Universidade de São Paulo
Brasil

Análise bioinformática de genomas de parasitas e organismos modelo

“Ferramentas e abordagens computacionais para estudos funcionais e análises da estrutura proteica”

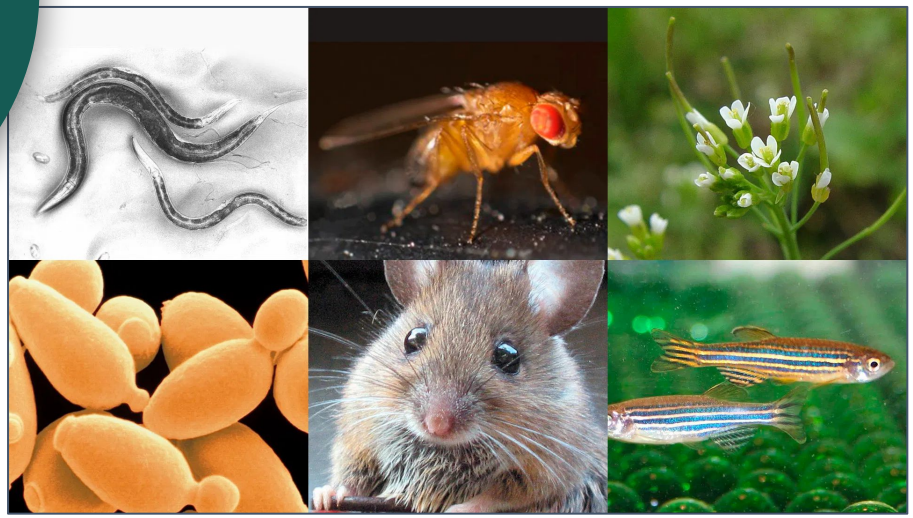
Ana Milena Murillo
Sabrina Marsiccobetre

ammurillog@usp.br
marsiccobetre@usp.br

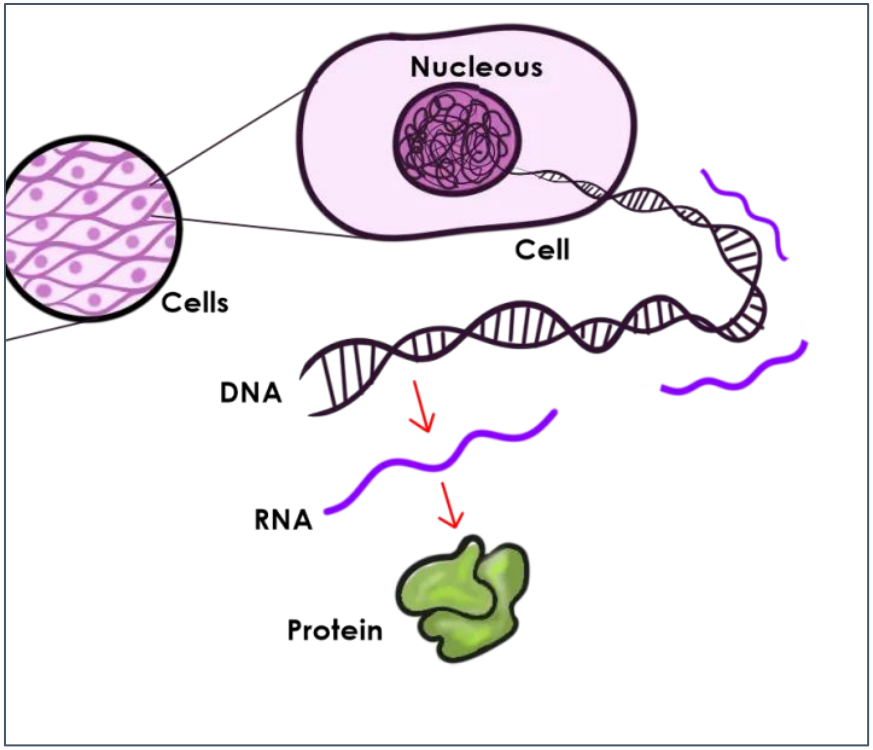
Junho, 2021

Organismos modelo na pesquisa

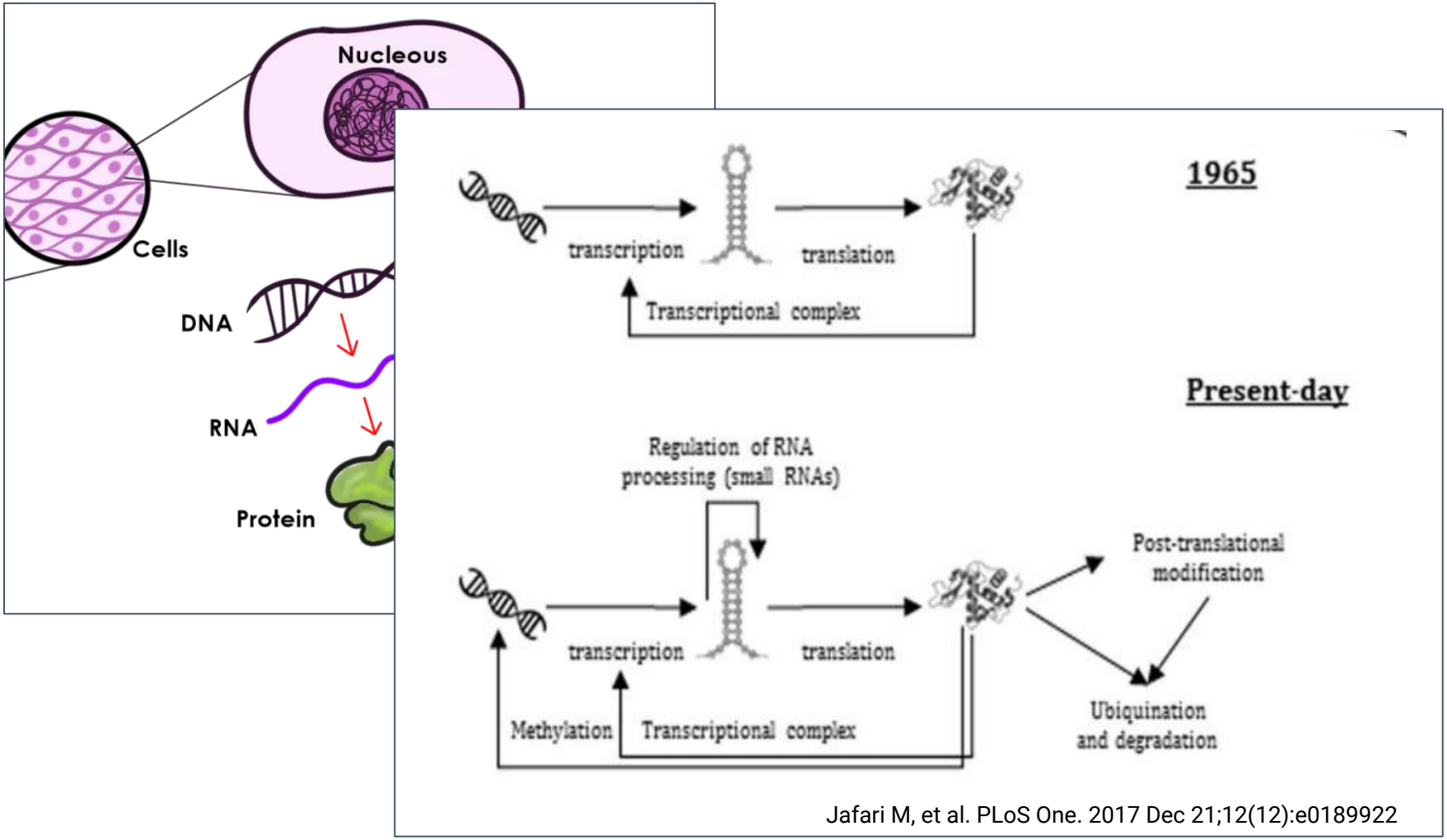
- Manutenção Reprodução
- Processos biológicos específicos
- Similaridade de características genéticas



