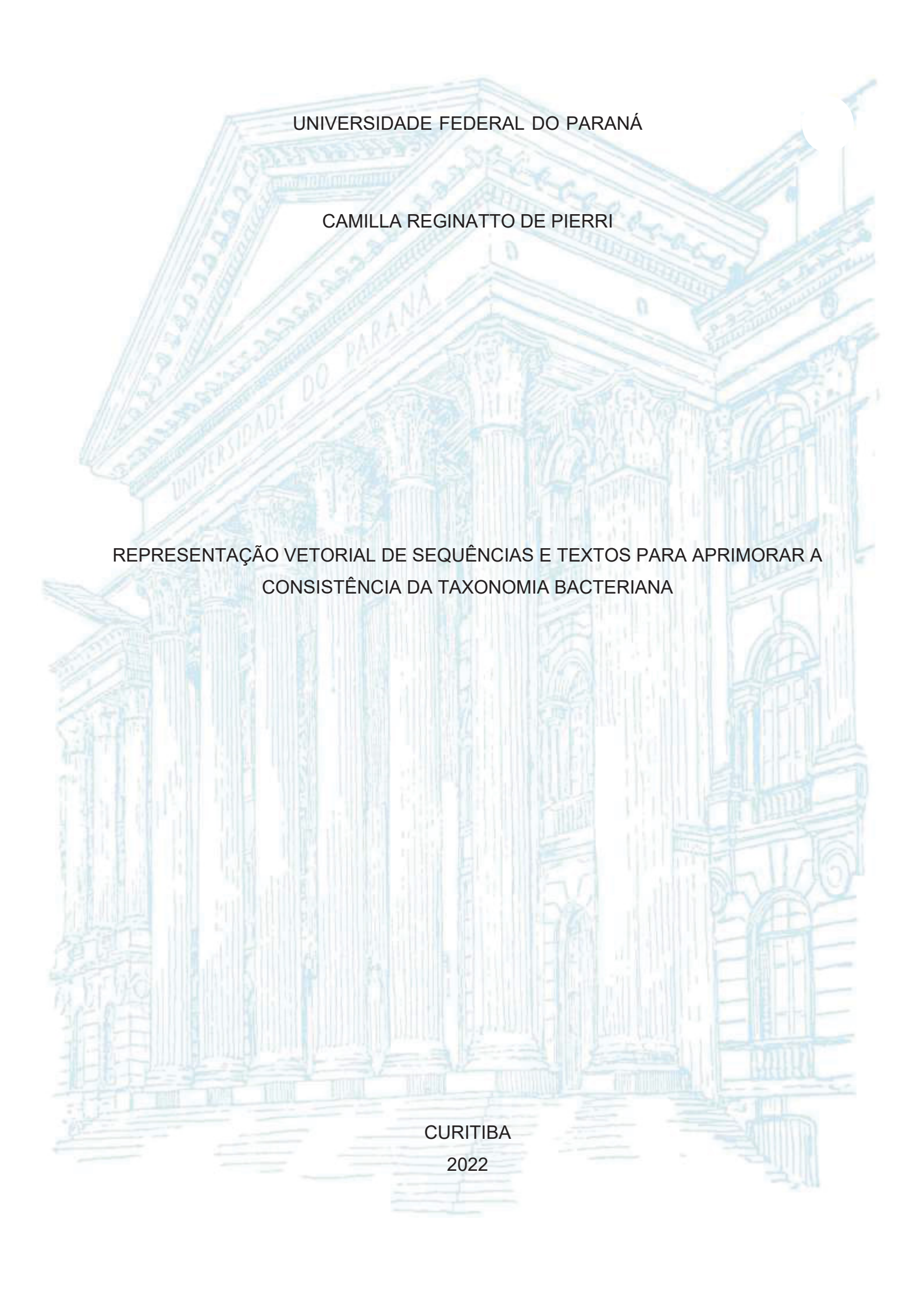




UNIVERSIDADE FEDERAL DO PARANÁ

CAMILLA REGINATTO DE PIERRI

REPRESENTAÇÃO VETORIAL DE SEQUÊNCIAS E TEXTOS PARA APRIMORAR A
CONSISTÊNCIA DA TAXONOMIA BACTERIANA

CURITIBA
2022

CAMILLA REGINATTO DE PIERRI

REPRESENTAÇÃO VETORIAL DE SEQUÊNCIAS E TEXTOS PARA APRIMORAR A
CONSISTÊNCIA DA TAXONOMIA BACTERIANA

Tese apresentada ao Programa de Pós-Graduação em Ciências – Bioquímica, Setor de Ciências Biológicas, Universidade Federal do Paraná, como requisito parcial para obtenção do título de Doutora em Ciências - Bioquímica.

Orientador: Prof. Dr. Fabio de Oliveira Pedrosa

Coorientador: Prof. Dr. Roberto Tadeu Raittz

CURITIBA

2022

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

Pierri, Camilla Reginatto de

Representação vetorial de sequências e textos para aprimorar a consistência da taxonomia bacteriana / Camilla Reginatto de Pierri. – Curitiba, 2022.

1 recurso on-line : PDF.

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

Orientador: Prof. Dr. Fabio de Oliveira Pedrosa.

Coorientador: Prof. Dr. Roberto Tadeu Raittz.

1. Bioinformática. 2. Inteligência artificial. 3. Mineração de dados (Computação). 4. Taxonomia. 5. Bactérias. 6. Nitrogênio – Fixação. I. Pedrosa, Fabio O.,-1947-. II. Raittz, Roberto Tadeu, 1966. III. Universidade Federal do Paraná. Setor de Ciências Biológicas. Programa de Pós-Graduação em Ciências (Bioquímica). IV. Título.

Bibliotecária: Giana Mara Seniski Silva. CRB-9/1406



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

ATA Nº337

ATA DE SESSÃO PÚBLICA DE DEFESA DE DOUTORADO PARA A OBTENÇÃO DO GRAU DE DOUTORA EM CIÊNCIAS (BIOQUÍMICA)

No dia dois de maio de dois mil e vinte e dois às 14:00 horas, na sala virtual, Plataforma Microsoft Teams, foram instaladas as atividades pertinentes ao rito de defesa de tese da doutoranda **CAMILLA REGINATTO DE PIERRI**, intitulada: **Representação Vetorial de Sequências e Textos para Aprimorar a Consistência da Taxonomia Bacteriana**, sob orientação do Prof. Dr. FABIO DE OLIVEIRA PEDROSA. A Banca Examinadora, designada pelo Colegiado do Programa de Pós-Graduação CIÊNCIAS (BIOQUÍMICA) da Universidade Federal do Paraná, foi constituída pelos seguintes Membros: FABIO DE OLIVEIRA PEDROSA (UNIVERSIDADE FEDERAL DO PARANÁ), ALEXANDRE ROSSI PASCHOAL (UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ), EMANUEL MALTEMPI DE SOUZA (UNIVERSIDADE FEDERAL DO PARANÁ), HELISSON FAORO (INSTITUTO CARLOS CHAGAS), DIEVAL GUIZELINI (UNIVERSIDADE FEDERAL DO PARANÁ). A presidência iniciou os ritos definidos pelo Colegiado do Programa e, após exarados os pareceres dos membros do comitê examinador e da respectiva contra argumentação, ocorreu a leitura do parecer final da banca examinadora, que decidiu pela APROVAÇÃO. Este resultado deverá ser homologado pelo Colegiado do programa, mediante o atendimento de todas as indicações e correções solicitadas pela banca dentro dos prazos regimentais definidos pelo programa. A outorga de título de doutora está condicionada ao atendimento de todos os requisitos e prazos determinados no regimento do Programa de Pós-Graduação. Nada mais havendo a tratar a presidência deu por encerrada a sessão, da qual eu, FABIO DE OLIVEIRA PEDROSA, lavrei a presente ata, que vai assinada por mim e pelos demais membros da Comissão Examinadora.

CURITIBA, 02 de Maio de 2022.

Assinatura Eletrônica

20/05/2022 17:50:50.0

FABIO DE OLIVEIRA PEDROSA

Presidente da Banca Examinadora

Assinatura Eletrônica

20/06/2022 10:01:58.0

ALEXANDRE ROSSI PASCHOAL

Avaliador Externo (UNIVERSIDADE TECNOLÓGICA FEDERAL DO

Assinatura Eletrônica

03/05/2022 10:29:46.0

EMANUEL MALTEMPI DE SOUZA

Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ) PARANÁ)

Assinatura Eletrônica

04/05/2022 13:35:14.0

HELISSON FAORO

Avaliador Externo (INSTITUTO CARLOS CHAGAS)

Assinatura Eletrônica

03/05/2022 10:37:12.0

DIEVAL GUIZELINI

Avaliador Externo (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: 180810

Para autenticar este documento/assinatura, acesse <https://www.prppg.ufpr.br/siga/visitante/autenticacaoassinaturas.jsp> e insira o código 180810



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 **CAMILLA REGINATTO DE PIERRI** intitulada: **Representação Vetorial de Sequências e Textos para Aprimorar a Consistência da Taxonomia Bacteriana**, sob orientação do Prof. Dr. FABIO DE OLIVEIRA PEDROSA, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de doutora 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, 02 de Maio de 2022.

Assinatura Eletrônica

20/05/2022 17:50:50.0

FABIO DE OLIVEIRA PEDROSA
Presidente da Banca Examinadora

Assinatura Eletrônica

20/06/2022 10:01:58.0

ALEXANDRE ROSSI PASCHOAL

Avaliador Externo (UNIVERSIDADE TECNOLÓGICA FEDERAL DO

Assinatura Eletrônica

03/05/2022 10:29:46.0

EMANUEL MALTEMPI DE SOUZA

Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ) PARANÁ)

Assinatura Eletrônica

04/05/2022 13:35:14.0

HELISSON FAORO

Avaliador Externo (INSTITUTO CARLOS CHAGAS)

Assinatura Eletrônica

03/05/2022 10:37:12.0

DIEVAL GUIZELINI

Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

*Dedico esta tese à
Maria Aparecida Lopes De Pierri
Delise Tesser Reginatto*

(In memoriam)

AGRADECIMENTOS

Aos meus Orientadores

Professor Dr. Fabio de Oliveira Pedrosa

Professor Dr. Roberto Tadeu Raittz

À minha Supervisora no exterior

Professora Dra. Maricel Kann

The James Hutton Institute

Professor Dr. Euan James

Aos membros da banca avaliadora

À minha família

Minha mãe, meu pai e minha irmã

Ao meu esposo e companheiro de pesquisa

Dr. Antonio Camilo da Silva Filho

Aos meus amigos

Me. Diogo de Jesus Soares Machado

Me. Camila Pereira Perico

Dr. Ricardo Voyceik

Professora Dra. Jeroniza Nunes Marchaukoski

Rosângela Raittz

Ao colegiado do Programa de Pós-graduação em Ciências – Bioquímica da UFPR

Aos membros do Laboratório de Inteligência Artificial aplicada à Bioinformática

Às Instituições

Universidade Federal do Paraná

University of Maryland – UMBC

Aos órgãos de fomento

CAPES

Fundação Araucária

“The view is often defended that sciences should be built up on clear and sharply defined basal concepts. In actual fact, no science, not even the most exact, begins with such definitions. The true beginning of scientific activity consists rather in describing phenomena and then in proceeding to group, classify and correlate them.”

Sigmund Freud

RESUMO

A grande quantidade de dados genômicos e literatura científica depositada nos bancos de dados atualmente leva ao entendimento de que as características taxonômicas de um organismo podem ser construídas a partir de uma estrutura evolutiva abrangente. Entretanto, mesmo com um legado importante de supervisão humana na taxonomia em geral e, particularmente, na taxonomia de bactérias, a classificação dos organismos nem sempre é consistente com os agrupamentos inferidos pela análise de dados moleculares. Além disso, não existem ferramentas para integrar dados de sequenciamento molecular e literatura científica, o que aponta para a fragilidade no suporte dos modelos taxonômicos atuais. Essa dificuldade é problemática e foi relatada em diversos estudos recentes. As técnicas de Mineração de dados (MD) e Mineração de textos (MT) baseadas em representação vetorial de sequências biológicas e de textos da literatura científica podem viabilizar a integração de forma consistente da expertise humana com informações moleculares para criar modelos taxonômicos. Nesta perspectiva, o objetivo desta tese é propor um modelo de suporte à taxonomia bacteriana baseado em técnicas de MD e MT para análise de Big Data. Este modelo explora os padrões de dados moleculares disponíveis nas bases de dados públicos, orientado por diretrizes gerais dos especialistas humanos. Foram utilizadas tecnologias vetoriais desenvolvidas *in-house* para explorar o Big Data na área biológica, tendo como principal elemento a representação vetorial. A nova metodologia proposta explora a união dos modelos SWeeP e BIOTEXT, que permite a comparação e exploração de dados genômicos com agilidade além de possibilitar execução de mineração de texto usando o arsenal de ferramentas de bioinformática. Foram conduzidos estudos de caso do domínio Bacteria, com foco no tema Fixação Biológica do Nitrogênio. Por meio da representação vetorial de sequências biológicas geradas pelo SWeeP, foram estudados os genomas completos dos gêneros de bactérias *Azoarcus*-*Aromatoleum* e foram identificadas inconsistências na classificação taxonômica que não haviam sido relatadas até o momento. Isto confirma que o SWeeP é uma forte ferramenta para apoiar a análise e discussão dos modelos taxonômicos. Aqui, foi proposto um modelo de suporte à taxonomia bacteriana, EGenera, que busca contemplar as regras gerais da expertise humana e explorar os recursos disponíveis de dados moleculares, realizando uma análise global do domínio Bacteria por meio de técnicas de aprendizado de máquina. Os resultados mostram que a máquina é capaz de aprender com a experiência humana. Por meio da análise automática de agrupamentos de gêneros, é possível incorporar o conhecimento a priori e o conhecimento em nível molecular à taxonomia bacteriana, de forma automática, rápida e consistente.

Palavras-chave: Bioinformática, Inteligência Artificial, Mineração de texto, Mineração de dados, Taxonomia bacteriana, Fixação Biológica do Nitrogênio.

ABSTRACT

The large amount of genomic data and scientific literature deposited in databases today leads to the understanding that the taxonomic characteristics of an organism can be constructed from a comprehensive evolutionary framework. However, even with an important legacy of human oversight in taxonomy in general and in the taxonomy of bacteria in particular, the classification of organisms is not always consistent with the groupings inferred by the analysis of molecular data. In addition, there are no tools to integrate molecular sequencing data and scientific literature, which points to the weakness in support of current taxonomic models. This difficulty is problematic and has been reported in several recent studies. Data Mining (MD) and Text Mining (MT) techniques based on vector representation of biological sequences and texts from scientific literature can enable the consistent integration of human expertise with molecular information to create taxonomic models. In this perspective, the objective of this thesis is to propose a model to support bacterial taxonomy based on MD and MT techniques for Big Data analysis. This model explores molecular data patterns available in public databases, guided by general guidelines from human experts. Vector technologies developed in-house were used to explore Big Data in the biological area, having vector representation as the main element. The proposed new methodology explores the union of SWeeP and BIOTEXT models, which allows the comparison and exploration of genomic data with agility, in addition to enabling the execution of text mining using the arsenal of bioinformatics tools. Case studies were conducted in the Bacteria domain, focusing on the topic Biological Nitrogen Fixation. Through the vector representation of biological sequences generated by SWeeP, the complete genomes of the *Azoarcus-Aromatoleum* bacteria genera were studied and inconsistencies in the taxonomic classification were identified that had not been reported so far. This confirms that SWeeP is a strong tool to support the analysis and discussion of taxonomic models. Here, a model to support bacterial taxonomy, EGenera, was proposed, which seeks to contemplate the general rules of human expertise and explore the available resources of molecular data, performing a global analysis of the Bacteria domain through machine learning techniques. The results show that the machine is capable of learning from human experience. Through automatic analysis of genera clusters, it is possible to incorporate a priori knowledge and knowledge at the molecular level into bacterial taxonomy, automatically, quickly and consistently.

Keywords: Bioinformatics, Artificial Intelligence, Text Mining, Data Mining, Bacterial Taxonomy, Biological Nitrogen Fixation.

LISTA DE FIGURAS

CAPÍTULO II: REVISÃO DA LITERATURA

FIGURA 1.2. Os 5 “Vs” do Big Data.....	23
--	----

CAPÍTULO III: O MODELO SWeeP PARA REPRESENTAÇÃO VETORIAL DE PROTEOMAS DE BACTÉRIAS

FIGURA 2.3. Representação gráfica dos proteomas bacterianos	41
FIGURA 3.3. Análise de dois componentes principais da projeção SWeeP para os três gêneros de organismos modelos.....	45
FIGURA S1.3. Resultados do teste de aprendizagem da máquina no software WEKA	49

CAPÍTULO IV: COMPARATIVE GENOMICS PROVIDES INSIGHTS INTO THE TAXONOMY OF *Azoarcus* AND REVEALS SEPARATE ORIGINS OF *nif* GENES IN THE PROPOSED *Azoarcus* AND *Aromatoleum* GENERA

FIGURE 1.4 Average nucleotide identity (ANI) of whole genome comparison of 32 <i>Azoarcus</i> and <i>Aromatoleum</i> genomes.....	61
FIGURE 2.4 Phylogenetic trees of whole genomes and clustering of core genomes	62
FIGURE 3.4. Nitrate reduction pathway analysis	66
FIGURE 4.4. Aromatic compound degradation pathway analysis	67
FIGURE 5.4 Single gene <i>nifH</i> phylogeny and heat map confirms the presence of two types of <i>nifH</i> genes in <i>Azoarcus</i> - <i>Aromatoleum</i> genomes	69
FIGURE 6.4 Analysis of the minimum set of <i>nif</i> genes for diazotrophy (<i>nifHDKENB</i>).....	70
FIGURE S1.4. Clustering of <i>Azoarcus</i> , <i>Aromatoleum</i> , and <i>Thauera</i> (branch 9468) in the global bacterial tree previously presented in the study by De Pierri <i>et al.</i> (2020), with 10,324 whole genomes.....	77
FIGURE S2.4. Whole genome comparison of 67 genomes analyzed based on sweep vector	77
FIGURE S3.4. Clustal Omega phylogenetic inference of 16S rRNA analysis for all 67 genomes	79
FIGURE S4.4. Core gene cluster comparisons for all 67 genomes.....	90
FIGURE S5.4. Nj phylogeny and heatmap illustrating the comparison of 67 genomes on SWeeP projection of the <i>nif</i> gene cluster set for each organism	91
FIGURE S6.4 Comparative analysis of the gene neighbourhoods of <i>nif</i> regulatory genes carried out with PATRIC 3.6.6.....	91

CAPÍTULO V: BIOTEXT: EXPLOITING BIOLOGICAL-TEXT FORMAT FOR TEXT MINING

FIGURE 1.5 Graphic representation of sequence headers	105
FIGURE S1.5 Text mining processes	130

FIGURE S2.5 BIOTEXT Strategy.....131

CAPÍTULO VI: O MODELO EGenera PARA CLASSIFICAÇÃO TAXONÔMICA AUTOMÁTICA DE BACTÉRIAS

FIGURA 1.6 Workflow dos processos do EGenera.....138

FIGURA 2.6. Representação do gênero *Clostridium* na árvore filogenética inferida a partir dos resultados do EGenera.....143

FIGURA 3.6. Representação do gênero *Tahuera* na árvore filogenética inferida a partir dos resultados do EGenera.....144

FIGURA 4.6. Representação do gênero *Polaribacter* na árvore filogenética inferida a partir dos resultados do EGenera.....145

FIGURA S1.6 Estrutura da rede neuronal artificial utilizada para a construção das redes do Ensemble...148

LISTA DE TABELAS

CAPÍTULO II: REVISÃO DA LITERATURA

TABELA 1.2 Demanda computacional para a análise de Big Data em Genômica	24
TABELA 2.2 Principais métodos de clusterização e suas definições	26
TABELA 3.2 Algoritmos de Mineração de Texto e suas definições.....	29
TABELA 4.2 Conjunto mínimo de genes essenciais da Nitrogenase de Molibdênio.....	33

CAPÍTULO III: O MODELO SWeeP PARA REPRESENTAÇÃO VETORIAL DE PROTEOMAS DE BACTÉRIAS

TABELA S1.3 Organismos e sequências de referência obtidas do NCBI usadas para a construção de heat maps	49
---	----

CAPÍTULO IV: COMPARATIVE GENOMICS PROVIDES INSIGHTS INTO THE TAXONOMY OF *Azoarcus* AND REVEALS SEPARATE ORIGINS OF *nif* GENES IN THE PROPOSED *Azoarcus* AND *Aromatoleum* GENERA

TABLE 1.4 Strains used in phylogenies and evidence for diazotrophy.....	54
TABLE 2.4 Nitrogenase (acetylene reduction) activity of selected strains of <i>Azoarcus</i> and <i>Aromatoleum</i> ..	71
TABLE S1.4 Accession numbers of sequences used in phylogenies	83
TABLE S2.4 Genome sequencing statistics for strains sequenced in this study.....	88
TABLE S3.4 Accession numbers of sequences used in the analysis of <i>nif</i> and other functional genes.....	89
TABLE S4.4 ANI percentage values for the 32 genomes of <i>Azoarcus</i> and <i>Aromatoleum</i> strains.....	93
TABLE S5.4 . Percentage similarity values between 16s rRNA sequences extracted from genomes of <i>Azoarcus</i> and <i>Aromatoleum</i> strains.....	94

CAPÍTULO V: BIOTEXT: EXPLOITING BIOLOGICAL-TEXT FORMAT FOR TEXT MINING

TABLE 1.5 AMINOcode and DNAbits Dictionary.....	102
TABLE S1.5 Text mining tools and approaches for Omics sciences.....	113
TABLE S2.5 Text mining algorithms/models and its purpose.....	122
TABLE S3.5 List of BIOTEXT functions.....	123
TABLE S4.5 Results for the blast search using the protein WP_011156533.1	126
TABLE S5.5 Demonstration of the difference between minimal and expanded encodings.....	130

CAPÍTULO VI: O MODELO EGenera PARA CLASSIFICAÇÃO TAXONÔMICA AUTOMÁTICA DE BACTÉRIAS

TABELA 1.6 Atributos considerados para o aprendizado de máquina.....	136
--	-----

TABELA 2.6. Teste de Performance.....137

TABELA 3.6 Resultados do EGenera 139

LISTA DE ABREVIATÖES

ANI	<i>Average Nucleotide Identity</i>
ANN	<i>Artificial Neural Network</i>
ARA	<i>Acetylene Reduction Assay</i>
BLAST	<i>Basic Local Alignment Search Tool</i>
CDS	<i>Coding Sequence</i>
DL	<i>Deep Learning</i>
EI	Extração de Informação
FNB	Fixação Biológica do Nitrogênio
FSB	Formato de Sequência Biológica
HDV	<i>High Dimension Vector</i>
ICNB	<i>International Code of Nomenclature of Bacteria</i>
MD	Mineração de Dados
ML	<i>Machine Learning</i>
MT	Mineração de Texto
NCBI	<i>National Center of Biotechnology Information</i>
N ₂	Nitrogênio Atmosférico
NH ₃	Amônia
NJ	<i>Neighbor-Joining</i>
PCA	<i>Principal components analysis</i>
REN	Reconhecimento de Entidade Nomeada
RI	Recuperação de Informação
rRNA	<i>Ribosomal Ribonucleic Acid</i>
SMV	<i>Support Vector Machine</i>
UPGMA	<i>Unweighted Pair Group Method with Arithmetic Mean</i>

SUMÁRIO

CAPÍTULO I: INTRODUÇÃO	19
1 CONSIDERAÇÕES INICIAIS.....	19
2 DESCRIÇÃO DO PROBLEMA	19
3 OBJETIVOS	20
3.1 <i>Objetivo geral</i>	20
3.2 <i>Objetivos específicos</i>	21
4 ORGANIZAÇÃO DA TESE	21
 CAPÍTULO II: REVISÃO DA LITERATURA.....	23
1 BIG DATA.....	23
2 MINERAÇÃO DE DADOS.....	25
2.1 <i>Técnicas baseadas em modelo preditivo</i>	25
2.2 <i>Técnicas baseadas em modelo descritivo</i>	25
3 APRENDIZADO DA MÁQUINA.....	27
3.1 <i>Aprendizagem supervisionada</i>	27
3.2 <i>Aprendizagem não supervisionada</i>	28
4 MINERAÇÃO DE TEXTO	28
5 TAXONOMIA BACTERIANA	30
5.1 <i>Contexto histórico da taxonomia procariótica</i>	30
5.2 <i>Taxonomia baseada em genoma</i>	30
5.3 <i>Fragilidades da taxonomia bacteriana</i>	31
6 FIXAÇÃO BIOLÓGICA DE NITROGÊNIO	32
6.1 <i>Enzima Nitrogenase</i>	32
6.2 <i>Cluster nif</i>	32
7 CONSIDERAÇÕES FINAIS	33
REFERÊNCIAS	34
 CAPÍTULO III: O MODELO SWeeP PARA REPRESENTAÇÃO VETORIAL DE PROTEOMAS DE BACTÉRIAS.....	39
1 INTRODUÇÃO	39
2 MÉTODOS	40
2.1 <i>Implementação</i>	41
2.2 <i>Definição de parâmetros</i>	41
3 RESULTADOS E DISCUSSÃO	41

3.1 Estudo de caso: Representação Vetorial de proteomas bacterianos	41
4 CONCLUSÃO	46

CAPÍTULO IV: COMPARATIVE GENOMICS PROVIDES INSIGHTS INTO THE TAXONOMY OF *Azoarcus* AND REVEALS SEPARATE ORIGINS OF *nif* GENES IN THE PROPOSED *Azoarcus* AND *Aromatoleum* GENERA

ABSTRACT	50
1 INTRODUCTION	51
2 MATERIALS AND METHODS	53
2.1 Bacterial Strains and Genomes	53
2.2 NCBI Sequence Datasets	56
2.3 Average Nucleotide Identity (ANI)	56
2.4 Protein Clusters	56
2.5 Core Genome Analysis Based on Core Protein Groups	57
2.6 Analysis of <i>nif</i> and Other Functional Genes	57
2.7 SWeeP Phylogenies	57
2.8 Heat Maps	58
2.9 16S rRNA Analysis	58
2.10 Nitrogenase Activity	59
3 RESULTS AND DISCUSSION	59
3.1 Whole-Genome Sequences and Identity of <i>Azoarcus</i> sp. Strain TTM-91	59
3.2 Average Nucleotide Identity (ANI) and Core Genome Analysis	59
3.3 Genes for Nitrate Reduction (<i>Nar/Nap</i> , <i>Nir</i> , <i>Nor</i> , <i>Nos</i>) and for the Anaerobic (<i>Bzd</i>) or Aerobic (<i>Box</i>) Degradation of Benzoate	65
3.4 <i>nif</i> genes and Nitrogenase Activity	68
4 CONCLUSIONS	72
SUPPLEMENTARY INFORMATION	77

CAPÍTULO V: BIOTEXT: EXPLOITING BIOLOGICAL-TEXT FORMAT FOR TEXT MINING

1 INTRODUCTION	99
2 METHODS	100
2.1 BIOTEXT Implementation	100
2.2 BIOTEXT application	101
3 RESULTS AND DISCUSSION	101
3.1 Converting texts into biological sequence format (BSF)	101
3.2 Interpreting BLAST results	103
3.3 BIOTEXT application experiment in large corpus	106

4 CONCLUSION.....	106
SUPPLEMENTARY INFORMATION	113

CAPÍTULO VI: O MODELO EGenera PARA CLASSIFICAÇÃO TAXONÔMICA AUTOMÁTICA DE BACTÉRIAS

1. INTRODUÇÃO	133
2. MÉTODOS	134
2.1 Conjunto de dados de origem	134
2.2 Avaliação de desempenho do modelo.....	134
3. RESULTADOS	135
3.1 Extração de características e Aprendizado de Máquina	135
3.2 Treinamento do ensemble.....	136
3.3 Aprendizado não supervisionado	137
4. DISCUSSÃO	139
5. CONSIDERAÇÕES FINAIS.....	146
MATERIAL SUPLEMENTAR	147
REFERÊNCIAS	148

CAPÍTULO VII: CONCLUSÃO

REFERÊNCIAS	153
-------------------	-----

CAPÍTULO I: INTRODUÇÃO

1 CONSIDERAÇÕES INICIAIS

Com o avanço das tecnologias genômicas, o volume de genomas completos de microrganismos depositados em bancos de dados públicos aumentou consideravelmente. Essa tendência também é observada nas bases de dados da literatura, onde milhões de pesquisas científicas são publicadas diariamente (SIMON *et al.*, 2019). Embora muitos microrganismos já tenham sido estudados experimentalmente de forma extensa, as informações a respeito estão dispersas na literatura científica. Isso reflete diretamente na área de taxonomia, pois a dificuldade de integrar a expertise humana, o conhecimento prévio e as descobertas mais recentes tornam o processo de classificação conservador, lento e com relevância limitada (GARRITY, 2016). Isso é observado principalmente no domínio Bacteria, onde a taxonomia ainda é fracamente baseada em informações moleculares. Sem um modelo taxonômico de consenso, a classificação torna-se imprecisa e redundante (ROSSELLÓ-MÓRA; WHITMAN, 2019).

A análise e interpretação deste grande volume de informações (Big Data) extrapola a capacidade humana, porém, é possível integrar e automatizar esses processos para descobrir a relação entre os microrganismos através das abordagens de Mineração de dados (MD) e Mineração de textos (MT) (ANANIADOU *et al.*, 2015). A partir disso, nos perguntamos: Seria possível a associação das abordagens de MD, MT e Inteligência Artificial substituir a expertise humana para validação da taxonomia bacteriana?

Para responder esta questão, utilizamos as ferramentas computacionais SWeeP (DE PIERRI *et al.*, 2020), BIOTEXT (MACHADO *et al.*, 2021) e scripts *in-house*, com o objetivo de explorar e analisar Big Data na área biológica. Para reforçar a eficácia da representação vetorial de proteomas bacterianos completos gerados pelo modelo SWeeP, realizamos um estudo de caso relacionada à Fixação Biológica de Nitrogênio (FBN). Foi possível identificar falhas na classificação taxonômica de organismos do gênero *Azoarcus-Aromatoleum* não identificadas até o momento. Isso mostra que mais estudos precisam ser conduzidos para destacar possíveis fragilidades na classificação de táxons no domínio Bacteria. Esses resultados reforçam a suposição da necessidade de novas abordagens computacionais para apoiar os modelos taxonômicos atuais.

2 DESCRIÇÃO DO PROBLEMA

A grande quantidade de dados genômicos nos bancos de dados leva ao entendimento de que uma taxonomia pode ser construída em uma estrutura evolutiva abrangente (HUGENHOLTZ *et al.*, 2021), no entanto, a falta de ferramentas para integrar dados biológicos e literatura traz à tona a fragilidade de modelos taxonômicos. Mesmo com um legado importante de supervisão humana na taxonomia em geral

e especificamente em bactérias, a classificação dos organismos nem sempre corresponde aos agrupamentos inferidos pela análise de dados moleculares. Essa imprecisão é problemática e foi relatada em estudos recentes (HUGENHOLTZ *et al.*, 2016; PARKS *et al.*, 2018; ROSSELLÓ-MÓRA; WHITMAN, 2019; MARGOS *et al.*, 2020; MURRAY *et al.*, 2020; HUGENHOLTZ *et al.*, 2021).

Abordagens que envolvem MD e MT são amplamente utilizadas para análise, processamento e extração de informações biológicas, mesmo assim, as ferramentas disponíveis hoje ainda são limitadas e não atendem a todas as demandas por agilidade no processamento de Big Data. As técnicas de MD e MT por representação vetorial podem integrar de forma consistente a expertise humana e as informações moleculares para criar modelos taxonômicos eficientes, respeitando o padrão atual e consistente. Nesta tese, estudamos e propomos um modelo de suporte à taxonomia bacteriana (EGenera) que busca, ao mesmo tempo, contemplar as diretrizes gerais dos especialistas humanos e explorar os recursos disponíveis de dados moleculares, aliados ao "melhor" em inteligência computacional. Assim, a motivação para esta pesquisa é a falta de ferramentas computacionais que atendam à demanda por analisar diferentes níveis de informação biológica com rapidez e precisão. As abordagens desenvolvidas podem contribuir:

- ✓ Fornecendo uma visão geral global da taxonomia bacteriana, apoiada por aprendizado de máquina e apropriada para a visão taxonômica atual
- ✓ Auxiliando na validação da taxonomia bacteriana global, de forma automatizada.
- ✓ Fornecendo suporte necessário para a especulação de todos os conceitos considerados importantes para a construção do mais alto nível de conhecimento, que é o estado da arte.
- ✓ Modernizando estudos e trazendo avanços para a área de Fixação Biológica de Nitrogênio e para a área de classificação taxonômica de bactérias.

Esta pesquisa também poderá auxiliar no refinamento e organização da produção acadêmica relacionada à Fixação Biológica de Nitrogênio, trazendo reflexões sobre a importância das pesquisas ao longo do tempo, tendências, ênfases, métodos, grupos de pesquisadores, diferenciando trabalhos entre si.

3 OBJETIVOS

3.1 Objetivo geral

- ✓ Propor novas abordagens de Mineração de Dados e Mineração de Textos para análise de Big Data e desenvolver um panorama global da taxonomia bacteriana, apoiado em aprendizado de máquina e adaptado à visão taxonômica atual.

3.2 Objetivos específicos

- ✓ Explorar e consolidar o modelo de representação vetorial de proteínas - SWeeP;
- ✓ Levantar bibliografia sobre MT e as principais abordagens utilizadas;
- ✓ Participar do desenvolvimento de uma abordagem de MT;
- ✓ Utilizar as abordagens de aprendizado de máquina como suporte na estruturação conceitual da classificação taxonômica de bactérias;
- ✓ Criar uma estrutura para validação da taxonomia bacteriana de maneira automatizada.

4 ORGANIZAÇÃO DA TESE

Esta dissertação está organizada em sete capítulos:

- ✓ Os capítulos I e II apresentam os principais conceitos relacionados a Big Data, Mineração de dados, Mineração de textos, Taxonomia Bacteriana e Fixação Biológica de Nitrogênio;
- ✓ Os capítulos III e V exploram as abordagens desenvolvidas *in-house* para lidar com a análise genômica e processamento automático de textos da literatura, em escala de Big Data (SWeeP e BIOTEXT);
- ✓ O capítulo IV é direcionado à exploração dessas ferramentas por meio de um estudo de caso do domínio Bactéria, com foco no tema fixação biológica do nitrogênio;
- ✓ O capítulo VI apresenta o modelo EGenera para suporte na classificação taxonômica;
- ✓ O capítulo VII apresenta a conclusão.

REFERENCES

- ANANIADOU, S., THOMPSON, P., NAWAZ, R., MCNAUGHT, J., KELL, D. B. Event-based text mining for biology and functional genomics. **Briefings in functional genomics**, v.14, 213–230, 2015.
- DE PIERRI, C.R.; VOYCEIK, R.; SANTOS DE MATTOS, L.G.C.; KULIK, M.G.; CAMARGO, J.O.; REPULA DE OLIVEIRA, A.M.; DE LIMA NICHIO, B.T.; MARCHAUKOSKI, J.N.; DA SILVA FILHO, A.C.; GUIZELINI, ORTAGEA, J.M.; PEDROSA, F.O. RAITTZ, R.T. SWeeP: Representing large biological sequences datasets in compact vectors. **Sci. Rep**, 10, 91, 2020.
- GARRITY G. M. A New Genomics-Driven Taxonomy of Bacteria and Archaea: Are We There Yet? **Journal of clinical microbiology**, v.54, 1956–1963, 2016.
- HUGENHOLTZ, P., CHUVOCHINA, M., OREN, A., PARKS, D. H., SOO, R. M. Prokaryotic taxonomy and nomenclature in the age of big sequence data. **The ISME journal**, Advance online publication, 2021. <https://doi.org/10.1038/s41396-021-00941-x>
- HUGENHOLTZ, P., SKARSHEWSKI, A., PARKS, D. H. Genome-Based Microbial Taxonomy Coming of Age. **Cold Spring Harbor perspectives in biology**, v.8, n.6, a018085, 2016.
- MACHADO, D.J.S.; DE PIERRI, C.R.; SANTOS, L.G.C.; PEDROSA, F.O.; RAITTZ, R.T. BIOTEXT: Exploiting Biological-like format for text mining. *BioRxiv* 2021.04.08.439078. doi: <https://doi.org/10.1101/2021.04.08.439078>.
- MARGOS, G., FINGERLE, V., CUTLER, S., GOFTON, A., STEVENSON, B., ESTRADA-PEÑA, A. Controversies in bacterial taxonomy: The example of the genus *Borrelia*. **Ticks and tick-borne diseases**, v.11, 2020.
- MURRAY, A. E., FREUDENSTEIN, J., GRIBALDO, S., HATZENPICHLER, R., HUGENHOLTZ, P., KÄMPFER, P., KONSTANTINIDIS, K. T., LANE, C. E., PAPKE, R. T., PARKS, *et al.* Roadmap for naming uncultivated Archaea and Bacteria. **Nature microbiology**, n.5, v.8, 987–994, 2020
- PARKS, D. H., CHUVOCHINA, M., WAITE, D. W., RINKE, C., SKARSHEWSKI, A., CHAUMEIL, P.-A., HUGENHOLTZ, P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. **Nature Biotechnology**, n.36, 996–1004, 2018.
- ROSSELLÓ-MÓRA, R., WHITMAN, W. B. Dialogue on the nomenclature and classification of prokaryotes. **Systematic and applied microbiology**, v.42, 5–14, 2019.
- SIMON, C., DAVIDSEN, K., HANSEN, C., SEYMOUR, E., BARNKOB, M. B., OLSEN, L. R. BioReader: a text mining tool for performing classification of biomedical literature. **BMC bioinformatics**, v.19, n. 57, 201

CAPÍTULO II: REVISÃO DA LITERATURA

1 BIG DATA

O termo Big Data foi utilizado pela primeira vez há mais de duas décadas por pesquisadores da NASA (COX; ELLSWORTH, 1997), em referência aos supercomputadores da época, que geravam grandes quantidades de informações que não podiam ser processadas e analisadas. Esse conceito continuou a ser explorado e atualmente é aplicado em diversas áreas do conhecimento (HASSANI *et al.*, 2020).

As tecnologias de Big Data podem ser definidas como abordagens projetadas para analisar, processar e extrair informações de conjuntos de dados extremamente complexos, heterogêneos e grandes, com os quais a capacidade humana nunca poderia lidar (BONNER *et al.*, 2017; GHARAJEH, 2018; NAVARRO *et al.*, 2019). A análise de Big Data transforma uma grande quantidade de dados em informações claras e úteis para a tomada de decisão. Pesquisas utilizando análise de dados baseada em inteligência artificial e aprendizado de máquina mostram resultados mais promissores (POPCHEV; OROZOVA, 2021).

Os cientistas de dados costumam descrever Big Data como dados com características principais, conhecidas como os "Vs" de Big Data: Volume, Velocidade, Variedade, Veracidade e Valor (WHITE, 2012; WAMBA *et al.*, 2015; GE *et al.*, 2018). Existem diferentes definições para essas características, dependendo da área de estudo. Uma breve visão geral do significado dessas características é apresentada abaixo levando em consideração o campo da Genômica (FIGURA 1.2).

FIGURA 1.2 Os 5 "Vs" do Big Data



Fonte: Autora (2022).

Hoje em dia, com a pesquisa científica e o sequenciamento do genoma aumentando sem precedentes, acumulando dados genômicos e textos em bancos de dados, a integração de dados está se tornando uma disciplina independente a medida que tecnologias estão sendo desenvolvidas para uma ampla variedade de contextos genômicos (ROSSI; GRIFANTINI, 2018). A genômica se destaca em relação ao número de dados disponíveis, porém, isso pode mudar no futuro, com o advento de novas tecnologias em outras áreas como; Processamento de imagens, estrutura macromolecular, análise de registros de saúde e proteômica (NAVARRO *et al.*, 2019).

De acordo com Stephens *et al.* (2015) a genômica é o mais exigente dos domínios de Big Data, onde as demandas computacionais de um conjunto de dados - aquisição, armazenamento, distribuição e análise - são as chaves tecnológicas para suportar soluções de Big Data (TABELA 1.2). Nesse sentido, várias técnicas podem ser utilizadas para a Análise de Big Data, tais como, abordagens de mineração de dados, aprendizado de máquina e mineração de texto para obter resultados de alto desempenho (GHARAJEH, 2018).

TABELA 1.2. Demanda computacional para a análise de Big Data em Genômica

Demandas	Descrição
Aquisição	Os dados coletados devem ser comparáveis a outros dados para que uma sequência do genoma seja útil. Para isso, esses dados devem ser pareados com métodos automatizados de coleta de metadados e dados fenotípicos, de acordo com padrões apropriados.
Armazenamento	Os algoritmos devem representar grandes conjuntos de genomas em formatos compactos, tornando mais eficiente e ágil comparar um genoma com muitos outros. Sistemas eficientes de compactação e indexação são essenciais para fazer o melhor uso dos disponíveis e torná-los altamente acessíveis.
Distribuição	Para distribuir sequências de genoma em escala global, a solução é usar sistemas de computação em nuvem para maximizar a associação entre os códigos.
Análise	É necessário integrar o conhecimento do domínio biológico, sistemas de aprendizado de máquina em grande escala e algoritmos computacionais que podem suportar consultas múltiplas para busca de padrões em conjuntos de dados muito grandes.

Fonte: Adaptado de Stephens *et al.*, (2015).

2 MINERAÇÃO DE DADOS

Mineração de Dados (MD) é um processo de descoberta de conhecimento em bancos de dados (do inglês, *Knowledge Discovery from texts* - KDD) que visa extrair padrões e conhecimento a partir das informações presentes em grandes conjuntos de dados por meios automáticos ou semiautomáticos (FAYYAD, 1997). Extrair informações úteis de bancos de dados é uma tarefa desafiadora, mas muito importante para a tomada de decisões.

A MD é considerada um campo interdisciplinar, que pertence ao mesmo tempo à estatística, inteligência artificial, aprendizado de máquina e sistemas de banco de dados (TSUI *et al.*, 2006). Muitas são as técnicas que podem ser utilizadas para realizar o MD, dependendo do objetivo a ser alcançado. Essas técnicas são essencialmente distribuídas em dois modelos, o preditivo e o descritivo (LEE *et al.*, 2008; HAND; ADAMS, 2015).

2.1 Técnicas baseadas em modelo preditivo

Técnicas preditivas são utilizadas para desenvolver ou treinar modelos de Classificação e/ou modelos de Regressão que podem ser usados para prever valores para dados, com base na significância estimada de um conjunto de variáveis de entrada (SHIRABAD, 2011).

Modelos de classificação são usados para classificar dados em diferentes classes predefinidas usando conceitos matemáticos, como árvores de decisão, programação linear, rede neural e estatística (TSUI *et al.*, 2006). Os algoritmos usados para classificação de dados são Regressão logística (EFROM, 1975), *Naive Bayes* (MCCALLUM; NIGAM, 1998), *k-Nearest Neighbors* (FIX; HODGES, 1951), Algoritmos Genéticos (HOLLAND, 1975) e *Support Vector Machine* (VAPNIK, 1998).

A técnica de regressão analisa a relação entre as variáveis, prevendo resultados com base em dados previamente conhecidos (SAUERBREI *et al.*, 2020). A regressão consiste basicamente em prever uma quantidade. O algoritmo de regressão mais utilizado é o Multivariado - que é uma extensão da Regressão Linear Múltipla (IZENMAN, 2013).

2.2 Técnicas baseadas em modelo descritivo

Os modelos descritivos determinam o que aconteceu no passado, analisando os dados armazenados e fornecendo dados aproximados (SHEIKH, 2013). As regras de associação e agrupamento são técnicas descritivas.

As regras de associação buscam encontrar associação entre os valores de uma variável específica

que forneça conhecimento sobre a distribuição da população (TSUI *et al.*, 2006). A clusterização é usada para identificar dados semelhantes entre si, ajudando a compreender as semelhanças e diferenças entre os dados (JAIN *et al.*, 1999).

De acordo com AGRAWAL *et al.* (1998), nenhuma técnica de clusterização atende a todos os pontos dos resultados. As abordagens MD e MT implementadas nesta tese apresentam critério de agrupamento que requerem a integração das várias técnicas. Portanto, usamos a estrutura de métodos de clusterização hierárquica e particionamento.

Os métodos de armazenamento em cluster hierárquico criam uma estrutura que descreve uma hierarquia de clusters nos dados; os algoritmos de particionamento dividem o banco de dados em k-grupos, onde o número k é fornecido pelo usuário (JAIN *et al.*, 1999). Os principais algoritmos desses métodos e suas definições são apresentados na quadro abaixo (TABELA 2.2).

TABELA 2.2. Principais métodos de clusterização e suas definições

Método	Algoritmo	Definição	Autores
Hierárquico	Single link	A distância entre os clusters é especificada pela distância do par de exemplos mais próximo, cada exemplo pertencendo a um desses clusters.	Sneath (1957) Johnson (1967)
	Complete link	Determina a distância entre os clusters de acordo com a maior distância entre um par de exemplos.	Sorensen (1948)
	Average link	A distância entre os clusters é definida como a média das distâncias entre todos os pares de exemplos em cada cluster.	Sokal & Michener (1957)
	Centroid link	Os agrupamentos identificados são representados por um vetor que armazena a média de cada atributo. Assim, a distância entre os clusters é definida em termos da distância no espaço euclidiano entre os representantes de cada cluster.	Good & Gaskings (1969)
	Ward	Minimiza a perda de informações associadas a cada cluster. Em cada etapa, todas as possíveis uniões entre pares de clusters são consideradas e os clusters que apresentam a perda mínima de informação são selecionados para o agrupamento. Essa perda de informação é definida em termos da soma dos quadrados dos erros.	Ward (1963)

(Continua)

Particional	K-means	Algoritmos parciais dividem o banco de dados em grupos k, onde o número k é fornecido pelo usuário. As K-médias utilizam a média dos objetos que pertencem a um determinado cluster, representando a distância média de cada objeto ao seu respectivo representante.	Steinhaus (1956)
	K-medoids	O objeto que está mais próximo do centro de gravidade do cluster é escolhido como o representante, onde os clusters são representados usando a mediana dos dados em vez da média. O elemento mais próximo do centro é chamado de “medoid”.	Kaufman & Rousseeuw (1987)
	Fuzzy C-means	Graus de associação são atribuídos a cada um dos pontos de dados, indicando o grau de pertencimento dos pontos de dados a cada cluster. Se os pontos estiverem próximos à borda do cluster, haverá um menor grau de associação dos pontos que estão no centro do cluster. No Fuzzy C-means, o centróide de um cluster será a média de todos os pontos ponderados pelo grau de pertencimento ao cluster.	Bezdek (1981)

Fonte: Autora (2022).

3 APRENDIZADO DA MÁQUINA

O Aprendizado de Máquina (AM) é um processo de Inteligência Artificial que busca definir um conjunto de abordagens para identificar padrões em dados, envolvendo métodos e algoritmos que têm a capacidade de prever comportamento extraindo informações automaticamente (ALLAHYARI *et al.*, 2017). O AM geralmente está no mesmo nível do desempenho humano preditivo, o que difere de ambos é que o uso do AM leva a uma redução *significativa* na carga de trabalho (O'MARA-EVES *et al.*, 2015). Além disso, as máquinas têm uma vantagem *significativa* sobre os humanos: podem ser conectadas em rede (KOTELUK *et al.*, 2021). Embora existam estratégias híbridas, o AM é dividido em duas categorias principais: aprendizagem supervisionada e aprendizagem não supervisionada.

3.1 Aprendizagem supervisionada

A aprendizagem supervisionada baseia-se no conhecimento prévio para fazer previsões sobre novos pontos de dados. Para isso, é necessário utilizar um conjunto de dados de treinamento rotulado para que os algoritmos “aprendam” como discriminar classes de dados. O AM supervisionado se preocupa

em prever o rótulo com base nos padrões de entrada (SCHRIDER, KERN, 2018).

Um subconjunto de AM supervisionado, aprendizado profundo (AP), é uma coleção de algoritmos em camadas que contêm funções complexas para descobrir automaticamente relacionamentos em dados brutos (LECUN *et al.*, 2015; JANIESCH *et al.*, 2021). AP explora uma arquitetura que imita a estrutura de um neurônio no cérebro humano, por um modelo chamado Redes Neurais Artificiais (KOTELUK *et al.*, 2021).

A estrutura de uma RNA consiste em uma camada de entrada, algumas camadas ocultas e uma camada de saída, à qual os neurônios são conectados. Essas conexões são ponderadas, calculadas durante o processo de treinamento, para que o algoritmo possa determinar a saída aproximada de cada camada e, assim, obter o resultado (LECUN *et al.*, 2015; KOTELUK *et al.*, 2021). As RNAs são inspiradas no modelo de neurônios, proposto por McCulloch e Pitts (1943).

3.2 Aprendizagem não supervisionada

A aprendizagem não supervisionada está relacionada à descoberta da estrutura dentro de um conjunto de dados, onde não há conhecimento prévio (SCHRIDER, KERN, 2018). Um exemplo de aprendizagem não supervisionada aplicada é a Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*), que, no contexto desta tese, é usada para descobrir relações de evidência entre organismos. Basicamente, o PCA usa uma matriz de vetores contendo as sequências biológicas como entrada e, em seguida, representa os dados da matriz em uma dimensão inferior (componentes).

4 MINERAÇÃO DE TEXTO

Mineração de Texto (MT) é um processo que se refere à extração de informações encontradas em textos. A grande vantagem das técnicas de MT é a capacidade de aprimorar a busca na literatura, facilitando a análise e o armazenamento dos dados, tornando o processo de pesquisa mais refinado e preciso (ANANIADOU *et al.*, 2015; ANTONS *et al.*, 2020). Vários processos e métodos automatizados são necessários para extrair conhecimento de textos (KODRATOFF, 1999; SINGHAL *et al.*, 2016; KRALLINGER *et al.*, 2017).

Embora várias técnicas e algoritmos possam ser usados, a mineração de textos pode ser classificado em três tarefas fundamentais: (1) Recuperação de Informação (RI), realizada para obter documentos relevantes para um determinado assunto de interesse; (2) Reconhecimento de entidade nomeada (REN), que analisa a ocorrência de palavras-chave específicas em documentos; (3) Extração de

informação (EI), que trata da detecção de links entre as palavras-chave encontradas, aplicando diferentes tipos de algoritmos (FLEUREN; ALKEMA, 2015). A seguir, na TABELA 3.2 é descrito o propósito dos algoritmos.

TABELA 3.2. Algoritmos de Mineração de Texto e suas definições

Tarefa	Tipos de Algoritmos	Propósito
Recuperação de Informação	Tokenização	Dividir o texto em palavras distintas, chamadas entidades textuais ou tokens
	<i>Stopping</i>	Remover palavras comuns, como “e”, “ou”
	<i>Steeming</i>	Remover prefixos e sufixos para normalizar palavras
	Normalização	Converter o texto em minúsculas ou maiúsculas
Reconhecimento de Entidade Nomeada	Indexação	Processar um texto para extrair estatísticas consideradas importantes por representar a informação e / ou permitir a busca rápida de seu conteúdo.
	Baseados em dicionário	Comparar um dicionário de termos contendo termos e sinônimos com o texto da consulta.
	TF-IDF	TF- IDF (<i>Term Frequency - Inverse Document Frequency</i>). armazena texto como vetores ponderados; A frequência de um termo é o número de vezes que o termo aparece em um documento.
	Baseados em Processamento de Linguagem Natural	Processamento de linguagem de baixo nível e tarefa de compreensão; etiquetagem de classes gramaticais.
Extração de Informação	Modelos preditivos	Classificação de entidades textuais em categorias predefinidas
	Modelos baseados em clusterização	Agrupamento de entidades textuais em grupos anteriormente indefinidos e desconhecidos.
	Regras de Associação	Encontrar associações interessantes ou correlações entre um grande conjunto de itens de dados.

Fonte: Autora (2022).

5 TAXONOMIA BACTERIANA

Para classificar procariontes é necessário para compreender e ordenar a diversidade biológica. Cientistas constroem taxonomias com base em um sistema hierárquico que reflete as relações naturais entre os organismos (ROSELLÓ-MÓRA; WHITMAN, 2019). Não existe uma estrutura de consenso; diversos sistemas de classificação e nomenclatura são utilizados, o que pode gerar classificações imprecisas (HUGENHOLTZ *et al.*, 2021).

5.1 Contexto histórico da taxonomia procariótica

A taxonomia biológica é um princípio organizacional originalmente proposto por Linnaeus (1758) para classificar os organismos de acordo com suas semelhanças, que consiste (atualmente) em uma hierarquia de oito níveis; Domínio, Reino, Filo, Classe, Ordem, Família, Gênero e Espécie, respectivamente (WOESE; FOX, 1977). A primeira taxonomia baseada em conceitos matemáticos que poderia apoiar a inferência filogenética foi proposta por Sokal e Sneath (1962), mas o modelo carecia de uma estrutura consistente para mostrar as relações evolutivas. Tentando contornar a situação, Zucherlandl e Pauling começaram a pesquisar uma maneira de usar macromoléculas para datar as relações evolutivas da mais antiga para a mais recente. Essas pesquisas motivaram Woese e Fox (1977) a usar os genes ribossomais 16S e 18S para inferência filogenética, pois são altamente conservados e pouco afetados por transferências gênicas. Ao mesmo tempo, foi possível integrar o fenótipo e os achados genômicos em nível de gênero por meio da Taxonomia Polifásica (VENDAMME *et al.*, 1996), entretanto, a classificação em nível de espécie foi melhor suportada por métodos quimiotaxonômicos (McCARTHY; BOLTON, 1963). Só então, com o advento das tecnologias de sequenciamento em larga escala, foi possível realizar a classificação taxonômica baseada em genomas completos (SANDERSON *et al.*, 1998), que é considerada por muitos pesquisadores o padrão-ouro para a classificação de bactérias (HUGENHOLTZ *et al.*, 2016; PARKS *et al.*, 2018; PAUL *et al.*, 2019; ZHU *et al.*, 2019; DE PIERRI *et al.*, 2020; PARKS *et al.*, 2020; HUGENHOLTZ *et al.*, 2021; RAITTZ *et al.*, 2021).

5.2 Taxonomia baseada em genoma

Uma taxonomia é considerada adequada quando descreve características morfológicas, fisiológicas, genéticas e moleculares, e a melhor forma de reunir e representar essas características é por meio de árvores filogenéticas.

