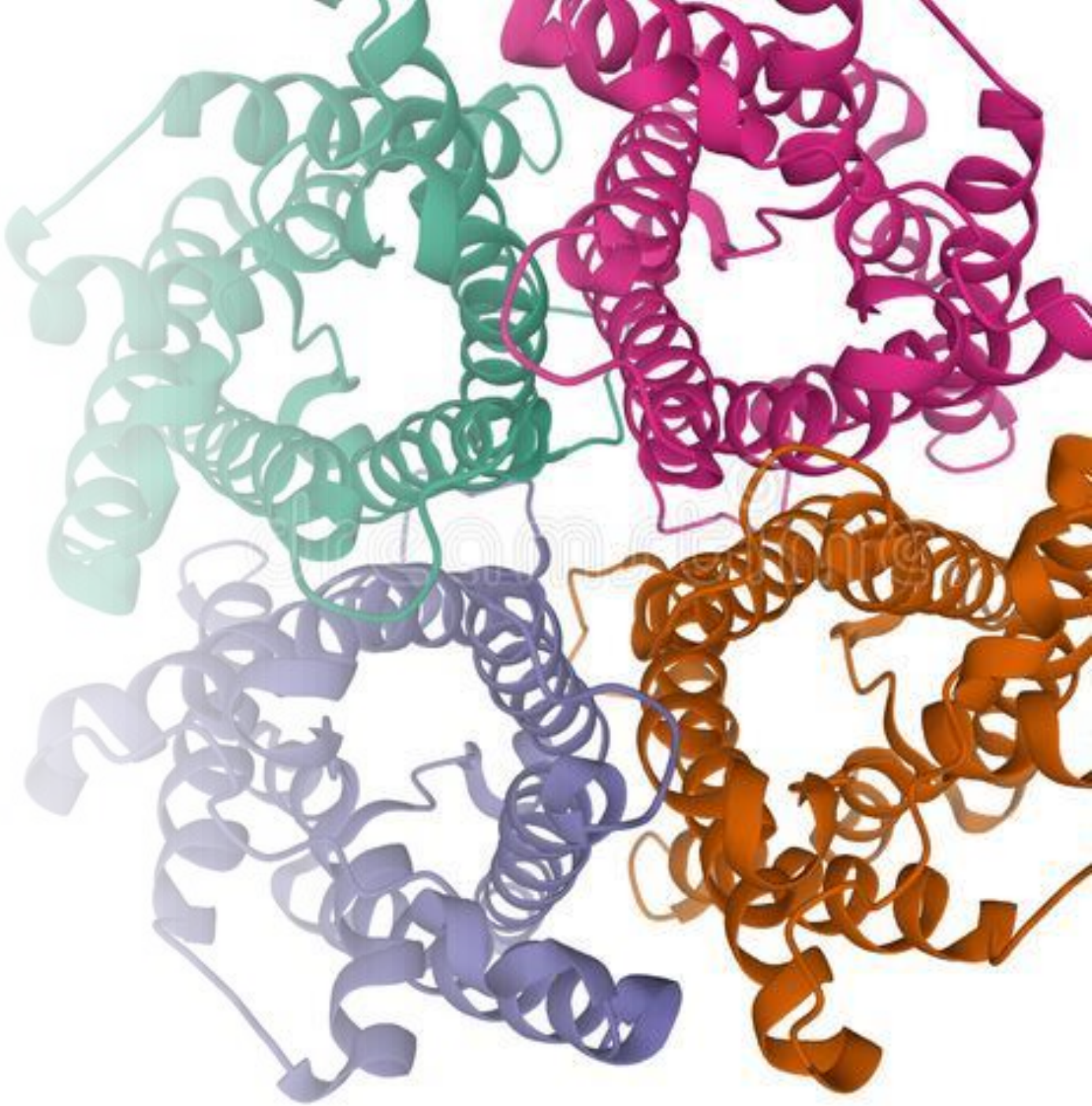
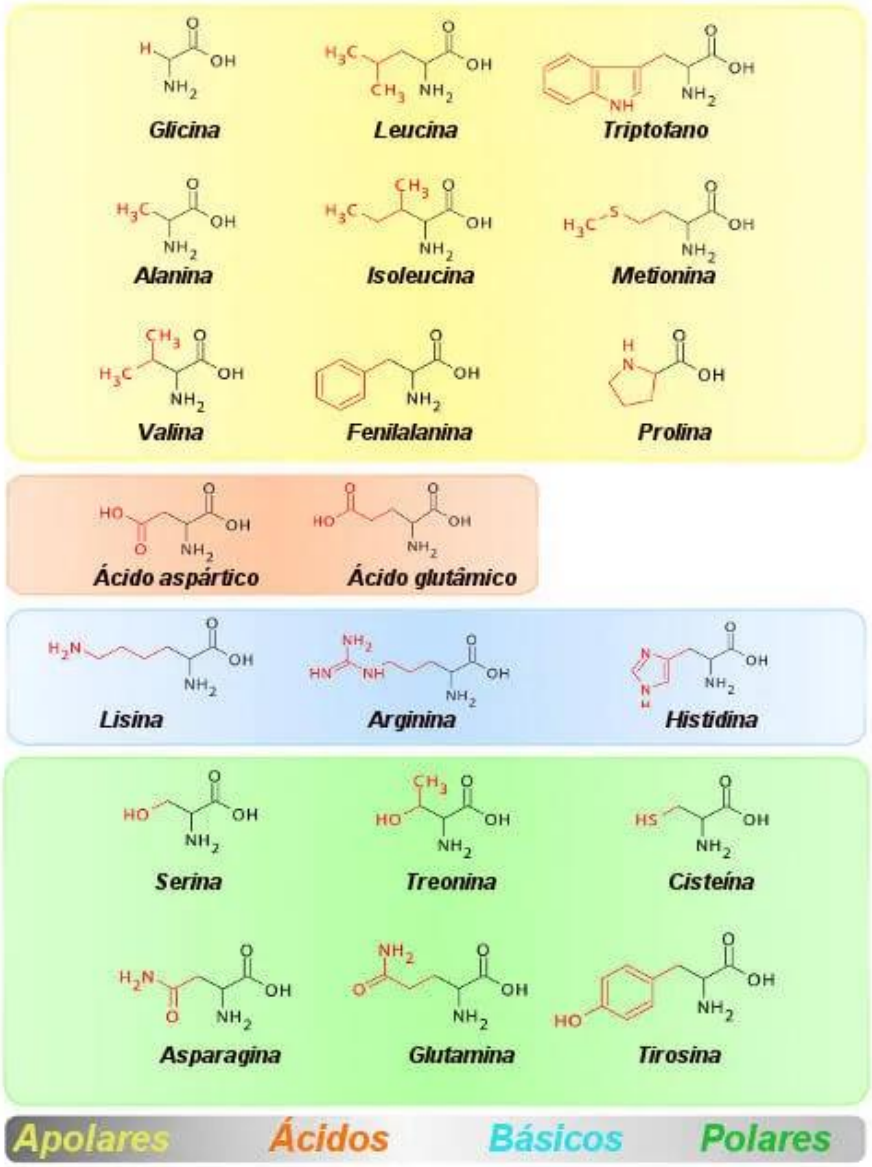
Introdução à Bioinformática de Proteínas

Ignacio G. Barroso

QBQ2507/IBI5035 Biologia Molecular Computacional

22/09/22

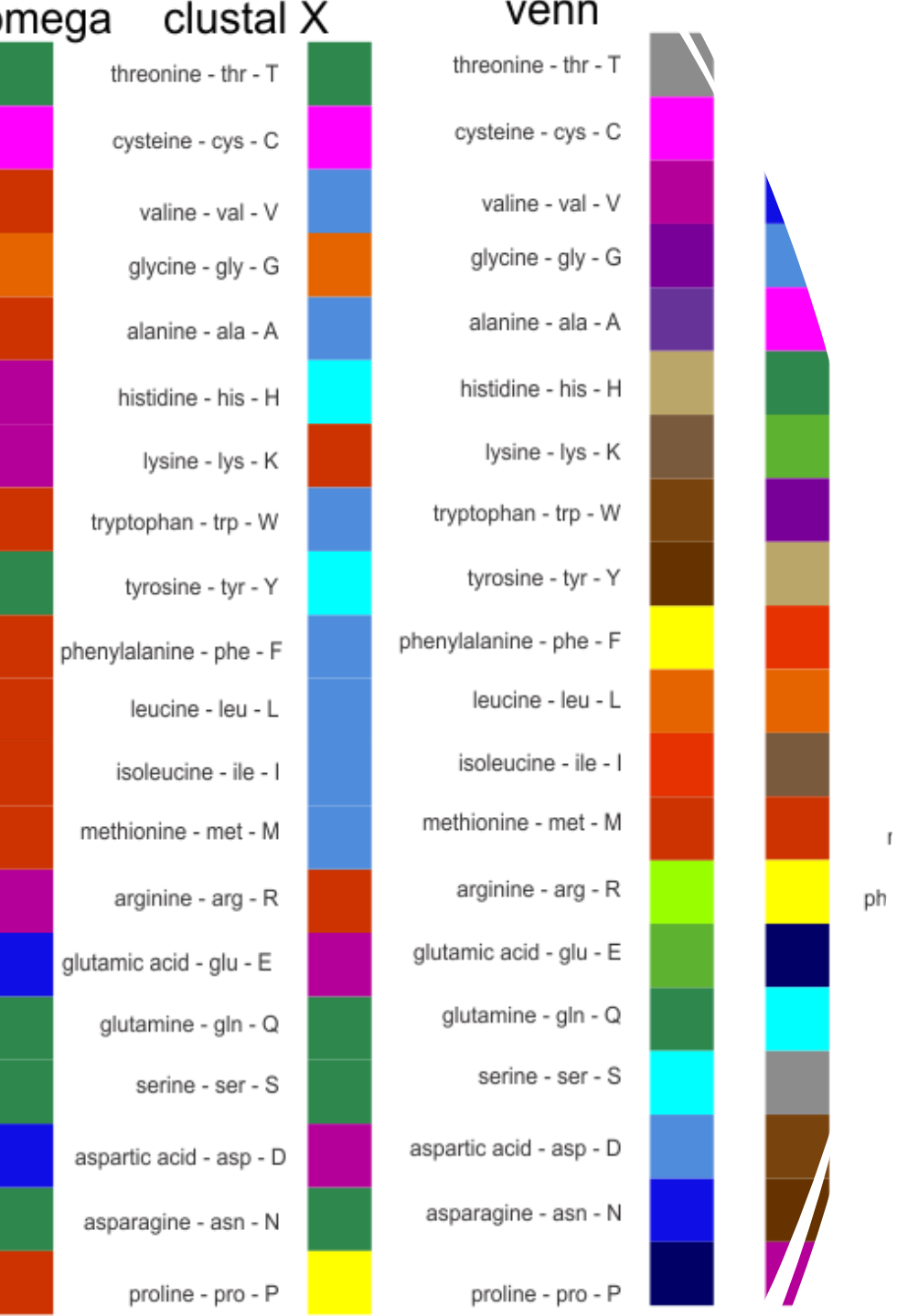




Proteínas

- O componente celular mais abundante.
- As moléculas mais diversas em estrutura e função.
- Ideais para estudos evolutivos comparados.

- Polímeros de aminoácidos
  - 20 aminoácidos diferenciados pela cadeia lateral.
  - Combinações infinitas.



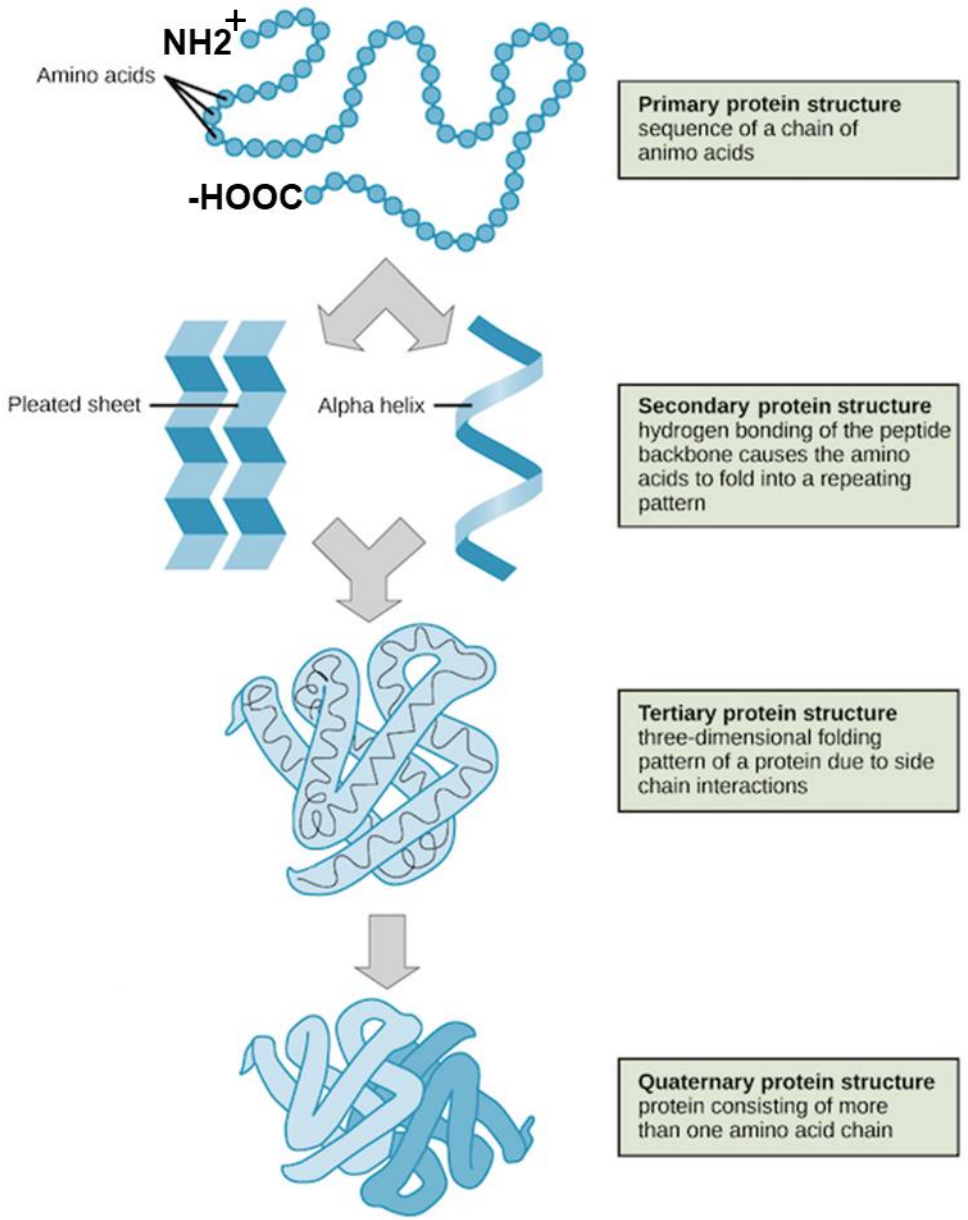
Aminoácidos podem ser escritos ou representados de diferentes maneiras.

- Nome: Triptofano
- Nome de três letras: Trp
- Nome de uma letra: W
- Cor: Clustal X

Estrutura de Proteínas

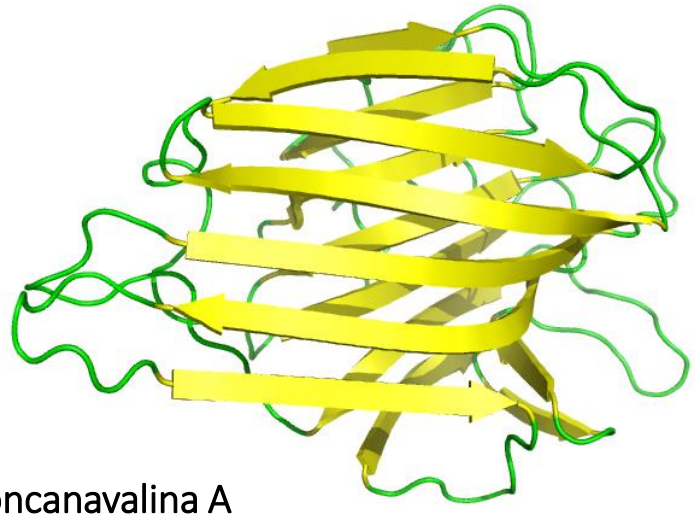
A sequência de aminoácidos determina a estrutura espacial da proteína.

- Quatro níveis de estrutura:
  - Primária: Sequência de aminoácidos.
    - Determinada geneticamente.
  - Secundária: Estruturas tridimensionais **regulares** de segmentos da cadeia polipeptídica.
    - Tipos:
      - α hélice
      - Folha β
    - Forças: Ligações de hidrogênio
  - Terciária: Dobramento final da cadeia polipeptídica por interação das estruturas secundárias e regiões de estrutura não definida. Proteína ativa.
    - Forças:
      - Ligações de Hidrogênio
      - Hidrofóbicas
      - Ligações iônicas ou salinas
      - Forças de van der Waals
  - Quaternária: Associação de duas ou mais cadeias polipeptídicas para compor a proteína funcional.