Dogma Central da Biologia Molecular

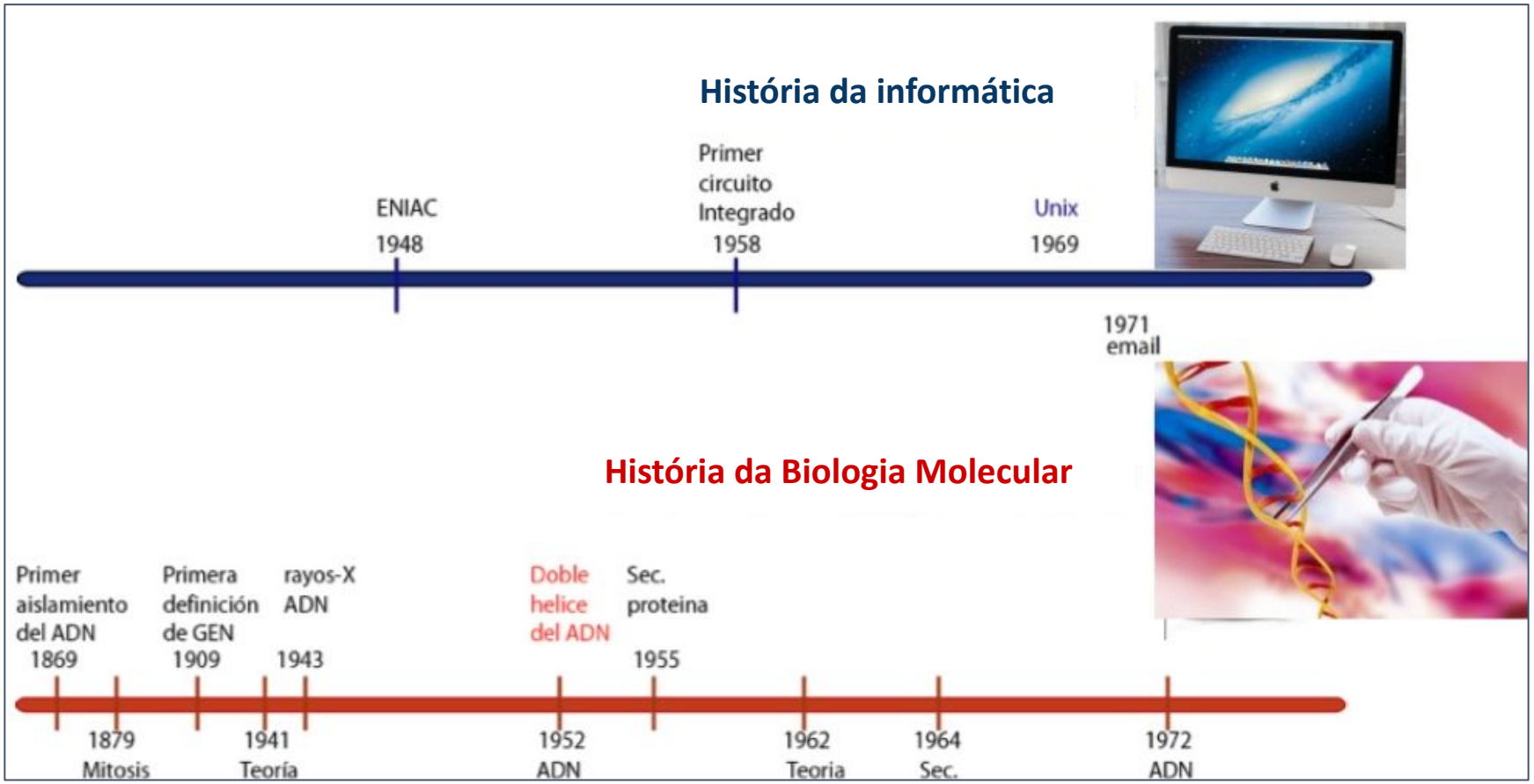


Dogma Central da Biologia Molecular

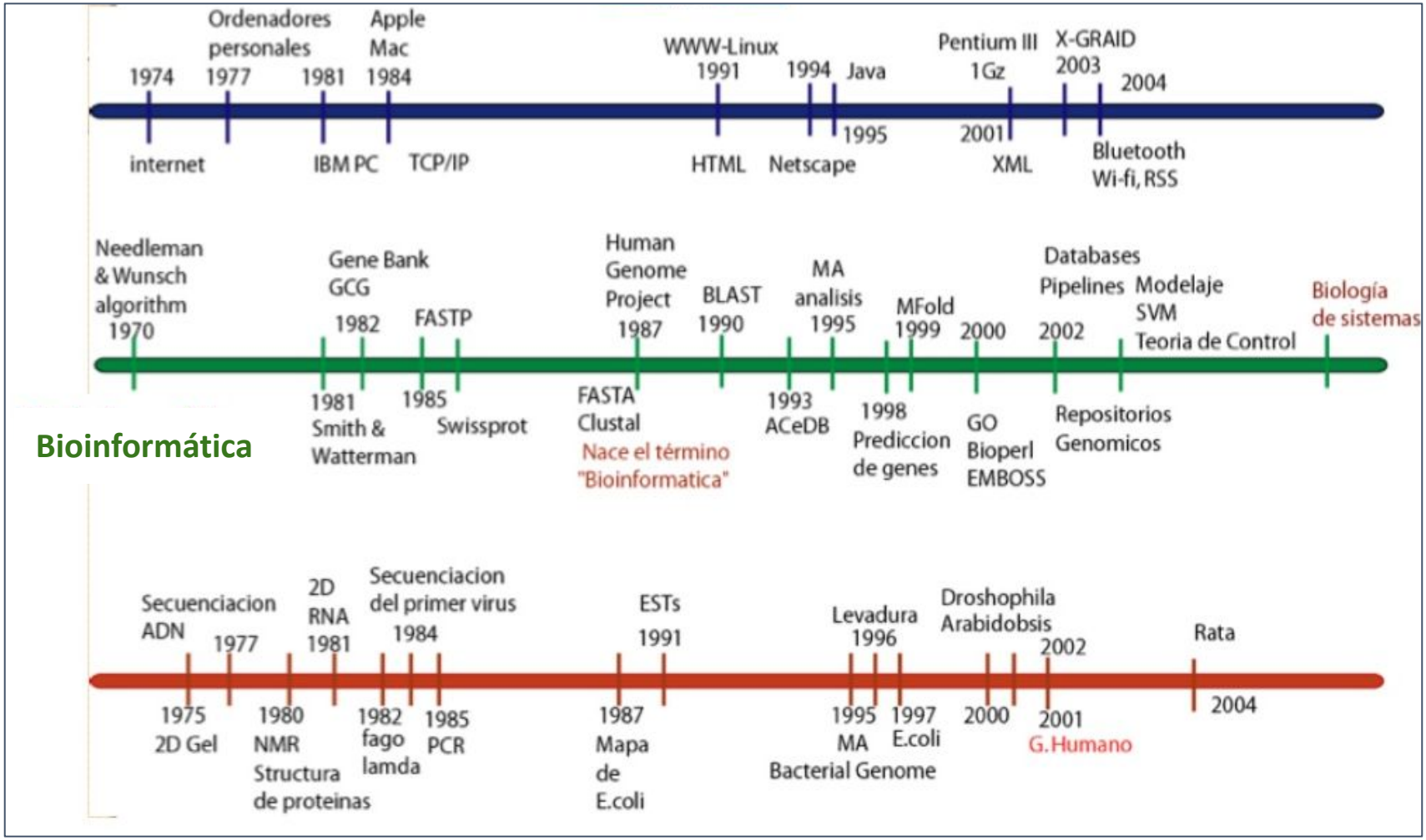


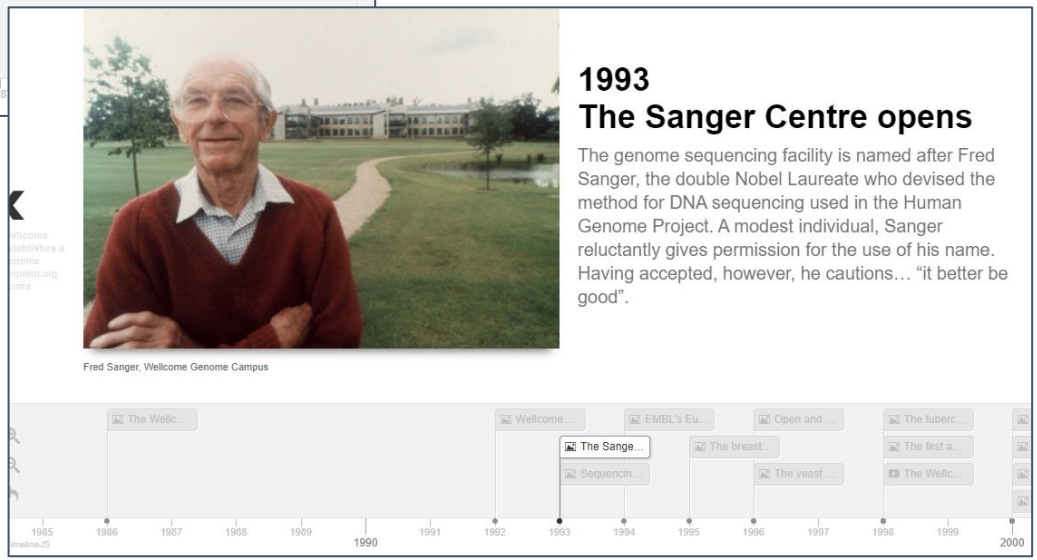
Jafari M, et al. PLoS One. 2017 Dec 21;12(12):e0189922

Desenvolvimento da Bioinformática



Desenvolvimento da Bioinformática





https://www.sanger.ac.uk/about/who-we-are/history-of-the-sanger-institute/

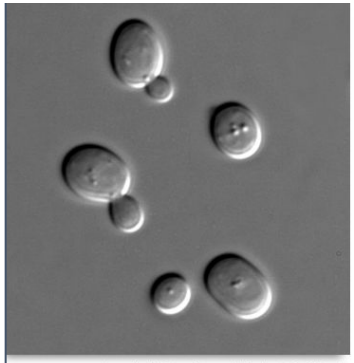


Image Credit: Masur. Licence: Public domain via Wikimedia

ccharomyces cerevisiae

1996
The yeast genome sequence

The Sanger Centre makes the single largest contribution to sequencing the first complete genome from a eukaryotic organism (an organism whose cells have nuclei) (12,000,000 bases). *Saccharomyces cerevisiae* is a yeast used for millennia in winemaking, brewing and baking. It is also an experimental model organism widely employed to study the functions and interactions of genes. Over half of the 6,000 genes revealed by the sequence are previously unknown.

Image Credit: CDC / Alissa Eckert and James Archer

cobacterium tuberculosis bacteria

1998
The tuberculosis bacterium genome sequence

Mycobacterium tuberculosis causes tuberculosis in hundreds of millions of people worldwide, most in low- and middle-income countries. It is one of the first bacteria sequenced and the first of many pathogens sequenced at the Sanger Centre. The sequence (4,000,000 bases) contains 4,000 genes and offers new strategies for diagnosis, prevention and treatment of tuberculosis.

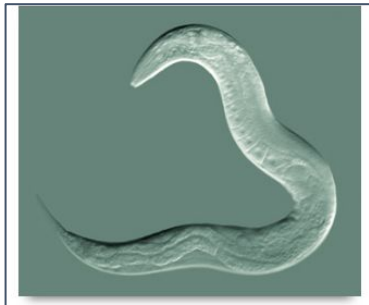
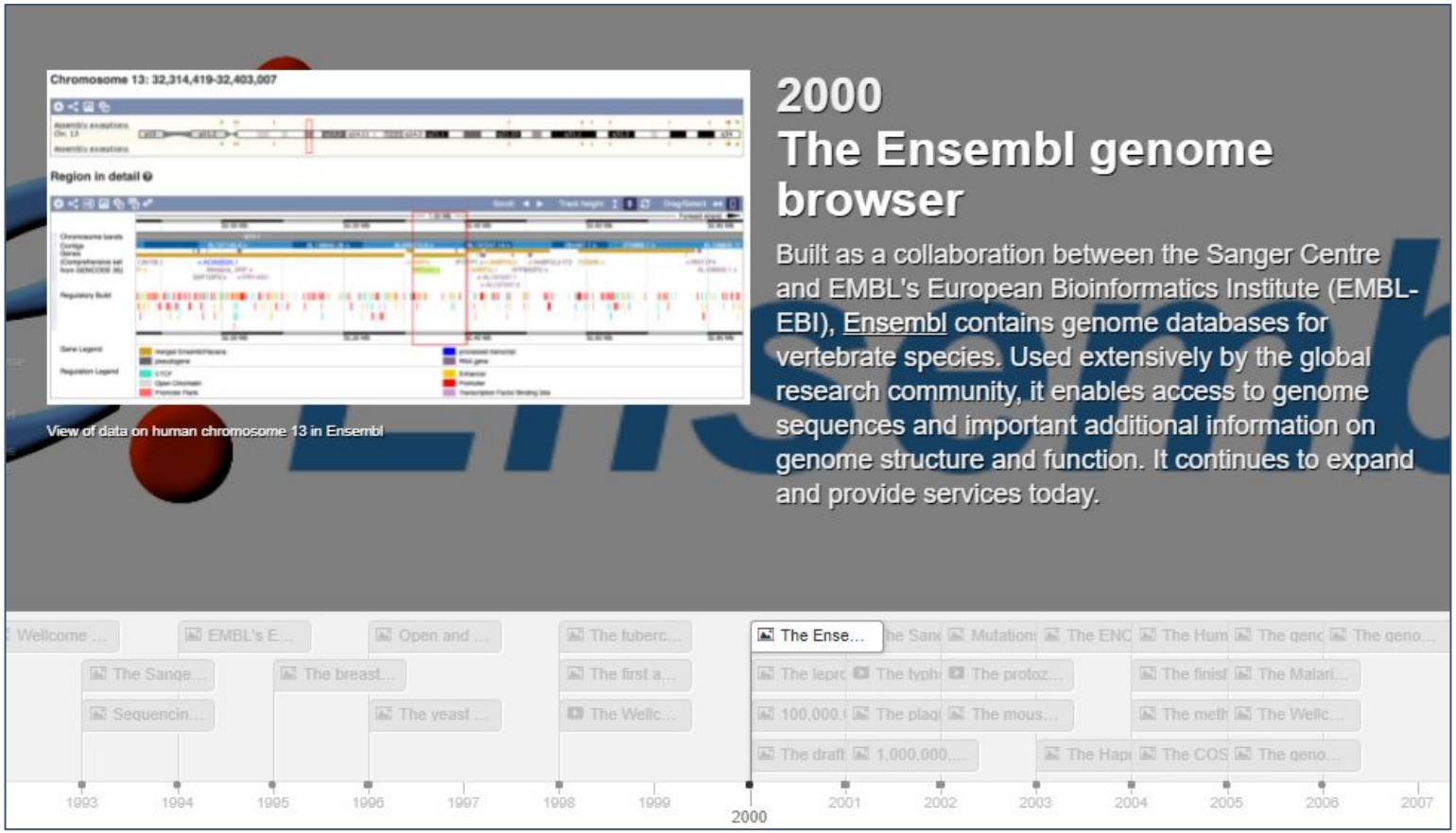


Image Credit: Bob Goldstein. Licence: CC BY-SA 3.0

Caenorhabditis elegans

1998
The first animal genome sequence

Caenorhabditis elegans (100,000,000 bases) is a nematode worm widely used as an experimental model organism. John Sulston, the founding Director of the Sanger Centre, won the Nobel Prize for describing its developmental lineage, and the sequencing of "the worm" was fundamental in developing the tools to sequence the human genome. The sequence has enabled *C.elegans* to provide profound insights into development, neurobiology and ageing.


