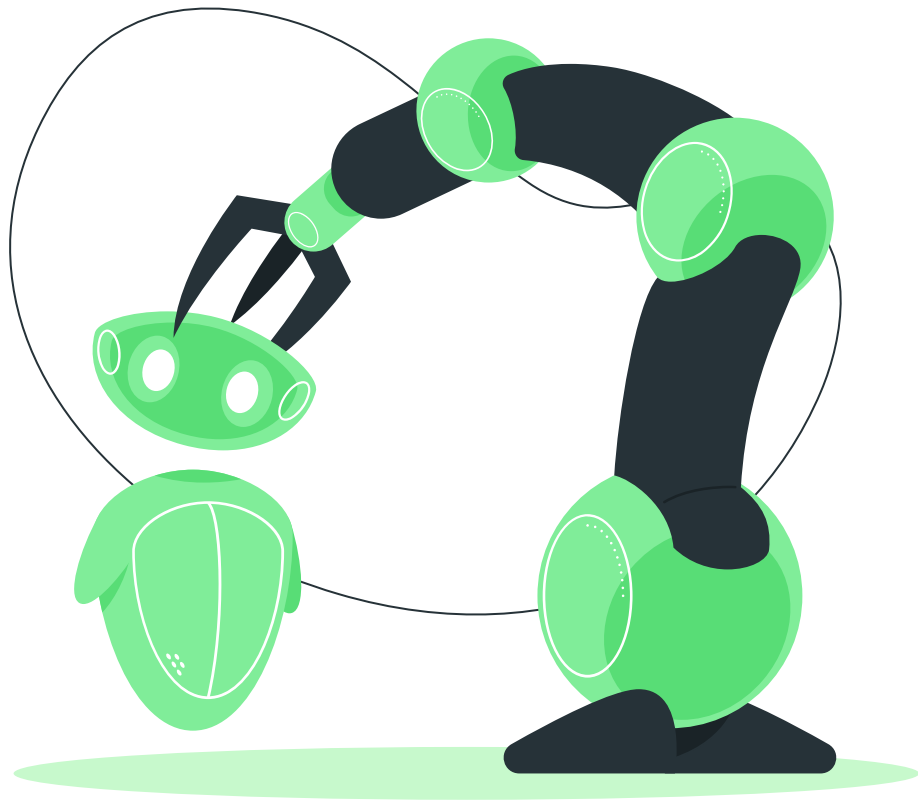


Aprendizado de Máquina

Um aliado na curadoria de
dados biológicos



THE
DEVELOPER'S
CONFERENCE



Quem somos?



Professora na UFVJM. Mestra em Ciência da Computação. Bacharela em Sistemas de Informação.

Doutoranda em Bioinformática na UFMG, em colaboração com a Fiocruz Minas, *University of Melbourne* e *Baker Institute* na Austrália.

Co-fundadora e voluntária do Código X, um projeto social que incentiva meninas a seguirem carreiras tecnológicas.

Co-fundadora do Laboratório de Bioinformática e Inteligência Artificial (BIA).

Graduada em Sistemas de Informação pela Universidade Estadual de Montes Claros (Unimontes).

Especialista em Redes pela Escola Superior Aberta do Brasil (ESAB).

Atualmente é técnica de laboratório de informática da Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM).

Colaboradora do Laboratório de Bioinformática e Inteligência Artificial (BIA).



De que vamos falar?

1

Bioinformática

3

**Facilitando a
curadoria**

2

Curadoria manual

4

Aprendizado Ativo



1

Bioinformática

*"...Um dos objetivos da bioinformática é usar softwares e computação para **descartar hipóteses equivocadas**, mas que seriam testadas, levando a um **desperdício de tempo e custos financeiros**".*

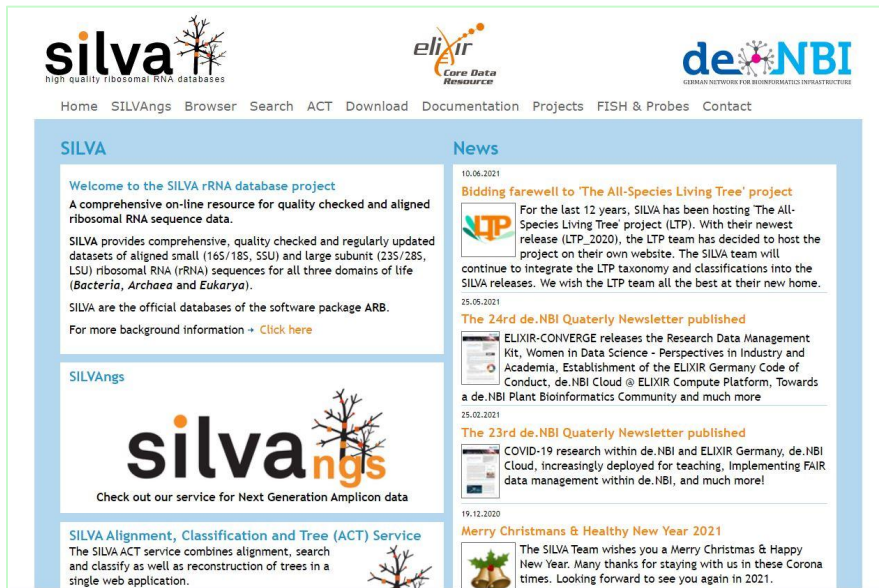
[Professor Hermes Amorim (ULBRA)]



