

Apresentação final

Eduardo Augusto Simão Vasconcellos	196240
Victória Pedrazzoli Ferreira	206664
Arthur Biscaino Fruch	164373
José Carlos Cieni Júnior	170859

Etapas 3: Análise SQL

UniProt - Human Diseases

Base de dados

<https://www.uniprot.org/diseases/>

- Banco de doenças;
- Nome das doenças, categoria e descrição;
- Necessário padronização;

The screenshot shows the UniProt Human diseases results page. At the top, there is a search bar with 'Human diseases' selected. Below the search bar, there are navigation links: BLAST, Align, Retrieve/ID mapping, and Peptide search. The main heading is 'Human diseases results'. A text box explains that human diseases in which proteins are involved are described in UniProtKB entries with a controlled vocabulary. It also provides information on how to search for diseases by name. Below this, there is a 'MapTo' section with a 'Download' button. The results list shows two entries: '2,4-dienoyl-CoA reductase deficiency' and '2-aminoadipic 2-oxoadipic aciduria'. Each entry includes a brief description and a link to the UniProtKB entry. At the bottom, there is a privacy notice banner stating that the site has updated its Privacy Notice to comply with the GDPR.

UniProt

Human diseases

BLAST Align Retrieve/ID mapping Peptide search

Human diseases results

The human diseases in which proteins are involved are described in UniProtKB entries with a controlled vocabulary.

Information about the usage of this controlled vocabulary in UniProtKB entries can be found in the user manual.

By default, searching the diseases will look for matches in both name and description. Example queries to search diseases only by name: `name:disorder` - `name:syndrome`

Help Tutorials and videos Downloads

MapTo

Download

1 to 25 of 5,444 Show 25

UniProtKB

Disease

2,4-dienoyl-CoA reductase deficiency

A rare, autosomal recessive, inborn error of polyunsaturated fatty acids and lysine metabolism, resulting in mitochondrial dysfunction. Affected individuals have a severe encephalopathy with neurologic and metabolic abnormalities beginning in early infancy. Laboratory studies show increased C10:2 carnitine levels and hyperlysinemia. UniProtKB (3)

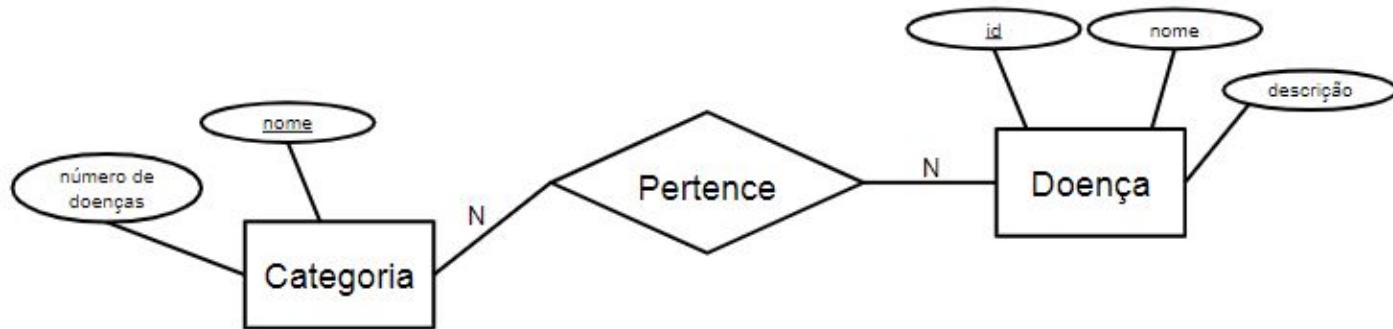
2-aminoadipic 2-oxoadipic aciduria

A metabolic disorder characterized by increased levels of 2-oxoadipate and 2-hydroxyadipate in the urine, and elevated 2-aminoadipate in the plasma. Patients can have mild to severe intellectual disability, muscular hypotonia, developmental delay, ataxia, and epilepsy. Most cases are asymptomatic.

We'd like to inform you that we have updated our [Privacy Notice](#) to comply with Europe's new General Data Protection Regulation (GDPR) that applies since 25 May 2018.

