

UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA MICROBIANA

SAMUEL GALVÃO ELIAS

**Ampliando o Potencial dos Dados Genômicos: Um Estudo
sobre o Enriquecimento de Metadados e a Classificação
Filogenética de Sequências Microbianas**

Brasília

2024

UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA MICROBIANA

SAMUEL GALVÃO ELIAS

**Ampliando o Potencial dos Dados Genômicos: Um Estudo
sobre o Enriquecimento de Metadados e a Classificação
Filogenética de Sequências Microbianas**

Tese apresentada ao Programa Pós-Graduação em Biologia Microbiana da Universidade de Brasília, como requisito à obtenção do título de Doutor em Biologia Microbiana.

Orientador: Helson Mário Martins do Vale

Brasília

2024

[Página reservada à ficha catalográfica]

Dr.(a) Helson Mario Martins do Vale, UnB
Presidente

Dr.(a) Gabriel Sergio Costa Alves, UnB
Examinador(a) Interno(a)

Dr.(a) Frederico Schmitt Kremer, UFPEL
Examinador(a) Externo(a) à Instituição

Dr.(a) José Miguel Ortega, UFMG
Examinador(a) Externo(a) à Instituição

Dr.(a) André Barros de Sales, UnB
Suplente

Aos meus avós Galdino e Elza e meu
irmão Mateus, *in memoriam*

Agradecimentos

Durante essa incrível jornada de estudos e dedicação, tive o prazer de conhecer grandes personas que tornaram esse incrível momento, possível. Primeiramente agradeço aos meus pais, Wânia Reis Galvão Elias e Evaldo Colossi Elias, os quais durante toda a minha vida me prepararam e incentivaram a conquistar o que sempre desejei com força e determinação.

Agradeço imensamente a minha esposa e inspiração, Dra. Débora Cervieri Guterres, pelo apoio incondicional durante todos os momentos que precisei me ausentar para me dedicar a construção dessa tese. Agradeço pela compreensão nas incontáveis e introspectiva manhãs, tardes, noites e madrugadas as quais precisei dedicar meu tempo na construção deste momento. Além da compreensão e apoio, agradeço por me ensinar sobre um dos grupos de organismos mais incríveis de já conheci, os fungos filacoróides.

Ao professor Dr. Helson Mário Martins do Vale, agradeço pela orientação e ensinamentos preciosos, sem os quais esse momento não seria possível. Ainda por me proporcionar uma visão engrandecedora sobre a dimensão de um dos maiores biomas brasileiros, o Cerrado.

Ao professor Dr. José Carmine Dianese, pelas parcerias científicas e principalmente pela oportunidade de aprender muito sobre micologia do planalto Central, e ainda me proporcionar acesso a uma das coleções mais incríveis de fungos fitopatogênicos brasileiras.

A pessoas como, Dr. Leandro Agra, Dr. Justino José Dias, Dra. Maria do Desterro Mendes, Jennifer Decloquement, Lincoln Vicente Araújo dos Santos Bizerra, Dra. Catharine Abreu Bomfim, Dra. Geisianny Moreira e Diego José da Silva, amigos que fiz durante minha passagem por Brasília, e que levarei para a vida.

A grandes pesquisadores como a professora Dra. Mercedes Bustamante, minha principal referência como pessoa e pesquisadora, que me acolheu em seu laboratório, me permitindo compreender a Ecologia em uma dimensão antes por mim inimaginável. O Dr. Guarino Rinaldi Colli, que mesmo que temporariamente, me acolheu em seu grupo de estudos e que permitiu a expansão dos meu horizontes sobre pesquisa em biodiversidade. Ao professor Dr. Robert Neil Gerard Miller, por sua consistente e incansável vontade de aprender. Ao professor Dr. Robert Weingart Barreto, por me acolher em seu grupo de pesquisas, me dando suporte para aprender ainda mais sobre fungos fitopatogênicos.

A Dra. Juliana Marcolino Gomes, constantemente me incentivando na conclusão desta tese. Sem a sua inspirando esse conteúdo não seria possível.

A todos os integrantes do, GEPPLANT (hoje não mais existente), por me permitir adquirir conhecimentos e parcerias das quais não os teria alcançado sob outras vias.

Finalmente, agradeço à Universidade de Brasília e ao Departamento de Biologia Celular

e Fitopatologia (que me acolheu mesmo que de forma extraoficial) e todos os seus membros. O presente trabalho foi realizado com apoio financeiro da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

Resumo

A presente tese aborda dois desafios cruciais na análise de dados genômicos: a agregação e complementação de metadados e a classificação filogenética de sequências biológicas. Para resolver o primeiro desafio, desenvolvemos o *GeneConnector*, uma ferramenta que agrega e complementa metadados de registros do GenBank, explorando informações compartilhadas entre diferentes sequências de um mesmo espécime. A aplicação do *GeneConnector* ao banco de dados GOPHY demonstrou sua eficácia na recuperação de informações valiosas sobre a origem, coleta e processamento das amostras, com ganhos de informação de até 60%. Adicionalmente, introduzimos os scores *Observed Completeness Score - OCS* e *Reachable Completeness Score - RCS* para avaliar a completude dos metadados e o potencial de enriquecimento de informações. Para o segundo desafio, desenvolvemos o *Classeq*, uma ferramenta de classificação de sequências biológicas baseada em posicionamento filogenético, rápida, precisa, independente de alinhamentos múltiplos de sequências e capaz de classificar sequências de genes inteiros. Nossos testes com o *Bacillus subtilis group* demonstraram a alta sensibilidade e especificidade da ferramenta, classificando corretamente quase todas as sequências do grupo em seus respectivos clados. Adicionalmente, o *Classeq* oferece uma interface de usuário amigável e uma API para facilitar sua integração em fluxos de trabalho existentes. Em suma, o *GeneConnector* e o *Classeq* representam avanços significativos na análise de dados genômicos, com potencial para impulsionar pesquisas em diversas áreas. Ao abordar os desafios de agregação de metadados e classificação filogenética, essas ferramentas oferecem novas perspectivas para a interpretação e utilização de dados genômicos, abrindo caminho para descobertas e aplicações inovadoras.

Palavras-chave: GeneConnector, Classeq, Micologia, Fitopatologia, GOPHY, Filogenia, *Bacillus subtilis group*, Posicionamento filogenético, Alignment-free, GenBank, NCBI.

Abstract

This thesis addresses two crucial challenges in genomic data analysis: metadata aggregation and complementation, and phylogenetic classification of biological sequences. To address the first challenge, we developed *GeneConnector*, a tool that aggregates and complements metadata from GenBank records by exploiting shared information among different sequences from the same specimen. The application of *GeneConnector* to the GOPHY database demonstrated its effectiveness in retrieving valuable information about the origin, collection, and processing of samples, with information gains of up to 60%. Additionally, we introduced the OCS (Observed Completeness Score) and RCS (Reachable Completeness Score) to assess metadata completeness and potential for information enrichment. For the second challenge, we developed *Classeq*, a tool for classifying biological sequences based on phylogenetic placement, which is fast, accurate, independent of multiple sequence alignments, and capable of classifying whole gene sequences. Our tests with the *Bacillus subtilis* group demonstrated the high sensitivity and specificity of the tool, correctly classifying almost all sequences of the group into their respective clades. Additionally, *Classeq* offers a user-friendly interface and an API to facilitate its integration into existing workflows. In summary, *GeneConnector* and *Classeq* represent significant advances in genomic data analysis, with the potential to drive research in various fields. By addressing the challenges of metadata aggregation and phylogenetic classification, these tools offer new perspectives for interpreting and utilizing genomic data, paving the way for innovative discoveries and applications.

Keywords: *GeneConnector, Classeq, Mycology, Phytopathology, GOPHY, Phylogeny, Bacillus subtilis group, Phylogenetic placement, Alignment-free, Genbank, NCBI.*

Lista de ilustrações

Figura 1 – Evolução dos metadados associados a <i>Ceratocystis fimbriata</i>	30
Figura 2 – O papel dos classificadores de sequencia biológicas no enriquecimento de informações	38
Figura 3 – Comparação da classificação de sequências orientadas a taxonomia e filogenia	43
Figure 4 – GeneConnector information diagram	54
Figure 5 – Information gain at the globe scale	59
Figure 6 – Information gain by genus of phytopathogenic fungi registered in GOPHY .	60
Figure 7 – The overall workflow of the Classeq life cycle is depicted	70
Figure 8 – Details of the prediction algorithm	71
Figure 9 – RefSeq database acquisition performed during the Classeq validation assay	73
Figure 10 – The Classeq prediction for the main <i>Bacillus subtilis</i> group clades	75

Lista de tabelas

Tabela 1 – Contraste dos termos utilizados nos arquivos GBFF e os utilizados como resposta XML fornecidos pelo Entrez	29
Table 2 – GeneConnector manuscript’s authors with affiliation.	49
Table 3 – GeneConnector’s Metadata Indicator Groups	53
Table 4 – The first three steps of the GeneConnector’s completeness scores calculation	56
Table 5 – Classeq manuscript’s authors with affiliation.	62
Tabela 6 – Qualificadores elegíveis como <i>Feature keys</i> do tipo <i>source</i>	115

Lista de abreviaturas e siglas

AC	Audubon Core
AFM	Alocação Filogenética de Metagenomas
BI	Bayesian Inference (Inferência Bayesiana)
BioCAsE	Biological Collection Access Service
BioPAX	Biological Pathway Exchange
BLAST	Basic Local Alignment Search Tool
CNN	Convolutional Neural Network
DBN	Deep Belief Network
DDBJ	DNA DataBank of Japan
DIF	Directory Interchange Format
DNA	ácido desoxirribonucleico
DTR	- Doença do Tronco e da Raiz de <i>Neofusicoccum</i>
DwC	Darwin Core
eDNA	Environmental DNA
EMBL	European Molecular Biology Laboratory
EML	Ecological Metadata Language
EPA	Evolutionary Placement Algorithm
Fkey	Feature keys
Fqual	Feature qualifiers
FqS	Feature qualifiers source
FT	Feature Table
GBIF	Global Biodiversity Information Facility
GFF	Genbank Flat File

GOPHY	Genera of Phytopathogenic Fungi
GSC	Genomic Standards Consortium
GUI	Graphical User Interface
HTS	High-Throughput Sequencing
ID	Identifier
IF	Inferência Filogenética
INSDC	International Nucleotide Sequence Database Collaboration
ITS	Internal Transcribed Spacer
JSON	Javascript Object Notation
KII	k-mers inverse indices
KNB	Knowledge Network for Biocomplexity
LCA	Last Common Ancestor
ML	Máxima Verossimilhança
MMS	Microbiome Metadata Standards
MOD-CO	Meta-omics Data and Collection Objects
NCBI	National Center for Biotechnology Information
NMDC	National Microbiome Data Collaborative
NOAA	Administração Nacional Oceânica e Atmosférica
OCS	Observed Completeness Score
OvR	One-vs-Rest
PEP	Portable Encapsulated Project
QS	query sequences
RA	reference alignments
RCS	Reachable Completeness Score
RDP	Ribosomal Database Project
RNA	ácido ribonucleico

rRNA	RNA ribossomal
RS	reference sequences
RT	reference tree
SiBBr	Sistema de Informação sobre a Biodiversidade Brasileira
SRA	Sequence Read Archive
TDWG	Taxonomic Databases Working Group
TCS	Taxonomic Concept Transfer Schema
TSV	Tab-separated values
URI	Uniform Resource Identifier
UUID	Universal Unique Identifiers
WGS	Whole Genome Sequence
XML	eXtensible Markup Language

Sumário

1	INTRODUÇÃO	17
2	OBJETIVOS	19
2.1	Objetivo geral	19
2.2	Objetivos específicos	19
3	FUNDAMENTAÇÃO TEÓRICA	21
3.1	Metadados: contextualizando dados	21
3.1.1	O que são? Como são organizados?	21
3.1.2	Esquemas de metadados	23
3.1.3	Metadados do Genbank	26
3.2	Classificação de seqüências biológicas	37
3.2.1	Taxonomia de metagenomas: a abordagem tradicional	37
3.2.2	Filogenias em metagenômica: a abordagem alternativa	44
4	GENECONNECTOR: UNLOCKING THE FULL POTENTIAL OF GENBANK METADATA	49
4.1	Introduction	50
4.2	Problem statement	52
4.3	Proposed solution	53
4.3.1	Concepts and Information Modelling	53
4.3.2	Technologies and Code Availability	56
4.3.3	Study case: GOPHY data completeness	57
4.4	Results	58
4.4.1	MIG's representativity and distribution	58
4.4.2	Phytopathogenic completeness along GOPHY genus	59
4.5	Conclusions	60
5	CLASSEQ: A CLADE-INFORMED SEQUENCE IDENTIFICATION TOOL	62
5.1	Introduction	63
5.2	Software Usage, Structure, and Algorithm	65
5.2.1	Usage Summary	65
5.2.2	Building the Phylogenetic Indices	65
5.2.3	Prediction	69

5.3	A proof-of-concept: Classifying specimens of <i>Bacillus subtilis</i> group using the <i>gyrB</i> gene	72
5.3.1	RefSeq data acquisition	72
5.3.2	Reference phylogeny reconstruction	73
5.3.3	Classeq predictions	74
5.3.4	Results and Discussions	76
5.4	Conclusions	77
5.5	Future remarks	77
6	CONSIDERAÇÕES FINAIS	79
	REFERÊNCIAS	80
	APÊNDICES	94
	APÊNDICE A – MANUSCRITO REFERENTE AO CAPÍTULO 4 . .	95
	APÊNDICE B – MANUSCRITO REFERENTE AO CAPÍTULO 5 . .	103
	ANEXOS	113
	ANEXO A – QUALIFICADORES ELEGÍVEIS AOS FEATURE KEYS DO TIPO <i>SOURCE</i>	114

1 Introdução

Os dados biológicos representam um pilar fundamental para a compreensão dos mecanismos que sustentam a vida, impulsionando avanços em diversas áreas do conhecimento, incluindo medicina, biotecnologia e agricultura (SHENDURE; FINDLAY; SNYDER, 2019; JEYASRI et al., 2021). Com o advento das tecnologias de sequenciamento de alto rendimento (HTS), a geração de dados genômicos tem crescido exponencialmente, resultando em um vasto acúmulo de informações em bancos de dados públicos, como o GenBank (BENSON et al., 2012). No entanto, a mera coleta de dados brutos não é suficiente para extrair conhecimento significativo. É nesse contexto que os metadados emergem como elementos cruciais, fornecendo o contexto essencial para a interpretação e análise precisa dos dados (MICHENER, 2015).

Definidos como "dados sobre dados", os metadados fornecem informações contextuais valiosas sobre a origem, o processamento e as características das sequências biológicas (DUVAL, 2001). Essas informações abrangem uma ampla gama de atributos, incluindo origem taxonômica, localização geográfica, características do hospedeiro, condições ambientais e protocolos experimentais (JONES et al., 2019). A riqueza e a qualidade dos metadados são essenciais para garantir a confiabilidade, a reprodutibilidade e a reutilização dos dados em pesquisas futuras (WILKINSON et al., 2016).

No entanto, a qualidade dos metadados em bancos de dados públicos, como o GenBank, muitas vezes é heterogênea e incompleta (CHEN; ZOBEL; VERSPOOR, 2017). A ausência de informações contextuais adequadas pode levar a interpretações errôneas, dificultar a comparação entre estudos e limitar o potencial de descoberta científica. Para superar esses desafios, ferramentas como o GeneConnector (ELIAS et al., 2024, fruto da presente tese) foram desenvolvidas para agregar e enriquecer metadados de registros do GenBank, permitindo uma análise mais abrangente e precisa dos dados.

Além dos metadados, a classificação de sequências biológicas desempenha um papel fundamental na análise de dados genômicos, especialmente em estudos de microbiomas (HMP, 2012b; HMP, 2012a). A identificação taxonômica e filogenética de sequências desconhecidas permite a caracterização da diversidade microbiana e a compreensão das relações evolutivas entre os organismos. Ferramentas como FASTA (PEARSON; LIPMAN, 1988), BLAST (ALTSCHUL et al., 1990) e CLARK (OUNIT et al., 2015) utilizam abordagens baseadas em similaridade e algoritmos de aprendizado de máquina para classificar sequências em diferentes níveis taxonômicos.

No entanto, a classificação baseada em taxonomia pode ser suscetível a erros devido a inconsistências e atualizações constantes na nomenclatura e nas relações filogenéticas (EDGAR, 2018; LYDON; LIPP, 2018). Para superar essa limitação, métodos de posicionamento filoge-

nético, como PPLACER (MATSEN; KODNER; ARMBRUST, 2010) e EPA-NG (BARBERA et al., 2019), utilizam árvores filogenéticas de referência para alocar sequências de forma mais precisa e informativa.

Em suma, a crescente disponibilidade de dados genômicos, juntamente com o desenvolvimento de ferramentas avançadas de análise, como GeneConnector e classificadores filogenéticos, tem impulsionado avanços significativos na compreensão da biodiversidade e na exploração do potencial biotecnológico dos microrganismos. A integração de metadados e informações filogenéticas representa um passo crucial para desvendar a complexidade dos ecossistemas microbianos e suas interações com o meio ambiente, abrindo caminho para novas descobertas e aplicações em diversas áreas da ciência.

A presente tese está estruturada em seis capítulos, incluindo esta introdução e objetivos que são apresentados nos dois primeiros. O Capítulo 3 fornece a fundamentação teórica essencial para a compreensão dos conceitos e métodos utilizados, abordando os temas de metadados e classificação de sequências biológicas. No Capítulo 4, é apresentado o GeneConnector, uma ferramenta desenvolvida para enriquecer metadados em registros do GenBank, juntamente com um estudo de caso demonstrando sua aplicação e resultados. O Capítulo 5 introduz o Classeq, uma ferramenta inovadora para classificação filogenética de sequências, e apresenta um estudo de caso para validar sua eficácia. Por fim, o Capítulo 6 tece as considerações finais, reunindo os principais resultados e conclusões desta pesquisa, e discutindo suas implicações para o futuro da análise de dados genômicos.

2 Objetivos

2.1 Objetivo geral

Aprimorar a análise e interpretação de dados genômicos microbianos, facilitando o acesso à informação contextual crucial para estudos de biodiversidade, taxonomia e filogenia molecular, com o intuito de aprofundar a compreensão da ecologia e do papel dos microrganismos em diversos ambientes.

2.2 Objetivos específicos

- **Desenvolver o GeneConnector:** Criar uma ferramenta que permita a agregação e o enriquecimento automático de metadados em registros de sequências nucleotídicas do GenBank, explorando informações compartilhadas entre diferentes marcadores genéticos de um mesmo espécime. Isso possibilitará a recuperação de informações valiosas sobre a origem, coleta e processamento das amostras, que muitas vezes estão fragmentadas ou incompletas nos registros individuais.
- **Avaliar a efetividade do GeneConnector:** Aplicar a ferramenta `GeneConnector` ao banco de dados *Genera of Phytopathogenic Fungi - GOPHY*, que contém informações sobre fungos fitopatogênicos, para demonstrar sua capacidade de melhorar a completude dos metadados e, conseqüentemente, a qualidade dos dados genômicos disponíveis para estudos em micologia e fitopatologia.
- **Desenvolver o Classeq:** Criar uma ferramenta de classificação de sequências biológicas baseada em posicionamento filogenético que seja rápida, precisa e independente de alinhamentos múltiplos de sequências, superando as limitações das ferramentas tradicionais de classificação. A ferramenta permitirá a inserção de sequências de genes em árvores filogenéticas de referência, possibilitando a identificação e classificação de microrganismos de forma mais precisa e informativa, mesmo em sequências altamente divergentes ou com informações taxonômicas incompletas.
- **Validar a performance do Classeq:** Testar a ferramenta `Classeq` utilizando sequências de *Bacillus subtilis group*, um importante grupo de bactérias com diversas aplicações biotecnológicas e industriais, para avaliar sua sensibilidade, especificidade e capacidade de classificar sequências em clados filogenéticos corretos. A validação da ferramenta em um conjunto de dados de referência garantirá sua confiabilidade e aplicabilidade em estudos microbiológicos.

- **Disponibilizar as ferramentas à comunidade científica:** Tornar os frutos da presente tese (`GeneConnector` e `Classeq`) acessíveis aos pesquisadores, por meio de interfaces de linha de comando e API's, facilitando a análise e interpretação de dados genômicos microbianos em diversas áreas da pesquisa, como microbiologia ambiental, microbiologia médica, biotecnologia e agricultura. A disponibilização das ferramentas contribuirá para o avanço do conhecimento científico e para o desenvolvimento de novas aplicações biotecnológicas e agrícolas.

3 Fundamentação teórica

3.1 Metadados: contextualizando dados

3.1.1 O que são? Como são organizados?

No âmbito das ciências da vida, a coleta de dados biológicos representa um pilar fundamental para a compreensão dos mecanismos que sustentam a vida. No entanto, a mera coleta de dados, por si só, não é suficiente para desvendar os fenômenos biológicos. É nesse contexto que os metadados assumem um papel crucial, fornecendo a contextualização necessárias para transformar os dados brutos em conhecimento científico de valor.

Os metadados, definidos como informações que descrevem outras informações (DUVAL, 2001), ou mesmo conforme definido por Patrick Lambe como "uma coleção de informações estruturadas sobre um documento ou conteúdo" (LAMBE, 2014), funcionam como a alma dos dados, fornecendo o contexto vital que permite a sua correta interpretação e análise. Ao associar os dados biológicos a informações como data e local de coleta, condições ambientais, métodos de análise e outros parâmetros relevantes, os metadados permitem que os pesquisadores compreendam a origem dos dados, sua representatividade e as nuances que influenciam os resultados obtidos (MICHENER, 2015).

Acima de tudo os metadados possuem um papel central na reutilização dos dados coletados ao longo da cadeia de produção científica. Sem eles não seríamos capazes de compreender os padrões do mundo que nos rodeia. Por exemplo, ao compararmos duas sequências de DNA, poderíamos apenas determinar parâmetros intrínsecos as mesmas, como (dis)similaridade, composição ou mesmo número de bases. Entretanto jamais poderíamos determinar sua origem geográfica, associação com hospedeiros, parâmetros de qualidade ou mesmo a veracidade do conteúdo das próprias sequências sem olhar o contexto, ou minimamente, sem confrontar nossos dados contra outros que possuam um contexto previamente determinado.

Os metadados por si só possuem uma grande amplitude de variações no cumprimento de sua missão em descrever um contexto dos dados. Por esse motivo autores propõe diferentes taxonomias para organização dos metadados. Seguindo a direção proposta por Patrick Lambe (LAMBE, 2014), os metadado podem ser agrupados de acordo com seu propósito com relação aos dados, estando presentes em quatro grupos:

- **Os que identificam conteúdo.** Aqui estão metadados descritivos, que comportam informações que tornam os dados únicos, possibilitam distinguir um dado de qualquer outro. Se utilizarmos referências biológicas, aqui estarão informações sobre ambiente de coleta

dos dados, relações ecológicas (hospedeiros, substratos, tipos de solo), marcos temporais e geográficos, tipo de vegetação, etc.

- **Os que conectam dados em sistemas de gerenciamento.** Aqui estão informações administrativas e estruturais, como números de versão em sistemas de versionamento de dados, datas relacionadas a aquisição do dado (sempre relativos ao sistema), ou mesmo informações sobre o arquivo que comporta o dado.
- **Os que permitem a recuperação do documento.** Esses metadados comportam informações que auxiliam na manutenção da robustez dos dados, como descrição(ões) do(s) documento(s) que contém os dados ou dos próprios dados, chaves de indexação ou mesmo sua taxonomia.
- **Os que permitem a incrementação dos dados.** Aqui estão metadados que permitem, após eventos de expansão de dados, conectar esses a partes anteriores, ou mesmo informações geradas em tempo real, como citações em uma literatura científica.

De maneira contínua a proposta de Lambe 2014, para Jenn Ryley 2017, os metadados são agrupados em quatro categorias nominais, conforme listagem abaixo:

- **Metadados descritivos.** Informações presentes nessa categoria permitem aos usuários que compreendam o conteúdo dos dados, incluindo indicações de onde, como, quando, por quem, etc., os dados foram coletados.
- **Estruturais.** Descrevem conexões existentes dentre e entre os dados. Metadados presentes nessa categoria podem indicar fontes externas de informação, como relações com bancos de dados de referência. A título de exemplo, o [Genbank](#) utiliza chaves do tipo `db_xref` para relacionar informações nucleotídicas a informações taxonômicas (e.g. [KY855514.1](#) tem como `db_xref` o metadado `taxon:5507`, indicando se tratar de um *Fusarium oxysporum*), sendo essa uma estratégia que permite a evolução independente de grandes coleções de dados.
- **Administrativos.** Aqui estão metadados relacionados aos tratamentos técnicos (renderização, recuperação ou acesso), elementos legais (licenças, distribuição e direitos autorais), assim como preservação/validade dos dados.
- **Linguagem de Marcação.** Por último, aqui estão informações acerca de referências internas ou elementos semânticos dos dados. Em dados biológicos metadados dessa categoria são menos empregados, porém ao se tratar de dados de categorias literárias por exemplo, elementos desse tipo são amplamente empregados. De acordo com Riley 2017 aqui estão mixados metadados e conteúdo propriamente ditos.

Em resumo, devemos destacar que os metadados são, sobretudo, tão importantes como os próprios dados, pois são esses que garantem a robustez das informações a longo prazo e primariamente a interoperabilidade e escalonamento tanto espacial como temporal de pesquisas científicas.

3.1.2 Esquemas de metadados

Os metadados podem ser delineados/modelados para servirem a domínios específicos de conhecimento, e para cada um desses domínios assumirem variações de sintaxe, cumprir regras de codificação, seguir modelos de conteúdo e ainda adotarem *frameworks* semânticos, nos chamados esquemas de metadados (RILEY, 2017).

Os principais esquemas de metadado dedicados a modelagem de informações sobre biodiversidade, atualmente são desenvolvidos e mantidos pela *Biodiversity Information Standards* (anteriormente conhecida como *Taxonomic Databases Working Group*, nome que originou o acrônimo TDWG), uma organização internacional engajada no desenvolvimento de padrões abertos para o compartilhamento de informações sobre biodiversidade. Entre os vários esquemas de dados disponibilizados pela TDWG, alguns dos mais notáveis são listados abaixo em ordem cronológica:

- **Taxonomic Concept Transfer Schema - TCS** (KENNEDY; KUKLA; PATERSON, 2005): Este é um padrão para a troca de informações sobre conceitos taxonômicos, ou seja, a interpretação de nomes biológicos.
- **Biological Collection Access Service - BioCAsE** (GÜNTSCH; BERENDSOHN; MERGEN, 2007): Este protocolo permite o acesso a dados de coleções biológicas e ambientais.
- **Access to Biological Collection Data - ABCD** (HOLETSCHEK et al., 2012): Este esquema permite o acesso a informações detalhadas sobre coleções biológicas, incluindo dados sobre espécimes e observações.
- **Darwin Core - DwC** (WIECZOREK et al., 2012): Com propósito biológico geral, o padrão é focado na descrição de informações associadas a ocorrência de organismos vivos e fósseis. Esse esquema é adotado pelo *Global Biodiversity Information Facility* (GBIF, 2020) e por consequência é adotado pelo *Sistema de Informação sobre a Biodiversidade Brasileira* (SiBBR).
- **Audubon Core - AC** (MRTG, 2020): Este é um conjunto de vocabulários projetados para a descrição de recursos de mídia de biodiversidade, como imagens, sons e vídeos.

Em adição aos cinco citados acima, a TDWG ainda conta com mais 13 esquemas de dados aprovados e documentados em sua [página oficial](#) além de disponibilizar detalhes sobre

implementação e revisões no seu repositório oficial do Github github.com/tdwg. Cada um desses esquemas desempenha um papel fundamental na promoção da interoperabilidade de dados e na facilitação do compartilhamento e reutilização de dados de biodiversidade. A adoção desses padrões pela comunidade científica tem o potencial de acelerar o progresso na compreensão e na conservação da biodiversidade global.

Apesar de TDWG representar uma instituição de alta relevância dentro de fora da comunidade científica, existem outros esquemas de metadados disponíveis para dados biológicos que não necessariamente são mantidos pela TDWG. Um exemplo amplamente conhecido e que no momento da escrita dessa tese completa quase três décadas, o *Directory Interchange Format* - DIF (BARTON, 1995), representa um padrão específico para armazenamento e compartilhamento de informações ambientais gerados e mantidos pela Administração Nacional Oceânica e Atmosférica dos Estados Unidos (NOAA), com foco em armazenamento de informações oceânicas, continentais e espaciais orientado a diretórios.

Em termos de informações ecológicas, podemos citar o *Ecological Metadata Language* - EML (JONES et al., 2019), um padrão baseado em XML¹ dedicado especificamente a metadados associados a dados ecológicos, utilizado principalmente na descrição de informações ambientais. O EML foi desenvolvido pela comunidade de ecologia, especificamente pelo *Knowledge Network for Biocomplexity* - KNB, uma rede que promove a colaboração interdisciplinar e a partilha de dados entre ecologistas, cientistas da computação e pesquisadores de outras áreas relacionadas.

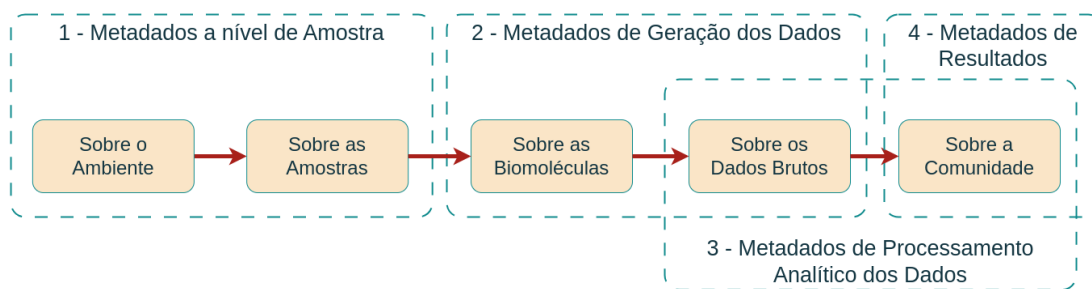
Tratando-se de informações moleculares podemos citar propostas como o *Meta-omics Data and Collection Objects* - MOD-CO (RAMBOLD et al., 2019), um padrão de metadados desenvolvido para aprimorar a interoperabilidade e a reutilização de dados em pesquisa metagênômica, que inclui metagenômica, metatranscriptômica, metaproteômica e metabolômica.

Temos ainda o *Biological Pathway Exchange* - BioPAX (DEMIR et al., 2010), sendo um esquema de dados que permite a representação e a integração de vias biológicas em diversos níveis, incluindo interações de proteínas, vias de sinalização e vias metabólicas. Representando um importante padrão para geneticistas, biólogos moleculares, bioquímicos e profissionais das ciências da vida como um todo.

Finalmente, temos o *Microbiome Metadata Standards* - MMS, uma iniciativa que surgiu em 2019 a partir de um workshop organizado pelo *National Microbiome Data Collaborative* - NMDC, com o objetivo de padronizar os metadados em estudos do microbioma (VANGAY et al., 2021), o qual fornece um conjunto de diretrizes para a coleta e padronização de metadados de estudos do tipo. O objetivo principal do MMS é possibilitar a integração de dados de estudos orientados a amostras ambientais. No Quadro 1 apresento uma adaptação da simplificação na hierarquia de metadados presentes em estudos ambientais apresentada no Workshop, incluindo exemplos de metadados nas diferentes hierarquias de informação em microbiomas.

¹ eXtensible Markup Language.

Quadro 1 | Simplificação da hierarquia de metadados em estudos ambientais



Uma adaptação da hierarquia de metadados apresentada por Vangay et. al. 2021. Na figura acima podemos identificar grupos de metadados (e em alguns casos, dados) coletados durante o fluxo de execução de um estudo em microbiomas ambientais.

- 1. Nível de amostra:** Esses metadados carregam informações que permitem identificar e caracterizar o ambiente de onde as amostras foram extraídas. A título de exemplo:

 - **Marcos temporais:** Datas de coleta, datas de armazenamento, datas da expedição, etc.
 - **Marcos geográficos:** Latitude, longitude, altitude/profundidade, nome do local e país.
 - **Substrato/Hospedeiro:** Solo, água do mar, fezes, raízes, folhas, ritidoma, etc.
 - **Ambiente:** Temperatura, salinidade, pH e estado da doença.
 - **Observação:** Informações gerais que independem do processamento da amostra.
- 2. Geração de Dados:** Informações referentes a extração de dados das amostras, também conhecidos como metadados de preparação.

 - **Meios de processamento:** Extração de DNA, sequenciamento e localização dos dados brutos.
 - **Sequenciamento de DNA:** Primers, kits de biblioteca, instrumentos e parâmetros.
- 3. Processamento analítico:** Aqui estão informações sobre processos dos dados gerados a partir das amostras e esses metadados possuem o objetivo de permitir que novos usuários a replicabilidade dos dados.

 - **Propriedades:** Comprimento da sequência, sequências por amostra, pares de bases totais.
 - **Controle de Qualidade:** Remoção de adaptadores, corte e filtragem de qualidade, desreplicação e remoção de quimera.
 - **Montagem:** Ferramenta, *binning* e finalização.
 - **Anotação de Genes:** Ferramenta e banco de dados.
 - **Parâmetros dos Softwares:** Versões e configurações utilizadas.
- 4. Sobre resultados:** Esse conjunto pode conter dados e metadados. Aqui estão informações sobre o conteúdo informacional das comunidades microbianas propriamente ditas.

 - **OTUs/ASVs:** Taxonomia, sequências de referência e identificadores. Esses metadados podem incluir matrizes de espécies x *pools* de dados e outras *features* que permitam aos usuários compreender componentes das amostras em análise.

Considerando a alta relevância das definições contidas no MMS, vários esquemas de metadados foram desenvolvidos em conformidade com o mesmo. Como exemplo dos esquemas mais importantes podemos citar o MIxS (YILMAZ et al., 2011) e MIMARKS (YILMAZ et al., 2011).

MIxS é um acrônimo para *Minimum Information about any (x) Sequence* e representa um esquema que fornece um conjunto de padrões de metadados para a descrição de sequências de genomas e metagenomas, incluindo detalhes sobre o ambiente de amostragem, a qualidade da

sequência e o processamento de dados. Desenvolvido como parte do *Genomic Standards Consortium* - GSC (YILMAZ et al., 2011). Já o MIMARKS, um acrônimo de *Minimum Information about a MARKer gene Sequence*, se diferencia do MIxS por especificar um conjunto mínimo de metadados necessários para a descrição de sequências de genes marcadores, como o 16S rRNA, usado frequentemente em estudos de microbioma. (YILMAZ et al., 2011).

Todos esses padrões/esquemas servem sobretudo ao propósito central de garantir a segurança de dados no que diz respeito aos princípios FAIR (WILKINSON et al., 2016), que constituem um conjunto de diretrizes que visam aumentar a utilidade dos dados científicos e de pesquisa. Este princípio é um acrônimo para *Findable, Accessible, Interoperable* e *Reusable*. O princípio *Findable* refere-se à capacidade de os dados serem facilmente descobertos através de identificadores únicos e descritivos. *Accessible* indica que, uma vez encontrado, os dados devem ser acessíveis e a maneira como acessá-los deve ser claramente definida. *Interoperable* sugere que os dados devem ser capazes de interagir com outros conjuntos de dados e serem integrados de forma útil e eficaz, o que é frequentemente facilitado por padrões comuns de dados e vocabulários. Por fim, *Reusable* sugere que os dados devem ser suficientemente bem descritos e preservados para que possam ser reutilizados por outros, além do seu propósito inicial. Este princípio é fundamental para a promoção do compartilhamento de dados abertos e da ciência de dados reproduzível.

Os parágrafos anteriores exemplificam a diversidade de esquemas disponíveis para o tratamento de metadados biológicos. Entretanto com o aumento exponencial do montante e importância dos dados e metadados na geração de conhecimento científico, é importante destacar iniciativas como o *Schema Playground*, uma plataforma online interativa desenvolvida para facilitar a criação, validação e compartilhamento de esquemas de metadados para conjuntos de dados científicos. Esta ferramenta permite aos usuários desenvolver e testar esquemas de metadados personalizados, bem como explorar e reutilizar esquemas existentes, contribuindo para a padronização e interoperabilidade dos metadados (CANO et al., 2023). Ao facilitar a criação e gestão de metadados de alta qualidade, o *Schema Playground* pode contribuir significativamente para a evolução e manutenção da qualidade das informações geradas pela comunidade científica. Metadados bem estruturados e padronizados são fundamentais para a descoberta, acesso, interpretação e reutilização eficazes dos dados, contribuindo para a transparência, replicabilidade e avanço da pesquisa científica (WIECZOREK et al., 2012).

3.1.3 Metadados do Genbank

Até aqui eu destaquei informações fundamentais sobre metadados, no intuito de apresentar ao leitor a importância de "informações sobre informações" para a ciência. Na presente sessão focarei no esquema de dados do Genbank, objetivando dar suporte ao leitor na compreensão do Capítulo 4 da presente tese.

Em uma breve introdução, o GenBank é um banco de dados de acesso livre mantido pelo

National Center for Biotechnology Information - NCBI, dos Estados Unidos. Seu objetivo inicial é centralizar e padronizar a crescente quantidade de dados de sequenciamento de DNA gerados por diversos laboratórios ao redor do mundo. A instituição faz parte do *International Nucleotide Sequence Database Collaboration* - INSDC (ARITA; KARSCH-MIZRACHI; COCHRANE, 2021), que é fruto do esforço compartilhado entre o *DNA DataBank of Japan* - DDBJ, o *European Molecular Biology Laboratory* - EMBL, e o próprio GenBank no NCBI (BENSON et al., 2012; WHEELER et al., 2003; COCHRANE et al., 2016).

Com o intuito de estabelecer uma interface comum para troca de dados entre instituições, o Genbank e o EMBL em 1986 (incluindo o DDBJ em 1987) realizaram um esforço colaborativo para estabelecer um formato de troca de informações de sequências biológicas e suas anotações, comum entre as instituições (ver seção 1 do referencial técnico do INSDC). Nesse momento, estabeleceu-se a *Feature Table* - FT (abordado nos próximos parágrafos), e os *Flat Files* - FF como padrões comuns de troca de dados os quais perduram até o momento de escrita desta tese. A definição dos padrões supracitados permitem que as instituições realizem sincronizações diárias de dados, o que faz com que sequências depositadas em uma das instituições pertencentes ao consórcio, sejam automaticamente replicadas para as demais regiões do globo, variando somente em detalhes de sintaxe para cada instituição (ver seção 7 do referencial técnico do INSDC para uma comparação entre os formatos). Por se tratar de um padrão comum e com poucas variações atribuídas somente a formato, aqui nos ateremos ao *Genbank Flat File* - GBFF (GENBANK, 2024).

Primariamente, é importante destacar que diferentemente dos esquemas descritos na seção 3.1.2 desse documento, o GBFF (GENBANK, 2024) não possui uma declaração oficial de esquema, disponibilizada aos usuários através de portais dedicados a esse propósito. Para os GBFF, podemos contar apenas com um documento de exemplo descrevendo os elementos básicos que contém cada seção do padrão, e que pode ser acessado através da URI² www.ncbi.nlm.nih.gov/genbank/samplerecord/ (GENBANK, 2024). Adicionalmente, podemos encontrar uma referência não oficial das informações utilizadas pelo GBFF na plataforma de compartilhamento de esquemas FAIRsharing (MCQUILTON et al., 2018) que pode ser acessado através do URI [FAIRsharing.rg2vmt](https://fairsharing.org/2vmt), porém esse não é suportado por mantenedores oficiais do Genbank.

