Sequence analysis

DNA, RNA & proteins

ILLINOIS INSTITUTE OF TECHNOLOGY

Jean-François Pombert, Ph.D.
Office PS 296 (Lab PS 340)

Sequence analysis – Topics

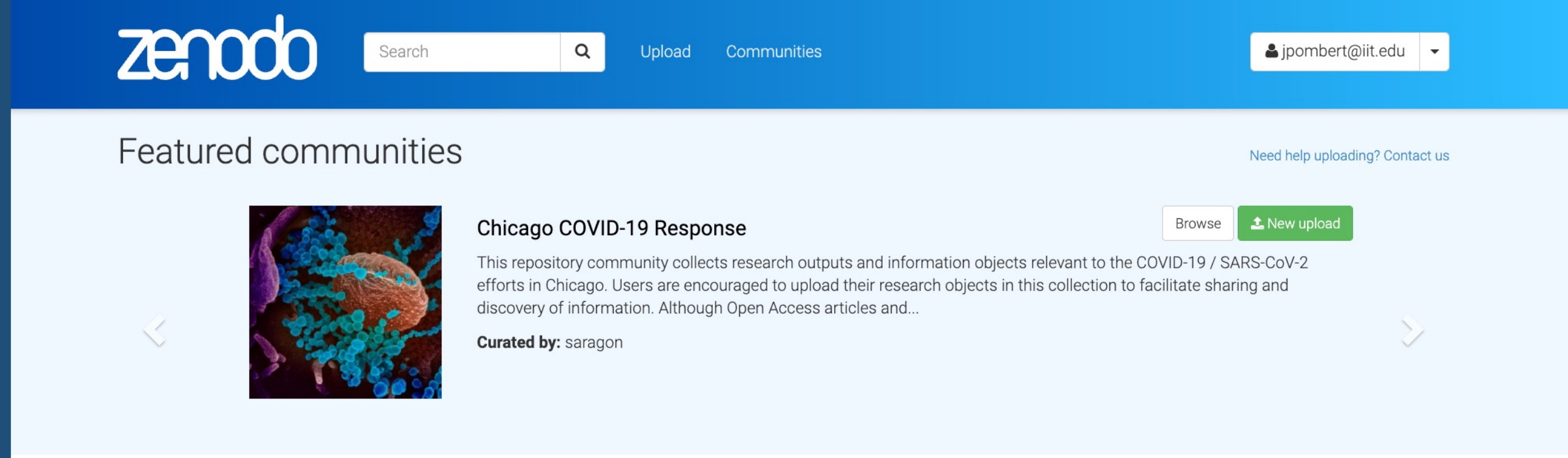
- Sequence alignment
- Homology search
- Types of sequencing data
- Read mapping + variant calling
- Genome assembly
- Genome annotation

Course section:

05 – Sequence analysis

06 – Sequencing data

07 – Genome Annotation



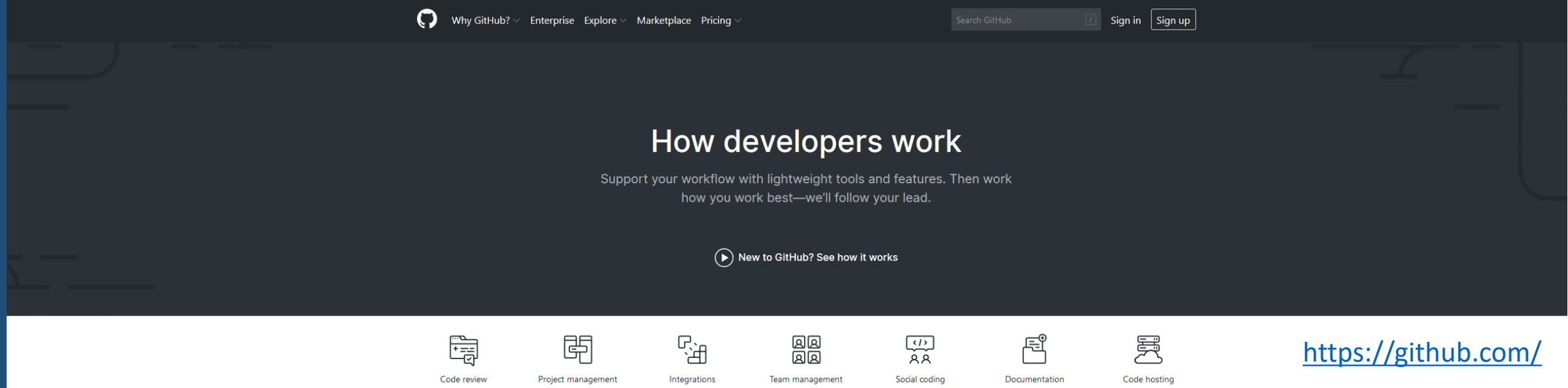
https://zenodo.org/

Open source data

Data accessibility is becoming essential for publication in high impact journals

NCBI databases - https://www.ncbi.nlm.nih.gov/ ## National Institutes of Health (NIH)

Zenodo - https://zenodo.org/ ## European Organization for Nuclear Research (CERN)



https://github.com/

Open source software (OSS)

The standard in scientific publications

Often required for publication in high impact journals

Many, many programs available

Free, various licenses (MIT, GNU…)

Source-code available <-> can be modified

EMBOSS Programs

https://www.ebi.ac.uk/Tools/emboss/

EMBOSS: the European Molecular Biology Open Software Suite
Rice P, Longden I, Bleasby A
Trends Genet. 2000. 16(6):276-7
DOI: 10.1016/s0168-9525(00)02024-2

http://emboss.sourceforge.net/

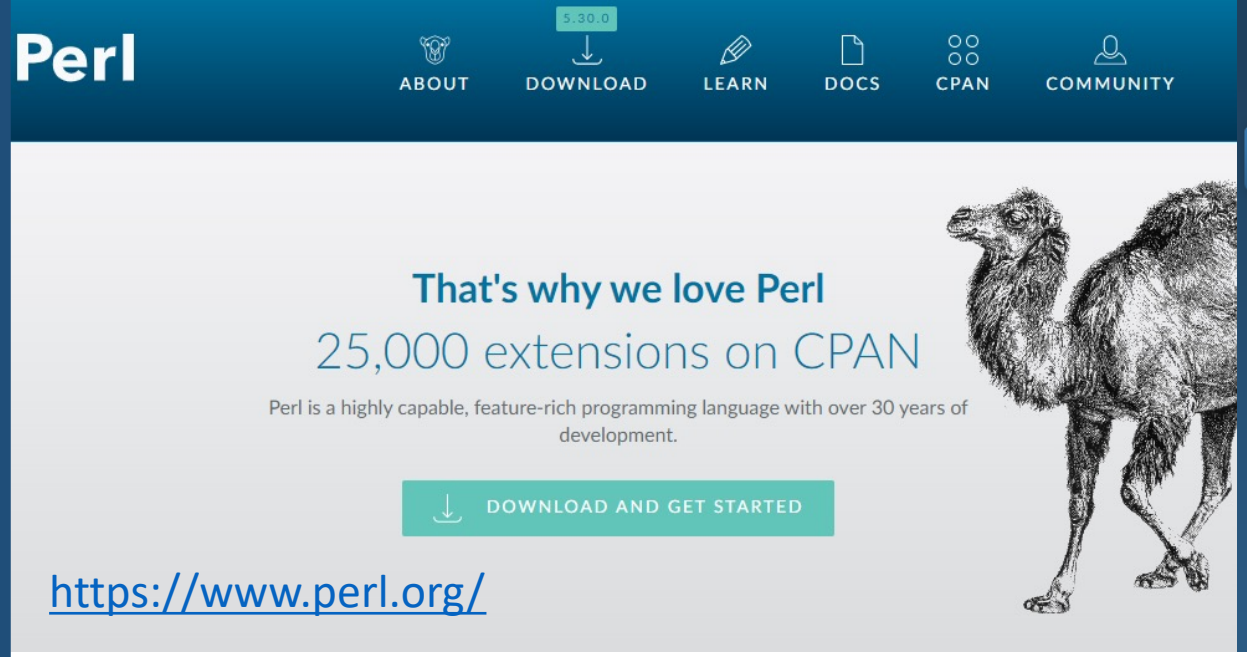
EMBOSS

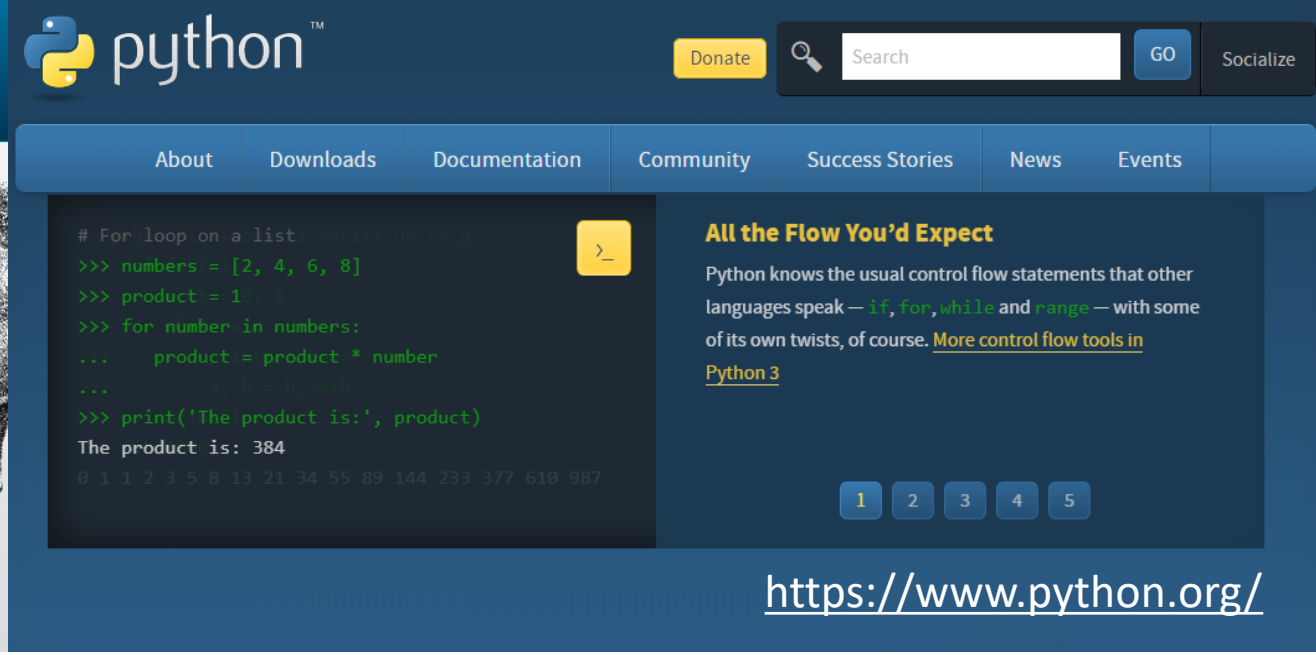
European **M**olecular **B**iology **O**pen **S**oftware **S**uite

Free ## Other software packages in the 1990s were often very expensive

On Redhat/Fedora: dnf install EMBOSS

On Debian/Ubuntu: apt-get install emboss





Perl & Python

Scripting languages

Bio-specific libraries are available

Helpful communities

Many tools are available on GitHub

Design/adapt software to your specific needs

BioPerl, BioPython

IUPAC – Standard codes

International **U**nion of **P**ure and **A**pplied **C**hemistry

https://iupac.org/

http://www.bioinformatics.org/sms2/iupac.html

About understanding algorithms

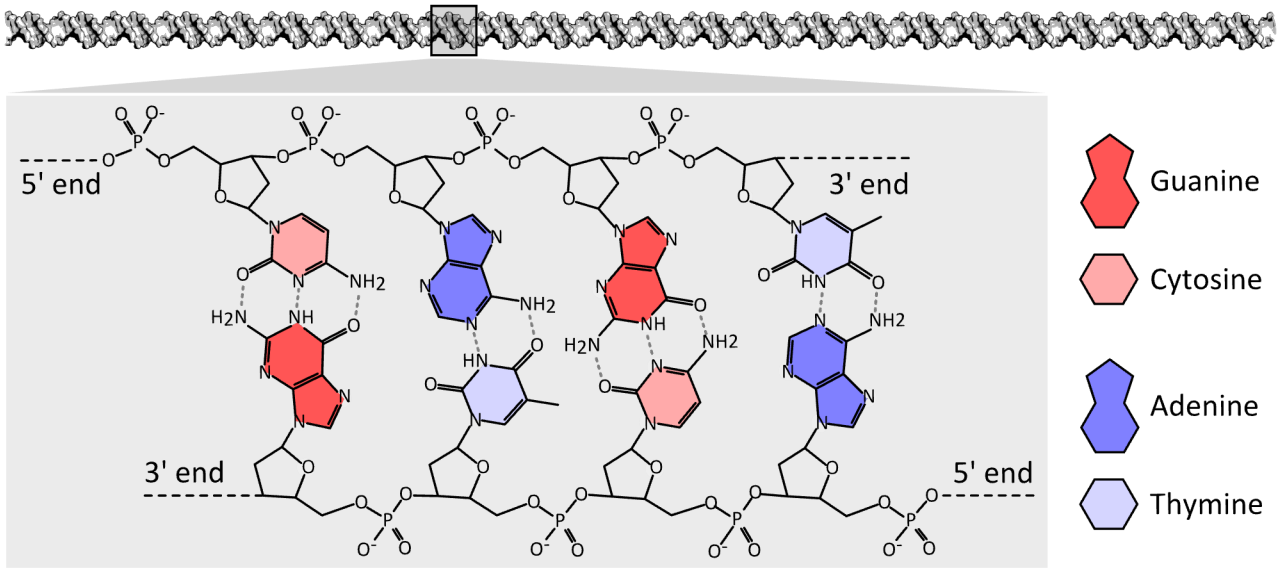
This is a hands-on class

We'll focus on the **general ideas**

Don't worry about the math

Alignments & homology

Chicken or the egg?



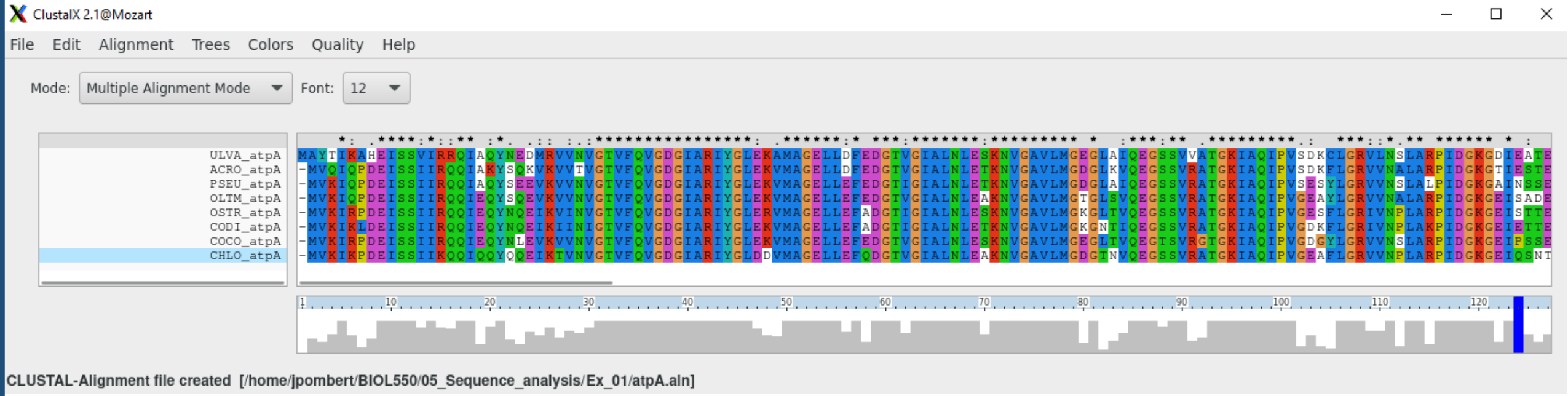
WARNING! – Orientation

Sequences can be reversed and/or complemented

5’ – 3’; common ## Most sequencers work from 5’ to 3’ (direction of synthesis)

3’ – 5’; very rare ## Maxam-Gilbert’s degradation sequencing

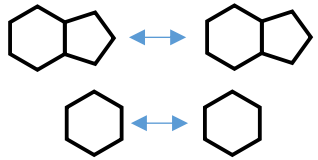
Wrong strand? ## Common in databases, especially for RNA-encoding genes



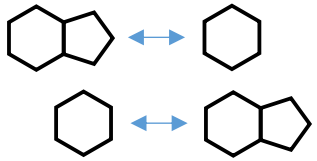
Identity *vs.* homology

DNA/RNA – transitions, transversions

Transitions



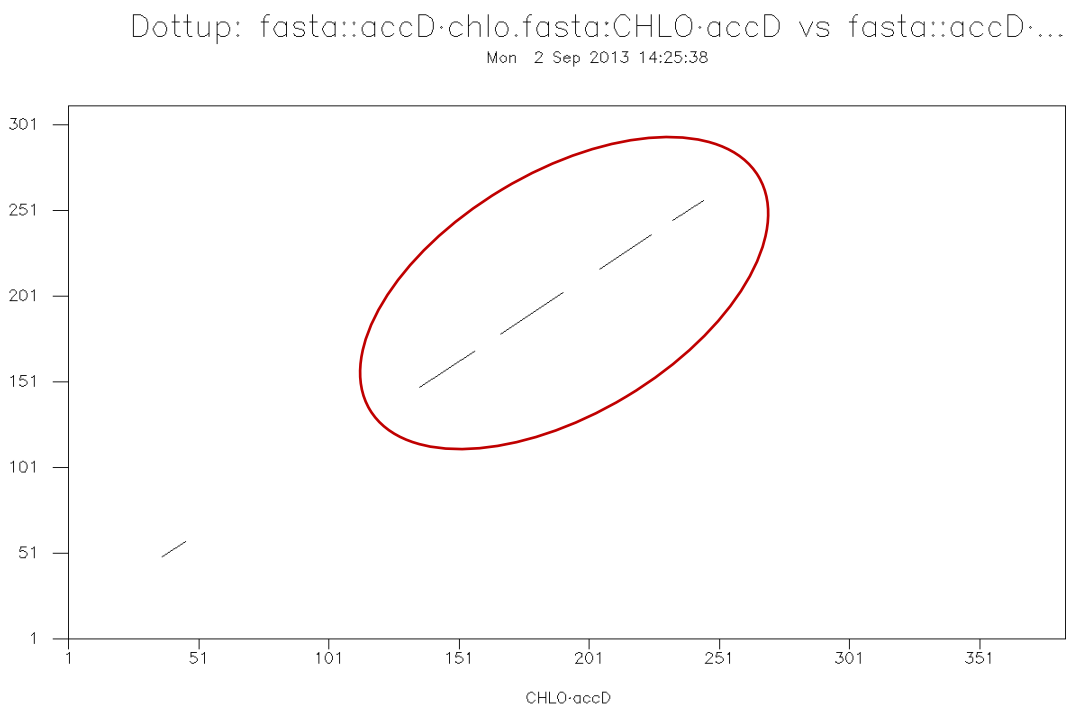
Transversions



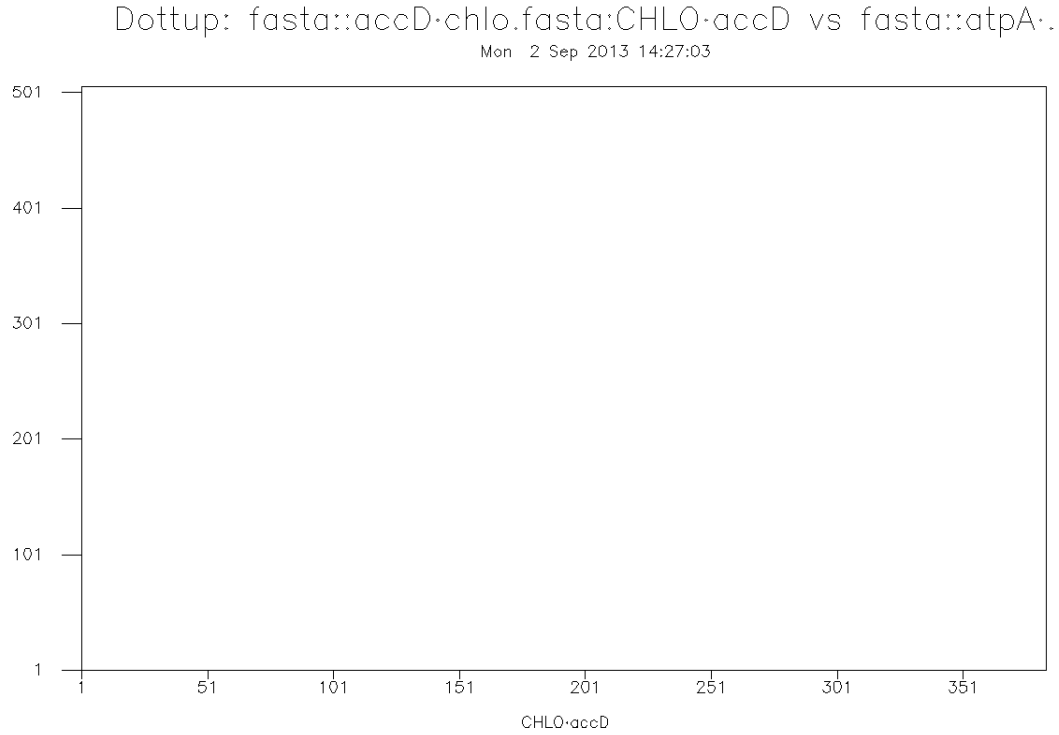
Proteins – amino acid (aa) properties:

Aromatic, aliphatic, hydrophobic, polar, acidic, basic? ## aa property chart from Sigma-Aldrich

Identity dot plots – Easy to visualize

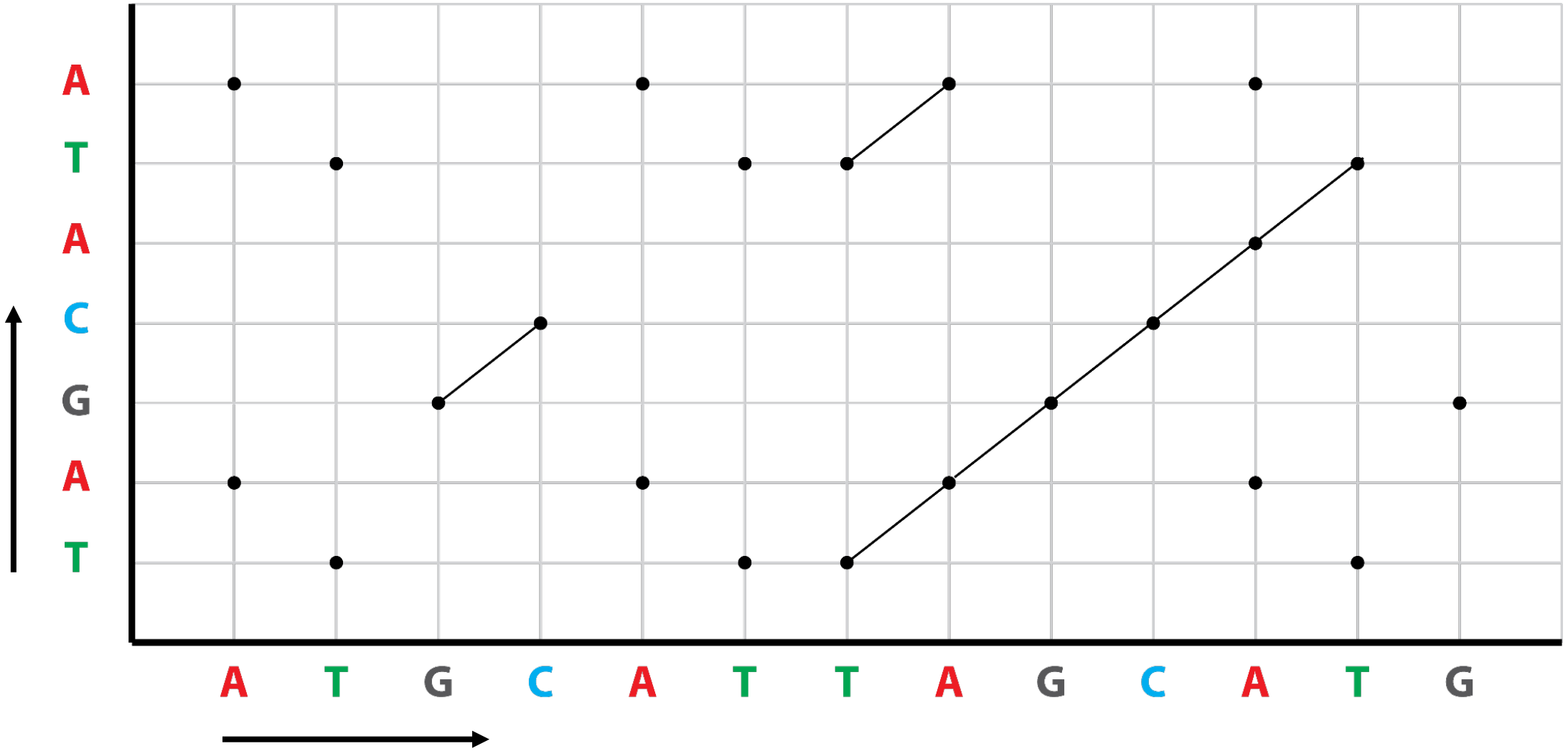


Decent similarity

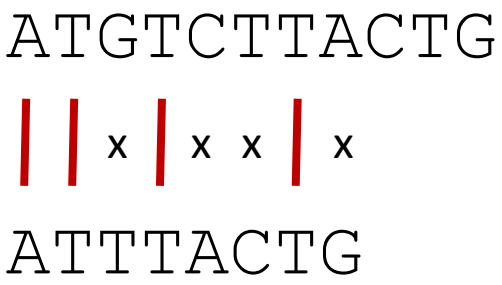


No similarity (obviously)

Dot plots – How do they work?



Gaps – Phased identity/homology increase

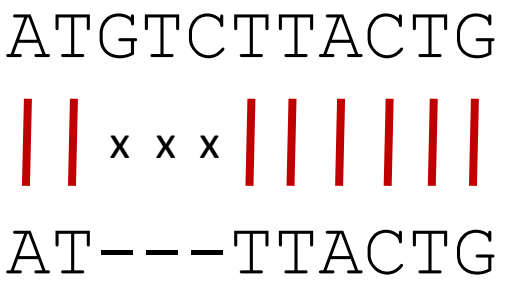


Not very similar

4 identities

4 mismatches

0 gap



Better

8 identities

0 mismatch

3 gaps (codon?)

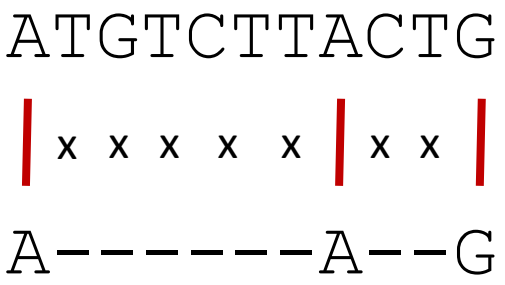
Gaps – Maximise homology, minimize holes

Better

2 identities

1 mismatch

0 gap



Illogical

3 identities

0 mismatch

7 gaps

We need scoring matrices

Why?

So that we can attribute a score to alignments

Algorithms will return the highest scoring alignments

A simple identity matrix

Points are awarded for a match

Mismatches get nothing (or penalties)

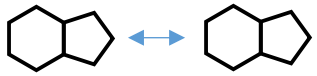
Values are arbitrary and can be modified

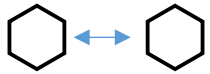
	A	T	G	C
A	1	0	0	0
T		1	0	0
G			1	0
C				1

A transition/trans-version matrix

Transitions are more likely to occur than transversions

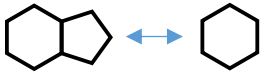
Transitions get lower penalties

Purine to purine (A <-> G) 

Pyrimidine to pyrimidine (T <-> C) 

Transversions get higher penalties

Purine <-> pyrimidine (A or G <-> C or T)



Values are arbitrary and can be modified

	A	T	G	C
A	**1**	**-1**	**0**	**-1**
T		**1**	**-1**	**0**
G			**1**	**-1**
C				**1**

Amino acids tables are more complex

BLOSUM62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	J	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	-1	-1	-4
R		5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	-2	0	-1	-4
N			6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	4	-3	0	-1	-4
D				6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	-3	1	-1	-4
C					9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-1	-3	-1	-4
Q						5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	-2	4	-1	-4
E							5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	-3	4	-1	-4
G								6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-4	-2	-1	-4
H									8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	-3	0	-1	-4
I										4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	3	-3	-1	-4
L											4	-2	2	0	-3	-2	-1	-2	-1	1	-4	3	-3	-1	-4
K												5	-1	-3	-1	0	-1	-3	-2	-2	0	-3	1	-1	-4
M													5	0	-2	-1	-1	-1	-1	1	-3	2	-1	-1	-4
F														6	-4	-2	-2	1	3	-1	-3	0	-3	-1	-4
P															7	-1	-1	-4	-3	-2	-2	-3	-1	-1	-4
S																4	1	-3	-2	-2	0	-2	0	-1	-4
T																	5	-2	-2	0	-1	-1	-1	-1	-4
W																		11	2	-3	-4	-2	-2	-1	-4
Y																			7	-1	-3	-1	-2	-1	-4
V																				4	-3	2	-2	-1	-4
B																					4	-3	0	-1	-4
J																						3	-3	-1	-4
z																							4	-1	-4
X																								-1	-4
*																									1

Extra non-IUPAC groups

1992. PNAS. Henikoff & Henikoff. 89:10915-10919

2004. Nature Biotechnology. Eddy. 22(8):1035-1036 https://www.nature.com/articles/nbt0804-1035

B Asx	asparagi(ne)c acid	L Lxe	(iso)leucine
Z Glx	glutami(ne)c acid	X Xaa	unknown

Properties of Common Amino Acids

Name	Abbr.		Molecular Weight	Molecular Formula	Residue Formula	Residue Weight (-H_2O)	pKa[1]	pKb[2]	pKx[3]	pI[4]
Alanine	Ala	A	89.10	$C_3H_7NO_2$	C_3H_5NO	71.08	2.34	9.69	–	6.00
Arginine	Arg	R	174.20	$C_6H_{14}N_4O_2$	$C_6H_{12}N_4O$	156.19	2.17	9.04	12.48	10.76
Asparagine	Asn	N	132.12	$C_4H_8N_2O_3$	$C_4H_6N_2O_2$	114.11	2.02	8.80	–	5.41
Aspartic acid	Asp	D	133.11	$C_4H_7NO_4$	$C_4H_5NO_3$	115.09	1.88	9.60	3.65	2.77
Cysteine	Cys	C	121.16	$C_3H_7NO_2S$	C_3H_5NOS	103.15	1.96	10.28	8.18	5.07
Glutamic acid	Glu	E	147.13	$C_5H_9NO_4$	$C_5H_7NO_3$	129.12	2.19	9.67	4.25	3.22
Glutamine	Gln	Q	146.15	$C_5H_{10}N_2O_3$	$C_5H_8N_2O_2$	128.13	2.17	9.13	–	5.65
Glycine	Gly	G	75.07	$C_2H_5NO_2$	C_2H_3NO	57.05	2.34	9.60	–	5.97
Histidine	His	H	155.16	$C_6H_9N_3O_2$	$C_6H_7N_3O$	137.14	1.82	9.17	6.00	7.59
Hydroxyproline	Hyp	O	131.13	$C_5H_9NO_3$	$C_5H_7NO_2$	113.11	1.82	9.65	–	–
Isoleucine	Ile	I	131.18	$C_6H_{13}NO_2$	$C_6H_{11}NO$	113.16	2.36	9.60	–	6.02
Leucine	Leu	L	131.18	$C_6H_{13}NO_2$	$C_6H_{11}NO$	113.16	2.36	9.60	–	5.98
Lysine	Lys	K	146.19	$C_6H_{14}N_2O_2$	$C_6H_{12}N_2O$	128.18	2.18	8.95	10.53	9.74
Methionine	Met	M	149.21	$C_5H_{11}NO_2S$	C_5H_9NOS	131.20	2.28	9.21	–	5.74
Phenylalanine	Phe	F	165.19	$C_9H_{11}NO_2$	C_9H_9NO	147.18	1.83	9.13	–	5.48
Proline	Pro	P	115.13	$C_5H_9NO_2$	C_5H_7NO	97.12	1.99	10.60	–	6.30
Pyroglutamatic	Glp	U	139.11	$C_5H_7NO_3$	$C_5H_5NO_2$	121.09	–	–	–	5.68
Serine	Ser	S	105.09	$C_3H_7NO_3$	$C_3H_5NO_2$	87.08	2.21	9.15	–	5.68
Threonine	Thr	T	119.12	$C_4H_9NO_3$	$C_4H_7NO_2$	101.11	2.09	9.10	–	5.60
Tryptophan	Trp	W	204.23	$C_{11}H_{12}N_2O_2$	$C_{11}H_{10}N_2O$	186.22	2.83	9.39	–	5.89
Tyrosine	Tyr	Y	181.19	$C_9H_{11}NO_3$	$C_9H_9NO_2$	163.18	2.20	9.11	10.07	5.66
Valine	Val	V	117.15	$C_5H_{11}NO_2$	C_5H_9NO	99.13	2.32	9.62	–	5.96

[1] pKa is the negative of the logarithm of the dissociation constant for the -COOH group.
[2] pKb is the negative of the logarithm of the dissociation constant for the -NH_2 group.
[3] pKx is the negative of the logarithm of the dissociation constant for any other group in the molecule.
[4] pI is the pH at the isoelectric point.
Reference: D.R. Lide, *Handbook of Chemistry and Physics, 72nd Edition*, CRC Press, Boca Raton, FL, 1991.

Amino Acids Reference Charts

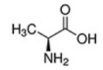
Hydrophobic - aliphatic
Hydrophobic - aromatic
Neutral - polar side chains
Acidic
Basic
Unique
Properties Table
Hydrophobicity Index
Technical Articles

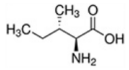
Amino acids are the compounds or building blocks that make up peptides and proteins. Each amino acid is structured from an amino group and a carboxyl group bound to a tetrahedral carbon. This carbon is designated as the α-carbon (alpha-carbon).

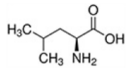
Amino acids differ from each other with respect to their side chains, which are referred to as R groups. The R group for each of the amino acids will differ in structure, electrical charge, and polarity.

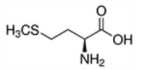
Refer to the charts and structures below to explore amino acid properties, types, applications, and availability.

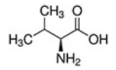
Amino Acids with Hydrophobic Side Chain – Aliphatic


Alanine, Ala, A

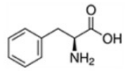

Isoleucine, Ile, I

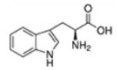

Leucine, Leu, L

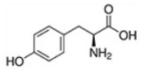

Methionine, Met, M


Valine, Val, V

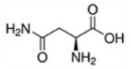
Amino Acids with Hydrophobic Side Chain – Aromatic

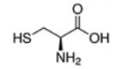

Phenylalanine, Phe, F

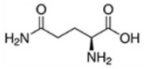

Tryptophan, Trp, W

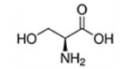

Tyrosine, Tyr, Y

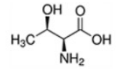
Amino Acids with Polar Neutral Side Chains


Asparagine, Asn, N

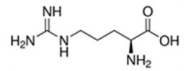

Cysteine, Cys, C

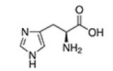

Glutamine, Gln, Q

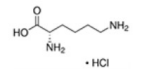




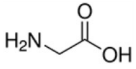
Amino Acids with Electrically Charged Side Chains – Basic

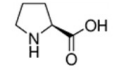

Arginine, Arg, R


Histidine, His, H


Lysine, Lys, K

Unique Amino Acids


Glycine, Gly, G


Proline, Pro, P

Hydrophobicity Index for Common Amino Acids

The hydrophobicity index is a measure of the relative hydrophobicity, or how soluble an amino acid is in water. In a protein, hydrophobic amino acids are likely to be found in the interior, whereas hydrophilic amino acids are likely to be in contact with the aqueous environment.

The values in the table below are normalized so that the most hydrophobic residue is given a value of 100 relative to glycine, which is considered neutral (0 value). The scales were extrapolated to residues which are more hydrophilic than glycine.

At pH 2[A]		At pH 7[B]	
Very Hydrophobic			
Leu	100	Phe	100
Ile	100	Ile	99
Phe	92	Trp	97
Trp	84	Leu	97
Val	79	Val	76
Met	74	Met	74
Hydrophobic			
Cys	52	Tyr	63
Tyr	49	Cys	49
Ala	47	Ala	41
Neutral			
Thr	13	Thr	13
Glu	8	His	8
Gly	0	Gly	0
Ser	-7	Ser	-5
Gln	-18	Gln	-10
Asp	-18		
Hydrophilic			
Arg	-26	Arg	-14
Lys	-37	Lys	-23
Asn	-41	Asn	-28
His	-42	Glu	-31
Pro	-46	Pro	-46 (used pH 2)
		Asp	-55

[A]pH 2 values: Normalized from Sereda et al., *J. Chrom.* 676: 139-153 (1994).
[B]pH 7 values: Monera *et al.*, *J. Protein Sci.* 1: 319-329 (1995).

https://www.sigmaaldrich.com/life-science/metabolomics/learning-center/amino-acid-reference-chart.html

Amino acid properties

Big or small?

Hydrophilic or hydrophobic? (*i.e.* polar or non-polar)

Acidic or basic?

Aromatic or aliphatic? (Benzene rings)

Gap penalties – In general

Opening	high
Extension	medium to high
Codons?	Lower (in protein-coding sequences, triplets are less penalized)

Behavior can be modified to reflect biological properties

Needleman-Wunsch

match = 1 mismatch = -1 gap = -1

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

Picture from: Kamil Slowikowski
https://gist.github.com/slowkow/508393#file-needlemanwunsch-png

Multiple Sequence Alignment Using a Genetic Algorithm and GLOCSA
Arenas-Diaz ED, Ochoterena H, Rodriguez-Vazquez K. JAEA. 2009.
https://doi.org/10.1155/2009/963150

TABLE 1: Number of possible alignments for two sequences; **m** and **n** are the respective sizes of two given sequences.

m, *n*	No. of possible alignments
1,1	3
2,2	13
3,3	63
4,4	321
5,5	1683
6,6	8989
7,7	48639
8,8	265729
9,9	1462563
10,10	8097453

Needleman-Wunsch – 1970

DNA sequencing wasn't invented yet!