Ideia Geral



Bases de dados biológicas



The screenshot shows the SILVA database homepage. At the top, there are logos for SILVA (high quality ribosomal RNA databases), ELIXIR Core Data Resource, and de.NBI (German Network for Bioinformatics Infrastructure). Below the logos is a navigation bar with links: Home, SILVAngs, Browser, Search, ACT, Download, Documentation, Projects, FISH & Probes, and Contact. The main content area is divided into two columns. The left column, titled 'SILVA', contains a welcome message, a description of the database, and information about the software package ARB. The right column, titled 'News', contains several news items, including a farewell to 'The All-Species Living Tree' project, the publication of the 24rd de.NBI Quaterly Newsletter, and the publication of the 23rd de.NBI Quaterly Newsletter. At the bottom, there is a section for 'SILVAngs' with a logo and a description of the service.

SILVA
high quality ribosomal RNA databases

Home SILVAngs Browser Search ACT Download Documentation Projects FISH & Probes Contact

Welcome to the SILVA rRNA database project
A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data.

SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (*Bacteria*, *Archaea* and *Eukarya*).

SILVA are the official databases of the software package ARB.

For more background information → [Click here](#)

SILVAngs

silvangs
Check out our service for Next Generation Amplicon data

SILVA Alignment, Classification and Tree (ACT) Service
The SILVA ACT service combines alignment, search and classify as well as reconstruction of trees in a single web application.

News

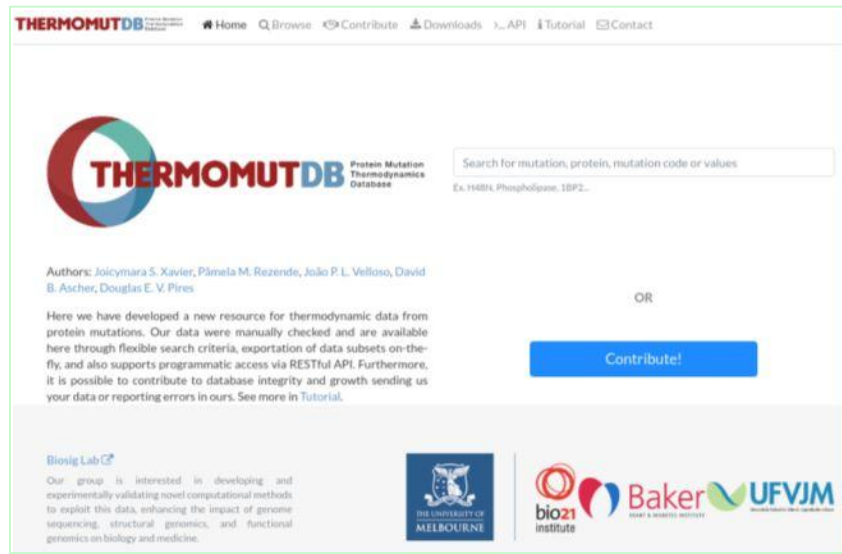
10.06.2021
Bidding farewell to 'The All-Species Living Tree' project
For the last 12 years, SILVA has been hosting The All-Species Living Tree' project (LTP). With their newest release (LTP_2020), the LTP team has decided to host the project on their own website. The SILVA team will continue to integrate the LTP taxonomy and classifications into the SILVA releases. We wish the LTP team all the best at their new home.

25.05.2021
The 24rd de.NBI Quaterly Newsletter published
ELIXIR-CONVERGE releases the Research Data Management Kit, Women in Data Science - Perspectives in Industry and Academia, Establishment of the ELIXIR Germany Code of Conduct, de.NBI Cloud @ ELIXIR Compute Platform, Towards a de.NBI Plant Bioinformatics Community and much more

25.02.2021
The 23rd de.NBI Quaterly Newsletter published
COVID-19 research within de.NBI and ELIXIR Germany, de.NBI Cloud, increasingly deployed for teaching, Implementing FAIR data management within de.NBI, and much more!