Finalmente, tratando-se do padrão GBFF, esse representa um formato bastante intuitivo e amigável a atores humanos. Cada arquivo GBFF pode conter um ou mais registros de sequência, delimitados pelo termo LOCUS ao início do documento e a dupla-barra (//) ao seu final. Esse tipo de arquivo é utilizado na troca de informações tanto com usuários finais, quanto via acesso programático. Apesar do formato apresentar algumas limitações quanto a performance quando utilizado para descrever sequências longas ou com anotações altamente complexas/comple-

² A URI (Uniform Resource Identifier, ou Identificador Uniforme de Recursos) é uma *string* (sequência de caracteres) que se refere a um recurso da web.

tas, esse é amplamente adotado por *softwares* de propósito geral em bioinformática, como o Biopython (COCK et al., 2009), Bioperl (STAJICH et al., 2002), EMBOSS (RICE; LONGDEN; BLEASBY, 2000) e GBParsy (LEE; KIM; NAHM, 2008, apresentam um *benchmark* de comparação entre as ferramentas existentes para processamento de GBFF).

Sobre a estrutura, internamente o GBFF é composto por uma série de termos que descrevem informações importantes sobre as sequências, incluindo metadados descritivos (DEFINITION, SOURCE), administrativos (ACCESSION, VERSION, KEYWORDS, REFERENCES), assim como campos que suportam informações mistas, como LOCUS e FEATURES. Os termos descritivos e administrativos são auto explicativos e, a sua definição básica pode ser acessado pelo leitor na Tabela 1. Já o termo LOCUS, é um campo composto altamente passível ao processamento de máquina, pois carrega informações completas sobre o nome do próprio locus, comprimento do fragmento, tipo de molécula, data de última modificação e [divisão do Genbank](#) da qual o organismo doador da sequencia pertence. Ele permite que mecanismos de busca possam ter ciência do conteúdo do registro sem necessariamente varrer o documento por completo - ou seja, altamente valioso para indexadores. Por último, o termo FEATURES representa um campo coringa, o que possibilita o carregamento de uma série de informações estruturais e descritivas sobre a sequência biológica.

Para o desenvolvimento da presente tese, a seção de maior importância é a FEATURES, pois essa é utilizada no enriquecimento de informações biológicas apresentado no Capítulo 4. Assim, abordaremos esta com maior detalhamento. A seção FEATURES é responsável por carrear dados da FT, uma peça fundamental em registros de sequencias biológicas, pois cada recurso listado na FT é responsável por referenciar regiões específicos da sequência alvo e conectar a esses, qualificadores que fornecem informações adicionais sobre sua origem e função. Uma sequência pode conter uma ou mais anotações, assim como as próprias anotações podem ser sobrepostas umas as outras. A sintaxe do conteúdo das FT's é definida nas especificações técnicas do INSDC, disponíveis no [portal oficial da instituição](#)³.

Sobretudo, cada registro presente nas FT contém elementos nomeados *Feature keys* - Fkey, que referenciam seções específicas da sequencia biológica, e os *Feature qualifiers* - Fqual, que trazem informações auxiliares sobre a porção referenciada por Fkey. As Fkey são agrupadas em famílias que espelham possíveis funções dos fragmentos referenciados por elas (ver seção 7.2 do referencial técnico do INSDC, para ter acesso a lista completa de Fkey's). Cada Fkey pode conter um (1) ou mais Fqual's, sendo que a Fkey *source* deve estar sempre presente (campo obrigatório), pois esse identifica a fonte biológica da sequência.

Os Fqual's pertencentes ao Fkey *source* (a partir desse ponto, nesse capítulo, trataremos qualificadores do grupo *source* como FqS) possuem uma sintaxe extritamente definida e representam uma rica fonte de informação. Por esse motivo representa o alvo do enriquecimento de

³ Repare que se desconsiderarmos as informações presentes nesse termo, os registros do Genbank não passariam de meras sequências de DNA anêmicas.

Tabela 1 – Contraste dos termos utilizados nos arquivos GBFF e os utilizados como resposta XML fornecidos pelo Entrez. Cada termo citado na coluna "Termo GBFF" possui o link para a referência oficial do Genbank, caso o leitor desejar. A coluna "Conteúdo" traz um resumo sobre o que devemos esperar de informação que estarão sob a tutela de cada termo.

Termo GBFF	Termo XML	Conteúdo
LOCUS	GBSeq_locus	Contém informações básicas, como o nome da sequência, o comprimento da sequência, o tipo de molécula (DNA, RNA, etc.), a forma da molécula (linear, circular), a divisão do Genbank onde a sequência está depositada e a data da última atualização.
DEFINITION	GBSeq_definition	Fornece uma descrição breve e concisa da sequência.
ACCESSION	GBSeq_primary-accession	Apresenta o número de acesso da sequência, que é um identificador único atribuído a cada sequência no GenBank.
VERSION	GBSeq_accession-version	Inclui o número de versão da sequência e o número de identificação da proteína (se aplicável).
KEYWORDS	GBSeq_keywords	Lista as palavras-chave associadas à sequência.
SOURCE	GBSeq_source	Indica a fonte biológica da sequência.
ORGANISM	GBSeq_organism	Inclui o nome científico do organismo de onde a sequência foi derivada e a classificação taxonômica. Esse é um sub-campo de SOURCE.
REFERENCES	GBSeq_references	Lista as referências bibliográficas associadas à sequência.
FEATURES	GBSeq_feature-table	Fornece anotações sobre regiões específicas da sequência, como genes, exons, regiões reguladoras, assim como fornece informações sobre interações ecológicas, marcos temporais, geográficos. Ou seja, metadados descritivos e analíticos.
ORIGIN	–	Representa um indicador local para o início da sequência, geralmente envolvendo um sítio de clivagem de restrição determinado experimentalmente ou o locus genético (se disponível). Esta informação está presente apenas em registros mais antigos e é um campo opcional.
–	GBSeq_sequence	Esta seção contém a sequência de ácidos nucleicos ou proteínas em si. Em arquivos GBFF essa seção existe porém não é declarada, sendo iniciada logo após o ORIGIN.

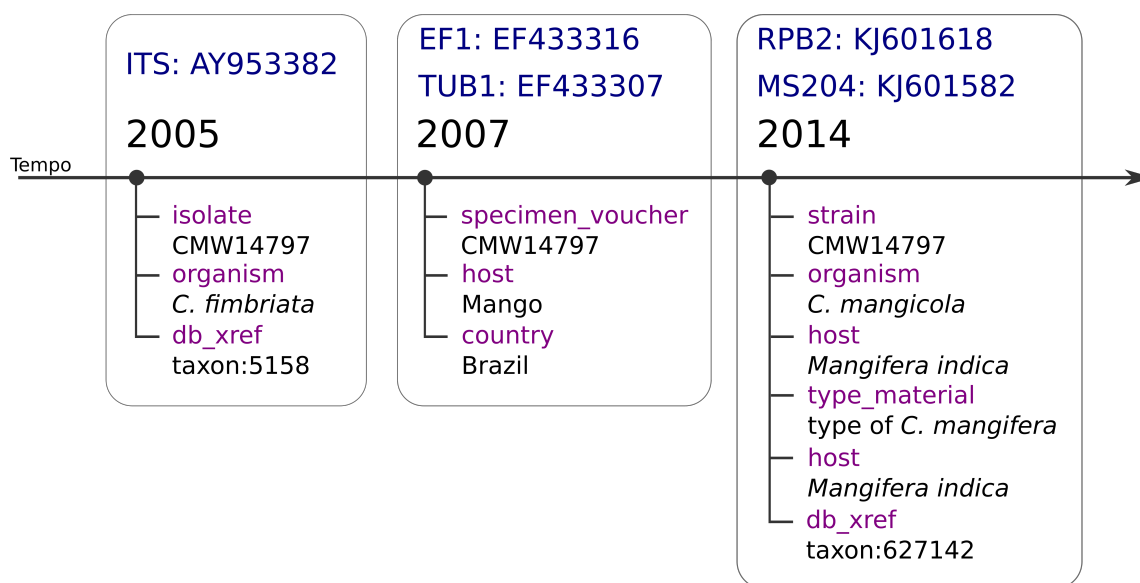


Figura 1 – Evolução dos metadados associados a *Ceratocystis fimbriata*. Da esquerda para a direita, são apresentados os eventos de incrementação de metadados conectados aos Fkey do tipo *source* (FqS). Aqui podemos observar que o primeiro evento de registro de informações ocorreu em 2005, com a publicação de informações básicas sobre a origem do DNA, permitindo apenas a identificação do isolado e a taxonomia original (*Ceratocystis fimbriata*) (WYK et al., 2005). Após dois anos as informações sofreram uma incrementação, com a inclusão de metadados sobre hospedeiro (através de seu nome vulgar) e país de origem (WYK et al., 2007). O último evento de incrementação só ocorreu após quase uma década do registro inicial, com a realocação taxonômica da espécie em *C. mangifera* (WYK et al., 2011)

dados realizado pelo GeneConnector (ver Capítulo 4). É aqui que estão os marcos geográficos, temporais, informações sobre ambiente, amostras, indivíduos, isolados, referências externas, métodos de sequenciamento, molécula sequenciada, entre outros. Ao todo são descritos pelo INSDC, 103 qualificadores (ver item 7.3), dos quais 55 são elegíveis para serem tratados como FqS (ver Anexo A para mais detalhes).

A grande variabilidade dos FqS's faz com que esse grupo de informações represente ao mesmo tempo, uma importantíssima fonte de dados para o desenvolvimento de automações sobre os registros do Genbank, como também um dos maiores débitos técnicos da plataforma. Isso porque, como podemos observar no Anexo A, apenas uma ínfima fração das informações elegíveis a pertencerem aos FqS são obrigatoriamente disponibilizados no momento da inclusão de novos registros por usuários da plataforma. Fato esse que faz com que as informações disponíveis no Genbank sejam altamente heterogêneas em termos cobertura de metadados acerca da fonte biológica das sequências.

Um forte atenuante a grande variabilidade na cobertura de metadados é o fato dos registros do Genbank serem orientados a sequências biológicas, e não aos próprios organis-

mos. Esse modelo de arquitetura de dados permite que as sequências biológicas, extraídas do mesmo organismo hospedeiro em momentos distintos, possam co-existir e evoluir de maneira independente umas das outras dentro da plataforma. Essa evolução independente pode levar a pulverização ou redundância de informações, gerando perfis semelhantes ao apresentado na Figura 1. Nela eu apresento o caso específico do isolado CMW14797 pertencente a espécie *Ceratocystis mangicola* (WYK et al., 2011), em que ao longo da quase uma década em que as informações do organismo foram incrementadas na plataforma, diversos traços de metadados do organismo evoluíram de forma independente, sem incluir ao menos informações "estruturais" que permitiriam a manutenção da conexão entre os registros independentes.

Casos semelhantes ao exposto para *Ceratocystis mangicola* são vastos na literatura científica, estando distribuídos ao longo dos mais diversos grupos taxonômicos. Abaixo eu trago casos adicionais, que exemplificam casos mistos de redundância e pulverização de informações associadas a organismos fitopatogênicos:

Neofusicoccum andinum

Agente causal de Doença do Tronco e da Raiz de *Neofusicoccum* (DTR-N) em *Vitis vinifera* (uva vinífera), *Citrus sinensis* (laranja), *Citrus limon* (limão), *Citrus reticulata* (tangerina), *Prunus persica* (pêssego), *Prunus armeniaca* (damasco), *Prunus domestica* (ameixa), *Eucalyptus* spp. (eucalipto), *Pinus* spp. (pinheiro), *Juglans regia* (noz-pecã), *Platanus* spp. (plátano) e diversas outras espécies ornamentais e florestais (referências citadas abaixo, nos cabeçalhos das listagens).

- Registro primário (MOHALI; SLIPPERS; WINGFIELD, 2006):

- **Marcadores:**

- * AY693976_{ITS rDNA}

- * AY693977_{EF-1}

- **Fqual's:**

- * organism *Neofusicoccum andinum*

- * isolate CMW13455

- Incrementação (YANG et al., 2017):

- **Marcadores:**

- * KX464002_{RFB2}

- * KX464923_{TUB2}

- **Fqual's:**

- * organism *Neofusicoccum andinum*

- * strain CMW 13455

- * isolation_source *Eucalyptus* sp.
- * culture_collection CBS:117453
- * country: Venezuela
- * collected_by S. Mohali
- * note strain co-identity: CBS 117453 = CMW 13455; sequence from ex-type culture.

Colletotrichum simmondsii

Espécie pertencente ao complexo de espécies *Colletotrichum acutatum*, que possui ampla distribuição geográfica assim como ampla gama de hospedeiros, sendo o principal agente causal da Doença da Antracnose em inúmeras culturas. Pode afetar *Mangifera indica* (Manga), *Persea americana* (abacate), *Musa* spp. (banana), *Citrus* spp. (citros), *Solanum lycopersicum* (tomate), *Capsicum annuum* (pimentão), *Phaseolus vulgaris* (Feijão), *Glycine max* (soja), *Cicer arietinum* (grão de bico), *Lens culinaris* (lentilha), *Lactuca sativa* (Alface), *Brassica oleracea var. capitata* (repolho), *Brassica oleracea var. italica* (brócolis), *Brassica oleracea var. botrytis* (couve-flor), *Rosa* spp. (Rosa), *Begonia* spp. (begônia), *Chrysanthemum* spp. (crisântemo), *Pelargonium* spp. (gerânio).

- Registro primário (PRIHASTUTI et al., 2009):

– **Marcadores:**

- * FJ972591_{GSI}
- * FJ917510_{CMD}

– **Equal's:**

- * organism *Colletotrichum acutatum*
- * host *Carica papaya*
- * specimen_voucher BRIP28519
- * country Australia

- Incrementação (DAMM et al., 2012):

– **Marcadores:**

- * JQ949927_{TUB2}
- * JQ948276_{ITS rDNA}
- * JQ949267_{HIS3}
- * JQ948937_{CHS-1}
- * JQ949597_{ACT}

- * JQ948606_{GAPDH}

– **Equal's:**

- * organism *Colletotrichum simmondsii*
- * strain CBS 122122
- * culture_collection CBS 122122
- * type_material culture from holotype of *Colletotrichum simmondsii*

Colletotrichum horii

Pertencente ao complexo de espécies *Colletotrichum gloeosporioides*, *Colletotrichum horii*, causa Podridão de Frutos e Caules de *Diospyros kaki* (Caqui) no leste asiático (China, Japão e Nova Zelândia). O fungo pode representar um possível patógeno de *Capsicum annuum* (pimentão), *Musa acuminata* (banana) e *Cucurbita pepo* (abobrinha) (XIE et al., 2010).

- Registro primário (WEIR; JOHNSTON, 2010):

– **Marcadores:**

- * GQ329681_{GAPDH}
- * GQ329690_{ITS rDNA}

– **Equal's:**

- * organism *Colletotrichum horii*
- * isolate C1180.1
- * isolation_source fruit
- * host *Diospyros kaki* (persimmon)
- * specimen_voucher PDD:98210
- * specimen_voucher TNS-F-26102
- * culture_collection ICMP:10492
- * culture_collection NBRC:7478
- * type_material culture from neotype of *Colletotrichum horii*
- * country Japan
- * collection_date 1959
- * note type strain of *Colletotrichum horii*

- Incrementação 1 (WEIR; JOHNSTON; DAMM, 2012):

– **Marcadores:**

- * JX010370_{SOD2}
- * JX009438_{ACT}

- * JX009604_{CAL}
- * JX009752_{CHS}
- * JX010137_{GS}
- * JX010450_{TUB2}

– **Fqual's:**

- * organism *Colletotrichum horii*
- * strain C1180.1
- * host *Diospyros kaki*
- * culture_collection ICMP:10492
- * culture_collection NBRC:7478
- * type_material ulture from neotype of *Colletotrichum horii*
- * country Japan
- * PCR_primers fwd_name: SODglo2-F, fwd_seq: cagatcatggagctgcacca, rev_name: SODglo2-R, rev_seq: tagtacgcgtgctcggacat
- * note ex type culture of *Colletotrichum horii*

- Incrementação 2 (SHARMA; SHENOY, 2014):

– **Marcadores:**

- * JQ807840_{Mat1-2-1}

– **Fqual's:**

- * organism *Colletotrichum horii*
- * strain ICMP 10492
- * host *Diospyros kaki*
- * culture_collection ICMP:10492
- * type_material culture from neotype of *Colletotrichum horii*
- * note type strain of *Colletotrichum horii*



Repare nas listagens acima que os Fqual's mol_type e db_xref foram omitidos para melhor leitura das informações.

Miotto et al. 2008 ilustram essa problemática em seu estudo com mais de 90.000 registros de proteínas do vírus influenza A, provenientes do Genbank. Os autores apontaram inconsistências associadas a metadados estruturais como nome de proteínas, como também informações de origem biológica (FqS: subtipo do vírus, isolado, hospedeiro, origem geográfica

e ano de isolamento), que impossibilitariam estudos de saúde pública por exemplo, devido principalmente ao grande volume de dados, demandando tempo e recursos excessivos.

Para superar esses desafios, Miotto et al. 2008 propuseram uma abordagem automatizada baseada em regras estruturais e semânticas. As regras estruturais, utilizando a linguagem XPath (CLARK; DEROSE et al., 1999), permitiram a extração eficiente de metadados de campos específicos dos registros. No entanto, a heterogeneidade semântica exigiu a combinação de múltiplas regras para cada propriedade, evidenciando a complexidade do problema. Essa abordagem reduziu significativamente a necessidade de curadoria manual, tornando a análise em larga escala viável e eficiente "para o grupo de estudo".

Esses são exemplos importantes, porém não singulares, que destacam algumas fragilidades do Genbank acerca da conexão de (meta)dados. Tais fragilidades por si só, justificam fortemente o desenvolvimento de novas tecnologias desenhadas para tratar a complexidade dos dados da plataforma no que diz respeito a criação de novas camadas de conectividade entre informações.

Diversas ferramentas podem ser utilizadas para acesso aos dados e metadados do Genbank, entretanto poucas delas permitem o enriquecimento de informações através de conexões de dados. A título de exemplo, a ferramenta de linha de comando `ffq` tem como objetivo facilitar a coleta de metadados e links para dados genômicos brutos em formato JSON a partir de diferentes bancos de dados, incluindo NCBI SRA, GEO, EMBL-EBI ENA, DDBJ GEA e ENCODE (GÁLVEZ-MERCHÁN et al., 2023). Com os dados em mãos, os usuários precisam necessariamente enfrentar a complexidade das informações por si mesmos.

Alternativamente, o utilitário `pysradb` (CHOUDHARY, 2019), é uma ferramenta que fornece uma interface de linha de comando para consulta primariamente de metadados de sequenciamento do grupo SRA. A ferramenta permite a recuperação de metadados e a conversão entre diferentes códigos de acesso baseando-se nas informações dos arquivos `SRAmetadb` (`SRAmetadb.sqlite`). Semelhante ao `ffq`, a ferramenta também não possibilita aos usuários visualizarem conexões entre os dados.

Posso citar ainda o `GEOfetch`, que viabiliza o acesso ao *Gene Expression Omnibus* - GEO (KHOROSHEVSKYI et al., 2023). O principal objetivo da ferramenta é facilitar a recuperação e a organização de (meta)dados num formato padronizado, o *Portable Encapsulated Project* - PEP, que resume o caminho a análises posteriores. `GEOfetch` pode ser utilizada para acessar tanto dados quanto metadados e da mesma forma que os utilitários anteriores, mantém ao encargo dos usuários realizarem o enriquecimento de informações.


Esses exemplos só reforçam a necessidade do desenvolvimento de novas estratégias voltadas ao enriquecimento de informações associadas aos registros do Genbank. Assim, seguindo a via pavimentada até aqui, eu proponho no Capítulo 4, a ferramenta nomeada `GeneConnector`, um utilitário de linha de comando que permite aos usuários realizarem o enriquecimento de

informações associadas a registros do Genbank. Como já informado ao leitor em momentos anteriores, o enriquecimento de informações é executado após o consumo de informações associados aos FqS, elementos obrigatoriamente presentes nos registros da plataforma e que representam importantes fontes de dados para esse propósito.

3.2 Classificação de sequências biológicas

Durante a Seção 3.1 da presente tese, eu entrego ao leitor o suporte teórico sobre os metadados, informações consideradas por muitos "a alma dos dados". Diferentemente, nesse capítulo eu apresento ao leitor informações sobre classificação de sequências biológicas, um ramo essencial das ciências da vida que nos permitem posicionar sequências de DNA ao longo dos ramos da árvore da vida utilizando métodos computacionais de comparação de sequências. Essa seção dá suporte ao Capítulo 5 dessa tese.



Considerando a relação muitas vezes não tão explícita entre os metadados e a classificação de sequências biológicas, utilizaremos o símbolo  para identificar ao leitor onde os metadados são necessários ou estão relacionados as informações apresentadas.

3.2.1 Taxonomia de metagenomas: a abordagem tradicional

Com o constante barateamento no custo do sequenciamento de DNA, impulsionado principalmente pelo advento das tecnologias de sequenciamento em massa (conhecidas como tecnologias de *High-Throughput Sequencing* - HTS) (PAREEK; SMOCZYNSKI; TRETYN, 2011; LOGARES et al., 2012; MARDIS, 2013; MIGNARDI; NILSSON, 2014; HEATHER; CHAIN, 2016; MARDIS, 2017), bancos públicos e privados de informações genéticas vem recebendo um aporte significativo de informações acerca de sequências nucleotídicas e outras informações relacionadas.

O aporte de informação vem sendo tão volumoso, que ainda em 2022 o *Sequence Read Archive* - SRA, o principal banco de referência para arquivos de leituras brutas do Genbank, alcançou a impressionante escala em petabytes de conteúdo armazenado e disponibilizado publicamente (KATZ et al., 2022). Essa explosão de dados foi totalmente esperada, visto que ainda no ano de 2005, os bancos de sequencias pertencentes aos projetos de *Whole Genome Sequence* - WGS, já haviam ultrapassado em proporção as divisões tradicionais de bancos de sequencias (NCBI, 2005).

A fotografia apresentada acima sobre o volume dos bancos é fruto principalmente da adoção em massa das tecnologias de HTS pela comunidade científica, que vem aplicando-a em estudos microbiológicos que acessam DNA ambiental [ou eDNA, acrônimo para *Environmental DNA* (DEINER et al., 2017; RUPPERT; KLINE; RAHMAN, 2019)] a partir de amostras coletadas em ambientes naturais como solo (DUPONT et al., 2016; MAHÉ et al., 2017), água (KARSENTI et al., 2011; LACOURSIÈRE-ROUSSEL et al., 2016) e ar (CLARE et al., 2022; HANSON et al., 2016), ou ainda ambientes artificiais como hospitais (ELRAKAIBY et al., 2019), metrô

(GOHLI et al., 2019) ou cozinhas domésticas (LORIMER et al., 2019). Os mesmos métodos também são amplamente empregados para estudo de microbioma humano, onde provavelmente está o maior pioneirismo da área (HMP, 2012a; HMP, 2012b; WANG et al., 2015).

É nesse contexto que devemos destacar a classificação de sequências biológicas como um importante pilar na execução dos estudos em microbioma utilizando HTS. Ao nos depararmos com produtos de sequenciamento de amostras de solo gerados através dessa tecnologia, somos confrontados a um universo de sequências de DNA totalmente anêmicas no que diz respeito a sua composição biológica. Assim, uma das mais importantes etapas⁴ em estudos de metagenoma se trará da comparação do eDNA contra sequências de referência⁵ (DESAI et al., 2012). Esse procedimento dá as sequências anêmicas um contexto biológico extrínseco a informação já presente nas próprias amostras + sequências (ver simplificação do procedimento na Figura 2).

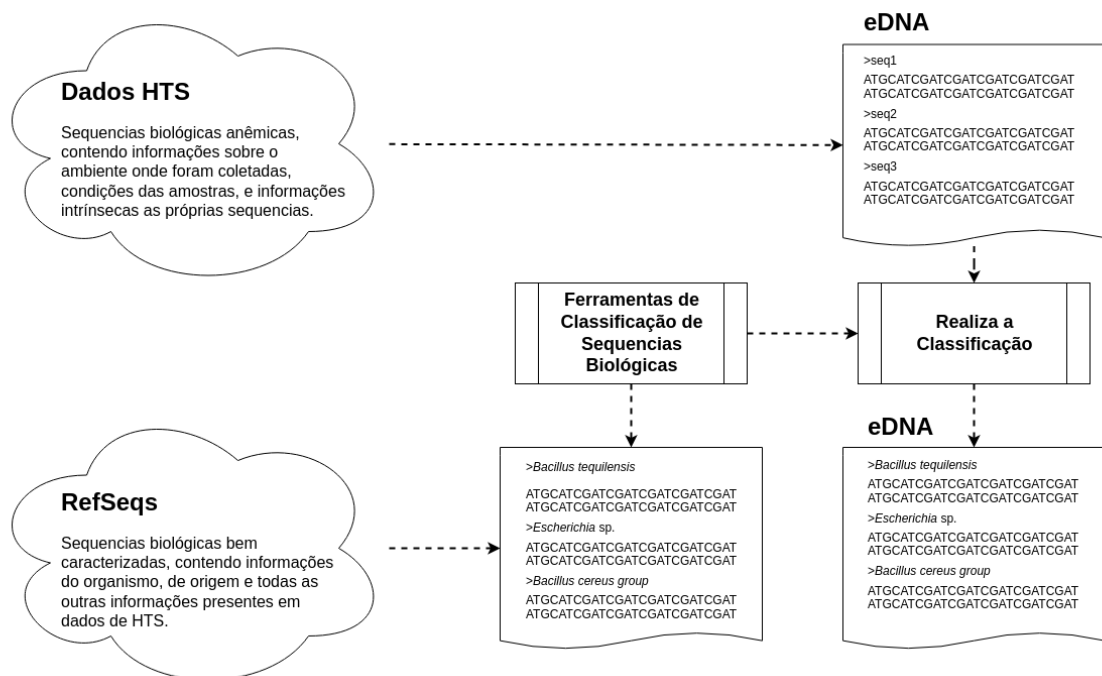


Figura 2 – O papel dos classificadores de sequencia biológicas no enriquecimento de informações de sequências associadas a amostra ambientais anêmicas.

Os métodos mais populares utilizados em procedimentos dessa natureza baseiam-se na busca por sequências presentes em bancos de referencia que possuam maior similaridade⁵ com as sequências presentes no eDNA. Nesse caso assume-se o pressuposto de que, caso a sequência a

⁴ Obviamente, a referida etapa é realizada a jusante da aplicação de métodos para controle de qualidade, junção de pares *forward-reverse*, *assembling*, montagem de *contigs*, remoção de ruídos e de quimeras, entre outros.

⁵ Aqui tratamos a similaridade como um conceito inespecífico. Entretanto esse só é válido quando tratamos da comparação entre sequências de amino-ácidos, em que a similaridade entre sequências é medida somando-se o número de resíduos idênticos (e homólogos) ao número de resíduos similares (que possuem características bioquímicas equivalentes). No caso da comparação entre sequências nucleotídicas, somente a identidade é passível a mensuração [ver Pearson 2013 para uma revisão sobre similaridade].

ser analisada não possui similaridade suficiente com referências já descritas, existe um indicativo de novidade científica em termos de espécie (TEMPERTON; GIOVANNONI, 2012).

Ferramentas que realizam tal procedimento de busca são abundantes e amplamente difundidas em meio a comunidade científica. Atualmente, um dos principais algoritmos utilizado na comparação de sequências por similaridade é o BLAST, com suas variações de algoritmo para alvos específicos (ALTSCHUL et al., 1990). O BLAST é um algoritmo de busca heurística composto por duas etapas principais, a busca e o refinamento. Durante a etapa de busca, *query words*⁶ são utilizadas para buscar no banco de referências as *subjects*⁷, que possuam o maior número de *matches*. As sequências identificadas na primeira etapa são então ranqueadas e enviadas para a etapa de refinamento. Nessa etapa, cada *query word* encontrado em comum entre *subject* e *query* passa pelo processo de extensão, que trata de adicionar resíduos nucleotídicos em ambas as extremidades das *query words* iniciais⁸, até o momento em que as sequências passam a ser divergentes. As sequências com a maior pontuação nessa etapa são novamente ranqueadas e utilizadas no cálculo de estatísticas descritivas da proximidade entre as sequências em cheque.

O algoritmo do BLAST é considerado um método dependente de alinhamento de sequências. Nessa mesma categoria podemos citar o FASTA⁹, uma ferramenta de alinhamento de sequências usada para identificar similaridades entre sequências de DNA e proteínas. O algoritmo opera em quatro etapas: Primeiro identifica segmentos altamente similares entre duas sequências, procurando por "palavras" (semelhantes as *words* utilizadas pelo BLAST) idênticas, de um determinado comprimento (geralmente dois para proteínas e seis para ácidos nucleicos). Essa busca inicial é realizada usando uma tabela bidimensional com foco em diagonais com alta densidade de palavras correspondentes¹⁰. Em seguida o algoritmo coleta as dez melhores diagonais baseando-se na matriz da etapa anterior. Após seleção, as diagonais são unidas seguindo a região da tabela (sempre dentro da diagonal principal) com maior pontuação, para finalmente, criar um alinhamento Smith-Waterman das diagonais identificadas na etapa anterior (PEARSON; LIPMAN, 1988; PEARSON, 2016).

O CLARK pode ser utilizado como ferramenta alternativa de buscar em bancos de dados (OUNIT et al., 2015). A ferramenta pertence a categoria dos classificadores independentes de alinhamento, realizando a busca de sequências utilizando espectros de k-mer's otimizado para cada alvo presente no banco de dados de referência. O CLARK utiliza um conceito de k-mer's discriminativos, na qual durante uma etapa de pré-processamento todos os k-mer's compartilhados entre os grupos a serem preditos são omitidos, restando somente elementos altamente específicos dos alvos. Essa estratégia garante maior eficiência do processo de busca, tornando a ferramenta altamente performática.

⁶ Fragmentos de tamanho fixo da sequência a ser identificada.

⁷ Sequências presentes no banco de referência.

⁸ Essas *words* são denominadas *seed words*.

⁹ O FASTA foi o primeiro software desenvolvido para busca por similaridade em banco de dados.

¹⁰ Em termos gráficos a matriz seria visualizada através de um gráfico de pontos (dot-plot) com foco em diagonais de alta densidade, comumente utilizados em bioinformática.

Ambos – FASTA, BLAST e CLARK – são algoritmos dedicados a busca por similaridades entre pares de sequências, podendo ser utilizados na busca por sequências em bancos de dados de texto ou binários. Sua aplicação prática em estudos de metagenômica se resume em realizar a busca pela identidade taxonômica de sequências coletadas em uma amostra ambiental diretamente contra sequências de espécimes equivalente a "ramos terminais" em uma árvore da vida⁹. Apesar da estratégia ser amplamente aceita em estudos do tipo, essa abordagem possui uma importante fragilidade: ela só permite identificar sequências de espécimes relativamente pouco divergentes aos presentes em bancos de dados de referência. O principal efeito colateral deixado por essa fragilidade é a grande quantidade de sequências sem anotação ao final de bateladas de processamento de amostras ambientais.

Baseando-se nessa fragilidade, novas estratégias foram desenvolvidas para possibilitar a anotação de sequências com maior divergência genética, em relação ao que conhecemos hoje através dos nossos bancos de dados de referência⁹. A principal estratégia adotada como medida mitigatória foi a migração da anotação de "ramos terminais" para os "prováveis ramos internos" da árvore da vida, representados através de nós da hierarquia Lineana. Um dos *softwares* proeminentes na aplicação da hierarquia Lineana na alocação de sequências é o classificador do *Ribosomal Database Project* - RDP (MAIDAK et al., 1997; MAIDAK et al., 2000). O RDP utiliza um algoritmo independente de alinhamentos, e realiza a anotação de sequências utilizando como base de referência, bancos de sequências previamente anotados taxonomicamente⁹. Alguns exemplos de bancos de dados que disponibilizam referência contendo anotações próprias para uso do RDP incluem o próprio banco de dados da ferramenta RDP, o Greengenes (DESANTIS et al., 2006), o Silva (QUAST et al., 2012), e o Unite (KÖLJALG et al., 2005; ABARENKOV et al., 2010; ABARENKOV et al., 2024).

Em termos de funcionamento algorítmico, o RDP utiliza do arcabouço Bayesiano para anotar sequências biológicas baseando-se na sua composição de k-mer's (*words*) de comprimento fixo de tamanho oito (8-mer's). O funcionamento da ferramenta se baseia em quatro etapas. Primeiramente – e não fugindo a regra de todo algoritmo do arcabouço Bayesiano – são calculados os *priors* específicos aos k-mer's, definidos como $[n(w_i) + 0.5]/(N + 1)$, em que $n(w_i)$ representa o número de sequências de DNA no banco de referências que contém o k-mer i , enquanto N , o número de sequências existentes no banco de dados. Já o 0.5 (presente no numerador) e 1 (no denominador) mantém o valor dos *priors* entre 0 e 1. A próxima etapa consiste em calcular a probabilidade condicional de cada grupo taxonômico¹¹ conter cada um dos k-mer's identificados para o banco de dados de referência. Com esses parâmetros gerados, partimos para a próxima etapa onde é calculada a probabilidade de que a sequência *query* pertença a um dado grupo taxonômico dada a probabilidade desta pertencer a qualquer outro grupo. A última etapa consiste simplesmente em estimar a confiança dos resultados obtidos por permutações de Bootstrap (MAIDAK et al., 1997; MAIDAK et al., 2000). Um algoritmo

¹¹ No artigo original somente o gênero foi citado como grupo taxonômico.

elegante e poderoso.

Alternativamente a classificação Bayesiana do RDP, temos o Kraken [versão 1 (WOOD; SALZBERG, 2014) e versão 2 (WOOD; LU; LANGMEAD, 2019)], que baseia o seu processo de classificação em tabelas *hash* que mapeiam para cada k-mer presente nas sequências de referência¹², os nós que conectam-o ao nó ancestral primordial¹³. Assim, a cada rodada de classificação uma única sequência biológica pode ser mapeada para um ou muitas taxa dependendo de sua composição de k-mer's, sendo uma ferramenta poderosa para geração do perfil de sequências. Uma aplicação importante da ferramenta, além da própria identificação da sequência alvo, é na identificação de possíveis contaminações em resultados arquivos multi-*fasta*(q), pois a ferramenta é capaz de gerar um perfil altamente informativo que inclui a proporção em que cada taxon identificado está presente na biblioteca em estudo.

Já na classificação de metagenomas via *shotgun*¹⁴, o Kaiju (MENZEL; NG; KROGH, 2015) realiza a anotação de conteúdo codificante presente nas amostras ambientais, comparando-o diretamente contra um banco de dados de proteínas anotadas de genomas microbianos de referência. O software utiliza a transformada Burrows-Wheeler [a mesma abordagem utilizada pelo software BWA para otimização no alinhamento de sequências (LI; DURBIN, 2009)] para realizar a busca de forma eficiente em grandes bancos de dados de proteínas. Semelhante ao Kraken, o Kaiju utiliza a estratégia de LCA (acrônimo para Last Common Ancestor) para resolução de ambiguidades taxonômicas. Diferentemente do classificador RDP e Kraken, que utilizam apenas *matches* exatos da sequência original durante as buscas, o Kaiju utiliza a busca completa, considerando seis *frames* gerados a partir da sequência original, permitindo a inclusão de substituições de aminoácidos, em uma abordagem gulosa usando a matriz de substituição BLOSUM62 (SONG et al., 2014) para tal. Essa estratégia garante que mesmo sequências com maior divergência do banco de dados possam ser anotadas taxonomicamente. A principal desvantagem de tal abordagem quando comparada aos demais é a sua performance.

Nas últimas décadas os algoritmos de aprendizado de máquina vem se mostrando altamente eficazes em realizar atividades relacionadas a microbiologia (QU et al., 2019; GHANNAM; TECHTMANN, 2021; JIANG et al., 2022) e bioinformática (AO et al., 2022), e a classificação de sequências biológicas é uma delas.

Tratando-se de classificadores que realizam a predição de sequências baseando-se nas hierarquias Lineanas temos o IDTAXA, uma ferramenta pertencente ao pacote R, DECIPHER (WRIGHT, 2016), que realiza a classificação de *amplicons* utilizando uma abordagem nomeada *tree descent*, um processo hierárquico em que as sequências são classificadas usando um conjunto de k-mer's que melhor distinguem os grupos taxonômicos em cada nível da hierarquia Lineana

¹² São usados k-mer's de tamanho 31 na versão 1 e 35 na versão 2 do software.

¹³ Considerando que a ferramenta utiliza como referência o banco de dados de taxonomia do Genbank, o ancestral primordial é representado pelo rank *cellular organisms* (taxid 131567)

¹⁴ Veja Tessler et al. 2017 para uma comparação do uso de técnicas de sequenciamento de *amplicons* e *shotgun* na análise de metagenomas.

(semelhante ao CLARK). Nesse processo, a cada nível hierárquico¹⁵ é calculada a probabilidade da sequência pertencer a cada uma das classes existentes no nível, então o subgrupo com a maior confiança é selecionado e o processo desce para o próximo nível da hierarquia. Essa etapa é repetida até que a confiança seja inferior a um limite predefinido (98% por padrão) (MURALI; BHARGAVA; WRIGHT, 2018).

Além do IDTAXA, que possui uma implementação formal na forma de pacote R, podemos citar ainda o protótipo apresentado por Fiannaca et al. 2018, que realiza a predição de sequências ao longo das hierarquias Lineanas utilizando duas abordagens de aprendizado de máquina profundo, a *Convolutional Neural Network* - CNN, e a *Deep Belief Network* - DBN. Os autores demonstraram que a aplicação principalmente da CNN na predição das taxonomias de *amplicons* de 16S, baseando-se na sua composição de k-mer's, representa uma abordagem promissora para classificação de sequências biológicas. O protótipo dos autores foi capaz de superar os resultados obtidos pelo classificador RDP (MAIDAK et al., 1997; MAIDAK et al., 2000).

Em termos de acurácia, evidências apontam que os métodos de classificação baseados em aprendizado de máquina são inferiores aos demais métodos, principalmente em situações em que os bancos de dados de referência possuem uma cobertura abrangente da diversidade a ser classificada. Entretanto, se tratando da detecção de entidades taxonômicas não mapeadas, o emprego de métodos de aprendizado de máquina pode produzir resultados mais completos, reduzindo o número de espécimes não posicionados taxonomicamente (TIAN et al., 2024).

Nos parágrafos anteriores da presente seção, destacamos os métodos utilizados na classificação de sequências pertencentes a três grupos: os independentemente de taxonomia e baseados em alinhamento [BLAST (ALTSCHUL et al., 1990) e FASTA (PEARSON; LIPMAN, 1988)], o independente de taxonomia e de alinhamentos CLARK (OUNIT et al., 2015), assim como os livres de alinhamento e orientados a taxonomia [RDP (MAIDAK et al., 1997; MAIDAK et al., 2000), Kraken (WOOD; SALZBERG, 2014; WOOD; LU; LANGMEAD, 2019), Kaiju (MENZEL; NG; KROGH, 2015), IDTAXA (MURALI; BHARGAVA; WRIGHT, 2018), protótipo CNN + DBN (FIANNACA et al., 2018)]. Reconheço que todos esses exemplos representam importantes ferramentas que impactaram fortemente a microbiologia nas últimas quase quatro décadas¹⁶. Apesar disso, quando tratamos de ferramentas desenvolvidas para posicionar sequências ao longo da árvore da vida baseando-se em informações taxonômicas, devemos de forma prudente destacar uma importante fragilidade dos métodos: a taxonomia propriamente dita (veja comparação apresentada na Figura 3 para uma simplificação da classificação baseada em taxonomia versus filogenia).