Chromosome 13: 32,314,419-32,403,007
Region in detail
View of data on human chromosome 13 in Ensembl
2000
The Ensembl genome browser
Built as a collaboration between the Sanger Centre and EMBL's European Bioinformatics Institute (EMBL-EBI), Ensembl contains genome databases for vertebrate species. Used extensively by the global research community, it enables access to genome sequences and important additional information on genome structure and function. It continues to expand and provide services today.
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007

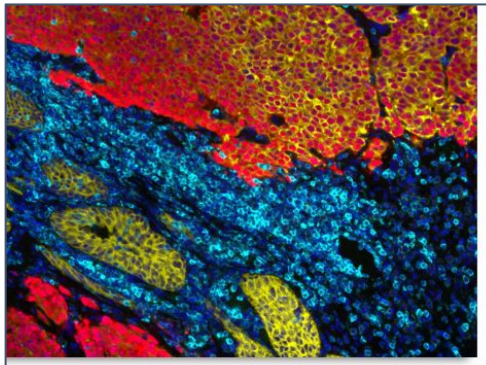
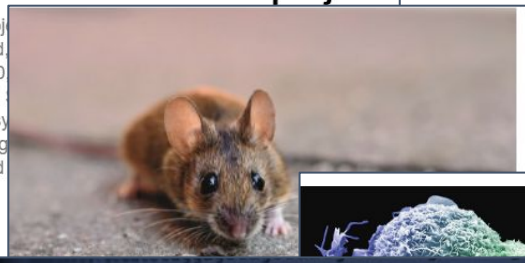


Image Credit: Mateus Crespo, The Institute of Cancer Research. Licence:CC BY 4.0

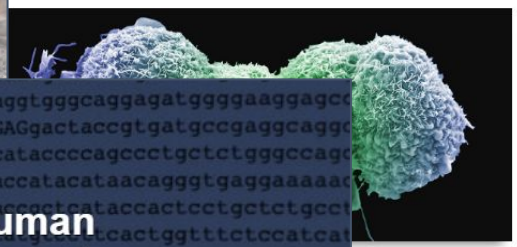
2000
The Cancer Genome project

The proj
mutated.
By 2020
through
world, sy
cancer g
mutated



2002
The mouse genome sequence

The Sanger Institute is a leader in the International



k Institute. Licence: CC-BY-NC-4.0

2009
The first complete cancer genomes are sequenced

Comprehensive genome sequences of a malignant melanoma and a lung cancer by Sanger Institute researchers reveal all the mutations in cancer genomes for the first time. Analysis of the lung cancer genome suggests a typical smoker acquires one mutation in a lung cell for every 15 cigarettes smoked.

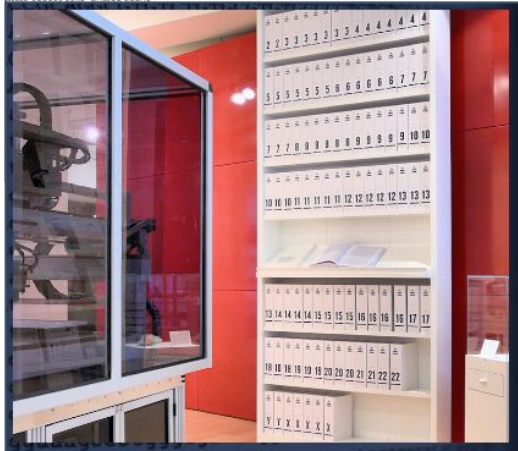


Image credit: Russ London at English Wikipedia. Licence: CC BY-SA 3.0

A printout of the human genome presented as a series of books at the Wellcome Collection, London. The 3.4 billion units of DNA code are written into more than a hundred volumes, each a thousand pages long.

2004
The finished human genome sequence is published

Analysis of the finished human genome is published
and
culm
accu
iden
new

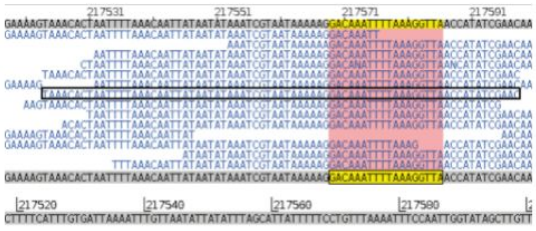


Image Credit: Adam Reid, Wellcome Sanger Institute

Reads aligned to a reference genome sequence using BWA

2009
Computational tools for analysing human genome sequences

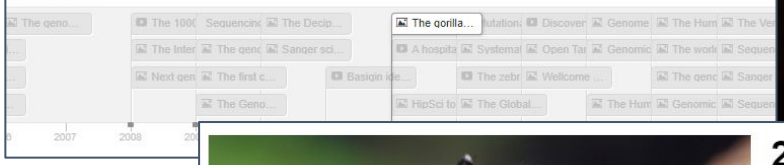
Sanger researchers develop the Burrows-Wheeler Alignment (BWA) tool. The freely available software aligns short reads of DNA sequence produced by next-generation technologies against a reference sequence, such as the human genome, more accurately and faster than existing tools. BWA remains in use today and by 2020 has been cited over 17,000 times.

Image Credit: San Diego Zoo

Kamilah the Gorilla

2012
The gorilla genome sequence

The DNA sequence (3,000,000,000 bases) of female gorilla, named Kamilah, provides insig the evolution of the great apes, which include *sapiens*. It sheds new light on human languag development, neurodegenerative illness, and information relevant to survival of mountain g species.



2013
The zebrafish genome sequence

The zebrafish is an important experimental model organism used to understand embryonic development in health and disease. Sanger's zebrafish genome sequence (1,600,000,000 bases) reveals that humans share about 70 per cent of their genome with the zebrafish. Zebrafish becomes a model of choice in screening and testing new treatments for a range of diseases.

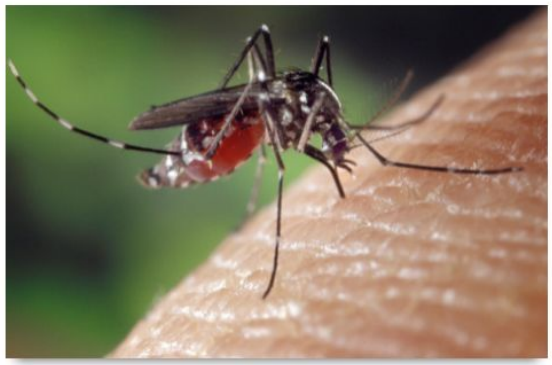
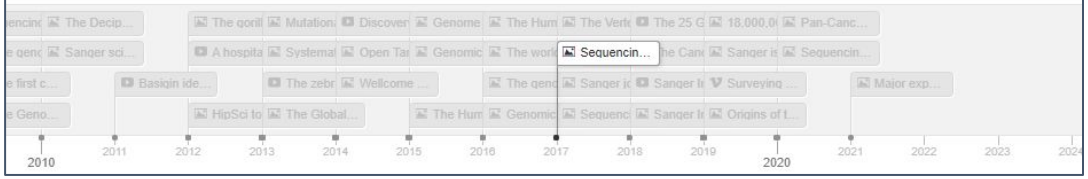


Image Credit: Pixabay

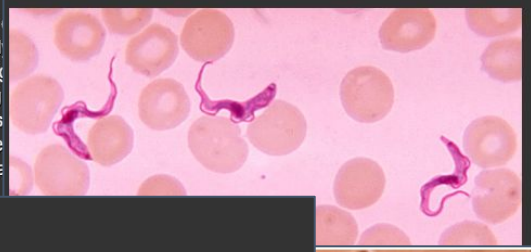
2017
Sequencing mosquito genomes reveals spread of insecticide resistance across Africa

Anopheles mosquitos transmit the malaria parasites that cause hundreds of thousands of deaths every year. In the largest study of Anopheles mosquito genomes to date, Sanger researchers with African collaborators discover that mosquitos are extremely genetically diverse, providing clues as to why they quickly become insecticide-resistant.



2002
The protozoan causing the deadliest form of malaria is sequenced

Plasmodium falciparum (
each year, predominantl
The genome sequence (
improves understanding
severe disease, enables
treatments, and become
spread of resistance to a



2005
The genome sequence of the protozoan causing sleeping sickness

Trypanosoma brucei, transmitted by tsetse flies, causes sleeping sickness, a common and deadly disease in sub-Saharan Africa. The genome

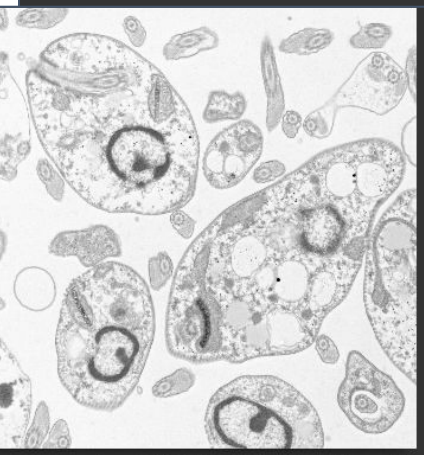
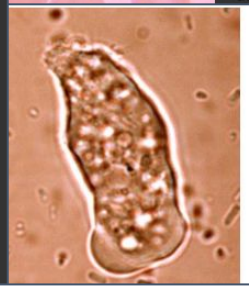


Image Credit: Cynthia Goldsmith and Luciana Flannery, CDC
smission electron microscopic image of Leishmania major

2006
The genome sequence of the protozoan causing leishmaniasis

Leishmania species cause a spectrum of disfiguring and sometimes fatal diseases. Two million people in 100 countries are affected each year. Scientists at the Sanger Institute sequence 24 of the 36 chromosomes in the (
(32,80
search



2005
The genome sequence of the protozoan causing amoebic dysentery

Entamoeba histolytica is an intestinal parasite which causes the bowel disease amoebiasis, a significant cause of illness and death in low- and middle-income tropical countries. The genome sequence (23,700,000
rchers provides new
ution of this major

Image Credit: Bruce Wetzel and Harry Schaefer, National Cancer Institute
Schistosome parasite

2009
The genome sequence of the worm causing schistosomiasis

Sanger Institute researchers publish the complete sequence of *Schistosoma mansoni* (363,000,000 bases), a parasitic worm transmitted in drinking water and responsible for schistosomiasis which affects 200 million people worldwide. The genome sequence provides a foundation for understanding the parasite's complex biology enabling further research into new treatments.

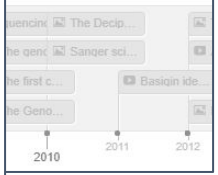


Image Credit: Dan Ross / Wellcome Sanger Institute

Coronavirus samples in storage at the Wellcome Sanger Institute

2020
Sequencing of Covid-19 genomes to monitor the pandemic at a national scale

Partners in the COVID-19 Genomics UK consortium (COG-UK), the Sanger Institute generates genomes at scale for national surveillance of the UK pandemic. The data enables tracking virus spread in real time, detection of new hypertransmissible strains, monitoring vaccine effectiveness and supports public health authorities in tackling local outbreaks.

