bioinformatics.ca

Canadian Bioinformatics Workshops

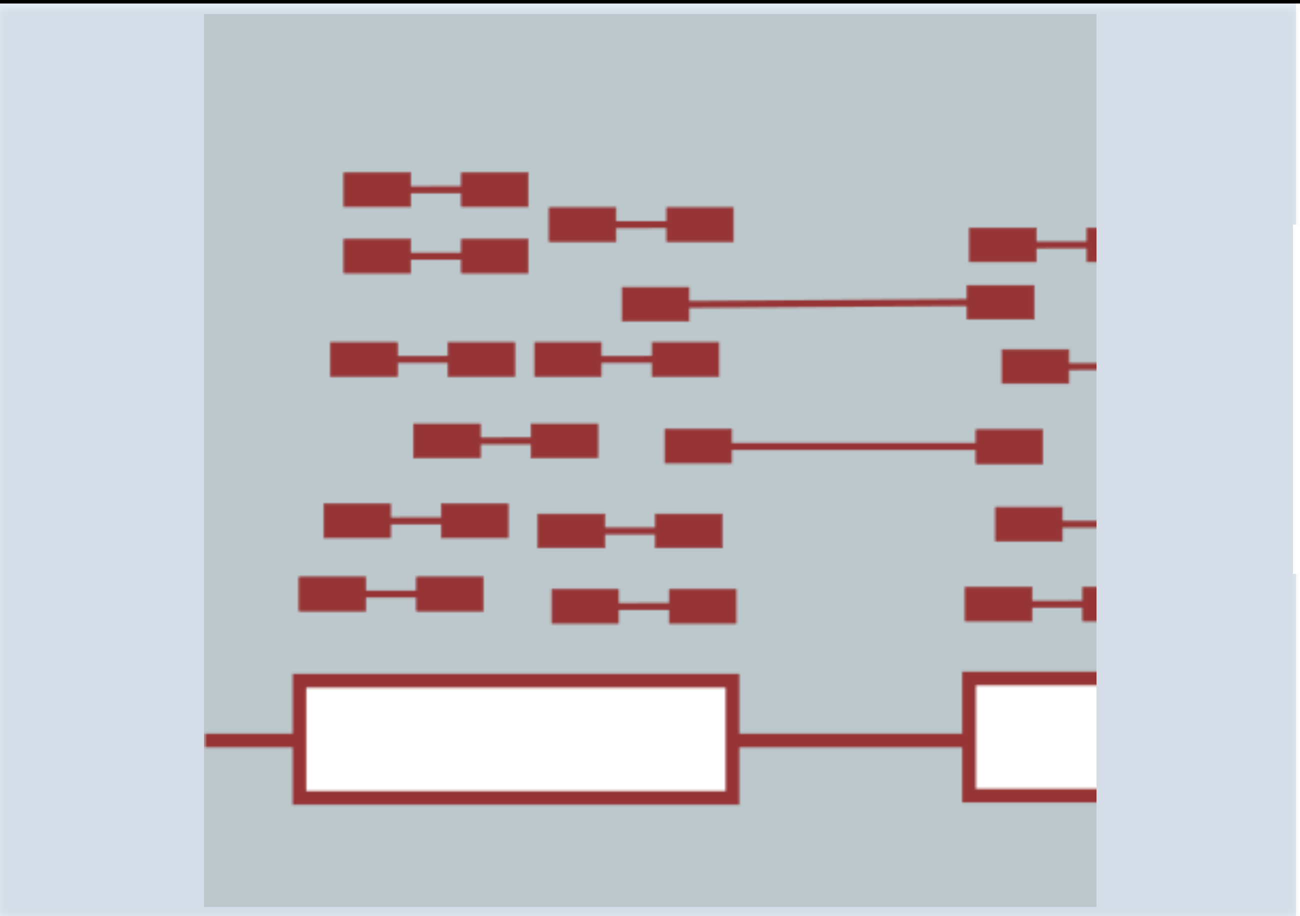
www.bioinformatics.ca

Functional Annotation and Analysis of Transcripts

Brian Haas

Informatics for RNA-Seq Analysis
July 10-12, 2017

bioinformatics.ca



Learning Objectives of Module

- Explore methods to glean biological function from transcript sequences.
- Differentiate between homology-based and sequence composition-based functional inference.

Transcript Functional Annotation

```
GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCGTGGTCCT
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGA                                                 TCGAC
TCTCCC                                                 TCCCA
AAAGAC                                                 CCTGG
GGCTTC                                                 CCTAA
TGACCT                                                 TGCTG
GAAAAC                                                 CAGCC
TTGTCA                                                 TTCCA
```

Can we gather hints of biological function from sequence?

```
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG
ATGTGGTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC
```

Methods used to predict function from sequence

- Sequence homology

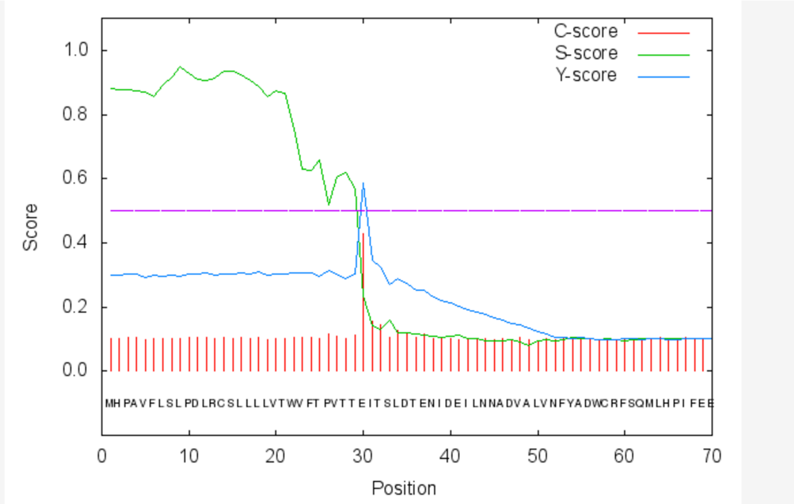
Searching protein database for sequence similarity

```
         Query THVHRPYNEHKSLSGTARYMSINTHLGREQSRRDDLESMGHVFMYFLRGSLPW--QGLKA
               T    P + K   GT  Y S + HLG    RR DLE +G      L   LPW  Q L A
Database Match TGDFKP-DPKKMHNGTIEYTSRDAHLG-VPTRRADLEILGYNLIEWLGAELPWVTQKLLA
```

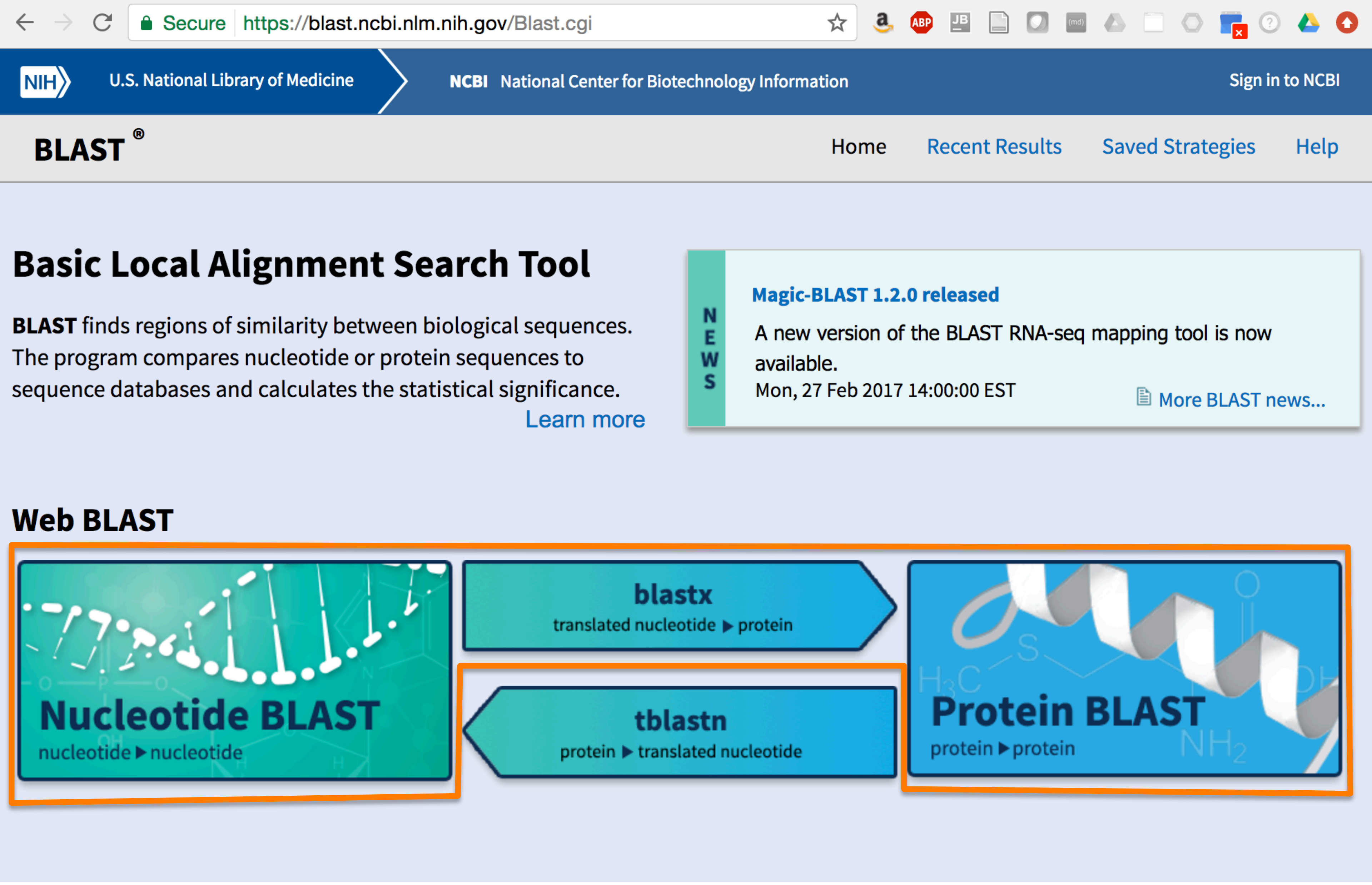
- Sequence composition

Predict functions of sequence using machine learning methods for pattern recognition.