Árvores filogenéticas são usadas para inferir relações evolutivas entre organismos, fornecendo uma estrutura para o desenvolvimento de uma taxonomia que leva em consideração as diferentes taxas de evolução (HUGENHOLTZ *et al.*, 2021). Uma filogenia ideal é aquela baseada na comparação de regiões

homólogas, destacando precisamente as semelhanças / dissimilaridades genéticas entre os organismos.

Uma escolha básica e fundamental antes da construção de árvores filogenéticas diz respeito ao método de comparação entre sequências biológicas, que pode ser realizada por alinhamento ou sem alinhamento. Nesta tese, utilizamos o método livre de alinhamento (Sweep), pois proporciona agilidade na análise e comparação de genomas completos por meio da representação de sequências em espaços vetoriais (DE PIERRI *et al.*, 2020; VIEIRA *et al.*, 2021; RAITTZ *et al.*, 2021; PERICO *et al.*, 2021, DA SILVA FILHO, *et al.*, 2021). Atualmente, esses tipos de métodos já estão consolidados e sua aplicação na análise de um grande número de genomas tem se mostrado, em diversos estudos filogenéticos, mais eficientes do que o alinhamento tradicional (LEIMEISTER *et al.*, 2014; ASGARI, MOFRAD, 2015; LI *et al.*, 2017; LEIMEISTER *et al.*, 2018; CHOKALINGAM *et al.*, 2020; DE PIERRI *et al.*, 2020; RAITTZ *et al.*, 2021).

Nesse contexto, para construir árvores filogenéticas baseadas em genomas completos, duas metodologias são utilizadas principalmente: (1) Superárvores - estabelecendo uma árvore de consenso a partir de outras árvores geradas; (2) Supermatrizes - construídas através da concatenação de sequências em matrizes.

Após a construção da matriz de distâncias (fundamental para comparar milhares de sequências, são aplicados modelos para inferência filogenética. Neste trabalho, para a inferência filogenética, foram utilizados modelos baseados em análise de agrupamento – UPGMA (MICHENER; SOKAL, 1957) e WARD (WARD, 1963) (TABELA 2) -, e evolução mínima – NJ (SAITOU, NEI, 1987).

Em resumo, a taxonomia baseada no genoma completo tem muitas vantagens sobre o uso de apenas alguns conjuntos de genes. Além do ganho óbvio em representar toda a maquinaria genética de um organismo, também é possível processar um grande número de genomas usando ferramentas computacionais e de bioinformática. As filogenias baseadas no genoma completo podem ser ágeis mesmo quando milhares de sequências são usadas. Por exemplo, nesta tese, a construção de uma árvore filogenética de 10.324 genomas bacterianos completos levou apenas algumas horas - o que nunca poderia ser feito usando métodos tradicionais como o alinhamento de sequências (ver Capítulo III). Até onde sabemos, esta é a maior árvore de genomas bacterianos completos gerados automaticamente até agora.

5.3 Fragilidades da taxonomia bacteriana

Mesmo com métodos robustos de inferência filogenética existentes hoje, inconsistências nas taxonomias atuais ainda são observadas com frequência. Essas fragilidades estão mais presentes no domínio Bacteria devido à falta de revisão, uma vez que muitas dessas taxonomias são arcaicamente baseadas apenas na classificação fenotípica (GARRITY, 2016; ZHU *et al.*, 2019; PARKS *et al.*, 2020;

HUGENHOLTZ *et al.*, 2021).

Os métodos desenvolvidos e utilizados nesta tese podem auxiliar na classificação da taxonomia bacteriana, fornecendo suporte embasado na literatura científica e reduzindo as fragilidades ocasionadas pela falta de revisão.

6 FIXAÇÃO BIOLÓGICA DE NITROGÊNIO

A Fixação Biológica de Nitrogênio (BNF) é um processo fundamental para adicionar nitrogênio ao ecossistema realizado por organismos que apresentam a enzima nitrogenase. Esses organismos, chamados diazotróficos, são capazes de catalisar a conversão do nitrogênio atmosférico (N₂) em amônia (NH₃) em condições ambientais, processo fundamental para a nutrição das plantas e para a vida em geral (WAGNER, 2011).

6.1 Enzima Nitrogenase

A conversão de N₂ em NH₃ pode ocorrer de diferentes maneiras, de acordo com o tipo de nitrogenase. Existem 3 tipos dessa enzima: a nitrogenase-1 é expressa apenas na presença de molibdênio; a nitrogenase-2 é expressa quando há falta de molibdênio e presença de vanádio; a nitrogenase-3 é expressa na falta de molibdênio e vanádio e na presença de ferro (DIXON; KAHN, 2004). De acordo com Li *et al.* (2016), o processo FBN mais comum entre os diazotróficos é aquele catalisado pela nitrogenase de molibdênio.

6.2 Cluster *nif*

Para que ocorra a formação do complexo enzimático nitrogenase, é necessário que haja um conjunto de genes fundamentais para a síntese, maturação e montagem do complexo, denominado *nif* (BATISTA *et al.*, 2015). Estudos mostram que a organização e o número de genes *nif* necessários para a síntese e montagem da enzima nitrogenase variam amplamente entre as espécies diazotróficas (BOYD *et al.*, 2011; PEDROSA *et al.*, 2011; DOS SANTOS *et al.*, 2012; TEMME ; ZHAO; VOIGT, 2012), embora as propriedades bioquímicas e a estrutura das nitrogenases sejam semelhantes em vários organismos (DOS SANTOS *et al.*, 2012). No *Herbaspirillum seropedicae*, por exemplo, os genes *nif* são agrupados, compreendendo uma região de 46 genes e pelo menos 7 operons (PEDROSA *et al.*, 2011), em comparação, em *Klebsiella oxytoca*, o cluster *nif* contém 20 genes e 7 operons (TEMME; ZHAO; VOIGT, 2012). No entanto, as variações no conteúdo dos genes *nif* são determinadas pelo estilo de vida ambiental

de cada organismo (LI *et al.*, 2016). Para que o processo de fixação de nitrogênio seja viável, é necessário um conjunto mínimo de genes *nif* (*nifHDKENB*) (DOS SANTOS *et al.*, 2012; BURÉN *et al.*, 2020). A TABELA 4.2 apresenta os genes essenciais, sua descrição e função.

TABELA 4.2. Conjunto mínimo de genes essenciais da Nitrogenase de Molibdênio

Gene	Descrição	Função
<i>nifH</i>	Proteína	Doador de elétrons obrigatório para a proteína MoFe. Também necessário para a biossíntese de FeMo-co e maturação apo- MoFe.
<i>nifD</i>	α -subunit da proteína MoFe	Subunidade de ligação FeMo-co. Forma um tetrâmero $\alpha_2\beta_2$ com <i>NifK</i> .
<i>nifK</i>	β -subunit da proteína MoFe	Forma um tetrâmero $\alpha_2\beta_2$ com <i>NifD</i>
<i>nifE</i>	Scaffold para <i>NifB</i> -co	Scaffold para maturação de <i>NifB</i> -co em FeMo-co. Forma um tetrâmero $\alpha_2\beta_2$ com <i>NifN</i> .
<i>nifN</i>	Scaffold para <i>NifB</i> -co	Scaffold para maturação de <i>NifB</i> -co em FeMo-co. Forma um tetrâmero $\alpha_2\beta_2$ com <i>NifE</i>
<i>nifB</i>	Essencial para síntese do FeMo-co	Essencial para a síntese de FeMo-co. Catalisa a formação de <i>NifB</i> -co, a primeira etapa comprometida na biossíntese de FeMo-co.

Fonte: Adaptado de Burén *et al.*, 2020.

7 CONSIDERAÇÕES FINAIS

O objetivo deste Capítulo foi fornecer embasamento teórico para auxiliar na compreensão da tese como um todo. Este Capítulo foi dividido em 3 conceitos centrais: Conceitos gerais de computação e Bioinformática aplicados à análise genômica (Análise de Big Data, Mineração de dados, Aprendizado de Máquina e Mineração de texto), Taxonomia bacteriana e Fixação Biológica de Nitrogênio.

REFERÊNCIAS

- AGRAWAL, R., GEHRKE, J., GUNOPULOS, D., AND RAGHAVAN, P. Automatic subspace clustering of high dimensional data for data mining applications. **Proc. of 1998 ACM SIGMOD Int. Conf. on Management of Data**, pp. 94–105, 1998.
- ALLAHYARI, M.; POURIYEH, S.; ASSEFI, M.; SAFAEI, S.; TRIPPE, E. D. GUTIERREZ, J.; KOCHUT, K. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques., v.13, 2017. Available at: <https://arxiv.org/abs/1707.02919>. Accessed on April, 18th, 2021.
- ANTONS, D.; GRÜNWALD, E.; CICHY, P.; SALGE, T.O. The application of text mining methods in innovation research: Current state, evolution patterns, and development priorities. **R&D Manag.** v.50, p.329–351, 2020.
- ANANIADOU, S., THOMPSON, P., NAWAZ, R., MCNAUGHT, J., KELL, D.B. Event-based text mining for biology and functional genomics. **Brief. Funct. Genomics** v.14, p. 213–230, 2015.
- ASGARI, E. & MOFRAD, M. R. K. Continuous distributed representation of biological sequences for deep proteomics and genomics. **PLoS One** 10, 1–15, 2015.
- BATISTA, M.B.; WASSEM, R.; PEDROSA F.O.; SOUZA, E.M.; DIXON, R.; MONTEIRO, R.A. Enhanced oxygen consumption in *Herbaspirillum seropedicae* fnr mutants leads to increased *NifA* mediated transcriptional activation. **BMC Microbiology**, vol.15 n.95, 2015.
- BEZDEK, J.C.; EHRLICH, R.; FULL, W. FCM: The fuzzy c-means clustering algorithm **Computers & Geosciences**, v.10, n. 2–3, p.191-203, 1984.
- BONNER, S., KURESHI, I., BRENNAN, J., & THEODOROPOULOS, G. EXPLORING the Evolution of Big Data Technologies. **Software Architecture for Big Data and the Cloud**, p. 253–283, 2017.
- BURÉN, S., JIMÉNEZ-VICENTE, E., ECHAVARRI-ERASUN, C.; RUBIO, L. M. Biosynthesis of Nitrogenase Cofactors. **Chemical reviews**, v.120, n.12, p. 4921–4968 2020.
- CHOCKALINGAM, S. P., PANNU, J., HOOSHMAND, S., THANKACHAN, S. V., & ALURU, S. An alignment- free heuristic for fast sequence comparisons with applications to phylogeny reconstruction. **BMC bioinformatics**, v,21, n. 404. 2020.
- COX, M.; ELLSWORTH, D. Report Application-Controlled Demand Paging for Out-of-Core Visualization. **NASA Ames Research Center**, 1997. Available at: <https://www.nas.nasa.gov/assets/pdf/techreports/1997/nas-97-010.pdf> . Accessed on April, 27th, 2021.
- DE PIERRI, C.R.; VOYCEIK, R.; SANTOS DE MATTOS, L.G.C.; KULIK, M.G.; CAMARGO, J.O.; REPULA DE OLIVEIRA, A.M.; DE LIMA NICHIO, B.T.; MARCHAUKOSKI, J.N.; DA SILVA FILHO., *et al.* SWeeP: Representing large biological sequences datasets in compact vectors. **Sci. Rep.**, 10, 91, 2020.
- DA SILVA FILHO, A.C.; MARCHAUKOSKI, J.N.; RAITTZ, R.T.; DE PIERRI, C.R.; MACHADO, D.J.S, FADEL- PICHETH, C.M.T.; PICHETH, G. Prediction and Analysis in silico of Genomic Islands in *Aeromonas hydrophila*. **Frontiers in microbiology**, v. 12, p. 769380-769380, 2021.
- DIXON, R., KAHN, D. Genetic regulation of biological nitrogen fixation. **Nature reviews. Microbiology**, v.2, n.8, p. 621–631, 2004.
- DOS SANTOS, P.C.; FANG, Z.; MASON, S.W.; SETUBAL, J.C.; DIXON, R. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. **BMC Genom.**, 13, 1471–2164, 2012.
- EFRON, B. The efficiency of logistic regression compared to normal. **Journal of the American Statistical Association.**, v. 70, n.352, p. 892-898, 1975.
- FAYYAD, U. Data mining and knowledge discovery in databases: implications for scientific databases. **Proceedings. Ninth International Conference on Scientific and Statistical Database Management**, p. 2-11,

1997.

FIX, E.; HODGES, J.L. Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties. **USAF School of Aviation Medicine**, 1951.

FLEUREN, W. W. M.; ALKEMA, W. Application of text mining in the biomedical domain. **Methods**, 74, 97–106, 2015. GARRITY G. M. A New Genomics-Driven Taxonomy of Bacteria and Archaea: Are We There Yet? **Journal of clinical microbiology**, v.54, 1956–1963, 2016.

GE, M.; BANGUI, H.; BUHNOVA, B. Big Data for Internet of Things: A Survey. **Future Generation Computer Systems**, v.87, p.601–614, 2018.

GHARAJEH, M. S. Biological Big Data Analytics. **A Deep Dive into NoSQL Databases: The Use Cases and Applications**, p.321–355, Elsevier: 2018.

GOOD, I. J.; GASKINS, A. The centroid method of numerical integration. **Numerische Mathematik**, v.16, n.4, p. 343–359, 1971.

HAND, D.J.; ADAMS, N.M. Data Mining. Data mining. **Wiley StatsRef: Statistics Reference Online**, 1-7, 2014.

HASSANI, H.; BENEKI, C.; UNGER, S.; MAZINANI, M.T.; YEGANEHI, M.R. Text Mining in Big Data Analytics. **Big Data Cogn. Comput.** v.4, n.1, 2020.

HOLLAND, J.H. Adaptation in Natural and Artificial Systems. **Univ. of Michigan Press.**, 1975.

HUGENHOLTZ, P., CHUVOCHINA, M., OREN, A., PARKS, D. H., SOO, R. M. Prokaryotic taxonomy and nomenclature in the age of big sequence data. **The ISME journal**, Advance online publication, 2021. <https://doi.org/10.1038/s41396-021-00941-x>

HUGENHOLTZ, P., SKARSHEWSKI, A., PARKS, D. H. Genome-Based Microbial Taxonomy Coming of Age **Cold Spring Harbor perspectives in biology**, v.8, n.6, a018085, 2016.

IZENMAN, A. J. Multivariate Regression. **Modern Multivariate Statistical Techniques**, p. 159–194, 2013.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Computing Surveys**, v.31, n.3, p.264–323, 1999.

JANIESCH, C.; ZSCHECH, P.; HEINRICH, K. Machine learning and deep learning. **Electron Markets**, 2021. JOHNSON, S. C. Hierarchical clustering schemes. **Psychometrika**, v.3, n. 32, p. 241–54, 1967.

KAUFMAN L, ROUSSEEUW PJ. Clustering by means of medoids. **Statistical data analysis based on the L1-norm and related methods**. Springer: 1987. p. 405–416.

KODRATOFF, Y. Knowledge discovery in texts: A definition, and applications. **Lect. Notes Comput. Sci.**, 1609, 16–29, 1999.

KOTELUK, O.; WARTECKI, A.; MAZUREK, S.; KOŁODZIEJCZAK, I.; MACKIEWICZ, A. How Do Machines Learn? Artificial Intelligence as a New Era in Medicine. **J. Pers. Med.**, v. 11, n. 32, 2021.

KRALLINGER, M., RABAL, O., LOURENÇO, A., OYARZABAL, J., VALENCIA, A. A. Information retrieval and text mining technologies for chemistry. **Chem. Rev.** v.117, p. 7673–7761, 2017.

LEE, J. K.; WILLIAMS, P. D.; CHEON, S. Data mining in genomics. **Clinics in laboratory medicine**, v.28, n.1, 145–148, 2008.

LEIMEISTER CA, SCHELLHORN J, DÖRRER S, GERTH M, BLEIDORN C, MORGENSTERN B. Prot-SpaM: fast alignment-free phylogeny reconstruction based on whole-proteome sequences. **Gigascience** 8, 1–14, 2018.

LEIMEISTER, C. A., BODEN, M., HORWEGE, S., LINDNER, S. & MORGENSTERN, B. Fast alignment-free sequence comparison using spaced-word frequencies. **Bioinformatics** 30, 1991–1999, 2014.

LI, X.X.; LIU, Q.; LIU, X.M.; SHI, H.W.; CHEN, S.F. Using synthetic biology to increase nitrogenase activity. *Microbial Cell Factories*, vol. 15, n. 43. 2016.

LI, Y., HE, L., LUCY HE, R. & YAU, S. S. T. A novel fast vector method for genetic sequence comparison. **Sci.Rep.** 7, 1–11, 2017.

LINNAEUS, C. **Systema naturae**, v.1. Holmiae: Impensis Direct. Laurentii Salvii, 1758.

MCCALLUM, A.; NIGAM, K. A comparison of event models for naive Bayes text classification. In **AAAI-98 Workshop on Learning for Text Categorization**. Tech. rep. WS-98-05, AAAI Press. <http://www.cs.cmu.edu/mccallum>, 1998.

MCCARTHY B.J., BOLTON E.T. An approach to the measurement of genetic relatedness among organisms. **Proc Natl Acad Sci U S A**. v.50, n.1, p.156-164. 1963.

MCCULLOCH, W.S., PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics** v.5, p. 115–133, 1943.

NAVARRO, F., MOHSEN, H., YAN, C., LI, S., GU, M., MEYERSON, W., GERSTEIN, M. Genomics and data science: an application within an umbrella. **Genome Biol**, v.20, n.1, 109 p., 2019.

O'MARA-EVES, A., THOMAS, J., MCNAUGHT, J., MIWA, M., ANANIADOU, S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. **Systematic Reviews**, v.4, n.5, 2015.

PARKS D.H., CHUVOCHINA M., WAITE D.W., RINKE C., SKARSHIEWSKI A, CHAUMEIL P.A., HUGENHOLTZ P. A. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. **Nat. Biotechnol.** 36, 996, 2018.

PARKS, D.H., CHUVOCHINA, M., CHAUMEIL, PA. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. **Nat Biotechnol** v.38, p.1079–1086, 2020.

PAUL, B., DIXIT, G., MURALI, T. S.; SATYAMOORTHY, K. Genome-based taxonomic classification. **Genome**, v.62, n.2, p. 45–52, 2019.

PANKIEVICZ, V.C.S., IRVING, T.B., MAIA, L.G.S. *et al.* Are we there yet? The long walk towards the development of efficient symbiotic associations between nitrogen-fixing bacteria and non-leguminous crops. **BMC Biol** v.17, n.99, 2019.

PEDROSA, F.O; MONTEIRO, R.A; WASSEM, R., *et al.* Genome of *Herbaspirillum seropedicae* Strain SmR1, a Specialized Diazotrophic Endophyte of Tropical Grasses. Richardson PM, ed. PLoS Genetics, vol. 7 n. 5, 2011.

PERICO, C.P.; DE PIERRI, CR.; NETO, G.P.; FERNANDES, D.R.; PEDROSA, F.O.; SOUZA, E.M.; RAITTZ, R.T. Genomic landscape of SARS-CoV-2 pandemic in Brazil suggests an external P. 1 variant origin. **medRxiv**, 2021.

POPCHEV, I.; OROZOVA, D. Data Science: Experience and Trends. **IOP Conf. Ser.: Mater. Sci. Eng.**, 1031, 1-8, 2021.

RAITTZ, R. T., DE PIERRI, C. R., MALUK, M., BUENO BATISTA, M., CARMONA, M., JUNGHARE, M., FAORO, H., CRUZ, L. M., BATTISTONI, F., SOUZA, E. *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. **Genes (Basel)**, v.12(1), n.71, 2021.

ROSSELLÓ-MÓRA, R., WHITMAN, W. B. Dialogue on the nomenclature and classification of prokaryotes. **Systematic and applied microbiology**, v.42, 5–14, 2019.

ROSSI, R.; GRIFANTINI, R. M. Big Data: Challenge and Opportunity for Translational and Industrial Research in Healthcare. **Frontiers in Digital Humanities**, n.5, 1-13, 2018.

SANDERSON, M.J., PURVIS, A., HENZE, C. Phylogenetic supertrees: assembling the trees of life. **Trends Ecol Evol.** v.13, p.105–9, 1998.

SAITOU, N., NEI, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. **Molecular Biology and Evolution** v.4, p.406–425, 1987.

STEINHAUS, H. Sur la division des corps materiels en parties. **Bulletin de l'Academie Polonaise des Sciences**, vol. IV, n. 12, p.801-804, 1956.

SAUERBREI, W., PERPEROGLOU, A., SCHMID, M., ABRAHAMOWICZ, M., BECHER, H., BINDER, H., DUNKLER, D., HARRELL, F., ROYSTON, P., HEIZE, G. State of the art in selection of variables and functional forms in multivariable analysis. **Diagn Progn Res.**, v.4, n.3, 2020.

SCHRIDER, D. R.; KERN, A. D. Supervised Machine Learning for Population Genetics: A New Paradigm. **Trends in Genetics**, v..34, n.4, 301–312, 2018.

SHEIKH, N. Performance Variables and Model Development. **Implementing Analytics**, p. 63–84, 2013.

SHIRABAD, J. Predictive Techniques in Software Engineering. **Encyclopedia of Machine Learning**. Springer: Boston, 2011.

SIMON, C., DAVIDSEN, K., HANSEN, C., SEYMOUR, E., BARNKOB, M. B., OLSEN, L. R. BioReader: A text mining tool for performing classification of biomedical literature. **BMC Bioinformatics** v.19, 2019.

SINGHAL, A.; SIMMONS, M.; LU, Z. Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine. **PLoS Comput. Biol.** 12, 1–19, 2016.

SNEATH, P. H. A. The applications of computers to taxonomy. **Gen. Microbiol**, v.1, n.17, p. 201-226, 1957.

SNEATH, P.H.A., SOKAL, R.R. Numerical taxonomy. **Nature**.v.193, p.855–60, 1962.

SOKAL, R.R.; MICHENER, C.D. A statistical method for evaluating systematic relationships". **University of Kansas Science Bulletin.**, v.38: p.1409–1438, 1958.

SORENSEN, T.A. Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species and Its Application to Analyses of the Vegetation on Danish Commons. **Biologiske Skrifter/Kongelige Danske Videnskabernes Selskab**, v.5, p.1-34,1948.

STEPHENS, Z.D., LEE, S.Y., FAGHRI, F., CAMPBELL, R.H., ZHAI, C., EFRON, M.J., RAVISHANKAR, I., SCHATZ, M., SINHA, S., ROBINSON, G. Big Data: Astronomical or Genomical? **PLOS Biology**, v.13, n.7, 2015.

TEMME, K.; ZHAO, D.; VOIGT, C.A. Refactoring the nitrogen fixation gene cluster from *Klebsiella oxytoca*. Proceedings of the National Academy of Sciences of the United States of America, vol. 109 n.18, p. 7085- 7090, 2012.

TSUI, K.L., CHEN, V., JIANG, W., ASLANDOGAN, Y. Data Mining Methods and Applications. **Springer Handbook of Engineering Statistics**, p.651–669, 2006.

VANDAMME, P., POT, B., GILLIS, M., DE VOS, P., KERSTERS, K., SWINGS, J. Polyphasic taxonomy, a consensus approach to bacterial systematics. **Microbiol Rev.** v.60, p.407–38, 1996.VAPNIK V. Statistical Learning Theory. **Wiley**: New York, 1998.

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**, v. 7, n. 4, 2021.

WAGNER, S. C. Biological Nitrogen Fixation. **Nature Education Knowledge** v.3, n.10, p.15, 2011.

WAMBA, S. F., AKTER, S., EDWARDS, A., CHOPIN, G.; GNANZOU, D. How “big data” can make big impact: Findings from a systematic review and a longitudinal case study. **International Journal of Production Economics**, v.165, p.

234–246, 2015.

WARD, J. H. Hierarchical grouping to optimize an objective function. **Journal of American Statistics Association**, v.1, n.58, p.236–244, 1963.

WHITE, M. Digital workplaces: Vision and reality. **Business Information Review**, v.29, n.4, 2012.

WOESE, C.R., FOX, G.E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. **Proc Natl Acad Sci USA**. v.74, p.5088–90, 1977.

YE, Z., TAFTI, A. P., HE, K. Y., WANG, K., HE, M. M. SparkText: Biomedical text mining on big data framework. **PLoS One** v.11, p.1–15, 2016.

ZHU, Q., MAI, U., PFEIFFER, W., JANSSEN, S., ASNICAR, F., SANDERS, J. G., BELDA-FERRE, P., AL-GHALITH, G. A., KOPYLOVA, E., MCDONALD, D. *et al.* Phylogenomics of 10,575 genomes reveal evolutionary proximity between domains Bacteria and Archaea. **Nat. Comm.**, v.10, n. 5477, 2019.

ZUCKERKANDL, E., PAULING, L. Molecules as documents of evolutionary history. **J Theor Biol**. v.8, p.357–5.

CAPÍTULO III: O MODELO SWeeP PARA REPRESENTAÇÃO VETORIAL DE PROTEOMAS DE BACTÉRIAS

Este capítulo diz respeito à consolidação e exploração do modelo SWeeP para a representação de proteomas bacterianos. Esses resultados foram adaptados do artigo originalmente publicado na *Scientific Reports*, no entanto, aqui só serão evidenciados os resultados obtidos da exploração dos proteomas bacterianos desenvolvidos para esta tese (o conceito matemático SWeeP e teste de desempenho são encontrados na versão completa do artigo).

Referência:

Camilla Reginatto De Pierri[†], Ricardo Voyceik[†], Letícia Graziela Costa Santos de Mattos, Mariane Gonçalves Kulik, Josué Oliveira Camargo, Aryel Marlus Repula de Oliveira, Bruno Thiago de Lima Nichio, Jeroniza Nunes Marchaukoski, Antonio Camilo da Silva Filho, Dieval Guizelini, J. Miguel Ortega, Fabio O. Pedrosa, Roberto Tadeu Raittz. SWeeP: representing large biological sequences datasets in compact vectors. *Sci Rep* 10, 91 (2020). <https://doi.org/10.1038/s41598-019-55627-4> ©.

[†]Estes autores contribuíram igualmente para este trabalho.

1 INTRODUÇÃO

As análises e comparações de sequências biológicas são tradicionalmente realizadas usando algoritmos de alinhamento, sendo o BLAST a ferramenta mais comumente usada (ALTSCHUL *et al.*, 1990). Mesmo com técnicas de computação dinâmica (OTU, SAYOOD, 2003; EDDY, 2004), o alinhamento de grandes conjuntos de dados exige muito tempo e se torna inviável quando é necessário analisar genomas completos (VINGA, ALMEIDA, 2003). Além disso, a aplicação de técnicas de alinhamento pode se tornar problemática quando a identidade de sequência é baixa porque as matrizes de substituição podem afetar *significativamente* os resultados de alinhamento (ZIELEZINSKI *et al.*, 2017).

Essa necessidade levou ao desenvolvimento de alternativas para acelerar comparações de dados estruturados (CHEN *et al.*, 2016; ZANG *et al.*, 2017; LI *et al.*, 2017). Vários estudos usaram com sucesso métodos sem alinhamento para as análises comparativas de genomas completos e outros grandes conjuntos de dados de sequência biológica (WU *et al.*, 2009; SIMS *et al.*, 2009; JUN *et al.*, 2010; HAUBOLD, 2014; LEIMEISTER *et al.*, 2018), mas a investigação dessas técnicas ainda é necessária para verificar sua eficácia. Portanto, abordagens baseadas no mapeamento de frequências de palavras relativas (k-mers) em espaços vetoriais têm sido objeto de vários estudos recentes (VINGA, ALMEIDA, 2003; WU *et al.*, 2009; SIMS *et al.*, 2009; JUN *et al.*, 2010; BODEN *et al.*, 2013; LEIMEISTER *et al.*, 2014; HORWEGE *et*

et al., 2014; NOÉ, MARTIN, 2014; VINGA, 2014; HAUBOLD, 2014; CHEN *et al.*, 2016; ZANG *et al.*, 2017; LI *et al.*, 2017; ZIELEZINSKI *et al.*, 2017; LEIMEISTER *et al.*, 2018). As representações vetoriais de proteínas facilitam o manuseio de dados e permitem o uso de mineração de dados para identificar características importantes escondidas em sequências biológicas (BASTOLLA, 2005; TEICHERT, PORTO, 2006; VINGA, 2014; ASGARI, MOFRAD, 2015).

Utilizamos o modelo SWeeP para criar representações vetoriais para todos os proteomas bacterianos completos, representando o conjunto de todas as proteínas traduzidas das sequências disponíveis no NCBI (*National Center of Biotechnology Information*). Em seguida, isolamos as representações dos gêneros bacterianos de alguns organismos modelo (*Corynebacterium*, *Klebsiella* e *Escherichia*) e utilizamos uma abordagem de aprendizado de máquina para demonstrar a capacidade de classificação do modelo SWeeP geral.

2 MÉTODOS

SWeeP é baseado na representação vetorial de sequências de proteínas como um modelo compacto baseado na projeção de conjuntos de k-mers em uma base quasi-ortonormal orientada aleatoriamente, com um número suficiente de coordenadas para manter comparações entre sequências. SWeeP usa palavras espaçadas (BODEN *et al.*, 2013) para escanear sequências e criar índices, que são empregados para criar um vetor de alta dimensão (HDV). O HDV permite a redução da dimensionalidade em sua projeção em um vetor de dimensão inferior e mantém a maior parte das informações de comparação, conforme proposto por Johnson e Lindenstrauss (JOHNSON; LINDENSTRAUSS, 1984).

A entrada para SWeeP consiste em um arquivo multiFASTA contendo sequências de aminoácidos. No caso de um proteoma contendo várias sequências, as proteínas são concatenadas, separadas por delimitadores - que são ignorados na construção dos vetores - para formar uma única sequência para cada conjunto de proteínas de um genoma. Cada proteoma é então representado por uma matriz bidimensional M que contém as palavras espaçadas de toda a sequência concatenada, a partir de uma máscara predefinida. A matriz M pode ser construída em dois formatos: (1) reversível e (2) binária. A forma binária foi usada para os estudos de caso descritos neste artigo, pois é econômica e de processamento rápido.

A matriz M , quando linearizada por colunas, é um vetor que reflete dados de sequência altamente representativos, que designamos como um vetor de dimensão superior (HDV). O modelo SWeeP projeta esses HDVs em uma base quasi-ortonormal, de acordo com o lema de Johnson - Lindenstrauss (JOHNSON; LINDENSTRAUSS, 1984) com o objetivo de permitir um desempenho eficiente com grandes conjuntos de dados.

2.1 Implementação

As funções específicas do SWeeP usadas neste estudo foram implementadas usando a linguagem de programação MATLAB. As etapas apresentadas na definição dos parâmetros para o caso de estudo bacteriano foram realizadas em um processador Intel Xeon de 40 núcleos com 256GB de RAM rodando Ubuntu 16.04.1 LTS. O SWeeP está disponível gratuitamente para ambos os sistemas operacionais, em: <https://sourceforge.net/projects/spacedwordsprojection/>. Para testes de aprendizado de máquina, usamos o Weka Software (HALL *et al.*, 2000).

A visualização e manipulação das árvores filogenéticas construídas usando SWeeP foram realizadas usando Dendroscope 3 (HUSON; SCORNAVACCA *et al.*, 2012). As CDSs dos genomas bacterianos usados para a representação gráfica da matriz M foram obtidos no NCBI. Os organismos e seus respectivos números de acesso são fornecidos na Tabela Suplementar S1.

2.2 Definição de parâmetros

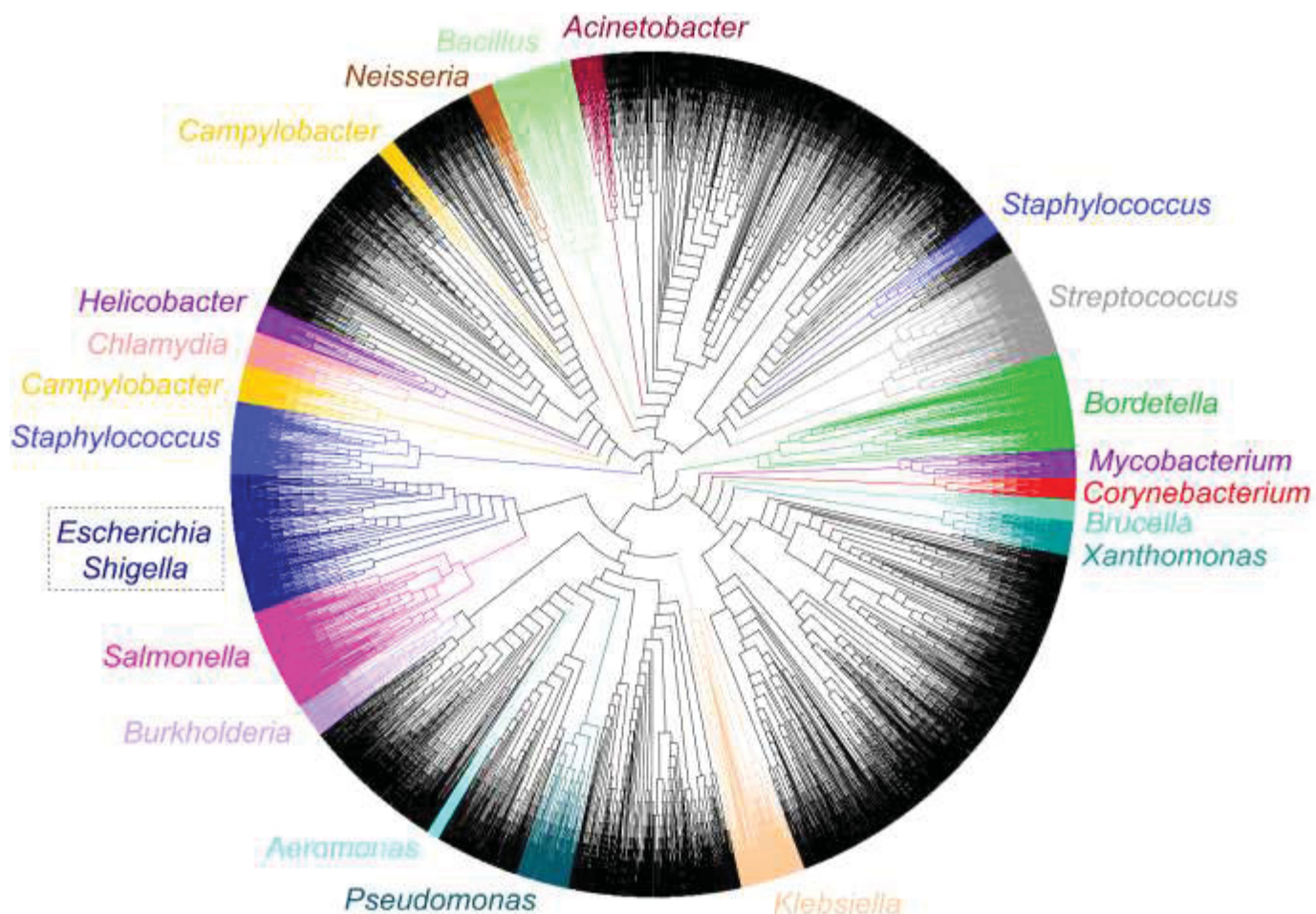
Neste estudo, os vetores HDV foram obtidos a partir de uma matriz binária. A projeção utilizada foi de 600 coordenadas; os k-mers espaçados foram selecionados aplicando-se a máscara “11011” por ser sugerida pela literatura como uma boa escolha para proteínas (MA *et al.*, 2002). No entanto, o método SWeeP é ajustável e permite várias projeções e vários k-mers (por meio da alteração e / ou adição de máscaras), permitindo que o modelo seja ajustado aos dados a serem minerados, visto que o tamanho do k-mer aumenta a dificuldade computacional.

3 RESULTADOS E DISCUSSÃO

3.1 Estudo de caso: Representação Vetorial de proteomas bacterianos

Um vetor SWeeP com projeção de 1369 coordenadas foi criado para as sequências codificadoras (CDS) de todos os genomas bacterianos completos disponíveis no NCBI (10.324 bactérias) e uma árvore filogenética foi construída a partir dos proteomas usando o método de Ward (WARD, 1963). Neste estudo de caso, o método SWeeP provou ser computacionalmente eficaz quando aplicado a um grande número de proteomas de tamanho considerável. A árvore filogenética global dos táxons bacterianos, apresentada na FIGURA 1.3, foi analisada manualmente, e os resultados foram consistentes com árvores relatadas na literatura (LANG *et al.*, 2013; HUG *et al.*, 2016; PARKS *et al.*, 2018).

FIGURA 1.3. Árvore filogenética gerada pelo SWeeP com 10,324 proteomas bacterianos



Fonte: Autora, 2022.

Legenda: Árvore filogenética com 10,324 organismos, apresentando gêneros com ramos completamente homogêneos. Em contraste (quadrado tracejado), os gêneros *Escherichia* e *Shigella* com distribuição heterogênea.

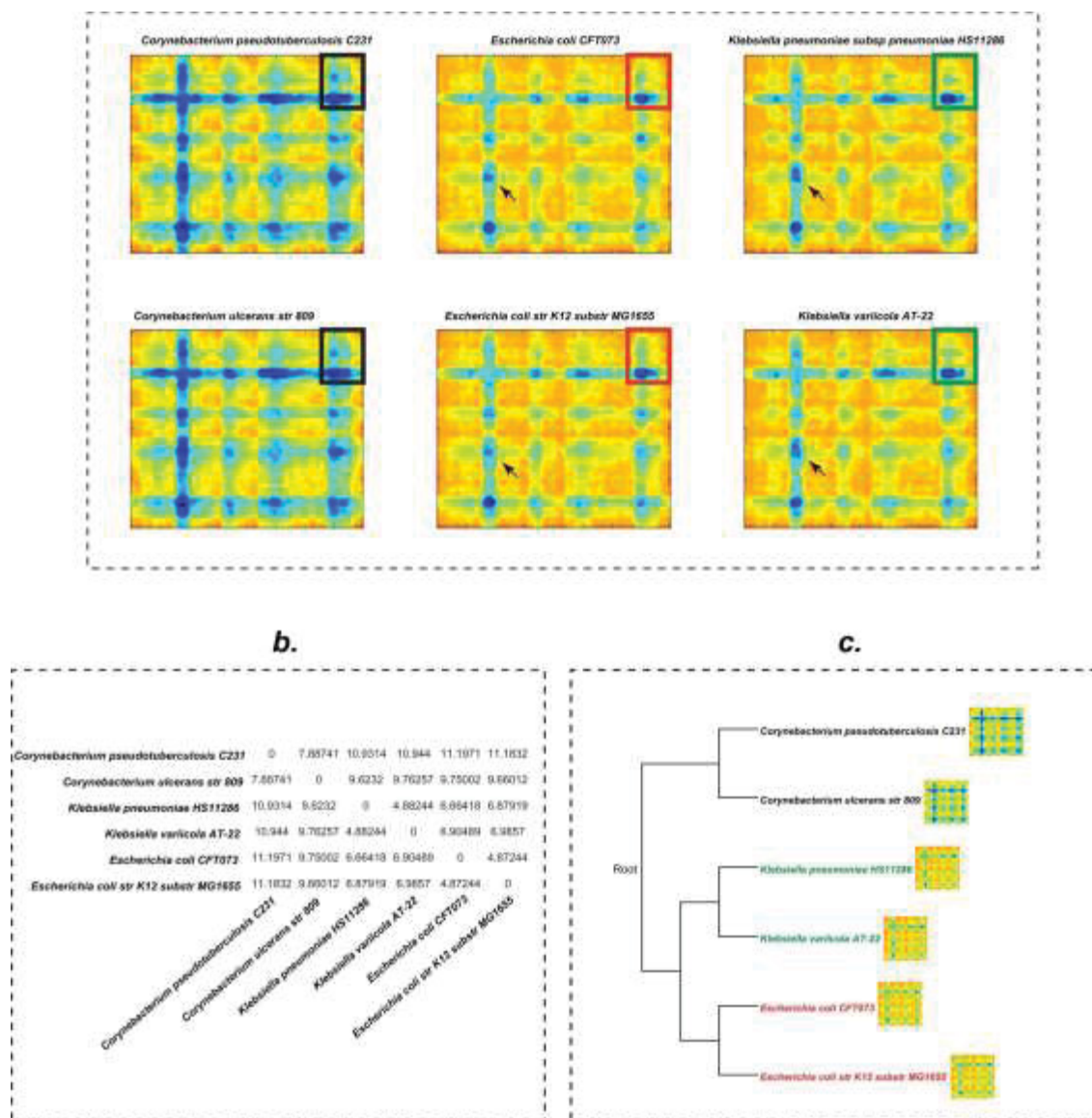
No estudo de Parks *et al.* (2018), foi inferida uma árvore filogenética a partir da concatenação de 120 proteínas (representando aproximadamente 4% de um genoma bacteriano), abrangendo 94,759 genomas bacterianos. Em contrapartida, em nosso estudo a inferência filogenética é realizada com 10,324 proteomas bacterianos completos, que além de representar todas as características genotípicas das bactérias estudadas, apresenta um maior volume de dados processados quando comparado ao estudo de Parks *et al.*, (2018).

Outro objetivo desta análise foi obter uma visualização gráfica e comparação detalhada dos genomas de *Corynebacterium*, *Klebsiella* e *Escherichia*. Também escolhemos cepas específicas desses organismos modelo, *Escherichia coli* K12 e CFT073, *Klebsiella pneumoniae* HS11286, *Klebsiella variicola*

AT-22, *Corynebacterium pseudotuberculosis* C231 e *Corynebacterium ulcerans* 809 (Tabela Suplementar S1) para análise.

Uma matriz de *Mb* binária e a respectiva projeção SWeeP de (aqui com projeção tamanho 600) foram criadas para cada bactéria. A Figura 2.3a representa *heat maps* (mapas de calor) das matrizes; cada coluna corresponde a bactéria do mesmo gênero (um gênero por coluna). Pode-se ver que bactérias do mesmo gênero são mais semelhantes entre si do que bactérias de gêneros diferentes. No entanto, detalhes revelam que mesmo os microrganismos pertencentes ao mesmo gênero diferem uns dos outros (retângulos). Este efeito fica mais claro quando a matriz de distância (Figura 2.3b) e o dendrograma (Figura 2.3c) construído a partir da matriz de distância SWeeP das seis bactérias são comparados. Microrganismos filogeneticamente mais próximos apresentam distâncias recíprocas menores e maior proximidade no dendrograma.

FIGURA 2.3. Representação gráfica dos proteomas bacterianos



Fonte: De Pierri *et al.*, (2020).

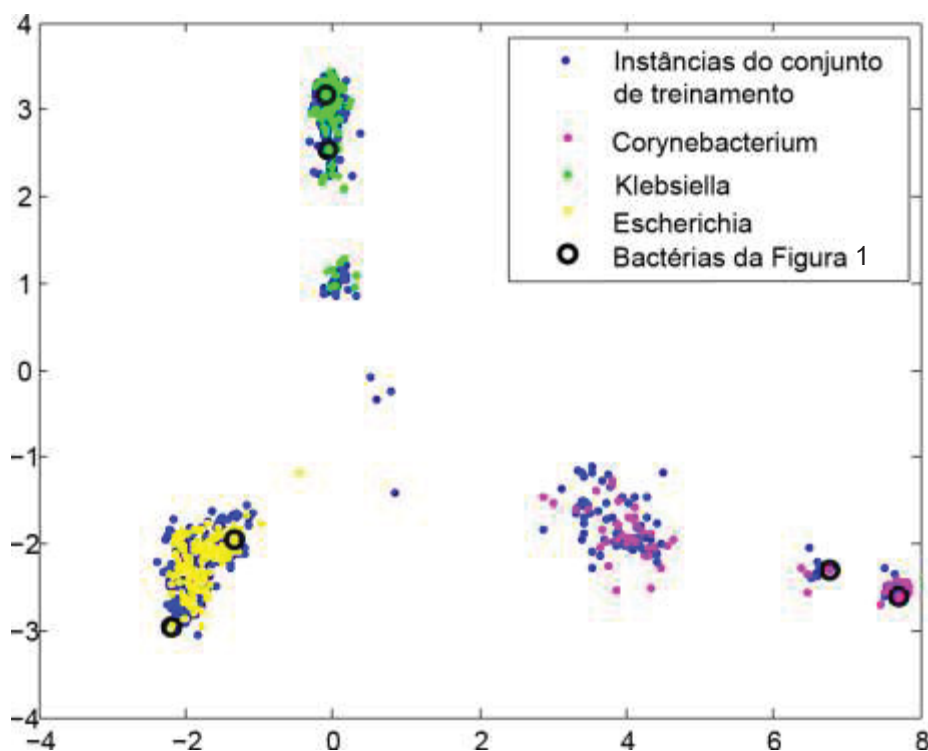
Legenda: *Heat maps* de proteomas bacterianos representados pela matriz M (400 × 400) (a) Os quadrados pretos destacados nas representações do proteoma bacteriano mostram uma das áreas de semelhança entre os proteomas das cepas de *Corynebacterium*. Os quadrados vermelhos mostram uma das semelhanças entre os proteomas das cepas de *Escherichia coli*. Os quadrados verdes mostram áreas de semelhança entre as espécies de *Klebsiella*. As setas apontam para regiões com menor similaridade entre os proteomas de *Escherichia coli* e *Klebsiella*. (b) Distância euclidiana entre as projeções de 600 coordenadas dos proteomas bacterianos em a. (c) A árvore filogenética de bactérias criada por uma projeção SWEEP de 600 coordenadas.

Dos 10.324 proteomas bacterianos projetados com SWEEP em 600 coordenadas, selecionamos 1001 organismos e os classificamos como 1 - *Corynebacterium*, 2 - *Klebsiella* e 3 - *Escherichia*. Um conjunto de treinamento com 700 instâncias foi utilizado para construir uma *Support Vector Machine* (SVM) treinada no software Weka (HALL, 2000). Em seguida, testamos o modelo treinado nas 301 instâncias

restantes. Os resultados são mostrados na Figura Suplementar S1. Nesse teste, todas as instâncias dos três gêneros foram classificadas corretamente. Pode-se observar que embora as diferenças entre *Escherichia* e *Klebsiella* sejam difíceis de perceber (FIGURA 2.3a), elas foram classificadas corretamente por meio do aprendizado de máquina.

Além disso, para a visualização dos dados, os dois componentes principais da análise de componentes principais (PCA) dos vetores do conjunto de treinamento foram plotados para cada instância nos conjuntos de treinamento e teste (FIGURA 3.3). Aqui, pode-se ver que todos os três gêneros estão separados de forma consistente, incluindo os organismos referenciados na Figura 2.3.

FIGURA 3.3. Análise de dois componentes principais da projeção SWeeP para os três gêneros de organismos modelos



Fonte: Adaptado de De Pierri *et al.*, (2020).

Legenda: A figura destaca o potencial de comparabilidade das informações contidas na projeção SWeeP com 600 coordenadas. Em azul os pontos do conjunto de treinamento (usado para criar a base do PCA). Em rosa, amarelo e verde os pontos de ajuste de teste. Os círculos pretos marcam as instâncias comparativas da Figura 1.

Seria computacionalmente inviável realizar PCA com a matriz SWeeP completa (160k) por causa da alta dimensionalidade, que no caso da máscara "11011" exigiria pelo menos 2×10^{11} bytes de memória, ou seja, aproximadamente 190 GB. Em contraste, a projeção SWeeP em coordenadas de 600

requer apenas 0,027 GB. Assim, propomos uma abordagem viável para PCA para grandes volumes de sequências.

4 CONCLUSÃO

Embora as técnicas de alinhamento sejam adequadas para a comparação de sequências de proteínas, sua aplicação à análise de grandes volumes de dados é limitada. Neste estudo, aplicamos o modelo SWeeP para analisar um grande número de sequências de proteomas bacterianos, e pela primeira vez, árvores filogenéticas foram construídas automaticamente. Também utilizamos o SWeeP para aprendizado de máquina e PCA, mostrando sua eficácia nessas aplicações.

Com o modelo SWeeP é possível realizar a construção rápida e sensível de árvores filogenéticas compactas, sendo uma alternativa eficiente para comparações de sequências de grandes conjuntos de dados. Embora existam outras técnicas alternativas de comparação de sequências, o alinhamento de sequências ainda é o método preferido para a maioria das comparações de sequências biológicas. SWeeP é um método novo e inovador sem alinhamento que pode substituir as técnicas de alinhamento tradicionais na comparação de sequências biológicas.

A representação vetorial de informações é usada universalmente e é bem definida na maioria das áreas. Propomos esta nova abordagem para que os recursos matemáticos e computacionais disponíveis possam ser estendidos e aplicados mais facilmente na análise e mineração de sequências biológicas.

REFERÊNCIAS

- ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. Basic local alignment search tool. **J. Mol. Biol.** 215, 403–410, 1990.
- ASGARI, E. & MOFRAD, M. R. K. Continuous distributed representation of biological sequences for deep proteomics and genomics. **PLoS One** 10, 1–15, 2015.
- BASTOLLA, U., PORTO, M., ROMAN, H. E. & VENDRUSCOLO, M. Principal eigenvector of contact matrices and hydrophobicity profiles in proteins. **Proteins Struct. Funct. Genet.** 58, 22–30, 2005.
- BODEN, M., SCHÖNEICH, M., HORWEGE, S., LINDNER, S., LEIMEISTER, C., MORGENSTERN, B. Alignment-free sequence comparison with spaced k-mers. **Ser. Inform.** 34, 24–34, 2013.
- CHEN, S., DENG, L.Y., BOWMAN, D. SHIAU, J.H.; WONG, T., MADAHIAN, B., LU, H.H. Phylogenetic tree construction using trinucleotide usage profile, TUP. **BMC Bioinformatics** 17, 119–130, 2016.
- EDDY, S.R. What is dynamic programming? 22, 909–910, 2004.
- DE PIERRI, C. R. Representações vetoriais de proteomas: Um estudo de caso. Com sequências mitocondriais. **Dissertação de Mestrado**. Programa de Pós-graduação em Bioinformática. Universidade Federal do Paraná: Curitiba, 2017.
- HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., WITTEN, I.H. The WEKA Data Mining Software: An Update. 11, 10–18, 2000.
- HAUBOLD, B. Alignment-free phylogenetics and population genetics. **Brief. Bioinform.** 15, 407–418, 2014.
- HORWEGE, S., LINDNER, S., BODEN, M., HATJE, K., KOLLMAR, M., LEIMEISTER, C., MORGENSTERN, B. Spaced words and kmacs: Fast alignment-free sequence comparison based on inexact word matches. **Nucleic Acids Res.** 42, 7–11, 2014.
- HUG LA, BAKER BJ, ANANTHARAMAN K, BROWN CT, PROBST AJ, CASTELLE CJ, BUTTERFIELD CN, HERNSDORF AW, AMANO Y, ISE K., *et al.* A new view of the tree of life. **Nat. Microbiol.** 1, 1–6, 2016.
- HUSON, D. H. & SCORNAVACCA, C. DENDROSCOPE 3: An interactive tool for rooted phylogenetic trees and networks. **Syst. Biol.** 61, 1061–1067, 2012.
- JOHNSON, W. B., LINDENSTRAUSS, J. Extensions of Lipschitz mappings into a Hilbert space. **Contemp. Math.** 26, 189–206, 1984. JUN, S.-R., SIMS, G. E., WU, G. A. & KIM, S.H. Whole-proteome phylogeny of prokaryotes by feature frequency profiles: An alignment-free method with optimal feature resolution. **Proc. Natl. Acad. Sci. USA** 107, 133–138, 2010.
- LANG, J. M., DARLING, A. E. & EISEN, J. A. Phylogeny of bacterial and archaeal genomes using conserved genes: Supertrees and supermatrices. **PLoS One** 8, 1–14, 2013.
- LEIMEISTER, C. A., SOHRABI-JAHROMI, S. & MORGENSTERN, B. Fast and accurate phylogeny reconstruction using candidate virulence factors. **BMC Genomics** 12, 2011.
- LEIMEISTER, C. A. & MORGENSTERN, B. Kmacs: The k-mismatch average common substring approach to alignment-free sequence comparison. **Bioinformatics** 30, 2000–2008, 2014.
- LEIMEISTER, C. A., BODEN, M., HORWEGE, S., LINDNER, S. & MORGENSTERN, B. Fast alignment-free sequence comparison using spaced-word frequencies. **Bioinformatics** 30, 1991–1999, 2014.
- LEIMEISTER, C.A., SCHELLHORN J, DÖRRER S, GERTH M, BLEIDORN C, MORGENSTERN B. Prot-SpaM: fast alignment-free phylogeny reconstruction based on whole-proteome sequences. **Gigascience** 8, 1–14, 2018.
- LI, Y., HE, L., LUCY HE, R. & YAU, S. S. T. A novel fast vector method for genetic sequence comparison. **Sci. Rep.** 7,

1–11, 2017.

LIU P, LI P, JIANG X, BI D, XIE Y, TAI C, DENG Z, RAJAKUMAR K, OU HY. Complete genome sequence of *Klebsiella pneumoniae* subsp. *pneumoniae* HS11286, a multidrug-resistant strain isolated from human sputum. *J. Bacteriol.* 194, 1841–1842, 2012.

MA, B., TROMP, J. & LI, M. PatternHunter: Faster and more sensitive homology search. *Bioinformatics* 18,440–445, 2002.

NOÉ, L. & MARTIN, D. E. K. A coverage criterion for spaced seeds and its applications to support vector machine string kernels and k -mer distances. *J. Comput. Biol.* 21, 947–963, 2014.

OTU, H. H.; SAYOOD, K. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics* 19, 2122–2130, 2003.

PARKS DH, CHUVOCHINA M, WAITE DW, RINKE C, SKARSHIEWSKI A, CHAUMEIL PA, HUGENHOLTZ P. A. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* 36, 996, 2018.

PINTO-TOMÁS AA, ANDERSON MA, SUEN G, STEVENSON DM, CHU FS, CLELAND WW, WEIMER PJ, CURRIE CR. Symbiotic nitrogen fixation in the fungus gardens of leaf-cutter ants. *Science*, 80. 326, 1120–1123, 2009.

RILEY M, ABE T, ARNAUD MB, BERLYN MK, BLATTNER FR, CHAUDHURI RR, GLASNER JD, HORIUCHI T, KESELER IM, KOSUGE T., *et al.* *Escherichia coli* K-12: A cooperatively developed annotation snapshot -2005. *Nucleic Acids Res.* 34, 1–9, 2006.

