<i>Chrysiogenetes</i>	ASM17763v2
11	<i>Desulfuromonas</i> sp DDH964	$\delta$ - <i>proteobacteria</i>	ASM161127v1
12	<i>Dissulfurispirilla thermophila</i> T55J	<i>Nitrospirae</i>	ASM1470123v1
13	<i>Dolichospermum compactum</i> NIES-806	<i>Cyanobacteria</i>	ASM236811v1
14	<i>Draconibacterium halophilum</i> [sp. M1]	<i>Bacterioidetes</i>	ASM1044883v1

15	<i>Geobacter pickeringii</i> G13	$\delta$ -proteobacteria	ASM81795v1
16	<i>Geobacter</i> sp M18	$\delta$ -proteobacteria	ASM17511v2
17	<i>Geobacter</i> sp M21	$\delta$ -proteobacteria	ASM2364v1
18	<i>Halothiobacillus</i> sp LS2	$\gamma$ -proteobacteria	ASM166381v1
19	<i>Methanobacterium</i> sp. BRmetb2	Euryarcheota	ASM349128v1
20	<i>Methanothermobacter</i> sp. KEPCO-1	Euryarcheota	ASM803370v1
21	<i>Methanothermobacter</i> sp. THM-2	Euryarcheota	ASM991766v1
22	<i>Methylacidiphilum</i> sp IT5	Verrucomicrobia	ASM1731052v1
23	<i>Methylacidiphilum</i> sp IT6	Verrucomicrobia	ASM1731050v1
24	<i>Nostoc</i> sp CENA543	Cyanobacteria	ASM289687v1
25	<i>Paenibacillus</i> sp FSL P4-0081	Firmicutes	ASM75856v1
26	<i>Paenibacillus</i> sp Izh-N1	Firmicutes	ASM287243v1
27	<i>Prosthecochloris</i> sp HL-130-GSB	Chlorobi	ASM211382v1
28	<i>Pseudoclostridium thermosuccinogenes</i> DSM 5807	Firmicutes	ASM289685v1
29	<i>Rhizobium bangladeshense</i> BLR175	$\alpha$ -proteobacteria	ASM1735724v1
30	<i>Rhizobium lentis</i> strain BLR27	$\alpha$ -proteobacteria	ASM1735213v1
31	<i>Rhizobium</i> sp ACO-34A	$\alpha$ -proteobacteria	ASM260063v1
32	<i>Rhizobium</i> sp Kim5	$\alpha$ -proteobacteria	ASM17269v2
33	<i>Rhizobium</i> sp N1314	$\alpha$ -proteobacteria	ASM166430v1
34	<i>Rhizobium</i> sp N1341	$\alpha$ -proteobacteria	ASM166408v1
35	<i>Rhizobium</i> sp N324	$\alpha$ -proteobacteria	ASM166448v1
36	<i>Rhizobium</i> sp N621	$\alpha$ -proteobacteria	ASM166432v1
37	<i>Rhizobium</i> sp N6212	$\alpha$ -proteobacteria	ASM166446v1
38	<i>Rhizobium</i> sp N731	$\alpha$ -proteobacteria	ASM166414v1
39	<i>Rhizobium</i> sp N741	$\alpha$ -proteobacteria	ASM166454v1
40	<i>Rhizobium</i> sp N871	$\alpha$ -proteobacteria	ASM166452v1
41	<i>Rhizobium</i> sp TAL182	$\alpha$ -proteobacteria	ASM211772v1
42	<i>Rhodobacter blasticus</i> 28-5	$\alpha$ -proteobacteria	ASM307140v1
43	<i>Sphaerospermopsis kisseleviana</i> NIES-73	Cyanobacteria	ASM236807v1
44	<i>Skermanella pratensis</i> strain W17	$\alpha$ -proteobacteria	ASM884314v1
45	<i>Skermanella</i> sp TT6	$\alpha$ -proteobacteria	ASM1665363v2
46	<i>Sulfurospirillum halospirans</i> DSM 13726	$\epsilon$ -proteobacteria	ASM172360v1
47	<i>Sulfurospirillum</i> sp JPD-1	$\epsilon$ -proteobacteria	ASM230953v1
48	<i>Sulfurospirillum</i> sp SL2-1	$\epsilon$ -proteobacteria	ASM216231v1
49	<i>Syntrophobacter fumaroxidans</i> MPOB	$\delta$ -proteobacteria	ASM1496v1
50	<i>Thermoanaerobacterium aotearoense</i> SCUT27	Firmicutes	ASM990525v1
51	<i>Thiospirochaeta perfilievii</i> P	Spirochaetes	ASM832994v1
52	Zoogloeaceae bacteirum Par-f-2	$\beta$ -proteobacteria	ASM305451v1

FONTE: O Autor (2022).

## 5.2 NIF IDENTIFICADAS POR HOMOLOGIA

Neste trabalho, utilizamos a ferramenta RAFTS<sup>3</sup>G em dois momentos: em dados de 118 proteomas obtidos da primeira revisão sistemática e nas informações dos 662 proteomas classificados como fixadores comprovados conterem o *nif* core. Em ambos tivemos como objetivo identificar as proteínas Nif através da criação de grupos homólogos. A partir dos 662 proteomas (totalizando 3.143.063 sequências de proteínas) foi gerado 1.021.121 agrupamentos. Destes, 327.558 sequências foram

agrupadas com, ao menos, duas sequências e o restante das proteínas não foram agrupadas (693.563 sequências). Das sequências agrupadas, selecionamos os 100 maiores agrupamentos contendo sequências Nif (APÊNDICE, TABELA 2). O número total de proteínas Nif core encontradas NifH (1001), NifD (680), NifK (681), NifE (686), NifN (667) e NifB (669) supera o número total de organismos estudados (662). Além disso, identificamos a duplicação das estruturas Nif core em *Aulosira laxa* NIES-50 (ASM236805v1), *Azoarcus* sp KH32C (ASM34994v1), *Calothrix brevissima* NIES-22 (ASM236799v1), *Nostoc carneum* NIES-2107 (ASM236815v1), *Tolypothrix tenuis* PCC 7101 (ASM236829v1) e *Trichormus variabilis* ATCC 29413 (ASM20407v1) o que sugere a existência de mais de um núcleo *nif* nessas espécies.

Ainda sobre os agrupamentos gerados, resgatamos informações de sequências Nif que não estavam anotadas nos proteomas disponíveis pelo NCBI. Nestes dados, houve significativo número de sequências hipotéticas (sem função ou gene relacionado) relacionadas aos grupos com proteínas Nif. Em paralelo, identificamos proteínas com funções correlatas as sequências Nif (tendo como referência as anotações dos genes *nif* de *Klebsiella variicola* 342 (ASM1956v1) e *Methanosarcina mazei* Go1 (ASM706v1). Na TABELA 1 sintetizamos o resultado obtido pelo agrupamento, descrição da proteína correlatas, número de sequências encontradas e o agrupamento Nif inserido.

TABELA 1 - PROTEÍNAS COM OUTRAS DEFINIÇÕES IDENTIFICADAS EM CLUSTERS NIF HOMÓLOGOS.

Descrição(ões) da proteína encontrada	Nº de Sequências encontradas*	Cluster associado
Proteína de biossíntese FeMoco, proteína da família NifB/NifX ligante de <i>clusters</i> FeMo.	578	NifX
Proteína esqueleto de montagem [Fe-S], NifU-1 C-terminal, proteína IscU;	292**	NifU
Aminotransferase Classe V, piridoxal-fosfato dependente aminotransferase, cisteína desulfurase	960	NifS
2-Isopropitil sintase, homoserina sintase, piruvato carboxiltransferase, R-citramalato sintase, trans-homoaconitato sintase;	520	NifV
Serina O-acetiltransferase, serina acetiltransferase; proteína CysE	221	NifP

Proteína FixU, proteína FixT	418	NifT
Arsenato redutase; proteína ArsC;	18	NifO
Hidrogenase	296	NifQ
Peptidil prolil cis-trans isomerase, peptidil prolil isomerase, rotamase;	110	NifM
Flavodoxina; NADP(H) dependente oxidoreductase, FMN redutase;	204	NifF
radical SAM, Fe-S oxidoreductase	647	NifB
ATPase AAA dependente, Fis family sigma54 dependente, Regulador positivo transcricional;	476	NifA
Histidina quinase; sensor híbrido PAS/PAC histidina quinase; sensor dois componentes, Regulador negativo transcricional;	108	NifL

FONTE: O Autor (2022).

LEGENDA: (\*) número total de sequências homólogas identificadas do top 100 maiores agrupamentos gerados por RAFTS<sup>3</sup>G com 0.5 similaridade. (\*\*) Sequências íntegras da proteína NifU, excluindo-se as variantes NifU-N, Nfu-C e NifU *partial*.

### 5.3 EXTRAÇÃO DE NIF CORE, PROTEOMA E CLUSTER NIF

Conforme a seção 4.8 da metodologia, na procura de proteínas Nif em dados de proteomas, foi desenvolvida uma estratégia de busca baseada em termos para extração de *clusters*, o que culminou no desenvolvimento da ferramenta *GBK2cluster* (ANEXO, FIGURA 3) e no preditor de proteínas Nif, a ferramenta *NifFINDER*. A identificação do *cluster nif* nos diazotrofos desta seção foi substituída pela busca de sintenia por SyntTax devido a ocorrência de significativo volume de dados genômicos fragmentados. As sequências do Nif *core* e do Proteoma foram obtidas e, através dessas informações, integramos às análises descritas nas seções subsequentes.

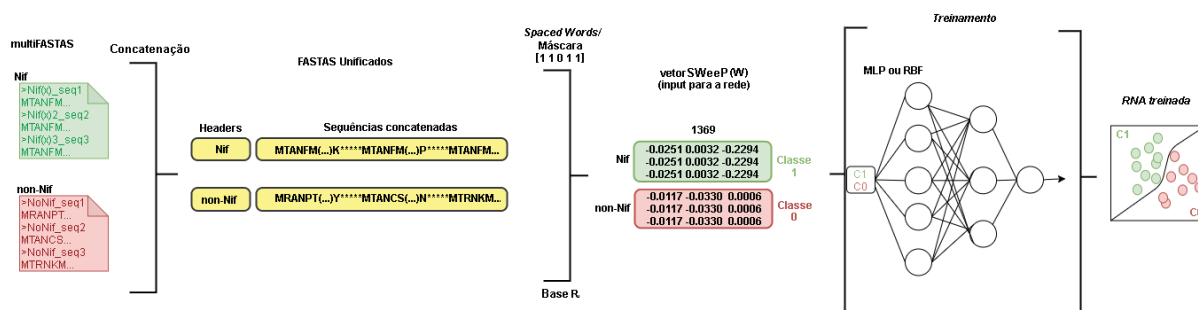
#### 5.3.1 NIFFINDER

Desenvolvemos NifFINDER para tornar-se uma ferramenta voltada à predição de até 27 proteínas Nif (22 classes Nif somados a NifEN, NifNB, Nifl<sub>1</sub> + Nifl<sub>2</sub>, NifXa e MonC). Em síntese, cada classe Nif foi treinada como conjunto de dados verdadeiros para o treinamento das redes neurais de maneira independente. Os conjuntos falsos



foram compostos por seqüências de clusters randômicos obtidos por RAFTS<sup>3</sup>G, numa proporção aproximada de dados amostrais 1:2 (para cada 1 seqüência Nif 2 seqüências foram não Nif). Os modelos obtidos basearam-se nos modelos de aprendizado *Radial Basis Function* (RBF) e MLP. A FIGURA 18 ilustra, de modo sucinto, a estratégia de desenvolvimento do modelo de rede Nif em NifFINDER para uma classe Nif (e.g.: para criação de rede preditora de seqüências NifH). Dentre os dois modelos, as redes com maior valor de *F1-score* foram selecionadas e, caso as métricas coincidissem, selecionamos os modelos de maior acurácia. NifFINDER apresentou resultados satisfatórios (*F1-score* médio de 0,952).

FIGURA 18 - DEFINIÇÃO DO MODELO PARA NIFFINDER.



FONTE: O Autor (2022).

LEGENDA: Cada classe de proteína Nif (multiFASTA) foi considerada como conjunto verdadeiro (C1 – Classe 1) e seqüências não Nif (C0 – Classe 0) para o treinamento da rede. Cada conjunto treinamento foi convertido e projetados em vetores SWeeP (W). Estes vetores compuseram os *inputs* de cada RNA treinada e avaliamos os resultados para modelo MLP e RBF.

Estratégia de aplicação: O algoritmo converte cada FASTA (de aminoácidos) de entrada para seus respectivos vetores SWeeP. NifFINDER busca dentro da estrutura por proteínas candidatas a proteínas Nif. Por fim, NifFINDER gera duas saídas, sendo a primeira um *struct* de extensão. mat (que pode ser lida e manipulada pelo MATLAB®) e, a segunda, em formato FASTA (de aminoácidos). NifFINDER foi compilado com MCR v8.0 (MATLAB®) e está disponível em: [https://sourceforge.net/projects/niffinder/files//?upload\\_just\\_completed=true](https://sourceforge.net/projects/niffinder/files//?upload_just_completed=true).

Realizamos teste comparativo (QUADRO 3) na predição de seqüências Nif obtidos por NifFINDER contra NifPRED (MEHER *et al.*, 2018) utilizando o genoma de referência *Aphanizomenon flos-aquae* DEX188 (ASM1734685v1). Das 19 seqüências Nif existentes, NifPRED identificou 10 (neste exemplo, a sensibilidade de NifPRED foi de 0,526), sendo que para as seqüências de NifS e NifJ, o preditor classificou como

NifN e NifE, respectivamente, e uma sequência de 2-isopropilmalato sintase e outra proteína inibidora de formação de septo Maf, foram classificadas como NifN e NifB. Enquanto que NifFINDER identificou um total de 19 sequências (100%), cada um dentro da sua própria classe.

QUADRO 3 – PREDIÇÃO DE NIFPRED E NIFFINDER PARA *APHANIZOMENON FLOS-AQUAE*

<i>DEX188</i>		
<i>ID PROTEIN</i>	<b>NIFPRED</b>	<b>NIFFINDER</b>
QSV67144.1	nifH	NifH
QSV67151.1	nifD	NifD
QSV67173.1	nifK	NifK
QSV67174.1	nifE	NifE
QSV67175.1	nifN	NifN
QSV67140.1	NifB	NifB
QSV67178.1	Non-nif (0.996)	NifW
QSV67134.1	Non-nif (0.996)	NifZ
QSV67902.1	Non-nif (0.98)	NifP
QSV67138.1	Non-nif (0.968)	NifP
QSV66065.1	Non-nif (0.995)	NifI <sub>1</sub>
QSV67358.1	Non-nif (0.616)	NifJ
QSV67143.1	Non-nif (0.694)	NifU
QSV69175.1	-	NifX
QSV67176.1	-	NifXa
QSV67133.1	-	NifT
QSV67135.1	-	NifV
QSV67142.1	nifN (cysteine desulfurase)	NifS
QSV67499.1	nifE (pyruvate ferredoxin oxidoreductase)	NifJ
QSV67565.1 2	nifE (2-isopropylmalate synthase)	-
QSV65539.1	nifB (septum formation inhibitor Maf)	-

FONTE: O Autor (2022).

LEGENDA: Proteínas ou genes *nif* preditos e corretos em cor verde. Identificados com probabilidade de não ser *nif* ("Non-nif"). Nif identificados erroneamente (em vermelho). ("-") significa que a proteína não foi encontrada pelo respectivo método.

#### 5.4 GRUPOS OBTIDOS POR HCPC

Os valores indicados de *k* foram obtidos para o conjunto Nif core utilizando o

método HCPC e comparado com os resultados da estatística gap, método do cotovelo e coeficiente da silhueta (TABELA 2). Os valores indicados de  $k$  para método cotovelo foram 5, 10, 13 e 16 para o coeficiente da silhueta foram 10, 13 e 18 e para estatística gap o  $k$  indicado foi 8 (APÊNDICE, FIGURA 4). Para o método HCPC, os valores obtidos de  $k$  foram superiores a 8 (consensus CDF). Os valores de  $k = 8, 10$  e 13 foram selecionados para investigação dos grupos (APÊNDICE, FIGURAS 5 e TABELA 7) comparando-se as informações com o mapa de calor obtido da matriz de dissimilaridade ordenada (APÊNDICE, FIGURA 9).

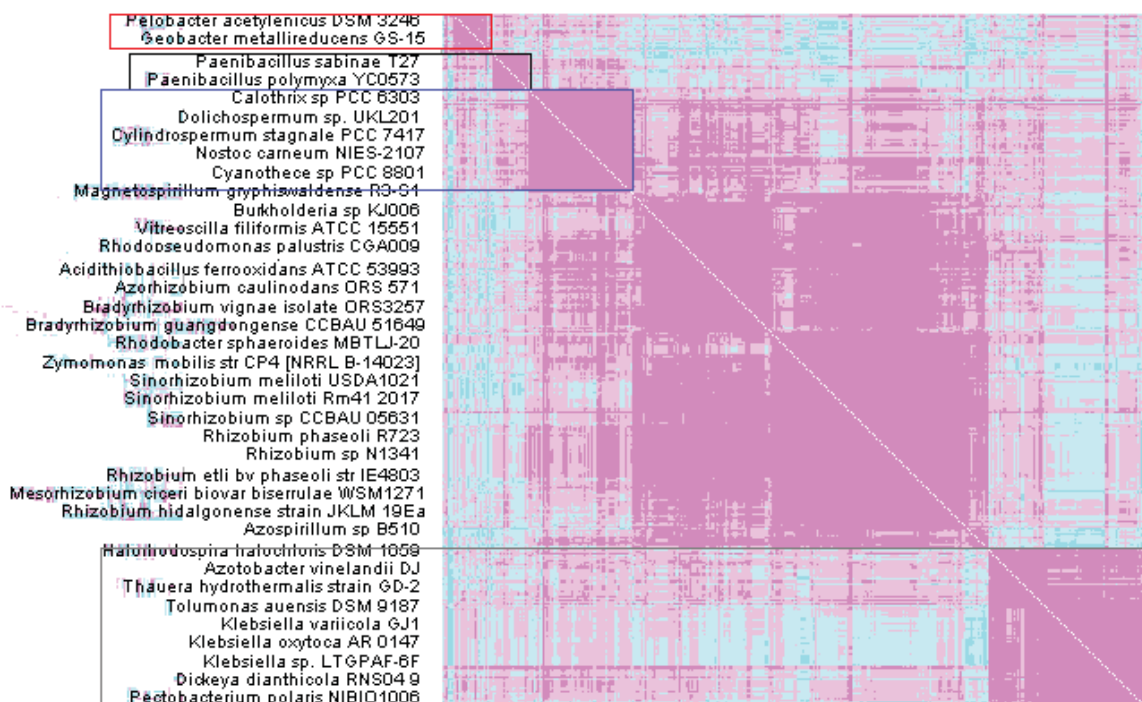
TABELA 2 – VALORES ESTATÍSTICOS PARA  $K$  GRUPOS INDICADOS

<i>Estatística</i>	Valor(es) de $K$ indicado(s)
<i>Método do cotovelo</i>	5, 10, 13 e 16
<i>Coeficiente da silhueta</i>	10, 13 e 18
<i>Estatística gap</i>	8
<i>HCPC</i>	$K$ superiores a 8 ( <i>consensus CDF</i> )

FONTE: O Autor (2022).

Observando dois clusters de organismos no mapa de calor ordenado: um formado pelas cianobactérias (FIGURA 19, em azul) e outro por *Paenibacillus* spp. (FIGURA 18, em preto), obtemos que para  $k=13$  as cianobactérias se subdividiram em dois novos grupos (grupos 6 e 7) (APÊNDICE, TABELA 7) e, em  $k=8$ , todo o gênero de *Paenibacillus* se aglutinou com *Azotobacter* spp. (grupo 3) (APÊNDICE, TABELA 7). Em  $k=10$  o grupo das cianobactérias se manteve coeso no grupo 6 e *Paenibacillus* spp. se agruparam no grupo 4 com *Geobacter* spp. (FIGURA 19, em vermelho). Por conseguinte, concluiu-se que o  $k$  ideal para o conjunto Nif core foi 10.

FIGURA 19 - MAPA DE CALOR ORDENADO PARA NIF CORE



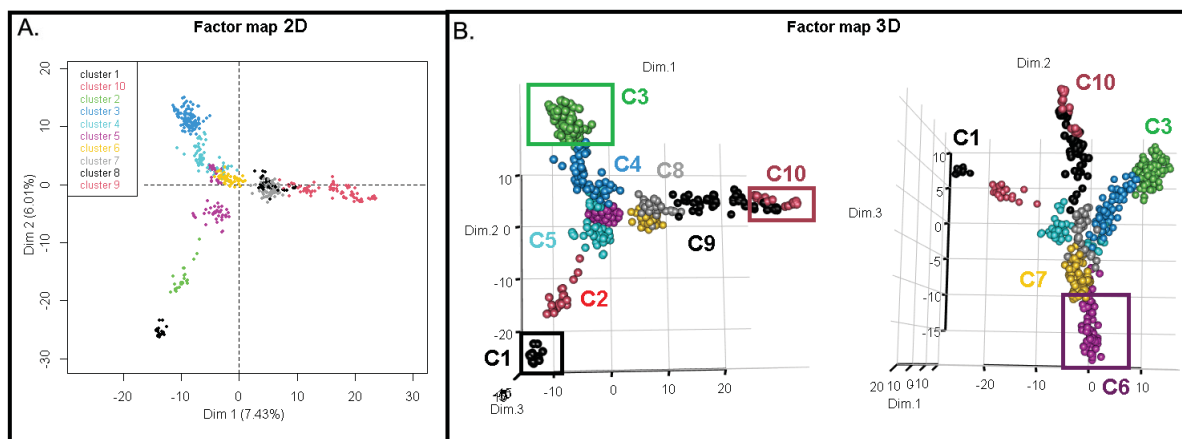
FONTE: O Autor (2022).

LEGENDA: No mapa de calor ordenado (obtido da matriz de dissimilaridade ordenada) temos *Pelobacter* spp. (destaque em vermelho), *Paenibacillus* spp. (preto) e cianobactérias (azul) em proximidade das informações. No grupo contendo *Azotobacter* spp. (cinza) nota-se significativa distância dos organismos supracitados. Nesta imagem, a cor rosa denota distância 0 e azul claro a distância máxima entre os dados observados.

## 5.5 MAPAS DE FATORES OBTIDOS POR PCA

Plotamos os mapas de fatores obtidos da PCA dos vetores SWeeP para os dados de Nif core mapeando a distribuição dos 10 grupos em 2D (FIGURA 20-A) e 3D (FIGURA 20-B). Uma parcela significativa da distribuição dos dados encontrou-se perpendicular ao plano, por este motivo, plotamos o mapa da visualização também em 3D. A partir das visualizações, observamos os grupos com os maiores auto valores em 1, 3, 6 e 10 (FIGURA 20 grupos delimitados em quadros).

FIGURA 20 – MAPA DE FATORES DOS 10 GRUPOS POR HCPC.



FONTE: O Autor (2022).

LEGENDA: Mapas de fatores obtidos com as sequências *Nif core* visualizados em 2D e 3D (A e B, respectivamente). Os clusters relativos aos maiores auto valores estão identificados em quadros, sendo eles: C1, C3, C6 e C10.

Obtemos a correlação dos 10 clusters gerados pela PCA-HCPC com grupos contendo os filos ou divisões bacterianas mais representativos (QUADRO 4): o grupo 1 (G1) constituído por proteobactérias gama (divisão *Klebsiella*), o grupo 2 (G2) por proteobactérias gama (divisão *Dickeya*), o grupo 3 (G3) por arqueias (divisão *Methanosarcina*), *Firmicutes* (divisão *Clostridium*), *Fusobacteria*, *Spirochaetes*, *Verricummicrobia* (*Coralimargarita akajimensis*), *Chloroflexi*, *Chlorobi*, *Bacteroidetes* e proteobactérias delta (divisão *Desulfovibrio*), o grupo 4 (G4) por *Archaea* (divisão *Methanobacterium*), *Firmicutes* (bacilares), *Nitrospirae*, *Aquificales*, proteobactérias épsilon e delta (divisão *Geobacter*), o grupo 5 (G5) por proteobactérias alfa (*Magnetococcus marinus* MC-1 e *Marteella endophytica* YC6887), beta (divisão *Azoarcus*) e gama (divisão *Azotobacter*), o grupo 6 (G6) por cianobactérias, o grupo 7 (G7) por proteobactérias alfa (divisão *Bradyrhizobium*), beta (divisão *Burkholderia*) e gama (divisão *Halothiobacillus*), *Acidithiobacilia*, *Nitrospirae* (*Leptospirillum* spp.) e *Verricummicrobiota* (*Methylacidiphilum* spp.), os grupos 8 (G8), 9 (G9) e 10 (G10) exclusivamente compostos por proteobactérias alfa, divisões *Rhodobacter*, *Sinorhizobium* e *Rhizobium*, respectivamente.

QUADRO 4 – FILOS CARACTERIZADOS PARA OS 10 GRUPOS OBTIDOS POR HCPC

Grupo	Filo(s):
G1	proteobactérias gama (divisão <i>Klebsiella</i> )

G2	proteobactérias gama (divisão <i>Dickeya</i> )
G3	<i>Euryarcheota</i> (divisão <i>Methanosarcina</i> ), <i>Firmicutes</i> (divisão <i>Clostridium</i> ), <i>Fusobacteria</i> , <i>Spirochaetes</i> , <i>Verricumicrobia</i> ( <i>Coraliomargarita akajimensis</i> ), <i>Chloroflexi</i> , <i>Chlorobi</i> , <i>Bacteroidetes</i> , proteobactérias delta (divisão <i>Desulfovibrio</i> )
G4	<i>Euryarcheota</i> (Divisão <i>Methanobacterium</i> ), <i>Firmicutes</i> (bacilares), <i>Nitrospirae</i> , <i>Aquificales</i> , proteobactérias épsilon e delta (divisão <i>Geobacter</i> )
G5	proteobactérias alfa ( <i>Magnetococcus marinus</i> MC-1 e <i>Marteella endophytica</i> YC6887), beta (divisão <i>Azoarcus</i> ) e gama (divisão <i>Azotobacter</i> )
G6	cianobactérias
G7	proteobactérias alfa (divisão <i>Bradyrhizobium</i> ), beta (divisão <i>Burkholderia</i> ) e gama (divisão <i>Halothiobacillus</i> ), <i>Acidithiobacilia</i> , <i>Nitrospirae</i> ( <i>Leptospirillum</i> spp.) e <i>Verricumicrobiota</i> ( <i>Methylacidiphilum</i> spp.)
G8	proteobactérias alfa (divisão <i>Rhodobacter</i> )
G9	proteobactérias alfa (divisão <i>Sinorhizobium</i> )
G10	proteobactérias alfa (divisão <i>Rhizobium</i> )

FONTE: O Autor (2022).

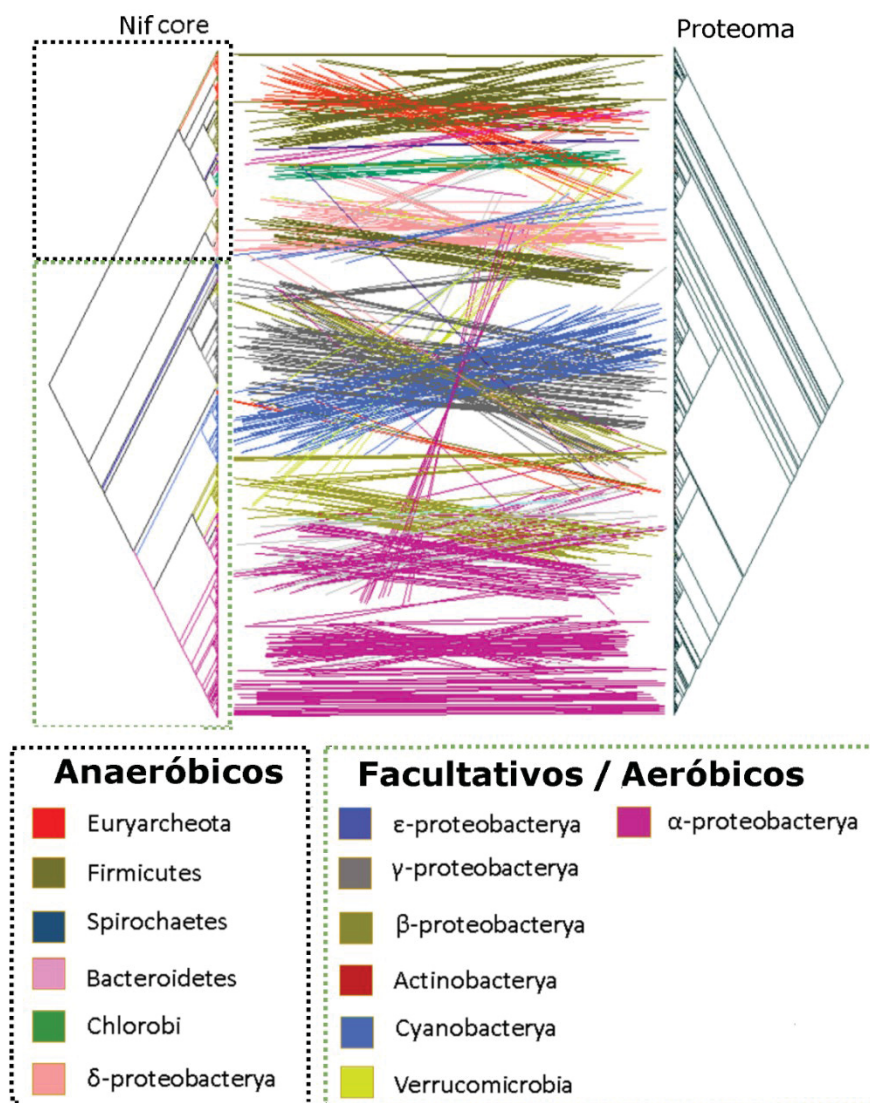
LEGENDA: Caracterização dos 10 grupos a nível de filo. Filos abordados mais de um grupo, descrevemos uma divisão com o gênero mais abundante e exclusivo no grupo correspondente.

## 5.6 FILOGENIA SWEEP

É possível a construção de dendrogramas utilizando os resultados do método HCPC. Entretanto, era desejável construir uma filogenia que nos permitisse uma árvore com, além das informações dos grupos obtidos, a aplicação de métricas filogenéticas (i.e.: NJ e *bootstrap*). Dessa maneira, obtemos as super-árvores para o conjunto Nif *core* e para o Proteoma (FIGURA 21). A primeira informação obtida entre a comparação dos dados da árvore Nif *core* da árvore proteoma é que foi estabelecida uma quase que perfeita divisão das espécies anaeróbias (FIGURA 21, linhas tracejadas em cor preta nos clados da filogenia Nif *core*) de aeróbias ou anaeróbias facultativas (FIGURA 21, linhas tracejadas em verde). Os filios anaeróbios se dispuseram no topo da filogenia Nif *core* contendo clados de *Euryarcheota* (arqueotes), *Firmicutes*, *Fusobacteria*, *Spirochaetes*, *Bacteroidetes*, *Chlorobi*, *Chloroflexi*, *Chrysiogenetes*, *Verricumicrobia* e  $\delta$ -proteobactérias. Os filios anaeróbios facultativos e aeróbios se ramificaram em espécies de actinobactérias, *Aquificales*, *Deferribacteres*, outras classes de *Firmicutes* (Bacilares), cianobactérias e

proteobactérias  $\alpha$ ,  $\beta$ ,  $\gamma$  e  $\epsilon$ .

FIGURA 21- SUPER-ÁRVORE NIF CORE VERSUS PROTEOMA DE ARCHAEA E BACTERIA.



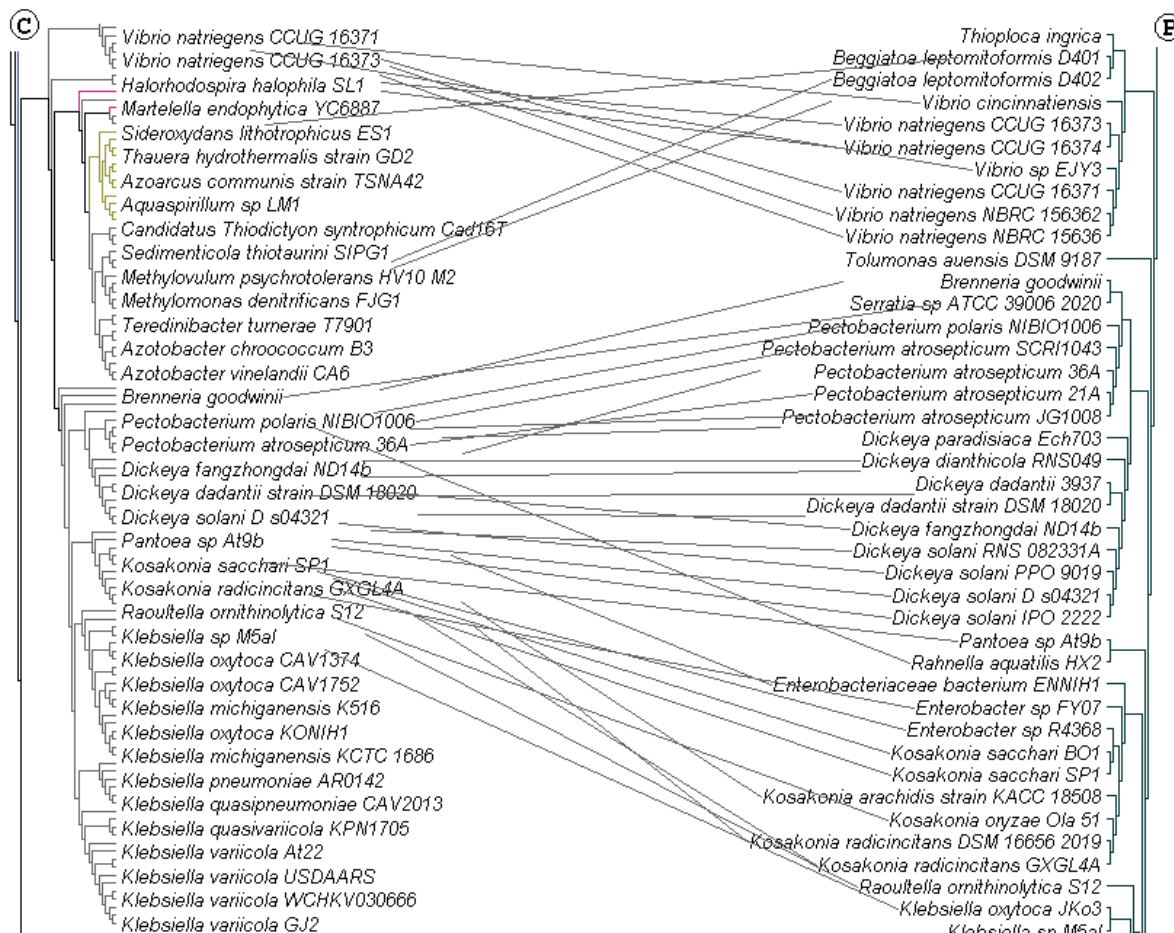
FONTE: O Autor (2022).

LEGENDA: Resultado do algoritmo tanglegram da Filogenia Nif core versus Proteoma. As posições de organismos relacionados pelas grades de cores (a nível de filo). Obs.: Os filios *Chloroflexi*, *Chrysiogenetes*, *Nitrospirae*, *Aquificales*, *Acidithiobacillia*, *Deferribacteres* e *Fusobacteria* fazem parte dessa construção.

Outra informação obtida entre as comparações, foi que a primeira filogenia apresentou várias inversões de clados em relação à segunda. No entanto, esses clados constituíram-se, em sua maioria, por organismos de proximidade taxonômica. A FIGURA 22 nos exemplifica a desorganização dos ramos, através do resultado observado para proteobactérias gama, em que o filo se agrupa na árvore do Proteoma (P). Todavia, existe a aparente desorganização de vários ramos com

espécies de beta e alfa-proteobactérias (e.g.: *Azoarcus communis* strain GD2 e *Marteella endophytica* YC6887, respectivamente) nos ramos da árvore Nif core (C).

FIGURA 22 – PROTEOBACTÉRIAS GAMA RELACIONADAS NAS ÁRVORES CORE E PROTEOMA



FONTE: O Autor (2022).

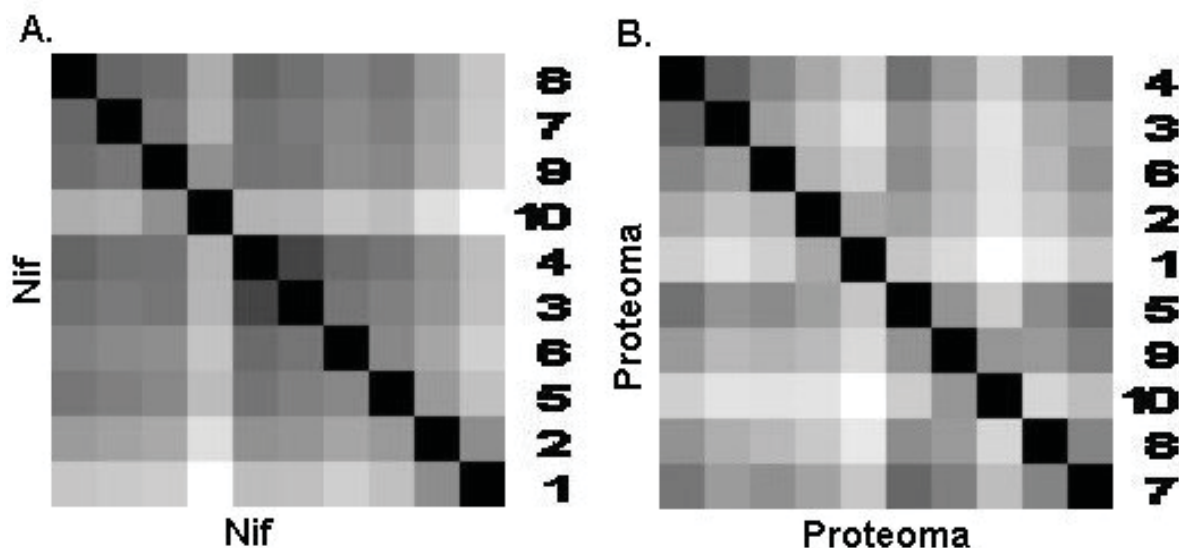
LEGENDA: Árvore obtida para Nif core (C) e Proteoma (P) com distância euclidiana e NJ. Bactérias alfa (ramos em cor rosa), beta (ramos em cor marrom) e gama (ramos em cor cinza) são destacadas na filogenia Nif core (C). As linhas em cinza foram obtidas pelo algoritmo tanglegram (Dendroscope) e conectaram as mesmas espécies de proteobactérias gama identificadas na filogenia Nif core e Proteoma (P).

O mapa de calor ordenado obtido da matriz de distância euclidiana ordenada é outra maneira de visualizar e avaliar as informações entre dados de organismos. Entre os grupos obtidos no mapa Proteoma por Nif core (FIGURA 23, A.) foram observadas alterações de posição dos grupos em relação a ordenação dos grupos no mapa de Proteoma versus Proteoma (FIGURA 23, B.). No mapa A o G10 se afastou consistentemente dos outros grupos e no mapa B o mesmo ocorreu para G1. Os



grupos do mapa A: G1, G2, G3, G4, G5 e G6 denotaram maior proximidade com espécies anaeróbicas e facultativas (com exceção de G6 – as cianobactérias) e G7, G8, G9 indicaram proximidade com espécies aeróbicas e simbiontes relacionados com as proteobactérias alfa. No mapa B: G7, G8, G9 e G10 formaram os grupos majoritariamente relacionados com as proteobactérias alfa, G5 predominantemente de proteobactérias beta e gama. G2 e G1 compuseram-se pelas proteobactérias gama. G3 e G4 pelas arqueias e gram variáveis (e.g.: Firmicutes) e, finalmente, o G6 composto pelas cianobactérias. Nos dois mapas de calor foi consenso a proximidade entre grupos em pares formados por G3 e G4 e outro por G7 e G8.

FIGURA 23 – MAPAS DE CALOR NIF CORE E PROTEOMA

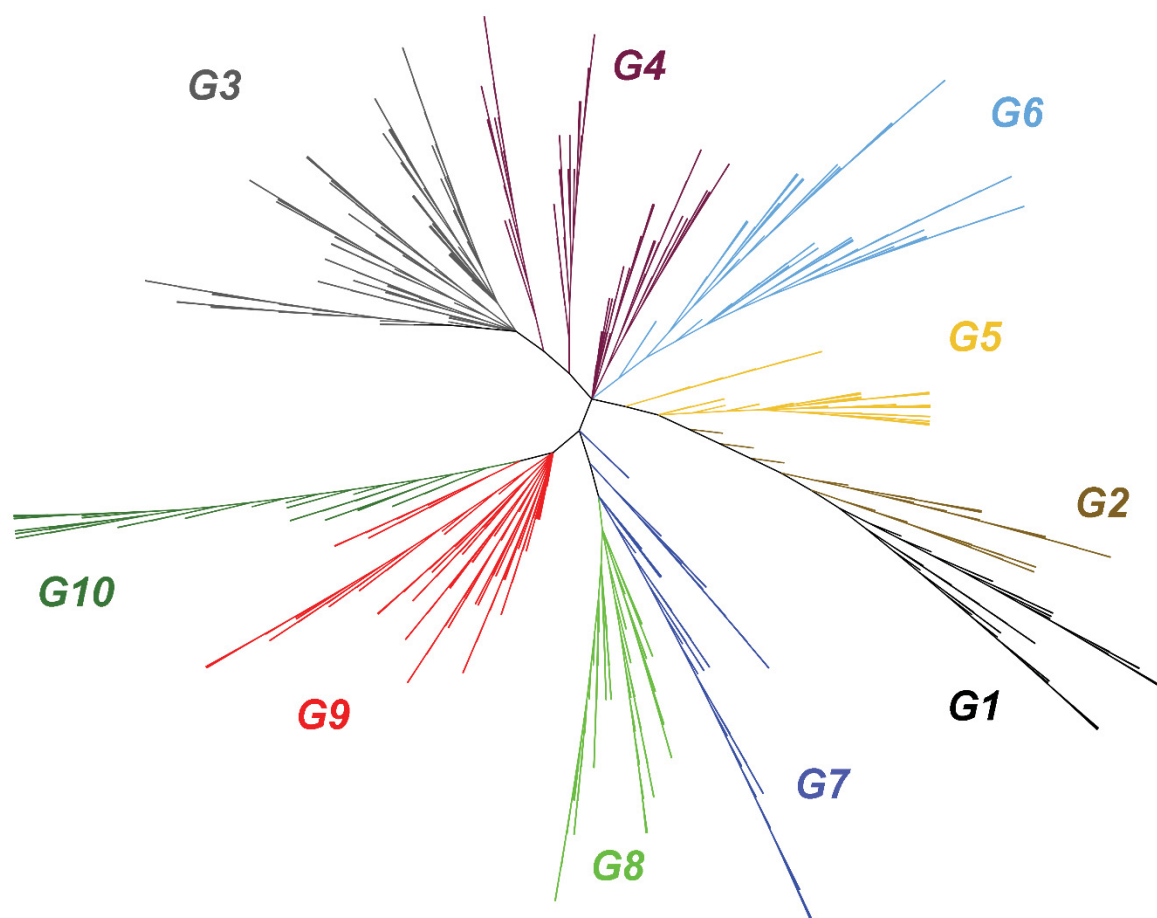


FONTE: O Autor (2022).

LEGENDA: O mapa de calor A foi constituído pelos dados do *Nif core* em relação aos G10 obtidos pela árvore de *Nif core* (NJ). O mapa B criado com as informações de proteoma em relação aos 10 grupos obtidos por *Nif core*. Nesta imagem, preto denota distância 0 e branco é a distância máxima.

Por fim, foi possível estabelecer as relações entre as informações dos grupos com a topologia da árvore *Nif core*. A ordenação dos grupos do sentido horário das espécies anaeróbicas até as aeróbicas, ficou estabelecida em: G3, G4, G5, G2, G1, G7, G8, G9 e G10 (FIGURA 24). Obs.: devido ao tamanho das árvores geradas, deixamos material complementar disponível em APÊNDICE – 8. Filogenia SWeeP.

FIGURA 24 – DISTRIBUIÇÃO DOS GRUPOS NA FILOGENIA NIF CORE



FONTE: O Autor (2022).

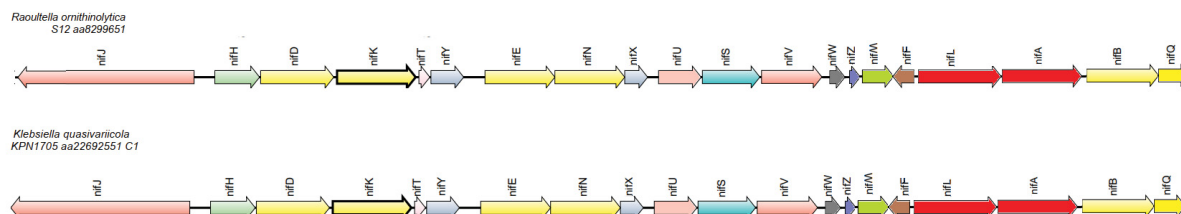
LEGENDA: A ordenação dos grupos na filogenia *Nif core* ficou estabelecido por: G3, G4, G6, G5, G2, G1, G7, G8, G9 e G10. Edição e visualização da filogenia pelo iTOL.

## 5.7 SINTENIA DOS NÚCLEOS NIF EM GRUPOS DIAZOTRÓFICOS

A sintenia genômica consistiu em mapear as regiões mais prováveis de conterem os núcleos *nif*. Os resultados foram sumarizados segundo a ordem crescente dos grupos e conforme a disponibilidade da base de dados fornecida por SyntTax.

O conjunto G1 compreendeu-se em proteobactérias gama, especificamente do gênero *Klebsiella* e *Raoutella ornithinolytica* S12. Representamos a sintenia para as espécies de *Klebsiella quasivariicola* KPN1705 e *Raoutella ornithinolytica* S12 (FIGURA 25).

FIGURA 25 - SINTENIA GENÔMICA PARA G1

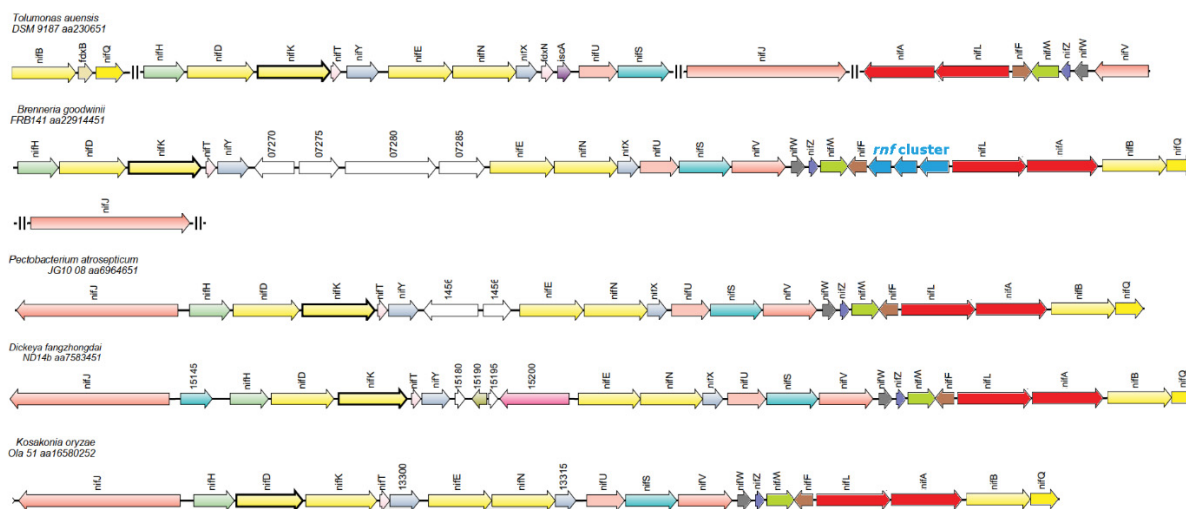
**GAMMA**

FONTE: O Autor (2022).

Conforme a ilustração, nestes organismos e em todas as demais espécies catalogadas em G1 identificamos como padrão de sintenia o núcleo *nifJHDKTYENXUSVWZMFLABQ*, compreendido em 20 genes.

No G2 identificamos outras espécies de  $\gamma$ -proteobactérias. A sintenia genômica para as espécies de *Tolomonas auensis*, *Brenneria goodwinii*, *Pectobacterium atrosepticum* JG10-08, *Dickeya fangzhongdai* ND14b e *Kosakonia oryzae* Ola 51 estão representadas na FIGURA 26. Identificamos núcleos *nif* com padrões similares ao G1, apresentando, no entanto, além do conjunto de genes *nif* descritos, genes relacionados ao processo da FBN, sendo eles: *fdxN*, *fdxB* e *iscA* (*iscN*).

FIGURA 26- SINTENIA GENÔMICA PARA G2

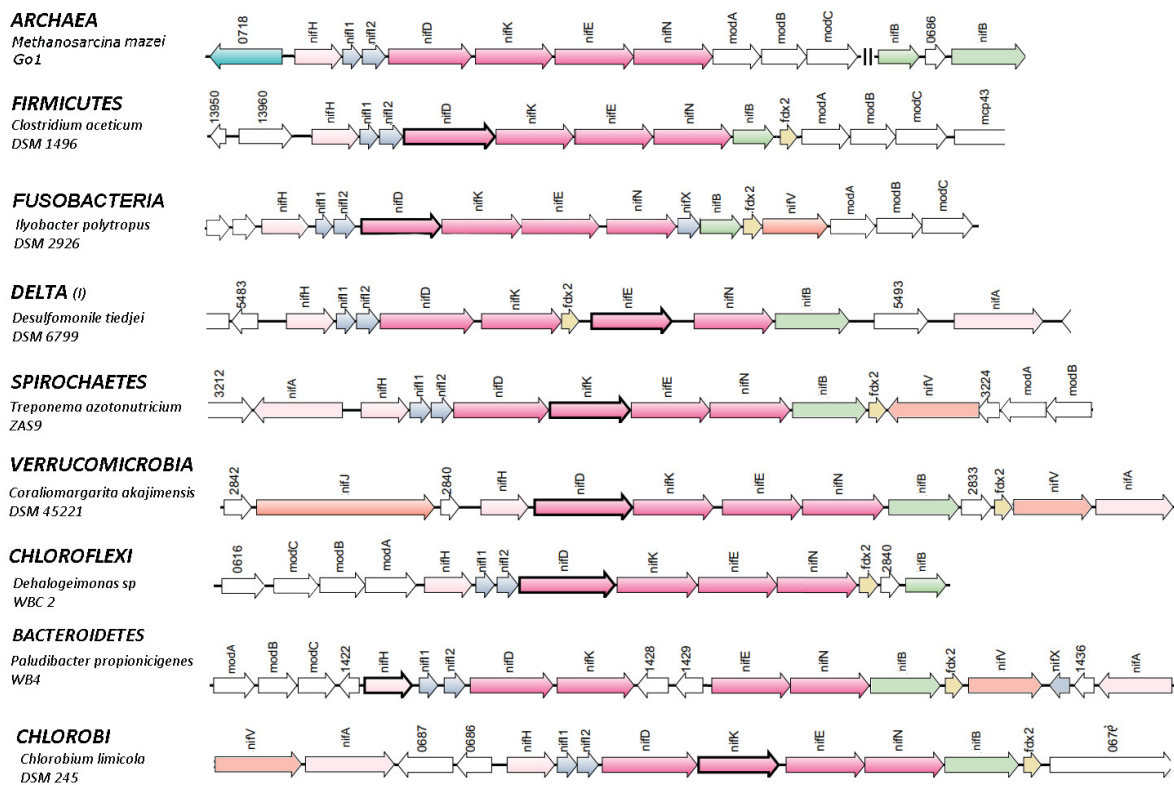
**GAMMA**

FONTE: O Autor (2022).

Em G3 foi agrupado uma miscelânea de espécies anaeróbicas obrigatórias.

Obtemos as sintenias genômicas e as representamos os filos desse grupo por: *Methanosarcina mazei* Go1 IH1 (*Euryarchaeota*-Archaea), *Clostridium acetivum* DSM1496 (*Firmicutes*), *Ilyobacter polytropus* DSM2926 (*Fusobacteria*), *Desulfomonile tiedjei* DSM6799 ( $\delta$ -proteobacteria), *Treponema azotonutricium* ZAS9 (*Spirochaetes*), *Coralimargarita akajimensis* DSM 45221 (*Verrucomicrobia*), *Dehalogeimonas* sp WBC 2 (*Chloroflexi*), *Paludibacter propionicigenes* WB4 (*Bacteroidetes*) e *Chlorobium limicola* DSM 245 (*Chlorobi*). Conforme ilustrado na FIGURA 27, identificamos que as espécies possuem o padrão mais compacto da constituição geral do núcleo *nif*.

FIGURA 27 – SINTENIA GENÔMICA PARA G3



FONTE: O Autor (2022).

LEGENDA: Sintenia de genes identificados para o grupo 3. Genes hipotéticos e hipotéticos conservados encontram-se enumerados.

A configuração do núcleo mais abundante nesse grupo foi representada pelo modelo clássico anaeróbico, *nifHI<sub>1</sub>I<sub>2</sub>DKENB*. Para certos organismos, há ocorrência das estruturas acessórias NifX e NifV junto ao núcleo (e.g.: *Ilyobacter polytropus* DSM2926). No entanto, em *Coralimargarita akajimensis* foi identificado o regulador positivo NifA na ausência dos parálogos NifI<sub>1</sub>I<sub>2</sub> caracterizando o núcleo por *nifHDKENBVA*. Obtemos sequências de NifA em simultâneo com os reguladores

anaeróbicos *Nifl<sub>1</sub>l<sub>2</sub>* nas espécies de *Bacteroidetes*, *Spirochaetes*, *Verrucomicrobia* e em  $\delta$ -proteobactérias. Outros genes relacionados à FBN foram identificados, tais como *fdx2* e o *cluster mod*, responsável no transporte e metabolismo de molibdênio.

O G4 se caracterizou pela abundância de filos heterogêneos, majoritariamente anaeróbicos obrigatórios e alguns organismos anaeróbicos facultativos ou microaeróbicos. As sintenias foram representadas na figura FIGURA 28. Identificamos, para certas classes de organismos, o aumento de genes *nif* em relação ao grupo anterior, nos quais destacamos *Methanotermococcus* sp. IH1 e *Heliobacterium modesticaldum* Ice1 (*nifl<sub>1</sub>l<sub>2</sub>HDKENXB*), *Kyrpidia spormannii* EA1 (*nifBVSHDKENXXaHesA*) e *Paenibacillus* spp. (*nifBHDKENXXahesAV*), *Nitrospirae*: *Dissulfurispira thermophila* T55J (*nifHDKENXI<sub>1</sub>l<sub>2</sub>B*) e *Thermodesulfovibrio yellowstonii* DSM 11347, *Aquificales*: *Hydrogenobacter thermophilus* TK6 e *Thermocrinis albus* DSM 14484 (*nifl<sub>1</sub>l<sub>2</sub>TBWZVSAHDKXaENX*), *Deferribacteres*: *Denitrovibrio acetiphilus* DSM 12809 (*nifAZHDKTENXXaVB*) e *Calditerrivibrio nitroreducens* DSM 19672 (*nifVBAZENXXaHDKT*), *Chrysogenetes*: *Desulfurispirillum indicum* S5 (*nifVBASUPHDKTZENX*), *Actinobacteria*: *Frankia alni* EA1 (*nifVHDKENXXaWZBHesAhesB*), proteobactérias delta: *Desulfuromonas soudanensis* WTL (*nifHDKENXI<sub>1</sub>l<sub>2</sub>B*), *Anaeromyxobacter* sp. K (*nifBHDKENXA*) e *Geobacter pickeringii* G13 (*nifVHDKENXB*) e proteobactérias épsilon: *Sulfurospirillum barnesii* SES3 e *Arcobacter nitrofigilis* DSM 7299 (*nifVBHDKTAENXXaFWZQ*).