- Neural Networks
- Hidden Markov Models



Use BLAST to search for sequence similarity to known proteins



Secure
https://blast.ncbi.nlm.nih.gov/Blast.cgi
NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
BLAST®
Home
Recent Results
Saved Strategies
Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

NEWS

Magic-BLAST 1.2.0 released

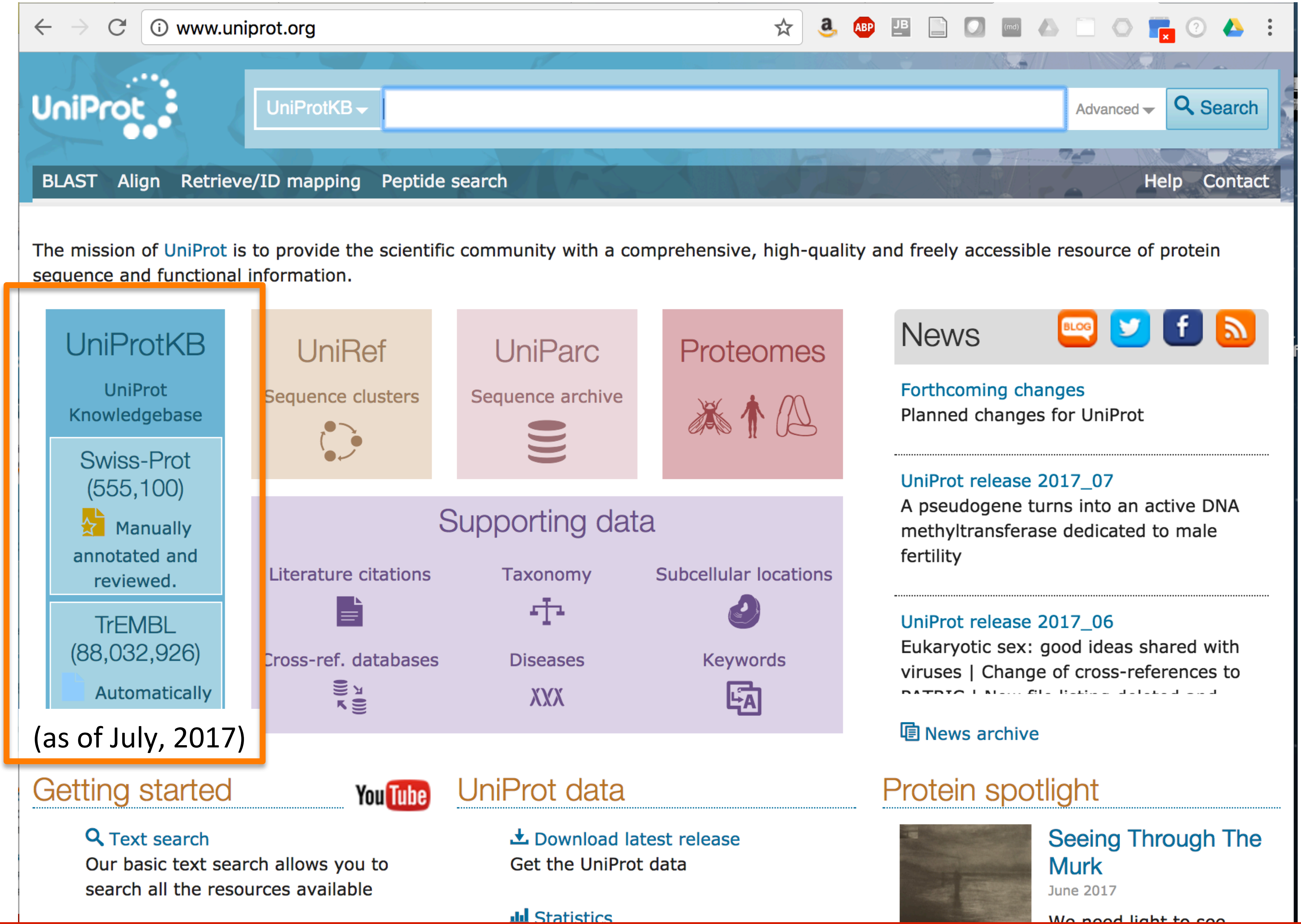
A new version of the BLAST RNA-seq mapping tool is now available.

Mon, 27 Feb 2017 14:00:00 EST

More BLAST news...

Web BLAST

The Swiss-Prot database is a valuable source of proteins with known functions



The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB

UniProt Knowledgebase

Swiss-Prot (555,100)

Manually annotated and reviewed.

TrEMBL (88,032,926)

Automatically

(as of July, 2017)

UniRef

Sequence clusters

UniParc

Sequence archive

Proteomes

Supporting data

Literature citations

Taxonomy

Subcellular locations

Cross-ref. databases

Diseases

Keywords

News

Forthcoming changes

Planned changes for UniProt

UniProt release 2017_07

A pseudogene turns into an active DNA methyltransferase dedicated to male fertility

UniProt release 2017_06

Eukaryotic sex: good ideas shared with viruses | Change of cross-references to PATRIC | New file listing deleted and

News archive

Getting started

Text search

Our basic text search allows you to search all the resources available

UniProt data

Download latest release

Get the UniProt data

Statistics

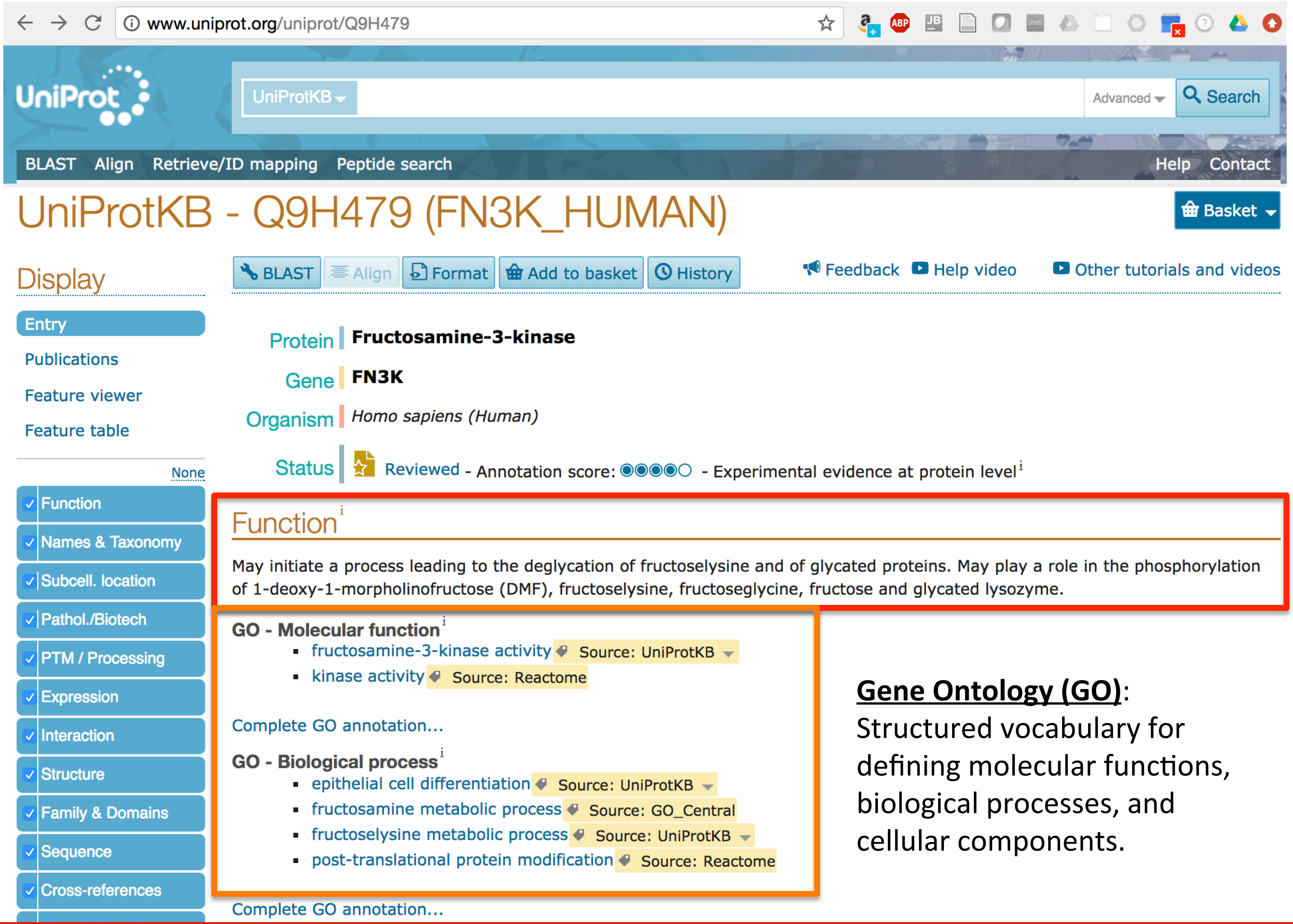
Protein spotlight

Seeing Through The Murk

June 2017

We need light to see

Example of a Swiss-Prot Record



UniProtKB - Q9H479 (FN3K_HUMAN)

Protein: **Fructosamine-3-kinase**

Gene: **FN3K**

Organism: *Homo sapiens (Human)*

Status: Reviewed - Annotation score: ●●●●○ - Experimental evidence at protein level[i]

Function[i]

May initiate a process leading to the deglycation of fructoselysine and of glycated proteins. May play a role in the phosphorylation of 1-deoxy-1-morpholinofructose (DMF), fructoselysine, fructoseglycine, fructose and glycated lysozyme.