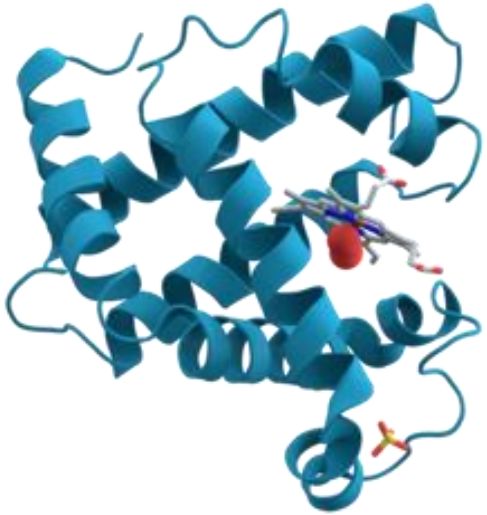


A proporção de estruturas secundárias nas proteínas varia muito

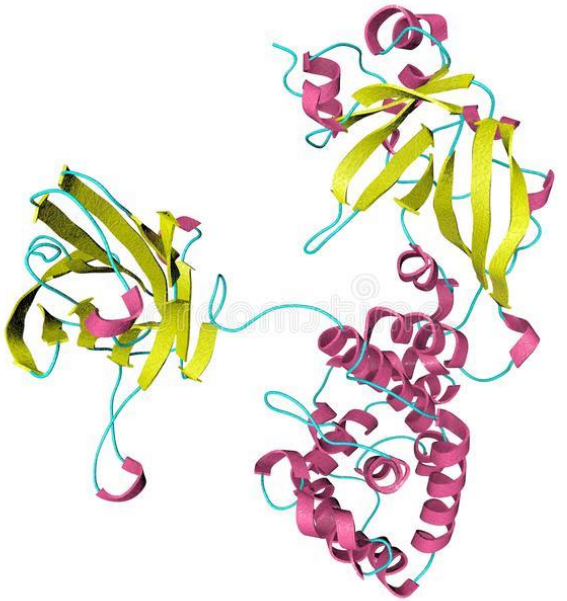
Estruturas terciárias de proteínas



Concanavalina A



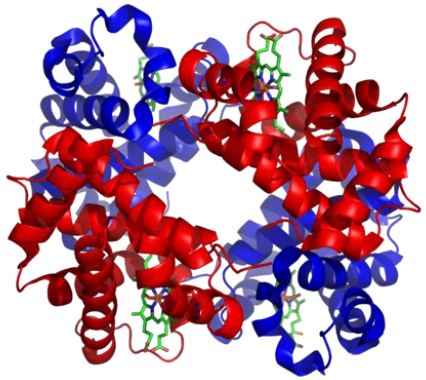
Mioglobina



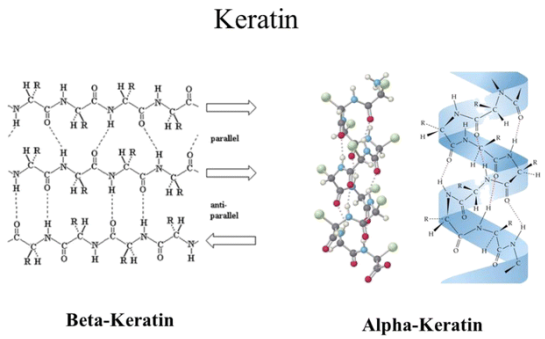
Toxina Diftérica

Proteínas podem ser classificadas estruturalmente

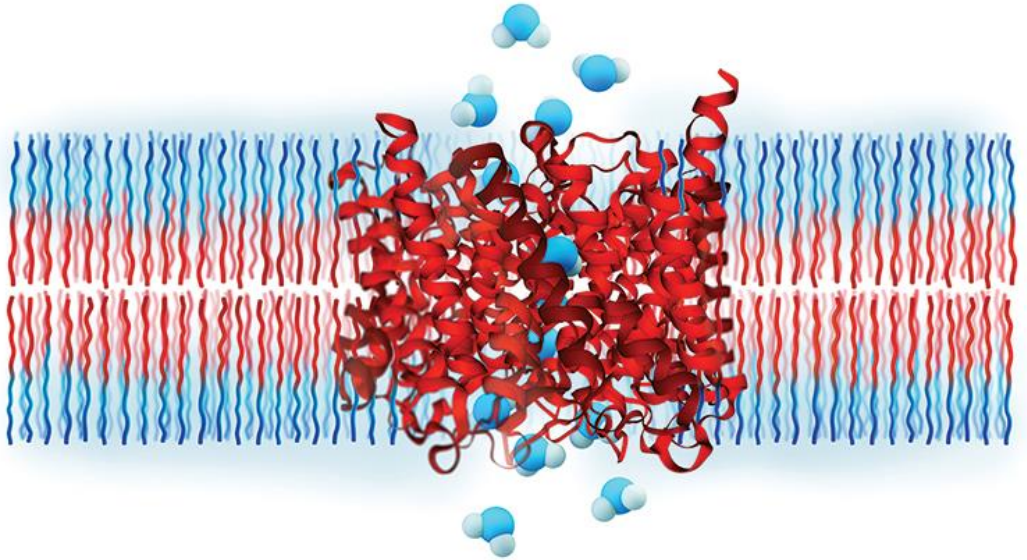
- Fibrosas: Papel estrutural.
- Globulares: Diferentes papeis.
  - *Enzimas*
  - *Transporte*
  - *Hormônios*
  - *Anticorpos*
- Transmembrana: Diferentes papeis.
  - *Enzimas*
  - *Transporte*
  - *Sinalização*



Hemoglobina



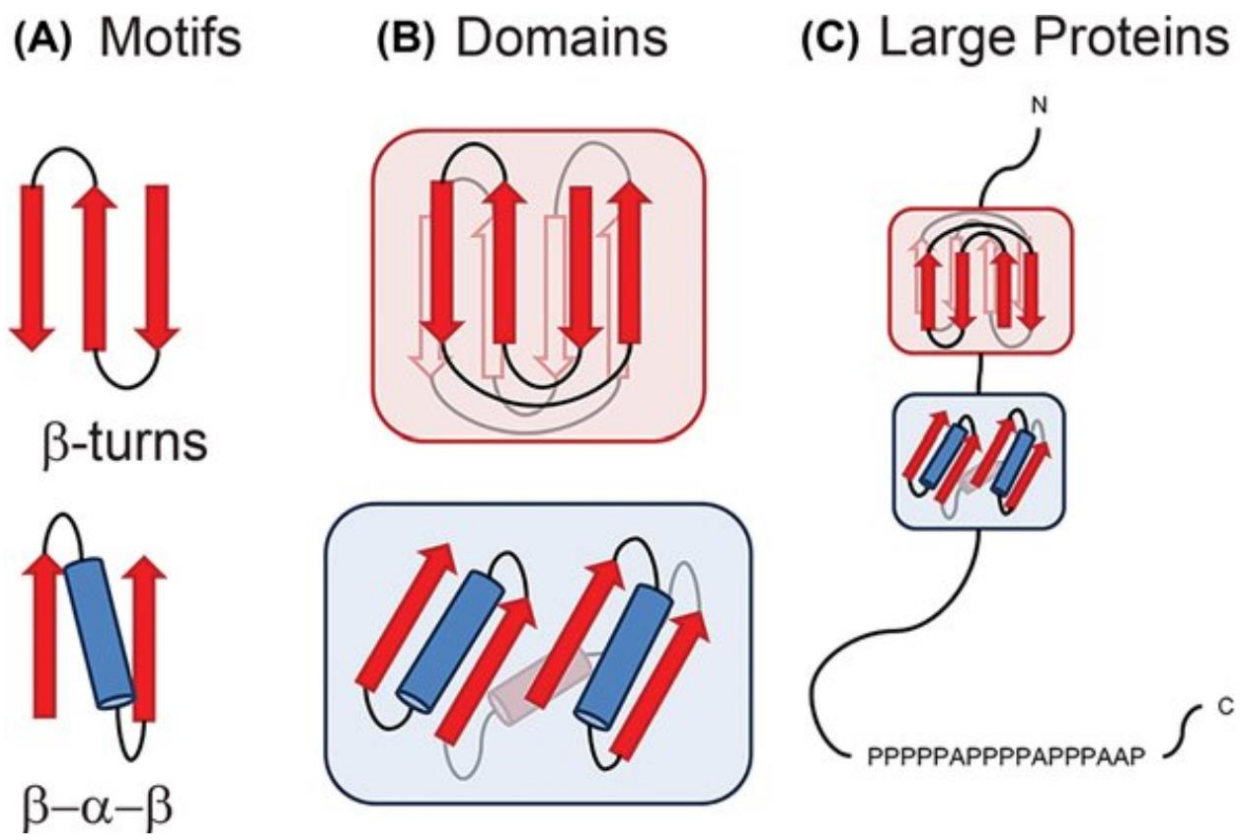
Queratina



Aquaporina

Estrutura

- A estrutura **terciaria** pode conter **domínios e motivos**.
- Padrões de elementos estruturais que se **repetem** em proteínas diferentes.
- Motivos são diferentes formas de organização da estrutura secundária.
- As proteínas longas se organizam em domínios, ou conjuntos estruturais definidos, formados por dobramentos da cadeia polipeptídica.
- Domínios podem estar ligados por segmentos flexíveis que permitem o movimento da proteína. Ex: Enzimas, transportadores de membrana.
- Domínios frequentemente são funcionais.
- Domínios com a mesma função apresentam estruturais similares em proteínas diferentes.
- **Motivos sequência**: segmentos curtos conservados presentes em muitas proteínas com significado funcional específico



Bancos de Dados de Proteínas

PIR
A UniProt CONSORTIUM MEMBER
Protein Information Resource

BRENDA
The Comprehensive Enzyme Information System

NCBI

Transporter Classification Database
TCDB is operated by the Saier Lab Bioinformatics

UniProt

WORLDWIDE PDB
PROTEIN DATA BANK

EXProt
CMBI

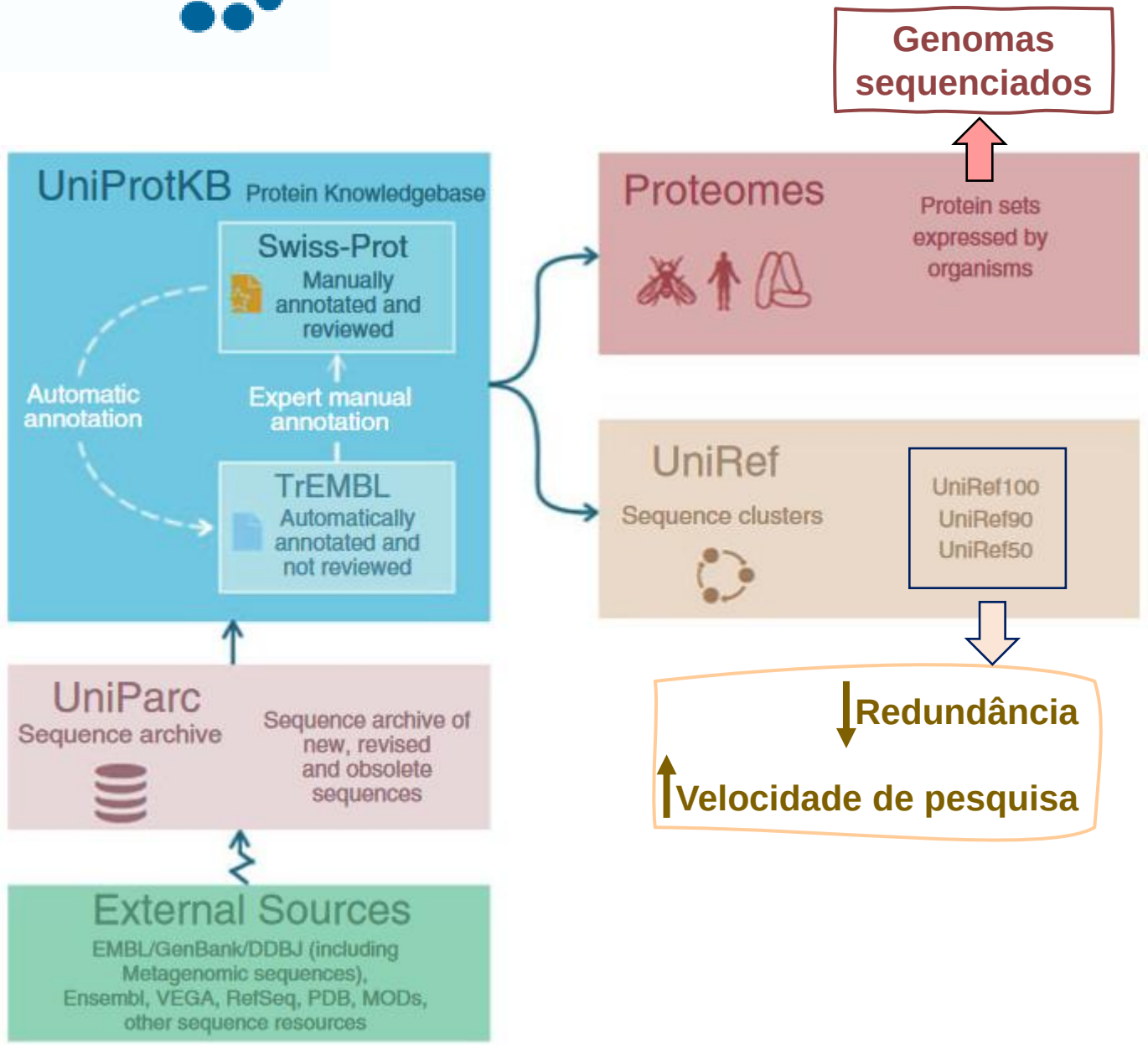
- **Banco de dados de anotação e sequências de proteínas gratuito**.

- Conjuntos de dados:

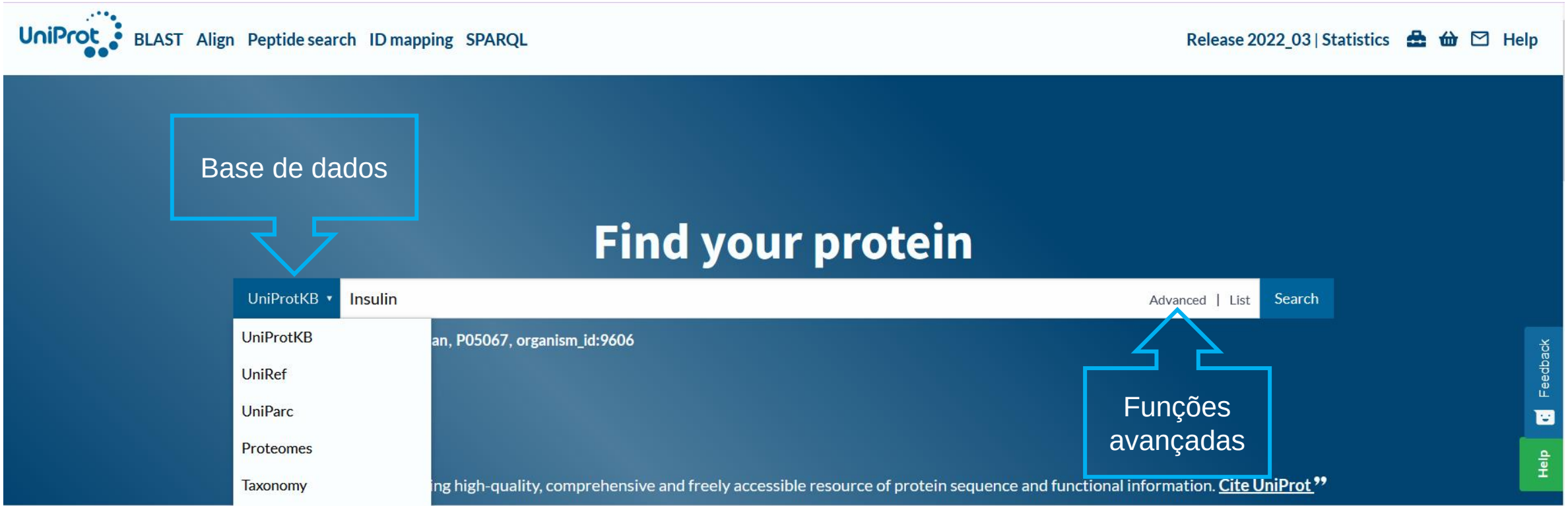
  - **UniProtKB** reúne todas as informações funcionais sobre proteínas em duas seções:
    - **Swiss-Prot:** Anotadas manualmente e revisadas.
    - TrEMBL: Computacionalmente anotadas e não revisadas.

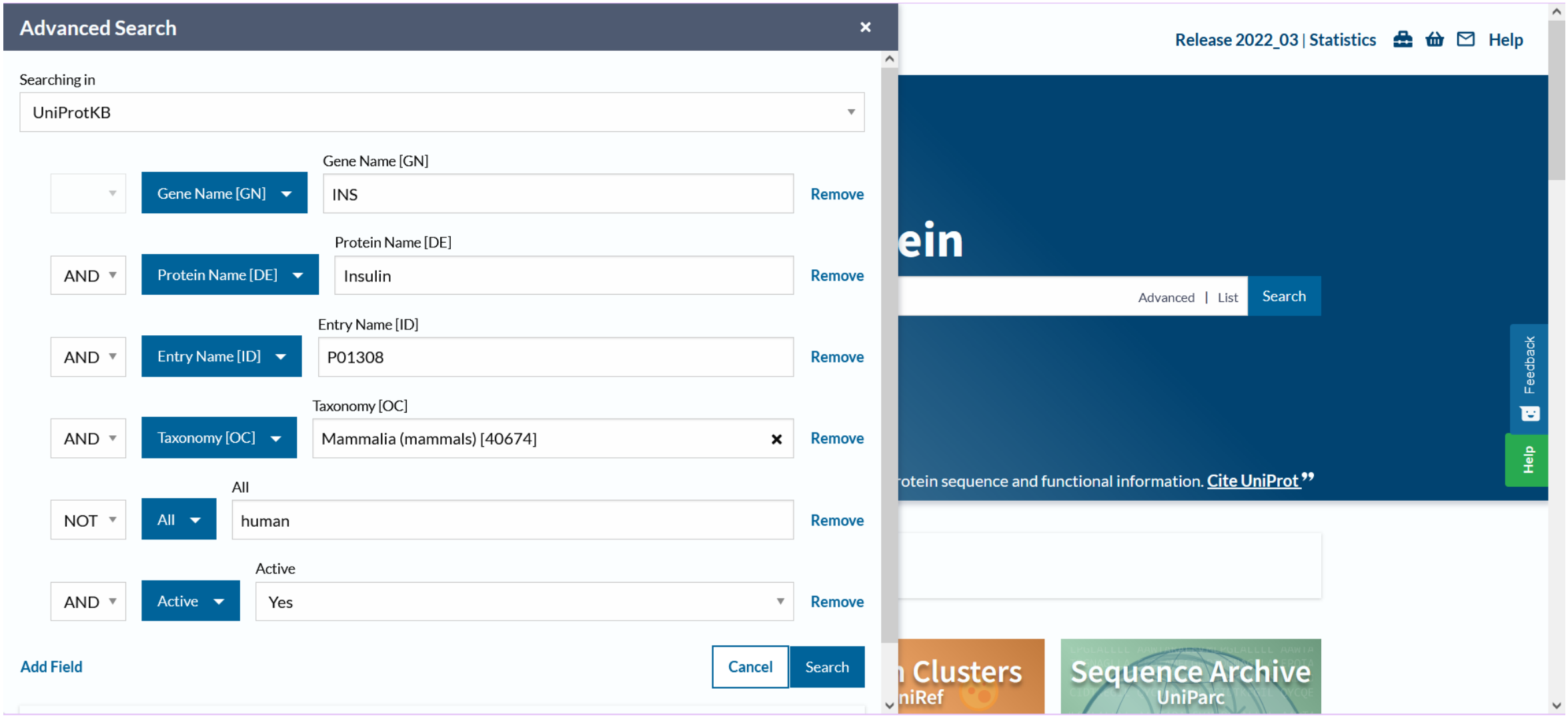
- Outros **recursos**:
  - Literatura
  - Taxonomia
  - Palavras-chave
  - Localizações subcelulares
  - Referências cruzadas
  - Doenças humanas