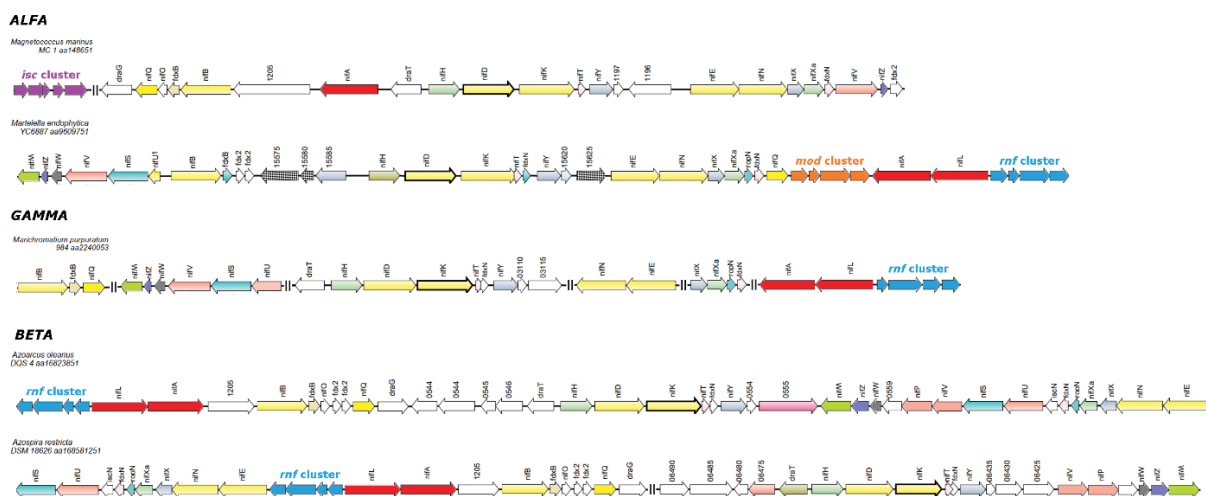
FONTE: O Autor (2022).

LEGENDA: as barras paralelas (-//-) representam os *gaps* de regiões genômicas. Genes hipotéticos e hipotéticos conservados encontram-se enumerados.

Conforme apresentado, identificamos para este grupo núcleos *nif* de tamanhos variáveis e o surgimento da proteína associada a NifX (sugerimos a nomenclatura do gene para *nifXa*).

No conjunto G5 identificamos espécies de proteobactérias alfa, beta e gama. Obtemos as sintenias e as representamos na FIGURA 29 os filios desse grupo por *Magnetococcus marinus* MC-1 e *Marteella endophytica* YC6887 (proteobactérias alfa), *Marichromatium purpuratum* 984 (proteobactéria gama) e *Azoarcus olearius* DQS4 e *Azospira restricta* DSM 18626 (proteobactérias beta). Em *M. marinus* o núcleo é o mais compacto em relação aos demais organismos do grupo (*nifABQHDKTYENXXaVZ*), com integração do *cluster isc*. Identificamos a proteína NifL apenas em uma proteobactéria alfa, a saber, *M. endophytica* (*nifUSVWZMBHDKTYENXXaQLA*). As espécies de  $\beta$ -proteobactéria deste grupo são exclusivas por conterem a proteína NifL. E em *Halorhodospira* spp. como o único organismo das espécies fixadoras das  $\gamma$ -proteobactérias por não o conter (e.g.: *H. halophila* SL1, *nifENXXaABQVWZMHDKY*). *Clusters* de genes vizinhos aos núcleos *nif* os quais integram outros sistemas foram reportados e que também possam atuar na FBN, tais como: *iscAUS* e *rnfABCDEH* relacionados à incorporação de Fe-S e de transferência de elétrons, respectivamente, fornecidos ao CEN.

FIGURA 29 - SINTENIA GENÔMICA PARA G5



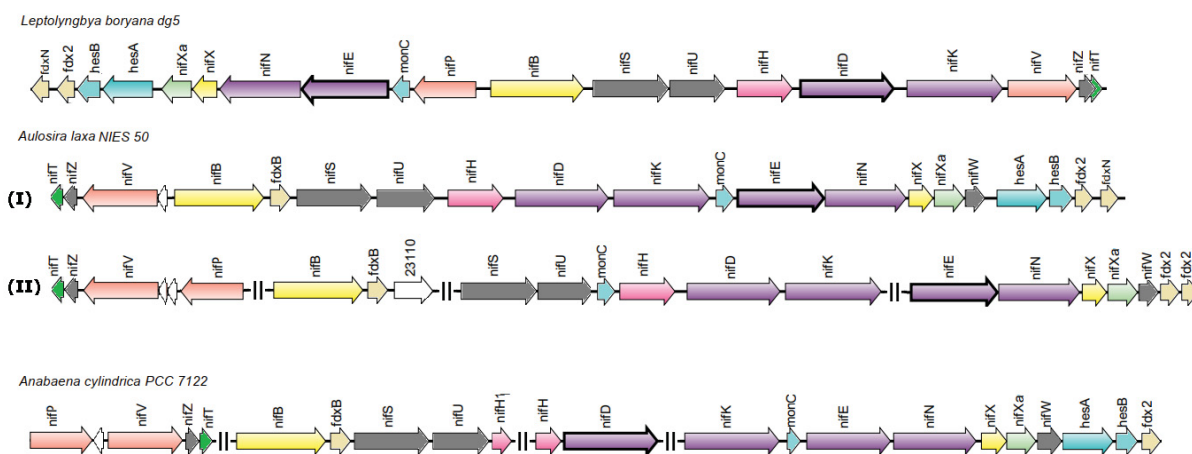
FONTE: O Autor (2022).

LEGENDA: as barras paralelas (-/-) representam os *gaps* de regiões genômicas. Genes hipotéticos e hipotéticos conservados encontram-se enumerados.

O G6 foi composto por todas as espécies de cianobactérias deste estudo. As sintonias foram representadas por *Aulosira laxa* NIES50, *Leptolyngbya boryana* dg5 e *Anabaena cylindrica* PCC 7122. Conforme ilustramos na FIGURA 30, identificamos a composição das proteínas Nif do núcleo *nif* com relativo padrão nas cianobactérias: *nifPVZTBSUHDKENXXaW*. Proteínas foram reportadas e aparentam desempenhar papéis significativos à FBN apresentados nesses núcleos, como: *fdxB*, *hesA*, *hesB* e Mo-nitrogenase C-terminal (sugerimos a nomenclatura de gene *monC*).

FIGURA 30 - SINTENIA GENÔMICA PARA G6

#### CYANOBACTERIA



FONTE: O Autor (2022).

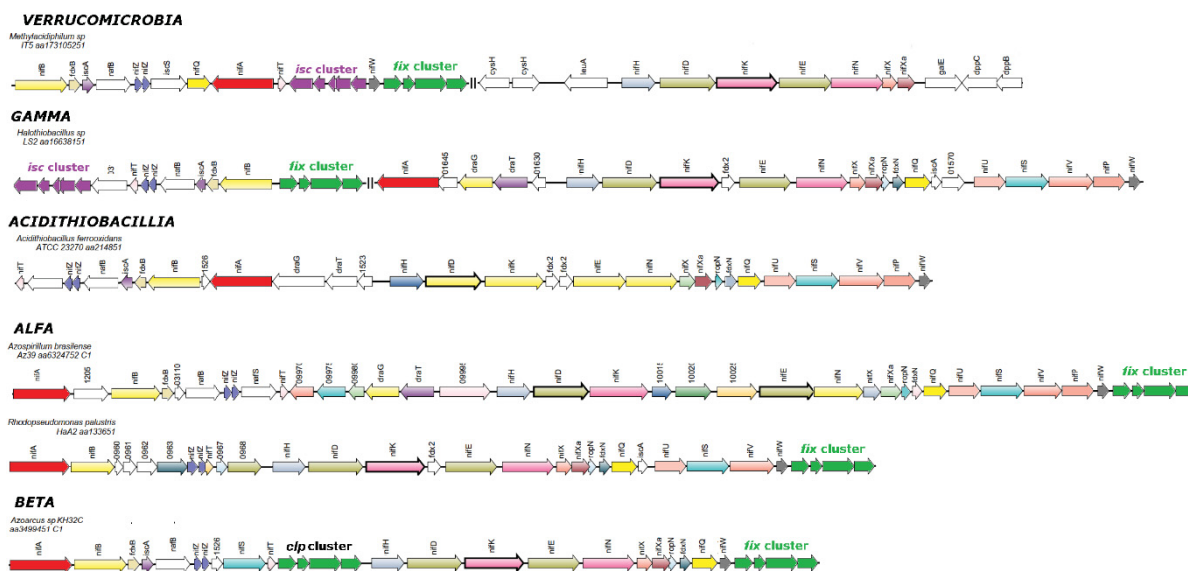
LEGENDA: os espaços com barras paralelas (-/-) representam o “corte” de regiões genômicas. Em *Aulosira laxa* NIES 50 foram identificados dois núcleos *nif* representados por I e II.

O G7 configurou-se por organismos pertencentes às classes de *Nitrospirae*, *Verricimicrobia* e pelas proteobactérias alfa, beta, gama e *Acidithiobacillia*. As sintonias foram representadas por *Halothiobacillus* sp. LS2 (proteobactéria gama), *Azospirillum brasilense* Az39 (proteobactéria alfa), *Azospira oryzae* OS e *Azoarcus* sp. KH32C (proteobactérias beta), *Acidithiobacillus ferrooxidans* ATCC 23270 (*Acidithiobacillia*), *Methylacidiphilum* sp. IT5 (*Verricimicrobia*) e *Leptospirillum ferrooxidans* C23 (*Nitrospirae*). Conforme reportamos na FIGURA 31, a composição das proteínas Nif do núcleo *nif*, apresentaram-se como:



*nifABZ(Z)THDKENXXaQUSVPW*. Existe para a maioria dos casos, a duplicação do gene *nifZ* (e.g.: *Azoarcus* sp. KH32, *nifABZ<sub>1</sub>Z<sub>2</sub>STHDKENXXaQW*). Clusters de genes *isc* composto por *hscA*, *iscA*, *iscN*, *iscU*, *iscS*, *rnf* composto por *rnfABCDEH*; *fix* composto por *fixABCX* foram reportados.

FIGURA 31 - SINTENIA GENÔMICA PARA G7

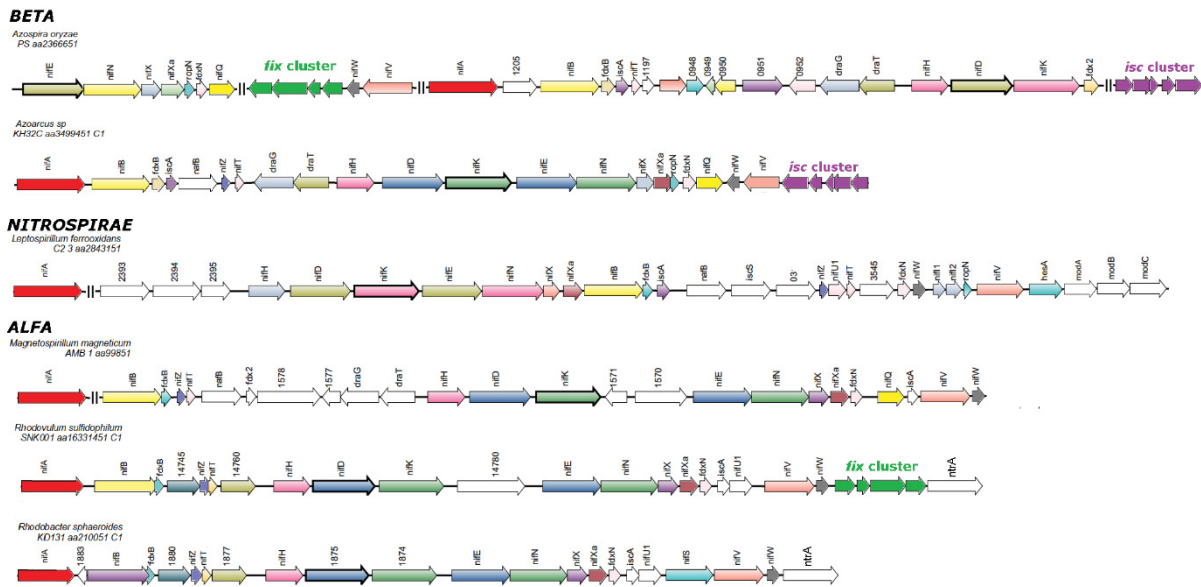


FONTE: O Autor (2022).

LEGENDA: as barras paralelas (//) representam os *gaps* de regiões genômicas. Genes hipotéticos e hipotéticos conservados encontram-se enumerados.

O G8 apresentou espécies de proteobactérias alfa, as quais representamos por *Blastochloris viridis* DSM 133, *Zymomonas mobilis* subsp *mobilis* ZM4 substr 8b, *Azoarcus* sp. CIB, *Rhodobacter blasticus* 285 e *Rhodovulum sulfidophilum* DSM 1374 (FIGURA 32). De modo geral, a composição dos elementos do núcleo *nif*, apresenta-se como: *nifABZTHDKENXXaQUSVW*. Clusters *isc* (*iscANUS*) e *fix* (*fixABCX*) também foram reportados na vizinhança de genes.

FIGURA 32 - SINTENIA GENÔMICA PARA G8

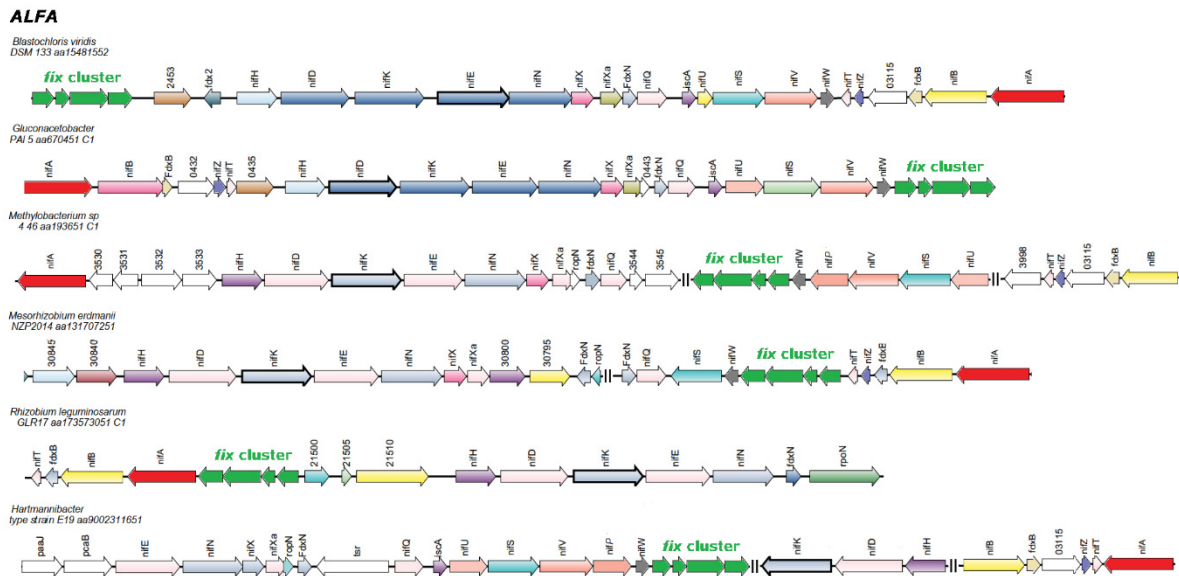


FONTA: O Autor (2022).

LEGENDA: as barras paralelas (---) representam os *gaps* de regiões genômicas. Genes hipotéticos e hipotéticos conservados encontram-se enumerados.

O G9 foi representado pelas  $\alpha$ -proteobactérias *Gluconacetobacter diazotrophicus* PA1 5, *Hartmannibacter diazotrophicus*, *Methylobacterium nodulans* ORS 2060, *Mesorhizobium erdmanii* NZP2014, *Rhizobium leguminosarum* GLR17 e *Sinorhizobium meliloti* (FIGURA 33). Não conseguimos identificar sintenia de genes *nif* por syntTax em *Neorhizobium* spp., *Sinorhizobium fredii*, *Ensifer sojae* e nas demais espécies de *Rhizobium* spp., possivelmente, por se tratarem de organismos com os genes da FBN em organelas plasmidiais. Os núcleos se assemelharam em constituição por *nifABZTHDKENXXaQUSV(P)W*. Cluster de genes *fix* vizinho ao núcleo *nif*. Várias sequências hipotéticas identificadas por RAFTS<sup>3</sup>G apresentaram homologia com insertases do grupo IS (e.g.: *iss3*).

FIGURA 33 SINTENIA GENÔMICA PARA G9



FONTE: O Autor (2022).

LEGENDA: as barras paralelas (-/-) representam os *gaps* de regiões genômicas. Genes hipotéticos e hipotéticos conservados encontram-se enumerados.

O G10 constituiu o maior número de espécies de *Rhizobium* spp., por *Sinorhizobium americanum* e *Sinorhizobium fredii* NXT3. As sequências Nif desse grupo foram identificadas em plasmídeos e, por este motivo, não foi possível o mapeamento da sintenia por SyntTax. No entanto, pelos resultados obtidos por RAFTS<sup>3</sup>G, o conjunto das proteínas Nif tende a se assemelhar ao G9 (*nifABZTHDKENXXa*). Entretanto, para *Rhizobium esperanzae* N561 há cópias dos componentes estruturais do arcabouço da Mo-nitrogenase: *nifHDK* (enumeramos aqui por H<sub>2</sub> D<sub>2</sub> e K<sub>2</sub>) e de outra cópia de *nifH* (H<sub>3</sub>) formando *nifQH<sub>3</sub>H<sub>2</sub>D<sub>2</sub>K<sub>2</sub>ABZTUSWHDKENXXa*. Identificadas proteínas Nod, Fix, proteínas relacionadas a insertases (e.g.: ISS) e citocromo P (CYP450).

## 6 DISCUSSÃO

### 6.1 PREDIÇÃO DE ORGANISMOS FIXADORES

O gênero *Skermanella* é descrito como gênero de bactérias incapaz de fixar nitrogênio (AN *et al.*, 2009; GUO *et al.*, 2020; SLY; STACKEBRANDT, 1999; WEON *et al.*, 2007). Entretanto, neste estudo, identificamos as sequências Nif core completas, além de Nif acessórias (NifA, NifW, NifP, NifS, NifU, NifX, NifXa, NifT e NifJ) nas espécies *Skermanella pratensis* (ASM884314v1) e *Skermanella* sp TT6 (ASM1665363v2). O presente estudo sugere que, ao menos para essas duas espécies, a FBN é viável o que não acarreta na incapacidade de fixação biológica para o gênero. Em contrapartida, Samaddar e colaboradores em 2011 reportaram evidência de nitrogenase funcional e de *nifH* para *Asaia bogorensis* NBRC 16594/MTCC 4041 (ASM154799v1) (SAMADDAR *et al.*, 2011). No entanto, não foi possível identificar nenhuma ocorrência *in silico* da Mo-nitrogenase nessa espécie.

De 167 espécies prováveis não fixadoras trazida pela estratégia de mineração, em 33 organismos foram identificadas cópias do gene *nifH* (e.g.: *Thermacetogenium phaeum* DSM 12270). Em casos particulares, pode-se encontrar evidências da Mo-nitrogenase pela identificação de quase totalidade das sequências Nif core. Vale salientar os casos de *Thermacetogenium phaeum* DSM 12270 (ASM30593v1) no qual identificamos as proteínas NifH, NifD, NifK, NifE e NifB ou *Roseiflexus castenholzii* DSM 13941 (ASM1780v1) e *Roseiflexus* sp. RS1 (ASM1666v1) nos quais identificamos as estruturas NifH, NifD, NifK e NifB. Em ambos, existe algum “elo perdido” destes organismos com a FBN. Portanto, o estudo desses dados poderia dar suporte a novas inferências evolutivas no contexto da Mo-nitrogenase, sobre, por exemplo, perda ou alterações funcionais da Mo-nitrogenase entre os organismos fixadores e não fixadores.

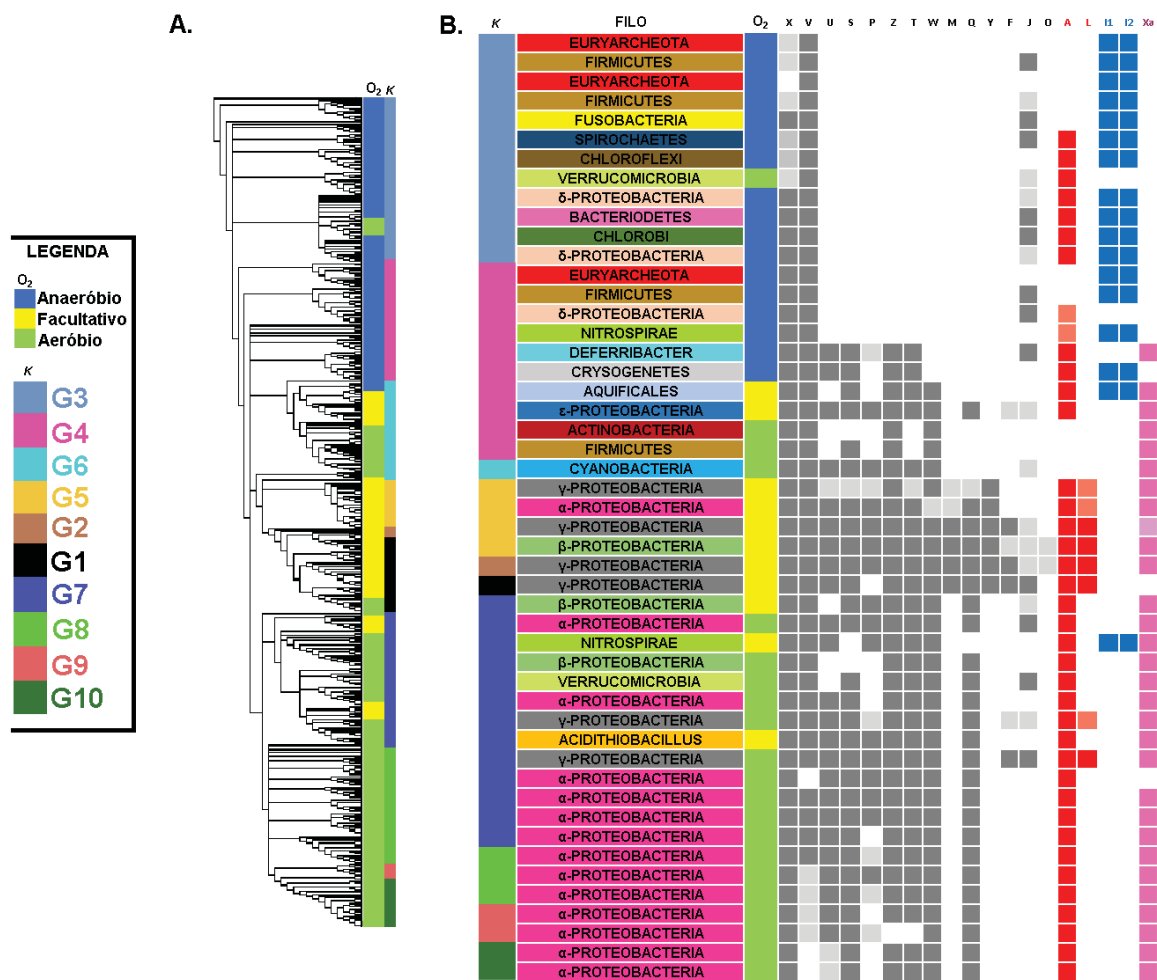
Reportamos também que a criação de NifFINDER como um preditor de sequências Nif possui grande potencial, visto que seu único e principal concorrente NifPRED não foi projetado para comportar o máximo de proteínas Nif estabelecidas. No entanto, maiores testes com NifFINDER precisam ser realizadas, visando publicação em trabalhos futuros. NifFINDER foi registrado sob no INPI (ANEXO, FIGURA 1), bem como, em parceria, a ferramenta GBK2cluster (Lobo *et al.* 2019 – não publicado) (ANEXO, FIGURA 2).

## 6.2 ANÁLISE DOS GRUPOS

O mecanismo da FBN parte de espécies anaeróbias até aeróbias, acompanhando o ciclo geológico da Terra (BOYD, E. S. *et al.*, 2011; MUS *et al.*, 2019). Ao analisarmos a classificação da Mo-nitrogenase proposta por Raymond e colaboradores (2004) em aeróbias (*Cluster I*) de anaeróbias (*Cluster II*), não seria possível a obtenção de informações sobre a dinâmica molecular das espécies fixadoras. Além disso, a herança da Mo-nitrogenase não segue a ordem taxonômica. Sugere-se que os eventos de THG tenham sido cruciais para o refinamento da FBN. A FIGURA 34 resume as informações obtidas para compreendermos melhor essa dinâmica através dos 10 grupos analisados.

Ademais, pelas informações obtidas com a PCA, foi possível identificar nos grupos gerados por HCPC aqueles com os maiores auto valores (G1, G3, G6 e G10). Anotamos percepções consideráveis na composição geral dos respectivos núcleos *nif* inter-grupos, ao passo que existe maior uniformidade intra-grupo na constituição molecular de genes *nif* em relação aos grupos fora desta lista. O G3 foi composto pelas arqueias e bactérias anaeróbicas, o núcleo *nif* é compacto, constituído por *nifHI<sub>1</sub>I<sub>2</sub>DKENB* (exceção para *nifHDKENBA* em *Coralimargarita akajimensis* DSM 45221). O G1 foi composto por espécies de proteobactéria gama e identificado o maior conjunto gênico de *nif*, a saber: *nifJHDKTYENXUSVWZMFLABQ*. G6 foi representado pelas cianobactérias e identificado o núcleo *nifPVZTBSUHDKENXXaW*. Por fim, o G10 foi constituído por espécies simbióticas fixadoras obrigatórias e, apesar de não identificarmos a sintenia através de SyntTax, sugere-se que exista proximidade com G9 em que obtemos: *nifABZTHDKENXXaQ*. Os grupos G3, G4, G5, G6, G7, G8 e G9 compuseram os conjuntos de maior proporção de filós taxonomicamente distintos e, ao mesmo tempo, formaram os grupos de maior variação constitutiva de genes *nif* intragrupo, a exemplo, G4 em que de *Nitrospirae* para *Deferribacter* identificamos uma diferença de 6 sequências Nif, a saber: NifU, NifS, NifP, NifZ, NifT e NifJ (FIGURA 34).

FIGURA 34 – ÁRVORE CORE E NIF IDENTIFICADAS



FONTE: O Autor (2022).

LEGENDA: Árvore Nif core (A.), caracterização dos organismos aeróbios em cor verde, anaeróbios facultativos e obrigatórios em amarelo e azul, respectivamente (coluna O<sub>2</sub>). E a distribuição dos grupos obtidos (coluna k). Tabela de Nif (excetuando-se o conjunto Nif core) identificados neste estudo (B.) coloridos por intensidade (cores mais claras indicam que nem todas as espécies do filo na linha identificamos o Nif respectivo na coluna).

As arqueias e firmicutes (G3) foram os organismos primitivos de maior domínio de THG do núcleo *nif* anaeróbio (*nifH1I2DKENB*). Em determinado período a evolução do CEN favoreceu a incorporação de *nifV*, antes relacionado a biossíntese de aminoácido L-leucina (LeuA) ao conjunto *nif*. Verificamos que δ-proteobactérias fixadoras anaeróbicas se relacionaram por THG com bactérias anaeróbicas gram positivas de Firmicutes (e.g.: *Clostridium* spp.) e *Chlorobi* (e.g.: *Chlorobium* spp.). Apesar de não ter sido expressivo o aumento de elementos *nif*, foi reportado grande interação destes com espécies anaeróbicas de filos distintos, tais como: *Fusobacteria*, *Spirochaetes*, *Verricumicrobia*, *Chloroflexi* e *Bacteroidetes*. A interação de organismos

anaeróbios tão diversos poderia muito bem favorecer a disseminação do conjunto *Nif core* para filios mais distantes de bactérias.

Observamos, em nossos resultados, que pela posição na árvore *Nif core* das arqueias de ordens *Methanococcales* e *Methanobacteriales* e por terem agrupado ao G4, é provável que tenham adquirido mais tardiamente a Mo-nitrogenase em relação aos demais arqueotes (e.g.: *methanosarcina* spp.). As espécies anaeróbicas de proteobactérias delta deste grupo participaram de THG com outros gêneros de Firmicutes (e.g.: *Dehalobacter* e *Desulfosporosinus*). Inclusive observamos transferência entre organismos sem a regulação descrita para os organismos anaeróbicos *nif<sub>1</sub>* (e.g.: *Geobacter*). A transição dos reguladores anaeróbicos para *NifA* ainda não está totalmente esclarecida, entretanto, o G4, em particular, permite a análise de certas evidências, tais como, a ocorrência de *NifA* em espécies anaeróbicas (e.g.: *nifH<sub>1</sub>*DKENBA *Desulfomonile tiedjei* DSM 6799 ) e a ocorrência em simultâneo dos reguladores *nifA* e *nif<sub>1</sub>* no filo Aquificales (I<sub>1</sub>2TBWZVSAHDKXaENX) conhecidas por espécies marítimas de vida livre submersas em condições extremas de temperatura e pressão (ARAI *et al.*, 2010; NISHIHARA *et al.*, 2018). Novamente a THG que, desta vez, englobou filios distintos anaeróbicos facultativos e/ou micro-aerófilos, a saber: Firmicutes (e.g.: *Kyrpidia spormannii* EA-1), Actinobacteria (e.g.: *Frankia alni* ACN14a), Aquificales (e.g.: *Hydrogenobacter thermophilus* TK-6), Nitrospirae (e.g.: *Dissulfurispira thermophila* T55J), Deferribacteres (e.g.: *Calditerrivibrio nitroreducens* DSM 19672) e as proteobactérias épsilon (e.g.: *Sulfurospirillum barnesii* SES-3) nos quais notamos o recrutamento de elementos *nif* em relação ao núcleo *nif* anaeróbio (e.g.: *nifXa*, *nifU*, *nifS*, *nifW*, *nifZ*, *nifT*). Outra possibilidade sobre as alterações da regulação da FBN de anaeróbios para aeróbios é a passagem do sistema pós traducional para pós transcricional sigma 70 dependente, mecanismo descrito em espécies fixadoras do gênero *Paenibacillus* (LI *et al.*, 2014; WANG, L. *et al.*, 2013) e que justificariam semelhanças moleculares da regulação desse grupo com espécies fixadoras delta dos gêneros *Geobacter* e *Pelobacter*, além da proximidade do grupo com as Cianobactérias (G6).

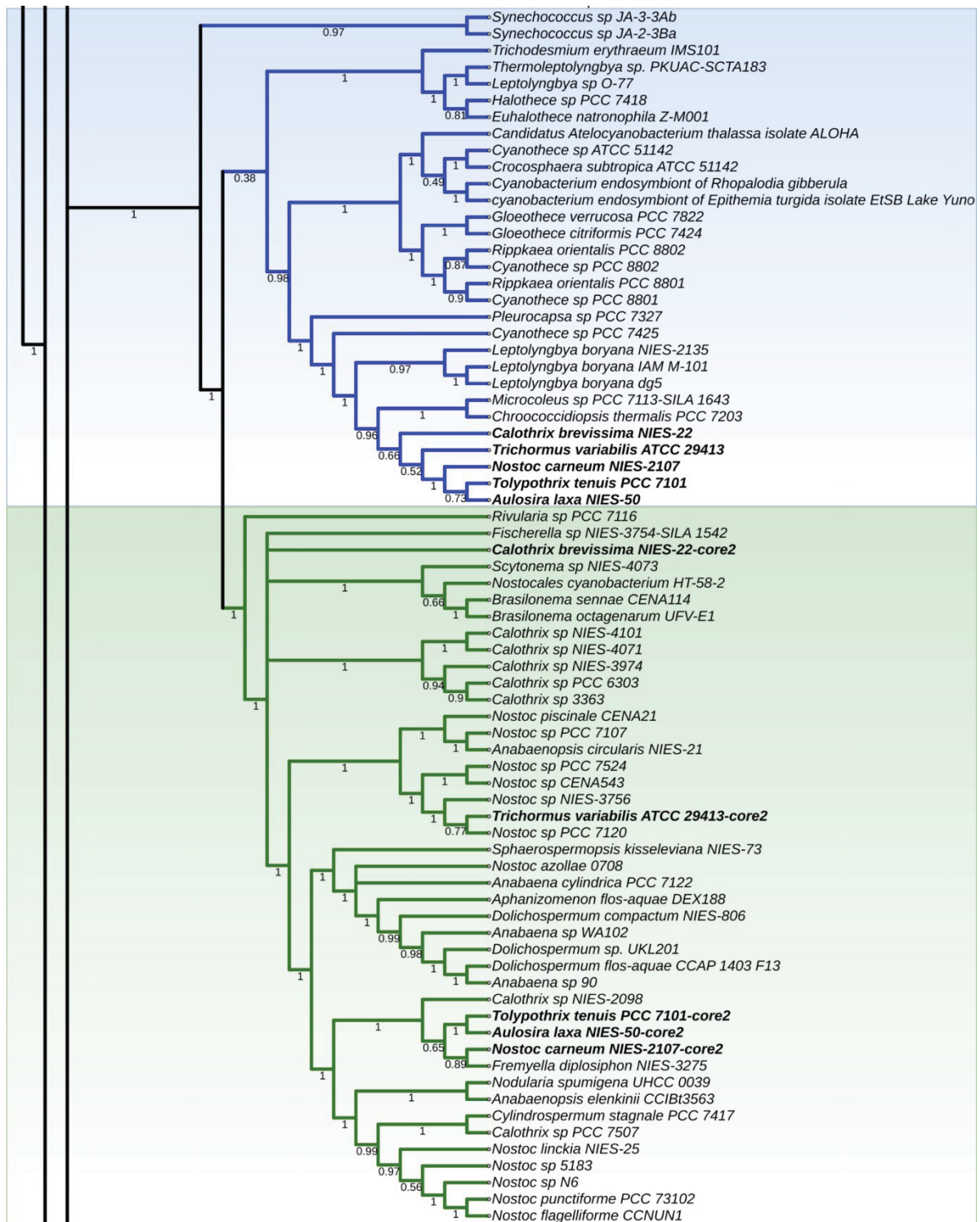
Segundo Watanabe e Horiike (2021), a aquisição da Mo-nitrogenase nas cianobactérias distribuem as espécies em um padrão descrito por 'mosaico' e, devido a estrutura de *operon* preservada, sugerem que a transferência da nitrogenase é restrita entre às cianobactérias (WATANABE; HORIIKE, 2021). Mesmo em condições adversas à FBN, como níveis celulares mais elevados de oxigênio e fotossíntese, as

cianobactérias desenvolveram mecanismos específicos de proteção da nitrogenase. Em algumas espécies de cianobactérias há a formação de um tipo celular especializado para que ocorra a FBN em condições de aerobiose celular e formado em resposta a falta de nitrogênio fixado ( $\text{NH}_4$  ou  $\text{NO}_3$ ), denominada de heterocisto. Estima-se que as cianobactérias formadoras de heterocistos tenham surgido há cerca de 1.000-720 milhões de anos (PANG *et al.*, 2018). Além disso, é descrito que em organismos que produzam heterocistos há produção de outro tipo de célula especializada móvel denominada hormogônio e não possuem o fotossistema II (responsável pela fotossíntese oxigenada na presença de luz). Os hormogônios são responsáveis pelo desenvolvimento das cianobactérias em simbiose em resposta a um fator indutor de hormogônio (HIF) secretado por plantas hospedeiras. As cianobactérias, então, diferenciam-se em heterocistos e entram em uma simbiose de trabalho com a planta hospedeira (CAMPBELL; MEEKS, 1989; HERRERO; STAVANS; FLORES, 2016). Em cianobactérias não formadoras de heterocistos, a fixação ocorre nas células vegetativas em baixa concentração de oxigênio, através da ativação da expressão do fator de transcrição, CnfR2 (C-terminal XRE-*type*), que ativa o gene *nifB2*. O gene *ntcA* regula a FBN em organismos formadores de heterocistos com a participação de outros genes, em que, por exemplo, a proteína PipX e 2-oxaglutarato se ligam a NtcA.

Identificamos a separação filoprotética das estirpes entre formadoras e não formadoras de heterocistos (FIGURA 35, cladogramas evidenciados em azul e em verde, respectivamente). No cluster dos não formadores (NFH), encontram-se os diazotrofos da ordem *Oscillatoriales*, *Chroococcidiopsidales*, *Pleurocapsales* e *Chroococcales*. No cluster dos formadores de heterocistos (FH) *Nostocales*, *Stigonematales* e *Synechococcales*. As estirpes *Calothrix brevissima* NIES22, *Trichormus variabilis* ATCC 29413, *Tolypothrix tenuis* PCC 7101, *Aulosira laxa* NIES50 e *Nostoc carneum* NIES2107 apresentaram duplicação do conjunto Nif core (conjunto representado na filogenia por *core-2*) e de sequências Nif acessórias. Além disso, devido a localização dos organismos na árvore dividindo NFH de FH, sugere-se que, nessas espécies, possam conter os dois sistemas da nitrogenase e que há uma derivação do cluster NFH para o cluster FH.



FIGURA 35 – FILOGENIA CORE DAS CIANOBACTÉRIAS (G6).



FONTE: O Autor (2022).

LEGENDA: Recorte da filogenia *Nif core* nas espécies cianobacterianas, divididas em não formadores de heterocistos (clado na cor azul) e formadores de heterocistos (clado na cor verde). As estirpes cianobacterianas *Calothrix brevisissima* NIES22, *Trichormus variabilis* ATCC 29413, *Tolypothrix tenuis* PCC 7101, *Aulosira laxa* NIES50 e *Nostoc carneum* NIES2107 apresentaram os dois tipos de sistemas da nitrogenase e duplicação do conjunto *Nif core* e de *Nif* acessórias.

As proteobactérias alfa, beta e gama, diferente das proteobactérias delta e épsilon, adquiriram a Mo-nitrogenase mais tardiamente (KOIRALA; BRÖZEL, 2021). Reportamos em G5, G2 e G1 como os grupos formados por espécies dos maiores detentores de genes *nif* (vide FIGURA 33). Outro fator característico dos três agrupamentos é a predominância do regulador negativo NifL e exclusividade no genoma das espécies destes grupos. Assim como em NifA, não se tem descrição da origem ou da perda do regulador negativo para a FBN. Em G5 a ocorrência de THG entre espécies de proteobactérias beta (e.g.: *Azoarcus communis* TSPY31), alfa (e.g.: *Marteella endophytica* YC6887) e gama (e.g.: *Halorhodospira halochloris* DSM 1059). Foi relatado a disposição de seis unidades regulatórias de genes *nif* inseridos em dois núcleos *nif*, estudados no organismo modelo *Azotobacter vinelandii* DJ: o cluster menor (*nifLA*, *nifB*, *nifOQ*) e o cluster maior (*nifHDKTY*, *nifENX*, *nifUSVPWZM*) (BURÉN *et al.*, 2020). A proximidade de várias espécies diferentes com *operons* similares a deste organismo (FIGURA 25) sugere que as THG (FIGURA 31) possam ter contribuído para o rearranjo estrutural dessas unidades regulatórias e foram transmitidos às espécies subsequentes de gama proteobactérias dos grupos G2 (e.g.: *Pectobacterium polaris* NIBIO1006) e G1 (e.g.: *Klebsiella variicola* 342) os organismos de maior número de genes *nif* em um único fragmento (e.g.: *nifJHDKTYENXUSVZMFLABQ* em *K. variicola* 342).

Os grupos G7, G8, G9 e G10 compuseram os fixadores de grande interesse biotecnológico, devido às características praticamente exclusivas que possibilitaram a esses organismos a simbiose com outros organismos (e.g.: leguminosas) e que, aparentemente, se relacionaram com a FBN. Em G7 foi reportado THG entre *Nitrospirae* (e.g.: *Leptospirillum ferriphilum* YSK), *Verricomicrobia* (*Methylacidiphilum fumariolicum* SolV), *Acidithiobacillia* (*Acidithiobacillus ferrivorans* SS3) e as proteobactérias alfa (e.g.: *Bradyrhizobium elkanii* USDA 61), beta (e.g.: *Herbaspirillum seropedicae* SmR1), gama (e.g.: *Halothiobacillus* sp. LS2). Ademais em G7 reportamos a duplicação de *nifZ* (*NifZ*<sub>1</sub> e *NifZ*<sub>2</sub>) para a maioria dos organismos, com um certo grau de conservação (FIGURA 27). Em G8 (FIGURA 28) a sintenia genômica se assemelha estruturalmente à G7, entretanto, a duplicação de *nifZ* supramencionada é ausente.

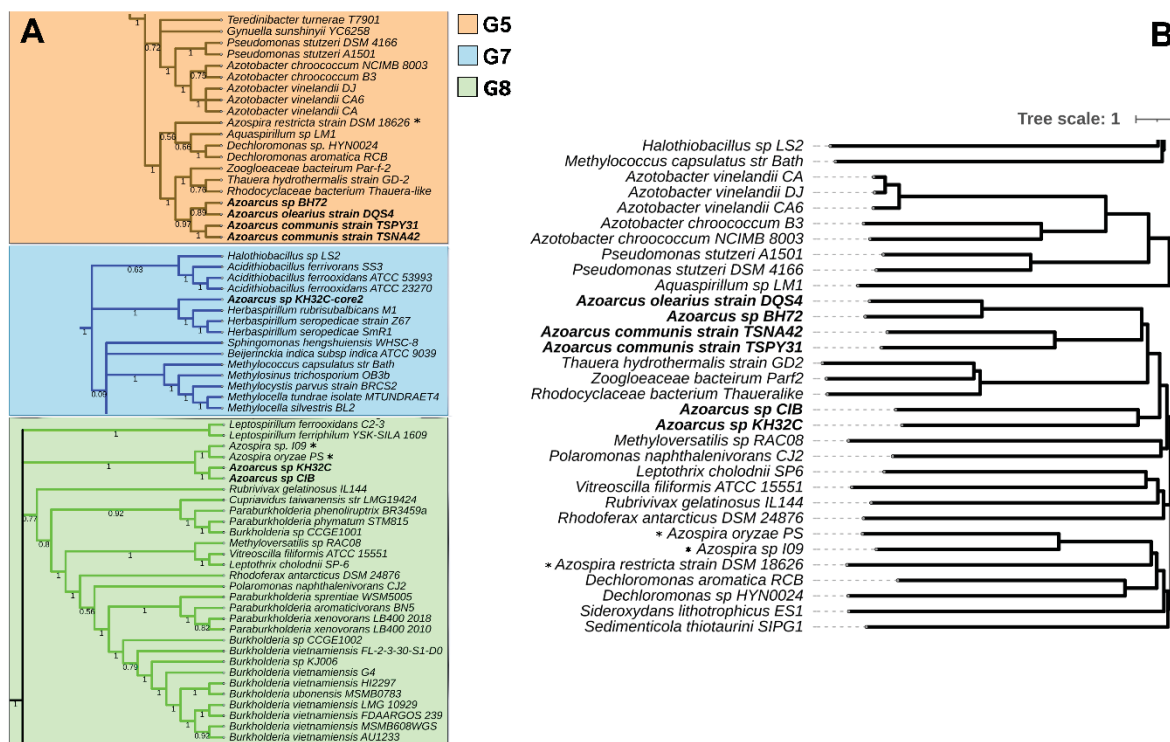
É importante destacar que há uma diminuição gradativa de genes *nif* que vai de G7 e termina em G10. Possivelmente, genes *nif* característicos à maturação da nitrogenase em organismos de vida livre são ‘selecionados’ por genes de interação

fixador-hospedeiro (e.g.: genes *fix* e *nod*). Aliado a isso, essa aparente redução possivelmente se relaciona a disponibilidade de outros produtos celulares, oriundos da alta variedade metabólica, incluindo sistemas fotossintéticos anaeróbios e magnetotáticos (WANG, Y. *et al.*, 2019). Em grande parcela das espécies fixadoras alfa simbiotes as sequências *nif* foram obtidas de cromossomos nucleares (G7 e G8). No entanto, a partir das espécies analisadas em G9 e mais expressivamente em G10 ocorreu aparente migração dos núcleos *nif* para as organelas plasmidiais (plasmídeo simbiótico). Este fenômeno é conhecido para as espécies do gênero *Rhizobium* e *Ensifer* (*Sinorhizobium*) (GONZÁLEZ *et al.*, 2003; YANG *et al.*, 2020). Além disso, foi reportado o gene *fen1*, que codifica para uma homocitrato sintase, em espécies de *Aeschynomene* e em *Lotus japonicus* (HAKOYAMA *et al.*, 2009; NOUWEN *et al.*, 2017). Portanto, devido às condições desfavoráveis de fixação ou de simbiose, os genes *nif* podem ter sido perdidos ou transferidos para outros tipos de bactérias e até mesmo, organismos taxonomicamente superiores.

### 6.3 THG: CASOS AZOARCUS E AZOSPIRA

Os casos dos gêneros *Azoarcus* e *Azospira* são os mais explícitos em nosso estudo sobre os impactos da THG para a divergência evolutiva da Mo-nitrogenase a nível de gênero. Em ambos tivemos provas de divergências intraespecíficas na aquisição da Mo-nitrogenase. Em *Azoarcus olearius*, as estirpes DQS4 e BH72, e em *Azoarcus communis*, as estirpes TSNA42 e TSPY31, foram agrupados em G5. Em *Azoarcus* sp. CIB e *Azoarcus* sp. KH32C, estes foram agrupados em G7. Ademais, divergências desses organismos na topologia da árvore *Nif core* (FIGURA 36, árvore da esquerda) *versus* árvore proteoma (FIGURA 36, árvore da direita), indicam que espécies do gênero *Azoarcus* pertençam a gêneros distintos. No trabalho de Raittz e colaboradores em 2021 reportam a divergência e, inclusive, sugerindo a reclassificação das espécies *Azoarcus* sp. CIB e *Azoarcus* sp. KH32 para o gênero *Aromatoleum* (RAITTZ *et al.*, 2021). De modo análogo, observamos em *Azospira restricta* DSM 18626, agrupado em G5, a divergência na aquisição da Mo-nitrogenase em relação a *Azospira oryzae* OS e *Azospira* sp. I09, os quais foram agrupados em G8.

FIGURA 36 – NIF CORE E PROTEOMA DOS GÊNEROS AZOARCUS E AZOSPIRA.



FONTE: O Autor (2022).

LEGENDA: Comparativo da topologia entre a árvore Nif core (A) versus a árvore do Proteoma (B) obtida para *Azoarcus* spp. e *Azospirae* spp. Observando a árvore Nif core, gênero *Azoarcus* encontra-se destacado em negrito e o gênero *Azospira* destacado com sinal gráfico \*. As espécies *Azoarcus* sp. BH72, *A. olearius* DQS4, *A. comunis* e *Azospira restricta* agruparam-se no G5, *Azoarcus* sp. KH32, *Azoarcus* sp. CIB, *Azospira oryzae* e sp *Azospira* sp. I09 agruparam-se em G8. A espécie *Azoarcus* sp. KH32C (core2) agrupou-se em G7. A divisão de espécies de *Azoarcus* obtida pela árvore Proteoma, sugere a subclassificação de algumas dessas espécies em *Aromatoleum* (*Azoarcus* sp. CIB e *Azoarcus* KH32) de acordo com o proposto em Raittz e colaboradores (2021).

A caracterização sintênica (FIGURA 29, FIGURA 31 e 32) e as proteínas identificadas por RAFTS<sup>3</sup>G fortalecem a distinção evolutiva na aquisição da Mo-nitrogenase em *Azoarcus* spp. e em *Azospira* spp. (TABELA 2). A exemplo disso, a presença de NifO, NifY, NifU, NifM e NifL nas espécies *Azoarcus comunis* TSNA42, *Azoarcus olearius* DQS4/BH72 e *Azospira restricta* DSM 18626 e ausência de Nif supramencionados nas espécies de *Azoarcus* sp. CIB/KH32 e *Azospira oryzae* e *Azospira* sp. I09. Além disso, para *Azoarcus* sp. KH32 existem dois conjuntos Nif core vinculados a dois tipos distintos de núcleos nif, os quais permitiram a separação da espécie em dois grupos distintos, a saber: G7 (*Azoarcus* sp. KH32 core-2) e G8.

TABELA 3 – NIF IDENTIFICADAS PARA AZOARCUS E AZOSPIRAE

ORGANISMO / <i>Nif</i>	X	V	U	S	P	Z	T	W	M	Q	Y	F	J	O	A	L	Xa
<i>Azoarcus communis</i> TSNA42	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Azoarcus olearius</i> DQS4/BH72	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Azospira restricta</i> DSM 18626	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Azospira oryzae</i> PS	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Azospira</i> sp I09	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Azoarcus [Aromatoleum]</i> sp CIB	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Azoarcus [Aromatoleum]</i> sp KH32C	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
<i>Azoarcus [Aromatoleum]</i> sp KH32C – core2	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■

FONTE: O Autor (2022).

LEGENDA: Identificação de *Nif* obtida *Methylacidiphilum* spp os por RAFTS<sup>3</sup>G para *Azoarcus* e *Azospirae*. Os organismos estão destacados em amarelo para o G5, em azul para G7 e em verde para o G8.

## 7 CONCLUSÕES

Conseguimos alcançar os objetivos propostos à construção e a interpretação das possíveis relações evolutivas entre grupos de organismos diazotróficos sob a análise das estruturas *Nif core* codificadas pelos genes *nif* em *Archaea* e *Bacteria*.

Listamos 662 genomas de diazotrofos. Destes, identificamos em 52 organismos como potenciais fixadores, incluindo o gênero estabelecido como não fixador *Skermanella*.

Os diazotrofos foram classificados em 10 grupos por *k-means* (HCPC). A análise destes grupos, sugere que, apesar da Mo-nitrogenase ser atualmente classificada no Cluster I (aeróbios) e II (anaeróbios) de Raymond, existem fortes evidências de que a subclassificação se faz crucial para entendimento evolutivo e molecular da Mo-nitrogenase nas espécies.

Existe o consenso literário do mecanismo da FBN partir de espécies anaeróbias até as aeróbias, acompanhando o ciclo geológico da Terra. Além disso, a herança da Mo-nitrogenase não obedece totalmente a lógica taxonômica. Os eventos de THG são mais comuns do que se imagina e, observando as transferências *versus* número de elementos *nif*, sugerimos que esse tipo de processo evolutivo tenha sido crucial para o refinamento da FBN.

Como produto em paralelo deste trabalho, desenvolvemos o preditor de proteínas *Nif*, *NifFINDER*, o qual possui grande potencial como nova estratégia para mineração de dados em amplo espectro de sequências *Nif*.

Por fim, este trabalho demonstra o potencial da integração do método *SWEEP* às técnicas de aprendizado de máquina na resolução de problemas biológicos complexos. Graças a este estudo, podemos expandir a compreensão sobre a evolução de organismos diazotróficos utilizando-se da bioinformática e dos dados livres disponíveis.

## REFERÊNCIAS

AGGARWAL, C. C. Data Mining. Cham, 2015. Available at: <https://doi.org/10.1007/978-3-319-14142-8>. Acesso em: 13 mar. 2022.

AN, H. *et al.* *Skermanella xinjiangensis* sp. nov., isolated from the desert of Xinjiang, China. **International Journal of Systematic and Evolutionary Microbiology**, [s. l.], v. 59, n. 6, p. 1531–1534, 2009. Available at: <https://doi.org/10.1099/IJS.0.003350-0/CITE/REFWORKS>. Acesso em: 7 mar. 2022.

ANDREOPOULOS, B. *et al.* A roadmap of clustering algorithms: finding a match for a biomedical application. **Briefings in Bioinformatics**, [s. l.], v. 10, n. 3, p. 297–314, 2009. Available at: <https://doi.org/10.1093/BIB/BBN058>. Acesso em: 15 fev. 2022.

ARAI, H. *et al.* Complete genome sequence of the thermophilic, obligately chemolithoautotrophic hydrogen-oxidizing bacterium *Hydrogenobacter thermophilus* TK-6. **Journal of Bacteriology**, [s. l.], v. 192, n. 10, p. 2651–2652, 2010. Available at: <https://doi.org/10.1128/JB.00158-10>

ARGÜELLES, M.; BENAVIDES, C.; FERNÁNDEZ, I. A new approach to the identification of regional clusters: hierarchical clustering on principal components. <http://dx.doi.org/10.1080/00036846.2014.904491>, [s. l.], v. 46, n. 21, p. 2511–2519, 2014. Available at: <https://doi.org/10.1080/00036846.2014.904491>. Acesso em: 18 fev. 2022.

ARNOLD, W. *et al.* Nucleotide sequence of a 24,206-base-pair DNA fragment carrying the entire nitrogen fixation gene cluster of *Klebsiella pneumoniae*. **Journal of Molecular Biology**, [s. l.], v. 203, n. 3, p. 715–738, 1988. Available at: [https://doi.org/10.1016/0022-2836\(88\)90205-7](https://doi.org/10.1016/0022-2836(88)90205-7)

ASGARI, E.; MOFRAD, M. R. K. Continuous distributed representation of biological sequences for deep proteomics and genomics. **PLoS ONE**, [s. l.], v. 10, n. 11, p. 1–15, 2015. Available at: <https://doi.org/10.1371/journal.pone.0141287>

BAHULIKAR, R. A. *et al.* Diversity of nitrogen-fixing bacteria associated with switchgrass in the native tallgrass prairie of Northern Oklahoma. **Applied and Environmental Microbiology**, [s. l.], v. 80, n. 18, p. 5636–5643, 2014. Available at: <https://doi.org/10.1128/AEM.02091-14>

BOCARD, D.; GILLET, F.; LEGENDRE, P. **Numerical Ecology with R**. 2nd.

ed. New York: Springer, 2018.

BOYD, E. S. *et al.* A late methanogen origin for molybdenum-dependent nitrogenase. **Geobiology**, [s. l.], v. 9, n. 3, p. 221–232, 2011. Available at: <https://doi.org/10.1111/j.1472-4669.2011.00278.x>

BOYD, Eric S. *et al.* Evolution of molybdenum nitrogenase during the transition from anaerobic to aerobic metabolism. **Journal of Bacteriology**, [s. l.], v. 197, n. 9, p. 1690–1699, 2015. Available at: <https://doi.org/10.1128/JB.02611-14>

BOYD, Eric S.; PETERS, J. W. **New insights into the evolutionary history of biological nitrogen fixation**. [S. l.]: Frontiers Research Foundation, 2013. Available at: <https://doi.org/10.3389/fmicb.2013.00201>. Acesso em: 12 jun. 2021.

BURÉN, S. *et al.* Biosynthesis of Nitrogenase Cofactors. **Chemical Reviews**, [s. l.], v. 120, n. 12, p. 4921–4968, 2020. Available at: <https://doi.org/10.1021/acs.chemrev.9b00489>. Acesso em: 14 jun. 2021.

CAMPBELL, E. L.; MEEKS, J. C. Characteristics of Hormogonia Formation by Symbiotic *Nostoc* spp. in Response to the Presence of *Anthoceros punctatus* or Its Extracellular Products. **Applied and Environmental Microbiology**, [s. l.], v. 55, n. 1, p. 125–131, 1989. Available at: <https://doi.org/10.1128/AEM.55.1.125-131.1989>. Acesso em: 12 jan. 2022.

CHAKRABORTY, I.; CHOUDHURY, A. A. Artificial Intelligence in Biological Data. **Journal of Information Technology & Software Engineering**, [s. l.], v. v.07, p. n.04, 2017. Available at: <https://doi.org/10.4172/2165-7866.1000207>

CHEUNG, S. *et al.* Diazotroph community structure in the deep oxygen minimum zone of the Costa Rica Dome. **Journal of Plankton Research**, [s. l.], v. 38, n. 2, p. 380–391, 2015. Available at: <https://doi.org/10.1093/plankt/fbw003>. Acesso em: 17 jun. 2021.