19.12.2020
Merry Christmas & Healthy New Year 2021
The SILVA Team wishes you a Merry Christmas & Happy New Year. Many thanks for staying with us in these Corona times. Looking forward to see you again in 2021.

Nível de RNA: Informações RNA ribossomal



The screenshot shows the THERMOMUTDB database homepage. At the top, there is a navigation bar with links: Home, Browse, Contribute, Downloads, API, Tutorial, and Contact. Below the navigation bar is a large search bar with the text 'Search for mutation, protein, mutation code or values'. Below the search bar is a section for 'Authors' with the names: Joicymara S. Xavier, Pamela M. Rezende, João P. L. Veloso, David B. Ascher, Douglas E. V. Pires. Below the authors is a section for 'Contribute!' with a blue button. At the bottom, there is a section for 'Biology Lab' with a logo and a description of the group's interests. To the right of the Biology Lab section are logos for the University of Melbourne, Baker Institute, and UFVJM.

THERMOMUTDB Protein Mutation Thermodynamics Database

Home Browse Contribute Downloads API Tutorial Contact

Search for mutation, protein, mutation code or values

Ex: 148B, Phospholipase, 1BP2...

Authors: Joicymara S. Xavier, Pamela M. Rezende, João P. L. Veloso, David B. Ascher, Douglas E. V. Pires

OR

Here we have developed a new resource for thermodynamic data from protein mutations. Our data were manually checked and are available here through flexible search criteria, exportation of data subsets on-the-fly, and also supports programmatic access via RESTful API. Furthermore, it is possible to contribute to database integrity and growth sending us your data or reporting errors in ours. See more in [Tutorial](#).

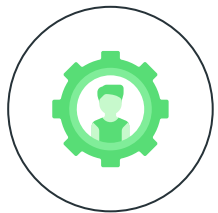
Contribute!

Biology Lab
Our group is interested in developing and experimentally validating novel computational methods to exploit this data, enhancing the impact of genome sequencing, structural genomics, and functional genomics on biology and medicine.

UNIVERSITY OF MELBOURNE
Baker
bio21 institute
UFVJM

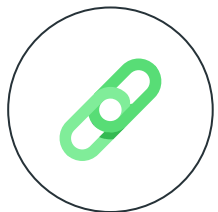
Especializadas: Dados experimentais que podem ser utilizadas para treinar preditores.

Desafios



Curadoria

Dados são inseridos por usuários e nem sempre são checados



Checagem humana

Quantidade de dados inviável para checagem humana.