Do not show this banner again

Modelo conceitual - ER





Análises

- Suportes

- Confianças

Predição



Propostas

- Retornar as 10 categorias de doenças mais comuns
- Retornar o suporte de cada categoria de doença
- Retornar o numero de ocorrência de doenças com duas categorias
- Retornar a confiança dessas associações
- Retornar à probabilidade de uma doença ser de uma categoria mais específica em relação a uma menos específica



Exemplo: Suporte das 10 categorias mais comuns

Categoria	Suporte
Síndrome	0.20334
Deficiência	0.08707
Autossomo Recessivo	0.06495
Displasia	0.04078
Congênita	0.03931

Categoria	Suporte
Autossomo Dominante	0.03637
X-Linked	0.02939
Mental	0.02663
Infantil	0.02094
Encefalopatia	0.01891



Exemplo: Ocorrências e Confianças

	Síndrome	Autossomo Dominante	Mental
Síndrome	-	9	26
Autossomo Dominante	9	-	1
Mental	26	1	-

	Síndrome	Autossomo Dominante	Mental
Síndrome	1	0,008	0,023
Autossomo Dominante	0,045	1	0,005
Mental	0,163	0,006	1



Exemplo: Probabilidades

P(A/B)	Autossomo	Total
Autossomo Dominante	0.35563	0.03637
Autossomo Recessivo	0.66197	0.06495

Etapas 3: Análise SQL

Breast Cancer Wisconsin (Diagnostic) Data Set



Base de dados

<https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>

<https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29>

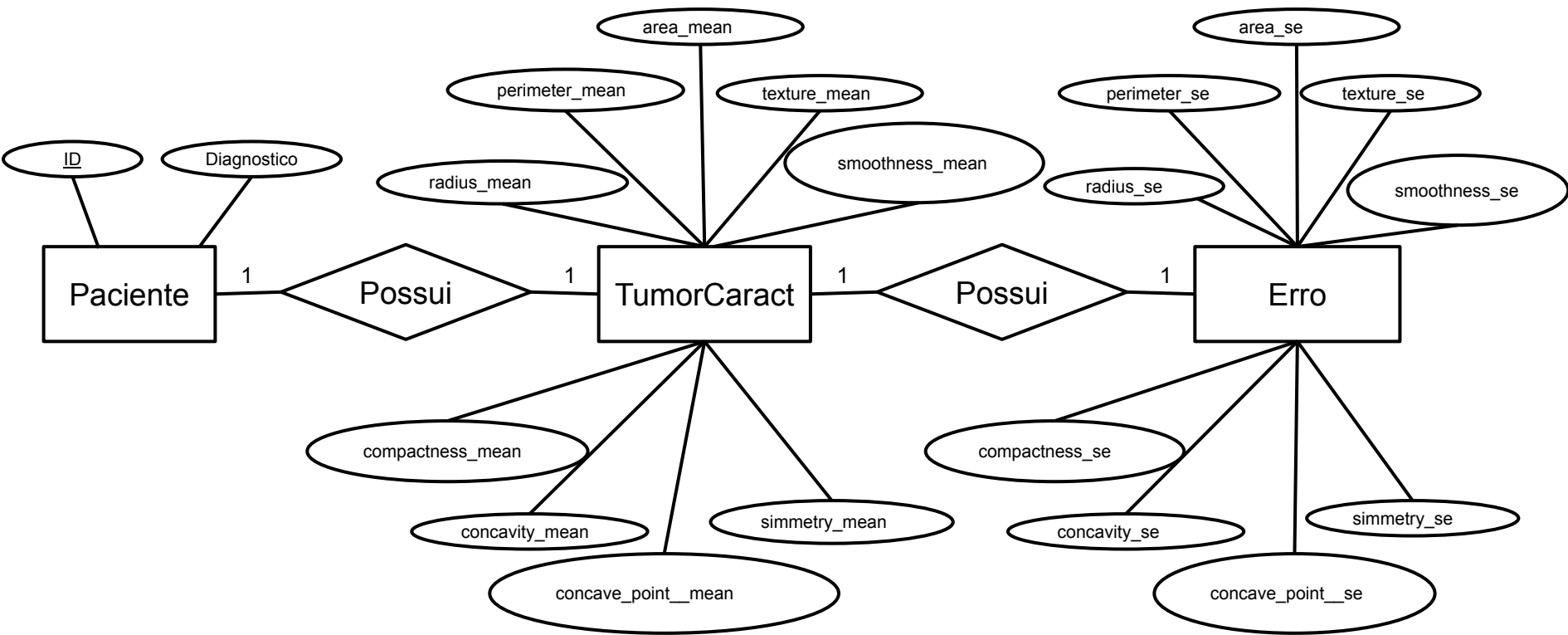
A base contém informações sobre diversos atributos físicos observados no núcleo celular de amostras de tumores na região da mama.