RUIZ JC, D'AFONSECA V, SILVA A, ALI A, PINTO AC, *et al.* Evidence for Reductive Genome Evolution and Lateral Acquisition of Virulence Functions in Two *Corynebacterium pseudotuberculosis* Strains. *PLoS ONE* v.6, n.4, p. e18551, 2011.

SIMS, G. E., JUN, S.-R., WU, G. A. & KIM, S.H. Alignment-free genome comparison with feature frequency profiles, FFP) and optimal resolutions. *Proc. Natl. Acad. Sci. USA* 106, 2677–2682, 2009.

TEICHERT, F. & PORTO, M. Vectorial representation of single- and multi-domain protein folds. *Eur. Phys. J.B.* 54, 131–136, 2006.

TROST, E., AL-DILAIMI, A., PAPAVALIOU, P. *et al.* Comparative analysis of two complete *Corynebacterium ulcerans* genomes and detection of 15/15 candidate virulence factors. *BMC Genomics* v.12, 383, 2011.

VINGA, S. & ALMEIDA, J. Alignment-free sequence comparison - A review. *Bioinformatics* 19, 513–523, 2003.

VINGA, S. Editorial: Alignment-free methods in computational biology. *Brief. Bioinform.* 15, 341–342, 2014.

WARD, J. H. Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc.* 58, 236–244, 1963.

WELCH RA, BURLAND V, PLUNKETT G RD, REDFORD P, ROESCH P, RASKO D, BUCKLES EL, LIOU SR, BOUTIN A, HACKETT J., *et al.* Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *PNAS*;1–5, 2002.

WU, G. A., JUN, S.-R., SIMS, G. E. & KIM, S.H. Whole-proteome phylogeny of large dsDNA virus families by an alignment-free method. *Proc. Natl. Acad. Sci. USA* 106, 12826–12831, 2009.

ZHANG, Q., JUN, S. R., LEUZE, M., USSERY, D. & NOOKAEW, I. Viral phylogenomics using an alignment-free method: A three-step approach to determine optimal length of k -mer. *Sci. Rep.* 7, 1–13, 2017.

ZIELEZINSKI, A., VINGA, S., ALMEIDA, J. & KARLOWSKI, W. M. Alignment-free sequence comparison: Benefits, applications, and tools. *Genome Biol.* 18, 1–17, 2017.

SUPPLEMENTARY INFORMATION

FIGURA S1.3. Resultados do teste de aprendizagem da máquina no software WEKA

```

==== Run information ====

Scheme: weka.classifiers.functions.LibSVM -S 0 -K 2 -D 3 -G 0.0 -R 0.0 -N 0.5 -M 40.0 -C 1.0 -E 0.001 -P 0.1 -model "C:\Program Files\Weka-3-8" -seed 1
Relation: VariavelMatlab
Instances: 700
Attributes: 601
[1 list of attributes omitted]
Test mode: user supplied test set: size unknown (reading incrementally)

==== Classifier model (full training set) ====

LibSVM wrapper, original code by Yasser EL-Manzalawy (= WLSVM)

Time taken to build model: 0.4 seconds

==== Evaluation on test set ====

Time taken to test model on supplied test set: 0.3 seconds

==== Summary ====

Correctly Classified Instances 301      100 %
Incorrectly Classified Instances 0      0 %
Kappa statistic 1
Mean absolute error 0
Root mean squared error 0
Relative absolute error 0 %
Root relative squared error 0 %
Total Number of Instances 301

==== Detailed Accuracy By Class ====

TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class
1,000  0,000  1,000  1,000  1,000  1,000  1,000  1,000  1
1,000  0,000  1,000  1,000  1,000  1,000  1,000  1,000  2
1,000  0,000  1,000  1,000  1,000  1,000  1,000  1,000  3
Weighted Avg. 1,000  0,000  1,000  1,000  1,000  1,000  1,000  1,000

==== Confusion Matrix ====

a b c <- classified as
62 0 0 | a = 1
0 92 0 | b = 2
0 0 147 | c = 3

```

Fonte: De Pierri *et al.*, (2020).

Legenda: A entrada é vetores SWeep com 600 coordenadas; A classe 1 representa organismos do gênero *Corynebacterium*; classe 2, *Klebsiella*; e classe 3 *Escherichia*. Todas as instâncias foram classificadas corretamente no conjunto de teste - 301 instâncias não usadas para treinar o modelo (consulte a matriz de confusão).

TABELA S1.3. Organismos e sequências de referência obtidas do NCBI usadas para a construção de heat maps

Organisms	NCBI sequence	Publication
<i>Corynebacterium pseudotuberculosis</i> C231	NC_017301.1	Ruiz <i>et al.</i> (2011)
<i>Corynebacterium ulcerans</i> str 809	NC_017317.1	Trost <i>et al.</i> (2011)
<i>Klebsiella pneumoniae</i> HS11286*	NC_016845.1	Liu <i>et al.</i> (2012)
<i>Klebsiella variicola</i> AT-22	NC_013850.1	Pinto-Tomás (2009)
<i>Escherichia coli</i> CFT073	NC_004431.1	Welch <i>et al.</i> (2002)
<i>Escherichia coli</i> str K12 substr MG1655	NC_000913.3	Riley <i>et al.</i> (2005)

* Chromosome sequence.

Fonte: De Pierri *et al.*, (2020).

CAPÍTULO IV: COMPARATIVE GENOMICS PROVIDES INSIGHTS INTO THE TAXONOMY OF *Azoarcus* AND REVEALS SEPARATE ORIGINS OF *nif* GENES IN THE PROPOSED *Azoarcus* AND *Aromatoleum* GENERA

Esses resultados foram publicados originalmente na *MDPI Genes*. Esta é uma pesquisa colaborativa, coordenada pelo Professor Dr. Euan K. James, do *The James Hutton Institute*. Nesta tese, a participação nos resultados diz respeito à aplicação do modelo SWeeP para apoio à análise taxonômica. O artigo está disponível em: <https://doi.org/10.3390/genes12010071>

Referência:

Roberto T. Raittz[†], Camilla R. De Pierri[†], Marta Maluk, Marcelo Bueno Batista, Manuel Carmona, Madan Junghare, Helisson Faoro, Leonardo M. Cruz, Federico Battistoni, Emanuel de Souza, Fabio O. Pedrosa, Wen-Ming Chen, Philip S Poole, Ray A Dixon, Euan K James. Comparative Genomics Provides Insights into the Taxonomy of *Azoarcus* and Reveals Separate Origins of *nif* Genes in the Proposed *Azoarcus* and *Aromatoleum* Genera. *Genes*, v.12, n.1, 2021. <https://doi.org/10.3390/genes12010071>©.

†These authors contributed equally.

ABSTRACT

Among other attributes, the Betaproteobacterial genus *Azoarcus* has biotechnological importance for plant growth-promotion and remediation of petroleum waste-polluted water and soils. It comprises at least two phylogenetically distinct groups. The “plant-associated” group includes strains that are isolated from the rhizosphere or root interior of the C4 plant Kallar Grass, but also strains from soil and/or water; all are considered to be obligate aerobes and all are diazotrophic. The other group (now partly incorporated into the new genus *Aromatoleum*) comprises a diverse range of species and strains that live in water or soil that is contaminated with petroleum and/or aromatic compounds; all are facultative or obligate anaerobes. Some are diazotrophs. A comparative genome analysis of 32 genomes from 30 *Azoarcus*-*Aromatoleum* strains was performed in order to delineate generic boundaries more precisely than the single gene, 16S rRNA, that has been commonly used in bacterial taxonomy. The origin of diazotrophy in *Azoarcus*-*Aromatoleum* was also investigated by comparing full-length sequences of *nif* genes, and by physiological measurements of nitrogenase activity using the acetylene reduction assay. Based on average nucleotide identity (ANI) and whole genome analyses, three major groups could be discerned: (i) *Azoarcus* comprising *Az. communis*, *Az. indigens* and *Az. olearius*, and two unnamed species complexes, (ii) *Aromatoleum* Group 1 comprising *Ar. anaerobium*, *Ar. aromaticum*, *Ar. bremense*, and *Ar. buckelii*, and (iii) *Aromatoleum*

Group 2 comprising *Ar. diolicum*, *Ar. evansii*, *Ar. petrolei*, *Ar. toluclasticum*, *Ar. tolulyticum*, *Ar. toluolicum*, and *Ar. toluvorans*. Single strain lineages such as *Azoarcus* sp. KH32C, *Az. pumilus*, and *Az. taiwanensis* were also revealed. Full length sequences of *nif*-cluster genes revealed two groups of diazotrophs in *Azoarcus-Aromatoleum* with *nif* being derived from *Dechloromonas* in *Azoarcus* sensu stricto (and two *Thauera* strains) and from *Azospira* in *Aromatoleum* Group 2. Diazotrophy was confirmed in several strains, and for the first time in *Az. communis* LMG5514, *Azoarcus* sp. TTM-91 and *Ar. toluolicum* T^T. In terms of ecology, with the exception of a few plant-associated strains in *Azoarcus* (s.s.), across the group, most strains/species are found in soil and water (often contaminated with petroleum or related aromatic compounds), sewage sludge, and seawater. The possession of *nar*, *nap*, *nir*, *nor*, and *nos* genes by most *Azoarcus-Aromatoleum* strains suggests that they have the potential to derive energy through anaerobic nitrate respiration, so this ability cannot be usefully used as a phenotypic marker to distinguish genera. However, the possession of *bzd* genes indicating the ability to degrade benzoate anaerobically plus the type of diazotrophy (aerobic vs. anaerobic) could, after confirmation of their functionality, be considered as distinguishing phenotypes in any new generic delineations. The taxonomy of the *Azoarcus-Aromatoleum* group should be revisited; retaining the generic name *Azoarcus* for its entirety, or creating additional genera are both possible outcomes.

1 INTRODUCTION

The genus *Azoarcus* (*Az.*) was first discovered after several strains were isolated from surface sterilized roots of the halo-tolerant C4 *Poaceae* species *Leptochloa fusca* L. (Kallar Grass) in Pakistan (REINHOLD *et al.*, 1986; REINHOLD *et al.*, 1987). The new species *Az. communis* and *Az. indigens* were subsequently described (REINHOLD *et al.*, 1993), and the modelendophytic *Azoarcus* strain, BH72, which was also isolated from Pakistani Kallar Grass, was recently incorporated into the new species, *Az. olearius* (CHEN *et al.*, 2013) on the basis of its high genomic similarity to the type strain, DQS-4^T (which was isolated from oil-contaminated soil (FAORO *et al.*, 2017). Subsequent to the discovery of the plant-associated *Azoarcus* spp., several new species were isolated from non-plant sources, particularly from soil and/or water contaminated with aromatic compounds such as *Az. evansii*, *Az. taiwanensis*, *Az. tolulyticus*, *Az. toluvorans*, *Az. toluclasticus*, and *Az. buckelii* (ANDERS *et al.*, 1995; ZHOU *et al.*, 1995; SPRINGER *et al.*, 1998; SONG *et al.*, 1999; MECHINI *et al.*, 2002; LEE *et al.*, 2014) as well as several strains that have not yet been classified/described to species level (RABUS *et al.*, 2005; NISHIZAWA *et al.*, 2012; FERNANDÉZ-LLAMOSAS *et al.*, 2014; JUNGHARE *et al.*, 2015; DENG *et al.*, 2019). These are taxonomically separate from the plant-associated species (RABUS *et al.*, 2005; LOY *et al.*, 2005; REINHOLD-HUREK *et al.*, 2000; CHEN *et al.*, 2013; FAORO *et al.*, 2017; FERNANDÉZ-LLAMOSAS *et al.*, 2014) due to their ability to metabolize aromatic compounds, many of them under strict anaerobic conditions, and hence may have biotechnological potential as agents for the remediation of oil-polluted soils (HUREK *et al.*, 2003; MARTIN-MOLDES *et al.*, 2015). In addition, unlike the plant-associated species

that are considered to be obligate aerobes, they are facultative anaerobes that can use nitrate as an alternative electron acceptor under anoxic conditions (RABUS *et al.*, 2019).

Both *Azoarcus* groups contain diazotrophs, but diazotrophy is not as common in the non-plant-associated group, and the *nif* genes of the two groups are unrelated (HURK *et al.*, 1997; FAORO *et al.*, 2017; FERNANDÉZ-LLAMOSAS *et al.*, 2014; MARTIN-MOLDES *et al.*, 2015).

So different are the two groups in terms of their molecular phylogeny and their lifestyles that there have been proposals to separate them into two genera (REINHOLD-HUREK *et al.*, 2000; RABUS *et al.*, 2005; MARTIN-MOLDES *et al.*, 2015; FAORO *et al.*, 2017). Accordingly, Rabus *et al.* (2019) have recently re-organised *Azoarcus* and described a new genus, *Aromatoleum*, comprising all species outside the “plant-associated” group including several newly-described species (*Ar. aromaticum*, *Ar. bremense*, *Ar. diolicum*, *Ar. petrolei*, and *Ar. toluolicum*). *Azoarcus* now consists only of the species *Az. communis*, *Az. indigenus*, and *Az. olearius*. However, the molecular taxonomy in the Rabus *et al.* (2019) study was based solely on 16S rRNA sequences and did not include several *Azoarcus* strains that were either not currently allocated to a species, or have been done so invalidly (e.g., *Az. taiwanensis*). Shortly after publication of Rabus *et al.* (2019), one of these strains, SY39, which was isolated from seawater, was described as the type strain of a new species, *Az. pumilus*, that clusters loosely with *Az. taiwanensis*, and which is intermediate between *Azoarcus* and *Aromatoleum* (FU *et al.*, 2019), thus increasing uncertainty about strict taxonomic divisions in *Azoarcus sensu lato*. Rapid technical advances combined with modern computational solutions have greatly improved the speed and quality of the analysis and sequencing of bacterial genomes, concomitantly with a huge reduction in costs. Such genome sequences can be used to more accurately discern the phylogeny/taxonomy of groups of bacteria as the full sequences of several core genes can be used to construct robust phylogenies. Recent examples are *Burkholderia-Paraburkholderia-Caballeronia* (BEUKES *et al.*, 2017), *Trinickia* and *Mycetohabitans* (ESTRADA-DE LOS SANTOS *et al.*, 2018), and *Bradyrhizobium* (AVONTUUR *et al.*, 2019; ORMEÑO-ORILLO, MARTÍNEZ-ROMERO, 2019).

The aim of the present study was to construct a robust phylogeny, utilising whole genome sequences of all described *Azoarcus* and *Aromatoleum* species together with genomes of strains that have not yet been allocated species status. A further aim was to use the genomes to investigate the lifestyle and ecology of the strains, with particular emphasis on the origin of the different types of *nif* genes in *Azoarcus-Aromatoleum*.

2 MATERIALS AND METHODS

2.1 Bacterial Strains and Genomes

Fourteen *Azoarcus-Aromatoleum* genomes were sequenced as part of this study (TABLE S1.4a,b). These, together with previously published genome sequences and with others available in the National Center for Biotechnology Information (NCBI) database, totalling 32 *AzoarcusAromatoleum* genomes (TABLE 1.4) allowed robust reanalysis of the *Azoarcus-Aromatoleum* group and its separation into subgroups by different methods such as average nucleotide identity (ANI) and core genome analysis. A further 35 genomes from related genera (TABLE S1.4a) were also included in the analysis in order to place the *Azoarcus-Aromatoleum* group into a wider taxonomic/phylogenetic context.

For the purpose of genome sequencing, the following strains were obtained from various culture collections (in parentheses): *Az. communis* Swub3T (DSM), *Az. communis* LMG 5514 (LMG), *Az. indigenus* VB32T (LMG), *Az. taiwanensis* NSC3 (LMG), *Ar. anaerobium* LuFRes1T (DSM), *Ar. aromaticum* pCyN1 (DSM), *Ar. bremense* PbN1T (DSM), *Ar. buckelii* U120T (DSM), *Ar. diolicum* 22LinT (DSM), *Ar. evansii* KB740T (DSM), *Ar. petrolei* ToN1T (DSM), *Ar. toluolicum* T T (DSM), and *Ar. toluvorans* Td21T (DSM). In addition, the *Azoarcus* sp. strain TTM-91 was described for the first time, and its genome was also sequenced as part of this study. Strain TTM-91 was isolated on R2A agar from uncontaminated water sampled from the Caohu River (24°04'024" N, 120°46'004" E) in the vicinity of Taichung City, Taiwan.

The genomes of the 14 strains in Table S1b were sequenced by MicrobesNG (UK) with the genomic DNA library prepared using the Nextera XT Library Prep Kit (Illumina, San Diego, CA, USA) following the manufacturer's protocol with the following modifications: 2 ng of DNA was used as input, and polymerase chain reaction (PCR) elongation time was increased to 1 min. DNA quantification and library preparation were carried out on a Microlab STAR automated liquid handling system (Hamilton, Birmingham, UK). Pooled libraries were quantified using the Kapa Biosystems Library Quantification Kit for Illumina on a Roche light cycler 96-qPCR machine (Roche, Basel, Switzerland). Libraries were sequenced on the Illumina HiSeq (Illumina) using a 250 bp paired end protocol. Read quality was verified with FastQC v0.11.5 (ANDREWS, 2021) and reads were adapter trimmed and quality filtered using Trimmomatic v0.39 (<http://www.usadellab.org/cms/index.php?page=trimmomatic>) with a sliding window quality cut-off of Q15 (BOLGER *et al.*, 2014). Possible contaminations were removed with the bbsplit.sh script from BBMAP v37.87 (BUSHNEEL, 2020) by removal of reads mapping to the contaminant's genomes (e.g., human genome). De novo genome assembly was carried out with SPAdes v3.11.1 (NURK *et al.*, 2013) and genes were annotated using Prokka v1.11 (SEEMANN, 2014). The assembly (contig) sizes were determined with QUAST v5.0.2 (GUREVITCH *et al.*, 2013).

TABLE 1.4. Strains used in phylogenies and evidence for diazotrophy

Strain Genome	Genome Size [bp]	Genome GC Content [%]	Nif Genes	Nitrogenase Activity Confirmed by Acetylene Reduction Assay (ARA)	Isolation Site and/or Type of		Country of Origin
					Habitat		
<i>Aromatoleum anaerobium</i> LuFRes1T	4,482,959	66.72	No	No	Sewage sludge		Germany
<i>Aromatoleum aromaticum</i> EbN1T	4,727,255	64.68	No	No	River water		Germany
<i>Aromatoleum aromaticum</i> pCyN1	4,208,498	65.23	No	No	River water		Germany
<i>Aromatoleum bremense</i> PbN1T	4,319,217	65.57	No	No	River water		Germany
<i>Aromatoleum buckelii</i> U120T	4,227,546	65.33	No	No	Oxic soil		Germany
<i>Aromatoleum diolicum</i> 22Lin2T	5,138,050	64.22	Yes	Yes	River water		Germany
<i>Aromatoleum evansii</i> KB740T	5,883,543	65.98	Yes	Yes	Creek sediment		USA
<i>Aromatoleum petrolei</i> Ton1T	5,364,035	65.73	No	No	River water		Germany
<i>Aromatoleum</i> sp. CIB	5,257,030	65.84	Yes	Yes	Diesel fuel-contaminated aquifer		Switzerland
<i>Aromatoleum</i> sp. DN11	4,956,835	66.30	Yes	Yes	Gasoline-contaminated groundwater		Japan
<i>Aromatoleum</i> sp. PA01	3,908,240	66.10	No	No	Sewage sludge, wastewater treatment plant		Germany
<i>Aromatoleum toluclasticum</i> MF63T	5,925,983	66.00	No	No	Shallow aquifer sediment		USA
<i>Aromatoleum tolulyticum</i> Tol-4T	5,110,391	66.43	Yes	Yes	Petroleum contaminated aquifer sediments		USA
<i>Aromatoleum toluolicum</i> T	6,025,652	65.93	Yes	Yes	Petroleum contaminated aquifer		Switzerland
<i>Aromatoleum toluvorans</i> Td21T	4,786,256	66.66	Yes	Yes	Muck soil		USA
<i>Azoarcus communis</i> LMG5514	4,996,403	62.45	Yes	Yes	Oily sludge, storage tank, petroleum refinery		France
<i>Azoarcus communis</i> SWub3T_1	4,976,326	62.36	Yes	Yes	Kallar Grass roots		Pakistan
<i>Azoarcus communis</i> SWub3T_2	5,004,685	62.50	Yes	Yes	Kallar Grass roots		Pakistan
<i>Azoarcus</i> sp. TSNA42	4,886,934	62.80	Yes	nt	Oil-contaminated sediment		South Korea
<i>Azoarcus</i> sp. TSPY31	4,572,081	63.20	Yes	nt	Oil-contaminated sediment		South Korea
							(Continue)

(Continue)

<i>Azoarcus</i> sp. TTM-91	5,393,782	67.70	Yes	Yes	River water	Taiwan
<i>Azoarcus indigenes</i> VB32T	5,562,509	67.31	Yes	Yes	Kallar Grass roots	Pakistan
<i>Azoarcus indigenes</i> VB32T	5,464,470	67.60	Yes	Yes	Kallar Grass roots	Pakistan
<i>Azoarcus olearius</i> BH72	4,376,040	67.90	Yes	Yes	Kallar Grass roots	Pakistan
<i>Azoarcus olearius</i> DQS-4T	4,451,751	67.83	Yes	Yes	Oil-contaminated soil	Taiwan
<i>Azoarcus pumilus</i> SY39T	3,225,512	66.53	No	No	Sea water	China
<i>Azoarcus</i> sp. BM101	4,904,245	62.90	Yes	nt	Estuary sediment	USA
<i>Azoarcus</i> sp. CC-YHH838	4,723,750	67.50	Yes	Yes	Termite nest	Taiwan
<i>Azoarcus</i> sp. CC-YHH848	4,602,270	68.30	Yes	Yes	Rhizosphere of <i>Ficus religiosa</i>	Taiwan
<i>Azoarcus</i> sp. DD4	5,400,077	66.70	Yes	nt	Activated sludge, wastewater treatment plant	USA
<i>Azoarcus</i> sp. KH32C	5,818,755	65.03	Yes	Yes	Soil from field with paddy rice and soybean crops	Japan
<i>Azoarcus taiwanensis</i> NSC3T	4,228,584	62.88	No	No	Hot spring	Taiwan

*nt: not tested.

REFERENCE: Adapted from Raittz *et al.*, (2021)

2.2 NCBI Sequence Datasets

All 67 (*Azoarcus-Aromatoleum* and related genera) genomes and nucleotide and amino acid coding sequences were downloaded from the NCBI GenBank in FASTA and GenBank formats and are presented in TABLE S1.4a.

2.3 Average Nucleotide Identity (ANI)

In order to relate our newly generated genomic resources to the pre-existing *Azoarcus-Aromatoleum* resources in the NCBI database, we applied whole genome comparisons by average nucleotide identity (ANI) using the alignment-free tool FastANI (version 1.31) (JAIN *et al.*, 2018). We conducted two independent analyses: first, the pairwise calculation of the ANI values was performed with the 32 genomes (see full-article); second, these calculations were performed with 52 genomes after the inclusion of 20 *Thauera* genomes (available in full-article). As a tool input requirement, the nucleotide coding sequences (CDS) were concatenated into a single multiFASTA. We performed the ANI pairwise values calculation for genomes using the FastANI default parameters. The FastANI output files (with a search summary and a lower triangular matrix containing the identity values in phylip format) were used to determine dissimilarity matrices and to construct a hierarchical neighbour-joining (NJ) tree cluster to reorder the lines and columns of the ANI matrix and construct a heat map. These analyses were performed using a Linux operating system (Ubuntu 20.4.1 LTS) and MATLAB® support programming.

2.4 Protein Clusters

The complete set of proteins translated from the CDS of each of the 67 genomes were joined in a FASTA file and clustered using the RAFTS3G tool (NICHIO *et al.*, 2019) (available at: <https://sourceforge.net/projects/rafts-g/>) applying a minimal self-score of 0.7 as a threshold parameter for a specific protein to be included in a cluster. Two groupings were performed separately, RAFTS3G-32 for the *Azoarcus-Aromatoleum* group and RAFTS3G-67 for the 67 genomes. For the *Azoarcus-Aromatoleum* group, from an input multiFASTA of 145,695 protein sequences, RAFTS3G generated 13,959 clusters with two or more proteins (89.68% of the sequences) and 15,043 clusters with only one protein (10.32% of the sequences). For the 67 genomes, from a multiFASTA with 271,578 protein sequences, RAFTS3G generated 19,666 clusters with two or more proteins (89.39% of the sequences) and 28,811 clusters with only one protein (10.61% of the sequences).

The protein clusters were the basis for several steps in the subsequent phylogenetic/taxonomic analyses including core genome determination, *nif* gene cluster analysis, partial SWeeP genome representations, and specific common genes analyses.

2.5 Core Genome Analysis Based on Core Protein Groups

The core genome in the *Azoarcus-Aromatoleum* group was extracted and 1044 gene clusters were found in common in all 32 genomes (25.08% of the sequences). The analysis of core genome for all 67 genomes revealed only 231 common gene clusters (6.91% of the sequences).

2.6 Analysis of *nif* and Other Functional Genes

Functional analyses of the *nif* clusters and *nifH* gene were performed for all 67 genomes. Analyses of the denitrification pathway and of aromatic compound degradation were performed only for the *Azoarcus-Aromatoleum* group.

The RAFTS3G-67 clusters were used to extract *nif* genes and genes for denitrification and benzoate metabolism. In this study, *nif* cluster analyses were performed using a manually curated “in-house” database containing 1,546 genes present in the *nif* gene clusters of 144 diazotrophic organisms (available in full-article). This is an expanded version based on the study by Dos Santos *et al.* (2012). After database searches and manual validation, all *nifH* genes were found in the same RAFTS3G-67 grouping.

Analysis of functional genes was performed for the nitrate reduction and the aromatic compound degradation pathways. For nitrate reduction pathway analysis, we searched the RAFTS3G-32 clusters to identify the presence of genes encoding nitrate reductase (*nar/nap*), nitrite reductase (*nir*), nitric oxide reductase (*nor*), and nitrous oxide reductase (*nos*) using *Aromatoleum* sp. ClB corresponding marker genes (RAFTS3) (VIALLE *et al.*, 2016; NICHIO *et al.*, 2019) and curated them manually. Aromatic compound degradation pathway genes including those for the anaerobic benzoate degradation pathway gene cluster (*bzd*), aerobic benzoate degradation pathway gene cluster (*box*), and genes from the “lower pathway” (*LP*) as marker genes were similarly searched in the RAFTS3G-32 groups.

2.7 SWeeP Phylogenies

Phylogenetic analyses were conducted using SWeeP (DE PIERRI *et al.*, 2020) for protein set representations (available at: <https://sourceforge.net/projects/spacedwordsprojection/>). The purpose of this method is to transform any set of amino acid sequences into a single vector that can, for example, represent all proteins in an organism. In practice, the input file to generate the SWeeP vectors consists of a multiFASTA file containing amino acid sequences, in which the protein sequences in a particular organism are concatenated, each protein flanked by delimiters, to get a single sequence (DE PIERRI *et al.*, 2020).

SWeeP vectors based on the complete set of proteins for each of the 32 *Azoarcus-Aromatoleum* genomes and for the wider group of 67 genomes were obtained to represent complete genomes. The parameters used were for spaced words, the “11011” mask, and for the projection into a size of 1369 for

visualization (37×37 , a perfect square number, is desired for some visualization tasks). These parameters were the same as those used for complete bacterial genomes in the study by De Pierri *et al.* (2020) (see FIGURE S1.4).

The NJ model based on Euclidean distance matrices of SWeeP vectors was used for four analyses: phylogenetic analyses of complete genomes, phylogenetic trees and heat maps for the analysis of the core genome, *nif* cluster, and *nifH* gene analyses. All phylogenetic trees were built using standard MATLAB functions and scripts (for an excerpt of script, see FILE S1.4), visualized with Dendroscope (version 3.7.2) (HUSON *et al.*, 2012) and arranged with Adobe Illustrator CC 2017 (version 21.0.0).

2.8 Heat Maps

The heatmap graphs were constructed using “in-house” MATLAB® scripts and arranged using Adobe Illustrator CC 2017 (version 21.0.0). For the FastANI heatmap, the pairwise ANI values matrix was normalized by the mean values: First, ANI values were converted into dissimilarity matrices, then, the lower and upper matrices were divided by their corresponding mean values. It was decided to normalize the mean to guarantee a better correspondence between the matrices in the visualization.

The results of FastANI were used to determine dissimilarities matrices to construct a hierarchical NJ tree cluster in which the order of lines and columns were established according to these clusters.

SWeeP vector heat maps were generated from the distance matrices normalized by the average values, and the order of the rows and columns were determined by corresponding NJ-based dendrograms (see FILE S1.4). This technique allowed us to compare different distance/dissimilarity matrices in the same heat map.

2.9 16S rRNA Analysis

For the 16S rRNA-based phylogeny, the genes were extracted automatically from 54 genomes from the NCBI GenBank using an “in-house” MATLAB® script and are listed in the full-article. From the SILVA database (QUAST *et al.*, 2012), 13 16S rRNA genes from *Dechloromonas* sp. strains H13, Dech2017 and CZR5, *D. hortensis* MA-1, *D. agitata* is5, *Thauera* sp. UPWRP, *T. butanivorans* NBRC 103042, *T. linaloolentis* 47Lol_1 and 47Lol_2, *T. hydrothermalis* GD-2, *T. phenolivorans* ZV1C, *Rhodocyclus tenuis* DSM109T, and *Rubrivivax gelatinosus* CBS were collected. The 16S rRNA gene phylogenetic analyses were performed using Clustal Omega (SIEVERS *et al.*, 2011) with default parameters. The phylogenetic trees were arranged with Adobe Illustrator CC 2017 (version 21.0.0).

2.10 Nitrogenase Activity

The acetylene reduction assay was used to determine nitrogenase activity on free-living cultures growing on semi-solid or liquid media () Growth conditions for all strains are detailed in the Supplementary Materials (File S2.4). Ethylene (C₂H₄) formation was determined after incubation of the cultures in acetylene by using a Perkin Elmer Clarus 480 gas chromatograph equipped with a HayeSep® N (80–100 MESH) column (Merck Life Science Ltd., Dorset, UK). The injector and oven temperatures were kept at 100 °C, while the flame ionisation detector was set at 150 °C. The carrier gas (nitrogen) flow was set at 8–10 mL min⁻¹. Nitrogenase activity is reported as nmol of C₂H₄ min⁻¹ mg protein⁻¹. The ethylene calibration curve was prepared from the chemical decomposition of ethephon (Merck Life Science Ltd., Dorset, UK) as described by Zhang and Wen (2012). Whole cell protein concentration was determined by the Bradford method (BRADFORD, 1976) after lysis in 0.1 mM NaOH.

3 RESULTS AND DISCUSSION

3.1 Whole-Genome Sequences and Identity of *Azoarcus* sp. Strain TTM-91

Features about the genomes of the 14 *Azoarcus-Aromatoleum* strains that were sequenced for this study are presented in TABLE 1.4 and TABLE S1.4a,b.. The genomes of an additional 18 strains were obtained from the NCBI, and these are listed in TABLE 1.4 (their original sources are listed in the full-article). *Azoarcus* sp.TTM-91 was the only strain that was isolated specifically for this study. The 16S rRNA sequence of *Azoarcus* sp. strain TTM-91 suggests that it is closely-related to *Az. indigenus* (>99% similarity), and it also has a very similar genome size (5,393,782 bp) and G + C content (67.70%) to *Az. indigenus* VB32^T. The genomes of the remaining strains sequenced for this study were relatively standard in their size (ranging from 4,227,546 for *Ar. buckelii* to 6,025,652 for *Ar. toluolicum*) and G + C content (ranging from 62.36% for *Az. communis* Swub3^T to 67.70% for *Azoarcus* sp. TTM-91) in the wider context of the *Azoarcus-Aromatoleum* group (Table 1.4 and Table S1.4b).

3.2 Average Nucleotide Identity (ANI) and Core Genome Analysis

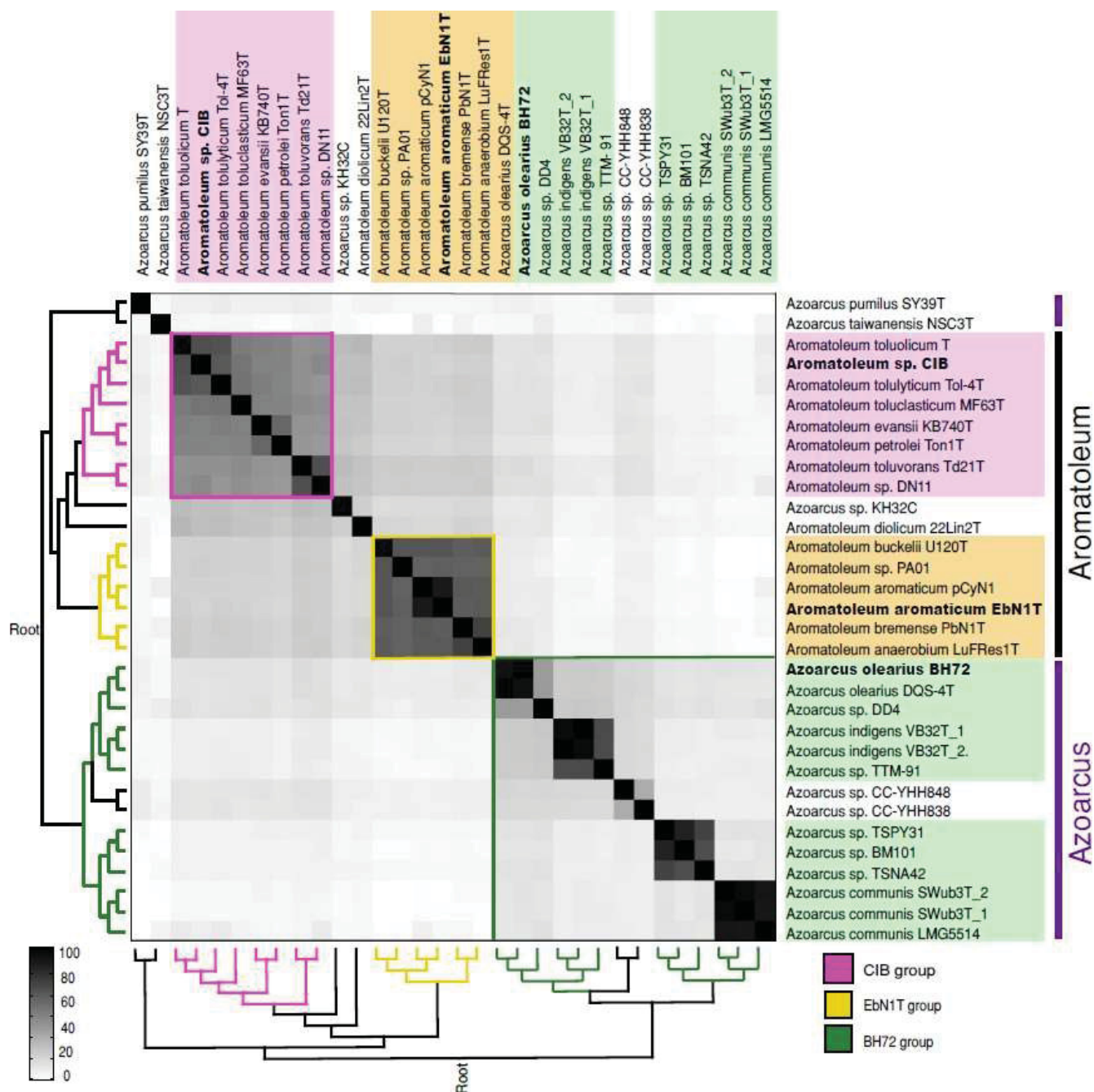
A matrix comparing the ANI of all 30 strains comprising 32 genomes (as two sequences were available for *Az. communis* Swub3^T and *Az. indigenus* VB32^T) is presented in Figure 1.4, and the ANI percentage values are given in Table S2.4a,b. The genomes of the 30 *Azoarcus-Aromatoleum* strains are broadly divided into the three main groups representing the genera *Azoarcus* and two large clades in *Aromatoleum*, but other groupings are also discernible. *Aromatoleum* contains two clearly separate groups: Group 1 (“EbN1 group”) comprising the species *Ar. anaerobium*, *Ar. aromaticum*, *Ar. buckelii*, and *Ar. bremense*, together with the strain PAO1; and Group 2 (“CIB group”) comprising the species *Ar. diolicum*, *Ar. evansii*, *Ar. petrolei*, *Ar. toluclasticum*, *Ar. tolulyticum*, *Ar. toluolicum*, and *Ar. toluvorans*, together with

the “*Azoarcus*” strains CIB and DN11 (MARTIN-MOLDES *et al.*, 2015; DEVANADERA *et al.*, 2019) *Aromatoleum diolicum* sits just outside Group 2, and the strain KH32C appears not to belong to either group, but is slightly closer to the *Aromatoleum* Group 2 than to Group 1. In the ANI matrix, the genus *Azoarcus* is divided into “species complexes” represented by the strains of the three formally described species, *Az. communis*, *Az. indigena*, and *Az. olearius*, plus a fourth, undescribed species related to *Az. communis* comprising two strains from South Korea (TSPY31, TSPN42) and the metagenome-derived strain BM101 from the USA. Outside these four well-defined, and related *Azoarcus* species complexes, are two Taiwanese strains “*Az. nasutitermitis*” CC-YHH838 and “*Az. rhizosphaerae*” CC-YHH848 (LIN *et al.*, 2020), and the type strains of the single strain species *Az. pumilus* and “*Az. taiwanensis*”, which are distantly related to each other.

A two principal components projection for k-means clusters ($k = 4$) generated with SWeeP vectors demonstrated that the wider group of 67 genome-sequenced strains was clustered into four groups (Figure S2.4a), and these were linked with a core genome phylogeny (FIGURE S2.4b). This Analysis allowed us to take a wider look at the *Azoarcus-Aromatoleum* group, particularly in the context of its neighbouring genus *Thauera*. It suggests that *Thauera* (or at least those 20 strains within it whose genomes have been fully sequenced) comprises two groups: a larger one in k3 consisting of 18 of the sequenced *Thauera* strains, and a much smaller subgroup containing only *T. hydrothermalis* GD-2T and *Thauera* sp. D20 that is placed in k2 (which also contains most of the *Azoarcus* strains). Both *Thauera* groups were more closely related to *Azoarcus* than to *Aromatoleum*, with *T. hydrothermalis* GD-2T and *Thauera* sp. D20 obviously nested within *Azoarcus*.

A phylogenetic analysis was performed based on the whole genome and the core genome (Figure 2.4A,B). Three major groups can be discerned comprising (1) *Azoarcus* consisting of 14 strains including all currently accepted *Azoarcus* species plus the informally named *Az. nasutitermitis* (CC-YHH838), *Az. rhizosphaerae* (CC-YHH848) and *Az. taiwanensis*; (2) The “EbN1-Group” of six *Aromatoleum* strains including *Ar. aromaticum* EbN1^T; and (3) the “CIB-Group” of 10 *Aromatoleum* strains plus “*Azoarcus*” sp. KH32C. Three small sub-groups are apparent in *Azoarcus*; the one closest to *Azoarcus sensu stricto* (s.s.) consists of two Taiwanese strains (CC-YHH838, CC-YHH848), and the other groups, which are more distant from *Azoarcus* (s.s.), consist of *Az. pumilus* and *Az. taiwanensis* from China and Taiwan, respectively. The EbN1-Group in *Aromatoleum* is monophyletic and could thus be termed *Aromatoleum sensu stricto* (s.s.), but the CIB-Group is looser and could be considered a polytomy (FIGURE 2.4A).

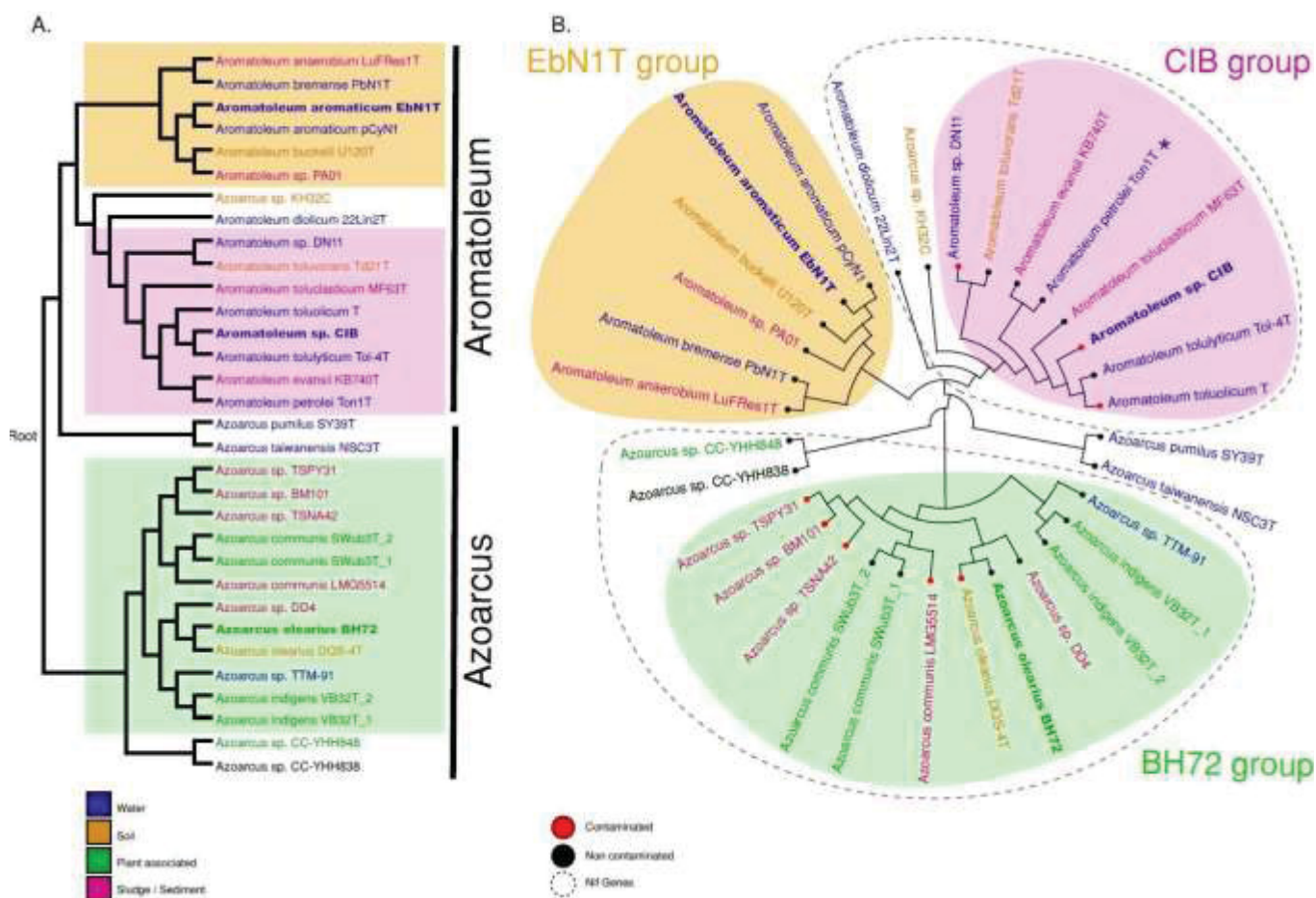
FIGURE 1.4. Average nucleotide identity (ANI) of whole genome comparison of 32 *Azoarcus* and *Aromatoleum* genomes



REFERENCE: Raittz *et al.*, (2021).

Legend: Average nucleotide identity (ANI) of whole genome comparison of 32 (18 database and 14 newly sequenced) *Azoarcus* and *Aromatoleum* genomes from 30 strains showing subgroups of the known genera *Azoarcus* and *Aromatoleum*. The depicted heatmap is a matrix generated by FastANI, ordered according to the dendrograms (left/below) obtained with the neighbour joining clustering method in MATLAB, considering upper and lower triangular complementary ANI dissimilarity matrices as input.

FIGURE 2.4. Phylogenetic trees of whole genomes and clustering of core genomes



REFERENCE: Raittz *et al.*, (2021).

Legend: Phylogenetic trees of whole genomes (A) and clustering of core genomes (B). These analyses confirm the grouping of 30 *Azoarcus* and *Aromatoleum* strains (represented by 32 genomes) into three major groups named the EBN1, CIB, and BH72 groups. Both trees were constructed based on distance matrices of SWeeP vectors. The colours in the names indicate the source of the initial isolation of the corresponding strain: water (blue), soil (brown), sludge/sediment (pink) or plant-associated (green). Whether the substrate was contaminated (red) by organic solvents or non-contaminated (black) is indicated by the tip-point colour in the circular tree. The presence of *nif* genes in a group is indicated with dashed outlines. Asterisk (*) indicates the absence of *nif* genes in *Ar. petrolei* Ton1^T.

In terms of ecology/habitat, no obvious pattern could be discerned in the genotaxonomy of *Azoarcus-Aromatoleum* (Figure 2.4B) (i.e., the 30 taxa come from a wide range of habitats, and no habitat is particular to any of the three major groups). These include water from widely different sources (oceans, rivers, hot springs, aquifers, and wastewater treatment plants) as the primary habitat for the genome-sequenced *Azoarcus-Aromatoleum* strains, comprising 13 of the strains, sludge/sediment as the secondary habitat with eight strains, and soil as the tertiary habitat with four strains. Of the remaining five strains, all of which are in *Azoarcus*, one came from a termite nest in Taiwan (*Az. nasutitermitis* CC-YHH838), one from the rhizosphere of *Ficus religiosa* in Taiwan (*Az. rhizosphaerae* CC-YHH848), and three from roots of

Kallar Grass including the model endophyte *Az. olearius* BH72, and the type strains of *Az. communis* (SwuB3T) and *Az. indigens* (VB32T). The genome-sequenced strains were also scored for habitats that were indicated as being contaminated by oil/petroleum; this showed that all the major groups in *Azoarcus-Aromatoleum* contained strains (seven of the genome-sequenced strains) that had an association with contaminated water or soils, but this association is by no means ubiquitous (Figure 2.4B).

In summary, although it should be stressed that the 30 strains presented in Figure 2B are only a “snapshot” of the habitats of *Azoarcus-Aromatoleum* strains, and their selection for our analysis is inevitably biased by the requirement for fully genome-sequenced strains, it is, nevertheless, clear from the present study and from perusing the databases for deposited gene sequences that water, sludge, and soil (contaminated and uncontaminated) are the principal habitats of all three *Azoarcus-Aromatoleum* groups. Plants are not a common habitat; only three strains in our analysis were actually isolated from plants (i.e., Kallar Grass), and all of these were from a single field in Pakistan (RENHOLD *et al.*, 1986; RENHOLD-HURECK, 1993). Moreover, in each case, these plant-associated strains (*Az. olearius* BH72, *Az. communis* SwuB3T, and *Az. indigens* VB32T) have sister strains in the same (or closely-related) species that were not isolated from plants such as *Az. olearius* DQS-4T (oil-contaminated soil; FAORO *et al.*, 2017), *Az. communis* LMG 5514 (refinery oil sludge; RENHOLD-HURECK, 1993), and *Azoarcus* sp. TTM-91 (river water; this study). Other examples are *Azoarcus* sp. DD4, which is quite closely related to *Az. olearius* and may belong to this species (Figure 2.4a,b); this was isolated from activated sludge from a wastewater treatment plant (DENG *et al.*, 2019). Moreover, the other genome-sequenced strains in the *Az. communis* species complex (BM101, TSPY31, TSNA42), were isolated from sediments. Therefore, even though *Azoarcus* was the only group in the present study that contained plant-associated strains/species, these were few in number, and the most intensively studied among them, *Az. olearius* BH72, did not appear to possess genes that make it any more specialised in terms of associating with plants than its non-plant-associated sister *Az. olearius* DQS4 T (FAORO *et al.*, 2017). Therefore, widely used terms in the literature like “the plant-associated *Azoarcus* genus/group” (see Introduction) should henceforth be avoided as these give a misleading impression about the much wider habitat preferences of *Azoarcus*.

Phylogenetic whole genome analyses including *Thauera* (Figure S2.4a,b) also indicated that *T. hydrothermalis* GD-2T and *Thauera* sp. D20 were actually nested within *Azoarcus*, and are apparently most closely related to the *Azoarcus* strains CC-YHH838 and CC-YHH848. *Thauera hydrothermalis* GD-2T was isolated from the sediment of a hot spring in Tibet (YANG *et al.*, 2018), and is a facultative anaerobe. Nothing has been published as yet about *Thauera* sp. D20, except that it was isolated from a saline lake in Inner Mongolia (ZHENG *et al.*, 2019).

The analyses of whole genomes and of the core genomes of the 30 *Azoarcus-Aromatoleum* strains (Figure 2.4) has closely supported the groupings revealed by the ANI analysis. In addition, the taxonomy revealed by these genome-based analyses does not conflict with that shown by 16S rRNA sequences

extracted from the genomes (Figure S3.4), which in itself is largely concordant with that of Rabus *et al.* (2019), who used fewer (and mainly type) strains to construct their 16S rRNA phylogeny. The main differences between the whole genome analysis (Figures S2.4b and S4.4) and that based on 16S rRNA (Figure S3) is that the former places *T. hydrothermalis* GD-2T and *Thauera* sp. D20 in *Azoarcus* (see above), but the latter places them in *Thauera* together with the *Azoarcus* strains CC-YHH838 and CC-YHH848 (Figure S3). Further studies including a polyphasic analysis should help determine if *T. hydrothermalis* GD-2T and *Thauera* sp. D20 actually belong to *Azoarcus* or if they belong to *Thauera* together with CC-YHH838 and CC-YHH848.

At present, the delineation of bacterial species and, to some extent, genera is still based upon the 16S rRNA gene with suggested cut-offs of <97% similarity between related taxa for species, and a more arbitrary value of 94–95% for genera (TINDALL *et al.*, 2010; QIN *et al.*, 2014; YARZA *et al.*, 2014; BARCO *et al.*, 2020). In addition, for generic delineations, it is recommended that phenotypic and lifestyle attributes be taken into account in addition to genetics. Accordingly, Rabus *et al.* (2019) described the new genus *Aromatoleum* on the basis that they are facultative aerobes that can use nitrate as an alternative electron acceptor under anoxia, and that they have the ability to degrade aromatic compounds under anaerobic conditions. Although they also recognised that 16S rRNA sequences showed that *Aromatoleum* contained two clades corresponding to the Groups 1 (EbN1T) and 2 (CIB) described in the present study, Rabus *et al.* (2019) recommended that in spite of this genetic separation that *Aromatoleum* should not be further divided as suggested by Martin-Moldes *et al.* (2015), as all members of the new genus as delineated by them had the overarching and prominent property of anaerobic biodegradation of aromatic compounds. However, they did not include several strains that are pertinent to this debate such as *Azoarcus* sp. KH32C, *Az. pumilus*, and *Az. taiwanensis*, which fit into neither *Aromatoleum* nor *Azoarcus* as delineated by Rabus *et al.*, (2019). Reinhold-Hurek *et al.* (2000) described three new genera within *Azoarcus* sensu lato as it stood at that time (viz. *Azonexus*, *Azopira*, and *Azovibrio*), at least partly on the basis that their 16S rRNA sequences were only 93–94% similar to *Azoarcus* (s.s.). If 16S rRNA sequences were to be used in this manner to discern genera within the *Azoarcus*-*Aromatoleum* strains in the present study, it would suggest that *Azoarcus*-*Aromatoleum* constitutes a single genus with the exception of *Az. pumilus* and *Az. taiwanensis*, which would be placed into separate genera (Figure S3.4, Table S3.4a,b). However, we are rapidly moving into an era of genome-based taxonomies, and so 16S rRNA sequences are likely to be used less and less frequently as the sole gene sequence-based criterion for describing new taxa.

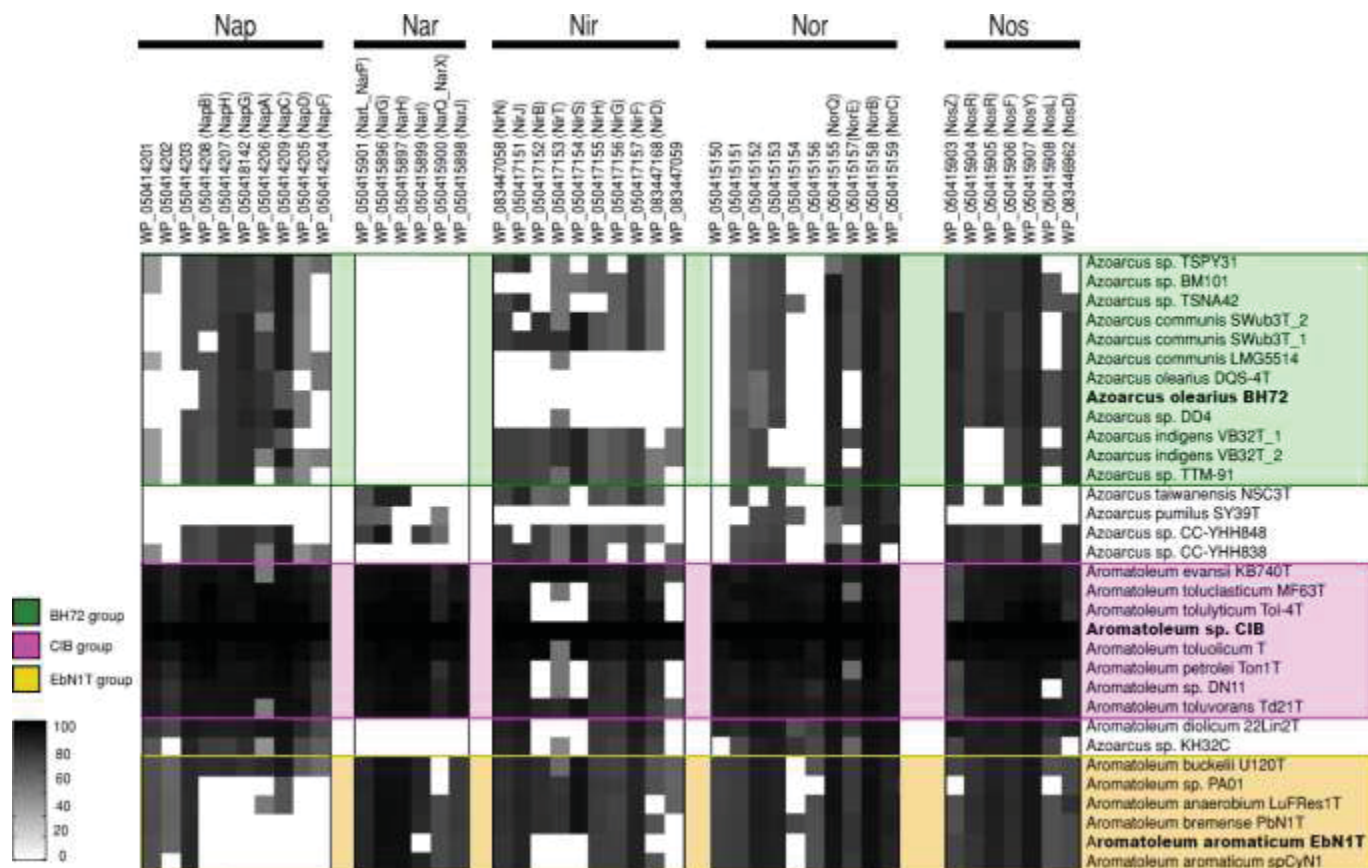
Although it has recently been demonstrated that ANI cannot be used for genus definition (QIN *et al.*, 2014), it can still give indications of where generic boundaries might genuinely lie. In the present study, together with whole genome and core phylogenetics analysis from large numbers of genes, it is suggested that in terms of their core genomes, the two clades in *Aromatoleum* could constitute separate genera, but

also that separate genera for *Az. pumilus* and *Az. taiwanensis* could be created, and possibly another for *Azoarcus* strain KH32C (Figures S2.4b and S4.4).