Erros de classificação taxonômica representam um importante ponto único de falha

¹⁵ O processo é iniciado pelo nível mais abrangente (o nó raiz).

¹⁶ Exatos 38 anos desde a criação da primeira ferramenta de classificação de sequências aqui citadas [FASTA (PEARSON; LIPMAN, 1988)].

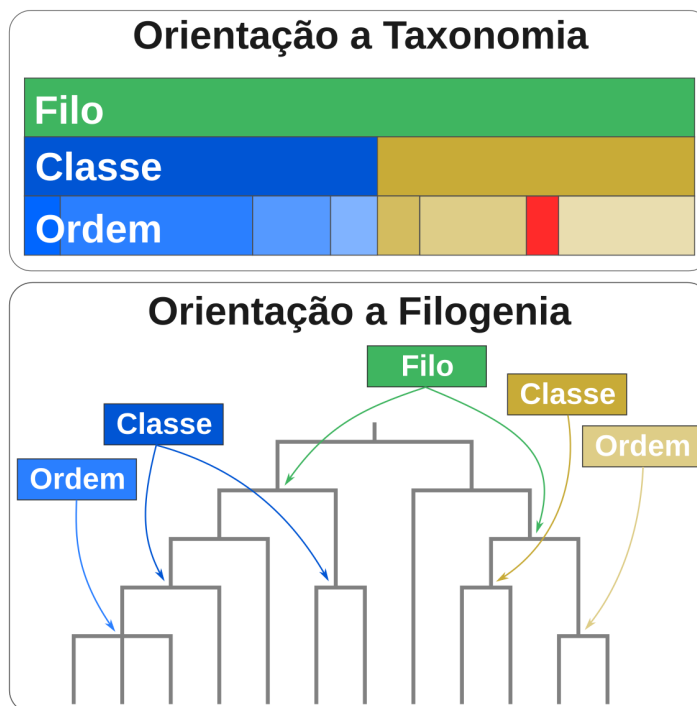


Figura 3 – Comparação das hierarquias para classificação de sequências orientadas a taxonomia (quadro superior) e filogenia (inferior). O retângulo marcado em vermelho no quadro superior representa um erro de posicionamento de um grupo taxonômico teórico.

presente nos nossos sistemas de classificação. Eles são responsáveis por inconsistências já reportadas em importantes bancos de dados de referência utilizados por classificadores amplamente adotados pela comunidade científica¹⁷. Tais erros, em muitos casos, são intrínsecos ao próprio ciclo de vida da taxonomia como processo científico, o que por consequência gerará inevitavelmente efeitos colaterais nas áreas que dependem das informações geradas e atualizadas diariamente por estudiosos da área.

Por esse motivo, novas pesquisas foram desenvolvidas na última década para amenizar o ponto único de falha gerado pela base taxonômica dos classificadores, por uma base menos volátil. Assim, surgiram os métodos de posicionamento filogenético de sequências biológicas (CZECH et al., 2022). Os métodos de posicionamento filogenético utilizam de filogenias pré-existentes como fonte de informação para realizar as "classificações". Por se tratar de um tópico de extrema importância para o desenvolvimento da presente tese, abordaremos o conteúdo na próxima seção.

¹⁷ Edgar 2018 evidenciou erros afetando os bancos Greengenes, RDP e SILVA; Lydon e Lipp 2018 reportaram erro afetando Pseudoalteromonadaceae completamente ao avaliarem o banco Greengenes; Muralidharan et al. 2024 destacaram o efeito de taxonomias transificas em bancos de referência como uma importante fonte de erros em classificadores de sequências.

3.2.2 Filogenias em metagenômica: a abordagem alternativa

Nesta seção, temos o objetivo principal de apresentar ao leitor informações sobre o que são os métodos de alocação de sequências em filogenias existentes e também quais as ferramentas disponíveis, além de realizarmos uma breve introdução sobre o que diferencia tais métodos da construção de filogenias *em si*. Essa última parte é importante para evitar equívocos na compreensão do nosso propósito.

Essa seção se baseia na revisão apresentada por Lucas Czech et al. 2022, que representa uma importante literatura para a área de estudo e é composta por relevantes pesquisadores no assunto:

- **Alexandros Stamatakis**: entre outros projetos, está envolvido diretamente no desenvolvimento da famosa ferramenta de estimação de filogenias RA_{xML} (STAMATAKIS, 2014).
- **Pierre Barbera**: entre outros, atua no desenvolvimento do $EPA-ng$ (BARBERA et al., 2019), ferramenta utilizada na alocação de sequência em filogenias existentes.

Posicionando metagenomas em filogenias

O posicionamento filogenético de sequências de metagenomas representam uma alternativa aos métodos citados ao longo da seção 3.2.1. Diferente das duas principais estratégias citadas (a comparação direta contra ramos terminais da árvore da vida ou a comparação contra nós da hierarquia Lineana), a alocação filogenética realiza o posicionamento das sequências alvo diretamente em filogenias pré-existentes (CZECH et al., 2022).

Na alocação filogenética de metagenomas (a partir daqui nos referiremos como AFM), são utilizadas filogenias previamente existentes e consolidadas, denominadas filogenias de referência - RT (acrônimo para *reference tree*), sobre as quais as sequências de busca - QS (acrônimo para *query sequences*), são confrontadas. Nesse processo, informações sobre as sequências de referência - RS (acrônimo para *reference sequences*), ou mesmo os alinhamentos de referência - RA (acrônimo para *reference alignments*), são utilizadas para alocar as QS (CZECH et al., 2022).

Não devemos confundir essa classe de métodos com a inferência filogenética (IF) propriamente dita (ver resumo dos métodos no Quadro 2), que buscam reconstruir a história evolutiva de um grupo de organismos, estimando as relações ancestrais entre eles. Tais métodos buscam inferir a árvore filogenética mais provável que explique as relações observadas entre as sequências biológicas. A inferência filogenética é particularmente útil para descobrir novas linhagens e entender a evolução de grupos de organismos, mas requer um conjunto de dados cuidadosamente

selecionado e curado para garantir a qualidade da árvore resultante (YANG; RANNALA, 2012).

Quadro 2 | Métodos de inferência filogenética

A inferência filogenética (IF) representa um campo inteiro da bioinformática. Métodos desenvolvidos para esse propósito incluem aproximações baseadas em matrizes de distância evolutivas pareadas entre sequências [o Neighbor-Joining (TREES, 1987) é um dos métodos mais famosos, sendo implementado em inúmeros softwares e pacotes de bioinformática e estatística de propósito geral], otimização por evolução mínima via Parcimônia [Fitch 1971 descreve o processo, sendo o PAUP (WILGENBUSCH; SWOFFORD, 2003) um dos softwares mais populares], otimização por Máxima Verossimilhança [Joseph Felsenstein foi o precursor do uso da verossimilhança em filogenias (FELSENSTEIN, 1981); softwares que realizam tal aproximação incluem o RAxML (STAMATAKIS, 2014) e PhyML (GUINDON et al., 2010)] e Inferência Bayesiana [(RANNALA; YANG, 1996; YANG; RANNALA, 1997); essa é a principal abordagem utilizada na reconstrução não somente de filogenias simples realizada por softwares como MrBayes (RONQUIST et al., 2012), mas também no cálculo de relógios moleculares pelo Beast (DRUMMOND; RAMBAUT, 2007; BOUCKAERT et al., 2014; BOUCKAERT et al., 2019)].

Na execução dos métodos de IF, a cada nova análise todas as relações evolutivas entre as sequências alvo + QS são revistas. Um processo minucioso e computacionalmente intensivo. Considerando a escala dos resultados obtidos em estudos de HTS, tais métodos são inviáveis. Assim, nos métodos de AFM tentamos obter somente o posicionamento das QS em meio aos ramos terminais das RT, saindo da escala exponencial da IF para a escala (quase)linear, relativa somente ao número de nós já existentes nas RT e a dimensão do eDNA a ser analisado [alguns algoritmos como o PPLACER (MATSEN; KODNER; ARMBRUST, 2010) entregam uma escala linear durante o processo de alocação].

Além da escalabilidade, o uso da tecnologia de AFM provê aos resultados um contexto mais informativo e preciso. Jamy et al. 2020 demonstram a superioridade da AFM [utilizando o software EPA-ng (BARBERA et al., 2019)] em relação ao método baseado em similaridade utilizado pelo VSEARCH (ROGNES et al., 2016) em vários aspectos. Os autores observaram que o método baseado em filogenia classificou 43,7% das sequências com baixa similaridade (<80%) a sequências de referência em linhagens diferentes do vsearch, chegando a divergir no nível de supergrupo em alguns casos. Essa diferença diminuiu para sequências com maior similaridade, mas ainda persistiu em menor grau. Um exemplo notável foi a classificação de uma sequência pertencente ao supergrupo Hemimastigophora, que o VSEARCH classificou incorretamente como planta terrestre devido à similaridade com sequências de Streptophyta [utilizando o banco de dados SILVA (QUAST et al., 2012)]. O método baseado em filogenia, por outro lado, a identificou como um "eucarioto não identificado", evidenciando sua maior precisão na ausência de referências próximas. Além disso, o método baseado em filogenia demonstrou

um comportamento mais conservador, classificando algumas sequências em níveis taxonômicos mais altos do que o *VSEARCH* na ausência de referências próximas.

A caixa de ferramentas

Dentre as ferramentas atualmente disponíveis para AFM, podemos destacar duas linhas principais de implementação. A primeira se baseia na busca pela posição evolutiva das QS em relação a RT, utilizando métodos de inferência por Máxima Verossimilhança - ML (acrônimo de *Maximum Likelihood*) e Inferência Bayesiana - BI (acrônimo de *Bayesian Inference*), já a segunda linha de implementação [aqui estão os métodos mais performáticos], estão os métodos baseados em distância evolutiva entre as sequências do RA e QS.

Os métodos pertencentes a primeira linha de softwares foram os primeiros a serem desenvolvidos, com o *PPLACER* sendo a mais antiga ferramenta descrita, há pouco mais de dez anos antes do presente ([MATSSEN; KODNER; ARMBRUST, 2010](#)). O *PPLACER* utiliza duas abordagens para estimar o posicionamento das QS: a aproximação por ML e estimativa da probabilidade posterior por BI. Durante a análise por ML, o *PPLACER* calcula a razão de verossimilhança para cada posição de inserção da QS, que é a razão entre a verossimilhança da melhor inserção em cada ramo e a soma das verossimilhanças de todas as inserções possíveis. Essas informações são utilizadas para quantificar a incerteza associada ao posicionamento das QS's ao final do processo de alocação. Alternativamente, durante a análise por BI, a ferramenta calcula a probabilidade posterior das QS's serem alocadas em cada ramo da árvore, condicionada a topologia e os comprimentos de ramo.

O *RAxML-EPA* é uma ferramenta bastante semelhante ao *PPLACER*, entretanto utiliza uma busca eurística para a melhor hipótese de ramo para alocação da QS ([BERGER; STAMATAKIS, 2011](#)). O algoritmo do *RAxML-EPA* inicia a busca inserindo a QS em um ramo arbitrário da RT, em seguida, avalia iterativamente as verossimilhanças de inserção em ramos vizinhos, movendo-se para o ramo com a maior verossimilhança em cada etapa. Parâmetros de topologia e comprimento de ramos são otimizados durante a busca.

Como ferramenta alternativa ao *PPLACER* e *RAxML-EPA*, o *EPA-NG* realiza o processo de busca pela melhor alocação da QS utilizando um algoritmo de duas etapas: A pré-alocação e a alocação completa. Na pré-alocação algoritmo seleciona rapidamente um subconjunto de ramos candidatos na RT onde a QS possui maior verossimilhança. Já na alocação completa, é realizada uma nova otimização considerando no cálculo da ML um modelo de substituição nucleotídica e a própria topologia da árvore. A ferramenta se difere das duas citadas anteriormente principalmente no que diz respeito a escalabilidade de execução em arquiteturas multi-core. Ideal para grandes volumes de dados. Em termos de semelhança, o *EPA-NG* realiza a seleção dinâmica de ramos candidatos do *RAxML-EPA* e a heurística de "mascaramento" do *PPLACER*, que remove partes não informativas das sequências para acelerar o processo ([BARBERA et al., 2019](#)).

A próxima ferramenta, o *RAPPAS* (acrônimo para *Rapid Alignment-free Phylogenetic*

Placement via Ancestral Sequences) foi contemporânea ao EPA-NG (LINARD; SWENSON; PARDI, 2019), porém atualmente a ferramenta foi depreciada em detrimento do EPICK (acrônimo para *Evolutionary Placement with Informative K-mers*) (ROMASHCHENKO et al., 2023). O software baseia seu algoritmo de busca no conceito de phylo-k-mer's, que são identificados a partir do mapeamento dos k-mer's presentes nas RS's para os nós internos da RT. Ao realizar a alocação o RAPPAS mapeia os k-mer's da QS contra os nós da RT, calculando a probabilidade (gerada a partir do conjunto de k-mer's mapeados) de cada sequência pertencer aos devidos nós. O EPICK, sucessor da ferramenta, implementa algumas melhorias tanto na realização da otimização, utilizando ML para busca da melhor posição para alocação ao longo da RT, além de realizar a filtragem de k-mer's informativos durante o processo de predição, ganhando performance no processo. Diferentemente do PPLACER, RAxML-EPA e EPA-NG, ambos o RAPPAS e EPA-NG são ferramentas livres de alinhamento.

Em adição as ferramentas de AFM baseadas em otimização por ML, temos as ferramentas que realizam o processo de busca baseando-se na distância genética entre sequências. Nessa categoria devo citar o APPLES (acrônimo para *Accurate Phylogenetic Placement using LEast Squares*), que conforme o próprio nome informa, utiliza quadrados mínimos para buscar o melhor nó de ancoragem na RT. Seu algoritmo possui duas etapas, a pré-computação e otimização. Na etapa de pré-computação, o algoritmo calcula um conjunto de valores fixos para cada nó da RT, contendo a distância da QS em relação as demais RS's, usados para acelerar os cálculos subsequentes. Na etapa de otimização, o algoritmo utiliza programação dinâmica para encontrar a posição ideal para alocar a QS, de maneira que o erro quadrático, entre as distâncias da QS em relação as demais RS, seja minimizado. O APPLES oferece flexibilidade na escolha do modelo de distância e da função objetivo, permitindo a otimização tanto pelo método de mínimos quadrados quanto pelo princípio de evolução mínima, além de permitir que os usuários utilizem alinhamentos múltiplos ou matrizes de distância como entrada de dados (BALABAN; SARMASHGHI; MIRARAB, 2020).

Por outro lado, o APP-SPAM (acrônimo para *Alignment-free Phylogenetic Placement Algorithm based on Spaced-word Matches*) utiliza o conceito de *spaced-word matches* (SpaMs) para calcular as distâncias entre QS e RS's. Um SpaM é definido por um padrão binário, que consiste em uma sequência de 0's e 1's, onde 1 indica uma posição de correspondência (*match*) e 0 uma posição de não-correspondência entre os pares a serem comparados. Os SpaM's pareados entre QS e RS são calculados na primeira de três etapas que compõe o algoritmo da ferramenta. Na segunda etapa é realizada a estimação da distância filogenética baseada em SpaM's. Com a matriz de distâncias estabelecida, o algoritmo utiliza busca eurística para identificar o nó correspondente da RT onde a QS irá ser posicionada. Na busca eurística podem ser utilizada a distância mínima entre QS e RS presentes no nó alvo, contagem de SpaM's ou ainda optar por utilizar o LCA compartilhado entre os componentes do nó alvo.

Algumas das técnicas citadas acima para AFM, representam os pilares centrais de impor-

tantes *frameworks* disponíveis à comunidade científica para realizarmos a agregação de valor científico. Os *frameworks* podem utilizar uma ou mais técnicas para identificarem características funcionais ou composicionais nas comunidades microbianas, e representam importantes peças da caixa de ferramentas de cientistas microbianos. Uma das mais proeminentes ferramentas com esse objetivo é o projeto `bioBakery 3` (BEGHINI et al., 2021), uma plataforma de análise metagenômica que integra um conjunto de ferramentas para caracterizar comunidades microbianas a partir de dados de sequenciamento. Ela inclui módulos para controle de qualidade (`KneadData`), perfil taxonômico (`MetaPhlAn`), perfil funcional (`HUMAnN`), perfil de linhagem (`StrainPhlAn 3` e `PanPhlAn 3`) e filogenética (`PhyloPhlAn 3`). Essas ferramentas utilizam um banco de dados atualizado de genomas microbianos e agrupamentos de famílias de genes (`ChocoPhlAn 3`) para realizar análises abrangentes e precisas. O objetivo principal da `bioBakery 3` é fornecer uma caracterização completa das comunidades microbianas, desde a identificação e quantificação de espécies até a análise funcional e a reconstrução de linhagens, permitindo uma compreensão mais profunda da ecologia e do papel dos microrganismos em diversos ambientes.

Ao longo dos parágrafos dessa seção, eu apresento ao leitor informações acerca do que é o processo de AFM, assim como eu busquei abordar a amplitude de métodos e estratégias utilizadas na anotação filogenética de sequências. Essas informações servirão de sustentação teórica para o Capítulo 5 da presente tese, na qual eu apresento o `Classeq`, uma ferramenta dedicada a AFM, livre de alinhamentos, baseada em k-mer's.

4 GeneConnector: Unlocking the Full Potential of Genbank Metadata

Table 2 – GeneConnector manuscript’s authors with affiliation.

Author	ORCID	Affiliation
Samuel Elias Galvão	0000-0001-9138-8845	University of Brasilia, Graduate Program in Microbial Biology, Institute of Biological Sciences, Brasilia, DF, Brazil and Bioinformatic Researcher in Biotrop, Solutions in Biological Technologies
Débora Guterres Cervieri	0000-0002-3902-8487	Post-doctoral Researcher in Federal University of Viçosa
Robert Barreto Weingart	0000-0001-8920-4760	Titular professor at the Department of Phytopathology at the Federal University of Viçosa
Helson Mário Martins do Vale	0000-0002-5452-3873	Associate Professor at the University of Brasília, Department of Phytopathology, Institute of Biological Sciences, Brasilia, DF, Brazil

Abstract

The Genbank database serves as a pivotal global repository for genetic information, housing an extensive and diverse array of data. Nonetheless, a significant proportion of its existing records suffer from fragmented and often inadequate metadata, thereby failing to furnish the requisite contextual information regarding their acquisition. In response to this challenge, we introduce GeneConnector, a novel tool designed to harness shared information within multiple records of the same specimen in Genbank, with the ultimate objective of augmenting the completeness of inadequately annotated nodes spanning various information domains. To exemplify the capabilities of this tool, we conducted a comprehensive review and aggregation of available data, utilizing the database for Genera of Phytopathogenic Fungi (GOPHY) as a testbed. Our evaluation revealed substantial improvements in information retrieval through the analysis of shared data among nodes that connect Genbank specimen records, yielding notable enhancements ranging from 2% to an astonishing 60%. Our approach equips users with the means to conduct precise, facile, and accurate assessments of the contextual associations of results, facilitated by two distinct metrics that assess the current level of data annotation and the potential information enhancement achievable through our evaluation, the Observed Completeness Score (OCS) and the Reachable Completeness Score (RCS).

Keywords: Genbank, NCBI, Gene-Connector, Mycology, Phytopathology, GOPHY.

4.1 Introduction

Genomic data has become increasingly important in many fields of research, including medicine (SHENDURE; FINDLAY; SNYDER, 2019), biotechnology, and agriculture (JEYASRI et al., 2021; JUMA, 2014; GILCHRIST; WANG; QUILICHINI, 2023). However, the sheer volume and complexity of this data can make it challenging to extract meaningful insights. Public databases, such as GenBank (BENSON et al., 2012), Ensembl (HOWE et al., 2021), and UniProt (CONSORTIUM, 2019), provide a wealth of information on genes, genomes, and their products. However, accessing and analyzing this information can be a time-consuming and daunting task. Therefore, the development of tools that can perform the aggregation of genomic metadata in public databases is critical to advancing research in genomics.

One example of such a tool is BioMart (SMEDLEY et al., 2009), a widely used data management system that allows users to query "data" (only) from multiple biological databases simultaneously. BioMart has been used in many studies, including the identification of genetic markers associated with disease and the exploration of gene expression patterns in different tissues. Another example is Ensembl's Biomart (see details), which allows users to query Ensembl's databases using the same interface as BioMart. These tools are just a few examples of the many resources available to researchers looking to access and analyze genomic "data".

Quadro 3 | The *Ceratocystis mangicola* (WYK et al., 2011) study-case.

ITS Submitted in Mar 4, 2005 [available under the Genbank accession nº [AY953382](#), (WYK et al., 2005)].

EF1 Feb 13, 2007 [[EF433316](#), (WYK et al., 2007)].

TUB1 Feb 13, 2007 [[EF433307](#), (WYK et al., 2007)].

RPB2 Mar 11, 2014 [[KJ601618](#), (FOURIE et al., 2015)].

MS204 Mar 11, 2014 [[KJ601582](#), (FOURIE et al., 2015)].

A pathogen originally described as member of the *Ceratocystis fimbriata sensu lato* complex causing the *Mangifera indica* disease, known as mango blight, murcha, or *seca da mangueira* in Brazil. Records of *C. mangicola* were registered in three different submission events, with a large time lag between the first (the Internal Transcribed Spacer [ITS] submission) and the latest submission events (RNA polymerase subunit II [RPB2] and guanine nucleotide-binding protein subunit beta-like protein [MS204], nine years after). Such time lag allowed the information associated to the *C. mangicola* to be gradually extended. Since the first registration of the ITS marker, the associated information was upgraded, starting from basic source modifiers as isolate and organism name to a well documented record including strain, specimen-voucher, type-materials, host, country, and others (see the Fig. 4 for details of the information gain associated to *C. mangicola*).

As attempted readers can see, the "data" has been the center of attention when it comes to data aggregation, while metadata is much more often overlooked. Therefore, the aggregation of genomic metadata is important because it enables researchers to integrate data from multiple sources and make more comprehensive and accurate analyses [important examples includes (CANAKOGLU et al., 2019; CHEN et al., 2022; KÖLJALG et al., 2005; ABARENKOV et al., 2010)]. For example, by combining genomic data with clinical data, researchers can identify genetic markers associated with disease and develop more effective treatments. The aggregation of genomic metadata also enables the identification of patterns and trends that may not be apparent when examining individual datasets. These patterns and trends can provide insights into the most variable scientific domains.

Despite the existence and importance of the tools that performing aggregation of genomic metadata from single records, and focused in high-throughput sequencing data [examples include Metagenote (QUIÑONES et al., 2020), and ffq (GÁLVEZ-MERCHÁN et al., 2023)], there are no tools that aggregate multi-loci data. There are still challenges associated with accessing and analyzing such data, and information consistency is maybe the most important of these [see (CHEN; ZOBEL; VERSPOOR, 2017) for a important study-case about Genbank information

consistency]. GeneConnector works around such challenges. Our proposal is to create connections between unique Genbank records and use the "unique + shared" information between records to improve single gene annotations.

When specifically dealing nucleotide data stored in Genbank, it is common to observe events of information increment associated with advancements in knowledge regarding taxonomic groups [see Box 3 for an example]. The phenomenon of information increment events can be attributed to the dynamic nature of scientific research. As researchers delve deeper into the genetic makeup of various organisms, they uncover novel data points and identify previously unrecognized patterns.

These discoveries, when incorporated into the Genbank database, enhance the breadth and depth of information available for taxonomic analysis. Consequently, with each advancement in our understanding of taxonomic groups, a ripple effect occurs, influencing future studies, expanding our knowledge base, and fostering further scientific breakthroughs [the natural stepping stones of science (REINING et al., 2022)].

Finally, GeneConnector was designed precisely to absorb this intrinsic characteristic of Genbank data during metadata acquisition campaigns. Therefore, our aim is to illustrate how this tool can enhance the metadata quality of a comprehensive database solely by leveraging the shared information within Genbank records. Furthermore, we introduce our novel approach, the Observed Completeness Score (OCS) and the Reachable Completeness Score (RCS), for quantifying the level of completeness in records associated with specimens with available information in Genbank. To accomplish this objective, we employed the comprehensive database for Genera of Phytopathogenic Fungi (GOPHY) as a case study (MARIN-FELIX et al., 2017a; MARIN-FELIX et al., 2019c; MARIN-FELIX et al., 2019a; CHEN et al., 2022a).

4.2 Problem statement

Genbank, a widely used repository for nucleotide sequence data, contains an immense amount of valuable genomic information. However, the lack of consistent and standardized metadata across the records poses a significant challenge for researchers aiming to extract meaningful insights from this vast collection. Existing approaches for metadata extraction and aggregation from Genbank records are often limited, inefficient, or require manual curation, hindering the ability to comprehensively exploit the data for scientific research.

To address this problem, a novel software solution has been developed to automate the process of populating and aggregating metadata from Genbank nucleotide records. The software aims to extract diverse metadata attributes, including taxonomy, organism properties, sequencing techniques, geographical location, and biological features, among others, from the extensive Genbank database. By automating this labor-intensive task, researchers will be empowered to efficiently access and analyze metadata associated with nucleotide sequences, enhancing their

ability to conduct comprehensive studies and accelerate advancements in various biological and genomic fields.

The research article aims to evaluate the effectiveness and reliability of the GeneConnector in extracting and aggregating metadata from a large-scale sample of Genbank nucleotide records. The outcomes of this research will contribute to improving the accessibility and quality of metadata associated with Genbank nucleotide records.

4.3 Proposed solution

4.3.1 Concepts and Information Modelling

Our tool was developed to modelling the information contained in Genbank records based in three basic data models: Metadata, Nodes, and Connections (see Fig. 4). A *Connection* is a top-level object centralizing *Nodes*. A single *Node* carries information of the accession number that originated the object, and the gene marker from which the record was extracted, connecting all metadata related to the original Genbank record. *Metadata* objects abstract the Genbank raw qualifiers information.

Raw Genbank qualifiers includes a list of key/value pairs describing the context associated to given nucleotide record. Our tool was developed to turn qualifiers into importance groups (from here on we will call them Metadata Indicator Groups, MIG) that mirrors the information relevance which turn a desired specimen unique (information importance are available in Table 3, and visually explored in Fig. 4). For example, a taxonomic key (e.g. organism [with value *Bipolaris victoriae*]) is shared between multiple real world specimens, so it must have less importance than a specimen related key (e.g. isolate [with value *CBS:327.64*]).

Table 3 – Metadata Indicator Groups used to rank keys and calculate the Gene Connector completeness scores.

Group	Score	Description
SPECIMEN	8	Unique identifiers of specimen.
TAXONOMY	5	Taxonomy related keys.
HOST_SUBSTRATE	3	Identifiers for host interactions.
TIME_REFERENCES	2	Time milestones.
GEO_REFERENCES	2	Geographical indicators.
ASSAY	0	Related to gene sequencing methods.
EXTERNAL_LINKS	0	References to external databases.
ACTORS	0	Human actors related to the record.
OTHER	0	Not already mapped keys.

Based on these principles, our tool systematically punctuates metadata from independent Genbank records, and calculates the information completeness associated to a set of records that represents real world specimens, improving the information usability.

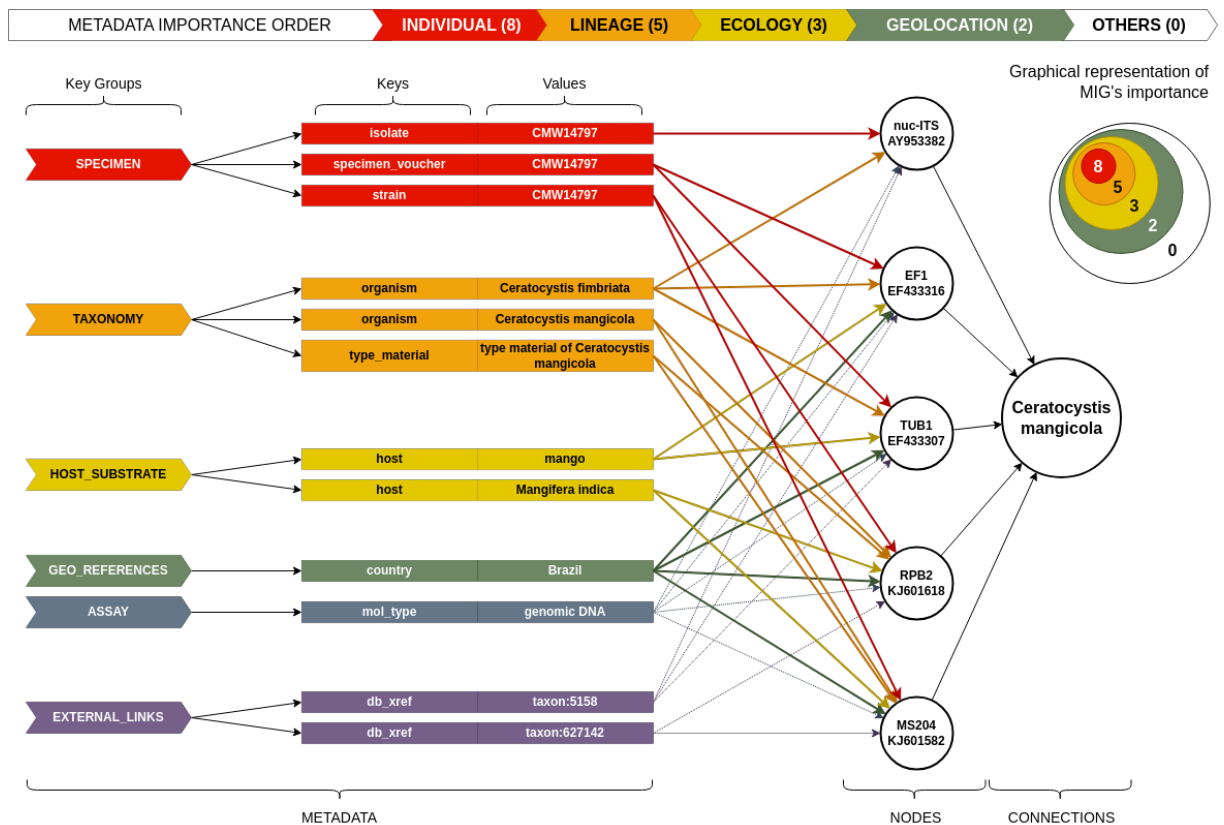


Figure 4 – GeneConnector information diagram. The diagram represents the main information models used by our tool to deal with Genbank records information: Metadata, Nodes, and Connections. Dotted arrows connecting metadata and nodes layers indicating zero scored MIG's. Warmer colors indicate more specific information about the specimen to which the nodes (genes) belong. Therefore, the closer to red indicates metadata with greater power to approximate nodes.

As already demonstrated in Table 3, the MIG importance score is expressed on a Fibonacci scale and enables the differentiation of important metadata from spurious ones, thereby facilitating the calculation of two completeness scores: the Observed Completeness Score (OCS) and the Reachable Completeness Score (RCS).

The OCS measures how well a connection is annotated in terms of information domains taking into account parameters as uniqueness (SPECIMEN), tree of life placement (TAXONOMY), ecological placement (HOST_SUBSTRATE), temporal marks (TIME_REFERENCES), spatial marks (GEO_REFERENCES), and less relevant ones (ASSAY, EXTERNAL_LINKS, ACTORS, and OTHER). The OCS is calculated independently for each node.

The RCS is a metric used to assess the completeness of connections based on the nodes they connect. Unlike the OCS, which considers all nodes independently, the RCS takes into account the dependencies between nodes that composes a connection. Specifically, the RCS is calculated as the ratio of the number of observed connections between nodes to the number of possible connections within a given MIG. This score reflects the degree to which the available

metadata within a given MIG is interconnected and should be used to complete another nodes.

The three main steps of an hypothetical calculation of the OCS and RCS are described below where the results of individual steps are shown in Table 4.

- Step 1. Finding nodes with at least one occurrence of qualifiers of each MIG with a score greater than zero:

Let N be the set of nodes.

Let Q be the set of qualifiers.

Let M be the set of MIGs.

The condition to find such nodes can be represented as:

$$\forall n \in N, \exists m \in M, \exists q \in Q \text{ with } \text{score}(m) > 0 \text{ such that } \text{occurs}(n, q)$$

- Step 2. Annotating nodes with the MIG score (no more than one key per MIG scored by a node):

Let S be the function that assigns a score to a node.

The annotation can be represented as:

$$\forall n \in N, \exists m \in M, S(n) = \text{score}(m)$$

- Step 3. Calculating the expected score during the second step:

Let E represent the expected score.

The calculation can be represented as the sum of products:

$$E = \sum_{m \in M} \text{score}(m) \cdot \text{number of nodes}$$

Finally we filtering expected scores to find MIGs with at least one member (penalizing MIGs with a zero score if not represented):

Let F be the set of MIGs with at least one member.

The filtering can be represented as:

$$F = \{m \in M \mid \exists n \in N \text{ such that occurs}(n, m)\}$$

The penalization of MIGs not represented can be represented as:

$$\forall m \in M \setminus F, \text{score}(m) = 0$$

Table 4 – The first three steps of the completeness scores calculation with hypothetical nodes A, B, C, and D.

Group	Step 1				Step 2	Step 3
	A	B	C	D	E-score [†]	0-score [‡]
SPECIMEN	8	8	-	8	32	32
TAXONOMY	5	5	5	5	20	20
HOST_SUBSTRATE	-	3	3	-	12	12
TIME_REFERENCES	-	-	-	-	8	0
GEO_REFERENCES	2	-	2	-	8	8
	15	16	10	13	80	72
Conn. Obs. score	54					
OCS	0.68					
RCS	0.90					

Step 1 contain four hypothetical nodes A, B, C, and D with group scores, respective. Dash indicate groups not represented in Node. Step 2 and Step 3 includes expected scores, and non-zero group scores, respectively. [†] Expected score by group. The product of the nodes number and the score value of the given group. [‡] Non-zero score by group. The same as expected score if at last one Node contains a given group. Otherwise is zero.

After execute the above steps we can calculate the Connection Observed Score by sum individual node scores (conn. obs. score = 54 in Table 4). Next, the OBS is calculated being the ratio between the Connection Observed Score and the sum of Expected scores ($54 / 80 = 0.68$ in Table 4), and finally the RCS should be calculated as the ratio of the step 3's values sum and the sum of Expected scores ($72 / 80 = 0.90$ in Table 4). This is a simple and elegant way to represents the information completeness of arbitrary Genbank records.

4.3.2 Technologies and Code Availability

GeneConnector was developed in Python [3.11+ (ROSSUM; DRAKE, 2009)] adopting the hexagonal architecture (COCKBURN, 2006). The complete logic for the calculation of scores,

data parsing, data validations, and the data collection from Genbank are centered at the package core sub-module. For curious readers, a complete metadata list by MIG should be found at the Github repository [sgelias/gene-connector-cli](#)) within the ‘metadata’ file [src/gcon/core/domain/dtos/metadata.py](#)). Our tools is Open Source and the codebase is available under the MIT license (see [details](#)).

4.3.3 Study case: GOPHY data completeness

To demonstrate the performance and value proposition of GeneConnector we downloaded and evaluate the complete GOPHY’s database containing seventeen gene markers and 1,246 specimen records. The complete database is available as a Supplementary material into the GeneConnector Github directory (files named `gcon-input-gophy.xlsx` in [docs/manuscript/supplementary-material](#)).

We value simplicity, so we make running the GeneConnector possible through a single command named **resolve** available after the tool installation on the host system. Currently our tool was tested only using Linux systems, thus, over Windows or Macintosh systems we recommend to run using a Docker environment ([MERKEL, 2014](#)). See below the execution command of GeneConnector CLI:

The Code snippet of Listing 4.1 exemplifies our package execution. After installed GeneConnector should be called using the `gcon` callable and the `resolve` command used to execute the full package pipeline. Required arguments are shown in lines 5, 6, and 7 of the previous code snippet. A comprehensive user guide is available at the GeneConnector [Github directory](#).

```
1 # GeneConnector execution in Linux environments
2 # using the ‘resolve’ command of the ‘gcon’ package.
3
4 $ gcon resolve \
5     --input-table input-table.tsv \
6     --temporary-directory /tmp/gcon/ \
7     --output-file gcon-out
```

Listing 4.1 – GeneConnector execution command example. Lines started with hashtag are code comments, so they are not executed.

The output generated by the aforementioned command encompasses a tabular file (TSV) that amalgamates several crucial components: (i) input table information, (ii) OCS and RCS scores, (iii) a statistical percentage depicting the information gain, which quantifies the quantity of information salvaged following the evaluation of metadata under the *Nodes* category, (iv) signatures, and ultimately, (v) all metadata associated with individual connections.

Signatures offer a streamlined mechanism enabling researchers to trace, index, or effortlessly compare results across multiple analyses conducted at disparate times. Our tool incorporates two distinct levels of signatures: the *connection-level* and *node-level* signatures, grounded in standard Universal Unique Identifiers (UUID) of version 3 hashes. These hashes are derived by compressing the most pivotal data elements that constitute *Nodes* (comprising Genbank accession, source genome, gene name, and metadata keys and values) and *Connections* (encompassing identifiers and node signatures). Such an approach empowers users with the capability to replicate results and swiftly compare records when necessary.

Metadata columns are composed of the MIG keys concatenated to metadata keys (as example SPECIMEN.isolate). Such way turn the further integration and indexing as a simple and natural way to store GeneConnector results. In addition to the above cited tabular file, the GeneConnector results includes at default a JSON¹ formatted output file as a optimal format to be inputs into ETL² pipelines and web integrations.

4.4 Results

4.4.1 MIG's representativity and distribution

Analysis of the complete GOPHY's database resulted in 414 events of information gain³ from the total of 1,246 specimen records. These amount comprises 33% of the database records suffering information gains. Gains ranged from 2% up to 60%, widely distributed along all fungal genus included in our analysis. Twenty-five of the twenty-nine genera present in GOPHY were contemplated with information gains. The complete tabular results is available as a supplementary material.

The most important MIG obviously was SPECIMEN, with *strain* and *culture_collection* as the most populated keys, with 86.3% and 69.3% of coverage. It was not surprisingly due to the nature of the GOPHY database proposal itself, including only high quality records, mainly belonging to type materials.

Next, the TAXONOMY MIG with *type_material* as the second⁴ most important key, with >69% of coverage in records. Both SPECIMEN and TAXONOMY are the most important keys to make each Connection record unique. And precisely for this reason that both are the best scored in our tool (see Table 3).

The third most important MIG for GeneConnector approach is HOST_SUBSTRACT. Both keys *host* and *isolation_source* covered approximately 62% and 40%, respectively, of the

¹ Javascript Object Notation format.

² Extract, Transform, and Load pipelines.

³ Calculated as the percentage of the Reachable Completeness Score which the Observed Completeness Score comprises.

⁴ Organism is a required field, so it has full coverage in Genbank nucleotide records.

full GOPHY dataset.

The next important MIG's is GEO_REFERENCES. It was present in about 76% of records, however in the most cases refereed to only as country. This key in most cases is not so geographically resolute when dealing with countries of continental dimensions, such as Brazil or Australia. From 1,246 records, just one included information of Latitude/Longitude, completely inhibiting the performance of geographic analyzes.