As informações foram agrupadas por paciente, contendo o diagnóstico (M = maligno, B = benigno) e as características da célula: raio (média da distância do centro da célula até o os pontos do perímetro), textura (desvio padrão dos valores em escala de cinza), perímetro, área, suavidade (variação local em raios diferentes), compactidade ($\text{perímetro}^2/\text{área} - 1$), concavidade (grau de concavidade em pontos do contorno), pontos côncavos (total no contorno), simetria e dimensão fractal (aproximação do contorno - 1). Cada registro possui o valor médio, erro padrão e valor para o pior caso observado para cada paciente.

Nossa análise (no notebook) leva em consideração apenas o valor médio e o erro padrão.



Modelo conceitual - ER



Resultados



Através de consultas em SQL, foram comparadas as características obtidas de forma a tentar identificar quais são maiores indicativos de um tumor maligno.

Com base nos valores médios das amostras malignas e benignas foram feitas algumas comparações para determinarmos qual destes atributos eram mais relevantes para predizer justamente se uma amostra seria maligna ou benigna. Após algumas consultas foi identificado que os atributos mais relevantes para predição seriam a área, perímetro, compacidade, concavidade e os pontos côncavos.

Foi calculada a correlação entre a área e as demais características (identificamos que a área pode ser um bom ponto de partida, pois há diferenciação mais perceptível entre diagnósticos benignos e malignos - o que também faz sentido biologicamente).

Encontramos correspondência maior entre área x raio (0,987), área x perímetro (0,986) e área x n de pontos côncavos (0,823), o que era de se esperar, já

RAIO_MEDIA	12.14652380952381
TEXTURA_MEDIA	17.914761904761903
PERIMETRO_MEDIA	78.07540616246499
AREA_MEDIA	462.7901960784314
SUAVIDADE_MEDIA	0.0924776470588235
COMPACIDADE_MEDIA	0.0800846218487394
CONCAVIDADE_MEDIA	0.0460576210084033
PONTOSCONCAVOS_MEDIA	0.0257174061624649
SIMETRIA_MEDIA	0.1741859943977591
DIMENSAOFRACTAL_MEDIA	0.0628673949579832

Figura 1 - médias dos valores atribuídos às características observadas em pacientes com diagnóstico benigno

RAIO_MEDIA	17.462830188679245
TEXTURA_MEDIA	21.60490566037736
PERIMETRO_MEDIA	115.36537735849056
AREA_MEDIA	978.3764150943397
SUAVIDADE_MEDIA	0.10289849056603774
COMPACIDADE_MEDIA	0.14518778301886792
CONCAVIDADE_MEDIA	0.16077471698113208
PONTOSCONCAVOS_MEDIA	0.08799
SIMETRIA_MEDIA	0.19290896226415094
DIMENSAOFRACTAL_MEDIA	0.06268009433962264

Figura 2 - médias dos valores atribuídos às características observadas em pacientes com diagnóstico maligno

Etapă 4: XQuery

UniProt - UniParc



Base de dados

<https://www.uniprot.org/uniparc/>

- Banco de proteínas;
- Informações de vacinas, taxonomia, genes;

The screenshot shows the UniParc website interface. At the top, there's a navigation bar with 'UniProt' logo and 'UniParc' dropdown. Below it, a search bar with 'Advanced' and 'Search' buttons. A secondary navigation bar includes 'BLAST', 'Align', 'Retrieve/ID mapping', 'Peptide search', 'Help', and 'Contact'. The main heading is 'UniParc results'. A text box explains that UniParc is a comprehensive and non-redundant database containing most of the publicly available protein sequences in the world, and that a UPI is never removed, changed, or reassigned. Below this, there's a table with columns: Entry, Organisms, UniProtKB, First seen, Last seen, and Length. The table shows two entries for Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR)) and one entry for Vaccinia virus (strain Copenhagen) (VACV) (Vaccinia virus (strain WR)).

