

LABORATÓRIO NACIONAL DE COMPUTAÇÃO CIENTÍFICA
PROGRAMA DE PÓS-GRADUAÇÃO DO LABORATÓRIO NACIONAL
DE COMPUTAÇÃO CIENTÍFICA
CURSO DE MESTRADO EM MODELAGEM COMPUTACIONAL
GA047 - BANCO DE DADOS DO PONTO DE VISTA BIOLÓGICO

Bancos de dados biológicos

Docente: Marisa Nicolas

Discente: Wellington Silva

Petrópolis - RJ
Fevereiro - 2021

Sumário

1	Banco de dados de sequência de nucleotídeos	1
1.1	Introdução	1
1.2	Visão geral do GenBank-NCBI no All Resources	1
1.2.1	Mycobacterium tuberculosis	1
1.2.1.1	O livro específico para MTB mais atual publicado	1
1.2.1.2	O artigo em Pubmed específico para tuberculosis mais atual publicado:	2
1.2.1.3	Quantas em PubmedCentral	2
1.2.1.4	Qual a entrada nucleotide sequences de M. tuberculosis mais atual depositada	3
1.2.2	BioProject e BioSample	4
1.2.3	Genome	5
1.2.4	PubMed	6
1.2.5	Taxonomy	6
1.2.6	GenBank	6
1.3	EMBL-EIB	7
1.3.1	Organização das seções	7
1.3.2	Ensembl Genomes	7
1.3.3	Europe Pubmed	7
1.3.4	ENA database	7
2	Banco de dados de sequência de Proteínas	7
	REFERÊNCIAS	10

Lista de ilustrações

Figura 1	– Livro mais recente sobre a espécie <i>Mycobacterium tuberculosis</i> é de Tettelin e Medini (2020)	1
Figura 2	– Artigo mais recente sobre a espécie <i>Mycobacterium tuberculosis</i> é de Oliveira et al. (2020)	2
Figura 3	– Existem 98799 resultados em PubMed Central sobre a espécie <i>Mycobacterium tuberculosis</i>	3
Figura 4	– O nucleotide sequences mais recente <i>Mycobacterium tuberculosis</i> 02_1987 atualizado em 2017	3
Figura 5	– Número de genomas são divididos entre a Eucariotas (15442); Procarionte (303290); Vírus (42078); Plasmídeo (26783); e Organelas (17828). Além da Kingdom de bactérias que contêm 29545.	5
Figura 6	– O maior genoma entre as bactérias é a <i>Bacterium</i>	5
Figura 7	– O menor genoma entre as bactérias é a <i>Bacterium</i>	6

1 Banco de dados de sequência de nucleotídeos

1.1 Introdução

Os dois principais bancos de dados públicos primários, GenBank-NCBI e EMBL-ENA contêm a Sequência de genes, proteínas e genomas completos atualizadas periodicamente (entre 2 a 3 meses), porém pouco trabalho manualmente é realizado nesses bancos. No entanto, seções especiais podem ser encontradas tanto no NCBI como no EMBL, as quais contêm ferramentas ou dados de Sequências analisados e classificados que são de grande utilidade para a pesquisa. Assim, esses dois principais bancos de dados podem ser utilizados como páginas líder para qualquer pesquisa envolvendo dados biológicos. Portanto, iniciaremos o roteiro com uma visão geral do conteúdo total desses dois bancos.

1.2 Visão geral do GenBank-NCBI no All Resources

1.2.1 Mycobacterium tuberculosis

Observar as grandes divisões: Databases, Download, Submission, Tools, How To. Fazer uma busca em “All Databases” para a espécie Mycobacterium tuberculosis (MTB). Listar para cada o maior número possível de seções dos hits mais atuais.