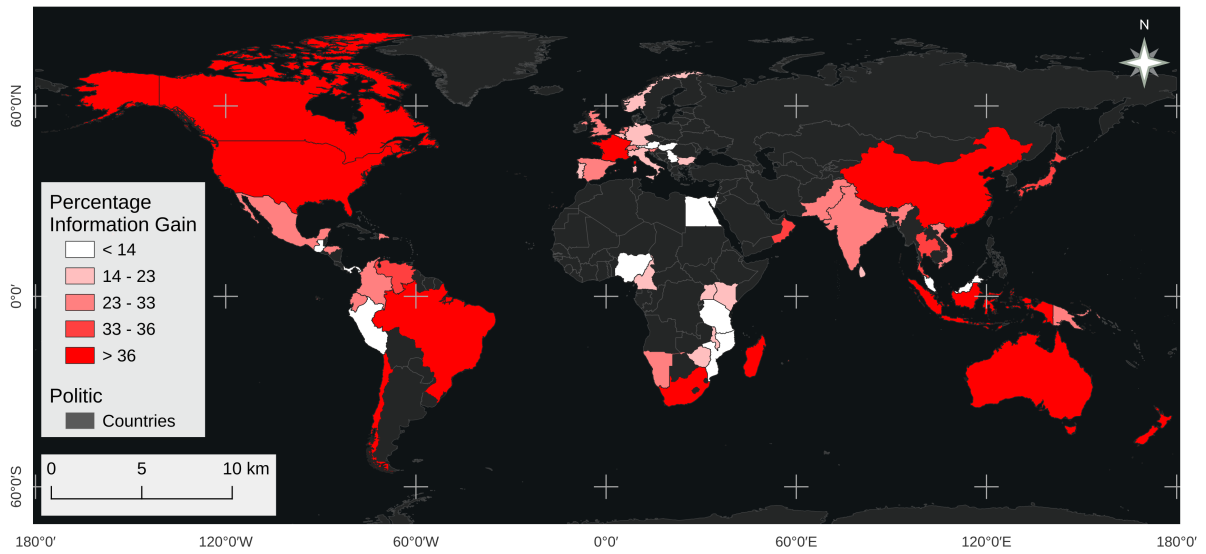


Figure 5 – Information gain at the globe scale. Records with some information gain registered in this study are highlighted in white-to-red scale (see scale legend).

A world scale map indicating the geographic range of the records, and including the maximum information gain reachable by country is shown in Figure 5. The 10 countries covered with the highest number of information gain events were China, United States, Australia, South Africa, Brazil, Thailand, Netherlands, Indonesia, Japan, and Ecuador. The maximum information gain reached by such countries ranges between 30% and 50%. This is proportional to the country contribution to the state of the art of the phytopathogenic fungi records, a fully expected scenario.

Different from the previous cited MIG's the TIME_REFERENCE was an exception. Only about 6.9% (86 records) of the GOPHY database contains time milestones. Despite such MIG is not highly scored in GeneConnector (score = 2), the absence of this information inhibits temporal interpretation of the collection effort on the phytopathogenic important fungi around the world.

4.4.2 Phytopathogenic completeness along GOPHY genus

Information gains by genus are shown in Figure 6. As above cited, information gains ranged from 2% up to 60%. The top ten phytopathogenic fungal genus with most number of specimen records suffering information gains were *Calonectria* with 73 events, *Diaporthe* (55), *Curvularia*

(48), *Colletotrichum* (41), *Ceratocystis* (23), *Bipolaris* (21), *Boeremia* (19), *Neofusicoccum* (17), *Phyllosticta* and *Huntiaella* with (16).

Using our approach 8 of 25 genus with information gains (GOPHY database include information of 29 genus) reached the full information completeness (100% of completeness, RCS = 1.0), grouping at last one of each MIG qualifier key per connection. A significant information gain in terms of the complete database. As can be seen in Fig. 6, median values of RCS were up to 90% in nine of ten most representative genus of GOPHY (cited in the previous paragraph).

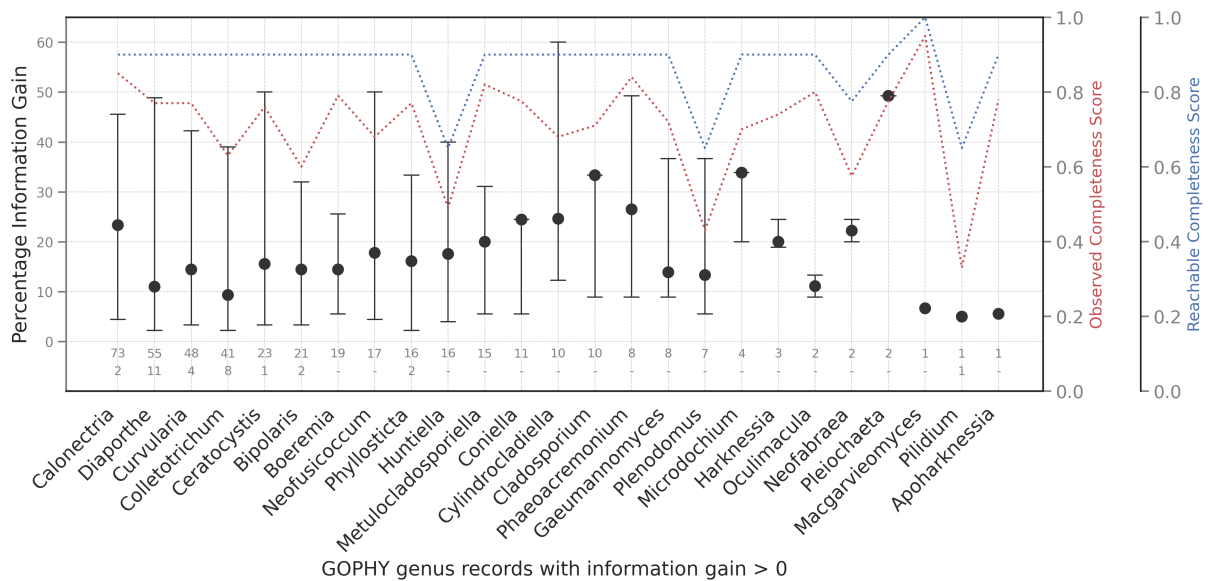


Figure 6 – Information gain by genus of phytopathogenic fungi registered in GOPHY database. Median with max/min values are presented in first Y-axis (left). Median values of Observed Completeness Scores and Reachable Completeness Scores are shown in 2nd and 3rd Y-axis (right), respectively. Only records with information gains greater zero were kept in chart. Numbers below zero in the X-axis indicates the number of records evaluated for each genus (upper number), and the number of record reaching the maximum reachable completeness (100%, lower number).

4.5 Conclusions

In this study, we showcase the remarkable ability of GeneConnector to substantially enhance the data completeness of specimens in Genbank by exclusively leveraging shared information within the records. Our findings demonstrate that utilizing our tool can yield gains of up to 60% in shared information among Genbank records, particularly for specific phytopathogenic genera. Furthermore, on a global scale, the data aggregation process holds the potential to benefit records from approximately 55 countries across the globe.

Moreover, our data aggregation process is both auditable and interpretable through two scores: the Observed Completeness Score (OCS) and the Reachable Completeness Score

(RCS). These scores provide insights into the current level of information completeness and the attainable information based on shared metadata among nodes of the same specimen in Genbank.

With these comprehensive metrics, our aim is to make a significant contribution to the ongoing improvement of the information accumulation process, benefiting scientists worldwide and fostering continuous advancements in knowledge acquisition.

Acknowledges

Thanks to the grant from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES). We also thank the anonymous reviewers who dedicated their valuable time to improving this work. Thanks to Biotrop, Solutions in Biological Technologies, for the support during the development of this work.

5 CLASSEQ: A clade-informed sequence identification tool

Table 5 – Classeq manuscript’s authors with affiliation.

Author	ORCID	Affiliation
Samuel Galvão Elias	0000-0001-9138-8845	University of Brasilia, Graduate Program in Microbial Biology, Institute of Biological Sciences, Brasilia, DF, Brazil and Bioinformatic Researcher in Biotrop, Solutions in Biological Technologies
Débora Guterres Cervieri	0000-0002-3902-8487	Post-doctoral Researcher in Federal University of Viçosa
Helson Mário Martins do Vale	0000-0002-5452-3873	Associate Professor at the University of Brasília, Department of Phytopathology, Institute of Biological Sciences, Brasilia, DF, Brazil

Abstract

Molecular phylogenetics is a powerful tool for the classification of biological sequences benefited by distinct knowledge fields as epidemiology, conservation biology, comparative biology, pharmacogenomics, forensics, and agriculture. Such an importance converged to a prominent new realm of classifiers, the phylogeny-based placers, as tools designed to place sequences into existing phylogenies, allong the large-scale classification of sequences with optimized performance when compared to traditional tools. Here, we present a novel clade-informed, alignment-free, and lightweight approach developed to classify gene-wide sequences. Our tool was tested to classify the *Bacillus subtilis* group RefSeq database, resulting in a highly sensitive and specific placer, capable of classifying almost all the sequences of the group into the correct clades. In addition to the traditional command line interface (CLI) available in bioinformatic tools, *classeq* provide an additional application programming interface allowing users to serve our classifier directly on the web, making the integration process of our algorithm with web-based tools easy.

Keywords: Phylogeny | *Bacillus subtilis* group | Phylogenetic placement | Alignment-free

5.1 Introduction

Molecular phylogenetics plays a crucial role in biology, examining the evolutionary connections between different biological species through their genetic traits. By scrutinizing genetic information such as DNA, RNA, and proteins, scientists can build phylogenetic trees or networks to illustrate these connections. Innumerable fields can benefit from phylogenetics, including the most common taxonomy classification (YANG; RANNALA, 2012), and have already been applied to epidemiology (HOLMES, 2004), conservation biology (ISAAC et al., 2007), comparative biology (HARVEY; PAGEL, 1991), pharmacogenomics (SQUASSINA et al., 2010), forensics (OGDEN, 2010), and agriculture [see (MORGANTE; SALAMINI, 2003) for genomic application and (MARIN-FELIX et al., 2017b; MARIN-FELIX et al., 2019d; MARIN-FELIX et al., 2019b; CHEN et al., 2022b) for examples of exploratory applications].

In this way, the phylogenetic placement of metagenomic sequences was an emerging field of bioinformatics just over a decade ago (CZECH et al., 2022). In contrast to the traditional approach used by metagenomics¹, phylogeny-based placement methods work around the comparison of query sequences² with internal or terminal nodes of previously existing trees (BRADY; SALZBERG, 2009; MATSEN; KODNER; ARMBRUST, 2010; BERGER; STAMATAKIS,

¹ Sequences of a given sample are identified by their direct comparison with reference sequences with further attribution of their taxonomy based on similarity (or not) with reference sequences (see (WOOD; SALZBERG, 2014; WOOD; LU; LANGMEAD, 2019; MENZEL; NG; KROGH, 2015; MAIDAK et al., 1997; MAIDAK et al., 2000) for example tools using such an approach and (AO et al., 2022; CALLAHAN et al., 2016; CAPORASO et al., 2010; BOLYEN et al., 2019; SCHLOSS et al., 2009) for frameworks using the tools cited above).

² Sequences derived from environmental samples or directly extracted from the target specimen.

2011; FIORAVANTI et al., 2018; BARBERA et al., 2019; LINARD; SWENSON; PARDI, 2019; BALABAN; SARMASHGHI; MIRARAB, 2020; BLANKE; MORGENSTERN, 2021; WASSAN; WANG; ZHENG, 2022; WASSAN; WANG; ZHENG, 2023). Unlike taxonomy-based placers, phylogenetic counterparts should provide highly accurate and rich in information results (BRADY; SALZBERG, 2009; SEGATA et al., 2012; TRUONG et al., 2015; JAMY et al., 2020; BEGHINI et al., 2021), allowing users to interpret the target diversity with additional complexity levels that include a most natural evolutionary past of the specimens (KEMBEL et al., 2011; SRIVASTAVA et al., 2012; CADOTTE, 2015).

Such an approach is highly advantageous for being 'almost' independent of the dynamicality of the taxonomic process, becoming immune to challenges intrinsic to taxonomic methods [see (EDGAR, 2018; LYDON; LIPP, 2018; MURALIDHARAN; FOX; POP, 2024) for examples of the database-derived errors impacting annotations, and the Taxallnomy (SAKAMOTO; ORTEGA, 2020) and SATIVA (KOZLOV et al., 2016) proposals to reduce the impact of such challenges].

It is important to note that the phylogenetic placement of metagenomic sequences is a different process from the traditional methods applied to infer the evolutionary history of a taxa. During a traditional placement workflow, the target sequence, or the query sequence (QS), is compared to all other reference sequences, and the full history of the QS is inferred using heuristic or exhaustive search [see (SWOFFORD, 1993; RONQUIST et al., 2012; STAMATAKIS, 2014)]. If multiple QS are included on the analysis, the relationships among this are also inferred.

Now, during the metagenome placement, the algorithms used for this purpose work to place QS into a reference tree (RT) by comparing sequences against the tree nodes independently from other QS. In other words, the relationship among the QS is not resolved. This process should be executed with or without solving multiple sequence alignments [see review of Czech et al. 2022]. This approach can significantly reduce the computation effort for high-dimensional data.

Phylogenetic placement methods commonly used for metagenomics adopt mostly two placement strategies, the maximum likelihood (ML) measurement, or the evolution distance calculation. Two of the most important tools (and the first developed for this purpose), PPLACER (MATSEN; KODNER; ARMBRUST, 2010) and RAxML-EPA (BERGER; STAMATAKIS, 2011) work around the ML framework. More recently, EPA-NG (BARBERA et al., 2019) was developed to combine the accuracy and performance of the two previous algorithms, which were already based on the ML framework. As an alternative to ML-based methods, a new phylo-k-mer-based method was introduced with the RAPPAS (LINARD; SWENSON; PARDI, 2019) algorithm and its successor EPIK (ROMASHCHENKO et al., 2023). All the above-cited software depends on the previous alignment of the query sequences to perform predictions, which leads to a significant increase in computational effort upstream the prediction phase and should be highly sensitive to high-divergent genome regions with no homologous positions as

noncoding portions of the DNA [as an example, the Internal Transcribed Spacer used as an important universal fungal barcoding (SCHUCH et al., 2012)].

Alongside the rise in the throughput of modern DNA sequencers, there has been a concurrent demand for placement methods that are not reliant on the multiple-sequence alignment techniques used in the preliminary stages of phylogenetic placement. In this way, new alignment-free methods were introduced with APPLES (BALABAN; SARMASHGHI; MIRARAB, 2020) and APP-SPAM (BLANKE; MORGENSTERN, 2021). Both perform distance-based placements calculated from the sequence k-mers, substantially increasing the placement performance when compared to the ML-based or phylo-k-mer-based alternative tools.

Still in APPLES, despite presenting itself as alignment-free (in the case of using the alignment-free option), the tool depends on a distance matrix as an input parameter, a solution that obviously suffers when non-equal size DNA sequences are compared. However, the solution APP-SPAM was developed primarily to work with short read sequences, being inadequate during the placement of sequences based on complete gene sequences.

Here, we present a novel clade-informed, alignment-free approach developed and tested to predict based on complete gene sequence trees. In the *first* section, we will present and discuss the *classeq* software structure and algorithm, introducing the data formats used during the indexation and prediction phases; during the *second* section we attempt to provide a proof of concept of our algorithm to classify sequences of *Bacillus subtilis* group RefSeq database.

5.2 Software Usage, Structure, and Algorithm

5.2.1 Usage Summary

Similarly to other classifiers the *classeq* pipeline flows through two stages: the initial *indexation* and the *prediction* (see Figure 7). *Indexation* is performed based on nucleotide sequences (as a FASTA format), the same used during the phylogeny reconstruction, and the phylogeny itself (as a NEWICK format). The follow *prediction* stage is performed based on *classeq* tarball artifact generated during the indexation phase and uses simple FASTA sequences as a query input. Both stages are described in the next sections.

5.2.2 Building the Phylogenetic Indices

As expected from a phylogeny-dependent classification tool, the main input of *classeq* during the indexation phase is a NEWICK-formatted phylogeny. Initially, the target phylogeny suffered re-rooting and sanitizing to remove low supported branches³. The re-root process is executed using the Environment for Tree Exploration, version 3 package [ETE3, (HUERTA-CEPAS;

³ Branches with low phylogenetic support. *Classeq* uses 95 as a default cutoff value, but this value can be changed through command line arguments at runtime.

(SERRA; BORK, 2016)]⁴. After the tree sanitization, branches with low phylogenetic support values are pruned and reconnected with the closely related parent node with sufficient support to be kept in phylogeny. In case the parent branch has no sufficient support, the search proceeds until the next supported branch is found (see Figures 7 A-B). The pruning process and all custom actions related to tree management are executed using extensions of the default Biopython (COCK et al., 2009) classes implemented as *classeq* elements.

During sequences sanitization, the nucleotide residuals are cleaned to remove ambiguous characters⁵ with the remaining ones being used to extract their k-mer contents. As default *classeq* take k-mers of size twelve, but this value should be changed during the index generation step (Figure 7C-D).

Next, k-mers serve as input to build the *k-mers inverse indices* (KII), a conceptual hash map containing k-mers as hash values, and an array containing the phylogeny leafs which the k-mers were mapped as values (Figure 7D). The KII comprises an intermediary object that speeds up the indexation phase. Next the k-mers content is mapped to the phylogeny internal nodes, originating the *Nodes to K-mers Index* (NKI in Figure 7E), a resulting object scheme is defined in Listing 5.1 (begin line 1), where the `nodePrior` type represents the phylogeny internal/external nodes. Each `nodePrior` carries information on all combinations of INGROUPE priors (a set of k-mers that match the target clade) and their SISTER clade priors (a set of k-mers shared between all clades not the INGROUPE). Both ingroup and sister priors are used during predictions (explained in 5.2.3 section).

At the end of the indexation process, the serialized KII plus the NKI object are available to users as a self-contained tarball artifact (Figure 7F) used during *predictions*. Such objects should be stored, retrieved, and shared among users to perform independent predictions.

The sanitized phylogeny (NEWICK) together with the sanitized sequences file (FASTA) persists as independent files, allowing users to audit the intermediary source results after indexation. Additionally, *classeq* produces an annotation artifact in the PHYLO.JSON format that contains the sanitized phylogeny. These artifacts are a web-friendly object and should be used during predictions by the CLI⁶ argument. The goal of the annotation artifact (as the name suggests) is to allow one to annotate individual nodes with an arbitrary name, their corresponding Genbank TaxID, and a related taxonomic rank. Annotations should be propagated up to the output artifact, increasing the user readability. The annotation artifacts are given in the schema of the Listing 5.1 (begin line 36).

```
1 ---
2 type: object # Priors artifact
3 properties :
```

⁴ We opted to specifically execute the re-root process with ETE3 due to the previously related inconsistency found in Biopython on execute tree managements (CZECH; HUERTA-CEPAS; STAMATAKIS, 2017).

⁵ Characters not in default DNA/RNA alphabet as A, T(U), C, and G.

⁶ Command Line Interface.

```
4   ingroup:
5     type: array
6     items:
7       $ref: '#/definitions/nodePrior'
8 definitions:
9   nodePrior:
10    type: object
11    properties:
12      parent:
13        type: string
14        format: uuid
15      clade_priors:
16        type: array
17        items:
18          $ref: '#/definitions/prior'
19 prior:
20   type: object
21   properties:
22     group:
23       type: string
24       enum: [INGROUP, SISTER]
25     kmers:
26       type: array
27       items:
28         type: string
29     labels:
30       type: array
31       items:
32         type: integer
33 ---
34 type: object # PHYLO.JSON Annotations artifact
35 properties:
36   id:
37     $ref: '#/definitions/nullString'
38   name:
39     $ref: '#/definitions/nullString'
40   rooted:
41     type: boolean
42   root:
43     $ref: '#/definitions/clade'
44 definitions:
45   nullString:
46     type: [string, null]
47   nullNumber:
48     type: [number, null]
49     minimum: 0
50   nullInteger:
```

```

51   type: [integer , null]
52   minimum: 1
53   clade:
54     type: object
55     properties:
56       id:
57         type: string
58         format: uuid
59       name:
60         $ref: '#/definitions/nullString'
61       confidence:
62         $ref: '#/definitions/nullNumber'
63       branch_length:
64         $ref: '#/definitions/nullNumber'
65       color:
66         $ref: '#/definitions/nullString'
67       width:
68         $ref: '#/definitions/nullNumber'
69       taxid:
70         $ref: '#/definitions/nullInteger'
71       related_rank:
72         $ref: '#/definitions/nullString'
73       clades:
74         $ref: '#/definitions/cladeArray'
75   cladeArray:
76     type: array
77     items:
78       $ref: '#/definitions/clade'

```

Listing 5.1 – Data schema of Priors and Annotation artifacts. Some double quotes were omitted to turn the reading better. The code snippets are present as a standard YAML format, but when needed users should convert to JSON schema using default web formatters from the original schema definitions files available on project Github documentation.

For basic/non-developer users, we provide a GUI⁷ interface available on execution of the CLI command *serve* (see code snippet in Listing 5.2), allowing users to annotate the internal nodes of the phylogeny without dealing with JSON specifications. For advanced/developer users, the annotation artifact can be easily parsed and edited using JavaScript interpreters like browser applications, or even manually edited if necessary, as long as they follow the standard schema of the Listing 5.1 (begin line 36).

Certain attributes within the schema require further consideration. Notably, the clade Identifier (ID, line 39) is a Universal Unique Identifier of version 3, which bears a nonstandard namespace derived from a nine sequence (e.g., 99999999-9999-9999-9999-999999999999),

⁷ Graphical User Interface.

which concatenates the clade name to the names of all its terminal children clades. Our approach permits multiple runs to converge in the same UUID's since their use the same clades configuration, facilitating future sharing and comparison of independent run annotations.

```
$ cls serve -t *.phylo.json
```

Listing 5.2 – Classeq Annotation Server startup example command.

5.2.3 Prediction

The prediction workflow is shown in Figure 8. The final users can make predictions using the CLI command *predict* with appropriate arguments or through the API⁸ port, which is readily available, allowing them to deploy their own classeq server on either local or public networks, thus facilitating predictions. The API option is not fully recommended for users with high performance interest because it is not already optimized for the web purpose. For developers, the prediction functions are available as *classeq* module named *predict-taxonomies-recursively* (accessible through a [use-case](#)⁹ with the same name).

As explained in Figure 7 (prediction section), after initial conversion of the query sequences into k-mers (Figure 7G and Figure 8), our algorithm starts the recursive query for the candidate node to place a given query sequence following the tree root to the leaf direction (Figure 7H).

Starting from the tree root node¹⁰ our algorithm tests the adherence of the given query sequence to the target node (including the tree root) using simple k-mers matching statistics, where the number of matches between the query sequence/target clade is compared to the matches with all sibling ones together. An strategy also known as *One-vs-Rest* (OvR, Figure 7I, and Figure 8B). At this phase, when a desired node is detected as a candidate, all siblings are discarded, and the search continues to the children's nodes, increasing the search depth. In case the current node includes children ones, AND the OvR comparison is unable to identify a candidate node, the search process stops with a young return including the path crossed along the phylogeny up to the current node plus the status code `MAX_RESOLUTION_REACHED`. Otherwise, in case the current node has just one candidate node and has no children, the status code should be `CONCLUSIVE_INGROUP` instead. If more than one candidate was found, the current node should be ignored and the returned status should be `INCONCLUSIVE`.

Exceptionally, in case during the first step of the search algorithm (depth 1) the k-mer match coverage between the query and the candidate node is lower than 50%¹¹, the search

⁸ Application Programming Interface.

⁹ A use-case represents the piece of intention of the software, often represented as a module, which users and developers can easily understand their particular goal within the software's tiers and layers.

¹⁰ As previously explained, the input phylogeny is re-rooted at middle-point as default and the root node serves as the entry-point for the classification flow.

¹¹ The match coverage is controlled with the argument `-matches-coverage` with default as 0.5 but ranging from 0.00001 to 1.

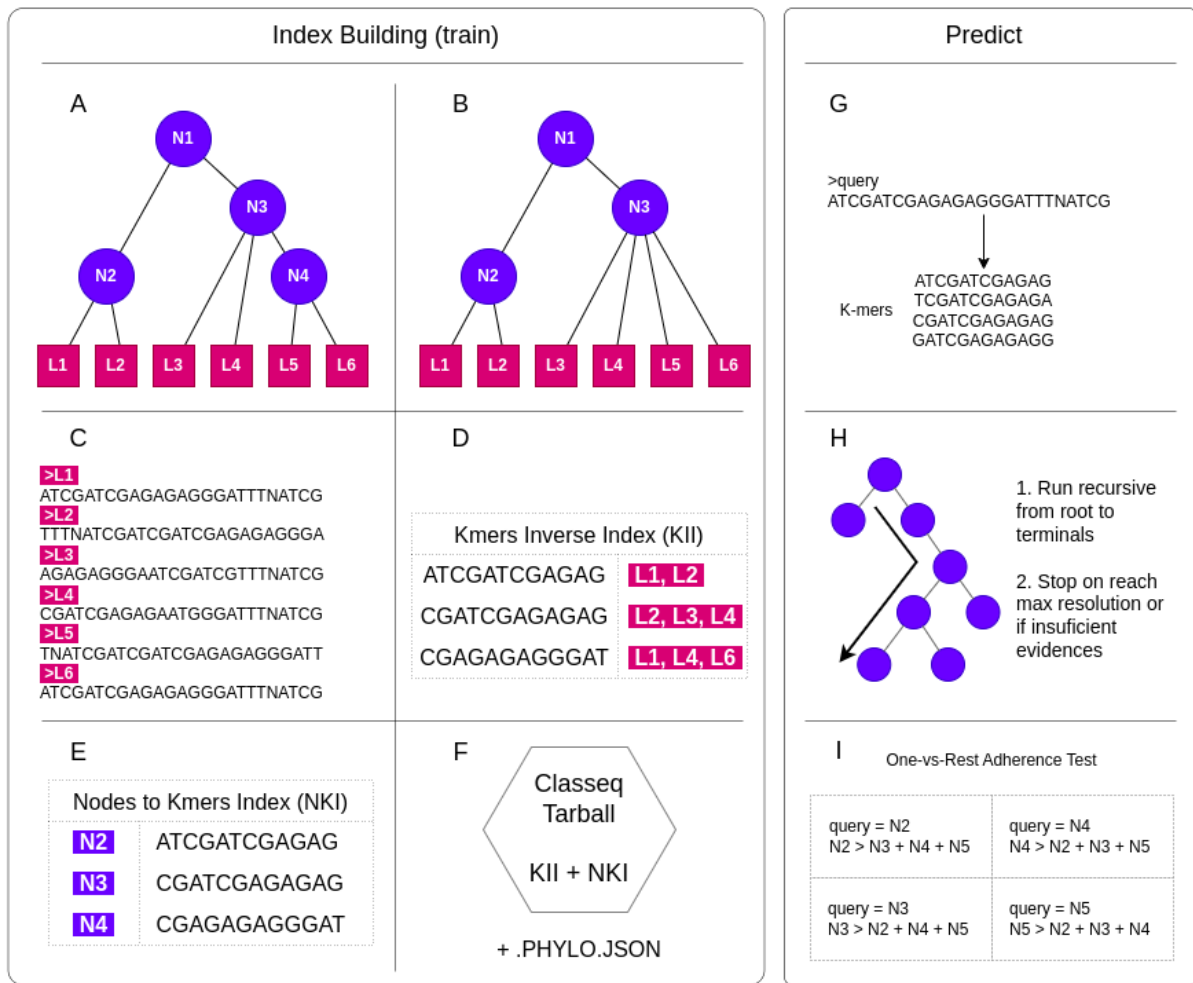


Figure 7 – The overall workflow of the *Classeq* life cycle is depicted. The Index Building column encompasses the train-like steps executed during the creation of the training tarball. Frame A illustrates a raw phylogeny provided to *Classeq* at runtime; frame B indicates the sanitization process, where node 4 (N4) is collapsed, and child nodes are connected to the parent node N3. Frame C contains the raw sequences (L1-L6), while frame D presents the basic structure of the k-mer inverse indices (KII) generated from these sequences. Frame E exemplifies the node-to-k-mer indices (NKI) mapping leaf k-mers to internal phylogeny nodes. Finally, frame F contains the final deliverables, including the tarball and the `phylo.json` object. The Predict column details the prediction process, with frame G showing the initial k-mer splitting of the query sequence. Frame H indicates the tree traversal performed during predictions, and frame I represents the one-vs-rest strategy executed on each tested phylogeny node.

suffering from young return with INCONCLUSIVE status. This rule prevents our algorithm from producing false positive predictions, thus increasing the specificity of each trained model.

Our choice for the OvR strategy already mirrors the conservative nature of the *classeq* search algorithm, which prevents greedy behaviors. The use of the OvR strategy means that a node needs not to be "only" the best node in the clade to become a candidate, but the best node

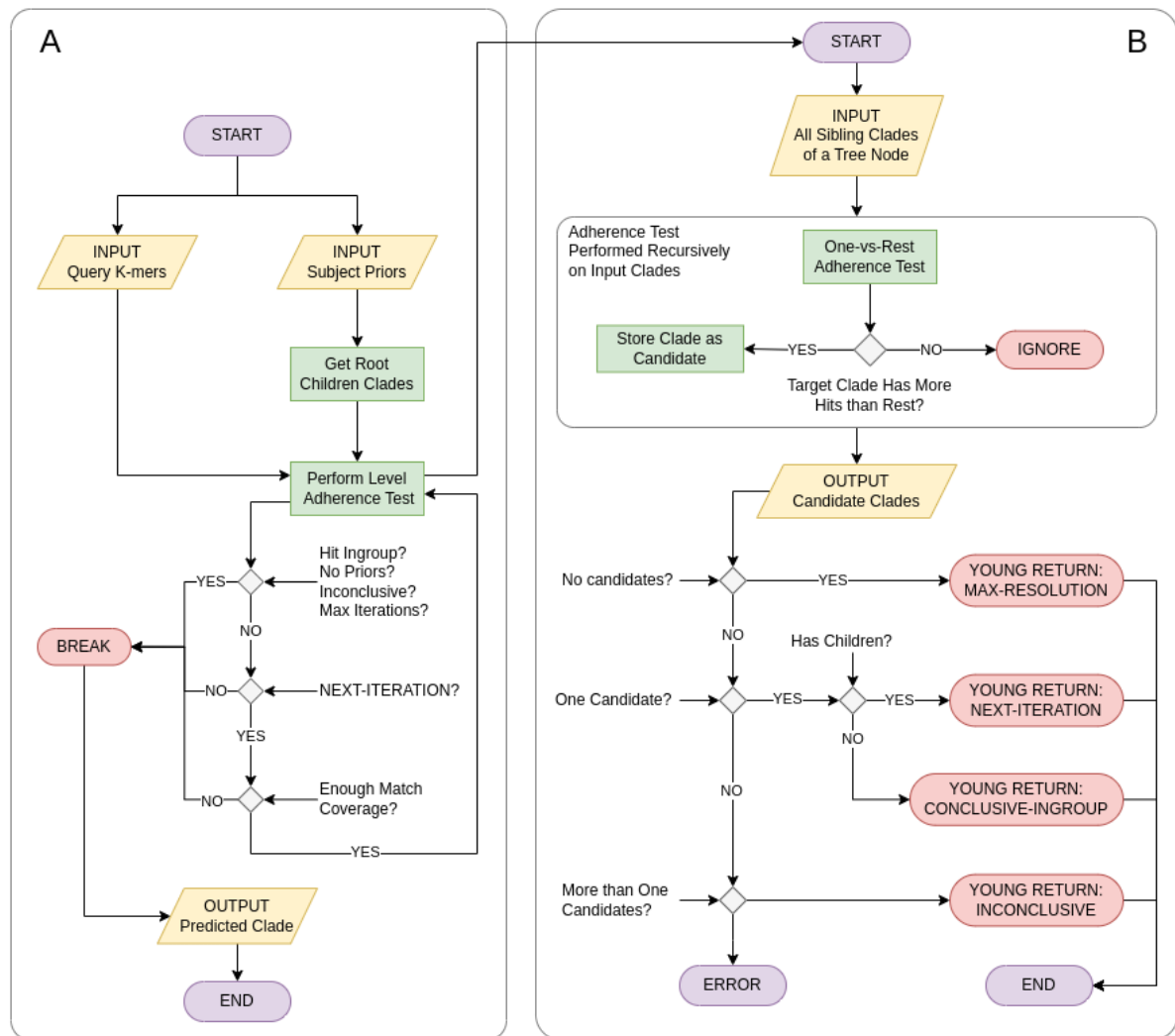


Figure 8 – Details of the prediction algorithm. Column A outlines the general prediction flow, beginning with the input of the query sequence and subject priors generated during index building. The first execution step starts from the root node, where child nodes are used for predictions, undergoing one-vs-rest (OvR) comparison between every node and its siblings, as shown in column B. Yellow boxes indicate input/output elements of the main algorithm use cases. Green boxes represent important procedures within the algorithm. Red boxes contain the algorithm statuses delivered to end-users. Gray diamonds indicate the main logic decisions made during the prediction workflow, while lilac boxes indicate the start, end, or raise of the process.

when compared to the sum of all the others. Conservative behavior is a convenient way to deal with incomplete sampled clades often verified in poorly explored taxonomic groups. This means that *classeq* favors the nonclassification of a desired sequence up to the lowest level as possible over their misplacement into a "wrong" clade.

5.3 A proof-of-concept: Classifying specimens of *Bacillus subtilis* group using the *gyrB* gene

To evaluate the performance of *classeq* on place real world sequences, we carried out a classification assay in an attempt to place specimens from the *Bacillus subtilis* group (WANG et al., 2007; ROONEY et al., 2009a; BHANDARI et al., 2013) in a comprehensive phylogeny of the group constructed from the *gyrB* gene (Figure 10A). The *gyrB*, which encodes the subunit B protein of the DNA gyrase gene, represents an alternative marker for the phylogenetic placement of the *B. subtilis* group specimens, providing mid/high resolution when compared with the 16S rRNA gene (WANG et al., 2007).

In performing the classification of specimens exclusively belonging to the *B. subtilis* group (from here called as Ingroup), we would only be able to test the conceptual *sensitivity* on the *classeq* results. Thus, aiming to be more precise on the test of the *classeq specificity*, we carried out a systematic sampling for the outgroup¹² specimens guided by the phylogenetic distance between the Ingroup and the independent outgroups. Our sampling considered three levels of divergence, including specimens from the closed related *Bacillus cereus* group (sharing the *Bacillus* [genus] as LCA¹³), followed by specimens of *Paenibacillus* (Bacilli [class] as LCA), and *Streptomyces* (Terrabacteria [clade] as LCA) as the most divergent group (see Figures 9 and 10B).

Finally, the confusion matrix of our assay was consolidated in Figure 10C, where both Ingroup and Outgroups classification metrics were used to calculate the *sensitivity* and *specificity* statistics (see (TREVETHAN, 2017) for a conceptual review).

5.3.1 RefSeq data acquisition

To guarantee the accuracy of the specimens identities included in our assay, both the train and the test were performed on the basis of sequences from the Genbank RefSeq database. Initially, accessions of the complete RefSeq database of *B. subtilis* group, *B. cereus* group, *Paenibacillus* spp. and *Streptomyces* spp. were retrieved (through the Genbank web interface) and sampled (using the `shuf` Linux functionality) to balance the input dataset. The samples were then downloaded and sanitized using custom Python scripts. The *B. subtilis* Group complete dataset (containing 2139 genome sequences) was sampled with $\approx 50\%$ of records (resulting in 1040 genomes) and used as a training dataset for the phylogeny reconstruction (see phylogeny reconstructoin section), where the remaining (1097 genomes) were used for testing purposes. Datasets of the remaining genus were sampled with a fixed size of 1000 records when available (see Figure 9).

¹² From here all specimens not belonging to the *B. subtilis* group will be refereed as outgroups.

¹³ The Last Common Ancestor shared between studied taxa.

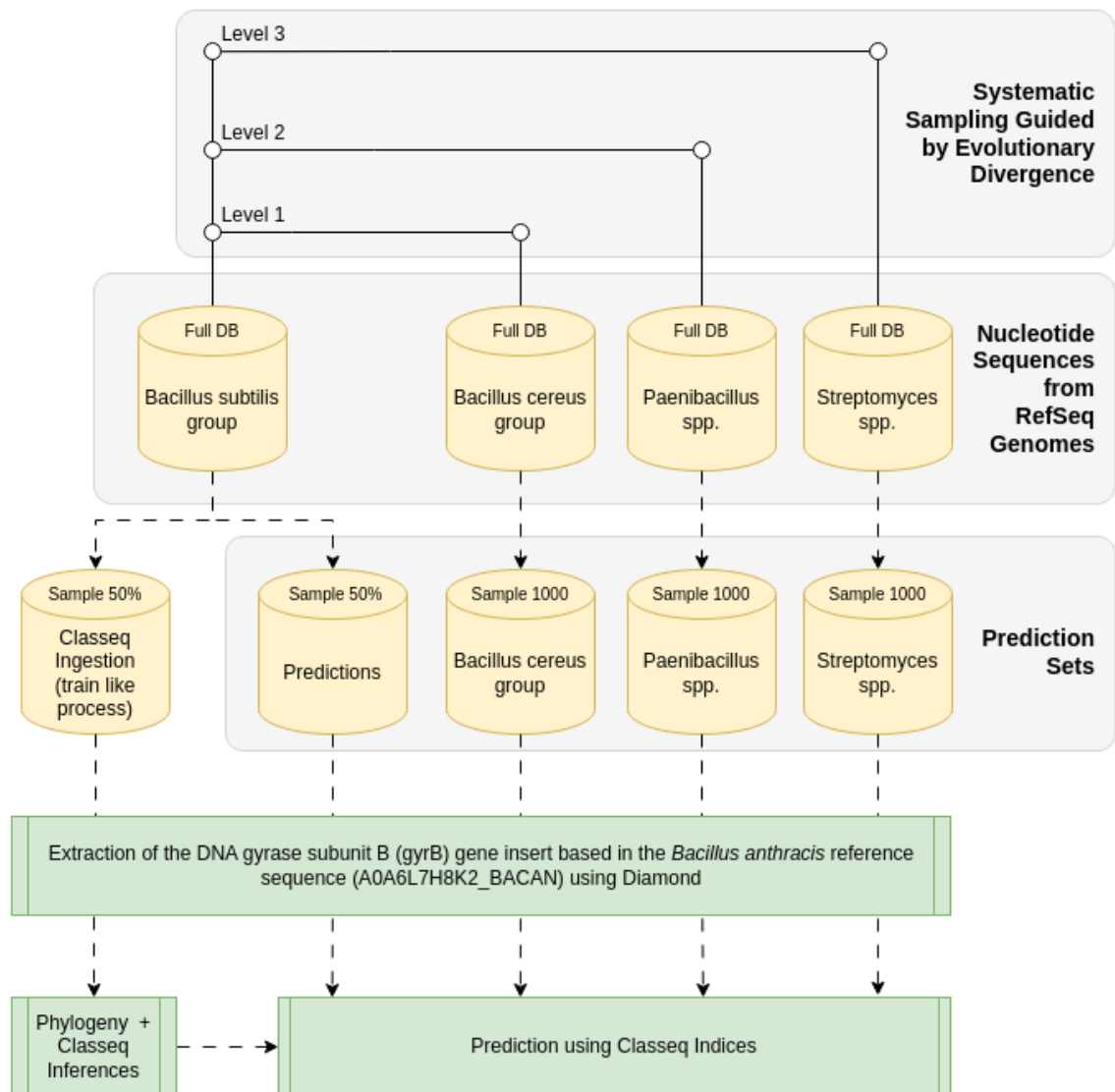


Figure 9 – RefSeq database acquisition performed during the `Classeq` validation assay.

Following the sampling steps, the `gyrB` gene was extracted from the reference genomes using the DIAMOND algorithm (BUCHFINK; XIE; HUSON, 2015) following default settings. We used `A0A6L7H8K2_BACAN` as a query sequence for gene extraction in the studied genomes. Genomes in which the reference gene did not match were discarded, resulting in a test dataset of 1000, 745, and 927 specimens of *B. cereus* group, *Paenibacillus* spp., and *Streptomyces* spp., respectively.