Dynamic programming ## Using past knowledge to solve future problems

Breaks down large problems into smaller overlapping subproblems

Alignments of long sequences would be near intractable otherwise

Struggles with very long sequences ## We can't align genomes with NW

Smith–Waterman algorithm (1981) differ in scoring scheme: **local** (SW) vs. **global** (NW)

Local *vs.* global alignments

- Most algorithms are fine-tuned for local alignments
- Local optima issues
- Global alignments sometimes make no sense ## motifs and domains

Example
LAGLIDADG
endonuclease

```
Query  12   LSYLAGFLDGDGCINAQIVRRSDYKLKFQIRVSITFFQKTNRHWFLIWLDKKLDCGTL-R  70
            L Y+AGF DG+G +  + VR   Y+  +++   + F QK      L  + + L  G L R
Sbjct  13   LDYIAGFFDGEGSVVVRFVRDGRYRAGYRVSTKVVFVQKERD--VLEEIHETLGMGHLYR  70

Query  71   KRPDGMSEYAIIGIASVRNLLSILKPYLKLKK---RQAILLLKIIEKMPHIQNDP  122
            +  DG+    I     +R  + ++     +K+    +   +L+++E   H   D
Sbjct  71   RGSDGVWYLEIYRREDLREFVELIGNRTMVKRDALERLATVLELLEGGVHGSRDG  125
```

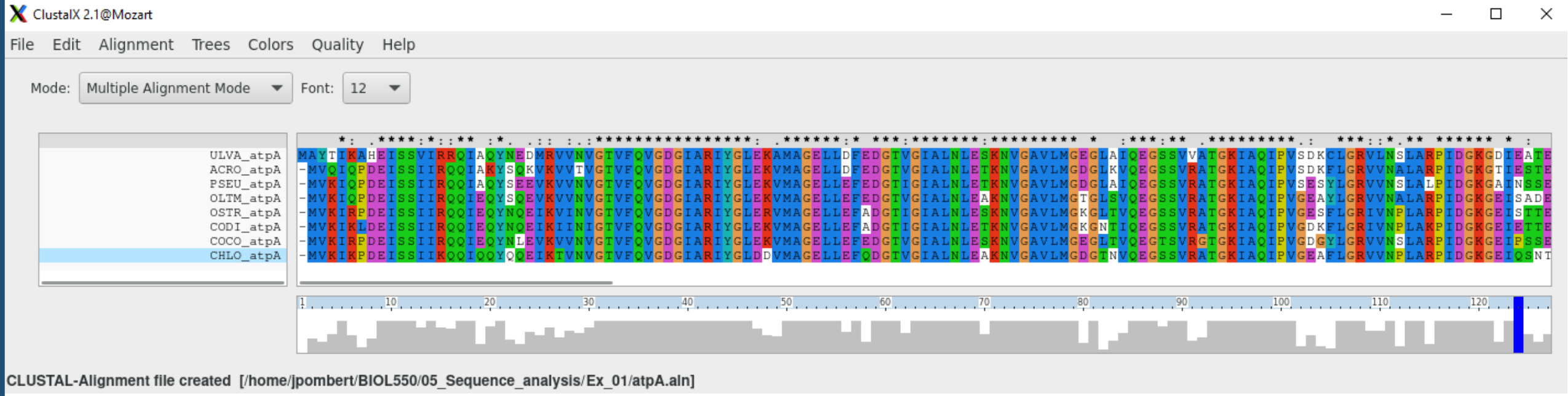
Locally? Good alignment of a known biological motif
Globally? Pretty bad alignment

Left -> Right ne Right -> Left

Alignments can differ if started from the **right** rather than the **left**!

An algorithm issue

Not a big problem with highly conserved sequences



MSAs – Multiple sequences align.

The most common scenario

MSA programs can be used for 2 or more sequences

Speed is key ## heuristics or bust

Common MSA tools

Some of the early ones; ClustalX is very visual and thus a good introduction

ClustalW, X, O	http://www.clustal.org/
T-COFFEE	http://www.tcoffee.org/

The recommended ones

MAFFT	http://mafft.cbrc.jp/alignment/software/
MUSCLE	http://www.drive5.com/muscle/

A special tool capable of aligning amino acid sequences from pseudogenes

MACSE	http://mbb.univ-montp2.fr/macse

Clustal W and Clustal X version 2.0
Larkin MA *et al.*
Bioinformatics. 2007. 23:2947-2948
DOI: 10.1093/bioinformatics/btm404

http://www.clustal.org/clustal2/

Exercise 1 – ClustalX

1) Connect to Mozart with -X enabled (ssh -X ID@216.47.151.148)
   ## -X is required to display the graphical interface (or -Y with MacOS/Xquartz)
2) Type /opt/clustal/clustalx (Note: can be slow depending on bandwidth)
3) Open the file atpA.fasta (File -> Open or ctrl+o)
4) Look at the options in the menu
5) Do a complete alignment (ctrl+l)
6) Noticed the improvements?

Exercise 2 – Alignment formats

Using the same alignment from exercise 1

1) File -> Save Sequences As (ctrl+s)
2) Save as CLUSTAL, NEXUS and PHYLIP formats
3) Look at the differences with less in the shell and compare them with the FASTA one
4) All three are interleaved formats, FASTA is sequential. Note that PHYLIP can also be written in a sequential format

Parallelization of MAFFT for large-scale multiple sequence alignments
Nakamura *T et al.*
Bioinformatics. 2018. 34(14):2490-2492
DOI: 10.1093/bioinformatics/bty121

https://mafft.cbrc.jp/alignment/software/

Exercise 3 – MAFFT

1) Type mafft -h, look at the options
2) Perform a mafft alignment of atpB.fasta
3) Look at the mafft output
4) Write a Perl script that will run mafft on all fasta files using 2 threads

MACSE v2: Toolkit for the Alignment of Coding Sequences Accounting for Frameshifts and Stop Codons

Ranwez V *et al.*
Molecular Biology and Evolution. 2018. 35(10):2582–2584
DOI: 10.1093/molbev/msy159

https://bioweb.supagro.inra.fr/macse/#

Exercise 4 – MACSE: codon based alignments

1) MACSE is a java program. To launch type:
   java -jar /opt/MACSE/macse_v2.05.jar ## New version with GUI
2) Look at the CMD line options with -help ## Not very helpful right?
3) Re-launch with the appropriate parameters for the standard genetic code: -gc_def 1. Use EC_0360.fasta as input.
4) Look at the outputs. AA => amino acids; NT => nucleotides. When present, putative frameshifts are indicated by !.

MACSE takes longer to compute but returns both amino acid and DNA alignments.
MACSE can align pseudogenes at the amino acid level <- Powerful tool

For a Perl script, see:
https://github.com/PombertLab/Publication_scripts/blob/master/2020_GBE_Hamilto_Ordospora_nonadaptive_processes/dNdS/run_macse.pl

Homology search

Sequence homology

Primary sequences	-	BLAST (alignment-based); MASH (alignment-free)
Domains and motifs	-	Pfam, CDD

Structural homology

3D structures

Sequence homology – primary sequences

BLAST

Basic Local Alignment Search Tool

Fast queries against databases (online/local)

http://blast.ncbi.nlm.nih.gov/Blast.cgi (searches against GenBank)

DIAMOND

Recent (2015); 500x to 20,000x faster than BLAST

Basic Local Alignment Search Tool
Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ
J. Mol. Biol. 1990. 215:403–410

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs
Altschul SF, *et al.*
Nucl. Acids Res. 1997. 25(17):3389–3402.

Fast and sensitive protein alignment using DIAMOND
Buchfink B, Xie C, Huson DH
Nature Methods. 2015. 12:59-60

BLAST: improvements for better sequence analysis
Ye J, McGinnis S, Madden TL
Nucleic Acids Res. 2006. 34:W6–W9

J. Mol. Biol. (1990) 215, 403–410

Basic Local Alignment Search Tool

Stephen F. Altschul[1], Warren Gish[1], Webb Miller[2]
Eugene W. Myers[3] and David J. Lipman[1]

[1]*National Center for Biotechnology Information*
National Library of Medicine, National Institutes of Health
Bethesda, MD 20894, U.S.A.

[2]*Department of Computer Science*
The Pennsylvania State University, University Park, PA 16802, U.S.A.

[3]*Department of Computer Science*
University of Arizona, Tucson, AZ 85721, U.S.A.

(Received 26 February 1990; accepted 15 May 1990)

A new approach to rapid sequence comparison, basic local alignment search tool (BLAST), directly approximates alignments that optimize a measure of local similarity, the maximal segment pair (MSP) score. Recent mathematical results on the stochastic properties of MSP scores allow an analysis of the performance of this method as well as the statistical significance of alignments it generates. The basic algorithm is simple and robust; it can be implemented in a number of ways and applied in a variety of contexts including straightforward DNA and protein sequence database searches, motif searches, gene identification searches, and in the analysis of multiple regions of similarity in long DNA sequences. In addition to its flexibility and tractability to mathematical analysis, BLAST is an order of magnitude faster than existing sequence comparison tools of comparable sensitivity.

1. Introduction

The discovery of sequence homology to a known protein or family of proteins often provides the first clues about the function of a newly sequenced gene. As the DNA and amino acid sequence databases continue to grow in size they become increasingly useful in the analysis of newly sequenced genes and proteins because of the greater chance of finding such homologies. There are a number of software tools for searching sequence databases but all use some measure of similarity between sequences to distinguish biologically significant relationships from chance similarities. Perhaps the best studied measures are those used in conjunction with variations of the dynamic programming algorithm (Needleman & Wunsch, 1970; Sellers, 1974; Sankoff & Kruskal, 1983; Waterman, 1984). These methods assign scores to insertions, deletions and replacements, and compute an alignment of two sequences that corresponds to the least costly set of such mutations. Such an alignment may be thought of as minimizing the evolutionary distance or maximizing the similarity between the two sequences compared. In either case, the cost of this alignment is a measure of similarity; the algorithm guarantees it is optimal, based on the given scores. Because of their computational requirements, dynamic programming algorithms are impractical for searching large databases without the use of a supercomputer (Gotoh & Tagashira, 1986) or other special purpose hardware (Coulson *et al.*, 1987).

Rapid heuristic algorithms that attempt to approximate the above methods have been developed (Waterman, 1984), allowing large databases to be searched on commonly available computers. In many heuristic methods the measure of similarity is not explicitly defined as a minimal cost set of mutations, but instead is implicit in the algorithm itself. For example, the FASTP program (Lipman & Pearson, 1985; Pearson & Lipman, 1988) first finds locally similar regions between two sequences based on identities but not gaps, and then rescores these regions using a measure of similarity between residues, such as a PAM matrix (Dayhoff *et al.*, 1978) which allows conservative replacements as well as identities to increment the similarity score. Despite their rather indirect approximation of minimal evolution measures, heuristic tools such as FASTP have been quite popular and have identified many distant but biologically significant relationships.

0022-2836/90/190403-08 $03.00/0

Sequence homology – motifs and domains

Not every portion of a protein has functional significance

Functional domains, or motifs, are conserved throughout evolution

Hidden Markov Models:

- Search for motifs
- Search for genes

nhmmer: DNA Homology Search With Profile HMMs
Wheeler TJ and Eddy SR
Bioinformatics. 2013. 29: 2487-2489

BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 19 2013, pages 2487–2489
doi:10.1093/bioinformatics/btt403

Sequence analysis

Advance Access publication July 9, 2013

nhmmer: DNA homology search with profile HMMs

Travis J. Wheeler* and Sean R. Eddy
HHMI Janelia Farm Research Campus, Ashburn, VA 20147, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Sequence database searches are an essential part of molecular biology, providing information about the function and evolutionary history of proteins, RNA molecules and DNA sequence elements. We present a tool for DNA/DNA sequence comparison that is built on the HMMER framework, which applies probabilistic inference methods based on hidden Markov models to the problem of homology search. This tool, called nhmmer, enables improved detection of remote DNA homologs, and has been used in combination with Dfam and RepeatMasker to improve annotation of transposable elements in the human genome.

Availability: nhmmer is a part of the new HMMER3.1 release. Source code and documentation can be downloaded from http://hmmer.org. HMMER3.1 is freely licensed under the GNU GPLv3 and should be portable to any POSIX-compliant operating system, including Linux and Mac OS/X.

Contact: wheelert@janelia.hhmi.org

Received and revised on June 17, 2013; accepted on July 5, 2013

1 INTRODUCTION

A widely used general purpose tool for DNA/DNA sequence comparison is blastn (Altschul *et al.*, 1990; Camacho *et al.*, 2009), which heuristically approximates the Smith–Waterman algorithm (Smith and Waterman, 1981) for recognizing local regions of similarity between two sequences. In recent years, most advances in DNA/DNA comparison have related to accelerating search for near-exact matches (Kent, 2002; Langmead *et al.*, 2009; Li and Durbin, 2009), and to improving whole-genome alignment (Kurtz *et al.*, 2004; Schwartz *et al.*, 2003). Another area that deserves attention is the development of methods that maximize the power of computational sequence comparison tools to detect remote homologies.

Profile hidden Markov models (profile HMMs) (Durbin *et al.*, 1998; Krogh *et al.*, 1994) represent an important advance in terms of sensitivity of sequence searches for remote homology. They provide a formal probabilistic framework for sequence comparison and improve detection of remote homologs by (i) enabling position-specific residue and gap scoring based on a query profile, and (ii) calculating the signal of homology based on the more powerful 'Forward/Backward' HMM algorithm that computes not just one best-scoring alignment, but a sum of support over all possible alignments. In the past, this improved sensitivity came at a significant computational cost, but recent advances in HMMER3 have increased speed for protein search by ~100-fold, reaching blastp-like speed through a combination of filtering heuristics (Eddy, 2008) and computer engineering (Eddy, 2011; Farrar, 2007). Tools based on profile HMMs (Eddy, 2009; Karplus *et al.*, 1998) have historically focused on protein search, with little concentration on the challenges presented by (i) chromosome-length target sequences, and (ii) the extreme composition bias often seen in genomic DNA. With attention to the details of DNA search, nhmmer builds upon the speed advances of HMMER3, bringing the power of profile HMMs to DNA homology search, at speeds nearly as fast as blastn with sensitive settings.

An example of a biological problem requiring sensitive detection of remote DNA homologs is the annotation of genomic sequence derived from ancient transposable element (TE) expansions. A prerelease version of the nhmmer tools has recently been shown to provide increased sensitivity over blastn and other single-sequence search methods, with reduced false discovery rate and reasonable runtime, in searching for TEs (Wheeler *et al.*, 2013). For example, when nhmmer was used within the recently released RepeatMasker 4.0 (Smit and Hubley, 2013), an additional 150 Mb (5%) of the human genome was reliably annotated as derived from TEs.

2 USAGE AND PERFORMANCE

Usage. The program nhmmer is used to search one or more nucleotide queries against a nucleotide sequence database. For each query, nhmmer searches the target database and outputs a ranked list of the hits with the most significant matches to the query. A query may consist of a single sequence, a multiple sequence alignment, or a profile HMM built using the HMMER program hmmbuild. Each hit represents a region of local similarity between a portion of the query and a subsequence of the full target database sequence, and is assigned a similarity score S in bits, along with an *E*-value (Eddy, 2008) indicating the expected number of false positives at a threshold of score S. Each hit is also accompanied by an alignment of the matched sequence to the model, with values indicating the confidence with which each position is aligned.

The final score, boundaries and alignment of a hit are computed based on filling in a Forward/Backward dynamic programming matrix, but the computational burden of doing this for the full target database is prohibitive. Therefore, nhmmer uses a series of acceleration filters that depend on simpler approximations of the final Forward score of a hit. These filters are based on those used in the HMMER3 protein search tools (Eddy, 2011), but have been modified to work in the context of long (potentially chromosome length) target sequences. The initial filter, called 'single segment ungapped Viterbi', scans along

*To whom correspondence should be addressed.

Structural homology – 3D structures

Sometimes the sequence itself is irrelevant

Structure is often key

If it looks like a wheel, it probably is…

Parallel CUDA-Based Protein 3D Structure Similarity Searching
Mrozek D
High-Performance Computational Solutions in Protein Bioinformatics. 2014. pp 49-79

Chapter 3
Parallel CUDA-Based Protein 3D Structure Similarity Searching

The structural alignment between two proteins: is there a unique answer?

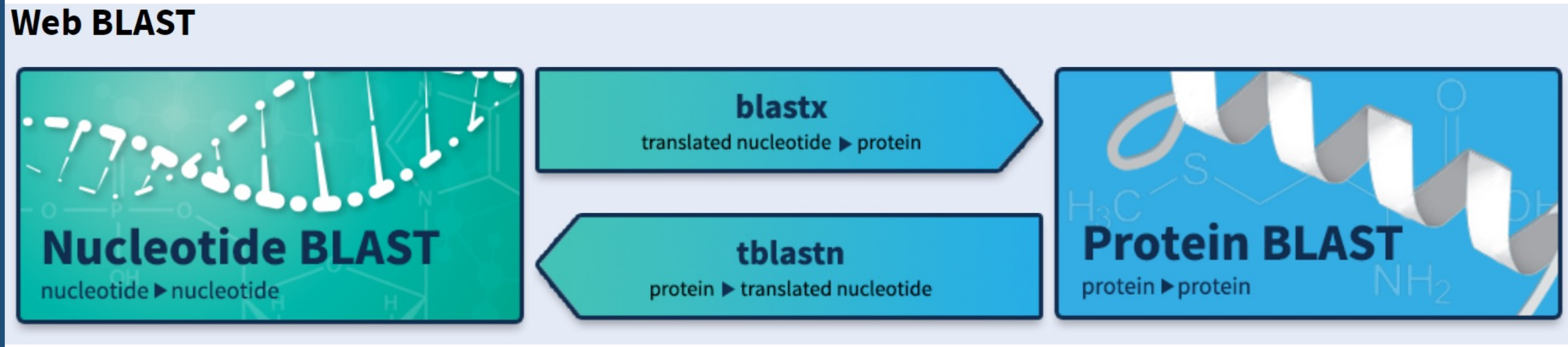
Adam Godzik, 1996

Abstract Finding common molecular substructures in complex 3D protein structures is still challenging. This is especially visible when scanning entire databases containing tens or even hundreds of thousands protein structures. Graphics processing units (GPUs) and general purpose graphics processing units (GPGPUs) promise to give a high speedup of many time-consuming and computationally demanding processes over their original implementations on CPUs. In this chapter, we will see that a massive parallelization of the 3D structure similarity searching on many core CUDA-enabled GPU devices leads to reduction of the execution time of the process and allows to perform it in real time.

Keywords Proteins · 3D protein structure · Tertiary structure · Similarity searching · Structure matching · Structure comparison · Structure alignment · Parallel computing · GPU · CUDA

3.1 Introduction

Protein 3D structure similarity searching is a process in which a given protein structure is compared to another protein structure or a set of protein structures collected in a database. The aim of the process is to find matching fragments of compared protein structures. On the basis of the similarities found during this process, scientists can draw useful conclusions about the common ancestry of the proteins, and thus the organisms (that the proteins came from), their evolutionary relationships, functional similarities, existence of common functional regions, and many other things [6]. This process is especially important in situations, where sequence similarity searches fail or deliver too few clues [15]. There are also other processes in which protein structure



NCBI BLAST

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

Basic Local Alignment Search Tool
Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ
J. Mol. Biol. 1990. 215:403–410

BLAST: improvements for better sequence analysis
Ye J, McGinnis S, Madden TL
Nucleic Acids Res. 2006. 34:W6–W9

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs
Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, & Lipman DJ
Nucl. Acids Res. 1997. 25(17):3389–3402.