1.2.1.1 O livro específico para MTB mais atual publicado

The screenshot shows the NCBI Bookshelf search results for 'Mycobacterium tuberculosis'. The search bar at the top contains the text 'Mycobacterium tuberculosis'. Below the search bar, there are options to 'Browse Titles', 'Create alert', and 'Advanced'. The results are displayed in a list format, sorted by relevance. The first result is 'The Pangenome: Diversity, Dynamics and Evolution of Genomes [Internet]' by Tettelin H, Medini D, editors, published by Cham (CH): Springer, 2020. This book is highlighted as the most recent. Below the book title, there is a 'Table of contents' link. The search details on the right side of the page show the search query: '"mycobacterium tuberculosis"[MeSH Terms] OR ("mycobacterium"[All Fields] AND "tuberculosis"[All Fields]) OR "mycobacterium tuberculosis"[All Fields]'. The recent activity section on the right shows the search history, including 'Mycobacterium tuberculosis (1750)' and 'Mycobacterium\ tuberculosis (1750)'. The bottom of the page shows the 'Recent activity' section with a list of recent searches and their results.

Figura 1 – Livro mais recente sobre a espécie Mycobacterium tuberculosis é de Tettelin e Medini (2020)



1.2.1.2 O artigo em Pubmed específico para tuberculosis mais atual publicado:



The screenshot shows the PubMed interface for a specific article. At the top, the NIH logo and 'National Library of Medicine' are visible. Below this is the 'PubMed.gov' logo and a search bar. The article title is 'Primary Prophylaxis to Prevent Tuberculosis Infection in Prison Inmates: A Randomized, Double-Blind, Placebo-Controlled Trial'. The authors listed are Roberto Dias de Oliveira, Andrea da Silva Santos, Cassia Barbosa Reis, Alessandra de Cássia Leite, Flávia Patussi Correia Sacchi, Rafaela Carla Pivetta de Araujo, Paulo César Pereira Dos Santos, Valeria Cavalcanti Rolla, Leonardo Martinez, Jason Andrews, and Julio Croda. The publication details are 'Am J Trop Med Hyg. 2020 Oct;103(4):1466-1472.' and the DOI is '10.4269/ajtmh.20-0110'. On the right side, there are buttons for 'Cite', 'Favorites', and 'Share'. The bottom of the page shows the PMID (32876010) and PMCID (PMC7543866).

Figura 2 – Artigo mais recente sobre a espécie *Mycobacterium tuberculosis* é de Oliveira et al. (2020)

1.2.1.3 Quantas em PubmedCentral

GENOME ASSEMBLY Was this helpful?  

ASM19595v2
Mycobacterium tuberculosis H37Rv (high GC Gram+)
 Sanger Institute (February 2013)
 RefSeq GCF_000195955.2
[PubMed \(3\)](#)

[Entrez Genome](#) [BLAST](#) [Download](#)

Use NCBI Datasets for bulk downloading of genome sequence and annotation data.
[NCBI Datasets](#) [Command-line tool](#) [API documentation](#)

Assembly statistics +

Literature		Genes	
Bookshelf	1,750	Gene	15,787
MeSH	254	GEO DataSets	10,186
NLM Catalog	411	GEO Profiles	429,123
PubMed	79,787	HomoloGene	5
PubMed Central	98,799	PopSet	412

Figura 3 – Existem 98799 resultados em PubMed Central sobre a espécie *Mycobacterium tuberculosis*

1.2.1.4 Qual a entrada nucleotide sequences de *M. tuberculosis* mais atual depositada

Observando a página da seção no menu direita. Results by táxon Organisms [Tree] *Mycobacterium tuberculosis*.

https://biocyc.org/organism-summary?object=GCF_000658295

BIOCYC
Database Collection

Enter a gene, protein, metabolite or pathway... [Quick Search](#) [Gene Set](#)

Searching *Mycobacterium tuberculosis* 02_1987 [change organism database](#)

Summary of *Mycobacterium tuberculosis*, Strain 02_1987, version 24.5

Authors: Pallavi Subhtraveti¹, Peter Midford, Ingrid Keseler¹, Anamika Kothari¹, Ron Caspi¹, Peter D Karp¹

¹SRI International

Summary:
 This Pathway/Genome Database (PGDB) was generated on 15-Feb-2019 from the annotated genome of *Mycobacterium tuberculosis* 02_1987, as obtained from RefSeq (annotation date: 08-MAR-2017). The PGDB was created computationally by the PathoLogic component of the Pathway Tools software (version 23.0) [Karp16, Karp11] using MetaCyc version 23.0 [Caspi18]. It has not undergone any manual curation or review, and may contain errors. Development of this PGDB was supported by grant GM080746 from the National Institutes of Health.