COELHO, M. R. R. *et al.* Molecular detection and quantification of *nifH* gene sequences in the rhizosphere of sorghum (*Sorghum bicolor*) sown with two levels of nitrogen fertilizer. **Applied Soil Ecology**, [s. l.], v. 42, n. 1, p. 48–53, 2009. Available at: <https://doi.org/10.1016/j.apsoil.2009.01.010>

COIMBRA, R. C. M. **Ferramenta de Visualização Interativa de Comparação entre Múltiplos Genomas para a Identificação de Sintenias Brasília 2010**. 111 f. 2010. [s. l.], 2010.

DE LIMA NICHIO, B. T. *et al.* RAFTS3G: An efficient and versatile clustering software to analyses in large protein datasets. **BMC Bioinformatics**, [s. l.], v. 20, n. 1,



p. 1–7, 2019. Available at: <https://doi.org/10.1186/s12859-019-2973-4>. Acesso em: 31 maio 2021.

DE PIERRI, C. R. *et al.* SWeeP: representing large biological sequences datasets in compact vectors. **Scientific Reports**, [s. l.], v. 10, n. 1, 2020. Available at: <https://doi.org/10.1038/s41598-019-55627-4>. Acesso em: 31 maio 2021.

DOBEREINER, J. Biological nitrogen fixation in the tropics: Social and economic contributions. *In:* , 1997. **Soil Biology and Biochemistry**. [S. l.]: Pergamon, 1997. p. 771–774. Available at: [https://doi.org/10.1016/S0038-0717\(96\)00226-X](https://doi.org/10.1016/S0038-0717(96)00226-X)

DOS SANTOS, P. C. *et al.* Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. **BMC Genomics**, [s. l.], v. 13, n. 1, p. 162, 2012. Available at: <https://doi.org/10.1186/1471-2164-13-162>

EHLERS, C. *et al.* Functional organization of a single nif cluster in the mesophilic archaeon *Methanosarcina mazei* strain Gö1. **Archaea**, [s. l.], v. 1, n. 2, p. 143–150, 2002. Available at: <https://doi.org/10.1155/2002/362813>

EINSLE, O. *et al.* Nitrogenase MoFe-protein at 1.16 Å resolution: A central ligand in the FeMo-cofactor. **Science**, [s. l.], v. 297, n. 5587, p. 1696–1700, 2002. Available at: <https://doi.org/10.1126/science.1073877>. Acesso em: 17 jun. 2021.

ESTEVEZ-FERREIRA, A. A. *et al.* Cyanobacterial nitrogenases: Phylogenetic diversity, regulation and functional predictions. **Genetics and Molecular Biology**, [s. l.], v. 40, n. 1, p. 261–275, 2017. Available at: <https://doi.org/10.1590/1678-4685-gmb-2016-0050>

GARCÍA KERDAN, I.; MORILLÓN GÁLVEZ, D. Artificial neural network structure optimisation for accurately prediction of exergy, comfort and life cycle cost performance of a low energy building. **Applied Energy**, [s. l.], v. 280, p. 115862, 2020. Available at: <https://doi.org/10.1016/J.APENERGY.2020.115862>

GAUTHIER, J. *et al.* A brief history of bioinformatics. **Briefings in Bioinformatics**, [s. l.], v. 20, n. 6, p. 1981–1996, 2019. Available at: <https://doi.org/10.1093/bib/bby063>

GHAHRAMANI, Z. Probabilistic machine learning and artificial intelligence. [s. l.], 2015. Available at: <https://doi.org/10.1038/nature14541>

GONZÁLEZ, V. *et al.* The mosaic structure of the symbiotic plasmid of *Rhizobium etli* CFN42 and its relation to other symbiotic genome compartments. **Genome biology**, [s. l.], v. 4, n. 6, p. 1–13, 2003. Available at: <https://doi.org/10.1186/GB-2003-4-6-R36/FIGURES/5>. Acesso em: 16 mar. 2022.

GUO, Q. *et al.* *Skermanella pratensis* sp. Nov., isolated from meadow soil, and emended description of the genus *Skermanella*. **International Journal of Systematic and Evolutionary Microbiology**, [s. l.], v. 70, n. 3, p. 1605–1609, 2020. Available at: <https://doi.org/10.1099/IJSEM.0.003944/CITE/REFWORKS>. Acesso em: 18 nov. 2021.

HAESLER, A. V. Do we still need supertrees? **BMC Biology**, [s. l.], v. 10, n. 1, p. 1–4, 2012. Available at: <https://doi.org/10.1186/1741-7007-10-13/FIGURES/1>. Acesso em: 15 fev. 2022.

HAKOYAMA, T. *et al.* Host plant genome overcomes the lack of a bacterial gene for symbiotic nitrogen fixation. **Nature**, [s. l.], v. 462, n. 7272, p. 514–517, 2009. Available at: <https://doi.org/10.1038/nature08594>. Acesso em: 28 jun. 2021.

HALGASWATHA, T. *et al.* Neural network based phylogenetic analysis. **2012 International Conference on Biomedical Engineering, ICoBE 2012**, [s. l.], p. 155–160, 2012. Available at: <https://doi.org/10.1109/ICoBE.2012.6178974>

HAMILTON, T. L. *et al.* Transcriptional profiling of nitrogen fixation in *Azotobacter vinelandii*. **Journal of Bacteriology**, [s. l.], v. 193, n. 17, p. 4477–4486, 2011. Available at: <https://doi.org/10.1128/JB.05099-11>

HAUBOLD, B. Alignment-free phylogenetics and population genetics. **Briefings in Bioinformatics**, [s. l.], v. 15, n. 3, p. 407–418, 2014. Available at: <https://doi.org/10.1093/BIB/BBT083>. Acesso em: 15 fev. 2022.

HERRERO, A.; STAVANS, J.; FLORES, E. **The multicellular nature of filamentous heterocyst-forming cyanobacteria**. [S. l.]: Oxford University Press, 2016. Available at: <https://doi.org/10.1093/femsre/fuw029>. Acesso em: 12 jun. 2021.

HOFFMAN, B. M. *et al.* Mechanism of nitrogen fixation by nitrogenase: The next stage. **Chemical Reviews**, [s. l.], v. 114, n. 8, p. 4041–4062, 2014. Available at: <https://doi.org/10.1021/cr400641x>

HOTELLING, H. Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology**, [s. l.], v. 24, n. 6, p. 417–441, 1933. Available at: <https://doi.org/10.1037/H0071325>

HU, Y. H. ; HWANG, J. **Handbook of Neural Network Signal Processing**. 1st. ed. New York: CRC Press, 2002.

HUSSON, F.; JOSSE, J.; PAGÈS, J. **Principal Component Methods Hierarchical Clustering Partitional Clustering Why Would We Need to Choose for Visualizing Data**. [S. l.], 2010. Available at:

<https://www.scirp.org/%28S%28351jmbntvnsjt1aadkposzje%29%29/reference/referencenpapers.aspx?referenceid=2950186>. Acesso em: 12 mar. 2022.

INOUE, J. *et al.* Distribution and Evolution of Nitrogen Fixation Genes in the Phylum &i&gt;Bacteroidetes&lt;/i&gt; **Microbes and environments**, [s. l.], v. 30, n. 1, p. 44–50, 2015. Available at: <https://doi.org/10.1264/jsme2.ME14142>

JACOBSON, M. R. *et al.* Physical and genetic map of the major nif gene cluster from *Azotobacter vinelandii*. **Journal of bacteriology**, [s. l.], v. 171, n. 2, p. 1017–1027, 1989.

JAMES, G. *et al.* **An Introduction to Statistical Learning: with Applications in R**. New York: Springer New York, 2013. *E-book*.

JOHNSON, W. B.; LINDENSTRAUSS, J.; SCHECHTMAN, G. Extensions of lipschitz maps into Banach spaces. **Israel Journal of Mathematics** 1986 54:2, [s. l.], v. 54, n. 2, p. 129–138, 1986. Available at: <https://doi.org/10.1007/BF02764938>. Acesso em: 16 mar. 2022.

JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, [s. l.], v. 374, n. 2065, 2016. Available at: <https://doi.org/10.1098/RSTA.2015.0202>. Acesso em: 2 ago. 2021.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. **Science**, [s. l.], v. 349, n. 6245, p. 255–260, 2015. Available at: <https://doi.org/10.1126/SCIENCE.AAA8415>. Acesso em: 3 mar. 2022.

KASSAMBARA, A. **Principal Component Methods in R: Practical Guide**. 1st. ed. [S. l.]: sthda.com, 2017.

KIRN, J.; REES, D. C. Crystallographic structure and functional implications of the nitrogenase molybdenum-iron protein from *Azotobacter vinelandii*. **Nature**, [s. l.], v. 360, n. 6404, p. 553–560, 1992. Available at: <https://doi.org/10.1038/360553a0>. Acesso em: 17 jun. 2021.

KOIRALA, A.; BRÖZEL, V. S. Phylogeny of Nitrogenase Structural and Assembly Components Reveals New Insights into the Origin and Distribution of Nitrogen Fixation across Bacteria and Archaea. **Microorganisms**, [s. l.], v. 9, n. 8, p. 1662, 2021. Available at: <https://doi.org/10.3390/microorganisms9081662>. Acesso em: 23 ago. 2021.

KUMAR, A. **Hold-out Method for Training Machine Learning Models - Data Analytics**. [S. l.], 2021. Available at: <https://vitalflux.com/hold-out-method-for-training->

machine-learning-model/. Acesso em: 3 mar. 2022.

LEMEY, P.; SALEMI, M.; VANDAMME, A.-M. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. **The Phylogenetic Handbook**, [s. l.], 2009. Available at: <https://doi.org/10.1017/CBO9780511819049>. Acesso em: 15 fev. 2022.

LI, X. *et al.* The genome of *Paenibacillus sabinæ* T27 provides insight into evolution, organization and functional elucidation of *nif* and *nif*-like genes. **BMC genomics**, [s. l.], v. 15, n. 2, p. 723, 2014. Available at: <https://doi.org/10.1186/1471-2164-15-723>

LIU, D.; HUNT, M.; TSAI, I. J. Inferring synteny between genome assemblies: A systematic evaluation. **BMC Bioinformatics**, [s. l.], v. 19, n. 1, p. 1–13, 2018. Available at: <https://doi.org/10.1186/S12859-018-2026-4/FIGURES/9>. Acesso em: 21 fev. 2022.

LYONS, T. W.; REINHARD, C. T.; PLANAVSKY, N. J. The rise of oxygen in Earth's early ocean and atmosphere. **Nature** **2014** **506:7488**, [s. l.], v. 506, n. 7488, p. 307–315, 2014. Available at: <https://doi.org/10.1038/nature13068>. Acesso em: 30 jul. 2021.

MARTINEZ-ARGUDO, I. *et al.* **The NifL-NifA System: A Multidomain Transcriptional Regulatory Complex That Integrates Environmental Signals**. [S. l.]: J Bacteriol, 2004. Available at: <https://doi.org/10.1128/JB.186.3.601-610.2004>. Acesso em: 14 jun. 2021.

MEHER, P. K. *et al.* nifPred: Proteome-wide identification and categorization of nitrogen-fixation proteins of diazotrophs based on composition-transition-distribution features using support vector machine. **Frontiers in Microbiology**, [s. l.], v. 9, n. MAY, 2018. Available at: <https://doi.org/10.3389/fmicb.2018.01100>. Acesso em: 1 jun. 2021.

MNIH, V. *et al.* Human-level control through deep reinforcement learning. **Nature** **2015** **518:7540**, [s. l.], v. 518, n. 7540, p. 529–533, 2015. Available at: <https://doi.org/10.1038/nature14236>. Acesso em: 14 abr. 2022.

MUS, F. *et al.* Exploring the alternatives of biological nitrogen fixation. **Metallomics**, [s. l.], v. 10, n. 4, p. 523–538, 2018. Available at: <https://doi.org/10.1039/c8mt00038g>. Acesso em: 6 dez. 2019.

MUS, F. *et al.* Geobiological feedbacks, oxygen, and the evolution of nitrogenase. **Free Radical Biology and Medicine**, [s. l.], v. 140, n. October 2018, p. 250–259, 2019. Available at: <https://doi.org/10.1016/j.freeradbiomed.2019.01.050>.

Acesso em: 9 jun. 2021.

NISHIHARA, A. *et al.* Nitrogenase Activity in Thermophilic Chemolithoautotrophic Bacteria in the Phylum *Aquificae* Isolated under Nitrogen-Fixing Conditions from Nakabusa Hot Springs. **Microbes and Environments**, [s. l.], v. 33, n. 4, p. 394–401, 2018. Available at: <https://doi.org/10.1264/jsme2.me18041>. Acesso em: 6 jun. 2021.

NOUWEN, N. *et al.* The role of rhizobial (NifV) and plant (FEN1) homocitrate synthases in Aeschynomene/photosynthetic Bradyrhizobium symbiosis. **Scientific Reports**, [s. l.], v. 7, n. 1, p. 1–10, 2017. Available at: <https://doi.org/10.1038/s41598-017-00559-0>. Acesso em: 6 jun. 2021.

OBERTO, J. SyntTax: A web server linking synteny to prokaryotic taxonomy. **BMC Bioinformatics**, [s. l.], v. 14, n. 1, p. 1, 2013. Available at: <https://doi.org/10.1186/1471-2105-14-4>

OVCHARENKO, I. *et al.* Evolution and functional classification of vertebrate gene deserts. **Genome Research**, [s. l.], v. 15, n. 1, p. 137–145, 2005. Available at: <https://doi.org/10.1101/GR.3015505>. Acesso em: 21 fev. 2022.

PANG, K. *et al.* Nitrogen-Fixing Heterocystous Cyanobacteria in the Tonian Period. **Current Biology**, [s. l.], v. 28, n. 4, p. 616-622.e1, 2018. Available at: <https://doi.org/10.1016/j.cub.2018.01.008>

PASSARGE, E.; HORSTHEMKE, B.; FARBER, R. A. Incorrect use of the term synteny. **Nature Genetics** 1999 23:4, [s. l.], v. 23, n. 4, p. 387–387, 1999. Available at: <https://doi.org/10.1038/70486>. Acesso em: 21 fev. 2022.

PAUL, S. **K-Means Clustering with scikit-learn**. [S. l.], 2018. Available at: [https://www.datacamp.com/community/tutorials/k-means-clustering-python?utm\\_source=adwords\\_ppc&utm\\_medium=cpc&utm\\_campaignid=1455363063&utm\\_adgroupid=65083631748&utm\\_device=c&utm\\_keyword=&utm\\_matchtype=&utm\\_network=g&utm\\_adpostion=&utm\\_creative=33260203436](https://www.datacamp.com/community/tutorials/k-means-clustering-python?utm_source=adwords_ppc&utm_medium=cpc&utm_campaignid=1455363063&utm_adgroupid=65083631748&utm_device=c&utm_keyword=&utm_matchtype=&utm_network=g&utm_adpostion=&utm_creative=33260203436). Acesso em: 3 mar. 2022.

PAULITSCH, F. *et al.* Phylogeny of symbiotic genes reveals symbiovars within legume-nodulating Paraburkholderia species. **Systematic and Applied Microbiology**, [s. l.], v. 43, n. 6, p. 126151, 2020. Available at: <https://doi.org/10.1016/j.syapm.2020.126151>

PERICO, C. P. *et al.* Genomic landscape of SARS-CoV-2 pandemic in Brazil suggests an external P.1 variant origin. **medRxiv**, [s. l.], p. 2021.11.10.21266084,

2021. Available at: <https://doi.org/10.1101/2021.11.10.21266084>. Acesso em: 19 fev. 2022.

PETERS, J. W. *et al.* Redox-dependent structural changes in the nitrogenase P-cluster. **Biochemistry**, [s. l.], v. 36, n. 6, p. 1181–1187, 1997. Available at: <https://doi.org/10.1021/bi9626665>. Acesso em: 17 jun. 2021.

POSTGATE, J. R. **The fundamental of nitrogen fixation**. Cambridge: Cambridge University Press, 1983.

POUDEL, S. *et al.* Origin and Evolution of Flavin-Based Electron Bifurcating Enzymes. **Frontiers in Microbiology**, [s. l.], v. 0, n. AUG, p. 1762, 2018. Available at: <https://doi.org/10.3389/FMICB.2018.01762>

PROCTER, J. B. *et al.* Visualization of multiple alignments, phylogenies and gene family evolution. **Nature Methods**, [s. l.], v. 7, n. 3, p. S25, 2010. Available at: <https://doi.org/10.1038/nmeth.1434>

RAITZ, R. T. *et al.* Comparative Genomics Provides Insights into the Taxonomy of Azoarcus and Reveals Separate Origins of Nif Genes in the Proposed Azoarcus and Aromatoleum Genera. [s. l.], 2021. Available at: <https://doi.org/10.3390/genes12010071>. Acesso em: 4 jun. 2021.

RANNALA, B.; YANG, Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. **J. Mol. Evol.**, [s. l.], v. 43, n. 3, p. 304–311, 1996. Available at: <https://doi.org/10.1007/bf02338839>

RAYMOND, J. *et al.* The Natural History of Nitrogen Fixation. **Molecular Biology and Evolution**, [s. l.], v. 21, n. 3, p. 541–554, 2004. Available at: <https://doi.org/10.1093/molbev/msh047>

RICHARD FORTEY. Systematics: Meaning, Branches and Its Application. **Biology Discussion**, [s. l.], 2016. Available at: <http://www.biologydiscussion.com/animals-2/systematics-meaning-branches-and-its-application/32374>. Acesso em: 15 fev. 2022.

ROMÁN-PONCE, B. *et al.* Rhizobium acidisoli sp. nov., isolated from root nodules of Phaseolus vulgaris in acid soils. **International Journal of Systematic and Evolutionary Microbiology**, [s. l.], v. 66, n. 1, p. 398–406, 2016. Available at: <https://doi.org/10.1099/ijsem.0.000732>. Acesso em: 17 jun. 2021.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. **Journal of Computational and Applied Mathematics**, [s. l.], v. 20, n. C, p. 53–65, 1987. Available at: <https://doi.org/10.1016/0377->

0427(87)90125-7

RUSSELL, S. J.; NORVIG, P. **Artificial intelligence, 3rd ed.** [S. l.]: Elsevier, 2013. v. 3rd

S, M.; Y, D. Principal component analysis based methods in bioinformatics studies. **Briefings in bioinformatics**, [s. l.], v. 12, n. 6, p. 714–722, 2011. Available at: <https://doi.org/10.1093/BIB/BBQ090>. Acesso em: 6 jul. 2021.

SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. **Molecular biology and evolution**, [s. l.], v. 4, n. 4, p. 406–425, 1987. Available at: <https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A040454>. Acesso em: 15 fev. 2022.

SAMADDAR, N. *et al.* Nitrogen fixation in *Asaia* sp. (Family Acetobacteraceae). **Current Microbiology**, [s. l.], v. 63, n. 2, p. 226–231, 2011. Available at: <https://doi.org/10.1007/s00284-011-9968-3>

SARKER, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. **SN Computer Science 2021 2:3**, [s. l.], v. 2, n. 3, p. 1–21, 2021. Available at: <https://doi.org/10.1007/S42979-021-00592-X>. Acesso em: 16 mar. 2022.

SAS. **SAS/STAT® 9.1 User's Guide**. [S. l.: s. n.], 2004.

SCHMIDT, A.; FORNE, I.; IMHOF, A. Bioinformatic analysis of proteomics data. **BMC systems biology**, [s. l.], v. 8, n. 2, p. S3, 2014. Available at: <https://doi.org/10.1186/1752-0509-8-S2-S3/FIGURES/1>. Acesso em: 21 fev. 2022.

SCHREINER, M. **Modelo computacional para comparação de táxons: Filogenia global e mineração de dados na família formicidae**. 82 f. 2021. [s. l.], 2021.

SCHUH, R. T.; BROWER, A. V. Z. (Andrew V. Z. Biological systematics : principles and applications. [s. l.], p. 311, 2009.

SIMON, P. **Too big to ignore : the business case for big data**. [S. l.: s. n.], 2015.

SISÓ, P.; DOLORS, M. **Bioinformatic tools for Big Data in Omic studies with application to genomic inversion calling and multiomic data integration**. [S. l.], 2020. Available at: <http://dspace.uvic.cat/handle/10854/6415>. Acesso em: 21 fev. 2022.

SLY, L. I.; STACKEBRANDT, E. Description of *Skermanella parooensis* gen. nov., sp. nov. to accommodate *Conglomeromonas largomobilis* subsp. *parooensis*

following the transfer of *Conglomeromonas largomobilis* subsp. *largomobilis* to the genus *Azospirillum*. **International Journal of Systematic Bacteriology**, [s. l.], v. 49, n. 2, p. 541–544, 1999. Available at: <https://doi.org/10.1099/00207713-49-2-541/CITE/REFWORKS>. Acesso em: 7 mar. 2022.

SOBER, E. The Contest Between Parsimony and Likelihood. **Systematic Biology**, [s. l.], v. 53, n. 4, p. 644–653, 2004. Available at: <https://doi.org/10.1080/10635150490468657>. Acesso em: 15 fev. 2022.

SPATZAL, T. *et al.* **Evidence for interstitial carbon in nitrogenase FeMo cofactor**. [S. l.]: American Association for the Advancement of Science, 2011. Available at: <https://doi.org/10.1126/science.1214025>. Acesso em: 17 jun. 2021.

STEIN, N. Synteny (Syntenic Genes). **Brenner's Encyclopedia of Genetics: Second Edition**, [s. l.], p. 623–626, 2013. Available at: <https://doi.org/10.1016/B978-0-12-374984-0.01508-4>

SYAKUR, M. A. *et al.* Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. **IOP Conference Series: Materials Science and Engineering**, [s. l.], v. 336, n. 1, p. 012017, 2018. Available at: <https://doi.org/10.1088/1757-899X/336/1/012017>. Acesso em: 11 fev. 2022.

THORNDIKE, R. L. Who belongs in the family? **Psychometrika**, [s. l.], v. 18, n. 4, p. 267–276, 1953. Available at: <https://doi.org/10.1007/BF02289263>

TIBSHIRANI, R.; WALTHER, G.; HASTIE, T. Estimating the number of clusters in a data set via the gap statistic. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, [s. l.], v. 63, n. 2, p. 411–423, 2001. Available at: <https://doi.org/10.1111/1467-9868.00293>. Acesso em: 12 fev. 2022.

TODOROV, H.; FOURNIER, D.; GERBER, S. Principal components analysis: theory and application to gene expression data analysis. **Genomics and Computational Biology**, [s. l.], v. 4, n. 2, p. e100041–e100041, 2018. Available at: <https://doi.org/10.18547/GCB.2018.VOL4.ISS2.E100041>. Acesso em: 2 ago. 2021.

VALAFAR, F. Neural Network Applications in Biological Sequencing. **Proceedings of the 2003 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences**, [s. l.], p. 24–27, 2003.

VAN DER MAATEN, L. J. P.; POSTMA, E. O.; VAN DEN HERIK, H. J. Dimensionality Reduction: A Comparative Review. **Journal of Machine Learning**



**Research**, [s. l.], v. 10, p. 1–41, 2009. Available at: <https://doi.org/10.1080/13506280444000102>

VIALLE, R. A. *et al.* RAFTS3: Rapid Alignment-Free Tool for Sequence Similarity Search. **bioRxiv**, [s. l.], p. 055269, 2016. Available at: <https://doi.org/10.1101/055269>. Acesso em: 8 jun. 2021.

VIEIRA, A. Z.; RAITTZ, R. T.; FAORO, H. Origin and evolution of nonulosonic acid synthases and their relationship with bacterial pathogenicity revealed by a large-scale phylogenetic analysis. **Microbial Genomics**, [s. l.], v. 7, n. 4, p. 000563, 2021. Available at: <https://doi.org/10.1099/MGEN.0.000563/CITE/REFWORKS>. Acesso em: 30 mar. 2022.

WANG, L. *et al.* A Minimal Nitrogen Fixation Gene Cluster from *Paenibacillus* sp. WLY78 Enables Expression of Active Nitrogenase in *Escherichia coli*. **PLoS Genetics**, [s. l.], v. 9, n. 10, p. 1–11, 2013. Available at: <https://doi.org/10.1371/journal.pgen.1003865>

WANG, Y. *et al.* Genomic evidence of the illumination response mechanism and evolutionary history of magnetotactic bacteria within the Rhodospirillaceae family. **BMC Genomics**, [s. l.], v. 20, n. 1, p. 1–13, 2019. Available at: <https://doi.org/10.1186/S12864-019-5751-9/FIGURES/6>. Acesso em: 26 nov. 2021.

WARNOW, T. Supertree Construction: Opportunities and Challenges. **arxiv**, [s. l.], 2018. Available at: <https://arxiv.org/abs/1805.03530v1>. Acesso em: 15 fev. 2022.

WATANABE, T.; HORIIKE, T. The evolution of molybdenum dependent nitrogenase in cyanobacteria. **Biology**, [s. l.], v. 10, n. 4, p. 1–9, 2021. Available at: <https://doi.org/10.3390/biology10040329>

WEISS, M.; GÖKER, M. Molecular Phylogenetic Reconstruction. **The Yeasts**, [s. l.], v. 1, p. 159–174, 2011. Available at: <https://doi.org/10.1016/B978-0-444-52149-1.00012-4>

WEON, H. Y. *et al.* *Skermanella aerolata* sp. nov., isolated from air, and emended description of the genus *Skermanella*. **International Journal of Systematic and Evolutionary Microbiology**, [s. l.], v. 57, n. 7, p. 1539–1542, 2007. Available at: <https://doi.org/10.1099/IJS.0.64676-0/CITE/REFWORKS>. Acesso em: 7 mar. 2022.

YANG, L. L. *et al.* Plasmids Related to the Symbiotic Nitrogen Fixation Are Not Only Cooperated Functionally but Also May Have Evolved over a Time Span in Family Rhizobiaceae. **Genome Biology and Evolution**, [s. l.], v. 12, n. 11, p. 2002–2014, 2020. Available at: <https://doi.org/10.1093/GBE/EVAA152>. Acesso em: 16 mar. 2022.

ZANI, S. *et al.* Expression of *nifH* genes in natural microbial assemblages in Lake George, New York, detected by reverse transcriptase PCR. **Applied and Environmental Microbiology**, [s. l.], v. 66, n. 7, p. 3119–3124, 2000. Available at: <https://doi.org/10.1128/AEM.66.7.3119-3124.2000>. Acesso em: 17 jun. 2021.

ZERKLE, A. L.; MIKHAIL, S. The geobiological nitrogen cycle: From microbes to the mantle. **Geobiology**, [s. l.], v. 15, n. 3, p. 343–352, 2017. Available at: <https://doi.org/10.1111/GBI.12228>. Acesso em: 30 jul. 2021.

## APÊNDICE

### 1. CONFIGURAÇÕES DE MÁQUINA

Base dos proteomas totais obtidos pelo NCBI foi manipulada e convertida em vetores SWEEP (parâmetros descritos em De Pierri et al 2020) com o uso do Servidor localizado com sede no PPG em Bioinformática da UFPR com as seguintes configurações:

TABELA 1 – Configurações do Servidor do PPG em Bioinformática da UFPR.

CPU	Xeon® 40 <i>threads</i>
RAM	251 Gigabytes
Sistema Operacional	CentOS versão 7

FONTE: O Autor (2022).

### 2. RESULTADOS OBTIDOS COM A RNA

#### 2.1 NA PREDIÇÃO DE ESPÉCIES FIXADORAS

Em nosso *script* foram inclusos, além dos 22 Nifs descritos, as sequências: NifEN, NifNB, fdxN, fdxB, fdx2, NifX-associated *protein* e *Mo-nitrogenase C-terminal*, expandindo a varredura para 29 sequências pelo algoritmo.

FIGURA 1 – SCRIPT CATANIFs DESENVOLVIDO EM MATLAB PARA A BUSCA DE NIFs EM GENOMAS CANDIDATOS

```

A.
1 function [list2 list1 xhits] = catanif(file1, varargin)
2 % File 1 é um genoma possível -
3 % file1 - multifasta org (aa) -
4 % file2 - multifasta base para RFT33 (fasta nifs)
5 % List2 - nifs encontrados
6 % List1 - Genes correspondentes organismo avaliado
7 % xhits - [a b c] -> a) índice org; b) índice genes fastanifs c) score
8 %%%
9 V = varargin;
10 if isempty(file1,'Sequence')
11     bicho = file1;
12 else
13     bicho = fastaread(file1);
14 end
15 %
16 BDnif = varargin{V,'varfas',fastaread('fixnif151.fa')};
17 BD = formatch(BDnif);
18 nm = length(bicho);
19 xhits = [];
20
21 for ii=1:nn
22     q= rafts3(bicho(ii),BD);
23     v = q.scores;
24     if numel(v)<2
25         continue
26     end
27     u= q.scores(:,1:2);
28     u= u(u(:,2)>0.2,:);
29     disp(num2str(u))
30     if numel(u)>2
31         xhits= [xhits; ii u];
32         disp(num2str([ii u]));
33     end
34 end
35 %bicho(xhits(:,2)).Header;
36 list1 = char(bicho(xhits(:,1)).Header)
37 list2 = char(BDnif(xhits(:,2)).Header)
38
B.
1 fst = RSMgr.fst;
2 GR = RSMgr.GR;
3 BD = RSMgr.BD;
4 LST = RSMgr.LST;
5 %
6 KO = 1;
7 iNet02 = allGoods02; % (1:22);
8 iGoods = [];
9 M = zeros(length(iNet02),29);
10 for K=KO:length(iNet02)
11     KK = iNet02(K);
12     nSZ(KK)
13     Fas = fastaread(['fastas\'+D(KK).name]);
14     [a, b, c] = catanifs(Fas,'BD',BD,'varfas',fst);
15     [t p] = unifysameinds(LST(c,:),c(:),3,@max);
16     M(K,p) = t;
17     disp(RSMgr.nifname(p));
18     nFort = sum(c(:),3)>0.5);
19     try
20         hash = sum(c(:),3)>0.45 & ismember(GR.igrp(c(:),2),GR.contg2{[72,1],1});
21         disp(['Força/H: ' num2str(inFort hash K)] '← iter');
22     catch
23         hash = 0;
24         nFort = 0;
25     end
26     if hash || nFort > 5
27         %break
28         iGoods = [iGoods KK];
29         save iGoods.dat iGoods -ascii
30     end
31 end
32

```

FONTE: O autor (2022).

LEGENDA: *Script*catanifs (A) e *catanifs* (B) para busca em proteomas selecionados como candidatos a fixadores a fixadores (*allGoods02*). pelas 22 classes de Nifs obtidos por RAFTS<sup>3</sup>G (*rafts3group*) em 118 proteomas obtidos por literatura. Por fim, o *script*catanifs salva os proteomas com, ao menos, o conjunto *core* de 5 sequências Nif (*iGoods*)

FIGURA 2 – ESTRUTURA DE DADOS (RSMGR) OBTIDOS POR RAFTS<sup>3</sup>G PARA BUSCA NOS CANDIDATOS PELA RNA

Field	Value
GR	1x1 struct
fst	2156x1 struct
D	10324x1 struct
nifname	29x1 cell
nifclust	29x1 cell
organismos	1x1 struct
LST	2156x1 double
BD	1x1 struct
M	711x29 double

FONTE: O Autor (2022)

LEGENDA: Dos 10.324 proteomas totais (D) extraímos 2156 sequencias Nif (RSMgr.fst), Base criada por RAFTS3 pela função *formatbd* (RSMgr.BD). Grupos contendo Nifs homólogos por RAFTS<sup>3</sup>G 0.5 similaridade (RSMgr.GR), Nome dos 29\* Nifs (RSMgr.nifname) relacionados ao número do clusters homólogos Nifs (RSMgr.nifclust), Matriz com o resultado da busca dos 29 Nifs nos 711 organismos candidatos (M). Visualização das informações no MATLAB®.

### 3. RESULTADOS OBTIDOS POR RAFTS<sup>3</sup>G E RAFTS3

O RESULTADO DE RAFTS<sup>3</sup>G PARA OS 662 PROTEOMAS PODE SER ACESSADO NO GDRIVE ATRAVÉS DO LINK DE ACESSO:  
[HTTPS://DRIVE.GOOGLE.COM/DRIVE/FOLDERS/14FJOO9TDDJPAGNqIUv3KF65AKJTU70?USP=SHARING](https://drive.google.com/drive/folders/14fjUo9tDDJPgAGNqIUv3KF65AKJTU70?usp=sharing)

TABELA 2 - LISTA DOS CLUSTERS OBTIDOS POR RAFTS<sup>3</sup>G PARA COMPOR A BASE NIFS UTILIZADA NA BUSCA POR RAFTS<sup>3</sup>.

<b>Nif</b>	<b>Nº Clusters</b>	<b>Cluster ID RAFTS<sup>3</sup>G</b>
NifH	3	26,68,582
NifD	16	24,70,260,314,322,328,342,344,352,534,574,584,586,596,610,806
NifK	10	22,262,316,324,326,348,350,422,572,594
NifE	14	20,198,208,240,264,338,340,370,474,560,584,590,592,612
NifN	25	18,42,98,266,300,310,366,372,384,404,432,460,474,494,536,568,590,600,604,614,684,694,750,764,832
NifX	42	16,44,58,60,96,102,126,168,254,280,284,354,374,376,378,398,414,416,434,458,470,472,478,486,496,500,542,552,578,588,598,606,656,682,718,726,762,768,810,812,814,834
NifB	26	30,104,268,278,298,306,308,332,334,346,362,382,386,392,394,406,448,480,508,510,608,686,708,768,820,850
NifO	1	176
NifQ	31	12,64,114,138,144,178,194,200,210,222,236,420,450,488,506,516,540,618,630,632,636,644,664,746,758,774,778,790,808,846,854
NifU	33	10,66,108,132,180,202,218,232,250,252,312,320,360,410,436,466,476,518,562,634,676,716,728,730,734,736,744,756,770,776,788,800,836
NifS	16	8,36,106,212,214,234,238,318,330,452,464,484,530,640,724,794
NifV	24	6,242,246,282,292,302,304,336,356,388,438,462,512,528,554,628,678,714,720,742,754,772,802,838
NifZ	21	32,34,54,90,122,188,226,230,358,442,454,468,502,526,544,556,688,692,722,738,842
NifW	19	2,56,76,92,140,440,456,492,504,524,538,626,740,752,780,786,804,822,840
NifT	5	38,88,116,428,522
NifY	4	118,156,430,830
NifM	11	52,120,164,190,286,408,444,648,792,818,844
NifJ	3	80,426,646
NifF	2	136,166
NifP	5	4,82,256,402,782
NifA	17	28,48,86,112,128,154,248,258,390,400,514,532,548,620,660,766,784
NifL	8	110,172,184,228,446,546,642,852
NifR	9	0,40,78,130,206,216,270,482,624
Nif.	10	14, 46, 74, 94, 142, 182, 424, 558, 672, 680
MonC	2	72, 668

FONTE: O Autor (2022).

TABELA 3 – 711 ORGANISMOS CANDIDATOS VERSUS SEQUÊNCIAS NIF CORE IDENTIFICADAS POR RAFTS3.

INDEX	NCBI	ORGANISMO	NifH	nifD	NifK	NifE	NifN	NifB
517	GCA_000018105.1_ASM1810v1	<i>Acaryochloris marina</i> MBIC11017	0	0,2	0	0	0	0
1714	GCA_000247605.1_ASM24760v1	<i>Acetobacterium woodii</i> DSM 1030	0,8	0,64	0,58	0,66	0	0
1483	GCA_000214095.3_ASM21409v3	<i>Acidithiobacillus ferrivorans</i> SS3	1	1	1	1	1	1
673	GCA_000021485.1_ASM2148v1	<i>Acidithiobacillus ferrooxidans</i> ATCC 23270	1	1	1	1	1	1
643	GCA_000020825.1_ASM2082v1	<i>Acidithiobacillus ferrooxidans</i> ATCC 53993	1	1	1	1	1	0,96
977	GCA_000092025.1_ASM9202v1	<i>Agrobacterium fabrum</i> str C58	0	0	0	0	0	0
438	GCA_000016265.1_ASM1626v1	<i>Agrobacterium radiobacter</i> K84	0	0	0	0	0	0
6750	GCA_002005205.2_ASM200520v2	<i>Agrobacterium rhizogenes</i>	0	0	0	0	0	0
1316	GCA_000192635.1_ASM19263v1	<i>Agrobacterium</i> sp H133	0	0	0	0	0	0
5661	GCA_001713475.1_ASM171347v1	<i>Agrobacterium</i> sp RAC06	0	0	0	0	0	0
3754	GCA_000971565.1_ASM97156v1	<i>Agrobacterium tumefaciens</i>	0	0	0	0	0	0
4801	GCA_001551895.1_ASM155189v1	<i>Agrobacterium tumefaciens</i>	0	0	0	0	0	0
8845	GCA_002943835.1_ASM294383v1	<i>Agrobacterium tumefaciens</i>	0	0	0	0	0	0
439	GCA_000016285.1_ASM1628v1	<i>Agrobacterium vitis</i> S4	0	0	0	0	0	0
471	GCA_000016985.1_ASM1698v1	<i>Alkaliphilus metalliredigens</i> QYMF	0,9	0,91	0,92	0,85	0,32	0,81
847	GCA_000025485.1_ASM2548v1	<i>Allochromatium vinosum</i> DSM 180	1	1	1	1	1	1
5008	GCA_001605015.1_ASM160501v1	<i>Aminobacter aminovorans</i>	0	0	0	0	0	0
9303	GCA_003071665.1_ASM307166v1	<i>Aminobacter</i> sp MSH1	0	0	0	0	0,24	0
2126	GCA_000317695.1_ASM31769v1	<i>Anabaena cylindrica</i> PCC 7122	0,9	1	1	1	1	1
2096	GCA_000312705.1_ASM31270v1	<i>Anabaena</i> sp 90	0,9	0,97	0,95	0,95	0,77	0,94
4220	GCA_001277295.1_ASM127729v1	<i>Anabaena</i> sp WA102	0,9	0,96	0,95	0,95	0,86	0,94
642	GCA_000020805.1_ASM2080v1	<i>Anaeromyxobacter</i> sp K	0,9	0,84	0,82	0	0,77	0,52
6739	GCA_002002905.1_ASM200290v1	<i>Aquaspirillum</i> sp LM1	0,9	0,87	0,8	0,77	0,6	0,72
984	GCA_000092245.1_ASM9224v1	<i>Arcobacter nitrofigilis</i> DSM 7299	1	1	1	1	1	1
1940	GCA_000284235.1_ASM28423v1	<i>Arcobacter</i> sp L	1	0,94	0,91	0,92	0,86	0,85

4771	GCA_001547995.1_ASM154799v1	<i>Asaia bogorensis</i> NBRC 16594	0	0	0	0	0	0	0	0	0	0
4428	GCA_001442755.1_ASM144275v1	<i>Aureimonas</i> sp AU20	0	0	0	0	0	0	0	0	0	0
9383	GCA_003111645.1_ASM311164v1	<i>Azoarcus communis</i>	0,9	0,94	0,9	0,9	0,89	0,89	0,9	0,89	0,9	0,9
9384	GCA_003111665.1_ASM311166v1	<i>Azoarcus communis</i>	0,9	0,94	0,9	0,89	0,89	0,89	0,9	0,89	0,9	0,9
5422	GCA_001682385.1_ASM168238v1	<i>Azoarcus olearius</i>	1	1	1	1	1	1	1	1	1	1
930	GCA_000061505.1_ASM6150v1	<i>Azoarcus</i> sp BH72	1	1	1	1	1	1	1	1	1	1
4137	GCA_001190925.1_ASM119092v1	<i>Azoarcus</i> sp CIB	0,9	0,83	0,68	0,72	0,61	0,61	0,58	0,61	0,58	0,58
2261	GCA_000349945.1_ASM34994v1	<i>Azoarcus</i> sp KH32C	0,9	0,9	0,81	0,86	0,69	0,69	0,86	0,69	0,86	0,86
178	GCA_000010525.1_ASM1052v1	<i>Azorhizobium caulinodans</i> ORS 571	1	1	1	1	1	1	1	1	1	1
2821	GCA_000632475.2_ASM63247v2	<i>Azospirillum brasilense</i>	1	1	0,99	1	0,96	0,99	0,99	0,96	0,99	0,99
4362	GCA_001315015.1_ASM131501v1	<i>Azospirillum brasilense</i>	1	1	1	1	1	1	1	1	1	1
1915	GCA_000283655.1_ASM28365v1	<i>Azospirillum lipoferum</i> 4B	1	1	1	1	1	1	1	1	1	0,97
188	GCA_000010725.1_ASM1072v1	<i>Azospirillum</i> sp B510	1	1	1	1	1	1	1	1	1	1
9558	GCA_003233655.1_ASM323365v1	<i>Azospirillum</i> sp M2T2B2	1	1	0,99	0,98	0,96	0,96	0,94	0,96	0,94	0,94
9396	GCA_003119115.1_ASM311911v1	<i>Azospirillum</i> sp TSH58	1	1	1	0,99	0,98	0,98	0,99	0,98	0,99	0,99
4316	GCA_001305595.1_ASM130559v1	<i>Azospirillum thiophilum</i>	1	1	1	1	1	1	1	1	1	1
9406	GCA_003130795.1_ASM313079v1	<i>Azospirillum thiophilum</i>	1	0,97	0,95	0,94	0,86	0,86	0,87	0,86	0,87	0,87
7565	GCA_002220155.1_ASM222015v1	<i>Azotobacter chroococcum</i>	1	1	1	1	1	1	1	1	1	1
3370	GCA_000817975.1_ASM81797v1	<i>Azotobacter chroococcum</i> NCIMB 8003	1	1	1	1	1	1	1	1	1	1
2289	GCA_000380335.1_ASM38033v1	<i>Azotobacter vinelandii</i> CA	1	1	1	1	1	1	1	1	1	1
2290	GCA_000380365.1_ASM38036v1	<i>Azotobacter vinelandii</i> CA6	1	1	1	1	1	1	1	1	1	1
652	GCA_000021045.1_ASM2104v1	<i>Azotobacter vinelandii</i> DJ	1	1	1	1	1	1	1	1	1	1
2173	GCA_000328745.1_ASM32874v1	<i>Bacillus subtilis</i> BEST7613	0	0	0	0	0	0	0	0	0	0
4315	GCA_001305575.2_ASM130557v2	<i>Beggiatoa leptomitiformis</i>	1	1	1	1	1	1	1	1	1	1
8537	GCA_002849755.1_ASM284975v1	<i>Beggiatoa leptomitiformis</i>	1	1	1	1	1	1	1	1	1	1
594	GCA_000019845.1_ASM1984v1	<i>Beijerinckia indica</i> subsp <i>indica</i> ATCC 9039	1	1	1	1	1	1	1	1	1	1
4372	GCA_001402875.1_ASM140287v1	<i>Blastochloris viridis</i>	0,9	0,8	0,79	0,72	0,64	0,64	0,66	0,64	0,66	0,66
4512	GCA_001459775.1_Blastochloris_viridis_genome	<i>Blastochloris viridis</i>	0,9	0,8	0,79	0,72	0,64	0,64	0,66	0,64	0,66	0,66
4777	GCA_001548155.2_BV133_assembly_1.0	<i>Blastochloris viridis</i>	0,9	0,8	0,79	0,72	0,64	0,64	0,66	0,64	0,66	0,66
4870	GCA_001562255.1_ASM156225v1	<i>Bosea</i> sp PAMC 26642	0	0	0	0	0	0	0	0	0	0
5660	GCA_001713455.1_ASM171345v1	<i>Bosea</i> sp RAC05	0	0	0	0	0,2	0	0	0	0	0
4797	GCA_001549695.1_ASM154969v1	<i>Bradyrhizobium diazoefficiens</i>	1	1	1	1	0,98	0,99	0,99	0,98	0,99	0,99



6222	GCA_001908315.1_ASM190831v1	<i>Bradyrhizobium diazoefficiens</i>	1	1	1	1	1	1	0,95
9465	GCA_003183845.1_ASM318384v1	<i>Bradyrhizobium diazoefficiens</i>	1	1	1	1	1	1	1
214	GCA_000011365.1_ASM1136v1	<i>Bradyrhizobium diazoefficiens USDA 110</i>	1	1	1	1	1	1	1
5211	GCA_001642675.1_ASM164267v1	<i>Bradyrhizobium diazoefficiens USDA 110</i>	1	1	1	1	1	1	0,97
5522	GCA_001693385.1_ASM169338v1	<i>Bradyrhizobium icense</i>	1	0,95	0,92	0,84	0,8	0,8	0,86
3298	GCA_000807315.1_ASM80731v1	<i>Bradyrhizobium japonicum</i>	1	1	1	1	1	1	0
6086	GCA_001887695.1_ASM188769v1	<i>Bradyrhizobium japonicum</i>	1	1	1	1	1	1	0,97
1947	GCA_000284375.1_ASM28437v1	<i>Bradyrhizobium japonicum USDA 6</i>	1	1	1	1	1	1	0,95
2235	GCA_000344805.1_ASM34480v1	<i>Bradyrhizobium oligotrophicum S58</i>	1	0,98	0,96	0,98	0,95	0,95	0,99
7741	GCA_002278135.2_ASM227813v2	<i>Bradyrhizobium ottawaense 0099</i>	1	1	1	1	1	1	1
9919	GCA_900011245.1_BF49.v1	<i>Bradyrhizobium sp</i>	0,2	0	0	0	0	0	0
7720	GCA_002266435.2_ASM226643v2	<i>Bradyrhizobium sp 2 39S1MB</i>	1	0,95	0,9	0,87	0,82	0,82	0,91
7721	GCA_002266465.2_ASM226646v2	<i>Bradyrhizobium sp 3 85S1MB</i>	0	0	0	0	0	0	0
392	GCA_000015165.1_ASM1516v1	<i>Bradyrhizobium sp BTAi1</i>	1	1	1	1	1	1	1
2000	GCA_000296215.2_ASM29621v2	<i>Bradyrhizobium sp CCGELA001</i>	1	0,93	0,88	0,78	0,79	0,79	0,78
873	GCA_000026145.1_ASM2614v1	<i>Bradyrhizobium sp ORS 278</i>	1	1	0,99	0,98	0,97	0,97	1
10015	GCA_900176205.1_BRAD285_PRJEB202261_v1	<i>Bradyrhizobium sp ORS 285</i>	1	0,99	0,99	0,98	0,96	0,96	0,99
10108	GCA_900324035.1_BRAD3257	<i>Bradyrhizobium sp ORS 3257</i>	1	0,93	0,96	0,9	0,89	0,89	0,83
1942	GCA_000284275.1_ASM28427v1	<i>Bradyrhizobium sp S23321</i>	1	0,95	0,91	0,87	0,8	0,8	0,92
8435	GCA_002831585.1_ASM283158v1	<i>Bradyrhizobium sp SK17</i>	0	0	0	0	0	0	0
4009	GCA_001039275.2_ASM103927v2	<i>Brevibacillus brevis</i>	0	0	0	0	0	0	0
160	GCA_000010165.1_ASM1016v1	<i>Brevibacillus brevis NBRC 100599</i>	0	0	0	0,2	0	0	0
7532	GCA_002215075.1_ASM221507v1	<i>Brevibacillus formosus</i>	0	0	0	0	0	0	0
464	GCA_000016845.1_ASM1684v1	<i>Brucella ovis ATCC 25840</i>	0	0	0	0	0	0	0
1197	GCA_000176935.3_ASM17693v3	<i>Burkholderia sp CCGE1001</i>	1	0,99	1	1	1	1	0,99
1014	GCA_000092885.1_ASM9288v1	<i>Burkholderia sp CCGE1002</i>	1	1	1	1	1	1	1
1800	GCA_000262695.1_ASM26269v1	<i>Burkholderia sp KJ006</i>	1	1	1	0,99	1	1	1
5702	GCA_001718695.1_ASM171869v1	<i>Burkholderia ubonensis</i>	1	1	1	0,99	1	1	1
5693	GCA_001718495.1_ASM171849v1	<i>Burkholderia vietnamiensis</i>	1	1	1	1	1	1	1
5704	GCA_001718775.1_ASM171877v1	<i>Burkholderia vietnamiensis</i>	1	1	1	0,99	1	1	1
5706	GCA_001718815.1_ASM171881v1	<i>Burkholderia vietnamiensis</i>	1	1	1	1	0,99	1	1
5711	GCA_001718915.1_ASM171891v1	<i>Burkholderia vietnamiensis</i>	1	1	1	0,99	1	1	1

9157	GCA_003019945.1_ASM301994v1	<i>Burkholderia vietnamiensis</i>	1	1	1	0,99	1	1
436	GCA_000016205.1_ASM1620v1	<i>Burkholderia vietnamiensis</i> G4	1	1	1	1	1	1
3672	GCA_000959445.1_ASM95944v1	<i>Burkholderia vietnamiensis</i> LMG 10929	1	1	1	0,99	1	1
1166	GCA_000166775.1_ASM16677v1	<i>Caldicellulosiruptor kronotskyensis</i> 2002	0,8	0,53	0,45	0	0	0,46
2961	GCA_000734895.2_ASM73489v2	<i>Calothrix</i> sp 3363	0,9	0,92	0,9	0,8	0,75	0,8
7983	GCA_002368175.1_ASM236817v1	<i>Calothrix</i> sp NIES2098	1	0,95	0,98	0,9	0,92	0,94
7987	GCA_002368395.1_ASM236839v1	<i>Calothrix</i> sp NIES3974	1	0,91	0,83	0,84	0,82	0,83
7988	GCA_002368455.1_ASM236845v1	<i>Calothrix</i> sp NIES4071	0,9	0,92	0,97	0,82	0,84	0,82
7986	GCA_002368375.1_ASM236837v1	<i>Calothrix</i> sp NIES4101	0,9	0,83	0,97	0,83	0,86	0,58
2116	GCA_000317435.1_ASM31743v1	<i>Calothrix</i> sp PCC 6303	0,9	1	1	0,97	1	0,8
2103	GCA_000316575.1_ASM31657v1	<i>Calothrix</i> sp PCC 7507	1	0,92	0,98	0,91	0,89	0,87
832	GCA_000025125.1_ASM2512v1	<i>Candidatus Atelocyanobacterium thalassa</i> isolate ALOHA	1	1	1	1	1	1
530	GCA_000018425.1_ASM1842v1	<i>Candidatus Desulforudis audaxviator</i> MP104C	0,8	0,23	0,43	0	0	0,47
8419	GCA_002813775.1_ASM281377v1	<i>Candidatus Thiodictyon syntrophicum</i>	0,9	0,88	0,85	0,82	0,59	0,68
9239	GCA_003047205.1_ASM304720v1	<i>Carboxydocella thermotrophica</i>	0,8	0,61	0,5	0,5	0	0,54
9260	GCA_003054495.1_ASM305449v1	<i>Carboxydocella thermotrophica</i>	0,8	0,62	0,5	0,5	0	0,54
348	GCA_000014245.1_ASM1424v1	<i>Chelatiorans</i> sp BNC1	0	0	0	0	0	0
6068	GCA_001887265.1_ASM188726v1	<i>Chelatococcus daeguensis</i>	0	0	0	0	0	0
4180	GCA_001271345.1_ASM127134v1	<i>Chelatococcus</i> sp CO6	0	0	0	0	0	0
5813	GCA_001747405.1_ASM174740v1	<i>Chlorobaculum limnaeum</i>	1	0,95	0,92	0,98	0,92	0,94
627	GCA_000020505.1_ASM2050v1	<i>Chlorobaculum parvum</i> NCIB 8327	1	1	1	1	1	1
268	GCA_000012585.1_ASM1258v1	<i>Chlorobium chlorochromatii</i> CaD3	1	0,9	0,84	0,89	0,77	0,77
625	GCA_000020465.1_ASM2046v1	<i>Chlorobium limicola</i> DSM 245	1	1	1	1	1	1
629	GCA_000020545.1_ASM2054v1	<i>Chlorobium phaeobacterioides</i> BS1	1	1	1	1	1	1
391	GCA_000015125.1_ASM1512v1	<i>Chlorobium phaeobacterioides</i> DSM 266	1	0,93	0,91	0	0,86	0,81
431	GCA_000016085.1_ASM1608v1	<i>Chlorobium phaeovibrioides</i> DSM 265	1	0,92	0,9	0,91	0,85	0,83
19	GCA_000006985.1_ASM698v1	<i>Chlorobium tepidum</i> TLS	1	1	1	1	1	1
628	GCA_000020525.1_ASM2052v1	<i>Chloroherpeton thalassium</i> ATCC 35110	1	1	1	1	1	0,73
2114	GCA_000317125.1_ASM31712v1	<i>Chroococcidiopsis thermalis</i> PCC 7203	1	1	1	0	1	1
5597	GCA_001701045.1_ASM170104v1	<i>Cloacibacillus porcorum</i>	0,5	0	0	0	0	0
4024	GCA_001042715.1_ASM104271v1	<i>Clostridium acetium</i>	1	1	1	1	0,32	1

96	GCA_000008765.1_ASM876v1	<i>Clostridium acetobutylicum</i> ATCC 824	1	1	1	1	1	1	0
1523	GCA_000218855.1_ASM21885v1	<i>Clostridium acetobutylicum</i> DSM 1731	1	1	1	1	1	1	0
1309	GCA_000191905.1_ASM19190v1	<i>Clostridium acetobutylicum</i> EA 2018	1	1	1	1	1	1	0
6951	<b>GCA_002074155.1_ASM207415v1</b>	<b><i>Clostridium argentinense</i></b>	0,9	0	0	0,71	0	0	0
2507	GCA_000484505.1_ASM48450v1	<i>Clostridium autoethanogenum</i> DSM 10061	0,9	0,89	0,89	0,85	0	0,81	0,81
4574	GCA_001484725.1_ASM148472v1	<i>Clostridium autoethanogenum</i> DSM 10061	0,9	0,89	0,89	0,85	0	0,81	0,81
3510	GCA_000833105.2_ASM83310v2	<i>Clostridium beijerinckii</i>	1	1	0,99	0,98	0	0,98	0,98
6745	GCA_002003345.1_ASM200334v1	<i>Clostridium beijerinckii</i>	1	1	1	0,99	0	0,99	0,99
9918	GCA_900010805.1_C._beijerinckii_DSM_6423	<i>Clostridium beijerinckii</i>	1	1	0,99	0,99	0,21	0,98	0,98
3180	GCA_000767745.1_ASM76774v1	<i>Clostridium beijerinckii</i> ATCC 35702	1	1	1	1	0	1	1
470	GCA_000016965.1_ASM1696v1	<i>Clostridium beijerinckii</i> NCIMB 8052	1	1	1	1	0	1	1
2592	GCA_000506785.3_ASM50678v3	<i>Clostridium beijerinckii</i> NRRL B598	1	1	1	1	0	0,99	0,99
2707	GCA_000577895.1_M2_40	<i>Clostridium bournimense</i>	0,9	0,89	0,9	0,85	0	0,8	0,8
3418	<b>GCA_000829015.1_ASM82901v1</b>	<b><i>Clostridium botulinum</i></b>	0,9	0	0	0	0	0	0
6016	<b>GCA_001879625.1_ASM187962v1</b>	<b><i>Clostridium botulinum</i></b>	0,9	0	0	0,7	0	0	0
6053	<b>GCA_001886775.1_ASM188677v1</b>	<b><i>Clostridium botulinum</i></b>	0,9	0	0	0,69	0	0	0
6108	<b>GCA_001889325.1_ASM188932v1</b>	<b><i>Clostridium botulinum</i></b>	0,9	0	0	0	0	0	0
8649	<b>GCA_002865765.1_ASM286576v1</b>	<b><i>Clostridium botulinum</i></b>	0,9	0	0	0,68	0	0	0
8650	<b>GCA_002865785.1_ASM286578v1</b>	<b><i>Clostridium botulinum</i></b>	0	0	0	0	0	0	0
8651	<b>GCA_002865805.1_ASM286580v1</b>	<b><i>Clostridium botulinum</i></b>	0,9	0	0	0	0	0	0
8652	<b>GCA_002865825.1_ASM286582v1</b>	<b><i>Clostridium botulinum</i></b>	0,9	0	0	0,71	0	0	0
8654	<b>GCA_002865885.1_ASM286588v1</b>	<b><i>Clostridium botulinum</i></b>	0,9	0	0	0	0	0	0
8656	<b>GCA_002866045.1_ASM286604v1</b>	<b><i>Clostridium botulinum</i></b>	0,9	0	0	0	0	0	0
9126	<b>GCA_003013615.1_ASM301361v1</b>	<b><i>Clostridium botulinum</i></b>	0,8	0	0	0,69	0	0	0
9128	<b>GCA_003013655.1_ASM301365v1</b>	<b><i>Clostridium botulinum</i></b>	0,8	0	0	0,69	0	0	0
9138	<b>GCA_003017145.1_ASM301714v1</b>	<b><i>Clostridium botulinum</i></b>	0,9	0	0	0,69	0	0	0
9139	<b>GCA_003017195.1_ASM301719v1</b>	<b><i>Clostridium botulinum</i></b>	0,9	0	0	0,69	0	0	0
9140	<b>GCA_003017225.1_ASM301722v1</b>	<b><i>Clostridium botulinum</i></b>	0,9	0	0	0	0	0	0
9141	<b>GCA_003017265.1_ASM301726v1</b>	<b><i>Clostridium botulinum</i></b>	0,9	0	0	0	0	0	0
9142	<b>GCA_003017335.1_ASM301733v1</b>	<b><i>Clostridium botulinum</i></b>	0	0	0	0	0	0	0
9269	<b>GCA_003058345.1_ASM305834v1</b>	<b><i>Clostridium botulinum</i></b>	0,9	0	0	0	0	0	0
9270	<b>GCA_003058445.1_ASM305844v1</b>	<b><i>Clostridium botulinum</i></b>	0,9	0	0	0	0	0	0