Entry	Organisms	UniProtKB	First seen	Last seen	Length
<input type="checkbox"/> UP10000000001	Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR)) Horsepox virus (HSPV) Rabbitpox virus Vaccinia virus Cowpox virus (CPV) synthetic construct	P07612 P07612.1 (obsolete) A0A2I2H0D11 Q0GN26 Q6R2L4 Q71TT2 Q76QK2 Q76ZT7.1 (obsolete)	1988-11-01	2019-10-16	250
<input type="checkbox"/> UP10000000002	Vaccinia virus (strain Copenhagen) (VACV) Vaccinia virus (strain Western Reserve) (VACV) (Vaccinia virus (strain WR)) Horsepox virus (HSPV) Rabbitpox virus Vaccinia virus Vaccinia virus WAU86/88-1	P68616 P68617 P21056.1 (obsolete) A0A2I2MCB0 Q0GN58 Q6R2F0 Q71TT1 V5Q2I4	1991-02-01	2019-10-16	185



Análises

Característica hereditária das informações dificulta busca por grupos com algo em comum



Pesquisas em camadas, por taxonomia, nome da proteína e afins



Proposta final:

Problema: Algumas pessoas possuem alergias a proteínas



Proposta: Analisar as proteínas e suas propriedades relacionadas a algum tipo de vacina



Exemplo

xquery

```
let $doc := doc("uniparc-250.xml")
```

```
for $p in $doc//*:dbReference
```

```
  where contains(data($p/*:property[@type= 'gene_name']/@value), 'VAC') and  
    $p/*:property[@type='NCBI_taxonomy_id']/@value != '10245'
```

```
return $p
```

```
<dbReference xmlns="http://uniprot.org/uniparc" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" type="UniProtKB/Swiss-Prot">  
  <property type="NCBI_GI" value="55977540"/>  
  <property type="NCBI_taxonomy_id" value="10254"/>  
  <property type="protein_name" value="Protein E7"/>  
  <property type="gene_name" value="VACWR063"/>  
</dbReference>  
<dbReference xmlns="http://uniprot.org/uniparc" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" type="EMBL" id="AA089342">  
  <property type="NCBI_GI" value="29692169"/>  
  <property type="NCBI_taxonomy_id" value="10254"/>  
  <property type="protein_name" value="soluble, myristylprotein"/>  
  <property type="gene_name" value="VACWR063"/>  
  <property type="proteome_id" value="UP000000344"/>  
  <property type="component" value="Genome"/>  
</dbReference>  
<dbReference xmlns="http://uniprot.org/uniparc" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" type="UniProtKB/Swiss-Prot">  
  <property type="NCBI_GI" value="137632"/>  
  <property type="NCBI_taxonomy_id" value="10254"/>  
  <property type="protein_name" value="Probable FAD-linked sulphhydryl oxidase E10"/>  
  <property type="gene_name" value="VACWR066"/>  
</dbReference>
```



Problema

A base escolhida não possui muitos níveis hierárquicos pois foi projetada para ser funcional também em uma abordagem relacional

Etap 4: XQuery

TB Database



Base de dados

http://tbdb.bu.edu/tbdb_sysbio/Downloads.html

- TBD (Tuberculosis Database)
- Banco de genes relacionados a tuberculose
- Necessário conversão de dados

The screenshot shows the TB Database website interface. At the top, there's a navigation bar with the logo 'TB Database' and the tagline 'AN INTEGRATED PLATFORM FOR TUBERCULOSIS RESEARCH'. Below the navigation bar, there's a 'QUICK SEARCH' box with a search button. The main content area is titled 'Download Data' and lists various Mycobacterium species and strains with links to download their data as zip files. The list includes:

Download Data by Organism	Download Link
M. tuberculosis H37Rv (GB:AL123456)	zip file
M. tuberculosis CDC1551	zip file
M. tuberculosis F11	zip file
M. tuberculosis C	zip file
M. tuberculosis Haarlem (draft)	zip file
M. tuberculosis H37Ra	zip file
M. africanum GHO411182	zip file
M. bovis AF2122/97	zip file
M. bovis BCG	zip file
M. leprae TN	zip file
M. marinum	zip file
M. ulcerans Agy99	zip file
M. avium 104	zip file
M. avium k10	zip file
M. sp. MCS	zip file
M. sp. KMS	zip file
M. smegmatis MC2 155	zip file
M. vanbaalenii PYR-1	zip file
M. gilvum PYR-GCK	zip file
M. abscessus	zip file
M. sp. JLS	zip file
R. jostii RHA1	zip file
N. farcinica IFM 10152	zip file
C. glutamicum ATCC 13032	zip file
C. efficiens YS-314	zip file
C. diphtheriae NCTC 13129	zip file
C. jejuni K411	zip file



Analises

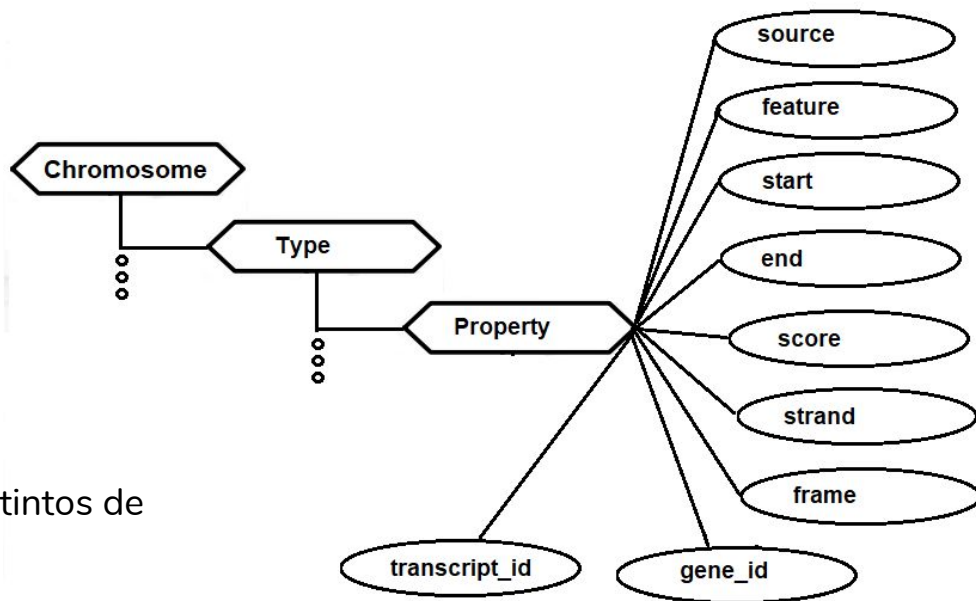
Análise dos genes da Rhodbacter Sphaeroides, um tipo notável de bactéria relacionada a Tuberculose



Um dos organismos mais importantes no estudo da fotossíntese bacteriana, não requer condições incomuns de crescimento e é incrivelmente eficiente em baixos níveis de O_2



Modelo



Pelo menos 6 tipos distintos de cromossomos

Incontáveis propriedades por tipo



Propostas

- Contar o número de propriedades relacionadas a um tipo de cromossomo
- Contar o número de códons de parada(feature = 'stop_codon')
- Retornar de todas as propriedades de um certo tipo que começam depois de uma certa posição (start >1200)
- Encontrar as propriedades de um certo tipo relativas a um gene específico
- Retornar propriedades com tamanho pequeno (start - end <10)
- Retornar todos os genes de um certo tipo que são códons de começo e parada