Taxonomic Lineage: cellular organisms, Bacteria <bacteria>, Terrabacteria group, Actinobacteria <actinobacteria>, Actinobacteria <high GC Gram+>, Corynebacteriales, Mycobacteriaceae, Mycobacterium, Mycobacterium tuberculosis complex, Mycobacterium tuberculosis, Mycobacterium tuberculosis 02_1987

Unification Links: BIOSAMPLE:SAMN02400330, NCBI BioProject:PRJNA224116, NCBI-Taxonomy:515616

Organism or Sample Properties	
Annotation Provider	NCBI
Annotation Date	2017-3-7 7:52:18
Annotation Pipeline	NCBI Prokaryotic Genome Annotation Pipeline
Annotation Pipeline Version	4.1
Annotation Comment	Best-placed reference protein set: GeneMarkS+

Figura 4 – O nucleotide sequences mais recente *Mycobacterium tuberculosis* 02_1987 atualizado em 2017

1.2.2 BioProject e BioSample

O BioProject conjunto de dados biológicos relacionados a uma única iniciativa, oriundos de uma única organização ou consórcio. O BioSample é um banco de dados que contém descrições de materiais de origem biológica usados em ensaios experimentais. Portanto, para identificar as entradas para o vírus *SARS-CoV-2* o BioProject disponibiliza 469 resultados Já o BioSample 231361 resultados, ambas as pesquisas realizadas no dia 21/01/2020. A diferença é dada pelo critério de armazenamento do banco de dados biológicos. Sendo o BioProject mantido pelas organizações, já BioSample armazena ensaios experimentais. Então, se o número de ensaios é maior que o número de organizações. Além disso, uma organização pode realizar diversos ensaios experimentais.

1.2.3 Genome

Esta seção organiza informações sobre genomas, incluindo sequências, mapas, cromossomos, montagens e anotações. A divisão dos genomas é representada na Figura 5.

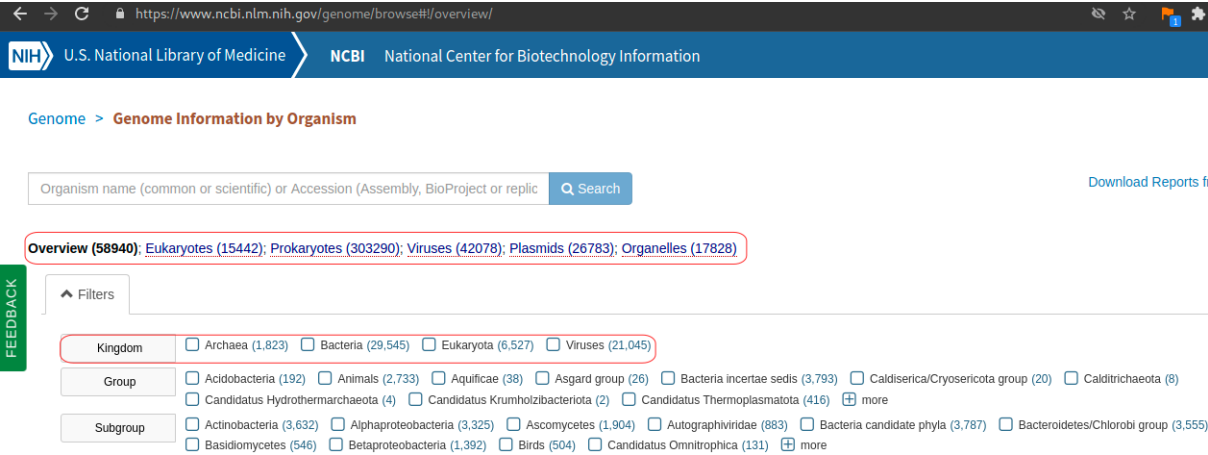


Figura 5 – Número de genomas são divididos entre a Eucariotas (15442); Procarionte (303290); Vírus (42078); Plasmídeo (26783); e Organelas (17828). Além da Kingdom de bactérias que contêm 29545.