5.3.2 Reference phylogeny reconstruction

The sequences of the *B. subtilis* group sampled in the previous step were clustered at 99.9% of similarity using the VSEARCH cluster_fast strategy (ROGNES et al., 2016) reducing the number of records used during phylogenetic reconstruction. The clustering strategy reduces the noise of non-informative/repetitive sequences and significantly reduces the search space for the

tree convergence. The final result is a lean and highly informative tree. The clustered dataset was composed of 293 records with sequences ranging from 1905 to 1911 bp (exception to record NZ_JALAQF010000173 [*Bacillus atrophaeus* strain CK3J3B4] reaching 1708 bp in length).

The resulting database was aligned using the Mafft algorithm (KATOHI; STANDLEY, 2013) with default parameters and the strategy `ginsi`, resulting in an MSA¹⁴ containing 752 variables (39.5%) with 722 informative sites (37.9%). The base composition of MSA was 31.8% of adenine, 21.2% of cytosine, 24.8% of guanine, and 22.3% of thymine, with an average transition/transversion ratio of 1.4 (ranging from 0 to 25% between all sequence pairs).

Finally, the MSA was used into the phylogeny reconstruction using the software RAxML-HPC v8 (STAMATAKIS, 2014), under the GTRCAT substitution model, using the Rapid Bootstrap Algorithm (STAMATAKIS; HOOVER; ROUGEMONT, 2008) with 1000 pseudo-replications. The parsimony seed value was 12345. The default settings were kept for the remaining parameters.

The resulting tree is shown in Figure 10A where all the *B. subtilis* group were successfully recovered, including: *B. sonorensis* (PALMISANO et al., 2001) (represented here by 2 specimens), *B. licheniformis* (WEIGMANN, 1898) (6 spp.), *B. paralicheniformis* (DUNLAP et al., 2015) (14 spp.), *B. atrophaeus* (NAKAMURA, 1989) (25 spp.), *B. amyloliquefaciens* group (FAN et al., 2017) including *B. siamensis* (SUMPAPAVAPOL et al., 2010) (7 spp.), *B. amyloliquefaciens* (PRIEST et al., 1987) with two clades which includes *B. velezensis* (RUIZ-GARCIA et al., 2005) (109 spp. as the total). Additionally, *B. halotolerans* (TINDALL, 2017) and *B. mojavensis* (ROBERTS; NAKAMURA; COHAN, 1994) within the *B. mojavensis* group (SCHOCH et al., 2020) (total 20 spp.), *B. tequilensis* (GATSON et al., 2006) (2 spp.), *B. spizizenii* (DUNLAP; BOWMAN; ZEIGLER, 2020) (4 spp.), *B. vallismortis* (ROBERTS; NAKAMURA; COHAN, 1996) (6 spp.), *B. stercoris* (DUNLAP; BOWMAN; ZEIGLER, 2020) (4 spp.), *B. inaquosorum* (ROONEY et al., 2009b; DUNLAP; BOWMAN; ZEIGLER, 2020) (23 spp.), and *B. subtilis stricto sensu* (VBD, 1980) (69 spp.).

5.3.3 Classeq predictions

With the phylogenetic reconstruction build on the previous step plus their MSA, we constructed the *classeq* indices (see section about phylogeny indexing) with default parameters for the k-mer size (`-kmer-size 12`) and strand (`-strad both`, taking into account k-mers generated with forward and reverse strands), and a non-default value for the nodes support cutoff parameter (`-support-value-cutoff 80`, default 95).

Next, independent predictions were performed on the target group and outgroups following the default parameters of the *classeq predict* command. A manual annotated phylogeny (such as PHYLO.JSON format) was used as a reference for the clades identity. Prediction times were

¹⁴ Multiple Sequence Alignment.

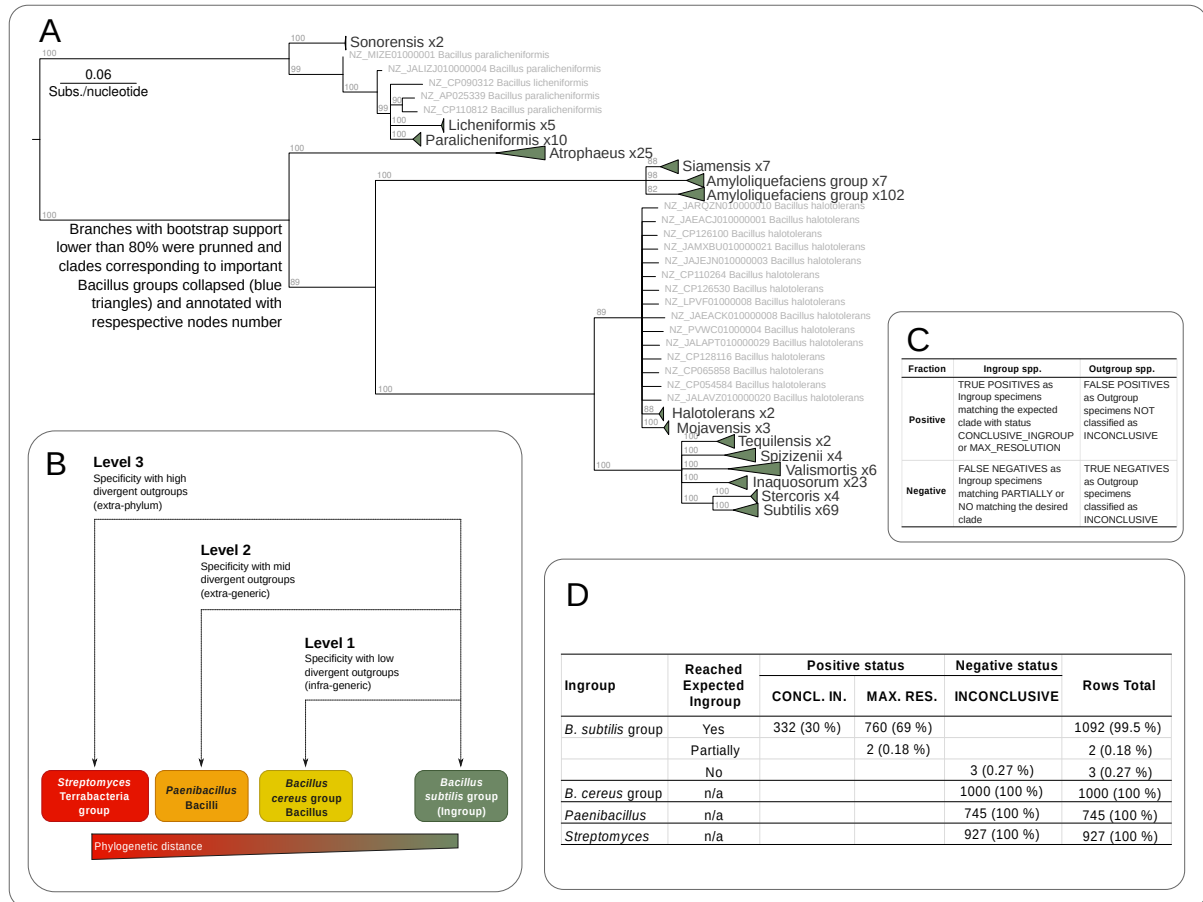


Figure 10 – The *Classeq* prediction for the main *Bacillus subtilis* group clades. The figure illustrates the components of the *Classeq* evaluation process. Section A presents the reference phylogeny used during predictions, encompassing all major clades of the phylogenetic group. Section B details the sampling strategy adopted to assess *Classeq*'s sensitivity and specificity, including specimens sampled both within and outside the *B. subtilis* group. Section C displays the confusion matrix consolidation used to guide data collection during algorithm evaluation, while section D presents the *Classeq* prediction results.

≈ 16 seconds for *B. subtilis* group ($n = 1097$), ≈ 4 s *B. cereus* group ($n = 1000$), ≈ 3 s *Paenibacillus* spp. ($n = 745$), and ≈ 4 s *Streptomyces* ($n = 927$). The complete assay was performed on a Linux-based workstation equipped with a 12th Gen Intel® Core™ i5-12500H×16, and 16,0 GiB of RAM memory. All files related to model training, annotation, and prediction are available as [Support Files 1-4](#) in project [Github Repository](#).

To provide a reference time for the *classeq* results evaluation, we performed a Blast search with default parameters to verify the identity of *B. subtilis* group using the clustered MSA (as the same of *classeq* input, see phylogeny reconstruction section for details) as a subject. The execution time of the Blast search was ≈ 18 s ($n = 1097$). The resulting identities were not shown due to the information incompatibility between the Blast and *Classeq* searches. We limit our discussion to inform readers that the Blast search matched among their top 50 results (when

available) samples already present in the clades identified by *classeq*, being fully compatible in terms of the final results.

Furthermore, to determine the compatibility of the *classeq* results with a traditional method for phylogenetic reconstruction, we generated a comparative result by placing the test sequences into the same phylogeny of the training set, using a Maximum Likelihood search. To perform this task, the sequences database of the test set was split into files with up to 220 sequences (1097 records of the training set resulting in four alignments of 220 sequences plus one of 217 sequences) being each chunk re-aligned against the original train dataset. The techniques for MSA calculation and phylogeny reconstruction were the same used to build the train dataset, see details in phylogeny reconstruction section.

Surprisingly, the sequences of *Bacillus subtilis* (strain DE0224, NZ_VTQV01000062) and *Bacillus amyloliquefaciens* (strain SN781, NZ_JAGFMA01000006) were highly divergent from other MSA sequences and were discarded. Both records are currently **suppressed** from the RefSeq database. Furthermore, the sequence of *Bacillus sonorensis* (strain MarseilleP3463, NZ_LT745775), was highly divergent from other *B. sonorensis* specimens, but it was kept in MSA because it was not considered enough divergent from another sequences from *B. subtilis* group, being a possible previous misplacement.

5.3.4 Results and Discussions

Our predictions with *classeq* on *B. subtilis* group specimens reached high values of *sensitivity* (>99.54%, see resolution of Equation 5.1) and *specificity* (100%, see Equation 5.2), where 1092 specimens from the 1097 *B. subtilis* group original specimens were correctly placed (see Figure 10D) on the expected clades, and all records of the outgroup correctly classified as INCONCLUSIVE¹⁵. The number of partially or not classified specimens summed was less than 1% (see Figure 10D for details).

$$TP/(TP + FN) = 1092/(1092 + 2 + 3) = 0.9954 \quad (5.1)$$

$$TN/(FP + TN) = 2672/(0 + 2672) = 1.0 \quad (5.2)$$

In the above equations, TP = True Positive results, where FN = False Negative, TN = True Negative, and FP, False Positives. Seeing the consolidation of Figure 10C, the TP values include all correctly classified *B. subtilis* group specimens, where FN includes specimens partially or not classified; TN includes the count of the Outgroup specimens truly classified (1000 + 745 + 927), where FP the number of Outgroup specimens classified as *B. subtilis* group.

All placements provided by *classeq* algorithm over the query sequences were compatible with identities recovered by our validation using the traditional method of maximum likelihood

¹⁵ The default prediction status code when a sequence does not belongs to the training phylogeny.

for reconstruction of the phylogeny. The proof phylogenies together with all intermediary files should be found as [Support File 6](#) on the project [Github repository](#).

Our assay results show that in addition to the high values of *sensitivity* and *specificity* of the *classeq* predictions — up to the limit of our investigation — our algorithm suffers a poor or null influence of the clade size during predictions. Clades containing the minimum number of specimens to be considered a conceptual "clade" [has at least two terminals, see ([QUEIROZ, 2013](#))] had the same prediction performance of highly dense ones, e.g. *Subtilis* and *Tequilensis* subclades, containing 69 and 2 specimens, respectively, and both with significantly lower error rates.

Conservative behavior represents an important aspect of the *classeq* algorithm. Unlike other algorithms that perform predictions based on a sparse *one-hot* representation of biological sequences using k-mers [e.g. ([DESAI et al., 2020a](#); [DESAI et al., 2020b](#); [FIANNACA et al., 2018](#))], our method is based on dense continuous vectors (DCV) as input that abstracts the sequence composition. This is the same strategy used by important classifiers as the RDP classifier ([MAIDAK et al., 1997](#); [MAIDAK et al., 2000](#)), Kraken (1 and 2) ([WOOD](#); [SALZBERG, 2014](#); [WOOD](#); [LU](#); [LANGMEAD, 2019](#)), and Kaiju ([MENZEL](#); [NG](#); [KROGH, 2015](#)).

The application of DCV representation on the classification of DNA sequences in opposition to *one-hot* encoding was investigated by Lo Bosco and Di Gangi in ([BOSCO](#); [GANGLI, 2017](#)), where it was concluded that the use of DCV representation generates the most fine-grained rank performance in the classification of biological sequences through long–short–term memory (LSTM) and convolutional neural network (CNN) models.

5.4 Conclusions

Here we present a highly *sensitive* and *specific*, alignment-free, and lightweight tool for the phylogenetic placement of sequences. We see the simplicity and clarity of the *classeq* solution as an advantage for future improvements in terms of performance and customization.

Unlike competing solutions, *classeq* is the first tool that provides a native server to perform classification through an API port. Such a feature provides a rapid way integrate our tool with web native systems, pipelines attached to the web commonly existing into departmental wide solutions.

5.5 Future remarks

Despite *classeq* algorithm representing a prominent solution to deal with gene-wide trees for sequences classification, our tool still represents an unoptimized prototype, with all steps of the indexing and prediction workflows executed in Python Virtual Machine ([IKE-NWOSU, 2015](#)).

This means that the obvious first way to achieve the performance of other classifiers is to translate the key point of our algorithm into a low-level programming language.

In the following, in order to provide a complete and informative report on the *classeq* functionality with optimizations, our tools should be directly compared with competing tools such as APPLES and APP-APAM to benchmark our algorithm against a global scenario of phylogenetic placers.

Acknowledges

Thanks to the grant from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES). We also thank the anonymous reviewers who dedicated their valuable time to improving this work. Thanks to Biotrop, Solutions in Biological Technologies, for the support during the development of this work.

6 Considerações Finais

A presente tese explorou dois desafios cruciais na análise de dados genômicos: a agregação e complementação de metadados e a classificação filogenética de sequências biológicas. Para abordar o primeiro desafio, desenvolvemos o `GeneConnector`, uma ferramenta que agrega e complementa metadados de registros do GenBank, explorando informações compartilhadas entre diferentes sequências de um mesmo espécime. A aplicação do `GeneConnector` ao banco de dados GOPHY demonstrou sua eficácia na recuperação de informações valiosas sobre a origem, coleta e processamento das amostras, com ganhos de informação de até 60%. Além disso, a introdução dos scores OCS e RCS oferece métricas para avaliar a completude dos metadados e o potencial de enriquecimento de informações, contribuindo para a melhoria contínua do processo de acumulação de dados e para a pesquisa científica em geral.

Para o segundo desafio, desenvolvemos o `Classeq`, uma ferramenta de classificação de sequências biológicas baseada em posicionamento filogenético, que se destaca por ser rápida, precisa, independente de alinhamentos múltiplos de sequências e capaz de classificar sequências de genes inteiros. Nossos testes com o grupo *Bacillus subtilis* demonstraram a alta sensibilidade e especificidade da ferramenta, classificando corretamente quase todas as sequências do grupo em seus respectivos clados. Além disso, a interface de usuário amigável e a disponibilidade de uma API facilitam a integração do `Classeq` em fluxos de trabalho existentes, tornando-o uma ferramenta valiosa para a comunidade científica.

Em suma, as ferramentas desenvolvidas nesta tese, `GeneConnector` e `Classeq`, representam avanços significativos na análise de dados genômicos, com potencial para impulsionar pesquisas em diversas áreas, desde a medicina e biotecnologia até a agricultura e conservação da biodiversidade. Ao abordar os desafios de agregação de metadados e classificação filogenética, essas ferramentas oferecem novas perspectivas para a interpretação e utilização de dados genômicos, abrindo caminho para descobertas e aplicações inovadoras.

Para trabalhos futuros, vislumbramos o aprimoramento contínuo dessas ferramentas, incluindo a expansão do `GeneConnector` para outros bancos de dados genômicos e a otimização do `Classeq` para lidar com árvores filogenéticas maiores e mais complexas. Além disso, a integração dessas ferramentas em plataformas de análise metagenômica, como o `bioBakery 3`, permitirá uma caracterização mais completa e precisa de comunidades microbianas, desde a identificação taxonômica até a análise funcional e reconstrução de linhagens, aprofundando nossa compreensão da ecologia e do papel dos microrganismos em diversos ambientes.

Referências

ABARENKOV, K. et al. The unite database for molecular identification of fungi—recent updates and future perspectives. *The New Phytologist*, JSTOR, v. 186, n. 2, p. 281–285, 2010. Citado 2 vezes nas páginas 40 e 51.

ABARENKOV, K. et al. The unite database for molecular identification and taxonomic communication of fungi and other eukaryotes: sequences, taxa and classifications reconsidered. *Nucleic Acids Research*, Oxford University Press, v. 52, n. D1, p. D791–D797, 2024. Citado na página 40.

ALTSCHUL, S. F. et al. Basic local alignment search tool. *Journal of molecular biology*, Elsevier, v. 215, n. 3, p. 403–410, 1990. Citado 3 vezes nas páginas 17, 39 e 42.

AO, C. et al. Biological sequence classification: A review on data and general methods. *Research*, v. 2022, 1 2022. ISSN 2639-5274. Citado 2 vezes nas páginas 41 e 63.

ARITA, M.; KARSCH-MIZRACHI, I.; COCHRANE, G. The international nucleotide sequence database collaboration. *Nucleic Acids Research*, Oxford University Press, v. 49, n. D1, p. D121–D124, 2021. Citado na página 27.

BALABAN, M.; SARMASHGHI, S.; MIRARAB, S. Apples: scalable distance-based phylogenetic placement with or without alignments. *Systematic Biology*, Oxford University Press, v. 69, n. 3, p. 566–578, 2020. Citado 4 vezes nas páginas 47, 63, 64 e 65.

BARBERA, P. et al. Epa-ng: massively parallel evolutionary placement of genetic sequences. *Systematic biology*, Oxford University Press, v. 68, n. 2, p. 365–369, 2019. Citado 6 vezes nas páginas 18, 44, 45, 46, 63 e 64.

BARTON, G. S. Directory interchange format: a metadata tool for the noaa earth system data directory. *The role of metadata in managing large environmental science datasets. Pacific Northwest Laboratory, Richland, Washington, USA*, p. 19–23, 1995. Citado na página 24.

BEGHINI, F. et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. *elife*, eLife Sciences Publications, Ltd, v. 10, p. e65088, 2021. Citado 2 vezes nas páginas 48 e 64.

BENSON, D. A. et al. Genbank. *Nucleic acids research*, Oxford University Press, v. 41, n. D1, p. D36–D42, 2012. Citado 3 vezes nas páginas 17, 27 e 50.

BERGER, S. A.; STAMATAKIS, A. Aligning short reads to reference alignments and trees. *Bioinformatics*, Oxford University Press, v. 27, n. 15, p. 2068–2075, 2011. Citado 3 vezes nas páginas 46, 63 e 64.

BHANDARI, V. et al. Molecular signatures for bacillus species: demarcation of the bacillus subtilis and bacillus cereus clades in molecular terms and proposal to limit the placement of new species into the genus bacillus. *International Journal of Systematic and Evolutionary Microbiology*, Microbiology Society, v. 63, n. Pt_7, p. 2712–2726, 2013. Citado na página 72.

- BLANKE, M.; MORGENSTERN, B. App-spam: phylogenetic placement of short reads without sequence alignment. *Bioinformatics Advances*, Oxford University Press, v. 1, n. 1, p. vbab027, 2021. Citado 3 vezes nas páginas 63, 64 e 65.
- BOLYEN, E. et al. Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature biotechnology*, Nature Publishing Group, v. 37, n. 8, p. 852–857, 2019. Citado na página 63.
- BOSCO, G. L.; GANGI, M. A. D. Deep learning architectures for dna sequence classification. In: SPRINGER. *Fuzzy Logic and Soft Computing Applications: 11th International Workshop, WILF 2016, Naples, Italy, December 19–21, 2016, Revised Selected Papers 11*. [S.l.], 2017. p. 162–171. Citado na página 77.
- BOUCKAERT, R. et al. Beast 2: a software platform for bayesian evolutionary analysis. *PLoS computational biology*, Public Library of Science San Francisco, USA, v. 10, n. 4, p. e1003537, 2014. Citado na página 45.
- BOUCKAERT, R. et al. Beast 2.5: An advanced software platform for bayesian evolutionary analysis. *PLoS computational biology*, Public Library of Science San Francisco, CA USA, v. 15, n. 4, p. e1006650, 2019. Citado na página 45.
- BRADY, A.; SALZBERG, S. L. Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nature methods*, Nature Publishing Group US New York, v. 6, n. 9, p. 673–676, 2009. Citado 2 vezes nas páginas 63 e 64.
- BUCHFINK, B.; XIE, C.; HUSON, D. H. Fast and sensitive protein alignment using diamond. *Nature methods*, Nature Publishing Group US New York, v. 12, n. 1, p. 59–60, 2015. Citado na página 73.
- CADOTTE, M. W. Phylogenetic diversity and productivity. *Functional Ecology*, JSTOR, v. 29, n. 12, p. 1603–1606, 2015. Citado na página 64.
- CALLAHAN, B. J. et al. Dada2: High-resolution sample inference from illumina amplicon data. *Nature methods*, Nature Publishing Group US New York, v. 13, n. 7, p. 581–583, 2016. Citado na página 63.
- CANAKOGLU, A. et al. Genosurf: metadata driven semantic search system for integrated genomic datasets. *Database*, Oxford Academic, v. 2019, 2019. Citado na página 51.
- CANO, M. A. et al. Schema playground: A tool for authoring, extending, and using metadata schemas to improve fairness of biomedical data. *BMC bioinformatics*, Springer, v. 24, n. 1, p. 159, 2023. Citado na página 26.
- CAPORASO, J. G. et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, Nature Publishing Group, v. 7, n. 5, p. 335–336, 2010. Citado na página 63.
- CHEN, Q. et al. Genera of phytopathogenic fungi: Gophy 4. *Studies in Mycology*, Westerdijk Fungal Biodiversity Institute, v. 101, n. 1, p. 417–564, 2022. Citado na página 52.
- CHEN, Q. et al. Genera of phytopathogenic fungi: Gophy 4. *Studies in Mycology*, Westerdijk Fungal Biodiversity Institute, v. 101, p. 417, 2022. Citado na página 63.

- CHEN, Q.; ZOBEL, J.; VERSPOOR, K. Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study. *Database*, Oxford Academic, v. 2017, 2017. Citado 2 vezes nas páginas 17 e 51.
- CHEN, Z. et al. Global landscape of sars-cov-2 genomic surveillance and data sharing. *Nature genetics*, Nature Publishing Group US New York, v. 54, n. 4, p. 499–507, 2022. Citado na página 51.
- CHOUDHARY, S. pysradb: A python package to query next-generation sequencing metadata and data from ncbi sequence read archive. *F1000Research*, Faculty of 1000 Ltd, v. 8, 2019. Citado na página 35.
- CLARE, E. L. et al. Measuring biodiversity from dna in the air. *Current Biology*, Elsevier, v. 32, n. 3, p. 693–700, 2022. Citado na página 37.
- CLARK, J.; DEROSE, S. et al. *XML path language (XPath)*. 1999. Citado na página 35.
- COCHRANE, G. et al. The international nucleotide sequence database collaboration. *Nucleic acids research*, Oxford University Press, v. 44, n. D1, p. D48–D50, 2016. Citado na página 27.
- COCK, P. J. et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, Oxford University Press, v. 25, n. 11, p. 1422, 2009. Citado 2 vezes nas páginas 28 e 66.
- COCKBURN, A. *Ports And Adapters Architecture*. 2006. <<http://wiki.c2.com/?PortsAndAdaptersArchitecture>> [Accessed: 2022-11-20]. Citado na página 56.
- CONSORTIUM, U. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, Oxford University Press, v. 47, n. D1, p. D506–D515, 2019. Citado na página 50.
- CZECH, L.; HUERTA-CEPAS, J.; STAMATAKIS, A. A critical review on the use of support values in tree viewers and bioinformatics toolkits. *Molecular biology and evolution*, Oxford University Press, v. 34, n. 6, p. 1535–1542, 2017. Citado na página 66.
- CZECH, L. et al. Metagenomic analysis using phylogenetic placement—a review of the first decade. *Frontiers in Bioinformatics*, v. 2, 5 2022. ISSN 2673-7647. Citado 4 vezes nas páginas 43, 44, 63 e 64.
- DAMM, U. et al. The colletotrichum acutatum species complex. *Studies in mycology*, Elsevier, v. 73, p. 37–113, 2012. Citado na página 32.
- DEINER, K. et al. Environmental dna metabarcoding: Transforming how we survey animal and plant communities. *Molecular ecology*, Wiley Online Library, v. 26, n. 21, p. 5872–5895, 2017. Citado na página 37.
- DEMIR, E. et al. The biopax community standard for pathway data sharing. *Nature biotechnology*, Nature Publishing Group US New York, v. 28, n. 9, p. 935–942, 2010. Citado na página 24.
- DESAI, H. P. et al. Comparative study using neural networks for 16s ribosomal gene classification. *Journal of Computational Biology*, Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New . . . , v. 27, n. 2, p. 248–258, 2020. Citado na página 77.

- DESAI, H. P. et al. Deep ensemble models for 16s ribosomal gene classification. In: SPRINGER. *Bioinformatics Research and Applications: 16th International Symposium, ISBRA 2020, Moscow, Russia, December 1–4, 2020, Proceedings 16*. [S.l.], 2020. p. 282–290. Citado na página 77.
- DESAI, N. et al. From genomics to metagenomics. *Current opinion in biotechnology*, Elsevier, v. 23, n. 1, p. 72–76, 2012. Citado na página 38.
- DESANTIS, T. Z. et al. Greengenes, a chimera-checked 16s rRNA gene database and workbench compatible with arb. *Applied and environmental microbiology*, Am Soc Microbiol, v. 72, n. 7, p. 5069–5072, 2006. Citado na página 40.
- DRUMMOND, A. J.; RAMBAUT, A. Beast: Bayesian evolutionary analysis by sampling trees. *BMC evolutionary biology*, Springer, v. 7, p. 1–8, 2007. Citado na página 45.
- DUNLAP, C. A.; BOWMAN, M. J.; ZEIGLER, D. R. Promotion of *Bacillus subtilis* subsp. *inaquosorum*, *Bacillus subtilis* subsp. *spizizenii* and *Bacillus subtilis* subsp. *stercoris* to species status. *Antonie Van Leeuwenhoek*, Springer, v. 113, p. 1–12, 2020. Citado na página 74.
- DUNLAP, C. A. et al. *Bacillus paralicheniformis* sp. nov., isolated from fermented soybean paste. *International journal of systematic and evolutionary microbiology*, Microbiology Society, v. 65, n. Pt_10, p. 3487–3492, 2015. Citado na página 74.
- DUPONT, A. et al. Differences in soil micro-eukaryotic communities over soil pH gradients are strongly driven by parasites and saprotrophs. *Environmental Microbiology*, Wiley Online Library, v. 18, n. 6, p. 2010–2024, 2016. Citado na página 37.
- DUVAL, E. Metadata standards: What, who & why. *J. Univers. Comput. Sci.*, v. 7, n. 7, p. 591–601, 2001. Citado 2 vezes nas páginas 17 e 21.
- EDGAR, R. Taxonomy annotation and guide tree errors in 16s rRNA databases. *PeerJ*, v. 6, p. e5030, 6 2018. ISSN 2167-8359. Citado 3 vezes nas páginas 17, 43 e 64.
- ELIAS, S. G. et al. Geneconnector: Unlocking the full potential of genbank metadata. *IEEE Latin America Transactions*, IEEE, v. 22, n. 2, p. 99–105, 2024. Citado na página 17.
- ELRAKAIBY, M. T. et al. Hospital microbiome variations as analyzed by high-throughput sequencing. *Omics: a journal of integrative biology*, Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New . . . , v. 23, n. 9, p. 426–438, 2019. Citado na página 37.
- FAN, B. et al. *Bacillus amyloliquefaciens*, *Bacillus velezensis*, and *Bacillus siamensis* form an “operational group *B. amyloliquefaciens*” within the *B. subtilis* species complex. *Frontiers in microbiology*, Frontiers Media SA, v. 8, p. 22, 2017. Citado na página 74.
- FELSENSTEIN, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, Springer, v. 17, p. 368–376, 1981. Citado na página 45.
- FIANNACA, A. et al. Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC bioinformatics*, Springer, v. 19, p. 61–76, 2018. Citado 2 vezes nas páginas 42 e 77.
- FIORAVANTI, D. et al. Phylogenetic convolutional neural networks in metagenomics. *BMC Bioinformatics*, v. 19, p. 49, 3 2018. ISSN 1471-2105. Citado 2 vezes nas páginas 63 e 64.

- FITCH, W. M. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Biology*, Society of Systematic Zoology, v. 20, n. 4, p. 406–416, 1971. Citado na página 45.
- FOURIE, A. et al. Molecular markers delimit cryptic species in *Ceratocystis* sensu stricto. *Mycological Progress*, Springer, v. 14, p. 1–18, 2015. Citado na página 51.
- GÁLVEZ-MERCHÁN, Á. et al. Metadata retrieval from sequence databases with ffq. *Bioinformatics*, Oxford University Press, v. 39, n. 1, p. btac667, 2023. Citado 2 vezes nas páginas 35 e 51.
- GATSON, J. W. et al. *Bacillus tequilensis* sp. nov., isolated from a 2000-year-old mexican shaft-tomb, is closely related to *Bacillus subtilis*. *International journal of systematic and evolutionary microbiology*, Society for General Microbiology, v. 56, n. 7, p. 1475–1484, 2006. Citado na página 74.
- GBIF, G. The global biodiversity information facility (2024) what is gbif. Available from [13 January 2020], 2020. Citado na página 23.
- GENBANK. *Sample Genbank Record*. [S.l.], 2024. Disponível em: <<https://www.ncbi.nlm.nih.gov/genbank/samplerecord/>>. Citado na página 27.
- GHANNAM, R. B.; TECHTMANN, S. M. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. *Computational and Structural Biotechnology Journal*, Elsevier, v. 19, p. 1092–1107, 2021. Citado na página 41.
- GILCHRIST, E. J.; WANG, S.; QUILICHINI, T. D. The impact of biotechnology and genomics on an ancient crop: *Cannabis sativa*. In: *Genomics and the Global Bioeconomy*. [S.l.]: Elsevier, 2023. p. 177–204. Citado na página 50.
- GOHLI, J. et al. The subway microbiome: seasonal dynamics and direct comparison of air and surface bacterial communities. *Microbiome*, Springer, v. 7, p. 1–16, 2019. Citado na página 38.
- GUINDON, S. et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phylml 3.0. *Systematic biology*, Oxford University Press, v. 59, n. 3, p. 307–321, 2010. Citado na página 45.
- GÜNTSCH, A.; BERENDSOHN, W. G.; MERGEN, P. The biocase project—a biological collections access service for Europe. *Ferrantia*, v. 51, p. 103–108, 2007. Citado na página 23.
- HANSON, B. et al. Characterization of the bacterial and fungal microbiome in indoor dust and outdoor air samples: a pilot study. *Environmental Science: Processes & Impacts*, Royal Society of Chemistry, v. 18, n. 6, p. 713–724, 2016. Citado na página 37.
- HARVEY, P. H.; PAGEL, M. D. *The comparative method in evolutionary biology*. [S.l.]: Oxford university press, 1991. Citado na página 63.
- HEATHER, J. M.; CHAIN, B. The sequence of sequencers: The history of sequencing DNA. *Genomics*, Elsevier, v. 107, n. 1, p. 1–8, 2016. Citado na página 37.
- HMP, T. H. M. P. C. A framework for human microbiome research. *Nature*, Nature Publishing Group UK London, v. 486, n. 7402, p. 215–221, 2012. Citado 2 vezes nas páginas 17 e 38.

HMP, T. H. M. P. C. Structure, function and diversity of the healthy human microbiome. *nature*, Nature Publishing Group UK London, v. 486, n. 7402, p. 207–214, 2012. Citado 2 vezes nas páginas 17 e 38.

HOLETSCHKEK, J. et al. The abcd of primary biodiversity data access. *Plant Biosystems-An International Journal Dealing with all Aspects of Plant Biology*, Taylor & Francis, v. 146, n. 4, p. 771–779, 2012. Citado na página 23.

HOLMES, E. C. The phylogeography of human viruses. *Molecular ecology*, Wiley Online Library, v. 13, n. 4, p. 745–756, 2004. Citado na página 63.

HOWE, K. L. et al. Ensembl 2021. *Nucleic acids research*, Oxford University Press, v. 49, n. D1, p. D884–D891, 2021. Citado na página 50.

HUERTA-CEPAS, J.; SERRA, F.; BORK, P. Ete 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution*, Society for Molecular Biology and Evolution, v. 33, n. 6, p. 1635–1638, 2016. Citado na página 66.

IKE-NWOSU, O. *Inside the python virtual machine*. [S.l.]: Lean Publishing, 2015. Citado na página 77.

ISAAC, N. J. et al. Mammals on the edge: conservation priorities based on threat and phylogeny. *PloS one*, Public Library of Science San Francisco, USA, v. 2, n. 3, p. e296, 2007. Citado na página 63.

JAMY, M. et al. Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. *Molecular Ecology Resources*, Wiley Online Library, v. 20, n. 2, p. 429–443, 2020. Citado 2 vezes nas páginas 45 e 64.

JEYASRI, R. et al. An overview of abiotic stress in cereal crops: negative impacts, regulation, biotechnology and integrated omics. *Plants*, MDPI, v. 10, n. 7, p. 1472, 2021. Citado 2 vezes nas páginas 17 e 50.

JIANG, Y. et al. Machine learning advances in microbiology: a review of methods and applications. *Frontiers in Microbiology*, Frontiers, v. 13, p. 925454, 2022. Citado na página 41.

JONES, M. et al. Ecological metadata language version 2.2.0. KNB Data Repository, 2019. Disponível em: <<https://eml.ecoinformatics.org>>. Citado 2 vezes nas páginas 17 e 24.

JUMA, C. *The gene hunters: Biotechnology and the scramble for seeds*. [S.l.]: Princeton University Press, 2014. v. 996. Citado na página 50.

KARSENTI, E. et al. A holistic approach to marine eco-systems biology. *PLoS biology*, Public Library of Science San Francisco, USA, v. 9, n. 10, p. e1001177, 2011. Citado na página 37.

KATO, K.; STANDLEY, D. M. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, Society for Molecular Biology and Evolution, v. 30, n. 4, p. 772–780, 2013. Citado na página 74.

KATZ, K. et al. The sequence read archive: a decade more of explosive growth. *Nucleic acids research*, Oxford University Press, v. 50, n. D1, p. D387–D390, 2022. Citado na página 37.

KEMBEL, S. W. et al. The phylogenetic diversity of metagenomes. *PLoS One*, Public Library of Science San Francisco, USA, v. 6, n. 8, p. e23214, 2011. Citado na página 64.

KENNEDY, J. B.; KUKLA, R.; PATERSON, T. Scientific names are ambiguous as identifiers for biological taxa: their context and definition are required for accurate data integration. In: SPRINGER. *International Workshop on Data Integration in the Life Sciences*. [S.l.], 2005. p. 80–95. Citado na página 23.

KHOROSHEVSKYI, O. et al. Geofetch: a command-line tool for downloading data and standardized metadata from geo and sra. *Bioinformatics*, Oxford University Press, v. 39, n. 3, p. btad069, 2023. Citado na página 35.

KÖLJALG, U. et al. Unite: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist*, Wiley Online Library, v. 166, n. 3, p. 1063–1068, 2005. Citado 2 vezes nas páginas 40 e 51.

KOZLOV, A. M. et al. Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Research*, v. 44, p. 5022–5033, 6 2016. ISSN 0305-1048. Citado na página 64.

LACOURSIÈRE-ROUSSEL, A. et al. Quantifying relative fish abundance with edna: a promising tool for fisheries management. *Journal of Applied Ecology*, Wiley Online Library, v. 53, n. 4, p. 1148–1157, 2016. Citado na página 37.

LAMBE, P. *Organising knowledge: taxonomies, knowledge and organisational effectiveness*. [S.l.]: Elsevier, 2014. Citado 2 vezes nas páginas 21 e 22.

LEE, T.-H.; KIM, Y.-K.; NAHM, B. H. Gbparsy: a genbank flatfile parser library with high speed. *BMC bioinformatics*, Springer, v. 9, p. 1–6, 2008. Citado na página 28.

LI, H.; DURBIN, R. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, Oxford University Press, v. 25, n. 14, p. 1754–1760, 2009. Citado na página 41.

LINARD, B.; SWENSON, K.; PARDI, F. Rapid alignment-free phylogenetic identification of metagenomic sequences. *Bioinformatics*, Oxford University Press, v. 35, n. 18, p. 3303–3312, 2019. Citado 3 vezes nas páginas 47, 63 e 64.

LOGARES, R. et al. Environmental microbiology through the lens of high-throughput dna sequencing: synopsis of current platforms and bioinformatics approaches. *Journal of microbiological methods*, Elsevier, v. 91, n. 1, p. 106–113, 2012. Citado na página 37.

LORIMER, J. et al. Making the microbiome public: participatory experiments with dna sequencing in domestic kitchens. *Transactions of the Institute of British Geographers*, Wiley Online Library, v. 44, n. 3, p. 524–541, 2019. Citado na página 38.

LYDON, K. A.; LIPP, E. K. Taxonomic annotation errors incorrectly assign the family pseudoalteromonadaceae to the order vibriionales in greengenes: implications for microbial community assessments. *PeerJ*, v. 6, p. e5248, 7 2018. ISSN 2167-8359. Citado 3 vezes nas páginas 17, 43 e 64.

MAHÉ, F. et al. Parasites dominate hyperdiverse soil protist communities in neotropical rainforests. *Nature ecology & evolution*, Nature Publishing Group UK London, v. 1, n. 4, p. 0091, 2017. Citado na página 37.

MAIDAK, B. L. et al. The rdp (ribosomal database project) continues. *Nucleic acids research*, Oxford University Press, v. 28, n. 1, p. 173–174, 2000. Citado 4 vezes nas páginas 40, 42, 63 e 77.