GO - Molecular function[i]

- fructosamine-3-kinase activity Source: UniProtKB
- kinase activity Source: Reactome

Complete GO annotation...

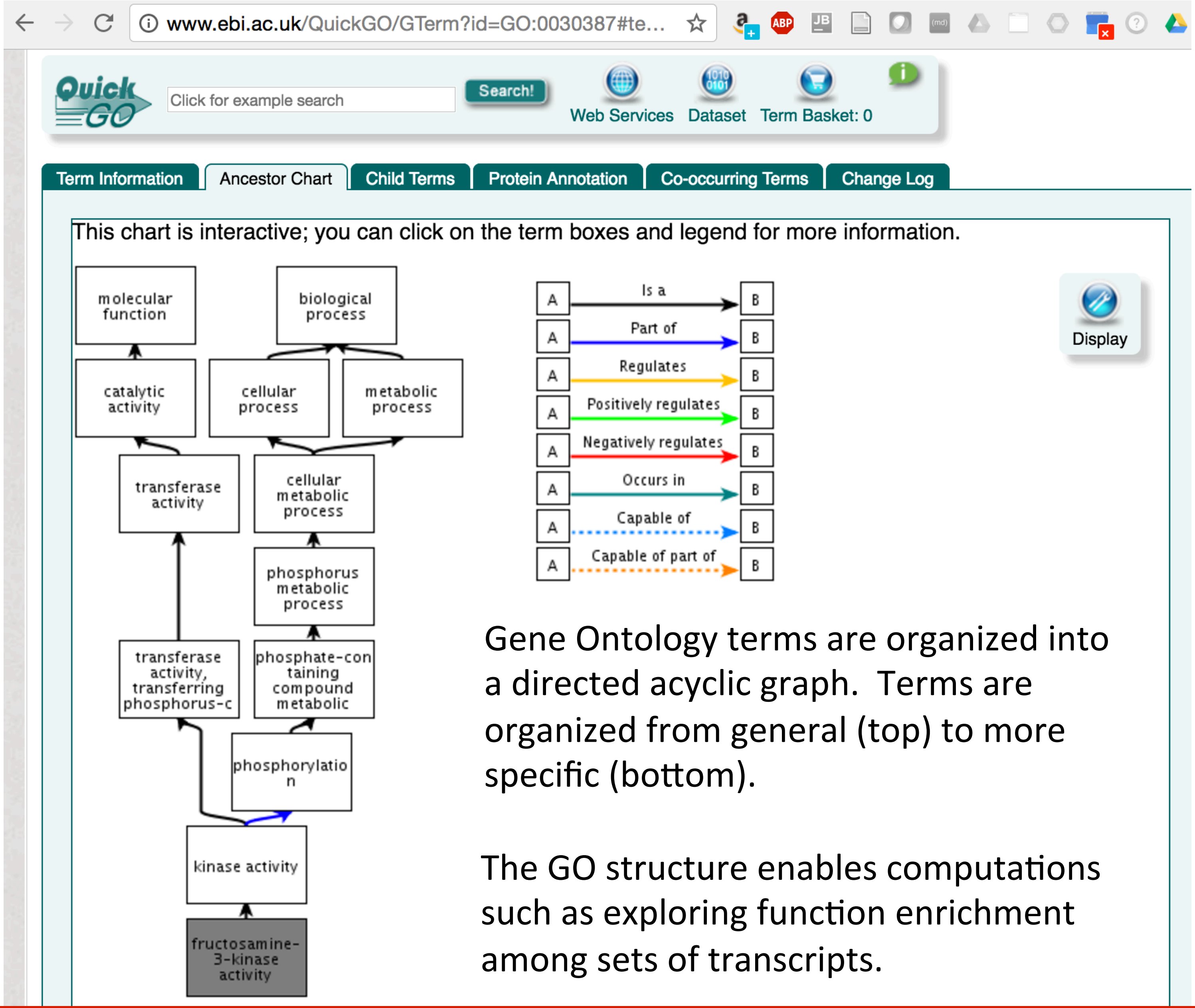
GO - Biological process[i]

- epithelial cell differentiation Source: UniProtKB
- fructosamine metabolic process Source: GO_Central
- fructoselysine metabolic process Source: UniProtKB
- post-translational protein modification Source: Reactome

Complete GO annotation...

Gene Ontology (GO): Structured vocabulary for defining molecular functions, biological processes, and cellular components.

Gene Ontology: a structured relational vocabulary for describing biological functions



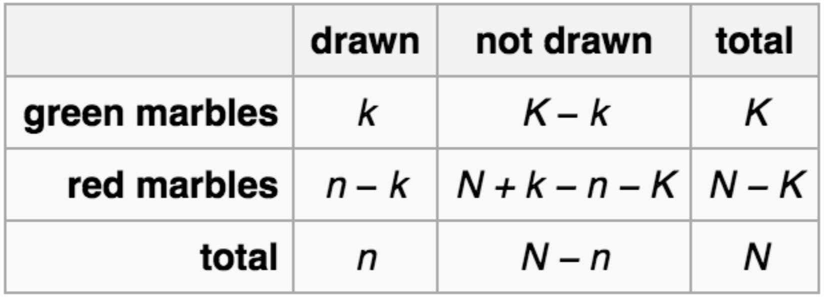
This chart is interactive; you can click on the term boxes and legend for more information.

Gene Ontology terms are organized into a directed acyclic graph. Terms are organized from general (top) to more specific (bottom).

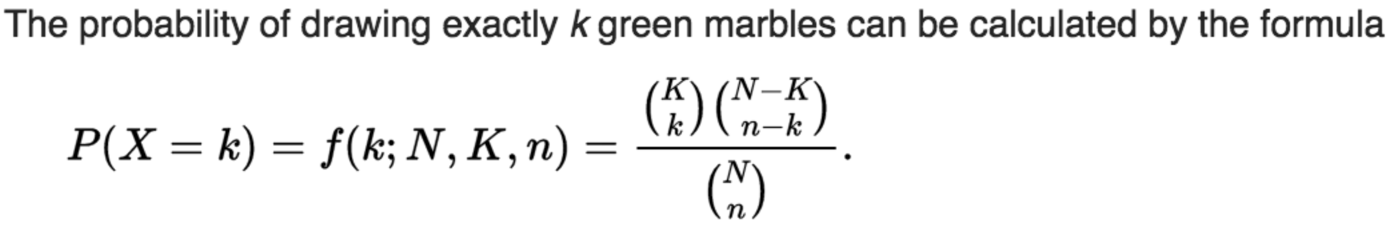
The GO structure enables computations such as exploring function enrichment among sets of transcripts.

Gene ontology functional enrichment

	(+) Differentially Expressed	(-) Not Differentially Expressed	Totals
+ Gene Ontology	50	200	250
- Gene Ontology	1950	17800	19750
Totals	2000	18000	20000



	drawn	not drawn	total
green marbles	k	$K - k$	K
red marbles	$n - k$	$N + k - n - K$	$N - K$
total	n	$N - n$	N



The probability of drawing exactly *k* green marbles can be calculated by the formula

$$P(X = k) = f(k; N, K, n) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

No significant sequence similarity… What else?

```
GGAGCTGGAGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCGTGGTCCT
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA
AAAGACAGCTCCAGTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG
GGCTTGGAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA
TGACCTTGGCCTACGATAATGGCATCAACCTGTTCGATACGGCGGAGGTCTACGCTGCTG
GAAAAGCTGAAGTGGTATTAGGGAACATCATTAAGAAGGGATGGAGACGGTCCAGCC
TTGTCATCACCACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG
ATGTGGTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC
```

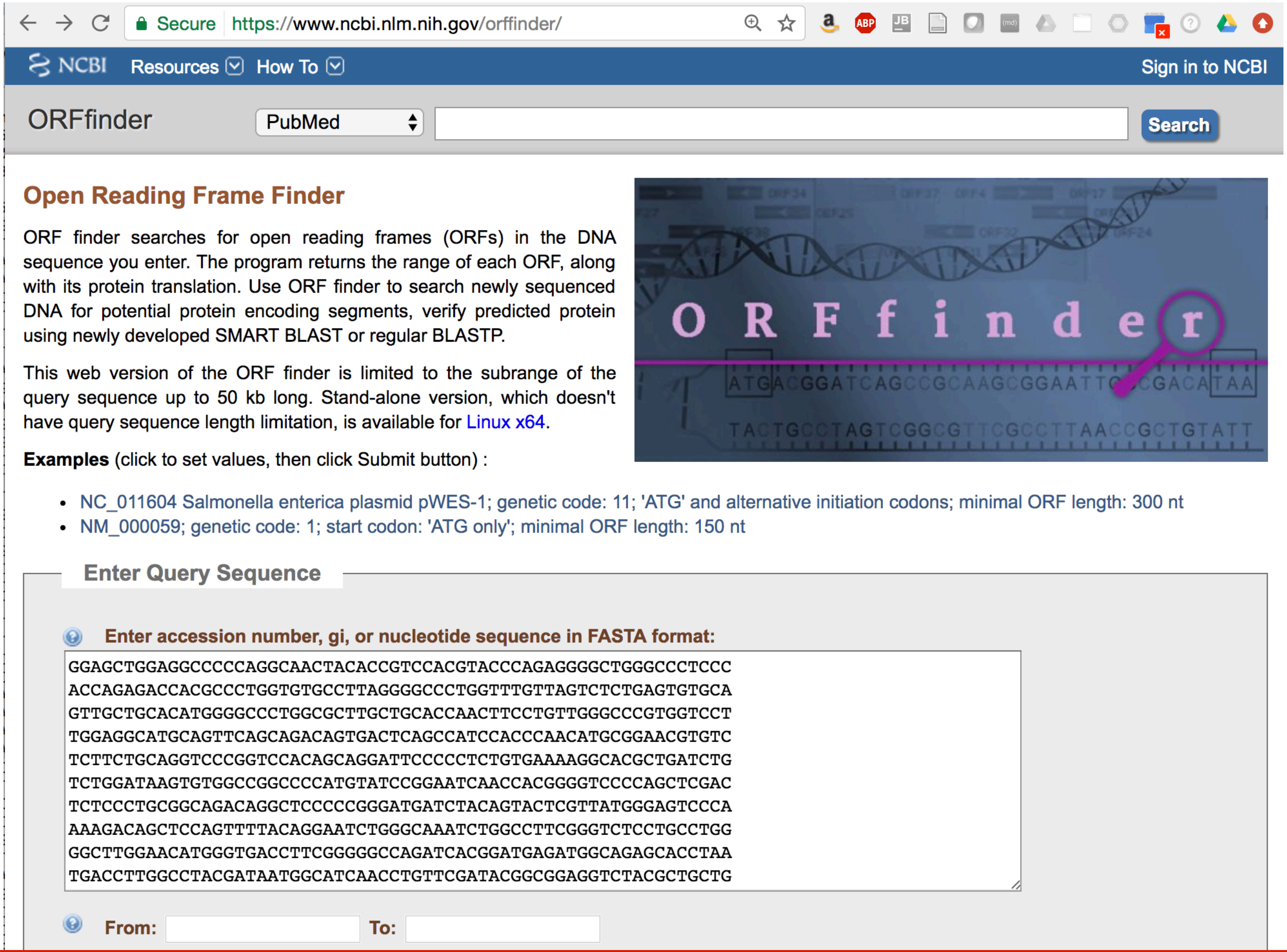
Is there an ORF for a potential Coding Region?

```
GGAGCTGGAGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCGTGGTCCT
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA
AAAGACAGCTCCAGTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG
GGCTTGGAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA
TGACCTTGGCCTACGATAATGGCATCAACCTGTTCGATACGGCGGAGGTCTACGCTGCTG
GAAAAGCTGAAGTGGTATTAGGGAACATCATTAAGAAGAAGGGATGGAGACGGTCCAGCC
TTGTCATCACCACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG
ATGTGGTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC
```