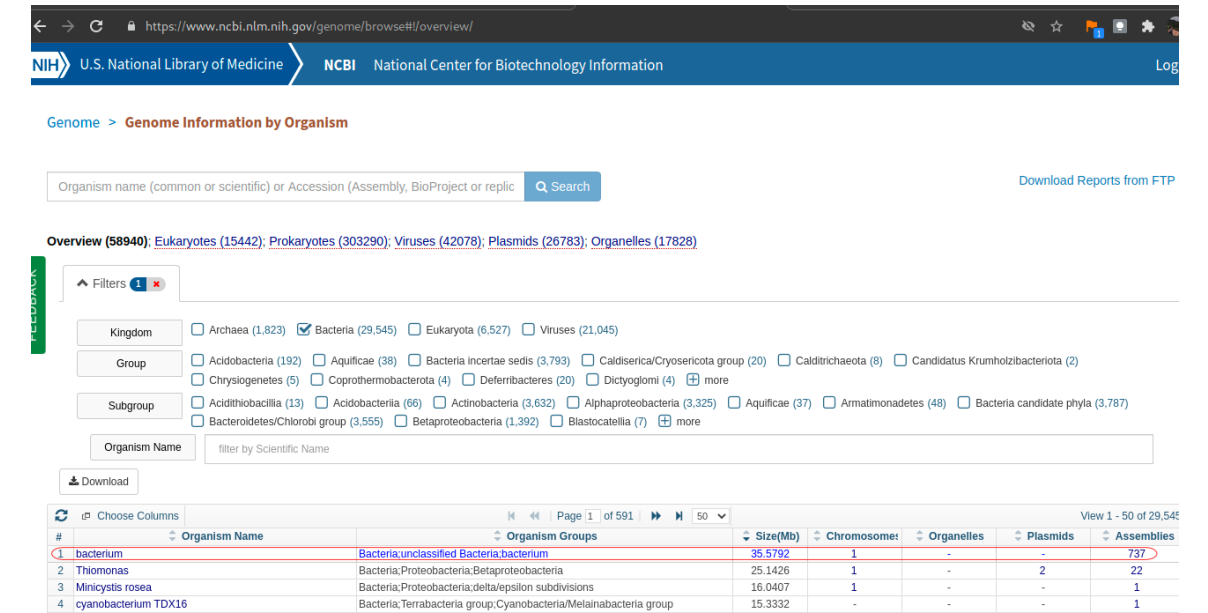


Figura 6 – O maior genoma entre as bactérias é a Bacterium

Overview (58940): Eukaryotes (15442); Prokaryotes (303290); Viruses (42078); Plasmids (26783); Organelles (17828)

Filters: 1

Kingdom: ☐ Archaea (1,823) ☒ Bacteria (29,545) ☐ Eukaryota (6,527) ☐ Viruses (21,045)

Group: ☐ Acidobacteria (192) ☐ Aquificae (38) ☐ Bacteria incertae sedis (3,793) ☐ Caldiserica/Cryosericota group (20) ☐ Caldtrichaeota (8) ☐ Candidatus Krumholzibacteriota (2) ☐ Chrysiogenetes (5) ☐ Coprothermobacterota (4) ☐ Deferribacteres (20) ☐ Dictyoglomi (4) ☐ more

Subgroup: ☐ Acidithiobacillia (13) ☐ Acidobacteria (66) ☐ Actinobacteria (3,632) ☐ Alphaproteobacteria (3,325) ☐ Aquificae (37) ☐ Armatimonadetes (48) ☐ Bacteria candidate phyla (3,787) ☐ Bacteroidetes/Chlorobi group (3,555) ☐ Betaproteobacteria (1,392) ☐ Blastocatellia (7) ☐ more