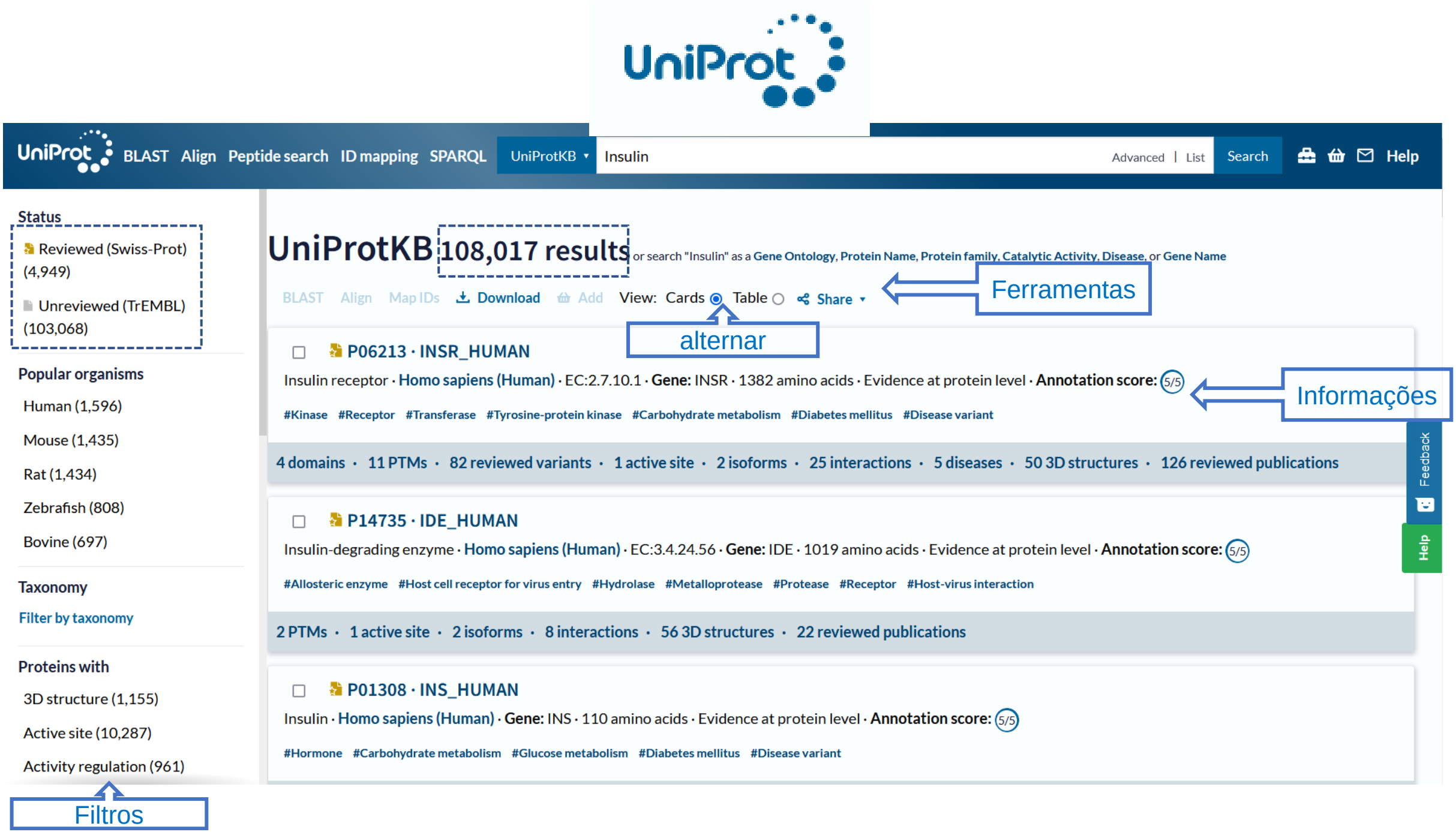
- **Ferramentas**:
  - Blast
  - Align: alinhamentos múltiplos
  - Retrieve/ID: Pacotes de dados



www.uniprot.org




Advanced Search
Searching in
UniProtKB
Gene Name [GN]
Gene Name [GN]
INS
Remove
Protein Name [DE]
AND
Protein Name [DE]
Insulin
Remove
Entry Name [ID]
AND
Entry Name [ID]
P01308
Remove
Taxonomy [OC]
AND
Taxonomy [OC]
Mammalia (mammals) [40674]
Remove
All
NOT
All
human
Remove
Active
AND
Active
Yes
Remove
Add Field
Cancel
Search
Release 2022_03 | Statistics
Help
ein
Advanced | List
Search
otein sequence and functional information. Cite UniProt
Feedback
Help
Clusters
niRef
Sequence Archive
UniParc


UniProt
BLAST Align Peptide search ID mapping SPARQL
UniProtKB
Insulin
Advanced | List
Search
Help
Status
Reviewed (Swiss-Prot) (4,949)
Unreviewed (TrEMBL) (103,068)
Popular organisms
Human (1,596)
Mouse (1,435)
Rat (1,434)
Zebrafish (808)
Bovine (697)
Taxonomy
Filter by taxonomy
Proteins with
3D structure (1,155)
Active site (10,287)
Activity regulation (961)
Filtros
UniProtKB 108,017 results
or search "Insulin" as a Gene Ontology, Protein Name, Protein family, Catalytic Activity, Disease, or Gene Name
BLAST Align Map IDs Download Add View: Cards Table Share
Ferramentas
alternar
P06213 · INSR_HUMAN
Insulin receptor · Homo sapiens (Human) · EC:2.7.10.1 · Gene: INSR · 1382 amino acids · Evidence at protein level · Annotation score: 5/5
#Kinase #Receptor #Transferase #Tyrosine-protein kinase #Carbohydrate metabolism #Diabetes mellitus #Disease variant
4 domains · 11 PTMs · 82 reviewed variants · 1 active site · 2 isoforms · 25 interactions · 5 diseases · 50 3D structures · 126 reviewed publications
Informações
P14735 · IDE_HUMAN
Insulin-degrading enzyme · Homo sapiens (Human) · EC:3.4.24.56 · Gene: IDE · 1019 amino acids · Evidence at protein level · Annotation score: 5/5
#Allosteric enzyme #Host cell receptor for virus entry #Hydrolase #Metalloprotease #Protease #Receptor #Host-virus interaction
2 PTMs · 1 active site · 2 isoforms · 8 interactions · 56 3D structures · 22 reviewed publications
P01308 · INS_HUMAN
Insulin · Homo sapiens (Human) · Gene: INS · 110 amino acids · Evidence at protein level · Annotation score: 5/5
#Hormone #Carbohydrate metabolism #Glucose metabolism #Diabetes mellitus #Disease variant
Feedback
Help

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB Insulin Advanced | List Search Help

Status

Reviewed (Swiss-Prot) (4,949)

Unreviewed (TrEMBL) (103,068)

Popular organisms

Human (1,596)

Mouse (1,435)

Rat (1,434)

Zebrafish (808)

Bovine (697)

Taxonomy

Filter by taxonomy

Proteins with

3D structure (1,155)

Active site (10,287)

Activity regulation (961)

Allergen (8)

UniProtKB 108,017 results

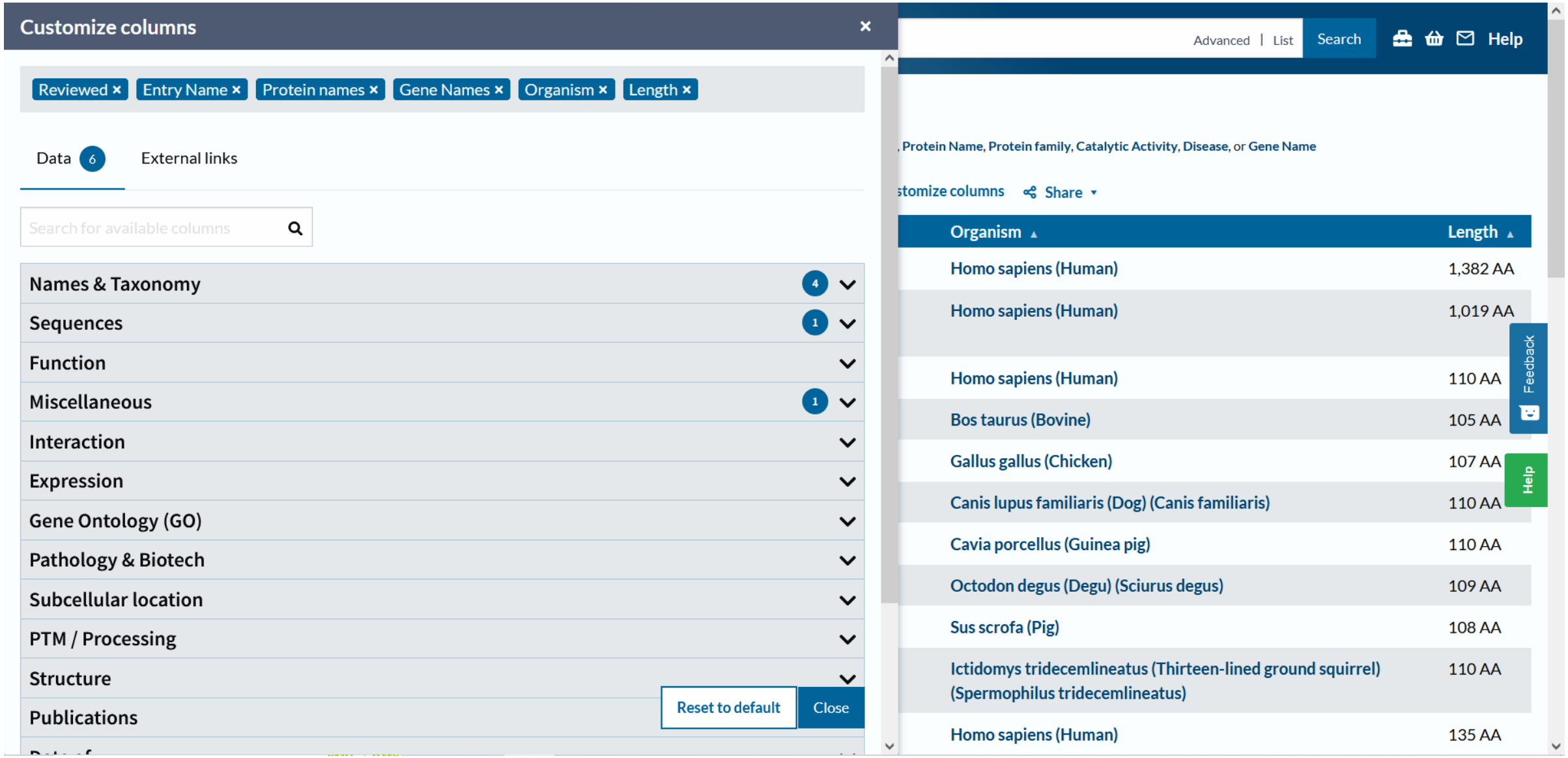
or search "Insulin" as a Gene Ontology, Protein Name, Protein family, Catalytic Activity, Disease, or Gene Name

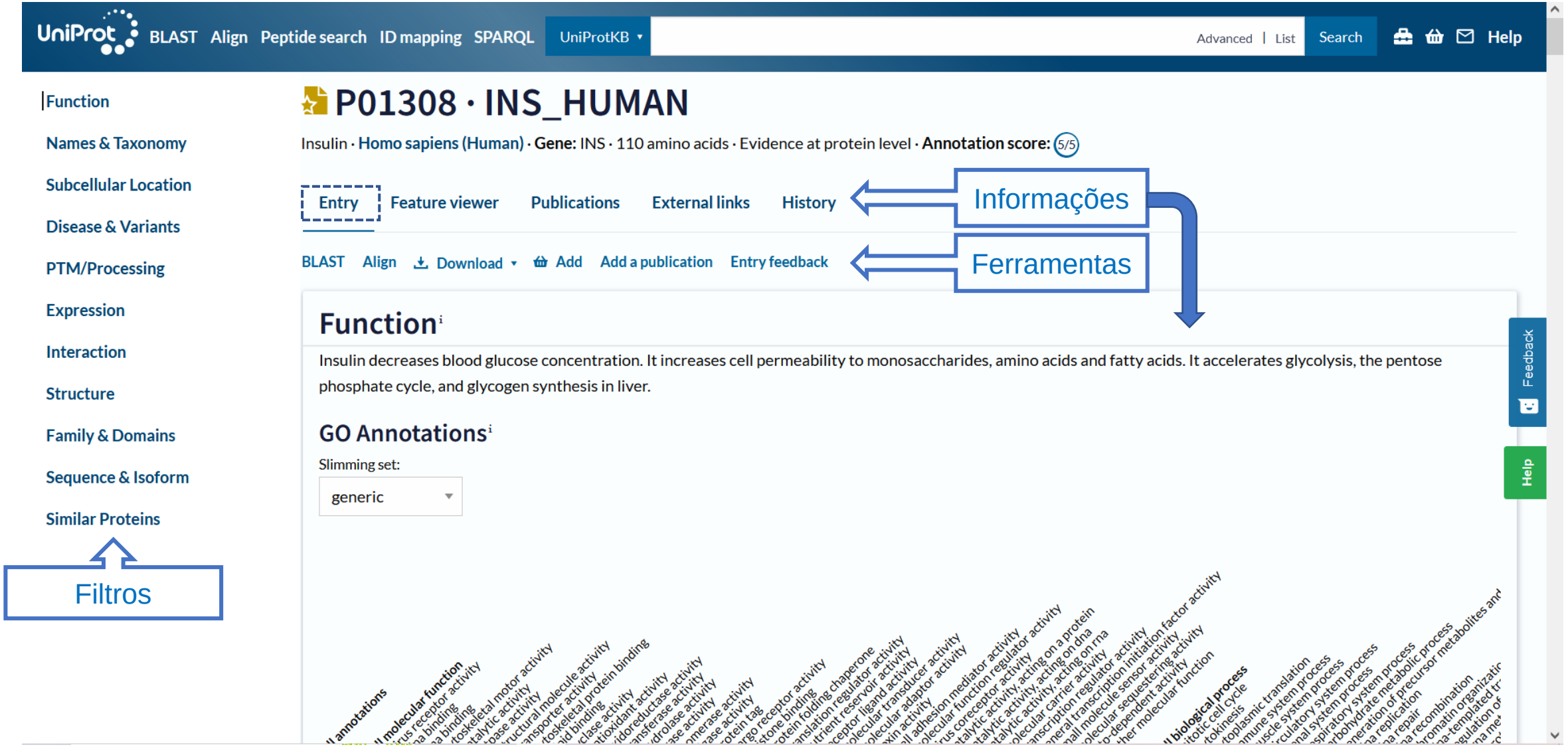
BLAST Align Map IDs Download Add View: Cards ○ Table ◉ Customize columns Share

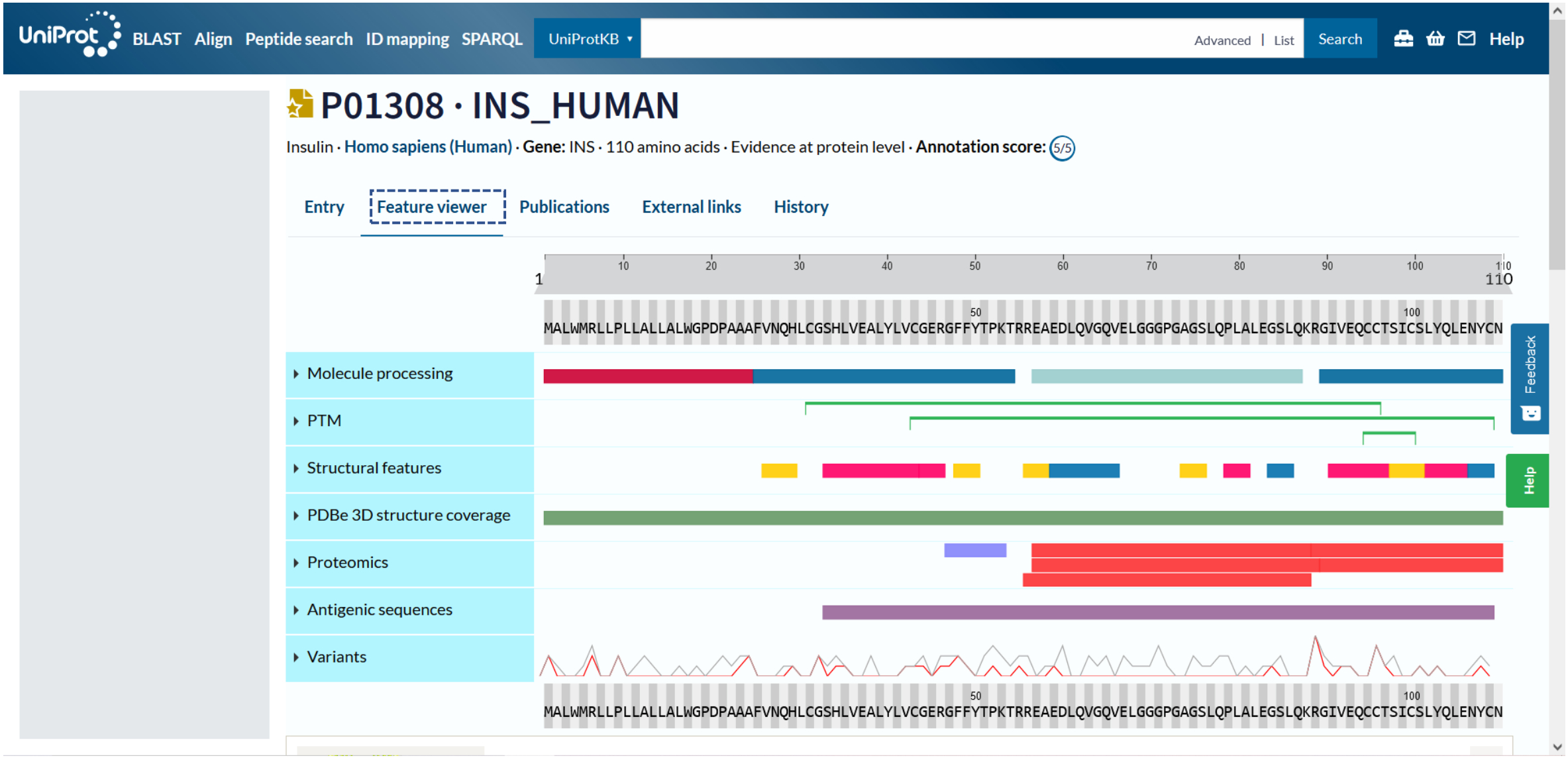
Entry	Entry Name	Protein Names	Gene Names	Organism	Length
P06213	INSR_HUMAN	Insulin receptor[...]	INSR	Homo sapiens (Human)	1,382 AA
P14735	IDE_HUMAN	Insulin-degrading enzyme[...]	IDE	Homo sapiens (Human)	1,019 AA
P01308	INS_HUMAN	Insulin	INS	Homo sapiens (Human)	110 AA
P01317	INS_BOVIN	Insulin	INS	Bos taurus (Bovine)	105 AA
P67970	INS_CHICK	Insulin	INS	Gallus gallus (Chicken)	107 AA
P01321	INS_CANLF	Insulin	INS	Canis lupus familiaris (Dog) (Canis familiaris)	110 AA
P01329	INS_CAVPO	Insulin	INS	Cavia porcellus (Guinea pig)	110 AA
P17715	INS_OCTDE	Insulin	INS	Octodon degus (Degu) (Sciurus degus)	109 AA
P01315	INS_PIG	Insulin	INS	Sus scrofa (Pig)	108 AA
Q91XI3	INS_ICTTR	Insulin	INS	Ictidomys tridecemlineatus (Thirteen-lined ground squirrel) (Spermophilus tridecemlineatus)	110 AA
Q9Y5Q6	INSL5_HUMAN	Insulin-like peptide	INSL5,	Homo sapiens (Human)	135 AA