Is there an ORF for a potential Coding Region?

GGAGCTGGAGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCAC**ATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCGTGGTCCT**
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA
AAAGACAGCTCCAGTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG
GGCTTGGAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA
TGACCTTGGCCTACGATAATGGCATCAACCTGTTCGATACGGCGGAGGTCTACGCTGCTG
GAAAAGCTGAAGTGGTATTAGGGAACATCATTAAGAAGAAGGGATGGAGACGGTCCAGCC
TTGTCATCACCACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG
ATGTGGTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC

Find all ORFs using ORFfinder



Secure
https://www.ncbi.nlm.nih.gov/orffinder/
NCBI
Resources
How To
Sign in to NCBI
ORFfinder
PubMed
Search
O R F f i n d e r
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT

Open Reading Frame Finder

ORF finder searches for open reading frames (ORFs) in the DNA sequence you enter. The program returns the range of each ORF, along with its protein translation. Use ORF finder to search newly sequenced DNA for potential protein encoding segments, verify predicted protein using newly developed SMART BLAST or regular BLASTP.

This web version of the ORF finder is limited to the subrange of the query sequence up to 50 kb long. Stand-alone version, which doesn't have query sequence length limitation, is available for Linux x64.

Examples (click to set values, then click Submit button) :

- NC_011604 Salmonella enterica plasmid pWES-1; genetic code: 11; 'ATG' and alternative initiation codons; minimal ORF length: 300 nt
- NM_000059; genetic code: 1; start codon: 'ATG only'; minimal ORF length: 150 nt

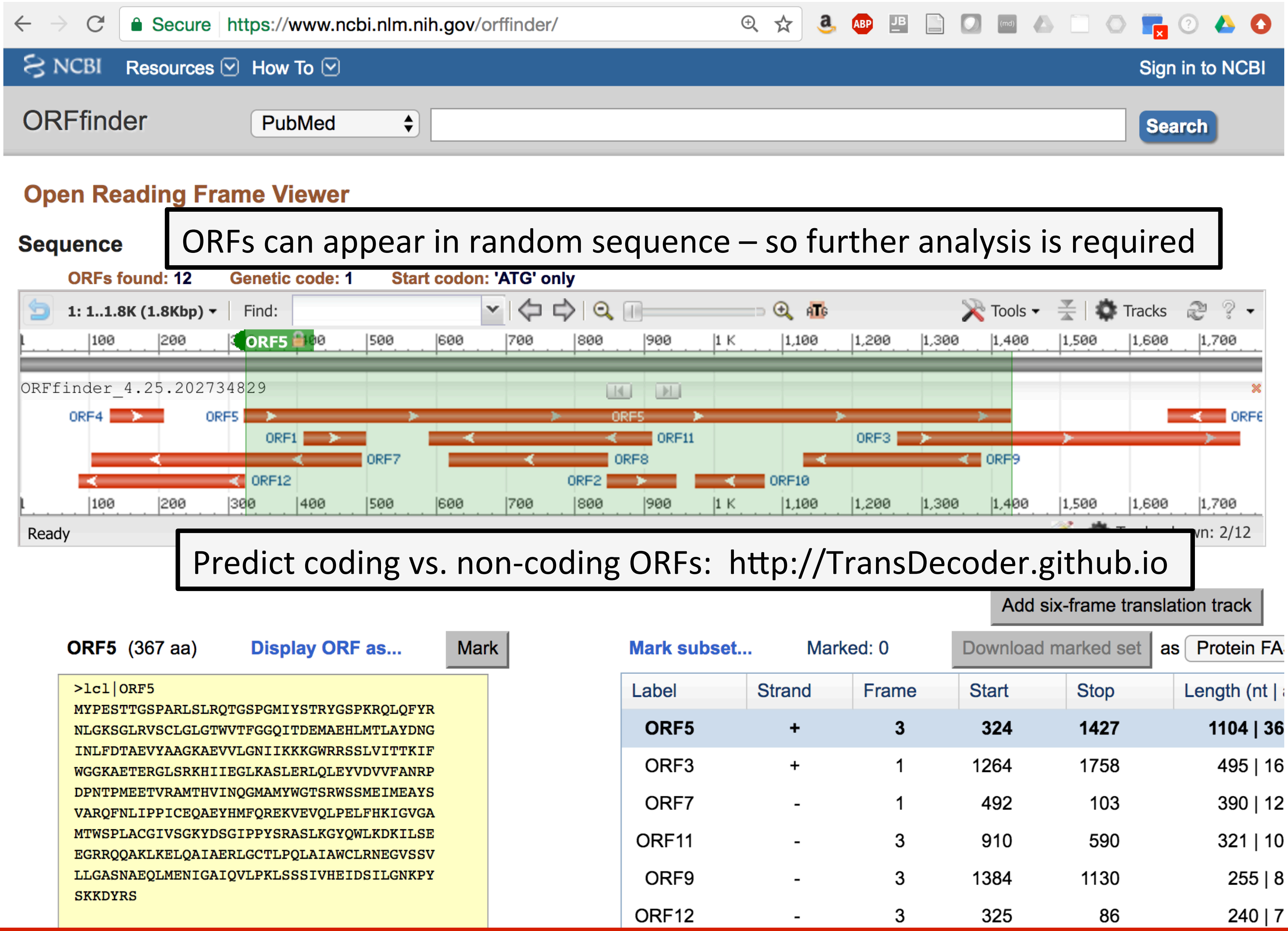
Enter Query Sequence

Enter accession number, gi, or nucleotide sequence in FASTA format:

```
GGAGCTGGAGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCGTGGTCCT
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA
AAAGACAGCTCCAGTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG
GGCTTGGAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA
TGACCTTGGCCTACGATAATGGCATCAACCTGTTCGATACGGCGGAGGTCTACGCTGCTG
```

From: **To:**

ORFfinder finds all open reading frames and provides translations



ORFs found: 12 Genetic code: 1 Start codon: 'ATG' only

ORFs can appear in random sequence – so further analysis is required

Predict coding vs. non-coding ORFs: http://TransDecoder.github.io

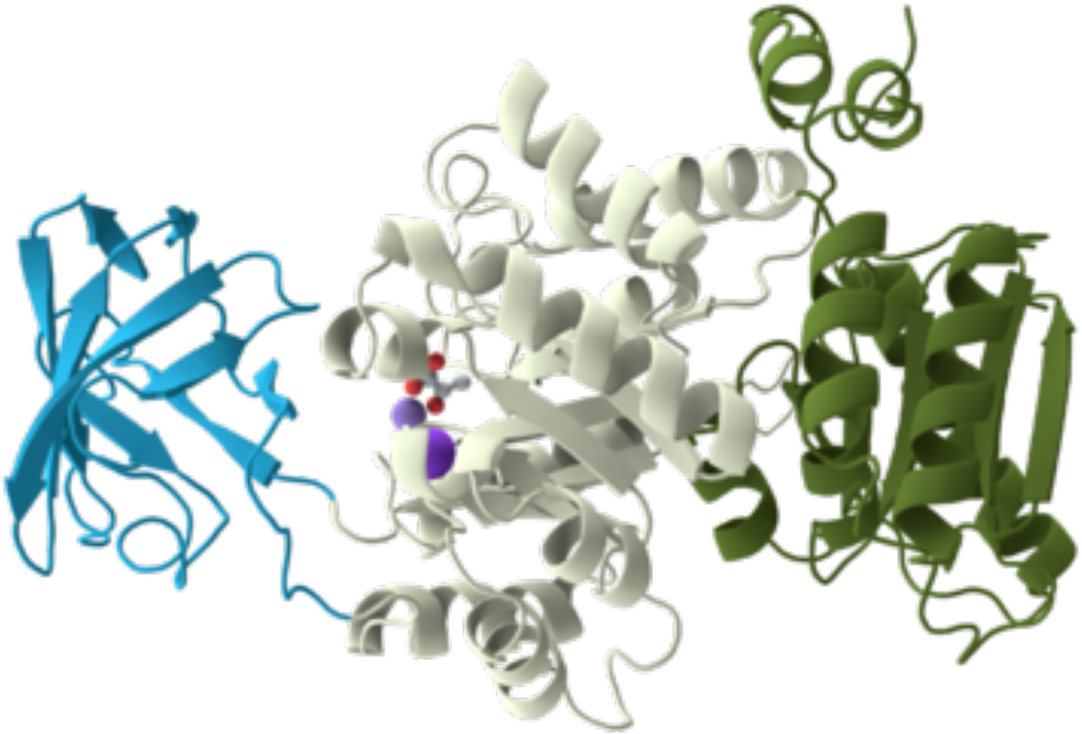
ORF5 (367 aa)

```
>lcl|ORF5
MYPESTTGSPARLSLRQTGSPGMIYSTRYGSPKRQLQFYR
NLGKSGLRVSCLGLGTWVTFGGQITDEMAEHLMTLAYDNG
INLFDTAEVYAAGKAEVVLGNIIKKKGWRRSSLVITTKIF
WGGKAETERGLSRKHIIEGLKASLERLQLEYVDVVFANRP
DPNTPMEETVRAMTHVINQGMAMYWGTSRWSSMEIMEAYS
VARQFNLIPPICEQAEYHMFQREKVEVQLPELFHKIGVGA
MTWSPLACGIVSGKYDSGIPPYSRASLKGYQWLKDKILSE
EGRRQQAKLKELQAIAERLGCTLPQLAIAWCLRNEGVSSV
LLGASNAEQLMENIGAIQVLPKLSSSIVHEIDSILGNKPY
SKKDYRS
```