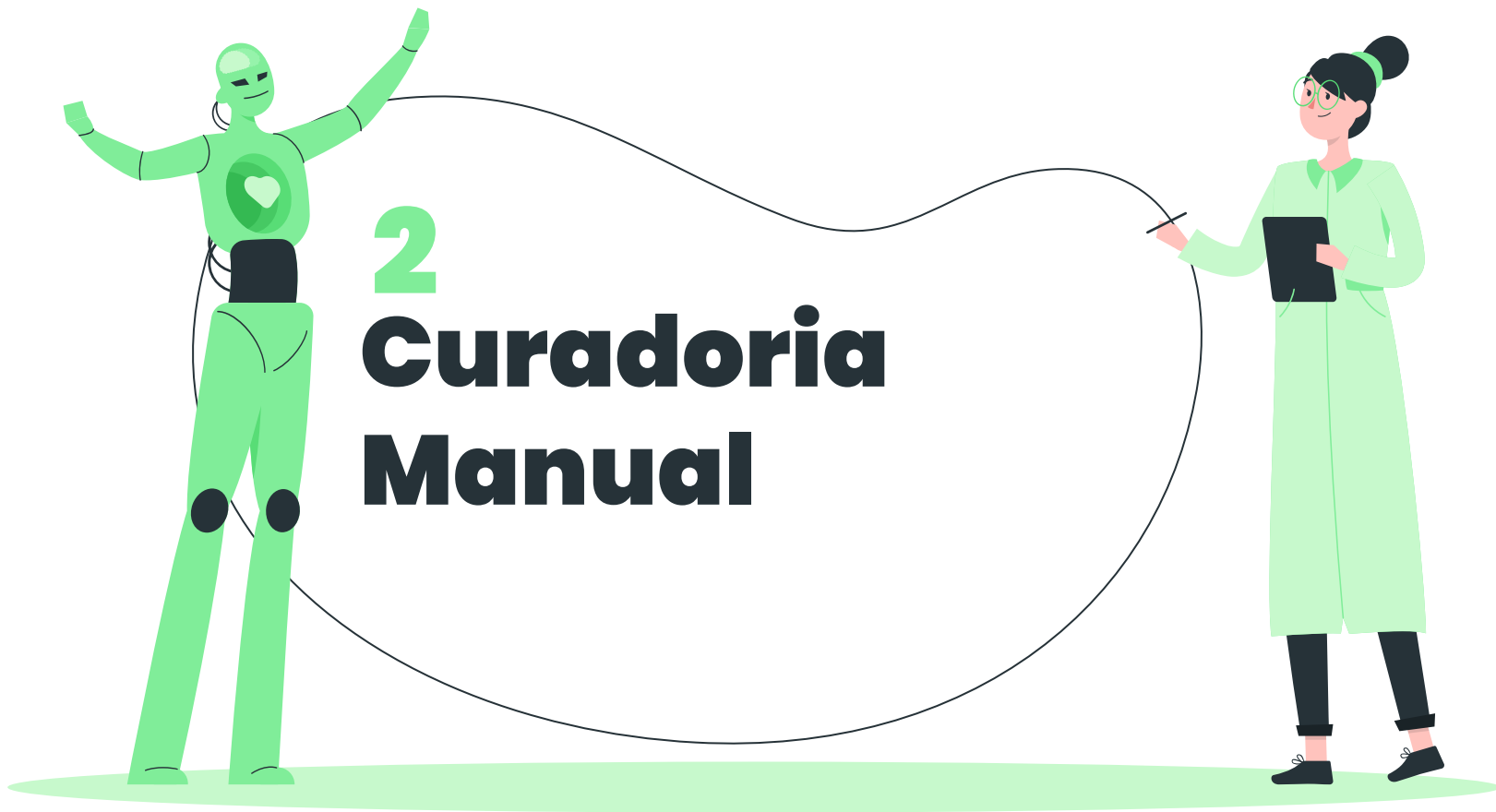
Volume de dados

Dados biológicos estão sendo produzidos o tempo todo

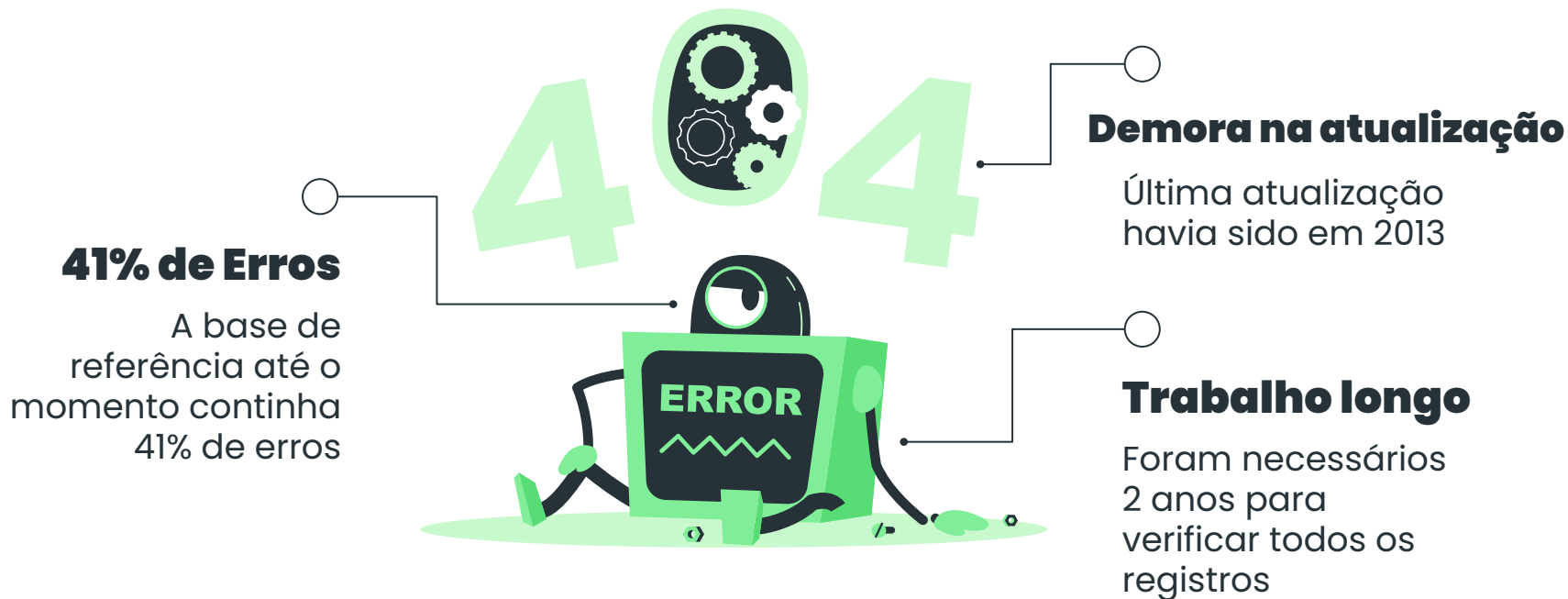


Modelos enviesados

Modelos são treinados com esses dados e reproduzem os vieses