Exemplo 1: Propriedades relativas a um mesmo gene

xquery

```
let $doc := doc("file2.xml")
```

```
for $c in ($doc//type[@value = 4]/property)
```

```
where $c/@gene_id = 'RSP_2483'
```

```
return $c
```

```
<property source="RSP1_INSERTASSEMBLYFROMGENBANK_1" feature="start_codon" start="71408" end="71410"
score="." strand="+" frame="0" gene_id="RSP_2483" transcript_id="null"/>
<property source="RSP1_INSERTASSEMBLYFROMGENBANK_1" feature="stop_codon" start="72713" end="72715"
score="." strand="+" frame="0" gene_id="RSP_2483" transcript_id="null"/>
<property source="RSP1_INSERTASSEMBLYFROMGENBANK_1" feature="exon" start="71408" end="72715" score="."
strand="+" frame="." gene_id="RSP_2483" transcript_id="null"/>
<property source="RSP1_INSERTASSEMBLYFROMGENBANK_1" feature="CDS" start="71408" end="72712" score="."
strand="+" frame="0" gene_id="RSP_2483" transcript_id="null"/>
```



Exemplo 2: Propriedades que ocupam uma porção pequena do Cromossomo

xquery

```
let $doc := doc("file2.xml")
```

```
for $c in ($doc//type[@value = 1]/property)
```

```
where $c/@end - $c/@start <= 10
```

```
return $c
```

```
<property source="RSP1_INSERTASSEMBLYFROMGENBANK_1" feature="start_codon" start="941208" end="941210"
score="." strand="+" frame="0" gene_id="RSP_1425" transcript_id="null"/>
<property source="RSP1_INSERTASSEMBLYFROMGENBANK_1" feature="start_codon" start="39397" end="39399"
score="." strand="+" frame="0" gene_id="RSP_3003" transcript_id="null"/>
<property source="RSP1_INSERTASSEMBLYFROMGENBANK_1" feature="start_codon" start="39958" end="39960"
score="." strand="+" frame="0" gene_id="RSP_3004" transcript_id="null"/>
<property source="RSP1_INSERTASSEMBLYFROMGENBANK_1" feature="stop_codon" start="40564" end="40566"
score="." strand="+" frame="0" gene_id="RSP_3004" transcript_id="null"/>
<property source="RSP1_INSERTASSEMBLYFROMGENBANK_1" feature="start_codon" start="41165" end="41167"
score="." strand="-" frame="0" gene_id="RSP_3005" transcript_id="null"/>
<property source="RSP1_INSERTASSEMBLYFROMGENBANK_1" feature="start_codon" start="41397" end="41399"
score="." strand="+" frame="0" gene_id="RSP_3006" transcript_id="null"/>
```


Etapas Final: Neo4j

Virus-Host DB

Base de dados

<https://www.genome.jp/virushostdb/>

- Virus-Host DB é uma base de dados sobre a relação entre os vírus e seus hospedeiros, com informações do genoma e mais.




Virus-Host DB

[About](#) | [Statistics](#) | [Browse](#) | [Virus Index](#) | [Host Index](#) | [Feedback](#)

Virus-Host DB organizes data about the relationships between viruses and their hosts, represented in the form of pairs of NCBI taxonomy IDs for viruses and their hosts. Virus-Host DB covers viruses with complete genomes stored in 1) NCBI/RefSeq and 2) GenBank whose accession numbers are listed in EBI Genomes. The host information is collected from RefSeq, GenBank (in free text format), UniProt, ViralZone, and manually curated with additional information obtained by literature surveys.

➡ [Browse all viruses](#)

➡  [Human viruses](#)

[Other major hosts](#)
[Animal | Plant | Other Eukaryotes | Bacteria | Archaea]

RefSeq GenBank (release 96, September 9, 2019)
(release 233.0, August 15, 2019)


Viral entries
(No. of sequence accessions)


14688


Source of host information (No. of evidence records)


RefSeq + GenBank	UniProt	Viral Zone	Manual annotation	Total
9304	1644	200	2760	14469