Marked: 0

Label	Strand	Frame	Start	Stop	Length (nt |
ORF5	**+**	**3**	**324**	**1427**	**1104 | 36**
ORF3	+	1	1264	1758	495 | 16
ORF7	-	1	492	103	390 | 12
ORF11	-	3	910	590	321 | 10
ORF9	-	3	1384	1130	255 | 8
ORF12	-	3	325	86	240 | 7

Can we recognize functional domains in putative coding regions?

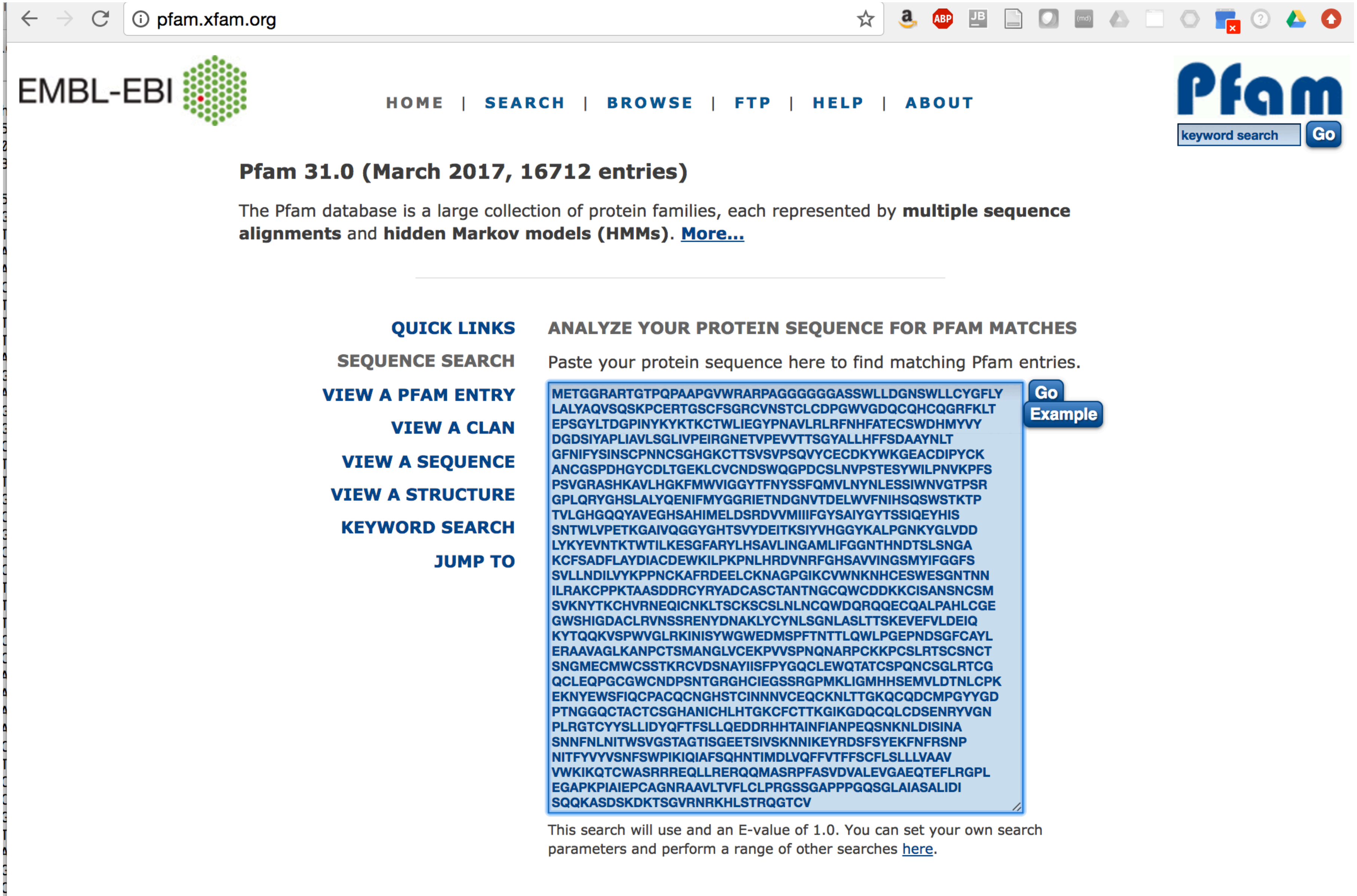


Hints at substrate binding or catalytic activity

DNA, RNA, calcium, phoshate, etc.

Glycoslase, methylase, kinase, nuclease, lipase, protease, etc.

Search the Pfam library of HMMs to identify potential functional domains



Pfam 31.0 (March 2017, 16712 entries)