9745	GCA_0033345335.1_ASM3334533v1	<i>Clostridium botulinum</i>	0,9	0	0	0,7	0	0
3259	GCA_000789355.1_ASM78935v1	<i>Clostridium botulinum</i> 202F	0	0	0	0	0	0
473	GCA_000017025.1_ASM1702v1	<i>Clostridium botulinum</i> A str ATCC 19397	0,9	0	0	0	0	0
581	GCA_000019545.1_ASM1954v1	<i>Clostridium botulinum</i> A3 str Loch Maree	0,9	0	0	0,7	0	0
2079	GCA_000307125.1_ASM30712v1	<i>Clostridium botulinum</i> B str Eklund 17B NRP	0	0	0	0	0	0
619	GCA_000020345.1_ASM2034v1	<i>Clostridium botulinum</i> Ba4 str 657	0,9	0	0	0,7	0	0
3368	GCA_000817935.1_ASM81793v1	<i>Clostridium botulinum</i> CDC1436	0,9	0	0	0,7	0	0
3361	GCA_000816945.1_ASM81694v1	<i>Clostridium botulinum</i> CDC297	0,8	0	0	0,7	0	0
4484	GCA_001456065.2_ASM145606v2	<i>Clostridium butyricum</i>	1	1	1	1	0	1
4518	GCA_001465175.1_ASM146517v1	<i>Clostridium butyricum</i>	1	1	1	1	0	1
5232	GCA_001646605.1_ASM164660v1	<i>Clostridium butyricum</i>	1	1	1	1	0	1
5799	GCA_001746555.1_ASM174655v1	<i>Clostridium butyricum</i>	1	1	1	1	0	1
6058	GCA_001886875.1_ASM188687v1	<i>Clostridium butyricum</i>	1	1	1	0,99	0	0,99
9663	GCA_003315755.1_ASM331575v1	<i>Clostridium butyricum</i>	1	1	1	1	0	0,99
4007	GCA_001038625.1_ASM103862v1	<i>Clostridium carboxidovorans</i> P7	0,4	0	0	0	0	0
699	GCA_000022065.1_ASM2206v1	<i>Clostridium cellulolyticum</i> H10	0,5	0	0	0	0	0
3626	GCA_000953215.1_DG5	<i>Clostridium cellulosi</i>	0,8	0,58	0	0	0	0,48
1056	GCA_000145275.1_ASM14527v1	<i>Clostridium cellulovorans</i> 743B	0,9	0,77	0,67	0,7	0	0
1653	GCA_000237085.1_ASM23708v1	<i>Clostridium clariflavum</i> DSM 19732	0,8	0,72	0,75	0,68	0	0,24
9371	GCA_003096175.1_ASM309617v1	<i>Clostridium drakei</i>	0,9	0,91	0,87	0	0	0,76
6007	GCA_001877035.1_ASM187703v1	<i>Clostridium estertheticum</i> subsp estertheticum	0,9	0,69	0,66	0,75	0	0
5929	GCA_001854185.1_ASM185418v1	<i>Clostridium formicaceticum</i>	1	0,98	0,96	0,93	0,44	0,93
6997	GCA_002080475.1_ASM208047v1	<i>Clostridium formicaceticum</i>	1	0,98	0,96	0,93	0,44	0,93
6208	GCA_001902295.1_ASM190229v1	<i>Clostridium kluuyveri</i>	1	0,99	0,98	0,98	0,6	0,99
449	GCA_000016505.1_ASM1650v1	<i>Clostridium kluuyveri</i> DSM 555	1	1	1	1	1	1
165	GCA_000010265.1_ASM1026v1	<i>Clostridium kluuyveri</i> NBRC 12016	1	1	1	1	1	1
1208	GCA_000178835.2_ASM17883v2	<i>Clostridium lentocellum</i> DSM 5427	0,8	0,64	0,63	0,67	0	0
1035	GCA_000143685.1_ASM14368v1	<i>Clostridium ijundahlii</i> DSM 13528	0,9	0,9	0,88	0,85	0	0,81
5946	GCA_001856695.1_ASM185669v1	<i>Clostridium pasteurianum</i>	0,9	0,77	0,8	0,68	0	0,68
2302	GCA_000389635.1_ASM38963v1	<i>Clostridium pasteurianum</i> BC1	0,9	0,77	0,72	0,63	0	0,69
3293	GCA_000807175.1_ASM80717v1	<i>Clostridium pasteurianum</i> DSM 525 ATCC 6013	0,9	0,77	0,8	0,63	0	0,68
3295	GCA_000807255.1_ASM80725v1	<i>Clostridium pasteurianum</i> DSM 525 ATCC 6013	0,9	0,77	0,8	0,63	0	0,68

5943	GCA_001856645.1_ASM185664v1	<i>Clostridium pasteurianum</i> DSM 525 ATCC 6013	0.9	0.77	0.8	0.68	0	0.68
6743	GCA_002003305.1_ASM200330v1	<i>Clostridium saccharoperbutylacetonicum</i>	1	0.93	0.91	0.92	0	0.85
2213	GCA_000340885.1_ASM34088v1	<i>Clostridium saccharoperbutylacetonicum</i> N14HMT	1	0.94	0.94	0.94	0	0.86
3742	GCA_000968375.1_ASM96837v1	<i>Clostridium scatologenes</i>	0.9	0.91	0.86	0	0	0.76
1706	GCA_000244875.1_ASM24487v1	<i>Clostridium</i> sp BNL1100	0.9	0.73	0.65	0.71	0	0.45
3783	GCA_000973705.1_ASM97370v1	<i>Clostridium sporogenes</i>	0.8	0	0	0.69	0	0
3921	GCA_001020205.1_ASM102020v1	<i>Clostridium sporogenes</i>	0.8	0	0	0.69	0	0
5541	GCA_001697285.1_ASM169728v1	<i>Clostridium stercorarium</i> subsp <i>leptospartum</i> DSM 9219	0.8	0.58	0	0	0	0.49
2188	GCA_000331995.1_ASM33199v1	<i>Clostridium stercorarium</i> subsp <i>stercorarium</i> DSM 8532	0.8	0.55	0	0	0	0.49
2259	GCA_000349885.2_ASM34988v2	<i>Clostridium stercorarium</i> subsp <i>stercorarium</i> DSM 8532	0.8	0.58	0	0	0	0.49
5539	GCA_001697245.1_ASM169724v1	<i>Clostridium stercorarium</i> subsp <i>thermolacticum</i> DSM 2910	0.8	0.58	0	0	0	0.49
5787	GCA_001735765.2_ASM173576v2	<i>Clostridium taenioporum</i>	0	0	0	0	0	0
5210	GCA_001642655.1_ASM164265v1	<i>Clostridium tyrobutyricum</i>	1	1	1	1	0	1
5402	GCA_001679705.1_ASM167970v1	<i>Clostridium tyrobutyricum</i>	1	1	1	1	0	1
4317	GCA_001305615.1_ASM130561v1	<i>Confluentimicrobium</i> sp EMB200NS6	0.9	0.85	0.78	0	0.75	0.74
862	GCA_000025905.1_ASM2590v1	<i>Coraliomargarita akajimensis</i> DSM 45221	0.8	0.62	0.62	0.69	0.52	0
2118	GCA_000317495.1_ASM31749v1	<i>Criminalium epipsammum</i> PCC 9333	0	0	0	0	0	0
3426	GCA_000829235.1_ASM82923v1	<i>cyanobacterium endosymbiont of Epithemia turgida</i> isolate EISB Lake Yunoko	1	0.94	0.88	0.87	0.83	0.87
506	GCA_000017845.1_ASM1784v1	<i>Cyanothece</i> sp ATCC 51142	1	1	1	1	1	1
1076	GCA_000147335.1_ASM14733v1	<i>Cyanothece</i> sp PCC 7822	0.9	0.93	0.92	0.83	0.61	0.79
688	GCA_000021825.1_ASM2182v1	<i>Cyanothece</i> sp PCC 8801*	0.9	0.95	0.93	0.79	0.62	0.78
698	GCA_000022045.1_ASM2204v1	<i>Cyanothece</i> sp PCC 8802*	0.9	0.92	0.9	0.85	0.77	0.82
261	GCA_000012425.1_ASM1242v1	<i>Dechloromonas aromatica</i> RCB	0.9	0.86	0.77	0.81	0.64	0.71
5132	GCA_001620265.1_ASM162026v1	<i>Defluviimonas alba</i>	0.9	0.9	0.86	0.81	0.78	0
2626	GCA_000512895.1_ASM51289v1	<i>Dehalobacter restrictus</i> DSM 9455	0	0	0	0	0	0
2061	GCA_000305815.1_ASM30581v1	<i>Dehalobacter</i> sp CF	0.8	0.81	0.6	0.76	0.38	0.59
2059	GCA_000305775.1_ASM30577v1	<i>Dehalobacter</i> sp DCA	0.8	0.81	0.6	0.76	0.38	0.59
237	GCA_000011905.1_ASM1190v1	<i>Dehalococcoides mccartyi</i> 195	1	1	1	1	1	1
3450	GCA_000830905.1_ASM83090v1	<i>Dehalococcoides mccartyi</i> CG4	1	1	1	1	0.99	1

3847	GCA_001005265.1_ASM100526v1	<i>Dehalogenimonas</i> sp WBC2	0,8	0,74	0,73	0,69	0,47	0
854	GCA_000025725.1_ASM2572v1	<i>Denitrovibrio acetiphilus</i> DSM 12809	0,8	0,62	0,53	0,55	0,37	0
692	GCA_000021905.1_ASM2190v1	<i>Desulfatibacillum alkenivorans</i> AK01	0,9	0,72	0,65	0,75	0,55	0,52
1705	GCA_000243155.3_ASM24315v3	<i>Desulfotobacterium dehalogenans</i> ATCC 51507	0,6	0	0	0	0	0
1704	GCA_000243135.3_ASM24313v3	<i>Desulfotobacterium dichloroeliminans</i> LMG P21439	0	0	0	0	0	0
693	GCA_000021925.1_ASM2192v1	<i>Desulfotobacterium hafniense</i> DCB2	1	1	1	1	1	0,47
154	GCA_000010045.1_ASM1004v1	<i>Desulfotobacterium hafniense</i> Y51	1	0,97	0,99	1	0,99	0,45
1341	GCA_000195295.1_ASM19529v1	<i>Desulfobacca acetoxidans</i> DSM 11109	0,9	0,69	0,7	0,77	0,58	0,5
620	GCA_000020365.1_ASM2036v1	<i>Desulfobacterium autotrophicum</i> HRM2	0,8	0,7	0,71	0,7	0,52	0,56
1272	GCA_000186885.1_ASM18688v1	<i>Desulfobulbus propionicus</i> DSM 2032	0,9	0,86	0,8	0,7	0,59	0,53
2219	GCA_000341395.1_ASM34139v1	<i>Desulfocapsa sulfexigens</i> DSM 10523	0,8	0,83	0,8	0,77	0,6	0
1467	GCA_000212675.2_ASM21267v2	<i>Desulfocurvibacter</i> [ <i>Desulfovibrio</i> ] <i>africanus</i> str Walvis Bay	1	1	1	1	1	1
748	GCA_000023225.1_ASM2322v1	<i>Desulfomicrobium baculatum</i> DSM 4028	1	1	1	1	1	1
1822	GCA_000266945.1_ASM26694v1	<i>Desulfomonile tiedjei</i> DSM 6799	0,8	0,71	0,64	0,77	0,59	0,51
1750	GCA_000255115.3_ASM25511v3	<i>Desulfosporosinus acidiphilus</i> SJ4	0,9	0,95	0,89	0,9	0,76	0,51
1616	GCA_000231385.3_ASM23138v3	<i>Desulfosporosinus meridiei</i> DSM 13257	0,9	0,88	0,87	0,89	0,83	0,46
1638	GCA_000235605.1_ASM23560v1	<i>Desulfosporosinus orientis</i> DSM 765	1	1	1	1	1	0,5
790	GCA_000024205.1_ASM2420v1	<i>Desulfotomaculum acetoxidans</i> DSM 771	0,7	0,81	0,84	0,56	0,73	0,67
1625	GCA_000233715.3_ASM23371v3	<i>Desulfotomaculum gibsoniae</i> DSM 7213	0,9	0,8	0,82	0,74	0,68	0,63
1492	GCA_000214435.1_ASM21443v1	<i>Desulfotomaculum nigrificans</i> CO1SRB	1	1	1	1	1	1
434	GCA_000016165.1_ASM1616v1	<i>Desulfotomaculum reducens</i> MI1	0,9	0,91	0,9	0,88	0,84	0,82
1500	GCA_000215085.1_ASM21508v1	<i>Desulfotomaculum ruminis</i> DSM 2154	1	1	1	1	1	1
1281	GCA_000189295.2_ASM18929v2	<i>Desulfovibrio desulfuricans</i> ND132	0,9	0,79	0,78	0,83	0,62	0,59
185	GCA_000010665.1_ASM1066v1	<i>Desulfovibrio magneticus</i> RS1	0,9	0,73	0,63	0,8	0,62	0,6
759	GCA_000023445.1_ASM2344v1	<i>Desulfovibrio salexigens</i> DSM 2638	0,9	0,82	0,65	0,83	0,61	0,59
406	GCA_000015485.1_ASM1548v1	<i>Desulfovibrio vulgaris</i> DP4	1	1	1	1	1	1
1154	GCA_000166115.1_ASM16611v1	<i>Desulfovibrio vulgaris</i> RCH1	1	1	1	0,8	0,94	0,81
1354	GCA_000195755.1_ASM19575v1	<i>Desulfovibrio vulgaris</i> str Hildenborough	1	1	1	0,89	0,94	0,81
668	GCA_000021385.1_ASM2138v1	<i>Desulfovibrio vulgaris</i> str Miyazaki F	0,9	0,8	0,83	0,83	0,48	0,26
1203	GCA_000177635.2_ASM17763v2	<i>Desulfurispirillum indicum</i> S5	0,8	0,72	0	0,56	0,34	0,44
982	GCA_000092205.1_ASM9220v1	<i>Desulfurivibrio alkaliphilus</i> AHT 2	0,9	0,84	0,8	0,76	0,47	0

4228	GCA_001278055.1_ASM127805v1	<i>Desulfuromonas soudanensis</i>	0,9	0,79	0,77	0,84	0,7	0,6
5082	GCA_001611275.1_ASM161127v1	<i>Desulfuromonas</i> sp DDH964	0,9	0,78	0,75	0,83	0,68	0,61
1073	GCA_000147055.1_ASM14705v1	<i>Dickeya dadantii</i> 3937	1	0,93	0,91	0,9	0,84	0,83
8133	GCA_002706485.1_ASM270648v1	<i>Dickeya dianthicola</i> RNS049	1	0,94	0,92	0,87	0,83	0,76
761	GCA_000023545.1_ASM2354v1	<i>Dickeya paradisiaca</i> Ech703	1	0,93	0,92	0,92	0,79	0,82
3118	GCA_000758345.1_ASM75834v1	<i>Dickeya solani</i>	1	0,94	0,92	0,9	0,85	0,72
8506	GCA_002846995.1_ASM284699v1	<i>Dickeya solani</i>	1	0,93	0,91	0,89	0,84	0,83
8505	GCA_002846975.1_ASM284697v1	<i>Dickeya solani</i> D s04321	1	0,93	0,91	0,89	0,84	0,83
5227	GCA_001644705.1_ASM164470v1	<i>Dickeya solani</i> IPO 2222	1	0,93	0,91	0,89	0,84	0,83
2606	GCA_000511285.2_ASM51128v2	<i>Dickeya solani</i> RNS 082331A	1	0,93	0,91	0,86	0,84	0,83
5165	GCA_001632845.1_ASM163284v1	<i>Ectothiorhodospira</i> sp BSL9	1	1	1	1	1	0,69
2877	GCA_000697965.2_ASM69796v2	<i>Ensifer adhaerens</i>	0	0	0	0	0	0
7815	GCA_002288525.1_ASM228852v1	<i>Ensifer sojae</i> CCB AU 05684	1	1	1	1	1	1
4935	GCA_001582075.1_ASM158207v1	<i>Enterobacter</i> sp FY07	1	0,98	0,94	0,88	0,7	0,84
2336	GCA_000410515.1_ASM41051v1	<i>Enterobacter</i> sp R4368	1	1	1	1	1	1
8802	GCA_002903065.1_ASM290306v1	<i>Enterobacteriaceae bacterium</i> ENNIH1	1	0,98	0,92	0,93	0,8	0,79
810	GCA_000024665.1_ASM2466v1	<i>Fibrobacter succinogenes</i> subsp <i>succinogenes</i> S85	0,7	0	0	0	0	0,46
1067	GCA_000146505.1_ASM14650v1	<i>Fibrobacter succinogenes</i> subsp <i>succinogenes</i> S85	0,7	0	0	0	0	0,46
4792	GCA_001548455.1_ASM154845v1	<i>Fischerella</i> sp NIES3754	1	0,92	0,89	0,81	0,81	0,83
927	GCA_000058485.1_ASM5848v1	<i>Frankia alni</i> ACN14a	1	1	1	1	1	1
305	GCA_000013345.1_ASM1334v1	<i>Frankia casuarinae</i>	1	1	1	1	1	1
512	GCA_000018005.1_ASM1800v1	<i>Frankia</i> sp EAN1pec	1	0,92	0,91	0,83	0,5	0
1202	GCA_000177615.2_ASM17761v2	<i>Frankia symbiont of Datisca glomerata</i>	0,9	0,67	0,68	0,7	0	0
7971	GCA_002356455.1_ASM235645v1	<i>Fusobacterium varium</i>	0,6	0	0	0	0	0
2110	GCA_000317045.1_ASM31704v1	<i>Geitlerinema</i> sp PCC 7407	0	0	0	0	0	0
5155	GCA_001628815.1_ASM162881v1	<i>Geobacter anodireducens</i>	0	0	0	0	0	0
638	GCA_000020725.1_ASM2072v1	<i>Geobacter bemidjensis</i> Bem	1	0,94	0,92	0	0,85	0,68
708	GCA_000022265.1_ASM2226v1	<i>Geobacter daltonii</i> FRC32	1	0,95	0,92	0	0,77	0,87
621	GCA_000020385.1_ASM2038v1	<i>Geobacter lovleyi</i> SZ	1	1	1	0	1	0,57
284	GCA_000012925.1_ASM1292v1	<i>Geobacter metallireducens</i> GS15	1	1	1	0	0,83	1
3369	GCA_000817955.1_ASM81795v1	<i>Geobacter pickeringii</i>	1	0,93	0,93	0	0,87	0,8
1186	GCA_000175115.2_ASM17511v2	<i>Geobacter</i> sp M18	1	0,94	0,93	0	0,85	0,67

766	GCA_000023645.1_ASM2364v1	<i>Geobacter</i> sp M21	1	0,95	0,92	0	0,85	0,68
1439	GCA_000210155.1_ASM21015v1	<i>Geobacter sulfurreducens</i> KN400	1	1	0,87	0	0,84	1
62	GCA_000007985.2_ASM798v2	<i>Geobacter sulfurreducens</i> PCA	1	1	0,87	0	0,84	1
459	GCA_000016745.1_ASM1674v1	<i>Geobacter uraniireducens</i> Rf4	1	1	1	0	1	1
5823	GCA_001750685.1_ASM175068v1	<i>Geosporobacter ferrireducens</i>	0,9	0,88	0,91	0,78	0,39	0,79
2120	GCA_000317555.1_ASM31755v1	<i>Gloeocapsa</i> sp PCC 7428	0,2	0	0	0	0	0
665	GCA_000021325.1_ASM2132v1	<i>Gluconacetobacter diazotrophicus</i> PA1 5	1	1	1	1	1	1
939	GCA_000067045.1_ASM6704v1	<i>Gluconacetobacter diazotrophicus</i> PA1 5	1	1	1	1	1	1
3597	GCA_000940805.1_ASM94080v1	<i>Gyvuella sunshinyi</i> YC6258	0,9	0,92	0,88	0,83	0,68	0,81
7972	GCA_002356555.2_HH1059_assembly_1.0	<i>Halorhodospira halochloris</i>	0,9	0,76	0,83	0,71	0,66	0,38
411	GCA_000015585.1_ASM1558v1	<i>Halorhodospira halophila</i> SL1	1	1	1	1	1	1
2123	GCA_000317635.1_ASM31763v1	<i>Halothece</i> sp PCC 7418	0,9	0,81	0,7	0,83	0,68	0,7
626	GCA_000020485.1_ASM2048v1	<i>Halothermothrix orenii</i> H 168	0	0	0	0	0	0
5339	GCA_001663815.1_ASM166381v1	<i>Halothiobacillus</i> sp LS2	0,9	0,93	0,88	0	0,75	0,86
10082	GCA_900231165.1_HDIA1	<i>Hartmannibacter diazotrophicus</i>	0,9	0,86	0,8	0,79	0,72	0,82
562	GCA_000019165.1_ASM1916v1	<i>Heliobacterium modesticaldum</i> Ice1	1	1	1	1	1	1
4562	GCA_001483945.1_ASM148394v1	<i>Herbaspirillum rubrisubalbicans</i> M1	1	1	1	1	1	1
4013	GCA_001040945.1_ASM104094v1	<i>Herbaspirillum seropedicae</i>	1	1	1	1	1	1
1032	GCA_000143225.1_ASM14322v1	<i>Herbaspirillum seropedicae</i> SmR1	1	1	1	1	1	1
3879	GCA_001011155.1_ASM101115v1	<i>Hoefflea</i> sp IMCC20628	0	0	0	0	0	0
191	GCA_000010785.1_ASM1078v1	<i>Hydrogenobacter thermophilus</i> TK6	1	1	0	1	1	0
1138	GCA_000164905.1_ASM16490v1	<i>Hydrogenobacter thermophilus</i> TK6	1	1	0	1	1	0
1744	GCA_000253295.1_ASM25329v1	<i>Hyphomicrobium</i> sp MC1	1	0,95	0,92	0,87	0,78	0,91
1143	GCA_000165505.1_ASM16550v1	<i>Ilyobacter polytropus</i> DSM 2926	0,8	0,74	0,73	0,65	0	0
4470	GCA_001454945.1_ASM145494v1	<i>Intestinimonas butyriciproducens</i>	0,5	0	0	0	0	0,44
2924	GCA_000724525.1_ASM72452v1	<i>Klebsiella michiganensis</i>	1	1	1	0,99	0,99	0
7554	GCA_002216835.1_ASM221683v1	<i>Klebsiella michiganensis</i>	1	1	1	1	1	0,99
7826	GCA_002290285.1_ASM229028v1	<i>Klebsiella michiganensis</i>	1	1	1	1	1	0,99
9342	GCA_003074075.1_ASM307407v1	<i>Klebsiella michiganensis</i>	1	1	1	1	0,99	0,99
1869	GCA_000276705.2_ASM27670v2	<i>Klebsiella michiganensis</i> E718	1	1	1	1	1	0,99
2819	GCA_000632415.1_ASM63241v1	<i>Klebsiella michiganensis</i> HKOPL1	1	1	1	0,99	1	0,99
1682	GCA_000240325.1_ASM24032v1	<i>Klebsiella michiganensis</i> KCTC 1686	1	1	1	0,99	1	0,99



3952	GCA_001022195.1_ASM102219v1	<i>Klebsiella oxytoca</i>	1	1	1	1	1	1	1
4787	GCA_001548355.1_ASM154835v1	<i>Klebsiella oxytoca</i>	1	1	0,99	0,98	0,97	0,98	0,98
6406	GCA_001970835.1_ASM197083v1	<i>Klebsiella oxytoca</i>	1	1	1	0,99	0,99	0,99	0,99
6905	GCA_002072655.1_ASM207265v1	<i>Klebsiella oxytoca</i>	1	1	1	1	1	0,99	0,99
8814	GCA_002906395.1_ASM290639v1	<i>Klebsiella oxytoca</i>	1	1	1	0,99	0,99	0,99	0,99
8815	GCA_002906415.1_ASM290641v1	<i>Klebsiella oxytoca</i>	1	1	1	0,99	0,99	0,99	0,99
8816	GCA_002906435.1_ASM290643v1	<i>Klebsiella oxytoca</i>	1	1	0,99	0,99	1	0,99	0,99
8883	GCA_002947505.1_ASM294750v1	<i>Klebsiella oxytoca</i>	1	1	1	1	1	0,99	0,99
2911	GCA_000714655.1_ASM71465v1	<i>Klebsiella oxytoca KONIH1</i>	1	1	1	0,99	0,99	0,99	0,99
4785	GCA_001548315.1_ASM154831v1	<i>Klebsiella pneumoniae</i>	1	1	1	1	1	0,99	0,99
7217	GCA_002156765.1_ASM215676v1	<i>Klebsiella pneumoniae</i>	1	1	1	1	1	0,99	0,99
9311	GCA_003073315.1_ASM307331v1	<i>Klebsiella pneumoniae</i>	1	0,99	0,99	0,99	0,97	0,97	0,96
582	GCA_000019565.1_ASM1956v1	<i>Klebsiella pneumoniae 342</i>	1	1	1	1	1	1	1
2910	GCA_000714635.1_ASM71463v1	<i>Klebsiella pneumoniae subsp pneumoniae KP51</i>	1	1	1	1	1	0,99	0,99
8960	GCA_002953295.1_ASM295329v1	<i>Klebsiella quasipneumoniae</i>	1	1	1	0,99	0,98	0,97	0,97
9414	GCA_003146635.1_ASM314663v1	<i>Klebsiella quasipneumoniae</i>	1	1	1	0,99	0,96	0,96	0,96
9415	GCA_003146655.1_ASM314665v1	<i>Klebsiella quasipneumoniae</i>	1	1	1	0,99	0,96	0,96	0,96
9416	GCA_003146685.1_ASM314668v1	<i>Klebsiella quasipneumoniae</i>	1	1	1	0,99	0,96	0,96	0,96
9604	GCA_003285165.1_ASM328516v1	<i>Klebsiella quasipneumoniae subsp quasipneumoniae</i>	1	1	1	0,99	0,97	0,96	0,96
7724	GCA_002269255.1_ASM226925v1	<i>Klebsiella quasivariicola</i>	1	1	1	1	1	0,99	0,99
5852	GCA_001753185.1_ASM175318v1	<i>Klebsiella sp LTGPAF6F</i>	1	1	0,99	0,99	1	0,99	0,99
7032	GCA_002090195.1_ASM209019v1	<i>Klebsiella sp M5a1</i>	1	1	0,99	0,98	0,97	0,98	0,98
3310	GCA_000812205.2_ASM81220v2	<i>Klebsiella variicola</i>	1	1	1	0,99	0,99	1	1
3397	GCA_000828055.2_ASM82805v2	<i>Klebsiella variicola</i>	1	1	1	1	1	0,99	0,99
6676	GCA_001989495.1_ASM198949v1	<i>Klebsiella variicola</i>	1	1	1	1	1	0,99	0,99
6677	GCA_001989515.1_ASM198951v1	<i>Klebsiella variicola</i>	1	1	1	1	1	0,99	0,99
6678	GCA_001989535.1_ASM198953v1	<i>Klebsiella variicola</i>	1	1	1	1	1	0,99	0,99
7272	GCA_002164805.2_ASM216480v2	<i>Klebsiella variicola</i>	1	0,99	1	1	1	1	0,99
8572	GCA_002852915.3_ASM285291v3	<i>Klebsiella variicola</i>	1	1	1	1	1	0,99	1
9605	GCA_003285185.1_ASM328518v1	<i>Klebsiella variicola</i>	1	1	1	1	1	0,99	0,99
9645	GCA_003290425.1_ASM329042v1	<i>Klebsiella variicola</i>	1	1	1	1	1	1	0,99
846	GCA_000025465.1_ASM2546v1	<i>Klebsiella variicola At22</i>	1	1	1	1	1	1	1

5303	GCA_001658025.1_ASM165802v1	<i>Kosakonia oryzae</i>	1	0,98	0,95	0,95	0,85	0,85	0,84
6085	GCA_001887675.1_ASM188767v1	<i>Kosakonia radincincitans</i>	1	0,98	0,95	0,95	0,85	0,85	0,84
1896	GCA_000280495.2_ASM28049v2	<i>Kosakonia radincincitans DSM 16656</i>	1	0,98	0,95	0,95	0,85	0,85	0,85
5438	GCA_001683395.1_ASM168339v1	<i>Kosakonia sacchari</i>	1	0,99	0,99	0,99	0,96	0,96	0,95
2042	GCA_000300455.4_ASM30045v4	<i>Kosakonia sacchari SP1</i>	1	0,99	0,99	0,99	0,96	0,96	0,95
8368	GCA_002804065.1_ASM280406v1	<i>Kyripidia sp EA1</i>	0,7	0,71	0,6	0,57	0,31	0,49	
7964	GCA_002356295.1_ASM235629v1	<i>Labilibaculum antarcticum [Marinifilaceae bacterium] SPP2</i>	0,8	0,71	0,74	0,71	0,52	0	
6710	GCA_001999245.1_ASM199924v1	<i>Labrenzia aggregata</i>	0	0	0	0	0	0	0
4922	GCA_001579905.1_ASM157990v1	<i>Labrenzia sp CP4</i>	0	0	0	0	0	0	0
7641	GCA_002237595.1_ASM223759v1	<i>Labrenzia sp VG12</i>	0	0	0	0	0	0	0
5486	GCA_001688665.2_ASM168866v2	<i>Lachnoclostridium sp YL32</i>	0,7	0	0	0	0	0	0
7190	GCA_002142495.1_ASM214249v1	<i>Leptolyngbya boryana dg5</i>	1	1	1	1	1	1	1
7189	GCA_002142475.1_ASM214247v1	<i>Leptolyngbya boryana IAM M101</i>	1	1	1	1	1	1	1
7984	GCA_002368255.1_ASM236825v1	<i>Leptolyngbya boryana NIES2135</i>	1	1	1	1	1	1	1
4791	GCA_001548435.1_ASM154843v1	<i>Leptolyngbya sp NIES3755</i>	0	0	0	0	0	0	0
4789	GCA_001548395.1_ASM154839v1	<i>Leptolyngbya sp O77</i>	0,9	0,82	0,74	0,76	0,48	0,74	
2870	GCA_000695975.1_ASM69597v1	<i>Leptospirillum ferriphilum YSK</i>	0,8	0,89	0,76	0,75	0,43	0,64	
1944	GCA_000284315.1_ASM28431v1	<i>Leptospirillum ferrooxidans C23</i>	0,9	0,88	0,77	0,75	0,43	0,63	
592	GCA_000019785.1_ASM1978v1	<i>Leptothrix cholodnii SP6</i>	1	0,96	0,87	0,84	0,63	0,83	
379	GCA_000014865.1_ASM1486v1	<i>Magnetococcus marinus MC1</i>	0,9	0,83	0,72	0,77	0,49	0,5	
9050	GCA_002995535.1_ASM299553v1	<i>Magnetospirillum gryphiswaldense</i>	1	0,81	0,83	0,84	0,68	0,77	
9049	GCA_002995515.1_ASM299551v1	<i>Magnetospirillum gryphiswaldense MSR1</i>	1	0,81	0,83	0,84	0,68	0,77	
2630	GCA_000513295.1_MGMSRV2	<i>Magnetospirillum gryphiswaldense MSR1 v2</i>	1	0,81	0,83	0,84	0,67	0,77	
151	GCA_000009985.1_ASM998v1	<i>Magnetospirillum magneticum AMB1</i>	1	1	1	1	1	1	1
7061	GCA_002105535.1_ASM210553v1	<i>Magnetospirillum sp ME1</i>	1	0,88	0,83	0,97	0,97	0,95	
4644	GCA_001511835.1_XM-1	<i>Magnetospirillum sp XM1</i>	1	0,87	0,84	0,97	0,97	0,95	
3682	GCA_000960975.1_ASM96097v1	<i>Martella endophytica</i>	0,9	0,83	0,73	0,73	0,57	0,74	
6872	GCA_002043005.1_ASM204300v1	<i>Martella mediterranea DSM 17316</i>	0	0	0	0	0	0	0
4907	GCA_001578105.1_ASM157810v1	<i>Martella sp AD3</i>	0	0	0	0	0	0	0
9101	GCA_003006415.1_ASM300641v1	<i>Megasphaera elsdenii</i>	0,5	0	0	0	0	0	0
4309	GCA_001304715.1_ASM130471v1	<i>Megasphaera elsdenii 1414</i>	0,5	0	0	0	0	0	0





603	GCA_000020025.1_ASM2002v1	<i>Nostoc punctiforme</i> PCC 73102	1	0	1	1	1	1	1
8785	GCA_002896875.1_ASM289687v1	<i>Nostoc</i> sp CENA543	0	0,97	0,99	0,93	0,92	0,92	0,92
8900	GCA_002949795.1_ASM294979v1	<i>Nostoc</i> sp <i>Lobaria pulmonaria</i> 5183 cyanobiont	1	0,92	0,96	0,98	0,96	0,95	0,95
4788	GCA_001548375.1_ASM154837v1	<i>Nostoc</i> sp NIES3756	1	0,98	0,98	0,97	0,95	0,95	0,95
2105	GCA_000316625.1_ASM31662v1	<i>Nostoc</i> sp PCC 7107	1	0,96	0,98	0,94	0,98	0,99	0,99
138	GCA_000009705.1_ASM970v1	<i>Nostoc</i> sp PCC 7120	1	1	1	1	1	1	1
2106	GCA_000316645.1_ASM31664v1	<i>Nostoc</i> sp PCC 7524	0,8	0,97	0	0,92	0,92	0,93	0,93
8899	GCA_002949735.1_ASM294973v1	<i>Nostoc</i> sp <i>Peltigera membranacea</i> cyanobiont N6	1	0,93	0,96	0,96	0,99	0,96	0,96
7269	GCA_002163975.1_ASM216397v1	<i>Nostocales cyanobacterium</i> HT582	1	0,91	0,92	0,79	0,74	0,82	0,82
486	GCA_000017405.1_ASM1740v1	<i>Ochrobactrum anthropi</i> ATCC 49188	0	0	0	0	0	0	0
5262	GCA_001652485.1_ASM165248v1	<i>Ochrobactrum pseudogrignonense</i>	0	0	0	0	0	0	0
7736	GCA_002278035.1_ASM227803v1	<i>Ochrobactrum</i> sp A44	0	0	0,2	0	0	0	0
1521	GCA_000218585.1_ASM21858v1	<i>Oligotropha carboxidovorans</i> OM4	0	0	0	0	0	0	0
667	GCA_000021365.1_ASM2136v1	<i>Oligotropha carboxidovorans</i> OM5	0	0	0	0	0	0	0
1520	GCA_000218565.1_ASM21856v1	<i>Oligotropha carboxidovorans</i> OM5	0	0	0	0	0	0	0
1701	GCA_000242935.3_ASM24293v3	<i>Optutaceae bacterium</i> TAV5	0,8	0,6	0,61	0,63	0,42	0	0
2113	GCA_000317105.1_ASM31710v1	<i>Oscillatoria acuminata</i> PCC 6304	0	0	0	0	0	0	0
2117	GCA_000317475.1_ASM31747v1	<i>Oscillatoria nigroviridis</i> PCC 7112	0	0	0	0	0	0	0
3129	GCA_000758665.1_ASM75866v1	<i>Paenibacillus borealis</i>	1	0,93	0,86	0,96	0,94	0,93	0,93
7341	GCA_002192415.1_ASM219241v1	<i>Paenibacillus donghaensis</i>	0,9	0,86	0,79	0,94	0,9	0,82	0,82
3104	GCA_000756615.1_ASM75661v1	<i>Paenibacillus durus</i>	1	1	1	1	1	1	1
3846	GCA_000993825.1_ASM99382v1	<i>Paenibacillus durus</i> ATCC 35681	1	0,99	0,98	1	1	0,99	0,99
3131	GCA_000758705.1_ASM75870v1	<i>Paenibacillus graminis</i>	1	0,97	0,93	0,95	0,92	0,94	0,94
7664	GCA_002240415.1_ASM224041v1	<i>Paenibacillus kribbensis</i>	1	0,97	0,95	0,97	0,96	0,93	0,93
2611	GCA_000511405.1_ASM51140v1	<i>Paenibacillus larvae</i> subsp <i>larvae</i> DSM 25430	0	0	0	0	0	0	0
3132	GCA_000758725.1_ASM75872v1	<i>Paenibacillus odorifer</i>	0,9	0,87	0,82	0,92	0,89	0,85	0,85
9570	GCA_003255855.1_ASM325585v1	<i>Paenibacillus odorifer</i>	0,9	0,86	0,82	0,92	0,89	0,85	0,85
4184	GCA_001272655.2_ASM127265v2	<i>Paenibacillus peoriae</i>	1	0,98	0,97	0,98	0,98	0,96	0,96
3383	GCA_000819665.1_ASM81966v1	<i>Paenibacillus polymyxa</i>	1	1	1	1	1	1	1
5715	GCA_001719045.1_ASM171904v1	<i>Paenibacillus polymyxa</i>	1	0,98	0,97	0,98	0,97	0,96	0,96
5995	GCA_001874405.3_ASM187440v3	<i>Paenibacillus polymyxa</i>	1	0,98	0,97	0,98	0,97	0,96	0,96
5996	GCA_001874425.3_ASM187442v3	<i>Paenibacillus polymyxa</i>	1	0,98	0,97	0,98	0,98	0,96	0,96

6258	GCA_001922145.1_ASM192214v1	<i>Paenibacillus polymyxa</i>	1	0,98	0,97	0,98	0,98	0,98	0,96
2593	GCA_000507205.2_ASM50720v2	<i>Paenibacillus polymyxa</i> CR1	1	0,97	0,96	0,98	0,98	0,98	0,96
3810	GCA_000981585.1_SBR5_T_	<i>Paenibacillus riograndensis</i> SBR5	1	0,96	0,93	0,95	0,93	0,92	0,92
2773	GCA_000612505.1_ASM61250v1	<i>Paenibacillus sabiniae</i> T27	1	0,98	0,96	0,97	0,98	0,96	0,96
3122	GCA_000758525.1_ASM75852v1	<i>Paenibacillus</i> sp FSL H70357	1	1	1	1	1	1	1
3124	GCA_000758565.1_ASM75856v1	<i>Paenibacillus</i> sp FSL P40081	1	0,94	0,85	0,97	0,94	0,94	0,94
3126	GCA_000758605.1_ASM75860v1	<i>Paenibacillus</i> sp FSL R50912	1	0,93	0,86	0,96	0,94	0,94	0,94
3127	GCA_000758625.1_ASM75862v1	<i>Paenibacillus</i> sp FSL R70273	1	0,97	0,94	0,95	0,94	0,92	0,92
3128	GCA_000758645.1_ASM75864v1	<i>Paenibacillus</i> sp FSL R70331	1	0,96	0,94	0,96	0,93	0,93	0,93
8663	GCA_002872435.1_ASM287243v1	<i>Paenibacillus</i> sp IzhN1	1	1	0,99	1	1	1	0,99
3130	GCA_000758685.1_ASM75868v1	<i>Paenibacillus stellifer</i>	1	0,93	0,91	0,95	0,95	0,95	0,95
1637	GCA_000235585.1_ASM23558v1	<i>Paenibacillus terrae</i> HPL003	1	0,97	0,95	0,98	0,97	0,94	0,94
4568	GCA_001484065.1_ASM148406v1	<i>Pannonibacter phragmitetus</i>	0	0	0	0	0	0	0
1191	GCA_000175935.2_ASM17593v2	<i>Pantoea</i> sp At9b	1	1	1	1	1	1	1
7738	GCA_002278075.1_ASM227807v1	<i>Paraburkholderia aromaticivorans</i> BN5	1	1	0,99	0,99	1	0,99	0,99
2037	GCA_000300095.1_ASM30009v1	<i>Paraburkholderia phenoliruptrix</i> BR3459a	1	1	1	1	1	1	0,99
604	GCA_000020045.1_ASM2004v1	<i>Paraburkholderia phymatum</i> STM815	1	1	1	1	1	1	1
5964	GCA_001865575.1_ASM186557v1	<i>Paraburkholderia sprentiae</i> WSM5005	1	0,97	0,95	0,93	0,92	0,94	0,94
319	GCA_000013645.1_ASM1364v1	<i>Paraburkholderia xenovorans</i> LB400	1	1	1	1	1	1	1
3089	GCA_000756045.1_ASM75604v1	<i>Paraburkholderia xenovorans</i> LB400	1	1	1	1	1	1	1
1949	GCA_000284415.2_ASM28441v2	<i>Parahodospirillum photometricum</i> DSM 122	0,9	0,8	0,81	0,66	0,45	0,74	0,74
2873	GCA_000696465.1_ASM69646v1	<i>Pectobacterium atrosepticum</i>	1	0,92	0,9	0,89	0,74	0,78	0,78
3005	GCA_000740965.1_ASM74096v1	<i>Pectobacterium atrosepticum</i>	1	0,92	0,9	0,89	0,74	0,78	0,78
8376	GCA_002804265.1_ASM280426v1	<i>Pectobacterium atrosepticum</i>	1	0,92	0,9	0,9	0,74	0,78	0,78
222	GCA_000011605.1_ASM1160v1	<i>Pectobacterium atrosepticum</i> SCR11043	1	0,92	0,9	0,9	0,74	0,77	0,77
7855	GCA_002307355.1_ASM230735v1	<i>Pectobacterium polaris</i>	1	0,94	0,91	0,91	0,76	0,77	0,77
6090	GCA_001888165.1_ASM188816v1	<i>Pelobacter acetylenicus</i>	0	0,92	0,86	0,89	0,82	0	0
283	GCA_000012885.1_ASM1288v1	<i>Pelobacter carbinolicus</i> DSM 2380	1	1	1	1	1	1	1
387	GCA_000015045.1_ASM1504v1	<i>Pelobacter propionicus</i> DSM 2379	1	1	1	0	1	1	1
6087	GCA_001887775.1_ASM188777v1	<i>Pelobacter</i> sp SFB93	1	0,86	0,71	0,88	0,79	0,62	0,62
264	GCA_000012485.1_ASM1248v1	<i>Pelodictyon [Chlorobium] luteolum</i> DSM 273	1	1	1	1	1	1	1
634	GCA_000020645.1_ASM2064v1	<i>Pelodictyon phaeoclathratiforme</i> BU1	1	0,92	0,91	0,9	0,87	0,84	0,84

1854	GCA_000271665.2_ASM27166v2	<i>Pelosinus fermentans</i> JBW45	0,9	0,83	0,81	0,83	0,41	0,47
2933	GCA_000725345.1_ASM72534v1	<i>Pelosinus</i> sp UFO1	0,9	0,83	0,81	0,76	0,38	0,59
180	GCA_000010565.1_ASM1056v1	<i>Pelotomaculum thermopropionicum</i> SI	0	0	0	0	0	0
3604	GCA_000940995.1_ASM94099v1	<i>Photobacterium gaetbulicola</i> Gung47	0,5	0	0	0	0	0,44
9104	GCA_003008515.1_ASM300851v1	<i>Phreatobacter</i> sp S12	0	0	0	0	0	0
8295	GCA_002764115.1_ASM276411v1	<i>Phyllobacterium</i> sp Tri48	0	0	0	0	0	0
2109	GCA_000317025.1_ASM31702v1	<i>Pleurocapsa</i> sp PCC 7327	0,9	0,86	0,81	0,8	0,73	0,8
407	GCA_000015505.1_ASM1550v1	<i>Polaromonas naphthalenivorans</i> CJ2	1	1	1	1	1	1
633	GCA_000020625.1_ASM2062v1	<i>Prosthecochloris aestuarii</i> DSM 271	1	1	1	1	1	1
5470	GCA_001687065.1_ASM168706v1	<i>Prosthecochloris</i> sp CIB 2401	1	0,93	0,93	0,92	0,84	0,84
7657	GCA_002240205.1_ASM224020v1	<i>Prosthecochloris</i> sp GSB1	1	0,87	0,9	0,91	0,9	0,82
7100	GCA_002113825.1_ASM211382v1	<i>Prosthecochloris</i> sp HL130GSB	0	0,88	0,9	0	0,88	0,82
2111	GCA_000317065.1_ASM31706v1	<i>Pseudanabaena</i> sp PCC 7367	0	0	0	0	0	0
8784	GCA_002896855.1_ASM289685v1	<i>PseudoClostridium thermosuccinogenes</i>	0,8	0,72	0,6	0,65	0,43	0,47
1196	GCA_000176915.2_ASM17691v2	<i>Pseudodesulfovibrio aespoensis</i> Aspo2	0,9	0,83	0,79	0,82	0,6	0,61
4874	GCA_001563225.1_ASM156322v1	<i>Pseudodesulfovibrio indicus</i>	0,9	0,85	0,79	0,83	0,52	0,61
10081	GCA_900217235.1_DPRO_PRJEB22548_v1	<i>Pseudodesulfovibrio profundus</i>	0,9	0,84	0,74	0,82	0,62	0,62
9236	GCA_003047145.1_ASM304714v1	<i>Pseudomonas stutzeri</i>	1	1	1	0,98	0,99	0,88
325	GCA_000013785.1_ASM1378v1	<i>Pseudomonas stutzeri</i> A1501	1	1	1	1	1	1
1339	GCA_000195105.1_ASM19510v1	<i>Pseudomonas stutzeri</i> DSM 4166	1	1	1	0,98	0,99	0,84
7151	GCA_002119765.1_ASM211976v1	<i>Pseudorhodoplanes sinuspersici</i>	0	0	0	0	0	0
1760	GCA_000255535.1_ASM25553v1	<i>Rahnella aquatilis</i> HX2	1	0,89	0,89	0,85	0,78	0,69
117	GCA_000009285.2_ASM928v2	<i>Ralstonia eutropha</i> H16	0	0	0	0	0	0
3437	GCA_000829965.1_ASM82996v1	<i>Raoultella ornithinolytica</i>	1	0,99	0,98	0,97	0,91	0,9
5352	GCA_001664265.1_ASM166426v1	<i>Rhizobium esperanzae</i> (etli) N561	1	1	0,99	0,99	0,99	1
7154	GCA_002119845.1_ASM211984v1	<i>Rhizobium etli</i>	1	0,97	0,96	0,88	0,85	0,89
6224	GCA_001908375.1_ASM190837v1	<i>Rhizobium etli</i> 8C3	1	1	1	1	0,99	1
2424	GCA_000442435.1_ASM44243v1	<i>Rhizobium etli</i> bv mimosae str Mimi1	1	0,97	0,96	0,88	0,85	0,92
3342	GCA_000816125.1_ASM81612v1	<i>Rhizobium etli</i> bv phaseoli str IE4803	1	0,99	0,99	0,93	0,97	0,98
978	GCA_000092045.1_ASM9204v1	<i>Rhizobium etli</i> CFN 42	1	1	1	1	1	1
615	GCA_000020265.1_ASM2026v1	<i>Rhizobium etli</i> CIAT 652	1	1	0,99	0,99	0,99	1
6231	GCA_001908615.1_ASM190861v1	<i>Rhizobium gallicum</i>	1	0,95	0,94	0,89	0,88	0,92

3359	GCA_000816845.1_ASM81684v1	<i>Rhizobium gallicum</i> bv <i>gallicum</i> R602	1	0,96	0,94	0,86	0,85	0,92
5405	GCA_001679785.1_ASM167978v1	<i>Rhizobium leguminosarum</i>	1	0,95	0,96	0,96	0,95	0,95
6132	GCA_001890425.1_ASM189042v1	<i>Rhizobium leguminosarum</i>	1	0,95	0,96	0,96	0,95	0,95
8973	GCA_002953715.1_ASM295371v1	<i>Rhizobium leguminosarum</i>	1	0,93	0,97	0,95	0,92	0,92
9644	GCA_003290405.1_ASM329040v1	<i>Rhizobium leguminosarum</i>	1	0,94	0,97	0,93	0,87	0,88
2647	GCA_000520875.1_ASM52087v1	<i>Rhizobium leguminosarum</i> bv <i>trifolii</i> CB782	0,9	0,91	0,93	0,81	0,76	0,3
746	GCA_000023185.1_ASM2318v1	<i>Rhizobium leguminosarum</i> bv <i>trifolii</i> WSM1325	1	1	1	1	1	1
2643	GCA_000517605.1_ASM51760v1	<i>Rhizobium leguminosarum</i> bv <i>trifolii</i> WSM1689	1	0,97	0,96	0,92	0,89	0,93
666	GCA_000021345.1_ASM2134v1	<i>Rhizobium leguminosarum</i> bv <i>trifolii</i> WSM2304	1	0,86	0,91	0,82	0,77	0
7674	GCA_002243365.1_ASM224336v1	<i>Rhizobium leguminosarum</i> bv <i>viciae</i>	1	0,94	0,98	0,97	0,94	0,96
8892	GCA_002948295.1_ASM294829v1	<i>Rhizobium leguminosarum</i> bv <i>viciae</i>	1	0,97	0,96	0,97	0,95	0,95
7656	GCA_002240185.1_ASM224018v1	<i>Rhizobium leguminosarum</i> bv <i>viciae</i>	0	0	0	0	0	0
116	GCA_000009265.1_ASM926v1	<i>Rhizobium leguminosarum</i> bv <i>viciae</i> 3841	1	1	1	1	1	1
5344	GCA_001664105.1_ASM166410v1	<i>Rhizobium phaseoli</i>	1	0,99	1	0,99	0,99	1
5345	GCA_001664125.1_ASM166412v1	<i>Rhizobium phaseoli</i>	1	1	1	0,99	0,99	1
5347	GCA_001664165.1_ASM166416v1	<i>Rhizobium phaseoli</i>	1	1	1	0,99	0,99	1
5349	GCA_001664205.1_ASM166420v1	<i>Rhizobium phaseoli</i>	1	1	1	0,99	0,99	1
5350	GCA_001664225.1_ASM166422v1	<i>Rhizobium phaseoli</i>	1	0,99	1	0,99	0,99	1
5351	GCA_001664245.1_ASM166424v1	<i>Rhizobium phaseoli</i>	1	1	1	0,99	0,99	0,99
5353	GCA_001664285.1_ASM166428v1	<i>Rhizobium phaseoli</i>	1	1	1	0,99	0,99	1
5357	GCA_001664365.1_ASM166436v1	<i>Rhizobium phaseoli</i>	1	1	1	0,99	0,99	1
5358	GCA_001664385.1_ASM166438v1	<i>Rhizobium phaseoli</i>	1	0,99	1	0,99	0,99	1
5359	GCA_001664405.1_ASM166440v1	<i>Rhizobium phaseoli</i>	1	1	1	0,99	0,99	1
5360	GCA_001664425.1_ASM166442v1	<i>Rhizobium phaseoli</i>	1	1	1	0,99	0,99	1
5361	GCA_001664445.1_ASM166444v1	<i>Rhizobium phaseoli</i>	1	0,99	1	0,99	0,99	1
5348	GCA_001664185.1_ASM166418v1	<i>Rhizobium phaseoli</i>	0	0	0	0	0	0
1181	GCA_000172715.2_ASM17271v2	<i>Rhizobium phaseoli</i> Brasil 5	1	1	0,99	0,99	0,99	0,99
2575	GCA_000499645.1_IRBG74	<i>Rhizobium pusense</i> IRBG74 (sp IRGB74)	1	0,94	0,92	0,82	0,88	0,9
7730	GCA_002277895.1_ASM227789v1	<i>Rhizobium</i> sp 11515TR	0	0	0	0	0	0
8118	GCA_002600635.1_ASM260063v1	<i>Rhizobium</i> sp ACO34A	0,9	0,89	0,83	0,8	0	0,83
1182	GCA_000172795.2_ASM17279v2	<i>Rhizobium</i> sp CIAT894	1	1	1	0,99	0,99	1
2886	GCA_000698845.1_ASM69884v1	<i>Rhizobium</i> sp IE4771	1	0,97	0,96	0,93	0,91	0,92