Feedback

Help


Customize columns
Reviewed ×
Entry Name ×
Protein names ×
Gene Names ×
Organism ×
Length ×
Data 6
External links
Search for available columns
Names & Taxonomy 4
Sequences 1
Function
Miscellaneous 1
Interaction
Expression
Gene Ontology (GO)
Pathology & Biotech
Subcellular location
PTM / Processing
Structure
Publications
Reset to default
Close
Advanced | List
Search
Help
, Protein Name, Protein family, Catalytic Activity, Disease, or Gene Name
stomize columns
Share
Organism
Length
Homo sapiens (Human) 1,382 AA
Homo sapiens (Human) 1,019 AA
Homo sapiens (Human) 110 AA
Bos taurus (Bovine) 105 AA
Gallus gallus (Chicken) 107 AA
Canis lupus familiaris (Dog) (Canis familiaris) 110 AA
Cavia porcellus (Guinea pig) 110 AA
Octodon degus (Degu) (Sciurus degus) 109 AA
Sus scrofa (Pig) 108 AA
Ictidomys tridecemlineatus (Thirteen-lined ground squirrel) (Spermophilus tridecemlineatus) 110 AA
Homo sapiens (Human) 135 AA
Feedback
Help


UniProt
BLAST
Align
Peptide search
ID mapping
SPARQL
UniProtKB
Advanced | List
Search
Help
Function
Names & Taxonomy
Subcellular Location
Disease & Variants
PTM/Processing
Expression
Interaction
Structure
Family & Domains
Sequence & Isoform
Similar Proteins
Filtros
P01308 · INS_HUMAN
Insulin · Homo sapiens (Human) · Gene: INS · 110 amino acids · Evidence at protein level · Annotation score: 5/5
Entry
Feature viewer
Publications
External links
History
Informações
BLAST
Align
Download
Add
Add a publication
Entry feedback
Ferramentas
Function
Insulin decreases blood glucose concentration. It increases cell permeability to monosaccharides, amino acids and fatty acids. It accelerates glycolysis, the pentose phosphate cycle, and glycogen synthesis in liver.
GO Annotations
Slimming set:
generic
Feedback
Help


UniProt
BLAST
Align
Peptide search
ID mapping
SPARQL
UniProtKB
Advanced | List
Search
Help
P01308 · INS_HUMAN
Insulin · Homo sapiens (Human) · Gene: INS · 110 amino acids · Evidence at protein level · Annotation score: 5/5
Entry
Feature viewer
Publications
External links
History
1
10
20
30
40
50
60
70
80
90
100
110
MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERGFFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLYQLENYCN
Molecule processing
PTM
Structural features
PDBe 3D structure coverage
Proteomics
Antigenic sequences
Variants
Feedback
Help

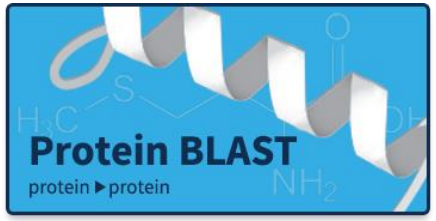
Práticas Análise Bioinformática de Proteínas
Bancos de Dados

1. Utilizando a base de dados de proteínas anotadas, responda as seguintes perguntas sobre as aquaporinas 3 e 4 de humanos.
   a) Onde podem ser encontradas (tecido e localização celular)?
   b) Apresentam ambas as aquaporinas a mesma função?
   c) De quais processos biológicos participam?
   d) Em que consiste a estrutura secundária? E a terciaria? Há estrutura quaternária?
   e) Quais são os motivos sequência? São conservados entre as aquaporinas?
   f) Há alguma mutação relacionada a alguma doença humana em alguma das aquaporinas? Qual?
   g) Quais resíduos são glicosilados? E fosforilados?

2. Pegue as sequências FASTA das Aquaporinas 3 e 4 e renomeie com o nome curto sugerido para cada proteína no banco de dados.

3. Procure o proteoma de um inseto que esteja o mais completo possível, com a menor porcentagem de genes duplicados e sem genes fragmentados. Copie o código do proteoma e o nome da espécie.

4. Ache duas sequências de aquaporinas de insetos utilizando a função avançada (faça print da tela):
   a) Excluir o canal iônico BIB que faz parte da mesma família de proteínas.
   b) As proteínas devem estar anotadas (Score ≥ 3).
   c) Baixe as sequências FASTA e renomeie com o nome sugerido para o gene.



An official website of the United States government Here's how you know

National Library of Medicine
National Center for Biotechnology Information

Log in

BLAST® » blastp suite

Home Recent Results Saved Strategies Help

Standard Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more...

Reset page Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange

From

To

Or, upload file Procurar... Nenhum arquivo selecionado.

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

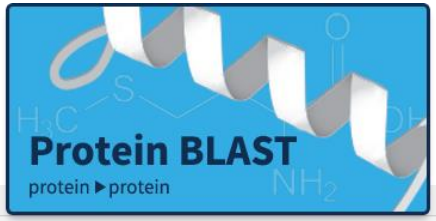
Choose Search Set

Databases: Standard databases (nr etc.): New Experimental databases

Try experimental clustered nr database

For more info see What is clustered nr?

Compare: Select to compare standard and experimental database



Standard

Database

UniProtKB/Swiss-Prot(swissprot)

- **Non-redundant protein sequences (nr)**
- RefSeq Select proteins (refseq_select)
- Reference proteins (refseq_protein)
- Model Organisms (landmark)
- UniProtKB/Swiss-Prot(swissprot)
- Patented protein sequences(pataa)
- Protein Data Bank proteins(pdb)
- Metagenomic proteins(env_nr)
- Transcriptome Shotgun Assembly proteins (tsa_nr)

Organism Optional

ed exclude Add organism

axa will be shown.

Exclude Optional

(WP) Uncultured/environmental sample sequences

Program Selection

Algorithm

- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST

Search database swissprot using Blastp (protein-protein BLAST)

Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

Organism Optional

Enter organism name or id--completions will be suggested exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional

Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences

Program Selection

Algorithm

- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

Prática Análise Bioinformática de Proteínas
Ferramentas de análise

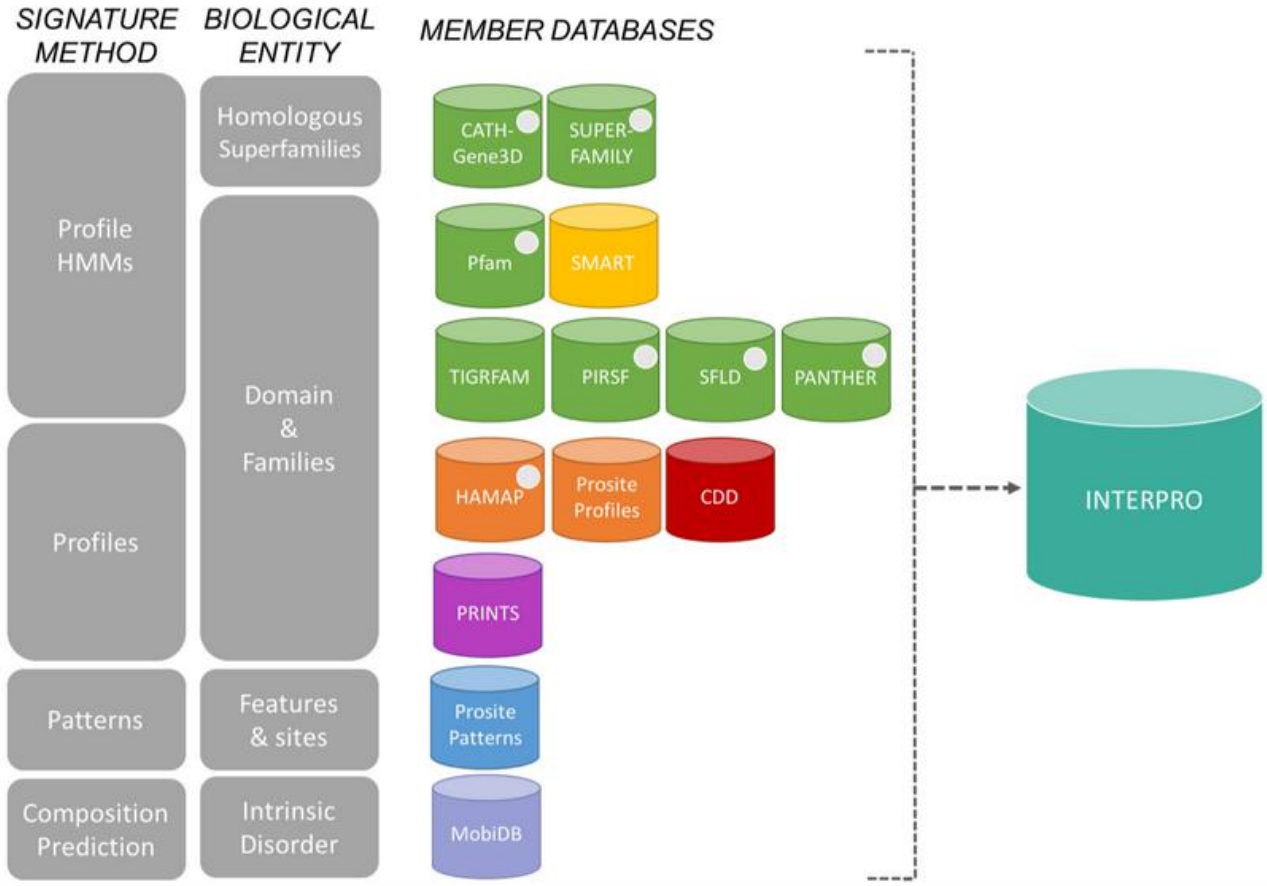
Utilizando Blastp e as sequências de aquaporinas de humanos e insetos do exercício anterior:

a) Baixar o arquivo FASTA das sequências das aquaporinas de *Pediculus humanus corporis* (piolho) e do inseto cujo proteoma foi analisado no exercício anterior?
b) Quais são as coberturas e identidades entre as aquaporinas isca e problema?
c) Baseado nos dados de identidade, quais poderiam ser homólogas?

Análise de Sequências de Proteínas

InterPro

- Recurso gratuito para à análise funcional de sequências de proteínas.
- Permite classificar proteínas em famílias pela presença de domínios e sítios importantes.
- Utiliza modelos preditivos fornecidos por 13 bancos de dados colaboradores (referidos como bancos de dados membros) que compõem coletivamente o consórcio InterPro.
- Fornece informações sobre famílias de proteínas individuais, domínios e sítios importantes.
- Permite a busca realizar pesquisas de sequência e navegar pelas anotações do InterPro.
- Intuitivo



InterPro Classification of protein families

Home | Search | Browse | Results | Release notes | Download | Help | About

Classification of protein families

InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites. To classify proteins in this way, InterPro uses predictive models, known as signatures, provided by several different databases (referred to as member databases) that make up the InterPro consortium. We combine protein signatures from these member databases into a single searchable resource, capitalising on their individual strengths to produce a powerful integrated database and diagnostic tool.

90.0 **InterPro 90.0**
3 August 2022

▸ Citing InterPro

Powering down the Pfam website

On October 5th, we will start redirecting the traffic from Pfam (pfam.xfam.org) to InterPro (www.ebi.ac.uk/interpro). The Pfam website will be available at legacy.pfam.xfam.org until January 2023, when it will be decommissioned. You can read more about the sunset period in our blog post.

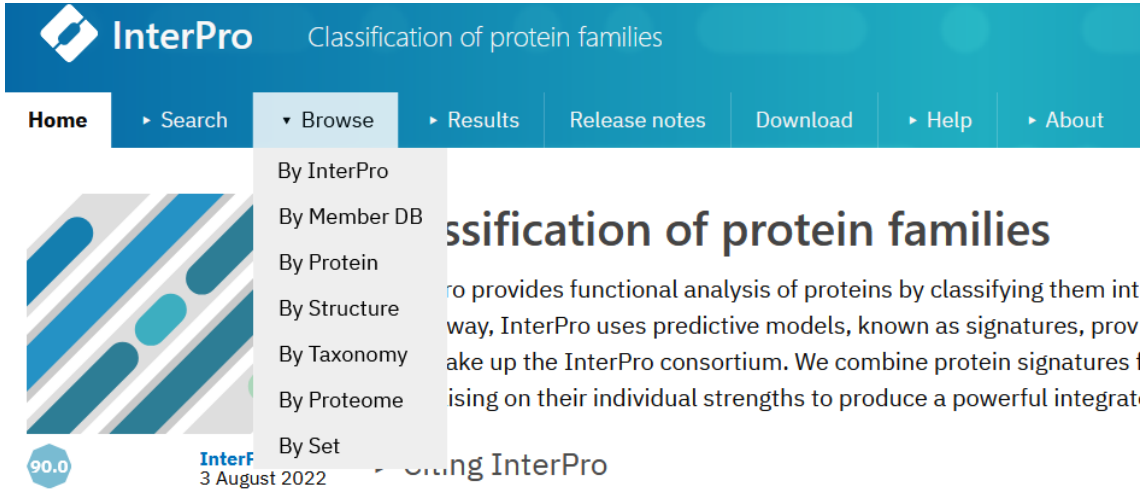
Search by sequence | Search by text | Search by Domain Architecture