3.3 Genes for Nitrate Reduction (*Nar/Nap*, *Nir*, *Nor*, *Nos*) and for the Anaerobic (*Bzd*) or Aerobic (*Box*) Degradation of Benzoate

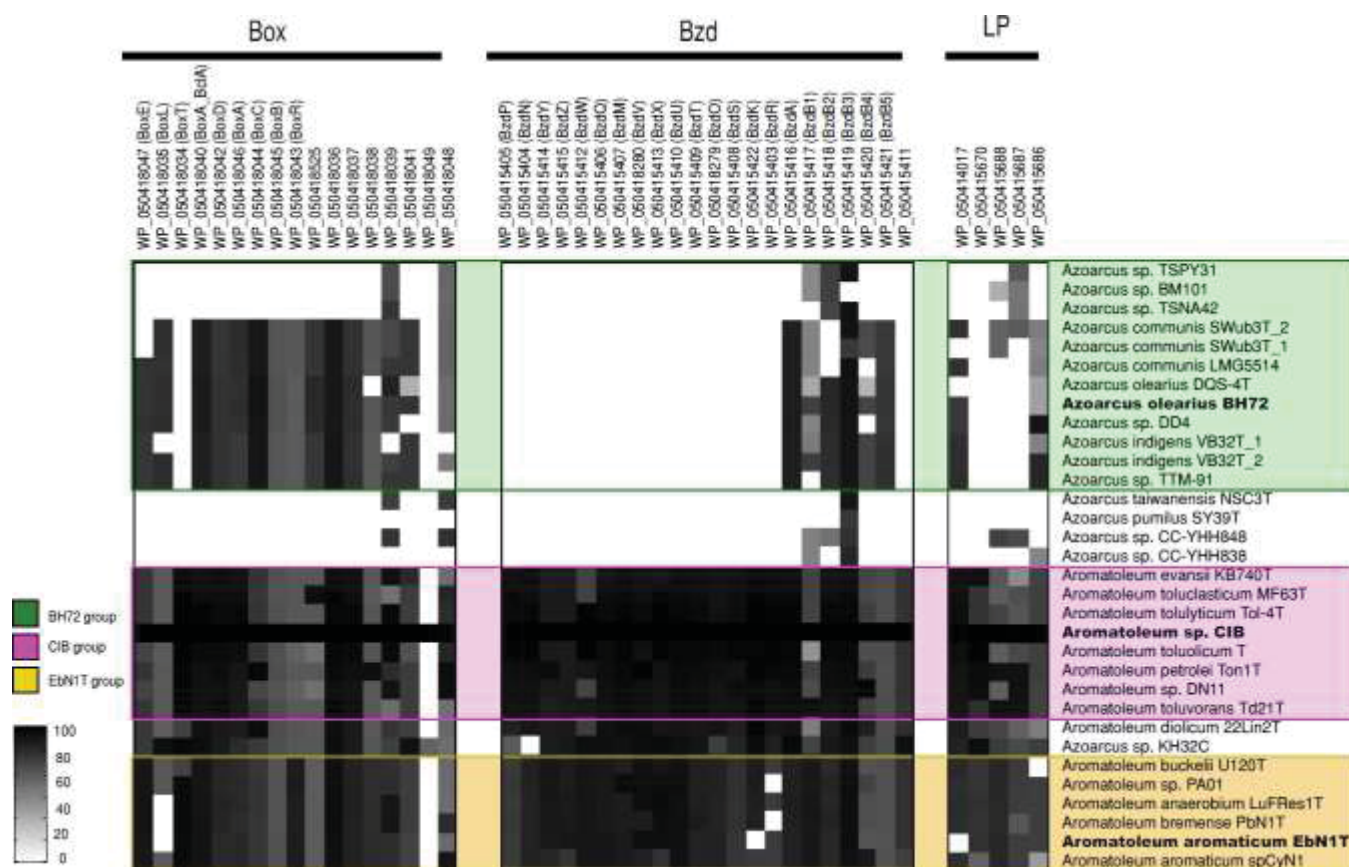
Further insights into the taxonomy of *Azoarcus-Aromatoleum* were gained by comparing sets of genes involved in the use of nitrate as an alternative electron acceptor thus allowing for anaerobic respiration (Figure 3.4) as well as genes involved in the anaerobic and aerobic catabolism of benzoate-*bzd* and *box*, respectively- (Figure 4.4). Using the genome of *Aromatoleum* sp. CIB as a reference, it was revealed that with the exception of *Az. pumilus* SY39T, all the genome-sequenced *Azoarcus* and *Aromatoleum* strains possessed genes for nitrate reduction (*nap* and/or *nar*), but *nar* appeared not to be present in the *Azoarcus* genomes, except for that of the strain CC-YHH848 and that of *Az. taiwanensis*. Moreover, although *Azoarcus* possessed *nap* genes, the *Aromatoleum* strains had a fuller complement of them; interestingly, strains in the CIB group were particularly well endowed with both *nap* and *nar* genes, whereas those in the EbN1 group (except for *Ar. buckelii*) were apparently lacking in key *nap* genes, and hence may be unable to utilize this pathway for nitrate reduction. The only other strains that lacked *nap* genes were *Az. pumilus* SY39T and *Az. taiwanensis*. All the examined genomes contained genes for nitrite reduction (*nir*), except for those of the three *Az. olearius* strains (DQS-4T, BH72, DD4), *Az. communis* LMG5514, and *Az. pumilus* SY39T (FIGURE 3.4). Genes for the final two steps in denitrification, *nor* (nitric oxide reductase) and *nos* (nitrous oxide reductase), were present in all the sequenced *Azoarcus-Aromatoleum* genomes, except that *nos* genes could not be detected in *Az. pumilus* SY39T.

FIGURE 3.4. Nitrate reduction pathway analysis

REFERENCE: Raittz *et al.*, (2021).

Legend: Heatmap revealing the presence/absence of genes from the *nap*, *nar*, *nir*, *nor*, and *nos* gene clusters from 32 genomes (from 30 strains) of the *Azoarcus*-*Aromatoleum* group. The indicated marker genes were identified in the RAFTS3G-32 clusters containing corresponding genes in *Aromatoleum* sp. CIB and verified manually. The homology of the identified genes to reference genes is indicated with a grey scale gradient from white (0%) to black (100%)

FIGURE 4.4. Aromatic compound degradation pathway analysis

REFERENCE: Raittz *et al.*, (2021).

Legend: Heatmap revealing the presence/absence of different genes from the benzoate (bzd) and benzoyl-coenzyme clusters (box) as well as of genes from the "lower pathway" (LP) present in 32 genomes (from 30 strains) of the *Azoarcus-Aromatoleum* group. The indicated marker genes were identified in the RAFTS3G-32clusters containing corresponding genes in *Aromatoleum* sp. CIB and verified manually. The homology of the identified genes to reference genes is indicated with a grey scale gradient from white (0%) to black (100%).

bzd genes were present in all of the *Aromatoleum* strains plus *Azoarcus* sp. KH32C (Figure 4.4), and some were also observed in *Az. olearius*, *Az. indigena*, and *Az. communis*, although it is unlikely that the latter has the capacity to metabolize benzoate under anaerobic conditions since the genes that encode for the key enzyme involved in the benzoyl-CoA reduction (bzdNOPQ) are absent in these strains. On the other hand, all the aforementioned strains also possess box genes, so it is probable that they might degrade benzoate aerobically (VALDERRAMA *et al.*, 2014), but the absence of bzd and box genes in the *Az. communis* strains TSPY31, TSPN42, and BM101 in *Az. pumilus* SY39T and in *Az. taiwanensis* suggests that they cannot metabolize benzoate at all.

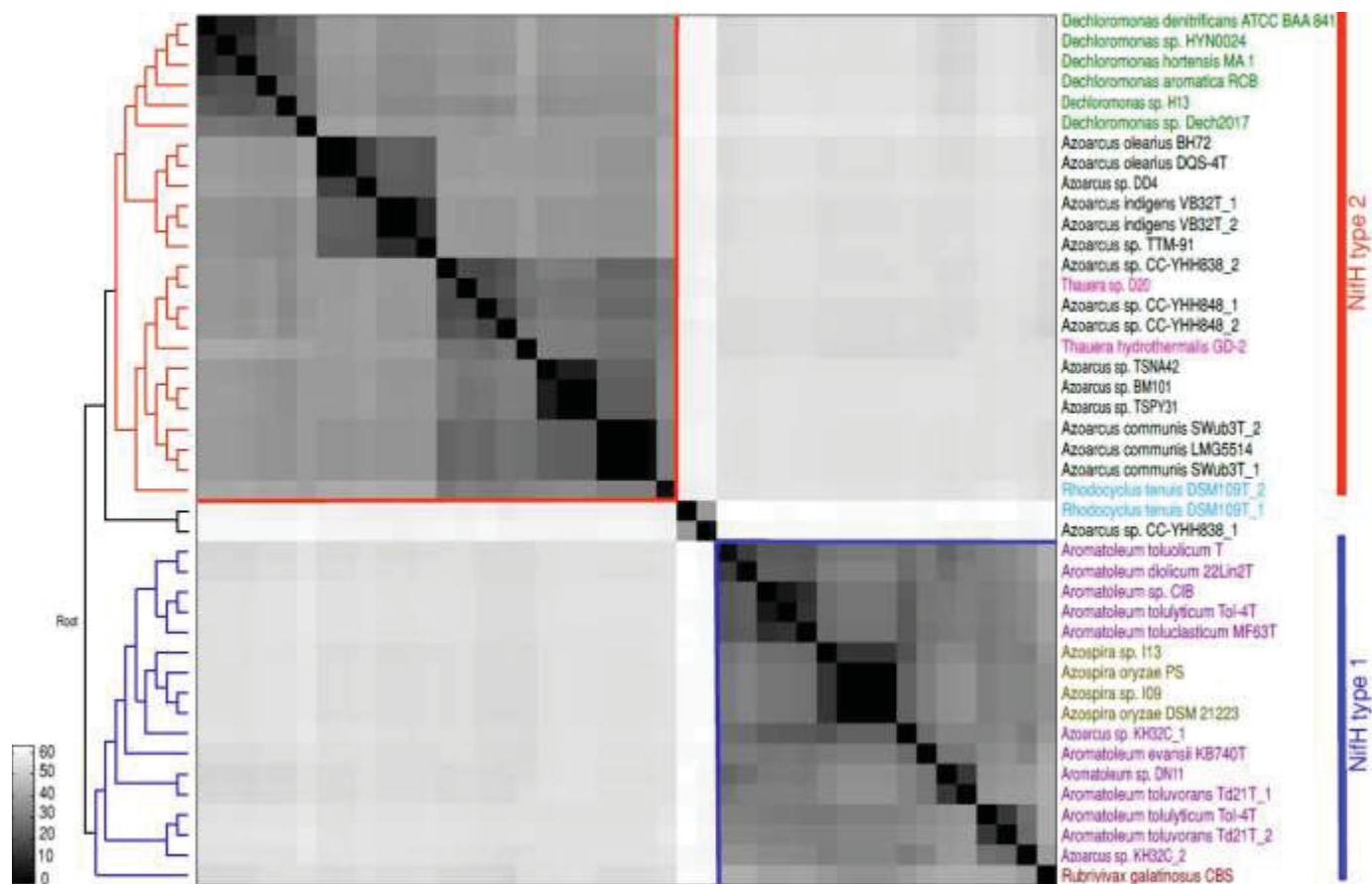
In summary, if the presence of nap and/or nar genes are indicative of a capacity to derive energy from respiratory nitrate reduction (KRAFT *et al.*, 2011) then it is possible that all the *Azoarcus*-

Aromatoleum strains in the present study possess it, with the exception of *Az. pumilus* SY39T. Indeed, Zamarro *et al.* (2017) experimentally demonstrated nitrate reductase-based anaerobic metabolism for a modified variant of *Az. communis* Swub3T with a *bzd* cassette inserted into its genome. In addition, it is also likely that all the strains, except for *Az. olearius*, *Az. communis* LMG5514, and *Az. pumilus* SY39T, have the genetic capacity to perform the complete denitrification of nitrate to N₂.

3.4 *nif* genes and Nitrogenase Activity

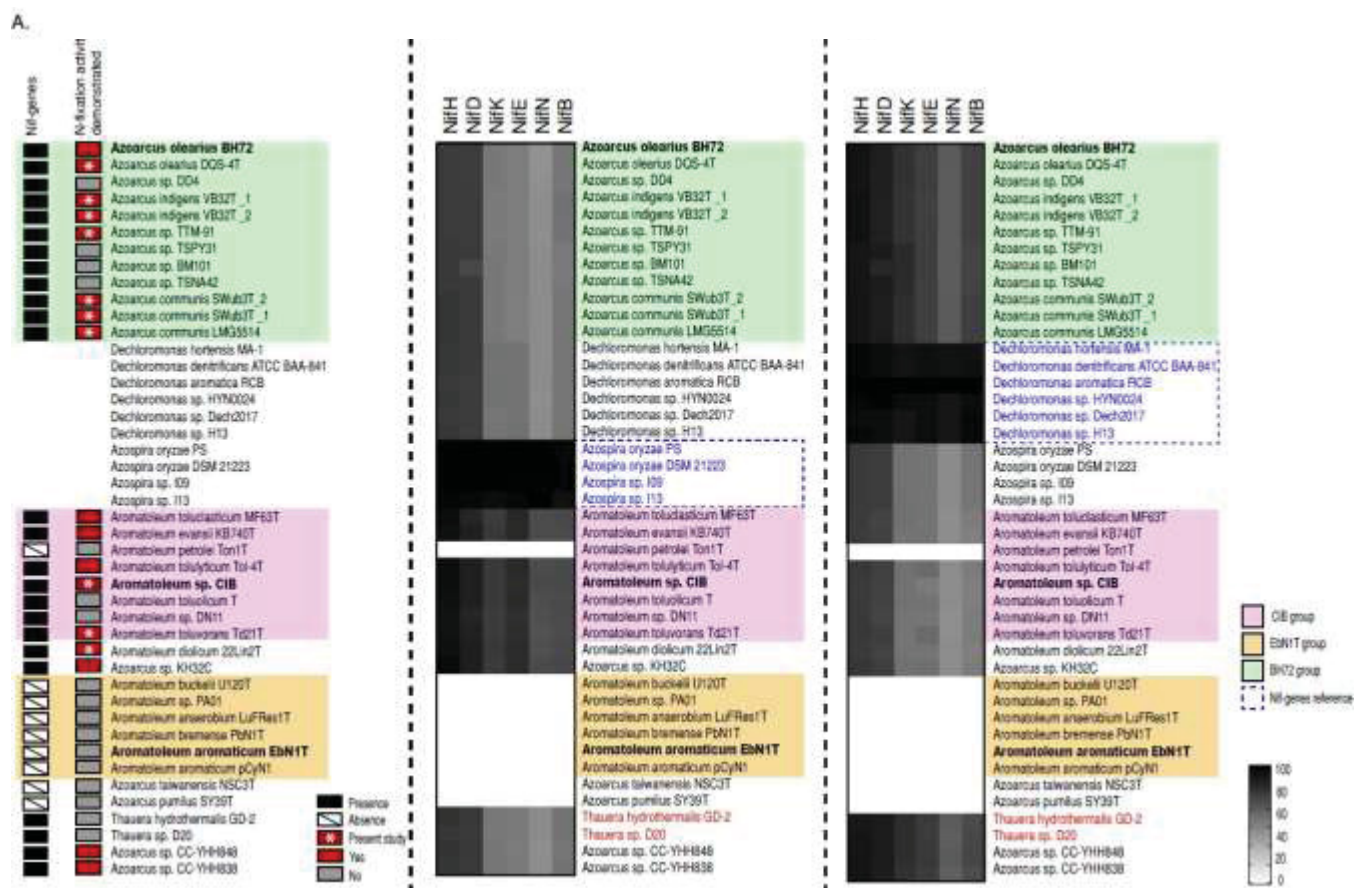
The presence of *nif* genes was analysed in all 30 *Azoarcus*-*Aromatoleum* genomes used in the present study as well as the related genera *Azospira*, *Dechloromonas*, and *Thauera* (Figure 5.4 and Figure 6.4, Figure S5.4; Table 1.4, Table S1.4a). The RAFTS3G-32 cluster analyses showed that *nifH* genes were present in 43 genomes in total. The following organisms lack the *nifH* gene and homology to all other *nif* genes (except *nifUS*): *Ar. buckelii* U120T, *Aromatoleum* sp. PA01, *Ar. anaerobium* LuFRes1, *Ar. bremense* PbN1T, *Ar. aromaticum* EbN1T, *Ar. aromaticum* pCyN1, *Az. taiwanensis* NSC3T, and *Az. pumilus* SY39T. In contrast, two copies of the *nifH* gene were found in *Azoarcus* sp. CC YHH838, *Azoarcus* sp. CC YHH848, *Rhodocyclus tenuis* DSM109T, *Azoarcus* sp. KH32C, and *Ar. toluvorans* Td21T. All 12 *Azoarcus* strains contained the minimal set of *nif* genes necessary for nitrogen fixation (ZHENG *et al.*, 2019) suggesting that all are diazotrophs including some not yet demonstrated experimentally for nitrogenase activity such as *Azoarcus* sp. strain DD4 (the nitrogenase activity of *Az. communis* LMG5514 and *Azoarcus* sp. TTM-91 was confirmed in the present study; Table 2.4). Interestingly, all members of the *Aromatoleum* CIB-Group, with the exception of *Ar. petrolei*, are diazotrophs, albeit with a very different *nif* gene complement to those in *Azoarcus*. This includes the recently described species *Ar. diolicum* and *Ar. toluolicum*; both have been demonstrated to express nitrogenase activity in the present study (Table 2.4), and *Ar. diolicum* by Rabus *et al.* (2019). Although we were unable to demonstrate nitrogenase activity for *Ar. evansii* under any of the conditions tested (Table 2.4), the original description of this species as *Az. evansii* stated that activity could be detected (ANDERS *et al.*, 1995), but did not give details about how the assays were performed. None of the *Aromatoleum* EbN1 Group possessed *nif* genes, as previously reported (RASBUS *et al.*, 2005; RASBUS *et al.*, 2019).

FIGURE 5.4. Single gene *nifH* phylogeny and heat map confirms the presence of two types of *nifH* genes in *Azoarcus*-*Aromatoleum* genomes



REFERENCE: Raittz *et al.*, (2021).

Legend: The *nifH* genes all belonged to a RATFTS3G-67 cluster that was identified by RAFTS3 searches and verified manually.

FIGURE 6.4. Analysis of the minimum set of *nif* genes for diazotrophy (*nifHDKENB*)REFERENCE: Raittz *et al.*, (2021).

Legend: Analysis of the minimum set of *nif* genes for diazotrophy (*nifHDKENB*) reveals the presence of *Azospira*-type *nif* genes in the CIB-Group and *Dechloromonas*-type *nif* genes in the BH72-Group in *Azoarcus*-*Aromatoleum*. The presence of *nif* genes in black and experimentally proven nitrogenase (acetylene reduction) activity (ARA) in red are shown in panel (A) with ARA data obtained during the present study indicated by *. Known *nif* genes from the reference strains *Azospira oryzae* DSM 21223 (B) or *Dechloromonas aromatica* RCB (C) were used to identify homologous genes in the 32 analysed *Azoarcus* (green) and *Aromatoleum* (yellow and pink) genomes as well as in 20 related *Thauera* (note that only the two *Thauera* genomes harbouring *nif* genes are shown). For comparison, four *Azospira* (A) and six *Dechloromonas* (B) genomes are also included. The heat maps indicate homology with a grey scale gradient from white (0%) to black (100%).

TABLE 2.4. Nitrogenase (acetylene reduction) activity of selected strains of *Azoarcus* and *Aromatoleum*.

Strain or Control	Activity $\pm$ SEM nmol C ₂ H ₄ mg Protein ⁻¹ min ⁻¹	Conditions ³
Control ¹	0.111 $\pm$ 0.088	SM-semi-solid
Control ²	not detected	SM-semi-solid
<i>Az. olearius</i> DQS-4 ^T	23.222 $\pm$ 0.826	SM-semi-solid
<i>Az. communis</i> SWub3 ^T	10.611 $\pm$ 0.778	SM-semi-solid
<i>Az. communis</i> LMG5514	26.025 $\pm$ 1.041	SM-semi-solid
<i>Az. indigenus</i> VB32 ^T	17.952 $\pm$ 0.653	1/2CVM-semi-solid
<i>Azoarcus</i> sp. TTM-91	3.960 $\pm$ 0.724	1/2CVM-semi-solid
<i>Aromatoleum</i> sp. CIB	1.163 $\pm$ 0.041	MNF + 3 mM Benzoate-liquid
<i>Ar. toluovorans</i> Td21 ^T	1.318 $\pm$ 0.354	MNF + 3 mM Benzoate-liquid
<i>Ar. diolicum</i> 22Lin2 ^T	0.479 $\pm$ 0.022	UMS + 30 mM Malate-liquid-1% O ₂
<i>Ar. toluolicum</i> T ^T	0.653 $\pm$ 0.049	UMS + 30 mM Malate-liquid-1% O ₂
<i>Ar. evansii</i> KB740 ^T	not detected	All above conditions tested

¹ DQS4 + 20 mM NH₄Cl; ² *nifA* deletion mutant derived from *Az. olearius* DQS-4T; ³ See Supplementary Materials for Details (File S2.4).

The potential origin of the *nif* genes as suggested by their closest relatives was investigated. On the basis of partial *nifD* and *nifH* gene sequences, Faoro *et al.* (2017) suggested that these might be *Dechloromonas* for *Azoarcus* and *Azospira* for *Aromatoleum*. *Dechloromonas*, *Azospira*, *Azoarcus*, *Aromatoleum*, and *Thauera* are all in the order *Rhodocyclales* (LOY *et al.*, 2005), but were in different families: *Dechloromonas* is in the *Azonexaceae*, *Azospira* in the *Rhodocyclaceae*, and *Azoarcus*, *Aromatoleum* and *Thauera* belong to the *Zoogloeaceae*. Phylogenies built from nearly full-length *nifH* (Figure 5.4) and from *nif* cluster genes (Figure S5.4) again confirmed the different *nif* origins of the two groups of diazotrophs in *Azoarcus*-*Aromatoleum*, and the relatedness of the *nif* cluster in the *Aromatoleum* CIB-Group to *Azospira* and that of *Azoarcus* to *Dechloromonas*. The minimal *nif* gene set required for nitrogen fixation (*nifH*, *nifD*, *nifK*, *nifE*, *nifN*, and *nifB*) (DOS SANTOS *et al.*, 2012) of *Azoarcus*-*Aromatoleum* were compared to *Azospira oryzae* 6a3T (HENHOLD-HURECK *et al.*, 2000) and to *D. aromatica* strain RCBT (SALINERO *et al.*, 2009) by RAFTS3G-32 cluster analyses. The heatmaps clearly show that the *nif* genes of the *Aromatoleum* CIB-Group strains are more closely related to *Azospira oryzae* (Figure 6.4B) and those of *Azoarcus* to *D. aromatica* (Figure 6.4C). They also illustrate the absence of *nif* genes in the *Aromatoleum* EbN1 Group and in *Thauera*. Interestingly, both the analyses using the single *nifH* phylogenies (Figure 5.4) and core *nif* genes (Figure 6.4) placed two *Thauera* strains (*T. hydrothermalis* GD-2T and *Thauera* sp. D20) within *Azoarcus*, and these were the same strains that were nested within *Azoarcus* in the whole-genome analysis (Figures S1.4 and S4.4). Rather than suggesting that the genus *Thauera* possessed some diazotrophic members, it lends support to our earlier contention that these “*Thauera*” strains should be more appropriately included within the genus *Azoarcus*.

Frequent gene duplications in the genome of *Azoarcus* sp. KH32C, *Ar. toluvorans* Td21T, and *Azoarcus* sp. CC-YHH848 were revealed by our analysis as well as the presence of two types of *nifH* genes in the genomes of *Azoarcus* sp. CC-YHH838 and *Rhodocyclus tenuis* DSM109T (Figure 5.4).

The separate origins of the *nif* genes in *Azoarcus* (s.s.) and in *Aromatoleum* CIB Group (*Aromatoleum* Group 2) suggests that they were obtained from different sources via horizontal gene transfer (HGT). It is not clear when and why this happened. *Azospira oryzae*, the possible donor of *nif* to *Aromatoleum* Group 2 is a facultative anaerobe associated with Kallar Grass and rice with an ecology similar to *Azoarcus* spp. (HENHOLD-HURECK *et al.*, 2000; ENGELHARD *et al.*, 2000) whereas *Dechloromonas*, the possible donor of *nif* to *Azoarcus*, is also a facultative anaerobe, but is not normally found associated with plants (SALINERO *et al.*, 2009) (i.e., the *nif* donors to the two groups of diazotrophs in *Azoarcus-Aromatoleum* have the opposite ecology to that which would have been intuitively expected). On the other hand, this makes more sense if the *nif* gene regulation in *Azoarcus* and *Aromatoleum* strains is compared (i.e., the *Az. olearius* BH72-related strains possess *nifL* and *nifA* linked to the *rnf1* gene cluster, whereas the *Aromatoleum* CIB group strains do not encode *nifL*, and in this case *nifA* is linked to *nifB*) (Figure S6.4). In contrast to the *Azoarcus* strains, the *Aromatoleum* CIB group strains encode a class of *nifA* proteins that possesses a cysteine-containing interdomain linker that confers oxygen sensitivity (MORETT *et al.*, 1998). The more elaborate *nifL-nifA* regulatory system, when linked to the Rnf1 complex, which can support electron transfer to nitrogenase under aerobic conditions (LEDBETTER *et al.*, 2017), is the one that is more likely to be associated with diazotrophic organisms that are more dependent on aerobic respiration such as *Azoarcus* (BUENO BATISTA *et al.*, 2019). This again illustrates how difficult it is to ascribe a definitive habitat, ecology, and lifestyle to any of the groups in *Azoarcus-Aromatoleum*. Good examples are the two *Az. olearius* strains, BH72 and DQS-4T, that were isolated, respectively, from plants and from oil-contaminated soil, and yet they are almost identical in terms of genes, putatively allowing them to have an endophytic lifestyle (FAORO *et al.*, 2017). In addition, the former *Azoarcus* sp. strain CIB, which is now placed in *Aromatoleum* Group 2, a group of organisms not previously known to be associated with plants, contains many “plant-associated” features in its genome, and can associate with rice endophytically and even express nitrogenase (and possibly produce indole acetic acid) in planta (FERNÁNDEZ-LIAMOSAS *et al.*, 2014; MARTIN-MOLDES *et al.*, 2015; FERNÁNDEZ-LIAMOSAS, *et al.*, 2020).

4 CONCLUSIONS

The aim of this study was neither to describe nor to re-circumscribe new taxa within *Azoarcus-Aromatoleum*, but it does provide an opportunity to suggest how these new genomic data may be used for this purpose in future studies (combined with appropriate phenotypic criteria). The paraphyletic nature of

the *Azoarcus-Aromatoleum* group revealed by the ANI analysis, and by phylogenetic analysis of the core genomes and the 16S rRNA sequences of 30 genome-sequenced strains suggests that its taxonomy could be revised solely on the basis of genetics to either (1) retain the generic name *Azoarcus* for its entirety, or (2) that if *Aromatoleum* is to be retained as a separate genus, it could be divided into two genera (the non-diazotrophic *Aromatoleum* sensu stricto and the mainly diazotrophic *Aromatoleum* Group 2 or “CIB-Group”), and three additional genera could be created comprising *Az. pumilus*, *Az. taiwanensis*, and the single strain lineage *Azoarcus* strain KH32C (which is so far undescribed to species level), respectively; thus creating six genera within *Azoarcus-Aromatoleum*. In terms of ecology, retaining the umbrella name *Azoarcus* might be justified, as with the exception of a few plant-associated strains in *Azoarcus* (s.s.), across the entire *Azoarcus-Aromatoleum* group as most strains/species are found in soil and water (often contaminated with petroleum or related compounds), sewage sludge, and seawater. On the other hand, if metabolism/lifestyle is considered to be the primary factor in describing bacterial genera, the ability of *Aromatoleum* to utilize nitrate as a terminal electron acceptor for the anaerobic degradation of aromatic compounds such as benzoate and the inability of *Azoarcus* (s.s.) to do this makes for an obvious division between *Azoarcus* and *Aromatoleum* as proposed by Rabus *et al.* (2019). However, it could be argued that if *Azoarcus* strains like *Az. communis* Swub3T can be induced to perform the anaerobic degradation of benzoate by the relatively simple insertion of the *bzd* gene cassette (ZAMARRO *et al.*, 2017) then even this phenotype is not so clear as a distinguishing feature. Additional complications are created by the presence of the aforementioned “intermediate” groups. *Azoarcus* sp. KH32C possesses the anaerobic benzoate degradation phenotype, but it clearly does not belong to *Aromatoleum*. Moreover, *Az. pumilus* and *Az. taiwanensis* are both facultative anaerobes (LEE *et al.*, 2014; FU *et al.*, 2019), and neither possess *nif* genes, suggesting a closer affiliation to the EbN1 group of *Aromatoleum*, but genomically, they are not close to either *Azoarcus* or *Aromatoleum* (this study). Similarly, the *Azoarcus* strains CC-YHH838 and CC-YHH848 from Taiwan occupy a peripheral position within the genus *Azoarcus*, and their metabolism (aerobic vs. anaerobic) is so-far undescribed, but they do appear to have fairly typical *Azoarcus*-type *nif* genes (Figure 6.4).

Clearly, more work is required to resolve the taxonomy of *Azoarcus-Aromatoleum*, both in terms of sequencing more genomes, and in terms of examining lifestyles and habitats of this highly diverse and potentially useful group of micro-organisms.

REFERENCES

- ANDREWS, S. *et al.* FastQC: a quality control tool for high throughput sequence data. 2010.
- ANDERS, H.J.; KAETZKE, A.; KAMPFER, P.; LUDWIG, W.; FUCHS, G. Taxonomic position of aromatic- degrading denitrifying pseudo-monad strains K 172 and KB 740 and their description as new members of the genera *Thauera*, as *Thauera aromatica* sp. nov., and *Azoarcus*, as *Azoarcus evansii* sp. nov., respectively, members of the beta subclass of the *Proteobacteria*. **Int. J. Syst. Bacteriol.**, 45, 327–333, 1995.
- AVONTUUR, J.R.; PALMER, M.; BEUKES, C.W.; CHAN, W.Y.; COETZEE, M.P.A.; BLOM, J.; STEPKOWSKI, T.; KYRPIDES, N.C.; WOYKE, T.; SHAPIRO, N.; *et al.* Genome-informed *Bradyrhizobium* taxonomy: Whereto from here? **Syst. Appl. Microbiol.** 42, 427–439. 2019.
- BEUKES, C.W.; PALMER, M.; MANYAKA, P.; CHAN, W.Y.; AVONTUUR, J.R.; VAN ZYL, E.; HUNTEMANN, M.; CLUM, A.; PILLAY, M.; PALANIAPPAN, K.; *et al.* Genome data provides high support for generic boundaries in *Burkholderia sensu lato*. **Front. Microbio.**, 18, 2017.
- BOLGER, A. M.; LOHSE, M; USADEL, B. Trimmomatic: a flexible trimmer for Illumina sequence data. **Bioinformatics**, v. 30, n. 15, p. 2114-2120, 2014.
- BUSHNELL, B. **BBMap: a fast, accurate, splice-aware aligner**. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States), 2014.
- BUENO BATISTA, M.; DIXON, R. Manipulating nitrogen regulation in diazotrophic bacteria for agronomic benefit. **Biochem. Soc. Trans**, 47, 603–614, 2019.
- CHEN, M.H.; SHEU, S.Y.; JAMES, E.K.; YOUNG, C.C.; CHEN, W.M. *Azoarcus olearius* sp. nov., a nitrogen-fixing bacterium isolated from oil-contaminated soil. **Int. J. Syst. Evol. Microbiol.**, 63, 3755–3761, 2013.
- DE LIMA NICHIO, B.T.; DE OLIVEIRA, A.M.R.; DE PIERRI, C.R.; SANTOS, L.G.C.; LEJAMBRE, A.Q.; VIALLE, R.A.; DA ROCHA COIMBRA, N.A.; GUIZELINI, D.; MARCCHAUKOSKI, J.N.; PEDROSA, F.O.; *et al.* RAFTS(3)G: An efficient and versatile clustering software to analyses in large protein datasets. **BMC Bioinform**, 20, 392, 2019
- DE PIERRI, C.R.; VOYCEIK, R.; SANTOS DE MATTOS, L.G.C.; KULIK, M.G.; CAMARGO, J.O.; REPULA DE OLIVEIRA, A.M.; DE LIMA NICHIO, B.T.; MARCCHAUKOSKI, J.N.; DA SILVA FILHO, A.C.; GUIZELINI, ORTAGEA, J.M; PEDROSA, F.O. RAITTZ, R.T. SWeeP: Representing large biological sequences datasets in compact vectors. **Sci. Rep**, 10, 91, 2020.
- DENG, D.; LI, F.; YE, L.; LI, M. Complete genome sequence of *Azoarcus* sp. strain DD4, a gram-negative propanotroph that degrades 1,4-dioxane and 1,1-dichloroethylene. **Microbiol. Resour. Announc.** 8, 2019.
- DEVANADERA, A.; VEJARANO, F.; ZHAI, Y.; SUZUKI-MINAKUCHI, C.; OHTSUBO, Y.; TSUDA, M.; KASAI, Y.; TAKAHATA, Y.; OKADA, K.; NOJIRI, H. Complete genome sequence of an anaerobic benzene-degrading bacterium, *Azoarcus* sp. Strain DN11. **Microbiol. Resour. Announc.** 8, 2019.
- DOS SANTOS, P.C.; FANG, Z.; MASON, S.W.; SETUBAL, J.C.; DIXON, R. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. **BMC Genom**, 13, 1471–2164, 2012.
- ESTRADA-DE LOS SANTOS, P.; PALMER, M.; CHAVEZ-RAMIREZ, B.; BEUKES, C.; STEENKAMP, E.T.;
- BRISCOE, L.; KHAN, N.; MALUK, M.; LAFOS, M.; HUMM, E.; *et al.* Whole genome analyses suggests that *Burkholderia sensu lato* contains two additional novel genera (*Mycetohabitans* gen. nov., and *Trinickia* gen. nov.): Implications for the evolution of diazotrophy and nodulation in the *Burkholderiaceae*. **Genes**, 9, 389, 2018.
- FAORO, H.; RENE MENEGAZZO, R.; BATTISTONI, F.; GYANESHWAR, P.; DO AMARAL, F.P.; TAULE, C.; RAUSCH, S.; GONCALVES GALVAO, P.; DE LOS SANTOS, C.; MITRA, S.; *et al.* The oil-contaminated soil diazotroph *Azoarcus olearius* DQS-4(T) is genetically and phenotypically similar to the model grass endophyte *Azoarcus* sp. BH72. **Environ. Microbiol. Rep**, 9, 223–238, 2017.

- FERNÁNDEZ, H.; PRANDONI, N.; FERNÁNDEZ-PASCUAL, M.; FAJARDO, S.; MORCILLO, C.; DÍAZ, E.; CARMONA, M. *Azoarcus* sp. CIB, an anaerobic biodegrader of aromatic compounds shows an endophytic lifestyle. **PLoS One**, 9, e110771, 2014.
- FERNÁNDEZ-LLAMOSAS, H.; IBERO, J.; THIJS, S.; IMPERATO, V.; VANGRONSVELD, J.; DÍAZ, E.; CARMONA, M. Enhancing the rice seedlings growth promotion abilities of *Azoarcus* sp. CIB by heterologous expression of ACC deaminase to improve performance of plants exposed to cadmium stress. **Microorganisms**, 8, 1453, 2020.
- FU, G.-Y.; YU, X.-Y.; YU, X.-D.; ZHAO, Z.; CHEN, C.; WANG, R.-J.; WU, M.; ZHANG, X.-Q. *Azoarcus pumilus* sp. nov., isolated from seawater in Sanya, China. **Int. J. Syst. Evol. Microbiol.**, 69, 1459–1464, 2019.
- GUREVICH, A. *et al.* QUAST: quality assessment tool for genome assemblies. **Bioinformatics**, v. 29, n. 8, p. 1072–1075, 2013.
- HUREK, T.; EGENER, T.; REINHOLD-HUREK, B. Divergence in nitrogenases of *Azoarcus* spp., *Proteobacteria* of the beta subclass. **J. Bacteriol.**, 179, 4172–4178, 1997.
- HUREK, T.; REINHOLD-HUREK, B. *Azoarcus* sp. strain BH72 as a model for nitrogen-fixing grassendophytes. **J. Bacteriol.**, 106, 169–178, 2003.
- HUSON, D.H.; SCORNAVACCA, C. Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. **Syst. Biol.**, 61, 1061–1067, 2012.
- JAIN, C.; RODRIGUEZ-R, L.M.; PHILLIPPY, A.M.; KONSTANTINIDIS, K.T.; ALURU, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. **Nat. Commun.**, 9, 5114, 2018.
- JUNGHARE, M.; PATIL, Y.; SCHINK, B. Draft genome sequence of a nitrate-reducing, o-phthalate degrading bacterium, *Azoarcus* sp. strain PA01(T). **Stand Genom. Sci.**, 10, 90, 2015.
- KRAFT, B.; STROUS, M.; TEGETMEYER, H.E. Microbial nitrate respiration—Genes, enzymes and environmental distribution. **J. Biotechnol.**, 155, 104–117, 2011.
- LEE, D.-J.; WONG, B.-T.; ADAV, S.S. *Azoarcus taiwanensis* sp. nov., a denitrifying species isolated from a hot spring. **Appl. Microbiol. Biotechnol.**, 98, 1301–1307, 2014.
- LOY, A.; SCHULZ, C.; LÜCKER, S.; SCHÖPFER-WENDELS, A.; STOECKER, K.; BARANYI, C.; LEHNER, A.; WAGNER, M. 16S rRNA gene-based oligonucleotide microarray for environmental monitoring of the betaproteobacterial order “*Rhodocyclales*”. **Appl. Environ. Microbiol.**, 71, 1373–1386, 2005.
- MARTIN-MOLDES, Z.; ZAMARRO, M.T.; DEL CERRO, C.; VALENCIA, A.; GOMEZ, M.J.; ARCAS, A.; UDAONDO, Z.; GARCIA, J.L.; NOGALES, J.; CARMONA, M.; *et al.* Whole-genome analysis of *Azoarcus* sp. strain CIB provides genetic insights to its different lifestyles and predicts novel metabolic features. **Syst. Appl. Microbiol.**, 38, 462–471, 2015.
- MECHICHI, T.; STACKEBRANDT, E.; GAD’ON, N.; FUCHS, G. Phylogenetic and metabolic diversity of bacteria degrading aromatic compounds under denitrifying conditions, and description of *Thauera phenylacetica* sp. nov., *Thauera aminoaromatica* sp. nov., and *Azoarcus buckelii* sp. nov. **Arch. Microbiol.**, 178, 26–35, 2002.
- NISHIZAWA, T.; TAGO, K.; OSHIMA, K.; HATTORI, M.; ISHII, S.; OTSUKA, S.; SENOO, K. Complete genome sequence of the denitrifying and N₂O-reducing bacterium *Azoarcus* sp. strain KH32C. **J. Bacteriol.**, 194, 06618–11, 2012.
- NURK, S. *et al.* Assembling genomes and mini-metagenomes from highly chimeric reads. In: **Annual international conference on research in computational molecular biology**. Springer, Berlin, Heidelberg, 2013. p. 158–170.
- ORMEÑO-ORRILLO, E.; MARTÍNEZ-ROMERO, E. A genomotaxonomy view of the *Bradyrhizobium* genus. **Front. Microbiol.**, 10, 2019.
- QIN, N.; YANG, F.; LI, A.; PRIFTI, E.; CHEN, Y.; SHAO, L.; GUO, J.; LE CHATELIER, E.; YAO, J.; WU, L.; *et al.* Alterations of the human gut microbiome in liver cirrhosis. **Nature**, 513, 59–64, 2014.

QIN, Q.-L.; XIE, B.-B.; ZHANG, X.-Y.; CHEN, X.-L.; ZHOU, B.-C.; ZHOU, J.; OREN, A.; ZHANG, Y.-Z. A proposed genus boundary for the prokaryotes based on genomic insights. **J. Bacteriol.**, 196, 2210–2215, 2014.

QUAST, C.; PRUESSE, E.; YILMAZ, P.; GERKEN, J.; SCHWEER, T.; YARZA, P.; PEPLIES, J.; GLÖCKNER, F.O. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. **Nucleic Acids Res.**, 41, D590–D596, 2012.

RABUS, R.; KUBE, M.; BECK, A.; WIDDEL, F.; REINHARDT, R. Genes involved in the anaerobic degradation of ethylbenzene in a denitrifying bacterium, strain EbN1. **Arch. Microbiol.**, 178, 506–516, 2002.

RABUS, R.; KUBE, M.; HEIDER, J.; BECK, A.; HEITMANN, K.; WIDDEL, F.; REINHARDT, R. The genome sequence of an anaerobic aromatic-degrading denitrifying bacterium, strain EbN1. **Arch. Microbiol.**, 183, 27–36, 2005.

REINHOLD, B.; HUREK, T.; FENDRIK, I.; POT, B.; GILLIS, M.; KERSTERS, K.; THIELEMANS, S.; DE LEY, J. *Azospirillum halopraeferens* sp. nov., a nitrogen-fixing organism associated with roots of Kallar Grass (*Leptochloa fusca* (L.) Kunth). **Int. J. Syst. Evol. Microbiol.**, 37, 43–51, 1987.

REINHOLD, B.; HUREK, T.; NIEMANN, E.-G.; FENDRIK, I. Close association of *Azospirillum* and diazotrophic rods with different root zones of Kallar Grass. **Appl. Environ. Microbiol.**, 52, 520–526, 1986.

REINHOLD-HUREK, B.; HUREK, T. Reassessment of the taxonomic structure of the diazotrophic genus *Azoarcus sensu lato* and description of three new genera and new species, *Azovibrio restrictus* gen. nov., sp. nov., *Azospira oryzae* gen. nov., sp. nov. and *Azonexus fungiphilus* gen. nov., sp. nov. **Int. J. Syst. Evol. Microbiol.**, 2, 649–659, 2000.

REINHOLD-HUREK, B.; HUREK, T.; GILLIS, M.; HOSTE, B.; VANCANNEYT, M.; KERSTERS, K.; LEY, J. *Azoarcus* gen. nov., nitrogen-fixing proteobacteria associated with roots of Kallar Grass (*Leptochloa fusca* (L.) Kunth), and description of two species, *Azoarcus indigens* sp. nov. and *Azoarcus communis* sp. nov. **Int. J. Syst. Evol. Microbiol.**, 43, 1993.

SALINERO, K.K.; KELLER, K.; FEIL, W.S.; FEIL, H.; TRONG, S.; DI BARTOLO, G.; LAPIDUS, A. Metabolic analysis of the soil microbe *Dechloromonas aromatica* str. RCB: Indications of a surprisingly complex life-style and cryptic anaerobic pathways for aromatic degradation. **BMC Genom.**, 10, 1471–2164, 2009.

SEEMANN, T. Prokka: rapid prokaryotic genome annotation. **Bioinformatics**, v. 30, n. 14, p. 2068–2069, 2014.

SIEVERS, F.; WILM, A.; DINEEN, D.; GIBSON, T.J.; KARPLUS, K.; LI, W.; LOPEZ, R.; MCWILLIAM, H.; REMMERT, M.; SÖDING, J.; *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. **Mol. Syst. Biol.**, 7, 539, 2011.

SONG, B.; HAGGBLOM, M.M.; ZHOU, J.; TIEDJE, J.M.; PALLERONI, N.J. Taxonomic characterization of denitrifying bacteria that degrade aromatic compounds and description of *Azoarcus toluvorans* sp. nov. and *Azoarcus toluclasticus* sp. nov. **Int. J. Syst. Bacteriol.**, 3, 1129–1140, 1999.

SPRINGER, N.; LUDWIG, W.; PHILIPP, B.; SCHINK, B. *Azoarcus anaerobius* sp. nov., a resorcinol-degrading, strictly anaerobic, denitrifying bacterium. **Int. J. Syst. Bacteriol.**, 48, 953–956, 1998.

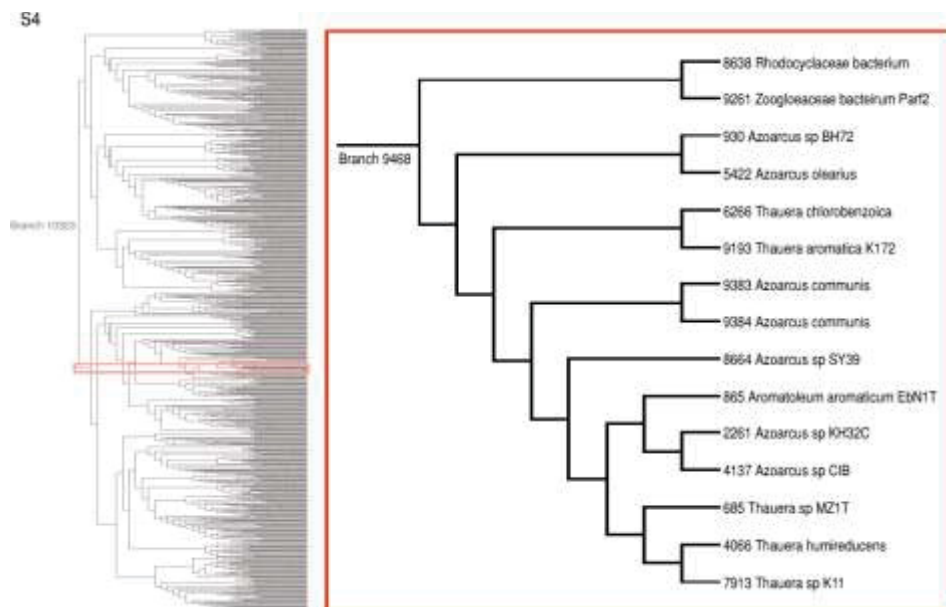
VIALLE, R.A.; DE OLIVEIRA PEDROSA, F.; WEISS, V.A.; GUIZELINI, D.; TIBAES, J.H.; MARCCHAUROSKI, J.N.; DE SOUZA, E.M.; RAITTZ, R.T. RAFTS3: Rapid Alignment-Free Tool for Sequence Similarity Search. **bioRxiv**, 2016.

ZHENG, W.-S.; LIANG, Q.-Y.; DU, Z.-J. *Thauera lacus* sp. nov., isolated from a saline lake in Inner Mongolia. **Int. J. Syst. Evol. Microbiol.**, 69, 3786–3791, 2019.

ZHOU, J.; FRIES, M.R.; CHEE-SANFORD, J.C.; TIEDJE, J.M. Phylogenetic analyses of a new group of denitrifiers capable of anaerobic growth on toluene and description of *Azoarcus tolulyticus* sp. nov. **Int. J. Syst. Bacteriol.**, 45, 500–506, 1995.

SUPPLEMENTARY INFORMATION

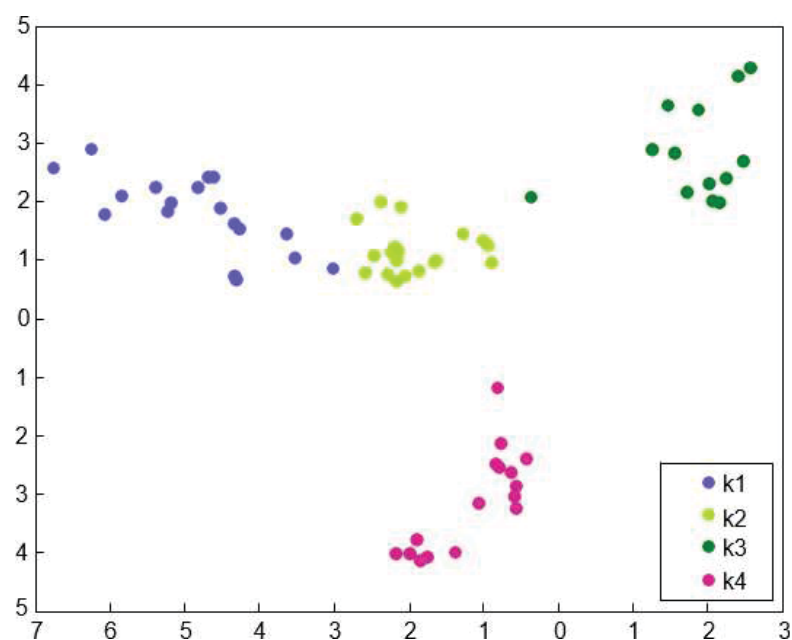
FIGURE S1.4. Clustering of *Azoarcus*, *Aromatoleum*, and *Thauera* (branch 9468) in the global bacterial tree previously presented in the study by De Pierri *et al.* (2020), with 10,324 whole genomes



REFERENCE: Raittz *et al.*, (2021).

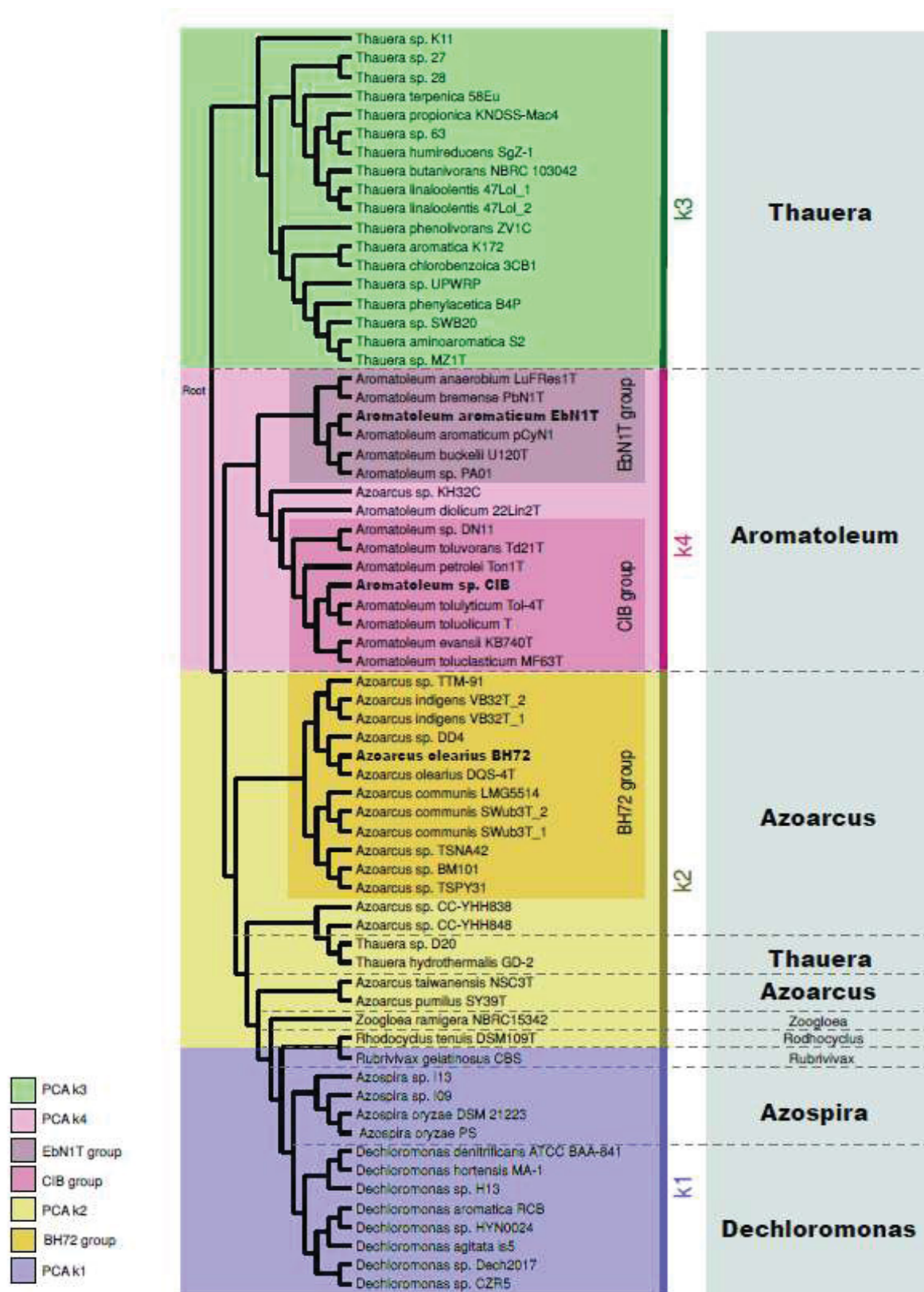
FIGURE S2.4. Whole genome comparison of 67 genomes analyzed in this study based on sweep vectors

a)



(Continue)

b)

REFERENCE: Raittz *et al.*, (2021).