Imagem: Getty Images

Em 48 horas, brasileiros sequenciam genoma do coronavírus

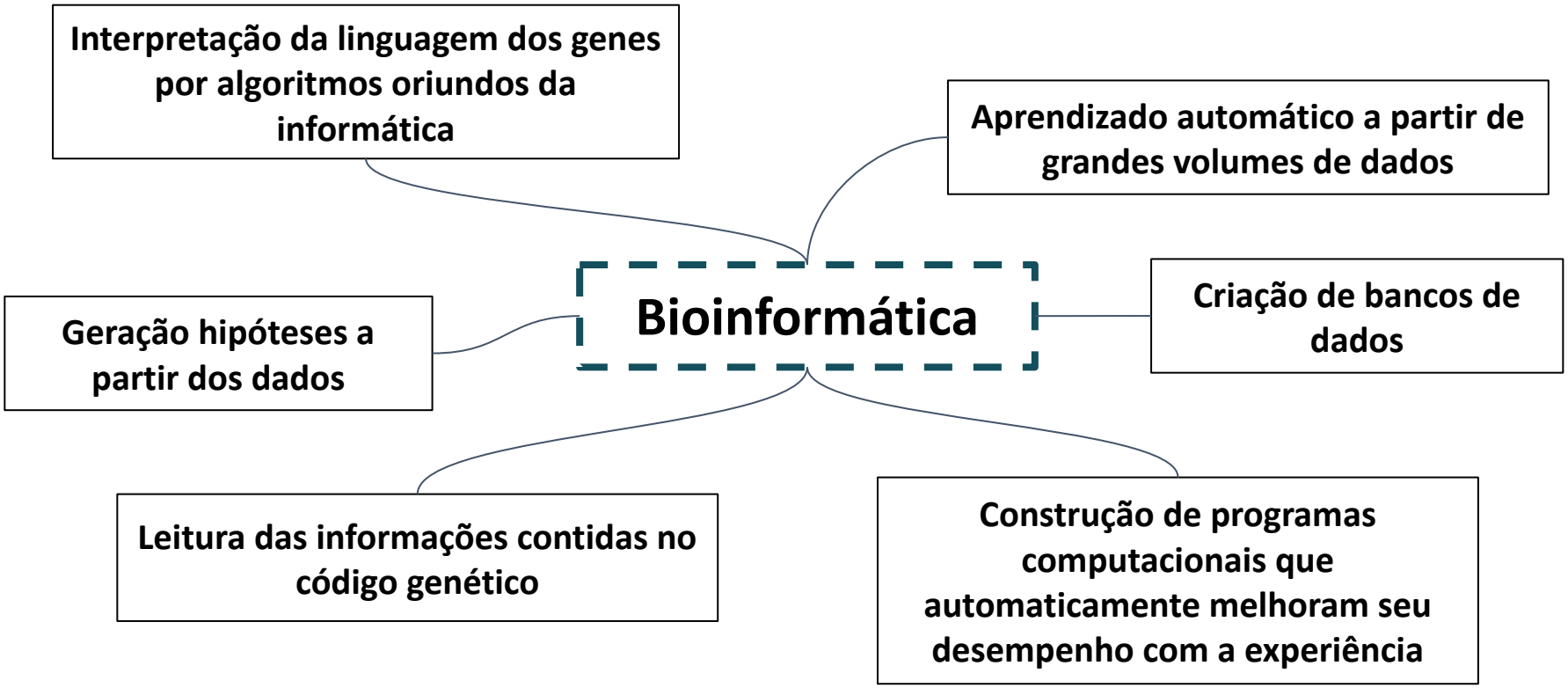
ESTADÃO conteúdo

Giovana Girardi

29/02/2020 11h20

Em só 48 horas desde a confirmação do primeiro caso no Brasil d pelo novo coronavírus, pesquisadores brasileiros conseguiram seq genoma do vírus que chegou ao País. O trabalho foi conduzido por cientistas do Instituto Adolfo Lutz, do Instituto de Medicina Tropical da

Fonte: ARAÚJO et al. (2008, p. 146)



- Ciência e tecnologia de aprendizagem, gestão e processamento de informação biológica - utilização de métodos computacionais, matemáticos e estatísticos para analisar dados biológicos, bioquímicos e biofísicos
- Biologia molecular, computação, estatística, matemática, engenharia

Bases de dados primárias

Nucleotídeos

- GenBank (NCBI)
- ENA (European Nucleotide Archive-EMBL-EBI)
- DDBJ (DNA Bank of Japan)
- NDB (Nucleic acid database)

Proteínas

- SWISS-PROT
- UniProtKB
- PDB (Protein Data Bank)

Bases de dados secundárias

Proteínas

- ProSite
- Pfam

Otras

- Gene-Ontology
- Refseq

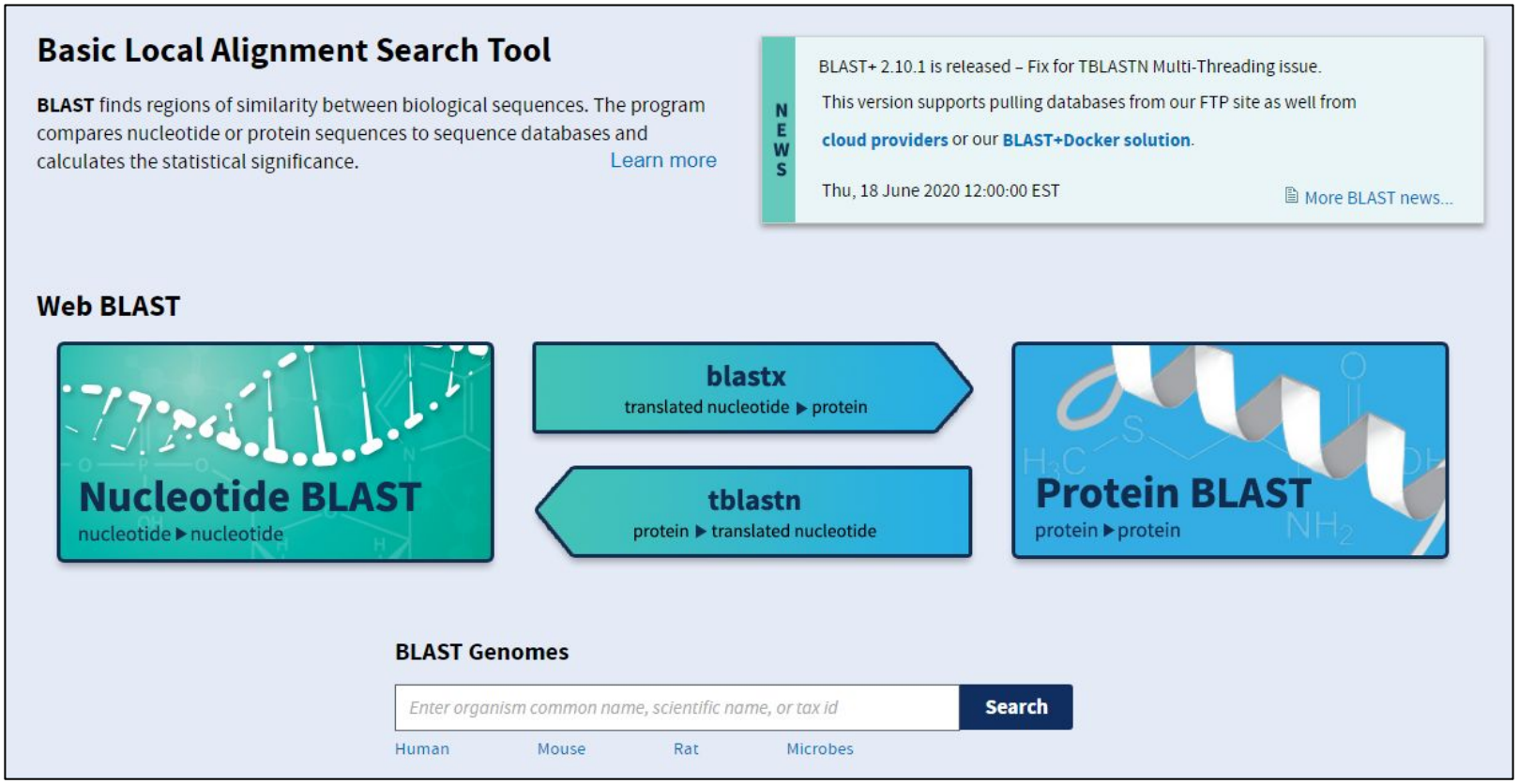
Bases de dados especializadas

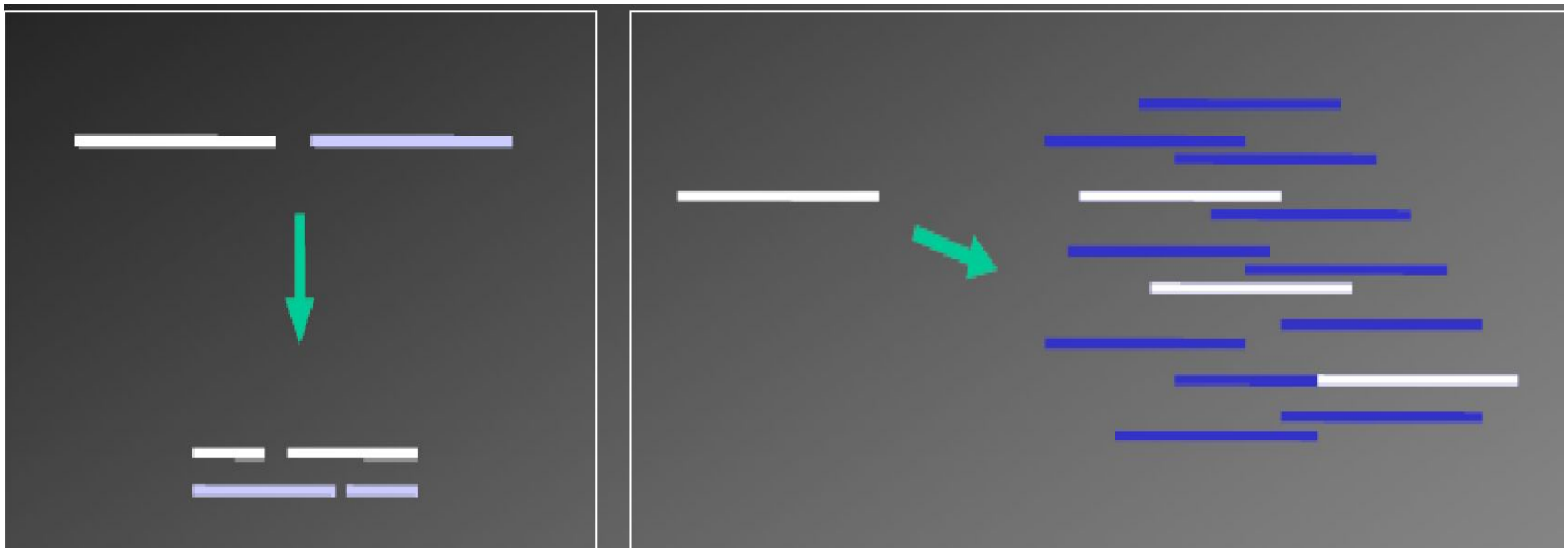
ZFIN (Zebra fish)
Flybase (*D. melanogaster*)
TAIR (*Arabidopsis*)
ENSEML (Homem, camundongo e outros)
KEGG (Metabolic pathways)
OMIM (Human genetic disorders)
SGD (*Saccharomyces*)

Entendendo as ferramentas!!!