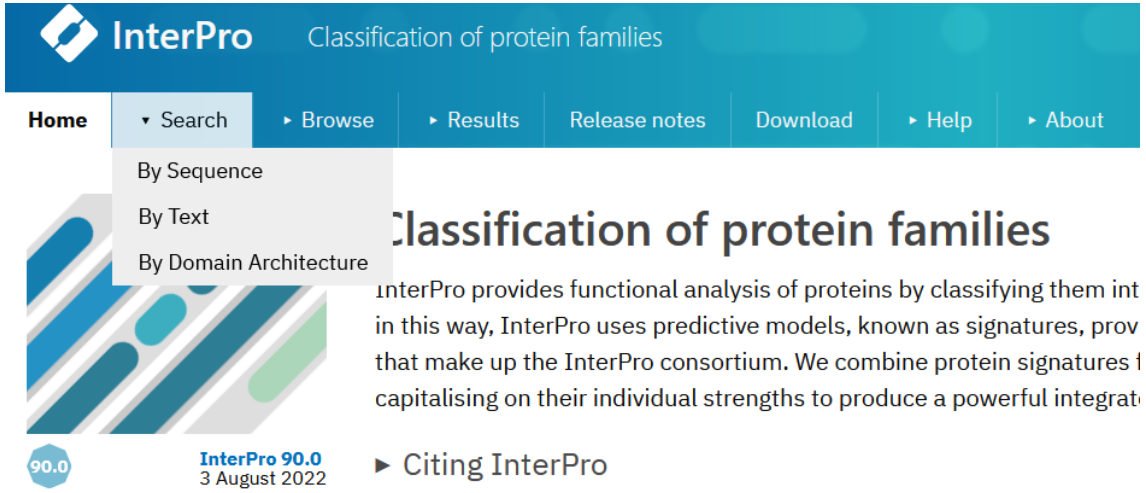
Sequence, in FASTA format

Enter your sequence

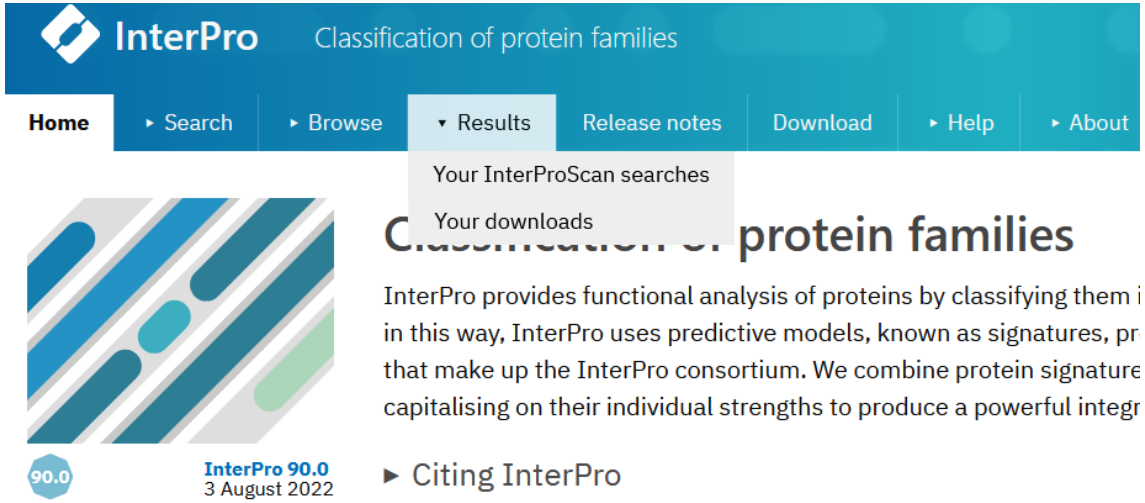
https://www.ebi.ac.uk/interpro/

InterPro





Mecanismos de busca
Resultados



InterPro Classification of protein families

Home | Search | Browse | Results | Release notes | Download | Help | About

/ Result / InterProScan / Iprscan5-R20220918-201621-0556-57558232-p2m / Overview

Overview | Entries 4 | Sequence

InterProScan Search Result

Title sp|P55087|AQP4_HUMAN Aquaporin-4 OS=Homo sapiens OX=9606 GN=AQP4 PE=1 SV=

Job ID iprscan5-R20220918-201621-0556-57558232-p2m

Length 323 amino acids

Actions

Status finished

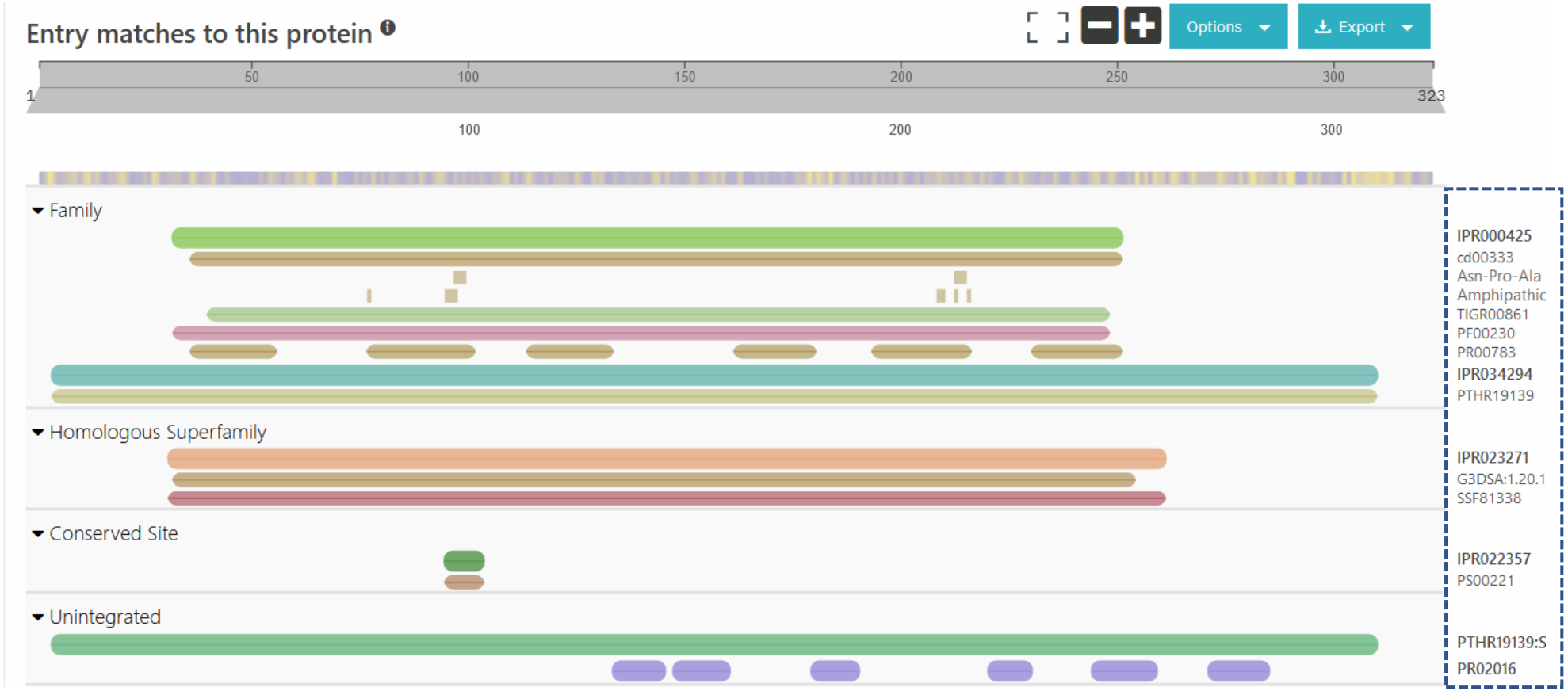
Expires Sun Sep 25 2022

Protein family membership

- F Major intrinsic protein (IPR000425)
  - F Aquaporin transporter (IPR034294)

InterPro

Predição de Motivos, Sítios e Domínios Funcionais

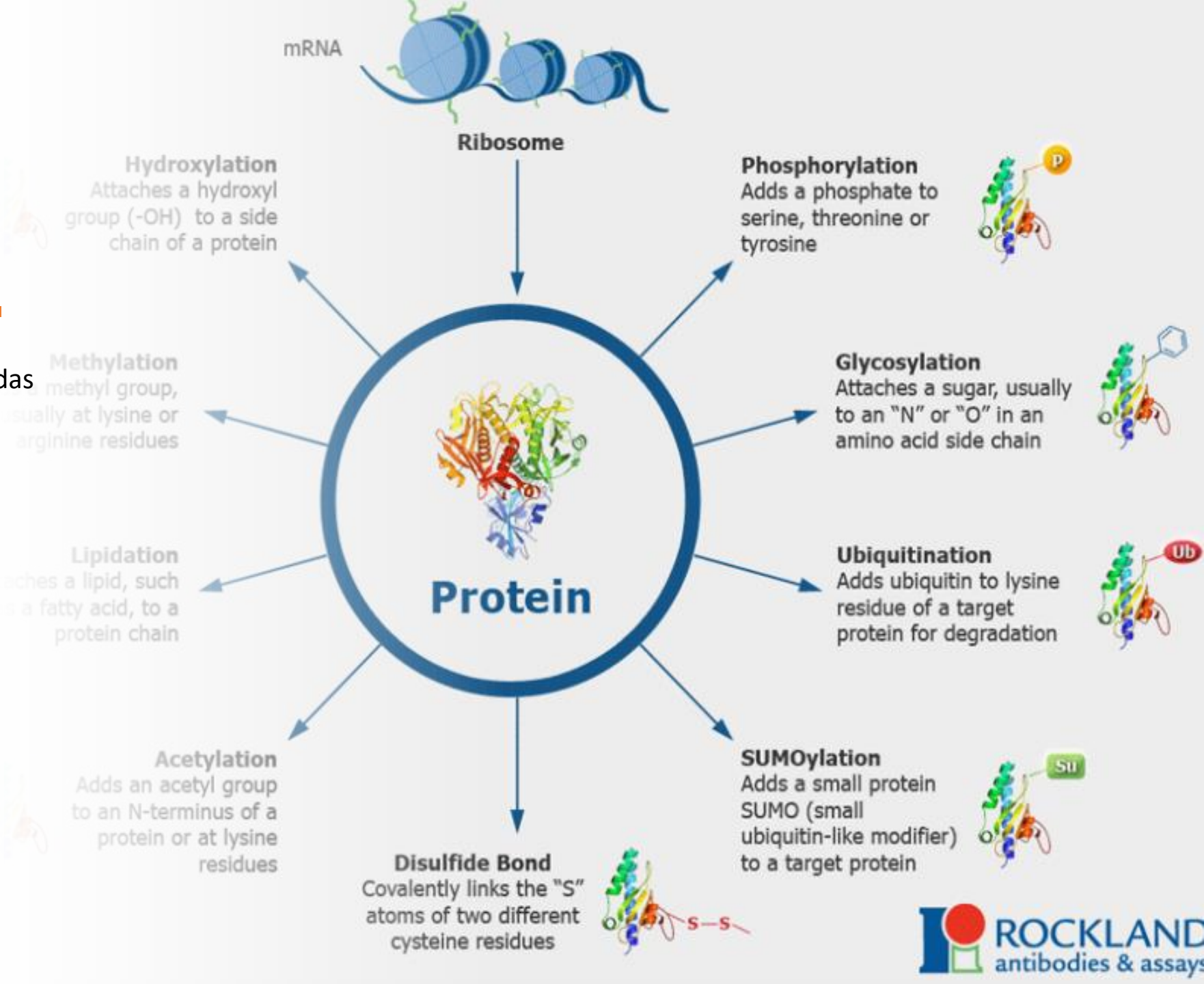


InterPro

Predição de Estrutura Secundária

Modificações pós-traducionais

- Alterações químicas após a síntese das proteínas nos ribossomos.
- A função de muitas proteínas depende delas.
  - Tipos:
    - **Adicionam grupos funcionais nos aminoácidos**
    - Ligam proteínas.



Preditor de N-Glicosilação

NetNGlyc - 1.0

N-linked glycosylation sites in human proteins

The NetNglyc server predicts N-Glycosylation sites in human proteins using artificial neural networks that examine the sequence context of Asn-Xaa-Ser/Thr sequons.

Submission | Instructions | Output format | Abstract | Downloads

Submission

Sequence submission: paste the sequence(s) *or* upload a local file

Paste a single sequence or several sequences in FASTA format into the field below:

Submit a file in FASTA format directly from your local disk:

Procurar... Nenhum arquivo selecionado.

Alternatively, type in Swiss-Prot ID/AC (e.g. CBG_HUMAN)

☑ **Generate graphics** ☐ **Show additional thresholds (0.32, 0.75, 0.90) in the graph(s)**

By default, predictions are done only on the Asn-Xaa-Ser/Thr sequons (incl. Asn-Pro-Ser/Thr)

☐ **Predict on all Asn residues - *use this only if you know what you are doing!***

Clear fields Send file

Notes: SignalP is automatically run on all sequences. A warning is displayed if a signal peptide is not detected. In transmembrane proteins, only extracellular domains may be N-glycosylated. This is currently not checked by the NetNGlyc server. Cytoplasmic and transmembrane sequence regions may be predicted to be glycosylated - this should, of course, be ignored. One transmembrane region predictor is TMHMM.

NetNGlyc-1.0 Server Output - DTU Health Tech

Asn-Xaa-Ser/Thr sequons in the sequence output below are highlighted in blue.
Asparagines predicted to be N-glycosylated are highlighted in red.

Output for 'sp_P55087_AQP4_HUMAN'

```
##############################################################################

   Warning: This sequence may not contain a signal peptide!!

   Proteins without signal peptides are unlikely to be exposed to
   the N-glycosylation machinery and thus may not be glycosylated
   (in vivo) even though they contain potential motifs.

   SignalP-NN euk predictions are as follows:

  # name               Cmax  pos ?  Ymax  pos ?  Smax  pos ?  Smean ?  D     ?
  sp_P55087_AQP4_HUMAN       0.224  41  0.155  41  0.203  56  0.133   0.146 N  0.500      SignalP-TM

   SignalP output is explained at https://services.healthtech.dtu.dk/services/SignalP-4.1/output.php

##############################################################################

Name:  sp_P55087_AQP4_HUMAN     Length:  323
MSDRPTARRWGKCGPLCTRENIMVAFKGVWTQAFWKAVTAEFLAMLIFVLLSLGSTINWGGTEKPLPVDMVLISLCFGLS      80
IATMVQCFGHISGGHINPAVTVAMVCTRKISIAKSVFYIAAQCLGAIIGAGILYLVTPPSVVGGLGVTMVHGNLTAGHGL     160
LVELIITFQLVFTIFASCDSKRTDVTGSIALAIGFSVAIGHLFAINYTGASMNPARSFGPAVIMGNWENHWIYWVGPIIG     240
AVLAGGLYEYVFCPDVEFKRRFKEAFSKAAQQTKGSYMEVEDNRSQVETDDLILKPGVVHVIDVDRGEEKKGKDQSGEVL     320
SSV
.....................................................................................      80
.....................................................................N.......     160
.....................................N.....................................     240
.....................................N.....................................     320
...                                                                                400

(Threshold=0.5)
--------------------------------------------------------------------------------
SeqName      Position  Potential   Jury    N-Glyc
                                   agreement result
--------------------------------------------------------------------------------
sp_P55087_AQP4_HUMAN  153 NLTA   0.6463     (9/9)   ++
sp_P55087_AQP4_HUMAN  206 NYTG   0.6799     (9/9)   ++
sp_P55087_AQP4_HUMAN  283 NRSQ   0.6479     (8/9)   +
--------------------------------------------------------------------------------
```

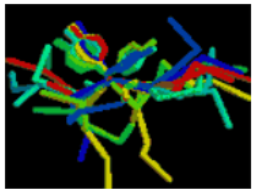
Preditor de Fosforilação

NetPhos - 3.1

Generic phosphorylation sites in eukaryotic proteins

The **NetPhos 3.1** server predicts serine, threonine or tyrosine phosphorylation sites in eukaryotic proteins using ensembles of neural networks. Both generic and kinase specific predictions are performed. The **generic** predictions are identical to the predictions performed by NetPhos 2.0. The **kinase specific** predictions are identical to the predictions by NetPhosK 1.0. Predictions are made for the following 17 kinases:

ATM, **CKI**, **CKII**, **CaM-II**, **DNAPK**, **EGFR**, **GSK3**, **INSR**, **PKA**, **PKB**, **PKC**, **PKG**, **RSK**, **SRC**, **cdc2**, **cdk5** and **p38MAPK**.



Submission | Instructions | Output format | PhosphoBase | Downloads

Submission

Sequence submission: paste the sequence(s) *and/or* upload a local file

Paste a single sequence or several sequences in FASTA format into the field below:

Submit a file in FASTA format directly from your local disk:

Procurar... Nenhum arquivo selecionado.

Residues to predict ○ serine ○ threonine ○ tyrosine ◉ all three

For each residue display only the best prediction ☐

Display only the scores higher than 0

Output format ◉ classical ○ GFF

Generate graphics ☑

Submit Clear fields

Submission | Instructions | Output format | PhosphoBase | Downloads

NetPhos-3.1 Server Output - DTU Health Tech

```
gawk: /tools/src/ape-1.0/disp/netphos.awk:125: warning: escape sequence `\/' treated as plain `/'
>sp_P55087_AQP4_HUMAN   323 amino acids
#
# netphos-3.1b prediction results
#
# Sequence                 # x   Context     Score   Kinase    Answer
# ---------------------------------------------------------------------
# sp_P55087_AQP4_HUMAN     2 S   ---MSDRPT   0.426   GSK3        .
# sp_P55087_AQP4_HUMAN     2 S   ---MSDRPT   0.426   cdc2        .
# sp_P55087_AQP4_HUMAN     2 S   ---MSDRPT   0.417   CaM-II      .
# sp_P55087_AQP4_HUMAN     2 S   ---MSDRPT   0.372   p38MAPK     .
# sp_P55087_AQP4_HUMAN     2 S   ---MSDRPT   0.368   CKII        .
# sp_P55087_AQP4_HUMAN     2 S   ---MSDRPT   0.366   CKI         .
# sp_P55087_AQP4_HUMAN     2 S   ---MSDRPT   0.353   DNAPK       .
# sp_P55087_AQP4_HUMAN     2 S   ---MSDRPT   0.344   PKG         .
# sp_P55087_AQP4_HUMAN     2 S   ---MSDRPT   0.266   ATM         .
# sp_P55087_AQP4_HUMAN     2 S   ---MSDRPT   0.244   RSK         .
# sp_P55087_AQP4_HUMAN     2 S   ---MSDRPT   0.234   PKC         .
# sp_P55087_AQP4_HUMAN     2 S   ---MSDRPT   0.196   cdk5        .
# sp_P55087_AQP4_HUMAN     2 S   ---MSDRPT   0.192   PKA         .
# sp_P55087_AQP4_HUMAN     2 S   ---MSDRPT   0.092   PKB         .
# sp_P55087_AQP4_HUMAN     2 S   ---MSDRPT   0.032   unsp        .
#
# sp_P55087_AQP4_HUMAN     6 T   SDRPTARRW   0.880   PKC        YES
# sp_P55087_AQP4_HUMAN     6 T   SDRPTARRW   0.754   unsp       YES
# sp_P55087_AQP4_HUMAN     6 T   SDRPTARRW   0.462   GSK3        .
# sp_P55087_AQP4_HUMAN     6 T   SDRPTARRW   0.428   CaM-II      .
# sp_P55087_AQP4_HUMAN     6 T   SDRPTARRW   0.391   CKI         .
# sp_P55087_AQP4_HUMAN     6 T   SDRPTARRW   0.387   PKG         .
# sp_P55087_AQP4_HUMAN     6 T   SDRPTARRW   0.352   cdc2        .
# sp_P55087_AQP4_HUMAN     6 T   SDRPTARRW   0.341   DNAPK       .
# sp_P55087_AQP4_HUMAN     6 T   SDRPTARRW   0.315   cdk5        .
# sp_P55087_AQP4_HUMAN     6 T   SDRPTARRW   0.302   p38MAPK     .
# sp_P55087_AQP4_HUMAN     6 T   SDRPTARRW   0.268   RSK         .
# sp_P55087_AQP4_HUMAN     6 T   SDRPTARRW   0.237   CKII        .
# sp_P55087_AQP4_HUMAN     6 T   SDRPTARRW   0.234   ATM         .
# sp_P55087_AQP4_HUMAN     6 T   SDRPTARRW   0.139   PKA         .
# sp_P55087_AQP4_HUMAN     6 T   SDRPTARRW   0.082   PKB         .
#
```

Prática Análise Bioinformática de Proteínas
Ferramentas de análise

1. Utilizando as ferramentas de predição introduzidas na aula, responda as seguintes perguntas a respeito das sequências problema das aquaporinas de insetos do exercício anterior.

    a) A qual família pertencem?
    b) Apresentam algum sitio, motivo ou domínio conservado com as de humanos? Qual? Escreva o código INTERPRO e o nome.
    c) Quantas regiões transmembrana teriam?
    d) Existe alguma modificação pós-traducional potencial conservada com humanos?