Curadoria do ThermoMutDB

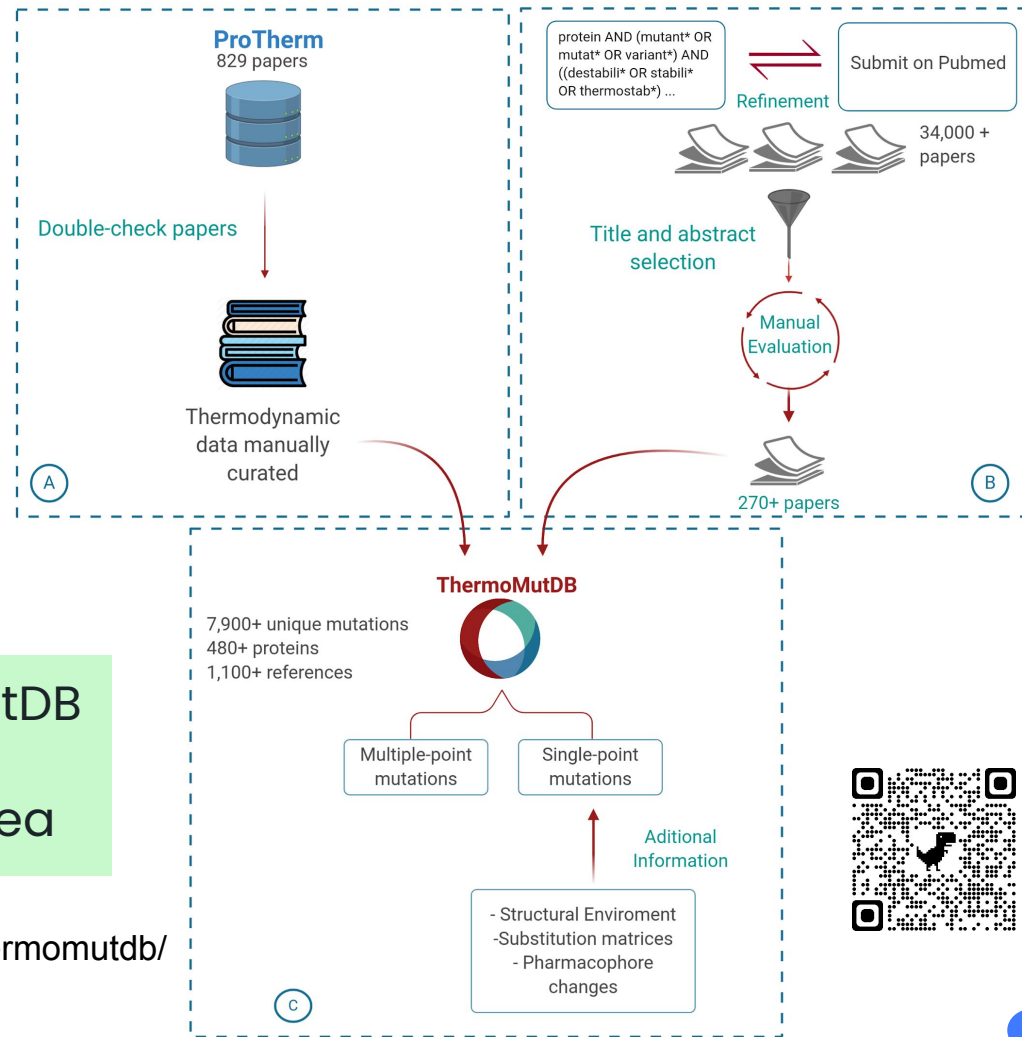


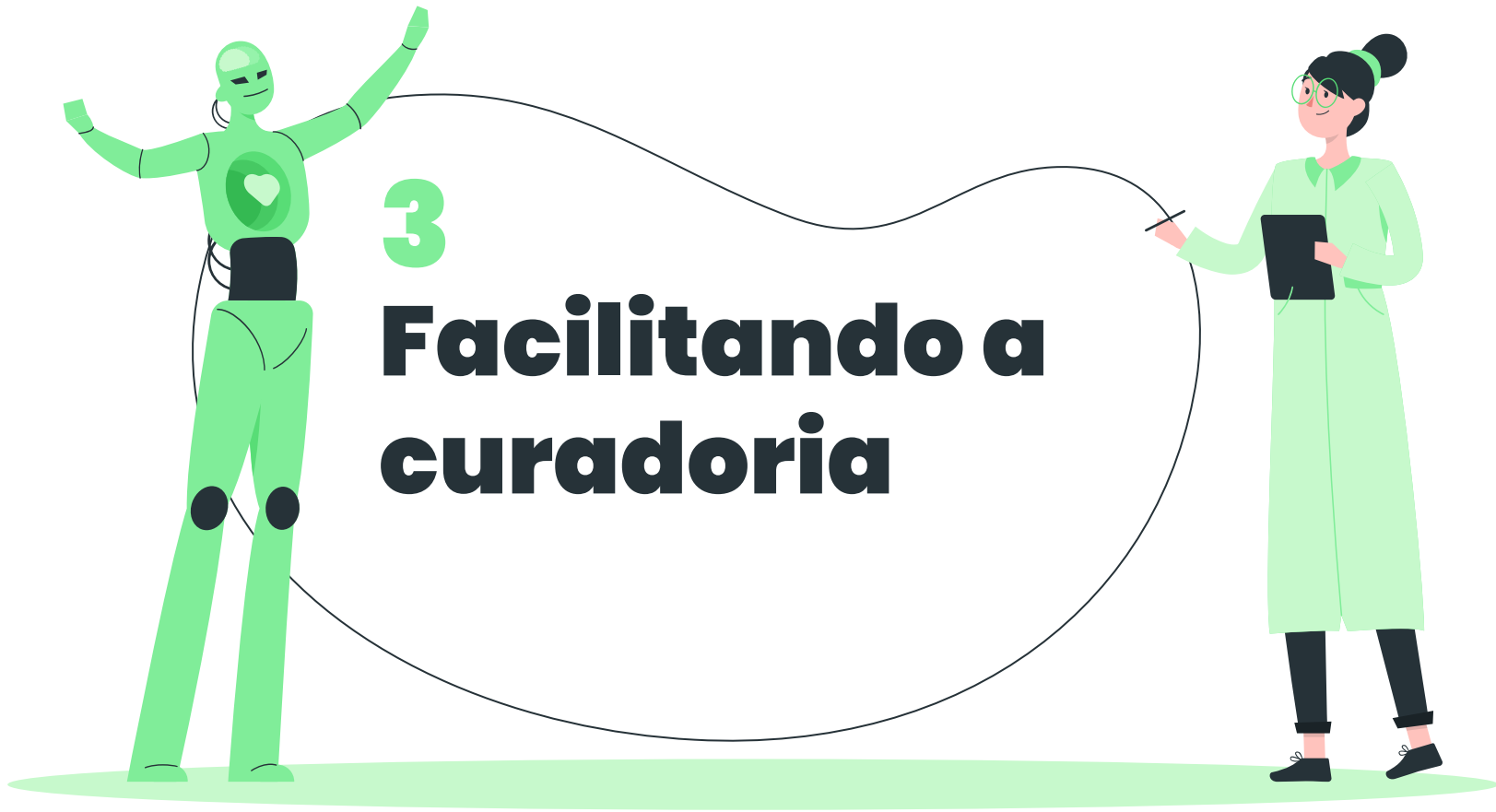
Metodologia



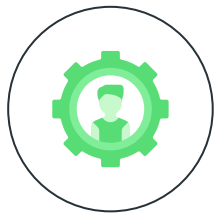
Hoje o ThermoMutDB
é a principal
referência na área

<http://biosig.unimelb.edu.au/thermomutdb/>





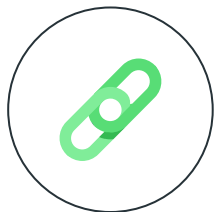
Problemas da curadoria manual



**Altamente
dependente de
especialistas**



**Humanamente
custoso**



**Não há garantia de
eliminação total dos
vieses**



Tempo

Checar manualmente?