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

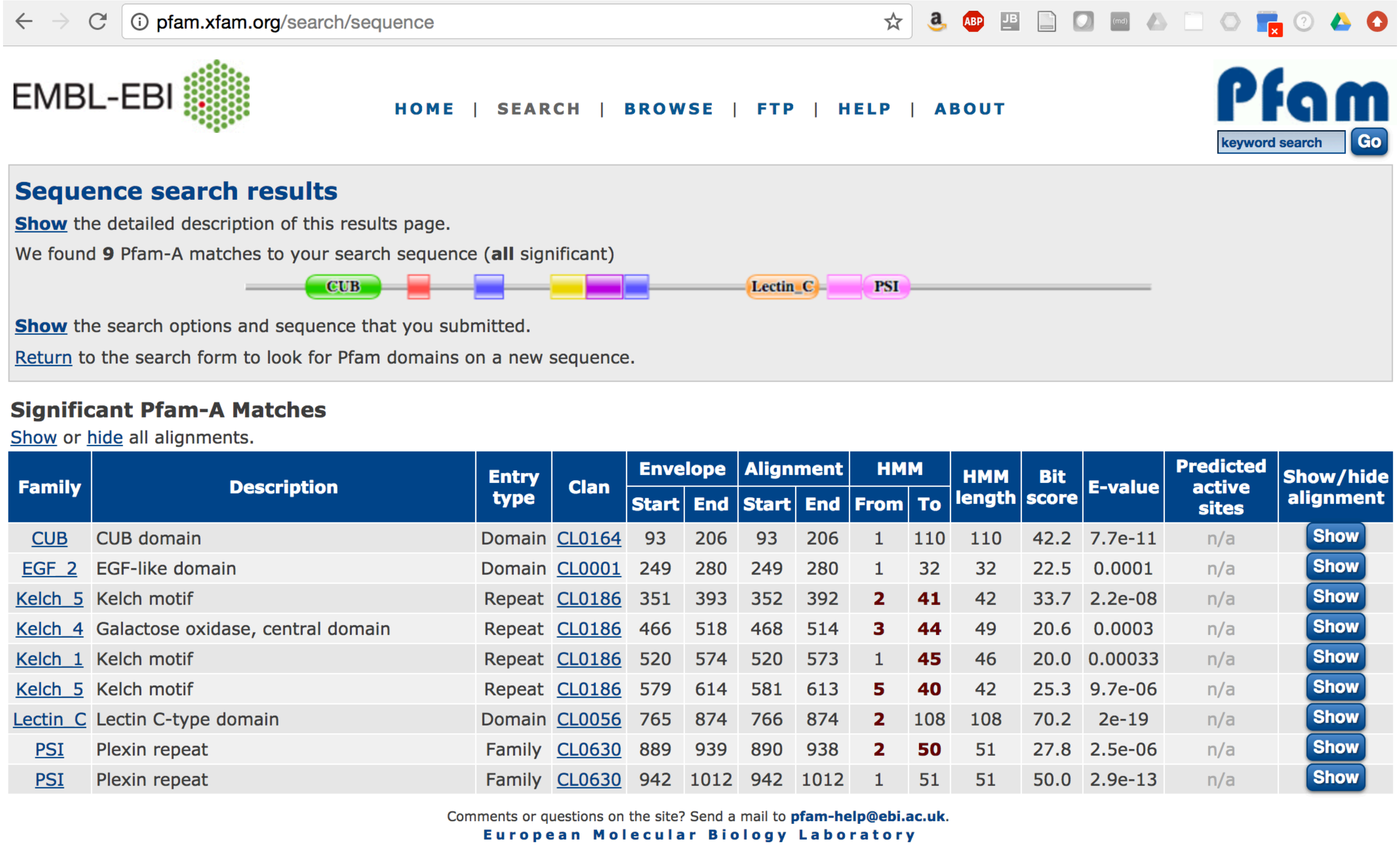
ANALYZE YOUR PROTEIN SEQUENCE FOR PFAM MATCHES

Paste your protein sequence here to find matching Pfam entries.

```
METGGRARTGTPQPAAPGVWRARPAGGGGGGASSWLLDGNSWLLCYGFLY
LALYAQVSQSKPCERTGSCFSGRCVNSTCLCDPGWVGDQCQHCQGRFKLT
EPSGYLTDGPINYKYKTKCTWLIEGYPNAVLRLRFNHFATECSWDHMYVY
DGDSIYAPLIAVLSGLIVPEIRGNETVPEVVTTSGYALLHFFSDAAYNLT
GFNIFYSINSCPNNCSGHGKCTTSVSVPSQVYCECDKYWKGEACDIPYCK
ANCGSPDHGYCDLTGEKLCVCNDSWQGPDCSLNVPSTESYWILPNVKPFS
PSVGRASHKAVLHGKFMWVIGGYTFNYSSFQMVLNYNLESSIWNVGTPSR
GPLQRYGHSLALYQENIFMYGGRIETNDGNVTDELWVFNIHSQSWSTKTP
TVLGHGQQYAVEGHSAHIMELDSRDVVMIIIFGYSAIYGYTSSIQEYHIS
SNTWLVPETKGAIVQGGYGHTSVYDEITKSIYVHGGYKALPGNKYGLVDD
LYKYEVNTKTWTILKESGFARYLHSAVLINGAMLIFGGNTHNDTSLSNGA
KCFSADFLAYDIACDEWKILPKPNLHRDVNRFGHSAVVINGSMYIFGGFS
SVLLNDILVYKPPNCKAFRDEELCKNAGPGIKCVWNKNHCESWESGNTNN
ILRAKCPPKTAASDDRCYRYADCASCTANTNGCQWCDDKKCISANSNCSM
SVKNYTKCHVRNEQICNKLTSCKSCSLNLNCQWDQRQQECQALPAHLCGE
GWSHIGDACLRVNSSRENYDNAKLYCYNLSGNLASLTTSKEVEFVLDEIQ
KYTQQKVSPWVGLRKINISYWGWEDMSPFTNTTLQWLPGEPNDSGFCAYL
ERAAVAGLKANPCTSMANGLVCEKPVVSPNQNARPCKKPCSLRTSCSNCT
SNGMECMWCSSTKRCVDSNAYIISFPYGQCLEWQTATCSPQNCSGLRTCG
QCLEQPGCGWCNDPSNTGRGHCIEGSSRGPMKLIGMHHSEMVLDTNLCPK
EKNYEWSFIQCPACQCNGHSTCINNNVCEQCKNLTTGKQCQDCMPGYYGD
PTNGGQCTACTCSGHANICHLHTGKCFCTTKGIKGDQCQLCDSENRYVGN
PLRGTCYYSLLIDYQFTFSLLQEDDRHHTAINFIANPEQSNKNLDISINA
SNNFNLNITWSVGSTAGTISGEETSIVSKNNIKEYRDSFSYEKFNFRSNP
NITFYVYVSNFSWPIKIQIAFSQHNTIMDLVQFFVTFFSCFLSLLLVAAV
VWKIKQTCWASRRREQLLRERQQMASRPFASVDVALEVGAEQTEFLRGPL
EGAPKPIAIEPCAGNRAAVLTVFLCLPRGSSGAPPPGQSGLAIASALIDI
SQQKASDSKDKTSGVRNRKHLSTRQGTCV
```

This search will use and an E-value of 1.0. You can set your own search parameters and perform a range of other searches here.

Example Pfam report illustrating modular domain architecture



pfam.xfam.org/search/sequence

Sequence search results

Show the detailed description of this results page.

We found **9** Pfam-A matches to your search sequence (**all** significant)

Show the search options and sequence that you submitted.

Return to the search form to look for Pfam domains on a new sequence.

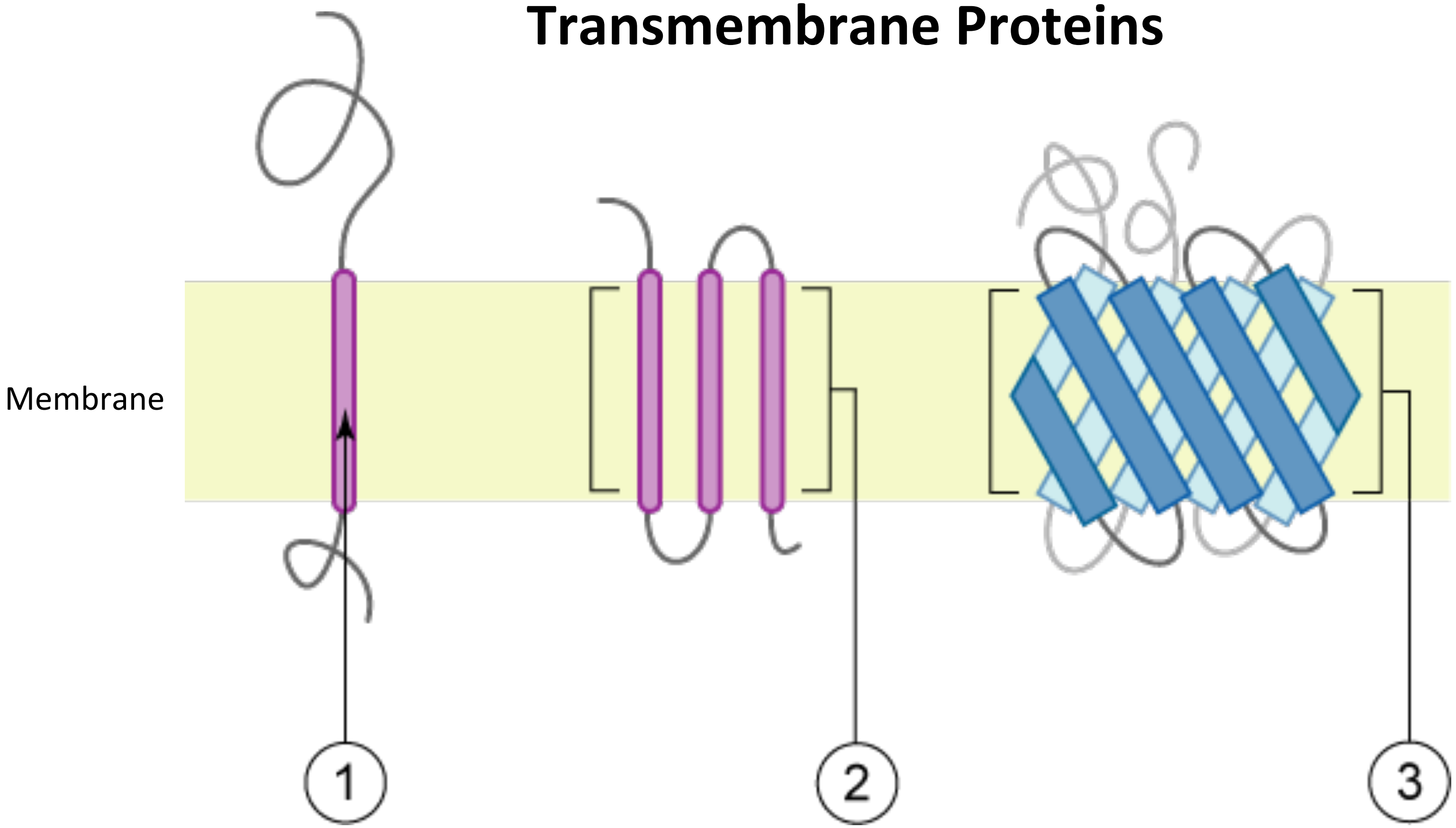
Significant Pfam-A Matches

Show or hide all alignments.

Family	Description	Entry type	Clan	Envelope		Alignment		HMM		HMM length	Bit score	E-value	Predicted active sites	Show/hide alignment
				Start	End	Start	End	From	To					
CUB	CUB domain	Domain	CL0164	93	206	93	206	1	110	110	42.2	7.7e-11	n/a	Show
EGF_2	EGF-like domain	Domain	CL0001	249	280	249	280	1	32	32	22.5	0.0001	n/a	Show
Kelch_5	Kelch motif	Repeat	CL0186	351	393	352	392	2	41	42	33.7	2.2e-08	n/a	Show
Kelch_4	Galactose oxidase, central domain	Repeat	CL0186	466	518	468	514	3	44	49	20.6	0.0003	n/a	Show
Kelch_1	Kelch motif	Repeat	CL0186	520	574	520	573	1	45	46	20.0	0.00033	n/a	Show
Kelch_5	Kelch motif	Repeat	CL0186	579	614	581	613	5	40	42	25.3	9.7e-06	n/a	Show
Lectin_C	Lectin C-type domain	Domain	CL0056	765	874	766	874	2	108	108	70.2	2e-19	n/a	Show
PSI	Plexin repeat	Family	CL0630	889	939	890	938	2	50	51	27.8	2.5e-06	n/a	Show
PSI	Plexin repeat	Family	CL0630	942	1012	942	1012	1	51	51	50.0	2.9e-13	n/a	Show

Comments or questions on the site? Send a mail to **pfam-help@ebi.ac.uk**.
European Molecular Biology Laboratory