Types of BLAST

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

Type	Query	Database
blastn	nucleotide	nucleotide
blastp	protein	protein
blastx	translated nucleotide	protein
tblastn	protein	translated nucleotide
tblastx	translated nucleotide	translated nucleotide

Note: To perform protein <-> nucleotide searches, the nucleotide sequences are translated to amino acids, then searches are done at the amino acid level

Sequences producing significant alignments Download Manage Columns Show 100

select all *94 sequences selected* GenPept Graphics Distance tree of results Multiple alignment

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
DNA-directed RNA polymerase [Mitosporidium daphniae]	1373	1373	99%	0.0	62.68%	XP_013238900.1
DNA-directed RNA polymerase II second largest subunit [Encephalitozoon cuniculi]	1281	1281	99%	0.0	55.63%	AGE96471.1
DNA-DIRECTED RNA POLYMERASE II SECOND LARGEST SUBUNIT [Encephalitozoon cuni	1280	1280	99%	0.0	55.63%	NP_586140.1
DNA-directed RNA polymerase subunit B [Encephalitozoon hellem ATCC 50504]	1272	1272	99%	0.0	55.54%	XP_003888100.1
subunit beta of DNA-directed RNA polymerase [Hamiltosporidium tvaerminnensis]	488	665	98%	9e-154	32.19%	TBU01157.1
subunit beta of DNA-directed RNA polymerase [Hamiltosporidium tvaerminnensis]	488	665	98%	1e-153	32.19%	TBU10981.1
DNA-directed RNA polymerase II subunit RPB2 [Vittaforma corneae ATCC 50505]	469	469	44%	1e-153	45.67%	XP_007605419.1
subunit beta of DNA-directed RNA polymerase [Hamiltosporidium magnivora]	488	665	98%	2e-153	32.19%	TBU08960.1

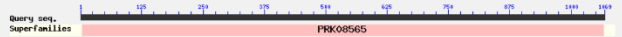
bit scores E-values

Descriptions | Graphic Summary | Alignments | Taxonomy

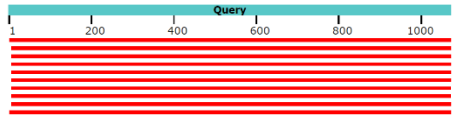
hover to see the title *click to show alignments* Show Conserved Domains

Alignment Scores < 40 | 40 - 50 | 50 - 80 | 80 - 200 | >= 200

94 sequences selected **Putative conserved domains have been detected, click on the image below for detailed results.**



Distribution of the top 116 Blast Hits on 94 subject sequences



Feedback

Alignment view Pairwise Restore defaults

94 sequences selected

Download GenPept Graphics

DNA-directed RNA polymerase [Mitosporidium daphniae]

Sequence ID: XP_013238900.1 Length: 1184 Number of Matches: 1

See 1 more title(s) See all Identical Proteins(IPG)

Range 1: 102 to 1184 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
1373 bits(3554)	0.0	Compositional matrix adjust.	687/1096(63%)	825/1096(75%)	43/1096(3%)

```
Query  4     LSRPTFVEADGETVVLFPKEARLRNLTYASPLYVDVERRVTT--RRRTPDTEEWEEETEV  61
             LS+P+  E+DG T  +FP EARLRNLTYA+PLYVD+ RR+ +     TP   +     EV
Sbjct  102   LSKPSMTESDGSTYPMFPNEARLRNLTYAAPLYVDITRRIVSGGDSTTPHGGDDVIVHEV  161

Query  62    TPYDKIFLGEVPTMLRSDYCNLHNTTDRDLTDLGECPYDQGGYFVINGSEKVLIAQERMA  121
                 K+F+G++P MLRS YC LH  +DR+L DLGECPYDQGGYF+I+GSEKVLIAQERMA
Sbjct  162   EELPKVFIGKIPIMLRSTYCILHGISDRELVDLGECPYDQGGYFIISGSEKVLIAQERMA  221

Query  122   NNHVYVFRKTAPSKWSYQAEIRSVVEGSSRRSSSMSVTMLARGGDGRRGGAGSVLRAQIP  181
              N VYVF K  PS  ++ AEIRS VE  S+ S+++++ +  RG  G  G  G V+R  +P
Sbjct  222   TNVVYVFSKAQPSPVTHIAEIRSSVEKGSKLSTTLTLKLQTRGAGGSNGEGGKVIRVAVP  281

Query  182   YVKADIPILVVFRALGFVADRDILEHIVYDSEDADMLDALRPSIEEAFPIQSQEVALDYI  241
             YVK+DIPI ++FRALGFV+DR+ILEHI +D EDA+ML AL+PSIEEA  IQ Q  ALD+I
Sbjct  282   YVKSDIPIGILFRALGFVSDREILEHICWDFEDAEMLAALKPSIEEAIVIQDQTNALDFI  341

Query  242   GRRGSVVGATKEARVRYARELLQRELLPHVSMEPHCETKKAYFFGYVVHRLLLVSLGRRE  301
             G+RG+  G  ++ R+ +AR +LQ+E+LPH+ +    ETKKAYF GY+ HRLLL +L RRE
Sbjct  342   GKRGTPPGTPQDRRISFARMILQKEMLPHMGVLEFNETKKAYFLGYMAHRLLLGALDRRE  401
```

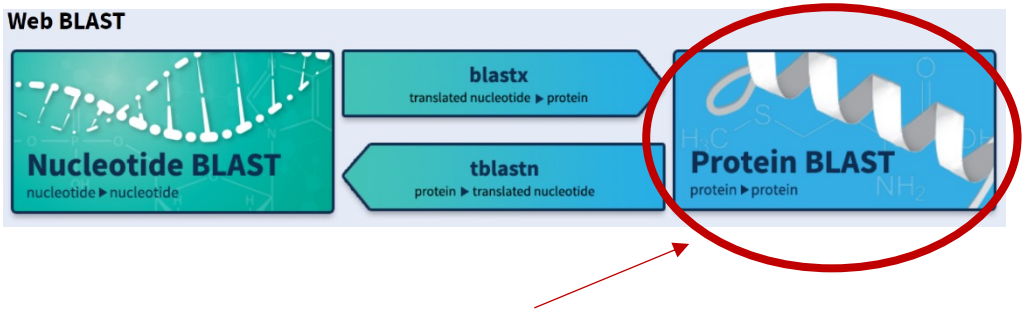
BLAST – *E*-values

BLAST analyses will return many hits

Not all of them are good (quite a few are garbage actually)

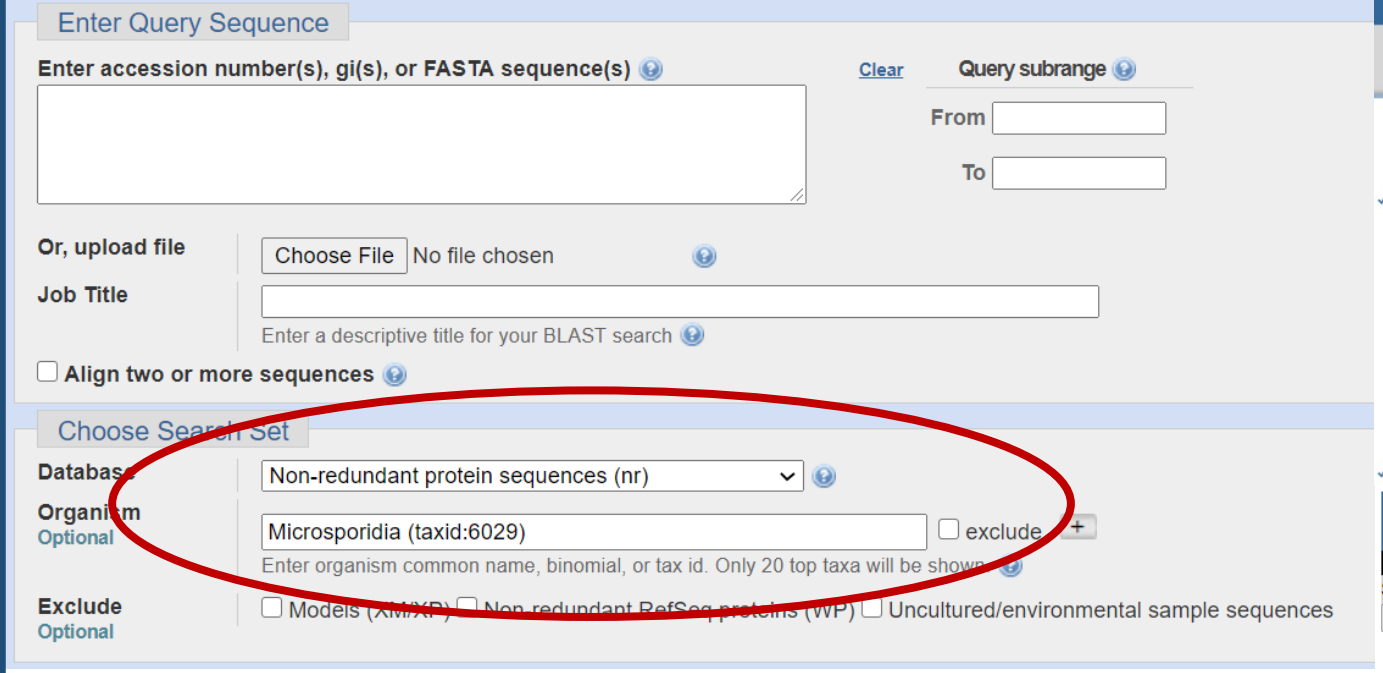
E-values are probability estimates ## Lower is better

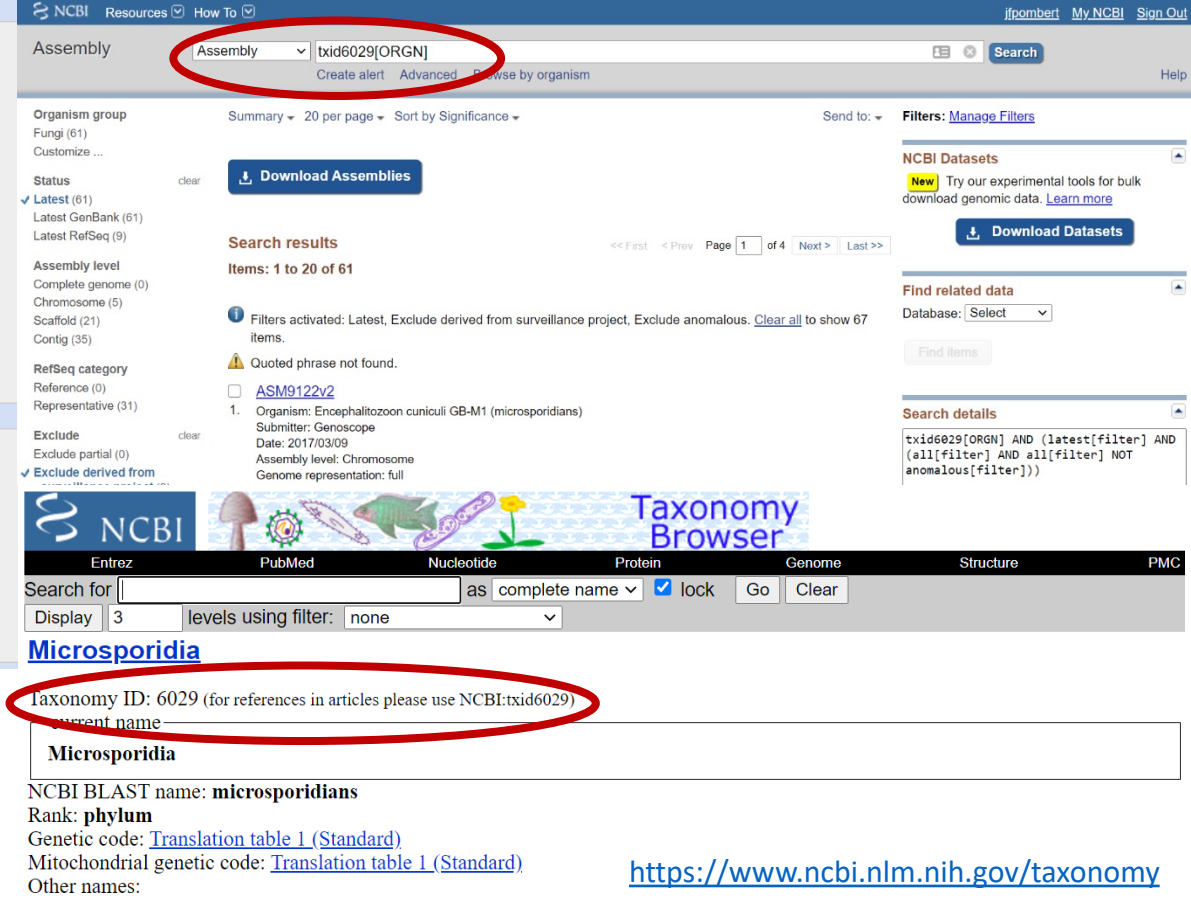
bit scores are the sum of all the aligned bits ## Higher is better



Exercise 5 – Online BLAST (BLASTP)

1) Goto http://blast.ncbi.nlm.nih.gov/Blast.cgi
2) Open BLAST_1.fasta, perform a BLASTP on it
3) Find anything good?
4) Open BLAST_2.fasta, perform a BLASTP on it
5) What about this time? Try with BLAST_3.fasta
6) Explore the various options by yourself, feel free to ask questions





https://www.ncbi.nlm.nih.gov/taxonomy

txidXXXX[ORGN]

We can restrict our searches with a taxonomic ID

The txid is a unique ID given by NCBI to an organism or group

It can be retrieved from the scrolling menu on the BLAST query page

The txid can also be found from the NCBI taxonomy database

BLAST – Command line

No web latency

Can run batch searches

Can be automated

Runs privately (you can make your own databases) ## Useful for data privacy

Makeblastdb – Private databases

```
makeblastdb \
  -in file.fasta \          ## FASTA input file (single/multifasta; DNA, RNA or protein)
  -dbtype nucl \            ## Database type: nucl (DNA/RNA) or prot (proteins)
  -out nickname \           ## Name of the database to be created
  -title 'short description'  ## Description (optional)
```

```
(t)blast(n,p,x) \                  ## BLAST program to use
  -num_threads 16 \                ## Number of threads to use
  -query input.fasta \             ## FASTA file to query
  -db database_name \              ## Database to query against
  -evalue 1e-40 \                  ## Desired minimum E-value
  -culling_limit 10 \              ## Desired number of top hits
  -outfmt 6 \                      ## Desired output format; 0 = pairwise alignments, 6 = tab-delimited
  -out test.blastp.6               ## Desired output name
```

Running local BLAST searches

Works against NCBI and/or custom databases

Great when running several queries

No CPU usage limit, unlike the web-based tool

https://www.ncbi.nlm.nih.gov/books/NBK279684/

Exercise 6 – Creating your own BLAST db

1) Concatenate all fasta files from the Alignments folder into a single file (my.fasta)
2) Create a folder BLAST and a subfolder db
3) Move (or copy) the my.fasta file into the BLAST/db folder
4) In db, create your own database called my, for title use PlastidProteins
5) Keep it, we will use it for the next exercise

Exercise 7 – Running BLAST on my db

1) Use the my database from exercise 7
2) Copy ARABthali.fasta into folder BLAST
3) From the BLAST folder, run a BLASTP with ARABthali.fasta against my with:
   2 threads, evalue of 1e-10, and outfmt 6 (tabular), save it as ARABthali.blastp
4) Browse the output with less
5) Let's parse it with Perl

Exercise 8 – A simple BLAST parsing

While simple, the script may be beyond your current abilities. If so, read the
answer and try to understand what every little piece of code does.

Using the ARABthali.blastp output from exercise 8, create a script that:

1) Takes the best hit for each protein query ## the top hit is the best for each query
   Hints: %protein = (), if (exists $protein{$query}){ do something; }
2) Assigns $query (e.g. NP_085480.1) and $gene values (e.g. CHLO_atpA)
3) Prints only the best hit to the output file *e.g.*:

   NP_085480.1 rpl16

   NP_085492.1 petB

   NP_085496.1 rbcL

taxdb.tar.gz	## Taxonomy information for the databases
nr.*tar.gz	## Non-redundant protein sequences
nt.*tar.gz	## Partially non-redundant nucleotide sequences
16SMicrobial.tar.gz	## Bacterial + Archaeal 16S rRNA sequences (BioProjects 33175 and 33117)
Representative_Genomes.*tar.gz	## Representative bacterial/archaeal genomes
human_genomic.*tar.gz	## Human RefSeq (NC_######) chromosome record
other_genomic.*tar.gz	## RefSeq chromosome records (NC_######) for non-human organisms
cdd_delta.tar.gz	## Conserved Domain Database sequences for use with standalone **deltablast**
env_nr.*tar.gz	## Protein sequences for metagenomes
env_nt.*tar.gz	## Nucleotide sequences for metagenomes
refseq_genomic.*tar.gz	## NCBI genomic reference sequences
refseq_protein.*tar.gz	## NCBI protein reference sequences
refseq_rna.*tar.gz	## NCBI Transcript reference sequences

Downloading NCBI databases

ftp://ftp.ncbi.nlm.nih.gov/blast/db/ ## See NCBI_scripts for automatic downloading

The files are compressed to save space ## Large files, NR > 28Gb and growing

Must be decompressed (tar -zxvf *.tar.gz)

```
blastdb_aliastool \
   -dbtype nucl \                    ## Database type (nucl or prot)
   -gilist file.gi \                 ## List of GIs (sequence identifiers)
   -db nr \                          ## Input database
   -out nickname \                   ## Name of the desired subset
   -title 'short description'        ## Description (optional)
```

Creating local db subsets

Creates smaller subsets of NCBI databases for to reduce computation time

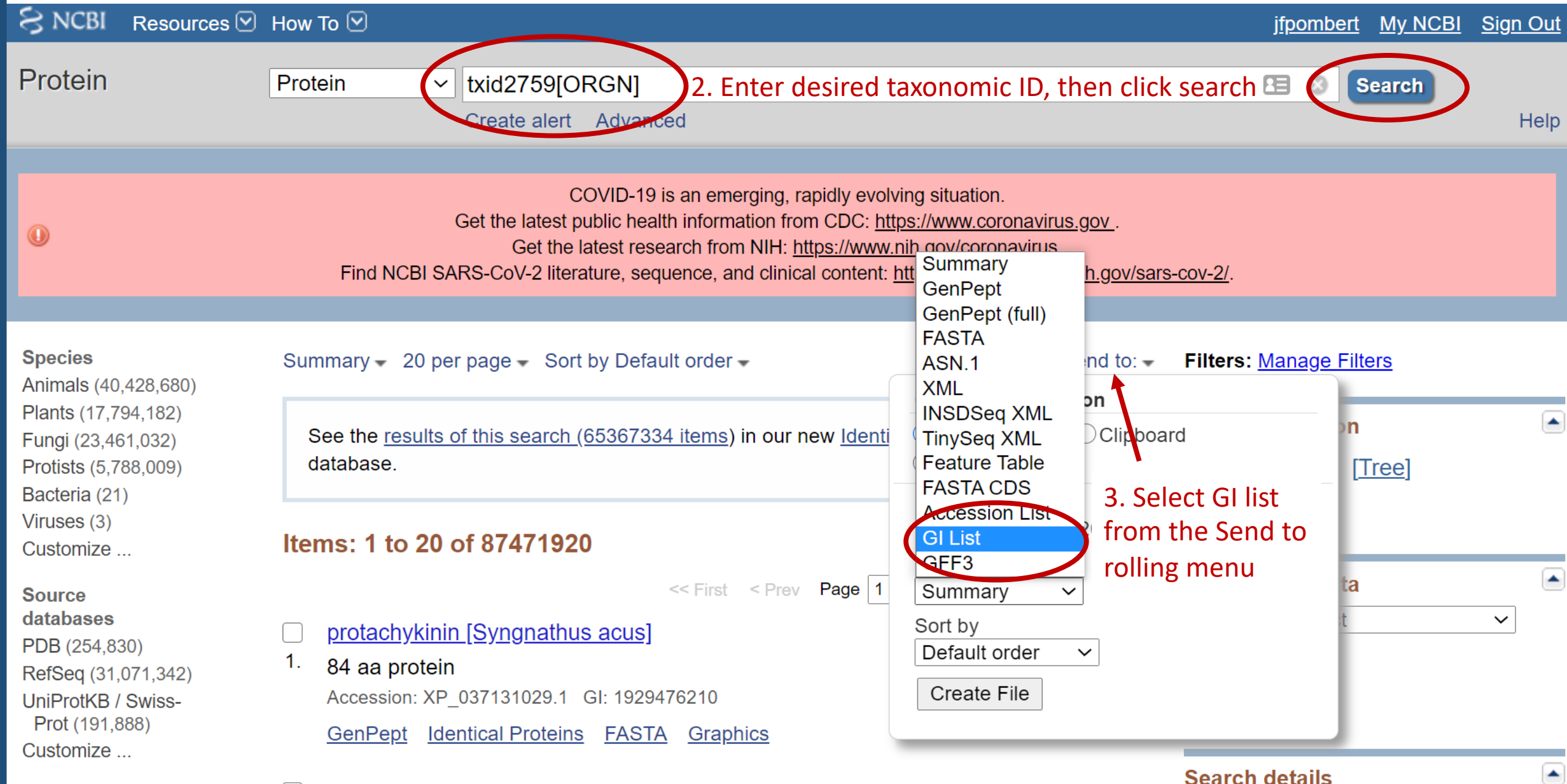
Works with NR and NT

Get taxonomic ID (e.g. txid2759[ORGN]) ## NCBI taxonomy or BLAST (Organism field)

Get GI list from Entrez: http://www.ncbi.nlm.nih.gov/protein/ - NR, /nucleotide/ - NT

How to get GI list from NCBI?