BLAST

https://blast.ncbi.nlm.nih.gov/Blast.cgi





Alinhamento de sequências **Busca de sítios homólogos**

- **Cadeias curtas de alta similaridade**
  - **Grandes bancos de dados**
- **E-value < 1 e^{-05} → muito confiável**

Homólogos → origem comum em organismos diferentes. Não indica a mesma função

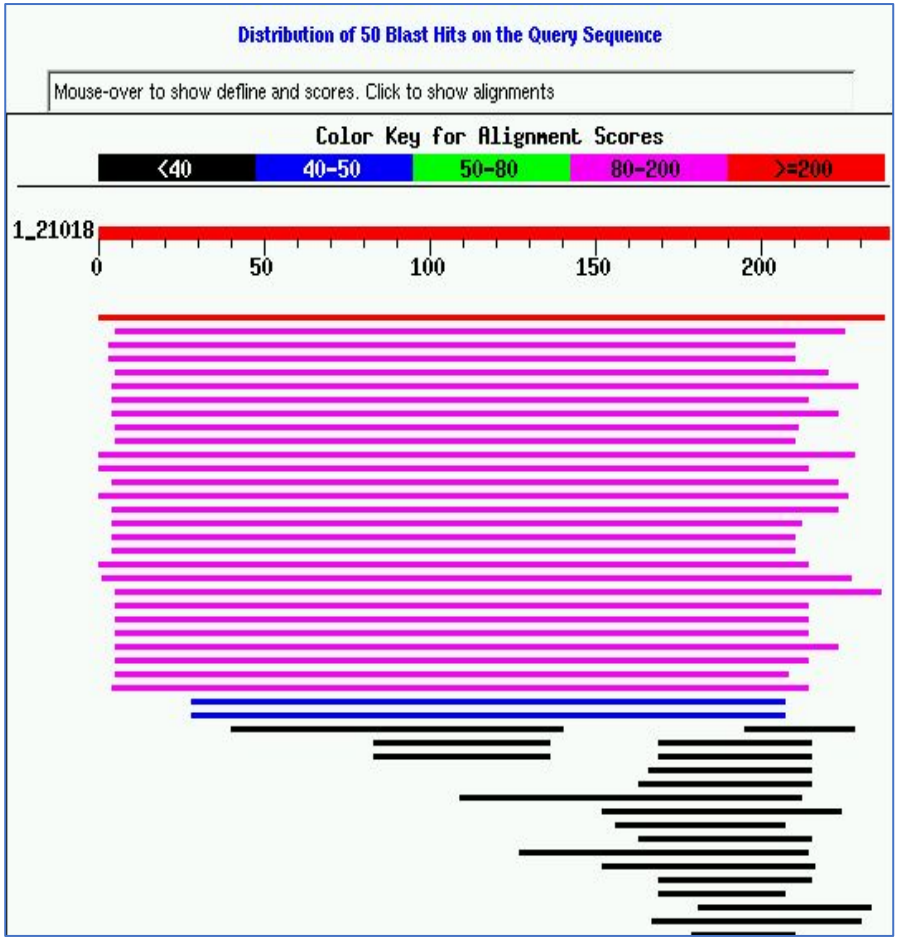
Descriptions | Graphic Summary | Alignments | Taxonomy

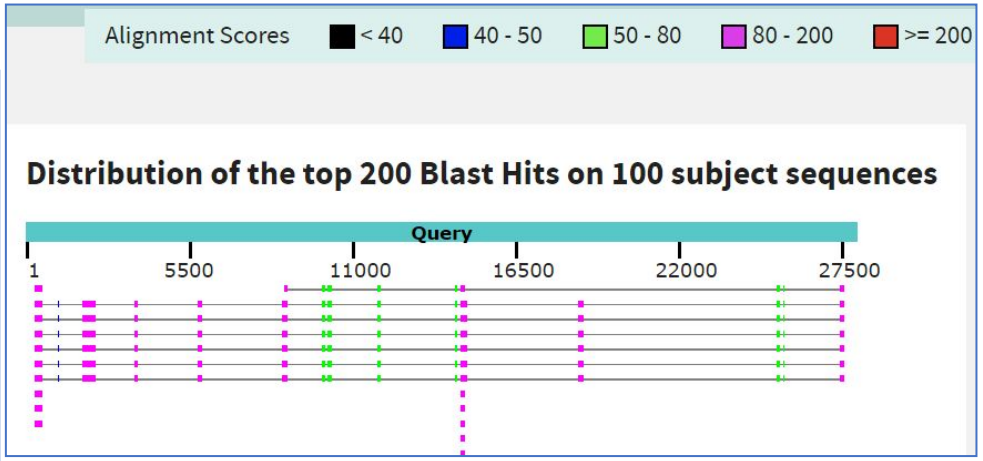
Sequences producing significant alignments Download Manage Columns Show 100

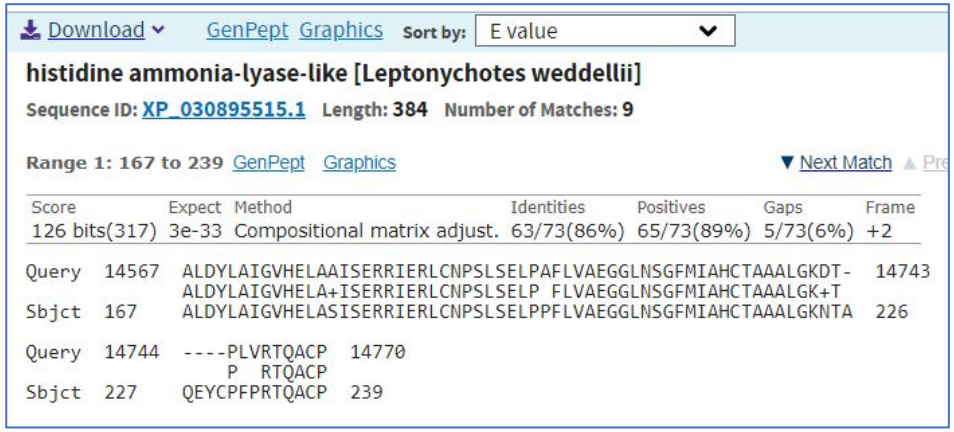
select all *100 sequences selected* GenPept Graphics

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
histidine ammonia-lyase [Mus musculus]	148	397	19%	8e-33	100.00%	NP_034531.1
histidine ammonia-lyase [Mus caroli]	147	396	19%	2e-32	98.78%	XP_021029505.1
histidine ammonia-lyase [Mastomys coucha]	143	389	19%	3e-31	96.34%	XP_031204984.1
histidine ammonia-lyase [Arvicanthis niloticus]	143	387	19%	3e-31	96.34%	XP_034343177.1
histidine ammonia-lyase [Meriones unguiculatus]	143	380	19%	4e-31	95.12%	XP_021511140.1
histidine ammonia-lyase [Grammomys surdaster]	142	382	19%	5e-31	96.34%	XP_028638877.1
histidine ammonia-lyase [Rattus norvegicus]	142	387	19%	9e-31	95.12%	NP_058855.1
histidine ammonia-lyase [Cricetulus griseus]	141	380	19%	1e-30	95.12%	XP_003501465.1
PREDICTED: histidine ammonia-lyase [Peromyscus maniculatus bairdii]	140	379	19%	3e-30	93.90%	XP_006988520.1
histidine ammonia-lyase [Mus pahari]	140	387	19%	3e-30	95.12%	XP_021060879.1
histidine ammonia-lyase [Nannospalax galili]	140	370	19%	4e-30	91.46%	XP_008830287.1
histidine ammonia-lyase isoform X2 [Leptonychotes weddellii]	132	354	19%	4e-30	86.59%	XP_030893358.1
histidine ammonia-lyase [Mesocricetus auratus]	139	378	19%	6e-30	92.68%	XP_005068059.1
histidine ammonia-lyase isoform X1 [Leptonychotes weddellii]	132	353	19%	1e-29	86.59%	XP_030893357.1
histidine ammonia-lyase [Microtus ochrogaster]	138	378	19%	1e-29	92.68%	XP_005358216.1
histidine ammonia-lyase isoform X2 [Pan troglodytes]	137	365	19%	1e-29	91.46%	XP_024203778.1
histidine ammonia-lyase isoform X2 [Nomascus leucogenys]	137	362	19%	2e-29	91.46%	XP_0

Feed

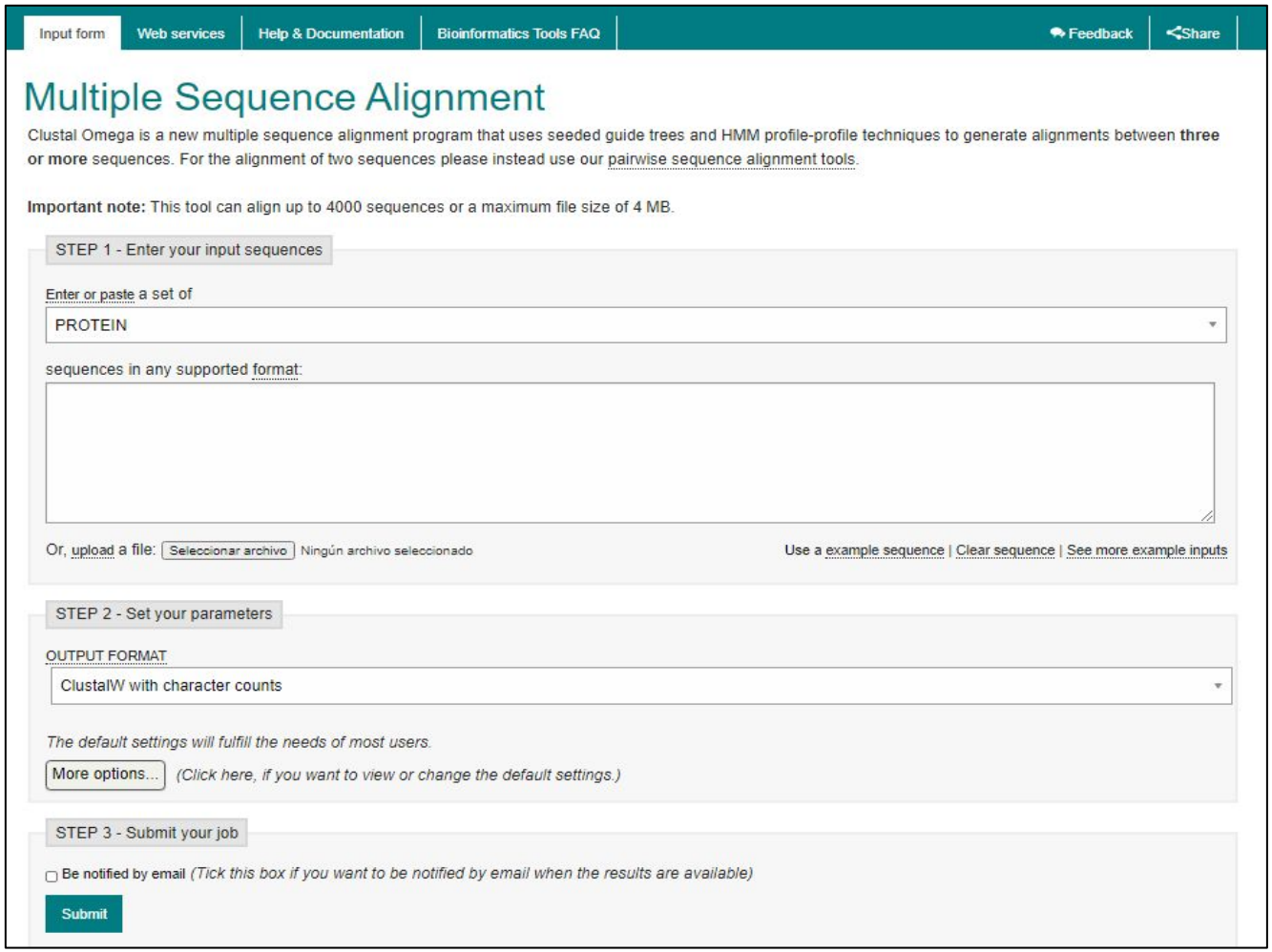

Distribution of 50 Blast Hits on the Query Sequence
Mouse-over to show defline and scores. Click to show alignments
Color Key for Alignment Scores
<40
40-50
50-80
80-200
>=200
1_21018
0
50
100
150
200