Selected viral families


 Adenoviridae

 Papillomaviridae

 Herpesviridae

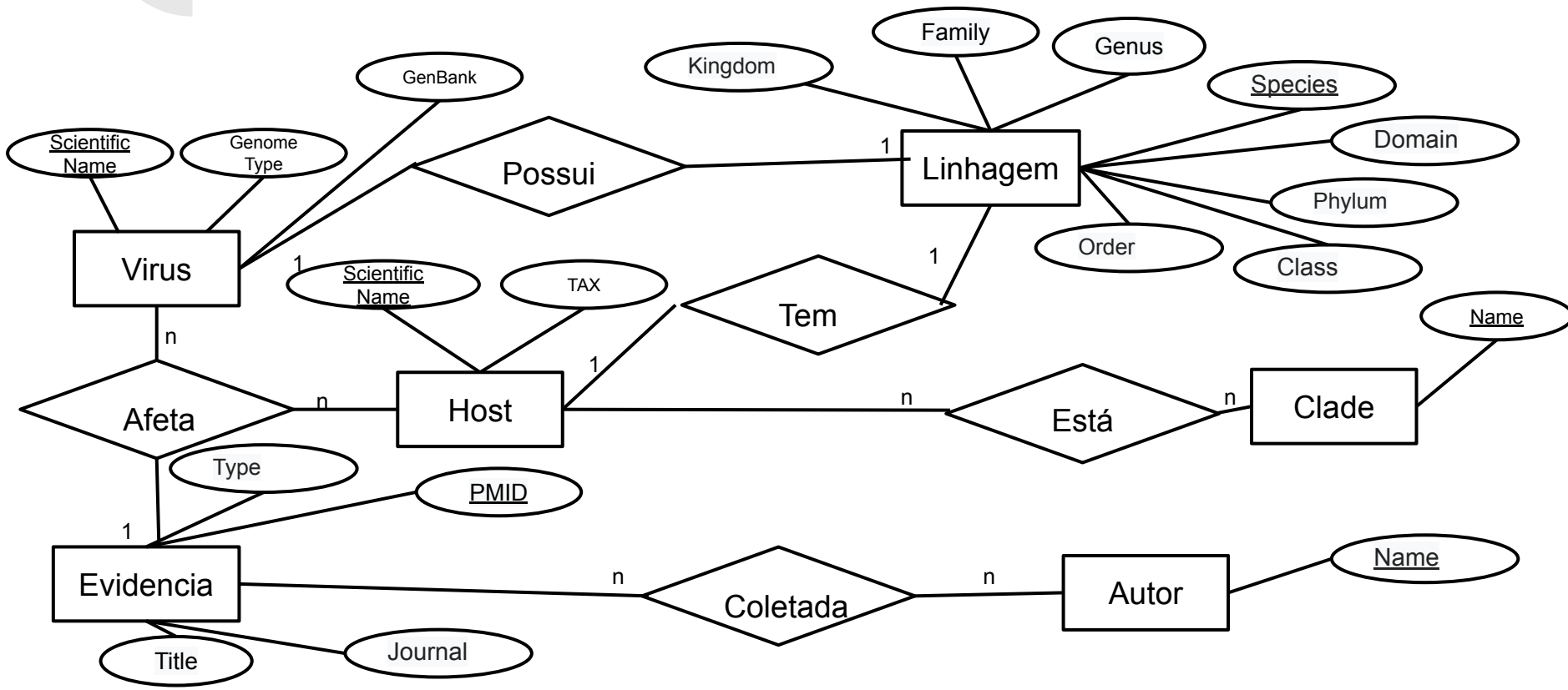
 Baculoviridae

 Poxviridae

 Parvoviridae



Modelo Conceitual - ER





Análises

```
USING PERIODIC COMMIT 10
LOAD CSV WITH HEADERS FROM
'file:///virushostdb.csv' AS line
WITH line
WHERE NOT line.host_tax_id IS NULL
MERGE (v:Virus {virus_name:line.virus_name,virus_id:line.virus_tax_id})
MERGE (h:Host{host_id:line.host_tax_id,host_name:line.host_name})
MERGE (v)-[i:Infects]->(h)
```

- Comandos usados na criação do grafo no Neo4j, foi necessário o uso do comando MERGE() para evitar duplicatas.
- É importante notar que não utilizamos todos campos da base de dados originais, apenas os campos Id e Nome tanto dos hospedeiros quanto dos virus.