1. Go to http://www.ncbi.nlm.nih.gov/protein/
or https://www.ncbi.nlm.nih.gov/nucleotide



Alternatives to db subsets

Since BLAST 2.4.0+, you can run blast searches with the following options:

-gilist Filename # Restrict search to GI's listed in this file. Local searches only

-negative_gilist Filename # Opposite. Restrict search to everything but the GI's listed

-seqidlist/-negative_seqidlist Filename # Restrict search to list of accession numbers

-query X.fasta -subject Y.fasta # Directly search query against subject

Taxonomy: Classification of living organisms according to their evolutionary relationships

Taxonomized BLASTs

Requires local NCBI Taxonomy database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/taxdb.tar.gz)

Databases must be in the path (export BLASTDB=/path/to/NCBI/TaxDB:/path/to/NCBI/NR:/path/to/NCBI/NT)

Uses a modified outfmt

e.g. -outfmt '6 qseqid sseqid qstart bitscore evalue staxids sscinames sskingdoms sblastnames'

Fields can be defined; http://www.metagenomics.wiki/tools/blast/blastn-output-format-6

Exercise 9 – Taxonomized BLAST

1) Add Mozart's local NCBI NR, NT and TaxDB databases located in /media/Data_1/NCBI/ to your .bash_profile:

```
BLASTDB=$BLASTDB:/media/Data_1/NCBI/TaxDB
BLASTDB=$BLASTDB:/media/Data_1/NCBI/NR
BLASTDB=$BLASTDB:/media/Data_1/NCBI/NT
export BLASTDB
```

2) Run a blastp analysis using TaxBlast.fasta as query against the NCBI nr database using 4 threads the following outfmt format:
   -outfmt '6 qseqid sseqid bitscore evalue staxids sskingdoms sscinames'

Fast and sensitive protein alignment using DIAMOND
Buchfink B *et al.*
Nat Methods. 2015. 12(1):59-60
doi: 10.1038/nmeth.3176

Diamond – Faster BLASTP/BLASTX

https://github.com/bbuchfink/diamond

Up to 20,000X times faster than BLASTP/BLASTX

Works from FASTA or FASTQ files ## Fast enough to work with illumina data!

Easy to use

Fast and sensitive protein alignment using DIAMOND
Buchfink B *et al.*
Nat Methods. 2015. 12(1):59-60
doi: 10.1038/nmeth.3176

Diamond – Running searches

```
diamond makedb --in input.fasta -d custom        ## Creating a custom database
diamond blastx -d custom -q reads.fna -o matches.m8    ## Running blastx equivalent
diamond blastp -d custom -q reads.fq -o matches.m8     ## Running blastp equivalent
```

```
### Downloading NCBI taxonomy accession numbers to taxon ids file
nice -n +15 wget ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/accession2taxid/prot.accession2taxid.gz

### Downloading NCBI taxonomy dmp files (nodes.dmp + names.dmp)
nice -n +15 wget ftp://ftp.ncbi.nlm.nih.gov/pub/taxonomy/taxdmp.zip
unzip -o taxdmp.zip

### Downloading NCBI nr FASTA file => huge file!!!
nice -n +15 wget ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz

### Creating diamond database from NCBI NR fasta files; this will take time!
diamond makedb \
    --threads 64 --in nr.gz \
    --db diamond_nr \
    --taxonnames names.dmp \
    --taxonmap prot.accession2taxid.gz \
    --taxonnodes nodes.dmp
```

Fast and sensitive protein alignment using DIAMOND
Buchfink B *et al.*
Nat Methods. 2015. 12(1):59-60
doi: 10.1038/nmeth.3176

Diamond – Setting up NCBI NR database

We can use the NCBI NR database with Diamond

Searches can be performed with taxonomy

Requires the NCBI NR fasta file + taxonomy files

Fast and sensitive protein alignment using DIAMOND
Buchfink B *et al.*
Nat Methods. 2015. 12(1):59-60
doi: 10.1038/nmeth.3176

```
### Running a taxonomized search with Diamond
diamond blastx \                                                                    ## Blastx or blastp
    -d /media/Data_1/NCBI/Diamond/diamond_nr \                                      ## Database
    -q test.fasta \                                                                 ## Query
    -o test.blastx.diamond \                                                        ## Output
    -f 6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore \   ## Desired fields
         staxids skingdoms sphylums sscinames sblastnames salltitles                ## Desired fields
```

Diamond – Taxonomized searches

We can use the NCBI NR database with Diamond

Searches can be performed with taxonomy

Requires the NCBI NR fasta file + taxonomy files

Annotations from homology searches

Don’t trust them without cross-validation! Why?

- **Online databases can contain many errors** (contributors are not always experts)
- The homology may be too weak to be meaningful
- The homology may be with a different domain that differs from the annotated function(s)

Bottom line? Use critical thinking (aka common sense)

Inferring homology using different tools is a good idea

- We must corroborate our predictions
- Using **different databases** and/or **tools** may yield different results
- Congruency across independent analyses increases confidence in our predictions

We can run BLAST searches against UniProtKB

Metrics as of 2020-10-19

UniProtKB

UniProt Knowledgebase

Swiss-Prot (563,552)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (195,104,019)

Automatically annotated and not reviewed.

Records that await full manual annotation.

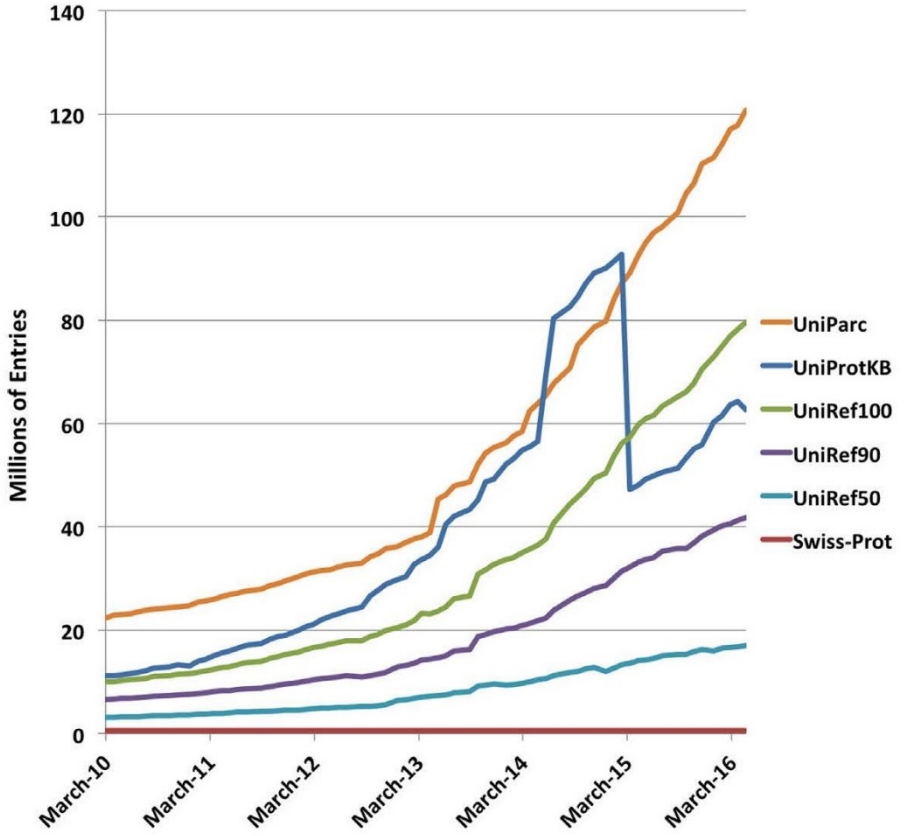


Figure 1. Growth of the number of sequences in UniProt databases. The blue line shows the growth in UniProtKB/TrEMBL entries from January 2010 to date. The sharp drop in UniProtKB entries corresponds to the proteome redundancy minimization (PRM) procedure implemented in March 2015. Note that the post-PRM growth in UniProtKB is no longer exponential.

UniProtKB

http://www.uniprot.org/

Swiss-Prot – Manually annotated and curated

TrEMBL – Automatically annotated, not curated

UniProt: a hub for protein information
UniProt Consortium
Nucleic Acids Res. 2015. 43:D204-D212 Database issue

UniProt: the universal protein knowledgebase
UniProt Consortium
Nucleic Acids Res. 2017. 45:D158-D169 Database issue

Downloads release 2020_05

https://www.uniprot.org/downloads

Basket

UniProt is updated every eight weeks (see FAQ on how to be notified automatically of updates). You can download small data sets and subsets directly from this website by following the download link on any search result page.

We can download and search UniProtKB

For a quick how-to, see:
https://github.com/PombertLab/A2GB#Performing-homology-searches-against-UniProt-databases

Downloading the SwissProt and TrEMBL databases

We can use get_UniProt.pl to download the SwissProt and/or TrEMBL databases from UniProt:

```
get_UniProt.pl -s -t -f $ANNOT/UNIPROT/ -n 20 -l download.log
```

Options for get_UniProt.pl are:

```
-s (--swiss)            Download Swiss-Prot
-t (--trembl)           Download trEMBL
-f (--folder)           Download folder [Default: ./]
-n (--nice)             Linux Process Priority [Default: 20] ## Runs downloads in the background
-l (--log)              Print download information to log file
-d (--decompress)       Decompresss downloaded files with gunzip ## trEMBL files will be huge, off by default
```

Creating tab-delimited product lists from UniProt databases

Homology searches against the UniProt databases will return positive matches against the corresponding accession numbers. However, these matches will not include product names. To facilitate downstream analyses, we can create tab-delimited lists of accession numbers and their products with get_uniprot_products.pl:

```
get_uniprot_products.pl $ANNOT/UNIPROT/uniprot_*.fasta.gz
```

Running DIAMOND or BLAST searches against UniProt databases

We can use DIAMOND to perform homology searches against the UniProt databases. Documentation on how to use DIAMOND can be found here.

First, let's create DIAMOND-formatted databases:

```
mkdir $ANNOT/DIAMOND/; mkdir $ANNOT/DIAMOND/DB/;

diamond makedb \
   --in $ANNOT/UNIPROT/uniprot_sprot.fasta.gz \
   -d $ANNOT/DIAMOND/DB/sprot

diamond makedb \
   --in $ANNOT/UNIPROT/uniprot_trembl.fasta.gz \
   -d $ANNOT/DIAMOND/DB/trembl
```

Second, let's perform protein-protein homology searches against the UniProt databases with a tabular output format (same as NCBI BLAST+'s -outfmt 6 format). Note that these searches will likely take a while depending the total number of proteins queried and the size of these databases:

```
diamond blastp \
   -d $ANNOT/DIAMOND/DB/sprot \
   -q $ANNOT/proteins.fasta \
   -o $ANNOT/DIAMOND/diamond.sprot.6 \
   -f 6

diamond blastp \
   -d $ANNOT/DIAMOND/DB/trembl \
   -q $ANNOT/proteins.fasta \
   -o $ANNOT/DIAMOND/diamond.trembl.6 \
   -f 6
```

We can also search custom reference datasets

Useful if you have access to:

- Well-curated references
- Your own datasets
- Datasets that must be kept private

With **DIAMOND**, we must pipe the data from the **STDIN**

Performing homology searches against reference datasets

If desired, reference datasets (custom or downloaded from NCBI) can also be used as databases in homology searches to help with annotations. NCBI datasets can be accessed from the NCBI genome database or directly from their new dataset repository.

For example, using two datasets downloaded from NCBI:

```
## Downloading data from NCBI
mkdir $ANNOT/REFERENCES/;
wget -O $ANNOT/REFERENCES/ref1.faa.gz https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/007/859/695/GCA_007859695.1
wget -O $ANNOT/REFERENCES/ref2.faa.gz https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/214/015/GCF_000214015.3

## Creating DIAMOND database; using zcat to concatenate the output to STDOUT, then feed it to diamond as input
zcat $ANNOT/REFERENCES/*.gz | diamond makedb -d $ANNOT/DIAMOND/DB/reference

## Running DIAMOND
diamond blastp \
   -d $ANNOT/DIAMOND/DB/reference \
   -q $ANNOT/proteins.fasta \
   -o $ANNOT/DIAMOND/diamond.reference.6 \
   -f 6
```

To create a tab-delimited list of accession numbers and their associated proteins from the downloaded NCBI .faa.gz files, we can use get_reference_products.pl.

```
get_reference_products.pl -f $ANNOT/REFERENCES/*.gz -l $ANNOT/REFERENCES/reference.list
```

For a quick how-to, see:
https://github.com/PombertLab/A2GB#Performing-homology-searches-against-reference-datasets

Domains and motifs-based searches

Most tools use Hidden Markov Model (HMM) approaches

Major tools are:

- Pfam
- CDD
- InterProScan 5

EMBL-EBI

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam

keyword search Go

Pfam 33.1 (May 2020, 18259 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. **More...**

Pfam

The Pfam website will be decommissioned in January 2023.

You will be redirected in 3 seconds to the corresponding data page on the InterPro website.

→ Now part of InterPro/InterProScan

Pfam

http://pfam.xfam.org/

A database of curated protein families

Probabilistic, based on Hidden Markov model (HMM)

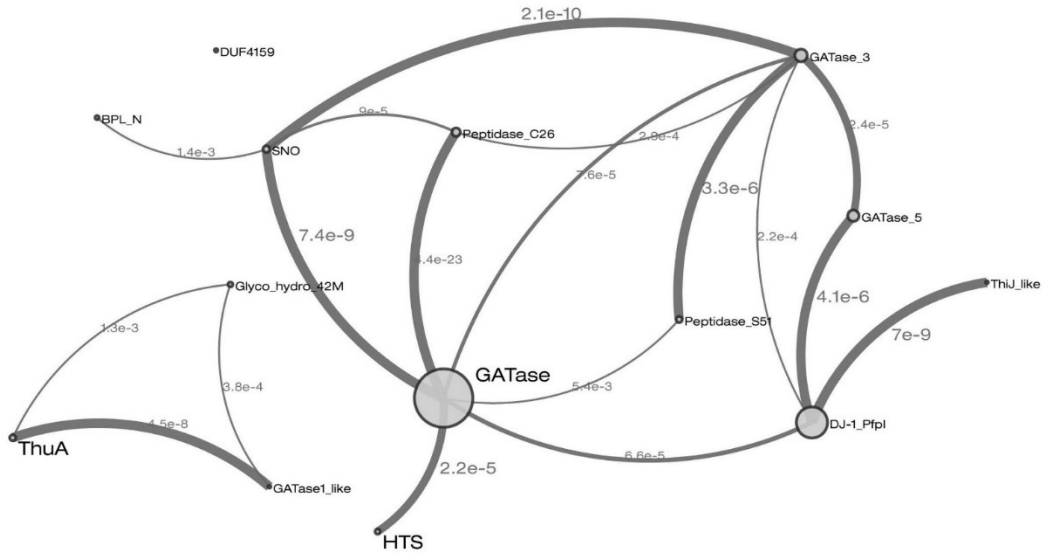
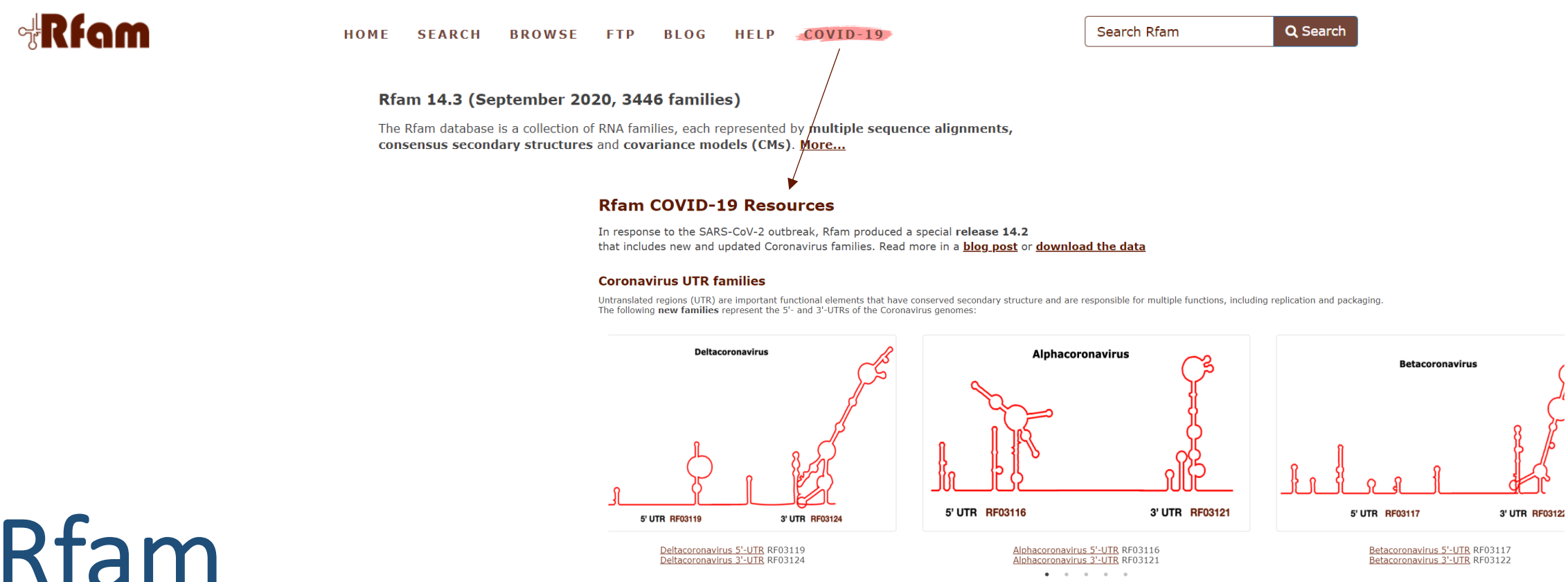


Figure 1. Example of the improved representation of relationships graph, indicating the similarity between the Pfam entries within a clan. This particular entry shows the relationship between the entries in the Glutaminase I clan (accession:CL0014). Each entry in the clan is a node in the graph and is represented as circle, with the diameter of the circle being proportional to the number of sequences in the *full* alignment. Nodes are connected (edges) based on the HHsearch results between the clan members, with the width of edges proportional to the E-value of the HHsearch similarity (E-values ≤ 0.01 are deemed significant). The clanviewer component has been included in the BioJS registry (http://biojs.io/d/clanviewer) and its code is freely available in github (https://github.com/ProteinsWebTeam/clanviewer). In this particular clan, there are three entries (ThuA (PF06283), GATase1_like (PF07090) and Glyco_hydro_42M (PF08532)) that from a disconnected sub-cluster. DUF4159 (PF13709) is also unconnected to any other entry. However, these entries are included as part of this clan based on the structural similarities to other entries in the clan.