1180	GCA_000172695.2_ASM17269v2	<i>Rhizobium</i> sp Kim5	1	1	1	0,93	0,97	0,98
5356	GCA_001664345.1_ASM166434v1	<i>Rhizobium</i> sp N113	1	1	0,99	0,99	0,99	1
5354	GCA_001664305.1_ASM166430v1	<i>Rhizobium</i> sp N1314	1	1	1	0,99	0,99	1
5343	GCA_001664085.1_ASM166408v1	<i>Rhizobium</i> sp N1341	1	1	0,99	0,99	0,99	1
5363	GCA_001664485.1_ASM166448v1	<i>Rhizobium</i> sp N324	1	1	1	0,99	0,99	0,99
5355	GCA_001664325.1_ASM166432v1	<i>Rhizobium</i> sp N621	1	1	0,99	0,99	0,99	1
5362	GCA_001664465.1_ASM166446v1	<i>Rhizobium</i> sp N6212	1	1	0,99	0,99	0,99	1
5346	GCA_001664145.1_ASM166414v1	<i>Rhizobium</i> sp N731	1	1	1	0,99	0,99	1
5365	GCA_001664545.1_ASM166454v1	<i>Rhizobium</i> sp N741	1	1	0,99	0,99	0,99	1
5364	GCA_001664525.1_ASM166452v1	<i>Rhizobium</i> sp N871	1	1	0,99	0,99	0,99	1
3731	GCA_000967425.1_ASM96742v1	<i>Rhizobium</i> sp NT26	0	0	0	0	0	0
7132	GCA_002117485.1_ASM211748v1	<i>Rhizobium</i> sp NXC14	1	0,96	0,96	0,88	0,85	0,9
8848	GCA_002944315.1_ASM294431v1	<i>Rhizobium</i> sp NXC24	1	0,98	0,99	0,98	0,98	0,96
5505	GCA_001691455.1_ASM169145v1	<i>Rhizobium</i> sp S41	0	0	0	0	0	0
7133	GCA_002117725.1_ASM211772v1	<i>Rhizobium</i> sp TAL182	1	1	1	0,99	0,99	1
8427	GCA_002814035.1_ASM281403v1	<i>Rhizobium</i> sp Y9	0	0	0	0	0	0
2179	GCA_000330885.1_ASM33088v1	<i>Rhizobium tropici</i> CIAT 899	1	1	1	1	1	1
690	GCA_000021865.1_ASM2186v1	<i>Rhodobacter capsulatus</i> SB 1003	1	1	1	1	1	1
4896	GCA_001576595.1_ASM157659v1	<i>Rhodobacter sphaeroides</i>	1	1	1	0,78	0,68	1
5452	GCA_001685625.1_ASM168562v1	<i>Rhodobacter sphaeroides</i>	1	1	1	0,78	0,68	1
8130	GCA_002706325.1_ASM270632v1	<i>Rhodobacter sphaeroides</i>	1	1	1	0,78	0,68	1
9291	GCA_003071405.1_ASM307140v1	<i>Rhodobacter sphaeroides</i>	1	0,87	0,77	0,76	0,76	0,77
9670	GCA_003324715.1_ASM332471v1	<i>Rhodobacter sphaeroides</i> 241	1	1	1	0,78	0,68	1
444	GCA_000016405.1_ASM1640v1	<i>Rhodobacter sphaeroides</i> ATCC 17025	0,9	0,97	0,95	0,69	0,67	0,95
427	GCA_000015985.1_ASM1598v1	<i>Rhodobacter sphaeroides</i> ATCC 17029	1	1	1	0	1	1
650	GCA_000021005.1_ASM2100v1	<i>Rhodobacter sphaeroides</i> KD131	1	0,99	1	0,78	0,66	0,99
8638	GCA_002863805.1_ASM286380v1	<i>Rhodocyclaceae bacterium</i>	0,9	0,93	0,9	0,88	0,8	0,83
6383	GCA_001955735.1_ASM195573v1	<i>Rhodoferax antarcticus</i>	1	0,93	0,91	0,9	0,84	0,84
1152	GCA_000166055.1_ASM16605v1	<i>Rhodomicrobium vannielii</i> ATCC 17100	1	1	1	1	1	1
4919	GCA_001579845.1_ASM157984v1	<i>Rhodoplanes</i> sp Z2YC6860	0	0	0	0	0	0
9198	GCA_003031245.1_ASM303124v1	<i>Rhodopseudomonas palustris</i>	1	0,99	0,97	0,98	0,96	0,99
9199	GCA_003031265.1_ASM303126v1	<i>Rhodopseudomonas palustris</i>	1	0,99	0,97	0,98	0,96	0,99

377	GCA_000014825.1_ASM1482v1	<i>Rhodopseudomonas palustris BisA53</i>	1	0,92	0,85	0,75	0,79	0,87
323	GCA_000013745.1_ASM1374v1	<i>Rhodopseudomonas palustris BisB18</i>	1	0,89	0,87	0,65	0,79	0,86
321	GCA_000013685.1_ASM1368v1	<i>Rhodopseudomonas palustris BisB5</i>	1	0,97	0,9	0,93	0,89	0,9
1355	GCA_000195775.1_ASM19577v1	<i>Rhodopseudomonas palustris CGA009</i>	1	1	1	1	1	1
1200	GCA_000177255.2_ASM17725v2	<i>Rhodopseudomonas palustris DX1</i>	1	0,99	0,98	0,98	0,95	0,97
306	GCA_000013365.1_ASM1336v1	<i>Rhodopseudomonas palustris HaA2</i>	1	0,97	0,93	0,94	0,66	0,92
624	GCA_000020445.1_ASM2044v1	<i>Rhodopseudomonas palustris TIE1</i>	1	1	1	1	0,99	1
435	GCA_000016185.1_ASM1618v1	<i>Rhodospirillum centenum SW</i>	1	1	1	0	1	1
292	GCA_000013085.1_ASM1308v1	<i>Rhodospirillum rubrum ATCC 11170</i>	1	1	1	1	1	1
1587	GCA_000225955.1_ASM22595v1	<i>Rhodospirillum rubrum F11</i>	1	1	1	1	1	1
6944	GCA_002073975.1_ASM207397v1	<i>Rhodovulum sp MB263</i>	1	0,86	0,81	0,73	0,67	0,62
6991	GCA_002079305.1_ASM207930v1	<i>Rhodovulum sp P5</i>	0,9	0,85	0,76	0,74	0,47	0,66
4773	GCA_001548075.1_ASM154807v1	<i>Rhodovulum sulfidophilum</i>	1	0,86	0,8	0,73	0,48	0,62
5167	GCA_001633145.1_ASM163314v1	<i>Rhodovulum sulfidophilum</i>	1	0,86	0,8	0	0,48	0,62
5168	GCA_001633165.1_ASM163316v1	<i>Rhodovulum sulfidophilum DSM 1374</i>	1	0,86	0,8	0	0,48	0,62
687	GCA_000021805.1_ASM2180v1	<i>Rippakea orientalis 8801</i>	1	1	1	1	1	1
783	GCA_000024045.1_ASM2404v1	<i>Rippakea orientalis 8802</i>	1	1	1	1	1	1
2107	GCA_000316665.1_ASM31666v1	<i>Rivularia sp PCC 7116</i>	0,9	0	0,82	0,83	0,79	0
505	GCA_000017805.1_ASM1780v1	<i>Roseiflexus castenholzii DSM 13941</i>	0,7	0,29	0,39	0	0	0,46
455	GCA_000016665.1_ASM1666v1	<i>Roseiflexus sp RS1</i>	0,7	0,35	0,38	0	0	0,39
1941	GCA_000284255.1_ASM28425v1	<i>Rubrivax gelatinosus IL 144</i>	0,9	0,92	0,83	0,83	0,59	0,8
1761	GCA_000255615.3_ASM25561v3	<i>Ruminiclostridium thermocellum AD2</i>	0,8	0,58	0	0	0,22	0,47
422	GCA_000015865.1_ASM1586v1	<i>Ruminiclostridium thermocellum ATCC 27405</i>	0,8	0,58	0	0	0	0,47
1250	GCA_000184925.1_ASM18492v1	<i>Ruminiclostridium thermocellum DSM 1313</i>	0,8	0,58	0	0	0	0,47
5516	GCA_001692755.1_ASM169275v1	<i>Ruminiclostridium thermocellum DSM 2360</i>	0,8	0,58	0	0	0,22	0,47
1216	GCA_000179635.2_ASM17963v2	<i>Ruminococcus albus 7 DSM 20455</i>	0,8	0,62	0	0	0	0,46
8428	GCA_002814095.1_ASM281409v1	<i>Sagittula sp P11</i>	0,9	0,88	0,8	0,76	0,48	0,76
3862	GCA_001007875.1_ASM100787v1	<i>Sedimenticola thiotaurini</i>	0,9	0,88	0,83	0,79	0,63	0,76
1043	GCA_000143985.1_ASM14398v1	<i>Sediminisprochaeta smaragdinae DSM 11293</i>	0,9	0,74	0,72	0,65	0	0
1933	GCA_000284095.1_ASM28409v1	<i>Selenomonas ruminantium subsp lactilytica TAM6421</i>	0,8	0,66	0	0	0	0,48
4657	GCA_001514455.1_ASM151445v1	<i>Serratia fonticola</i>	0	0	0	0	0	0
2454	GCA_000463345.3_ASM46334v3	<i>Serratia sp ATCC 39006</i>	1	1	1	1	1	1

8507	GCA_002847015.1_ASM284701v1	<i>Serratia</i> sp ATCC 39006	1	1	1	1	1	1	1	1	1	1
5264	GCA_001652565.1_ASM165256v1	<i>Shinella</i> sp HZN7	0	0	0	0	0	0	0	0	0	0
6101	GCA_001889105.1_ASM188910v1	<i>Sinorhizobium americanum</i>	1	0,99	0,98	0,95	0,98	0,98	0,98	0,99	0,98	0,99
2895	GCA_000705595.2_ASM70559v2	<i>Sinorhizobium americanum</i> CCGM7	1	0,99	0,99	0,94	0,98	0,98	0,98	0,98	0,98	0,98
8849	GCA_002944405.1_ASM294440v1	<i>Sinorhizobium fredii</i>	1	1	1	1	1	1	1	1	1	1
9449	GCA_003177055.1_ASM317705v1	<i>Sinorhizobium fredii</i> CCBAU 25509	1	1	1	1	1	1	1	1	1	1
9381	GCA_003100575.1_ASM310057v1	<i>Sinorhizobium fredii</i> CCBAU 45436	1	1	1	1	1	1	1	1	1	1
7813	GCA_002288485.1_ASM228848v1	<i>Sinorhizobium fredii</i> CCBAU 83666	1	1	1	1	1	1	1	1	1	1
534	GCA_000018545.1_ASM1854v1	<i>Sinorhizobium fredii</i> NGR234	1	1	1	1	1	1	1	1	1	1
479	GCA_000017145.1_ASM1714v1	<i>Sinorhizobium medicae</i> WSM419	1	1	1	1	1	1	1	1	1	1
3033	GCA_000747295.1_ASM74729v1	<i>Sinorhizobium melliloti</i>	1	1	1	0,99	1	1	1	1	1	1
7352	GCA_002197025.1_ASM219702v1	<i>Sinorhizobium melliloti</i>	1	1	1	0,99	1	1	1	1	1	1
7353	GCA_002197045.1_ASM219704v1	<i>Sinorhizobium melliloti</i>	1	1	1	1	1	0,99	1	1	1	1
7354	GCA_002197065.1_ASM219706v1	<i>Sinorhizobium melliloti</i>	1	1	1	1	1	1	1	1	1	1
7355	GCA_002197085.1_ASM219708v1	<i>Sinorhizobium melliloti</i>	1	0,96	0,98	0,92	0,93	0,93	0,93	0,93	0,93	0,93
7356	GCA_002197105.1_ASM219710v1	<i>Sinorhizobium melliloti</i>	1	1	1	0,99	1	1	1	1	1	1
7358	GCA_002197145.1_ASM219714v1	<i>Sinorhizobium melliloti</i>	1	0,99	1	0	1	1	1	1	1	1
7359	GCA_002197165.1_ASM219716v1	<i>Sinorhizobium melliloti</i>	1	1	1	0,99	1	1	1	1	1	1
7371	GCA_002197445.1_ASM219744v1	<i>Sinorhizobium melliloti</i>	1	1	1	1	1	0,99	1	1	1	1
7372	GCA_002197465.1_ASM219746v1	<i>Sinorhizobium melliloti</i>	1	1	1	0,99	1	1	1	1	1	1
7538	GCA_002215195.1_ASM221519v1	<i>Sinorhizobium melliloti</i>	1	1	1	0,98	1	1	1	1	1	1
7837	GCA_002302355.1_ASM230235v1	<i>Sinorhizobium melliloti</i>	1	1	1	1	1	1	1	1	1	1
7838	GCA_002302375.1_ASM230237v1	<i>Sinorhizobium melliloti</i>	1	1	1	1	1	1	1	1	1	1
7357	GCA_002197125.1_ASM219712v1	<i>Sinorhizobium melliloti</i>	0	0	0,21	0	0	0	0	0	0	0
18	GCA_000006965.1_ASM696v1	<i>Sinorhizobium melliloti</i> 1021	1	1	1	1	1	1	1	1	1	1
2236	GCA_000346065.1_ASM34606v1	<i>Sinorhizobium melliloti</i> 2011	1	1	1	1	1	1	1	1	1	1
1082	GCA_000147795.3_ASM14779v3	<i>Sinorhizobium melliloti</i> AK83	1	1	1	0,99	1	1	1	1	1	1
1081	GCA_000147775.3_ASM14777v3	<i>Sinorhizobium melliloti</i> BL225C	1	1	1	0,98	1	1	1	1	1	1
2158	GCA_000320385.2_ASM32038v2	<i>Sinorhizobium melliloti</i> GR4	1	1	1	0,99	0,99	0,99	0,99	0,99	0,99	0,99
2052	GCA_000304415.1_RM41	<i>Sinorhizobium melliloti</i> Rm41	1	1	1	1	1	0,99	1	1	0,99	1
4060	GCA_001050915.2_ASM105091v2	<i>Sinorhizobium melliloti</i> RU11001	1	0,99	1	0,98	1	0,98	1	1	1	1
1517	GCA_000218265.1_ASM21826v1	<i>Sinorhizobium melliloti</i> SM11	1	0,99	1	0,98	1	0,98	1	1	1	1

7814	GCA_002288505.1_ASM228850v1	<i>Sinorhizobium</i> sp CCBAU 05631	1	1	1	0,99	0,99	0,99	0,93
5657	GCA_008843145.1_ASM884314v1	<i>Skermanella pratensis</i> strain W17	0,99	0,98	0,97	0,97	0,97	0,99	0,99
5658	GCA_016653635.2_ASM1665363v2	<i>Skermanella</i> sp TT6	0,99	0,98	0,97	0,97	0,99	0,99	0,98
3594	GCA_000935025.1_ASM93502v1	<i>Sphingomonas hengshuiensis</i>	0,9	0,92	0,87	0,49	0,74	0,81	
1074	GCA_000147075.1_ASM14707v1	<i>Spirochaeta thermophila</i> DSM 6192	0,8	0,68	0,61	0,43	0	0,48	
2121	GCA_000317575.1_ASM31757v1	<i>Stanieria cyanosphaera</i> PCC 7437	0	0	0	0	0	0	0
7927	GCA_002355455.1_ASM235545v1	<i>Stanieria</i> sp NIES3757	0	0	0	0	0	0	0
1811	GCA_000265295.1_ASM26529v1	<i>Sulfurospirillum barnesii</i> SES3	0,9	0,9	0,81	0,77	0	0,67	
5759	GCA_001723605.1_ASM172360v1	<i>Sulfurospirillum halorespirans</i> DSM 13726	0,9	0,9	0,8	0,76	0	0,68	
2687	GCA_000568815.1_ASM56881v1	<i>Sulfurospirillum multivorans</i> DSM 12446	0,9	0,89	0,8	0,76	0	0,68	
7856	GCA_002309535.1_ASM230953v1	<i>Sulfurospirillum</i> sp JPD1	0,9	0,89	0,82	0,75	0,33	0,69	
7249	GCA_002162315.1_ASM216231v1	<i>Sulfurospirillum</i> sp SL21	0,9	0,89	0,81	0,65	0,38	0,69	
299	GCA_000013225.1_ASM1322v1	<i>Synechococcus</i> sp JA23Ba213	1	1	1	1	1	1	1
298	GCA_000013205.1_ASM1320v1	<i>Synechococcus</i> sp JA33Ab	1	0,99	1	0,98	1	1	1
2108	GCA_000316685.1_ASM31668v1	<i>Synechococcus</i> sp PCC 6312	0	0	0	0	0	0	0
383	GCA_000014965.1_ASM1496v1	<i>Syntrophobacter fumaroxidans</i> MPOB	0,9	0,73	0,69	0,75	0,54	0	
1296	GCA_000190635.1_ASM19063v1	<i>Syntrophobolus glycolicus</i> DSM 8271	0,8	0,8	0,63	0,73	0,42	0,6	
991	GCA_000092405.1_ASM9240v1	<i>Syntrophothermus lipocalidus</i> DSM 12680	0,9	0,56	0,36	0	0	0,5	
738	GCA_000023025.1_ASM2302v1	<i>Teredinibacter turnerae</i> T7901	1	1	1	1	1	1	1
2062	GCA_000305935.1_ASM30593v1	<i>Thermacetogenium phaeum</i> DSM 12270	0,9	0,33	0,37	0,4	0	0,5	
1017	GCA_000092945.1_ASM9294v1	<i>Thermincola potens</i> JR	0,9	0,82	0,82	0,79	0,72	0,38	
2083	GCA_000307585.2_ASM30758v2	<i>Thermoanaerobacterium saccharolyticum</i> JW5L YS485	0,8	0,58	0,45	0,63	0,42	0,54	
7691	GCA_002250075.1_ASM225007v1	<i>Thermoanaerobacterium thermosaccharolyticum</i>	0,8	0,58	0,33	0,7	0,28	0,45	
1058	GCA_000145615.1_ASM14561v1	<i>Thermoanaerobacterium thermosaccharolyticum</i> DSM 571	0,8	0,58	0,32	0,63	0,28	0,44	
2168	GCA_000328545.1_ASM32854v1	<i>Thermoanaerobacterium thermosaccharolyticum</i> M0795	0,8	0,76	0,45	0,7	0,42	0,55	
851	GCA_000025605.1_ASM2560v1	<i>Thermocrinis albus</i> DSM 14484	1	1	0	1	1	1	1
1605	GCA_000227745.3_ASM22774v3	<i>Thiocystis violascens</i> DSM 198	1	0,93	0,91	0,88	0,58	0,82	
3409	GCA_000828835.1_ASM82883v1	<i>Thioploca ingraca</i>	0,8	0,87	0,8	0,8	0,63	0,6	
740	GCA_000023065.1_ASM2306v1	<i>Tolomonas auensis</i> DSM 9187	0,9	0,85	0,75	0,79	0,63	0,44	
1489	GCA_000214355.1_ASM21435v1	<i>Treponema azotonutricium</i> ZAS9	0,8	0,68	0,6	0,7	0,49	0,44	

1490	GCA_000214375.1_ASM21437v1	<i>Treponema primitia</i> ZAS2	0,8	0,67	0,62	0,63	0,51	0
349	GCA_000014265.1_ASM1426v1	<i>Trichodesmium erythraeum</i> IMS101	1	1	1	0	1	1
1421	GCA_000204075.1_ASM20407v1	<i>Trichormus variabilis</i> ATCC 29413	1	1	1	1	1	1
7922	GCA_002355335.1_ASM235533v1	<i>Variibacter gojawalensis</i>	0	0,2	0	0	0	0
5408	GCA_001680045.1_ASM168004v1	<i>Vibrio natriegens</i>	1	1	1	0,99	0,99	0,99
5409	GCA_001680065.1_ASM168006v1	<i>Vibrio natriegens</i>	1	1	1	0,99	0,99	0,99
5410	GCA_001680085.1_ASM168008v1	<i>Vibrio natriegens</i>	1	1	1	0,99	0,99	1
4493	GCA_001456255.1_ASM145625v1	<i>Vibrio natriegens</i> NBRC 15636 ATCC 14048 DSM 759	1	1	1	0,99	0,99	0,99
5407	GCA_001680025.1_ASM168002v1	<i>Vibrio natriegens</i> NBRC 15636 ATCC 14048 DSM 759	1	1	1	0,99	0,99	0,99
1686	GCA_000241385.1_ASM24138v1	<i>Vibrio</i> sp EJY3	1	1	1	1	1	1
7598	GCA_002222655.1_ASM222265v1	<i>Vitreoscilla filiformis</i>	0,9	0,89	0,86	0,85	0,61	0,82
497	GCA_000017645.1_ASM1764v1	<i>Xanthobacter autotrophicus</i> Py2	1	1	1	1	1	1
5471	GCA_001687105.1_ASM168710v1	<i>Yangia</i> sp CCBMM3	0,9	0,89	0,8	0,78	0,76	0,77
9261	GCA_003054515.1_ASM305451v1	<i>Zoogloaceae</i> <i>bacteirum</i> Parf2	0,9	0,93	0,9	0,88	0,8	0,83
7232	GCA_002158845.1_ASM215884v1	<i>Zymomonas mobilis</i> subsp <i>mobilis</i>	0,9	0,8	0,83	0,66	0,54	0,6
9251	GCA_003052025.1_ASM305202v1	<i>Zymomonas mobilis</i> subsp <i>mobilis</i>	0,9	0,8	0,83	0,67	0,54	0,6
9252	GCA_003052045.1_ASM305204v1	<i>Zymomonas mobilis</i> subsp <i>mobilis</i>	0,9	0,8	0,83	0,67	0,54	0,6
1188	GCA_000175255.2_ASM17525v2	<i>Zymomonas mobilis</i> subsp <i>mobilis</i> ATCC 10988	0,9	0,8	0,83	0,66	0,54	0,6
792	GCA_000024245.1_ASM2424v1	<i>Zymomonas mobilis</i> subsp <i>mobilis</i> NCIMB 11163	0,9	0,8	0,83	0,66	0,54	0,6
2701	GCA_000576125.1_ASM57612v1	<i>Zymomonas mobilis</i> subsp <i>mobilis</i> NRRL B12526	0,9	0,8	0,83	0,66	0,54	0,6
2570	GCA_000498655.1_ASM49865v1	<i>Zymomonas mobilis</i> subsp <i>mobilis</i> str CP4 NRRL B14023	0,9	0,8	0,83	0,66	0,54	0,6
2703	GCA_000576165.1_ASM57616v1	<i>Zymomonas mobilis</i> subsp <i>mobilis</i> str CP4 NRRL B14023	0,9	0,8	0,83	0,66	0,47	0,6
23	GCA_000007105.1_ASM710v1	<i>Zymomonas mobilis</i> subsp <i>mobilis</i> ZM4 ATCC 31821	0,9	0,8	0,83	0,67	0,54	0,6
9264	GCA_003054575.1_ASM305457v1	<i>Zymomonas mobilis</i> subsp <i>mobilis</i> ZM4 ATCC 31821	0,9	0,8	0,83	0,67	0,54	0,6

FONTE: O Autor (2022).

LEGENDA: Organismos destacados em **vermelho** não identificamos as Nifs core.

QUADRO 1- LISTA DOS 662 DIAZOTROFOS OBTIDOS PELO RESULTADO DA RNA E VALIDADOS PELA LITERATURA.

<b>Organismo</b>	<b>Código de acesso NCBI</b>	<b>Referência</b>	<b>DOI/ PMID/LINK</b>
<i>Acetobacterium woodii</i> DSM 1030	GCA_000247605.1_ASM24760v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Acidithiobacillus ferrooxidans</i> SS3	GCA_000214095.3_ASM21409v3	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Acidithiobacillus ferrooxidans</i> ATCC 23270	GCA_000021485.1_ASM2148v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Acidithiobacillus ferrooxidans</i> ATCC 53993	GCA_000020825.1_ASM2082v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Alkaliphilus metalliredigens</i> QYMF	GCA_000016985.1_ASM1698v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Alkalitalea saponilacus</i>	GCF_002201795.1_ASM220179v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Allochromatium vinosum</i> DSM 180	GCA_000025485.1_ASM2548v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Anabaena cylindrica</i> (ATCC 27899 / PCC 7122)	GCA_000317695.1_ASM31769v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Anabaena</i> sp 90	GCA_000312705.1_ASM31270v1	Wang et al, 2012	10.1186/1471-2164-13-613
<i>Anabaena</i> sp WA102	GCA_001277295.1_ASM127729v1	Esteves-Ferreira, 2017	10.1590/1678-4685-GMB-2016-0050
<i>Anabaenopsis circularis</i> NIES-21	GCF_002367975.1_ASM236797v1	Fay, 1976	10.1128/AEM.31.3.376-379.1976
<i>Anabaenopsis elenkinii</i> CC/IB3563	GCA_015245355.1_ASM1524535v1	Delbaje, 2021	10.1099/ijsem.0.004648
<i>Anaeromyxobacter</i> sp. Fw109-5	GCF_000017505.1_ASM1750v1	Masuda, 2020	10.1128/AEM.00956-20
<i>Anaeromyxobacter</i> sp. K	GCF_000020805.1_ASM2080v1	Masuda, 2020	10.1128/AEM.00956-20
<i>Aphanizomenon flos-aquae</i> DEX188	GCA_017346855.1_ASM1734685v1	Kuznecova, 2020	10.3389/ijmcb.2020.02010
<i>Aquaspirillum</i> sp LM1	GCA_002002905.1_ASM200290v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Arcobacter nitrofigilis</i> DSM 7299	GCA_000092245.1_ASM9224v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Arcobacter</i> sp L	GCA_000284235.1_ASM28423v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Aulosira laxa</i> NIES-50	GCF_002368055.1_ASM236805v1	Kobbia & El-Sharouny, 1985	10.1002/jobm.3620250608
<i>Azoarcus communis</i> strain TSNA42	GCA_003111665.1_ASM311166v1	Zorraquino, 2018	10.1128/MRA.01080-18
<i>Azoarcus communis</i> strain TSPY31	GCA_003111645.1_ASM311164v1	Zorraquino, 2018	10.1128/MRA.01080-18
<i>Azoarcus olearius</i> DQS4	GCA_001682385.1_ASM168238v1	Chen, 2013	10.1099/ijfs.0.050609-0
<i>Azoarcus olearius</i> BH72 (sp BH72 )	GCA_000061505.1_ASM6150v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Azoarcus</i> sp CIB	GCA_001190925.1_ASM119092v1	Fernández, 2014	10.1371/journal.pone.0110771
<i>Azoarcus</i> sp KH32C	GCA_000349945.1_ASM34994v1	Nishizawa, 2011	10.1128/JB.06618-11
<i>Azorhizobium caulinodans</i> ORS 571	GCA_000010525.1_ASM1052v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Azospira oryzae</i> PS	GCF_000236665.1_ASM23666v1	Bae, 2007	10.1099/ijfs.0.64965-0
<i>Azospira stricta</i> strain DSM 18626	GCA_016858125.1_ASM1685812v1	Bae, 2007	10.1099/ijfs.0.64965-0

<i>Azospira</i> sp 109	GCF_009177135.1_ASM917713v1	Raitz, 2021	10.3390/genes12010071
<i>Azospirillum brasilense</i> Az39	GCA_000632475.2_ASM63247v2	Young, 1992	10.10071-4020-3054-1_14
<i>Azospirillum brasilense</i> sp7	GCA_001315015.1_ASM131501v1	Young, 1992	10.10071-4020-3054-1_14
<i>Azospirillum humicireducens</i> Strain SgZ-5	GCA_001639105.2_ASM163910v2	Zhou, 2013	10.1099/ijis.0.046813-0
<i>Azospirillum lipoferum</i> 4B	GCA_000283655.1_ASM28365v1	Young, 1992	10.10071-4020-3054-1_14
<i>Azospirillum</i> sp B510	GCA_000010725.1_ASM1072v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Azospirillum</i> sp CFH 70021	GCA_003130795.1_ASM313079v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Azospirillum</i> sp M2T2B2	GCA_003233655.1_ASM323365v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Azospirillum</i> sp TSH58	GCA_003119115.1_ASM311911v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Azospirillum thiophilum</i> BV-S	GCA_001305595.1_ASM130559v1	Steenhoudt, 2000	10.1111/ij.1574-6976.2000.tb00552.x
<i>Azotobacter chroococcum</i> B-3	GCA_002220155.1_ASM222015v1	Robson, 1986	10.1038/322388a0
<i>Azotobacter chroococcum</i> NC1MB 8003	GCA_000817975.1_ASM81797v1	Robson, 1986	10.1038/322388a0
<i>Azotobacter vinelandii</i> CA	GCA_000380335.1_ASM38033v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Azotobacter vinelandii</i> CA6	GCA_000380365.1_ASM38036v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Azotobacter vinelandii</i> DJ	GCA_000021045.1_ASM2104v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Beggiatoa leptomitiformis</i> D-401	GCA_002849755.1_ASM284975v1	Dubiniha, 2017	10.1099/ijsem.0.001584#ab2
<i>Beggiatoa leptomitiformis</i> D-402	GCA_001305575.2_ASM130557v2	Dubiniha, 2017	10.1099/ijsem.0.001584#ab2
<i>Beijerinckia indica</i> subsp <i>indica</i> ATCC 9039	GCA_000019845.1_ASM1984v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Blastochloris (Rhodopseudomonas viridis) viridis</i> ATCC 19567	GCA_001459775.1_Blastochloris_viridis_genome	Howard, 1983	0021-9193/63/070107-06802.000
<i>Blastochloris (Rhodopseudomonas) viridis</i> 1	GCA_001402875.1_ASM140287v1	Howard, 1983	0021-9193/63/070107-06802.000
<i>Blastochloris viridis</i> DSM 133	GCA_001548155.2_BV133_assembly_1.0	Howard, 1983	0021-9193/63/070107-06802.000
<i>Bradyrhizobium amphicarpaeeae</i> 39S1MB	GCF_002266435.2_ASM226643v2	Bromfield, 2019	10.1099/ijsem.0.003569
<i>Bradyrhizobium arachidis</i> CCBAU 051107	GCF_015291705.1_ASM1529170v1	Wang, 2012	10.1016/ij.syapm.2012.10.009
<i>Bradyrhizobium diazoefficiens</i> NK6	GCA_001549695.1_ASM154969v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Bradyrhizobium diazoefficiens</i> USDA 110	GCA_000011365.1_ASM1136v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Bradyrhizobium diazoefficiens</i> USDA 110 2016	GCA_001642675.1_ASM164267v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Bradyrhizobium diazoefficiens</i> USDA 122	GCA_001908315.1_ASM190831v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Bradyrhizobium diazoefficiens</i> XF7	GCA_003183845.1_ASM318384v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Bradyrhizobium elkanii</i> USDA 61	GCF_012871055.1_ASM1287105v1	Souza, 2012	10.1186/1471-2164-13-162
<i>Bradyrhizobium guangdongense</i> strain CCBAU 51649	GCA_004114975.1_ASM411497v1	Li, 2015	10.1128/JB.00563-12
<i>Bradyrhizobium guangxiense</i> strain CCBAU 53363	GCA_004114915.1_ASM411491v1	Li, 2015	10.1099/ijsem.0.000629
<i>Bradyrhizobium guangzhouense</i> CCBAU 51670	GCF_004114955.1_ASM411495v1	Li, 2019	10.1099/ijsem.0.000629
<i>Bradyrhizobium icense</i> LMTR 13	GCA_001693385.1_ASM169338v1	Durán, 2014	10.1016/ij.syapm.2019.1.26002
<i>Bradyrhizobium japonicum</i> E109	GCA_000807315.1_ASM80731v1	Dos Santos, 2012	10.1099/ijis.0.060426-0
<i>Bradyrhizobium japonicum</i> J5	GCA_001887695.1_ASM188769v1	Dos Santos, 2012	10.1186/1471-2164-13-162

<i>Bradyrhizobium japonicum</i> USDA 6	GCA_000284375.1_ASM28437v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Bradyrhizobium oligotrophicum</i> S58	GCA_000344805.1_ASM34480v1	Okubo, 2013	10.1128/AEM.00009-13
<i>Bradyrhizobium ottawaense</i> OO99 [LMG 26739 / HAMB1 3284]	GCA_002278135.2_ASM227813v2	Yu, 2014	10.1099/ijis.0.065540-0
<i>Bradyrhizobium</i> sp 2 39S1MB	GCA_002266435.2_ASM226643v2	Bromfield, 2019	10.1099/ijsem.0.003569
<i>Bradyrhizobium</i> sp CCGELA001	GCA_000296215.2_ASM29621v2	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Bradyrhizobium</i> sp ORS 278	GCA_000026145.1_ASM2614v1	Lorquin, 1993	10.1007/978-94-017-2416-6_57
<i>Bradyrhizobium</i> sp ORS 285	GCA_900176205.1_BRAD285_PRJEB20226_1_v1	Lorquin, 1993	10.1007/978-94-017-2416-6_57
<i>Bradyrhizobium</i> sp ORS 3257	GCA_900324035.1_BRAD3257	Lorquin, 1993	10.1007/978-94-017-2416-6_57
<i>Bradyrhizobium</i> sp S23321	GCA_000284275.1_ASM28427v1	Lorquin, 1993	10.1007/978-94-017-2416-6_57
<i>Bradyrhizobium</i> sp. (strain BTAi1 / ATCC BAA-1182)	GCA_000015165.1_ASM1516v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Bradyrhizobium vignae</i> ORS3257	GCF_900324035.1_BRAD3257	Gronemeyer, 2016	10.1099/ijsem.0.000674
<i>Bradyrhizobium zhanjiangense</i> CCBAU 51778	GCF_004114935.1_ASM411493v1	Li, 2019	10.1016/ij.syapm.2019.126002
<i>Brasilonema octagenarum</i> UFV-E1	GCF_006968915.1_ASM696891v1	Alvarenga, 2020	10.7717/peerj.9158
<i>Brasilonema sennae</i> CENA114	GCF_006968745.1_ASM696874v1	Alvarenga, 2020	10.7717/peerj.9158
<i>Brenneria goodwinii</i> FRB141	GCF_002291445.1_ASM229144v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Burkholderia</i> sp CCGE1001	GCA_000176935.3_ASM17693v3	Ormeño-Orrillo, 2012	10.1128/JB.01756-12
<i>Burkholderia</i> sp CCGE1002	GCA_000092885.1_ASM9288v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Burkholderia</i> sp KJ006	GCA_000262695.1_ASM26269v1	Cho, 2007	10.1111/ij.1574-6941.2007.00280.x
<i>Burkholderia ubonensis</i> MSMB0783	GCA_001718695.1_ASM171869v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Burkholderia vietnamiensis</i> AU1233	GCA_001718775.1_ASM171877v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Burkholderia vietnamiensis</i> FDAARGOS_239	GCA_003019945.1_ASM301994v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Burkholderia vietnamiensis</i> FL-2-3-30-S1-D0	GCA_001718495.1_ASM171849v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Burkholderia vietnamiensis</i> G4	GCA_000016205.1_ASM1620v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Burkholderia vietnamiensis</i> HI2297	GCA_001718815.1_ASM171881v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Burkholderia vietnamiensis</i> LMG 10929	GCA_000959445.1_ASM95944v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Burkholderia vietnamiensis</i> MSMB608WGS	GCA_001718915.1_ASM171891v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Calditerrivibrio nitroreducens</i> DSM 19672	GCF_000183405.1_ASM18340v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Calothrix brevissima</i> NIES-22	GCA_002367995.1_ASM236799v1	Bell-Doyon, 2019	10.1002/edn3.66
<i>Calothrix</i> sp 336/3	GCA_000734895.2_ASM73489v2	Allahverdiyeva, 2009	10.1016/ij.ijhydene.2009.12.030
<i>Calothrix</i> sp NIES2098	GCA_002368175.1_ASM236817v1	Hartman, 2010	10.1007/s00239-010-9365-8
<i>Calothrix</i> sp NIES3974	GCA_002368395.1_ASM236839v1	Hartman, 2010	10.1007/s00239-010-9365-8
<i>Calothrix</i> sp NIES4071	GCA_002368455.1_ASM236845v1	Hartman, 2010	10.1007/s00239-010-9365-8
<i>Calothrix</i> sp NIES-4101	GCA_004296455.1_ASM429645v1	Hartman, 2010	10.1007/1-4020-3054-1_14
<i>Calothrix</i> sp PCC 6303	GCA_000317435.1_ASM31743v1	Singh, 2018	10.1007/s12298-018-0504-9



<i>Calothrix</i> sp PCC 7507	GCA_000316575.1_ASM31657v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Candidatus Atelocyanobacterium thalassa</i> isolate ALOHA [UCYN-A]	GCA_000025125.1_ASM2512v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Candidatus Azobacteroides pseudotriconymphae</i>	GCF_002091395.1_ASM209139v1	Boyd, 2011	10.3389/fmicb.2011.00207
<i>Candidatus Thiodictyon syntrophicum</i>	GCA_002813775.1_ASM281377v1	Esteves-Ferreira, 2017	10.1590/1678-4685-GMB-2016-0050
<b><i>Carboxydocella thermautotrophica</i> 019</b>	<b>GCA_003047205.1_ASM304720v1</b>	<b>INCONCLUSIVO</b>	-
<b><i>Carboxydocella thermautotrophica</i> strain 41</b>	<b>GCA_003054495.1_ASM305449v1</b>	<b>INCONCLUSIVO</b>	-
<i>Chlorobaculum</i> [Chlorobium] tepidum TLS	GCA_000006985.1_ASM698v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Chlorobaculum limnaeum</i> DSM 1677	GCA_001747405.1_ASM174740v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Chlorobaculum parvum</i> NCIB 8327	GCA_000020505.1_ASM2050v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Chlorobium chlorochromatii</i> CaD3	GCA_000012585.1_ASM1258v1	Cerqueda-García, 2013	10.1038/ismej.2013.207
<i>Chlorobium limicola</i> (strain DSM 245 / NBRC 103803 / 6330)	GCA_000020465.1_ASM2046v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Chlorobium phaeobacteroides</i> BS1	GCA_000020545.1_ASM2054v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Chlorobium phaeobacteroides</i> DSM 266	GCA_000015125.1_ASM1512v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Chlorobium phaeovibrioides</i> DSM 265	GCA_000016085.1_ASM1608v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Chloroherpeton thalassium</i> ATCC 35110	GCA_000020525.1_ASM2052v1	Hartman, 2010	10.1007/s00239-010-9365-8
<i>Chroococcidiopsis thermalis</i> PCC 7203	GCA_000317125.1_ASM31712v1	Almon, 1988	10.1111/j.1574-6968.1988.tb02773.x
<i>Clostridium acetatum</i> DMS 1496	GCA_001042715.1_ASM104271v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Clostridium acetobutylicum</i> ATCC 824	GCA_000008765.1_ASM876v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Clostridium acetobutylicum</i> DSM 1731	GCA_000218855.1_ASM21885v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Clostridium acetobutylicum</i> EA 2018	GCA_000191905.1_ASM19190v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Clostridium autoethanogenum</i> DSM 10061	GCA_000484505.1_ASM48450v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Clostridium autoethanogenum</i> DSM 10061-2	GCA_001484725.1_ASM148472v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Clostridium beijerinckii</i> ATCC 35702	GCA_000767745.1_ASM76774v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Clostridium beijerinckii</i> BAS/B3//124	GCA_002003345.1_ASM200334v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Clostridium beijerinckii</i> DSM 6423	GCA_900010805.1_C._beijerinckii_DSM_6423	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Clostridium beijerinckii</i> NCIMB 14988	GCA_000833105.2_ASM83310v2	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Clostridium beijerinckii</i> NCIMB 8052	GCA_000016965.1_ASM1696v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Clostridium beijerinckii</i> NRRL B-598	GCA_000506785.4_ASM50678v4	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Clostridium bornimense</i> M2-40	GCA_000577895.1_M2_40	Koirala e Brözel, 2021	10.1186/1471-2164-13-162
<i>Clostridium butyricum</i> 29-1	GCA_005145065.1_ASM514506v1	Young, 1992	10.3390/microorganisms9081662
<i>Clostridium butyricum</i> 4-1	GCA_005145085.1_ASM514508v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Clostridium butyricum</i> CDC_51208	GCA_001886875.1_ASM188687v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Clostridium butyricum</i> JKY6D1	GCA_001465175.1_ASM146517v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Clostridium butyricum</i> KNU-L09	GCA_001456065.2_ASM145606v2	Young, 1992	10.1007/1-4020-3054-1_14
<i>Clostridium butyricum</i> TK520	GCA_001746555.1_ASM174655v1	Young, 1992	10.1007/1-4020-3054-1_14

<i>Clostridium butyricum</i> TOA	GCA_001646605.1_ASM164660v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Clostridium cellulovorans</i> 743B	GCA_000145275.1_ASM14527v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Clostridium clariflavum</i> DSM 19732	GCA_000237085.1_ASM23708v1	Li, 2016	10.1038/srep19015
<i>Clostridium estertheticum</i> subsp <i>estertheticum</i> DSM 8809	GCA_001877035.1_ASM187703v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Clostridium formicaceticum</i> ATCC 27076	GCA_001854185.1_ASM185418v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Clostridium formicaceticum</i> DSM 92	GCA_002080475.1_ASM208047v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Clostridium kluveri</i> DSM 555	GCA_000016505.1_ASM1650v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Clostridium kluveri</i> JZZ	GCA_001902295.1_ASM190229v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Clostridium kluveri</i> NBRC 12016	GCA_000010265.1_ASM1026v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<b><i>Clostridium lentocellum</i> DSM 5427</b>	<b>GCA_000178835.2_ASM17883v2</b>	<b>INCONCLUSIVO</b>	-
<i>Clostridium ljungdahlii</i> DSM 13528	GCA_000143685.1_ASM14368v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Clostridium pasteurianum</i> BC1	GCA_000389635.1_ASM38963v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Clostridium pasteurianum</i> DSM 525 = ATCC 6013 2015	GCA_000807255.1_ASM80725v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Clostridium pasteurianum</i> DSM 525 = ATCC 6013 2015-2	GCA_000807175.1_ASM80717v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Clostridium pasteurianum</i> DSM 525 = ATCC 6013 2016	GCA_001856645.1_ASM185664v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Clostridium pasteurianum</i> M150B	GCA_001856695.1_ASM185669v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Clostridium saccharoperbutylacetonicum</i> N1-4[HMT]	GCA_000340885.1_ASM34088v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Clostridium saccharoperbutylacetonicum</i> N1-504	GCA_002003305.1_ASM200330v1	Young, 1992	10.1007/1-4020-3054-1_14
<b><i>Clostridium</i> sp BNL 1100</b>	<b>GCA_000244875.1_ASM24487v1</b>	<b>INCONCLUSIVO</b>	-
<i>Clostridium tyrobutyricum</i> KCTC 5387	GCA_001642655.1_ASM164265v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Clostridium tyrobutyricum</i> W428	GCA_001679705.1_ASM167970v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Confluentimicrobium</i> sp <i>EMB200-NS6</i>	GCA_001305615.1_ASM130561v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Coraliomargarita akajimensis</i> DSM 45221	GCA_000025905.1_ASM2590v1	Mavromatis, 2010	10.4056/sigs.952166
<i>Crocospaera subtropica</i> ATCC 51142	GCF_000017845.1_ASM1784v1	Saito, 2011	10.1073/pnas.1006943108
<i>Cupriavidus taiwanensis</i> str LMG 19424	GCA_000069785.1_ASM6978v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Cyanobacterium endosymbiont of Epithemia turgida</i> isolate <i>EtSB Lake Yunoko</i>	GCA_000829235.1_ASM82923v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Cyanobacterium endosymbiont of Rhopalodia gibberula</i>	GCA_003574135.1_ASM357413v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Cyanothece</i> sp (strain <i>PCC 7425</i> / ATCC 29141)	GCA_000022045.1_ASM2204v1	Bandyopadhyay, 2011	10.1128/mBio.00214-11.
<i>Cyanothece</i> sp ATCC 51142	GCA_000017845.1_ASM1784v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Cyanothece</i> sp <i>PCC 8801</i>	GCA_000021805.1_ASM2180v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Cyanothece</i> sp <i>PCC 8802</i>	GCA_000024045.1_ASM2404v1	Bandyopadhyay, 2011	10.1128/mBio.00214-11.
<i>Cylindrospermum stagnale</i> <i>PCC 7417</i>	GCA_000317535.1_ASM31753v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Dechloromonas aromatica</i> <i>RCB</i>	GCA_000012425.1_ASM1242v1	Salinero, 2009	10.1186/1471-2164-10-351
<i>Dechloromonas</i> sp. <i>HYN0024</i>	GCF_003441615.1_ASM344161v1	Raittz, 2021	10.3390/genes12010071

<b>Defluviococcus sp. isolate SSA4</b>	GCA_013414705.1_ASM1341470v1	INCONCLUSIVO	
<b>Defluviococcus vanus strain Ben 114</b>	GCF_014672695.1_ASM1467269v1	INCONCLUSIVO	
<i>Defluviimonas alba</i> cai42	GCA_001620265.1_ASM162026v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Dehalobacter</i> sp CF	GCA_000305815.1_ASM30581v1	Ferrando, 2015	10.1093/femsec/fiv104
<i>Dehalobacter</i> sp DCA	GCA_000305775.1_ASM30577v1	Ferrando, 2015	10.1093/femsec/fiv104
<i>Dehalococcoides ethenogenes</i> mccartyi 195	GCA_000011905.1_ASM1190v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Dehalococcoides ethenogenes</i> mccartyi CG4	GCA_000830905.1_ASM83090v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<b>Dehalogenimonas sp WBC-2</b>	GCA_001005265.1_ASM100526v1	INCONCLUSIVO	
<i>Denitrovibrio acetiphilus</i> DSM 12809	GCA_000025725.1_ASM2572v1	Petersen, 2016	10.1038/nmicrobiol.2016.195
<i>Desulfatibacillum aliphaticivorans</i> AK-01	GCA_000021905.1_ASM2190v1	Bahulikar, 2014	10.1128/AEM.02091-14
<i>Desulfotobacterium hafniense</i> DCB-2	GCA_000021925.1_ASM2192v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Desulfotobacterium hafniense</i> Y51	GCA_000010045.1_ASM1004v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Desulfobacca acetoxidans</i> DSM 11109	GCA_000195295.1_ASM19529v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Desulfobacterium autotrophicum</i> HRM2	GCA_000020365.1_ASM2036v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Desulfobulbus propionicus</i> DSM 2032	GCA_000186885.1_ASM18688v1	Thajudeen, 2017	10.1007/s11274-017-2205-x
<i>Desulfocapsa sulfexigens</i> DSM 10523	GCA_000341395.1_ASM34139v1	Finster, 2013	10.4056/sgs.3777412
<i>Desulfocurvibacter</i> [ <i>Desulfovibrio</i> ] <i>africanus</i> subsp <i>africanus</i> str <i>Walvis Bay</i>	GCA_000212675.2_ASM21267v2	Young, 1992	10.1007/1-4020-3054-1_14
<i>Desulfomicrobium baculatum</i> DSM 4028	GCA_000023225.1_ASM2322v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Desulfomonile tiedjei</i> DSM 6799	GCA_000266945.1_ASM26694v1	Shelton and Tiedje, 1984	0099-2240/84/100840-09\$02.00/0
<b>Desulfosarcina alkanivorans PL12</b>	GCF_009688945.1_ASM968894v1	INCONCLUSIVO	
<b>Desulfosarcina ovata subsp ovata oXyS1</b>	GCF_009689005.1_ASM968900v1	INCONCLUSIVO	
<i>Desulfosporosinus acidiphilus</i> SJ4	GCA_000255115.3_ASM25511v3	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Desulfosporosinus meridiei</i> DSM 13257	GCA_000231385.3_ASM23138v3	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Desulfosporosinus orientis</i> DSM 765	GCA_000235605.1_ASM23560v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Desulfotomaculum acetoxidans</i> DSM 771	GCA_000024205.1_ASM2420v1	Spring, 2009	10.4056/sgs.39508
<i>Desulfotomaculum gibsoniae</i> DSM 7213	GCA_000233715.3_ASM23371v3	Kuever, 2014	10.4056/sgs.5209235
<i>Desulfotomaculum nigrificans</i> CO-1-SRB	GCA_000214435.1_ASM21443v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Desulfotomaculum reducens</i> MI-1	GCA_000016165.1_ASM1616v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Desulfotomaculum ruminis</i> DSM 2154	GCA_000215085.1_ASM21508v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Desulfovibrio desulfuricans</i> ND132	GCA_000189295.2_ASM18929v2	Young, 1992	10.1007/1-4020-3054-1_14
<i>Desulfovibrio gigas</i> DSM 1382 = ATCC 19364	GCA_000468495.1_ASM46849v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Desulfovibrio magneticus</i> RS-1	GCA_000010665.1_ASM1066v1	Jing, 2015	10.3389/fmicb.2015.011172
<i>Desulfovibrio salexigens</i> DSM 2638	GCA_000023445.1_ASM2344v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Desulfovibrio vulgans</i> DP4	GCA_000015485.1_ASM1548v1	Dos Santos, 2012	10.1186/1471-2164-13-162

<i>Desulfiovibrio vulgaris</i> RCH1	GCA_000166115.1_ASM16611v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Desulfiovibrio vulgaris</i> str Hildenborough	GCA_000195755.1_ASM19575v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Desulfiovibrio vulgaris</i> str Miyazaki F	GCA_000021385.1_ASM2138v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<b><i>Desulfurispirillum indicum</i> S5</b>	<b>GCF_000177635.2_ASM17763v2</b>	<b>INCONCLUSIVO</b>	-
<i>Desulfurivibrio alkaliphilus</i> AHT 2	GCA_000092205.1_ASM9220v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Desulfuromonas soudanensis</i> WTL	GCA_001278055.1_ASM127805v1	Badalamenti, 2016	10.3389/fmicb.2016.00913
<b><i>Desulfuromonas</i> sp DDH964</b>	<b>GCA_001611275.1_ASM161127v1</b>	<b>INCONCLUSIVO</b>	-
<i>Dickeya dadantii</i> 3937	GCA_000147055.1_ASM14705v1	Dos Santos, 2020	10.1002/cbic.202000022
<i>Dickeya dadantii</i> strain DSM 18020	GCF_003049785.1_ASM304978v1	Dos Santos, 2020	10.1002/cbic.202000022
<i>Dickeya dianthicola</i> RNS049	GCA_002706485.1_ASM270648v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Dickeya fangzhongdai</i> ND14b	GCA_000758345.1_ASM75834v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Dickeya paradisiaca</i> Ech703	GCA_000023545.1_ASM2354v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Dickeya solani</i> D s0432-1	GCA_002846975.1_ASM284697v1	Golanowska, 2015	http://theses.insa-lyon.fr/publication/2015ISAL0087/these.pdf
<i>Dickeya solani</i> IPO 2222	GCA_001644705.1_ASM164470v1	Golanowska, 2015	http://theses.insa-lyon.fr/publication/2015ISAL0087/these.pdf
<i>Dickeya solani</i> PPO 9019	GCA_002846995.1_ASM284699v1	Golanowska, 2015	http://theses.insa-lyon.fr/publication/2015ISAL0087/these.pdf
<i>Dickeya solani</i> RNS 08_23_3_1_A	GCA_000511285.2_ASM51128v2	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<b><i>Dissulfurispira thermophila</i> T55J</b>	<b>GCA_014701235.1_ASM1470123v1</b>	<b>INCONCLUSIVO</b>	-
<b><i>Dolichospermum compactum</i> NIES-806</b>	<b>GCF_002368115.1_ASM236811v1</b>	<b>INCONCLUSIVO</b>	-
<i>Dolichospermum flos-aquae</i> CCAP 1403 F13	GCF_012516395.1_ASM1251639v1	Yema, 2016	10.1016/j.hal.2016.11.007
<i>Dolichospermum</i> sp. UKL201	GCA_017346795.1_ASM1734679v1	Eigeman, 2019	10.1371/journal.pone.0223294
<b><i>Draconibacterium halophilum</i> [ sp. M1]</b>	<b>GCF_010448835.1_ASM1044883v1</b>	<b>INCONCLUSIVO</b>	-
<i>Dysgonomonas capnocytophagoides</i> DSM 22835	GCF_000426485.1_ASM42648v1	Inoue, 2015	10.1264/j.sme2.ME14142
<i>Ectothiorhodospira</i> sp BSL-9	GCA_001632845.1_ASM163284v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Ensifer sojae</i> CCBAU 05684	GCA_002288525.1_ASM228852v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Enterobacter</i> sp FY07	GCA_001582075.1_ASM158207v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Enterobacter</i> sp R4-368	GCA_000410515.1_ASM41051v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Enterobacteriaceae bacterium</i> ENNIH1	GCA_002903065.1_ASM290306v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Ethanoligenens harbinense</i> W1	GCF_005222505.1_ASM522250v1	Dos Santos, 2020	10.1002/cbic.202000022
<i>Euhalothece natronophila</i> Z M001	GCF_007904085.1_ASM790408v1	Fernández-Juárez, 2019	10.3389/fmicb.2019.01903
<i>Fischerella</i> sp NIES3754	GCA_001548455.1_ASM154845v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Frankia alhi</i> ACN14a	GCA_000058485.1_ASM5848v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Frankia casuarinae</i> strain Ccl3	GCA_000013345.1_ASM1334v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Frankia</i> sp EAN1pec	GCA_000018005.1_ASM1800v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Frankia symbiont of Datisca glomerata</i>	GCA_000177615.2_ASM17761v2	Young, 1992	10.1007/1-4020-3054-1_14
<i>Fremyella [Microchaete] diplosiphon</i> NIES3275	GCF_002368275.1_ASM236827v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662

<i>Geobacter bemidjensis</i> Bem	GCA_000020725.1_ASM2072v1	Yun, 2011	10.1128/AEM.00114-11
<i>Geobacter daltonii</i> FRC32	GCA_000022265.1_ASM2226v1	Yun, 2011	10.1128/AEM.00114-11
<i>Geobacter lovleyi</i> SZ	GCA_000020385.1_ASM2038v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Geobacter metallireducens</i> GS-15	GCA_000012925.1_ASM1292v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<b><i>Geobacter pickeringii</i> G13</b>	<b>GCA_000817955.1_ASM81795v1</b>	<b>INCONCLUSIVO</b>	-
<b><i>Geobacter</i> sp M18</b>	<b>GCA_000175115.2_ASM17511v2</b>	<b>INCONCLUSIVO</b>	-
<b><i>Geobacter</i> sp M21</b>	<b>GCA_000023645.1_ASM2364v1</b>	<b>INCONCLUSIVO</b>	-
<i>Geobacter sulfurreducens</i> KN400	GCA_000210155.1_ASM21015v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Geobacter sulfurreducens</i> PCA	GCA_000007985.2_ASM798v2	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Geobacter uraniireducens</i> Rf4	GCA_000016745.1_ASM1674v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Geosporobacter ferrireducens</i> IRF9	GCA_001750685.1_ASM175068v1	Jung, 2017	10.1007/s12275-018-7451-6
<i>Gloeothece citriformis</i> PCC 7424	GCF_000021825.1_ASM2182v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Gloeothece verrucosa</i> PCC 7822	GCF_000147335.1_ASM14733v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Gluconacetobacter diazotrophicus</i> PA1 5	GCA_000067045.1_ASM6704v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Gluconacetobacter diazotrophicus</i> PA1 5 2014	GCA_000021325.1_ASM2132v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Gyvuella sunshinyii</i> YC6258	GCA_0000940805.1_ASM94080v1	König, 2016	10.1038/nmicrobiol.2016.193
<i>Halorhodospira halochloris</i> DSM 1059	GCA_002356555.2_HH1059_assembly_1.0	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Halorhodospira halophila</i> (strain DSM 244 / SL1)	GCA_000015585.1_ASM1558v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Halothece</i> sp PCC 7418	GCA_000317635.1_ASM31763v1	Fernández-Juárez, 2019	10.3389/fmicb.2019.01903
<b><i>Halothece</i> sp PCC 7418</b>	<b>GCA_001663815.1_ASM166381v1</b>	<b>INCONCLUSIVO</b>	-
<i>Hartmannibacter diazotrophicus</i> E19T	GCA_900231165.1_HDIA1	Suarez, 2014	10.1099/jls.0.064154-0
<i>Hellobacterium modesticaldum</i> (strain ATCC 51547 / Ice1)	GCA_000019165.1_ASM1916v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Herbaspirillum rubrisubalbicans</i> M1	GCA_001483945.1_ASM148394v1	Serrato, 2010	10.1139/w10-011
<i>Herbaspirillum seropedicae</i> SmR1	GCA_000143225.1_ASM14322v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Herbaspirillum seropedicae</i> Z67	GCA_001040945.1_ASM104094v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Hydrogenobacter thermophilus</i> TK-6 2010	GCA_000010785.1_ASM1078v1	Nishihara, 2018	10.1264/jseme2.ME18041
<i>Hydrogenobacter thermophilus</i> TK-6 2011	GCA_000164905.1_ASM16490v1	Nishihara, 2018	10.1264/jseme2.ME18041
<i>Hyphomicrobium</i> sp MC1	GCA_000253295.1_ASM25329v1	Fesefeldt, 1998	10.1139/w97-139
<i>Ilyobacter polytropus</i> DSM 2926	GCA_000165505.1_ASM16550v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Klebsiella michiganensis</i> AR375	GCA_003074075.1_ASM307407v1	Dantur, 2018	10.3389/fmicb.2018.01548
<i>Klebsiella michiganensis</i> E718	GCA_000276705.2_ASM27670v2	Dantur, 2018	10.3389/fmicb.2018.01548
<i>Klebsiella michiganensis</i> HKOPL1	GCA_000632415.1_ASM63241v1	Dantur, 2018	10.3389/fmicb.2018.01548
<i>Klebsiella michiganensis</i> K516	GCA_002216835.1_ASM221683v1	Dantur, 2018	10.3389/fmicb.2018.01548
<i>Klebsiella michiganensis</i> K518	GCA_002290285.1_ASM229028v1	Dantur, 2018	10.3389/fmicb.2018.01548
<i>Klebsiella michiganensis</i> KCTC 1686	GCA_000240325.1_ASM24032v1	Dantur, 2018	10.3389/fmicb.2018.01548