Transmembrane Proteins

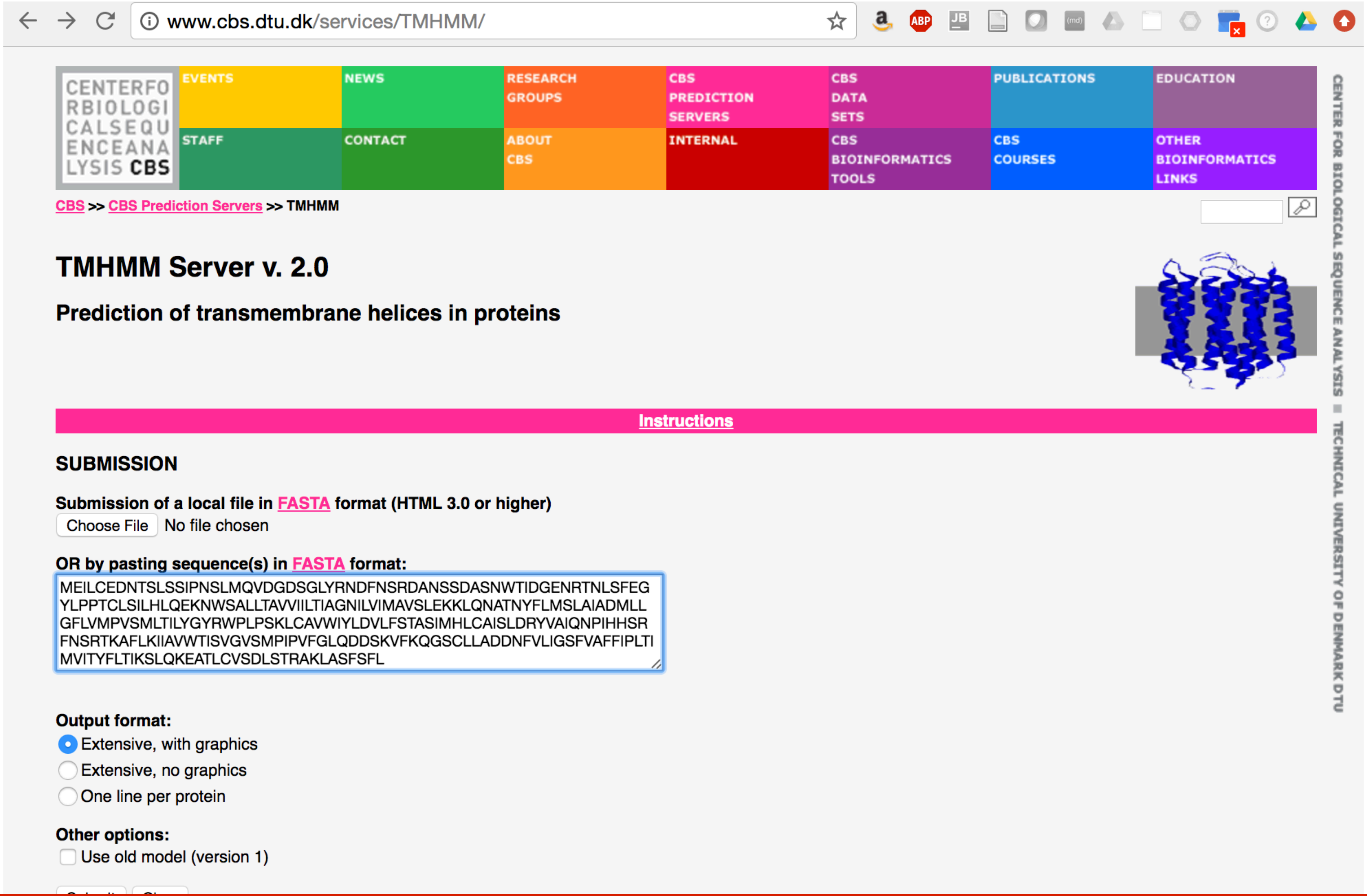


Single transmembrane α-helix (bitopic membrane protein)

Polytopic transmembrane α-helical protein

Polytopic transmembrane β-sheet protein

Using TMHMM to identify putative transmembrane proteins



www.cbs.dtu.dk/services/TMHMM/
CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS
EVENTS
NEWS
RESEARCH GROUPS
CBS PREDICTION SERVERS
CBS DATA SETS
PUBLICATIONS
EDUCATION
STAFF
CONTACT
ABOUT CBS
INTERNAL
CBS BIOINFORMATICS TOOLS
CBS COURSES
OTHER BIOINFORMATICS LINKS
CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS ■ TECHNICAL UNIVERSITY OF DENMARK DTU

CBS >> CBS Prediction Servers >> TMHMM

TMHMM Server v. 2.0

Prediction of transmembrane helices in proteins

Instructions

SUBMISSION

Submission of a local file in FASTA format (HTML 3.0 or higher)

Choose File No file chosen

OR by pasting sequence(s) in FASTA format:

```
MEILCEDNTSLSSIPNSLMQVDGDSGLYRNDFNSRDANSSDASNWTIDGENRTNLSFEG
YLPPTCLSILHLQEKNWSALLTAVVIILTIAGNILVIMAVSLEKKLQNATNYFLMSLAIADMLL
GFLVMPVSMLTILYGYRWPLPSKLCAVWIYLDVLFSTASIMHLCAISLDRYVAIQNPIHHSR
FNSRTKAFLKIIAVWTISVGVSMPIPVFGLQDDSKVFKQGSCLLADDNFVLIGSFVAFFIPLTI
MVITYFLTIKSLQKEATLCVSDLSTRAKLASFSFL
```

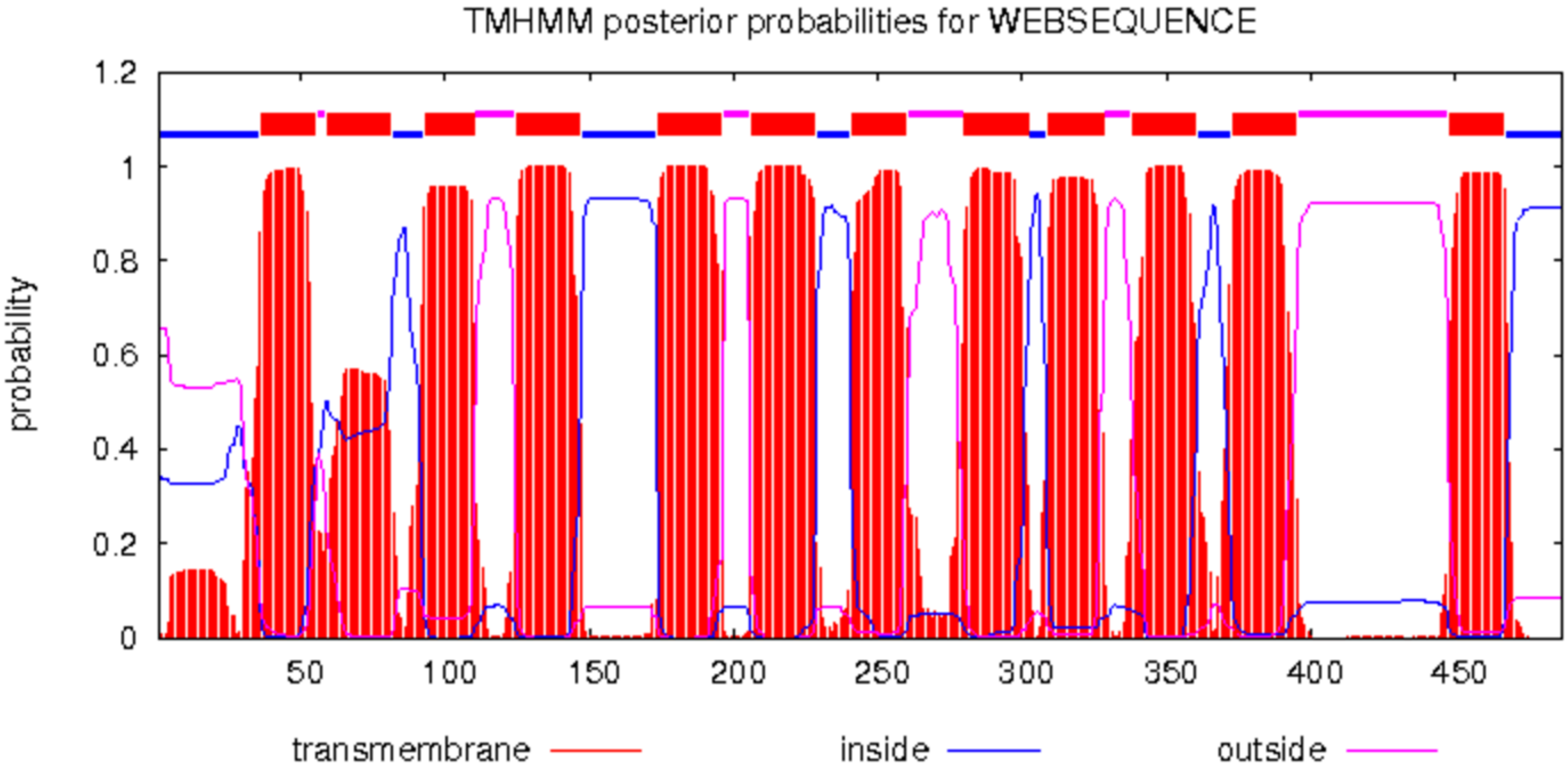
Output format:

● Extensive, with graphics
○ Extensive, no graphics
○ One line per protein

Other options:

□ Use old model (version 1)