Tabela 1 - Resumo das bases de dados utilizadas

Base de dados	Versão	Instâncias	Níveis	Max. de classes por nível	Desafios
Silva	128	1.719.541	7	32.122	
KEGG	17/07/2019	225.976	4	497	
BioLip	15/07/2019	24.321	4	207	
CATH	4.2	24.765	4	6119	

Legenda:

Desbalanceamento

Número elevado
de classes

Níveis
profundos

Previsão

Não curada

Curadoria

1.719.541

Quantidade total de instâncias.

200 mil

Após pré processamento,
eliminando alguns vieses.

10%

Semente inicial (amostra
aleatória)

2332

Curadoria manual em 7 níveis.
Tempo médio: 6 meses.



Anotação

Manual annotation of SILVA



Discordâncias em Classificação Taxonômica

3 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Thiopseudomonas;s__Pseudomonas sp.
2 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Vibrionales;f__Vibrionaceae;g__Vibrio;s__Pseudomonas fluorescens
1 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Aquimonas;s__Pseudomonas sp.
1 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Luteimonas;s__Pseudomonas geniculata
14 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Pseudoxanthomonas;s__Pseudomonas sp.
2 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas aeruginosa
1 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas boreopolis
97 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas geniculata
1 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas helmanticensis
36 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas hibiscicola
1 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas nitroreducens
4 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas pictorum
79 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas sp.
1 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Xanthomonas;s__Pseudomonas cissicola
1 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Xanthomonas;s__Pseudomonas geniculata
7 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Xylella;s__Pseudomonas boreopolis
10 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Xylella;s__Pseudomonas sp.
2 k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadales Incertae Sedis;g__Steroidobacter;s__Pseudomonas sp.

(QUAST et al., 2013)

Discordâncias em Classificação Taxonômica



3 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Thiopseudomonas;s__Pseudomonas sp.
2 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Vibrionales;f__Vibrionaceae;g__Vibrio;s__Pseudomonas fluorescens
1 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Aquimonas;s__Pseudomonas sp.
1 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Luteimonas;s__Pseudomonas geniculata
14 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Pseudoxanthomonas;s__Pseudomonas sp.
2 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas aeruginosa
1 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas boreopolis
97 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas geniculata
1 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas helmanticensis
36 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas hibiscicola
1 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas nitroreducens
4 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas pictorum
79 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas sp.
1 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Xanthomonas;s__Pseudomonas cissicola
1 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Xanthomonas;s__Pseudomonas geniculata
7 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Xylella;s__Pseudomonas boreopolis
10 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Xylella;s__Pseudomonas sp.
2 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadales Incertae Sedis;g__Steroidobacter;s__Pseudomonas sp.

(QUAST et al., 2013)

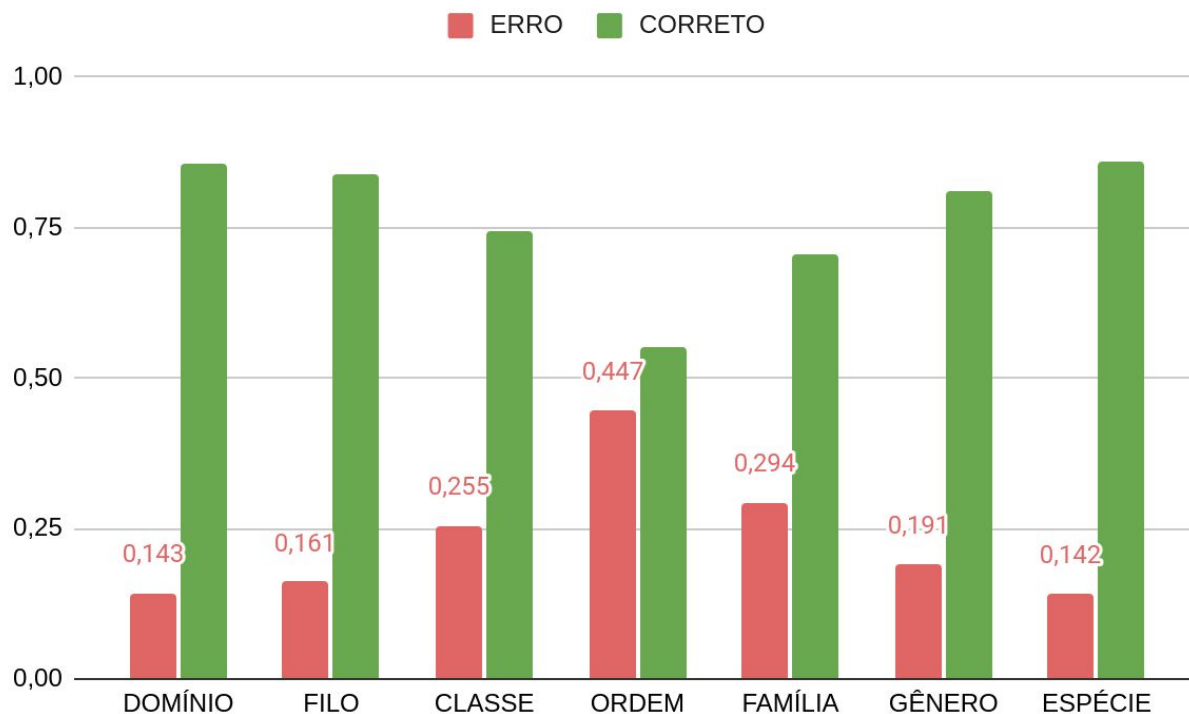
Discordâncias em Classificação Taxonômica