Exemplo 1

```
MATCH p=(v:Virus)-[r:Infects]->(h:Host {host_name:"Homo sapiens"})  
RETURN count(v)
```

- Consulta simples que retorna todos vírus que tem como hospedeiro o ser humano (espécie “Homo sapiens”).

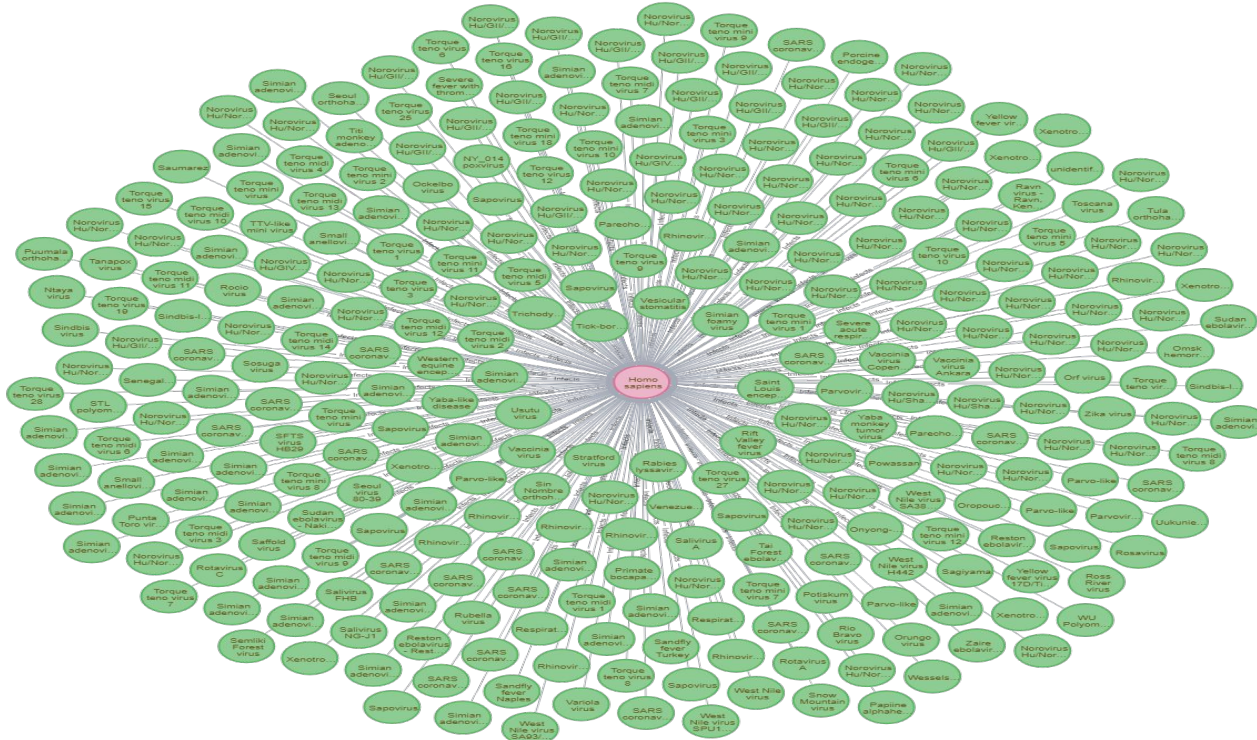


Foto
ilustrativa da
consulta do
exemplo 1.
Para apenas
300 valores
(são mais de
1300).



Exemplo 2: Teste de comunidades.

```
CALL algo.unionFind('Host','Infects', {  
  write: true,  
  writeProperty: 'community'  
})  
YIELD nodes, setCount, loadMillis, computeMillis, writeMillis
```

- Aqui utilizamos um dos algoritmos disponibilizados pelo Neo4j para fazer a análise por comunidades como um teste e, pudemos constatar que nosso grafo não se adaptou bem a este tipo de problema, tendo em vista que o número de comunidades encontradas foi o mesmo do número de nós de hospedeiros.