Organism Name: filter by Scientific Name

Download

#	Organism Name	Organism Groups	Size (Mb)	Chromosome	Organelles	Plasmids	Assemblies
1	bacterium AG-323-L21	Bacteria:unclassified Bacteria;bacterium AG-323-L21	0.101587	-	-	-	1
2	bacterium AG-316-A20	Bacteria:unclassified Bacteria;bacterium AG-316-A20	0.111261	-	-	-	1
3	bacterium AG-345-J16	Bacteria:unclassified Bacteria;bacterium AG-345-J16	0.111391	-	-	-	1
4	Xanthomonadaceae bacterium JGI 0001002-D18	Bacteria:Proteobacteria;Gammaproteobacteria	0.112613	-	-	-	1

Figura 7 – O menor genoma entre as bactérias é a Bacterium

1.2.4 PubMed

Na seção PubMed que inclui links para artigos completos e outras fontes relacionadas. O NCBI oferece você criar uma conta pessoal através do MyNCBI, um recurso de utilidade para poder salvar qualquer busca no banco.

- Fazer uma busca cruzando cada um dos termos agr (D, B, C ou A) com *Staphylococcus aureus* e identificar quantos artigos foram achados em cada busca.

1.2.5 Taxonomy

Na seção Taxonomy: Procurar as Sequências de genes *agrD*, *agrB*, *agrC* e *agrA* em *Staphylococcus aureus*. Para isso iniciar a busca da espécie no Taxonomy browser e depois localizar na tabela “Entrez records” em Direct link: em Nucleotide.

- quantas entradas são recuperadas para cada busca?

1.2.6 GenBank

Na seção GenBank. GenBank Submission Types: descrever os tipos de submissões permitidas. Em Submission tools: investigar como submeter uma Sequência de nucleotídeos no GenBank.

A submissão para o GenBank é aceita em mRNA ou dados de sequência genômica determinados diretamente pelo solicitante. O envio deve incluir informações sobre o organismo de origem e anotações fornecidas pelo remetente¹.

¹ https://www.ncbi.nlm.nih.gov/genbank/submit_types/

1.3 EMBL-EIB

Visão geral do site map do EMBL-EIB ²

1.3.1 Organização das seções

As seções estão organizadas alfabeticamente. Na seção EMBL-Bank encontra-se o banco de nucleotídeos ENA e contém as informações para submissões, semelhantes às descritas no NCBI. Procurar a seção para submissão de uma Sequência.

As tecnologias fornecem opções apropriadas para a escala e frequência de envio, a experiência e capacidade do solicitante e a natureza dos dados a serem transferidos. Sendo por intermédio de formulário, por terminal de comando ou pelo formato XML ³.

1.3.2 Ensembl Genomes

A seção Ensembl Genomes tem como objetivo desenvolver um sistema que mantém anotação automática de genomas. Os genomas disponíveis em ⁴ são de planta, animal, alga, fungo e bactéria.

1.3.3 Europe Pubmed

A seção Europe Pubmed que também forma parte do PubMed, contém search para buscas bibliográficas. Fazer uma busca como indicado na Seção 1.2.5 através do formulário SRS MEDLINE.

1.3.4 ENA database

Procurar no ENA database⁵ as mesmas sequências de *S. aureus* indicadas na Seção 1.2.5

2 Banco de dados de sequência de Proteínas

1. Quais os principais bancos de dados de sequências de proteínas? Quais outros bancos os compõem?
2. O que é o RefSeq Database e quais suas principais características?
3. Quais as diferenças entre o GenBank e o RefSeq?
4. Qual a descrição dos seguintes “status” de uma entrada no RefSeq: model, predicted, inferred, provisional, reviewed, validated, WGS.

² <https://www.ebi.ac.uk/services/all>

³ <https://www.ebi.ac.uk/ena/browser/submit?src=wizard&wiztype=quicklink&wizid=2b8303c5-9938-42d9-b485-868eacf22195>

⁴ ensemblgenomes.org

⁵ www.ebi.ac.uk/ena/browser/home

5. Fazendo uma busca no All databases, em quantos bancos de dados do NCBI podemos encontrar resultados para a proteína “myosin”? Quantas estruturas 3D existem relacionadas com esta proteína?