Análise Manual de Sequências de Proteínas

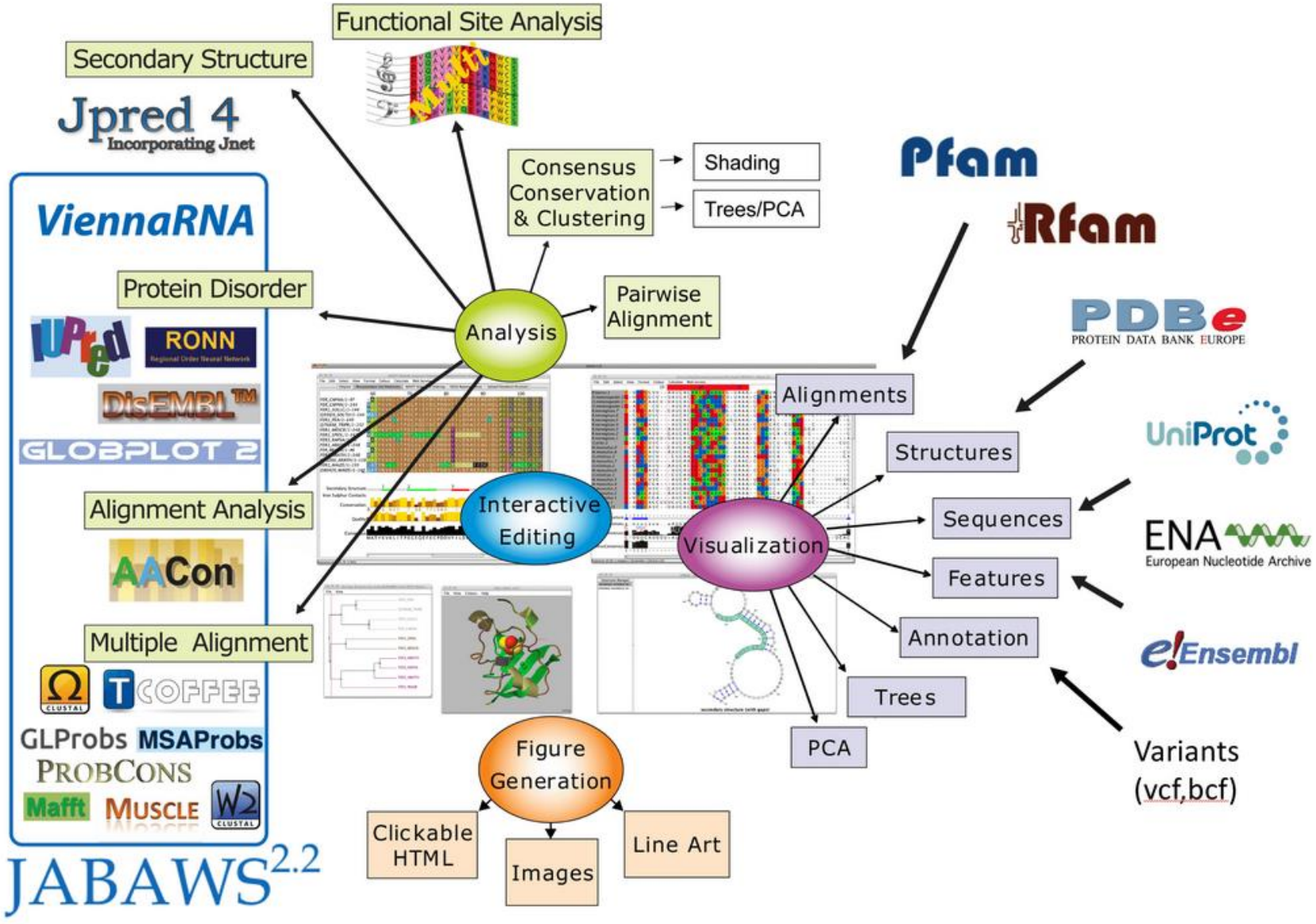
- **Programa multiplataforma gratuito para edição, visualização e análise de alinhamento de múltiplas sequências.**

- Utilizado para alinhar, visualizar e editar alinhamentos de sequências, analisá-los com árvores filogenéticas e gráficos de análise de componentes principais (PCA) e explorar estruturas moleculares e anotações.
- Permite o uso de sequências de DNA, RNA e proteínas.
- Apresenta recursos de visualização e análise de estrutura de ácidos nucleicos e proteínas.
- Visualização de produtos alinhados de DNA e proteína.
- Utiliza:
  - Jmol para visualizar estruturas 3D
  - VARNA para exibir a estrutura secundária de RNA
  - Chimera, ChimeraX e Pymol para visualização de estrutura 3D de proteínas.

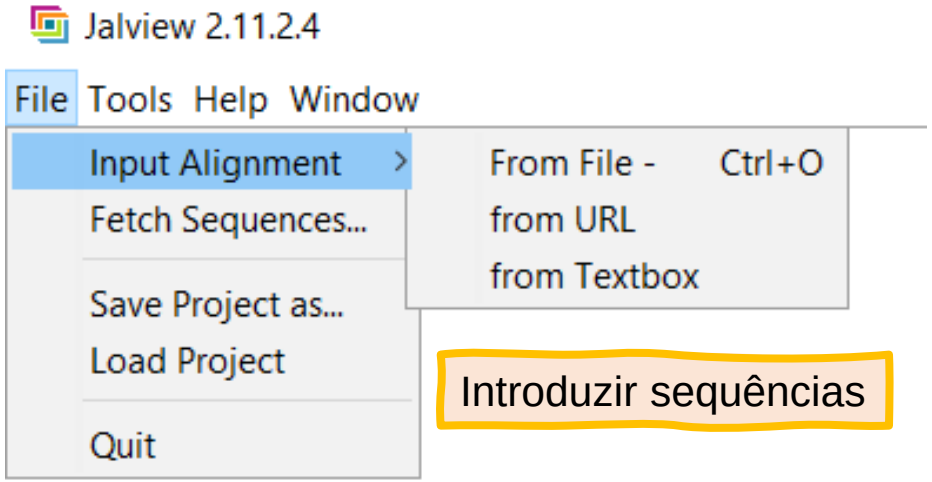
Análise Manual de Sequências de Proteínas

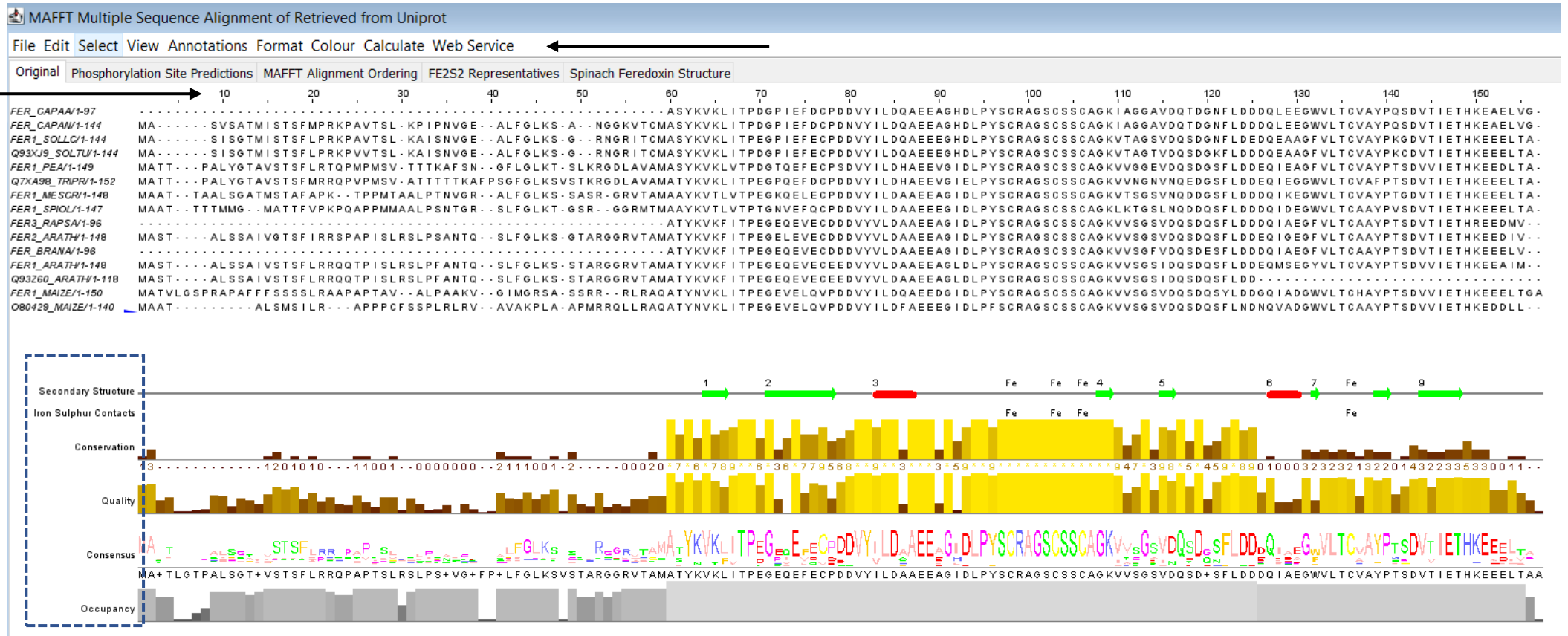
Jalview

A workbench for multiple sequence alignment and analysis

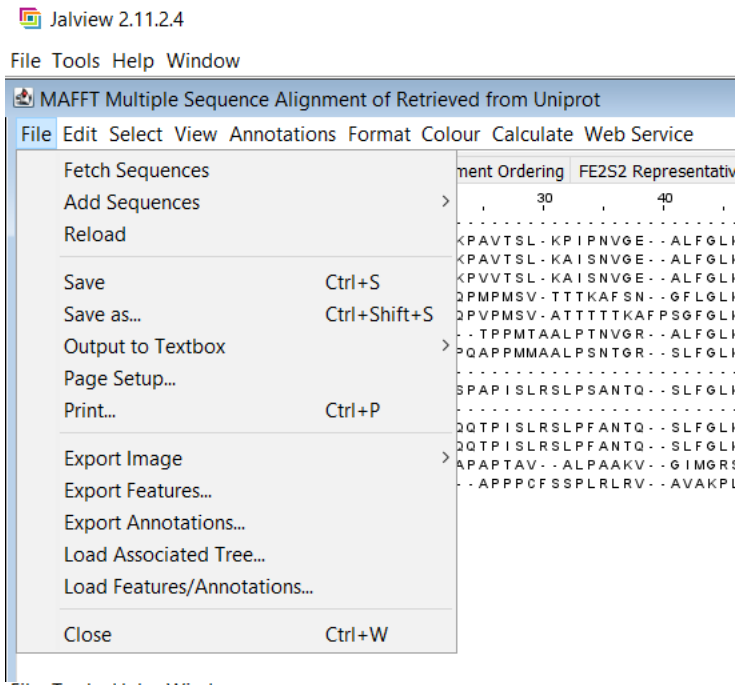


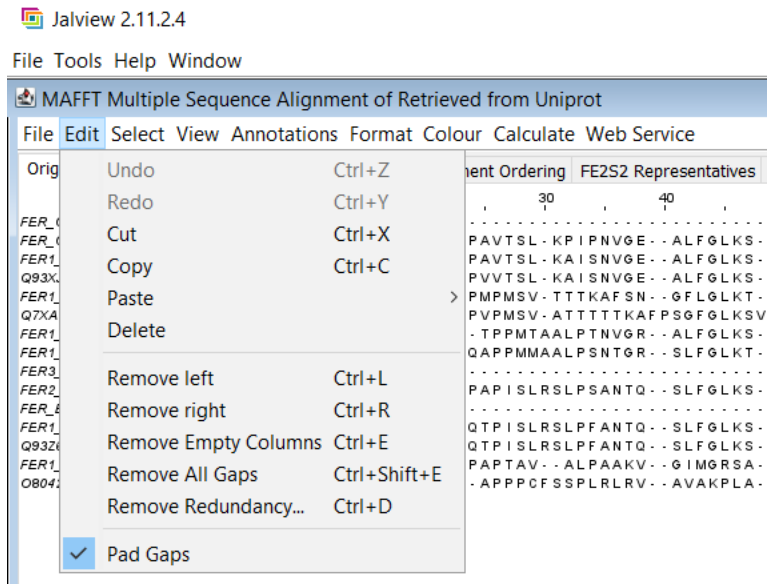
Jalview

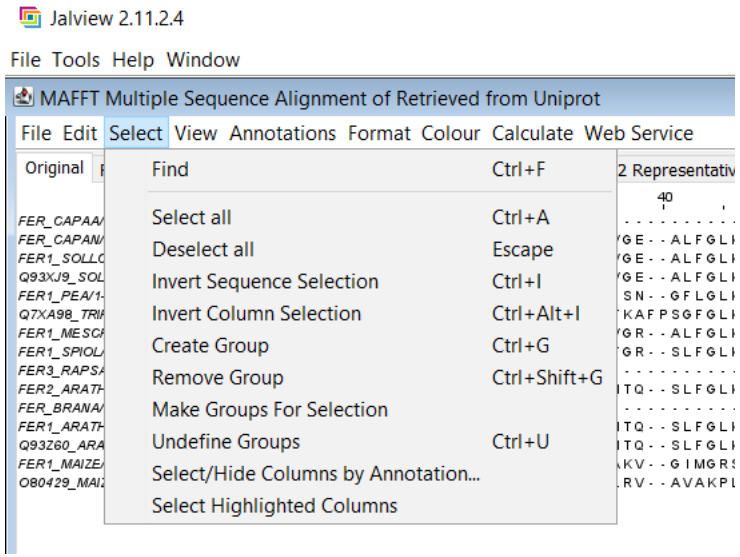

Jalview 2.11.2.4
File Tools Help Window
Input Alignment
Fetch Sequences...
Save Project as...
Load Project
Quit
From File - Ctrl+O
from URL
from Textbox
Introduzir sequências

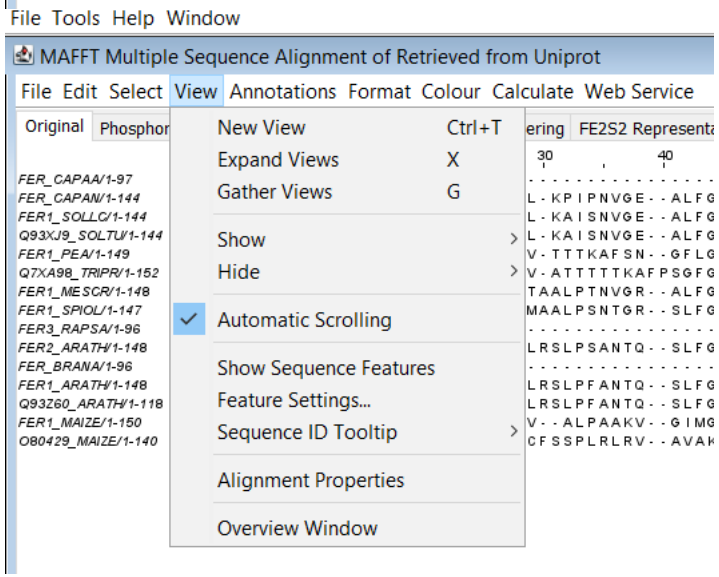

MAFFT Multiple Sequence Alignment of Retrieved from Uniprot
File Edit Select View Annotations Format Colour Calculate Web Service
Original
Phosphorylation Site Predictions
MAFFT Alignment Ordering
FE2S2 Representatives
Spinach Feredoxin Structure
FER_CAPAA/1-97
FER_CAPAN/1-144
FER1_SOLLC/1-144
Q93XJ9_SOLTU/1-144
FER1_PEA/1-149
Q7XA98_TRIPR/1-152
FER1_MESCR/1-148
FER1_SPIOL/1-147
FER3_RAPSA/1-96
FER2_ARATH/1-148
FER_BRANA/1-96
FER1_ARATH/1-148
Q93Z60_ARATH/1-118
FER1_MAIZE/1-150
O80429_MAIZE/1-140
Secondary Structure
Iron Sulphur Contacts
Conservation
Quality
Consensus
Occupancy

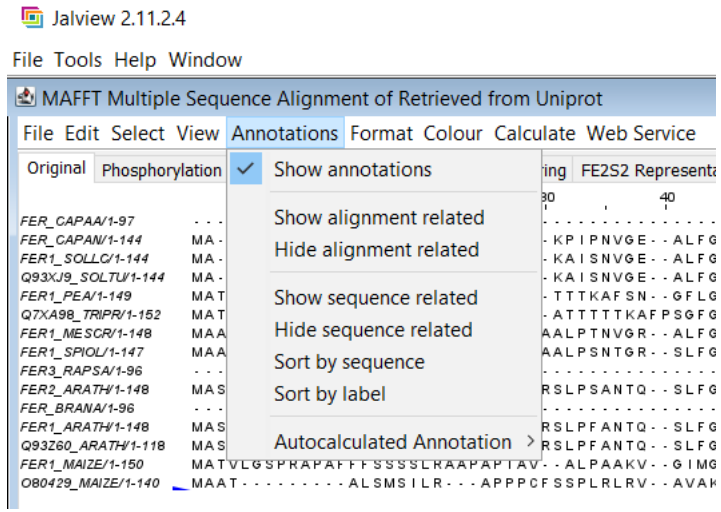
Jalview

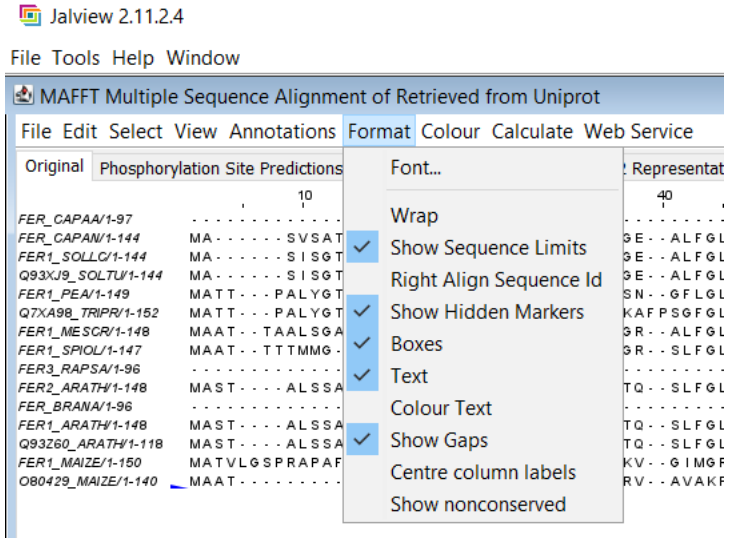

Jalview 2.11.2.4
File Tools Help Window
MAFFT Multiple Sequence Alignment of Retrieved from Uniprot
File Edit Select View Annotations Format Colour Calculate Web Service
Fetch Sequences
Add Sequences
Reload
Save Ctrl+S
Save as... Ctrl+Shift+S
Output to Textbox
Page Setup...
Print... Ctrl+P
Export Image
Export Features...
Export Annotations...
Load Associated Tree...
Load Features/Annotations...
Close Ctrl+W


Jalview 2.11.2.4
File Tools Help Window
MAFFT Multiple Sequence Alignment of Retrieved from Uniprot
File Edit Select View Annotations Format Colour Calculate Web Service
Undo Ctrl+Z
Redo Ctrl+Y
Cut Ctrl+X
Copy Ctrl+C
Paste
Delete
Remove left Ctrl+L
Remove right Ctrl+R
Remove Empty Columns Ctrl+E
Remove All Gaps Ctrl+Shift+E
Remove Redundancy... Ctrl+D
Pad Gaps