<i>Klebsiella michiganensis</i> M1	GCA_000724525.1_ASM72452v1	Dantur, 2018	10.3389/fmicb.2018.01548
<i>Klebsiella oxytoca</i> AR_0028	GCA_002947505.1_ASM294750v1	Dantur, 2018	10.1007/BF02277362
<i>Klebsiella oxytoca</i> AR_0147	GCA_002072655.1_ASM207265v1	Dantur, 2018	10.1007/BF02277362
<i>Klebsiella oxytoca</i> CAV1374	GCA_001022195.1_ASM102219v1	Cakmakci, 1981	10.1007/BF02277362
<i>Klebsiella oxytoca</i> CAV1752	GCA_001970835.1_ASM197083v1	Cakmakci, 1981	10.1007/BF02277362
<i>Klebsiella oxytoca</i> JK03	GCA_001548355.1_ASM154835v1	Cakmakci, 1981	10.1007/BF02277362
<i>Klebsiella oxytoca</i> KONIH1	GCA_000714655.1_ASM71465v1	Cakmakci, 1981	10.1007/BF02277362
<i>Klebsiella oxytoca</i> KONIH2	GCA_002906415.1_ASM290641v1	Cakmakci, 1981	10.1007/BF02277362
<i>Klebsiella oxytoca</i> KONIH4	GCA_002906395.1_ASM290639v1	Cakmakci, 1981	10.1007/BF02277362
<i>Klebsiella oxytoca</i> KONIH5	GCA_002906435.1_ASM290643v1	Cakmakci, 1981	10.1007/BF02277362
<i>Klebsiella pneumoniae</i> AR_0142	GCA_003073315.1_ASM307331v1	MacNeil, 1981	10.1128/JB.145.1.348-357.1981
<i>Klebsiella pneumoniae</i> YH43	GCA_001548315.1_ASM154831v1	MacNeil, 1981	10.1128/JB.145.1.348-357.1981
<i>Klebsiella quasipneumoniae</i> ATCC 700603	GCA_002953295.1_ASM295329v1	Chen, 2015	10.1002/jobm.201500415
<i>Klebsiella quasipneumoniae</i> CAV1947	GCA_003146685.1_ASM314668v1	Chen, 2015	10.1002/jobm.201500415
<i>Klebsiella quasipneumoniae</i> CAV2013	GCA_003146655.1_ASM314665v1	Chen, 2015	10.1002/jobm.201500415
<i>Klebsiella quasipneumoniae</i> CAV2018	GCA_003146635.1_ASM314663v1	Chen, 2015	10.1002/jobm.201500415
<i>Klebsiella quasipneumoniae</i> subsp <i>quasipneumoniae</i> A708	GCA_003285165.1_ASM328516v1	Chen, 2015	10.1002/jobm.201500415
<i>Klebsiella quasivariicola</i> KPN1705	GCA_002269255.1_ASM226925v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Klebsiella</i> sp L.TGPAF6F	GCA_001753185.1_ASM175318v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Klebsiella</i> sp M5a1 ( <i>oxytoca/pneumoniae</i> M5a1)	GCA_002090195.1_ASM209019v1	Temme, 2012	10.1073/jpnas.1120788109
<i>Klebsiella variicola</i> 13450	GCA_003285185.1_ASM328518v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Klebsiella variicola</i> 342	GCA_000019565.1_ASM1956v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Klebsiella variicola</i> At-22	GCA_000025465.1_ASM2546v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Klebsiella variicola</i> DSM 15968	GCA_000828055.2_ASM82805v2	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Klebsiella variicola</i> DX120E	GCA_000812205.2_ASM81220v2	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Klebsiella variicola</i> E57-7	GCA_003290425.1_ASM329042v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Klebsiella variicola</i> GJ1	GCA_001989495.1_ASM198949v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Klebsiella variicola</i> GJ2	GCA_001989515.1_ASM198951v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Klebsiella variicola</i> GJ3	GCA_001989535.1_ASM198953v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Klebsiella variicola</i> KPN1481 [ <i>pneumoniae</i> KPN1481]	GCA_002156765.1_ASM215676v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Klebsiella variicola</i> USDAARS / <i>Klebsiella pneumoniae</i> subsp <i>pneumoniae</i> KP51	GCA_000714635.1_ASM71463v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Klebsiella variicola</i> WCHKP19	GCA_002164805.2_ASM216480v2	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Klebsiella variicola</i> WCHKV030666	GCA_002852915.3_ASM285291v3	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Kosakonia arachidis</i> KACC 18508 (2020)	GCF_009363135.1_ASM936313v1	Dos Santos, 2020	10.1002/ebic.202000022

<i>Kosakonia oryzae</i> Ola 51	GCA_001658025.1_ASM165802v1	Dos Santos, 2020	10.1002/cbic.202000022
<i>Kosakonia radicinifans</i> DSM 16656 2019	GCF_000280495.2_ASM28049v2	Dos Santos, 2020	10.1002/cbic.202000022
<i>Kosakonia radicinifans</i> GXGL-4A	GCA_001887675.1_ASM188767v1	Dos Santos, 2020	10.1002/cbic.202000022
<i>Kosakonia sacchari</i> BO-1	GCA_001683395.1_ASM168339v1	Shinjo, 2016	10.1128/genomeA.00868-16
<i>Kosakonia sacchari</i> SP1	GCA_000300455.4_ASM30045v4	Zhu, 2012	10.1128/JB.01933-12
<i>Kyripidia spormannii</i> EA1	GCA_002804065.1_ASM280406v1	Reiner, 2018	10.1099/ijsem.0.003037
<i>Leptolyngbya boryana</i> dg5	GCA_002142495.1_ASM214249v1	Dos Santos, 2012	10.1002/cbic.202000022
<i>Leptolyngbya boryana</i> IAM M101	GCA_002142475.1_ASM214247v1	Dos Santos, 2012	10.1002/cbic.202000022
<i>Leptolyngbya boryana</i> NIES2135	GCA_002368255.1_ASM236825v1	Dos Santos, 2012	10.1002/cbic.202000022
<i>Leptolyngbya</i> sp. O-77	GCA_001548395.1_ASM154839v1	Yoon, 2018	10.1264/ismc2.ME17015
<i>Leptospirillum ferriphilum</i> YSK	GCA_000695975.1_ASM69597v1	Liu, 2020	10.1016/S1003-6326(20)65326-2
<i>Leptospirillum ferrooxidans</i> C23	GCA_000284315.1_ASM28431v1	Patro, 2004	10.1016/j.resmic.2004.05.010
<i>Leptothrix cholodnii</i> SP-6	GCA_000019785.1_ASM1978v1	Gupta, 2019	10.3389/fmolb.2019.00115
<i>Magnetococcus marinus</i> MC1	GCA_000014865.1_ASM1486v1	Bazylinski, 2013	10.1099/ijls.0.038927-0
<i>Magnetospirillum gryphiswaldense</i> MSR-1	GCA_002995515.1_ASM299551v1	Jiang, 2002	10.1360/021b9455
<i>Magnetospirillum gryphiswaldense</i> MSR-1 v2	GCA_000513295.1_MGMSRV2	Jiang, 2002	10.1360/021b9455
<i>Magnetospirillum gryphiswaldense</i> R3/S1	GCA_002995535.1_ASM299553v1	Jiang, 2002	10.1360/021b9455
<i>Magnetospirillum magneticum</i> AMB-1	GCA_000009985.1_ASM998v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Magnetospirillum</i> sp ME-1	GCA_002105535.1_ASM210553v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Magnetospirillum</i> sp XM-1	GCA_001511835.1_XM-1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Marichromatium purpuratum</i> 984	GCF_000224005.2_ASM22400v3	Affourtit, 2001	10.1007/s00248000090
<i>Labiliibaculum antarcticum</i> [Marinifilaceae bacterium SPP2]	GCA_002356295.1_ASM235629v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Martellella endophytica</i> YC6887	GCA_000960975.1_ASM96097v1	König, 2016	10.1038/nmicrobiol.2016.193
<i>Mesorhizobium amorphae</i> CCNWGS0123	GCA_001686985.1_ASM168698v1	Wang, 2002	10.1007/s00203-002-0448-9
<i>Mesorhizobium australicum</i> WSM2073	GCA_000230995.3_ASM23099v3	Reeve, 2013	10.4056/sgs.4568282
<i>Mesorhizobium ciceri</i> biovar biserrulae WSM1271	GCA_000185905.1_ASM18590v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Mesorhizobium ciceri</i> biovar biserrulae WSM1284	GCA_001618845.1_ASM161884v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Mesorhizobium ciceri</i> CC1192	GCA_001618825.1_ASM161882v1	Jiang, 2002	10.1186/1471-2164-13-162
<i>Mesorhizobium erdmanii</i> NZP2014	GCF_013170725.1_ASM1317072v1	Perry, 2020	10.1360/021b9455
<i>Mesorhizobium huakuii</i> 7653R	GCF_000709395.1_ASM70939v1	Dos Santos, 2012	10.1099/mgen.0.000426
<i>Mesorhizobium japonicum</i> MAFF 303099	GCF_000009625.1_ASM962v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Mesorhizobium loti</i> NZP2037	GCA_001676765.1_ASM167676v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Mesorhizobium opportunisticum</i> WSM2075	GCA_000176035.2_ASM17603v2	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Mesorhizobium</i> sp WSM1497	GCA_001672455.2_ASM167245v2	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanobacterium congolense</i> isolate Buetzberg	GCA_900095295.1_MCBBB	Young, 1992	10.1007/1-4020-3054-1_14

<i>Methanobacterium formicicum</i> strain BRM9	GCA_000762265.1_ASM76226v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanobacterium lacus</i> AL-21	GCA_000191585.1_ASM19158v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanobacterium paludis</i> strain SWAN1	GCA_000214725.1_ASM21472v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanothermobacter</i> [ <i>Methanobacterium</i> ] sp. BAmetb5	GCA_003491305.1_ASM349130v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<b><i>Methanobacterium</i> sp. BRmetb2</b>	<b>GCA_003491285.1_ASM349128v1</b>	<b>INCONCLUSIVO</b>	-
<i>Methanobacterium</i> sp. MB1	GCF_000499765.1_Methanobacterium	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Methanobacterium</i> sp. MZ-A1	GCF_002813675.1_ASM281367v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Methanobacterium subterraneum</i> strain A8p	GCA_002813695.1_ASM281369v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanocella avoryzae</i> MRE50	GCF_000063445.1_ASM6344v1	Lü, 2012	10.1128/JB.00207-12
<i>Methanocella conradii</i> HZ254	GCF_000251105.1_ASM25110v1	Lü, 2012	10.1128/JB.00207-12
<i>Methanococcus aeolicus</i> Nankai-3	GCA_000017185.1_ASM1718v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanococcus maripaludis</i> C5 uid58741	GCA_000016125.1_ASM1612v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanococcus vanniellii</i> SB	GCA_000017165.1_ASM1716v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanolacinia petrolearia</i> DSM 11571	GCF_000147875.1_ASM14787v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Methanobolus psychrophilus</i> R15	GCA_000306725.1_ASM30672v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanobolus zinderi</i> strain DSM 21339	GCF_013388255.1_ASM1338825v1	Zinder, 2016	10.1186/1471-2164-13-162
<i>Methanoregula boonei</i> 6A8	GCF_000017625.1_ASM1762v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Methanosarcina acetivorans</i> C2A	GCA_000007345.1_ASM734v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanosarcina barkeri</i> Fusaro	GCA_000195895.1_ASM19589v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanosarcina horonobensis</i> HB-1	GCA_000970285.1_ASM97028v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanosarcina mazei</i> Go1	GCA_000007065.1_ASM706v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanosarcina siciliae</i> T4/M	GCA_000970085.1_ASM97008v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanosarcina</i> sp. Kolksee	GCF_000969945.1_ASM96994v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Methanosarcina</i> sp. WH1	GCF_000970005.1_ASM97000v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Methanosarcina</i> sp. WWM596	GCF_000969965.1_ASM96996v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Methanosarcina vacuolata</i> Z-761	GCA_000969905.1_ASM96990v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanosphaerula palustris</i> E1-9c	GCF_000021965.1_ASM2196v1	Boyd, 2011	10.3389/fmicb.2011.00206
<i>Methanothermobacter marburgensis</i> str. Marburg	GCA_000145295.1_ASM14529v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanothermobacter</i> sp. EMTCatA1	GCA_002356395.1_ASM235639v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<b><i>Methanothermobacter</i> sp. KEPCO-1</b>	<b>GCF_008033705.1_ASM803370v1</b>	<b>INCONCLUSIVO</b>	-
<b><i>Methanothermobacter</i> sp. THM-2</b>	<b>GCF_009917665.1_ASM991766v1</b>	<b>INCONCLUSIVO</b>	-
<i>Methanothermobacter thermautotrophicus</i> Delta H	GCA_000008645.1_ASM864v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanothermobacter wolfeii</i> isolate SIV6	GCA_900095815.1_SIV6	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanothermococcus okinawensis</i> IH1	GCA_000179575.2_ASM17957v2	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methanotherx soehngenii</i> GP6	GCF_000204415.1_ASM20441v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662



<i>Methanotorris igneus</i> Kol 5	GCF_000214415.1_ASM21441v1	Nobu, 2016	10.1038/smej.2016.33
<i>Methylacidiphilum fumarolicum</i> SoIV	GCF_000297415.1_ASM29741v1	Khadem, 2010	10.1099/mic.0.036061-0
<i>Methylacidiphilum infernorum</i> V4	GCA_000019665.1_ASM1966v1	Hou, 2008	10.1186/1745-6150-3-26
<i>Methylacidiphilum kamchatkense</i> Kam1	GCF_007475525.1_ASM747552v1	Kruse, 2019	10.1186/s12864-019-5995-4
<b><i>Methylacidiphilum</i> sp IT5</b>	<b>GCF_017310525.1_ASM1731052v1</b>	<b>INCONCLUSIVO</b>	-
<b><i>Methylacidiphilum</i> sp IT6</b>	<b>GCF_017310505.1_ASM1731050v1</b>	<b>INCONCLUSIVO</b>	-
<i>Methylobacterium nodulosum</i> ORS 2060	GCA_000022085.1_ASM2208v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methylobacterium</i> sp. 4-46	GCA_000019365.1_ASM1936v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>M. silvestris</i> (DSM 15510 / CIP 108128 / LMG 27833 / NCIMB 13906 / BL2)	GCA_000021745.1_ASM2174v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methylotella tundrae</i> isolate MTUNDRASET4	GCF_900749825.1_MTUNDRASET4_v1	Haque, 2020	10.1099/mic.0.000977
<i>Methylococcus capsulatus</i> str Bath	GCA_000008325.1_ASM832v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methylocystis bryophila</i> S285	GCA_002117405.1_ASM211740v1	Belova, 2013	10.1099/ijis.0.043505-0
<i>Methylocystis parvus</i> strain BRCS2	GCF_009685195.1_ASM968519v1	Young, 1992	10.1007/1-4020-3064-1_14
<i>Methylocystis</i> sp SC2	GCA_000304315.1_Msp_SC2	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methylomonas denitrificans</i> FJG1	GCA_000785705.2_ASM78570v2	Kiils, 2015	10.1111/1462-2920.12772
<i>Methylomonas koyamae</i> LM6	GCA_002441955.1_ASM244195v1	König, 2016	10.1038/nmicrobiol.2016.193
<i>Methylomonas methanica</i> MC09	GCA_000214665.1_ASM21466v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methylomonas</i> sp DH1	GCA_001644685.1_ASM164468v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Methylomusa anaerophila</i> MMFC1	GCF_003966895.1_ASM396689v1	Dos Santos, 2020	10.1002/cbic.202000022
<i>Methylosinus trichosporium</i> OB3b	GCA_002752655.1_ASM275265v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Methyloversatilis</i> sp RAC08	GCA_001713355.1_ASM171335v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Methylovulum psychrotolerans</i> HV10	GCA_002209385.1_ASM220938v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Microcoleus</i> sp PCC 7113	GCA_000317515.1_ASM31751v1	Affourtit, 2001	10.1007/s002480000090
<i>Neorhizobium galegae</i> bv <i>officinale</i> str HAMB1 1141	GCA_000731295.1_RG1141_Ch_pA_pB	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Neorhizobium galegae</i> bv <i>orientalis</i> str HAMB1 540	GCA_000731315.1_RG540_Ch_pA	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Nitrospirillum amazonense</i> CBAmc	GCA_000730165.2_ASM73016v2	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Nodularia spumigena</i> UHCC 0039	GCA_003054475.1_ASM305447v1	Young, 1992	10.1007/1-4020-3064-1_14
<i>Nostoc azollae</i> 0708	GCA_000196515.1_ASM19651v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Nostoc carneum</i> NIES-2107	GCA_002368155.1_ASM236815v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Nostoc flagelliforme</i> CCNUN1	GCA_002813575.1_ASM281357v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Nostoc linckia</i> NIES-25	GCA_002368035.1_ASM236803v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Nostoc piscinale</i> CENA21	GCA_001298445.1_ASM129844v1	Leão, 2016	10.1128/genomeA.00189-16
<i>Nostoc punctiforme</i> PCC 73102	GCA_000020025.1_ASM2002v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Nostoc</i> sp [ <i>Lobaria pulmonaria</i> 5183 cyanobiont] 5183	GCA_002949795.1_ASM294979v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662

<i>Nostoc</i> sp. [ <i>Peltigera membranacea cyanobiont</i> ] N6	GCA_002949735.1_ASM294973v1	Gagunashvili, 2018	10.1186/s12864-018-4743-5
<b><i>Nostoc</i> sp. CENA543</b>	<b>GCA_002896875.1_ASM289687v1</b>	<b>INCONCLUSIVO</b>	-
<i>Nostoc</i> sp. NIES-3756	GCA_001548375.1_ASM154837v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Nostoc</i> sp. PCC 7107	GCA_000316625.1_ASM31662v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Nostoc</i> sp. PCC 7120 (SAG 25.82 / UTEX 2576)	GCA_000009705.1_ASM970v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Nostoc</i> sp. PCC 7524 / ATCC 29411	GCA_000316645.1_ASM31664v1	Lindberg, 2002	10.1016/S0360-3199(02)00121-0
<i>Nostocales cyanobacterium HT-58-2</i>	GCA_002163975.1_ASM216397v1	Hughes, 2017	10.1128/AEM.01068-17
<i>Opitutaceae bacterium TAV5</i>	GCA_000242935.3_ASM24293v3	Kotak, 2015	10.1128/genomeA.00060-15
<i>Paenibacillus borealis</i> DSM 13188 2020	GCF_000758665.1_ASM75866v1	Elo, 2001	10.1099/0207713-51-2-535
<i>Paenibacillus donghaensis</i> KCTC 13049 (JH8T/LMG 23780T)	GCA_002192415.1_ASM219241v1	Choi, 2008	PMID: 18309260.
<i>Paenibacillus durus</i> ATCC 35681	GCA_000993825.1_ASM99382v1	-	10.1007/1-4020-3064-1_14
<i>Paenibacillus durus</i> DSM 1735 [ <i>bacillus azotofixans</i> ]	GCA_000756615.1_ASM75661v1	Young, 1992	10.1007/1-4020-3064-1_14
<i>Paenibacillus graminis</i> DSM 15220	GCA_000758705.1_ASM75870v1	Li, 2019	10.3390/ijms20051145
<i>Paenibacillus kribbensis</i> AM49	GCA_002240415.1_ASM224041v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Paenibacillus odorifer</i> CBA7130	GCA_003255855.1_ASM325585v1	Beno, 2020	10.1128/mSphere.00739-19
<i>Paenibacillus odorifer</i> DSM 15391	GCA_000758725.1_ASM75872v1	-	10.1128/mSphere.00739-19
<i>Paenibacillus peoriae</i> HS311	GCA_001272655.2_ASM127265v2	von der Weid, 2002	10.1099/0207713-52-6-2147
<i>Paenibacillus polymyxa</i> ATCC 15970	GCA_001922145.1_ASM192214v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Paenibacillus polymyxa</i> CR1	GCA_000507205.2_ASM50720v2	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Paenibacillus polymyxa</i> J	GCA_001719045.1_ASM171904v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Paenibacillus polymyxa</i> Sb3-1	GCA_000819665.1_ASM81966v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Paenibacillus polymyxa</i> YC0136	GCA_001874405.3_ASM187440v3	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Paenibacillus polymyxa</i> YC0573	GCA_001874425.3_ASM187442v3	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Paenibacillus riograndensis</i> SBR5	GCA_000981585.1_SBR5_T_	Beneduzi, 2010	10.1099/ijis.0.011973-0
<i>Paenibacillus sabiniae</i> T27	GCA_000612505.1_ASM61250v1	Ma, 2007	10.1099/ijis.0.64519-0
<i>Paenibacillus</i> sp. FSL H7-0357	GCA_000758525.1_ASM75852v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<b><i>Paenibacillus</i> sp. FSL P4-0081</b>	<b>GCA_000758565.1_ASM75856v1</b>	<b>INCONCLUSIVO</b>	-
<i>Paenibacillus</i> sp. FSL R5-0912	GCA_000758605.1_ASM75860v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Paenibacillus</i> sp. FSL R7-0273	GCA_000758625.1_ASM75862v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Paenibacillus</i> sp. FSL R7-0331	GCA_000758645.1_ASM75864v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<b><i>Paenibacillus</i> sp. Izh-N1</b>	<b>GCA_002872435.1_ASM287243v1</b>	<b>INCONCLUSIVO</b>	-
<i>Paenibacillus stellifer</i> DSM 14472	GCF_000758685.1_ASM75868v1	Jin, 2011	PMID: 21513217
<i>Paenibacillus terrae</i> HPL003	GCA_000235585.1_ASM23558v1	Xie, 2014	10.1371/journal.pgen.1004231
<i>Paludibacter propionigenes</i> WB4	GCF_000183135.1_ASM18313v1	Inoue, 2015	10.1264/ijme2.ME14142
<i>Pantoea</i> sp. At-9b	GCA_000175935.2_ASM17593v2	Dos Santos, 2012	10.1186/1471-2164-13-162

<i>Paraburkholderia aromaticivorans</i> BN5 (sp BN5)	GCA_002278075.1_ASM227807v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Paraburkholderia phenoliruptrix</i> BR3459a	GCA_000300095.1_ASM30009v1	Cunha, 2012	10.1128/JB.01821-12
<i>Paraburkholderia phymatum</i> STM815	GCA_000020045.1_ASM2004v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Paraburkholderia spreintiae</i> WSM5005	GCA_001865575.1_ASM186557v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Paraburkholderia xenovorans</i> LB400 2010	GCA_000013645.1_ASM1364v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Paraburkholderia xenovorans</i> LB400 2018	GCA_000756045.1_ASM75604v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Pararhodospirillum photometricum</i> DSM 122	GCA_000284415.2_ASM28441v2	Dos Santos, 2020	10.1002/cbic.202000022
<i>Pectobacterium atrosepticum</i> 21A	GCA_000740965.1_ASM74096v1	Cheung, 2016	10.1093/plankt/fbw003
<i>Pectobacterium atrosepticum</i> 36A	GCA_002804265.1_ASM280426v1	Cheung, 2016	10.1093/plankt/fbw003
<i>Pectobacterium atrosepticum</i> JG10-08	GCA_000696465.1_ASM69646v1	Cheung, 2016	10.1093/plankt/fbw003
<i>Pectobacterium atrosepticum</i> SCR11043	GCA_000011605.1_ASM1160v1	Cheung, 2016	10.1093/plankt/fbw003
<i>Pectobacterium polaris</i> NIBIO1006	GCA_002307355.1_ASM230735v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Pelobacter acetylenicus</i> DSM 3246	GCA_001888165.1_ASM188816v1	Akob, 2017	10.1128/AEM.01198-17
<i>Pelobacter carbinolicus</i> DSM 2380	GCA_000012885.1_ASM1288v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Pelobacter propionicus</i> DSM 2379	GCA_000015045.1_ASM1504v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Pelobacter</i> sp SFB93	GCA_001887775.1_ASM188777v1	Akob, 2017	10.1128/AEM.01198-17
<i>Pelodictyon [Chlorobium] luteolum</i> DSM 273	GCA_000012485.1_ASM1248v1	Young, 1992	10.1007/1-4020-3054-1_14
<i>Pelodictyon phaeoclathratiforme</i> BU-1	GCA_000020645.1_ASM2064v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Pelosinus fermentans</i> JBW45	GCF_000271665.2_ASM27166v2	Dos Santos, 2020	10.1002/cbic.202000022
<i>Pelosinus</i> sp UFO1	GCA_000725345.1_ASM72534v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Pleurocapsa</i> sp PCC 7327	GCA_000317025.1_ASM31702v1	Rippka, 1979	10.1099/00221287-111-1-1
<i>Polaromonas naphthalenivorans</i> CJ2	GCA_000015505.1_ASM1550v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Prosthecochloris aestuarii</i> (strain DSM 271 / SK 413)	GCA_000020625.1_ASM2062v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Prosthecochloris</i> sp CIB 2401	GCA_001687065.1_ASM168706v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Prosthecochloris</i> sp GSB1 TY	GCA_002240205.1_ASM224020v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<b><i>Prosthecochloris</i> sp HL-130-GSB</b>	<b>GCA_002113825.1_ASM211382v1</b>	<b>INCONCLUSIVO</b>	-
<b><i>Pseudoclostridium thermosuccinogenes</i> DSM 5807</b>	<b>GCA_002896855.1_ASM289685v1</b>	<b>INCONCLUSIVO</b>	-
<i>Pseudodesulfobivibrio aespoensis</i> Asp-2	GCA_000176915.2_ASM17691v2	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Pseudodesulfobivibrio indicus</i> J2	GCA_001563225.1_ASM156322v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Pseudodesulfobivibrio profundus</i> 500-1	GCA_900217235.1_DPRO_PRJEB22548_v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Pseudomonas stutzeri</i> A1501	GCA_000013785.1_ASM1378v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Pseudomonas stutzeri</i> DSM 4166	GCA_000195105.1_ASM19510v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Rahnella aquatilis</i> HX2	GCA_000255535.1_ASM25553v1	Berge, 1991	10.1139/m91-030
<i>Raoultella omithinolytica</i> S12	GCA_000829965.1_ASM82996v1	Bentzon-Tilia, 2015	10.1128/mBio.00929-15
<i>Rhizobium acidisoli</i> strain FH23	GCF_002531755.2_ASM253175v2	Román-Ponce, 2016	10.1099/jisem.0.000732

**Rhizobium bangladeshense strain BLR175**

<i>Rhizobium esperanzae</i> (etli) N561	
<i>Rhizobium etli</i> 8C3	
<i>Rhizobium etli</i> <i>bv. mimosae</i> str Mim1	
<i>Rhizobium etli</i> <i>bv. phaseoli</i> str. IE4803	
<i>Rhizobium etli</i> CFN 42	
<i>Rhizobium etli</i> CIAT 652	
<i>Rhizobium etli</i> NXC12	
<i>Rhizobium gallicum</i> <i>bv. gallicum</i> R602	
<i>Rhizobium gallicum</i> IE4872	
<i>Rhizobium hidalgonense</i> strain JKLM 19E	
<i>Rhizobium indicum</i> strain JKLM 12A2	
<i>Rhizobium jaguaris</i> strain CCGE525	
<i>Rhizobium leguminosarum</i> ATCC 14479	
<i>Rhizobium leguminosarum</i> <i>bv. trifolii</i> CB782	
<i>Rhizobium leguminosarum</i> <i>bv. trifolii</i> WSM1325	
<i>Rhizobium leguminosarum</i> <i>bv. trifolii</i> WSM1689	
<i>Rhizobium leguminosarum</i> <i>bv. trifolii</i> WSM2304	
<i>Rhizobium leguminosarum</i> <i>bv. viciae</i> 3841	
<i>Rhizobium leguminosarum</i> <i>bv. viciae</i> BIHB 1217	
<i>Rhizobium leguminosarum</i> <i>bv. viciae</i> UPM791	
<i>Rhizobium leguminosarum</i> Norway	
<i>Rhizobium leguminosarum</i> Vaf10	
<i>Rhizobium leguminosarum</i> Vaf-108	
<b>Rhizobium lentis strain BLR27</b>	
<i>Rhizobium phaseoli</i> Brasil 5	
<i>Rhizobium phaseoli</i> N161	
<i>Rhizobium phaseoli</i> N261	
<i>Rhizobium phaseoli</i> N671	
<i>Rhizobium phaseoli</i> N771	
<i>Rhizobium phaseoli</i> N831	
<i>Rhizobium phaseoli</i> N841	
<i>Rhizobium phaseoli</i> N931	
<i>Rhizobium phaseoli</i> R611	
<i>Rhizobium phaseoli</i> R620	

**GCF\_017357245.1\_ASM1735724v1**

GCA_001664265.1_ASM166426v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_001908375.1_ASM190837v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_000442435.1_ASM44243v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_000816125.1_ASM81612v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_000092045.1_ASM9204v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_000020265.1_ASM2026v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_002119845.1_ASM211984v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_000816845.1_ASM81684v1	Young, 1992	10.1007/1-4020-3054-1_14
GCA_001908615.1_ASM190861v1	Young, 1992	10.1007/1-4020-3054-1_14
GCF_005862185.2_ASM586218v2	Yan, 2016	10.1007/s00203-016-1281-x
GCF_005862305.2_ASM586230v2	Rahi, 2020	10.1016/ij.syepm.2020.126127
GCF_003627755.1_ASM362775v1	Rincón-Rosales, 2013	10.1099/ijls.0.048249-0
GCA_003290405.1_ASM329040v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_000520875.1_ASM52087v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_000023185.1_ASM2318v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_000517605.1_ASM51760v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_000021345.1_ASM2134v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_000009265.1_ASM926v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_002243365.1_ASM224336v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_002948295.1_ASM294829v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_002953715.1_ASM295371v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_001679785.1_ASM167978v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_001890425.1_ASM189042v1	Dos Santos, 2012	10.1186/1471-2164-13-162

**GCF\_017352135.1\_ASM1735213v1**

GCA_000172715.2_ASM17271v2	Koirala e Brözel, 2021	10.3390/microorganisms9081662
GCA_001664165.1_ASM166416v1	Matinez, 1988	10.1007/BF02370113
GCA_001664105.1_ASM166410v1	Matinez, 1989	10.1007/BF02370114
GCA_001664365.1_ASM166436v1	Matinez, 1990	10.1007/BF02370115
GCA_001664405.1_ASM166440v1	Matinez, 1991	10.1007/BF02370116
GCA_001664125.1_ASM166412v1	Matinez, 1992	10.1007/BF02370117
GCA_001664425.1_ASM166442v1	Matinez, 1993	10.1007/BF02370118
GCA_001664285.1_ASM166428v1	Matinez, 1994	10.1007/BF02370119
GCA_001664445.1_ASM166444v1	Matinez, 1995	10.1007/BF02370120
GCA_001664245.1_ASM166424v1	Matinez, 1996	10.1007/BF02370121

**INCONCLUSIVO**

Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162
Young, 1992	10.1007/1-4020-3054-1_14
Young, 1992	10.1007/1-4020-3054-1_14
Yan, 2016	10.1007/s00203-016-1281-x
Rahi, 2020	10.1016/ij.syepm.2020.126127
Rincón-Rosales, 2013	10.1099/ijls.0.048249-0
Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162
Dos Santos, 2012	10.1186/1471-2164-13-162

**INCONCLUSIVO**

Koirala e Brözel, 2021	10.3390/microorganisms9081662
Matinez, 1988	10.1007/BF02370113
Matinez, 1989	10.1007/BF02370114
Matinez, 1990	10.1007/BF02370115
Matinez, 1991	10.1007/BF02370116
Matinez, 1992	10.1007/BF02370117
Matinez, 1993	10.1007/BF02370118
Matinez, 1994	10.1007/BF02370119
Matinez, 1995	10.1007/BF02370120
Matinez, 1996	10.1007/BF02370121

<i>Rhizobium phaseoli</i> R630	GCA_001664205.1_ASM166420v1	Matinez, 1997	10.1007/BF02370122
<i>Rhizobium phaseoli</i> R650	GCA_001664385.1_ASM166438v1	Matinez, 1998	10.1007/BF02370123
<i>Rhizobium phaseoli</i> R723	GCA_001664225.1_ASM166422v1	Matinez, 1999	10.1007/BF02370124
<i>Rhizobium pusense</i> IRBG74 (sp IRGB74)	GCF_000499645.1_IRBG74	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Rhizobium ruizarguesonis</i> strain NZLR24	GCF_017357285.1_ASM1735728v1	Jorriñ, 2020	10.1016/j.syepm.2020.126090
<i>Rhizobium</i> sp ACO-34A	GCA_002600635.1_ASM260063v1	INCONCLUSIVO	-
<i>Rhizobium</i> sp CIAT894	GCA_000172795.2_ASM17279v2	Koirlala e Brözel, 2021	10.3390/microorganisms9081662
<i>Rhizobium</i> sp IE4771	GCA_000698845.1_ASM69884v1	Koirlala e Brözel, 2021	10.3390/microorganisms9081662
<i>Rhizobium</i> sp Kim5	GCA_000172695.2_ASM17269v2	INCONCLUSIVO	-
<i>Rhizobium</i> sp N113	GCA_001664345.1_ASM166434v1	Koirlala e Brözel, 2021	10.3390/microorganisms9081662
<i>Rhizobium</i> sp N1314	GCA_001664305.1_ASM166430v1	INCONCLUSIVO	-
<i>Rhizobium</i> sp N1341	GCA_001664085.1_ASM166408v1	INCONCLUSIVO	-
<i>Rhizobium</i> sp N324	GCA_001664485.1_ASM166448v1	INCONCLUSIVO	-
<i>Rhizobium</i> sp N621	GCA_001664325.1_ASM166432v1	INCONCLUSIVO	-
<i>Rhizobium</i> sp N6212	GCA_001664465.1_ASM166446v1	INCONCLUSIVO	-
<i>Rhizobium</i> sp N731	GCA_001664145.1_ASM166414v1	INCONCLUSIVO	-
<i>Rhizobium</i> sp N741	GCA_001664545.1_ASM166454v1	INCONCLUSIVO	-
<i>Rhizobium</i> sp N871	GCA_001664525.1_ASM166452v1	INCONCLUSIVO	-
<i>Rhizobium</i> sp NXC14	GCA_002117485.1_ASM211748v1	Koirlala e Brözel, 2021	10.3390/microorganisms9081662
<i>Rhizobium</i> sp NXC24	GCA_002944315.1_ASM294431v1	Koirlala e Brözel, 2021	10.3390/microorganisms9081662
<i>Rhizobium</i> sp TAL182	GCA_002117725.1_ASM211772v1	INCONCLUSIVO	-
<i>Rhizobium tropici</i> CIAT 899	GCA_000330885.1_ASM33088v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<b><i>Rhodobacter blasticus</i> 28-5</b>	<b>GCA_003071405.1_ASM307140v1</b>	INCONCLUSIVO	-
<i>Rhodobacter capsulatus</i> SB1003 (strain ATCC BAA-309 / NBRC 16581)	GCA_000021865.1_ASM2186v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Rhodobacter sphaeroides</i> 2.4.1 (strain ATCC 17023 / NCIB 8253 / DSM 158)	GCA_003324715.1_ASM332471v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Rhodobacter sphaeroides</i> ATCC 17025	GCA_000016405.1_ASM1640v1	Yoch, 1988	10.1007/BF00409708
<i>Rhodobacter sphaeroides</i> ATCC 17029 (strain / ATH 2.4.9)	GCA_000015985.1_ASM1598v1	Yoch, 1988	10.1007/BF00409708
<i>Rhodobacter sphaeroides</i> KD131 (strain KCTC 12085)	GCA_000021005.1_ASM2100v1	Kim, 2012	10.1016/j.biotech.2012.04.011
<i>Rhodobacter sphaeroides</i> MBTLJ-13	GCA_001685625.1_ASM168562v1	Kim, 2012	10.1016/j.biotech.2012.04.011
<i>Rhodobacter sphaeroides</i> MBTLJ-20	GCA_002706325.1_ASM270632v1	Kim, 2012	10.1016/j.biotech.2012.04.011
<i>Rhodobacter sphaeroides</i> MBTLJ-8	GCA_001576595.1_ASM157659v1	Kim, 2012	10.1016/j.biotech.2012.04.011
<i>Rhodocyclaceae bacterium</i> Thauera-like	GCA_002863805.1_ASM286380v1	Koirlala e Brözel, 2021	10.3390/microorganisms9081662
<i>Rhodoferax antarcticus</i> DSM 24876	GCA_001955735.1_ASM195573v1	Baker, 2017	10.3390/microorganisms5010008

*Rhodomicrobium vannielii* (strain ATCC 17100 / ATH 3.1.1 / DSM 162 / LMG 4299)  
*Rhodopseudomonas palustris* BisA53  
*Rhodopseudomonas palustris* BisB18  
*Rhodopseudomonas palustris* BisB5  
*Rhodopseudomonas palustris* CGA009 (strain ATCC BAA-98)  
*Rhodopseudomonas palustris* DX-1  
*Rhodopseudomonas palustris* HaA2  
*Rhodopseudomonas palustris* PS3  
*Rhodopseudomonas palustris* TIE-1  
*Rhodopseudomonas palustris* YSC3  
*Rhodospirillum centenum* SW (strain ATCC 51521)  
*Rhodospirillum rubrum* ATCC 11170 (strain ATH 1.1.1 / DSM 467 / LMG 4362 / NCIB 8255 / S1)  
*Rhodospirillum rubrum* F11  
*Rhodovulum* sp MB263  
*Rhodovulum* sp P5  
*Rhodovulum sulfidophilum* DSM 1374  
*Rhodovulum sulfidophilum* DSM 2351  
*Rhodovulum sulfidophilum* SNK001  
*Rippakea orientalis* 8801  
*Rippakea orientalis* 8802  
*Rivularia* sp PCC 7116  
*Rubrivivax gelatinosus* IL144  
*Sagittula* sp P11  
*Scytonema* sp. NIES-4073  
*Sedimenticola thioaurini* SIP-G1  
*Sedimimispichoaeta smaragdinae* DSM 11293 (JCM 15392/SEBR 4228)  
*Serratia* sp ATCC 39006 2020  
*Sideroxydans lithotrophicus* ES-1 (ATCC 700298T/ JCM 14762/ DSMZ 22444/ NCMA B100)  
*Sinorhizobium americanum* CCGM7  
*Sinorhizobium americanum* CFNEI 73  
*Sinorhizobium fredii* CCBAU 25509  
*Sinorhizobium fredii* CCBAU 45436

GCA\_000166055.1\_ASM16605v1  
 GCA\_000014825.1\_ASM1482v1  
 GCA\_000013745.1\_ASM1374v1  
 GCA\_000013685.1\_ASM1368v1  
 GCA\_000195775.1\_ASM19577v1  
 GCA\_000177255.2\_ASM17725v2  
 GCA\_000013365.1\_ASM1336v1  
 GCA\_003031265.1\_ASM303126v1  
 GCA\_000020445.1\_ASM2044v1  
 GCA\_003031245.1\_ASM303124v1  
 GCA\_000016185.1\_ASM1618v1  
 GCA\_000013085.1\_ASM1308v1  
 GCA\_000225955.1\_ASM22595v1  
 GCA\_002073975.1\_ASM207397v1  
 GCA\_002079305.1\_ASM207930v1  
 GCA\_001633165.1\_ASM163316v1  
 GCA\_001548075.1\_ASM154807v1  
 GCA\_001633145.1\_ASM163314v1  
 GCF\_000021805.1\_ASM2180v1  
 GCF\_000024045.1\_ASM2404v1  
 GCA\_000316665.1\_ASM31666v1  
 GCA\_000284255.1\_ASM28425v1  
 GCA\_002814095.1\_ASM281409v1  
 GCF\_002368435.1\_ASM236843v1  
 GCA\_001007875.1\_ASM100787v1  
 GCA\_000143985.1\_ASM14398v1  
 GCF\_000463345.2\_ASM46334v3  
 GCA\_000025705.1\_ASM2570v1  
 GCA\_000705595.2\_ASM70559v2  
 GCA\_001889105.1\_ASM188910v1  
 GCA\_003177055.1\_ASM317705v1  
 GCA\_003100575.1\_ASM310057v1  
 Dos Santos, 2012  
 McKinlay, 2014  
 McKinlay, 2014  
 McKinlay, 2014  
 Dos Santos, 2012  
 Xing, 2009  
 McKinlay, 2014  
 McKinlay, 2014  
 McKinlay, 2014  
 McKinlay, 2014  
 Dos Santos, 2012  
 Dos Santos, 2012  
 Dos Santos, 2012  
 Koirala e Brözel, 2021  
 Hallenbeck, 2016  
 Dos Santos, 2012  
 Dos Santos, 2012  
 Koirala e Brözel, 2021  
 Hallenbeck, 2016  
 Dos Santos, 2012  
 Dos Santos, 2012  
 Dos Santos, 2012  
 Young, 1992  
 Bandyopadhyay, 2011  
 Hilton, 2016  
 Nagashima, 2012  
 Martínez-Pérez, 2018  
 Hartman, 2010  
 Flood, 2015  
 Koirala e Brözel, 2021  
 Koirala e Brözel, 2021  
 Emerson, 2013  
 Mnasri, 2012  
 Mnasri, 2012  
 Dos Santos, 2012  
 Dos Santos, 2012

10.1186/1471-2164-13-162

10.1074/jbc.M113.527515

10.1074/jbc.M113.527515

10.1074/jbc.M113.527515

10.1186/1471-2164-13-162

10.1016/j.bios.2009.06.013

10.1074/jbc.M113.527515

10.1074/jbc.M113.527515

10.1074/jbc.M113.527515

10.1074/jbc.M113.527515

10.1186/1471-2164-13-162

10.1186/1471-2164-13-162

10.1186/1471-2164-13-162

10.3390/microorganisms9081662

10.1016/j.ijhygiene.2015.11.090

10.1186/1471-2164-13-162

10.1186/1471-2164-13-162

10.1186/1471-2164-13-162

10.10071-4020-3054-1\_14

10.1128/mBio.00214-11.

10.1371/journal.pone.0156034

10.1128/JB.00511-12

10.1111/1462-2920.14008

10.1007/s00239-010-9365-8

10.1128/genomeA.00671-15

10.3390/microorganisms9081662

10.3390/microorganisms9081662

10.3389/fmicb.2013.00254

10.1016/j.syapm.2012.04.003

10.1016/j.syapm.2012.04.003

10.1186/1471-2164-13-162

10.1186/1471-2164-13-162

<i>Sinorhizobium fredii</i> CCBAU 83666	GCA_002288485.1_ASM228848v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium fredii</i> HH103	GCA_000283895.1_ASM28389v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium fredii</i> NGR234	GCA_000018545.1_ASM1854v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium fredii</i> NXT3	GCA_002944405.1_ASM294440v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium medicae</i> WSM419	GCA_000017145.1_ASM1714v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> 1021	GCA_000006965.1_ASM696v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> 2011	GCA_000346065.1_ASM34606v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> AK83	GCA_000147795.3_ASM14779v3	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> B399	GCA_002302375.1_ASM230237v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> B401	GCA_002302355.1_ASM230235v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> BL225C	GCA_000147775.3_ASM14777v3	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> CCM B554	GCA_002215195.1_ASM221519v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> GR4	GCA_000320385.2_ASM32038v2	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> HM006	GCA_002197165.1_ASM219716v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> KH35c	GCA_002197105.1_ASM219710v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> KH46	GCF_002197465.1_ASM219746v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> M270	GCA_002197085.1_ASM219708v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> Rm41 2015	GCA_000304415.1_RM41	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> Rm41 2017	GCA_002197045.1_ASM219704v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> RMO17	GCA_000747295.1_ASM74729v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> RU11001	GCA_001050915.2_ASM105091v2	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> SM11	GCA_000218265.1_ASM21826v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> T073	GCA_002197145.1_ASM219714v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> USDA1021	GCA_002197445.1_ASM219744v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> USDA1106	GCA_002197065.1_ASM219706v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium meliloti</i> USDA1157	GCA_002197025.1_ASM219702v1	Dos Santos, 2012	10.1186/1471-2164-13-162
<i>Sinorhizobium</i> sp CCBAU 05631	GCA_002288505.1_ASM228850v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<b><i>Skermanella pratensis</i> strain W17</b>	GCF_008843145.1_ASM884314v1	<b>NAO FIXADOR, Guo 2020</b>	10.1099/ijsem.0.003944
<b><i>Skermanella</i> sp TT6</b>	GCA_016653635.2_ASM1665363v2	<b>NAO FIXADOR, Guo 2020</b>	10.1099/ijsem.0.003944
<b><i>Sphaerospermopsis [Anabaena] kisseleviana</i> NIES-73</b>	<b>GCF_002366075.1_ASM236607v1</b>	<b>INCONCLUSIVO</b>	-
<i>Sphingomonas hengshuiensis</i> WHSC-8	GCA_000935025.1_ASM93502v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Spirochaeta thermophila</i> DSM 6192	GCF_000147075.1_ASM14707v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Sulfurospirillum barnesii</i> SES3	GCA_000265295.1_ASM26529v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
<i>Sulfurospirillum cavolei</i> strain UCH003	GCF_001548055.1_ASM154805v1	Ross, 2016	10.1371/journal.pone.0151214

**Sulfurospirillum halorespirans DSM 13726**  
*Sulfurospirillum multivorans* DSM 12446 2020  
**Sulfurospirillum sp JPD-1**  
**Sulfurospirillum sp SL2-1**  
*Synechococcus* sp. (strain JA-2-3B'a(2-13))  
*Synechococcus* sp. (strain JA-3-3Ab)  
**Syntrophobacter fumaroxidans MPOB**  
*Syntrophobotulus glycolicus* DSM 8271  
*Teredinibacter tumerae* T7901  
*Thauera hydrothermalis* strain GD-2  
*Thermincola potens* JR  
**Thermoanaerobacterium aotearoense SCUT27**  
*Thermoanaerobacterium saccharolyticum* JWSL-YS485  
*Thermoanaerobacterium thermosaccharolyticum* DSM 571  
*Thermoanaerobacterium thermosaccharolyticum* M0795  
*Thermoanaerobacterium thermosaccharolyticum* TG57  
*Thermocrinis albus* DSM 14484  
*Thermodesulfovibrio yellowstonii* DSM 11347  
*Thermoleptolyngbya* sp. PKUAC-SCTA183  
*Thiocystis violascens* DSM 198  
*Thioploca ingraca* 2015  
**Thiospirochaeta perfilevii strain P**  
*Tolomonas auensis* DSM 9187  
*Tolypothrix tenuis* PCC 7101  
*Treponema azotonutricium* ZAS-9  
*Treponema primitia* ZAS-2  
*Trichodesmium erythraeum* IMS101  
*Trichormus (Anabaena) variabilis* ATCC 29413  
*Vibrio cincinnatiensis* F8054  
*Vibrio natriegens* CCUG 16371  
*Vibrio natriegens* CCUG 16373  
*Vibrio natriegens* CCUG 16374  
*Vibrio natriegens* NBRC 15636 = ATCC 14048DSM 759  
*Vibrio natriegens* NBRC 15636 = ATCC 14048DSM 759-2  
*Vibrio* sp EJY3

GCA_001723605.1_ASM172360v1	INCONCLUSIVO	
GCF_000568815.1_ASM56881v1	Ju, 2007	10.1111/j.1462-2920.2006.01199.x
GCA_002309535.1_ASM230953v1	INCONCLUSIVO	
GCA_002162315.1_ASM216231v1	INCONCLUSIVO	
GCA_000013225.1_ASM1322v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_000013205.1_ASM1320v1	Young, 1992	10.1007/1-4020-3054-1_14
GCA_000014965.1_ASM1496v1	INCONCLUSIVO	
GCF_000190635.1_ASM19063v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_000023025.1_ASM2302v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCF_003345255.1_ASM334525v1	Raittz, 2021	10.3390/genes12010071
GCA_000092945.1_ASM9294v1	Koirala e Brözel, 2021	10.3390/microorganisms9081662
GCF_009905255.1_ASM990525v1	INCONCLUSIVO	
GCA_000307585.2_ASM30758v2	Currie, 2015	10.1186/12918-015-0159-x
GCA_000145615.1_ASM14561v1	Bogdahn, 1986	10.1007/bf00454964
GCA_000328545.1_ASM32854v1	Bogdahn, 1986	10.1007/bf00454964
GCA_002250075.1_ASM225007v1	Bogdahn, 1986	10.1007/bf00454964
GCA_000025605.1_ASM2560v1	Nishihara, 2018	10.1264/jsme2.ME18041
GCF_000020985.1_ASM2098v1	Boyd, 2011	10.3389/fmicb.2011.00205
GCF_013177315.1_ASM1317731v1	Watanabe, 2021	10.3390/biology10040329
GCA_000227745.3_ASM22774v3	Young, 1992	10.1007/1-4020-3054-1_14
GCA_000828835.1_ASM82883v1	Kojima, 2014	10.1038/ismej.2014.209
GCF_008329945.1_ASM832994v1	INCONCLUSIVO	
GCA_000023065.1_ASM2306v1	Cheung, 2016	10.1093/plankt/fbw003
GCA_002368295.1_ASM236829v1	Hamdi, 1970	PMID: 4986828
GCA_000214355.1_ASM21435v1	Graber, 2004	10.1128/AEM.70.3.1315-1320.2004
GCA_000214375.1_ASM21437v1	Graber, 2004	10.1128/AEM.70.3.1315-1320.2004
GCA_000014265.1_ASM1426v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCA_000204075.1_ASM20407v1	Dos Santos, 2012	10.1186/1471-2164-13-162
GCF_009763885.1_ASM976388v1	Young, 1992	10.1007/1-4020-3054-1_14
GCA_001680045.1_ASM168004v1	Young, 1992	10.1007/1-4020-3054-1_14
GCA_001680065.1_ASM168006v1	Young, 1992	10.1007/1-4020-3054-1_14
GCA_001680085.1_ASM168008v1	Young, 1992	10.1007/1-4020-3054-1_14
GCA_001680025.1_ASM168002v1	Young, 1992	10.1007/1-4020-3054-1_14
GCA_001456255.1_ASM145625v1	Young, 1992	10.1007/1-4020-3054-1_14
GCA_000241385.1_ASM24138v1	Dos Santos, 2012	10.1186/1471-2164-13-162



**Vitreoscilla filiformis ATCC 15551**

Wolinella succinogenes DSM 1740

Xanthobacter autotrophicus Py2

**Yangia sp CCB-MM3****Zoogloeaceae bacterium Par-f-2**

Zymomonas mobilis subsp mobilis ATCC 10988

Zymomonas mobilis subsp mobilis NCIMB 11163

Zymomonas mobilis subsp mobilis NRRL B-12526

Zymomonas mobilis subsp mobilis NRRL B-1960

Zymomonas mobilis subsp mobilis str CP4 = NRRL B-14023

Zymomonas mobilis subsp mobilis str CP4 = NRRL B-14023-2

Zymomonas mobilis subsp mobilis ZM4 = ATCC 31821

Zymomonas mobilis subsp mobilis ZM4 = ATCC 31821-2

Zymomonas mobilis subsp mobilis ZM4 substr. 2032

Zymomonas mobilis subsp mobilis ZM4 substr. 8b

GCA\_002222655.1\_ASM222265v1

GCF\_000196135.1\_ASM19613v1

GCA\_000017645.1\_ASM1764v1

GCA\_001687105.1\_ASM168710v1

GCA\_003054515.1\_ASM305451v1

GCA\_000175255.2\_ASM17525v2

GCA\_000024245.1\_ASM2424v1

GCA\_000576125.1\_ASM57612v1

GCA\_002158845.1\_ASM215884v1

GCA\_000498655.1\_ASM49865v1

GCA\_000576165.1\_ASM57616v1

GCA\_000007105.1\_ASM710v1

GCA\_003054575.1\_ASM305457v1

GCA\_003052045.1\_ASM305204v1

GCA\_003052025.1\_ASM305202v1

FONTE: O AUTOR (2022).

Koirala e Brözel, 2021

Hartman, 2010

Dos Santos, 2012

Koirala e Brözel, 2021

**INCONCLUSIVO**

Kremer, 2015

Kremer, 2015

Kremer, 2015

Kremer, 2015

Kremer, 2015

Kremer, 2015

Kremer, 2015

Kremer, 2015

Kremer, 2015

Kremer, 2015

10.3390/microorganisms9081662

10.1007/s00239-010-9365-8

10.1186/1471-2164-13-162

10.3390/microorganisms9081662

10.1073/pnas.1420663112

10.1073/pnas.1420663112

10.1073/pnas.1420663112

10.1073/pnas.1420663112

10.1073/pnas.1420663112

10.1073/pnas.1420663112

10.1073/pnas.1420663112

10.1073/pnas.1420663112

10.1073/pnas.1420663112

10.1073/pnas.1420663112

10.1073/pnas.1420663112

LEGENDA: Cepas em **negrito** e sem referência são **inconclusivos** ou sem referência na validação da ocorrência da FBN.

TABELA 4 - RELAÇÃO DOS NIFS IDENTIFICADOS POR RAFTS<sup>3</sup>G EM DADOS DE 662 PROTEOMAS COMPLETOS.

Proteína	Nº de Seqs	Agrupamento Nº
<b>NifH</b>	1001	19; 7934
<b>NifD</b>	680	194; 2363
<b>NifK</b>	681	192; 2362
<b>NifE</b>	673	228; 2349; 19006;
<b>NifEN</b>	13	36292;
<b>NifN</b>	605	488; 6423; 13736; 18155; 21147; 21495; 34326; 68118; 72650;
		74873

<b>NifNB</b>	49	10563
<b>NifB</b>	620	403; 7196; 10391; 12434; 17157; 15532*(SAM); 34170; 46009
<b>NifX</b>	510	1280; 2664; 7537; 20179; 21146; 22124; 22984; 26883; 45861; 46394
<b>NifV</b>	491	911; 1934; 15984* (alfa); 25539; 29600; 29861; 32348; 49250
<b>NifU</b>	272	812; 18111
<b>NifS</b>	783	33; 27390; 40669
<b>NifP</b>	221	11801; 4080
<b>NifZ</b>	476	139; 5628; 7253; 14214; 40913; 55180; 65372; 68116; 79452
<b>NifT</b>	418	139; 6422; 12246; 49248
<b>NifW</b>	379	2803; 5916; 7536; 17320; 19555; 40916; 48768
<b>NifM</b>	110	10456; 33122; 61146; 68115; 89688; 114130; 164003
<b>NifQ</b>	242	6947; 7165; 13286; 25229; 37252; 39732; 62056; 68113; 76981; 94175
<b>NifY</b>	109	6831; 15069; 61499
<b>NifF</b>	204	2810; 4563; 9622; 19740
<b>NifJ</b>	485	261
<b>NifO</b>	41	62055
<b>NifA</b>	366	1746; 2792; 17216; 31078; 39463; 47939; 54832; 55173; 88757
<b>NifL</b>	108	7252; 61097; 68111; 115907; 118489

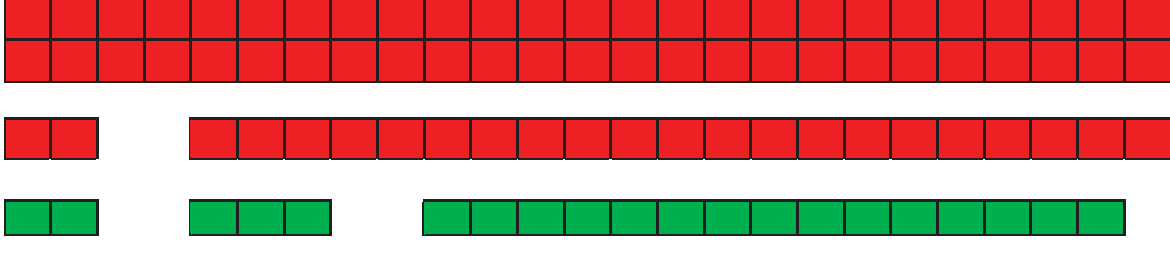
FONTE: O Autor (2022).