- MAIDAK, B. L. et al. The rdp (ribosomal database project). *Nucleic acids research*, Oxford University Press, v. 25, n. 1, p. 109–110, 1997. Citado 4 vezes nas páginas 40, 42, 63 e 77.
- MARDIS, E. R. Next-generation sequencing platforms. *Annual review of analytical chemistry*, Annual Reviews, v. 6, p. 287–303, 2013. Citado na página 37.
- MARDIS, E. R. Dna sequencing technologies: 2006–2016. *Nature protocols*, Nature Publishing Group UK London, v. 12, n. 2, p. 213–218, 2017. Citado na página 37.
- MARIN-FELIX, Y. et al. Genera of phytopathogenic fungi: Gophy 1. *Studies in mycology*, Elsevier, v. 86, p. 99–216, 2017. Citado na página 52.
- MARIN-FELIX, Y. et al. Genera of phytopathogenic fungi: Gophy 1. *Studies in mycology*, Elsevier, v. 86, p. 99–216, 2017. Citado na página 63.
- MARIN-FELIX, Y. et al. Genera of phytopathogenic fungi: Gophy 3. *Studies in mycology*, Elsevier, v. 94, p. 1–124, 2019. Citado na página 52.
- MARIN-FELIX, Y. et al. Genera of phytopathogenic fungi: Gophy 3. *Studies in mycology*, Elsevier, v. 94, p. 1–124, 2019. Citado na página 63.
- MARIN-FELIX, Y. et al. Genera of phytopathogenic fungi: Gophy 2. *Studies in mycology*, Elsevier, v. 92, p. 47–133, 2019. Citado na página 52.
- MARIN-FELIX, Y. et al. Genera of phytopathogenic fungi: Gophy 2. *Studies in mycology*, Elsevier, v. 92, p. 47–133, 2019. Citado na página 63.
- MATSEN, F. A.; KODNER, R. B.; ARMBRUST, E. V. pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics*, Springer, v. 11, p. 1–16, 2010. Citado 5 vezes nas páginas 18, 45, 46, 63 e 64.
- MCQUILTON, P. et al. Fairsharing, a cohesive community approach to the growth in standards, repositories and policies. 2018. Citado na página 27.
- MENZEL, P.; NG, K. L.; KROGH, A. Kaiju: Fast and sensitive taxonomic classification for metagenomics. *Biorxiv*, Cold Spring Harbor Laboratory, p. 031229, 2015. Citado 4 vezes nas páginas 41, 42, 63 e 77.
- MERKEL, D. Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, v. 2014, n. 239, p. 2, 2014. Citado na página 57.
- MICHENER, W. K. Ecological data sharing. *Ecological Informatics*, v. 29, p. 33–44, 9 2015. ISSN 15749541. Citado 2 vezes nas páginas 17 e 21.
- MIGNARDI, M.; NILSSON, M. Fourth-generation sequencing in the cell and the clinic. *Genome medicine*, Springer, v. 6, p. 1–4, 2014. Citado na página 37.
- MIOTTO, O.; TAN, T. W.; BRUSIC, V. Rule-based knowledge aggregation for large-scale protein sequence analysis of influenza a viruses. In: SPRINGER. *BMC bioinformatics*. [S.l.], 2008. v. 9, p. 1–14. Citado 2 vezes nas páginas 34 e 35.
- MOHALI, S.; SLIPPERS, B.; WINGFIELD, M. J. Two new fusicoccum species from acacia and eucalyptus in venezuela, based on morphology and dna sequence data. *mycological research*, Elsevier, v. 110, n. 4, p. 405–413, 2006. Citado na página 31.

- MORGANTE, M.; SALAMINI, F. From plant genomics to breeding practice. *Current Opinion in Biotechnology*, Elsevier, v. 14, n. 2, p. 214–219, 2003. Citado na página 63.
- MRTG, M. R. T. G. *Audubon Core Structure*. [S.l.], 2020. Citado na página 23.
- MURALI, A.; BHARGAVA, A.; WRIGHT, E. S. Idtaxa: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome*, Springer, v. 6, p. 1–14, 2018. Citado na página 42.
- MURALIDHARAN, H. S.; FOX, N. Y.; POP, M. The impact of transitive annotation on the training of taxonomic classifiers. *Frontiers in Microbiology*, v. 14, 1 2024. ISSN 1664-302X. Citado 2 vezes nas páginas 43 e 64.
- NAKAMURA, L. Taxonomic relationship of black-pigmented bacillus subtilis strains and a proposal for bacillus atrophaeus sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, Microbiology Society, v. 39, n. 3, p. 295–300, 1989. Citado na página 74.
- NCBI. *GenBank® Passes the 100 Gigabase Mark: in NCBI News*. [S.l.]: National Center for Biotechnology Information, 2005. Citado na página 37.
- OGDEN, R. Forensic science, genetics and wildlife biology: getting the right mix for a wildlife dna forensics lab. *Forensic science, medicine, and pathology*, Springer, v. 6, n. 3, p. 172–179, 2010. Citado na página 63.
- OUNIT, R. et al. Clark: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC genomics*, Springer, v. 16, p. 1–13, 2015. Citado 3 vezes nas páginas 17, 39 e 42.
- PALMISANO, M. M. et al. Bacillus sonorensis sp. nov., a close relative of bacillus licheniformis, isolated from soil in the sonoran desert, arizona. *International Journal of Systematic and Evolutionary Microbiology*, Microbiology Society, v. 51, n. 5, p. 1671–1679, 2001. Citado na página 74.
- PAREEK, C. S.; SMO CZYNSKI, R.; TRETYN, A. Sequencing technologies and genome sequencing. *Journal of applied genetics*, Springer, v. 52, p. 413–435, 2011. Citado na página 37.
- PEARSON, W. R. An introduction to sequence similarity (“homology”) searching. *Current protocols in bioinformatics*, Wiley Online Library, v. 42, n. 1, p. 3–1, 2013. Citado na página 38.
- PEARSON, W. R. Finding protein and nucleotide similarities with fasta. *Current protocols in bioinformatics*, Wiley Online Library, v. 53, n. 1, p. 3–9, 2016. Citado na página 39.
- PEARSON, W. R.; LIPMAN, D. J. Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 85, n. 8, p. 2444–2448, 1988. Citado 3 vezes nas páginas 17, 39 e 42.
- PRIEST, F. et al. Bacillus amyloliquefaciens sp. nov., nom. rev. *International journal of systematic and evolutionary microbiology*, Microbiology Society, v. 37, n. 1, p. 69–71, 1987. Citado na página 74.
- PRIHASTUTI, H. et al. Characterization of colletotrichum species associated with coffee berries in northern thailand. *Fungal Diversity*, Kunming, China, v. 39, n. 1, p. 89–109, 2009. Citado na página 32.

- QU, K. et al. Application of machine learning in microbiology. *Frontiers in microbiology*, Frontiers, v. 10, p. 451710, 2019. Citado na página 41.
- QUAST, C. et al. The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic acids research*, Oxford University Press, v. 41, n. D1, p. D590–D596, 2012. Citado 2 vezes nas páginas 40 e 45.
- QUEIROZ, K. de. Nodes, branches, and phylogenetic definitions. *Systematic biology*, Oxford University Press, v. 62, n. 4, p. 625–632, 2013. Citado na página 77.
- QUIÑONES, M. et al. Metagenote: a simplified web platform for metadata annotation of genomic samples and streamlined submission to ncbi’s sequence read archive. *BMC bioinformatics*, Springer, v. 21, p. 1–12, 2020. Citado na página 51.
- RAMBOLD, G. et al. Meta-omics data and collection objects (mod-co): a conceptual schema and data model for processing sample data in meta-omics research. *Database*, Oxford University Press, v. 2019, p. baz002, 2019. Citado na página 24.
- RANNALA, B.; YANG, Z. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of molecular evolution*, Springer, v. 43, p. 304–311, 1996. Citado na página 45.
- REINING, S. et al. Knowledge accumulation in design science research: ways to foster scientific progress. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, ACM New York, NY, USA, v. 53, n. 1, p. 10–24, 2022. Citado na página 52.
- RICE, P.; LONGDEN, I.; BLEASBY, A. Emboss: the european molecular biology open software suite. *Trends in genetics*, Elsevier, v. 16, n. 6, p. 276–277, 2000. Citado na página 28.
- RILEY, J. Understanding metadata. *Washington DC, United States: National Information Standards Organization (<http://www.niso.org/publications/press/UnderstandingMetadata.pdf>)*, v. 23, p. 7–10, 2017. Citado 2 vezes nas páginas 22 e 23.
- ROBERTS, M. S.; NAKAMURA, L.; COHAN, F. M. *Bacillus mojavensis* sp. nov., distinguishable from *Bacillus subtilis* by sexual isolation, divergence in dna sequence, and differences in fatty acid composition. *International Journal of Systematic and Evolutionary Microbiology*, Microbiology Society, v. 44, n. 2, p. 256–264, 1994. Citado na página 74.
- ROBERTS, M. S.; NAKAMURA, L. K.; COHAN, F. M. *Bacillus vallismortis* sp. nov., a close relative of *Bacillus subtilis*, isolated from soil in death valley, california. *International Journal of Systematic and Evolutionary Microbiology*, Microbiology Society, v. 46, n. 2, p. 470–475, 1996. Citado na página 74.
- ROGNES, T. et al. Vsearch: a versatile open source tool for metagenomics. *PeerJ*, PeerJ Inc., v. 4, p. e2584, 2016. Citado 2 vezes nas páginas 45 e 73.
- ROMASHCHENKO, N. et al. Epik: precise and scalable evolutionary placement with informative k-mers. *Bioinformatics*, Oxford University Press, v. 39, n. 12, p. btad692, 2023. Citado 2 vezes nas páginas 47 e 64.
- RONQUIST, F. et al. Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, Oxford University Press, v. 61, n. 3, p. 539–542, 2012. Citado 2 vezes nas páginas 45 e 64.

ROONEY, A. P. et al. Phylogeny and molecular taxonomy of the bacillus subtilis species complex and description of bacillus subtilis subsp. inaquosorum subsp. nov. *International journal of systematic and evolutionary microbiology*, Society for General Microbiology, v. 59, n. 10, p. 2429–2436, 2009. Citado na página 72.

ROONEY, A. P. et al. Phylogeny and molecular taxonomy of the bacillus subtilis species complex and description of bacillus subtilis subsp. inaquosorum subsp. nov. *International journal of systematic and evolutionary microbiology*, Society for General Microbiology, v. 59, n. 10, p. 2429–2436, 2009. Citado na página 74.

ROSSUM, G. V.; DRAKE, F. L. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN 1441412697. Citado na página 56.

RUIZ-GARCIA, C. et al. Bacillus velezensis sp. nov., a surfactant-producing bacterium isolated from the river vélez in Málaga, southern Spain. *International Journal of Systematic and Evolutionary Microbiology*, Microbiology Society, v. 55, n. 1, p. 191–195, 2005. Citado na página 74.

RUPPERT, K. M.; KLINE, R. J.; RAHMAN, M. S. Past, present, and future perspectives of environmental DNA (edna) metabarcoding: A systematic review in methods, monitoring, and applications of global edna. *Global Ecology and Conservation*, Elsevier, v. 17, p. e00547, 2019. Citado na página 37.

SAKAMOTO, T.; ORTEGA, J. M. Taxallnomy: an extension of ncbi taxonomy that produces a hierarchically complete taxonomic tree. 2020. Disponível em: <<http://bioinfo.icb.ufmg.br/taxallnomy>> Citado na página 64.

SCHLOSS, P. D. et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, Am Soc Microbiol, v. 75, n. 23, p. 7537–7541, 2009. Citado na página 63.

SCHOCH, C. L. et al. Ncbi taxonomy: a comprehensive update on curation, resources and tools. *Database*, Oxford University Press, v. 2020, p. baaa062, 2020. Citado na página 74.

SCHOCH, C. L. et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. *Proceedings of the national academy of Sciences*, National Acad Sciences, v. 109, n. 16, p. 6241–6246, 2012. Citado na página 65.

SEGATA, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, Nature Publishing Group US New York, v. 9, n. 8, p. 811–814, 2012. Citado na página 64.

SHARMA, G.; SHENOY, B. D. Colletotrichum fructicola and C. siamense are involved in chili anthracnose in India. *Archives of Phytopathology and Plant Protection*, Taylor & Francis, v. 47, n. 10, p. 1179–1194, 2014. Citado na página 34.

SHENDURE, J.; FINDLAY, G. M.; SNYDER, M. W. Genomic medicine—progress, pitfalls, and promise. *Cell*, Elsevier, v. 177, n. 1, p. 45–57, 2019. Citado 2 vezes nas páginas 17 e 50.

SMEDLEY, D. et al. Biomart—biological queries made easy. *BMC genomics*, v. 10, n. 1, p. 1–12, 2009. Citado na página 50.

- SONG, D. et al. Parameterized blosum matrices for protein alignment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, IEEE, v. 12, n. 3, p. 686–694, 2014. Citado na página 41.
- SQUASSINA, A. et al. Realities and expectations of pharmacogenomics and personalized medicine: impact of translating genetic knowledge into clinical practice. *Pharmacogenomics, Future Medicine*, v. 11, n. 8, p. 1149–1167, 2010. Citado na página 63.
- SRIVASTAVA, D. S. et al. Phylogenetic diversity and the functioning of ecosystems. *Ecology letters*, Wiley Online Library, v. 15, n. 7, p. 637–648, 2012. Citado na página 64.
- STAJICH, J. E. et al. The bioperl toolkit: Perl modules for the life sciences. *Genome research*, Cold Spring Harbor Lab, v. 12, n. 10, p. 1611–1618, 2002. Citado na página 28.
- STAMATAKIS, A. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, Oxford University Press, v. 30, n. 9, p. 1312–1313, 2014. Citado 4 vezes nas páginas 44, 45, 64 e 74.
- STAMATAKIS, A.; HOOVER, P.; ROUGEMONT, J. A rapid bootstrap algorithm for the raxml web servers. *Systematic biology*, Taylor & Francis, v. 57, n. 5, p. 758–771, 2008. Citado na página 74.
- SUMPAVAPOL, P. et al. *Bacillus siamensis* sp. nov., isolated from salted crab (poo-khem) in thailand. *International journal of systematic and evolutionary microbiology*, Society for General Microbiology, v. 60, n. 10, p. 2364–2370, 2010. Citado na página 74.
- SWOFFORD, D. L. Paup, phylogenetic analysis using parsimony. *version 3.1. Computer program distributed by the Illinois Natural History Survey*, Champaign, 1993. Citado na página 64.
- TEMPERTON, B.; GIOVANNONI, S. J. Metagenomics: microbial diversity through a scratched lens. *Current opinion in microbiology*, Elsevier, v. 15, n. 5, p. 605–612, 2012. Citado na página 39.
- TESSLER, M. et al. Large-scale differences in microbial biodiversity discovery between 16s amplicon and shotgun sequencing. *Scientific reports*, Nature Publishing Group UK London, v. 7, n. 1, p. 6589, 2017. Citado na página 41.
- TIAN, Q. et al. Application and comparison of machine learning and database-based methods in taxonomic classification of high-throughput sequencing data. *Genome Biology and Evolution*, Oxford University Press, p. evae102, 2024. Citado na página 42.
- TINDALL, B. The consequences of *Bacillus axarquiensis* Ruiz-García et al. 2005, *Bacillus malacitensis* Ruiz-García et al. 2005 and *Brevibacterium halotolerans* Delaporte and Sasson 1967 (approved lists 1980) being treated as heterotypic synonyms. *International journal of systematic and evolutionary microbiology*, Microbiology Society, v. 67, n. 1, p. 175–176, 2017. Citado na página 74.
- TREES, R. P. The neighbor-joining method: a new method for. *Mol Biol Evol*, v. 4, n. 4, p. 406–425, 1987. Citado na página 45.
- TREVETHAN, R. Sensitivity, specificity, and predictive values: foundations, pliabilitys, and pitfalls in research and practice. *Frontiers in public health*, Frontiers Media SA, v. 5, p. 307, 2017. Citado na página 72.

- TRUONG, D. T. et al. Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nature methods*, Nature Publishing Group, v. 12, n. 10, p. 902–903, 2015. Citado na página 64.
- VANGAY, P. et al. Microbiome metadata standards: Report of the national microbiome data collaborative’s workshop and follow-on activities. *Msystems*, Am Soc Microbiol, v. 6, n. 1, p. 10–1128, 2021. Citado 2 vezes nas páginas 24 e 25.
- VBD, S. Approved lists of bacterial names. *Int J Syst Bacteriol*, v. 30, p. 225–420, 1980. Citado na página 74.
- WANG, L.-T. et al. Comparison of gyrB gene sequences, 16s rrna gene sequences and dna–dna hybridization in the bacillus subtilis group. *International journal of systematic and evolutionary microbiology*, Society for General Microbiology, v. 57, n. 8, p. 1846–1850, 2007. Citado na página 72.
- WANG, W.-L. et al. Application of metagenomics in the human gut microbiome. *World journal of gastroenterology: WJG*, Baishideng Publishing Group Inc, v. 21, n. 3, p. 803, 2015. Citado na página 38.
- WASSAN, J. T.; WANG, H.; ZHENG, H. A new phylogeny-driven random forest-based classification approach for functional metagenomics. In: . IEEE, 2022. p. 32–37. ISBN 978-1-6654-6819-0. Disponível em: <<https://ieeexplore.ieee.org/document/9995554/>>. Citado 2 vezes nas páginas 63 e 64.
- WASSAN, J. T.; WANG, H.; ZHENG, H. Developing a new phylogeny-driven random forest model for functional metagenomics. *IEEE Transactions on NanoBioscience*, v. 22, p. 763–770, 10 2023. ISSN 1536-1241. Disponível em: <<https://ieeexplore.ieee.org/document/10144805/>>. Citado 2 vezes nas páginas 63 e 64.
- WEIGMANN, H. Über zwei an der käsereifung beteiligte bakterien. *Zentralbl. Bakteriol. Hyg. II*, v. 4, p. 820–834, 1898. Citado na página 74.
- WEIR, B.; JOHNSTON, P.; DAMM, U. The colletotrichum gloeosporioides species complex. *Studies in mycology*, Elsevier, v. 73, p. 115–180, 2012. Citado na página 33.
- WEIR, B. S.; JOHNSTON, P. R. Characterisation and neotypification of gloeosporium kaki hori as colletotrichum horii nom. nov. *Mycotaxon*, Mycotaxon, v. 111, n. 1, p. 209–219, 2010. Citado na página 33.
- WHEELER, D. L. et al. Database resources of the national center for biotechnology. *Nucleic acids research*, Oxford University Press, v. 31, n. 1, p. 28–33, 2003. Citado na página 27.
- WIECZOREK, J. et al. Darwin core: an evolving community-developed biodiversity data standard. *PloS one*, Public Library of Science San Francisco, USA, v. 7, n. 1, p. e29715, 2012. Citado 2 vezes nas páginas 23 e 26.
- WILGENBUSCH, J. C.; SWOFFORD, D. Inferring evolutionary trees with paup. *Current protocols in bioinformatics*, Wiley Online Library, n. 1, p. 6–4, 2003. Citado na página 45.
- WILKINSON, M. D. et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, Nature Publishing Group, v. 3, n. 1, p. 1–9, 2016. Citado 2 vezes nas páginas 17 e 26.

- WOOD, D. E.; LU, J.; LANGMEAD, B. Improved metagenomic analysis with kraken 2. *Genome biology*, Springer, v. 20, p. 1–13, 2019. Citado 4 vezes nas páginas 41, 42, 63 e 77.
- WOOD, D. E.; SALZBERG, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, BioMed Central, v. 15, n. 3, p. 1–12, 2014. Citado 4 vezes nas páginas 41, 42, 63 e 77.
- WRIGHT, E. S. Using decipher v2. 0 to analyze big biological sequence data in r. *R Journal*, v. 8, n. 1, 2016. Citado na página 41.
- WYK, M. V. et al. *Ceratocystis manginecans* sp. nov., causal agent of a destructive mango wilt disease in oman and pakistan. *Fungal Divers*, v. 27, p. 213–230, 2007. Citado 2 vezes nas páginas 30 e 51.
- WYK, M. V. et al. Dna based characterization of *ceratocystis fimbriata* isolates associated with mango decline in oman. *Australasian Plant Pathology*, Springer, v. 34, p. 587–590, 2005. Citado 2 vezes nas páginas 30 e 51.
- WYK, M. V. et al. Two new *ceratocystis* species associated with mango disease in brazil. *Mycotaxon*, Mycotaxon, v. 117, n. 1, p. 381–404, 2011. Citado 3 vezes nas páginas 30, 31 e 51.
- XIE, L. et al. Biology of *colletotrichum horii*, the causal agent of persimmon anthracnose. *Mycology*, Taylor & Francis, v. 1, n. 4, p. 242–253, 2010. Citado na página 33.
- YANG, T. et al. Families, genera, and species of botryosphaeriales. *Fungal biology*, Elsevier, v. 121, n. 4, p. 322–346, 2017. Citado na página 31.
- YANG, Z.; RANNALA, B. Bayesian phylogenetic inference using dna sequences: a markov chain monte carlo method. *Molecular biology and evolution*, v. 14, n. 7, p. 717–724, 1997. Citado na página 45.
- YANG, Z.; RANNALA, B. Molecular phylogenetics: principles and practice. *Nature reviews genetics*, Nature Publishing Group UK London, v. 13, n. 5, p. 303–314, 2012. Citado 2 vezes nas páginas 45 e 63.
- YILMAZ, P. et al. Minimum information about a marker gene sequence (mimarks) and minimum information about any (x) sequence (mixs) specifications. *Nature biotechnology*, Nature Publishing Group US New York, v. 29, n. 5, p. 415–420, 2011. Citado 2 vezes nas páginas 25 e 26.

Apêndices

APÊNDICE A – Manuscrito referente ao Capítulo 4

GeneConnector: Unlocking the Full Potential of Genbank Metadata

Samuel Galvão Elias , Débora Cervieri Guterres , Robert Weingart Barreto , and Helson Mário Martins do Vale 

Abstract—The Genbank database serves as a pivotal global repository for genetic information, housing an extensive and diverse array of data. Nonetheless, a significant proportion of its existing records suffer from fragmented and often inadequate metadata, thereby failing to furnish the requisite contextual information regarding their acquisition. In response to this challenge, we introduce GeneConnector, a novel tool designed to harness shared information within multiple records of the same specimen in Genbank, with the ultimate objective of augmenting the completeness of inadequately annotated nodes spanning various information domains. To exemplify the capabilities of this tool, we conducted a comprehensive review and aggregation of available data, utilizing the database for Genera of Phytopathogenic Fungi (GOPHY) as a testbed. Our evaluation revealed substantial improvements in information retrieval through the analysis of shared data among nodes that connect Genbank specimen records, yielding notable enhancements ranging from 2% to an astonishing 60%. Our approach equips users with the means to conduct precise, facile, and accurate assessments of the contextual associations of results, facilitated by two distinct metrics that assess the current level of data annotation and the potential information enhancement achievable through our evaluation, the Observed Completeness Score (OCS) and the Reachable Completeness Score (RCS).

Link to graphical and video abstracts, and to code: <https://latamt.ieeer9.org/index.php/transactions/article/view/8241>

Index Terms—Genbank, NCBI, Gene-Connector, Mycology, Phytopathology, GOPHY.

I. INTRODUCTION

Genomic data has become increasingly important in many fields of research, including medicine [1], biotechnology, and agriculture [2]–[4]. However, the sheer volume and complexity of this data can make it challenging to extract meaningful insights. Public databases, such as GenBank [5], Ensembl [6], and UniProt [7], provide a wealth of information on genes, genomes, and their products. However, accessing and analyzing this information can be a time-consuming and daunting task. Therefore, the development of tools that can perform the aggregation of genomic metadata in public databases is critical to advancing research in genomics.

One example of such a tool is BioMart [8], a widely used data management system that allows users to query

Samuel Galvão Elias and Helson Mário Martins do Vale are with University of Brasília, Brasília, Brazil (e-mail: sgelias@outlook.com and helson@unb.br).

Débora Cervieri Guterres and Robert Weingart Barreto are with Federal University of Viçosa, Brazil (e-mail: debora.guterres@gmail.com and rbarreto@ufv.br).

"data" (only) from multiple biological databases simultaneously. BioMart has been used in many studies, including the identification of genetic markers associated with disease and the exploration of gene expression patterns in different tissues. Another example is Ensembl's Biomart (see details), which allows users to query Ensembl's databases using the same interface as BioMart. These tools are just a few examples of the many resources available to researchers looking to access and analyze genomic "data".

Box 1 | The *Ceratomyces mangicola* [9] study-case.

ITS Submitted in Mar 4, 2005 (available under the Genbank accession n° AY953382, [10]).
EF1 Feb 13, 2007 (EF433316, [11]).
TUB1 Feb 13, 2007 (EF433307, [11]).
RPB2 Mar 11, 2014 (KJ601618, [12]).
MS204 Mar 11, 2014 (KJ601582, [12]).

A pathogen originally described as member of the *Ceratomyces fimbriata sensu lato* complex causing the *Mangifera indica* disease, known as mango blight, murcha, or *seca da mangueira* in Brazil. Records of *C. mangicola* were registered in three different submission events, with a large time lag between the first (the Internal Transcribed Spacer [ITS] submission) and the latest submission events (RNA polymerase subunit II [RPB2] and guanine nucleotide-binding protein subunit beta-like protein [MS204], nine years after). Such time lag allowed the information associated to the *C. mangicola* to be gradually extended. Since the first registration of the ITS marker, the associated information was upgraded, starting from basic source modifiers as isolate and organism name to a well documented record including strain, specimen-voucher, type-materials, host, country, and others (see the Fig. 1 for details of the information gain associated to *C. mangicola*).

As attempted readers can see, the "data" has been the center of attention when it comes to data aggregation, while metadata is much more often overlooked. Therefore, the aggregation of genomic metadata is important because it enables researchers to integrate data from multiple sources and make more comprehensive and accurate analyses (important examples includes [13]–[16]). For example, by combining

genomic data with clinical data, researchers can identify genetic markers associated with disease and develop more effective treatments. The aggregation of genomic metadata also enables the identification of patterns and trends that may not be apparent when examining individual datasets. These patterns and trends can provide insights into the most variable scientific domains.

Despite the existence and importance of the tools that performing aggregation of genomic metadata from single records, and focused in high-throughput sequencing data (examples include Metagenote [17], and ffq [18]), there are no tools that aggregate multi-loci data. There are still challenges associated with accessing and analyzing such data, and information consistency is maybe the most important of these (see [19] for a important study-case about Genbank information consistency). GeneConnector works around such challenges. Our proposal is to create connections between unique Genbank records and use the "unique + shared" information between records to improve single gene annotations.

When specifically dealing nucleotide data stored in Genbank, it is common to observe events of information increment associated with advancements in knowledge regarding taxonomic groups (see Box 1 for an example). The phenomenon of information increment events can be attributed to the dynamic nature of scientific research. As researchers delve deeper into the genetic makeup of various organisms, they uncover novel data points and identify previously unrecognized patterns.

These discoveries, when incorporated into the Genbank database, enhance the breadth and depth of information available for taxonomic analysis. Consequently, with each advancement in our understanding of taxonomic groups, a ripple effect occurs, influencing future studies, expanding our knowledge base, and fostering further scientific breakthroughs (the natural stepping stones of science [20]).

Finally, GeneConnector was designed precisely to absorb this intrinsic characteristic of Genbank data during metadata acquisition campaigns. Therefore, our aim is to illustrate how this tool can enhance the metadata quality of a comprehensive database solely by leveraging the shared information within Genbank records. Furthermore, we introduce our novel approach, the Observed Completeness Score (OCS) and the Reachable Completeness Score (RCS), for quantifying the level of completeness in records associated with specimens with available information in Genbank. To accomplish this objective, we employed the comprehensive database for Genera of Phytopathogenic Fungi (GOPHY) as a case study ([21]–[24]).

II. PROBLEM STATEMENT

Genbank, a widely used repository for nucleotide sequence data, contains an immense amount of valuable genomic information. However, the lack of consistent and standardized metadata across the records poses a significant challenge for researchers aiming to extract meaningful insights from this vast collection. Existing approaches for metadata extraction and aggregation from Genbank records are often limited, inefficient, or require manual curation, hindering the ability to comprehensively exploit the data for scientific research.

To address this problem, a novel software solution has been developed to automate the process of populating and aggregating metadata from Genbank nucleotide records. The software aims to extract diverse metadata attributes, including taxonomy, organism properties, sequencing techniques, geographical location, and biological features, among others, from the extensive Genbank database. By automating this labor-intensive task, researchers will be empowered to efficiently access and analyze metadata associated with nucleotide sequences, enhancing their ability to conduct comprehensive studies and accelerate advancements in various biological and genomic fields.

The research article aims to evaluate the effectiveness and reliability of the GeneConnector in extracting and aggregating metadata from a large-scale sample of Genbank nucleotide records. The outcomes of this research will contribute to improving the accessibility and quality of metadata associated with Genbank nucleotide records.

III. PROPOSED SOLUTION

A. Concepts and Information Modelling

Our tool was developed to modelling the information contained in Genbank records based in three basic data models: Metadata, Nodes, and Connections (see Fig. 1). A *Connection* is a top-level object centralizing *Nodes*. A single *Node* carries information of the accession number that originated the object, and the gene marker from which the record was extracted, connecting all metadata related to the original Genbank record. *Metadata* objects abstract the Genbank raw qualifiers information.

Raw Genbank qualifiers includes a list of key/value pairs describing the context associated to given nucleotide record. Our tool was developed to turn qualifiers into importance groups (from here on we will call them Metadata Indicator Groups, MIG) that mirrors the information relevance which turn a desired specimen unique (information importance are available in Table I, and visually explored in Fig. 1). For example, a taxonomic key (e.g. organism [with value *Bipolaris victoriae*]) is shared between multiple real world specimens, so it must have less importance than a specimen related key (e.g. isolate [with value *CBS:327.64*]).

TABLE I
METADATA INDICATOR GROUPS USED TO RANK KEYS
AND CALCULATE THE GENE CONNECTOR COMPLETENESS
SCORES

Group	Score	Description
SPECIMEN	8	Unique identifiers of specimen.
TAXONOMY	5	Taxonomy related keys.
HOST_SUBSTRATE	3	Identifiers for host interactions.
TIME_REFERENCES	2	Time milestones.
GEO_REFERENCES	2	Geographical indicators.
ASSAY	0	Related to gene sequencing methods.
EXTERNAL_LINKS	0	References to external databases.
ACTORS	0	Human actors related to the record.
OTHER	0	Not already mapped keys.

Based on these principles, our tool systematically punctuates metadata from independent Genbank records, and calculates

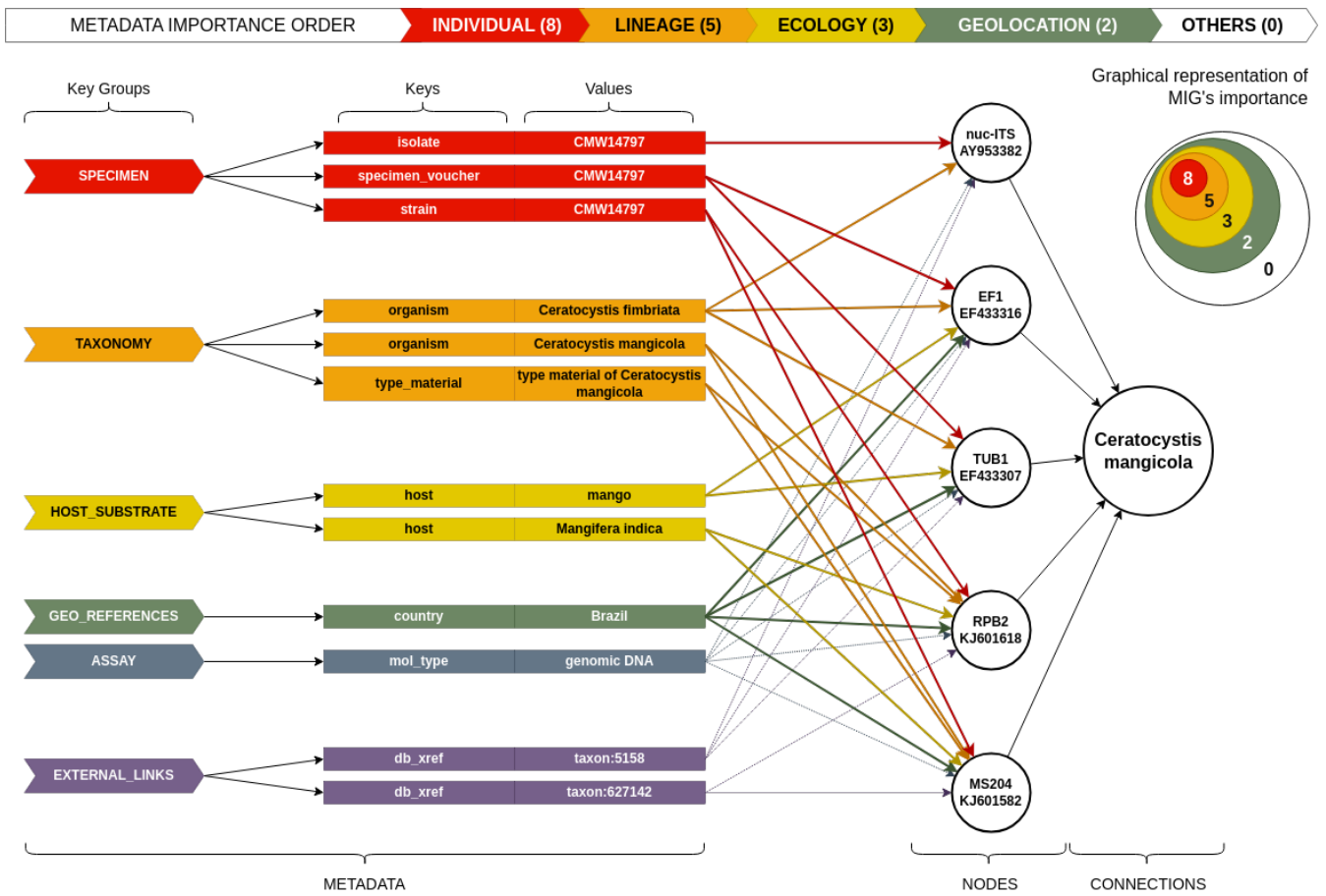


Fig. 1. GeneConnector information diagram. The diagram represents the main information models used by our tool to deal with Genbank records information: Metadata, Nodes, and Connections. Dotted arrows connecting metadata and nodes layers indicating zero scored MIG's. Warmer colors indicate more specific information about the specimen to which the nodes (genes) belong. Therefore, the closer to red indicates metadata with greater power to approximate nodes.

the information completeness associated to a set of records that represents real world specimens, improving the information usability.

As already demonstrated in Table I, the MIG importance score is expressed on a Fibonacci scale and enables the differentiation of important metadata from spurious ones, thereby facilitating the calculation of two completeness scores: the Observed Completeness Score (OCS) and the Reachable Completeness Score (RCS).

The OCS measures how well a connection is annotated in terms of information domains taking into account parameters as uniqueness (SPECIMEN), tree of life placement (TAXONOMY), ecological placement (HOST_SUBSTRATE), temporal marks (TIME_REFERENCES), spatial marks (GEO_REFERENCES), and less relevant ones (ASSAY, EXTERNAL_LINKS, ACTORS, and OTHER). The OCS is calculated independently for each node.

The RCS is a metric used to assess the completeness of connections based on the nodes they connect. Unlike the OCS, which considers all nodes independently, the RCS takes into account the dependencies between nodes that composes a connection. Specifically, the RCS is calculated as the ratio of the number of observed connections between nodes to the

number of possible connections within a given MIG. This score reflects the degree to which the available metadata within a given MIG is interconnected and should be used to complete another nodes.

The three main steps of an hypothetical calculation of the OCS and RCS are described below where the results of individual steps are shown in Table II.

- Step 1. Finding nodes with at least one occurrence of qualifiers of each MIG with a score greater than zero:

Let N be the set of nodes.

Let Q be the set of qualifiers.

Let M be the set of MIGs.

The condition to find such nodes can be represented as:

$$\forall n \in N, \exists m \in M, \exists q \in Q \text{ with } \text{score}(m) > 0 \text{ such that } \text{occurs}(n, q)$$

- Step 2. Annotating nodes with the MIG score (no more than one key per MIG scored by a node):

Let S be the function that assigns a score to a node.

The annotation can be represented as:

$$\forall n \in N, \exists m \in M, S(n) = \text{score}(m)$$

- Step 3. Calculating the expected score during the second step:

Let E represent the expected score.

The calculation can be represented as the sum of products:

$$E = \sum_{m \in M} \text{score}(m) \cdot \text{number of nodes}$$

Finally we filtering expected scores to find MIGs with at least one member (penalizing MIGs with a zero score if not represented):

Let F be the set of MIGs with at least one member.

The filtering can be represented as:

$$F = \{m \in M \mid \exists n \in N \text{ such that occurs}(n, m)\}$$

The penalization of MIGs not represented can be represented as:

$$\forall m \in M \setminus F, \text{score}(m) = 0$$

TABLE II

THE FIRST THREE STEPS OF THE COMPLETENESS SCORES CALCULATION WITH HYPOTHETICAL NODES A, B, C, AND D

Group	Step 1				Step 2	Step 3
	A	B	C	D	E-score [†]	0-score [‡]
SPECIMEN	8	8	-	8	32	32
TAXONOMY	5	5	5	5	20	20
HOST_SUBSTRATE	-	3	3	-	12	12
TIME_REFERENCES	-	-	-	-	8	0
GEO_REFERENCES	2	-	2	-	8	8
	15	16	10	13	80	72
Conn. Obs. score	54					
OCS	0.68					
RCS	0.90					

Step 1 contain four hypothetical nodes A, B, C, and D with group scores, respective. Dash indicate groups not represented in Node. Step 2 and Step 3 includes expected scores, and non-zero group scores, respectively. [†] Expected score by group. The product of the nodes number and the score value of the given group. [‡] Non-zero score by group. The same as expected score if at last one Node contains a given group. Otherwise is zero.

After execute the above steps we can calculate the Connection Observed Score by sum individual node scores (conn. obs. score = 54 in Table II). Next, the OBS is calculated being the ratio between the Connection Observed Score and the sum of Expected scores ($54 / 80 = 0.68$ in Table II), and finally the RCS should be calculated as the ratio of the step 3's values sum and the sum of Expected scores ($72 / 80 = 0.90$ in Table II). This is a simple and elegant way to represents the information completeness of arbitrary Genbank records.

B. Technologies and Code Availability

GeneConnector was developed in Python (3.11+ [25]) adopting the hexagonal architecture [26]. The complete logic for the calculation of scores, data parsing, data validations, and the data collection from Genbank are centered at the package core sub-module. For curious readers, a complete metadata list by MIG should be found at the Github repository `sgelias/geneconnector-cli` within the 'metadata' file `src/gcon/core/domain/dtos/metadata.py`. Our tools is Open Source and the codebase is available under the MIT license (see details).

C. Study Case: GOPHY Data Completeness

To demonstrate the performance and value proposition of GeneConnector we downloaded and evaluate the complete GOPHY's database containing seventeen gene markers and 1,246 specimen records. The complete database is available as a Supplementary material into the GeneConnector Github directory (files named `gcon-input-gophy.xlsx` in `docs/manuscript/supplementary-material`).

We value simplicity, so we make running the GeneConnector possible through a single command named **resolve** available after the tool installation on the host system. Currently our tool was tested only using Linux systems, thus, over Windows or Macintosh systems we recommend to run using a Docker environment [27]. See below the execution command of GeneConnector CLI:

The Code snippet of Listing 1 exemplifies our package execution. After installed GeneConnector should be called using the `gcon` callable and the `resolve` command used to execute the full package pipeline. Required arguments are shown in lines 5, 6, and 7 of the previous code snippet. A comprehensive user guide is available at the GeneConnector Github directory.

```

1 # GeneConnector execution in Linux environments
2 # using the 'resolve' command of the 'gcon' package.
3
4 $ gcon resolve \
5   --input-table input-table.tsv \
6   --temporary-directory /tmp/gcon/ \
7   --output-file gcon-out

```

Listing 1: GeneConnector execution command example. Lines started with hashtag are code comments, so they are not executed.

The output generated by the aforementioned command encompasses a tabular file (TSV) that amalgamates several crucial components: (i) input table information, (ii) OCS and RCS scores, (iii) a statistical percentage depicting the information gain, which quantifies the quantity of information salvaged following the evaluation of metadata under the *Nodes* category, (iv) signatures, and ultimately, (v) all metadata associated with individual connections.

Signatures offer a streamlined mechanism enabling researchers to trace, index, or effortlessly compare results across multiple analyses conducted at disparate times. Our tool incorporates two distinct levels of signatures: the *connection-level* and *node-level* signatures, grounded in standard Universal

Unique Identifiers (UUID) of version 3 hashes. These hashes are derived by compressing the most pivotal data elements that constitute *Nodes* (comprising Genbank accession, source genome, gene name, and metadata keys and values) and *Connections* (encompassing identifiers and node signatures). Such an approach empowers users with the capability to replicate results and swiftly compare records when necessary.