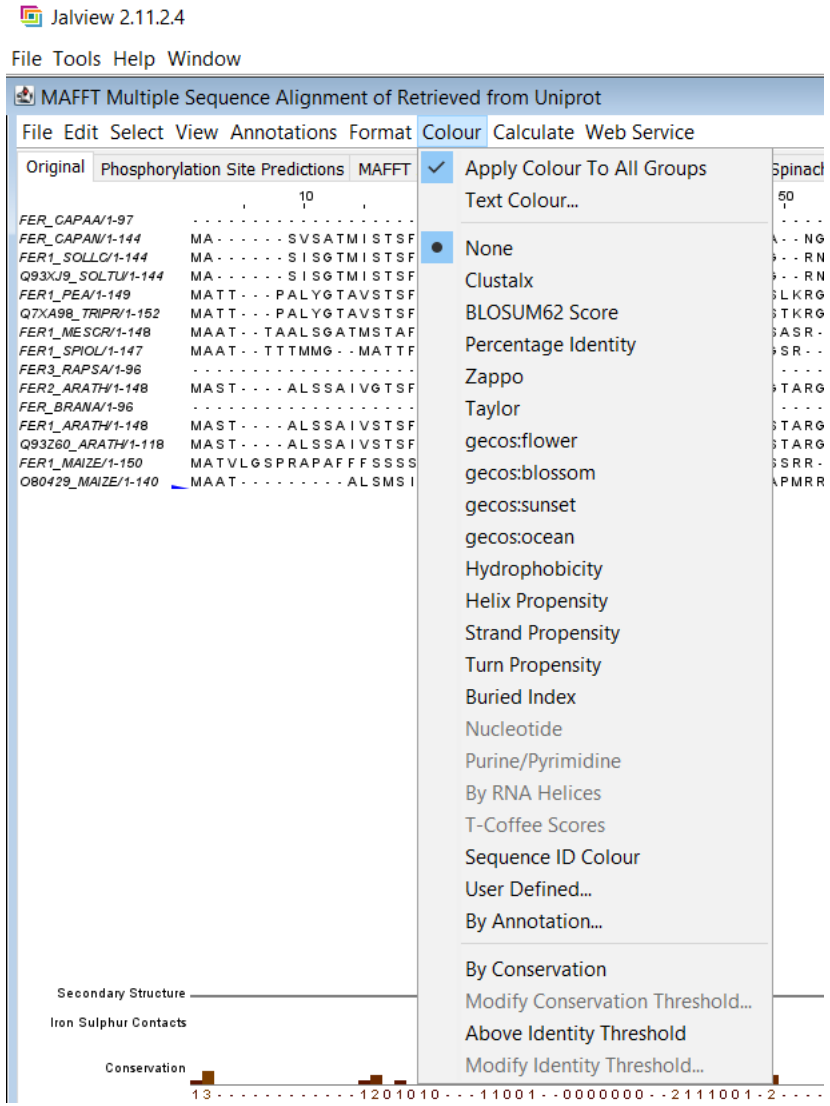

Jalview 2.11.2.4
File Tools Help Window
MAFFT Multiple Sequence Alignment of Retrieved from Uniprot
File Edit Select View Annotations Format Colour Calculate Web Service
Find Ctrl+F
Select all Ctrl+A
Deselect all Escape
Invert Sequence Selection Ctrl+I
Invert Column Selection Ctrl+Alt+I
Create Group Ctrl+G
Remove Group Ctrl+Shift+G
Make Groups For Selection
Undefine Groups Ctrl+U
Select/Hide Columns by Annotation...
Select Highlighted Columns

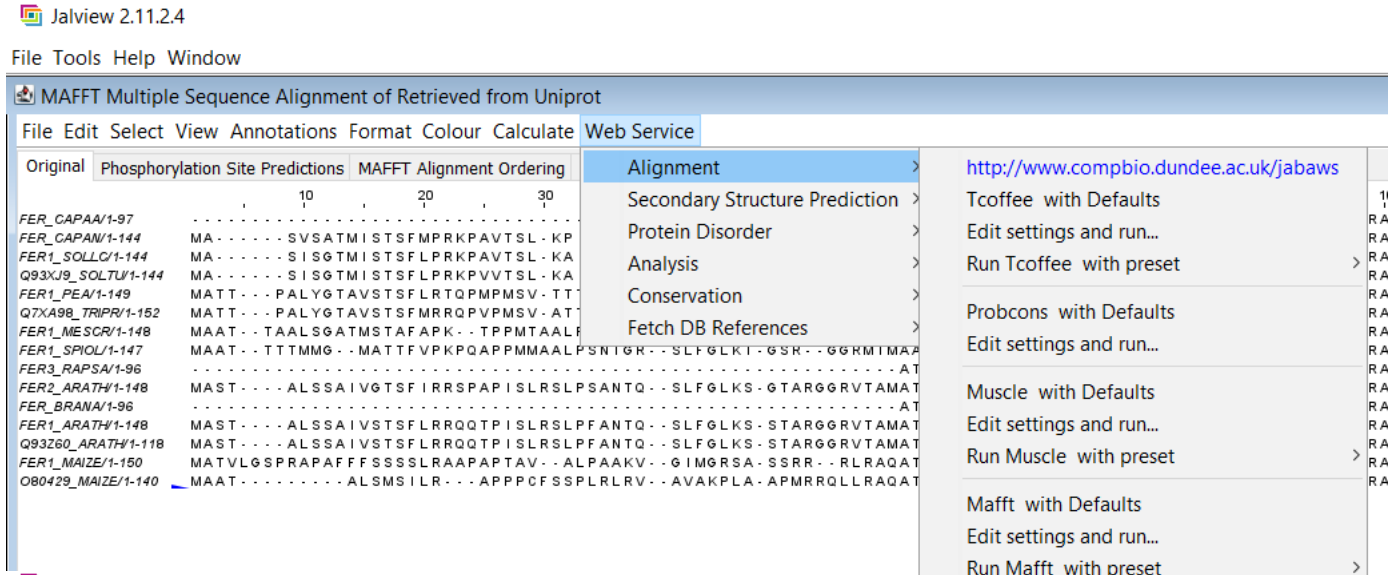

File Tools Help Window
MAFFT Multiple Sequence Alignment of Retrieved from Uniprot
File Edit Select View Annotations Format Colour Calculate Web Service
New View Ctrl+T
Expand Views X
Gather Views G
Show
Hide
Automatic Scrolling
Show Sequence Features
Feature Settings...
Sequence ID Tooltip
Alignment Properties
Overview Window

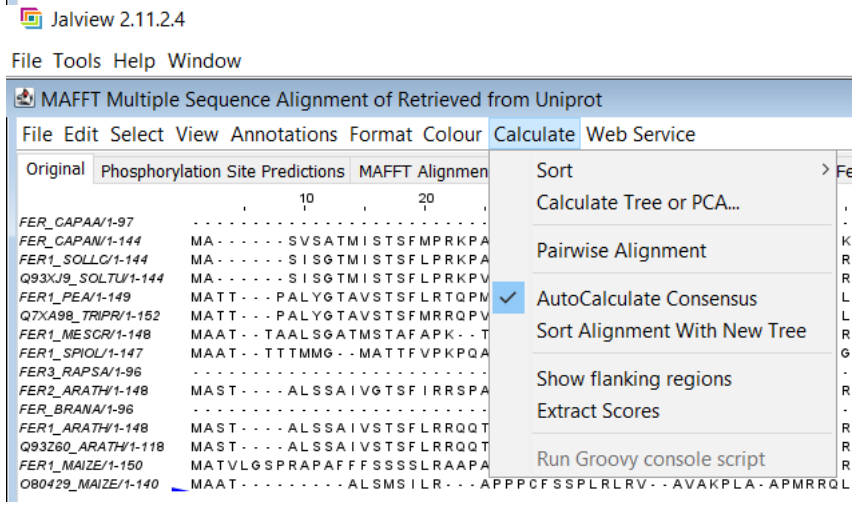

Jalview 2.11.2.4
File Tools Help Window
MAFFT Multiple Sequence Alignment of Retrieved from Uniprot
File Edit Select View Annotations Format Colour Calculate Web Service
Show annotations
Show alignment related
Hide alignment related
Show sequence related
Hide sequence related
Sort by sequence
Sort by label
Autocalculated Annotation


Jalview 2.11.2.4
File Tools Help Window
MAFFT Multiple Sequence Alignment of Retrieved from Uniprot
File Edit Select View Annotations Format Colour Calculate Web Service
Font...
Wrap
Show Sequence Limits
Right Align Sequence Id
Show Hidden Markers
Boxes
Text
Colour Text
Show Gaps
Centre column labels
Show nonconserved

Jalview


Jalview 2.11.2.4
File Tools Help Window
MAFFT Multiple Sequence Alignment of Retrieved from Uniprot
File Edit Select View Annotations Format Colour Calculate Web Service
Original
Phosphorylation Site Predictions
MAFFT
Apply Colour To All Groups
Text Colour...
None
Clustalx
BLOSUM62 Score
Percentage Identity
Zappo
Taylor
gecos:flower
gecos:blossom
gecos:sunset
gecos:ocean
Hydrophobicity
Helix Propensity
Strand Propensity
Turn Propensity
Buried Index
Nucleotide
Purine/Pyrimidine
By RNA Helices
T-Coffee Scores
Sequence ID Colour
User Defined...
By Annotation...
By Conservation
Modify Conservation Threshold...
Above Identity Threshold
Modify Identity Threshold...
Secondary Structure
Iron Sulphur Contacts
Conservation


Jalview 2.11.2.4
File Tools Help Window
MAFFT Multiple Sequence Alignment of Retrieved from Uniprot
File Edit Select View Annotations Format Colour Calculate Web Service
Original
Phosphorylation Site Predictions
MAFFT Alignment Ordering
Alignment
Secondary Structure Prediction
Protein Disorder
Analysis
Conservation
Fetch DB References
http://www.compbio.dundee.ac.uk/jabaws
Tcoffee with Defaults
Edit settings and run...
Run Tcoffee with preset
Probcons with Defaults
Edit settings and run...
Muscle with Defaults
Edit settings and run...
Run Muscle with preset
Mafft with Defaults
Edit settings and run...
Run Mafft with preset
MSAprobs with Defaults
Edit settings and run...
GLprobs with Defaults
Edit settings and run...
Clustal
Realign with Clustal
ClustalO
Realign with ClustalO


Jalview 2.11.2.4
File Tools Help Window
MAFFT Multiple Sequence Alignment of Retrieved from Uniprot
File Edit Select View Annotations Format Colour Calculate Web Service
Original
Phosphorylation Site Predictions
Sort
Calculate Tree or PCA...
Pairwise Alignment
AutoCalculate Consensus
Sort Alignment With New Tree
Show flanking regions
Extract Scores
Run Groovy console script

Prática Análise Bioinformática de Proteínas
Ferramentas de análise

Utilizando as sequências de aquaporinas de humanos e insetos vistas até agora:

1. Alinhe as sequências utilizando o método MAFFT.
2. Aplique a cor de identidade. Quais regiões são similares ou idênticas?
3. Aplique a cor de hidrofobicidade e identifique as regiões transmembrana.
4. Que tipo de estrutura secundária apresentam?
5. Compare o resultado anterior com a predição de estrutura secundária. Elas batem?
6. Os motivos NPA e os resíduos envolvidos na seletividade à água (posições 77, 201, 210, 216 em hAQP4) e de transporte de glicerol (219 em hAQP3) são conservados nas aquaporinas de insetos? Em quais?
7. Faça uma árvore por similaridade e alinhe as sequências segundo a árvore.
8. Considerando a região C-terminal das sequências, quais são homólogas das aquaporinas **anotadas** de insetos.
9. Baseado na conservação de resíduos envolvidos na função da proteína. Há alguma aquaporina homóloga de hAQP3 em insetos?

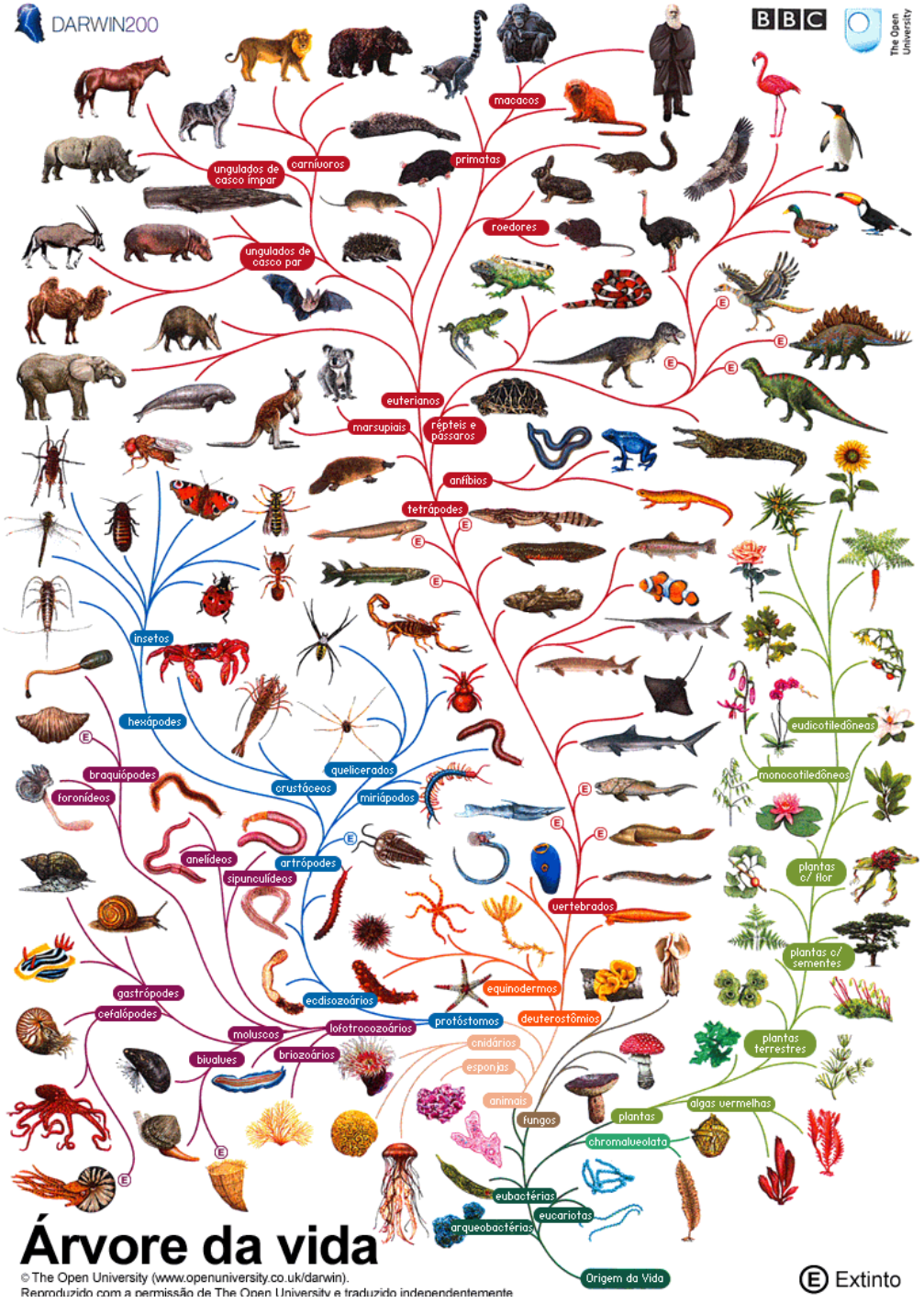
Estudos Evolutivos

Por que proteínas?

Segunda base do códon

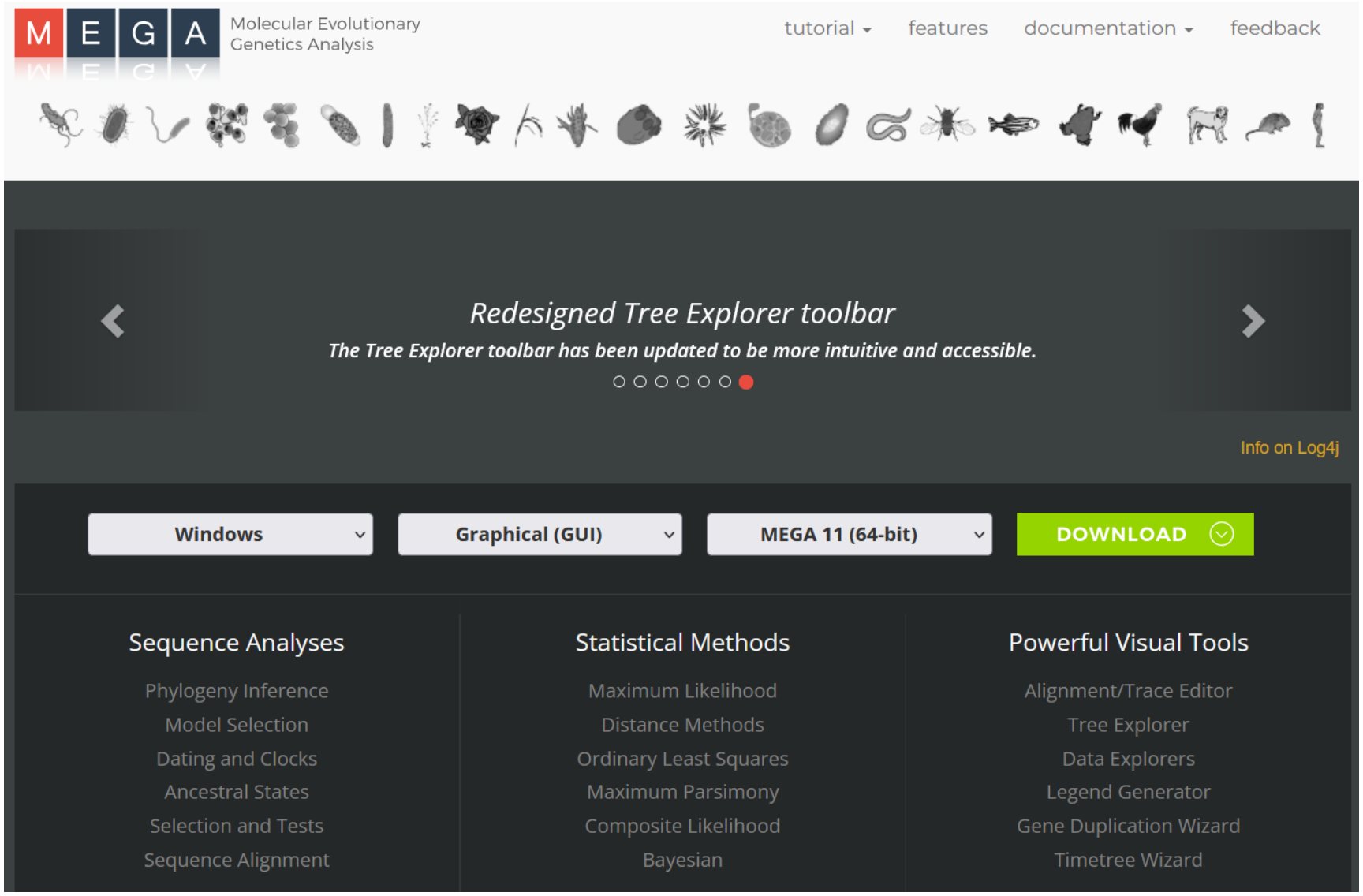
Primeira base do códon	U	C	A	G	Terceira base do códon
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr UAA stop UAG stop	UGU Cys UGC Cys UGA stop UGG Trp	U C A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	U C A G
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G