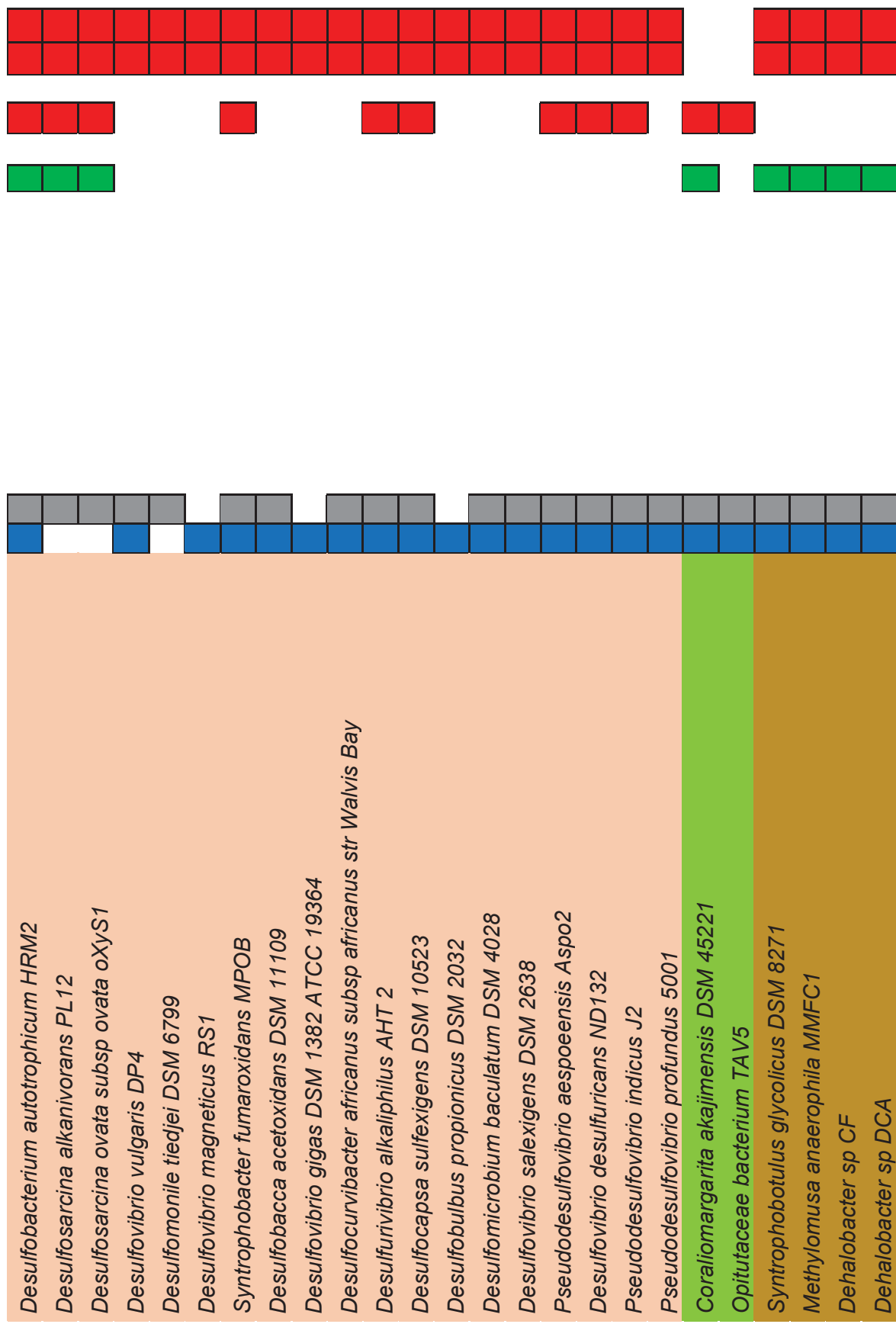






<i>Treponema primitia</i> ZAS2			
<i>Sediminispirochaeta smaragdinae</i> DSM 11293			
<i>Dehalococcoides mccartyi</i> 195			
<i>Dehalogenimonas</i> sp. WBC2			
<i>Labilibaculum antarcticum</i> SPP2			
<i>Draconibacterium</i> sp. M1			
<i>Paludibacter propionigenes</i> WB4			
<i>Akalitalea saponilacus</i> SC			
<i>Dysgonomonas capnocytophagoides</i> DSM 22835			
<i>Candidatus Azobacteroides pseudotriconymphae</i>			
<i>Chloroherpeton thalassium</i> ATCC 35110			
<i>Prosthecochloris</i> sp. GSB1			
<i>Prosthecochloris</i> sp. HL-130-GSB			
<i>Prosthecochloris aestuarii</i> DSM 271			
<i>Prosthecochloris</i> sp. CIB 2401			
<i>Chlorobium phaeobacteroides</i> BS1			
<i>Chlorobium chlorochromatii</i> CaD3			
<i>Chlorobaculum limnaeum</i> DSM 1677			
<i>Chlorobaculum parvum</i> NCIB 8327			
<i>Chlorobaculum tepidum</i> TLS			
<i>Chlorobium phaeobacteroides</i> DSM 266			
<i>Pelodictyon phaeoclathratiforme</i> BU1			
<i>Chlorobium limicola</i> DSM 245			
<i>Pelodictyon</i> [ <i>Chlorobium</i> ] <i>luteolum</i> DSM 273			
<i>Desulfatibacillum aliphaticivorans</i> AK-01			







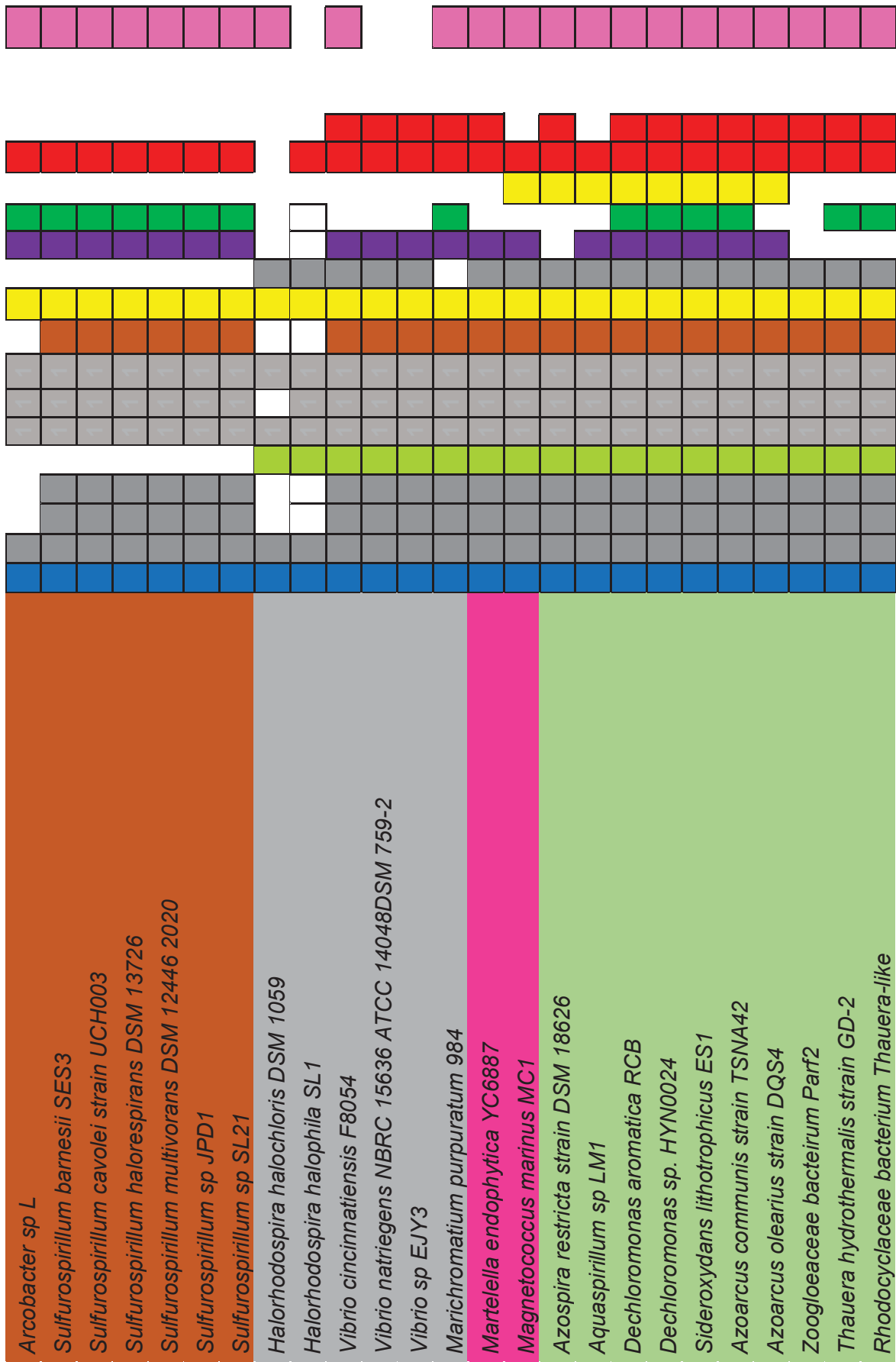


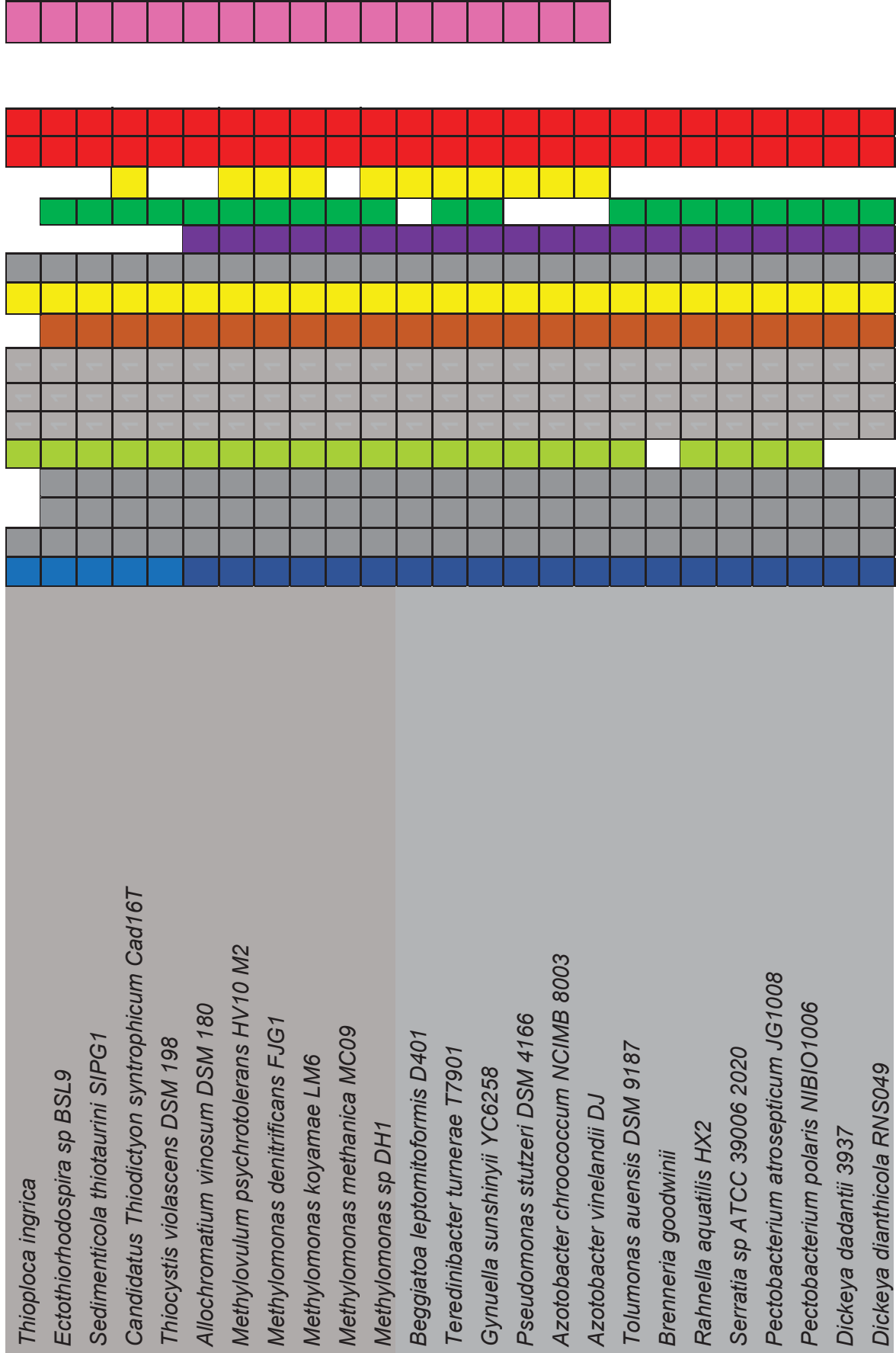


<i>Leptolyngbya</i> sp O77																						
<i>Euhalothece natronophila</i> Z M001																						
<i>Halothece</i> sp PCC 7418																						
<i>Gloeothece citriformis</i> PCC 7424																						
<i>Gloeothece verrucosa</i> PCC 7822																						
<i>Rippakea orientalis</i> 8802																						
<i>Rippakea orientalis</i> 8801																						
<i>Cyanothece</i> sp ATCC 51142																						
<i>Cyanothece</i> sp PCC 7425																						
<i>Candidatus Atelocyanobacterium thalassa</i> isolate ALOHA cyanobacterium endosymbiont of <i>Epithemia turgida</i> isolate EtSB Lake Yunoko																						
<i>Cyanobacterium endosymbiont of Rhopalodia gibberula</i>																						
<i>Crocospaera subtropica</i> ATCC 51142																						
<i>Pleurocapsa</i> sp PCC 7327																						
<i>Leptolyngbya boryana</i> dg5																						
<i>Microcoleus</i> sp PCC 7113																						
<i>Chroococcidiopsis thermalis</i> PCC 7203																						
<i>Calothrix brevis</i> NIES22																						
<i>Trichormus [Anabaena] variabilis</i> ATCC 29413																						
<i>Nostoc carneum</i> NIES2107																						
<i>Tolypothrix tenuis</i> PCC 7101																						
<i>Aulosira laxa</i> NIES50																						
<i>Calothrix</i> sp 3363																						
<i>Fischerella</i> sp NIES3754																						
<i>Nostoc</i> sp CENA543																						

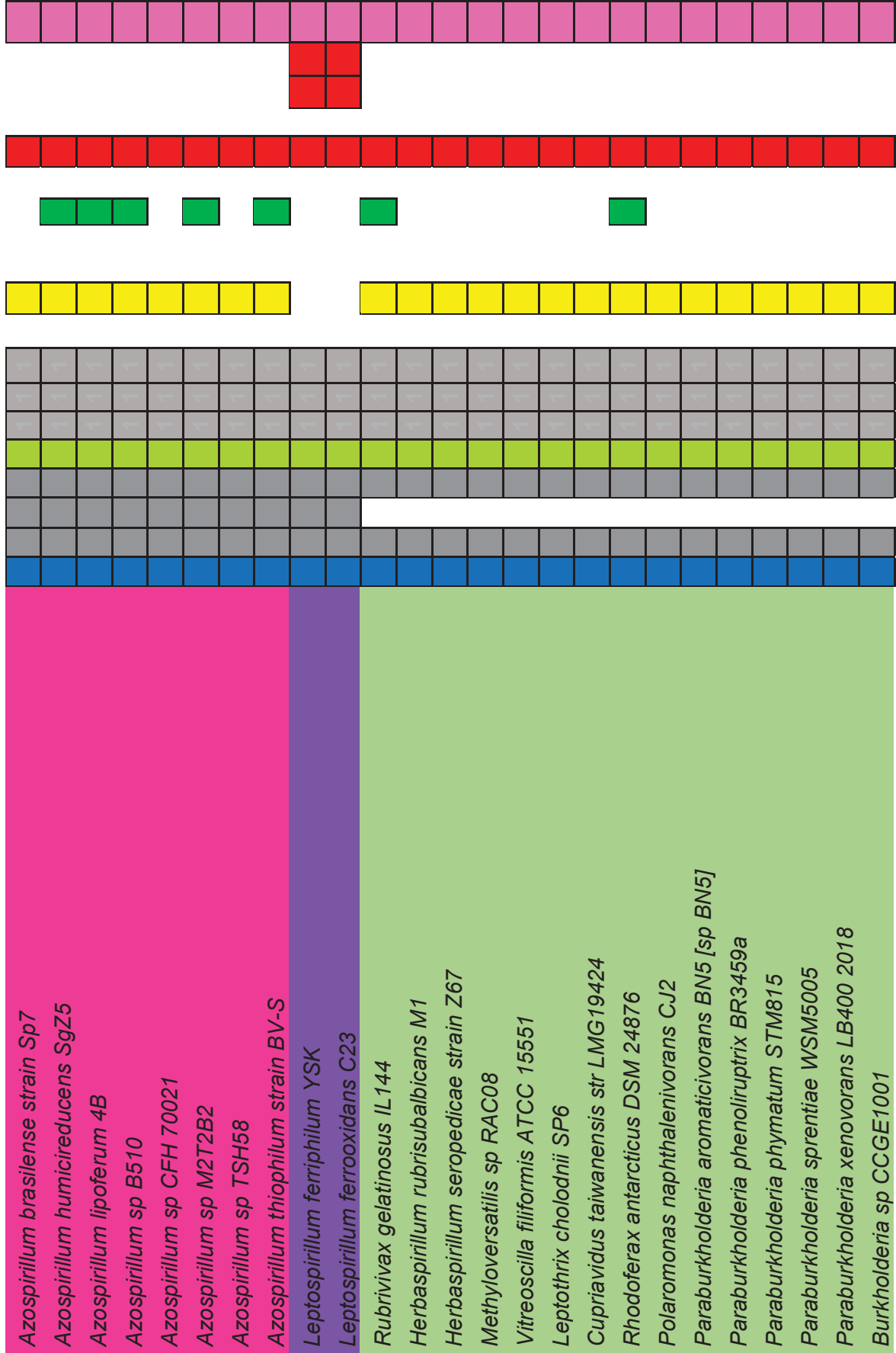






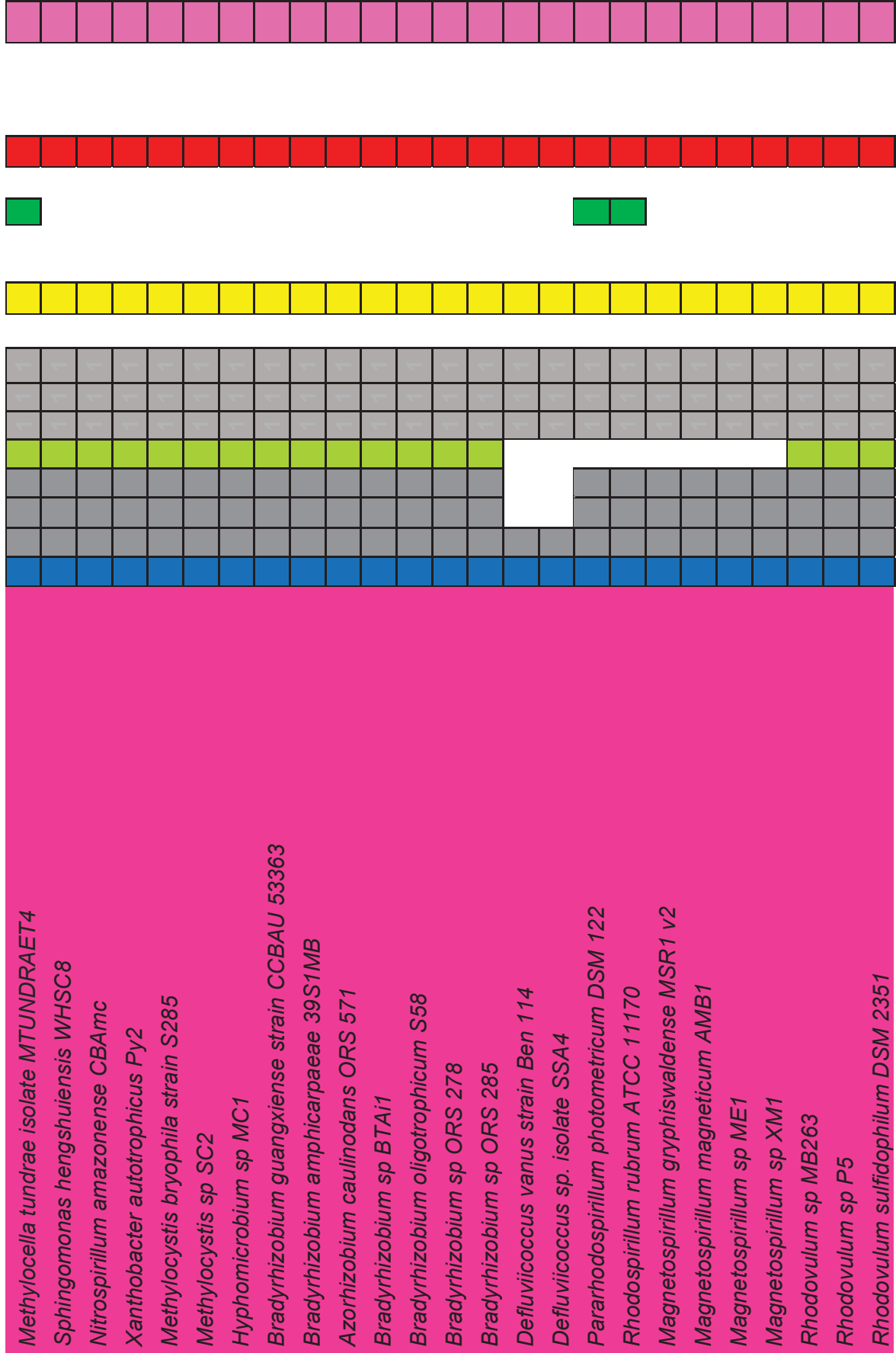


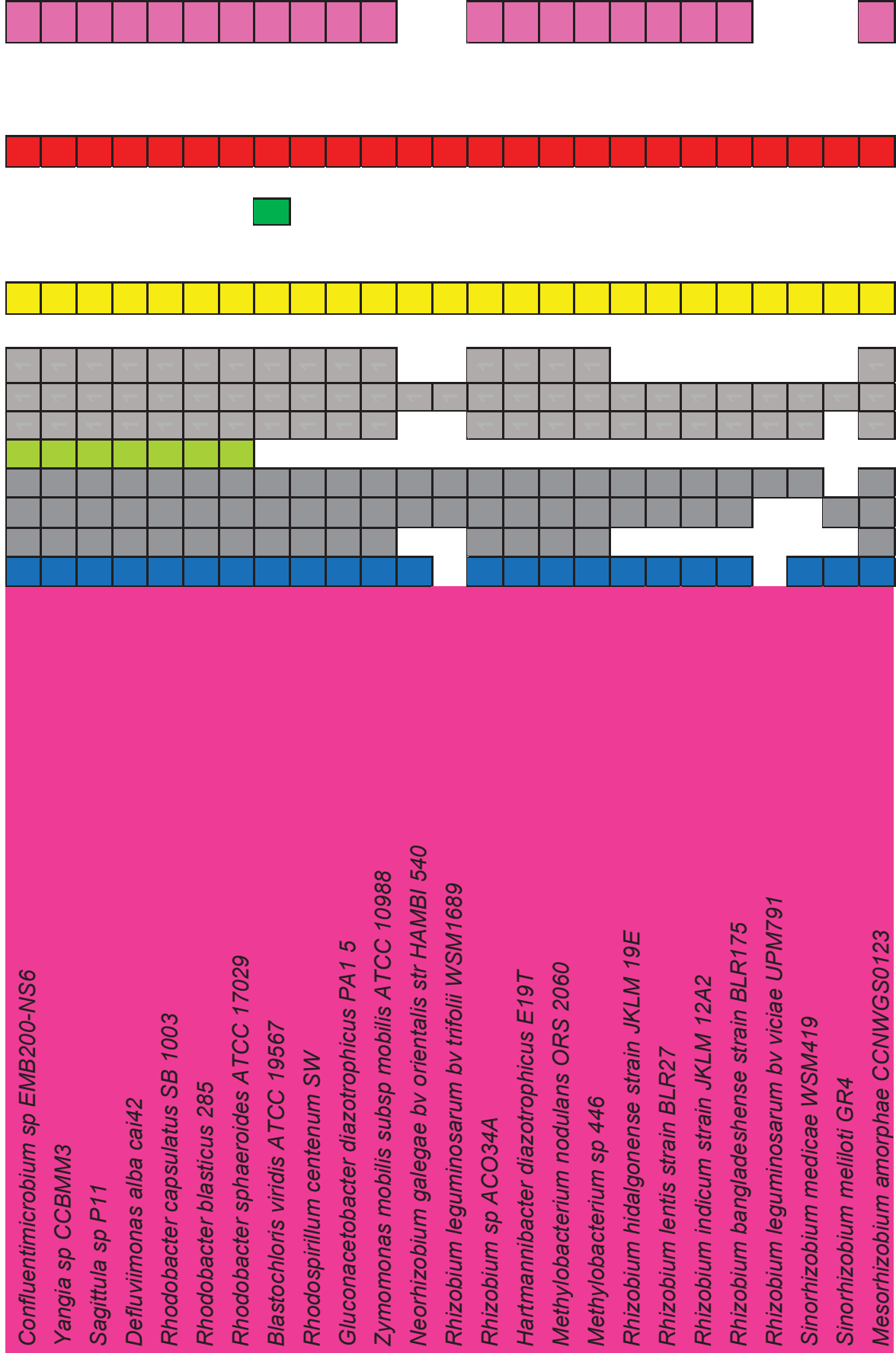




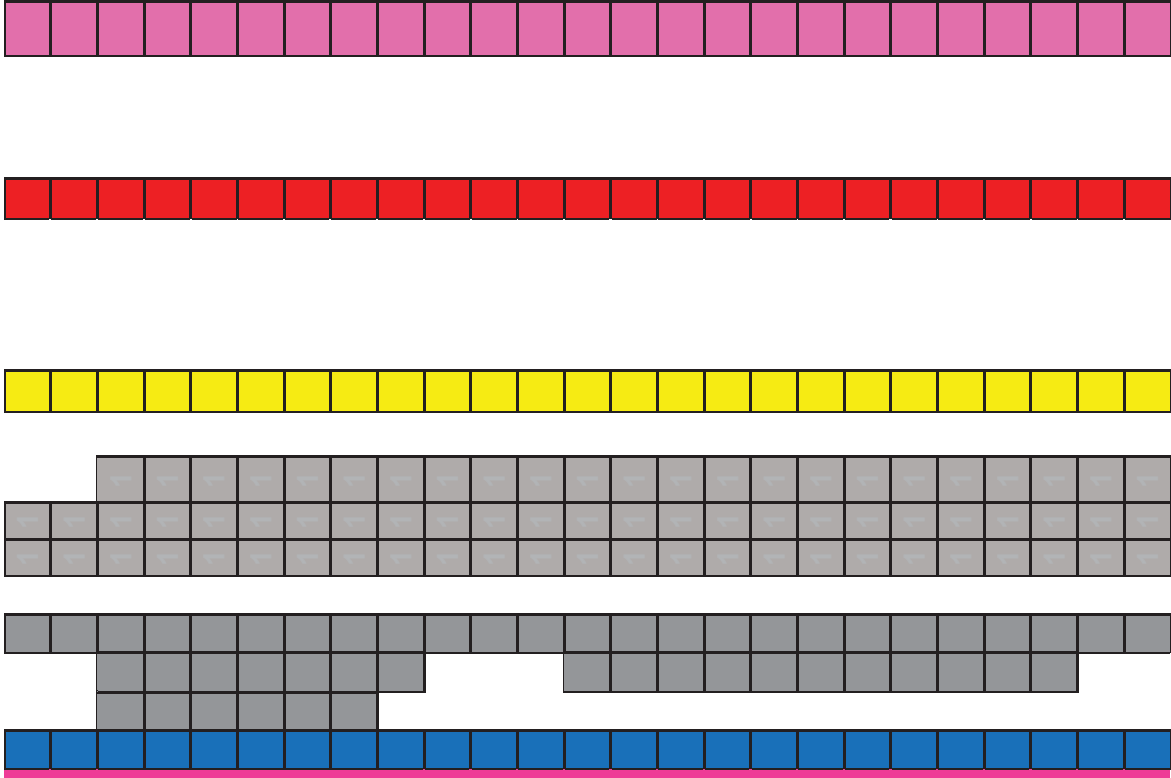








*Mesorhizobium huakuii* 7653R  
*Mesorhizobium japonicum* MAFF 303099  
*Mesorhizobium loti* NZP2037  
*Mesorhizobium erdmanii* NZP2014  
*Mesorhizobium australicum* WSM2073  
*Mesorhizobium ciceri* biovar *biserrulae* WSM1271  
*Mesorhizobium opportunistum* WSM2075  
*Mesorhizobium* sp WSM1497  
*Rhizobium pusense* (sp IRBG74)  
*Sinorhizobium* sp CCBAU 05631  
*Sinorhizobium fredii* NGR234  
*Ensifer sojae* CCBAU 05684  
*Rhizobium jaguaris* strain CCGE525  
*Rhizobium tropici* CIAT 899  
*Rhizobium* sp NXC24  
*Rhizobium gallicum* IE4872 plasmid pRgallE4872c  
*Sinorhizobium americanum* CCGM7  
*Rhizobium etli* bv *mimosae* str Mim1  
*Rhizobium* sp TAL182  
*Rhizobium* sp IE4771  
*Rhizobium acidisoli* strain FH23  
*Rhizobium* sp Kim5  
*Rhizobium phaseoli* Brasil 5  
*Rhizobium* sp CIAT894  
*Rhizobium* sp N621



*Rhizobium esperanzae* (etii) N561



FONTE: O Autor (2022).

## 4. RESULTADOS PARA NIFFINDER

TABELA 6 – RESULTADO DOS TREINAMENTOS PARA AS REDES NIF EM NIFFINDER

<b>NifH</b>	<b>MLP</b>	<b>RBF</b>
<i>S</i>	1,000	0,993
<i>E</i>	0,997	1,000
<i>A</i>	0,999	0,995
<i>P</i>	0,998	1,000
<i>F1</i>	0,999	0,996
<b>NifD</b>	<b>MLP</b>	<b>RBF</b>
<i>S</i>	0,997	0,997
<i>E</i>	1,000	1,000
<i>A</i>	0,998	0,998
<i>P</i>	1,000	1,000
<i>F1</i>	0,999	0,999
<b>NifK</b>	<b>MLP</b>	<b>RBF</b>
<i>S</i>	0,988	0,988
<i>E</i>	0,995	1,000
<i>A</i>	0,990	0,992
<i>P</i>	0,998	1,000
<i>F1</i>	0,993	0,994
<b>NifE</b>	<b>MLP</b>	<b>RBF</b>
<i>S</i>	0,995	0,990
<i>E</i>	0,995	1,000
<i>A</i>	0,995	0,993
<i>P</i>	0,997	1,000
<i>F1</i>	0,996	0,995
<b>NifN</b>	<b>MLP</b>	<b>RBF</b>
<i>S</i>	0,959	0,930
<i>E</i>	0,924	1,000
<i>A</i>	0,948	0,950
<i>P</i>	0,962	1,000
<i>F1</i>	0,961	0,963
<b>NifX</b>	<b>MLP</b>	<b>RBF</b>
<i>S</i>	0,870	0,891
<i>E</i>	0,850	0,985
<i>A</i>	0,864	0,916
<i>P</i>	0,934	0,994
<i>F1</i>	0,901	0,940
<b>NifB</b>	<b>MLP</b>	<b>RBF</b>
<i>S</i>	0,985	0,987

<i>E</i>	1,000	1,000
<i>A</i>	0,990	0,991
<i>P</i>	1,000	1,000
<i>F1</i>	0,992	0,994
<b>NifO</b>	<b>MLP</b>	<b>RBF</b>
<i>S</i>	1,000	0,973
<i>E</i>	0,947	1,000
<i>A</i>	0,981	0,981
<i>P</i>	0,972	1,000
<i>F1</i>	0,986	0,986
<b>NifQ</b>	<b>MLP</b>	<b>RBF</b>
<i>S</i>	0,978	0,974
<i>E</i>	0,843	1,000
<i>A</i>	0,933	0,981
<i>P</i>	0,926	1,000
<i>F1</i>	0,951	0,987
<b>NifU</b>	<b>MLP</b>	<b>RBF</b>
<i>S</i>	0,982	0,994
<i>E</i>	0,919	1,000
<i>A</i>	0,958	0,996
<i>P</i>	0,953	1,000
<i>F1</i>	0,967	0,997
<b>NifS</b>	<b>MLP</b>	<b>RBF</b>
<i>S</i>	0,988	0,978
<i>E</i>	0,968	1,000
<i>A</i>	0,981	0,985
<i>P</i>	0,984	1,000
<i>F1</i>	0,986	0,989
<b>NifV</b>	<b>MLP</b>	<b>RBF</b>
<i>S</i>	1,000	0,984
<i>E</i>	0,992	1,000
<i>A</i>	0,995	0,989
<i>P</i>	0,987	1,000
<i>F1</i>	0,994	0,992
<b>NifW</b>	<b>MLP</b>	<b>RBF</b>
<i>S</i>	0,985	0,978
<i>E</i>	0,831	1,000
<i>A</i>	0,924	0,985
<i>P</i>	0,898	1,000
<i>F1</i>	0,940	0,989
<b>NifZ</b>	<b>MLP</b>	<b>RBF</b>
<i>S</i>	0,990	0,983
<i>E</i>	0,998	0,993
<i>A</i>	0,992	0,987
<i>P</i>	0,999	0,996
<i>F1</i>	0,994	0,989

<b>NifP</b>	<b>MLP</b>	<b>RBF</b>
S	0,992	0,988
E	1,000	1,000
A	0,995	0,992
P	1,000	1,000
F1	0,996	0,994
<b>NifT</b>	<b>MLP</b>	<b>RBF</b>
S	0,977	0,983
E	0,873	0,993
A	0,934	0,987
P	0,915	0,996
F1	0,945	0,989
<b>NifY</b>	<b>MLP</b>	<b>RBF</b>
S	0,996	0,988
E	0,992	1,000
A	0,995	0,992
P	0,996	1,000
F1	0,996	0,994
<b>NifM</b>	<b>MLP</b>	<b>RBF</b>
S	0,985	0,987
E	0,909	1,000
A	0,960	0,991
P	0,956	1,000
F1	0,970	0,993
<b>NifF</b>	<b>MLP</b>	<b>RBF</b>
S	0,983	0,992
E	0,953	1,000
A	0,973	0,995
P	0,975	1,000
F1	0,979	0,996
<b>NifJ</b>	<b>MLP</b>	<b>RBF</b>
S	0,993	0,996
E	1,000	1,000
A	0,995	0,998
P	1,000	1,000
F1	0,996	0,998
<b>NifA</b>	<b>MLP</b>	<b>RBF</b>
S	0,972	0,983
E	1,000	1,000
A	0,981	0,989
P	1,000	1,000
F1	0,986	0,992
<b>NifL</b>	<b>MLP</b>	<b>RBF</b>
S	0,923	0,923
E	0,970	0,970
A	0,939	0,939

<i>P</i>	0,984	0,984
<i>F1</i>	0,952	0,952
<b>Nif1</b>	<b>MLP</b>	<b>RBF</b>
<i>S</i>	0,976	0,980
<i>E</i>	0,898	0,993
<i>A</i>	0,948	0,984
<i>P</i>	0,946	0,996
<i>F1</i>	0,960	0,988
<b>Nif2</b>	<b>MLP</b>	<b>RBF</b>
<i>S</i>	0,947	0,962
<i>E</i>	0,863	0,996
<i>A</i>	0,918	0,973
<i>P</i>	0,928	0,998
<i>F1</i>	0,937	0,980
<b>NifXa</b>	<b>MLP</b>	<b>RBF</b>
<i>S</i>	0,972	0,987
<i>E</i>	0,826	1,000
<i>A</i>	0,917	0,991
<i>P</i>	0,902	1,000
<i>F1</i>	0,936	0,994

FONTE: O Autor (2022).

LEGENDA: Resultados dos treinamentos para cada preditor Nif em modelo MLP e RBF. Calculamos as estatísticas sensibilidade (*S*), especificidade (*E*), acurácia (*A*), precisão (*P*) e F1-score (*F1*). Foram selecionados os modelos com melhor F1-score. Caso as métricas fossem muito próximas elencamos os modelos com maior número de acerto (acurácia).



QUADRO 2 – RESULTADOS OBTIDOS COM NIFPRED E NIFFINDER PARA APHANIZOMENON FLOS-AQUAE DEX188

**NifPRED**

*Sequência identificada por NifPRED:*

QSV67144.1 AAA family ATPase [Aphanizomenon flos-aquae DEX188]	<b>nifH</b>
QSV67151.1 nitrogenase molybdenum-iron protein alpha chain [Aphanizomenon flos-aquae DEX188]	<b>nifD</b>
QSV67173.1 nitrogenase molybdenum-iron protein subunit beta [Aphanizomenon flos-aquae DEX188]	<b>nifK</b>
QSV67174.1 nitrogenase iron-molybdenum cofactor biosynthesis protein NifE [Aphanizomenon flos-aquae DEX188]	<b>nifE</b>
QSV67175.1 nitrogenase iron-molybdenum cofactor biosynthesis protein NifN [Aphanizomenon flos-aquae DEX188]	<b>nifN</b>
QSV69175.1 nitrogen fixation protein NifX [Aphanizomenon flos-aquae DEX188]	<b>non-nif</b>
QSV67140.1 nitrogenase cofactor biosynthesis protein NifB [Aphanizomenon flos-aquae DEX188]	<b>nifB</b>
QSV67142.1 cysteine desulfurase NifS [Aphanizomenon flos-aquae DEX188]	<b>nifN</b>
QSV67565.1 2-isopropylmalate synthase [Aphanizomenon flos-aquae DEX188]	<b>nifE</b>
QSV65539.1 septum formation inhibitor Maf [Aphanizomenon flos-aquae DEX188]	<b>nifB</b>
QSV67178.1 nitrogenase-stabilizing/protective protein NifW [Aphanizomenon flos-aquae DEX188]	<b>non-nif</b>
QSV67134.1 nitrogen fixation protein NifZ [Aphanizomenon flos-aquae DEX188]	<b>non-nif</b>
QSV67176.1 NifX-associated nitrogen fixation protein [Aphanizomenon flos-aquae DEX188]	<b>non-nif</b>
QSV67138.1 serine O-acetyltransferase [Aphanizomenon flos-aquae DEX188]	<b>non-nif</b>
QSV67902.1 serine O-acetyltransferase [Aphanizomenon flos-aquae DEX188]	<b>non-nif</b>
QSV67358.1 pyruvate:ferredoxin (flavodoxin) oxidoreductase [Aphanizomenon flos-aquae DEX188]	<b>nifE</b>
QSV67499.1 pyruvate:ferredoxin (flavodoxin) oxidoreductase [Aphanizomenon flos-aquae DEX188]	<b>non-nif</b>
QSV66065.1 P-II family nitrogen regulator [Aphanizomenon flos-aquae DEX188]	<b>non-nif</b>
QSV67133.1 putative nitrogen fixation protein NifT [Aphanizomenon flos-aquae DEX188]	<b>non-nif</b>

**NifFINDER**

*Sequência identificada por NifFinder:*

QSV67144.1 AAA family ATPase [Aphanizomenon flos-aquae DEX188]	<b>NifH</b>
QSV67151.1 nitrogenase molybdenum-iron protein alpha chain [Aphanizomenon flos-aquae DEX188]	<b>NifD</b>
QSV67173.1 nitrogenase molybdenum-iron protein subunit beta [Aphanizomenon flos-aquae DEX188]	<b>NifK</b>
QSV67133.1 putative nitrogen fixation protein NifT [Aphanizomenon flos-aquae DEX188]	<b>NifT</b>
QSV67174.1 nitrogenase iron-molybdenum cofactor biosynthesis protein NifE [Aphanizomenon flos-aquae DEX188]	<b>NifE</b>
QSV67175.1 nitrogenase iron-molybdenum cofactor biosynthesis protein NifN [Aphanizomenon flos-aquae DEX188]	<b>NifN</b>
QSV69175.1 nitrogen fixation protein NifX [Aphanizomenon flos-aquae DEX188]	<b>NifX</b>
QSV67140.1 nitrogenase cofactor biosynthesis protein NifB [Aphanizomenon flos-aquae DEX188]	<b>NifB</b>
QSV67143.1 Fe-S cluster assembly protein NifU [Aphanizomenon flos-aquae DEX188]	<b>NifU</b>
QSV67142.1 cysteine desulfurase NifS [Aphanizomenon flos-aquae DEX188]	<b>NifS</b>
QSV67135.1 homocitrate synthase [Aphanizomenon flos-aquae DEX188]	<b>NifV</b>
QSV67178.1 nitrogenase-stabilizing/protective protein NifW [Aphanizomenon flos-aquae DEX188]	<b>NifW</b>
QSV67134.1 nitrogen fixation protein NifZ [Aphanizomenon flos-aquae DEX188]	<b>NifZ</b>
QSV67176.1 NifX-associated nitrogen fixation protein [Aphanizomenon flos-aquae DEX188]	<b>NifXa</b>
QSV67138.1 serine O-acetyltransferase [Aphanizomenon flos-aquae DEX188]	<b>NifP</b>
QSV67902.1 serine O-acetyltransferase [Aphanizomenon flos-aquae DEX188]	<b>NifP</b>
QSV67358.1 pyruvate:ferredoxin (flavodoxin) oxidoreductase [Aphanizomenon flos-aquae DEX188]	<b>NifJ</b>
QSV67499.1 pyruvate:ferredoxin (flavodoxin) oxidoreductase [Aphanizomenon flos-aquae DEX188]	<b>NifJ</b>
QSV66065.1 P-II family nitrogen regulator [Aphanizomenon flos-aquae DEX188]	<b>NifII</b>

FONTE: O Autor (2022).

LEGENDA: em NifPRED, as seqüências em verde, são as seqüências identificadas corretamente, em amarelo são seqüências que o preditor trouxe como não Nif (non-nif) e em vermelho seqüências trazidas como Nif, porém de outra classe (erro). Nesse exemplo, NifFINDER conseguiu predizer todas as Nifs corretamente.

## 5. PCA

FIGURA 3 – SCRIPT PLOTPCA PARA PLOTAR PCA EM 3D

```

library(rgl)
plotPCA <- function(x, nGroup) {
  n <- ncol(x)
  if(!(n %in% c(2,3))) { # check if 2d or 3d
    stop("x must have either 2 or 3 columns")
  }

  #Usando Hclust:
  #fit <- hclust(dist(x), method="complete") # cluster
  #groups <- cutree(fit, k=nGroup)

  #usando HCPC:
  groups <- res.HCPC$call$X$clust

  if(n == 3) { # 3d plot points
    plot3d(x, col=groups, type="s", size=1, axes=F)
    axes3d(edges=c("x--", "y--", "z"), lwd=3, axes.len=2, labels=T)
    grid3d("x")
    grid3d("y")
    grid3d("z")
  } else { # 2d plot
    maxes <- apply(abs(x), 2, max)
    rangeX <- c(-maxes[1], maxes[1])
    rangeY <- c(-maxes[2], maxes[2])
    plot(x, col=groups, pch=29, xlab=colnames(x)[1], ylab=colnames(x)[2], xlim=rangeX, ylim=rangeY)
    lines(c(0,0), rangeX*2)
    lines(rangeY*2, c(0,0))
  }
}

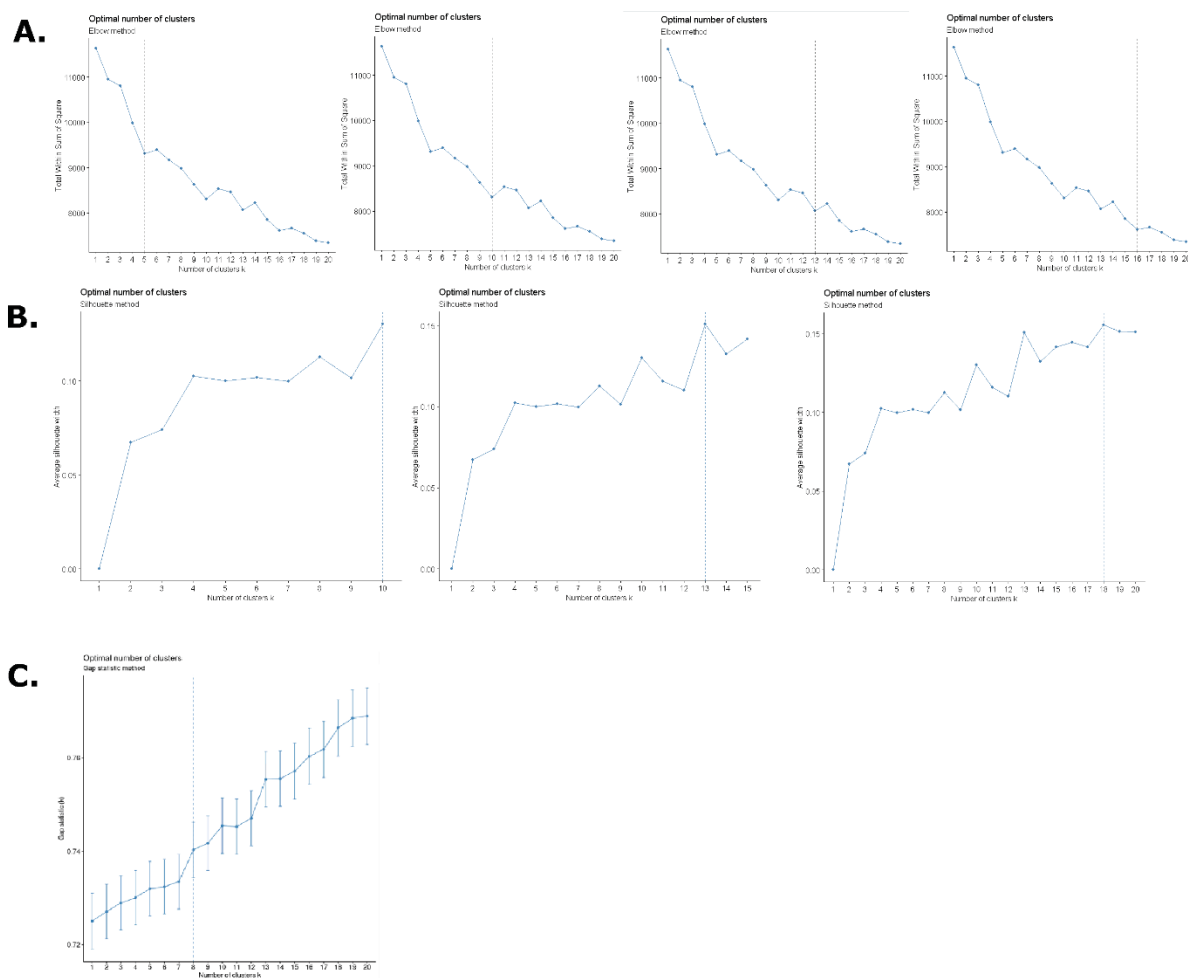
```

FONTE: O Autor (2022).

## 6. OBTENÇÃO DE K-GRUPOS

### 6.1 POR MÉTODO DO COTOVELO, COEFICIENTE DA SILHUETA E ESTATÍSTICA GAP

FIGURA 4 -RESULTADOS PARA (A) MÉTODO DO COTOVELO, ( B) COEFICIENTE DA SILHUETA E (C) ESTATÍSTICA GAP PARA OBTENÇÃO DE K EM DADOS DE NIF CORE

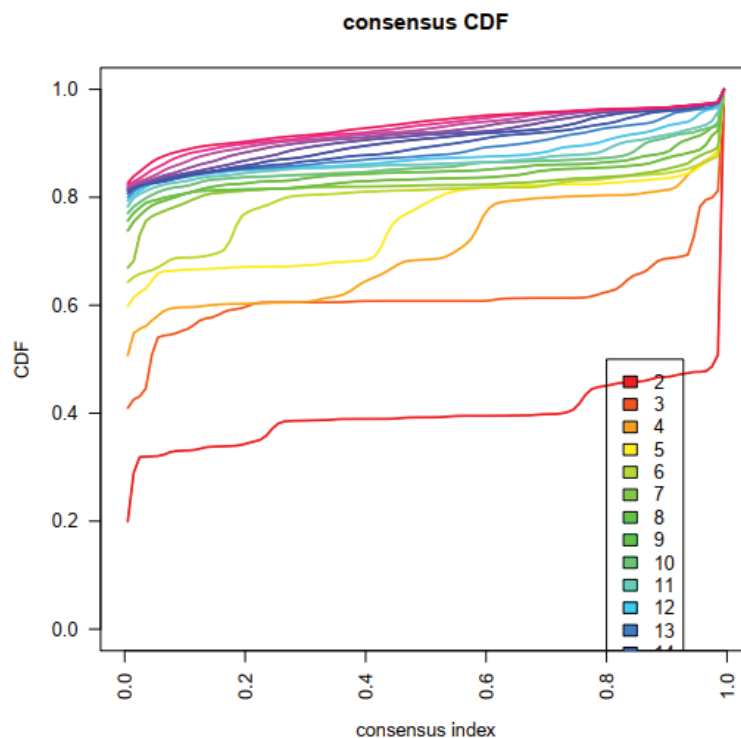


FONTE: O Autor (2022).

## 6.2 POR HCPC

*Consensus CDF*: o gráfico mostra as funções de distribuição cumulativa do consenso matriz para cada  $k$  (indicado por cores), estimado por um histograma. Esta figura permite que o usuário determine em qual número de cluster ( $k$ ), o CDF atinge um máximo aproximado, portanto, o consenso e a confiança do cluster estão em um máximo neste  $k$ .

FIGURA 5 – CONSENSUS CDF



FONTE: O Autor (2022).

**Consensus Matrix:** este gráfico mostra a mudança relativa na área sob a curva CDF comparando  $k$  e  $k - 1$ . Para  $k = 2$ , não há  $k - 1$ , então a área total sob a curva em vez do aumento relativo. Este gráfico permite que um usuário determine o aumento relativo no consenso e determine  $k$  em que há nenhum aumento apreciável.

TABELA 7 - DISTRIBUIÇÃO DOS DIAZOTROFOS PARA OS VALORES  $k$  6, 8, 10 E 13.