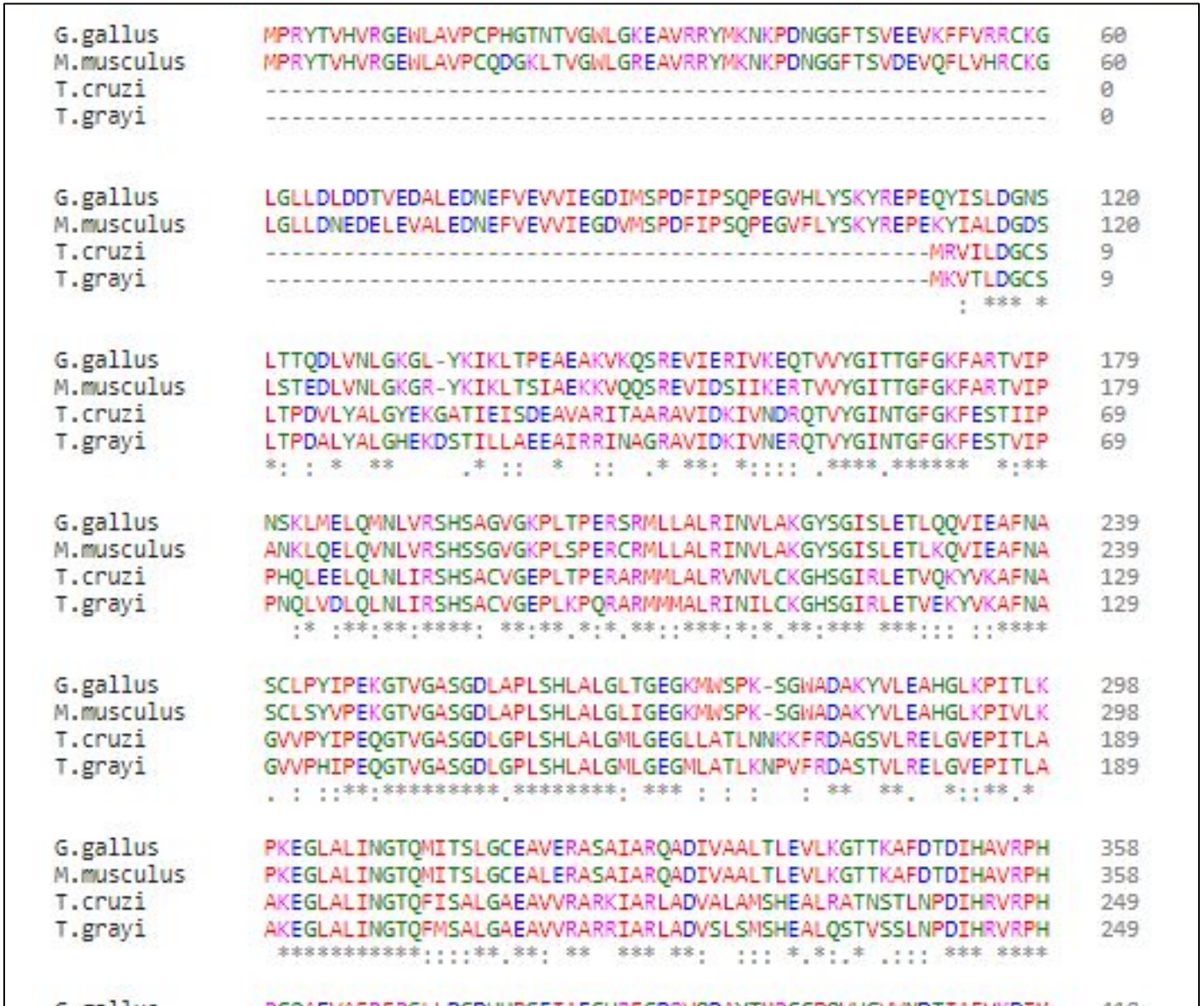

Alignment Scores
< 40
40 - 50
50 - 80
80 - 200
>= 200
Distribution of the top 200 Blast Hits on 100 subject sequences
Query
1
5500
11000
16500
22000
27500


Download
GenPept
Graphics
Sort by:
E value
histidine ammonia-lyase-like [Leptonychotes weddellii]
Sequence ID: XP_030895515.1 Length: 384 Number of Matches: 9
Range 1: 167 to 239 GenPept Graphics
Next Match
Score Expect Method Identities Positives Gaps Frame
126 bits(317) 3e-33 Compositional matrix adjust. 63/73(86%) 65/73(89%) 5/73(6%) +2
Query 14567 ALDYLAIGVHELAAISERRIERLCNPSLSELPAFLVAEGGLNSGFMIAHCTAAALGKDT- 14743
ALDYLAIGVHELA+ISERRIERLCNPSLSELP FLVAEGGLNSGFMIAHCTAAALGK+T
Sbjct 167 ALDYLAIGVHELASISERRIERLCNPSLSELPPFLVAEGGLNSGFMIAHCTAAALGKNTA 226
Query 14744 ----PLVRTQACP 14770
P RTQACP
Sbjct 227 QEYCPFPRTQACP 239

Clustal omega

https://www.ebi.ac.uk/Tools/msa/clustalo/

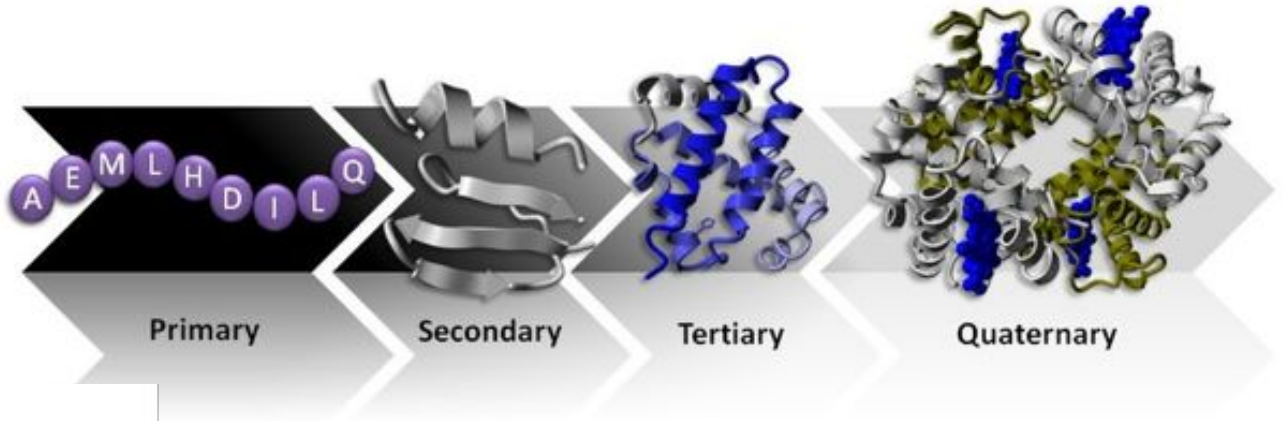



G.gallus MPRYTVHVRGEWLAVPCPHGTNTVGWLGKEAVRRYMKNKPDNGGFTSVEEVKFFVRRCKG 60
M.musculus MPRYTVHVRGEWLAVPCQDGKLTVGWLGREAVRRYMKNKPDNGGFTSVDEVQFLVHRCKG 60
T.cruzi ---------- 0
T.grayi ---------- 0
G.gallus LGLLDLDDTVEDALEDNEFVEVVIEGDIMSPDFIPSQPEGVHLYSKYREPEQYISLDGNS 120
M.musculus LGLLDNEDELEVALEDNEFVEVVIEGDVMSPDFIPSQPEGVFLYSKYREPEKYIALDGDS 120
T.cruzi ----------MRVILDGCS 9
T.grayi ----------MKVTLDGCS 9
G.gallus LTTQDLVNLGKGL-YKIKLTPEAEAKVKQSREVIERIVKEQTVVYGITTGFGKFARTVIP 179
M.musculus LSTEDLVNLGKGR-YKIKLTSIAEKKVQQSREVIDSIIKERTVVYGITTGFGKFARTVIP 179
T.cruzi LTPDVLYALGYEKGATIEISDEAVARITAARAVIDKIVNDRQTVYGINTGFGKFESTIIP 69
T.grayi LTPDALYALGHEKDSTILLAEEAIRRINAGRAVIDKIVNERQTVYGINTGFGKFESTVIP 69
G.gallus NSKLMELQMNLVRSHSAGVGKPLTPERSRMLLALRINVLAKGYSGISLETLQQVIEAFNA 239
M.musculus ANKLQELQVNLVRSHSSGVGKPLSPERCRMLLALRINVLAKGYSGISLETLKQVIEAFNA 239
T.cruzi PHQLEELQLNLIRSHSACVGEPLTPERARMMLALRVNVLCKGHSGIRLETVQKYVKAFNA 129
T.grayi PNQLVDLQLNLIRSHSACVGEPLKPQRARMMMALRINILCKGHSGIRLETVEKYVKAFNA 129
G.gallus SCLPYIPEKGTVGASGDLAPLSHLALGLTGEGKMNSPK-SGNADAKYVLEAHGLKPITLK 298
M.musculus SCLSYVPEKGTVGASGDLAPLSHLALGLIGEGKMNSPK-SGNADAKYVLEAHGLKPIVLK 298
T.cruzi GVVPYIPEQGTVGASGDLGPLSHLALGMLGEGLLATLNNKKFRDAGSVLRELGVEPITLA 189
T.grayi GVVPHIPEQGTVGASGDLGPLSHLALGMLGEGMLATLKNPVFRDASTVLRELGVEPITLA 189
G.gallus PKEGLALINGTQMITSLGCEAVERASAIARQADIVAALTLEVLKGTTKAFDTDIHAVRPH 358
M.musculus PKEGLALINGTQMITSLGCEALERASAIARQADIVAALTLEVLKGTTKAFDTDIHAVRPH 358
T.cruzi AKEGLALINGTQFISALGAEAVVRARKIARLADVALAMSHEALRATNSTLNPDIHRVRPH 249
T.grayi AKEGLALINGTQFMSALGAEAVVRARRIARLADVSLSMSHEALQSTVSSLNPDIHRVRPH 249