Clans = groups of related families

The Pfam protein families database
Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al.
Nucleic Acids Res. 2014. 42:D222-D230 Database Issue

The Pfam protein families database: towards a more sustainable future
Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al.
Nucleic Acids Res. 2016. 44:D279-D285 Database Issue



Rfam

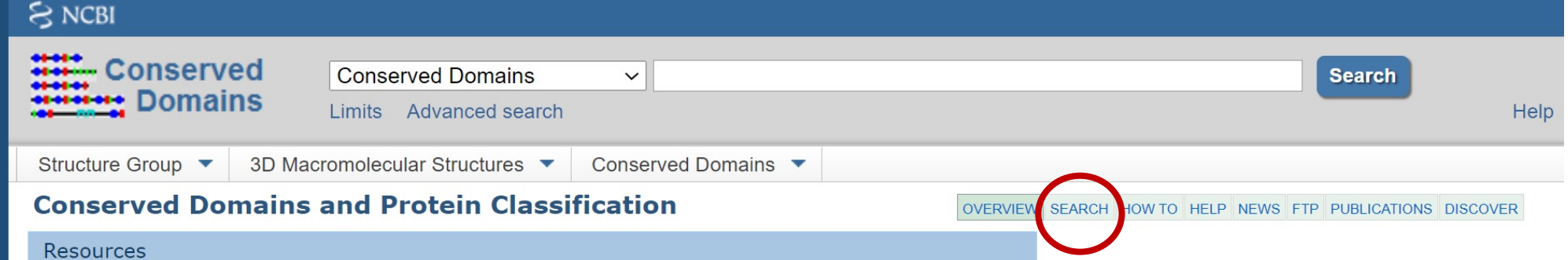
http://rfam.xfam.org/

A database of curated RNA families [http://rfam.xfam.org/families]

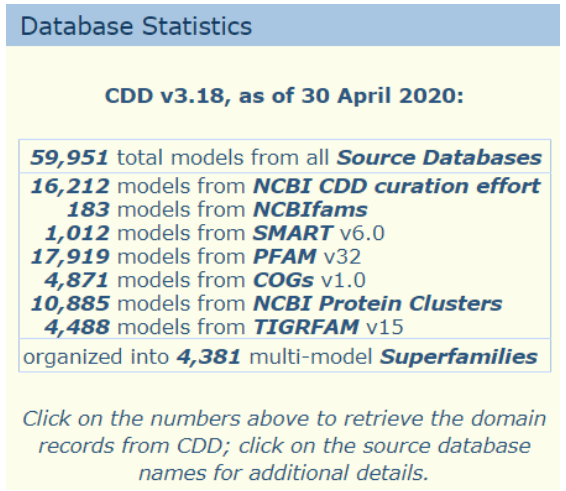
BLAST queries, then Infernal HMM alignments [http://infernal.janelia.org/]

Rfam 12.0: updates to the RNA families database
Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR, et al.
Nucleic Acids Res. 2015. 43:D130-D137 Database issue

Infernal 1.1: 100-fold faster RNA homology searches
Nawrocki EP, Eddy SR
Bioinformatics 2013. 29:2933-2935



CDD: NCBI's conserved domain database
Marchler-Bauer A *et al.*
Nucl. Acids Res. 2015. 43:D222–D226
Database issue



NCBI CDD & CD-Search

Conserved **D**omain **D**atabase

http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

https://www.ncbi.nlm.nih.gov/books/NBK279685/

```
deltablast -query query.fsa -db cdd_delta        ## We can also run it from the command line with NCBI BLAST+
```

DB available at: https://ftp.ncbi.nlm.nih.gov/blast/db/cdd_delta.tar.gz

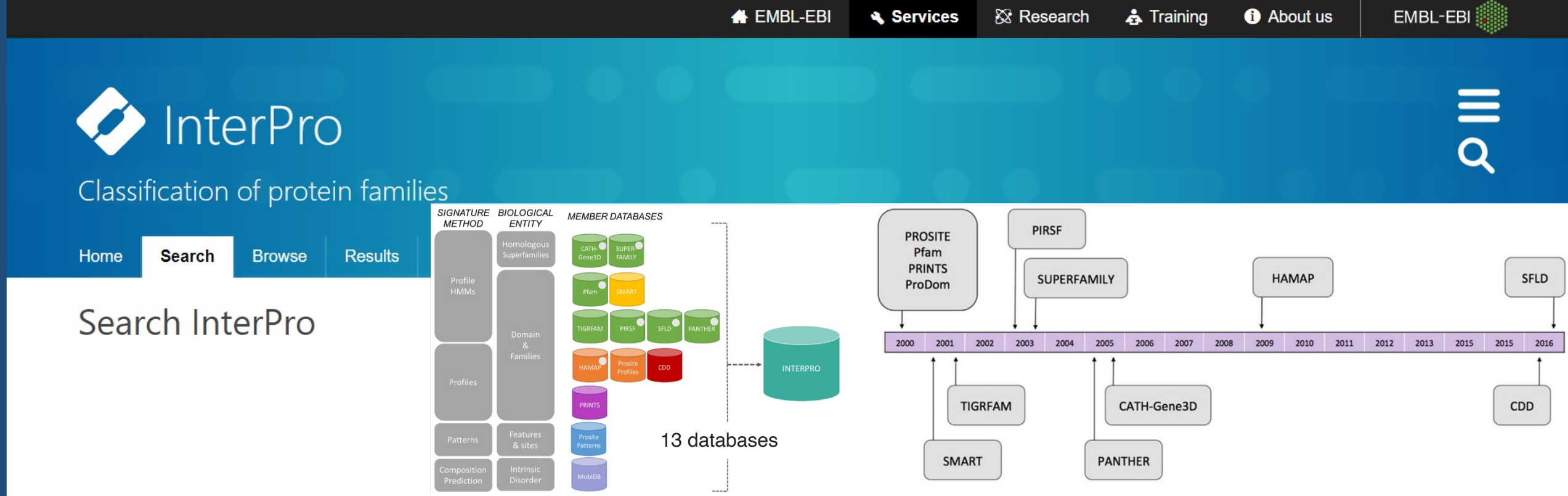


Figure 2. Timeline showing the member databases that have joined InterPro since version 1.0, released in 2000.

InterPro in 2017—beyond protein family and domain annotations
Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, et al.
Nucleic Acids Res. 2017. 45:D190-D199 Database issue

InterProScan

http://www.ebi.ac.uk/interpro/search/sequence-search

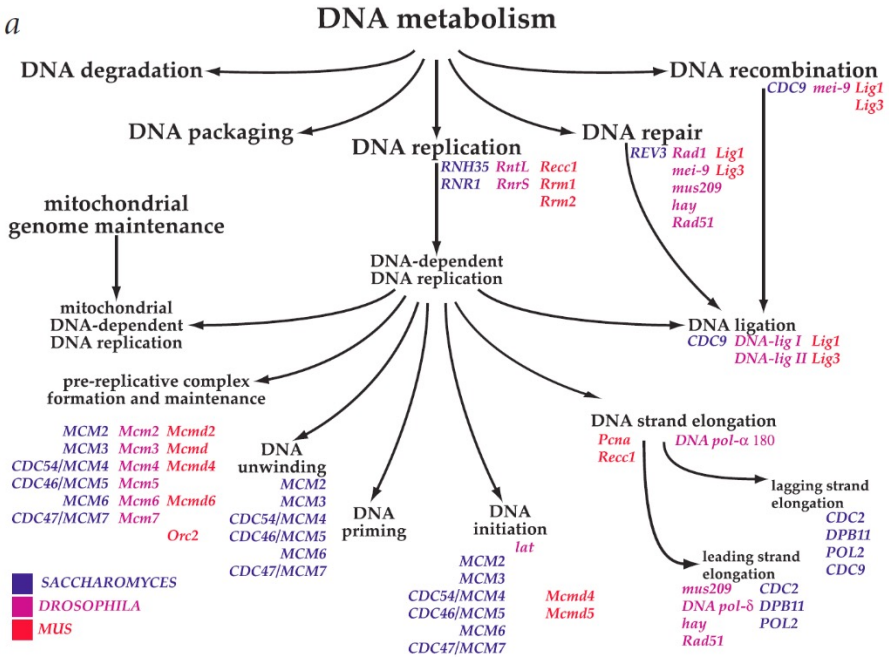
Searches for multiple evidences (acts as an aggregator)

Queries Pfam, CDD, Gene ontologies, KEGG pathways, and more

The InterPro protein families database: the classification resource after 15 years
Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, et al.
Nucleic Acids Res. 2015. 43:D213-D221 Database issue

InterProScan 5: genome-scale protein function classification
Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al.
Bioinformatics 2014. 30(9):1236-1240

Cellular Component	• What is it part of? • Organelle, macromolecular structure, etc.
Biological Process	• A series of events that accomplishes something • Not a pathway, a specific result
Molecular Function	• Specific functional activity • Catalytic activity, transporter activity, etc.



Gene ontology: tool for the unification of biology
Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, et al.
Nat Genet. 2000. 25(1):25-29

Gene ontologies (GOs)

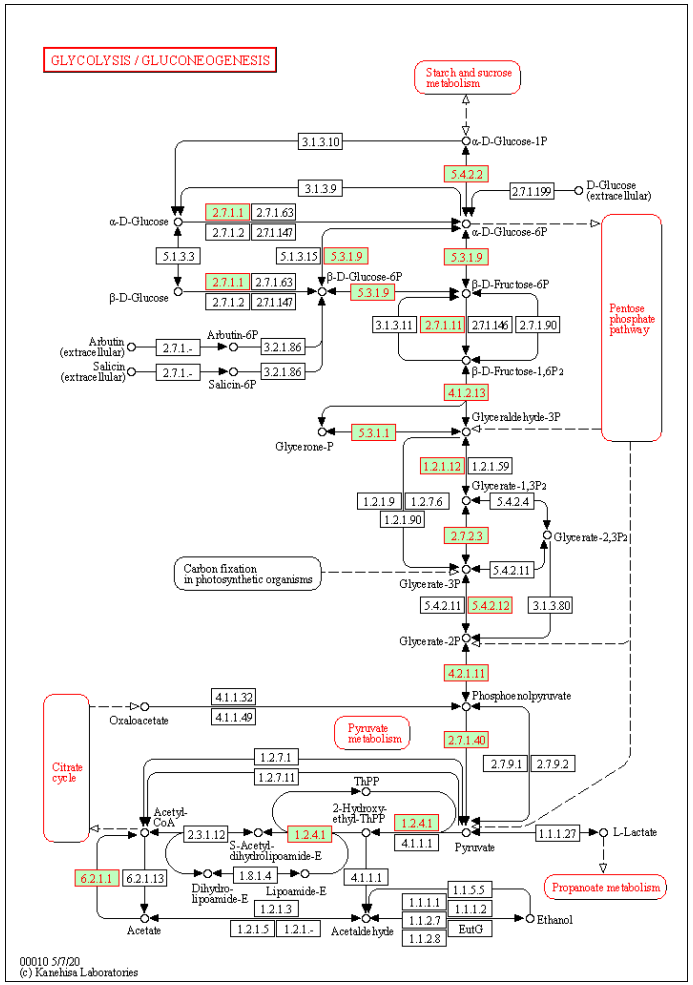
http://geneontology.org/

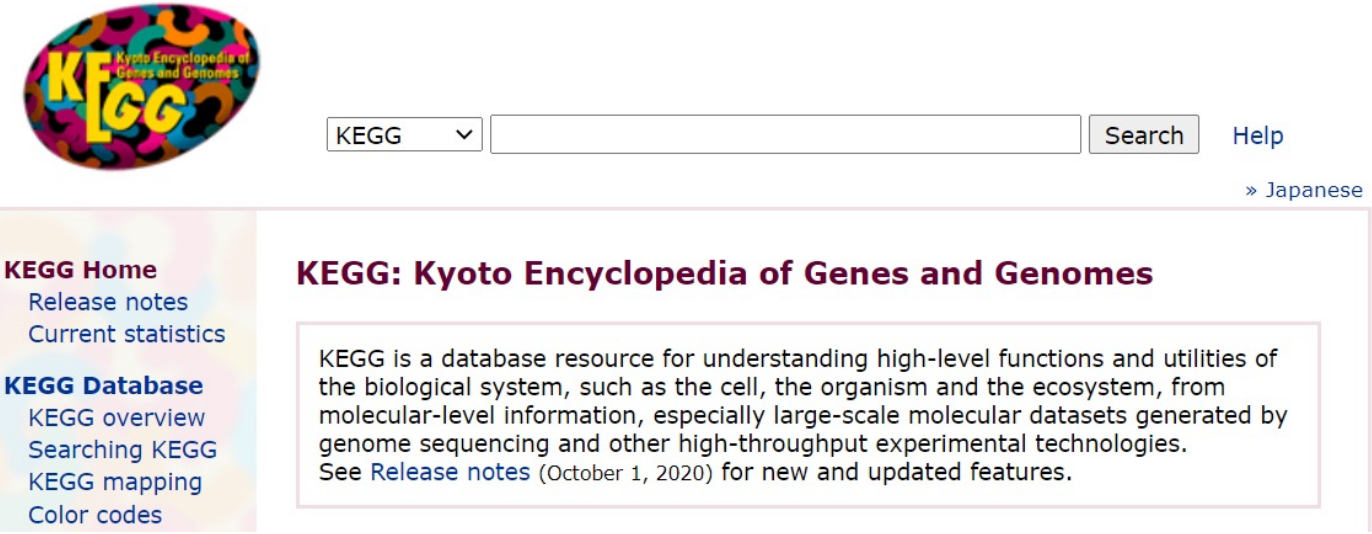
Structured product descriptions

Consistent across databases

Directed Acyclic Graphs

Example of a KEGG metabolic pathway





The KEGG databases

http://www.genome.jp/kegg/kegg2.html

Kyoto Encyclopedia of Genes and Genomes

The reference for metabolic pathways

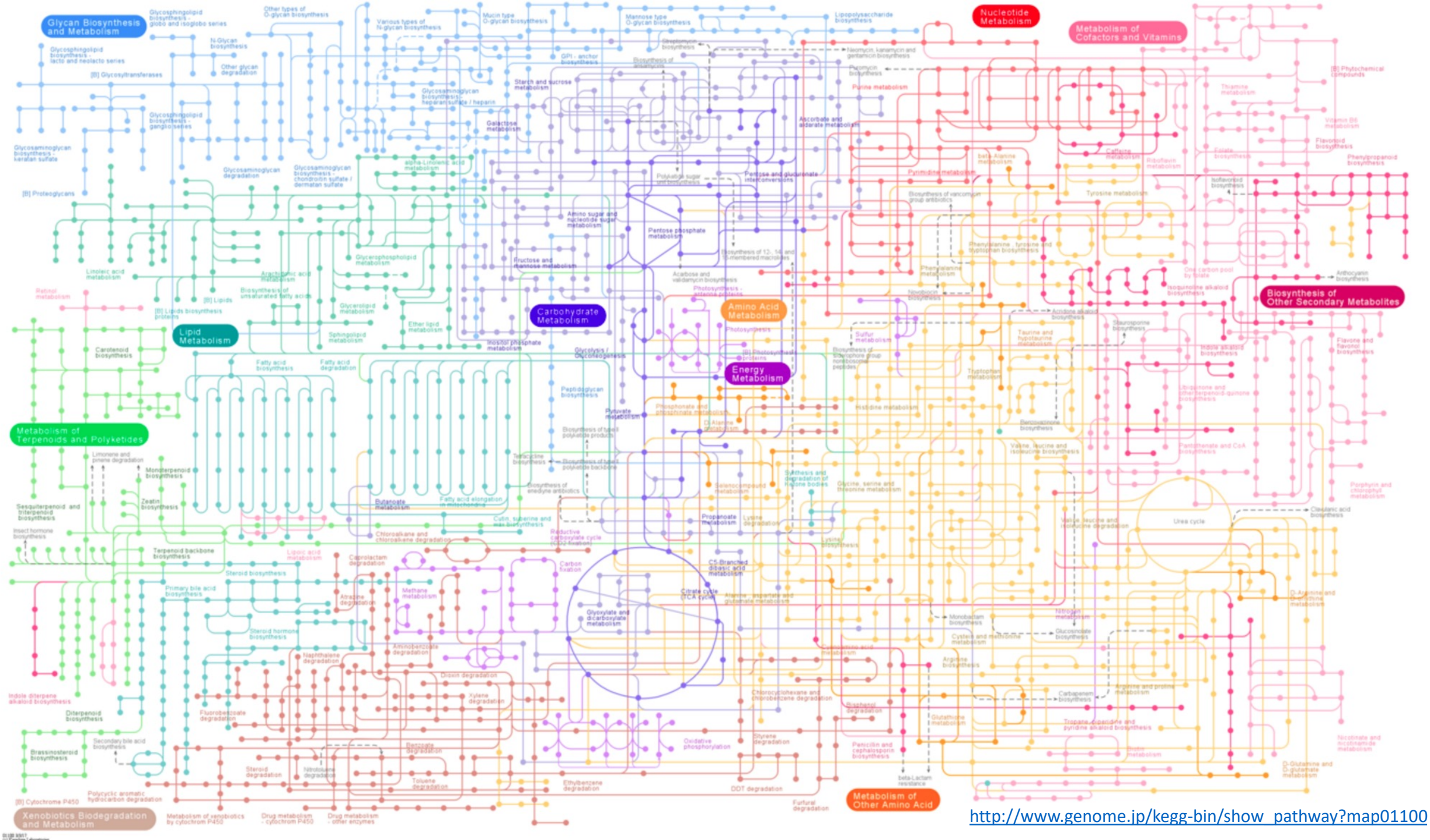
KOs instead of GOs; the idea of orthology is very important for pathways

Pathways - http://www.genome.jp/kegg/pathway.html

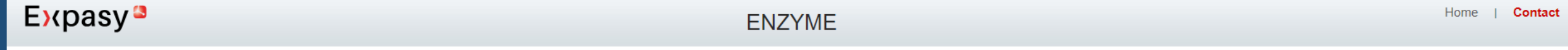
Orthology - http://www.genome.jp/kegg/ko.html ## Stores molecular-level functions

Organisms - http://www.genome.jp/kegg/catalog/org_list.html

KEGG: new perspectives on genomes, pathways, diseases and drugs
Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K
Nucleic Acids Res. 2017. 45:D353-D361 Database issue



http://www.genome.jp/kegg-bin/show_pathway?map01100



ENZYME
Enzyme nomenclature database

ENZYME is a repository of information relative to the nomenclature of enzymes. It is primarily based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB) and it describes each type of characterized enzyme for which an EC (Enzyme Commission) number has been provided [More details / References].
ENZYME now includes entries with preliminary EC numbers. Preliminary EC numbers include an 'n' as part of the fourth (serial) digit (e.g. EC 3.5.1.n3).

Release of 07-Oct-20 (6538 active entries)

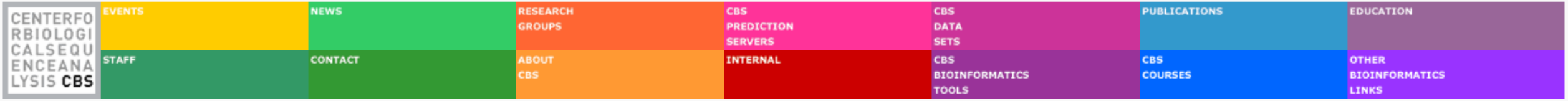
EC (Enzyme Commission) numbers

http://enzyme.expasy.org/

EC numbers are like IP addresses (4 digits separated by periods), e.g. EC 1.14.17.3

Unified enzyme nomenclature

The ENZYME database in 2000
Bairoch A
Nucleic Acids Res. 2000. 28:304-305



Signal motifs

Signal peptide – secretion

Target peptide – location

Locating proteins in the cell using TargetP, SignalP, and related tools
Emanuelsson O, Brunak S, von Heijne G, Nielsen H
Nat Protoc. 2007. 2:953-971

SignalP 4.0: discriminating signal peptides from transmembrane regions
Petersen TN, Brunak S, von Heijne G, Nielsen H
Nature Methods. 201.1 8:785-786

TMpred

Home | Contact

Prediction of Transmembrane Regions and Orientation

The TMpred program makes a prediction of membrane-spanning regions and their orientation. The algorithm is based on the statistical analysis of TMbase, a database of naturally occurring transmembrane proteins. The prediction is made using a combination of several weight-matrices for scoring.