Legend: (a) Two principal components projection for k-means clusters (k = 4) generated with SWeeP vectors; k1, k2, k3, and k4 are related to groups in the neighbour joining phylogeny presented in (b)

FIGURE S3.4. Clustal Omega phylogenetic inference of 16S rRNA analysis for all 67 genomes in this study

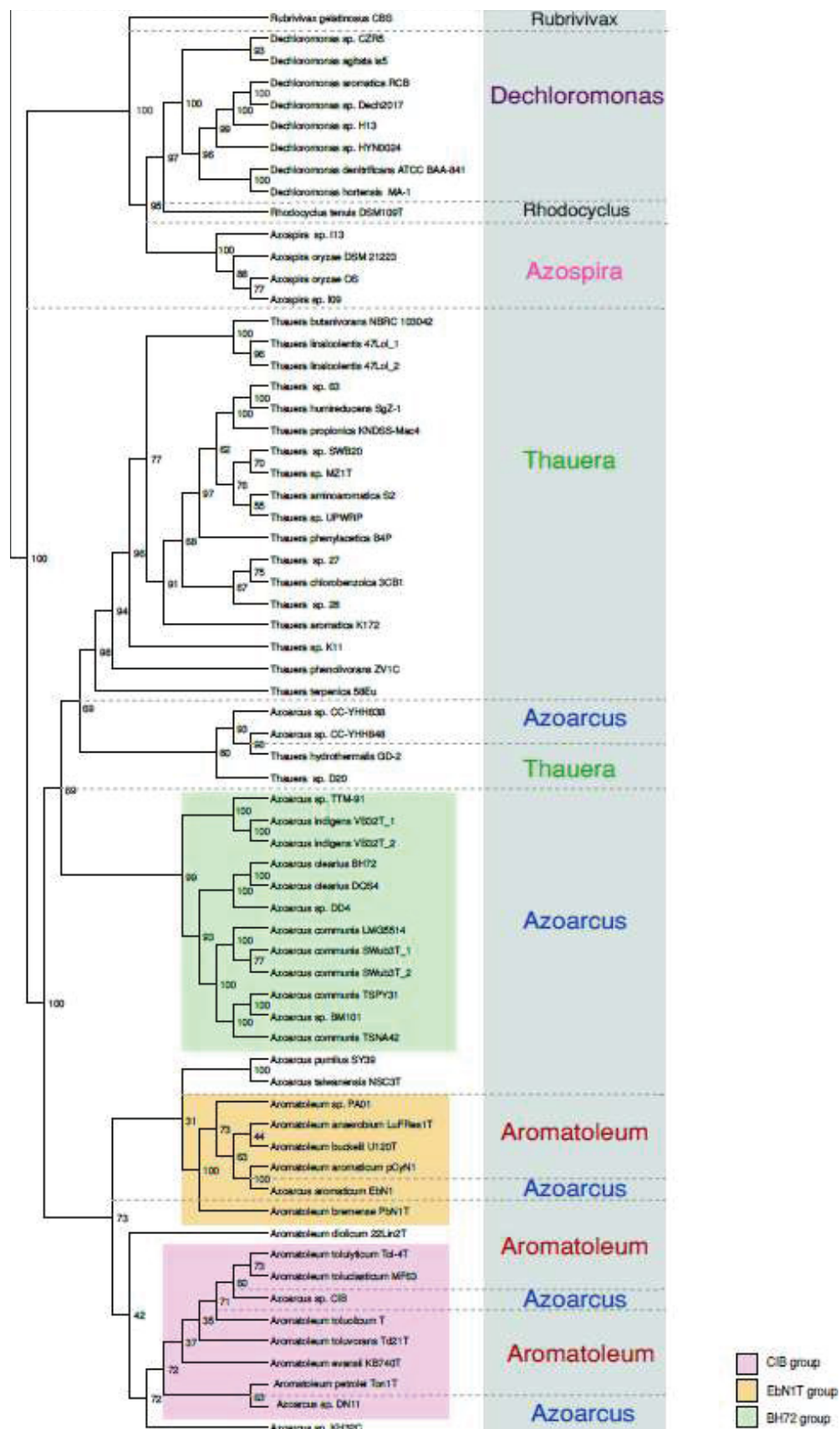
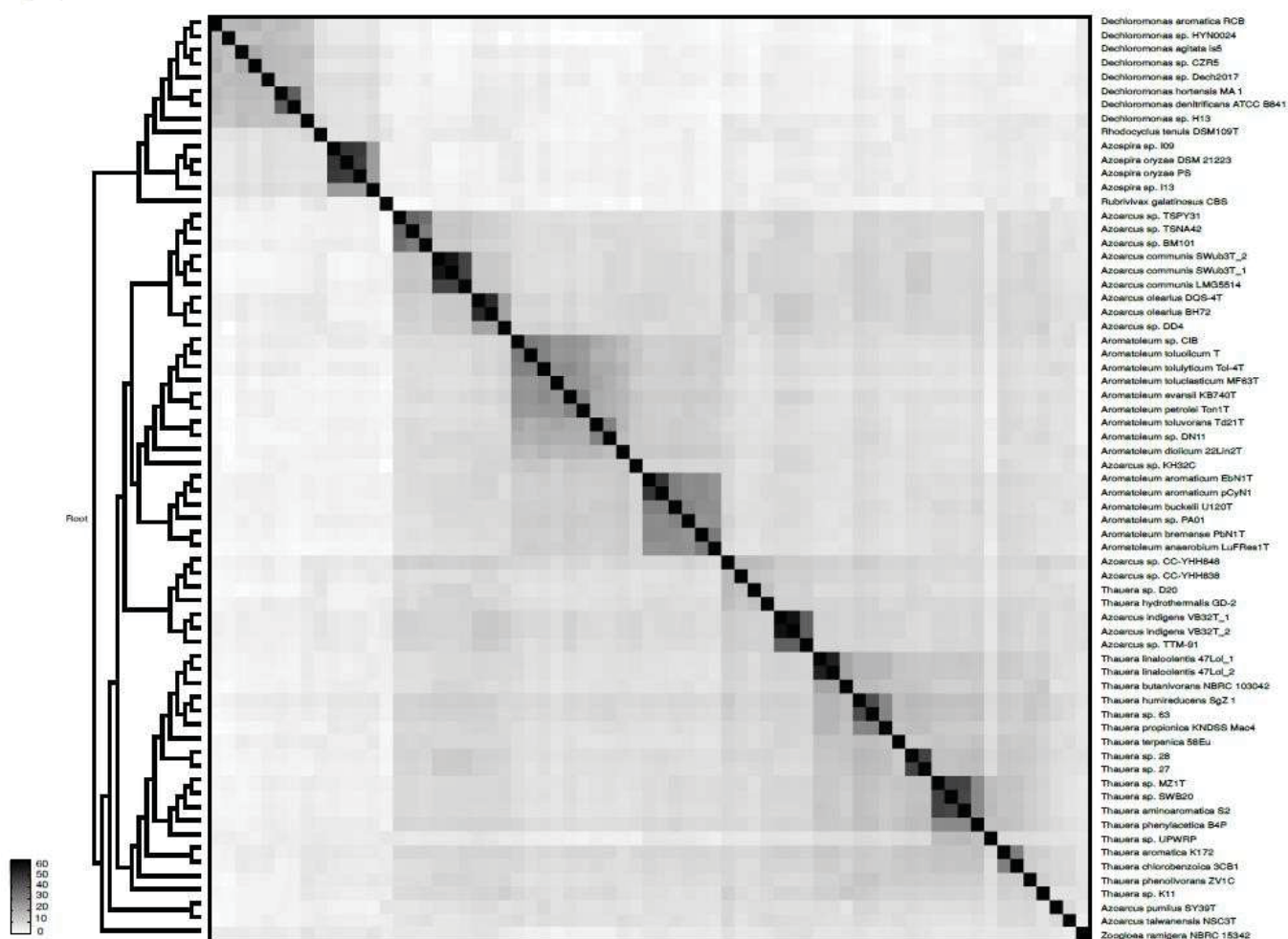
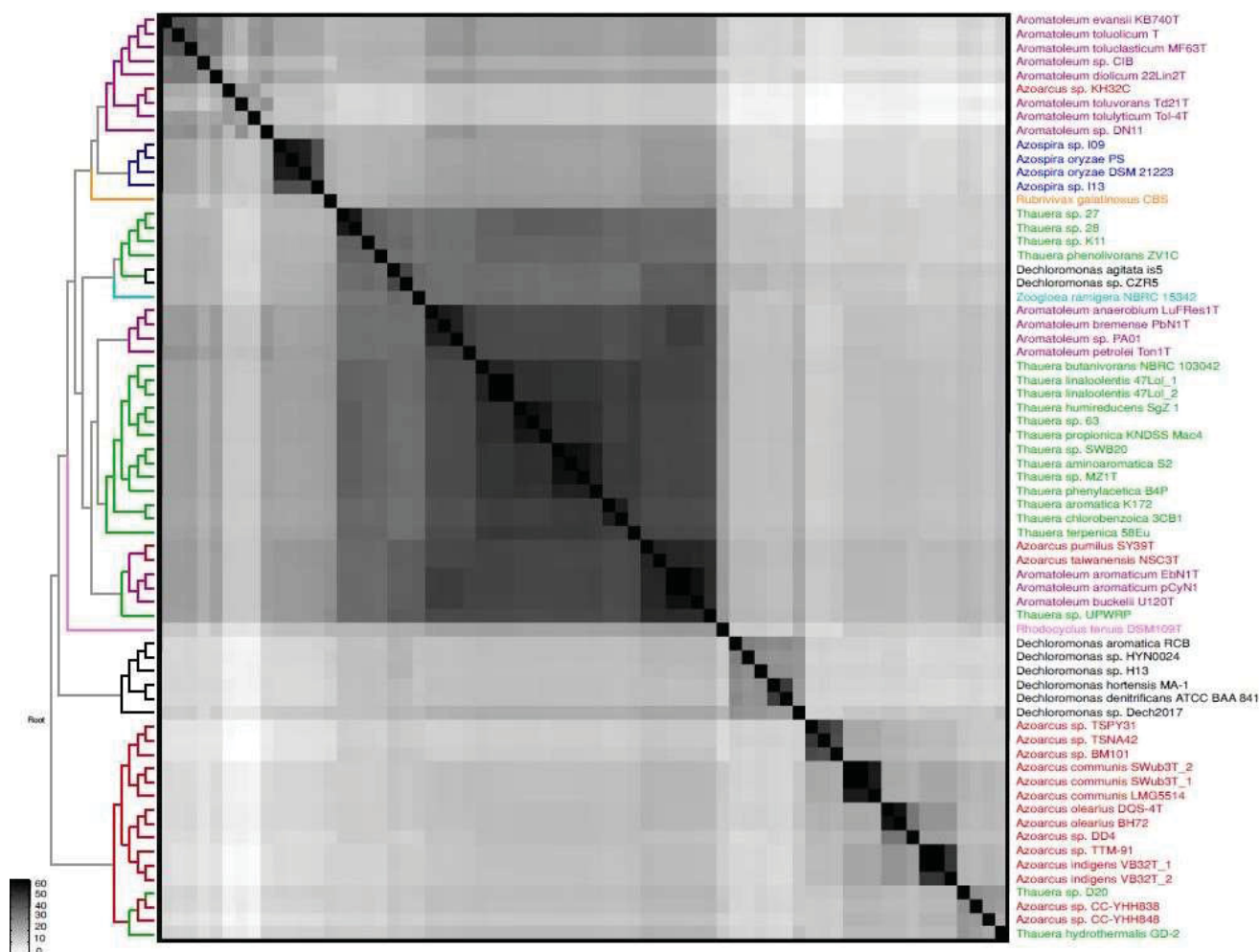


FIGURE S4.4. Core gene cluster comparisons for all 67 genomes



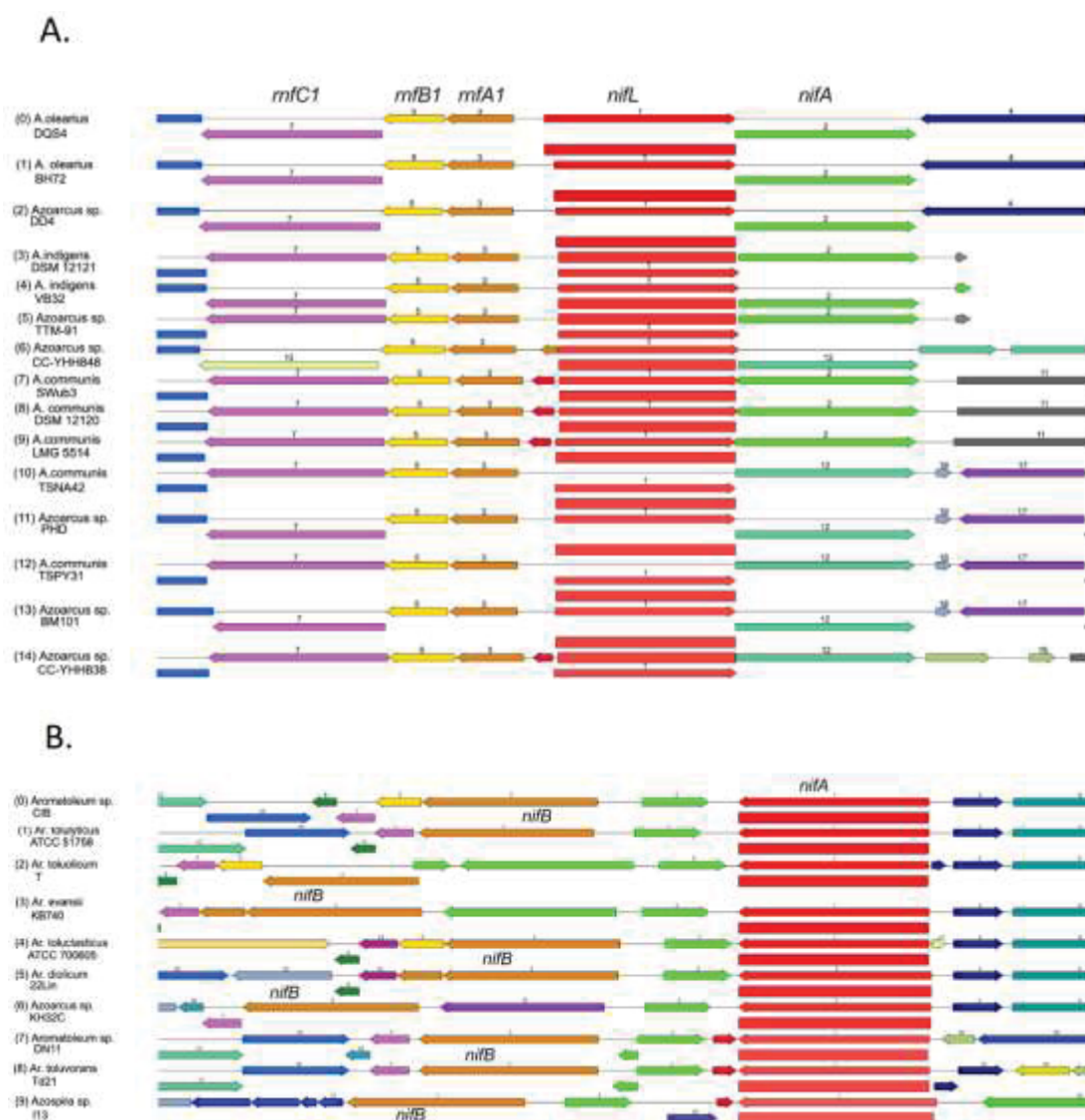
Legend: As observed in the complete genome analysis, *Azospira* and *Dechloromonas* were separated from the other genomes in the NJ phylogeny based on SWeeP projection and the ordered vectors distance matrix heatmap.

FIGURE S5.4. Nj phylogeny and heatmap illustrating the comparison of 67 genomes in this study based on SWEEP projection of the *nif* gene cluster set for each organism



REFERENCE: Raittz *et al.*, (2021).

FIGURE S6.4. Comparative analysis of the gene neighbourhoods of *nif* regulatory genes carried out with PATRIC 3.6.6



REFERENCE: Raittz *et al.*, 2021.

Legend: (A) Search conducted with *NifL* (ANQ83604.1) from *Azoarcus olearius* strain DQS4 as reference. *nifL*, its protein product and homologs are shown in red, with *nifA* homologs (labelled as either 2 or 12 according to product annotation) in green. *nifA1*, *nifB1*, and *nifC1* are depicted in brown, yellow and mauve, respectively. The gene depicted in light green (number 13) from *Azoarcus* sp. CC-YHH848 is also a homolog of *nifC1*. The gene depicted in light blue at the left of the figure is *nifD1*. (B) Search conducted with *NifA* (AKU10620.1) from *Aromatoleum* sp. CIB as reference. *nifA*, its protein product and homologs are shown in red. Genes shown in light green to the left of *nifA* encode hypothetical proteins. *nifB* is depicted in brown. Genes located to the right of *nifA*, depicted in dark blue (number 4) and cyan (number 6) are adjacent to *nifA* in seven of the strains shown and encode a nucleoside diphosphate kinase and a 23S RNA adenine methyl transferase, respectively.

TABLE S1.4. Accession numbers of sequences used in phylogenies

Strain	Other collection No	Assembly Accession	Genome size [bp]	Genome GC content [%]	<i>nif</i> genes	Nitrogenase activity confirmed by ARA	Isolation site and/or type of habitat	Country of origin	Genome submitter
<i>Aromatoleum anaerobium</i> LuFRes1 ^T	DSM 12081, LMG 30943	WTVG000000000	4,482,959	66.72	No	No	Sewage sludge	Germany	This study
<i>Aromatoleum aromaticum</i> EbN1 ^T	DSM 19018, LMG 30748	GCA_000025965.1	4,727,255	64.68	No	No	River water	Germany	Max Planck Institute
<i>Aromatoleum aromaticum</i> pCyN1	DSM 19016, LMG31004	WTV000000000	4,208,498	65.23	No	No	River water	Germany	This study
<i>Aromatoleum bremense</i> PbN1 ^T	DSM 19017, LMG 31005	WTVP000000000	4,319,217	65.57	No	No	River water	Germany	This study
<i>Aromatoleum buckelii</i> U120 ^T	DSM 14744, LMG 26916	WTVH000000000	4,227,546	65.33	No	No	Oxic soil	Germany	This study
<i>Aromatoleum diolicum</i> 22Lin2 ^T	DSM 15408, LMG 30750	WTVQ000000000	5,138,050	64.22	Yes	Yes	River water	Germany	This study
<i>Aromatoleum evansii</i> KB740 ^T	DSM 6898, CIP 109473, NBRC 107771	WTVK000000000	5,883,543	65.98	Yes	Yes	Creek sediment	USA	This study
<i>Aromatoleum petrolei</i> Ton1 ^T	DSM 19019, LMG 30746	WTVR000000000	5,364,035	65.73	No	No	River water	Germany	This study
<i>Aromatoleum</i> sp. CIB	CECT #5669	GCA_001190925.1	5,257,030	65.84	Yes	Yes	Diesel fuel-contaminated aquifer	Switzerland	CIB-CSIC
<i>Aromatoleum</i> sp. DN11	NBRC 108892	GCA_003628555.1	4,956,835	66.30	Yes	Yes	Gasoline-contaminated groundwater	Japan	The University of Tokyo
<i>Aromatoleum</i> sp. PA01		GCA_001274695.1	3,908,240	66.10	No	No	Sewage sludge; wastewater treatment plant, Konstanz	Germany	University of Konstanz
<i>Aromatoleum toluclasticum</i> MF63 ^T	ATCC 700605	GCA_0000378245.1	5,925,983	66.00	No	No	Shallow aquifer sediment, Moffett Field, California	USA	DOE Joint Genome Institute

(Continue)

<i>Aromatoleum toluolicum</i> Tol-4 ^T	ATCC 51758	GCA_900156155.1	5,110,391	66.43	Yes	Yes	Petroleum contaminated aquifer sediments, northern Michigan	USA	DOE Joint Genome Institute
<i>Aromatoleum toluolicum</i> T ^T	DSM 19020, LMG 30751	WTVS000000000	6,025,652	65.93	Yes	nt	Petroleum contaminated aquifer	Switzerland	This study
<i>Aromatoleum toluorans</i> Td21 ^T	ATCC 700604, DSM 15124	WTVN000000000	4,786,256	66.66	Yes	Yes	Muck soil	USA	This study
<i>Azoarcus communis</i> LMG5514	BPD2	WTVJ000000000	4,996,403	62.45	Yes	Yes	Oily sludge coming from a storage tank, Elf petroleum refinery, Donge France	France	This study
<i>Azoarcus communis</i> SWub3 ^T	DSM12120	WTVI000000000	4,976,326	62.36	Yes	Yes	Kallar grass roots	Pakistan	This study
<i>Azoarcus communis</i> SWub3 ^T	DSM 12120, LMG 22127, ATCC 51397, LMG 9095, NCIMB 13565	GCA_0032226565.1	5,004,685	62.50	Yes	Yes	Kallar grass roots	Pakistan	University of California, Davis
<i>Azoarcus</i> sp. TSN42	KCCM 11872P	GCA_003111665.1	4,886,934	62.80	Yes	nt	Oil-contaminated sediment, Taean	South Korea	Gyeongbuk Institute for Marine Bioindustry
<i>Azoarcus</i> sp. TSPY31	KCCM 11873P	GCA_003111645.1	4,572,081	63.20	Yes	nt	Oil-contaminated sediment, Taean	South Korea	Gyeongbuk Institute for Marine Bioindustry (GIMB)
<i>Azoarcus</i> sp. TTM-91		WUEI000000000	5,393,782	67.70	Yes	Yes	River water	Taiwan	This study
<i>Azoarcus indigenis</i> VB32 ^T		WTVL000000000	5,562,509	67.31	Yes	Yes	Kallar grass roots	Pakistan	This study
<i>Azoarcus indigenis</i> VB32 ^T	DSM12121	GCA_004361735.1	5,464,470	67.60	Yes	Yes	Kallar grass roots	Pakistan	DOE Joint Genome Institute
<i>Azoarcus olearius</i> BH72		GCA_000061505.1	4,376,040	67.90	Yes	Yes	Kallar grass roots	Pakistan	Bielefeld University
<i>Azoarcus olearius</i> DQS-4 ^T	BCRC 80407, KCTC 23918, LMG 26893	GCA_001682385.1	4,451,751	67.83	Yes	Yes	Oil-contaminated soil	Taiwan	UFPR
<i>Azoarcus pumilus</i> SY39 ^T	MCCC:1K03430, KCTC 62157	GCA_002872475.1	3,225,512	66.53	No	No	Sea water	China	Zhejiang University
<i>Azoarcus</i> sp. BM101		GCA_002869045.1	4,904,245	62.90	Yes	nt	Estuary sediment	USA	University of California, Berkeley (Continue)

[illegible]

<i>Thauera butanivorans</i> NBRC 103042	ATCC 43655, DSM 2080, JCM 20651, IAM 12574, Bu B1211	GCF_001591165_1	4,792,019	66.11		Activated sludge, wastewater	Japan	National Institute of Technology and Evaluation
<i>Thauera chlorobenzoica</i> 3CB1		GCF_001922305_1	3,772,491	67.30		3-chlorobenzoate-degrading denitrifying enrichment culture inoculated with sediment from the Hudson River near Albany, NY, USA	USA	Friedrich Schiller University of Jena
<i>Thauera humireducens</i> SgZ-1	KACC 16524, CCTCC M2011497	GCF_001051995_2	4,171,754	66.42		Anode biofilm of a sediment microbial fuel cell	China	Guangdong Institute of Eco-Environmental and Soil Sciences
<i>Thauera hydrothermalis</i> GD-2		GCF_003345255_1	3,059,321	63.58	Yes	Hot spring	China	Chinese Academy of Science, Institute of Microbiology
<i>Thauera linaloolentis</i> 47LoI	DSM 12138	GCF_000310205_1	4,199,939	66.60		Culture collection DSM12138	Germany	Norwegian University of Life Sciences
<i>Thauera linaloolentis</i> 47LoI		GCF_000621305_1	4,259,404	66.60		Activated sludge	Germany	DOE Joint Genome Institute
<i>Thauera phenolivorans</i> ZV1C	CGMCC 1.15497, NCBR 112379	GCF_001696715_1	4,199,939	66.58		Activated sludge, wastewater	China	Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences
<i>Thauera phenylacetica</i> B4P	DSM 14743, LMG 26918	GCF_000310225_1	5,022,393	68.63		Anaerobic sewage sludge	Germany	Norwegian University of Life Sciences
<i>Thauera propionica</i> KNDSS-Mac4		GCF_002245655_1	3,927,168	65.87		Sediment sample	India	Institute of Microbial Technology
<i>Thauera terpenica</i> 58Eu	DSM 12139	GCF_000443165_1	4,140,314	64.30		Ditch in a forest	Germany	Norwegian University of Life Sciences
<i>Azospira</i> sp. 109		GCF_009177135_1	4,167,394	64.75	Yes	Aquatic biome	Japan	Tokyo University of Agriculture and Technology
					nt			(Continue)

<i>Azospira</i> sp. I13	GCF_003112695_1	3,785,995	64.00	Yes	nt	Wastewater treatment plant	Japan	Tokyo University of Agriculture and Technology	
<i>Azospira</i> <i>oryzae</i> DSM 21223	6a3, CCUG 45016, LMG 9096	GCF_004217225_1	3,450,622	66.31	Yes	Yes	Surface sterilized roots of kallar grass	Pakistan	DOE Joint Genome Institute
<i>Azospira</i> <i>oryzae</i> PS		GCF_000236665_1	3,806,980	65.37	Yes	nt	Swine waste primary lagoon sediment	USA	US DOE Joint Genome Institute (JGI-PGF)
<i>Dechloromonas</i> sp. CZR5		GCF_008711155_1	4,287,931	61.40			Soil	China	Nanjing Agriculture University
<i>Dechloromonas</i> sp. Dech2017		GCF_009469585_1	3,900,498	62.90	Yes	nt	Activated sludge from full- scale wastewater treatment	China	Xiamen University
<i>Dechloromonas</i> sp. H13		GCF_009469615_1	3,672,233	64.30	Yes	nt	Activated sludge from full- scale wastewater treatment	China	Xiamen University
<i>Dechloromonas</i> sp. HYN0024		GCF_003441615_1	3,258,300	59.35	Yes	nt	Fresh water	South Korea	Korea Advanced Institute of Science and Technology
<i>Dechloromonas</i> <i>agitata</i> is5		GCF_000519045_1	3,627,226	62.82			Hanford 300 Area sediments	USA	DOE Joint Genome Institute
<i>Dechloromonas</i> <i>aromatica</i> RCB		GCF_000012425_1	4,501,104	59.25	Yes	nt	Enriched river sediments	USA	DOE Joint Genome Institute
<i>Dechloromonas</i> <i>denitrificans</i> ATCC BAA-841	ED1, DSM 15892	GCF_001551835_1	4,228,015	61.76	Yes	nt	Earthworm gut	Germany	Korea University
<i>Dechloromonas</i> <i>hortensis</i> MA-1	DSM 15637, ATCC BAA-776	GCF_009469765_1	4,303,504	61.40	Yes	nt	Garden soil	Netherlands	Xiamen University
<i>Zoogloea</i> <i>ramigera</i> NBRC 15342	ATCC 19623, NCIB 10340, NCTC 10482	GCA_006539865.1	4605859	67.2			Activated sludge	USA	National Institute of Technology and Evaluation Biological Resource Center
<i>Rhodocyclus</i> <i>tenuis</i> DSM109 ^T	Pfennig 2761, ATCC 25093, DSMZ 109, LMG 4367	GCA_009469755.1	3854478	64.5			Forest pond	Unknown	Xiamen University
<i>Rubrivivax</i> <i>gelatinosus</i> CBS		GCA_000257145.1	5010880	71.2			Soil	USA	Michigan State University

REFERENCE: Adapted from Ralitz *et al.*, 2021.

TABLE S2.4. Genome sequencing statistics for strains sequenced in this study

Strain	Az. communis	Az. communis	Az. indigens	Az. sp.	Az. taiwanensis	Ar. anaerobium	Ar. aromaticum	Ar. bremse	Ar. buckelii	Ar. difolicum	Ar. evansii	Ar. petrolei	Ar. toluolicum	Ar. toluvorans
	Swub3 ^T = DSM 12120 ^T = ATCC 51397 ^T = LMG 9095 ^T	LMG 5514 = BPD2 = S2	VB32 ^T = ATCC 51398 ^T = LMG 9092 ^T	TTM-91	NSC3 ^T = DSM 24109 ^T = BCRC 80111 ^T	LuFRes1 ^T = DSM 12081 ^T = LMG 30943 ^T	pCYN1 = DSM 19016 = LMG 31004	PbN1 ^T = DSM 19017 ^T = LMG 31005 ^T	U120 ^T = DSM 14744 ^T = LMG 26916 ^T	22Lin ^T = DSM 15408 ^T = LMG 30750 ^T	KB740 ^T = DSM 6898 ^T = CIP 109473 ^T = NBRC 107771 ^T	ToN1 ^T = DSM 19019 ^T = LMG 30746 ^T	T ^T = DSM 19020 ^T = LMG 30751 ^T	Td21 ^T = ATCC 700604 ^T = DSM 15124 ^T
NCBI taxID	41977	41977	29545	2691581	666964	182180	551760	76115	200254	75796	59406	76116	90060	92002
NCBI BioProject	PRJNA594342	PRJNA594346	PRJNA594494	PRJNA596905	PRJNA594497	PRJNA59430	PRJNA594500	PRJNA594503	PRJNA594332	PRJNA594504	PRJNA594348	PRJNA594506	PRJNA594509	PRJNA59499
NCBI BioSample	SAMN13513232	SAMN13513262	SAMN13519732	SAMN13642130	SAMN13519749	SAMN13512727	SAMN13519788	SAMN13519790	SAMN13513114	SAMN13519793	SAMN13513296	SAMN13519794	SAMN13519821	SAMN13519781
Number of reads	2,912,790	589,909	3,786,157	681,151	443,520	511,031	436,039	537,610	1,682,146	418,357	1,941,512	604,510	560,791	1,008,767
Coverage	205.9X	50.4X	225.0X	55.1X	41.2X	44.2X	46.0X	50.7X	174.4X	32.3X	124.5X	51.9X	40.5X	80.9X
N50	114,813	82,640	70,676	39,942	91,316	55,632	49,727	57,288	83,257	104,601	80,793	88,779	70,006	103,748
L50	12	17	18	38	17	25	29	24	16	17	24	21	28	15
Largest contig [bp]	388,275	289,023	285,731	190,024	230,466	189,990	141,587	178,692	288,931	251,187	215,611	209,025	235,864	387,972
Contigs number	548	129	962	338	152	280	278	202	407	151	334	177	293	118
Genome size [bp]	4,976,326	4,996,403	5,562,509	5,393,782	4,228,584	4,482,959	4,208,498	4,319,217	4,227,546	5,138,050	5,883,543	5,364,035	6,025,652	4,786,256
G + C content	62.36%	62.45%	67.31%	67.70%	62.88%	65.72%	65.23%	65.57%	65.33%	64.22%	65.98%	65.73%	65.93%	66.66%
Accession Number	WTVJ00000000	WTVJ00000000	WTVL00000000	WUEI00000000	WTVM00000000	WTVG00000000	WTVQ00000000	WTVP00000000	WTVH00000000	WTVQ00000000	WTVK00000000	WTVR00000000	WTVS00000000	WTVN00000000

REFERENCE: Raittz *et al.*, 2021.

TABLE S3.4. Accession numbers of sequences used in the analysis of *nif* and other functional genes

Strain	GenBank accession	Genome Assembly
<i>Acidithiobacillus ferrivorans</i> SS3	GCA_000214095.3	ASM21409v3
<i>Allochromatium vinosum</i> ATCC 17899	GCA_000025485.1	ASM2548v1
<i>Anabaena cylindrica</i> ATCC 27899	GCA_000317695.1	ASM31769v1
<i>Anabaena</i> sp. 90	GCA_000312705.1	ASM31270v1
<i>Arcobacter nitrofigilis</i> DSM 7299	GCA_000092245.1	ASM9224v1
<i>Aromatoleum</i> sp. CIB	GCA_001190925.1	ASM119092v1
<i>Asaia bogorensis</i> MTCC 4041T	GCA_001547995.1	ASM154799v1
<i>Atelocyanobacterium thalassa</i> ALOHA	GCA_000025125.1	ASM2512v1
<i>Azorhizobium caulinodans</i> ORS 571	GCA_000010525.1	ASM1052v1
<i>Azospirillum brasilense</i> Sp7	GCA_001315015.1	ASM131501v1
<i>Azospirillum lipoferum</i> 4B	GCA_000283655.1	ASM28365v1
<i>Azospirillum</i> sp. B510	GCA_000010725.1	ASM1072v1
<i>Azospirillum thiophilum</i> BV-S	GCA_001305595.1	ASM130559v1
<i>Azotobacter chroococcum</i> NCIMB 8003	GCA_000817975.1	ASM81797v1
<i>Azotobacter vinelandii</i> DJ	GCA_000021045.1	ASM2104v1
<i>Beggiatoa leptomitiformis</i> D-402	GCA_001305575.2	ASM130557v2
<i>Beijerinckia indica</i> ATCC 9039	GCA_000019845.1	ASM1984v1
<i>Bradyrhizobium diazoefficiens</i> USDA 110	GCA_000011365.1	ASM1136v1
<i>Bradyrhizobium oligotrophicum</i> S58	GCA_000344805.1	ASM34480v1
<i>Bradyrhizobium</i> sp. BTAi1	GCA_000015165.1	ASM1516v1
<i>Bradyrhizobium</i> sp. ORS 278	GCA_000026145.1	ASM2614v1
<i>Burkholderia vietnamiensis</i> G4	GCA_000016205.1	ASM1620v1
<i>Burkholderia vietnamiensis</i> LMG 10929	GCA_000959445.1	ASM95944v1
<i>Calothrix</i> sp. 336/3	GCA_000734895.2	ASM73489v2
<i>Calothrix</i> sp. PCC 6303	GCA_000317435.1	ASM31743v1
<i>Candidatus Atelocyanobacterium thalassa</i> ALOHA	GCA_000025125.1	ASM2512v1
<i>Chlorobaculum parvum</i> NCIB 8327	GCA_000020505.1	ASM2050v1
<i>Chlorobaculum tepidum</i> ATCC 49652	GCA_000006985.1	ASM698v1
<i>Chlorobium chlorochromatii</i> CaD3	GCA_000012585.1	ASM1258v1
<i>Chlorobium limicola</i> DSM 245	GCA_000020465.1	ASM2046v1
<i>Chlorobium luteolum</i> DSM 273	GCA_000012485.1	ASM1248v1
<i>Chlorobium phaeobacteroides</i> BS1	GCA_000020545.1	ASM2054v1
<i>Chlorobium tepidum</i> TLS	GCA_000006985.1	ASM698v1
<i>Chloroherpeton thalassium</i> ATCC 35110	GCA_000020525.1	ASM2052v1

(Continue)

<i>Chroococcidiopsis thermalis</i> PCC 7203	GCA_000317125.1	ASM31712v1
<i>Clostridium aceticum</i> DSM 1496	GCA_001042715.1	ASM104271v1
<i>Clostridium acetobutylicum</i> ATCC 824	GCA_000008765.1	ASM876v1
<i>Clostridium beijerinckii</i> NCIMB 14988	GCA_000833105.2	ASM83310v2
<i>Clostridium beijerinckii</i> NCIMB 8052	GCA_000016965.1	ASM1696v1
<i>Clostridium kluyveri</i> DSM 555	GCA_000016505.1	ASM1650v1
<i>Clostridium tyrobutyricum</i> KCTC 5387	GCA_001642655.1	ASM164265v1
<i>Cyanothece</i> sp. ATCC 51142	GCA_000017845.1	ASM1784v1
<i>Cyanothece</i> sp. PCC 7424	GCA_000021825.1	ASM2182v1
<i>Cyanothece</i> sp. PCC 7425	GCA_000022045.1	ASM2204v1
<i>Cyanothece</i> sp. PCC 7822	GCA_000147335.1	ASM14733v1
<i>Cyanothece</i> sp. PCC 8801	GCA_000021805.1	ASM2180v1
<i>Cyanothece</i> sp. PCC 8802	GCA_000024045.1	ASM2404v1
<i>Dehalococcoides mccartyi</i> 195	GCA_000011905.1	ASM1190v1
<i>Desulfitobacterium hafniense</i> DCB-2	GCA_000021925.1	ASM2192v1
<i>Desulfomicrobium baculatum</i> DSM 4028	GCA_000023225.1	ASM2322v1
<i>Desulfosporosinus orientis</i> DSM 765	GCA_000235605.1	ASM23560v1
<i>Desulfotomaculum nigrificans</i> CO-1-SRB	GCA_000214435.1	ASM21443v1
<i>Desulfotomaculum ruminis</i> DSM 2154	GCA_000215085.1	ASM21508v1
<i>Desulfovibrio africanus</i> Walvis Bay	GCA_000212675.2	ASM21267v2
<i>Desulfovibrio vulgaris</i> DP4	GCA_000015485.1	ASM1548v1
<i>Ectothiorhodospira</i> sp. BSL-9	GCA_001632845.1	ASM163284v1
<i>Ensifer sojae</i> CCBAU 05684	GCA_002288525.1	ASM228852v1
<i>Enterobacter</i> sp. R4-368	GCA_000410515.1	ASM41051v1
<i>Frankia alni</i> ACN14A	GCA_000058485.1	ASM5848v1
<i>Frankia casuarinae</i> Ccl3	GCA_000013345.1	ASM1334v1
<i>Geobacter lovleyi</i> SZ	GCA_000020385.1	ASM2038v1
<i>Geobacter metallireducens</i> GS-15	GCA_000012925.1	ASM1292v1
<i>Geobacter sulfurreducens</i> PCA	GCA_000007985.2	ASM798v2
<i>Geobacter uraniireducens</i> Rf4	GCA_000016745.1	ASM1674v1
<i>Gluconacetobacter diazotrophicus</i> PA1 5	GCA_000021325.1	ASM2132v1
<i>Halorhodospira halophila</i> DSM 244	GCA_000015585.1	ASM1558v1
<i>Heliobacterium modesticaldum</i> ATCC 51547	GCA_000019165.1	ASM1916v1
<i>Herbaspirillum rubrisubalbicans</i> M1	GCA_001483945.1	ASM148394v1
<i>Herbaspirillum seropedicae</i> SmR1	GCA_000143225.1	ASM14322v1
<i>Herbaspirillum seropedicae</i> Z67	GCA_001040945.1	ASM104094v1
<i>Klebsiella oxytoca</i> CAV1374	GCA_001022195.1	ASM102219v1

(Continue)

<i>Klebsiella pneumoniae</i> 342	GCA_000019565.1	ASM1956v1
<i>Klebsiella variicola</i> At-22	GCA_000025465.1	ASM2546v1
<i>Leptolyngbya boryana</i> dg5	GCA_002142495.1	ASM214249v1
<i>Leptolyngbya</i> sp. O-77	GCA_001548395.1	ASM154839v1
<i>Magnetospirillum magneticum</i> AMB 1	GCA_000009985.1	ASM998v1
<i>Mesorhizobium ciceri</i> WSM1271	GCA_000185905.1	ASM18590v1
<i>Mesorhizobium loti</i> NZP2037	GCA_001676765.1	ASM167676v1
<i>Mesorhizobium opportunistum</i> WSM2075	GCA_000176035.2	ASM17603v2
<i>Methylobacterium nodulans</i> ORS 2060	GCA_000022085.1	ASM2208v1
<i>Methylobacterium</i> sp. 4-46	GCA_000019365.1	ASM1936v1
<i>Methylocella silvestris</i> DSM 15510	GCA_000021745.1	ASM2174v1
<i>Methylococcus capsulatus</i> Bath	GCA_000008325.1	ASM832v1
<i>Methylocystis</i> sp. SC2	GCA_000304315.1	Msp_SC2
<i>Methylomonas methanica</i> MC09	GCA_000214665.1	ASM21466v1
<i>Methylosinus trichosporium</i> OB3b	GCA_002752655.1	ASM275265v1
<i>Neorhizobium galegae</i> HAMBI 1141	GCA_000731295.1	RG1141_Ch_pA_pB
<i>Neorhizobium galegae</i> HAMBI 540	GCA_000731315.1	RG540_Ch_pA
<i>Nitrospirillum amazonense</i> CBAmc	GCA_000730165.2	ASM73016v2
<i>Niveispirillum cyanobacteriorum</i>	GCA_002868735.1	ASM286873v1
<i>Nostoc flagelliforme</i> CCNUN1	GCA_002813575.1	ASM281357v1
<i>Nostoc punctiforme</i> ATCC 29133	GCA_000020025.1	ASM2002v1
<i>Nostoc</i> sp. ATCC 29411	GCA_000316645.1	ASM31664v1
<i>Nostoc</i> sp. PCC 7120	GCA_000009705.1	ASM970v1
<i>Nostoc</i> sp. CENA543	GCA_002896875.1	ASM289687v1
<i>Nostoc</i> sp. PCC 7107	GCA_000316625.1	ASM31662v1
<i>Nostoc</i> sp. N6	GCA_002949735.1	ASM294973v1
<i>Nostocales cyanobacterium</i> HT-58-2	GCA_002163975.1	ASM216397v1
<i>Paenibacillus durus</i> DSM 1735	GCA_000756615.1	ASM75661v1
<i>Paenibacillus polymyxa</i> sb3-1	GCA_000819665.1	ASM81966v1
<i>Paenibacillus</i> sp. FSL H7-0357	GCA_000758525.1	ASM75852v1
<i>Pantoea</i> At-9b	GCA_000175935.2	ASM17593v2
<i>Paraburkholderia phymatum</i> STM815	GCA_000020045.1	ASM2004v1
<i>Paraburkholderia xenovorans</i> LB400	GCA_000013645.1	ASM1364v1
<i>Pelobacter carbinolicus</i> DSM 2380	GCA_000012885.1	ASM1288v1
<i>Pelobacter propionicus</i> DSM 2379	GCA_000015045.1	ASM1504v1
<i>Pleurocapsa</i> sp. PCC 7327	GCA_000317025.1	ASM31702v1
<i>Prosthecochloris aestuarii</i> DSM 271	GCA_000020625.1	ASM2062v1

(Continue)

<i>Pseudomonas stutzeri</i> A1501	GCA_000013785.1	ASM1378v1
<i>Rhizobium etli</i> CFN 42	GCA_000092045.1	ASM9204v1
<i>Rhizobium gallicum</i> R602	GCA_000816845.1	ASM81684v1
<i>Rhizobium leguminosarum</i> WSM1325	GCA_000023185.1	ASM2318v1
<i>Rhizobium leguminosarum</i> 3841	GCA_000009265.1	ASM926v1
<i>Rhizobium tropici</i> CIAT 899	GCA_000330885.1	ASM33088v1
<i>Rhodobacter capsulatus</i> ATCC BAA-309	GCA_000021865.1	ASM2186v1
<i>Rhodobacter sphaeroides</i> ATCC 17029	GCA_000015985.1	ASM1598v1
<i>Rhodobacter sphaeroides</i> KD131	GCA_000021005.1	ASM2100v1
<i>Rhodomicrobium vannielii</i> ATCC 17100	GCA_000166055.1	ASM16605v1
<i>Rhodopseudomonas palustris</i> ATCC BAA-98	GCA_000195775.1	ASM19577v1
<i>Rhodopseudomonas palustris</i> BisA53	GCA_000014825.1	ASM1482v1
<i>Rhodopseudomonas palustris</i> BisB18	GCA_000013745.1	ASM1374v1
<i>Rhodopseudomonas palustris</i> BisB5	GCA_000013685.1	ASM1368v1
<i>Rhodopseudomonas palustris</i> DX-1	GCA_000177255.2	ASM17725v2
<i>Rhodopseudomonas palustris</i> HaA2	GCA_000013365.1	ASM1336v1
<i>Rhodopseudomonas palustris</i> TIE-1	GCA_000020445.1	ASM2044v1
<i>Rhodospirillum centenum</i> ATCC 51521	GCA_000016185.1	ASM1618v1
<i>Rhodospirillum rubrum</i> ATCC 11170	GCA_000013085.1	ASM1308v1
<i>Rhodovulum</i> sp. P5	GCA_002079305.1	ASM207930v1
<i>Rhodovulum sulfidophilum</i>	GCA_001633145.1	ASM163314v1
<i>Serratia</i> sp. ATCC 39006	GCA_000463345.3	ASM46334v3
<i>Sinorhizobium fredii</i> CCBAU 83666	GCA_002288485.1	ASM228848v1
<i>Sinorhizobium fredii</i> NGR234	GCA_000018545.1	ASM1854v1
<i>Sinorhizobium fredii</i> NXT3	GCA_002944405.1	ASM294440v1
<i>Sinorhizobium medicae</i> WSM419	GCA_000017145.1	ASM1714v1
<i>Sinorhizobium meliloti</i> 1021	GCA_000006965.1	ASM696v1
<i>Synechococcus</i> sp. JA-2-3Ba(2-13)	GCA_000013225.1	ASM1322v1
<i>Synechococcus</i> sp. JA-3-3Ab	GCA_000013205.1	ASM1320v1
<i>Teredinibacter turnerae</i> T7901	GCA_000023025.1	ASM2302v1
<i>Thiocystis violascens</i> DSM 198	GCA_000227745.3	ASM22774v3
<i>Trichodesmium erythraeum</i> IMS101	GCA_000014265.1	ASM1426v1
<i>Trichormus azollae</i> 0708	GCA_000196515.1	ASM19651v1
<i>Trichormus variabilis</i> ATCC 29413	GCA_000204075.1	ASM20407v1
<i>Vibrio</i> sp. EJY3	GCA_000241385.1	ASM24138v1
<i>Xanthobacter autotrophicus</i> Py2	GCA_000017645.1	ASM1764v1

REFERENCE: Raittz *et al.*, 2021.

TABLE S4.4. ANI percentage values for the 32 genomes of *Azoarcus* and *Aromatoleum* strains

<i>Aromatoleum anaerobium</i> LuFRes1T	100	92.37	92.16	94.33	92.26	81.93	82.23	81.83	82.26	82.63	91.55	82.39	82.44	82.53	82.58	78.42	78.53	78.61	78.92	78.99	78.95	80.00	79.12	80.12	78.61	79.64	79.63	79.75	81.46	78.74	78.59	78.86
<i>Aromatoleum aromaticum</i> EBN1T	92.37	100	97.84	92.30	92.78	81.43	81.82	81.66	82.02	82.37	91.89	82.10	82.42	82.38	82.44	78.88	78.46	78.39	78.85	78.93	79.01	79.89	79.02	79.82	78.64	79.43	79.53	79.85	81.36	78.72	78.48	78.76
<i>Aromatoleum aromaticum</i> pCYN1	92.16	97.84	100	92.43	92.80	81.45	81.93	81.86	82.10	82.39	91.64	82.11	82.16	82.08	82.39	79.39	78.41	78.46	78.70	79.14	79.10	79.90	78.86	79.78	78.69	79.40	79.41	79.89	81.32	78.72	79.80	78.79
<i>Aromatoleum bremense</i> pBN1T	94.33	92.20	92.43	100	91.98	82.04	82.12	81.86	82.24	82.45	91.62	82.30	82.54	82.43	82.61	78.99	78.53	78.57	79.05	79.05	79.11	80.25	78.86	80.24	78.82	79.62	79.68	80.04	81.50	78.80	79.14	78.99
<i>Aromatoleum bucheii</i> U120T	92.26	92.78	92.80	91.98	100	81.63	82.19	81.85	82.26	82.49	92.74	81.94	82.24	82.13	82.44	78.34	78.28	78.31	78.79	78.76	78.69	79.81	79.10	79.85	78.47	79.43	79.48	79.59	81.14	78.67	78.25	78.68
<i>Aromatoleum diolicum</i> 221.m2T	81.93	81.43	81.45	82.04	81.63	100	83.17	82.72	83.58	83.02	81.73	83.25	83.79	84.34	83.32	78.93	78.85	78.93	79.29	79.11	79.15	80.22	79.07	80.16	78.97	79.47	79.68	80.29	82.39	79.10	78.48	79.21
<i>Aromatoleum evansii</i> KB740T	82.23	81.82	81.93	82.12	82.19	83.17	100	91.17	88.90	87.14	82.05	89.04	89.24	89.12	87.24	78.84	78.98	79.10	79.63	79.67	79.67	80.70	79.38	80.73	79.39	80.36	80.52	80.83	83.33	79.38	78.58	79.53
<i>Aromatoleum petrolei</i> Tol1T	81.83	81.66	81.86	81.86	81.85	82.72	91.17	100	88.71	86.74	81.76	88.80	88.59	88.67	87.09	78.72	78.79	78.79	79.32	79.44	79.40	80.50	79.27	80.49	79.19	80.28	80.22	80.37	83.03	79.23	78.52	79.38
<i>Aromatoleum</i> sp. CIB	82.26	82.02	82.10	82.24	82.26	83.58	88.90	88.71	100	88.17	82.28	90.07	92.34	93.30	87.68	79.69	78.87	78.92	79.60	79.64	79.61	80.60	79.34	80.63	79.33	80.27	80.29	80.64	83.57	79.30	78.82	79.40
<i>Aromatoleum</i> sp. DN11	82.63	82.37	82.39	82.45	82.49	83.02	87.14	86.74	88.17	100	82.43	87.69	87.26	87.76	93.10	79.92	78.98	79.09	79.71	79.91	79.95	80.94	79.49	80.89	79.55	80.56	80.60	80.94	83.60	79.29	79.16	79.50
<i>Aromatoleum</i> sp. PA01	91.55	91.89	91.64	91.62	92.74	81.73	82.05	81.76	82.28	82.43	100	82.10	82.29	82.14	82.65	78.41	78.24	78.27	78.79	78.72	78.74	79.85	78.90	79.72	78.55	79.54	79.65	79.47	81.26	78.69	78.20	78.73
<i>Aromatoleum toluclasticum</i> MF63T	82.39	82.10	82.11	82.30	81.94	83.25	89.04	88.80	90.07	87.69	82.10	100	90.15	89.53	87.92	79.10	78.78	78.86	79.68	79.59	79.62	80.60	79.41	80.63	79.30	80.30	80.37	80.77	83.24	79.21	78.97	79.29
<i>Aromatoleum tolucliticum</i> Tol.4T	82.44	82.42	82.16	82.54	82.24	83.79	89.24	88.59	92.34	87.26	82.29	90.15	100	93.00	87.52	79.86	78.71	78.73	79.70	79.61	79.75	80.66	79.67	80.74	79.26	80.50	80.44	80.81	83.68	79.34	78.94	79.32
<i>Aromatoleum toluolicum</i> T	82.53	82.28	82.08	82.43	82.13	84.34	89.12	88.67	93.30	87.76	82.14	89.53	93.00	100	87.44	79.73	78.82	78.82	79.69	79.69	79.67	80.57	79.37	80.61	79.28	80.36	80.41	80.83	83.86	79.45	78.97	79.36
<i>Aromatoleum toluovans</i> Td21T	82.58	82.44	82.39	82.61	82.44	83.32	87.24	87.09	87.68	93.10	82.65	87.92	87.52	87.44	100	79.05	79.01	79.15	79.93	79.92	80.01	81.09	79.69	81.17	79.30	80.61	80.72	80.78	83.86	79.37	78.90	79.46
<i>Azoarcus communis</i> s LMG5514	78.42	78.88	79.39	78.99	78.34	78.93	78.84	78.72	79.69	79.92	78.41	79.10	79.86	79.73	79.05	100	98.06	98.21	79.71	79.70	79.63	80.27	78.22	80.42	80.43	79.64	79.73	80.97	78.75	80.41	79.38	80.55
<i>Azoarcus communis</i> s SWub3T_1	78.53	78.46	78.41	78.53	78.28	78.85	78.98	78.79	78.87	78.98	78.24	78.78	78.71	78.82	79.01	98.06	100	99.34	79.73	79.79	79.65	80.29	78.25	80.52	80.07	79.57	79.62	80.40	78.75	80.50	78.47	80.60
<i>Azoarcus communis</i> s SWub3T_2	78.61	78.39	78.46	78.57	78.31	78.93	79.10	78.79	78.92	79.09	78.27	78.86	78.73	78.82	79.15	98.21	99.34	100	79.78	79.87	79.80	80.42	78.25	80.68	80.22	79.70	79.58	80.47	78.67	80.52	78.45	80.55
<i>Azoarcus</i> sp. TTM-91	78.92	78.85	78.70	79.05	78.79	79.29	79.63	79.32	79.60	79.71	78.79	79.68	79.70	79.69	79.93	79.71	79.73	79.78	100	93.60	93.57	82.50	78.89	82.53	79.42	81.26	81.51	82.68	79.13	79.58	78.61	79.62
<i>Azoarcus indigenis</i> VB32T_1	78.99	78.93	79.14	79.05	78.76	79.11	79.67	79.44	79.64	79.91	78.72	79.59	79.61	79.69	79.92	79.70	79.79	79.87	93.60	100	99.33	82.55	78.77	82.60	79.64	81.26	81.59	82.94	79.07	79.63	78.71	79.72
<i>Azoarcus indigenis</i> VB32T_2	78.95	79.01	79.10	79.11	78.69	79.15	79.67	79.40	79.61	79.95	78.74	79.62	79.75	79.67	80.01	79.63	79.65	79.80	93.57	99.33	100	82.57	78.70	82.63	79.67	81.32	81.58	82.88	79.12	79.57	78.71	79.73
<i>Azoarcus olearius</i> DQ5-4T	80.00	79.89	79.90	80.25	79.81	80.22	80.70	80.50	80.60	80.94	79.85	80.60	80.66	80.57	81.09	80.27	80.29	80.42	82.50	82.55	82.57	100	79.50	98.77	80.34	81.55	81.76	96.59	80.05	80.62	78.89	80.59
<i>Azoarcus pumilus</i> SY39T	79.12	79.02	78.86	78.86	79.10	79.07	79.38	79.27	79.34	79.49	78.90	79.41	79.67	79.37	79.69	78.22	78.25	78.25	78.89	78.77	78.70	79.50	100	79.47	78.47	79.95	79.90	79.19	78.85	78.43	78.84	78.58
<i>Azoarcus olearius</i> BH72	80.12	79.82	79.78	80.24	79.85	80.16	80.73	80.49	80.63	80.89	79.72	80.63	80.74	80.61	81.17	80.42	80.52	80.68	82.53	82.60	82.63	98.77	79.47	100	80.34	81.47	81.74	86.62	80.09	80.62	78.83	80.62
<i>Azoarcus</i> sp. BM101	78.61	78.64	78.69	78.82	78.47	78.97	79.39	79.19	79.33	79.55	78.55	79.30	79.26	79.28	79.30	80.43	80.07	80.22	79.42	79.64	79.67	80.34	79.47	100	79.74	79.76	80.50	78.87	93.36	78.45	96.98	
<i>Azoarcus</i> sp. CCYHH838	79.64	79.43	79.40	79.62	79.43	79.47	79.07	80.36	80.28	80.27	80.56	79.54	80.30	80.50	80.36	80.61	79.64	79.57	79.70	81.26	81.32	81.55	79.95	81.47	79.74	100	85.30	81.35	79.49	79.58	79.13	79.74
<i>Azoarcus</i> sp. CCYHH848	79.63	79.53	79.41	79.68	79.48	79.68	80.52	80.22	80.29	80.60	79.65	80.37	80.44	80.41	80.72	79.73	79.62	79.58	81.51	81.59	81.58	81.76	79.90	81.74	79.76	85.30	100	81.74	79.63	79.72	78.88	79.87
<i>Azoarcus</i> sp. DD4	79.75	79.85	79.89	80.04	79.59	80.29	80.83	80.37	80.64	80.94	79.47	80.77	80.81	80.83	80.78	80.97	80.40	80.47	82.68	82.94	82.88	86.59	79.19	86.62	80.50	81.35	81.74	100	80.21	80.70	79.57	80.70
<i>Azoarcus</i> sp. KH32C	81.46	81.36	81.32	81.50	81.14	82.39	83.33	83.03	83.57	83.60	81.26	83.24	83.68	83.86	83.86	78.75	78.75	78.67	79.13	79.07	79.12	80.05	78.85	80.09	78.87	79.49	79.63	80.21	100	79.18	78.52	79.01
<i>Azoarcus</i> sp. TSN442	78.74	78.72	78.72	78.80	78.67	79.10	79.38	79.23	79.30	79.29	78.69	79.21	79.34	79.45	79.37	80.41	80.50	80.52	79.58	79.63	79.57	80.62	78.43	80.62	93.36	79.58	79.72	80.70	79.18	100	78.73	94.32
<i>Azoarcus taiwanensis</i> s NSC3T	78.59	78.48	79.80	79.14	78.25	78.48	78.58	78.52	78.82	79.16	78.20	78.97	78.94	78.97	78.90	79.38	78.47	78.45	78.61	78.71	78.89	78.84	78.83	78.45	79.13	78.88	79.57	78.52	78.73	100	78.41	
<i>Azoarcus</i> sp. TSPY31	78.86	78.76	78.79	78.99	78.68	79.21	79.53	79.38	79.40	79.50	78.73	79.29	79.32	79.36	79.46	80.55	80.60	80.55	79.62	79.72	79.73	80.59	78.58	80.62	96.98	79.74	79.87	80.70	79.01	94.32	78.41	100

REFERENCE: Raittz et al., 2021

Legend: Analogous relationships between *Aromatoleum-Azospira* and *Azoarcus-Dechloromonas* wereobserved, confirming the *nif H* gene analysis.