Metadata columns are composed of the MIG keys concatenated to metadata keys (as example SPECIMEN.isolate). Such way turn the further integration and indexing as a simple and natural way to store GeneConnector results. In addition to the above cited tabular file, the GeneConnector results includes at default a JSON¹ formatted output file as a optimal format to be inputs into ETL² pipelines and web integrations.

IV. RESULTS

A. MIG's Representativity and Distribution

Analysis of the complete GOPHY's database resulted in 414 events of information gain³ from the total of 1,246 specimen records. These amount comprises 33% of the database records suffering information gains. Gains ranged from 2% up to 60%, widely distributed along all fungal genus included in our analysis. Twenty-five of the twenty-nine genera present in GOPHY were contemplated with information gains. The complete tabular results is available as a supplementary material.

The most important MIG obviously was SPECIMEN, with *strain* and *culture_collection* as the most populated keys, with 86.3% and 69.3% of coverage. It was not surprisingly due to the nature of the GOPHY database proposal itself, including only high quality records, mainly belonging to type materials.

Next, the TAXONOMY MIG with *type_material* as the second⁴ most important key, with >69% of coverage in records. Both SPECIMEN and TAXONOMY are the most important keys to make each Connection record unique. And precisely for this reason that both are the best scored in our tool (see Table I).

The third most important MIG for GeneConnector approach is HOST_SUBSTRACT. Both keys *host* and *isolation_source* covered approximately 62% and 40%, respectively, of the full GOPHY dataset.

The next important MIG's is GEO_REFERENCES. It was present in about 76% of records, however in the most cases refereed to only as country. This key in most cases is not so geographically resolute when dealing with countries of continental dimensions, such as Brazil or Australia. From 1,246 records, just one included information of Latitude/Longitude, completely inhibiting the performance of geographic analyzes.

A world scale map indicating the geographic range of the records, and including the maximum information gain reachable by country is shown in Figure 2. The 10 countries covered with the highest number of information gain events

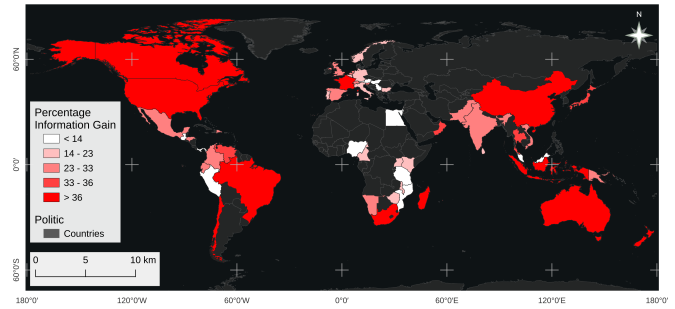


Fig. 2. Information gain at the globe scale. Records with some information gain registered in this study are highlighted in white-to-red scale (see scale legend).

were China, United States, Australia, South Africa, Brazil, Thailand, Netherlands, Indonesia, Japan, and Ecuador. The maximum information gain reached by such countries ranges between 30% and 50%. This is proportional to the country contribution to the state of the art of the phytopathogenic fungi records, a fully expected scenario.

Different from the previous cited MIG's the TIME_REFERENCE was an exception. Only about 6.9% (86 records) of the GOPHY database contains time milestones. Despite such MIG is not highly scored in GeneConnector (score = 2), the absence of this information inhibits temporal interpretation of the collection effort on the phytopathogenic important fungi around the world.

B. Phytopathogenic Completeness Along GOPHY Genus

Information gains by genus are shown in Figure 3. As above cited, information gains ranged from 2% up to 60%. The top ten phytopathogenic fungal genus with most number of specimen records suffering information gains were *Calonectria* with 73 events, *Diaporthe* (55), *Curvularia* (48), *Colletotrichum* (41), *Ceratocystis* (23), *Bipolaris* (21), *Boeremia* (19), *Neofusicoccum* (17), *Phyllosticta* and *Huntliella* with (16).

Using our approach 8 of 25 genus with information gains (GOPHY database include information of 29 genus) reached the full information completeness (100% of completeness, RCS = 1.0), grouping at last one of each MIG qualifier key per connection. A significant information gain in terms of the complete database. As can be seen in Fig. 3, median values of RCS were up to 90% in nine of ten most representative genus of GOPHY (cited in the previous paragraph).

V. CONCLUSIONS

In this study, we showcase the remarkable ability of GeneConnector to substantially enhance the data completeness of specimens in Genbank by exclusively leveraging shared information within the records. Our findings demonstrate that utilizing our tool can yield gains of up to 60% in shared information among Genbank records, particularly for specific phytopathogenic genera. Furthermore, on a global scale, the data aggregation process holds the potential to benefit records from approximately 55 countries across the globe.

¹Javascript Object Notation format.

²Extract, Transform, and Load pipelines.

³Calculated as the percentage of the Reachable Completeness Score which the Observed Completeness Score comprises.

⁴Organism is a required field, so it has full coverage in Genbank nucleotide records.

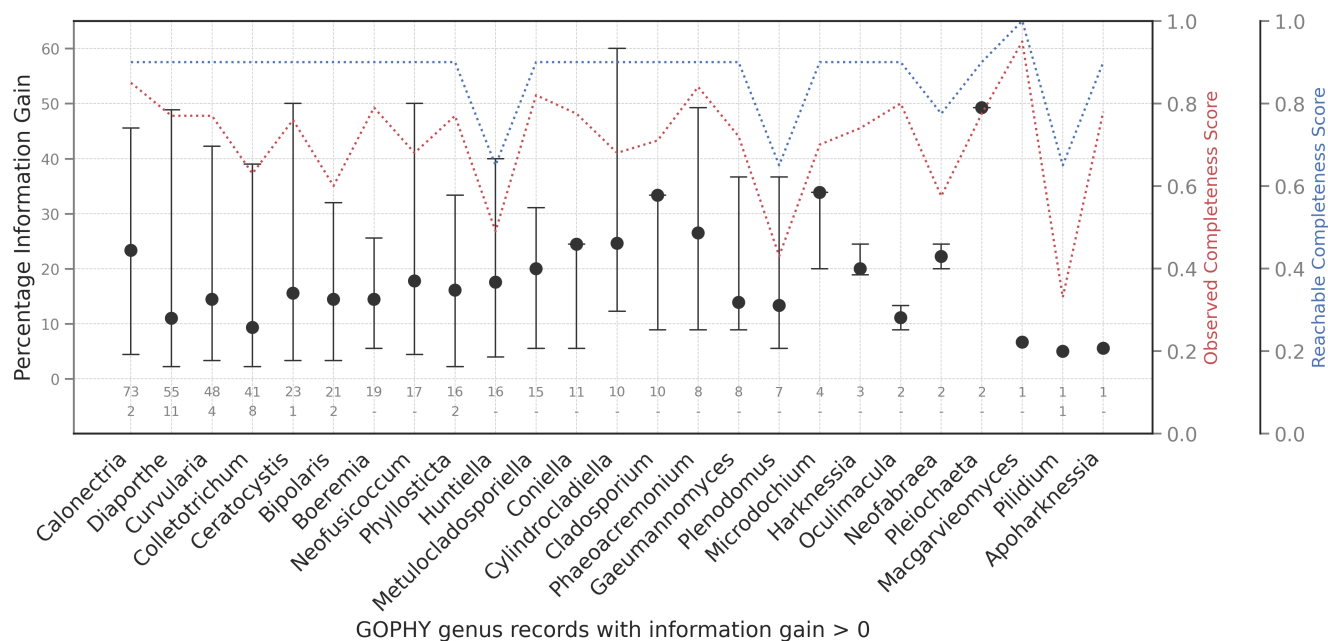


Fig. 3. Information gain by genus of phytopathogenic fungi registered in GOPHY database. Median with max/min values are presented in first Y-axis (left). Median values of Observed Completeness Scores and Reachable Completeness Scores are shown in 2nd and 3rd Y-axis (right), respectively. Only records with information gains greater zero were kept in chart. Numbers below zero in the X-axis indicates the number of records evaluated for each genus (upper number), and the number of record reaching the maximum reachable completeness (100%, lower number).

Moreover, our data aggregation process is both auditable and interpretable through two scores: the Observed Completeness Score (OCS) and the Reachable Completeness Score (RCS). These scores provide insights into the current level of information completeness and the attainable information based on shared metadata among nodes of the same specimen in Genbank.

With these comprehensive metrics, our aim is to make a significant contribution to the ongoing improvement of the information accumulation process, benefiting scientists worldwide and fostering continuous advancements in knowledge acquisition.

ACKNOWLEDGES

Thanks to the grant from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES). We also thank the anonymous reviewers who dedicated their valuable time to improving this work. Thanks to Biotrop, Solutions in Biological Technologies, for the support during the development of this work.

REFERENCES

- [1] J. Shendure, G. M. Findlay, and M. W. Snyder, “Genomic medicine—progress, pitfalls, and promise,” *Cell*, vol. 177, no. 1, pp. 45–57, 2019.
- [2] R. Jeyasri, P. Muthuramalingam, L. Satish, S. K. Pandian, J.-T. Chen, S. Ahmar, X. Wang, F. Mora-Poblete, and M. Ramesh, “An overview of abiotic stress in cereal crops: negative impacts, regulation, biotechnology and integrated omics,” *Plants*, vol. 10, no. 7, p. 1472, 2021.
- [3] C. Juma, *The gene hunters: Biotechnology and the scramble for seeds*, vol. 996. Princeton University Press, 2014.
- [4] E. J. Gilchrist, S. Wang, and T. D. Quilichini, “The impact of biotechnology and genomics on an ancient crop: *Cannabis sativa*,” in *Genomics and the Global Bioeconomy*, pp. 177–204, Elsevier, 2023.
- [5] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, “Genbank,” *Nucleic acids research*, vol. 41, no. D1, pp. D36–D42, 2012.
- [6] K. L. Howe, P. Achuthan, J. Allen, J. Allen, J. Alvarez-Jarreta, M. R. Amodè, I. M. Armean, A. G. Azov, R. Bennett, J. Bhai, *et al.*, “Ensembl 2021,” *Nucleic acids research*, vol. 49, no. D1, pp. D884–D891, 2021.
- [7] U. Consortium, “Uniprot: a worldwide hub of protein knowledge,” *Nucleic acids research*, vol. 47, no. D1, pp. D506–D515, 2019.
- [8] D. Smedley, S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson, and A. Kasprzyk, “Biomart—biological queries made easy,” *BMC genomics*, vol. 10, no. 1, pp. 1–12, 2009.
- [9] M. Van Wyk, B. D. Wingfield, A. O. Al-Adawi, C. J. Rossetto, M. F. Ito, and M. J. Wingfield, “Two new ceratocystis species associated with mango disease in brazil,” *Mycotaxon*, vol. 117, no. 1, pp. 381–404, 2011.
- [10] M. Van Wyk, A. Al-Adawi, B. Wingfield, A. Al-Subhi, M. Deadman, and M. Wingfield, “Dna based characterization of ceratocystis fimbriata isolates associated with mango decline in oman,” *Australasian Plant Pathology*, vol. 34, pp. 587–590, 2005.
- [11] M. Van Wyk, A. O. Al Adawi, I. A. Khan, M. L. Deadman, A. A. Al Jahwari, B. D. Wingfield, R. Ploetz, and M. J. Wingfield, “Ceratocystis manginecans sp. nov., causal agent of a destructive mango wilt disease in oman and pakistan,” *Fungal Divers*, vol. 27, pp. 213–230, 2007.
- [12] A. Fourie, M. J. Wingfield, B. D. Wingfield, and I. Barnes, “Molecular markers delimit cryptic species in ceratocystis sensu stricto,” *Mycological Progress*, vol. 14, pp. 1–18, 2015.
- [13] A. Canakoglu, A. Bernasconi, A. Colombo, M. Masseroli, and S. Ceri, “Genosurf: metadata driven semantic search system for integrated genomic datasets,” *Database*, vol. 2019, 2019.
- [14] Z. Chen, A. S. Azman, X. Chen, J. Zou, Y. Tian, R. Sun, X. Xu, Y. Wu, W. Lu, S. Ge, *et al.*, “Global landscape of sars-cov-2 genomic surveillance and data sharing,” *Nature genetics*, vol. 54, no. 4, pp. 499–507, 2022.
- [15] U. Köljalg, K.-H. Larsson, K. Abarenkov, R. H. Nilsson, I. J. Alexander, U. Eberhardt, S. Erland, K. Høiland, R. Kjølner, E. Larsson, *et al.*, “Unite: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi,” *New Phytologist*, vol. 166, no. 3, pp. 1063–1068, 2005.
- [16] K. Abarenkov, R. H. Nilsson, K.-H. Larsson, I. J. Alexander, U. Eberhardt, S. Erland, K. Høiland, R. Kjølner, E. Larsson, T. Pennanen, *et al.*,

"The unite database for molecular identification of fungi—recent updates and future perspectives," *The New Phytologist*, vol. 186, no. 2, pp. 281–285, 2010.

- [17] M. Quñones, D. T. Liou, C. Shyu, W. Kim, I. Vujkovic-Cvijin, Y. Belkaid, and D. E. Hurt, "Metagenote: a simplified web platform for metadata annotation of genomic samples and streamlined submission to ncbi's sequence read archive," *BMC bioinformatics*, vol. 21, pp. 1–12, 2020.
- [18] Á. Gálvez-Merchán, K. H. Min, L. Pachter, and A. S. Booesaghhi, "Metadata retrieval from sequence databases with fq," *Bioinformatics*, vol. 39, no. 1, p. btac667, 2023.
- [19] Q. Chen, J. Zobel, and K. Verspoor, "Duplicates, redundancies and inconsistencies in the primary nucleotide databases: a descriptive study," *Database*, vol. 2017, 2017.
- [20] S. Reining, F. Ahlemann, B. Mueller, and R. Thakurta, "Knowledge accumulation in design science research: ways to foster scientific progress," *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, vol. 53, no. 1, pp. 10–24, 2022.
- [21] Y. Marin-Felix, J. Groenewald, L. Cai, Q. Chen, S. Marinowitz, I. Barnes, K. Bensch, U. Braun, E. Camporesi, U. Damm, *et al.*, "Genera of phytopathogenic fungi: Gophy 1," *Studies in mycology*, vol. 86, pp. 99–216, 2017.
- [22] Y. Marin-Felix, M. Hernández-Restrepo, M. J. Wingfield, A. Akulov, A. Carnegie, R. Cheewangkoon, D. Gramaje, J. Z. Groenewald, V. Guaraccia, F. Halleen, *et al.*, "Genera of phytopathogenic fungi: Gophy 2," *Studies in mycology*, vol. 92, pp. 47–133, 2019.
- [23] Y. Marin-Felix, M. Hernández-Restrepo, I. Iturrieta-González, D. García, J. Gené, J. Z. Groenewald, L. Cai, Q. Chen, W. Quaedvlieg, R. Schumacher, *et al.*, "Genera of phytopathogenic fungi: Gophy 3," *Studies in mycology*, vol. 94, pp. 1–124, 2019.
- [24] Q. Chen, M. Bakhshi, Y. Balci, K. Broders, R. Cheewangkoon, S. Chen, X. Fan, D. Gramaje, F. Halleen, M. Horta Jung, *et al.*, "Genera of phytopathogenic fungi: Gophy 4," *Studies in Mycology*, vol. 101, no. 1, pp. 417–564, 2022.
- [25] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.
- [26] A. Cockburn, "Ports and adapters architecture," 2006. <http://wiki.c2.com/?PortsAndAdaptersArchitecture> [Accessed: 2022-11-20].
- [27] D. Merkel, "Docker: lightweight linux containers for consistent development and deployment," *Linux journal*, vol. 2014, no. 239, p. 2, 2014.



Samuel Galvão Elias I am a biologist/microbiologist and I value multi- and interdisciplinary approaches. As a biologist, my main focus is on mycology, and I have a strong knowledge of bacteriology as well. As a bioinformatician, I have experience in analyzing molecular data of various types. I also have expertise in analyzing microbial diversity, including community experimentation across a wide range of taxonomic groups. Additionally, I have extensive knowledge in molecular phylogenetics of eukaryotes and prokaryotic groups, along with ex-

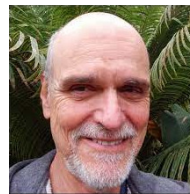
perience in post-phylogenetics. In terms of Data Science, I have expertise in analyzing diverse classes of data, including univariate to multivariate, unifactorial to multifactorial, and categorical to continuous data.

As a developer, I have experience in web application development (monolithic and distributed), embedded systems, desktop applications, and data pipelines both within and outside the field of bioinformatics. My language stack includes Python, R, Rust, Golang, JavaScript, and TypeScript, with experience in single and multithreaded development, single and multicore programming, concurrent programming, and parallel programming. I am involved in the architecture and development of stateful and stateless applications, native to cloud environments, with a focus on Kubernetes. Some of my main open-source projects include Mycelium (Rust), an API gateway currently under development that focuses on permissioning in distributed environments, and Blutils (Rust), a tool for optimizing the execution process and analysis of Blast results.



Débora Cervieri Guterres holds a Ph.D. in Phytopathology from the University of Brasília (UnB, 2018), a Master's degree in Environmental Sciences from the Federal University of Bahia (UFBA, 2013), and specializes in Environmental Management from FJC (2009). Additionally, earned a degree in Agronomist Engineering from FASB (2012) and holds a Bachelor's degree in Business Administration with a focus on Foreign Trade from FASB (2007).

With a diverse academic background, Debora has expertise in the fields of Phytopathology, Mycology, Etiology, and the Diversity and Taxonomy of Fungi. She has conducted extensive research in these areas, contributing to the understanding and management of plant diseases. Currently, Debora is engaged in a postdoctoral internship at the Federal University of Viçosa, further expanding her knowledge and expertise in the field.



Robert Weingart Barreto is an agronomist (UFRRJ) with a strong academic background in mycology. He obtained a MSc in Pure and Applied Taxonomy (Mycology) from the University of Reading in 1986 and went on to complete his Ph.D. in Botany (Mycology) at the same institution, along with the International Institute of Mycology (currently CAB International) in 1991. Following his doctoral studies, he pursued postdoctoral research focusing on molecular taxonomy of fungi at the esteemed Centraalbureau voor Schimmelcultures.

Currently, he holds the position of full professor in the Department of Phytopathology at the Federal University of Viçosa, where he is actively involved in teaching various courses in the field of mycology, plant disease diagnosis, and biological control. Since its establishment in 1998, he has been the dedicated Coordinator of the Plant Disease Clinic - DFP/UFV, ensuring effective management and treatment of plant diseases.

With extensive experience in mycology, his research interests encompass a wide range of topics. His expertise lies in the areas of biological control of weeds, fungal taxonomy, phytopathology, diagnosis of fungal diseases in plants, and the study of fungal biodiversity in Brazilian ecosystems. As an accomplished researcher, he has made significant contributions to these fields and is recognized as a leading figure in the discipline.

Furthermore, he holds the esteemed position of the current president of the Brazilian Society of Mycology, where he actively promotes collaboration and advances in mycological research. Through his leadership, he plays a pivotal role in shaping the direction of mycology in Brazil and fostering connections within the scientific community.



Helson Mário Martins do Vale holds a degree in Agricultural Sciences from the Federal Rural University of Rio de Janeiro (2002), a master's degree in Agricultural Microbiology from the Federal University of Lavras (2005) and a PhD in Agricultural Microbiology from the Federal University of Viçosa (2009), post-doctorate in metagenomics of endophytic fungi at the Ruhr-Universität Bochum, Germany. He is currently Associate Professor D, Level II at the University of Brasília (UnB) and Head of the Department of Phytopathology. He

works in undergraduate disciplines of Agronomy courses (Microbiology and Phytopathogenic Micro-organisms); Biology (Mycology) and Environmental Sciences (Microbial Diversity and Biological Collections) and postgraduate disciplines in the Phytopathology (Molecular Techniques) and Microbial Biology (Microbial Ecology) courses at UnB. He has experience in the area of Agronomy and Biology, with emphasis on Agricultural Microbiology, working mainly on the following topics: Biological Nitrogen Fixation, Microbial Ecology, Metagenomics, Next Generation Sequencing (NGS), Yeast Diversity in Brazilian Ecosystems, Molecular Diversity and Characterization of Epiphytic and Endophytic Microorganisms.

APÊNDICE B – Manuscrito referente ao Capítulo 5

CLASSEQ: A clade-informed sequence identification tool. Leveraging phylogenetic insights for biological sequences classification

Samuel Galvão Elias^{1,2,✉}, Debora Cervieri Guterres³, and Helson Mário Martins do Vale¹

¹Universidade de Brasília, UnB, Brasília, Brazil

²Biotrop, Solutions in Biological Technologies

³Universidade Federal de Viçosa, Viçosa, MG, Brazil

Molecular phylogenetics is a powerful tool for the classification of biological sequences benefited by distinct knowledge fields as epidemiology, conservation biology, comparative biology, pharmacogenomics, forensics, and agriculture. Such an importance converged to a prominent new realm of classifiers, the phylogeny-based placers, as tools designed to place sequences into existing phylogenies, along the large-scale classification of sequences with optimized performance when compared to traditional tools. Here, we present a novel clade-informed, alignment-free, and lightweight approach developed to classify gene-wide sequences. Our tool was tested to classify the *Bacillus subtilis* group RefSeq database, resulting in a highly sensitive and specific placer, capable of classifying almost all the sequences of the group into the correct clades. In addition to the traditional command line interface (CLI) available in bioinformatic tools, *classeq* provide an additional application programming interface allowing users to serve our classifier directly on the web, making the integration process of our algorithm with web-based tools easy.

Phylogeny | *Bacillus subtilis* group | Phylogenetic placement | Alignment-free

Correspondence: sgelias@outlook.com and samuel.elias@biotrop.com.br

Introduction

Molecular phylogenetics plays a crucial role in biology, examining the evolutionary connections between different biological species through their genetic traits. By scrutinizing genetic information such as DNA, RNA, and proteins, scientists can build phylogenetic trees or networks to illustrate these connections. Innumerable fields can benefit from phylogenetics, including the most common taxonomy classification (1), and have already been applied to epidemiology (2), conservation biology (3), comparative biology (4), pharmacogenomics (5), forensics (6), and agriculture [see (7) for genomic application and (8–11) for examples of exploratory applications].

In this way, the phylogenetic placement of metagenomic sequences was an emerging field of bioinformatics just over a decade ago (12). In contrast to the traditional approach used by metagenomics¹, phylogeny-based placement meth-

ods work around the comparison of query sequences² with internal or terminal nodes of previously existing trees (23–32). Unlike taxonomy-based placers, phylogenetic counterparts should provide highly accurate and rich in information results (23, 33–36), allowing users to interpret the target diversity with additional complexity levels that include a most natural evolutionary past of the specimens (37–39).

Such an approach is highly advantageous for being 'almost' independent of the dynamicality of the taxonomic process, becoming immune to challenges intrinsic to taxonomic methods [see (40–42) for examples of the database-derived errors impacting annotations, and the Taxallnomy (43) and SATIVA (44) proposals to reduce the impact of such challenges].

It is important to note that the phylogenetic placement of metagenomic sequences is a different process from the traditional methods applied to infer the evolutionary history of a taxa. During a traditional placement workflow, the target sequence, or the query sequence (QS), is compared to all other reference sequences, and the full history of the QS is inferred using heuristic or exhaustive search [see (45–47)]. If multiple QS are included on the analysis, the relationships among this are also inferred.

Now, during the metagenome placement, the algorithms used for this purpose work to place QS into a reference tree (RT) by comparing sequences against the tree nodes independently from other QS. In other words, the relationship among the QS is not resolved. This process should be executed with or without solving multiple sequence alignments [see review of Czech et al. 12]. This approach can significantly reduce the computation effort for high-dimensional data.

Phylogenetic placement methods commonly used for metagenomics adopt mostly two placement strategies, the maximum likelihood (ML) measurement, or the evolution distance calculation. Two of the most important tools (and the first developed for this purpose), PPLACER (24) and RAXML-EPA (25) work around the ML framework. More recently, EPA-NG (27) was developed to combine the accuracy and performance of the two previous algorithms, which were already based on the ML framework. As an alternative to ML-based methods, a new phylo-k-mer-based method was

above).

²Sequences derived from environmental samples or directly extracted from the target specimen.

¹Sequences of a given sample are identified by their direct comparison with reference sequences with further attribution of their taxonomy based on similarity (or not) with reference sequences (see (13–17) for example tools using such an approach and (18–22) for frameworks using the tools cited

introduced with the RAPPAS (28) algorithm and its successor EPIK (48). All the above-cited software depends on the previous alignment of the query sequences to perform predictions, which leads to a significant increase in computational effort upstream the prediction phase and should be highly sensitive to high-divergent genome regions with no homologous positions as noncoding portions of the DNA [as an example, the Internal Transcribed Spacer used as an important universal fungal barcoding (49)].

Alongside the rise in the throughput of modern DNA sequencers, there has been a concurrent demand for placement methods that are not reliant on the multiple-sequence alignment techniques used in the preliminary stages of phylogenetic placement. In this way, new alignment-free methods were introduced with APPLES (29) and APP-SPAM (30). Both perform distance-based placements calculated from the sequence k-mers, substantially increasing the placement performance when compared to the ML-based or phylo-k-mer-based alternative tools.

Still in APPLES, despite presenting itself as alignment-free (in the case of using the alignment-free option), the tool depends on a distance matrix as an input parameter, a solution that obviously suffers when non-equal size DNA sequences are compared. However, the solution APP-SPAM was developed primarily to work with short read sequences, being inadequate during the placement of sequences based on complete gene sequences.

Here, we present a novel clade-informed, alignment-free approach developed and tested to predict based on complete gene sequence trees. In the *first* section, we will present and discuss the *classeq* software structure and algorithm, introducing the data formats used during the indexation and prediction phases; during the *second* section we attempt to provide a proof of concept of our algorithm to classify sequences of *Bacillus subtilis* group RefSeq database.

Software Usage, Structure, and Algorithm

Usage Summary. Similarly to other classifiers the *classeq* pipeline flows through two stages: the initial *indexation* and the *prediction* (see Figure 1). *Indexation* is performed based on nucleotide sequences (as a FASTA format), the same used during the phylogeny reconstruction, and the phylogeny itself (as a NEWICK format). The follow *prediction* stage is performed based on *classeq* tarball artifact generated during the indexation phase and uses simple FASTA sequences as a query input. Both stages are described in the next sections.

Building the Phylogenetic Indices. As expected from a phylogeny-dependent classification tool, the main input of *classeq* during the indexation phase is a NEWICK-formatted phylogeny. Initially, the target phylogeny suffered re-rooting and sanitizing to remove low supported branches³. The re-root process is executed using the Environment for Tree Ex-

ploration, version 3 package (ETE3, (50))⁴. After the tree sanitization, branches with low phylogenetic support values are pruned and reconnected with the closely related parent node with sufficient support to be kept in phylogeny. In case the parent branch has no sufficient support, the search proceeds until the next supported branch is found (see Figures 1 A-B). The pruning process and all custom actions related to tree management are executed using extensions of the default Biopython (52) classes implemented as *classeq* elements.

During sequences sanitization, the nucleotide residuals are cleaned to remove ambiguous characters⁵ with the remaining ones being used to extract their k-mer contents. As default *classeq* take k-mers of size twelve, but this value should be changed during the index generation step (Figure 1C-D).

Next, k-mers serve as input to build the *k-mers inverse indices* (KII), a conceptual hash map containing k-mers as hash values, and an array containing the phylogeny leafs which the k-mers were mapped as values (Figure 1D). The KII comprises an intermediary object that speeds up the indexation phase. Next the k-mers content is mapped to the phylogeny internal nodes, originating the *Nodes to K-mers Index* (NKI in Figure 1E), a resulting object scheme is defined in Listing 1 (begin line 1), where the *nodePrior* type represents the phylogeny internal/external nodes. Each *nodePrior* carries information on all combinations of INGROUP priors (a set of k-mers that match the target clade) and their SISTER clade priors (a set of k-mers shared between all clades not the INGROUP). Both ingroup and sister priors are used during predictions (explained in section).

At the end of the indexation process, the serialized KII plus the NKI object are available to users as a self-contained tarball artifact (Figure 1F) used during *predictions*. Such objects should be stored, retrieved, and shared among users to perform independent predictions.

The sanitized phylogeny (NEWICK) together with the sanitized sequences file (FASTA) persists as independent files, allowing users to audit the intermediary source results after indexation. Additionally, *classeq* produces an annotation artifact in the PHYLO.JSON format that contains the sanitized phylogeny. These artifacts are a web-friendly object and should be used during predictions by the CLI⁶ argument. The goal of the annotation artifact (as the name suggests) is to allow one to annotate individual nodes with an arbitrary name, their corresponding Genbank TaxID, and a related taxonomic rank. Annotations should be propagated up to the output artifact, increasing the user readability. The annotation artifacts are given in the schema of the Listing 1 (begin line 36).

```
1 ———
2 type: object # Priors artifact
3 properties:
4   ingroup:
5     type: array
6     items:
7     $ref: '#/definitions/nodePrior'
```

⁴We opted to specifically execute the re-root process with ETE3 due to the previously related inconsistency found in Biopython on execute tree managements (51).

⁵Characters not in default DNA/RNA alphabet as A, T(U), C, and G.

⁶Command Line Interface.

³Branches with low phylogenetic support. *Classeq* uses 95 as a default cutoff value, but this value can be changed through command line arguments at runtime.

```

8 definitions:
9   nodePrior:
10     type: object
11     properties:
12       parent:
13         type: string
14         format: uuid
15       clade_priors:
16         type: array
17         items:
18           $ref: '#/definitions/prior'
19   prior:
20     type: object
21     properties:
22       group:
23         type: string
24         enum: [INGROUP, SISTER]
25       kmers:
26         type: array
27         items:
28           type: string
29       labels:
30         type: array
31         items:
32           type: integer
33
34 type: object # PHYLO.JSON Annotations artifact
35 properties:
36   id:
37     $ref: '#/definitions/nullString'
38   name:
39     $ref: '#/definitions/nullString'
40   rooted:
41     type: boolean
42   root:
43     $ref: '#/definitions/clade'
44 definitions:
45   nullString:
46     type: [string, null]
47   nullNumber:
48     type: [number, null]
49     minimum: 0
50   nullInteger:
51     type: [integer, null]
52     minimum: 1
53   clade:
54     type: object
55     properties:
56       id:
57         type: string
58         format: uuid
59       name:
60         $ref: '#/definitions/nullString'
61       confidence:
62         $ref: '#/definitions/nullNumber'
63       branch_length:
64         $ref: '#/definitions/nullNumber'
65       color:
66         $ref: '#/definitions/nullString'
67       width:
68         $ref: '#/definitions/nullNumber'
69       taxid:
70         $ref: '#/definitions/nullInteger'
71       related_rank:
72         $ref: '#/definitions/nullString'
73       clades:
74         $ref: '#/definitions/cladeArray'
75   cladeArray:
76     type: array
77     items:

```

```

78     $ref: '#/definitions/clade'

```

Listing 1. Data schema of Priors and Annotation artifacts. Some double quotes were omitted to turn the reading better. The code snippets are present as a standard YAML format, but when needed users should convert to JSON schema using default web formatters from the original schema definitions files available on project Github documentation.

For basic/non-developer users, we provide a GUI⁷ interface available on execution of the CLI command *serve* (see code snippet in Listing 2), allowing users to annotate the internal nodes of the phylogeny without dealing with JSON specifications. For advanced/developer users, the annotation artifact can be easily parsed and edited using JavaScript interpreters like browser applications, or even manually edited if necessary, as long as they follow the standard schema of the Listing 1 (begin line 36).

Certain attributes within the schema require further consideration. Notably, the clade Identifier (ID, line 39) is a Universal Unique Identifier of version 3, which bears a nonstandard namespace derived from a nine sequence (e.g., 99999999-9999-9999-9999-999999999999), which concatenates the clade name to the names of all its terminal children clades. Our approach permits multiple runs to converge in the same UUID's since their use the same clades configuration, facilitating future sharing and comparison of independent run annotations.

```
$ cls serve -t *.phylo.json
```

Listing 2. Classeq Annotation Server startup example command.

Prediction. The prediction workflow is shown in Figure 2. The final users can make predictions using the CLI command *predict* with appropriate arguments or through the API⁸ port, which is readily available, allowing them to deploy their own classeq server on either local or public networks, thus facilitating predictions. The API option is not fully recommended for users with high performance interest because it is not already optimized for the web purpose. For developers, the prediction functions are available as *classeq* module named *predict-taxonomies-recursively* (accessible through a [use-case](#)⁹ with the same name).

As explained in Figure 1 (prediction section), after initial conversion of the query sequences into k-mers (Figure 1G and Figure 2), our algorithm starts the recursive query for the candidate node to place a given query sequence following the tree root to the leaf direction (Figure 1H).

Starting from the tree root node¹⁰ our algorithm tests the adherence of the given query sequence to the target node (including the tree root) using simple k-mers matching statistics, where the number of matches between the query sequence/target clade is compared to the matches with all

⁷Graphical User Interface.

⁸Application Programming Interface.

⁹A use-case represents the piece of intention of the software, often represented as a module, which users and developers can easily understand their particular goal within the software's tiers and layers.

¹⁰As previously explained, the input phylogeny is re-rooted at middle-point as default and the root node serves as the entry-point for the classification flow.

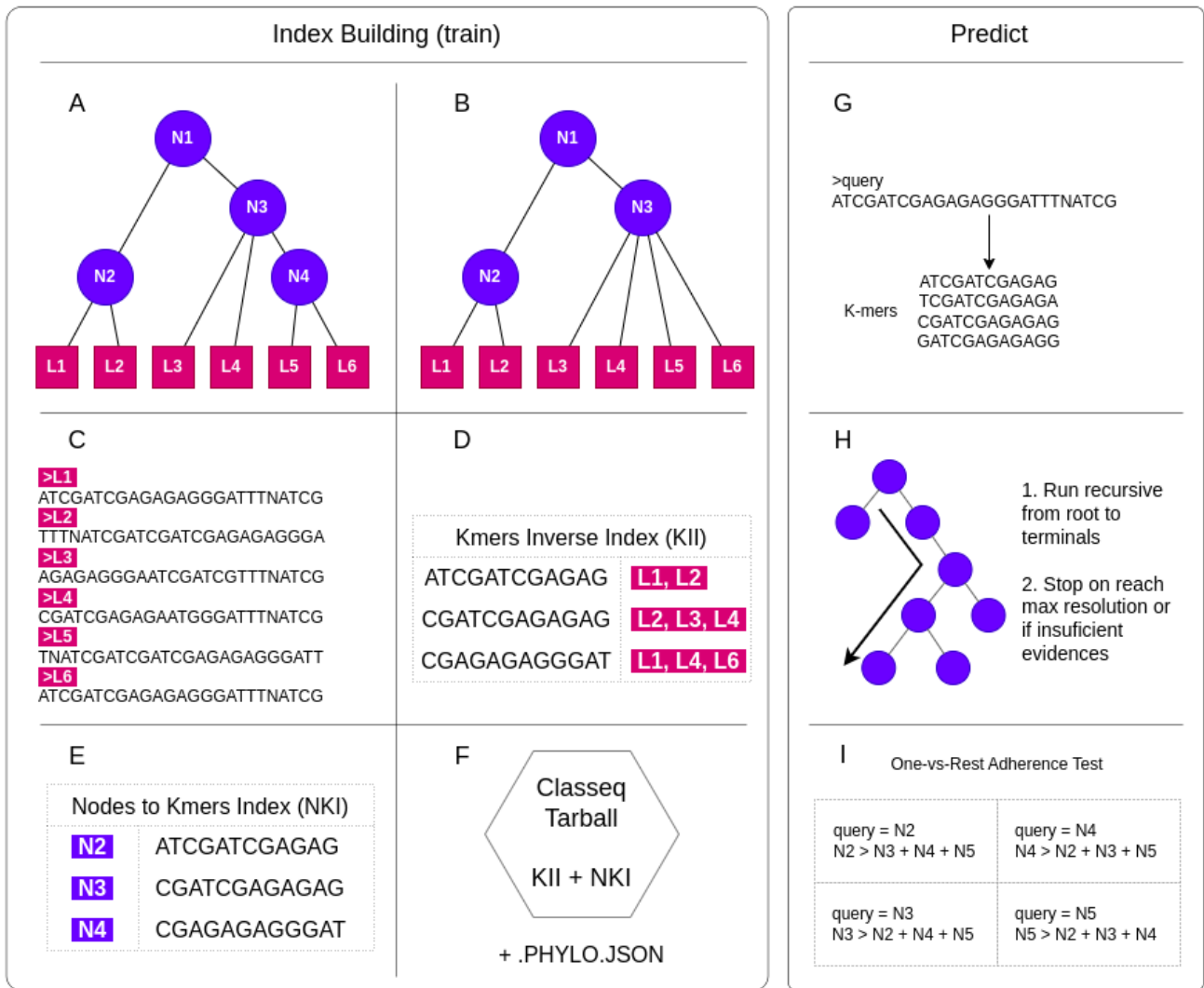


Fig. 1. Detailed description.

sibling ones together. An strategy also known as *One-vs-Rest* (OvR, Figure 1I, and Figure 2B). At this phase, when a desired node is detected as a candidate, all siblings are discarded, and the search continues to the children's nodes, increasing the search depth. In case the current node includes children ones, AND the OvR comparison is unable to identify a candidate node, the search process stops with a young return including the path crossed along the phylogeny up to the current node plus the status code `MAX_RESOLUTION_REACHED`. Otherwise, in case the current node has just one candidate node and has no children, the status code should be `CONCLUSIVE_INGROUP` instead. If more than one candidate was found, the current node should be ignored and the returned status should be `INCONCLUSIVE`.

Exceptionally, in case during the first step of the search algorithm (depth 1) the k-mer match coverage between the query and the candidate node is lower than 50%¹¹, the search suf-

¹¹The match coverage is controlled with the argument

fering from young return with `INCONCLUSIVE` status. This rule prevents our algorithm from producing false positive predictions, thus increasing the specificity of each trained model.

Our choice for the OvR strategy already mirrors the conservative nature of the *classeq* search algorithm, which prevents greedy behaviors. The use of the OvR strategy means that a node needs not to be "only" the best node in the clade to become a candidate, but the best node when compared to the sum of all the others. Conservative behavior is a convenient way to deal with incomplete sampled clades often verified in poorly explored taxonomic groups. This means that *classeq* favors the nonclassification of a desired sequence up to the lowest level as possible over their misplacement into a "wrong" clade.

`-matches-coverage` with default as 0.5 but ranging from 0.00001 to 1.