<b>Organismos</b>	k-8	k-10	k-13
<i>Acetobacterium woodii</i> DSM 1030	2	3	3
<i>Acidithiobacillus ferrivorans</i> SS3	6	7	8
<i>Acidithiobacillus ferrooxidans</i> ATCC 23270	6	7	8
<i>Acidithiobacillus ferrooxidans</i> ATCC 53993	6	7	8
<i>Alkaliphilus metalliredigens</i> QYMF	2	3	3
<i>Alkalitalea saponilacus</i>	2	3	3
<i>Allochromatium vinosum</i> DSM 180	5	5	5
<i>Anabaena cylindrica</i> PCC 7122	4	6	6
<i>Anabaena</i> sp 90	4	6	6
<i>Anabaena</i> sp WA102	4	6	6
<i>Anabaenopsis circularis</i> NIES-21	4	6	6

<i>Anabaenopsis elenkinii</i> CCIBt3563	4	6	6
<i>Anaeromyxobacter</i> sp Fw109-5	3	4	4
<i>Anaeromyxobacter</i> sp K	3	4	4
<i>Aphanizomenon flos-aquae</i> DEX188	4	6	6
<i>Aquaspirillum</i> sp LM1	5	5	5
<i>Arcobacter nitrofigilis</i> DSM 7299	3	4	4
<i>Arcobacter</i> sp L	3	4	4
<i>Aulosira laxa</i> NIES-50	4	6	6
<i>Aulosira laxa</i> NIES-50-core2	4	6	6
<i>Azoarcus communis</i> strain TSNA42	5	5	5
<i>Azoarcus communis</i> strain TSPY31	5	5	5
<i>Azoarcus olearius</i> strain DQS4	5	5	5
<i>Azoarcus</i> sp BH72	5	5	5
<i>Azoarcus</i> sp CIB	6	8	10
<i>Azoarcus</i> sp KH32C	5	8	10
<i>Azoarcus</i> sp KH32C-core2	6	7	8
<i>Azorhizobium caulinodans</i> ORS 571	6	7	9
<i>Azospira oryzae</i> PS	6	7	8
<i>Azospira restricta</i> strain DSM 18626	5	5	5
<i>Azospira</i> sp. I09	6	7	8
<i>Azospirillum brasilense</i> Az39	6	8	10
<i>Azospirillum brasilense</i> strain Sp7	6	8	10
<i>Azospirillum humicireducens</i> SgZ-5	6	8	10
<i>Azospirillum lipoferum</i> 4B	6	8	10
<i>Azospirillum</i> sp B510	6	8	10
<i>Azospirillum</i> sp CFH 70021	6	8	10
<i>Azospirillum</i> sp M2T2B2	6	8	10
<i>Azospirillum</i> sp TSH58	6	8	10
<i>Azospirillum thiophilum</i> strain BV-S	6	8	10
<i>Azotobacter chroococcum</i> B3	5	5	5
<i>Azotobacter chroococcum</i> NCIMB 8003	5	5	5
<i>Azotobacter vinelandii</i> CA	5	5	5
<i>Azotobacter vinelandii</i> CA6	5	5	5
<i>Azotobacter vinelandii</i> DJ	5	5	5
<i>Beggiatoa leptomitiformis</i> D-401	5	5	5
<i>Beggiatoa leptomitiformis</i> D-402	5	5	5
<i>Beijerinckia indica</i> subsp <i>indica</i> ATCC 9039	6	7	8
<i>Blastochloris viridis</i> 2015	7	9	11
<i>Blastochloris viridis</i> ATCC 19567	7	9	11
<i>Blastochloris viridis</i> DSM 133	7	9	11
<i>Bradyrhizobium amphicarpaeae</i> strain 39S1MB	6	7	9
<i>Bradyrhizobium arachidis</i> strain CCBAU 051107	6	7	9
<i>Bradyrhizobium diazoefficiens</i> NK6	6	7	9
<i>Bradyrhizobium diazoefficiens</i> USDA 110	6	7	9
<i>Bradyrhizobium diazoefficiens</i> USDA 110 2016	6	7	9
<i>Bradyrhizobium diazoefficiens</i> USDA 122	6	7	9

<i>Bradyrhizobium diazoefficiens</i> XF7	6	7	9
<i>Bradyrhizobium elkanii</i> USDA 61	6	7	9
<i>Bradyrhizobium guangdongense</i> strain CCBAU 51649	6	7	8
<i>Bradyrhizobium guangxiense</i> strain CCBAU 53363	6	7	9
<i>Bradyrhizobium guangzhouense</i> strain CCBAU 51670	6	7	8
<i>Bradyrhizobium icense</i> LMTR 13	6	7	9
<i>Bradyrhizobium japonicum</i> E109	6	7	9
<i>Bradyrhizobium japonicum</i> J5	6	7	9
<i>Bradyrhizobium japonicum</i> USDA 6	6	7	9
<i>Bradyrhizobium oligotrophicum</i> S58	6	7	9
<i>Bradyrhizobium ottawaense</i> OO99	6	7	9
<i>Bradyrhizobium</i> sp 2 39S1MB	6	7	9
<i>Bradyrhizobium</i> sp BTAi1	6	7	9
<i>Bradyrhizobium</i> sp CCGE-LA001	6	7	8
<i>Bradyrhizobium</i> sp ORS 278	6	7	9
<i>Bradyrhizobium</i> sp ORS 285	6	7	9
<i>Bradyrhizobium</i> sp ORS 3257	6	7	9
<i>Bradyrhizobium</i> sp S23321	6	7	9
<i>Bradyrhizobium vignae</i> isolate ORS3257	6	7	9
<i>Bradyrhizobium zhanjiangense</i> strain CCBAU 51778	6	7	9
<i>Brasilonema octagenarum</i> UFV-E1	4	6	6
<i>Brasilonema sennae</i> CENA114	4	6	6
<i>Brenneria goodwinii</i>	1	2	2
<i>Burkholderia</i> sp CCGE1001	6	7	8
<i>Burkholderia</i> sp CCGE1002	6	7	9
<i>Burkholderia</i> sp KJ006	6	7	9
<i>Burkholderia ubonensis</i> MSMB0783	6	7	9
<i>Burkholderia vietnamiensis</i> AU1233	6	7	9
<i>Burkholderia vietnamiensis</i> FDAARGOS 239	6	7	9
<i>Burkholderia vietnamiensis</i> FL-2-3-30-S1-D0	6	7	9
<i>Burkholderia vietnamiensis</i> G4	6	7	9
<i>Burkholderia vietnamiensis</i> HI2297	6	7	9
<i>Burkholderia vietnamiensis</i> LMG 10929	6	7	9
<i>Burkholderia vietnamiensis</i> MSMB608WGS	6	7	9
<i>Calditerrivibrio nitroreducens</i> DSM 19672	3	4	4
<i>Calothrix brevissima</i> NIES-22	4	6	6
<i>Calothrix brevissima</i> NIES-22-core2	4	6	7
<i>Calothrix</i> sp 3363	4	6	7
<i>Calothrix</i> sp NIES-2098	4	6	6
<i>Calothrix</i> sp NIES-3974	4	6	7
<i>Calothrix</i> sp NIES-4071	4	6	7
<i>Calothrix</i> sp NIES-4101	4	6	7
<i>Calothrix</i> sp PCC 6303	4	6	7
<i>Calothrix</i> sp PCC 7507	4	6	6
<i>Candidatus Atelocyanobacterium thalassa</i> isolate ALOHA	4	6	7
<i>Candidatus Azobacteroides pseudotrichonymphae</i>	2	3	3

<i>Candidatus Thiodictyon syntrophicum</i> Cad16T	5	5	5
<i>Carboxydocella thermautotrophica</i> 019	2	3	3
<i>Carboxydocella thermautotrophica</i> strain 41	2	3	3
<i>Chlorobaculum limnaeum</i> DSM 1677	2	3	3
<i>Chlorobaculum parvum</i> NCIB 8327	2	3	3
<i>Chlorobaculum tepidum</i> TLS	2	3	3
<i>Chlorobium chlorochromatii</i> CaD3	2	3	3
<i>Chlorobium limicola</i> DSM 245	2	3	3
<i>Chlorobium phaeobacteroides</i> BS1	2	3	3
<i>Chlorobium phaeobacteroides</i> DSM 266	2	3	3
<i>Chlorobium phaeovibrioides</i> DSM 265	2	3	3
<i>Chloroherpeton thalassium</i> ATCC 35110	2	3	3
<i>Chroococciopsis thermalis</i> PCC 7203	4	6	6
<i>Clostridium aceticum</i> DSM 1496	2	3	3
<i>Clostridium acetobutylicum</i> ATCC 824	2	3	3
<i>Clostridium acetobutylicum</i> DSM 1731	2	3	3
<i>Clostridium acetobutylicum</i> EA 2018	2	3	3
<i>Clostridium autoethanogenum</i> DSM 10061	2	3	3
<i>Clostridium autoethanogenum</i> DSM 10061-2	2	3	3
<i>Clostridium beijerinckii</i> ATCC 35702	2	3	3
<i>Clostridium beijerinckii</i> BASB31124	2	3	3
<i>Clostridium beijerinckii</i> DSM 6423	2	3	3
<i>Clostridium beijerinckii</i> NCIMB 14988	2	3	3
<i>Clostridium beijerinckii</i> NCIMB 8052	2	3	3
<i>Clostridium beijerinckii</i> NRRL B-598	2	3	3
<i>Clostridium bornimense</i> M2-40T	2	3	3
<i>Clostridium butyricum</i> 29-1	2	3	3
<i>Clostridium butyricum</i> 4-1	2	3	3
<i>Clostridium butyricum</i> CDC 51208	2	3	3
<i>Clostridium butyricum</i> JKY6D1	2	3	3
<i>Clostridium butyricum</i> KNU-L09	2	3	3
<i>Clostridium butyricum</i> TK520	2	3	3
<i>Clostridium butyricum</i> TOA	2	3	3
<i>Clostridium cellulovorans</i> 743B	2	3	3
<i>Clostridium clariflavum</i> DSM 19732	2	3	3
<i>Clostridium estertheticum</i> subsp <i>estertheticum</i> DSM 8809	2	3	3
<i>Clostridium formicaceticum</i> ATCC 27076	2	3	3
<i>Clostridium formicaceticum</i> DSM 92	2	3	3
<i>Clostridium kluyveri</i> DSM 555	2	3	3
<i>Clostridium kluyveri</i> JZZ	2	3	3
<i>Clostridium kluyveri</i> NBRC 12016	2	3	3
<i>Clostridium lentocellum</i> DSM 5427	2	3	3
<i>Clostridium ljungdahlii</i> DSM 13528	2	3	3
<i>Clostridium pasteurianum</i> BC1	2	3	3
<i>Clostridium pasteurianum</i> DSM 525 ATCC 6013 2015-2	2	3	3
<i>Clostridium pasteurianum</i> DSM 525 [ATCC 6013] 2015	2	3	3



<i>Clostridium pasteurianum</i> DSM 525 [ATCC 6013] 2016	2	3	3
<i>Clostridium pasteurianum</i> M150B	2	3	3
<i>Clostridium saccharoperbutylaceticum</i> N1-4[HMT]	2	3	3
<i>Clostridium saccharoperbutylaceticum</i> N1-504	2	3	3
<i>Clostridium</i> sp BNL1100	2	3	3
<i>Clostridium tyrobutyricum</i> strain KCTC 5387	2	3	3
<i>Clostridium tyrobutyricum</i> W428	2	3	3
<i>Confluentimicrobium</i> sp EMB200-NS6	6	8	10
<i>Coralimargarita akajimensis</i> DSM 45221	2	3	3
<i>Crocospaera subtropica</i> ATCC 51142	4	6	7
<i>Cupriavidus taiwanensis</i> str LMG19424	6	7	8
<i>cyanobacterium endosymbiont of Epithemia turgida isolate EtSB Lake Yunoko</i>	4	6	7
<i>Cyanobacterium endosymbiont of Rhopalodia gibberula</i>	4	6	7
<i>Cyanothece</i> sp ATCC 51142	4	6	7
<i>Cyanothece</i> sp PCC 7425	4	6	6
<i>Cyanothece</i> sp PCC 8801	4	6	7
<i>Cyanothece</i> sp PCC 8802	4	6	7
<i>Cylindrospermum stagnale</i> PCC 7417	4	6	6
<i>Dechloromonas aromatica</i> RCB	5	5	5
<i>Dechloromonas</i> sp. HYN0024	5	5	5
<i>Defluviicoccus</i> sp. isolate SSA4a	6	8	10
<i>Defluviicoccus vanus</i> strain Ben 114	6	8	10
<i>Defluviimonas alba</i> cai42	6	8	10
<i>Dehalobacter</i> sp CF	3	4	4
<i>Dehalobacter</i> sp DCA	3	4	4
<i>Dehalococcoides mccartyi</i> 195	2	3	3
<i>Dehalococcoides mccartyi</i> CG4	2	3	3
<i>Dehalogenimonas</i> sp WBC-2	2	3	3
<i>Denitrovibrio acetiphilus</i> DSM 12809	3	4	4
<i>Desulfitobacterium hafniense</i> DCB-2	3	4	4
<i>Desulfitobacterium hafniense</i> Y51	3	4	4
<i>Desulfobacca acetoxidans</i> DSM 11109	2	3	3
<i>Desulfobacterium autotrophicum</i> HRM2	2	3	3
<i>Desulfobulbus propionicus</i> DSM 2032	2	3	3
<i>Desulfocapsa sulfexigens</i> DSM 10523	2	3	3
<i>Desulfocurvibacter africanus</i> subsp africanus str Walvis Bay	2	3	3
<i>Desulfomicrobium baculatum</i> DSM 4028	2	3	3
<i>Desulfomonile tiedjei</i> DSM 6799	2	3	3
<i>Desulfosarcina alkanivorans</i> PL12a	2	3	3
<i>Desulfosarcina ovata</i> subsp ovata oXyS1a	2	3	3
<i>Desulfosporosinus acidiphilus</i> SJ4	3	4	4
<i>Desulfosporosinus meridiei</i> DSM 13257	3	4	4
<i>Desulfosporosinus orientis</i> DSM 765	3	4	4
<i>Desulfotomaculum acetoxidans</i> DSM 771	2	3	3
<i>Desulfotomaculum gibsoniae</i> DSM 7213	2	3	3
<i>Desulfotomaculum nigrificans</i> CO-1-SRB	2	3	3

<i>Desulfotomaculum reducens</i> MI-1	2	3	3
<i>Desulfotomaculum ruminis</i> DSM 2154	2	3	3
<i>Desulfovibrio desulfuricans</i> ND132	2	3	3
<i>Desulfovibrio gigas</i> DSM 1382 [ATCC 19364]	2	3	3
<i>Desulfovibrio magneticus</i> RS-1	2	3	3
<i>Desulfovibrio salexigens</i> DSM 2638	2	3	3
<i>Desulfovibrio vulgaris</i> DP4	2	3	3
<i>Desulfovibrio vulgaris</i> RCH1	2	3	3
<i>Desulfovibrio vulgaris</i> str Hildenborough	2	3	3
<i>Desulfovibrio vulgaris</i> str Miyazaki F	2	3	3
<i>Desulfurispirillum indicum</i> S5	3	4	4
<i>Desulfurivibrio alkaliphilus</i> AHT 2	2	3	3
<i>Desulfuromonas soudanensis</i> WTL	3	4	4
<i>Desulfuromonas</i> sp. DDH964	3	4	4
<i>Desutibacillum aliphaticivorans</i> AK-01	2	3	3
<i>Dickeya dadantii</i> 3937	1	2	2
<i>Dickeya dadantii</i> strain DSM 18020	1	2	2
<i>Dickeya dianthicola</i> RNS04 9	1	2	2
<i>Dickeya paradisiaca</i> Ech703	1	2	2
<i>Dickeya solani</i> D s0432-1	1	2	2
<i>Dickeya solani</i> IPO 2222	1	2	2
<i>Dickeya solani</i> PPO 9019	1	2	2
<i>Dickeya solani</i> RNS 08 23 3 1 A	1	2	2
<i>Dickeyangzhongdai</i> ND14b	1	2	2
<i>Dissulfurispira thermophila</i> T55J	3	4	4
<i>Dolichospermum compactum</i> NIES-806	4	6	6
<i>Dolichospermum flos-aquae</i> CCAP 1403 F13	4	6	6
<i>Dolichospermum</i> sp. UKL201	4	6	6
<i>Draconibacterium</i> sp M1	2	3	3
<i>Dysgonomonas capnocytophagoides</i> DSM 22835	2	3	3
<i>Ectothiorhodospira</i> sp BSL-9	5	5	5
<i>Ensifer sojae</i> CCBAU 05684	7	9	11
<i>Enterobacter</i> sp FY-07	1	2	2
<i>Enterobacter</i> sp R4-368	1	2	2
<i>Enterobacteriaceae</i> bacterium ENNIH1	1	2	2
<i>Ethanoligenens harbinense</i> strain W1	2	3	3
<i>Eubalthece natronophila</i> Z-M001	4	6	7
<i>Fischerella</i> sp NIES-3754-	4	6	6
<i>Frankia alni</i> ACN14a	3	4	4
<i>Frankia casuarinae</i> Ccl3	3	4	4
<i>Frankia</i> sp EAN1pec	3	4	4
<i>Frankia</i> symbiont of <i>Datisca glomerata</i>	3	4	4
<i>Fremyella</i> [Microchaete] diplosiphon NIES-3275	4	6	6
<i>Geobacter bemidjiensis</i> Bem	3	4	4
<i>Geobacter daltonii</i> FRC-32	3	4	4
<i>Geobacter lovleyi</i> SZ	3	4	4

<i>Geobacter metallireducens</i> GS-15	3	4	4
<i>Geobacter pickeringii</i> G13	3	4	4
<i>Geobacter</i> sp M18	3	4	4
<i>Geobacter</i> sp M21	3	4	4
<i>Geobacter sulfurreducens</i> KN400	3	4	4
<i>Geobacter sulfurreducens</i> PCA	3	4	4
<i>Geobacter uraniireducens</i> Rf4	3	4	4
<i>Geosporobacter ferrireducens</i> IRF9	2	3	3
<i>Gloeotheca citrififormis</i> PCC 7424	4	6	7
<i>Gloeotheca verrucosa</i> PCC 7822	4	6	7
<i>Gluconacetobacter diazotrophicus</i> PA1 5	7	9	11
<i>Gluconacetobacter diazotrophicus</i> PA1 5 2014	7	9	11
<i>Gynuella sunshinyii</i> YC6258	5	5	5
<i>Halorhodospira halochloris</i> DSM 1059	5	5	5
<i>Halorhodospira halophila</i> SL1	5	5	5
<i>Halothece</i> sp PCC 7418	4	6	7
<i>Halothiobacillus</i> sp LS2	6	7	8
<i>Hartmannibacter diazotrophicus</i> E19T	7	9	11
<i>Heliobacterium modesticaldum</i> Ice1	3	4	4
<i>Herbaspirillum rubrisubalbicans</i> M1	6	7	8
<i>Herbaspirillum seropedicae</i> SmR1	6	7	8
<i>Herbaspirillum seropedicae</i> strain Z67	6	7	8
<i>Hydrogenobacter thermophilus</i> TK-6 2010	3	4	4
<i>Hydrogenobacter thermophilus</i> TK-6 2011	3	4	4
<i>Hyphomicrobium</i> sp. MC1	6	7	9
<i>Ilyobacter polytropus</i> DSM 2926	2	3	3
<i>Klebsiella michiganensis</i> AR375	1	1	1
<i>Klebsiella michiganensis</i> E718	1	1	1
<i>Klebsiella michiganensis</i> HKOPL1	1	1	1
<i>Klebsiella michiganensis</i> K516	1	1	1
<i>Klebsiella michiganensis</i> K518	1	1	1
<i>Klebsiella michiganensis</i> KCTC 1686	1	1	1
<i>Klebsiella michiganensis</i> M1	1	1	1
<i>Klebsiella oxytoca</i> AR 0028	1	1	1
<i>Klebsiella oxytoca</i> AR 0147	1	1	1
<i>Klebsiella oxytoca</i> CAV1374	1	1	1
<i>Klebsiella oxytoca</i> CAV1752	1	1	1
<i>Klebsiella oxytoca</i> JKo3	1	1	1
<i>Klebsiella oxytoca</i> KONIH1	1	1	1
<i>Klebsiella oxytoca</i> KONIH2	1	1	1
<i>Klebsiella oxytoca</i> KONIH4	1	1	1
<i>Klebsiella oxytoca</i> KONIH5	1	1	1
<i>Klebsiella pneumoniae</i> AR 0142	1	1	1
<i>Klebsiella pneumoniae</i> YH43	1	1	1
<i>Klebsiella quasipneumoniae</i> ATCC 700603	1	1	1
<i>Klebsiella quasipneumoniae</i> CAV1947	1	1	1

<i>Klebsiella quasipneumoniae</i> CAV2013	1	1	1
<i>Klebsiella quasipneumoniae</i> CAV2018	1	1	1
<i>Klebsiella quasipneumoniae</i> subsp <i>quasipneumoniae</i> A708	1	1	1
<i>Klebsiella quasivariicola</i> KPN1705	1	1	1
<i>Klebsiella</i> sp. LTGPAF-6F	1	1	1
<i>Klebsiella</i> sp. M5a1	1	1	1
<i>Klebsiella variicola</i> 13450	1	1	1
<i>Klebsiella variicola</i> 342	1	1	1
<i>Klebsiella variicola</i> At-22	1	1	1
<i>Klebsiella variicola</i> DSM 15968	1	1	1
<i>Klebsiella variicola</i> DX120E	1	1	1
<i>Klebsiella variicola</i> E57-7	1	1	1
<i>Klebsiella variicola</i> GJ1	1	1	1
<i>Klebsiella variicola</i> GJ2	1	1	1
<i>Klebsiella variicola</i> GJ3	1	1	1
<i>Klebsiella variicola</i> KPN1481	1	1	1
<i>Klebsiella variicola</i> USDA-ARS	1	1	1
<i>Klebsiella variicola</i> WCHKP19	1	1	1
<i>Klebsiella variicola</i> WCHKV030666	1	1	1
<i>Kosakonia arachidis</i> strain KACC 18508	1	2	2
<i>Kosakonia oryzae</i> Ola 51	1	2	2
<i>Kosakonia radicincitans</i> DSM 16656 2019	1	2	2
<i>Kosakonia radicincitans</i> GXGL-4A	1	2	2
<i>Kosakonia sacchari</i> BO-1	1	2	2
<i>Kosakonia sacchari</i> SP1	1	2	2
<i>Kyrpidia spormannii</i> EA-1	3	4	4
<i>Leptolyngbya boryana</i> dg5	4	6	6
<i>Leptolyngbya boryana</i> IAM M-101	4	6	6
<i>Leptolyngbya boryana</i> NIES-2135	4	6	6
<i>Leptolyngbya</i> sp O-77	4	6	7
<i>Leptospirillum ferriphilum</i> YSK-	6	7	8
<i>Leptospirillum ferrooxidans</i> C2-3	6	7	8
<i>Leptothrix cholodnii</i> SP-6	6	7	8
<i>Magnetococcus marinus</i> MC-1	5	5	5
<i>Magnetospirillum gryphiswaldense</i> MSR-1	6	8	10
<i>Magnetospirillum gryphiswaldense</i> MSR-1 v2	6	8	10
<i>Magnetospirillum gryphiswaldense</i> R3-S1	6	8	10
<i>Magnetospirillum magneticum</i> AMB-1	6	8	10
<i>Magnetospirillum</i> sp ME-1	6	8	10
<i>Magnetospirillum</i> sp XM-1	6	8	10
<i>Marichromatium purpuratum</i> 984	5	5	5
<i>Marinifilaceae</i> bacterium SPP2	2	3	3
<i>Martelella endophytica</i> YC6887	5	5	5
<i>Mesorhizobium amorphae</i> CCNWGS0123	7	9	11
<i>Mesorhizobium australicum</i> WSM2073	7	9	11
<i>Mesorhizobium ciceri</i> biovar <i>biserrulae</i> WSM1271	7	9	11

<i>Mesorhizobium ciceri</i> biovar <i>biserrulae</i> WSM1284	7	9	11
<i>Mesorhizobium ciceri</i> CC1192	7	9	11
<i>Mesorhizobium erdmanii</i> strain NZP2014a	7	9	11
<i>Mesorhizobium huakuii</i> 7653R	7	9	11
<i>Mesorhizobium japonicum</i> MAFF 303099	7	9	11
<i>Mesorhizobium loti</i> NZP2037-	7	9	11
<i>Mesorhizobium opportunistum</i> WSM2075	7	9	11
<i>Mesorhizobium</i> sp WSM1497	7	9	11
<i>Methanobacterium congolense</i> isolate <i>Buetzberg</i>	3	3	4
<i>Methanobacterium formicicum</i> strain <i>BRM9</i>	3	3	4
<i>Methanobacterium lacus</i> AL-21	3	3	4
<i>Methanobacterium paludis</i> strain <i>SWAN1</i>	3	3	4
<i>Methanobacterium</i> sp <i>BAmetb5</i>	3	3	4
<i>Methanobacterium</i> sp <i>BRmetb2</i>	3	3	4
<i>Methanobacterium</i> sp <i>MB1</i>	3	3	4
<i>Methanobacterium</i> sp <i>MZ-A1</i>	3	3	4
<i>Methanobacterium subterraneum</i> strain <i>A8p</i>	3	3	4
<i>Methanocella arvoryzae</i> <i>MRE50</i>	2	3	3
<i>Methanocella conradii</i> <i>HZ254</i>	2	3	3
<i>Methanococcus aeolicus</i> <i>Nankai-3</i>	3	3	4
<i>Methanococcus maripaludis</i> <i>C5</i>	3	3	4
<i>Methanococcus vannielii</i> <i>SB</i>	3	3	4
<i>Methanolacinia petrolearia</i> <i>DSM 11571</i>	2	3	3
<i>Methanolobus psychrophilus</i> <i>R15</i>	2	3	3
<i>Methanolobus zinderi</i> strain <i>DSM 21339</i>	2	3	3
<i>Methanoregula boonei</i> <i>6A8</i>	2	3	3
<i>Methanosarcina acetivorans</i> <i>C2A</i>	2	3	3
<i>Methanosarcina barkeri</i> str <i>Fusaro</i>	2	3	3
<i>Methanosarcina horonobensis</i> <i>HB-1</i>	2	3	3
<i>Methanosarcina mazei</i> <i>Go1</i>	2	3	3
<i>Methanosarcina siciliae</i> <i>T4-M</i>	2	3	3
<i>Methanosarcina</i> sp <i>Kolksee</i>	2	3	3
<i>Methanosarcina</i> sp <i>WH1</i>	2	3	3
<i>Methanosarcina</i> sp <i>WWM596</i>	2	3	3
<i>Methanosarcina vacuolata</i> <i>Z-761</i>	2	3	3
<i>Methanosphaerula palustris</i> <i>E1-9c</i>	2	3	3
<i>Methanothermobacter marburgensis</i> str <i>Marburg</i>	3	3	4
<i>Methanothermobacter</i> sp <i>EMTCatA1</i>	3	3	4
<i>Methanothermobacter</i> sp <i>THM-2</i>	3	3	4
<i>Methanothermobacter</i> sp. <i>KEPCO-1</i>	3	3	4
<i>Methanothermobacter thermautotrophicus</i> str <i>Delta H</i>	3	3	4
<i>Methanothermobacter wolfeii</i> isolate <i>SIV6</i>	3	3	4
<i>Methanothermococcus okinawensis</i> <i>IH1</i>	3	3	4
<i>Methanotherx soehngenii</i> <i>GP6</i>	2	3	3
<i>Methanotorris igneus</i> <i>Kol 5</i>	3	3	4
<i>Methylacidiphilum fumarolicum</i> <i>SoIV</i>	6	7	8

<i>Methylacidiphilum inferorum</i> V4	6	7	8
<i>Methylacidiphilum kamchatkense</i> Kam1	6	7	8
<i>Methylacidiphilum</i> sp IT5	6	7	8
<i>Methylacidiphilum</i> sp IT6	6	7	8
<i>Methylobacterium nodulans</i> ORS 2060	7	9	11
<i>Methylobacterium</i> sp 4-46	7	9	11
<i>Methylocella silvestris</i> BL2	6	7	8
<i>Methylocella tundrae</i> isolate MTUNDRAET4	6	7	8
<i>Methylococcus capsulatus</i> str Bath	6	7	8
<i>Methylocystis bryophila</i> strain S285	6	7	9
<i>Methylocystis parvus</i> strain BRCS2	6	7	8
<i>Methylocystis</i> sp SC2	6	7	9
<i>Methylomonas denitrificans</i> FJG1	5	5	5
<i>Methylomonas koyamae</i> LM6	5	5	5
<i>Methylomonas methanica</i> MC09	5	5	5
<i>Methylomonas</i> sp DH-1	5	5	5
<i>Methylomusa anaerophila</i> strain MMFC1	3	3	4
<i>Methylosinus trichosporium</i> OB3b	6	7	8
<i>Methyloversatilis</i> sp RAC08	6	7	8
<i>Methylovulum psychrotolerans</i> HV10 M2	5	5	5
<i>Microcoleus</i> sp PCC 7113-	4	6	6
<i>Neorhizobium galegae</i> bv <i>officinalis</i> str HAMBI 1141	7	9	11
<i>Neorhizobium galegae</i> bv <i>orientalis</i> str HAMBI 540	7	9	11
<i>Nitrospirillum amazonense</i> CBAmc	6	7	8
<i>Nodularia spumigena</i> UHCC 0039	4	6	6
<i>Nostoc azollae</i> 0708	4	6	6
<i>Nostoc carneum</i> NIES-2107	4	6	6
<i>Nostoc carneum</i> NIES-2107-core2	4	6	6
<i>Nostoc flagelliforme</i> CCNUN1	4	6	6
<i>Nostoc linckia</i> NIES-25	4	6	6
<i>Nostoc piscinale</i> CENA21	4	6	6
<i>Nostoc punctiforme</i> PCC 73102	4	6	6
<i>Nostoc</i> sp [ <i>Lobaria pulmonaria</i> 5183 cyanobiont] 5183	4	6	6
<i>Nostoc</i> sp [ <i>Peltigera membranacea</i> cyanobiont] N6	4	6	6
<i>Nostoc</i> sp CENA543	4	6	6
<i>Nostoc</i> sp NIES-3756	4	6	6
<i>Nostoc</i> sp PCC 7107	4	6	6
<i>Nostoc</i> sp PCC 7120	4	6	6
<i>Nostoc</i> sp PCC 7524	4	6	6
<i>Nostocales cyanobacterium</i> HT-58-2	4	6	6
<i>Opiritaceae bacterium</i> TAV5	2	3	3
<i>Paenibacillus borealis</i> DSM 13188 2020	3	4	5
<i>Paenibacillus donghaensis</i> KCTC 13049	3	4	5
<i>Paenibacillus durus</i> ATCC 35681	3	4	5
<i>Paenibacillus durus</i> DSM 1735 2015	3	4	5
<i>Paenibacillus graminis</i> DSM 15220	3	4	5

<i>Paenibacillus kribbensis</i> AM49	3	4	5
<i>Paenibacillus odorifer</i> CBA7130	3	4	5
<i>Paenibacillus odorifer</i> DSM 15391	3	4	5
<i>Paenibacillus peoriae</i> HS311	3	4	5
<i>Paenibacillus polymyxa</i> ATCC 15970	3	4	5
<i>Paenibacillus polymyxa</i> CR1	3	4	5
<i>Paenibacillus polymyxa</i> J	3	4	5
<i>Paenibacillus polymyxa</i> Sb3-1	3	4	5
<i>Paenibacillus polymyxa</i> YC0136	3	4	5
<i>Paenibacillus polymyxa</i> YC0573	3	4	5
<i>Paenibacillus riograndensis</i> SBR5	3	4	5
<i>Paenibacillus sabinae</i> T27	3	4	5
<i>Paenibacillus</i> sp FSL H7-0357	3	4	5
<i>Paenibacillus</i> sp FSL P4-0081	3	4	5
<i>Paenibacillus</i> sp FSL R5-0912	3	4	5
<i>Paenibacillus</i> sp FSL R7-0273	3	4	5
<i>Paenibacillus</i> sp FSL R7-0331	3	4	5
<i>Paenibacillus</i> sp lzh-N1	3	4	5
<i>Paenibacillus stellifer</i> DSM 14472	3	4	5
<i>Paenibacillus terrae</i> HPL-003	3	4	5
<i>Paludibacter propionigenes</i> WB4	2	3	3
<i>Pantoea</i> sp At-9b	1	2	2
<i>Paraburkholderia aromaticivorans</i> BN5	6	7	8
<i>Paraburkholderia phenoliruptrix</i> BR3459a	6	7	8
<i>Paraburkholderia phymatum</i> STM815	6	7	8
<i>Paraburkholderia sprentiae</i> WSM5005	6	7	8
<i>Paraburkholderia xenovorans</i> LB400 2010	6	7	9
<i>Paraburkholderia xenovorans</i> LB400 2018	6	7	9
<i>Pararhodospirillum photometricum</i> DSM 122	6	8	10
<i>Pectobacterium atrosepticum</i> 21A	1	2	2
<i>Pectobacterium atrosepticum</i> 36A	1	2	2
<i>Pectobacterium atrosepticum</i> JG10-08	1	2	2
<i>Pectobacterium atrosepticum</i> SCRI1043	1	2	2
<i>Pectobacterium polaris</i> NIBIO1006	1	2	2
<i>Pelobacter acetylenicus</i> DSM 3246	3	4	4
<i>Pelobacter carbinolicus</i> DSM 2380	3	4	4
<i>Pelobacter propionicus</i> DSM 2379	3	4	4
<i>Pelobacter</i> sp SFB93	3	4	4
<i>Pelodictyon luteolum</i> DSM 273	2	3	3
<i>Pelodictyon phaeoclathratiforme</i> BU-1	2	3	3
<i>Pelosinus fermentans</i> JBW45	2	3	3
<i>Pelosinus</i> sp UFO1	2	3	3
<i>Pleurocapsa</i> sp PCC 7327	4	6	7
<i>Polaromonas naphthalenivorans</i> CJ2	6	7	8
<i>Prosthecochloris aestuarii</i> DSM 271	2	3	3
<i>Prosthecochloris</i> sp CIB 2401	2	3	3

<i>Prosthecochloris</i> sp GSB1 TY	2	3	3
<i>Prosthecochloris</i> sp HL-130-GSB	2	3	3
<i>Pseudoclostridium thermosuccinogenes</i> DSM 5807	2	3	3
<i>Pseudodesulfovibrio aespoeensis</i> Aspo-2	2	3	3
<i>Pseudodesulfovibrio indicus</i> J2	2	3	3
<i>Pseudodesulfovibrio profundus</i> 500-1	2	3	3
<i>Pseudomonas stutzeri</i> A1501	5	5	5
<i>Pseudomonas stutzeri</i> DSM 4166	5	5	5
<i>Rahnella aquatilis</i> HX2	1	2	2
<i>Raoultella ornithinolytica</i> S12	1	1	1
<i>Rhizobium acidisoli</i> strain FH23a	8	10	13
<i>Rhizobium bangladeshense</i> strain BLR175	7	9	11
<i>Rhizobium esperanzae</i> N561	8	10	13
<i>Rhizobium etli</i> 8C-3	8	10	13
<i>Rhizobium etli</i> bv <i>mimosae</i> str Mim1	8	10	13
<i>Rhizobium etli</i> bv <i>phaseoli</i> str IE4803	8	10	13
<i>Rhizobium etli</i> CFN 42	8	10	13
<i>Rhizobium etli</i> CIAT 652	8	10	13
<i>Rhizobium etli</i> NXC12 plasmid pRetNXC12d	8	10	13
<i>Rhizobium gallicum</i> bv <i>gallicum</i> R602	8	10	13
<i>Rhizobium gallicum</i> IE4872	8	10	13
<i>Rhizobium hidalgonense</i> strain JKLM 19Ea	7	9	11
<i>Rhizobium indicum</i> strain JKLM 12A2a	7	9	11
<i>Rhizobium jaguaris</i> strain CCGE525a	7	9	11
<i>Rhizobium leguminosarum</i> ATCC 14479	7	9	11
<i>Rhizobium leguminosarum</i> bv <i>trifolii</i> CB782	7	9	11
<i>Rhizobium leguminosarum</i> bv <i>trifolii</i> WSM1325	7	9	11
<i>Rhizobium leguminosarum</i> bv <i>trifolii</i> WSM1689	7	9	11
<i>Rhizobium leguminosarum</i> bv <i>trifolii</i> WSM2304	7	9	11
<i>Rhizobium leguminosarum</i> bv <i>viciae</i> 3841	7	9	11
<i>Rhizobium leguminosarum</i> bv <i>viciae</i> BIHB 1217	7	9	11
<i>Rhizobium leguminosarum</i> bv <i>viciae</i> UPM791	7	9	11
<i>Rhizobium leguminosarum</i> Norway	7	9	11
<i>Rhizobium leguminosarum</i> Vaf10	7	9	11
<i>Rhizobium leguminosarum</i> Vaf-108	7	9	11
<i>Rhizobium lentis</i> strain BLR27a	7	9	11
<i>Rhizobium phaseoli</i> Brasil 5	8	10	13
<i>Rhizobium phaseoli</i> N161	8	10	13
<i>Rhizobium phaseoli</i> N261 plasmid pRphaN261c	8	10	13
<i>Rhizobium phaseoli</i> N671 plasmid pRphaN671d	8	10	13
<i>Rhizobium phaseoli</i> N771	8	10	13
<i>Rhizobium phaseoli</i> N831	8	10	13
<i>Rhizobium phaseoli</i> N841	8	10	13
<i>Rhizobium phaseoli</i> N931	8	10	13
<i>Rhizobium phaseoli</i> R611	8	10	13
<i>Rhizobium phaseoli</i> R620	8	10	13



<i>Rhizobium phaseoli</i> R630	8	10	13
<i>Rhizobium phaseoli</i> R650	8	10	13
<i>Rhizobium phaseoli</i> R723	8	10	13
<i>Rhizobium pusense</i> IRBG74a	7	9	11
<i>Rhizobium pusense</i> strain 76 a	7	9	11
<i>Rhizobium</i> sp ACO-34A	7	9	11
<i>Rhizobium</i> sp CIAT894	8	10	13
<i>Rhizobium</i> sp IE4771	8	10	13
<i>Rhizobium</i> sp Kim5	8	10	13
<i>Rhizobium</i> sp N113	8	10	13
<i>Rhizobium</i> sp N1314	8	10	13
<i>Rhizobium</i> sp N1341	8	10	13
<i>Rhizobium</i> sp N324	8	10	13
<i>Rhizobium</i> sp N621	8	10	13
<i>Rhizobium</i> sp N6212	8	10	13
<i>Rhizobium</i> sp N731	8	10	13
<i>Rhizobium</i> sp N741	8	10	13
<i>Rhizobium</i> sp N871	8	10	13
<i>Rhizobium</i> sp NXC14	8	10	13
<i>Rhizobium</i> sp NXC24	7	9	11
<i>Rhizobium</i> sp TAL182	8	10	13
<i>Rhizobium tropici</i> CIAT 899	7	9	11
<i>Rhodobacter blasticus</i> 28-5	6	8	10
<i>Rhodobacter capsulatus</i> SB 1003	6	8	10
<i>Rhodobacter sphaeroides</i> 241	6	8	10
<i>Rhodobacter sphaeroides</i> ATCC 17025	6	8	10
<i>Rhodobacter sphaeroides</i> ATCC 17029	6	8	10
<i>Rhodobacter sphaeroides</i> KD131	6	8	10
<i>Rhodobacter sphaeroides</i> MBTLJ-13	6	8	10
<i>Rhodobacter sphaeroides</i> MBTLJ-20	6	8	10
<i>Rhodobacter sphaeroides</i> MBTLJ-8	6	8	10
<i>Rhodocyclaceae</i> bacterium <i>Thauera</i> -like	5	5	5
<i>Rhodoferax antarcticus</i> DSM 24876	6	7	8
<i>Rhodomicrobium vanniellii</i> ATCC 17100	6	7	8
<i>Rhodopseudomonas palustris</i> BisA53	6	7	8
<i>Rhodopseudomonas palustris</i> BisB18	6	7	8
<i>Rhodopseudomonas palustris</i> BisB5	6	7	8
<i>Rhodopseudomonas palustris</i> CGA009	6	7	8
<i>Rhodopseudomonas palustris</i> DX-1	6	7	8
<i>Rhodopseudomonas palustris</i> HaA2	6	7	8
<i>Rhodopseudomonas palustris</i> PS3	6	7	8
<i>Rhodopseudomonas palustris</i> TIE-1	6	7	8
<i>Rhodopseudomonas palustris</i> YSC3	6	7	8
<i>Rhodospirillum centenum</i> SW	7	9	11
<i>Rhodospirillum rubrum</i> ATCC 11170	6	8	10
<i>Rhodospirillum rubrum</i> F11	6	8	10

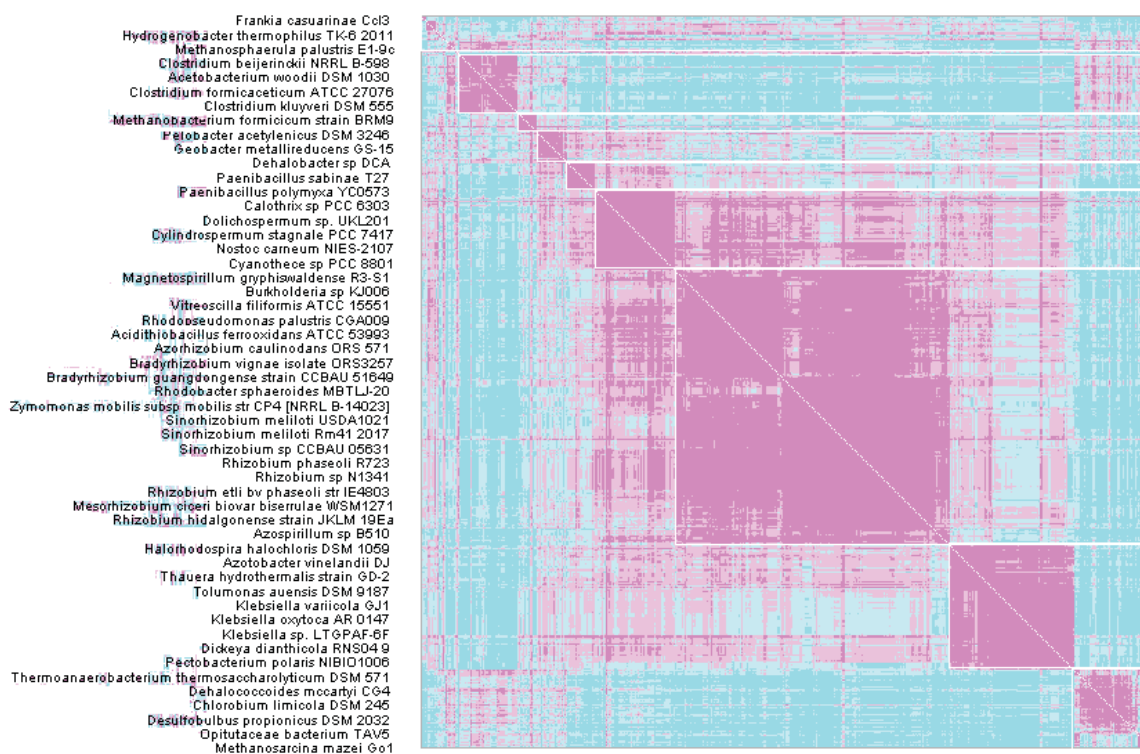
<i>Rhodovulum</i> sp MB263	7	8	10
<i>Rhodovulum</i> sp P5	7	8	10
<i>Rhodovulum sulfidophilum</i> DSM 1374	7	8	10
<i>Rhodovulum sulfidophilum</i> DSM 2351	7	8	10
<i>Rhodovulum sulfidophilum</i> SNK001	7	8	10
<i>Rippkaea orientalis</i> PCC 8801	4	6	7
<i>Rippkaea orientalis</i> PCC 8802	4	6	7
<i>Rivularia</i> sp PCC 7116	4	6	7
<i>Rubrivivax gelatinosus</i> IL144	6	7	8
<i>Sagittula</i> sp P11	6	8	10
<i>Scytonema</i> sp NIES-4073	4	6	6
<i>Sedimenticola thiotaurini</i> SIP-G1	5	5	5
<i>Sediminispirochaeta smaragdinae</i> DSM 11293	2	3	3
<i>Serratia</i> sp ATCC 39006 2020	1	2	2
<i>Sideroxydans lithotrophicus</i> ES-1	5	5	5
<i>Sinorhizobium americanum</i> CCGM7	8	10	13
<i>Sinorhizobium americanum</i> CFNEI 73	8	10	13
<i>Sinorhizobium fredii</i> CCBAU 25509	7	9	11
<i>Sinorhizobium fredii</i> CCBAU 45436	7	9	11
<i>Sinorhizobium fredii</i> CCBAU 83666	7	9	11
<i>Sinorhizobium fredii</i> HH103	7	9	11
<i>Sinorhizobium fredii</i> NGR234	7	9	11
<i>Sinorhizobium fredii</i> NXT3	8	10	13
<i>Sinorhizobium medicae</i> WSM419	7	9	12
<i>Sinorhizobium meliloti</i> 1021	7	9	12
<i>Sinorhizobium meliloti</i> 2011	7	9	12
<i>Sinorhizobium meliloti</i> AK83	7	9	12
<i>Sinorhizobium meliloti</i> B399	7	9	12
<i>Sinorhizobium meliloti</i> B401	7	9	12
<i>Sinorhizobium meliloti</i> BL225C	7	9	12
<i>Sinorhizobium meliloti</i> CCMM B554	7	9	12
<i>Sinorhizobium meliloti</i> GR4	7	9	12
<i>Sinorhizobium meliloti</i> HM006	7	9	12
<i>Sinorhizobium meliloti</i> KH35c	7	9	12
<i>Sinorhizobium meliloti</i> KH46	7	9	12
<i>Sinorhizobium meliloti</i> M270	7	9	12
<i>Sinorhizobium meliloti</i> Rm41 2015	7	9	12
<i>Sinorhizobium meliloti</i> Rm41 2017	7	9	12
<i>Sinorhizobium meliloti</i> RMO17	7	9	12
<i>Sinorhizobium meliloti</i> RU11001	7	9	12
<i>Sinorhizobium meliloti</i> SM11	7	9	12
<i>Sinorhizobium meliloti</i> T073	7	9	12
<i>Sinorhizobium meliloti</i> USDA1021	7	9	12
<i>Sinorhizobium meliloti</i> USDA1106	7	9	12
<i>Sinorhizobium meliloti</i> USDA1157	7	9	12
<i>Sinorhizobium</i> sp CCBAU 05631	7	9	11

<i>Skermanella pratensis</i> strain W17a	6	8	10
<i>Skermanella</i> sp TT6a	6	8	10
<i>Sphaerospermopsis kisseleviana</i> NIES-73	4	6	6
<i>Sphingomonas hengshuiensis</i> WHSC-8	6	7	8
<i>Spirochaeta thermophila</i> DSM 6192a	2	3	3
<i>Sulfurospirillum barnesii</i> SES-3	3	4	4
<i>Sulfurospirillum cavolei</i> strain UCH003	3	4	4
<i>Sulfurospirillum halorespirans</i> DSM 13726	3	4	4
<i>Sulfurospirillum multivorans</i> DSM 12446 2020	3	4	4
<i>Sulfurospirillum</i> sp JPD-1	3	4	4
<i>Sulfurospirillum</i> sp SL2-1	3	4	4
<i>Synechococcus</i> sp JA-2-3Ba[2-13]	4	6	7
<i>Synechococcus</i> sp JA-3-3Ab	4	6	7
<i>Syntrophobacter fumaroxidans</i> MPOB	2	3	3
<i>Syntrophobotulus glycolicus</i> DSM 8271	3	4	4
<i>Teredinibacter turnerae</i> T7901	5	5	5
<i>Thauera hydrothermalis</i> strain GD-2	5	5	5
<i>Thermincola potens</i> JR	2	3	3
<i>Thermoanaerobacterium aotearoense</i> strain SCUT27	2	3	3
<i>Thermoanaerobacterium saccharolyticum</i> JWSL-YS485	2	3	3
<i>Thermoanaerobacterium thermosaccharolyticum</i> DSM 571	2	3	3
<i>Thermoanaerobacterium thermosaccharolyticum</i> M0795	2	3	3
<i>Thermoanaerobacterium thermosaccharolyticum</i> TG57	2	3	3
<i>Thermocrinis albus</i> DSM 14484	3	4	4
<i>Thermodesulfovibrio yellowstonii</i> DSM 11347	3	4	4
<i>Thermoleptolyngbya</i> sp. PKUAC-SCTA183	4	6	7
<i>Thiocystis violascens</i> DSM 198	5	5	5
<i>Thioploca ingrlica</i>	5	5	5
<i>Thiospirochaeta perfilievii</i> strain Pa	2	3	3
<i>Tolomonas auensis</i> DSM 9187	1	2	2
<i>Tolypothrix tenuis</i> PCC 7101	4	6	6
<i>Tolypothrix tenuis</i> PCC 7101-core2	4	6	6
<i>Treponema azotonutricium</i> ZAS-9	2	3	3
<i>Treponema primitia</i> ZAS-2	2	3	3
<i>Trichodesmium erythraeum</i> IMS101	4	6	7
<i>Trichormus [Anabaena] variabilis</i> ATCC 29413	4	6	6
<i>Trichormus [Anabaena] variabilis</i> ATCC 29413-core2	4	6	6
<i>Vibrio cincinnatiensis</i>	5	5	5
<i>Vibrio natriegens</i> CCUG 16371	5	5	5
<i>Vibrio natriegens</i> CCUG 16373	5	5	5
<i>Vibrio natriegens</i> CCUG 16374	5	5	5
<i>Vibrio natriegens</i> NBRC 15636	5	5	5
<i>Vibrio</i> sp EJY3	5	5	5
<i>Vitreoscilla filiformis</i> ATCC 15551	6	7	8
<i>Wolinella succinogenes</i> DSM 1740	3	4	4
<i>Xanthobacter autotrophicus</i> Py2	6	7	9

<i>Yangia sp CCB-MM3</i>	6	8	10
<i>Zoogloeaceae bacterium Par-f-2</i>	5	5	5
<i>Zymomonas mobilis subsp mobilis ATCC 10988</i>	7	8	10
<i>Zymomonas mobilis subsp mobilis NCIMB 11163</i>	7	8	10
<i>Zymomonas mobilis subsp mobilis NRRL B-12526</i>	7	8	10
<i>Zymomonas mobilis subsp mobilis NRRL B-1960</i>	7	8	10
<i>Zymomonas mobilis subsp mobilis str CP4 [NRRL B-14023]</i>	7	8	10
<i>Zymomonas mobilis subsp mobilis str CP4 [NRRL B-14023]-2</i>	7	8	10
<i>Zymomonas mobilis subsp mobilis ZM4 [ATCC 31821]</i>	7	8	10
<i>Zymomonas mobilis subsp mobilis ZM4 substr 2032</i>	7	8	10
<i>Zymomonas mobilis subsp mobilis ZM4 substr 8b</i>	7	8	10

FONTE: O Autor (2022).

FIGURA 6 – MATRIZ DE DISSIMILARIDADE ORDENADA PARA NIF CORE.



FONTE: O Autor (2022).

## 7. MATRIZ DE DISTÂNCIA

FIGURA 7 – SCRIPT *COLDISS* (GILLET, 2009.)

```

# coldiss()
# Color plots of a dissimilarity matrix, without and with ordering
#
# License: GPL-2
# Author: Francois Gillet, August 2009
#
# cálculo da distancia euclidiana mdist <- dist(Funicore_sweep_matrix)
library(gclus)
library(cluster)
#
"coldiss" <- function(D, nc = 8, byrank = F, diag = FALSE)
{
  require(gclus)

  if (max(D)>1) D <- D/max(D)

  if (byrank) {
    spe.color = dmat.color(1-D, cm.colors(nc))
  }
  else {
    spe.color = dmat.color(1-D, byrank=FALSE, cm.colors(nc))
  }

  spe.o = order.single(1-D)
  speo.color = spe.color[spe.o, spe.o]

  op = par(mfrow=c(1,2), pty="s")

  if (diag) {
    plotcolors(spe.color, rlabels=attributes(D)$Labels,
               main="Dissimilarity Matrix",
               dlabels=attributes(D)$Labels)
    plotcolors(speo.color, rlabels=attributes(D)$Labels[spe.o],
               main="Ordered Dissimilarity Matrix",
               dlabels=attributes(D)$Labels[spe.o])
  }
  else {
    plotcolors(spe.color, rlabels=attributes(D)$Labels,
               main="Dissimilarity Matrix")
    plotcolors(speo.color, rlabels=attributes(D)$Labels[spe.o],
               main="Ordered Dissimilarity Matrix")
  }

  par(op)
}

# Usage:
# coldiss(D = dissimilarity.matrix, nc = 4, byrank = TRUE, diag = FALSE)
# If D is not a dissimilarity matrix (max(D) > 1), then D is divided by max(D)
# nc                number of colours (classes)
# byrank= TRUE      equal-sized classes
# byrank= FALSE     equal-length intervals
# diag = TRUE       print object labels also on the diagonal

# Example:
# coldiss(spe.dj, nc=9, byrank=F, diag=T)

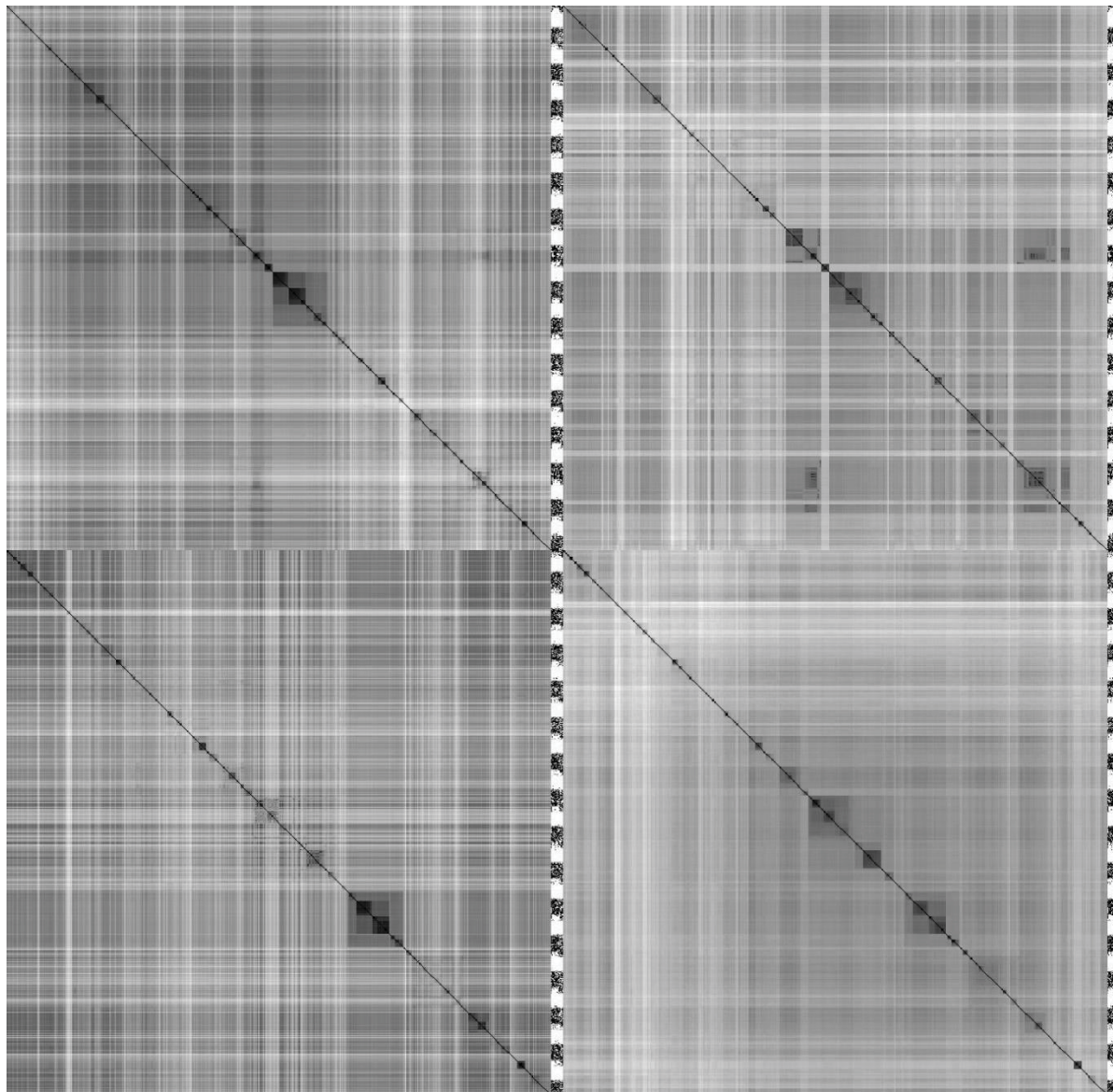
```

FONTE: Gillet (2009).

Todos os mapas de calor (Heatmaps) gerados neste estudo podem ser consultados no Gdrive atrás do link:

[https://drive.google.com/drive/folders/1lzGNOG0\\_OITFx\\_ZTWZ7TaoKzpCV00\\_Rm?usp=sharing](https://drive.google.com/drive/folders/1lzGNOG0_OITFx_ZTWZ7TaoKzpCV00_Rm?usp=sharing).

FIGURA 8 – MAPA DE CALOR 2X2 EM 4 BLOCOS



FONTE: O Autor (2022).

LEGENDA: Matriz (4x4) obtida para Nif *core* x Proteoma total e do proteoma total x Nif *core* (matrizes à esquerda) e ao lado (direita) são as matrizes ordenadas.

## 8. FILOGENIA SWEEP

As filogenias Nif *core*, Proteoma, núcleo (*cluster*) *nif* podem ser consultadas pelo Gdrive em:

<https://drive.google.com/drive/folders/1NaWA5PB4n3u2mTh9pKasYHDTGsu-06ec?usp=sharing>. Nesta pasta temos: as árvores obtidas (pasta TREE).

## ANEXO

FIGURA 9 - CERTIFICADO DE REGISTRE DE SOFTWARE PARA NIFFINDER



## Certificado de Registro de Programa de Computador

Processo Nº: **BR512021001201-1**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1º de janeiro subsequente à data de 10/04/2021, em conformidade com o §2º, art. 2º da Lei 9.609, de 19 de Fevereiro de 1998.

**Título:** NiffINDER - Rede Neuronal Artificial aplicada à busca de proteínas Nif

**Data de publicação:** 10/04/2021

**Data de criação:** 10/04/2021

**Titular(es):** UNIVERSIDADE FEDERAL DO PARANA

**Autor(es):** ROBERTO TADEU RAITTZ; JERONIZA NUNES MARCHAUKOSKI; BRUNO THIAGO DE LIMA NICHIO; FABIO DE OLIVEIRA PEDROSA; ROXANA BEATRIZ RIBEIRO CHAVES

**Linguagem:** MATLAB

**Campo de aplicação:** BL-01; BL-02; BL-04; BL-07

**Tipo de programa:** FA-01; FA-04; GI-04; IT-04

**Algoritmo hash:** SHA-512

**Resumo digital hash:**  
 272E1B7DAD340839ED8DD9D9704FCA2BFC9C7D9A7C59A13648F356A6118D623E52BC3BD2670FA1E883FE03F  
 141480DC2A8504A3C1F1A307F9BB5CE678CE571DA

**Expedido em:** 08/06/2021

**Aprovado por:**  
 Carlos Alexandre Fernandes Silva  
 Chefe da DIPTO

FONTE: O Autor (2022).



FIGURA 10 -CERTIFICADO DE REGISTRE DE SOFTWARE PARA GBK2CLUSTER







**REPÚBLICA FEDERATIVA DO BRASIL**  
 MINISTÉRIO DA ECONOMIA  
**INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL**  
 DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS INTEGRADOS

## Certificado de Registro de Programa de Computador

Processo Nº: **BR512019001240-2**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1º de janeiro subsequente à data de 22/04/2019, em conformidade com o §2º, art. 2º da Lei 9.609, de 19 de Fevereiro de 1998.

**Título:** GBK2CLUSTERS

**Data de publicação:** 22/04/2019

**Data de criação:** 22/04/2019

**Titular(es):** UNIVERSIDADE FEDERAL DO PARANA

**Autor(es):** ROBERTO TADEU RAITTZ; BRUNO THIAGO DE LIMA NICHIO; RAFAEL PRAZERES DE SOUZA LOBO

**Linguagem:** MATLAB

**Campo de aplicação:** BL-01; BL-02; BL-04; BL-07

**Tipo de programa:** TC-01; UT-01

**Algoritmo hash:** SHA-512

**Resumo digital hash:**  
 8F5B382256968835CF5CA027665C2FE9B1BFAB2630F25135982687310725B02C0A7CA6D3A8F1B1FE0839E15499  
 30C1A98735CCB639F909811BAC8E750A439CEC

**Expedido em:** 25/06/2019



FONTE: O Autor (2022).