TABLE S5.4. Percentage similarity values between 16s rRNA sequences extracted from genomes of *Azoarcus* and *Aromatoleum* strains

	<i>Azoarcus punilus</i> SY391	<i>Azoarcus</i> sp. CC-YHH838	<i>Azoarcus</i> sp. CC-YHH848	<i>Azoarcus</i> sp. TTM-91	<i>Azoarcus indigenus</i> VB32T_1	<i>Azoarcus indigenus</i> VB32T_2	<i>Azoarcus olearius</i> BH72	<i>Azoarcus olearius</i> DQS-4T	<i>Azoarcus communis</i> LMG5514	<i>Azoarcus communis</i> SWub3T_1	<i>Azoarcus communis</i> SWub3T_2	<i>Azoarcus communis</i> TSPY31	<i>Azoarcus communis</i> TSNM42	<i>Azoarcus</i> sp. BM101	<i>Azoarcus taiwanensis</i> NSC3T	<i>Aromatoleum dialicum</i> 22Lm2T	<i>Aromatoleum toluyliticum</i> Tol-4T	<i>Aromatoleum toluclasticum</i> MF63	<i>Azoarcus</i> sp. KH32C	<i>Aromatoleum tololicum</i> T	<i>Aromatoleum</i> sp. CIB	<i>Aromatoleum</i> sp. DN11	<i>Aromatoleum toluovorans</i> Td21T	<i>Aromatoleum evansii</i> KB740T	<i>Aromatoleum petrolei</i> Ton1T	<i>Aromatoleum</i> sp. PA01	<i>Aromatoleum bremense</i> PbN1T	<i>Aromatoleum anaerobium</i> LuFRes1T	<i>Aromatoleum buckelii</i> U120T	<i>Aromatoleum aromaticum</i> pCyN1	<i>Aromatoleum aromaticum</i> Ebn1	
	100	93.8	93.2	92.1	92.1	92.1	91.8	92.3	92.3	91.7	91.7	91.7	92.7	92.8	92.8	95	93.2	92.8	92.2	92.8	93	93.1	92.8	92.9	92.9	92.9	93.2	93.3	93.3	93.4	93.3	93.2
	93.8	100	96.8	95.7	95.7	95.7	93.8	94.4	94.4	94	94.1	94.1	94.8	94.8	94.7	94.4	95.3	95.3	94.1	95.1	95.3	95.3	95.4	95.7	95.5	95.5	95.1	95.1	94.7	94.9	95.1	95
	93.2	96.8	100	94.8	94.9	94.9	93.5	94	94.3	93.8	93.8	93.8	92.9	92.9	92.8	93.5	95.1	95	94.5	95	95.5	95.2	95.5	95.3	95.1	95.7	95.5	95.7	95.9	96.4	96.3	
	92.1	95.7	94.8	100	99.6	99.6	97	97.4	97.3	95.3	95.3	95.3	95.1	95.1	95	93.3	95.4	94.1	95.6	95.4	95.6	95.5	95.1	95.4	95.5	95.1	95.1	94.8	95.1	95		
	92.1	95.7	94.9	99.6	100	100	97	97.4	97.3	95.1	95.2	95.2	95.1	95.1	95	93.2	95.3	94.2	94.2	95.7	95.3	95.4	95.5	95.2	95.4	95.5	95.3	94.9	95.3	95.1	95.1	
	91.8	93.8	93.5	97	97	97	100	99.4	99.1	96	96.1	96.1	95.9	95.8	95.8	93.1	95.6	95.6	95.3	95.2	95.3	95.5	95.2	94.9	95.1	95.1	94.7	94.7	94.5	94.8	94.6	
	92.3	94.4	94	97.4	97.4	97.4	99.4	100	99.7	96.6	96.7	96.7	96.4	96.3	96.3	93.7	96.2	96.2	95.9	95.8	95.9	96	95.8	95.5	95.7	95.8	95.3	95.3	95.1	95.4	95.1	
	92.3	94.4	94.3	97.3	97.3	97.3	99.1	99.7	100	96.6	96.7	96.7	96.3	96.3	96.3	93.7	96.4	96.3	95.2	95.9	96	96.1	96.2	95.7	95.8	95.5	95.5	95.6	95.9	95.7	95.6	
	91.7	94	93.8	95.3	95.1	95.1	96	96.6	96.6	100	99.9	99.9	97.3	97.2	97.3	93.5	95.7	95.7	94.6	95.3	95.3	95.3	95.3	95.3	95.4	95.3	95.6	95.5	95	95.4	95.1	
	91.7	94.1	93.8	95.3	95.2	95.2	96.1	96.7	96.7	99.9	100	100	97.3	97.3	97.3	93.5	95.8	95.7	94.7	95.3	95.4	95.3	95.4	95.4	95.5	95.4	95.7	95.6	95.1	95.5	95.2	
	91.7	94.1	93.8	95.3	95.2	95.2	96.1	96.7	96.7	99.9	100	100	97.3	97.3	97.3	93.5	95.8	95.7	94.7	95.3	95.4	95.3	95.4	95.4	95.5	95.4	95.7	95.6	95.1	95.5	95.2	
	92.7	94.8	92.9	95.1	95.1	95.1	95.9	96.4	96.3	97.3	97.3	97.3	100	99.9	100	93.3	95.7	95.2	94.2	95.3	95.1	95.3	94.9	94.9	95.1	94.7	95	94.6	94.7	94.3	94.3	
	92.8	94.8	92.9	95.1	95.1	95.1	95.8	96.3	96.3	97.2	97.3	97.3	99.9	100	99.9	93.3	95.7	95.3	94.2	95.4	95.1	95.4	94.9	95	95.1	94.6	95	94.6	94.8	94.4	94.4	
	92.8	94.7	92.8	95	95	95	95.8	96.3	96.3	97.3	97.3	97.3	100	99.9	100	93.2	95.6	95.2	94.1	95.3	95	95.3	94.8	94.9	95	95.1	94.6	95	94.5	94.6	94.3	
	95	94.4	95.5	93.3	93.2	93.2	93.1	93.7	93.7	93.5	93.5	93.5	93.3	93.3	93.2	100	95.1	95.2	94.6	95	95.3	95.1	95	95.5	95.5	95.3	94.6	94.7	94.4	94.7	94.7	
	93.2	95.3	95.1	95.4	95.3	95.3	95.6	96.2	96.4	95.7	95.8	95.8	95.7	95.7	95.6	95.1	100	97.8	96.9	98.3	98.3	98.3	98.1	98.3	98.3	98.2	97.7	97.9	97.6	97.6	97.2	
	92.8	95.3	95	95.4	95.5	95.5	95.6	96.2	96.3	95.7	95.7	95.7	95.2	95.3	95.2	95.2	97.8	100	98.7	97.9	99	99	98.7	98.7	98.9	99	97.5	97.4	97.3	97.5	97.3	
	92.2	94.1	94.5	94.1	94.2	94.2	95.3	95.9	95.2	94.6	94.7	94.7	94.2	94.2	94.1	94.6	96.9	98.7	100	96.8	98	97.9	97.6	97.7	97.9	97.9	96.9	96.9	96.7	97	96.9	
	92.8	95.1	95	95.6	95.7	95.7	95.2	95.8	95.9	95.3	95.3	95.3	95.3	95.3	95.3	95	98.3	97.9	96.8	100	98.6	98.6	98.6	98.8	98.8	96.9	97.1	97.1	96.8	96.3	96.3	
	92.9	95.3	95.5	95.4	95.3	95.3	95.3	95.9	96	95.3	95.4	95.4	95.1	95.1	95	95.3	98.3	99	98	98.6	100	99.7	99.2	99.6	99.7	99.5	97.1	96.9	97	97	96.7	
	93	95.3	95.2	95.6	95.4	95.4	95.5	96	96.1	95.3	95.4	95.4	95.3	95.4	95.3	95.1	98.3	99	97.9	98.6	99.7	100	99.2	99.3	99.4	99.5	96.8	96.9	96.8	96.7	96.4	
	93.1	95.4	95.5	95.5	95.5	95.5	95.2	95.8	96.2	95.3	95.3	95.3	94.9	94.9	94.8	95	98.1	98.7	97.6	98.6	99.2	99.2	100	99.4	99.6	99.6	96.8	96.7	97.1	97.1	96.8	
	92.8	95.7	95.5	95.1	95.2	95.2	94.9	95.5	95.7	95.3	95.4	95.4	94.9	95	94.9	95.5	98.3	98.7	97.7	98.6	99.6	99.3	99.4	100	99.7	99.6	97	96.8	96.7	96.8	96.6	
	92.9	95.5	95.3	95.4	95.4	95.4	95.1	95.7	95.8	95.4	95.5	95.5	95.1	95.1	95	95.5	98.3	98.9	97.9	98.8	99.7	99.4	99.6	99.7	100	99.8	97	96.9	96.8	96.9	96.6	
	92.9	95.5	95.1	95.5	95.5	95.5	95.1	95.8	95.9	95.3	95.4	95.4	95.1	95.1	95.1	95.3	98.2	99	97.9	98.8	99.5	99.5	99.6	99.6	99.8	100	96.8	96.8	96.8	96.7	96.4	
	93.2	95.1	95.7	95.1	95.3	95.3	94.7	95.3	95.5	95.6	95.7	95.7	94.7	94.7	94.6	94.6	97.7	97.5	96.9	96.9	97.1	96.8	96.8	97	97	96.8	100	98.9	98.7	99.3	98.8	
	93.3	95.1	95.5	95.1	95.3	95.3	94.7	95.3	95.5	95.5	95.6	95.6	95	95	94.7	94.7	97.9	97.4	96.9	97.1	96.9	96.9	96.7	96.8	96.9	96.8	98.9	100	98.9	98.8	98.3	
	93.3	94.7	95.7	94.8	94.9	94.9	94.5	95.1	95.6	95	95.1	95.1	94.6	94.6	94.5	94.4	97.6	97.3	96.7	97.1	97	96.8	97.1	96.7	96.8	96.8	98.7	98.9	100	99.2	98.6	
	93.4	94.9	95.9	95.1	95.3	95.3	94.8	95.4	95.9	95.4	95.5	94.7	94.8	94.6	94.7	94.7	97.6	97.5	97	96.8	97	96.7	97.1	96.8	96.9	96.7	99.3	98.8	99.2	100	99.2	
	93.3	95.1	96.4	95	95.1	95.1	94.6	95.1	95.7	95.1	95.2	95.2	94.3	94.4	94.3	94.7	97.2	97.3	96.9	96.3	96.7	96.4	96.8	96.6	96.6	96.4	98.8	98.3	98.6	99.2	100	
	93.2	95	96.3	95	95.1	95.1	94.6	95.1	95.6	95.1	95.1	94.3	94.4	94.2	94.6	94.6	97.2	97.2	96.9	96.3	96.7	96.4	96.8	96.5	96.6	96.4	98.8	98.3	98.6	99.2	100	

REFERENCE: Raittz et al., (2021).

FILE S1.4

In this study, not all results were performed on specific scripts. In some cases, the results were obtained step by step in the MATLAB® command window and in other data of interchange between the various applied software. Below, is included excerpts from scripts with details and the links to the tools used that can assist in understanding the methods.

Function to calculate the distance of SWEEP Vectors:

```
pdist()
```

pdist() is a standard function in MATLAB® that calculate the pairwise distance (by default Euclidean distances) between observations (vectors);

Excerpt of script to obtain neighbour joining:

```
treeout = seqneighjoin(xdist,'equivar',xnames);
```

where *xdist* is a *pdist* vector (MATLAB standard result from a *pdist()* function call) and *xname* is a cell var with the nodes names.

Normalization by the mean of ANI matrices:

Here, the normalization was performed for the upper and lower ANI matrices separately (see the script below). First, ANI was converted into dissimilarity matrices and then the lower and upper matrices were divided by their corresponding mean values. It was decided to normalize the mean to guarantee a better correspondence between the matrices in the visualization. This technique allowed us to compare different distance / dissimilarity matrices in the same heat map.

```
% ANI matrix MANI
```

```
% Transforms lower and upper ANI matrices in dissimilarity matrices
```

```
D1 = 100 - tril(MANI)+tril(MANI)'; D2 = 100 - triu(MANI)+triu(MANI)'
```

```
% iani1 - Order for Lower triangular part of the ANI matrix
```

```
[~ iani1] = HFclus(distmat2pdist(D1),'Method','Nj');
```

```
% iani2 - Order for upper triangular part of the ANI matrix
```

```
[~ iani2] = HFclus(distmat2pdist(D2),'Method','Nj');
% HFclus is an 'in house' function used for several purposes; here it receives a MATLAB pdist vector and
returns de order of the leaves in the corresponding neighbour joining phylogenetic tree (exploiting the
MATLAB bioinformatics toolbox functions)
% Normalization and matrix reordering for Heat map
Mx = tril(D1)/mean(D1(:))+triu(D2)/mean(D2(:));
Mx = Mx(iani1,iani2);
% Transforms matrix into image
xHeat = mat2im3(Mx);
```

Links to all the programs used in this study:

1. FastQC

<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

2. Trimmomatic

<http://www.usadellab.org/cms/?page=trimmomatic>

3. BBMap

<https://sourceforge.net/projects/bbmap/>

4. SPAdes

<https://cab.spbu.ru/software/spades/>

5. Prokka

<https://github.com/tseemann/prokka>

6. QUAST

<http://quast.sourceforge.net/>

7. FastANI

<https://github.com/ParBLiSS/FastANI>

8. RAFTS3G

<https://sourceforge.net/projects/rafts-g/>

9. SWeeP

<https://sourceforge.net/projects/spacedwordsprojection/>

10. Dendroscope

<https://software-ab.informatik.unituebingen.de/download/dendroscope3/welcome.html>

11. Clustal Omega

<https://www.ebi.ac.uk/Tools/msa/clustalo/>

12. IQ-Tree

<http://www.iqtree.org/>

FILE S2.4

Growth conditions for assessment of nitrogenase activity using the acetylene reduction assay. *Azoarcus olearius* DQS-4 T, *Az. communis* Swub3T, *Az. communis* LMG 5514 (LMG), *Az. indigens* VB32T, and *Azoarcus* sp. TTM-91 were grown on SM or CVM media, prepared following the composition described in [1]. The SM medium was supplemented with a vitamin solution (1 mg L⁻¹ thiamine hydrochloride, 2 mg L⁻¹ D-pantothenic acid, 1 mg L⁻¹ nicotinic acid, and 0.1 mg L⁻¹ biotin) modified from UMS medium [2]. Semi solid medium (SM or ½CVM) was prepared at the moment of use by mixing pre-warmed solid (1.5% agar) and liquid media to reach a final concentration of 0.18% agar. ½CVM is CVM [1] but with half of the amount of bactopectone and yeast extract i.e. 1.5 and 0.5 g L⁻¹, respectively. *Aromatoleum* sp. CIB and *Ar. toluovorans* Td21T were grown on MNF medium [3] supplemented with 3 mM sodium benzoate as carbon source.

Prior to the nitrogenase activity assay, *Az. olearius*, DQS-4 T, *Az. communis* Swub3T, *Az. communis* LMG 5514 (LMG), *Az. indigens* VB32T, and *Azoarcus* sp. TTM-91 were streaked from glycerol stocks into CVM amended with 20 mM NH₄Cl (CVM+N) and incubated at 30°C for 3–5 days. Isolated colonies were then inoculated into 3 mL of liquid CVM+N (in 30 mL plastic universal tubes) and incubated at 30°C and 120 rpm. Once the culture reached saturation, a further inoculation (1/100) in liquid medium was performed before the inoculation into semisolid medium (SM amended with 0.5 mM sodium glutamate for *Az. olearius*, DQS-4 T, *Az. communis* Swub3T, and *Az. communis* LMG 5514, or ½CVM for *Az. indigens* VB32T, and *Azoarcus* sp. TTM-91). The semi-solid medium (3 mL) was set in a 7 mL glass vial (Fisherbrand™ #14873562). Liquid cultures (15–60 µL), pre-grown as above, were used for inoculation into the center-bottom of vials of semi-solid media which was aimed at achieving approximately the same number of cells (OD_{600nm} = 0.01 approximately) across the different inocula. Cells were incubated for 18–24 hours, or until the bacterial pellicle reached the meniscus. Once the pellicle was formed, the plastic cap from the vial was replaced by a suba-seal rubber stopper (n° 21) and the substrate, acetylene (10% of the gas phase), for the nitrogenase activity assay was injected. A similar preinoculation procedure was used to assay the nitrogenase activity in *Aromatoleum* sp. CIB and *Ar. toluovorans* Td21T except that MNF medium was supplemented with 3 mM sodium benzoate as carbon source and 5 mM KNO₃ as nitrogen source. The second re-inoculation into liquid medium was performed on a higher volume, 20 mL of medium in a 100 mL Erlenmeyer flask at 30°C and 120 rpm. Cells were then washed in MNF + 3 mM sodium benzoate to remove excess nitrate and then re-inoculated into fresh medium without nitrate to an OD_{600nm} of approximately 0.01. Ten mL cells (OD_{600nm} of 0.01), were transferred to 20 mL glass universal tubes, sealed with suba-seal rubber stopper n°41 and incubated for 24 hours at 30°C and 120 rpm before injection of the acetylene (10% of the gas phase). *Ar. diolicum* 22LinT was grown on UMS medium supplemented with 30 mM malate as carbon source and 20 mM NH₄Cl. Prior to the acetylene reduction assay, 50 mL

cells were collected by centrifugation (4000 rpm, 8 min) and re-suspended to an initial O.D600 of 0.4 (20 mL in an 100 mL Pyrex Erlenmeyer flask) in UMS plus 30 mM malate without added ammonium. The flask was stoppered with a suba-seal (nº 37) and the internal flask atmosphere was adjusted to approximately 1% O₂. After 6 hours incubation, the acetylene reduction assay was performed as above.

References

1. Krause, A.; Julich, H.; Mankar, M.; Reinhold-Hurek, B. The regulatory network controlling ethanol-induced expression of alcohol dehydrogenase in the endophyte *Azoarcus* sp. strain BH72. *Mol. Plant-Microbe Interact.* 2017, 30, 778-785, doi:MPMI-01-17-0013-R.
2. Pini, F.; East, A.K.; Appia-Ayme, C.; Tomek, J.; Karunakaran, R.; Mendoza-Suarez, M.; Edwards, A.; Terpolilli, J.; Rowoth, J.; Downie, J.A.; *et al.* Bacterial biosensors for in vivo spatiotemporal mapping of root secretion. *Plant Physiol.* 2017, 174, 1289-1306, doi:10.1104/pp.16.01302.
3. Fernández, H.; Prandoni, N.; Fernández-Pascual, M.; Fajardo, S.; Morcillo, C.; Díaz, E.; Carmona, M. *Azoarcus* sp. CIB, an anaerobic biodegrader of aromatic compounds shows an endophytic lifestyle. *PLoS One* 2014, 9, e110771, doi:10.1371/journal.pone.0110771.

CAPÍTULO V: BIOTEXT: EXPLOITING BIOLOGICAL-TEXT FORMAT FOR TEXT MINING

O objetivo neste capítulo é mostrar a aplicação da mineração de textos para a área das ciências ômicas, sob a forma de levantamento de bibliografia e estudo de caso da ferramenta BIOTEXT - desenvolvida em colaboração com Machado (2021). O BIOTEXT é um conjunto de scripts para processar e minerar textos que contém um núcleo de recursos original - com algoritmos inteiramente desenvolvidos *in-house* - que visa a resolução de questões pontuais de Bioinformática. Um preprint está disponível em <https://doi.org/10.1101/2021.04.08.439078>.

ABSTRACT

The large amount of existing textual data justifies the development of new text mining tools. Bioinformatics tools can be brought to Text Mining, increasing the arsenal of resources. Here, we present BIOTEXT, a package of strategies for converting natural language text into biological-like information data, providing a general protocol with standardized functions, allowing to share, encode and decode textual data for amino acid and DNA. The package was used to encode the arbitrary information present in the headings of the biological sequences found in a BLAST survey. The protocol implemented in this study consists of 12 steps, which can be easily executed and/ or changed by the user, depending on the study area. BIOTEXT empowers users to perform text mining using bioinformatics tools. BIOTEXT is freely available at <https://pypi.org/project/BIOTEXT/> (Python package) and https://sourceforge.net/projects/BIOTEXTtools/files/AMINOcode_GUI/ (Standalone tool).

1 INTRODUCTION

Texts are the most natural way of storing information and are responsible for the most extensive inventory of world scientific knowledge (TSHITOYAN *et al.*, 2019). In sharp contrast to the analysis options, the mass of textual information in public databases demonstrates the need to develop tools to meet the growing demand for data manipulation and analysis (KOCH *et al.*, 2020). For the omics area, whose primary focus is the analysis of complex biological events, Text Mining (TM) combined with bioinformatics techniques represents an advance for exploring and analyzing unstructured data (ZAFEIROPOULOS *et al.*, 2022).

Today's TM tools address different methodologies, with potential for research in the most diverse scenarios (TABLE S1.5 – for details of review, see FILE S1.5). These approaches involve the main tasks

such as Information Retrieval (IR), Named Entity Recognition (NER), and Information extraction (IE) (FIGURE S1.5), and a series of algorithms can be run depending on the type of analysis (TABLE S2.5) (FLEUREN; ALKEMA, 2015). However, we note that even with several aspects of TM well resolved in the literature, it is still possible to implement these processes advantageously. One of these ways is directly transposing Bioinformatics technologies to the TM area.

The exploration of biological sequences has a strong analogy with TM, since it works fundamentally with character string analysis (HARMSTON *et al.*, 2010). From more traditional techniques such BLAST for sequence alignment (ALTSCHUL *et al.*, 1990), to more recent ones such as the representation of biological sequences in vector spaces (ASGARI, MOFRAD, 2015; LEIMEISTER *et al.*, 2018; DE PIERRI *et al.*, 2020; RAITTZ *et al.*, 2021) can be brought to the TM area, adding to the arsenal of existing text analysis tools. In addition, as in the field of Bioinformatics, vector representation of texts is also a trend in text processing (MIKOLOV *et al.*, 2013; LEVY, GOLDENBERG, 2014; PENNINGTON *et al.*, 2014), since TM vectoral technologies are equally promising (MA *et al.*, 2015; LILLEBERG *et al.*, 2015; DEVLIN *et al.*, 2018; HASSANI *et al.*, 2020).

However, the Bioinformatics tools available today do not process normal texts. It is necessary to use the Biological Sequence Format (BSF) to perform analysis. Thus, here we present an approach that adapts texts in natural language to be processed with several Bioinformatics tools, including the main resources for the comparison of sequences added to the agility and robustness of vectorization technologies. BIOTEXT is a package to perform text mining strategies using bioinformatics tools, which provides AMINOcode and DNAbits features to encode text in BSF.

We explored tools for encoding arbitrary information present in the headers of biological sequences found in a typical BLAST search. We demonstrate that texts can be encoded into biological information using computational strategies based on bioinformatics. The benefit of representing texts in the form of biological sequences can be extended to the level of physical information storage, as the use of DNA for this task allows the storage of information in high density in much smaller volumes (TSHITOYAN *et al.*, 2019; KOCH *et al.*, 2020).

2 METHODS

2.1 BIOTEXT Implementation

The Python package is available from the PyPI repository at <https://pypi.org/project/BIOTEXT/>. The case study is accessible through the script "fastaHeader2plot.py, available at <https://github.com/diogomachado-bioinfo/BIOTEXT-examples/tree/master/fastaHeader2plot>. In addition, a complete list containing all BIOTEXT functions is available in Table S3.

2.2 BIOTEXT application

To demonstrate one of the applications of BIOTEXT using traditional methods in bioinformatics, we explored the local alignment for direct comparison between texts. We randomly selected the protein hypothetical "WP_011156533.1" of a set of sequences from a previous study (RUBEL *et al.*, 2015) for BLAST research, using default parameters; the first 100 hits were considered (for details, see TABLE S4.5).

3 RESULTS AND DISCUSSION

3.1 Converting texts into biological sequence format (BSF)

The BIOTEXT package includes two functions – and respective reverse functions - to code conventional texts to a valid biological format.

AMINOcode consists of replacing text characters with letters used to represent amino acids, according to a coded list. Table 1 shows the character substitution rules in AMINOcode. In addition, we present a minimal and an expanded encoding table that includes number and punctuation details representations. Whereas the 20 letters used to represent the amino acids are less than 26 needed to represent the whole alphabet, we define a conversion rule. For consonants, except for W and Y, the replacement is carried by a single letter of the amino acid representation (TABLE1.5). The coding replaces the vowels with two characters. The first is Y, a 'key' that defines the subsequent letter. Next, in the same perspective, all kinds of spaces in the texts are coded with YS. Finally, we propose the expanded version for digits and punctuation. In the expanded one, we use a key with two key characters, YP followed by the coding type. The minimal version can cause slight information, but it may be helpful in some instances. Also, we define a wildcard representation, YK, for any exception to Table 1.5.

TABLE 1.5. AMINOcode and DNAbits Dictionary

AMINOcode					
Letters	Minimal/ Expanded		Numbers	Expanded	Minimal
a	YA		0	YDA	YD
b	E		1	YDQ	
c	C		2	YDT	
d	D		3	YDH	
e	YE		4	YDF	
f	F		5	YDI	
g	G		6	YDS	
h	H		7	YDE	
i	YI		8	YDG	
j	I		9	YDN	
k	K				
l	L				
m	M				
n	N				
o	YQ				
p	P				
q	Q				
r	R				
s	S				
t	T				
u	YV				
v	V				
x	W				
z	A				
w	YW				
y	YY				
			Punctuation	Expanded	Minimal
			.	YPE	YP
			,	YPC	
			;	YPS	
			!	YPW	
			?	YPQ	
			:	YPT	
			Other	Minimal/Expanded	
			Space	YS	
			Any different	YK	
DNAbits					
Bits		Nucleotide Representation			
00		A			
10		C			
01		G			
11		T			

Legend: DNAbits dictionary is represented by the nucleotides A (adenine), C (cytosine) G (guanine) and T (thymine);

DNAbits performs the conversion of texts characters splitting every byte into four pairs of bits. The coding maintains the information according to the "American Standard Code for Information Interchange" (ASCII), replacing every couple of bits by A, C, G, and T according to the predefined rule (Table 1). In DNAbits, the characters are encoded from binary form by the function in Eq.(1), and the decoding of the DNAbits is translatable to the original text.

$$f(b) = \begin{cases} b = 00, A \\ b = 10, C \\ b = 01, G \\ b = 11, T \end{cases} \quad (1)$$

where b is any pair of bits, for example, the binary form for the character representing the letter 'a' is 10000110, resulting in CAGC after DNAbits encoding.

3.2 Interpreting BLAST results

We randomly selected the hypothetical protein WP_011156533.1" from a set of sequences from a previous study (RUBEL *et al.*, 2015) for research on BLAST (ALTSCHUL *et al.*, 1990) and run with the default parameters; the first 100 hits were considered (TABLE S4.5). Next, the headers of the sequences were converted to BSF using AMINOCODE, resulting in a FASTA file containing the text information. For the header's contents analyses, the headers sequences were converted to vectors by the SWEEP function (DE PIERRI *et al.*, 2020) Then, the headers groups were defined by agglomerative clustering (FIGURE 1.5a). Finally, we aligned the sequences from each cluster separately with Clustal Omega (SIVIERS *et al.*, 2018) and obtained the sequence consensus decoded through reverse AMINOCODE, as shown in Figure 1.5b. The entire process consists of 12 steps (see Figure S2.5):

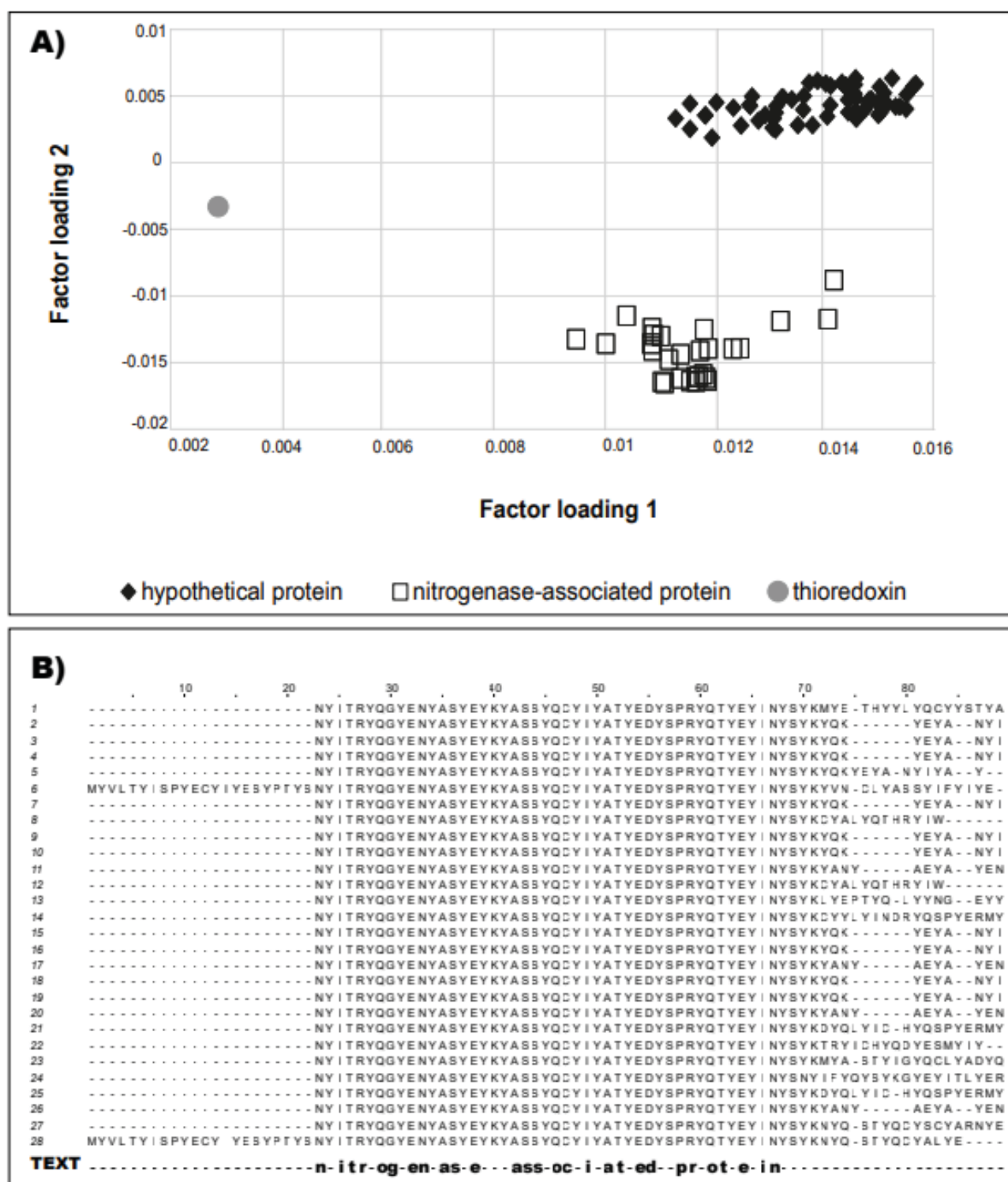
1. Extract headers from files;
2. Use expanded version with number and punctuation details (optional);
3. Find and remove patterns using regular expression for ID from all encoded texts;
4. Vectorize encoded header in a matrix with a line for each sequence;
5. Obtain the first two dimensions of loading of the PCA from SWEEP matrix;
6. Hierarchical clustering for two dimensions of loading, using the agglomerative clustering (scikit-learn python library);
7. Agglomerative clustering, using the 0,015 threshold (optional);
8. Group the encoded texts according to the clusters;
9. Obtain the consensus of sequences in each cluster from Clustal Omega alignment;
10. Find and remove patterns using regular expression for species name from consensus strings;
11. Decode the profile strings in texts;

12. Plot a 2D graph with two loading dimensions with the equivalent texts derived from the profiles as a legend.

According to the exploited data, the use of expanded or minimal versions is optional. Steps 3 and 10 correspond to removing unique patterns from each header (ID and species), which do not provide helpful information for the clustering and analysis of the texts.

Figure 1a shows the graph of the two main components of a vector PCA that represent the header sequences. Clusters are labeled by the consensus header automatically generated for the clusters: "Hypothetical protein", "Nitrogenase-associated protein," and "Thioredoxin". Distances among points (graph in FIGURE 1.5a) result from variations in the headers' text. The more variations between texts, the further apart the dots appear on the chart. Even points of the same cluster are not in the same position, as the texts represented are not identical. The exploitation of the BSF provided by BIOTEXT functions allowed us to derive information from a hypothetical BLAST query automatically. Interpreting a hypothetical protein is a challenge in bioinformatics, and comparisons with various known sequences are often adopted to understand them (IJAQ *et al.*, 2019). We seek to explore the relationship of information present in the header of a hypothetical protein using standard bioinformatics tools to accomplish this task.

FIGURE 1.5. Graphic representation of sequence headers



Reference: Authors, 2022.

Legend: a) Each dot represents a header obtained by the BLAST search using the WP_011156533.1 protein as a query. Consensus Headers are defined by: Hypothetical protein (lozangue), Nitrogenase-associated protein (square) and Thioredoxin (circle). The PCA was generated using the scikit-learn library (PEDREGOSA *et al.*, 2011). b) Representations of alignments with coded texts from a cluster. Alignment positions are shown at the top of the figure. Cuts were made in the alignments to highlight the region of greatest importance. Headers are displayed with identification numbers for each sequence (left). The last line, represented by the "TEXT" header is the decoded consensus text, properly aligned according to the coded texts

3.3 BIOTEXT application experiment in large corpus

With the string "thioredoxin", we searched on PubMed using the Biopython package "Bio.Entrez" (COCK *et al.*, 2009) (on March 18, 2022), resulting in 13,734 entries. We extract only the titles retrieved from the search and apply AMINOcode encoding using the expanded version (with detailed digits and punctuation). Then, the coded texts were vectorized and derived in a dendrogram with Euclidean distance based on the distance matrix - implementation of the SciPy library (VIRTANEN *et al.*, 2020). We use the scikit-bio library to transform the dendrogram into Newick format. The result Newick file, input text, and script to re-run all processes are available at <https://github.com/diogomachado-bioinfo/BIOTEXT-examples/tree/master/titleSet2tree>.

It is essential to highlight that the execution was quickly on an ordinary computer (i5-3470 processor and 16 Gb of DDR3 RAM). The entire process took 317.64 seconds. In particular, encoding with AMINOcode took 1.12 seconds, vectoring with SWeeP 244.38 seconds and creating the dendrogram 72.14 seconds.

4 CONCLUSION

Here we present BIOTEXT and run an example in interpreting a result of BLAST. The library can be applied in various issues involving word processing and text mining, employing FASTA files and Bioinformatics tools. In addition, this study can contribute to research on the vector representation of texts.

We encode texts (FASTA headers) in BSF and, therefore, could be treated similarly to what would be with biological FASTA. The case study shows that it is easy to perform TM tasks directly in BSF coded texts using bioinformatics tools and strategies, such as Clustering, PCA, alignment, and BLAST for consensus identification and even pattern recognition or machine learning. However, the fact that the available tools are dependent on vocabulary makes text handling complex. BIOTEXT incorporates compact vector technology to streamline analysis and eliminate the need for predefined vocabulary.

BIOTEXT combined with SWeeP vectors technology provided handling of large amounts of text, with fast execution time even on a personal computer. Moreover, our proposal is versatile, allowing the use of the entire arsenal of algorithms/methods in bioinformatics for analyzes in which texts are transformed into BSF. BIOTEXT is a general purpose tool for text mining that allows one to use different bioinformatics algorithms for multipurpose analyses and perform the main text mining tasks (IR, NER, and IE). In short, our approach can be handy for bioinformatics, given the analysts' experience with these features.

REFERENCES

- AKELLA, L. M., NORTON, C. N., MILLER, H. NetiNeti: discovery of scientific names from text using machine learning methods. **BMC bioinformatics**, 13, 211, 2012.
- ALLOT, A., PENG, Y., WEI, C. H., LEE, K., PHAN, L., LU, Z. LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. **Nucleic acids research**, v.46(W1), p. W530–W536, 2018.
- ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. Basic local alignment search tool. **J. Mol. Biol.** 215, 403–410, 1990.
- ASGARI, E. & MOFRAD, M. R. K. Continuous distributed representation of biological sequences for deep proteomics and genomics. **PLoS One** 10, 1–15, 2015.
- BARBOSA-SILVA, A., FONTAINE, J. F., DONNARD, E. R., STUSSI, F., ORTEGA, J. M., ANDRADE-NAVARRO, M. A. PESCADOR, a web-based tool to assist text-mining of biointeractions extracted from PubMed queries. **BMC bioinformatics**, v.12, n.435, 2011.
- BARBOSA-SILVA, A., SOLDATOS, T.G., MAGALHÃES, I.L. *et al.* LAITOR - Literature Assistant for Identification of Terms co-Occurrences and Relationships. **BMC Bioinformatics** v.11, n. 70, 2010.
- BECKER, K. G., HOSACK, D. A., DENNIS, G., JR, LEMPICKI, R. A., BRIGHT, T. J., CHEADLE, C., ENGEL, J. PubMatrix: a tool for multiplex literature mining. **BMC bioinformatics**, v.4, n.61, 2003.
- BRANCOTTE, B., BITON, A., BERNARD-PIERROT, I., RADVANYI, F., REYAL, F., COHEN-BOULAKIA, S. Gene List significance at-a-glance with GeneValorization. **Bioinformatics** (Oxford, England), v.27, n.8, p.1187–1189, 2011.
- CAÑADA, A., CAPELLA-GUTIERREZ, S., RABAL, O., OYARZABAL, J., VALENCIA, A., KRALLINGER, M. LimTox: a web tool for applied text mining of adverse event and toxicity associations of compounds, drugs and genes. **Nucleic acids research**, v.45(W1), p. W484–W489, 2017.
- CEJUELA, J. M., BOJCHEVSKI, A., UHLIG, C., BEKMUKHAMETOV, R., KUMAR KARN, S., MAHMUTI, S., BAGHUDANA, A., DUBEY, A., SATAGOPAM, V. P., Rost, B. nala: text mining natural language mutation mentions. **Bioinformatics** (Oxford, England), v.33, n.12, p.1852–1858, 2017.
- CHAIX, E., DELÉGER, L., BOSSY, R. NÉDELLEC, C. Text mining tools for extracting information about microbial biodiversity in food. **Food microbiology**, v.81, p.63–75, 2019.
- CHEN, C. C., HO, C. L. PubstrackHelper: A Web-based Text-Mining Tool for Marking Sentences in Abstracts from PubMed Using Multiple User-Defined Keywords. **Bioinformation**, v.10, n.11, p.708–710, 2014.
- CHEN, H., SHARP, B. M. Content-rich biological network constructed by mining PubMed abstracts. **BMC bioinformatics**, v.5, n.147, 2004.
- CHENG, D., KNOX, C., YOUNG, N., STOTHARD, P., DAMARAJU, S., WISHART, D. S. PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. **Nucleic acids research**, n.36, W399–W405, 2008.
- COMEAU, D. C., BATISTA-NAVARRO, R. T., DAI, H. J., DOĞAN, R. I., YEPES, A. J., KHARE, R., LU, Z., MARQUES, H., MATTINGLY, C. J., NEVES, M., PENG, Y. *et al.* BioC interoperability track overview. **Database: the journal of biological databases and curation**, bau053, 2014.
- CRISAN, A., MUNZNER, T., GARDY, J. L. Adjutant: an R-based tool to support topic discovery for systematic and literature reviews. **Bioinformatics** (Oxford, England), v.35, n.6, p.1070–1072. 2019.
- DAVID J. STATES, ALEX S. ADE, ZACHARY C. WRIGHT, AARON V. BOOKVICH, BRIAN D. Athey, MiSearch adaptive PubMed search tool, **Bioinformatics**, v. 25, Issue 7, p. 974–976, 2009

DE PIERRI, C.R.; VOYCEIK, R.; SANTOS DE MATTOS, L.G.C.; KULIK, M.G.; CAMARGO, J.O.; REPULA DE OLIVEIRA, A.M.; DE LIMA NICHIO, B.T.; MARCHAUKOSKI, J.N.; DA SILVA FILHO, A.C.; GUIZELINI, ORTAGEA, J.M.; PEDROSA, F.O. RAITTZ, R.T. SWeeP: Representing large biological sequences datasets in compact vectors. **Sci. Rep.**, 10, 91, 2020.

DEVLIN, J.; CHANG, M. W.; LEE, K.; TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

DING, R., BOUTET, E., LIEBERHERR, D., SCHNEIDER, M., TOGNOLLI, M., WU, C. H., VIJAY-SHANKER, K., ARIGHI, C. N. eGenPub, a text mining system for extending computationally mapped bibliography for UniProt Knowledgebase by capturing centrality. **Database: the journal of biological databases and curation**, bax081, 2017.

DOUGHTY, E.; KERTESZ-FARKAS, A.; BODENREIDER, O.; THOMPSON, G.; ADADEY, A.; PETERSON, T.; KANN, M.G. Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature, **Bioinformatics**, v.27, n. 3, 408–415, 2011.

ELSWORTH, B., GAUNT, T. R. MELODI Presto: a fast and agile tool to explore semantic triples derived from biomedical literature. **Bioinformatics** (Oxford, England), v.37, n.4, p. 583–585, 2021

FONTAINE, J. F., BARBOSA-SILVA, A., SCHAEFER, M., HUSKA, M. R., MURO, E. M., ANDRADE- NAVARRO, M. A. MedlineRanker: flexible ranking of biomedical literature. **Nucleic acids research**, v.37 (Web Server issue), p. W141–W146, 2009.

FURRER, L., JANCOS, A., COLIC, N., RINALDI, F. OGER++: hybrid multi-type entity recognition. **Journal of cheminformatics**, n.11(1), v.7, 2019.

GERNER, M., NENADIC, G., BERGMAN, C. M. LINNAEUS: a species name identification system for biomedical literature. **BMC bioinformatics**, v.11, n.85., 2010

GERNER, M., SARAFRAZ, F., BERGMAN, C. M., NENADIC, G. BioContext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events. **Bioinformatics** (Oxford, England), v.28, n.16, p.2154–2161, 2012.

GIGLIA E. Quertle and KNALIJ: searching PubMed has never been so easy and effective. **European journal of physical and rehabilitation medicine**, v.47, n.4, p.687–690, 2011.

GIJÓN-CORREAS, J. A., ANDRADE-NAVARRO, M. A., FONTAINE, J. FALKemio: association of chemicals with biomedical topics by text and data mining. **Nucleic acids research**, v.42(Web Server issue), p.W422– W429, 2014.

GUPTA, S., ROSS, K. E., TUDOR, C. O., WU, C. H., SCHMIDT, C. J., VIJAY-SHANKER, K. miRiaD: A Text Mining Tool for Detecting Associations of microRNAs with Diseases. **Journal of biomedical semantics**, v.7(1), n.9, 2016.

HARMSTON, N.; FILSELL, W.; STUMPF, M.P.H. What the papers say: Text mining for genomics and systems biology. **Human Genomics**, v.5, n.1, p.17-29, 2010.

HASSANI, H.; BENEKI, C.; UNGER, S.; MAZINANI, M. T.; YEGANEHI, M.R. Text mining in big data analytics. **Big Data and Cognitive Computing**, v. 4, n. 1, p. 1, 2020.

HE, X., LI, Y., KHETANI, R., SANDERS, B., LU, Y., LING, X., ZHAI, C., SCHATZ, B. BSQA: integrated text mining using entity relation semantics extracted from biological literature of insects. **Nucleic acids research**, n.38(Web Server issue), p. W175–W181, 2010.

HIRSCHMAN, L., YEH, A., BLASCHKE, C., VALENCIA, A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. **BMC bioinformatics**, v.6. Suppl 1, 2005.

IJAQ, J.; MALIK, G.; KUMAR, A.; DAS, P. S.; MEENA, N.; BETHI, N.; SURAVAJHALA, P. A model to predict the function of hypothetical proteins through a nine-point classification scoring schema. **BMC bioinformatics**, v. 20, n. 1, p. 1-8, 2019.

- JESSOP, D. M., ADAMS, S. E., WILLIGHAGEN, E. L., HAWIZY, L., MURRAY-RUST, P. OSCAR4: a flexible architecture for chemical text-mining. **Journal of cheminformatics**, v.3(1), n.41, 2011.
- KOCH J., GANTENBEIN S., MASANIA K., STARK W.J., ERLICH Y., GRASS R.N. A DNA-of-things storage architecture to create materials with embedded memory. **Nat Biotechnol** 38, 39–43, 2020.
- KONING, D., SARKAR, I. AND MORITZ, T. TaxonGrab: Extracting taxonomic names from text', **BiodiversityInform.** v. 2, pp. 79–82, 2005.
- KRALLINGER, M., RABAL, O., LOURENÇO, A., OYARZABAL, J., VALENCIA, A. A. Information retrieval and text mining technologies for chemistry. **Chem. Rev.** v.117, p. 7673–7761, 2017.
- KRALLINGER, M., RODRIGUEZ-PENAGOS, C., TENDULKAR, A., VALENCIA, A. PLAN2L: a web tool for integrated text mining and literature-based bioentity relation extraction. **Nucleic acids research**, v.37(WebServer issue), p.W160–W165, 2009.
- KUHN, M., VON MERING, C., CAMPILLOS, M., JENSEN, L. J., BORK, P. STITCH: interaction networks of chemicals and proteins. **Nucleic acids research**, v.36(Database issue), D684–D688, 2008.
- KWON, D., KIM, S., WEI, C. H., LEAMAN, R., LU, Z. ezTag: tagging biomedical concepts via interactive learning. **Nucleic acids research**, v.46(W1), p.W523–W529, 2018.
- LE, D. H., TRAN, T. RWRMTN: a tool for predicting disease-associated microRNAs based on a microRNA-target gene network. **BMC bioinformatics**, v.21(1), n.244, 2020.
- LEARY, P. R., REMSEN D. P, NORTON, C.N., PATTERSON, D.J. SARKAR, I.N. uBioRSS: Tracking taxonomic literature using RSS, **Bioinformatics**, v.23, n.11, June, 1p.434–1436 2007.
- LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature** v.521, p.436–444, 2015.
- LEE, H., YI, G. S., PARK, J. C. E3Miner: a text mining tool for ubiquitin-protein ligases. **Nucleic acids research**, v.36, W416–W422, 2008.
- LEIMEISTER CA, SCHELLHORN J, DÖRRER S, GERTH M, BLEIDORN C, MORGENSTERN B. Prot-SpaM: fast alignment-free phylogeny reconstruction based on whole-proteome sequences. **Gigascience** 8, 1–14, 2018.ASGARI
- LEVY, O.; GOLDBERG, Y. Linguistic regularities in sparse and explicit word representations. In: **Proceedings of the eighteenth conference on computational natural language learning**. 2014. p. 171-180.
- LI, G., ROSS, K. E., ARIGHI, C. N., PENG, Y., WU, C. H., VIJAY-SHANKER, K. miRTex: A Text Mining System for miRNA-Gene Relation Extraction. **PLoS computational biology**, v.11, n.9, e1004391, 2015.
- LI, M., HE, Q., MA, J., HE, F., ZHU, Y., CHANG, C., CHEN, T. PPICurator: A Tool for Extracting Comprehensive Protein-Protein Interaction Information. **Proteomics**, v.19(4), e1800291, 2019.
- LIM, K. M. K.; LI, C.; CHING, K. R.; NAGARAJAN, N. @MInter: Automated text-mining of microbial interactions. **Bioinformatics** 32, 2981–2987, 2016.
- LILLEBERG, J.; ZHU, Y.; ZHANG, Y. Support vector machines and word2vec for text classification with semantic features. In: **2015 IEEE 14th International**
- LIU, W., LAULEDERKIND, S. J., HAYMAN, G. T., WANG, S. J., NIGAM, R., SMITH, J. R., DE PONS, J., DWINELL, M. R., SHIMOYAMA, M. OntoMate: a text-mining tool aiding curation at the Rat GenomeDatabase. **Database: the journal of biological databases and curation**, bau129, 2015.
- LOURENÇO, A., CARREIRA, R., CARNEIRO, S., MAIA, P., GLEZ-PEÑA, D., FDEZ-RIVEROLA, F., FERREIRA, E. C., ROCHA, I., ROCHA, M. @Note: a workbench for biomedical text mining. **Journal of biomedical informatics**, v.42, n.4, p.710–720, 2009.

MA, Long; ZHANG, Yanqing. Using Word2Vec to process big text data. In: **2015 IEEE International Conference on Big Data (Big Data)**. IEEE, 2015. p. 2895-2897.

MACHADO, D.J.S. BIOTEXT: mineração de textos inspirada em técnicas de bioinformática. **Dissertação de Mestrado**. Programa de Pós-graduação em Bioinformática. Universidade Federal do Paraná: Curitiba, 2020.

MACHADO, D.J.S.; DE PIERRI, C.R.; SANTOS, L.G.C.; PEDROSA, F.O.; RAITTZ, R.T. BIOTEXT: Exploiting Biological-like format for text mining. *BioRxiv* 2021.04.08.439078. doi: 10.1101/2021.04.08.439078.

MAHMOOD, A., RAO, S., MCGARVEY, P., WU, C., MADHAVAN, S., VIJAY-SHANKER, K.. eGARD: Extracting associations between genomic anomalies and drug responses from text. **PloS one**, v.12, n.12,e0189663, 2017.

MCDONALD, R., PEREIRA, F. Identifying gene and protein mentions in text using conditional randomfields. **BMC Bioinformatics** v.6, S6, 2005.

MIKOLOV, T.; CHEN, K.; CORRADO, G.; DEAN, J. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

MÜLLER, H. M., KENNY, E. E., STERNBERG, P. W. Textpresso: an ontology-based information retrieval and extraction system for biological literature. **PLoS biology**, v. 2, n.11, e309, 2004.

NADERI, N., KAPPLER, T., BAKER, C. J., WITTE, R. OrganismTagger: detection, normalization and grounding of organism entities in biomedical documents. **Bioinformatics** (Oxford, England), v.27, .19), p.2721–2729, 2011.

NUNES, T., CAMPOS, D., MATOS, S., OLIVEIRA, J. L. BeCAS: biomedical concept recognition services and visualization. **Bioinformatics** (Oxford, England), v.29, n.15, p.1915–1916, 2013.

PAFILIS, E., FRANKILD, S. P., FANINI, L., FAULWETTER, S., PAVLOUDI, C., VASILEIADOU, A., ARVANITIDIS, C., JENSEN, L. J. The SPECIES and ORGANISMS Resources for Fast and Accurate Identification of Taxonomic Names in Text. **PloS one**, v.8(6), e65390, 2013.

PAPANIKOLAOU, N., PAVLOPOULOS, G. A., THEODOSIOU, T., ILIOPOULOS, I. Protein-protein interaction predictions using text mining methods. **Methods** v.74, p.47–53, 2015.

PEDREGOSA, F., VAROQUAUX, G., GRAMFORT, A., MICHAEL, V., THIRION, B. Scikit-learn: Machine Learning in Python. **Journal of Machine Learning Research** 12, 2825-2830, 2011.