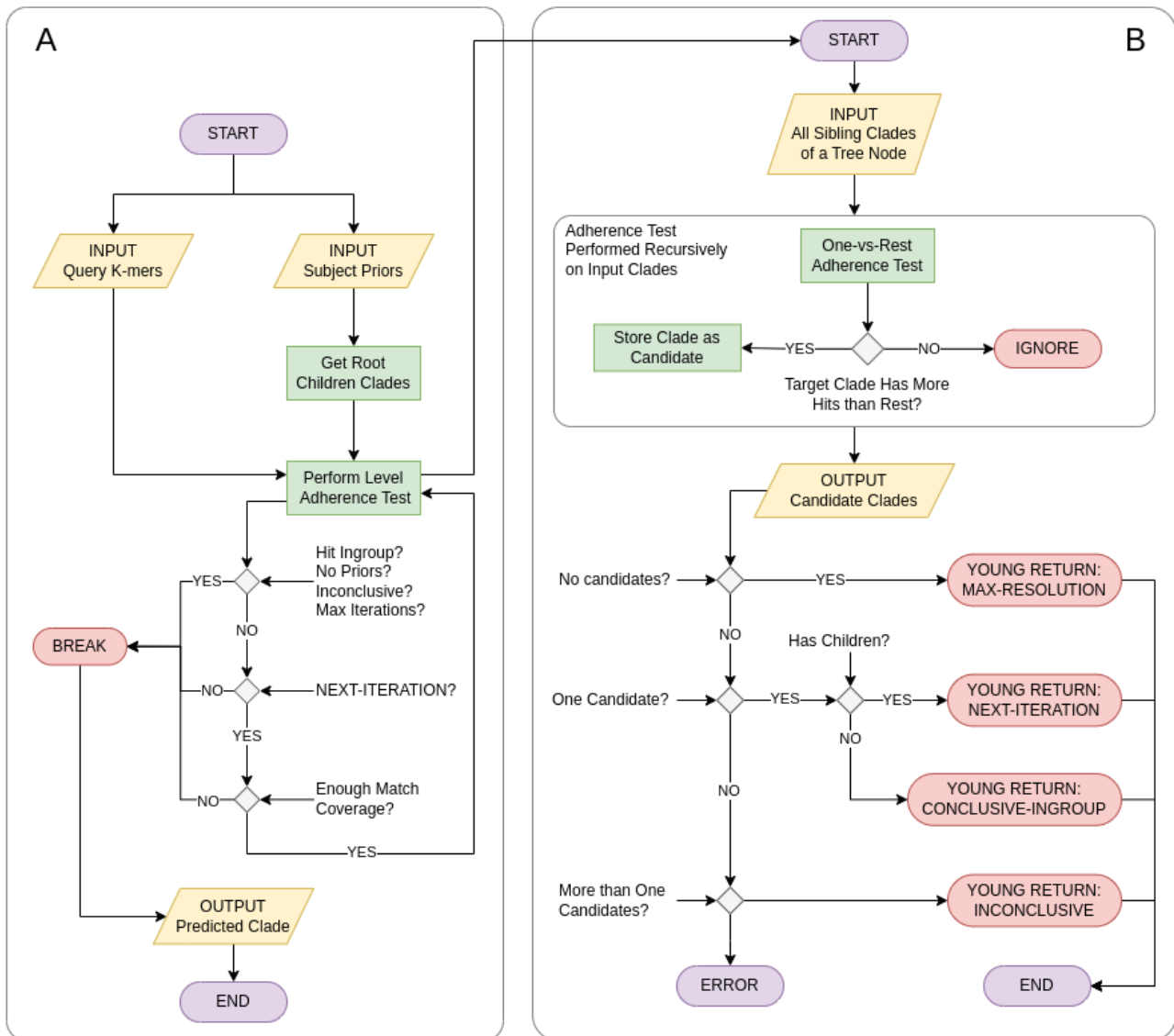


Fig. 2. Prediction algorithm.

A proof-of-concept: Classifying specimens of *Bacillus subtilis* group using the *gyrB* gene

To evaluate the performance of *classeq* on place real world sequences, we carried out a classification assay in an attempt to place specimens from the *Bacillus subtilis* group (53–55) in a comprehensive phylogeny of the group constructed from the *gyrB* gene (Figure 3A). The *gyrB*, which encodes the subunit B protein of the DNA gyrase gene, represents an alternative marker for the phylogenetic placement of the *B. subtilis* group specimens, providing mid/high resolution when compared with the 16S rRNA gene (53).

In performing the classification of specimens exclusively belonging to the *B. subtilis* group (from here called as Ingroup), we would only be able to test the conceptual *sensitivity* on the *classeq* results. Thus, aiming to be more precise on the test of the *classeq* performance, we carried out a systematic

sampling for the outgroup¹² specimens guided by the phylogenetic distance between the Ingroup and the independent outgroups. Our sampling considered three levels of divergence, including specimens from the closed related *Bacillus cereus* group (sharing the *Bacillus* [genus] as LCA¹³), followed by specimens of *Paenibacillus* (Bacilli [class] as LCA), and *Streptomyces* (Terrabacteria [clade] as LCA) as the most divergent group (see Figure 3B).

Finally, the confusion matrix of our assay was consolidated in Figure 3C, where both Ingroup and Outgroups classification metrics were used to calculate the *sensitivity* and *specificity* statistics (see (56) for a conceptual review).

RefSeq data acquisition. To guarantee the accuracy of the specimens identities included in our assay, both the train and

¹²From here all specimens not belonging to the *B. subtilis* group will be referred as outgroups.

¹³The Last Common Ancestor shared between studied taxa.

the test were performed on the basis of sequences from the Genbank RefSeq database. Initially, accessions of the complete RefSeq database of *B. subtilis* group, *B. cereus* group, *Paenibacillus* spp. and *Streptomyces* spp. were retrieved (through the Genbank web interface) and sampled (using the `shuf` Linux functionality) to balance the input dataset. The samples were then downloaded and sanitized using custom Python scripts. The *B. subtilis* Group complete dataset (containing 2139 genome sequences) was sampled with $\approx 50\%$ of records (resulting in 1040 genomes) and used as a training dataset for the phylogeny reconstruction (see phylogeny reconstruction section), where the remaining (1097 genomes) were used for testing purposes. Datasets of the remaining genus were sampled with a fixed size of 1000 records when available.

Following the sampling steps, the *gyrB* gene was extracted from the reference genomes using the DIAMOND algorithm (57) following default settings. We used [A0A6L7H8K2_BACAN](#) as a query sequence for gene extraction in the studied genomes. Genomes in which the reference gene did not match were discarded, resulting in a test dataset of 1000, 745, and 927 specimens of *B. cereus* group, *Paenibacillus* spp., and *Streptomyces* spp., respectively.

Reference phylogeny reconstruction. The sequences of the *B. subtilis* group sampled in the previous step were clustered at 99.9% of similarity using the VSEARCH cluster_fast strategy (58) reducing the number of records used during phylogenetic reconstruction. The clustering strategy reduces the noise of non-informative/repetitive sequences and significantly reduces the search space for the tree convergence. The final result is a lean and highly informative tree. The clustered dataset was composed of 293 records with sequences ranging from 1905 to 1911 bp (exception to record NZ_JALAQF010000173 [*Bacillus atrophaeus* strain CK3J3B4] reaching 1708 bp in length).

The resulting database was aligned using the Mafft algorithm (59) with default parameters and the strategy `ginsi`, resulting in an MSA¹⁴ containing 752 variables (39.5%) with 722 informative sites (37.9%). The base composition of MSA was 31.8% of adenine, 21.2% of cytosine, 24.8% of guanine, and 22.3% of thymine, with an average transition/transversion ratio of 1.4 (ranging from 0 to 25% between all sequence pairs).

Finally, the MSA was used into the phylogeny reconstruction using the software RAXML-HPC v8 (47), under the GTR-CAT substitution model, using the Rapid Bootstrap Algorithm (60) with 1000 pseudo-replications. The parsimony seed value was 12345. The default settings were kept for the remaining parameters.

The resulting tree is shown in Figure 3A where all the *B. subtilis* group were successfully recovered, including: *B. sonorensis* (61) (represented here by 2 specimens), *B. licheniformis* (62) (6 spp.), *B. paralicheniformis* (63) (14 spp.), *B. atrophaeus* (64) (25 spp.), *B. amyloliquefaciens* group (65) including *B. siamensis* (66) (7 spp.), *B. amyloliquefaciens*

(67) with two clades which includes *B. velezensis* (68) (109 spp. as the total). Additionally, *B. halotolerans* (69) and *B. mojavensis* (70) within the *B. mojavensis* group (71) (total 20 spp.), *B. tequilensis* (72) (2 spp.), *B. spizizenii* (73) (4 spp.), *B. vallismortis* (74) (6 spp.), *B. stercoris* (73) (4 spp.), *B. in-aquosorum* (73, 75) (23 spp.), and *B. subtilis stricto sensu* (76) (69 spp.).

Classeq predictions. With the phylogenetic reconstruction build on the previous step plus their MSA, we constructed the *classeq* indices (see section about phylogeny indexing) with default parameters for the k-mer size (`-kmer-size 12`) and strand (`-strad both`, taking into account k-mers generated with forward and reverse strands), and a non-default value for the nodes support cutoff parameter (`-support-value-cutoff 80`, default 95).

Next, independent predictions were performed on the target group and outgroups following the default parameters of the *classeq predict* command. A manual annotated phylogeny (such as PHYLO.JSON format) was used as a reference for the clades identity. Prediction times were ≈ 16 seconds for *B. subtilis* group ($n = 1097$), $\approx 4s$ *B. cereus* group ($n = 1000$), $\approx 3s$ *Paenibacillus* spp. ($n = 745$), and $\approx 4s$ *Streptomyces* ($n = 927$). The complete assay was performed on a Linux-based workstation equipped with a 12th Gen Intel® Core™ i5-12500H×16, and 16,0 GiB of RAM memory. All files related to model training, annotation, and prediction are available as [Support Files 1-4](#) in project [Github Repository](#).

To provide a reference time for the *classeq* results evaluation, we performed a Blast search with default parameters to verify the identity of *B. subtilis* group using the clustered MSA (as the same of *classeq* input, see phylogeny reconstruction section for details) as a subject. The execution time of the Blast search was $\approx 18s$ ($n = 1097$). The resulting identities were not shown due to the information incompatibility between the Blast and *Classeq* searches. We limit our discussion to inform readers that the Blast search matched among their top 50 results (when available) samples already present in the clades identified by *classeq*, being fully compatible in terms of the final results.

Furthermore, to determine the compatibility of the *classeq* results with a traditional method for phylogenetic reconstruction, we generated a comparative result by placing the test sequences into the same phylogeny of the training set, using a Maximum Likelihood search. To perform this task, the sequences database of the test set was split into files with up to 220 sequences (1097 records of the training set resulting in four alignments of 220 sequences plus one of 217 sequences) being each chunk re-aligned against the original train dataset. The techniques for MSA calculation and phylogeny reconstruction were the same used to build the train dataset, see details in phylogeny reconstruction section.

Surprisingly, the sequences of *Bacillus subtilis* (strain DE0224, NZ_VTQV01000062) and *Bacillus amyloliquefaciens* (strain SN781, NZ_JAGFMA01000006) were highly divergent from other MSA sequences and were discarded. Both records are currently [suppressed](#) from the RefSeq database. Furthermore, the sequence of *Bacillus sonorensis*

¹⁴Multiple Sequence Alignment.

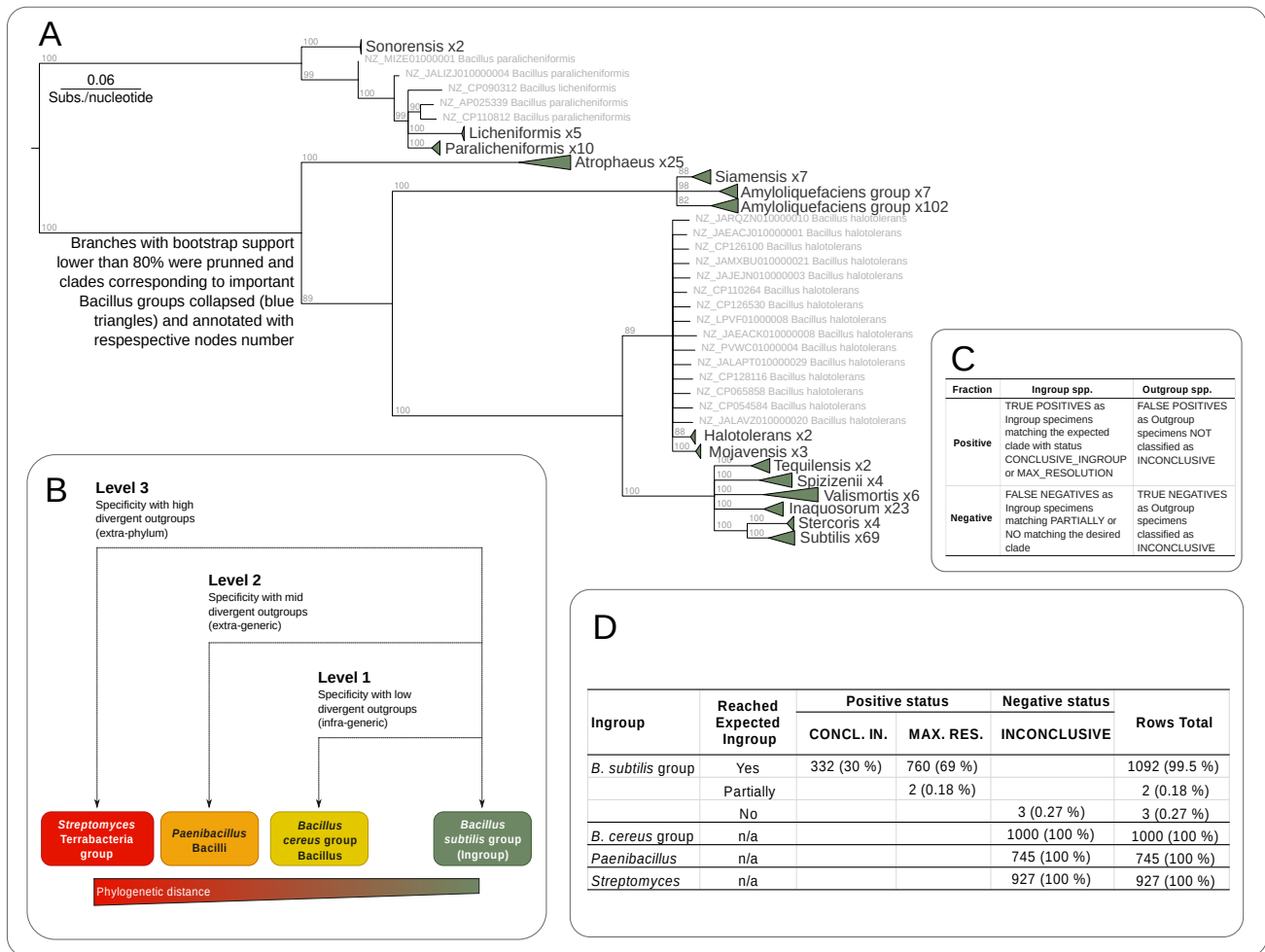


Fig. 3. Detailed description.

sis (strain MarseilleP3463, NZ_LT745775), was highly divergent from other *B. sonorensis* specimens, but it was kept in MSA because it was not considered enough divergent from another sequences from *B. subtilis* group, being a possible previous misplacement.

Discussed results. Our predictions with *classeq* on *B. subtilis* group specimens reached high values of *sensitivity* (>99.54%, see resolution of Equation 1) and *specificity* (100%, see Equation 2), where 1092 specimens from the 1097 *B. subtilis* group original specimens were correctly placed (see Figure 3D) on the expected clades, and all records of the outgroup correctly classified as INCONCLUSIVE¹⁵. The number of partially or not classified specimens summed was less than 1% (see Figure 3D for details).

$$TP/(TP + FN) = 1092/(1092 + 2 + 3) = 0.9954 \quad (1)$$

$$TN/(FP + TN) = 2672/(0 + 2672) = 1.0 \quad (2)$$

In the above equations, TP = True Positive results, where FN = False Negative, TN = True Negative, and FP, False Positives. Seeing the consolidation of Figure 3C, the TP values

¹⁵The default prediction status code when a sequence does not belongs to the training phylogeny.

include all correctly classified *B. subtilis* group specimens, where FN includes specimens partially or not classified; TN includes the count of the Outgroup specimens truly classified (1000 + 745 + 927), where FP the number of Outgroup specimens classified as *B. subtilis* group.

All placements provided by *classeq* algorithm over the query sequences were compatible with identities recovered by our validation using the traditional method of maximum likelihood for reconstruction of the phylogeny. The proof phylogenies together with all intermediary files should be found as [Support File 6](#) on the project [Github repository](#).

Our assay results show that in addition to the high values of *sensitivity* and *specificity* of the *classeq* predictions — up to the limit of our investigation — our algorithm suffers a poor or null influence of the clade size during predictions. Clades containing the minimum number of specimens to be considered a conceptual "clade" [has at least two terminals, see (77)] had the same prediction performance of highly dense ones, e.g. *Subtilis* and *Tequilensis* subclades, containing 69 and 2 specimens, respectively, and both with significantly lower error rates.

Conservative behavior represents an important aspect of the *classeq* algorithm. Unlike other algorithms that perform pre-

dictions based on a sparse *one-hot* representation of biological sequences using k-mers [e.g. (78–80)], our method is based on dense continuous vectors (DCV) as input that abstracts the sequence composition. This is the same strategy used by important classifiers as the RDP classifier (16, 17), Kraken (1 and 2) (13, 14), and Kaiju (15).

The application of DCV representation on the classification of DNA sequences in opposition to *one-hot* encoding was investigated by Lo Bosco and Di Gangi in (81), where it was concluded that the use of DCV representation generates the most fine-grained rank performance in the classification of biological sequences through long–short–term memory (LSTM) and convolutional neural network (CNN) models.

Conclusions

Here we present a highly *sensitive* and *specific*, alignment-free, and lightweight tool for the phylogenetic placement of sequences. We see the simplicity and clarity of the *classeq* solution as an advantage for future improvements in terms of performance and customization.

Unlike competing solutions, *classeq* is the first tool that provides a native server to perform classification through an API port. Such a feature provides a rapid way integrate our tool with web native systems, pipelines attached to the web commonly existing into departmental wide solutions.

Future remarks

Despite *classeq* algorithm representing a prominent solution to deal with gene-wide trees for sequences classification, our tool still represents an unoptimized prototype, with all steps of the indexing and prediction workflows executed in Python Virtual Machine (82). This means that the obvious first way to achieve the performance of other classifiers is to translate the key point of our algorithm into a low-level programming language.

In the following, in order to provide a complete and informative report on the *classeq* functionality with optimizations, our tools should be directly compared with competing tools such as APPLES and APP-APAM to benchmark our algorithm against a global scenario of phylogenetic placers.

ACKNOWLEDGEMENTS

Thanks to the grant from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES). We also thank the anonymous reviewers who dedicated their valuable time to improving this work. Thanks to Biotrop, Solutions in Biological Technologies, for the support during the development of this work.

Bibliography

- Ziheng Yang and Bruce Rannala. Molecular phylogenetics: principles and practice. *Nature reviews genetics*, 13(5):303–314, 2012.
- Edward C Holmes. The phylogeography of human viruses. *Molecular ecology*, 13(4):745–756, 2004.
- Nick JB Isaac, Samuel T Turvey, Ben Collen, Carly Waterman, and Jonathan EM Baillie. Mammals on the edge: conservation priorities based on threat and phylogeny. *PLoS one*, 2(3):e296, 2007.
- Paul H Harvey and Mark D Pagel. *The comparative method in evolutionary biology*. Oxford university press, 1991.
- Alessio Squassina, Mirko Manchia, Vangelis G Manolopoulos, Mehmet Artac, Christina Lappa-Manakou, Sophia Karkabouna, Konstantinos Mitropoulos, Maria Del Zompo, and George P Patrinos. Realities and expectations of pharmacogenomics and personalized

- medicine: impact of translating genetic knowledge into clinical practice. *Pharmacogenomics*, 11(8):1149–1167, 2010.
- Rob Ogden. Forensic science, genetics and wildlife biology: getting the right mix for a wildlife dna forensics lab. *Forensic science, medicine, and pathology*, 6(3):172–179, 2010.
- Michele Morgante and Francesco Salamini. From plant genomics to breeding practice. *Current Opinion in Biotechnology*, 14(2):214–219, 2003.
- Y Marin-Felix, JZ Groenewald, L Cai, Qian Chen, S Marincowitz, I Barnes, K Bensch, U Braun, E Camporesi, U Damm, et al. Genera of phytopathogenic fungi: Gophy 1. *Studies in mycology*, 86:99–216, 2017.
- Yasmina Marin-Felix, Margarita Hernández-Restrepo, Michael J Wingfield, A Akulov, AJ Carnegie, R Cheewangkoon, David Gramaje, Johannes Zacharias Groenewald, Vladimiro Guarnaccia, F Halleen, et al. Genera of phytopathogenic fungi: Gophy 2. *Studies in mycology*, 92:47–133, 2019.
- Yasmina Marin-Felix, Margarita Hernández-Restrepo, I Iturrieta-González, D García, J Gené, Johannes Zacharias Groenewald, L Cai, Q Chen, W Quaadvlieg, RK Schumacher, et al. Genera of phytopathogenic fungi: Gophy 3. *Studies in mycology*, 94:1–124, 2019.
- Q Chen, Mounes Bakhshi, Y Balci, KD Broders, R Cheewangkoon, SF Chen, XL Fan, David Gramaje, F Halleen, M Horta Jung, et al. Genera of phytopathogenic fungi: Gophy 4. *Studies in Mycology*, 101:417, 2022.
- Lucas Czech, Alexandros Stamatakis, Micah Dunthorn, and Pierre Barbera. Metagenomic analysis using phylogenetic placement—a review of the first decade. *Frontiers in Bioinformatics*, 2, 5 2022. ISSN 2673-7647. doi: 10.3389/fbinf.2022.871393.
- Derrick E Wood and Steven L Salzberg. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):1–12, 2014.
- Derrick E Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with kraken 2. *Genome biology*, 20:1–13, 2019.
- Peter Menzel, Kim Lee Ng, and Anders Krogh. Kaiju: Fast and sensitive taxonomic classification for metagenomics. *Biorxiv*, page 031229, 2015.
- Bonnie L Maidak, Gary J Olsen, Niels Larsen, Ross Overbeek, Michael J McCaughey, and Carl R Woese. The rdp (ribosomal database project). *Nucleic acids research*, 25(1):109–110, 1997.
- Bonnie L Maidak, James R Cole, Timothy G Lilburn, Charles T Parker Jr, Paul R Saxman, Jason M Stredwick, George M Garrity, Bing Li, Gary J Olsen, Sakti Pramanik, et al. The rdp (ribosomal database project) continues. *Nucleic acids research*, 28(1):173–174, 2000.
- Chunyan Ao, Shihu Jiao, Yansu Wang, Liang Yu, and Quan Zou. Biological sequence classification: A review on data and general methods. *Research*, 2022, 1 2022. ISSN 2639-5274. doi: 10.34133/research.0011.
- Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han, Amy Jo A Johnson, and Susan P Holmes. Dada2: High-resolution sample inference from illumina amplicon data. *Nature methods*, 13(7):581–583, 2016.
- J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, et al. Qiime allows analysis of high-throughput community sequencing data. *Nature methods*, 7(5):335–336, 2010.
- Evan Bolyen, Jai Ram Rideout, Matthew R Dillon, Nicholas A Bokulich, Christian C Abnet, Gabriel A Al-Ghalith, Harriet Alexander, Eric J Alm, Manimozhyan Arumugam, Francesco Asnicar, et al. Reproducible, interactive, scalable and extensible microbiome data science using qiime 2. *Nature biotechnology*, 37(8):852–857, 2019.
- Patrick D Schloss, Sarah L Westcott, Thomas Ryabin, Justine R Hall, Martin Hartmann, Emily B Hollister, Ryan A Lesniewski, Brian B Oakley, Donovan H Parks, Courtney J Robinson, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and environmental microbiology*, 75(23):7537–7541, 2009.
- Arthur Brady and Steven L Salzberg. Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nature methods*, 6(9):673–676, 2009.
- Frederick A Matsen, Robin B Kodner, and E Virginia Armbrust. pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC bioinformatics*, 11(1):1–16, 2010.
- Simon A Berger and Alexandros Stamatakis. Aligning short reads to reference alignments and trees. *Bioinformatics*, 27(15):2068–2075, 2011.
- Diego Fioravanti, Ylenia Giarratano, Valerio Maggio, Claudio Agostinelli, Marco Chierici, Giuseppe Jurman, and Cesare Furlanello. Phylogenetic convolutional neural networks in metagenomics. *BMC Bioinformatics*, 19:49, 3 2018. ISSN 1471-2105. doi: 10.1186/s12859-018-2033-5.
- Pierre Barbera, Alexey M Kozlov, Lucas Czech, Benoit Morel, Diego Darriba, Tomáš Flouri, and Alexandros Stamatakis. Epa-ng: massively parallel evolutionary placement of genetic sequences. *Systematic biology*, 68(2):365–369, 2019.
- Benjamin Linard, Krister Swenson, and Fabio Pardi. Rapid alignment-free phylogenetic identification of metagenomic sequences. *Bioinformatics*, 35(18):3303–3312, 2019.
- Metin Balaban, Shahab Sarmashghi, and Siavash Mirarab. Apples: scalable distance-based phylogenetic placement with or without alignments. *Systematic Biology*, 69(3):566–578, 2020.
- Matthias Blanke and Burkhard Morgenstern. App-spam: phylogenetic placement of short reads without sequence alignment. *Bioinformatics Advances*, 1(1):vbab027, 2021.
- Jyotsna Talreja Wassan, Haiying Wang, and Huiru Zheng. A new phylogeny-driven random forest-based classification approach for functional metagenomics. pages 32–37. IEEE, 12 2022. ISBN 978-1-6654-6819-0. doi: 10.1109/BIBM55620.2022.9995554.
- Jyotsna Talreja Wassan, Haiying Wang, and Huiru Zheng. Developing a new phylogeny-driven random forest model for functional metagenomics. *IEEE Transactions on NanoBioScience*, 22:763–770, 10 2023. ISSN 1536-1241. doi: 10.1109/TNB.2023.3283462.
- Nicola Segata, Levi Waldron, Annalisa Ballarini, Vagheesh Narasimhan, Olivier Jousson, and Curtis Huttenhower. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, 9(8):811–814, 2012.
- Duy Tin Truong, Eric A Franzosa, Timothy L Tickle, Matthias Scholz, George Weingart, Edoardo Pasoli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. Metaphlan2 for enhanced metagenomic taxonomic profiling. *Nature methods*, 12(10):902–903, 2015.

35. Mahwash Jamy, Rachel Foster, Pierre Barbera, Lucas Czech, Alexey Kozlov, Alexandros Stamatakis, Gary Bending, Sally Hilton, David Bass, and Fabien Burki. Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. *Molecular Ecology Resources*, 20(2):429–443, 2020.
36. Francesco Beghini, Lauren J McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, Paolo Manghi, Matthias Scholz, Andrew Maltz Thomas, et al. Integrating taxonomic, functional, and strain-level profiling of diverse microbial communities with biobakery 3. *elife*, 10:e65088, 2021.
37. Steven W Kembel, Jonathan A Eisen, Katherine S Pollard, and Jessica L Green. The phylogenetic diversity of metagenomes. *PLoS One*, 6(8):e23214, 2011.
38. Diane S Srivastava, Marc W Cadotte, A Andrew M MacDonald, Robin G Marushia, and Nicholas Mirochnick. Phylogenetic diversity and the functioning of ecosystems. *Ecology Letters*, 15(7):637–648, 2012.
39. Marc W Cadotte. Phylogenetic diversity and productivity. *Functional Ecology*, 29(12):1603–1606, 2015.
40. Robert Edgar. Taxonomy annotation and guide tree errors in 16s rRNA databases. *PeerJ*, 6:e5030, 6 2018. ISSN 2167-8359. doi: 10.7717/peerj.5030.
41. Keri Ann Lydon and Erin K. Lipp. Taxonomic annotation errors incorrectly assign the family pseudoalteromonadaceae to the order vibriales in greengenes: implications for microbial community assessments. *PeerJ*, 6:e5248, 7 2018. ISSN 2167-8359. doi: 10.7717/peerj.5248.
42. Hariharu Subrahmanian Muralidharan, Noam Y. Fox, and Mihai Pop. The impact of transitive annotation on the training of taxonomic classifiers. *Frontiers in Microbiology*, 14, 1 2024. ISSN 1664-302X. doi: 10.3389/fmicb.2023.1240957.
43. Tetsu Sakamoto and J Miguel Ortega. Taxallomy: an extension of ncbi taxonomy that produces a hierarchically complete taxonomic tree. 2020. doi: 10.1186/s12859-021-04304-3.
44. Alexey M. Kozlov, Jiajie Zhang, Pelin Yilmaz, Frank Oliver Glöckner, and Alexandros Stamatakis. Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Research*, 44:5022–5033, 6 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw396.
45. David L Swofford. Paup, phylogenetic analysis using parsimony. version 3.1. *Computer program distributed by the Illinois Natural History Survey*, 1993.
46. Fredrik Ronquist, Maxim Teslenko, Paul Van Der Mark, Daniel L Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A Suchard, and John P Huelsenbeck. MrBayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3):539–542, 2012.
47. Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.
48. Nikolai Romashchenko, Benjamin Linard, Fabio Pardi, and Eric Rivals. Epik: precise and scalable evolutionary placement with informative k-mers. *Bioinformatics*, 39(12):btad692, 2023.
49. Conrad L Schoch, Keith A Seifert, Sabine Huhndorf, Vincent Robert, John L Spouge, C André Levesque, Wen Chen, Fungal Barcoding Consortium, Fungal Barcoding Consortium Author List, Elena Bolchacova, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for fungi. *Proceedings of the national academy of sciences*, 109(16):6241–6246, 2012.
50. Jaime Huerta-Cepas, François Serra, and Peer Bork. Ete 3: reconstruction, analysis, and visualization of phylogenomic data. *Molecular biology and evolution*, 33(6):1635–1638, 2016.
51. Lucas Czech, Jaime Huerta-Cepas, and Alexandros Stamatakis. A critical review on the use of support values in tree views and bioinformatics toolkits. *Molecular biology and evolution*, 34(6):1535–1542, 2017.
52. Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.
53. Li-Ting Wang, Fwu-Ling Lee, Chun-Ju Tai, and Hiroaki Kasai. Comparison of gyrB gene sequences, 16S rRNA gene sequences and DNA–DNA hybridization in the bacillus subtilis group. *International journal of systematic and evolutionary microbiology*, 57(8):1846–1850, 2007.
54. Alejandro P Rooney, Neil PJ Price, Christopher Ehrhardt, James L Swezey, and Jason D Bannan. Phylogeny and molecular taxonomy of the bacillus subtilis species complex and description of bacillus subtilis subsp. inaquosorum subsp. nov. *International journal of systematic and evolutionary microbiology*, 59(10):2429–2436, 2009.
55. Vaibhav Bhandari, Nadia Z Ahmad, Haroun N Shah, and Radhey S Gupta. Molecular signatures for bacillus species: demarcation of the bacillus subtilis and bacillus cereus clades in molecular terms and proposal to limit the placement of new species into the genus bacillus. *International Journal of Systematic and Evolutionary Microbiology*, 63(Pt_7):2712–2726, 2013.
56. Robert Trevethan. Sensitivity, specificity, and predictive values: foundations, plabilities, and pitfalls in research and practice. *Frontiers in public health*, 5:307, 2017.
57. Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using diamond. *Nature methods*, 12(1):59–60, 2015.
58. Torbjørn Rognes, Tomáš Flouri, Ben Nichols, Christopher Quince, and Frédéric Mahé. Vsearch: a versatile open source tool for metagenomics. *PeerJ*, 4:e2584, 2016.
59. Kazutaka Katoh and Daron M Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.
60. Alexandros Stamatakis, Paul Hoover, and Jacques Rougemont. A rapid bootstrap algorithm for the raxml web servers. *Systematic biology*, 57(5):758–771, 2008.
61. Margaret M Palmisano, Lawrence K Nakamura, Kathleen E Duncan, Conrad A Istock, and Frederick M Cohan. Bacillus sonorensis sp. nov., a close relative of bacillus licheniformis, isolated from soil in the sonoran desert, arizona. *International Journal of Systematic and Evolutionary Microbiology*, 51(5):1671–1679, 2001.
62. H Weigmann. Über zwei an der käserzeugung beteiligte bakterien. *Zentralbl. Bakteriol. Hyg. II*, 4:820–834, 1898.
63. Christopher A Dunlap, Soon-Wo Kwon, Alejandro P Rooney, and Soo-Jin Kim. Bacillus paralicheniformis sp. nov., isolated from fermented soybean paste. *International journal of systematic and evolutionary microbiology*, 65(Pt_10):3487–3492, 2015.
64. LK Nakamura. Taxonomic relationship of black-pigmented bacillus subtilis strains and a proposal for bacillus atrophaeus sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, 39(3):295–300, 1989.
65. Ben Fan, Jochen Blom, Hans-Peter Klenk, and Rainer Borriss. Bacillus amyloliquefaciens, bacillus velezensis, and bacillus siamensis form an “operational group b. amyloliquefaciens” within the b. subtilis species complex. *Frontiers in microbiology*, 8:22, 2017.
66. Punnane Sumpavapool, Linna Tongyongk, Somboon Tanasupawat, Nipa Chokesajjawatee, Plearnpis Luxananil, and Wonnop Viessanguan. Bacillus siamensis sp. nov., isolated from salted crab (poo-khem) in thailand. *International journal of systematic and evolutionary microbiology*, 60(10):2364–2370, 2010.
67. FG Priest, M Goodfellow, LA Shute, and RCW Berkeley. Bacillus amyloliquefaciens sp. nov., nom. rev. *International journal of systematic and evolutionary microbiology*, 37(1):69–71, 1987.
68. Cristina Ruiz-García, Victoria Bejar, Fernando Martínez-Checa, Inmaculada Llamas, and Emilia Quesada. Bacillus velezensis sp. nov., a surfactant-producing bacterium isolated from the river vélez in Málaga, southern Spain. *International Journal of Systematic and Evolutionary Microbiology*, 55(1):191–195, 2005.
69. BJ Tindall. The consequences of bacillus axarquiensis ruiz-garcía et al. 2005, bacillus malacitensis ruiz-garcía et al. 2005 and brevibacterium halotolerans delaporte and sasson 1967 (approved lists 1980) being treated as heterotypic synonyms. *International journal of systematic and evolutionary microbiology*, 67(1):175–176, 2017.
70. Michael S Roberts, LK Nakamura, and Frederick M Cohan. Bacillus mojavensis sp. nov., distinguishable from bacillus subtilis by sexual isolation, divergence in DNA sequence, and differences in fatty acid composition. *International Journal of Systematic and Evolutionary Microbiology*, 44(2):256–264, 1994.
71. Conrad L Schoch, Stacy Ciuffo, Mikhail Domrachev, Carol L Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef Leipe, Richard McVeigh, Kathleen O'Neill, Barbara Robertse, et al. Ncbi taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020:baaa062, 2020.
72. Joshua W Gatson, Bruce F Benz, Chitra Chandrasekaran, Masataka Satomi, Kasthuri Venkateswaran, and Mark E Hart. Bacillus tequilensis sp. nov., isolated from a 2000-year-old mexican shaft-tomb, is closely related to bacillus subtilis. *International journal of systematic and evolutionary microbiology*, 56(7):1475–1484, 2006.
73. Christopher A Dunlap, Michael J Bowman, and Daniel R Zeigler. Promotion of bacillus subtilis subsp. inaquosorum, bacillus subtilis subsp. spizizenii and bacillus subtilis subsp. stercoris to species status. *Antonie Van Leeuwenhoek*, 113:1–12, 2020.
74. Michael S Roberts, Lawrence K Nakamura, and Frederick M Cohan. Bacillus vallismortis sp. nov., a close relative of bacillus subtilis, isolated from soil in death valley, California. *International Journal of Systematic and Evolutionary Microbiology*, 46(2):470–475, 1996.
75. Alejandro P Rooney, Neil PJ Price, Christopher Ehrhardt, James L Swezey, and Jason D Bannan. Phylogeny and molecular taxonomy of the bacillus subtilis species complex and description of bacillus subtilis subsp. inaquosorum subsp. nov. *International journal of systematic and evolutionary microbiology*, 59(10):2429–2436, 2009.
76. Skerman Vdb. Approved lists of bacterial names. *Int J Syst Bacteriol*, 30:225–420, 1980.
77. Kevin de Queiroz. Nodes, branches, and phylogenetic definitions. *Systematic biology*, 62(4):625–632, 2013.
78. Heta P Desai, Anuja P Parameshwaran, Rajshekhar Sunderraman, and Michael Weeks. Comparative study using neural networks for 16S ribosomal gene classification. *Journal of Computational Biology*, 27(2):248–258, 2020.
79. Heta P Desai, Anuja P Parameshwaran, Rajshekhar Sunderraman, and Michael Weeks. Deep ensemble models for 16S ribosomal gene classification. In *Bioinformatics Research and Applications: 16th International Symposium, ISBRA 2020, Moscow, Russia, December 1–4, 2020, Proceedings 16*, pages 282–290. Springer, 2020.
80. Antonino Fiannaca, Laura La Paglia, Massimo La Rosa, Giosue' Lo Bosco, Giovanni Renda, Riccardo Rizzo, Salvatore Gaglio, and Alfonso Urso. Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC bioinformatics*, 19:61–76, 2018.
81. Giosuè Lo Bosco and Mattia Antonino Di Gangi. Deep learning architectures for DNA sequence classification. In *Fuzzy Logic and Soft Computing Applications: 11th International Workshop, WILF 2016, Naples, Italy, December 19–21, 2016, Revised Selected Papers 11*, pages 162–171. Springer, 2017.
82. Obi Ike-Nwosu. Inside the python virtual machine, 2015.

Anexos

ANEXO A – Qualificadores elegíveis aos Feature Keys do tipo *source*

O conteúdo da tabela abaixo representa um compilado das informações disponíveis no documento técnico de referências sobre as Feature Keys utilizados para caracterizar os registros do Genbank (veja mais detalhes no [item 7.2](#) da documentação oficial do INSDC).

#	Qualificador	Tipo	Obrigatório
1	/organism	Texto	Sim
2	/mol_type	Vocabulário controlado	Sim
3	/altitude	Texto	Não
4	/bio_material	[<institution-code>:[<collection-code>:]]<material_id>	Não
5	/cell_line	Texto	Não
6	/cell_type	Texto	Não
7	/chromosome	Texto	Não
8	/clone	Texto	Não
9	/clone_lib	Texto	Não
10	/collected_by	text	Não
11	/collection_date	Texto	Não
12	/country	Vocabulário controlado	Não
13	/cultivar	Texto	Não
14	/culture_collection	Vocabulário controlado	Não
15	/db_xref	Vocabulário controlado	Não
16	/dev_stage	Texto	Não
17	/ecotype	Texto	Não
18	/environmental_sample	Binário	Não
19	/focus	Binário	Não
20	/geo_loc_name	<country_value>[:<region>][, <locality>]	Não
21	/germline	Binário	Não
22	/haplogroup	Texto	Não
23	/haplotype	Texto	Não
24	/host	Texto	Não
25	/identified_by	Texto	Não
26	/isolate	Texto	Não
27	/isolation_source	Texto	Não

28	/lab_host	Texto	Não
29	/lat_lon	Texto	Não
30	/macronuclear	Binário	Não
31	/map	Texto	Não
32	/mating_type	Texto	Não
33	/metagenome_source	Texto	Não
34	/note	Texto	Não
35	/organelle	Vocabulário controlado	Não
36	/PCR_primers	[fwd_name: XXX,]fwd_seq: xxxxx	Não
37	/plasmid	Texto	Não
38	/pop_variant	Texto	Não
39	/proviral	Binário	Não
40	/rearranged	Binário	Não
41	/segment	Texto	Não
42	/serotype	Texto	Não
43	/serovar	Texto	Não
44	/sex	Texto	Não
45	/specimen_voucher	Vocabulário controlado	Não
46	/strain	Texto	Não
47	/sub_clone	Texto	Não
48	/submitter_seqid	Texto	Não
49	/sub_species	Texto	Não
50	/sub_strain	Texto	Não
51	/tissue_lib	Texto	Não
52	/tissue_type	Texto	Não
53	/transgenic	Binário	Não
54	/type_material	Vocabulário controlado	Não
55	/variety	Texto	Não

Tabela 6 – Qualificadores elegíveis como *Feature keys* do tipo *source*.