6. Buscar uma entrada para a proteína “calmodulin-1” de humano no banco de dados Protein.

a) Selecionar aquelas que estão presentes no RefSeq. Existem isoformas para essa proteína? Em caso afirmativo, forneça o número de acesso para as entradas no RefSeq.

b) Escolha uma das isoformas e responda:.

c) Como é o formato de uma sequência FASTA? Forneça a sequência no formato FASTA desta entrada.

d) Qual o “status” desta entrada e o que esse status significa?

e) Qual a função dessa proteína, caso esteja descrito

f) Existe algum domínio conservado associado a esta proteína? Em caso afirmativo, citar a família e o(s) banco(s) de dados que corroboram a informação.

g) Utilizando a ferramenta BLASTp encontre sequências com 100% de identidade no banco de dados UniProtKb/Swiss-Prot para o mesmo organismo. Qual o número de acesso no Swiss-Prot?

7. Buscar proteínas homólogas á entrada angiotensin converting enzyme 2 de humano no banco “nr”.

a) Filtrar as entradas entre 90 e 100% de identidade da sequência.

b) Quais organismos estão representados nesse conjunto selecionado?

c) Qual o resultado do alinhamento com a sequência homóloga de Macaca mulatta?

1. Quais bancos de dados ou conjunto de dados podem ser acessados pelo UniProt?

2. Diferencie o UniProtKb/trEMBL e o UniProtKb/Swiss-Prot.

3. Em que consiste a anotação manual?

4. Quantas entradas estão presentes no UniProtKB relacionadas com bomba de efluxo (“efflux pump”)?

a) Quantas destas estão presentes no Swiss-Prot e quantas estão no trEMBL?

b) Quantas entradas relacionadas com bomba de efluxo presentes no Swiss-Prot pertencem ao organismo *Bacillus subtilis* (BACSU)?

5. Procurar a entrada Q4R9Z3 no UniProtKB.

a) Em qual seção do UniProtKB encontra-se esta entrada?

b) A qual organismo pertence?

c) A qual família essa proteína supostamente pertenceria?

d) Essa entrada poderia ser integrada no Swiss-Prot? Justifique sua resposta.

6- Procurar a proteína “toll-like receptor 4” de humanos no UniProtKB/Swiss-Prot.

a) quais outros nomes podem designar esta proteína?

b) Quais domínios e repetições estão associados a esta proteína? Cite as referências cruzadas (bancos de dados) que suportam sua resposta.

c) Qual a função desta proteína? d) Existem sequências alternativas para esta proteína? Em caso positivo, listar os identificadores.

7-Encontrar uma entrada no Swiss-Prot que seja fosforilada (dica: usar uma KW) e que tenha domínio transmembrana (dica: usar uma KW) no organismo *Staphylococcus aureus* strain N315.

- a) Indique o número de acessos de proteínas.
- b) Qual o nome e sinônimos dessas proteínas e suas funções?
- c) Fornecer o número EC, caso sejam enzimas.
- d) Qual o nome dos genes?
- e) Qual o código usado para designar a espécie *S. aureus* strain N315 no Swiss-Prot?
- f) Qual tipo de modificação essas proteínas sofrem? Em qual resíduo? Indique a posição na sequência de cada.
- g) Caso as proteínas tenham ligação à metal(is), citar o(s) metal(is) e em qual(is) posição(ões)?
- h) Existem estruturas 3D para essas proteínas?

Referências

OLIVEIRA, R. D. de et al. Primary prophylaxis to prevent tuberculosis infection in prison inmates: A randomized, double-blind, placebo-controlled trial. *The American Journal of Tropical Medicine and Hygiene*, ASTMH, v. 103, n. 4, p. 1466–1472, 2020.

TETTELIN, H.; MEDINI, D. *The pangenome: Diversity, dynamics and evolution of genomes*. [S.l.]: Springer Nature, 2020.