- **K. Hofmann & W. Stoffel (1993)**
  TMbase - A database of membrane spanning proteins segments
  Biol. Chem. Hoppe-Seyler 374,166

Transmembrane domains

TMHMM	https://services.healthtech.dtu.dk/service.php?TMHMM-2.0
TopPred	http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::toppred
Tmpred	http://www.ch.embnet.org/software/TMPRED_form.html

Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes
Krogh A, Larsson B, von Heijne G, Sonnhammer EL
J Mol Biol. 2001. 305:567–580

Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule
von Heijne G
J Mol Biol. 1992. 225:487–494

Exercise 10 – CDD

1) Go to: http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
2) Paste the content of CDD.fasta in the window
3) Click on Browse results for a detailed view
4) Select the desired query in the left panel. Click Show selected query.
5) In the main window, click on one of the domains display to inspect MSA of your query against the curated database sequences.

   ## MSA = Multiple sequence alignment

Exercise 11 – Pfam

1) Go to: http://pfam.xfam.org/
2) You cannot use multifasta files in the sequence search box
3) Instead, click on the search link at the top of the page
4) Then, select the batch search option on the left
5) Upload Pfam.fasta. Enter your email address (some searches may take a long time). Submit.
6) Look up your emails. Results will be sent to your inbox.
7) You can also do the analyses one by one. Try with the 1st one. (The web display is easier to look at)

Exercise 12 – Web InterProScan

1) Go to: http://www.ebi.ac.uk/interpro/
2) The search only works for single fasta files.
3) Paste the content of interpro.fasta, then submit the job.
4) Look at the detailed signatures.
5) You can export your search results at the top right of the screen.
6) The GO at the bottom of the page are gene ontologies. You can learn more about GOs from http://geneontology.org/.

Exercise 13 – Local InterProScan 5

Running InterProScan locally is great for large protein datasets.

1) From the Ex_13 folder, run:

```
interproscan.sh \
  -i inter.fasta \          ## FASTA input file
  -iprlookup \              ## InterPro annotations
  -goterms \                ## Gene ontologies
  -pa \                     ## Pathways (KEGG)
  -b output_interpro \      ## Output name
  -f GFF3 HTML TSV          ## Produces HTML output
```

2) You will obtain GFF3, HTML and TSV files. The TSV (tab-separated values) file is the one you really want. This tab-delimited text file can be parsed easily with Perl. It can also be opened with MS Excel or similar spreadsheet tools. Look at it.

Structural homology

- 3D structures often confer functions in biology
- Structural homology can be used as a complement to other analyses

Mandatory car analogy! :

Rubber wheels on cars and wooden wheels on ox carts serve the same function

The data points showed in the figure illustrate the mean server response time, where larger points correspond to faster servers. The vertical bars correspond to the standard deviation of a server for the given time frame (only model-1 data is incorporated). Please refer to the Help section for further details about the assessment procedure.

Dataset composition 13 weeks (2016-10-14 - 2017-01-07)

	All Targets	Easy	Medium	Hard
Monomer	96	20	42	34
Homo-Oligomers	96	25	52	19
containing ligand(s)	114	28	60	26
Total	**249**	**52**	**132**	**65**

Total Number of Models: 3135

The Protein Model Portal--a comprehensive resource for protein structure and model information
Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, Schwede T
Database. 2013:bat031. doi: 10.1093/database/bat031.

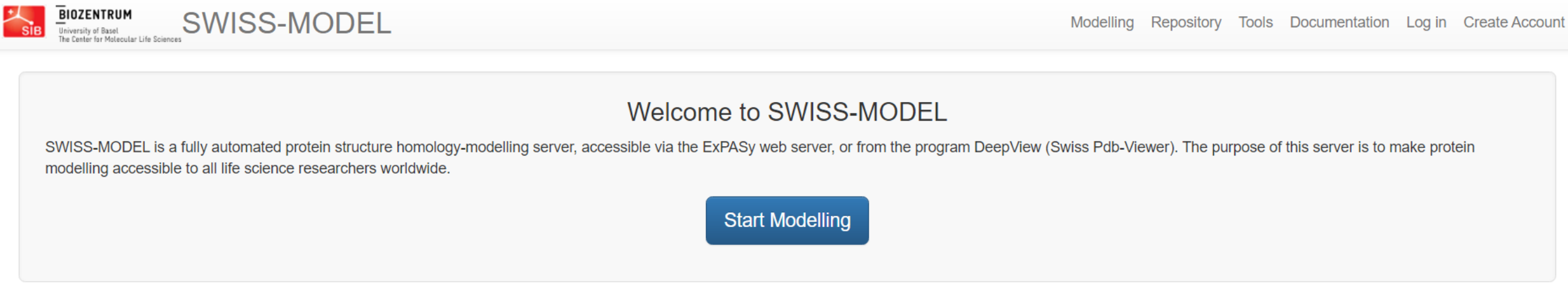
Predicting 3D structures

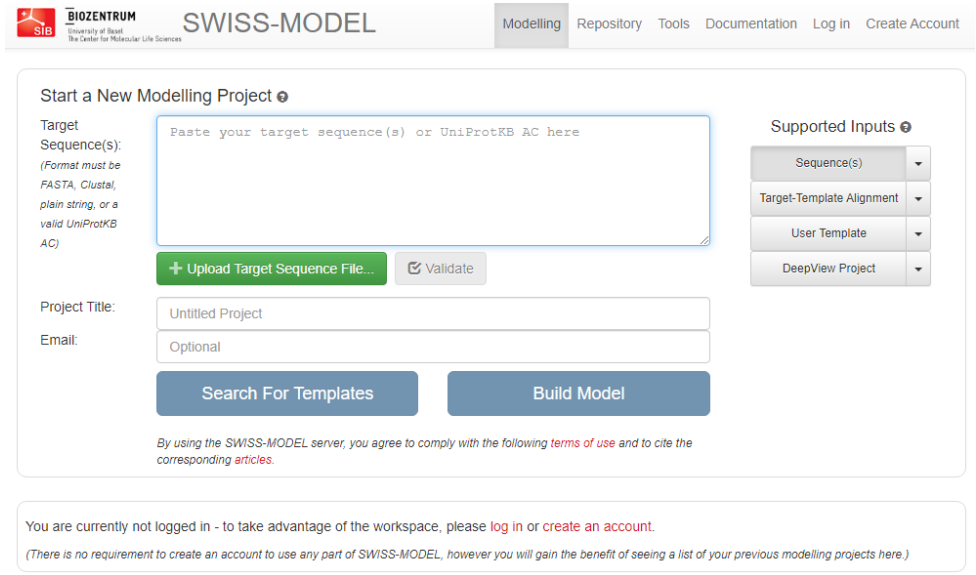
Many protein predictors exist

Not every predictor is reliable

No predictor is the best at everything

CAMEO Continuous Automated Model EvaluatiOn – http://cameo3d.org/





SWISS-MODEL

https://swissmodel.expasy.org/

Predicts 3D structures of proteins from FASTA files

Quick and easy to use

Web-server is hosted by the Swiss Institute of Bioinformatics

All Projects

Untitled Project Created: today at 16:50

To download results

Summary Templates 50 Models 1

Model Results

Order by: GMQE

Oligo-State: Homo-trimer (matching prediction)
Ligands: None
GMQE: 0.70
QMEAN: -3.57

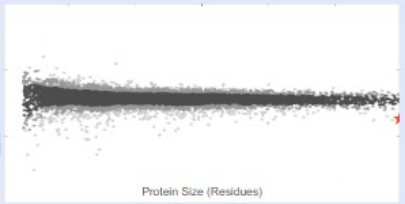
Model 01
Structure Assessment

Global Quality Estimate

QMEAN	-3.57
Cβ	-1.93
All Atom	-0.69
solvation	0.94
torsion	-3.47

Local Quality Estimate

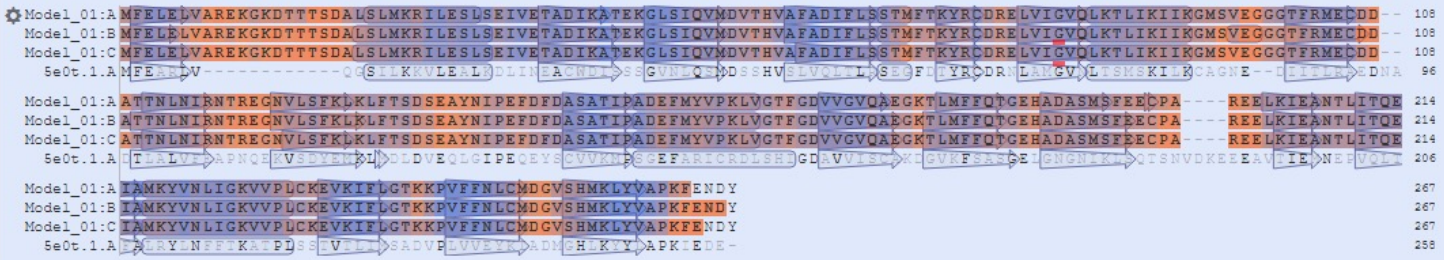
Comparison

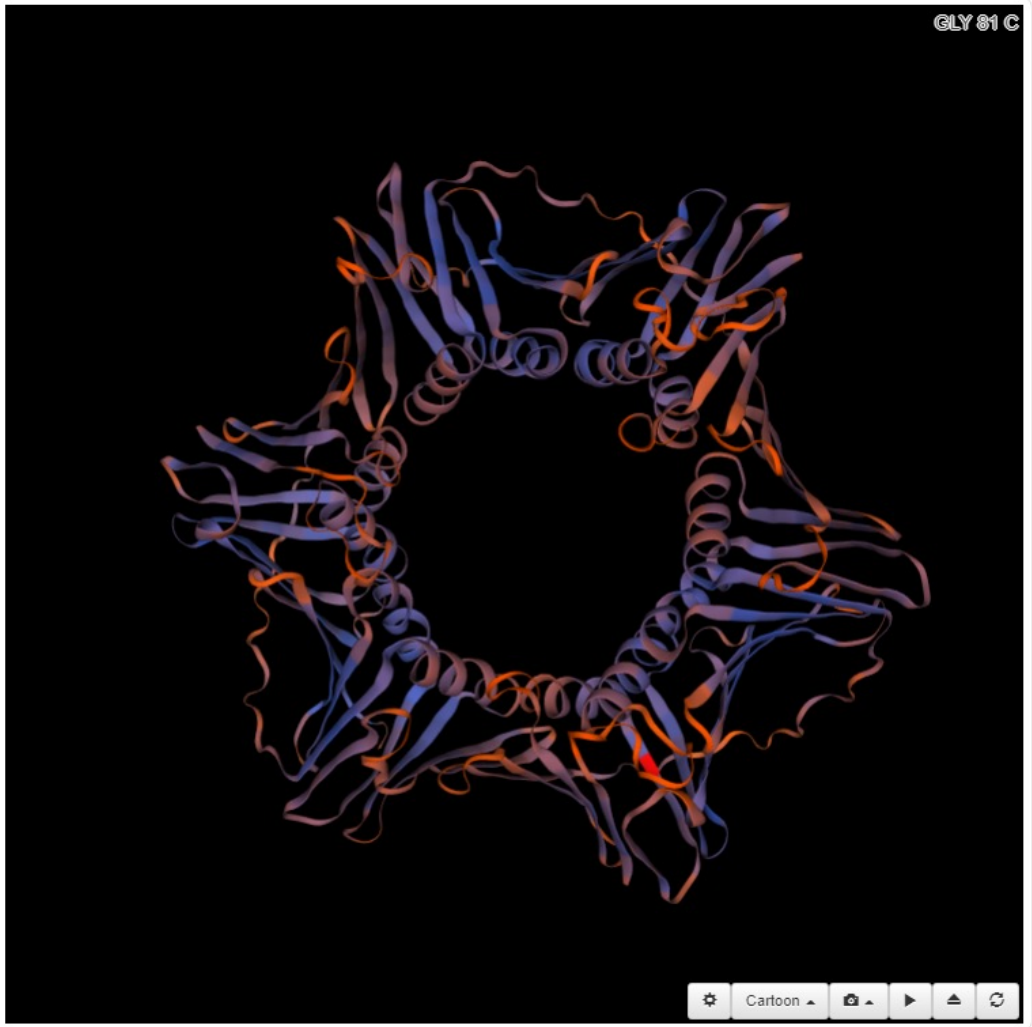


Template: 5e0t.1.A
Seq Identity: 28.57%
Coverage

Description: Proliferating cell nuclear antigen
Human PCNA mutant - S228I

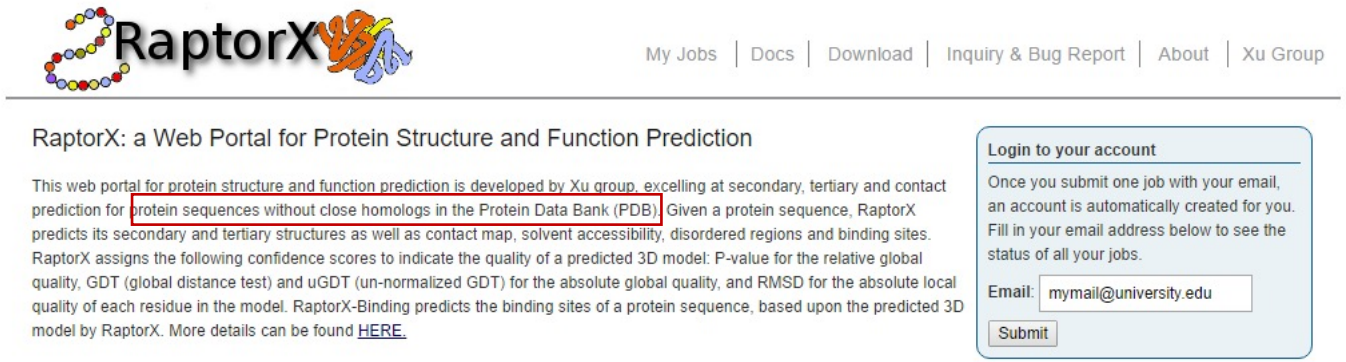
Model-Template Alignment





We can play with structure here

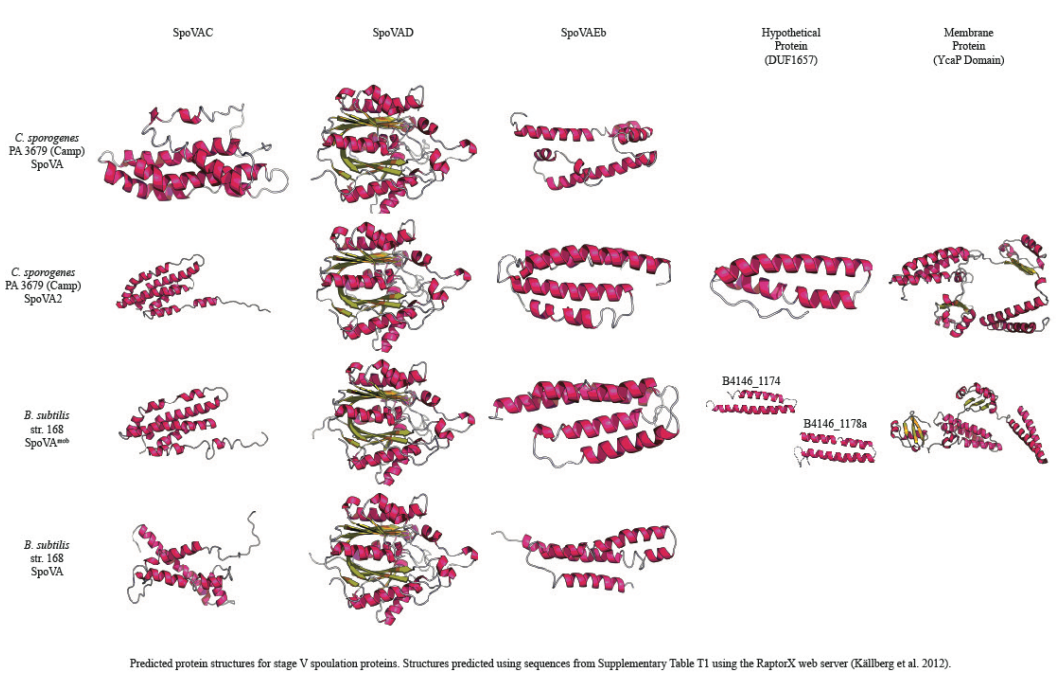
Example of results with SWISS-MODEL



Template-based protein structure modeling using the RaptorX web server
Källberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, Xu J
Nat Protoc. 2012. 7(8):1511-22. doi: 10.1038/nprot.2012.085

Protein structure alignment beyond spatial proximity
Wang S, Ma J, Peng J, Xu J
Sci Rep. 2013. 3:1448. doi: 10.1038/srep01448

Alignment of distantly related protein structures: algorithm, bound and implications to homology modeling
Wang S, Peng J, Xu J
Bioinformatics. 2011. 27(18):2537-45



Predicted protein structures for stage V spoulation proteins. Structures predicted using sequences from Supplementary Table T1 using the RaptorX web server (Källberg et al. 2012).