Arg – Arginina
Asn – Aspargina
Asp – Ácido aspártico
Cys – Cisteína
Gln – Glutamina
Glu – Ácido glutâmino
Gly – Glicina
His – Histidina
Ile – Isoleucina
Leu – Leucina
Lys – Lisina
Met – Metionina (códon de iní
Phe – Fenilalamina
Pro – Prolina
Ser – Serina
Stop – Códon de parada
Thr – Treonina
Tyr – Tirosina
Val – Valina

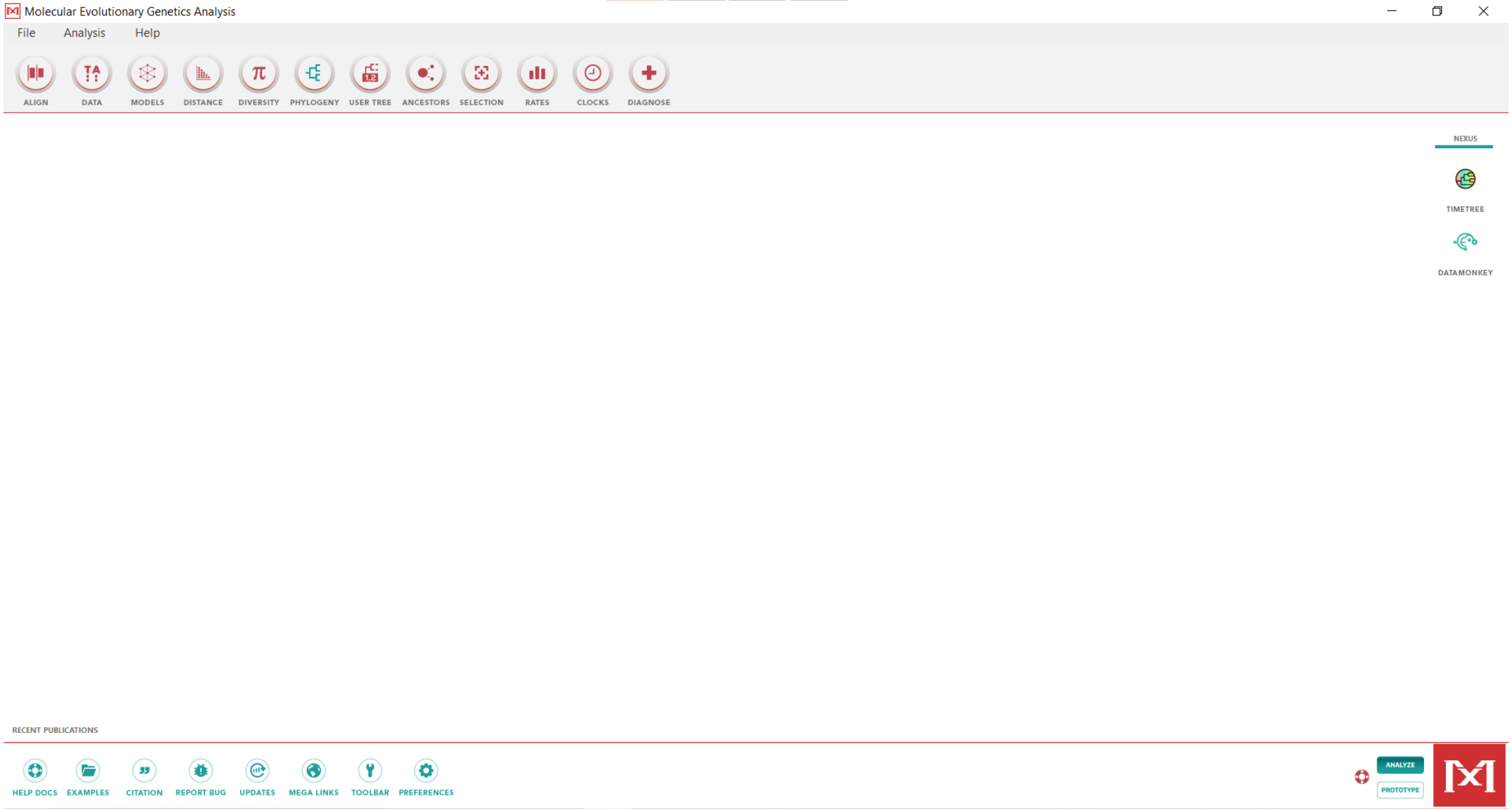


Estudos Evolutivos

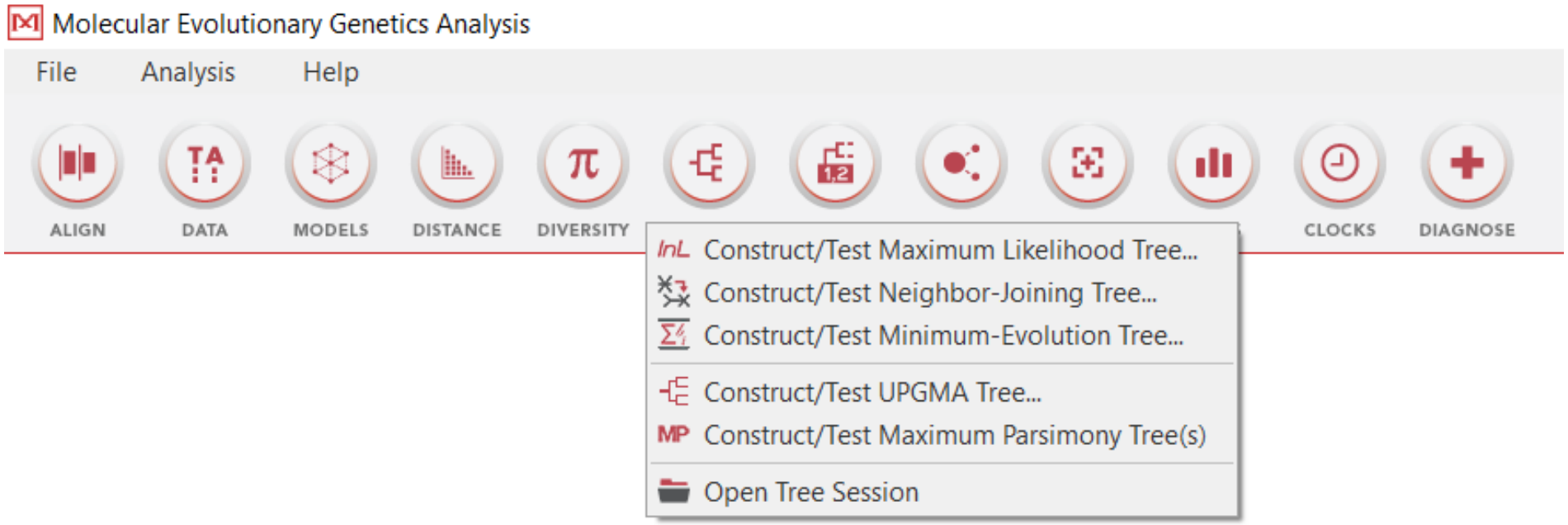
Software para a realização de análises estatísticas da evolução molecular e para a construção de árvores filogenéticas.
Inclui muitos métodos e ferramentas sofisticados para filogenia.



MEGA


Molecular Evolutionary Genetics Analysis
File
Analysis
Help
ALIGN
DATA
MODELS
DISTANCE
DIVERSITY
PHYLOGENY
USER TREE
ANCESTORS
SELECTION
RATES
CLOCKS
DIAGNOSE
NEXUS
TIMETREE
DATAMONKEY
RECENT PUBLICATIONS
HELP DOCS
EXAMPLES
CITATION
REPORT BUG
UPDATES
MEGA LINKS
TOOLBAR
PREFERENCES
ANALYZE
PROTOTYPE

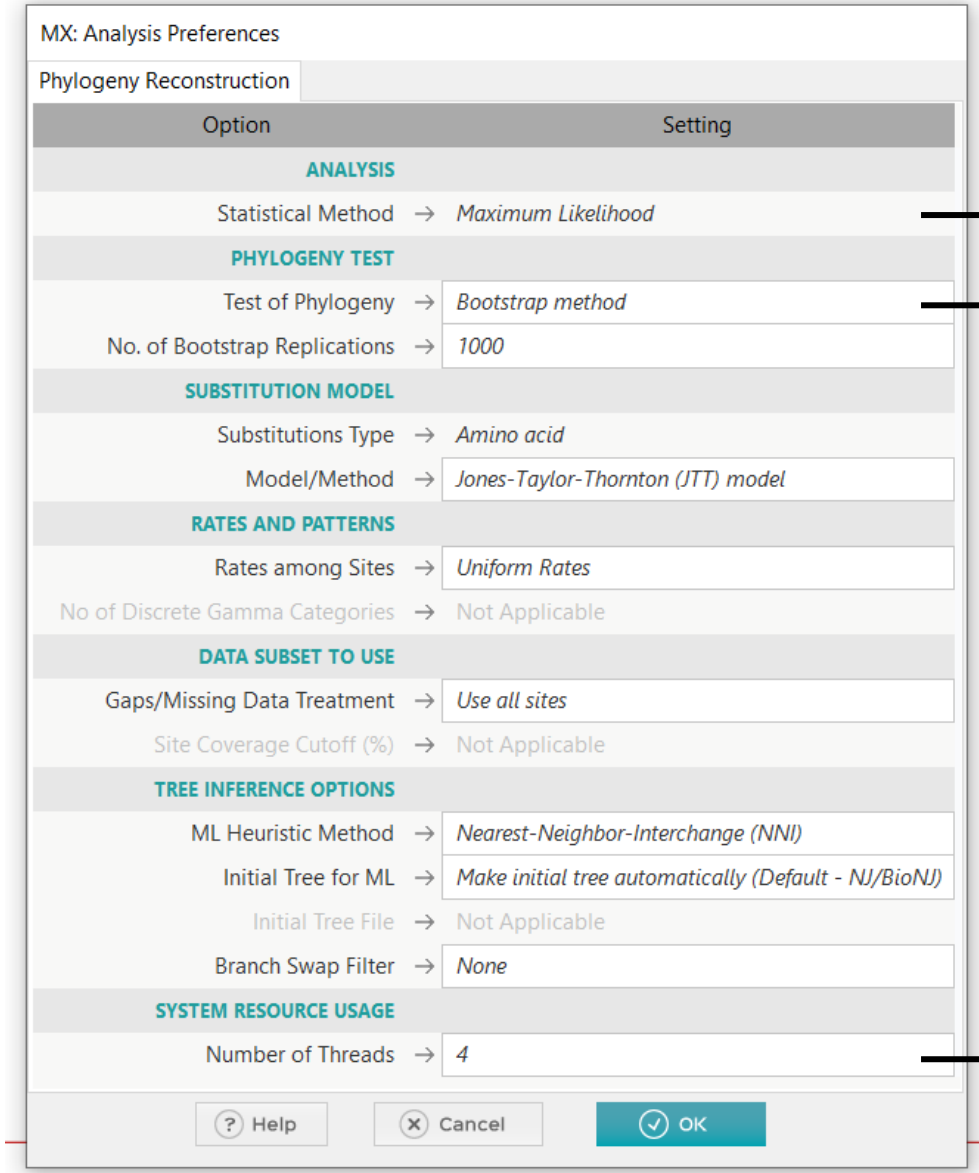
MEGA



Diferentes algoritmos/métodos para estudar a filogenia de um conjunto de sequências de proteínas:

Neighbor-joining e UPGMA: Por similaridade. Utilizado para testes rápidos. Não confiável quando as sequências são muito diferentes. Não aceito para publicação.
Maximum Parsimony e Minimum Evolution: Minimizam o tamanho dos ramos, ao diminuir o número de mutações ou distâncias evolutivas, respectivamente. Ignora fenômenos evolutivos, como convergência.
Maximum Likelihood: Utilizado para modelos evolutivos, ideal para construir a filogenia utilizando sequências. O mais utilizado para publicação. Computacionalmente custoso.

MEGA



Conferir

- O teste de bootstrap mede a consistência interna de um conjunto de dados moleculares analisando se alinhamentos ligeiramente modificados suportam os mesmos clados.

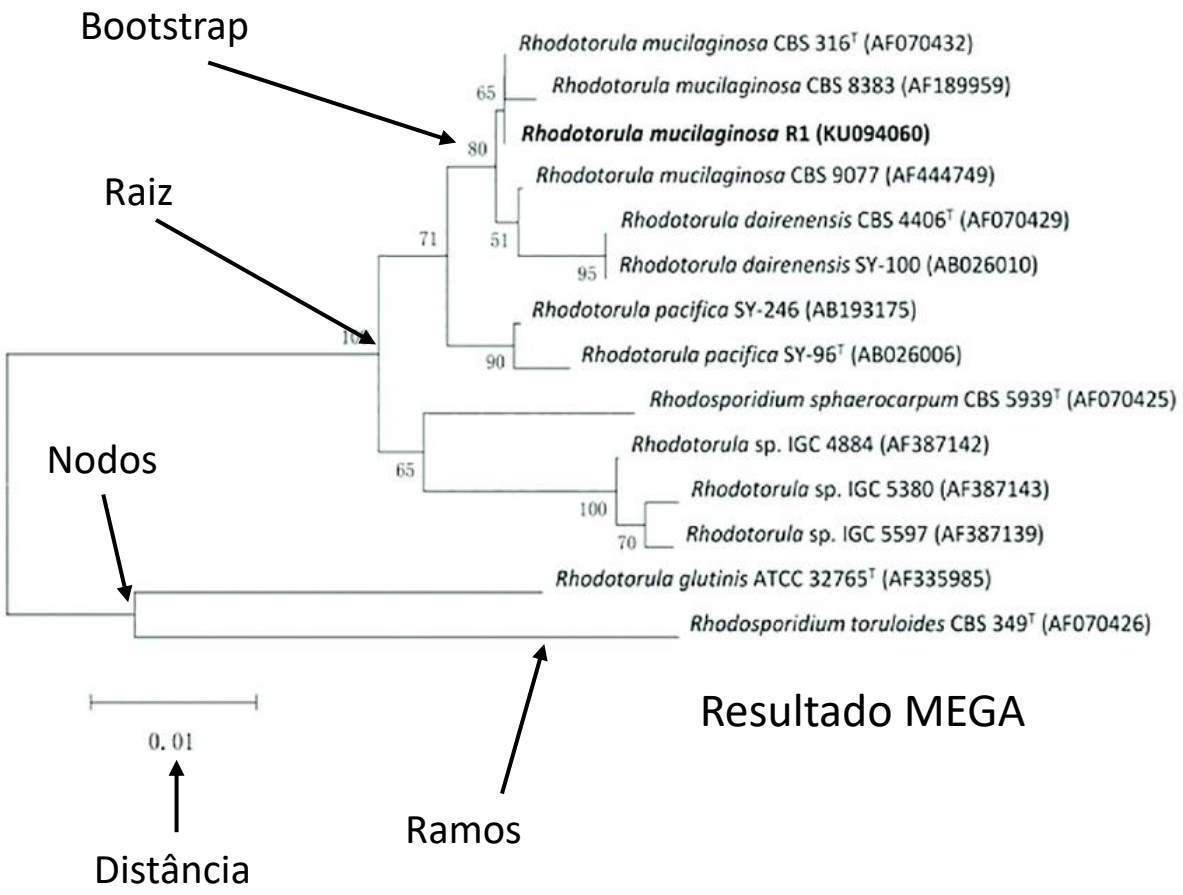
- Valores de bootstrap: Ex. 1000, quer dizer que de 1000, quantas vezes o mesmo ramo é observado ao repetir a geração de uma árvore filogenética em um conjunto de dados. Se obtivermos essa observação 1000 vezes de 1000, isso apóia nosso resultado.

Default

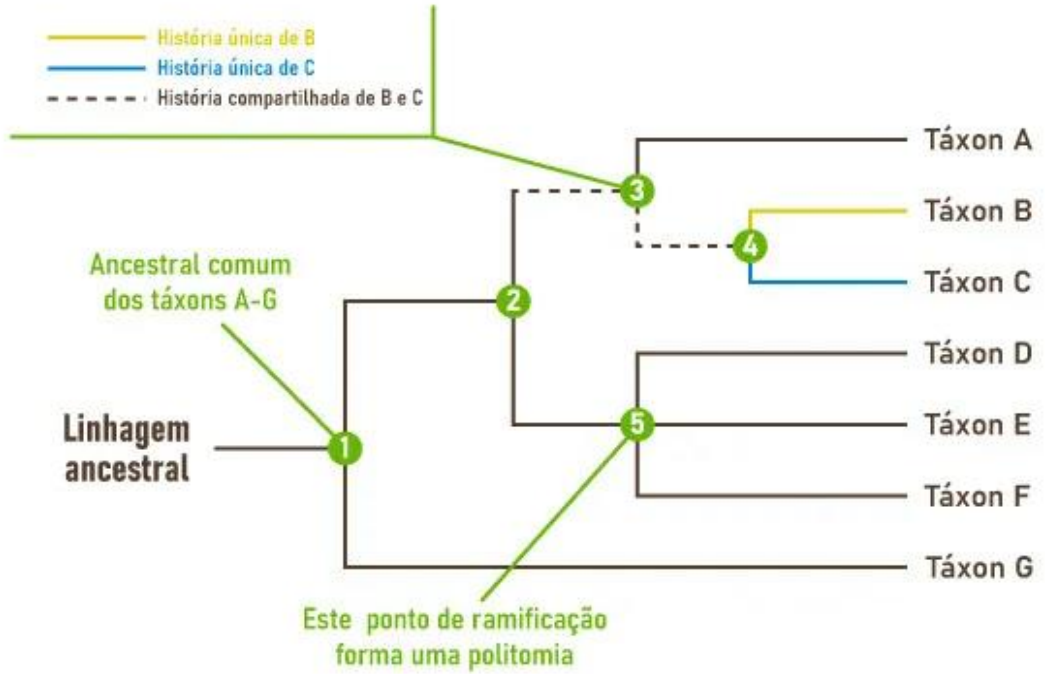
Atenção:

- Capacidade
- Tempo

Árvore Filogenética



Resultado MEGA



Prática Análise Bioinformática de Proteínas
Ferramentas de análise

1. A partir do alinhamento MAFFT gerado no programa Jalview:
   a) Gere diferentes árvores no programa MEGA utilizando diferentes métodos e bootstrap não maior de 5.
   b) Qual método foi mais rápido?
   c) Qual foi o resultado mais diferente?
2. Na árvore gerada pelo método Maximum Likelihood. Há algum ramo/s que tenham algum sentido evolutivo considerando à análise manual das sequências realizada no exercício anterior?