RCSB PDB
PROTEIN DATA BANK

https://www.rcsb.org

1971 Criação do PDB.
Em 1974 tinha 12 estruturas
24/06/21→ 179210

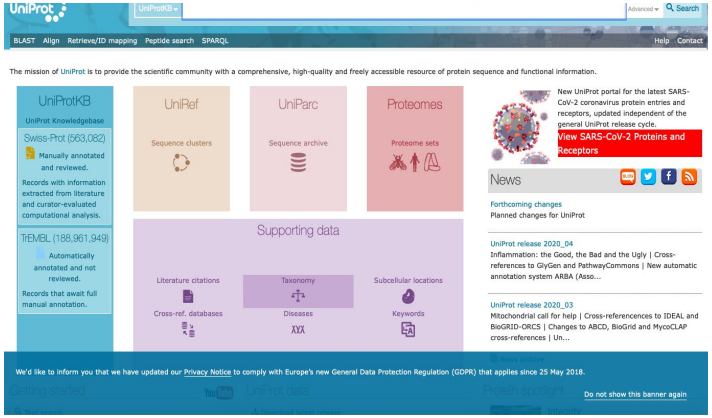



myoglobin


hemoglobin

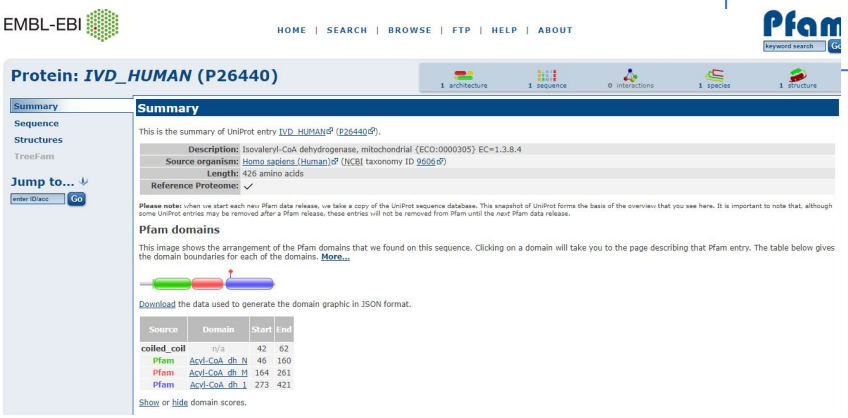

carboxypeptidase A


lactate dehydrogenase

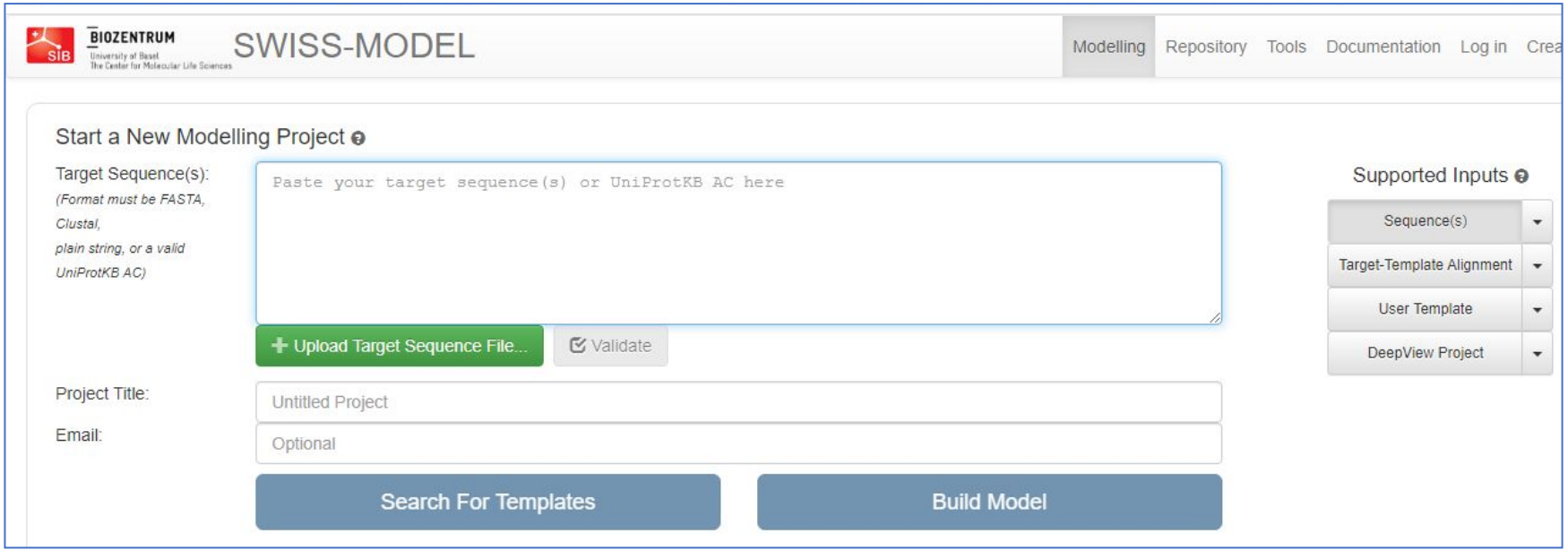


Uniprot
https://www.uniprot.org

Pfam
https://pfam.xfam.org



Modelagem de proteínas por homologia



https://swissmodel.expasy.org

	Max score	Total score	Query cover	E value	Ident	Accession
isovaleryl-coA dehydrogenase [Trypanosoma rangeli SC58]	707	707	98%	0.0	83%	ESL07937.1
isovaleryl-coA dehydrogenase [Trypanosoma theileri]	654	654	92%	0.0	81%	ORC91678.1
isovaleryl-coA dehydrogenase, putative [Trypanosoma brucei gambiense DAL972]	627	627	94%	0.0	76%	XP_011779312.1
isovaleryl-coA dehydrogenase, putative [Trypanosoma brucei brucei TREU927]	626	626	94%	0.0	76%	XP_828243.1
unnamed protein product [Trypanosoma congolense IL3000]	595	595	98%	0.0	72%	CCD16413.1
putative isovaleryl-coA dehydrogenase [Leishmania infantum JPCM5]	580	580	99%	0.0	71%	XP_001466486.1
putative isovaleryl-coA dehydrogenase [Leishmania major strain Friedlin]	579	579	95%	0.0	71%	XP_003721899.1
putative mitochondrial isovaleryl-coA dehydrogenase [Leptomonas pyrrhocoris]	579	579	96%	0.0	71%	XP_015659347.1
putative isovaleryl-coA dehydrogenase [Leishmania mexicana MHOM/GT/2001/U1103]	577	577	98%	0.0	71%	XP_003876667.1
putative isovaleryl-coA dehydrogenase [Leishmania braziliensis MHOM/BR/75/M2904]	575	575	99%	0.0	70%	XP_001565860.1
isovaleryl-coA dehydrogenase, putative [Leishmania panamensis]	575	575	95%	0.0	71%	XP_010700182.1
isovaleryl-coA dehydrogenase, putative [Leishmania guyanensis]	573	573	99%	0.0	70%	CCM16690.1
isovaleryl-coA dehydrogenase, putative [Bodo saltans]	573	573	94%	0.0	69%	CUE72386.1
putative isovaleryl-coA dehydrogenase [Leptomonas seymouri]	566	566	91%	0.0	72%	KPI87702.1
putative isovaleryl-coA dehydrogenase, fragment [Trypanosoma vivax Y486]	554	554	85%	0.0	78%	CCC52663.1
isovaleryl-CoA dehydrogenase [Sandaracinus amylolyticus]	496	496	91%	2e-172	64%	WP_053234518.1
isovaleryl-CoA dehydrogenase [Nannocystis exedens]	494	494	90%	1e-171	64%	WP_096327858.1
isovaleryl-CoA dehydrogenase [Plesiocystis pac				3e-170	62%	WP_006
isovaleryl-CoA dehydrogenase [Enhygromyxa sa				2e-165	62%	WP_106

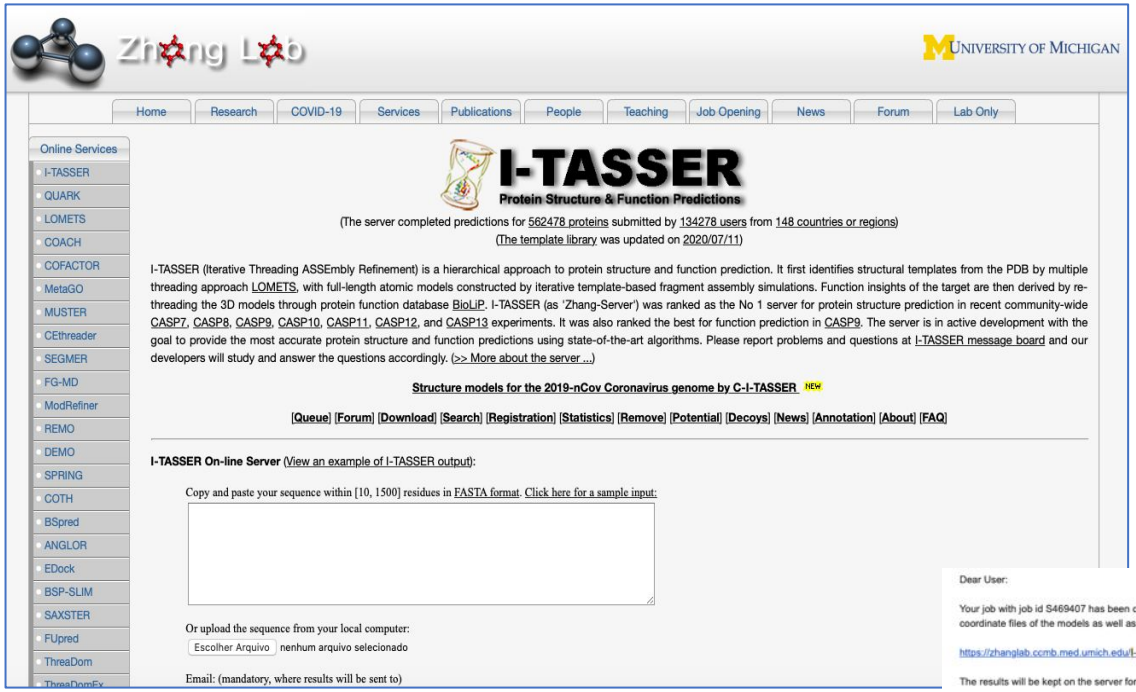
Model 1 □ 1jqi: Crystal Structure of Rat Short Chain Acyl-CoA Dehydrogenase Complexed With Acetoacetyl-CoA

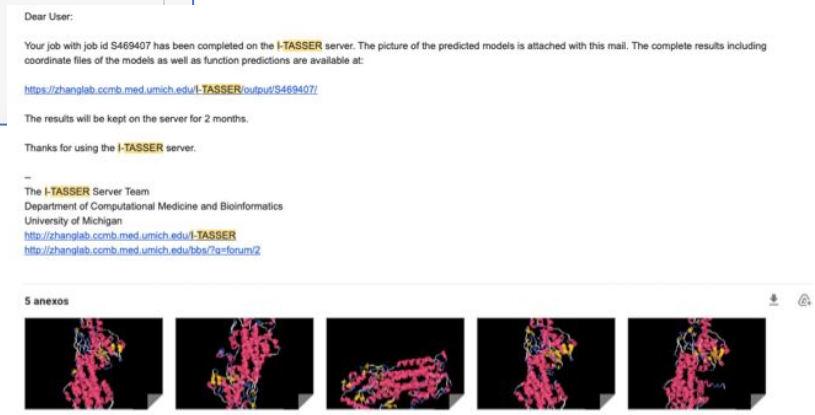
Model 2 □ ivh: Structure of human isovaleryl-coa dehydrogenase at 2.6 angstroms resolution: structural basis for substrate specificity