PIERECK, B., OLIVEIRA-LIMA, M., BENKO-ISEPPON, A. M., DIEHL, S., SCHNEIDER, R., BRASILEIRO-VIDAL, A. C., BARBOSA-SILVA, A. LAITOR4HPC: A text mining pipeline based on HPC for building interaction networks. **BMC bioinformatics**, v.21(1), n.365, 2020.

PLAKE, C., ROYER, L., WINNENBURG, R., HAKENBERG, J., SCHROEDER, M. GoGene: gene annotation in the fast lane. **Nucleic acids research**, v.37 (Web Server issue), p.W300–W304, 2009.

PLAKE, C., SCHIEMANN, T., PANKALLA, M., HAKENBERG, J., LESER, U. AliBaba: PubMed as a graph. **Bioinformatics** (Oxford, England), v.22, n.19, p. 2444–2445, 2006.

PLETSCHER-FRANKILD, S., PALLEJÀ, A., TSAFOU, K., BINDER, J. X., JENSEN, L. J. DISEASES: Text mining and data integration of disease-gene associations. **Methods** v.74, p.83–89, 2015.

PRICE, M. N.; ARKIN, A. P. PaperBLAST: Text Mining Papers for Information about Homologs. **Systems** 2,1–10, 2017.

RAVIKUMAR, K. E., RASTEGAR-MOJARAD, M., LIU, H. BELMiner: adapting a rule-based relation extractionsystem to extract biological expression language statements from bio-medical literature evidence sentences. **Database: the journal of biological databases and curation**, baw156, 2017.

REBHOLZ-SCHUHMANN, D., ARREGUI, M., GAUDAN, S., KIRSCH, H., JIMENO, A. Text processing through Web services: calling Whatizit. **Bioinformatics** (Oxford, England), v.24, n.2, p.296–298, 2008.

REN, J., LI, G., ROSS, K., ARIGHI, C., MCGARVEY, P., RAO, S., COWART, J., MADHAVAN, S., VIJAY-SHANKER, K., WU, C. H. iTextMine: integrated text-mining system for large-scale knowledge extraction from the literature. **Database: the journal of biological databases and curation**, bay128, 2018.

RINALDI, F., CLEMATIDE, S., MARQUES, H., ELLENDORFF, T., ROMACKER, M., RODRIGUEZ-ESTEBAN, R. OntoGene web services for biomedical text mining. **BMC bioinformatics**, v.15 Suppl 14(Suppl 14), S6, 2014.

RUBEL E.T., RAITTZ RT, COIMBRA N.A., GEHLEN M.A., PEDROSA F.O. ProClaT, a new bioinformatics tool for *in silico* protein reclassification: case study of DraB, a protein coded from the *draTGB* operon in *Azospirillum brasilense*. **BMC Bioinformatics** 17, 455, 2016.

SALGADO, D., KRALLINGER, M., DEPAULE, M., DRULA, E., TENDULKAR, A. V., LEITNER, F., VALENCIA, A., MARCELLE, C. MyMiner: a web application for computer-assisted biocuration and text annotation. **Bioinformatics** (Oxford, England), v.28, n.17, 2285–2287, 2012.

SEN SARMA, M., ARCOLEO, D., KHETANI, R. S., CHEE, B., LING, X., HE, X., JIANG, J., MEI, Q., ZHAI, C., SCHATZ, B. BeeSpace Navigator: exploratory analysis of gene function using semantic indexing of biological literature. **Nucleic acids research**, v.39(Web Server issue), p.W462–W469. 2011.

SETTLES B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. **Bioinformatics** (Oxford, England), v.21, n.14, p.3191–3192, 2005.

SHAO, W., ADAMS, C. E., COHEN, A. M., DAVIS, J. M., MCDONAGH, M. S., THAKURTA, S., YU, P. S., SMALHEISER, N. R. Aggregator: A machine learning approach to identifying MEDLINE articles that derive from the same underlying clinical trial. **Methods** v.74, p.65–70, 2015.

SHATKAY, H.; BRADY, S.; WONG, A. Text as data: Using text-based features for proteins representation and for computational prediction of their characteristics. **Methods** 74, 54–64, 2015.

SIEVERS F., HIGGINS D.G. Clustal Omega for making accurate alignments of many protein sequences. **Protein Sci**; 27:135–145, 2018.

SIMON, C., DAVIDSEN, K., HANSEN, C., SEYMOUR, E., BARNKOB, M. B., OLSEN, L. R. BioReader: a text mining tool for performing classification of biomedical literature. **BMC bioinformatics**, v.19(Suppl 13), n.57, 2019.

SINGHAL, A.; SIMMONS, M.; LU, Z. Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine. **PLoS Comput. Biol.** 12, 1–19, 2016.

SMALHEISER, N. R., ZHOU, W., TORVIK, V. I. Anne O'Tate: A tool to support user-driven summarization, drill down and browsing of PubMed search results. **Journal of biomedical discovery and collaboration**, v.3,n.2, 2008.

SNEL, B., LEHMANN, G., BORK, P., HUYNEN, M. A. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. **Nucleic acids research**, v.28, n.18, p. 3442–3444, 2000.

SRINIVASAN P. MeSHmap: a text mining tool for MEDLINE. **Proceedings. AMIA Symposium**, p.642–646, 2001.

STEFFEN, P., WU, J., HARIHARAN, S., VOSS, H., RAGHUNATH, V., MOLLOY, M. P., SCHLÜTER, H. OmixLitMiner: A Bioinformatics Tool for Prioritizing Biological Leads from 'Omics Data Using Literature Retrieval and Data Mining. **International journal of molecular sciences**, v.21(4), 1374, 2020.

TANABE, L., SCHERF, U.; SMITH, L. H.; LEE, J. K.; HUNTER, L.; WEINSTEIN, J. N. MedMiner: An Internet text-mining tool for biomedical information, with application to gene expression profiling. **BioTechniques**, v. 27, N. 6. pp. 1210-1217, 1999.

TORII, M., ARIGHI, C. N., LI, G., WANG, Q., WU, C. H., VIJAY-SHANKER, K. RLIMS-P 2.0: A Generalizable Rule-Based Information Extraction System for Literature Mining of Protein Phosphorylation Information. **IEEE/ACM transactions on computational biology and bioinformatics**, v.12, n.1, p.17–29, 2015.

TSEYTLIN, E., MITCHELL, K., LEGOWSKI, E., CORRIGAN, J., CHAVAN, G., JACOBSON, R. S. NOBLE – Flexible

concept recognition for large-scale biomedical natural language processing. **BMC bioinformatics**, v.17, n.32, 2016.

TSHITOYAN V, DAGDELEN J, WESTON L, DUNN A, RONG Z, KONONOVA O, PERSSON KA, CEDER G, JAIN A. Unsupervised word embeddings capture latent knowledge from materials science literature. **Nature** 571, 95–98, 2019.

TSURUOKA, Y., MIWA, M., HAMAMOTO, K., TSUJII, J., ANANIADOU, S. Discovering and visualizing indirect associations between biomedical concepts. **Bioinformatics** (Oxford, England), v.27, n.13, p.i111–i119, 2011.

TURINA, P., FARISELLI, P., CAPRIOTTI, E. ThermoScan: Semi-automatic Identification of Protein Stability Data from PubMed. **Frontiers in molecular biosciences**, v.8, 620475, 2021.

URAMOTO, N., MATSUZAKA, H., NAGANO, T., MURAKAMI, A., TAKEUCHI, H., TAKEDA, K. text-mining system for knowledge discovery from biomedical documents. **IBM Syst. J.**, v.43, p. 516–533, 2010.

VAN LANDEGHEM, S., BJÖRNE, J., WEI, C. H., HAKALA, K., PYYSALO, S., ANANIADOU, S., KAO, H. Y., LU, Z., SALAKOSKI, T., VAN DE PEER, Y., GINTER, F. Large-scale event extraction from literature with multi-level gene normalization. **PloS one**, v.8(4), e55814, 2013.

VIRTANEN, Pauli *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. **Nature methods**, v. 17, n. 3, p. 261–272, 2020.

WEI, C. H., KAO, H. Y., LU, Z. PubTator: a web-based text mining tool for assisting biocuration. **Nucleic acids research**, v.41(Web Server issue), p.W518–W522, 2013.

WOLFSON, M., WALLACE, S. E., MASCA, N., ROWE, G., SHEEHAN, N. A., FERRETTI, V., LAFLAMME, P., TOBIN, M. D., MACLEOD, J., LITTLE, J. *et al.* RDataSHIELD: resolving a conflict in contemporary bioscience-performing a pooled analysis of individual-level data without sharing the data. **International journal of epidemiology**, v.39, n.5, p. 1372–1382, 2010.

YE, Z., TAFTI, A. P., HE, K. Y., WANG, K., HE, M. M. SparkText: Biomedical Text Mining on Big Data Framework. **PloS one**, v.11(9), e0162721, 2016.

YU, H., KIM, T., OH, J., KO, I., KIM, S., HAN, W. S. Enabling multi-level relevance feedback on PubMed by integrating rank learning into DBMS. **BMC bioinformatics**, v.11 Suppl 2, S6. 2010.

ZAFEIROPOULOS, H.; PARAGKAMIAN, S.; NINIDAKIS, S.; PAVLOPOULOS, G. A.; JENSEN, L. J.; PAFILIS, E. PREGO: A Literature and Data-Mining Resource to Associate Microorganisms, Biological Processes, and Environment Types. **Microorganisms**, v. 10, n. 2, p. 293, 2022.

ZHANG, L., BERLEANT, D., DING, J., CAO, T., SYRKIN WURTELE, E. PathBinder--text empirics and automatic extraction of biomolecular interactions. **BMC bioinformatics**, v.10 Suppl 11, S18, 2009.

ZHU, F., PATUMCHAROENPOL, P., ZHANG, C., YANG, Y., CHAN, J., MEECHAI, A., VONGSANGNAK, W. SHEN, B. Biomedical text mining and its applications in cancer research. **J. Biomed. Inform.**, v.46, p. 200–211, 2013.

SUPPLEMENTARY INFORMATION

TABLE S1.5. Text mining tools and approaches for Omics sciences

Tool	Description	Purpose	Task	Input	Input	Authors and Year
MedMiner	Web-based tool that filters the literature and returns organized information.	Biocuration	IR	List of genes	List of genes	TANABE <i>et al.</i> , 1999
STRING	Tool for recurring instances of neighboring genes, retrieving and displaying the genes that in clusters on the genome.	Biocuration	IR	Query	Query	SNEL <i>et al.</i> , 2000
MeshMap	Text mining system to exploit the MeSH indexing accompanying the Medline records and compare entities of the same type.	General purpose	IR NER	Query	Query	SRINIVASAN, 2001
Pubmatrix	Comparison tool for systematic and automatic search on PubMed.	General purpose	IR	Query	Query	BECKER <i>et al.</i> , 2003
Chillibot	NLP-based text-mining approach which construct content-rich relationship networks among biological concepts, genes, proteins, or drugs.	General purpose	IR NER IE	List of genes	List of genes	CHEN; SHARP, 2004
Textpresso	A text Mining tool that splits papers into sentences, and sentences intowords or phrases, including all categories of the Gene Ontology database	General purpose	IR NER IE	Texts	Texts	MULLER <i>et al.</i> , 2004
BioCreActive	A multitasking Text Mining tool to assess the state of the art applied tobiological problems.	General purpose	IR NER IE	Texts	Texts	HIRSHMAN <i>et al.</i> , 2005
TAXONGRAB	A tool for extract taxonomic names from English, Spanish, French and German texts, using a lexicon of words.	Taxonomy	NER	Texts	List of scientific names	KONING <i>et al.</i> , 2005
CRF	A tool for tagging gene and protein mentions from text using the probabilistic sequence tagging framework.	General purpose	NER	List of terms	List of terms	MACDONALD; PEREIRA, 2005
ABNER	A tool for molecular biology text mining based on machine learning system to extract features and tag entities.	General purpose	NER	List of terms	List of terms	SETTLES, 2005

(Continue)

AliBaba	An interactive tool for graphical summarization parsing Abstracts of PubMed searches.	General purpose	IR	Query	Query	PLAKE <i>et al.</i> , 2006
STITCH	A TM tool for interactions of chemicals that integrates information about interactions from metabolic pathways, crystal structures, binding experiments and drug–target relationships.	Chemicals	IR NER IE	Query	Query	KUHN <i>et al.</i> , 2007
Whatizit	A Tool for tag concepts in texts based on thesaurus.	General purpose	NER	Texts	Texts	REBHOLZ-SCHUHMANN <i>et al.</i> , 2007
TaxonFinder	A tool designed to find scientific names from text with the help of separate dictionaries for species and genus names	Taxonomy	NER	Texts	List of scientific names	LEARY <i>et al.</i> , 2007
PolySearch	Web-tool system which makes relationships between several queries in the biomedical sciences.	General purpose	IR NER IE	Query	Query	CHENG <i>et al.</i> , 2008
E3Miner	Tool that extracts comprehensive knowledge about the ubiquitin-protein ligase (E3) from the scientific literature.	Molecular function	IR IE	Abstracts	Abstracts	LEE <i>et al.</i> , 2008
ANNE O'TATE	An integrated, generic tool for summarization, drill-down and browsing of PubMed search results	General purpose	IR NER	Query	Query	SMALHEISER <i>et al.</i> , 2008
MedlineRanker	Allows flexible and fast ranking of Medline abstracts for a topic of interest without expert knowledge.	General purpose	IR	Query	Query	FONTAINE <i>et al.</i> , 2009
MiSearch	Search tool that ranks citations based on a statistical model for the likelihood that a user will choose to view them.	General Purpose	IR	Query	Query	STATES <i>et al.</i> , 2009

(Continue)

GoGene	Performs high-throughput text mining to complement annotation of genes.	Biocuration	IR IE	List of Genes	List of Genes	PLAKE <i>et al.</i> , 2009
@Note	Its main functional contributions are the ability to process abstracts and full-texts.	General Purpose	IR	Abstracts / full-texts	Abstracts / full-texts	LOURENÇO <i>et al.</i> , 2009
PLAN2L	A web-based tool that integrates text mining and information extraction techniques to access useful information for analyzing genetic, cellular and molecular aspects of Arabidopsis thaliana.	Molecular function	IR NER IE	Query	Query	KRALLINGER <i>et al.</i> , 2009
PathBinder	Plant-based tool to explore literature, and to annotate metabolic and regulatory interactions and biomolecules.	Molecular function	IR NER IE	Query	Query	ZHANG <i>et al.</i> , 2009
RefMed	A multi-level relevance feedback system for PubMed.	General Purpose	IR	Query	Query	YU <i>et al.</i> , 2010
LINNAEUS	A system that uses a dictionary and a set of regular expressions to identify species mentions in text. LINNAEUS uses dictionaries for scientific and common names to construct a DFA (Deterministic Finite Automaton).	Taxonomy	NER	Texts	Query List of scientific names	GERNER <i>et al.</i> , 2010
LAITOR	A tool for analyze co-occurrences of bioentities, biointeractions, and other biological terms in MEDLINE abstracts.	Molecular function	IR NER	Query	Query	SILVA <i>et al.</i> 2010
DataShield	Allows the analysis of sensitive individual-level data from one study. It can be applied to post-publication sensitive data analysis or text data analysis.	General Purpose	IR NER	Texts	Texts	WOLFSON <i>et al.</i> , 2010
BSQA	Web server that performs integrated text mining, recognizing several entities and relations in Medline documents about <i>Drosophila melanogaster</i> .	Molecular function	IR NER	Query	Query	HE <i>et al.</i> , 2010

(Continue)

EMU	A Text mining tool for identify mutations and their associated genes.	Molecular function	IR NER IE	Texts	Texts	DOUGHTY <i>et al.</i> , 2010
Quertle	Allows a semantic search in multiple biomedical databases and runs aquery via relationships between concepts.	General Purpose	IR NER IE	Query	Query	GIGLIA, 2011
GeneValorization	A web-based tool that gives an overview of the bibliography available	Molecular function	IR NER	List of Genes	List of Genes	BRANCOTTE <i>et al.</i> , 2011
FACTA+	Real-time text-mining system for finding and visualizing indirect associations between biomedical concepts from MEDLINE abstracts.	General Purpose	IR NER IE	Query	Query	TSUROUKA <i>et al.</i> , 2011
OrganismTagger	A hybrid rule-based/machine learning system to extract organism mentions from the literature	Taxonomy	NER	List of names	Query List of scientific names	NADERI <i>et al.</i> , 2011
BeeSpace Navigator	Tool for exploratory analysis and extract gene functions.	Molecular function	IR NER IE	Query	Query	SARM, <i>et al.</i> , 2011
PESCADOR	A web tool that extracts a network of interactions from a set of PubMedabstracts, filtering according to user-defined concepts.	General purpose	IR NER IE	Query	Query	SILVA <i>et al</i> 2011
OSCAR4	The tool recognizes sections of reported experimental and use regular expressions to match the journal-mandated formats in which are reportedin the literature.	Chemicals	IR NER	text	text	JESSOP <i>et al.</i> , 2011
BioContext	TM system which extracts, extends and integrates results from a numberof tools performing entity recognition, biomolecular event extraction and contextualization.	Molecular function	IR NER	Query	Query	GERNER <i>et al.</i> 2012

(Continue)

NetiNeti	A tool focused on recognition/discovery of scientific names of organisms from various text sources.	Taxonomy	NER	Texts	List of scientific names	AKELLA <i>et al.</i> , 2012
MyMiner	MyMiner is a free and user-friendly text annotation tool aimed to assist in carrying out the main biocuration tasks and to provide labelled data for the development of text mining systems.	Biocuration	NER	Texts	Texts	SALGADO <i>et al.</i> 2012
BioC	BioC allows a large number of different annotations to be represented such as genes or diseases and relationships between named entities.	General Purpose	IR NER	Text	Text	COMEAU <i>et al.</i> , 2013
PubTator	Web server tool for text mining that allows multiple challenge-winning to assist biocuration, by featuring a PubMed-like interface.	Biocuration	IR NER	Query	Query	WEI <i>et al.</i> , 2013
SPECIES	Dictionary-based approaches, being able to recognize vernacular names in text in addition to scientific ones.	Taxonomy	NER	Texts	List of scientific names	PAFILIS <i>et al.</i> , 2013
EVEX	Is a text mining resource which focuses on biomedical event extraction and gene interactions, covering the whole biomedical literature available in PubMed.	Molecular function	IR NER IE	Query	Query	LANDEGHEM <i>et al.</i> , 2013
BeCAS	It's a web-based tool for annotation of biomedical concepts that can be integrated on larger text-processing pipelines or incorporated on external web pages.	General Purpose	NER	Abstracts	Abstracts	NUNES <i>et al.</i> , 2013
OntoGene	A TM system specialized in the detection of named entities as proteins, genes, and others, allowing a deeper connection between the unstructured information.	Molecular function	IR NER IE	Query	Query	RINALDI <i>et al.</i> 2014
Taxamatch	Employs a custom distance algorithm in tandem with a phonetic algorithm, together with a rule-based approach to produce improved levels of recall, precision and execution time over the n-grams algorithms.	Taxonomy	IR NER	Query	Query List of scientific names	REES, 2014

(Continue)

Alkemio	A tool that ranks all annotated chemicals (including drugs) in the literature for any biomedical topics.	Chemicals	IR NER	Query	Query	GIJÓN-CORREAS et al., 2014
PubstrackHelper	Allows the researcher to put up to ten different keywords and highlight sentences with co-occurring keyword in different colors.	General purpose	IR	Query	Query	CHEN; HO, 2014
RLIMS-P 2.0	Rule-based text-mining program specifically designed to extract protein phosphorylation information on protein kinase, substrate and phosphorylation sites from biomedical literature.	Molecular function	IR NER IE	Texts	Texts	TORII et al., 2015
miRTex	Text-mining system that extracts miRNA-gene regulation relations and miRNA-target relations from literature	Molecular function	IR NER IE	Texts	Texts	LI et al., 2015
OntoMate	Ontology-driven system used as a replacement for the PubMed searchengine in the Rat Genome Database that works on the gene curation workflow.	Molecular function	IR NER IE	Abstracts	Abstracts	LIU et al. 2015
DISEASES	A system for extracting disease-gene associations from biomedical abstracts, using a dictionary-based tagger for named entity recognition of human genes and diseases.	Molecular function	IR NER IE	Abstracts	Abstracts	PLETSCHER-FRANKILD et al., 2015.
Aggregator	Machine learning approach to identify articles that derived from the same clinical trial.	General Purpose	IR NER IE	Train sets	Train sets	SHAO et al., 2015
NOBLE	A recognition component for NLP (Natural Language Processing) pipelines.	General Purpose	NER	Query	Query	TSEYTLIN et al., 2016
@Minter	An automated information extraction system based on Support Vector Machines to analyze paper abstracts and infer microbial interactions.	Molecular function	IR NER IE	Abstracts	Abstracts	LIM et al., 2016.

(Continue)

MIRiad	A text mining tool that automatically extracts associations between microRNAs and diseases from the literature.	Molecular function	IR NER IE	Query	Query	GUPTA <i>et al.</i> , 2016
SparkText	Text mining framework, on a Big Data infrastructure, which is composed of Apache Spark data streaming and machine learning methods. An online text mining application that automatically extracts toxicology relevant information from text, with special emphasis on drug-induced adverse hepatobiliary reactions.	General Purpose	IR NER IE	Abstracts/full-texts	Abstracts/full-texts	YE <i>et al.</i> 2016
LimTox		Chemicals	NER	Preprocessed documents	Preprocessed documents	CANADA <i>et al.</i> , 2017
PaperBlast	PaperBLAST uses Europe PMC to search the full text of scientific articles for references to genes, to find similar proteins presents in snippets of text.	Biocuration	IR	Query	Query	PRICE <i>et al.</i> , 2017.
EGenPub	Selects articles about a certain protein and identify an additional bibliography from the UniProt knowledgebase, based in SVM.	Molecular function	IR NER IE	UNIPROT accession number	UNIPROT accession number	DING <i>et al.</i> , 2017.
BELMiner	Semi-automated Text mining system used to facilitate the curation tasks on building biological networks from unstructured scientific texts.	General Purpose	IR NER IE	Query	Query	RAVILKUMAR <i>et al.</i> , 2017
Nala	Nala capture natural language and mention of the target pattern mutation by combining conditional random fields with word incorporation features learned without supervision.	Molecular function	IR NER IE	Texts	Texts	CAJUELA <i>et al.</i> , 2017
Egad	A natural language processing-based text mining to extracting genomic anomalies association with Response to Drugs.	Chemicals	IR NER IE	Abstracts	Abstracts	MAHMOOD <i>et al.</i> , 2017
EzTag	Allows users to perform annotation. EzTag provides training data interactively. It supports all PubMed abstracts and PMC open access articles.	General Purpose	IR NER	Texts	Texts	KWON <i>et al.</i> , 2018.

(Continue)

ITextMine	A system which uses an automated workflow that integrates multiple textmining tools and run it on large-scale text for knowledge extraction.	General Purpose	IR NER IE	Text	Text	REN <i>et al.</i> , 2018
LitVar	A tool for the accurate search of variants and related information from unstructured human-related biomedical literature	Molecular function	IR NER IE	Query	Query	ALLOT <i>et al.</i> , 2018
Adjutant	R-based tool to support topic discovery for systematic and literature reviews.	General Purpose	IR NER IE	Query	Query	CRISAN <i>et al.</i> , 2019
AlvisIR	Text mining tool for extracting information about microbial biodiversity info.	General Purpose	IR NER IE	Abstracts	Abstracts	CHAIX <i>et al.</i> , 2019
PPIcurator	Tool for extracting comprehensive protein-protein interaction information.	Biocuration	IR NER IE	Query	Query	LI <i>et al.</i> , 2019
BioReader	Enables users to perform classification of scientific literature by text mining-based classification of article abstracts	General purpose	IR NER IE	Abstracts	Abstracts	SIMON <i>et al.</i> , 2019
OGER++	Performs document annotation by parsing, entity recognition and serialization, supporting concept recognition.	General purpose	IR NER	Abstracts	Abstracts	FURRER <i>et al.</i> , 2019
CoCoScore	An unsupervised counting approach with a machine learning approach based on distant supervision to score sentence-level co-mentions.	General purpose	IR NER IE	Abstracts	Abstracts	JUNGE; JENSEN, 2020
RWRMTN	A TM tool to predict novel disease-associated miRNAs.	Molecular function	IR NER IE	Query	Query	LI; TRAN, 2020

(Continue)

OmixLitMiner	The tool automates the retrieval of literature from PubMed based on UniProt protein identifiers, gene names and their synonyms, combined with user defined contextual keyword search	General purpose	IR NER IE	UNIPROT accession number	Query	STEFFEN <i>et al.</i> , 2020
Melodi Presto	A tool that links risk factors and diseases by identifying overlapping elements between two query lists using semantic triples.	General purpose	IR NER	Query	Query	ELSWORTH; GAUNT, 2020
LAITOR4HPC	Uses high performance computing aiming to access an unlimited number of abstracts to create new interaction networks or improve the ones that already exist.	Molecular function	IR NER IE	Query	Query	PIERECK <i>et al.</i> 2020
ThermoScan	A text mining approach for the identification of relevant thermodynamic data on protein stability from full-text articles	Molecular function	IR NER IE	Full-texts	Full-texts	TURINA <i>et al.</i> , 2021
PREGO	PREGO combines text mining and data integration techniques to explore associations from data and metadata scattered in the scientific literature and in public omics repositories.	Taxonomy	IR NER IE	Taxon Query	Tab of Organisms	ZAFEIROPOULOS <i>et al.</i> , 2022

Reference: Authors, 2022.

TABLE S2.5. Text mining algorithms/models and its purpose

Process	Type of Algorithms	Purpose
Information Retrieval	Tokenization	Break text into discrete words, called textual entities or tokens
	Stopping	Remove common words, such as “the”, “or”
	Stemming	Remove prefixes and suffixes to normalize words
	Normalize case	Convert the text to either all lower or all upper case
Named Entity Recognition	Indexing	Processing a text to extract statistics considered important for representing the information and/or to allow fast search on its content.
	Dictionary-based methods	Matching a dictionary of terms, containing terms and synonyms, against the query text.
	TF-IDF	TF- IDF (Term Frequency - Inverse Document Frequency). stores text as weighted vectors; The frequency for a term is the number of times the term appears in a document.
	NLP methods	Low-level language processing and comprehension task; grammatical class tagging.
Information Extraction	Prediction models	Classifying of textual entities into predefined categories
	Clustering models	Clustering of textual entities into formerly undefined and unknown groups
	Association rules	Finding interesting association or correlation relationships among a large set of data items

Reference: Authors, 2022.

TABLE S3.5. List of BIOTEXT functions

Function Name	Description	Input	Output
BIOTEXT.aminocode.encodeText	Encodes a string with AMINOCODE. Decodes a string with reverse AMINOCODE.	text: natural language text string to be encoded;	enc_text: encode text in string format.
BIOTEXT.aminocode.encodeText		detailing: details in coding. 'd' for details in digits. 'p' for details on the punctuation. 'dp' or 'pd' for both.	
BIOTEXT.aminocode.et			
BIOTEXT.aminocode.decodeText		text: text string encoded using the encodefile function to be decode;	
BIOTEXT.aminocode.decodeText		detailing: details used in the text to be decoded. 'd' for details in digits. 'p' for details on the punctuation. 'dp' or 'pd' for both.	
BIOTEXT.aminocode.dt			
BIOTEXT.aminocode.encodeFile	Encodes a text file or a list of strings with AMINOCODE.	input_file_name: text file name or list of string. It can also be used in a list of SeqRecord*, in which case the function will automatically extract the headers to do the encoding; output_file_name: the name for the output file. If not defined, the result will only be returned as a variable;	records: list of SeqRecord; if defined output_file_name a file will be saved.
BIOTEXT.aminocode.encodeFile		detailing: same as in the encodeText function;	
BIOTEXT.aminocode.ef		header_format: format for the headers of the generated FASTA. It can be 'number+originaltext', 'number' or 'originaltext'. 'number' is a count of the lines in the input file. Blank lines are considered in the count, but are not added to the FASTA file. 'originaltext' is the input text itself; verbose: if True displays progress.	
		input_file_name: file name or list of SeqRecord;	
		output_file_name: the name for the output file. If not defined, the result will only be returned as a variable;	
		detailing: same as in the decodeText function; verbose: if True displays progress; output: string list. If defined output_file_name a file will be saved.	
BIOTEXT.aminocode.decodeFile	Decodes a fasta file or a SeqRecord list with the reverse amino acid.		dec_list: string list; if defined output_file_name a file will be saved.
BIOTEXT.aminocode.decodeFile			
BIOTEXT.aminocode.df			
BIOTEXT.dnabits.encodeText	Encodes a string with DNABits.	text: natural language text string to be encoded.	enc_text: encode text in string format.
BIOTEXT.dnabits.encodeText			
BIOTEXT.dnabits.et			
BIOTEXT.dnabits.decodeText	Decodes a string with reverse DNABits.		dec_text: decode text in string format.
BIOTEXT.dnabits.decodeText		text: text string encoded using the encodefile function to be decode.	
BIOTEXT.dnabits.dt			
(Continue)			

BIOTEXT.dnabits.encodeFile BIOTEXT.dnabits.encodefile BIOTEXT.dnabits.ef	Encodes a text file or a list of strings with DNABits.	input_file_name: text file name or list of string. It can also be used in a list of SeqRecord, in which case the function will automatically extract the headers to do the encoding; output_file_name: the name for the output file. If not defined, the result will only be returned as a variable; header_format: format for the headers of the generated FASTA. It can be 'number+originaltext', 'number' or 'originaltext'. 'number' is a count of the lines in the input file. Blank lines are considered in the count, but are not added to the FASTA file. 'originaltext' is the input text itself; verbose: if True displays progress.	records: list of SeqRecord. if defined output_file_name a file will be saved.
	Decodes a text file or a SeqRecord list with reverse DNABits.	input_file_name: file name or list of SeqRecord; output_file_name: the name for the output file. If not defined, the result will only be returned as a variable; verbose: if True displays progress.	dec_list: string list; if defined output_file_name a file will be saved.
	Converts a list of strings to a list of SeqRecord.	seq: list of biological sequences in string format;	records: list of SeqRecord.
BIOTEXT.fasttools.list2SeqRecord BIOTEXT.fasttools.list2seqrecord BIOTEXT.fasttools.list2bioSeqRecord BIOTEXT.fasttools.list2bioseqrecord BIOTEXT.fasttools.list2fasta		header: list of headers in string format, if set to 'None' the headers will be automatically defined with an increasing number.	
BIOTEXT.fasttools.fasta Read BIOTEXT.fasttools.fastaread	Uses biopython to import a FASTA file.	input_file_name: input fasta file name.	records: list of SeqRecord.
BIOTEXT.fasttools.fasta Write BIOTEXT.fasttools.fastawrite	Create a file using a SeqRecord list.	records: list of SeqRecord; output_file_name: output fasta file name.	records: a file is saved with the defined name.
BIOTEXT.fasttools.getHeader BIOTEXT.fasttools.getHeader	Get the header from a list of SeqRecord.	records: list of SeqRecord.	headers: list with headers.
BIOTEXT.fasttools.getSeq BIOTEXT.fasttools.getseq	Get the string from a list of SeqRecord.	records: list of SeqRecord.	seqs: list with sequences.
BIOTEXT.fasttools.removePattern BIOTEXT.fasttools.removepattern	Removes patterns from a SeqRecord range based on regular expression.	records: list of SeqRecord; rex: regular expression.	new_records: list of SeqRecord with removal applied.

(Continue)

BIOTEXT.fastatools.clustalOmega BIOTEXT.fastatools.clustalomega BIOTEXT.fastatools.clustalo	Uses the Clustal Omega to align the strings in a FASTA file.	input_file_name: input FASTA file name.	align: list with strings aligned in string format.
BIOTEXT.fastatools.getCons BIOTEXT.fastatools.getcons	Save a temporary file with the sequences from the SeqRecord list, apply the clustalo function and obtain alignment consensus. Perform a vectorization of a list of SeqRecord using the SWeeP, a method developed by our group for FASTA vectorization.	records: list of SeqRecord.	consensus: consensus for alignment in string format; align: list with strings aligned in string format.
BIOTEXT.fastatools.fastatext2vect BIOTEXT.fastatools.fastaText2vect BIOTEXT.fastatools.fasta2vect		fastatext: list of SeqRecord.	vect: matrix with the generated vectors, in ndarray** format.
BIOTEXT.huflus.cluster BIOTEXT.huflus.clus	Perform clustering using Huflus, a method developed by our group that includes the cluster number prediction.	mat: matrix in numpy.ndarray format. For example, the output of the BIOTEXT.fastatools.fastatext2vect function can be used; loop: number of loops for smoothing that is used in the Huflus method, the default is 100.	cluster_idx: ndarray with the found clusters index.

REFERENCE: Machado *et al.* (2021).

Note: *SeqRecord: Biopython object to store biological sequences and its information, as described in <https://biopython.org/DIST/docs/api/Bio.SeqRecord.SeqRecord-class.html> **ndarray: Numpy package object to represent array, as described in <https://numpy.org/doc/stable/reference/generated/numpy.ndarray.html>.

TABLE S4.5. Results for the blast search using the protein WP_011156533.1

ID	% identity	e-value	Description	Specie
WP_011156533.1	100.000	1.16e-91	hypothetical protein	Rhodopseudomonas palustris
WP_012494740.1	99.237	1.40e-90	hypothetical protein	Rhodopseudomonas palustris
WP_107357400.1	96.183	1.24e-88	hypothetical protein	Rhodopseudomonas palustris
NEW97766.1	96.947	1.66e-88	hypothetical protein	Rhodopseudomonas sp. BR0G17
NEV79437.1	96.947	3.42e-88	hypothetical protein	Rhodopseudomonas sp. BR0C11
WP_013500942.1	93.130	8.92e-86	hypothetical protein	Rhodopseudomonas palustris
WP_107343178.1	94.656	1.65e-85	hypothetical protein	Rhodopseudomonas palustris
WP_047308448.1	91.603	2.95e-85	hypothetical protein	Rhodopseudomonas palustris
NEW90013.1	91.603	2.35e-84	hypothetical protein	Rhodopseudomonas sp. WA056
NEW92855.1	88.550	1.82e-81	hypothetical protein	Rhodopseudomonas sp. BR0M22
WP_110787918.1	89.313	5.04e-81	hypothetical protein	Rhodopseudomonas palustris
WP_114358707.1	88.550	8.27e-80	MULTISPECIES: hypothetical protein	Rhodopseudomonas
WP_092681992.1	76.923	3.42e-66	hypothetical protein	Rhodopseudomonas pseudopalustris
WP_011439953.1	77.519	3.99e-66	hypothetical protein	Rhodopseudomonas palustris
WP_011501608.1	76.923	7.97e-66	hypothetical protein	Rhodopseudomonas palustris
WP_022723522.1	73.846	1.28e-64	hypothetical protein	Rhodopseudomonas sp. B29
WP_054161643.1	76.154	6.84e-62	hypothetical protein	Rhodopseudomonas sp. AAP120
WP_111360672.1	61.832	2.62e-52	hypothetical protein	Rhodoplanes elegans
WP_108683060.1	63.780	2.71e-52	hypothetical protein	Methyloceanibacter sp. wino2
WP_116132402.1	61.069	3.85e-52	hypothetical protein	Tropicimonas sp. IMCC34043
WP_011474719.1	62.992	5.06e-52	hypothetical protein	Rhodopseudomonas palustris
WP_011665752.1	60.000	6.30e-52	hypothetical protein	Rhodopseudomonas palustris
WP_069443161.1	62.992	1.22e-51	hypothetical protein	Methyloceanibacter stevinii
WP_142861444.1	60.305	1.52e-51	hypothetical protein	Methylosinus sporium
WP_108916584.1	60.305	1.85e-51	hypothetical protein	Methylosinus sporium
WP_110806343.1	61.069	2.46e-51	hypothetical protein	Rhodobacter viridis
WP_085770093.1	60.630	1.91e-50	hypothetical protein	Methylocystis bryophila

(Continue)

WP_132695580.1	59.843	2.70e-50	hypothetical protein	Rhodovulum steppense
WP_018264372.1	59.843	3.09e-50	hypothetical protein	Methylosinus sp. LW4
WP_036289079.1	59.843	3.09e-50	hypothetical protein	Methylosinus sp. PW1
WP_138166377.1	61.417	3.46e-50	hypothetical protein	Methylocystis sp. B8
KAF0130244.1	59.843	3.87e-50	nitrogenase-associated protein	Methylocystaceae bacterium
WP_014889984.1	59.843	3.99e-50	hypothetical protein	Methylocystis sp. SC2
WP_123175947.1	59.843	4.22e-50	hypothetical protein	Methylocystis hirsuta
WP_024878936.1	59.843	4.44e-50	MUL TISPECIES: hypothetical protein	Methylocystaceae
WP_124738045.1	60.630	5.14e-50	hypothetical protein	Methylocystis rosea
NEP86806.1	62.393	5.37e-50	nitrogenase-associated protein	Okeania sp. SIO2C2
WP_037454643.1	59.843	7.43e-50	hypothetical protein	Skermanella stibiresistens
WP_111417306.1	68.702	8.14e-50	hypothetical protein	Rhodoplanes roseus
NEP77170.1	61.538	8.50e-50	nitrogenase-associated protein	Okeania sp. SIO3B3
NEO54746.1	62.393	9.26e-50	nitrogenase-associated protein	Okeania sp. SIO3B5
WP_159730261.1	59.843	1.14e-49	hypothetical protein	Methylosinus sp. Ce-a6
JC4746	59.690	1.20e-49	hypothetical 15k protein - rice	Oryza sativa
WP_036287787.1	57.480	1.38e-49	hypothetical protein	Methylocystis sp. ATCC 49242
WP_109328306.1	66.379	1.57e-49	hypothetical protein	Azospirillum sp. CFH 70021
WP_109026782.1	59.055	1.79e-49	MUL TISPECIES: hypothetical protein	Methylocystis
WP_017716702.1	60.000	2.06e-49	hypothetical protein	Oscillatoria sp. PCC 10802
WP_124142967.1	61.538	2.06e-49	nitrogenase-associated protein	Okeania hirsuta
WP_029651706.1	59.055	2.06e-49	hypothetical protein	Methylocystis sp. SB2
WP_083240244.1	64.957	2.13e-49	hypothetical protein	Methyloceanibacter methanicus
WP_018408038.1	59.055	2.13e-49	hypothetical protein	Methylocystis rosea
WP_086758352.1	64.103	2.19e-49	MUL TISPECIES: nitrogenase-associated protein	unclassified Nostoc
WP_073598916.1	64.103	2.29e-49	hypothetical protein	Hydrococcus rivularis
PPD46134.1	57.692	2.39e-49	hypothetical protein CTY15_01795	Methylocystis sp.
HAG82093.1	62.185	2.53e-49	hypothetical protein	Cyanobacteria bacterium UBA12227
NEN90905.1	61.538	2.93e-49	nitrogenase-associated protein	Okeania sp. SIO3H1
WP_096594850.1	64.957	3.32e-49	nitrogenase-associated protein	Calothrix sp. NIES-2098
WP_113892645.1	60.630	3.63e-49	hypothetical protein	Roseiarcus fermentans
NET29449.1	61.538	3.89e-49	nitrogenase-associated protein	Okeania sp. SIO117

(Continue)

NES67966.1	60.684	4.07e-49	nitrogenase-associated protein	Okeania sp. SIO2D1
MTJ12972.1	60.684	5.44e-49	nitrogenase-associated protein	Anabaena sp. UHCC 0187
WP_161912950.1	59.055	6.56e-49	hypothetical protein	Methylostinus sp. C49
WP_015126499.1	62.393	6.98e-49	nitrogenase-associated protein	Calothrix sp. PCC 7507
WP_026610005.1	64.706	8.77e-49	hypothetical protein	Methylocaldum szegediense
WP_003610087.1	57.031	1.12e-48	MULTISPECIES: hypothetical protein	Methylostinus
NER79243.1	58.678	1.14e-48	nitrogenase-associated protein	Leptolyngbya sp. SIO1D8
WP_015210749.1	63.248	1.66e-48	nitrogenase-associated protein	Cylindrospermum stagnale
NER08466.1	60.684	1.82e-48	nitrogenase-associated protein	Okeania sp. SIO3C4
WP_149468965.1	58.915	1.85e-48	hypothetical protein	Roseomonas genomospecies 6
WP_015079413.1	60.684	2.18e-48	thioredoxin	Anabaena sp. 90
WP_145622405.1	58.140	2.27e-48	hypothetical protein	Azospirillum brasilense
WP_044431862.1	58.268	2.34e-48	hypothetical protein	Skermanella aerolata
WP_133718238.1	61.240	2.42e-48	hypothetical protein	Methylocaldum sp. 0917
NEQ71961.1	58.974	2.82e-48	nitrogenase-associated protein	Okeania sp. SIO2C9
WP_134077353.1	60.630	3.13e-48	hypothetical protein	Rhodovulum visakhapatnamense
MTJ55644.1	59.829	3.39e-48	nitrogenase-associated protein	Anabaena sp. UHCC 0253
NET45175.1	60.684	3.71e-48	nitrogenase-associated protein	Okeania sp. SIO2B3
WP_045368263.1	60.976	3.77e-48	hypothetical protein	Methyloceanibacter caenitepidi
WP_156879325.1	59.231	4.14e-48	hypothetical protein	Methylocaldum sp. BRCS4
NEP04475.1	60.684	4.28e-48	nitrogenase-associated protein	Okeania sp. SIO4D6
MTJ09605.1	58.974	4.40e-48	nitrogenase-associated protein	Anabaena sp. UHCC 0204
WP_019499581.1	61.538	6.36e-48	hypothetical protein	Pseudanabaena sp. PCC 6802
WP_069968843.1	60.169	6.97e-48	hypothetical protein	Desertifilum sp. IPPAS B-1220
MTJ25652.1	58.974	7.23e-48	nitrogenase-associated protein	Dolichospermum planctonicum UHCC 0167
WP_086135375.1	60.465	8.81e-48	hypothetical protein	Methylocaldum sp. SAD2
WP_083620450.1	59.829	8.86e-48	hypothetical protein	Planktothrix seta
WP_007357421.1	60.504	9.36e-48	MULTISPECIES: hypothetical protein	Kamptomena
WP_011613151.1	59.829	1.17e-47	nitrogenase-associated protein	Trichodesmium erythraeum
WP_017315443.1	61.538	1.19e-47	nitrogenase-associated protein	Mastigocladopsis repens
WP_097070467.1	58.871	1.47e-47	hypothetical protein	Rhodobacter maris
PPT11162.1	55.118	1.54e-47	Nitrogenase-associated protein <i>NifO</i>	Geitlerinema sp. FC II

(Continue)

WP_028083562.1	58.974	1.57e-47	nitrogenase-associated protein	Dolichospermum circinale
OBQ05589.1	59.829	1.58e-47	nitrogenase-associated protein	Anabaena sp. LE011-02
WP_016917984.1	57.692	1.61e-47	hypothetical protein	Methylocystis parvus
WP_112315187.1	60.630	1.63e-47	hypothetical protein	Rhodovulum viride
WP_096727879.1	62.393	1.88e-47	nitrogenase-associated protein	Nostoc carneum
WP_119507507.1	58.140	1.96e-47	hypothetical protein	Azospirillum brasilense
WP_039727361.1	58.120	2.26e-47	MULTISPECIES: hypothetical protein	Cyanobacteria
WP_075784184.1	59.843	2.27e-47	hypothetical protein	Rhodovulum sulfidophilum
WP_045873077.1	62.393	2.28e-47	MULTISPECIES: nitrogenase-associated protein	Nostocales

Reference: MACHADO *et al.*, (2020)

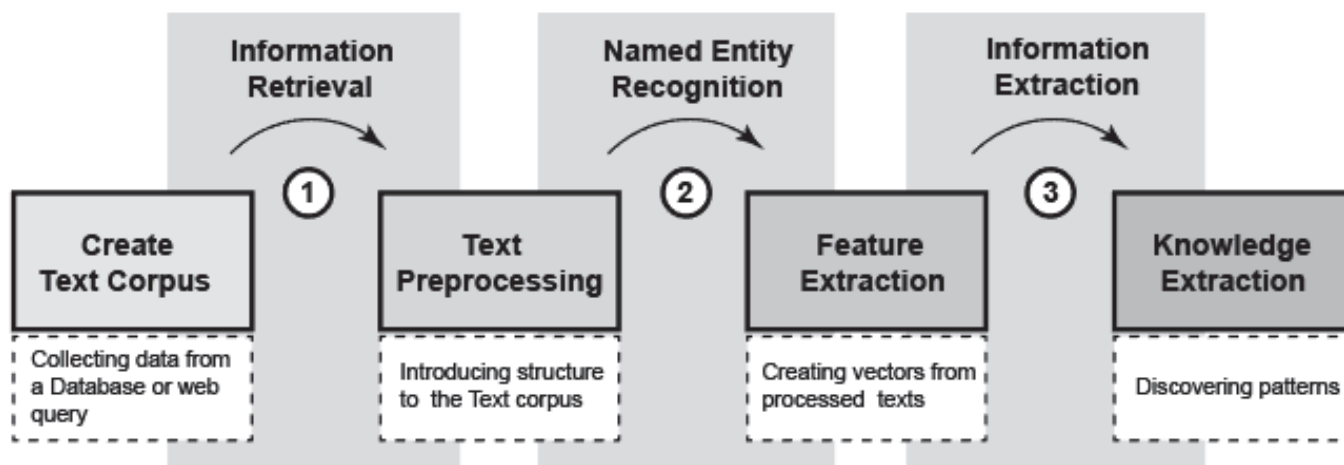
TABLE S5.5. Demonstration of the difference between minimal and expanded encodings

original text	hypothetical 15k protein - rice	MULTISPECIES: hypothetical protein
expanded encode	HYYPYQTHYETYICYALYSYDQYDIK YSPRYQTYEYINYSYKYSRYICYE	MYVLTYISPYECYIYESYPTYSHYYPYQTH YETYICYALYSPRYQTYEYIN
expanded decode	hypothetical 15k protein - rice	multispecies: hypothetical protein
minimal encode	HYYPYQTHYETYICYALYSYDYDKY SPRYQTYEYINYSYKYSRYICYE	MYVLTYISPYECYIYESYPYSHYYPYQTHY ETYICYALYSPRYQTYEYIN
minimal decode	hypothetical 99k protein - rice	multispecies. hypothetical protein

REFERENCE: Machado *et al.* (2021).

Note: The expanded version allows coding that maintains details of numbers and punctuation. Note change in numbers and punctuation mark in the minimal version, representing the information sensitivity loss.

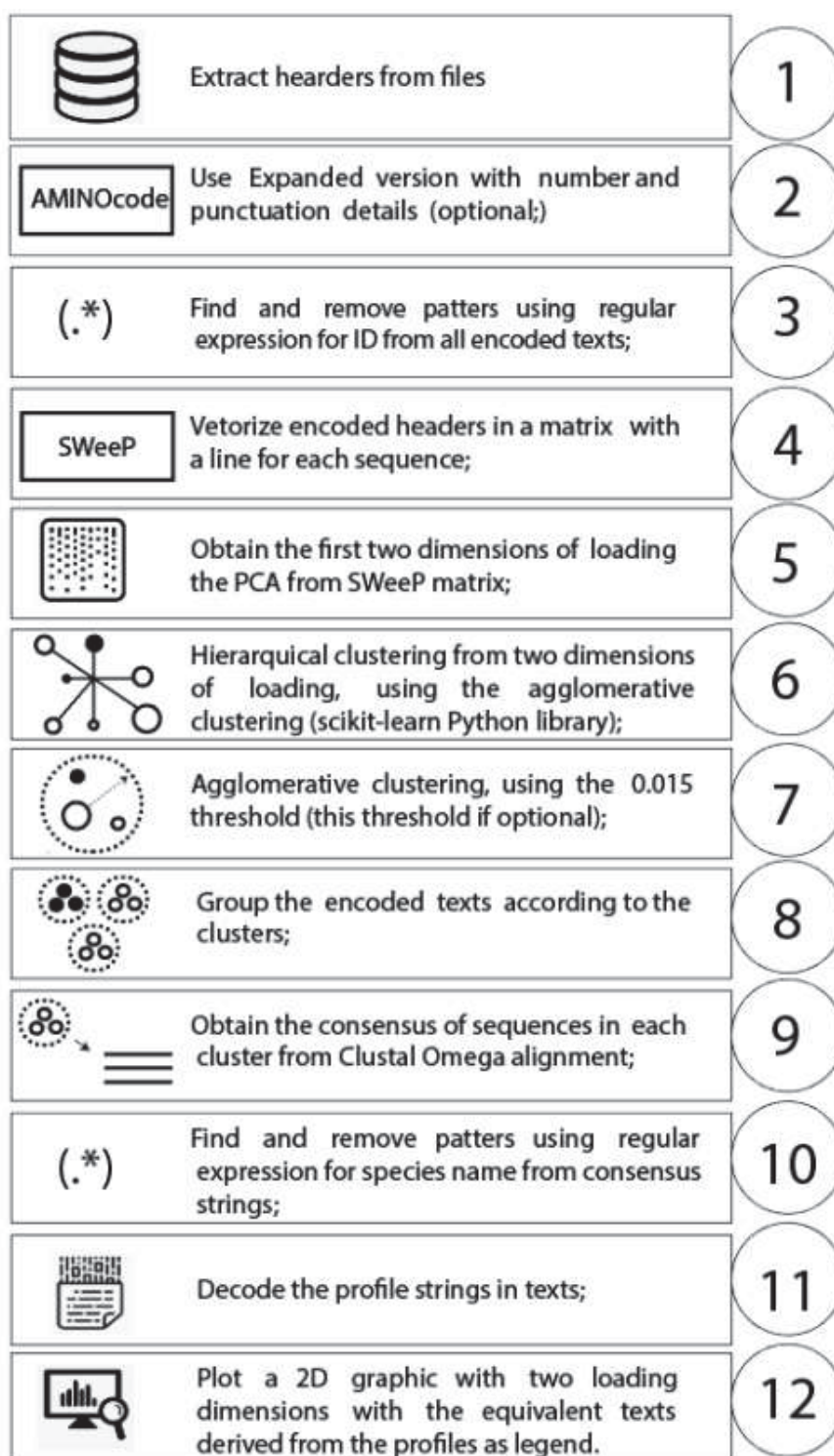
FIGURE S1.5. Text mining processes



REFERENCE: Authors, 2022.

Legend: (1) Information Retrieval (IR), performed to obtain documents relevant to a particular subject of interest; (2) Named Entity Recognition (NER), which analyzes the occurrence of specific keywords in documents; (3) Information extraction (IE).

FIGURE S2.5. BIOTEXT strategy

REFERENCE: Machado *et al.*, 2021.

FILE S1.5

Literature review details:

a) Description of the application;

b) Purpose for the field of biology, classified into 4 categories:

- General Purpose: Tools focused on describing/predicting events presented in texts in general, that can be applied for many purposes;
- Biocuration: Tools specialized in biocuration task, matching genes, proteins and associating them with functions;
- Molecular function: Tools that describe/predict molecular events, associating genes/mutations/diseases/organisms to the query texts;
- Chemicals: Tools that mainly describe/predict events related to chemicals/drugs

b) TM Tasks, according to Fleuren and Alkema (2015), which are:

- Information Retrieval (IR)
 - Named Entity Recognition (NER)
- Information Extraction (IE);

d) Input:

- Lists (genes, terms)
 - Query (Keywords, sentences or concepts)
 - Abstracts (generally, in PubMed/MEDLINE format)
- Full-article and Texts (XML format);

e) Output: Each tool has its own output, depending on the type of analysis. In most cases, it consists on:

- List of genes/terms/abstracts
- Relationships between concepts/terms/genes
- Interaction/Graphical networks
- Preprocessed/tagged documents
- Statistics, predictions, matrices, scores;

f) Authors and year of publication, from 1999 to 2022.

CAPÍTULO VI: O MODELO EGenera PARA CLASSIFICAÇÃO TAXONÔMICA AUTOMÁTICA DE BACTÉRIAS

1. INTRODUÇÃO

A classificação taxonômica é baseada em um sistema hierárquico que infere as relações naturais entre organismos (ROSELLÓ-MORA; WHITMAN, 2019), no entanto, não há uma abordagem consenso para realizar esta tarefa. Vários sistemas de classificação e nomenclatura têm sido utilizados, tornando as análises de gêneros imprecisas (PARKS *et al.*, 2018). A dificuldade na classificação taxonômica é observada principalmente no domínio Bacteria, onde a taxonomia ainda é fracamente baseada em informações moleculares (PARKS *et al.*, 2108; ROSSELLÓ-MORA; WHITMAN, 2019; HUGENHOLTZ *et al.*, 2021), pois grande parte das classificações de microrganismos ainda são realizadas apenas por meio da análise do gene rRNA 16S.

O gene 16S rRNA pode ser idêntico ou altamente homólogo entre alguns gêneros e espécies diferentes, o que torna difícil as classificações de nível mais alto, como gênero e filo (WHITMAN, 2016; GWAK; RHO, 2020). Além disso, mesmo com um nível alto de identidade entre organismos da mesma espécie, não é possível diferenciar divergência evolutiva recente, devido ao alto grau de conservação deste gene (JANDA; SHARON, 2007). Isso pode levar a taxonomias incorretas e, como consequência, gera um grande transtorno para os pesquisadores, pois podem gerar inconsistências em uma variedade de análises (PARKS *et al.*, 2018).

Mesmo com um importante legado de supervisão humana em geral para a taxonomia de bactérias, a classificação dos organismos nem sempre corresponde aos agrupamentos inferidos. Essa imprecisão tem sido relatada em estudos recentes (HUGENHOLTZ *et al.*, 2016; PARKS *et al.*, 2018; NOUIOUI *et al.*, 2018; ROSSELLÓ-MÓRA; WHITMAN, 2019; MARGOS *et al.*, 2020; MURRAY *et al.*, 2020; HUGENHOLTZ *et al.*, 2020; HUGENHOLTZ *et al.*, 2021; RAITTZ *et al.*, 2021). Existem diversos procariotos com nomenclatura incorreta depositadas em bancos de dados públicos, e muitas delas permanecem sem revisão até o momento (SCHOCH *et al.*, 2020). Além disso, não há uma abordagem padronizada para atribuir classificações taxonômicas. Sem um modelo taxonômico de consenso, a classificação torna-se imprecisa e redundante (ROSELLÓ-MORA; WHITMAN, 2019).