3 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Thiopseudomonas;s__Pseudomonas sp.
2 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Vibrionales;f__Vibrionaceae;g__Vibrio;s__Pseudomonas fluorescens
1 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Aquimonas;s__Pseudomonas sp.
1 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Luteimonas;s__Pseudomonas geniculata
14 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Pseudoxanthomonas;s__Pseudomonas sp.
2 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas aeruginosa
1 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas boreopolis
97 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas geniculata
1 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas helmanticensis
36 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas hibiscicola
1 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas nitroreducens
4 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas pictorum
79 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas;s__Pseudomonas sp.
1 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Xanthomonas;s__Pseudomonas cissicola
1 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Xanthomonas;s__Pseudomonas geniculata
7 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Xylella;s__Pseudomonas boreopolis
10 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Xylella;s__Pseudomonas sp.
2 k_Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadales Incertae Sedis;g__Steroidobacter;s__Pseudomonas sp.

(QUAST et al., 2013)

Discordâncias em Classificação Taxonômica





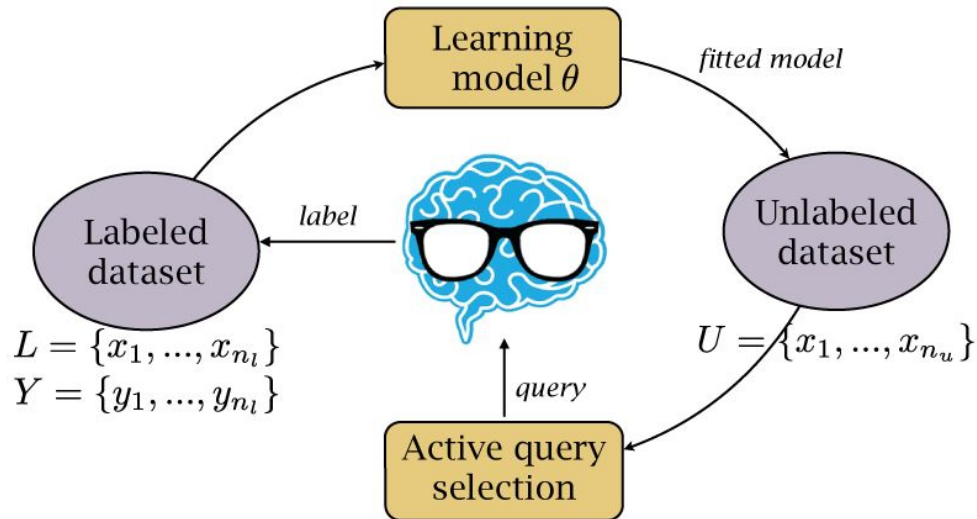
E agora?

- Criar uma base de dados de sequências curadas;
- Utilizar aprendizado de máquina;
- Utilizar metodologias independentes de alinhamento.
- Disponibilizar publicamente essa base curada.





Aprendizado Ativo



- Verificação manual de uma pequena amostra da base (10%);
- 6 meses de trabalho manual;
- Restante do trabalho é aprendizado de máquina;

Começando na Bioinformática

Eventos do Curso de Verão da UFMG

WOB20

I WORKSHOP ONLINE DE BIOINFORMÁTICA



@bioinfo_ufmg

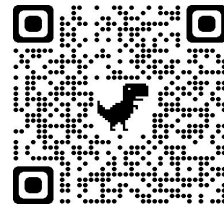
Revista Bioinfo



bioinfo

<https://bioinfo.com.br/>

Artigo: Inteligência Artificial aplicada à Bioinformática



@revista-bioinfo



Obrigada!

✉ joicy.xavier@ufvjm.edu.br

🌐 <https://linktr.ee/joicyxavier>



✉ quezia.bernardes@ufvjm.edu.br

🌐 Insta: @queziabernardes