Model 1

Model 2

Modelagem de proteínas por homologia

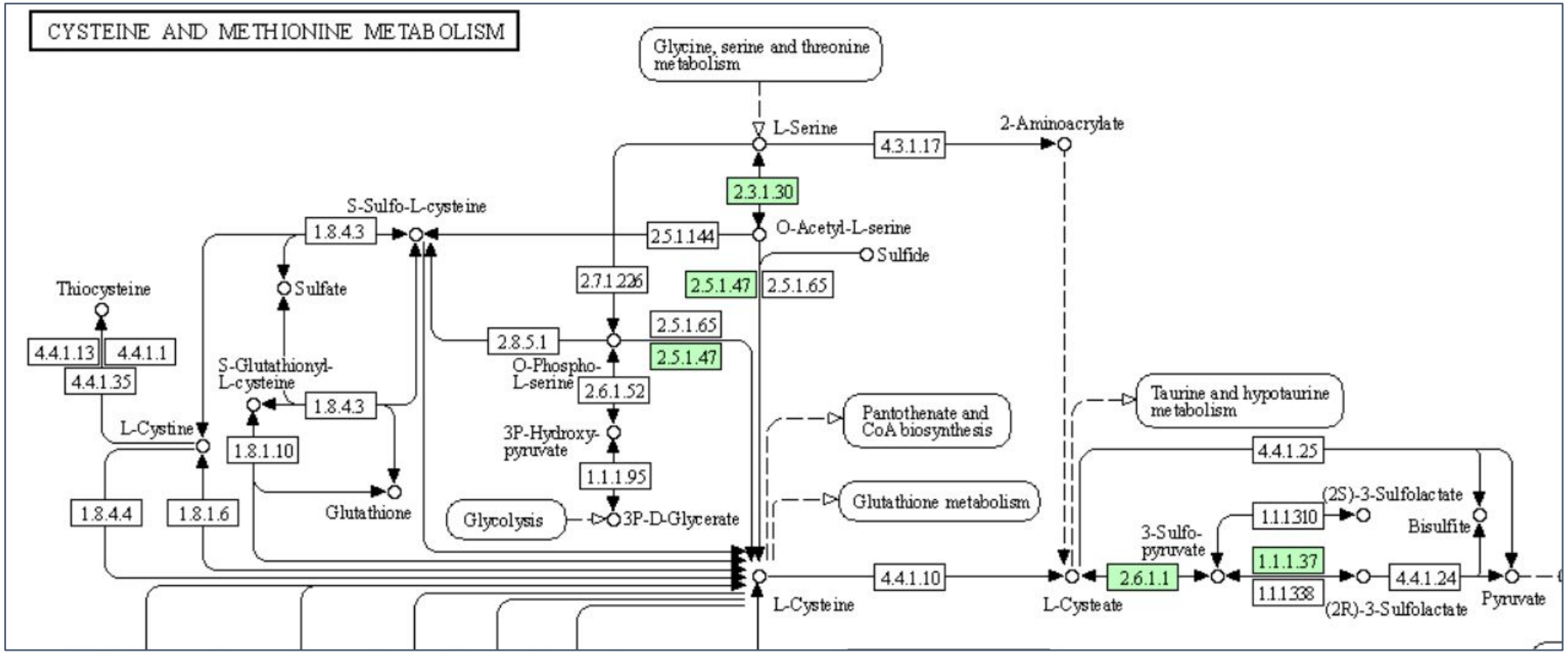




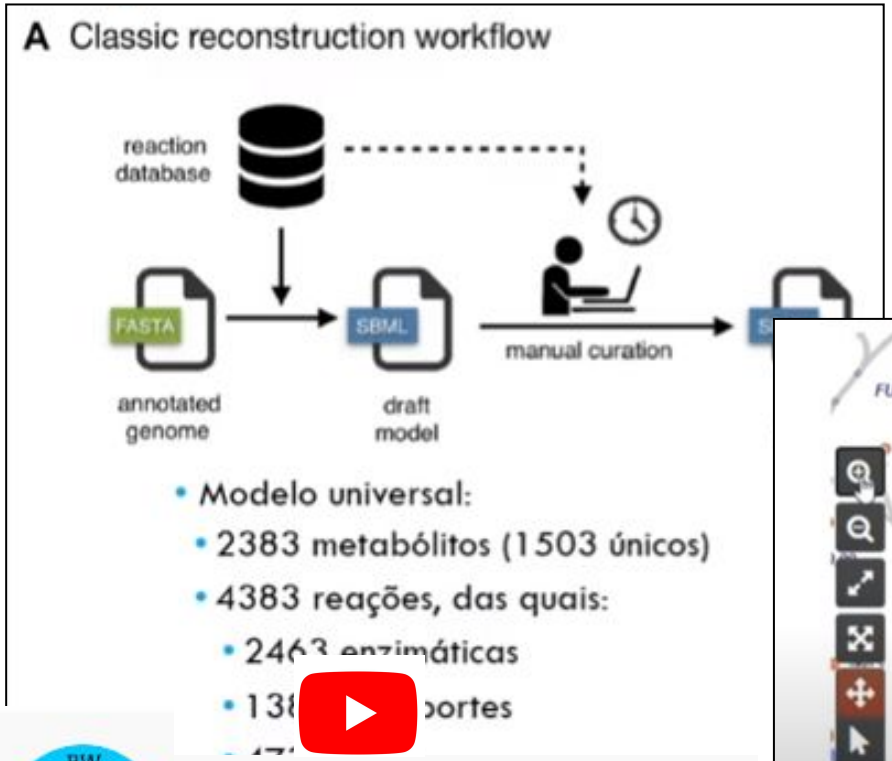
https://zhanglab.ccmb.med.umich.edu/I-TASSER/

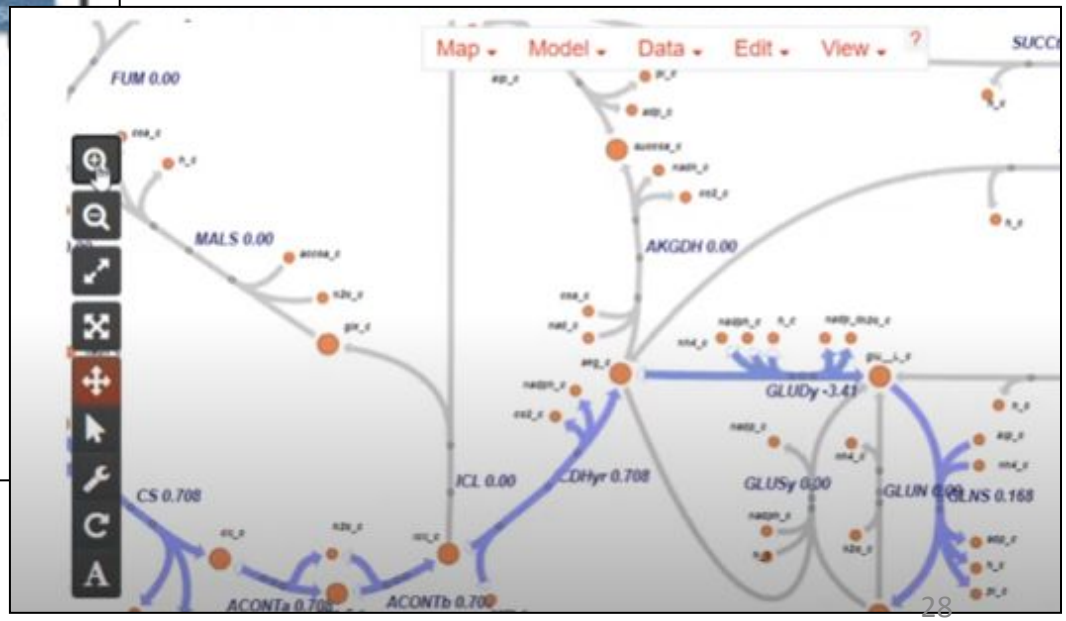
Bases de dados reconstruções metabólicas

Metabolismo da Cisteína: KEGG (*T. cruzi*)



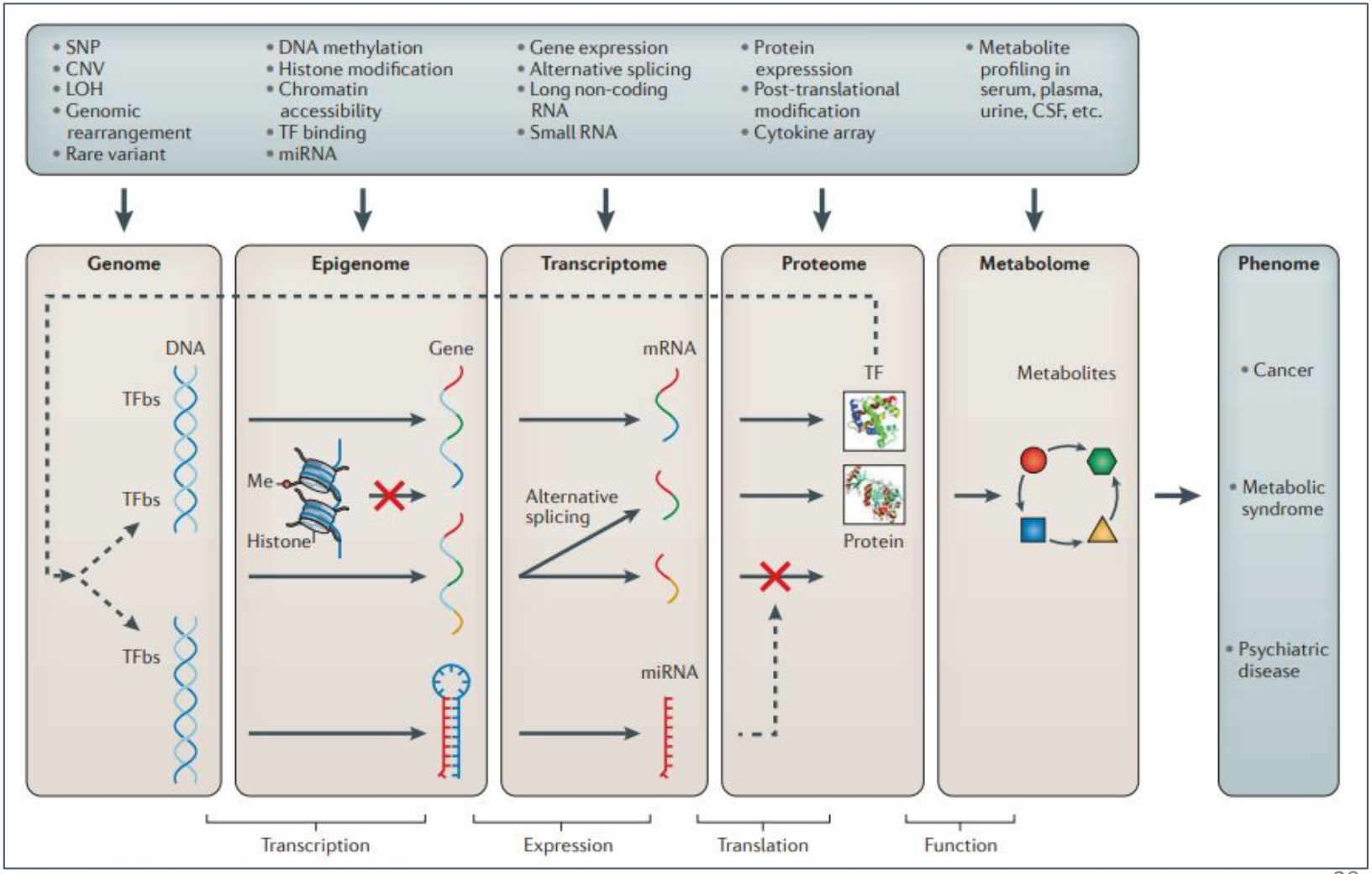
Bases de dados + evidencia experimental
→ Construção de modelos metabólicos





BioWinformática Tutoriais

Ômicas



Ritchie, M. D. et al. *Nature Reviews Genetics, 16(2), 85–97*

VEuPathDB
Release 52
20 May 2021
Eukaryotic Pathogen, Vector & Host Informatics Resources

BRC
NIAID Bioinformatics Resource Centers

AmoebaDB CryptoDB FungiDB GiardiaDB MicrosporidiaDB

PiroplasmaDB PlasmoDB ToxoDB TrichDB TriTrypDB OrthoMCL

Exemplos para pesquisa em bases de dados:

Nome	Abrev.	NCBI (GenBank)	Uniprot	PDB	EC	TriTrypsDB (TcCLB)	TriTrypsDB (Dm28)
Histidine Ammonia Lyase	HAL	ESS65584	V5BD20 V5BD20_TRYCR	6V6H	4.3.1.3	TcCLB.506247.220	BCY84_07468
Serine AcetylTransferase	SAT	RNC54559	Q4DJX5 Q4DJX5_TRYCC	4H7O*	2.3.1.30	TcCLB.510879.80	BCY84_07439
Enoyl-CoA Hydratase	ECH	ESS70836	Q4E679 Q4E679_TRYCC	5KJP*	4.2.1.17	TcCLB.508153.130	C4B63_20g212
Cysteine Synthase	CS	XP_805193	Q4CST7 Q4CST7_TRYCC	4AEC*	2.5.1.47	TcCLB.507165.50	TCDM_02303
Isovaleryl-CoA dehydrogenase	IVDH	PBJ72985	Q4DQC2 Q4DQC2_TRYCC	4O5M*	1.3.8.4	TcCLB.506629.220	TCDM_08009

Vamos usar/provar as ferramentas para responder as dúvidas!!!