Genetic characterization of the exceptionally high heat resistance of the non-toxic surrogate *Clostridium sporogenes* PA 3679
Butler RR 3rd, Schill KM, Wang Y, Pombert JF
Front Microbiol. 2017 Apr 3;8:545. doi: 10.3389/fmicb.2017.00545. eCollection 2017

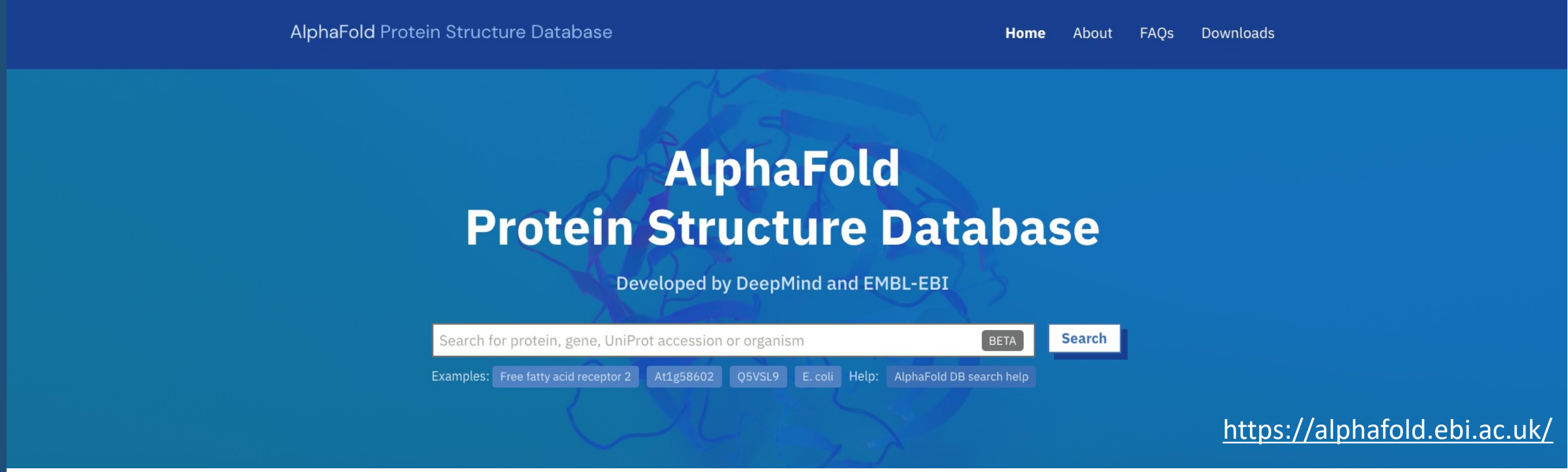
RaptorX

Developed at University of Chicago – http://raptorx.uchicago.edu/

One of the top predictors

Easy to use web interface

Can align protein to references



AlphaFold2

Highly accurate protein structure prediction with AlphaFold
Jumper J. *et al.*
Nature. 2021. PMID: 34265844

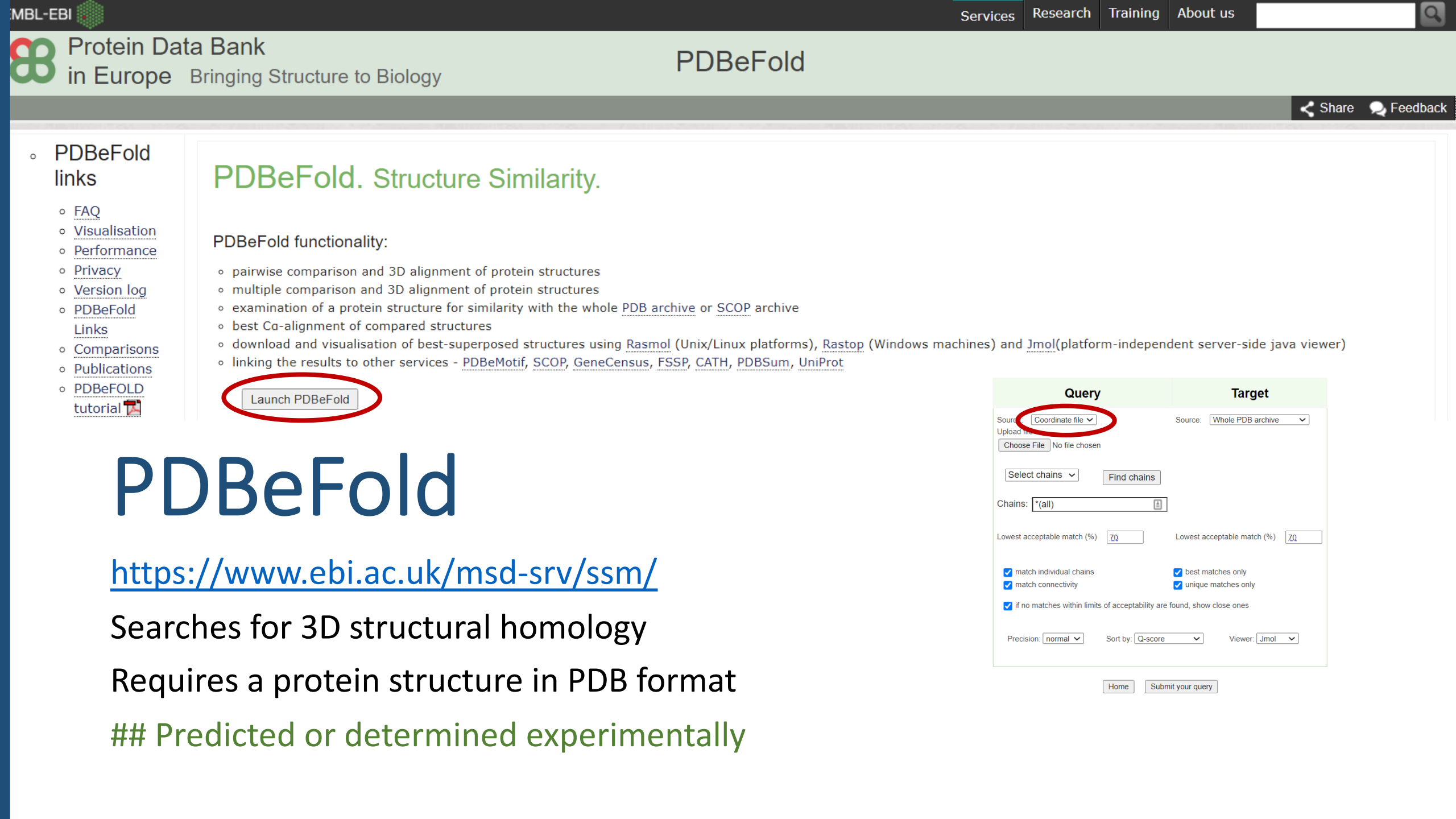
Developed by DeepMind – https://deepmind.com/

Available as a docker image

Based on deep learning ## The best tool in CASP14

Leverages TensorFlow

https://github.com/deepmind/alphafold



PDBeFold

https://www.ebi.ac.uk/msd-srv/ssm/

Searches for 3D structural homology

Requires a protein structure in PDB format

Predicted or determined experimentally

Structure Alignment Results.

Query: ECU05_1030_1sxjF.raptorx.pdb, 267 residues

Examined 163976 entries, (442651 chains). **Displaying Matches 1-20 of 379.**

Back to query | next | last page | Sort by Q-score | arrange by SCOP family | match 1 | jump

Example of results with PDBeFold

##	Scoring			RMSD	N_{align}	N_g	$\%_{seq}$	Query	Target (PDB entry)				
	Q	P	Z					$\%_{sse}$	Match	$\%_{sse}$	N_{res}	×	Title
1	0.90	41.8	19.3	0.42	252	2	22	91	1sxj:F	91	258	☐	CRYSTAL STRUCTURE OF THE EUKARYOTIC CLAMP LOADER (REPLICATION FACTOR C, RFC) BOUND TO THE DNA SLIDING CLAMP (PROLIFERATING CELL NUCLEAR ANTIGEN, PCNA)
2	0.84	41.8	19.4	0.91	250	3	22	100	6e49:B	96	255	☐	PIF1 PEPTIDE BOUND TO PCNA TRIMER
3	0.83	41.4	19.3	0.90	249	3	22	100	6b8i:B	96	255	☐	ROLE OF THE PIF1-PCNA COMPLEX IN POL DELTA DEPENDENT STRAND DISPLACEMENT DNA SYNTHESIS AND BREAK INDUCED REPLICATION
4	0.83	40.7	19.1	0.97	250	3	22	100	6e49:A	96	255	☐	PIF1 PEPTIDE BOUND TO PCNA TRIMER
5	0.83	41.3	19.3	0.84	246	4	22	100	6e49:C	96	254	☐	PIF1 PEPTIDE BOUND TO PCNA TRIMER
6	0.83	40.4	19.1	0.92	248	4	22	100	6b8i:A	96	255	☐	ROLE OF THE PIF1-PCNA COMPLEX IN POL DELTA DEPENDENT STRAND DISPLACEMENT DNA SYNTHESIS AND BREAK INDUCED REPLICATION

The maximum possible Q-score is 1.0 ## Anything above 0.7 is very similar structurally

RCSB PDB Deposit Search Visualize Analyze Download Learn More MyPDB

RCSB PDB PROTEIN DATA BANK

170172 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

PDB – A database of experimentally determined structures

Advanced Search | Browse Annotations

PDB-101 | Worldwide PDB | EMDataResource | NDB Nucleic Acid Database | Worldwide Protein Data Bank Foundation

Welcome

Deposit

Search

Visualize

Analyze

Download

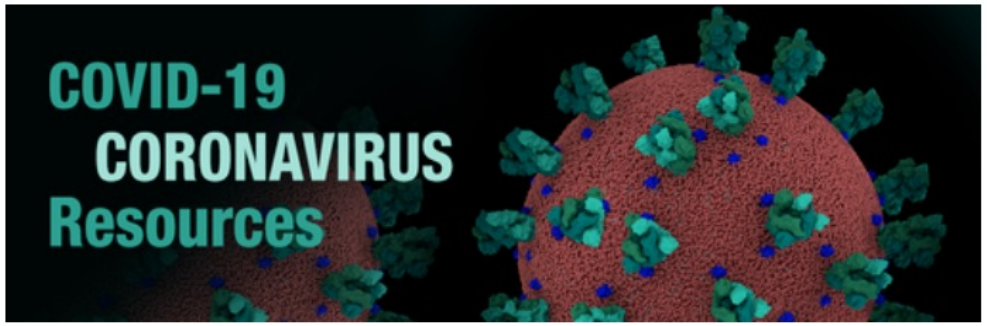
Learn

A Structural View of Biology

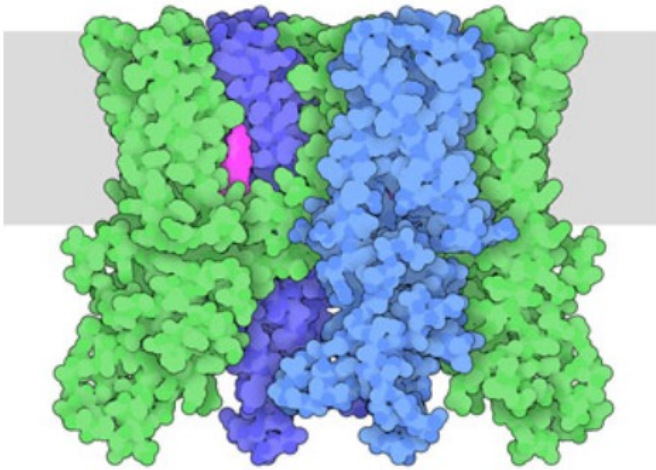
This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.



October **Molecule of the Month**



Capsaicin Receptor TRPV1

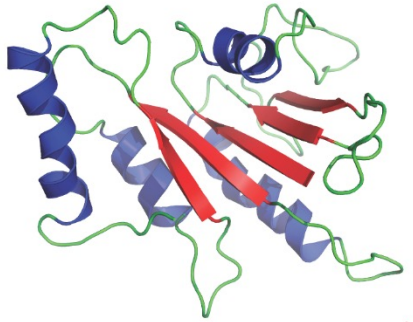
https://www.rcsb.org/

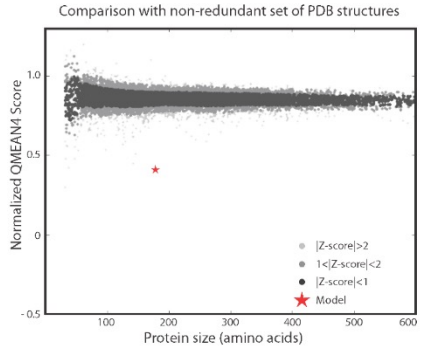
Structures can be downloaded from: https://www.rcsb.org/pages/download/ftp

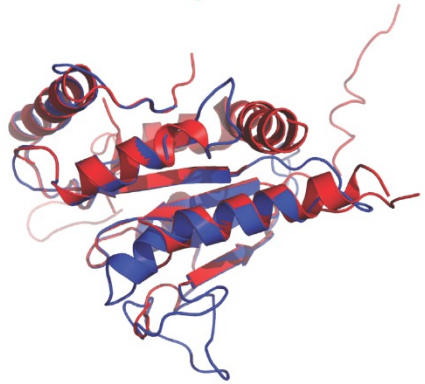
SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling
Guex N, Peitsch MC
Electrophoresis. 1997. 18:2714-2723

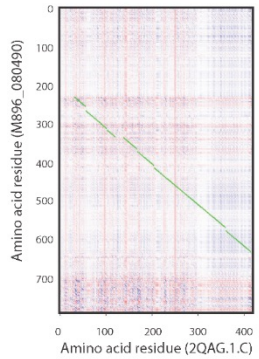
Defining and searching for structural motifs using DeepView/Swiss-PdbViewer
Johansson MU, Zoete V, Michielin O, Guex N
BMC Bioinformatics. 2012. 13:173

UCSF Chimera--a visualization system for exploratory research and analysis
Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE
J Comput Chem. 2004. 25(13):1605-12









Viewing structures

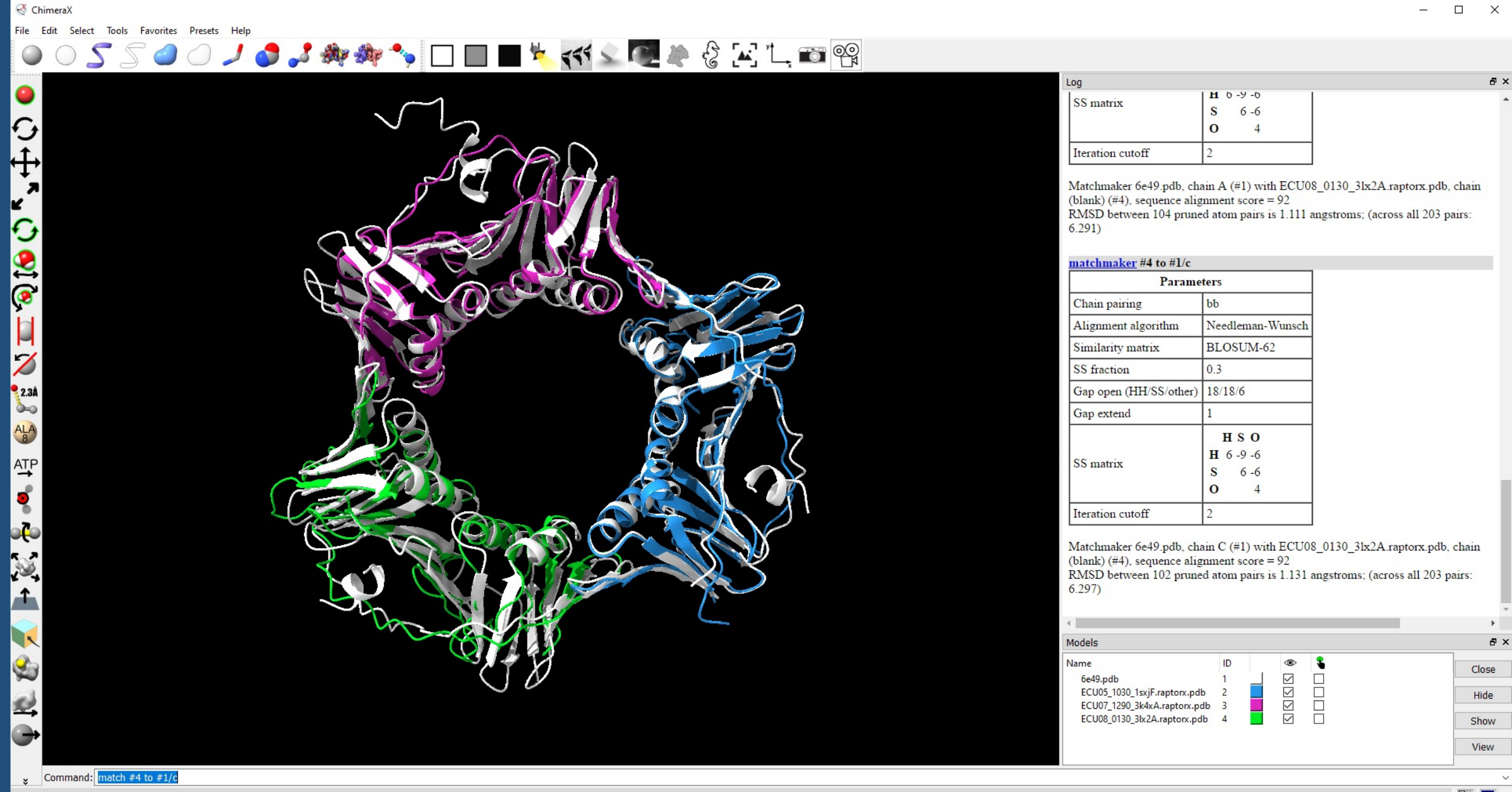
The *Ordospora colligata* genome; evolution of extreme reduction in microsporidia and host-to-parasite horizontal gene transfer
Pombert JF, Haag KL, Beidas S, Ebert D, Keeling PJ
mBio. 2015. 6(1):e02400-14.

Swiss-PdbViewer http://spdbv.vital-it.ch/

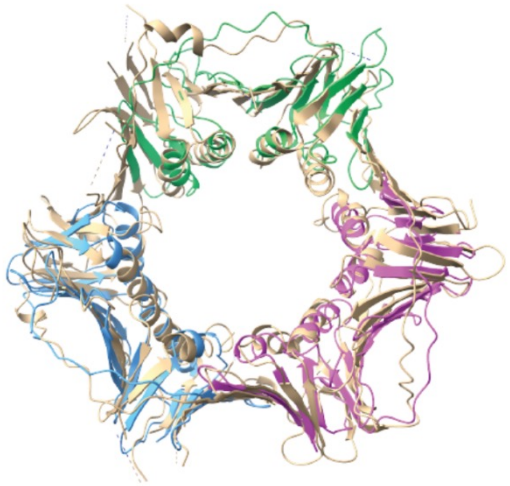
ChimeraX https://www.cgl.ucsf.edu/chimerax/

PyMOL https://www.pymol.org/

Some features are limited to
the non-free version



We can match structures against others with the match function; We can rotate, edit, make videos... ## Powerful visualization of 3D structural homology



3DFI

Three-dimensional functional inference
using structural homology

Enhanced fold recognition using efficient short fragment clustering
Krissinel E
J Mol Biochem. 2012;1(2):76-85. PMID: 27882309 PMCID: PMC5117261

```
### GPU96_01g00180-m1-3se6A
GPU96_01g00180-m1-3se6A       1   3SE6   B     0.8635   0.4638   0.2713   811   860   pdb3se6.ent.gz   CRYSTAL STRUCTURE OF THE HUMAN ENDOPLASMIC RETICULUM AMINOPEPTIDASE 2
GPU96_01g00180-m1-3se6A       2   3SE6   A     0.8611   0.4749   0.2699   815   870   pdb3se6.ent.gz   CRYSTAL STRUCTURE OF THE HUMAN ENDOPLASMIC RETICULUM AMINOPEPTIDASE 2
### GPU96_01g00200-m1-1j3iC
GPU96_01g00200-m1-1j3iC       1   5H39   A     0.8836   0.8169   0.6064   282   285   pdb5h39.ent.gz   STRUCTURAL ANALYSIS OF KSHV THYMIDYLATE SYNTHASE
GPU96_01g00200-m1-1j3iC       2   3IHI   B     0.8829   0.7784   0.6464   280   283   pdb3ihi.ent.gz   CRYSTAL STRUCTURE OF MOUSE THYMIDYLATE SYNTHASE
### GPU96_01g00840-m1-5hvmA
GPU96_01g00840-m1-5hvmA       1   5HVM   A     0.8841   0.4410   0.4465   430   446   pdb5hvm.ent.gz   STRUCTURE OF ASPERGILLUS FUMIGATUS TREHALOSE-6-PHOSPHATE SYNTHASE A INCOMPLEX WITH UDP AND VALIDOXYLAMINE A
GPU96_01g00840-m1-5hvmA       2   5HVM   B     0.8639   0.5333   0.4419   430   452   pdb5hvm.ent.gz   STRUCTURE OF ASPERGILLUS FUMIGATUS TREHALOSE-6-PHOSPHATE SYNTHASE A INCOMPLEX WITH UDP AND VALIDOXYLAMINE A
```

3DFI

https://github.com/PombertLab/3DFI

Automates structural homology searches at the 3D level ## A Perl pipeline

Leverages **RaptorX/AlphaFold2**, **GESAMT**, and **ChimeraX**

Can be used to query against both predicted and/or experimental structures

Optional

Additional slides and/or exercises for the curious

Exercise 14 – SignalP

1) Go to http://www.cbs.dtu.dk/services/SignalP/
2) Upload signal_1.fasta. Use the Eukaryotes group (default)
3) Look at the output
4) Now upload signal_2.fasta and look at the results
5) Notice the difference?

- Signal and target peptides when present are located in the N-terminal portions of the protein
- A signal peptide motif suggests that the protein is likely secreted

Exercise 15 – TargetP

1) Go to http://www.cbs.dtu.dk/services/TargetP/
2) Upload targetP_1.fasta. Use the Non-plant group (default)
3) Look at the output
4) Now re-upload targetP_1.fasta using the Plant group instead.
5) Notice a difference?
   - Target peptides are not always conserved and differ between various groups
   - Keep in mind that the localization results you obtain are *in silico* predictions only.

Exercise 16 – TMHMM

1) Go to http://www.cbs.dtu.dk/services/TMHMM/
2) Upload tmhmm.fasta
3) Look at the output
4) Here, outside means outside the membrane
5) In this protein, only the C-terminal domain is anchored to the membrane