Recentemente, Parks e colaboradores (2018) propuseram uma taxonomia bacteriana, *Genome Taxonomy Database* (GTDB), baseada na análise filogenética de 120 proteínas codificadas por genes universais, incluindo metagenomas para aumentar a diversidade de espécies bacterianas. O GTDB baseia-se na divergência evolutiva relativa (RED), para delinear táxons de classificação mais alta, e identidade nucleotídica média (ANI) para delinear agrupamentos de espécies (PARKS *et al.*, 2018; PARKS

et al., 2022). Entretanto, o GTDB utiliza métodos baseados em alinhamento que podem resultar em instabilidade nas árvores filogenéticas geradas. Isso se deve ao fato de que, ao inserir novos genomas profundamente enraizados, ocorrerá modificações nos padrões de ramificação ao longo da árvore, o que implicaria em uma mudança na nomenclatura de táxon (CEVALLOS; DEGLI ESPOSTI, 2022).

Dessa forma, a falta de ferramentas para integrar dados de informações biológicas e taxonômicas traz à tona as fragilidades dos modelos taxonômicos existentes. Aqui, propomos o EGenera, um modelo de suporte à taxonomia bacteriana que busca contemplar as regras gerais da expertise humana e explorar os recursos disponíveis de dados moleculares, realizando uma análise global do domínio Bacteria por meio de técnicas de aprendizado de máquina. Nossa proposta é baseada em aprendizado supervisionado e não supervisionado, tendo como elemento principal a representação vetorial de padrões de proteomas bacterianos (DE PIERRI *et al.*, 2020; RAITTZ *et al.*, 2021). Por meio da análise automática de agrupamentos de gêneros, mostramos que é possível incorporar à taxonomia bacteriana o conhecimento molecular a priori decorrente do sequenciamento molecular, fornecendo evidências de que a máquina é capaz de aprender com a experiência humana.

2. MÉTODOS

O modelo EGenera é baseado em aprendizado supervisionado e não supervisionado. O método foi implementado em linguagem de programação MATLAB, e os processos foram executados em uma máquina com processador i5 e 16Gb RAM.

2.1 Conjunto de dados de origem

Para explorar o modelo EGenera, utilizamos como conjunto de dados origem (*W*) previamente descrito no estudo de De Pierri *et al.* (2020). Este conjunto é composto por 10.324 proteomas bacterianos, obtidos do NCBI (download em julho, 2019), onde para cada bactéria deste conjunto, as proteínas foram concatenadas formando um proteoma.

Os proteomas foram representados em vetores, utilizando o SWeeP (DE PIERRI *et al.*, 2020), com tamanho de projeção 1369 - tamanho do vetor com redução de dimensionalidade - e máscara "11011" - K-mer espaçado, de tamanho 5.

2.2 Avaliação de desempenho do modelo

Para avaliar o desempenho do modelo EGenera, foram aplicadas as métricas de Acurácia, Precisão, Recall e F1Score, conforme as equações a seguir:

$$Acurácia = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

$$Precisão = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$F1\ Score = \frac{Precisão \cdot Recall}{Precisão + Recall} \quad (4)$$

onde TP é um verdadeiro positivo, FP é um falso positivo, TN é um verdadeiro negativo e FN é um falso negativo.

3. RESULTADOS

Os resultados são divididos em 3 etapas principais (1) Extração de características dos dados de origem (2) Treinamento do Ensemble, que discorre sobre a estruturação do conjunto de treinamento utilizado no aprendizado supervisionado e (3) Aprendizado não supervisionado, que apresenta os critérios da tomada de decisão do Ensemble e a clusterização.

3.1 Extração de características e Aprendizado de Máquina

Para cada instancia W_i no conjunto de dados W , que representa proteomas, são selecionadas outras 250 instancias (WQ) por meio de sorteios tendenciosos baseados em distâncias euclidianas. Assim, para cada W_i foram selecionados outros 250 proteomas WQ , de acordo com um viés que determina maior chance de selecionar instâncias de W que estejam mais próximos vetorialmente de W_i . Definimos Wq como uma instância específica em WQ .

Para comparação entre W_i e Wq foram utilizadas 7 medidas (TABELA 1.6) que foram definidas como atributos para o aprendizado de máquina derivando em um conjunto de características (Wf), contendo 2.500.000 padrões de comparações entre proteomas, ou seja, 250 instâncias para cada proteoma (W_i) do conjunto de dados de origem W .

TABELA 1.6. Atributos considerados para o aprendizado de máquina

Medida	Descrição	Fórmula
Distância Euclideana	A raiz quadrada da soma dos quadrados das diferença das coordenadas correspondentes nos dois vetores.	$D(x, y) = \sqrt{\sum_{i=1}^n (y_i - x_i)^2}$
Cosseno entre os vetores	O cosseno do ângulo entre dois vetores; corresponde ao produto escalar dos vetores dividido pelo produto de seus módulos.	$\cos \theta = \frac{\vec{u} \cdot \vec{v}}{ \vec{u} \vec{v} }$
Quadrado da diferença entre as projeções recíprocas	Cálculo de acordo com a fórmula ao lado;	$\left(\left(\frac{\vec{v} \cdot \vec{u}}{u} \right) - \left(\frac{\vec{v} \cdot \vec{u}}{v} \right) \right)^2$
Relação entre as normas	Sendo $\vec{u}$ correspondente a W_i , de acordo com a fórmula ao lado;	$\frac{ \vec{v} + 1}{ \vec{u} + 1}$
Projeção de W_i em W_q	Sendo $\vec{u}$ correspondente a W_i , de acordo com a fórmula ao lado;	$\left(\vec{u} \cdot \frac{\vec{v}}{v} \right)$
Norma de W_i	Comprimento do vetor W_i	$ \vec{u} $
Norma de W_q	Comprimento do vetor W_q	$ \vec{v} $
Classe	---	(1) = Verdadeiro, (0) = Falso

Fonte: Autora, 2022.

3.2 Treinamento do ensemble

Foram utilizadas 100 amostras aleatórias contendo cada uma 20% de Wf para aprendizado supervisionado e a construção de um ensemble (E) com 100 redes neuronais artificiais *multilayer perceptron* (MLP), com duas camadas intermediárias com 5 e 3 neurônios, respectivamente (FIGURA S1.6). As redes foram treinadas independentemente, com conjuntos diferentes de dados, obtidos na mesma proporção a partir do conjunto Wf .

Construímos então, uma matriz binária, Me , contendo decisão do ensemble para cada possibilidade de par de organismos em W . O critério para a definição do conteúdo de uma posição em Me é a maioria de votos no conjunto de redes do Ensemble. O valor 1, portanto, quando os organismos relacionados ao par linha/coluna na matriz são do mesmo gênero e zero caso contrário.

O teste do aprendizado supervisionado é realizado no conjunto total Wf , e medidas estatísticas são apresentadas na (TABELA 2.6).

TABELA 2.6. Teste de Performance

Métrica	Resultado
Acurácia	0.99
F1Score	0.96
<i>Recall</i>	0.93
Precisão	0.98

Fonte: Autora, 2022.

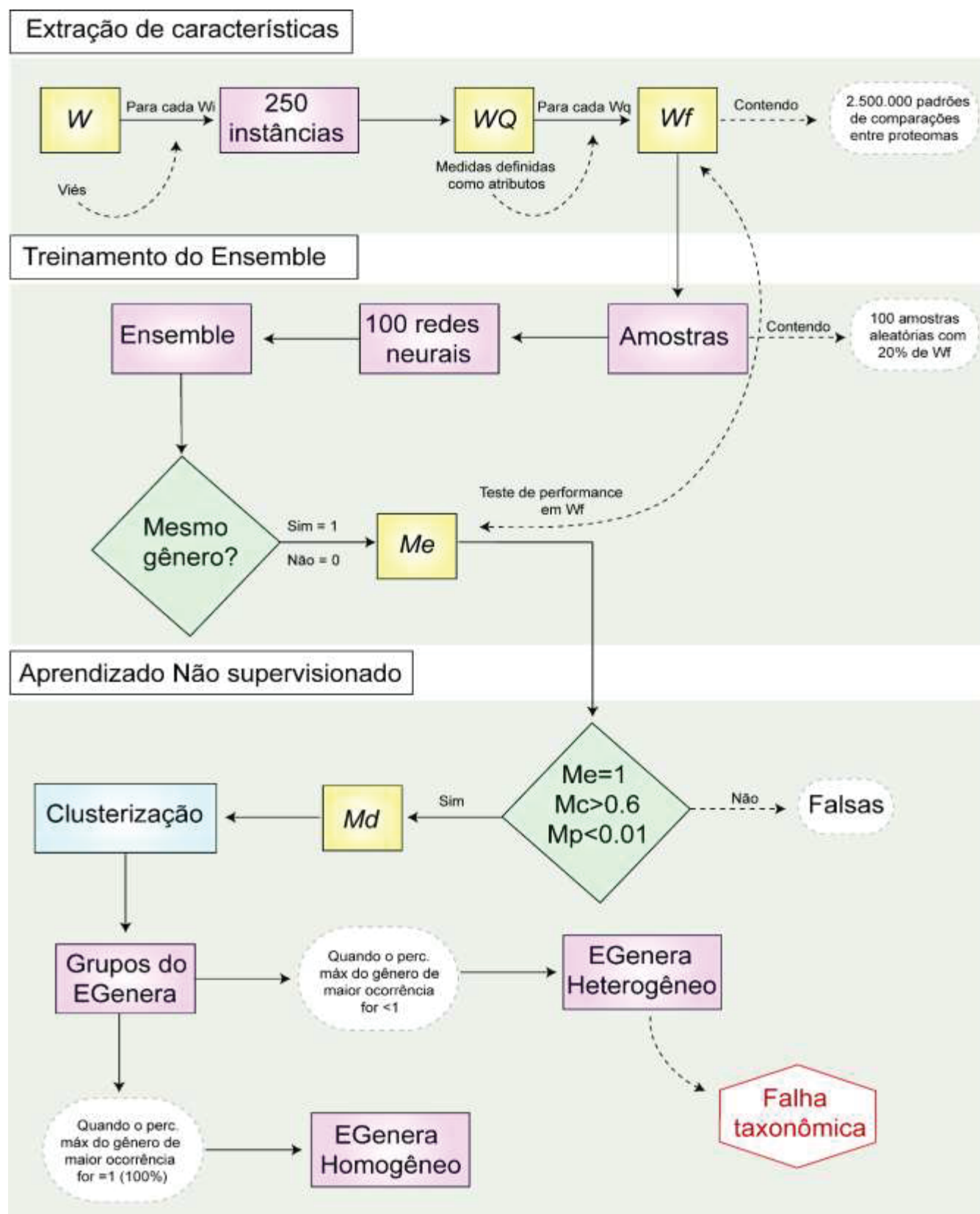
3.3 Aprendizado não supervisionado

Para explorar os resultados do ML obtidos no tópico anterior e definir uma estratégia de agrupamento de gêneros automática propomos uma etapa de aprendizado não supervisionado (clusterização) que reúne, além do critério relacionado à experiência humana resumida na matriz Me , outros critérios de consistência que são a alta correlação ($Mc = corr(Me)$) com os congêneres e com baixo p-valor (Mp). Para o passo seguinte do algoritmo, é gerada uma matriz binária (Md), onde valor de uma determinada posição é verdadeiro (1) para uma relação quando:

- ✓ Me for 1,
- ✓ a matriz de correlação Mc de Me for >0.6 e
- ✓ Mp for <0.01

As relações entre gêneros que não se enquadram nos parâmetros definidos em Md são setadas como falsas (0). Para realizar os agrupamentos foi desenvolvido um algoritmo específico inspirado em DBSCAN (ESTER *et al.*, 1996) e RAFTS3G (NICHIO *et al.*, 2019), usando como critério para agregação entre instâncias e as informações de Md . A Figura 1.6 apresenta uma visão global dos processos do EGenera.

Figura 1.6. Workflow de processos do EGenera



Fonte: Autora, 2022.

4. DISCUSSÃO

Este estudo identificou no conjunto original (*W*, 10k) 1614 EGenera, sendo 1123 composto por um só organismo. O modelo EGenera sugere a existência de problemas de classificação de gênero e taxa (detalhes disponíveis em <https://sourceforge.net/projects/egenera/>). Dentre 491 representantes de gêneros com pelo menos 2 organismos da lista inicial (10k), EGenera identificou 111 com possíveis inconsistências taxonômicas. A Tabela 3.6 apresenta os organismos cujo percentual de organismos no gênero de máxima ocorrência em cada EGenera inferior à 1, ou seja, não são completamente homogêneos em termos de gêneros anotados. Importante destacar que estes achados são entendidos como falhas no conjunto de dados de origem, tais como, erros de nomenclatura, classificação e anotação.

Geramos uma árvore filogenética pelo método *Neighbor-joining* (SAITOU; NEI, 1987) com os 491 representantes de gêneros (não singletons) para compararmos com os dados da Tabela 3.6. Como já apontado em estudos recentes (GARRITY, 2016; ROSSELLÓ-MÓRA, WHITMAN, 2018; HUGENHOLTZ *et al.*, 2021); existem diversas falhas envolvendo a classificação baseada apenas em rRNA 16S no domínio Bacteria, e muitos gêneros são complicados de serem identificados.

TABLE 3.6. Resultados do EGenera

Índice do representante em <i>W</i>	Gênero representativo do EGenera	Índice no EGenera	Número de organismos no EGenera	Número de gêneros diferentes no grupo	Percentual max. gênero/ EGenera
4970	<i>Bacillus</i>	55	385	2	0.9974026
3362	<i>Pseudomonas</i>	9	211	2	0.9952607
7188	<i>Staphylococcus</i>	30	379	2	0.9920844
6799	<i>Corynebacterium</i>	469	112	2	0.9910714
4340	<i>Bordetella</i>	433	406	2	0.9876847
3525	<i>Francisella</i>	67	67	2	0.9850746
7217	<i>Klebsiella</i>	88	299	3	0.9799331
3386	<i>Streptomyces</i>	83	103	3	0.9708738
3033	<i>Sinorhizobium</i>	272	32	2	0.96875
7930	<i>Methylobacterium</i>	278	22	2	0.9545455
2726	<i>Chlamydia</i>	6	147	2	0.952381
3187	<i>Enterobacter</i>	234	79	2	0.9367089
2874	<i>Halomonas</i>	426	14	2	0.9285714
4455	<i>Corynebacterium</i>	100	28	2	0.9285714
7317	<i>Brucella</i>	21	91	4	0.9230769
3818	<i>Cronobacter</i>	257	12	2	0.9166667
7858	<i>Caulobacter</i>	14	11	2	0.9090909
4358	<i>Sphingopyxis</i>	183	11	2	0.9090909
139	<i>Synechocystis</i>	81	10	2	0.9

(Continua)

3150	<i>Thermoanaerobacter</i>	19	9	2	0.8888889
4932	<i>Lysinibacillus</i>	261	9	2	0.8888889
3472	<i>Rickettsia</i>	18	58	2	0.8793103
6001	<i>Polynucleobacter</i>	235	8	2	0.875
3801	<i>Thermotoga</i>	59	15	2	0.8666667
6847	<i>Xanthomonas</i>	22	164	2	0.8536585
4212	<i>Mycobacterium</i>	471	270	6	0.8481481
7950	<i>Thermus</i>	47	13	3	0.8461538
5519	<i>Synechococcus</i>	284	6	2	0.8333333
3754	<i>Agrobacterium</i>	444	6	2	0.8333333
7798	<i>Planktophila</i>	1399	16	2	0.8125
1250	<i>Ruminiclostridium</i>	225	5	2	0.8
5187	<i>Hafnia</i>	850	5	2	0.8
756	<i>Geobacillus</i>	84	29	3	0.7931034
4728	<i>Pantoea</i>	407	22	2	0.7727273
6793	<i>Mycoplasma</i>	49	140	5	0.7714286
667	<i>Oligotropha</i>	151	4	2	0.75
5929	<i>Clostridium</i>	246	4	2	0.75
7100	<i>Prosthecochloris</i>	302	4	2	0.75
2822	<i>Dyella</i>	650	4	2	0.75
2583	<i>Saccharibacteria</i>	807	4	2	0.75
8801	<i>Enterobacteriaceae</i>	1002	4	2	0.75
3812	<i>Microcystis</i>	111	8	2	0.75
4256	<i>Arthrobacter</i>	205	24	4	0.75
7996	<i>Phaeobacter</i>	136	48	10	0.75
8726	<i>Elizabethkingia</i>	355	35	3	0.7428571
5438	<i>Kosakonia</i>	716	7	2	0.7142857
6009	<i>Virgibacillus</i>	875	7	3	0.7142857
3297	<i>Borrelia</i>	63	40	3	0.7
627	<i>Chlorobaculum</i>	17	3	2	0.6666667
117	<i>Ralstonia</i>	75	3	2	0.6666667
191	<i>Hydrogenobacter</i>	115	3	2	0.6666667
618	<i>Sulfurihydrogenibium</i>	297	3	2	0.6666667
3983	<i>Endomicrobium</i>	484	3	2	0.6666667
2037	<i>Paraburkholderia</i>	522	3	2	0.6666667
1509	<i>Thermodesulfobacterium</i>	623	3	2	0.6666667
6028	<i>Synechococcus</i>	757	3	2	0.6666667
2162	<i>Echinicola</i>	780	3	2	0.6666667
4012	<i>Streptococcus</i>	195	676	7	0.6642012
4500	<i>Achromobacter</i>	511	17	2	0.6470588
2124	<i>Cyanobacterium</i>	773	5	2	0.6
5663	<i>Acidovorax</i>	215	19	7	0.5789474
78	<i>Propionibacterium</i>	54	21	2	0.5714286

(Continua)

5411	<i>Gluconobacter</i>	132	9	4	0.5555556
6324	<i>Thermosipho</i>	244	9	2	0.5555556
5	<i>Yersinia</i>	264	141	3	0.5390071
8401	<i>Escherichia</i>	2	982	5	0.5213849
2622	<i>Thermocrinis</i>	62	2	2	0.5
1421	<i>Trichormus</i>	80	2	2	0.5
1234	<i>Calditerrivibrio</i>	118	2	2	0.5
8220	<i>Synechococcus</i>	125	2	2	0.5
257	<i>Pelagibacter</i>	140	2	2	0.5
872	<i>Herminiimonas</i>	176	2	2	0.5
9715	<i>Rhodobiaceae</i>	253	2	2	0.5
536	<i>Petrogoga</i>	274	2	2	0.5
698	<i>Cyanothece</i>	327	2	2	0.5
699	<i>cellulolyticum</i>	328	2	2	0.5
1815	<i>Terriglobus</i>	332	2	2	0.5
1014	<i>Burkholderia</i>	457	2	2	0.5
5737	<i>Bacillus</i>	463	2	2	0.5
1217	<i>Terriglobus</i>	528	2	2	0.5
1298	<i>Desulfurobacterium</i>	542	2	2	0.5
4874	<i>Pseudodesulfovibrio</i>	550	2	2	0.5
8897	<i>Devosia</i>	649	2	2	0.5
2123	<i>Halothece</i>	772	2	2	0.5
1019	<i>Paucimonas</i>	839	2	2	0.5
3326	<i>Coxiella</i>	928	2	2	0.5
8018	<i>Cedecea</i>	971	2	2	0.5
4022	<i>Parascardovia</i>	1014	2	2	0.5
9223	<i>Halomonas</i>	1225	2	2	0.5
6242	<i>Marivivens</i>	1273	2	2	0.5
9932	<i>Mikella</i>	1607	2	2	0.5
771	<i>Methylovorus</i>	354	4	2	0.5
9708	<i>Paraburkholderia</i>	493	4	2	0.5
1386	<i>Nostoc</i>	582	4	2	0.5
7231	<i>Oceanisphaera</i>	675	4	3	0.5
3640	<i>Fermentimonas</i>	962	4	3	0.5
9190	<i>Plantactinospora</i>	480	8	4	0.5
1013	<i>Cellulomonas</i>	366	7	3	0.4285714
4066	<i>Thauera</i>	321	14	5	0.4285714
8015	<i>Pasteurella</i>	11	110	4	0.4272727
5719	<i>Erythrobacter</i>	160	12	3	0.4166667
5240	<i>Haemophylus</i>	43	54	7	0.3703704
1929	<i>Blastococcus</i>	406	3	3	0.3333333
847	<i>Allochrocatium</i>	408	3	3	0.3333333
9365	<i>Limnobaculum</i>	1004	3	3	0.3333333

(Continua)

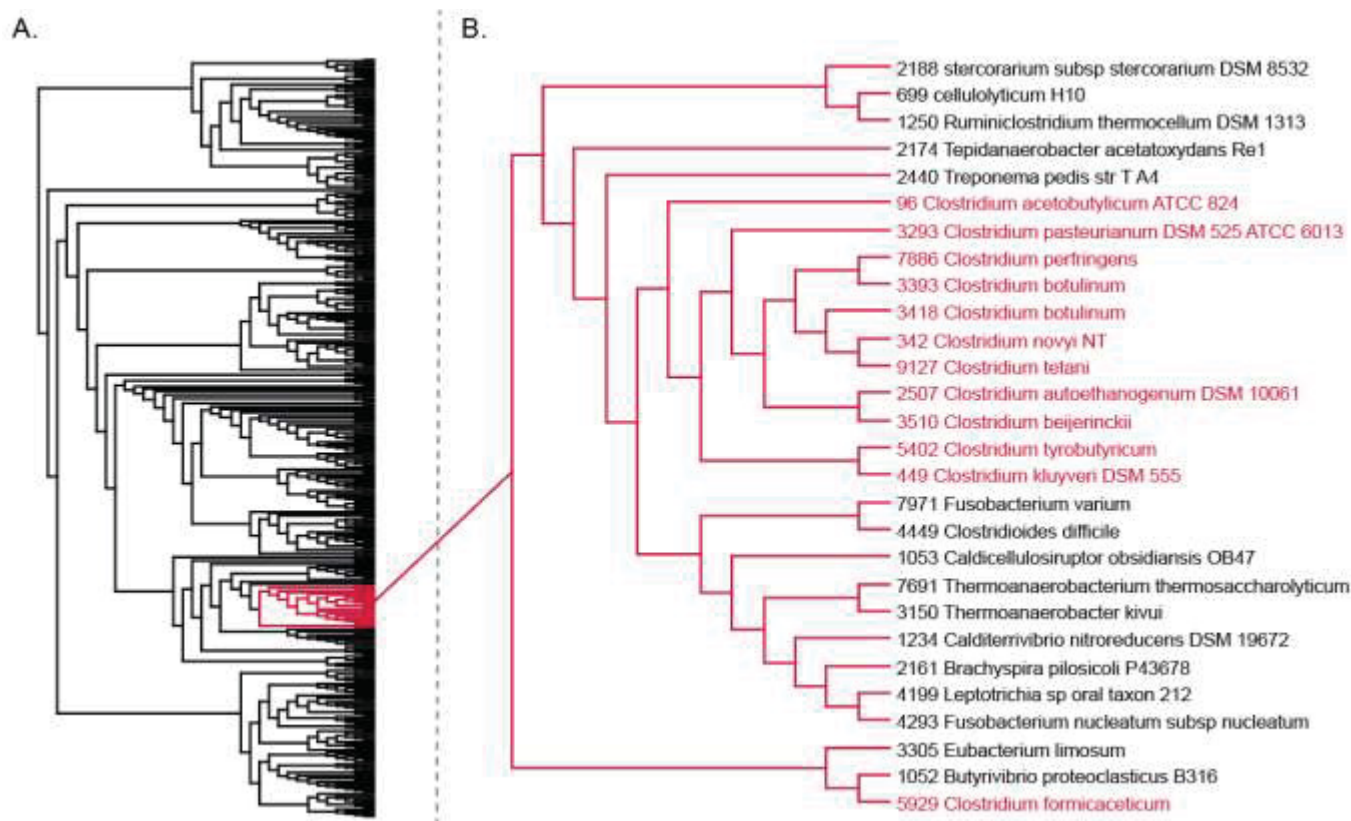
4559	<i>Roseateles</i>	1090	3	3	0.3333333
7250	<i>Cellulosimicrobium</i>	395	6	4	0.3333333
5261	<i>Microbacterium</i>	31	42	15	0.3095238
4278	<i>Thioglobus</i>	107	4	4	0.25
947	<i>Phytoplasma</i>	87	5	5	0.2
1106	<i>Polaribacter</i>	428	37	17	0.1621622

Fonte: Autora, 2022.

O gênero *Clostridium* tem sido extensivamente estudado e amplamente reconhecido como polifilético, desde então, muitas espécies têm sido realocadas para outros gêneros (COLLINS *et al.*, 1994; YUTIN *et al.*, 2013; LAWSON *et al.*, 2016; SEDLAR *et al.*, 2017; PRESCOTT; MACINNES; WU 2016; CRUZ-MORÁLES *et al.*, 2019; HAAS, BLANCHARD, 2020; KOBAYASHI *et al.*, 2021). Mesmo assim, até o momento não há padronização classificação destes organismos.

As dificuldades de classificação do gênero *Clostridium* foram apontadas pelo EGenera. A Figura 2.6 mostra que existem heterogeneidades filogenéticas no gênero; enquanto *C. acetobutylicum*, *C. pasteurianum*, *C. perfringens*, *C. botulinum*, *C. novyi*, *C. tetani*, *C. beijerinckii*, *C. tyrobutyricum*, *C. kluyveri* e *C. Autoethanogenum*, apesar de apresentar considerável distâncias taxonômicas entre si, de acordo com o EGenera, se encontram em um grupo maior consistente. Por outro lado, no entanto, o *C. formicaceticum* apresenta maior possibilidade de estar classificado de forma inconsistente. Em particular, o *C. formicaceticum* - identificado por Andreseen e colaboradores (1970) - não teve a classificação validada pelo *International Code of Nomenclature of Bacteria* (ICNB) e permanece sem revisão até o momento.

FIGURA 2.6. Representação do gênero *Clostridium* na árvore filogenética inferida a partir dos resultados do EGenera



Fonte: Autora, 2022.

Legenda: A. Árvore filogenética *Neighbour-Joining* inferida a partir dos resultados do EGenera, contendo 491 representantes de gênero. B. Em vermelho, o ramo ampliado contendo os representantes do gênero *Clostridium*.

De acordo com ALLARD *et al.* (2015), a classificação do gênero *Clostridium* deve ser revisitada e simplificada usando análises de genomas completos, combinando abordagens clássicas com filogenômicas. Neste sentido o EGenera apresenta uma proposta mais assertiva para a classificação taxonômica de bactérias em comparação à utilização o gene rRNA 16S, que poderia ser usado para a revisão da classificação do domínio bactéria de maneira geral.

Outro grupo que nos chama à atenção é o EGenera 321 – representado na Tabela 3.6 pelo gênero *Thauera* – e que está muito relacionado ao nosso estudo apresentado no Capítulo IV. Lá descrevemos a questão em aberto representada pela separação – ou não - entre os gêneros *Azoarcus* e *Aromatoleum*. Apresentamos também a relação próxima destes gêneros com o grupo dos *Thauera*, observando que existem alguns sobreposições em algumas classificações de organismos destes gêneros. Neste estudo constatamos que há um EGenera predominante (EGenera 321) que engloba a maioria dos organismos referenciados da lista analisada aqui (FIGURA 3.6). Desta forma, de acordo com EGenera, a questão seria resolvida pela integração dos gêneros em foco incluindo *Thauera*.

FIGURA 3.6. Representação do gênero *Thauera* na árvore filogenética inferida a partir dos resultados do EGenera

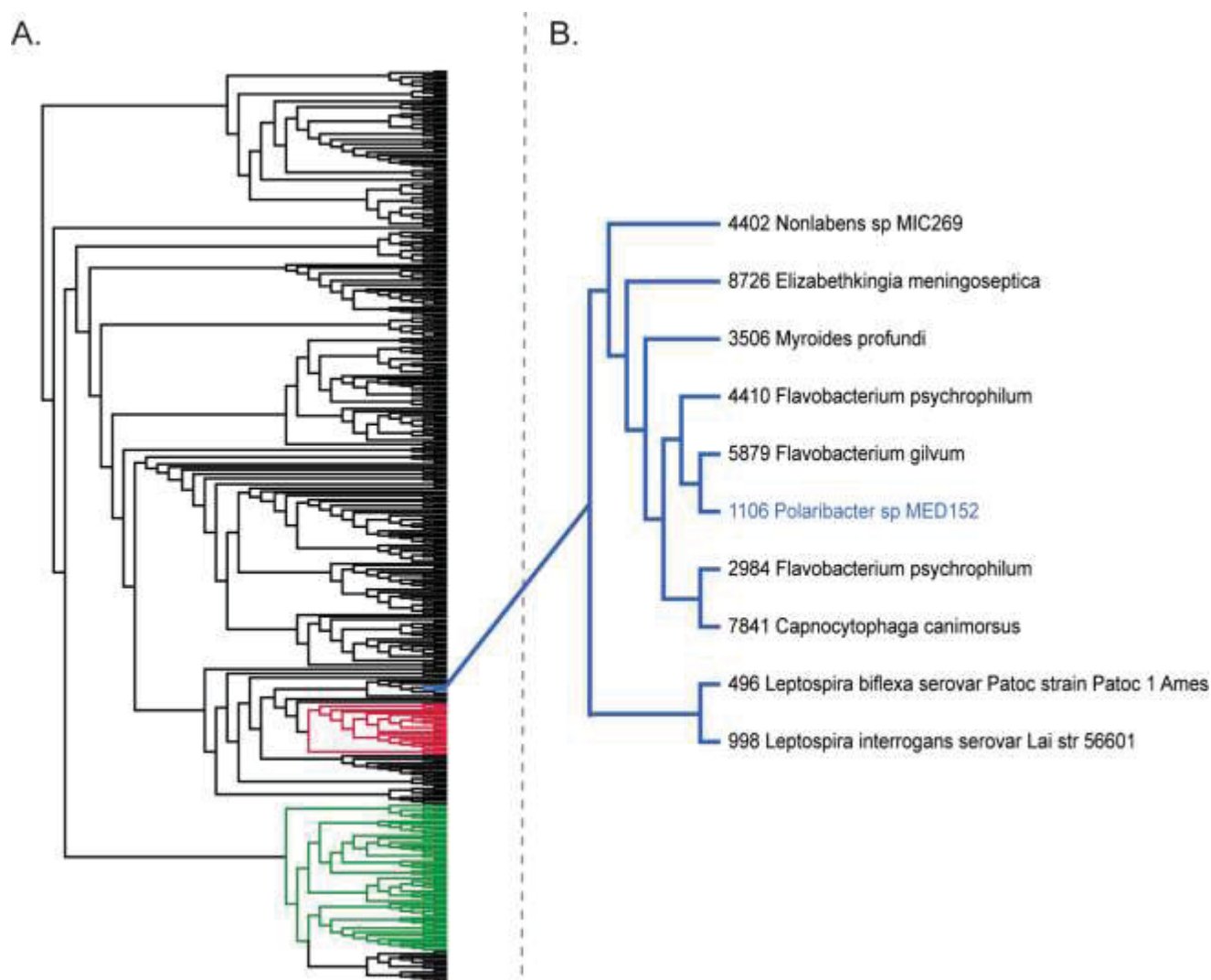


Fonte: Autora, 2022.

Legenda: A. Árvore filogenética *Neighbour-Joining* inferida a partir dos resultados do EGenera, contendo 491 representantes de gênero; em vermelho os organismos apresentados na Figura 3.5 B. Em verde, o ramo ampliado contendo os gêneros vizinhos do representante *Thauera humireducens*.

De acordo com o percentual máximo de EGenera, a maior discrepância de classificação entre os EGenera e os gêneros anotados ocorreu no grupo EGenera-428, representado pelo gênero *Polaribacter* (FIGURA 4.6). A *Polaribacter* sp. MED 152 foi identificado por Gonzales *et al.*, (2008) e o método de classificação foi baseada no gene rRNA 16S. A MED 152. apresenta alta similaridade com a *Polabribacter dokdonensis*, entretanto, foram identificadas diferenças fenotípicas importantes; ao contrário de *P. dokdonensis*, MED152 é β -galactosidase positivo e é capaz de degradar a gelatina (GONZALES *et al.*, 2008). Isso demonstra que a análise baseada apenas no gene rRNA 16S foi insuficiente para a classificação. *Polaribacter* é considerado um dos gêneros mais abundantes encontrados na vida marinha e por isso, deve ser estudado mais a fundo.

FIGURA 4.6. Representação do gênero *Polaribacter* na árvore filogenética inferida a partir dos resultados do EGenera



Fonte: Autora, 2022.

Legenda: A. Árvore filogenética *Neighbour-Joining* inferida a partir dos resultados do EGenera, contendo 491 representantes de gênero; em vermelho os organismos apresentados na Figura 3.5 e em verde os organismos apresentados na Figura 3.6 B. Em azul, o ramo ampliado contendo os gêneros vizinhos do representante *Polaribacter* sp. MED152.

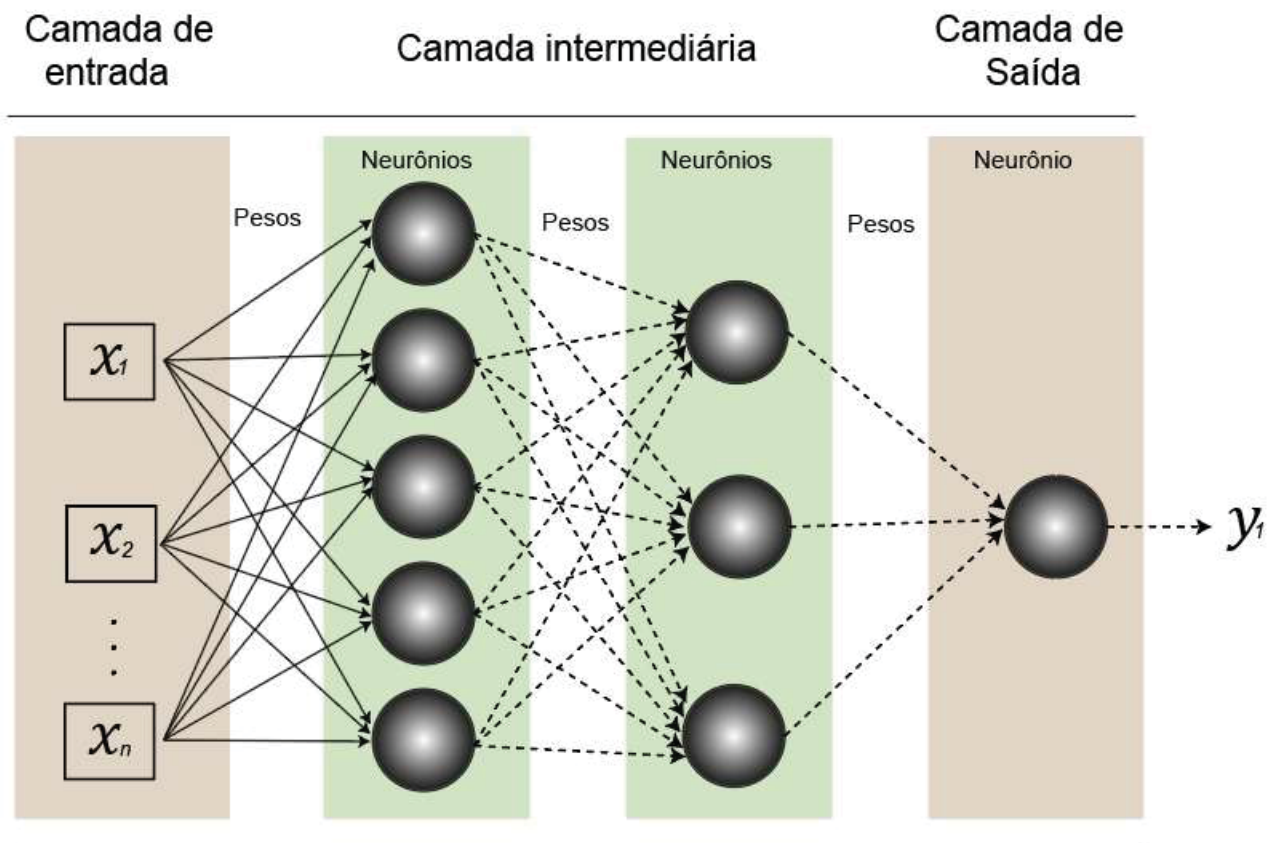
5. CONSIDERAÇÕES FINAIS

A maior parte dos gêneros do EGenera obtidos pelo aprendizado supervisionado, é consistente. O EGenera pode ajudar no rumo da consolidação da taxonomia, apresentando uma abordagem automatizada e universal para definir gêneros de bactéria. Por meio das redes treinadas do *Ensemble*, é possível tratar todos os organismos do mesmo modo, evitando discrepâncias de classificação geradas por diferentes métodos e análises. Ao incorporar um modelo humano para fazer taxonomia de bactérias, estudamos as consequências e sugerimos revisões de classificação.

Neste estudo, mostramos que o conhecimento acumulado pela expertise humana pode ser utilizado para treinar a máquina para fazer o processo de classificação taxonômica de forma rápida e robusta e possivelmente mais consistente. No entanto, ressaltamos que isso não significa que os modelos existentes não sejam válidos, mas sugerimos que as técnicas de classificação taxonômica atualmente disponíveis não apresentam um modelo consensual mesmo entre a comunidade de especialistas humanos e que, a criação de padrões taxonômicos consistentes pode facilitar os estudos comparativos entre bactérias.

MATERIAL SUPLEMENTAR

Figura S1.6. Estrutura da rede neuronal artificial utilizada para a construção das redes do Ensemble



Fonte: Autora, 2022.

Legenda: Rede neuronal artificial *Multi-layer Perceptron* (MLP), que consiste em: Uma camada de entrada, contendo os padrões de Wf ; camadas intermediárias, contendo 5 e 3 neurônios, respectivamente; e uma camada de saída, à qual os neurônios são conectados. Os pesos das conexões são calculadas durante o processo de treinamento, para que o algoritmo possa determinar a saída aproximada de cada camada, que será representada em Md .

REFERÊNCIAS

- ALLARD, G. *et al.* SPINGO: a rapid species-classifier for microbial amplicon sequences. *BMC bioinformatics*, v. 16, n. 1, p. 1-8, 2015.
- CEVALLOS, M. A.; DEGLI ESPOSTI, M. New Alphaproteobacteria Thrive in the Depths of the Ocean with Oxygen Gradient. *Microorganisms*, v. 10, n. 2, p. 455, 2022.
- COLLINS, M. D. *et al.* The phylogeny of the genus *Clostridium*: proposal of five new genera and eleven new species combinations. *International Journal of Systematic and Evolutionary Microbiology*, v. 44, n. 4, p. 812-826, 1994.
- CRUZ-MORALES, P. *et al.* Revisiting the evolution and taxonomy of Clostridia, a phylogenomic update. *Genome biology and evolution*, v. 11, n. 7, p. 2035-2044, 2019.
- ESTER, M. KRIEGEL, H.P.; SANDER, J.; XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **kdd**. 1996. p. 226-231.
- GARRITY G. M. A New Genomics-Driven Taxonomy of Bacteria and Archaea: Are We There Yet? **Journal of clinical microbiology**, v.54, 1956–1963, 2016.
- GONZÁLEZ, José M. *et al.* Genome analysis of the proteorhodopsin-containing marine bacterium *Polaribacter* sp. MED152 (Flavobacteria). *Proceedings of the National Academy of Sciences*, v. 105, n. 25, p. 8724-8729, 2008.
- HAAS, K. N.; BLANCHARD, J. L. Reclassification of the *Clostridium clostridioforme* and *Clostridium sphenoides* clades as *Enterocloster* gen. nov. and *Lacrimispora* gen. nov., including reclassification of 15 taxa. **International Journal of Systematic and Evolutionary Microbiology**, v. 70, n. 1, p. 23-34, 2020.
- HUGENHOLTZ, P., CHUVOCHINA, M., OREN, A., PARKS, D. H., SOO, R. M. Prokaryotic taxonomy and nomenclature in the age of big sequence data. **The ISME journal**, Advance online publication, 2021. <https://doi.org/10.1038/s41396-021-00941-x>
- KOBAYASHI, H. *et al.* Taxonomic status of the species *Clostridium methoxybenzovorans* Mechichi *et al.* 1999. **International Journal of Systematic and Evolutionary Microbiology**, v. 71, n. 8, p. 004951, 2021.
- KONSTANTINIDIS, K.T.; TIEDJE, J.M. Towards a genome-based taxonomy for prokaryotes. **Journal of bacteriology**, v. 187, n. 18, p. 6258-6264, 2005.
- LAWSON, P. A. *et al.* Reclassification of *clostridium difficile* as *clostridioides difficile* (Hall and O'Toole 1935) Prévot 1938. **Anaerobe**, v. 40, p. 95-99, 2016.
- NICHIO, B.T. L.; DE OLIVEIRA, A.M.R.; DE PIERRI, C.R.; SANTOS, L.G.C.; LEJAMBRE, A.Q.; VIALLE, R.A.; DA ROCHA COIMBRA, N.A.; GUIZELINI, D.; MARCHAUKOSKI, J.N.; PEDROSA, F.O.; *et al.* RAFTS(3)G: An efficient and versatile clustering software to analyses in large protein datasets. **BMC Bioinform**, 20, 392, 2019.
- PARKS D.H., CHUVOCHINA M, WAITE DW, RINKE C, SKARSHEWSKI A, CHAUMEIL PA, HUGENHOLTZ P. A. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. **Nat.Biotechnol**. 36, 996, 2018.
- PARKS, D.H., CHUVOCHINA, M., CHAUMEIL, PA. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. **Nat Biotechnol** v.38, p.1079–1086, 2020.
- PARKS, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. **Nucleic acids research**, v. 50, n. D1, p. D785-D794, 2022.
- PIERRI, C.R.; VOYCEIK, R.; SANTOS DE MATTOS, L.G.C.; KULIK, M.G.; CAMARGO, J.O.; REPULA DE OLIVEIRA, A.M.; DE LIMA NICHIO, B.T.; MARCHAUKOSKI, J.N.; DA SILVA FILHO, A.C. *et al.* SWEEP: Representing large biological sequences datasets in compact vectors. **Sci. Rep**, 10, 91, 2020.

PRESCOTT, John F.; MACINNES, Janet I.; WU, Anson KK. Taxonomic relationships among the clostridia. *Clostridial Diseases of Animals*. Ames, IA: Wiley-Blackwell, v. 2016, p. 3-5, 2016.

RAITZ, R. T., DE PIERRI, C. r., MALUK, M., BUENO BATISTA, M., CARMONA, M., JUNGHARE, M., FAORO, H., CRUZ, L. M., BATTISTONI, F., SOUZA, E., PEDROSA, F. O., CHEN, W. M., POOLE, P. S., DIXON, R. A., & JAMES, E. K. Comparative Genomics Provides Insights into the Taxonomy of *Azoarcus* and Reveals Separate Origins of *Nif* Genes in the Proposed *Azoarcus* and *Aromatoleum* Genera. **Genes (Basel)**, v.12(1), n.71, 2021.

ROSSELLÓ-MÓRA, R., WHITMAN, W. B. Dialogue on the nomenclature and classification of prokaryotes. **Systematic and applied microbiology**, v.42, 5–14, 2019.

SAITOU, N., NEI, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. **Molecular Biology and Evolution** v.4, p.406-425, **1987**.

SEDLAR, K. *et al.* Reclassification of non-type strain *Clostridium pasteurianum* NRRL B-598 as *Clostridium beijerinckii* NRRL B-598. **Journal of biotechnology**, v. 244, p. 1-3, 2017.

SCHOCH, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. **Database**, v. 2020, 2020

WHITMAN, William B. Modest proposals to expand the type material for naming of prokaryotes. **International Journal of Systematic and Evolutionary Microbiology**, v. 66, n. 5, p. 2108-2112, 2016.

YUTIN, N.; GALPERIN, M.Y. A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia. **Environmental microbiology**, v. 15, n. 10, p. 2631-2641, 2013.

CAPÍTULO VII: CONCLUSÃO

Nesta tese, exploramos e consolidamos o modelo de representação vetorial de proteomas – SweeP (DE PIERRI *et al.*, 2020). A representação e exploração dos proteomas bacterianos foi uma indispensável contribuição para a construção desta pesquisa (Capítulo III). Apresentamos um conjunto de dados de 10.324 proteomas bacterianos vetorizados pelo modelo SWeeP: Esta representação é consistente, podendo ser utilizada para análises de aprendizado de máquina de maneira geral.

Com o modelo SweeP é possível realizar a construção rápida e sensível de uma árvore filogenética compacta de proteomas bacteriano. Até onde sabemos, esta é a maior árvore de proteomas bacterianos construída de maneira automatizada até o momento. Além disso, o SWeeP também possibilitou a comparação dos proteomas completos por meio de heat maps e PCA. A representação de proteomas bacterianos que propomos neste estudo é inédita e facilita a manipulação e visualização dos dados, atendendo as principais demandas do Big Data na área das ciências ômicas, incluindo a agilidade no processamento.

A partir da árvore filogenética construída identificamos que existem questões taxonômicas a serem resolvidas em relação ao domínio Bacteria; diversos clados ainda não estão bem resolvidos. Constatamos que essas incosistências ocorrem principalmente devido à falta de acurácia dos métodos atuais utilizados para classificação taxonômica e falta de revisão – fatos que também foram levantados em estudos recentes (PARKS *et al.*, 2018; ROSSELLÓ-MÓRA; WHITMAN *et al.*, 2019; ZHU *et al.*, 2019; MURRAY *et al.*, 2020; MARGOS *et al.*, 2020; PARKS *et al.*, 2020; HUGENHOLTZ *et al.*, 2021; RAITTZ *et al.*, 2021; PARKS *et al.*, 2022).

Exploramos mais a fundo essa questão e observamos que abordagens de Mineração de Textos são estratégias *significativas* para realizar a curadoria textos científicos e, portanto, a associação dessas abordagens com a tecnologia vetorial do SweeP é *acertiva*. Pudemos observar pelo levantamento bibliográfico apresentado no Capítulo V que a MT é amplamente utilizada para a área das ciências ômicas, e que existem ferramentas que podem ser usadas para auxiliar no estudo da taxonomia (seja como ferramentas de suporte ou para realizar teste complementares). Em particular, identificamos 78 ferramentas que podem contribuir: (1) Fornecendo auxílio aos pesquisadores na melhor escolha da ferramenta de MT, de acordo com a classificação sugerida aqui (de acordo com a aplicação da ferramenta, propósito e contexto dentro de MT); (2) Atualizando o estado da arte da MT para as ciências ômicas. Contudo, notamos que mesmo com vários aspectos de TM bem resolvidos na literatura, ainda é possível implementar esses processos de maneira vantajosa.

Dessa forma, nesta pesquisa houve a colaboração no desenvolvimento de uma abordagem de

Mineração de texto, o BIOTEXT (MACHADO *et al.*, 2020 - submetido e em fase de revisão). O BIOTEXT tem como elemento fundamental a transposição direta das tecnologias de Bioinformática para a área de TM. Trata-se de um pacote para executar estratégias de MT que contém um núcleo de recursos original, com algoritmos inteiramente desenvolvidos *in-house* - AMINOcode, DNAbits (MACHADO *et al.*, 2021), SWeeP (DE PIERRI *et al.*, 2020) e HUFLUS (TABELA S3.5 – Capítulo V) - que visa a resolução de questões pontuais de Bioinformática. Mostramos que textos podem ser codificados em informações biológicas usando estratégias computacionais baseadas em bioinformática. Com isso, a tecnologia vetorial pôde ser explorada como recurso principal no processamento de textos e na comparação de sequências de maneira ágil e robusta.

Utilizamos as abordagens desenvolvidas como suporte na estruturação conceitual da classificação taxonômica de bactérias. O SWeeP proporcionou análises robustas e demonstrou que o uso de proteomas completos para análises filogenéticas e taxonômicas pode ser superior e mais efetivo que genomas ou 16S rRNA, evidenciados no estudo de caso dos gêneros *Azoarcus-Aromatoleum* apresentado (Capítulo IV). Foram identificadas questões taxonômicas novas nos gêneros *Azoarcus* e *Aromatoleum*, que ainda não haviam sido sugeridas na literatura.

As análises do SWeeP em conjunto com análises de cluster, core genoma, ANI e 16S (e análises de bancada que não são relatadas aqui, mas podem ser encontradas no artigo completo – (RAITZ *et al.*, 2021), sugeriram que a taxonomia dos gêneros *Azoarcus-Aromatoleum* deve ser revisada. A possível saída seria firmar o gênero *Azoarcus* e sugerir novos gêneros adicionais. O Capítulo IV forneceu evidências necessárias de que o SWeeP tem potencial para auxiliar em análises taxonômicas de bactérias. Importante destacar que foram utilizados a base de dados de cluster e o RAFTS3G (NICHIO *et al.*, 2019), desenvolvidos pelo grupo, além de serem criados algoritmos e novos scripts com aplicação específica em análises taxonômicas.

Com base nesses achados, criamos um modelo, EGenera, para validação da taxonomia bacteriana de maneira automatizada, evidenciados no Capítulo VI. Esta última etapa permitiu realizar um panorama do domínio bacteria por meio de abordagens de aprendizado supervisionado e não supervisionado, contextualizando aspectos históricos, temporais e taxonômicos. O EGenera contempla as regras gerais da expertise humana e explora os recursos disponíveis de dados moleculares, mostrando que a máquina é capaz de aprender com a experiência humana. Através da análise automática de agrupamentos de gêneros, é possível incorporar o conhecimento a priori e o conhecimento em nível molecular na taxonomia bacteriana.

Por fim, fomos capazes de responder à questão norteadora desta tese: Seria possível a associação das abordagens de MD, MT e Inteligência Artificial substituir a expertise humana para validação da taxonomia bacteriana? Sugerimos que sim, que é possível tratar as classificações taxonômicas da mesma

forma sem os vieses de interpretação gerados por diferentes métodos e análises. A partir da incorporação de um modelo humano para fazer taxonomia, foi possível treinar a máquina para fazer o processo de classificação taxonômica de forma rápida e consistente e adaptado à visão taxonômica atual.

Entendemos que ainda há trabalho pela frente na consolidação do EGenera. Entretanto, a maior parte dos resultados obtidos ao longo desta tese estão robustos e com amplo potencial para aplicação nas ciências ômicas. A representação vetorial de informações é usada universalmente e é bem definida na maioria das áreas. Propomos esta nova abordagem para que os recursos matemáticos e computacionais disponíveis possam ser estendidos e aplicados mais facilmente na análise e mineração de sequências biológicas e textos científicos.

REFERÊNCIAS

- DE PIERRI, C.R.; VOYCEIK, R.; SANTOS DE MATTOS, L.G.C.; KULIK, M.G.; CAMARGO, J.O.; REPULA DE OLIVEIRA, A.M.; DE LIMA NICHIO, B.T.; MARCCHAUKOSKI, J.N.; DA SILVA FILHO, A.C. *et al.* SWeeP: Representing large biological sequences datasets in compact vectors. **Sci. Rep.**, 10, 91, 2020.
- HUGENHOLTZ, P., CHUVOCHINA, M., OREN, A., PARKS, D. H., SOO, R. M. Prokaryotic taxonomy and nomenclature in the age of big sequence data. **The ISME journal**, Advance online publication, 2021. <https://doi.org/10.1038/s41396-021-00941-x>
- MACHADO, D.J.S.; DE PIERRI, C.R.; SANTOS, L.G.C.; PEDROSA, F.O.; RAITTZ, R.T. BIOTEXT: Exploiting Biological-like format for text mining. *BioRxiv* 2021.04.08.439078. doi: <https://doi.org/10.1101/2021.04.08.439078>.
- MARGOS, G., FINGERLE, V., CUTLER, S., GOFTON, A., STEVENSON, B., ESTRADA-PEÑA, A. Controversies in bacterial taxonomy: The example of the genus *Borrelia*. **Ticks and tick-borne diseases**, v.11, 2020.
- MURRAY, A. E., FREUDENSTEIN, J., GRIBALDO, S., HATZENPICHLER, R., HUGENHOLTZ, P., KÄMPFER, P., KONSTANTINIDIS, K. T., LANE, C. E., PAPKE, R. T., PARKS, *et al.* Roadmap for naming uncultivated Archaea and Bacteria. **Nature microbiology**, n.5, v.8, 987–994, 2020.
- NICHIO, B.T. L.; DE OLIVEIRA, A.M.R.; DE PIERRI, C.R.; SANTOS, L.G.C.; LEJAMBRE, A.Q.; VIALLE, R.A.; DA ROCHA COIMBRA, N.A.; GUIZELINI, D.; MARCCHAUKOSKI, J.N.; PEDROSA, F.O.; *et al.* RAFTS(3)G: An efficient and versatile clustering software to analyses in large protein datasets. **BMC Bioinform.**, 20, 392, 2019.
- PARKS D.H., CHUVOCHINA M, WAITE DW, RINKE C, SKARSHEWSKI A, CHAUMEIL PA, HUGENHOLTZ P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. **Nat. Biotechnol.** 36, 996, 2018.
- PARKS, D.H., CHUVOCHINA, M., CHAUMEIL, PA. *et al.* A complete domain-to-species taxonomy for Bacteria and Archaea. **Nat Biotechnol** v.38, p.1079–1086, 2020.
- PARKS, D. H. *et al.* GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. **Nucleic acids research**, v. 50, n. D1, p. D785–D794, 2022.
- RAITTZ, R. T., DE PIERRI, C. r., MALUK, M., BUENO BATISTA, M., CARMONA, M., JUNGHARE, M., FAORO, H., CRUZ, L. M., BATTISTONI, F., SOUZA, E., PEDROSA, F. O., CHEN, W. M., POOLE, P. S., DIXON, R. A., & JAMES, E. K. Comparative Genomics Provides Insights into the Taxonomy of *Azoarcus* and Reveals Separate Origins of *Nif* Genes in the Proposed *Azoarcus* and *Aromatoleum* Genera. **Genes (Basel)**, v.12(1), n.71, 2021.
- ROSSELLÓ-MÓRA, R., WHITMAN, W. B. Dialogue on the nomenclature and classification of prokaryotes. **Systematic and applied microbiology**, v.42, 5–14, 2019.
- ZHU, Q., MAI, U., PFEIFFER, W., JANSSEN, S., ASNICAR, F., SANDERS, J. G., BELDA-FERRE, P., ALGHALITH, G. A., KOPYLOVA, E., MCDONALD, D. *et al.* Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. **Nat. Comm.**, v.10, n. 5477, 2019.