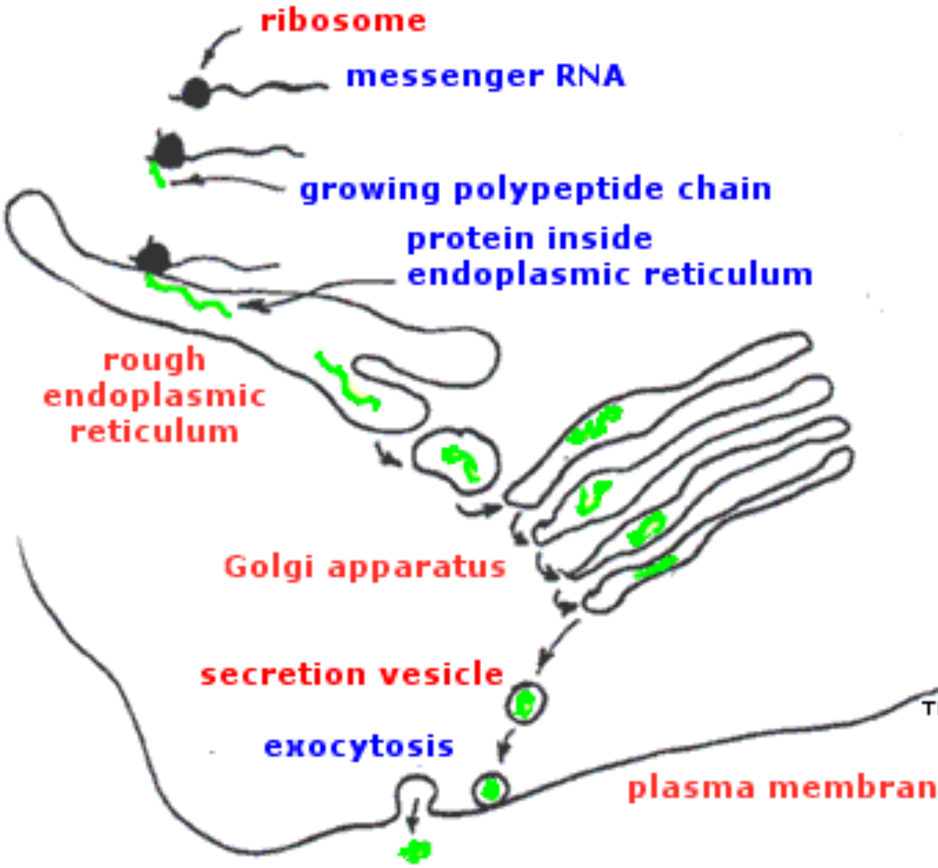
Trans-membrane Domains via TmHMM

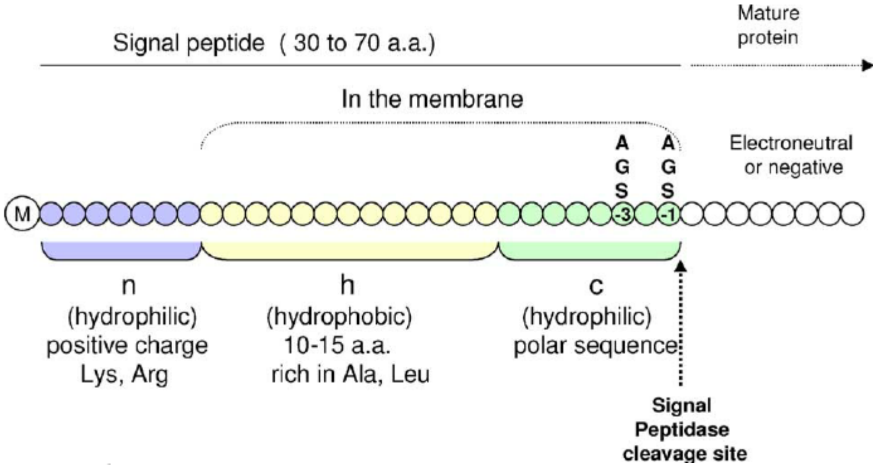


Topology=i36-55o59-81i93-110o125-147i174-196o206-228i241-260o280-302i309-328o338-360i373-395o448-467i

http://www.cbs.dtu.dk/services/TMHMM/

Predicting Secreted Proteins



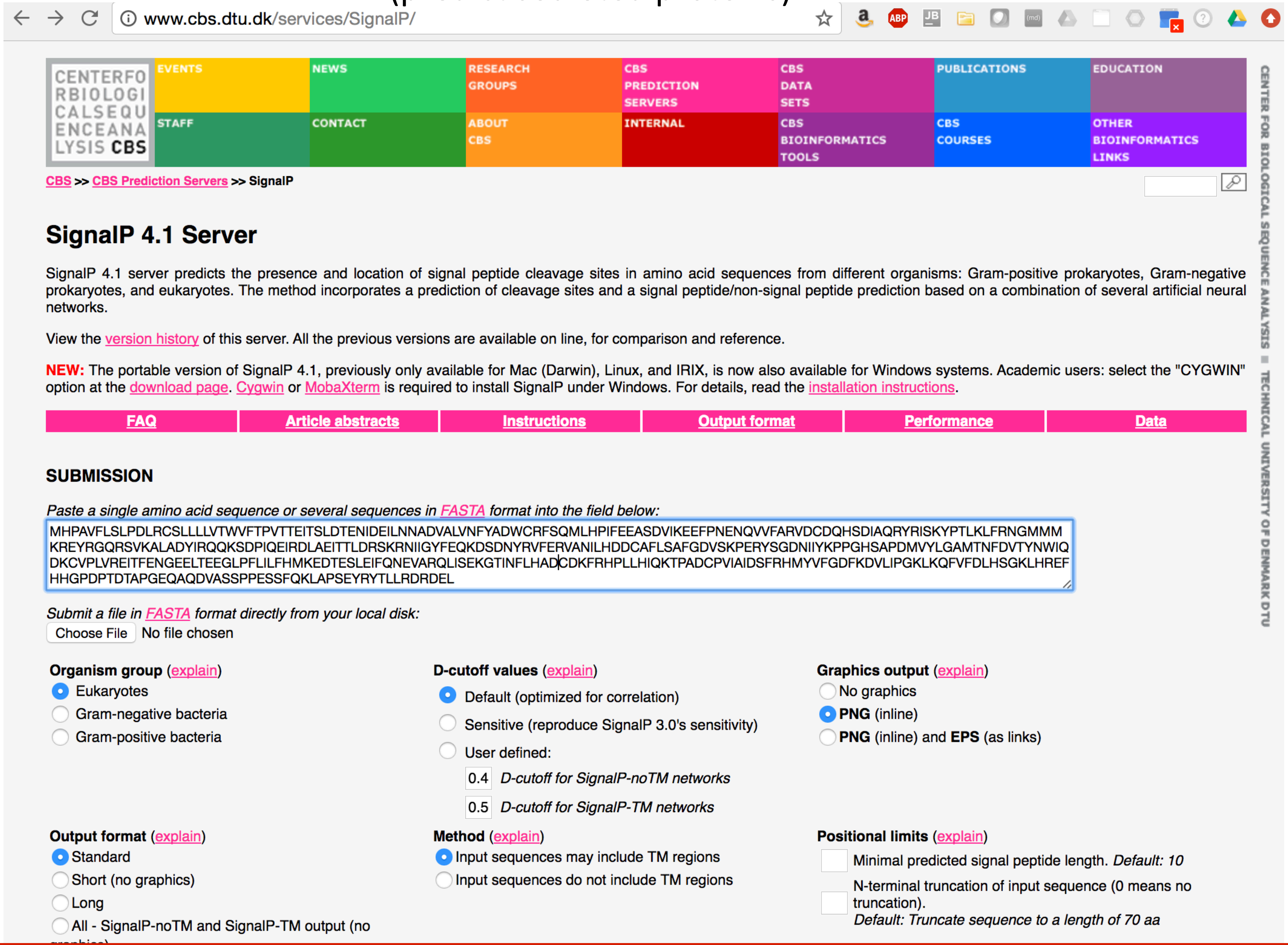


(from: Vaccine 23(15):1770-8)

(from: https://courses.washington.edu/conj/cell/secretion.htm)

SignalP: Prediction of N-terminal signal peptides

(predict secreted proteins)


www.cbs.dtu.dk/services/SignalP/
CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS CBS
EVENTS
NEWS
RESEARCH GROUPS
CBS PREDICTION SERVERS
CBS DATA SETS
PUBLICATIONS
EDUCATION
STAFF
CONTACT
ABOUT CBS
INTERNAL
CBS BIOINFORMATICS TOOLS
CBS COURSES
OTHER BIOINFORMATICS LINKS
CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS ▪ TECHNICAL UNIVERSITY OF DENMARK DTU

CBS >> CBS Prediction Servers >> SignalP

SignalP 4.1 Server

SignalP 4.1 server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms: Gram-positive prokaryotes, Gram-negative prokaryotes, and eukaryotes. The method incorporates a prediction of cleavage sites and a signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks.

View the version history of this server. All the previous versions are available on line, for comparison and reference.

NEW: The portable version of SignalP 4.1, previously only available for Mac (Darwin), Linux, and IRIX, is now also available for Windows systems. Academic users: select the "CYGWIN" option at the download page. Cygwin or MobaXterm is required to install SignalP under Windows. For details, read the installation instructions.

SUBMISSION

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

```
MHPAVFLSLPDLRCSLLLLVTWVFTPVTTEITSLDTENIDEILNNADVALVNFYADWCRFSQMLHPIFEEASDVIKEEFPNENQVVFARVDCDQHSDIAQRYRISKYPTLKLFRNGMMM
KREYRGQRSVKALADYIRQQKSDPIQEIRDLAEITTLDRSKRNIIGYFEQKDSDNYRVFERVANILHDDCAFLSAFGDVSKPERYSGDNIIYKPPGHSAPDMVYLGAMTNFDVTYNWIQ
DKCVPLVREITFENGEELTEEGLPFLILFHMKEDTESLEIFQNEVARQLISEKGTINFLHADCDKFRHPLLHIQKTPADCPVIAIDSFRHMYVFGDFKDVLIPGKLKQFVFDLHSGKLHREF
HHGPDPTDTAPGEQAQDVASSPPESSFQKLAPSEYRYTLLRDRDEL
```

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

Choose File No file chosen

Organism group (explain)

● Eukaryotes
○ Gram-negative bacteria
○ Gram-positive bacteria

D-cutoff values (explain)

● Default (optimized for correlation)
○ Sensitive (reproduce SignalP 3.0's sensitivity)
○ User defined:
0.4 *D-cutoff for SignalP-noTM networks*
0.5 *D-cutoff for SignalP-TM networks*

Graphics output (explain)

○ No graphics
● **PNG** (inline)
○ **PNG** (inline) and **EPS** (as links)

Output format (explain)

● Standard
○ Short (no graphics)
○ Long
○ All - SignalP-noTM and SignalP-TM output (no graphics)

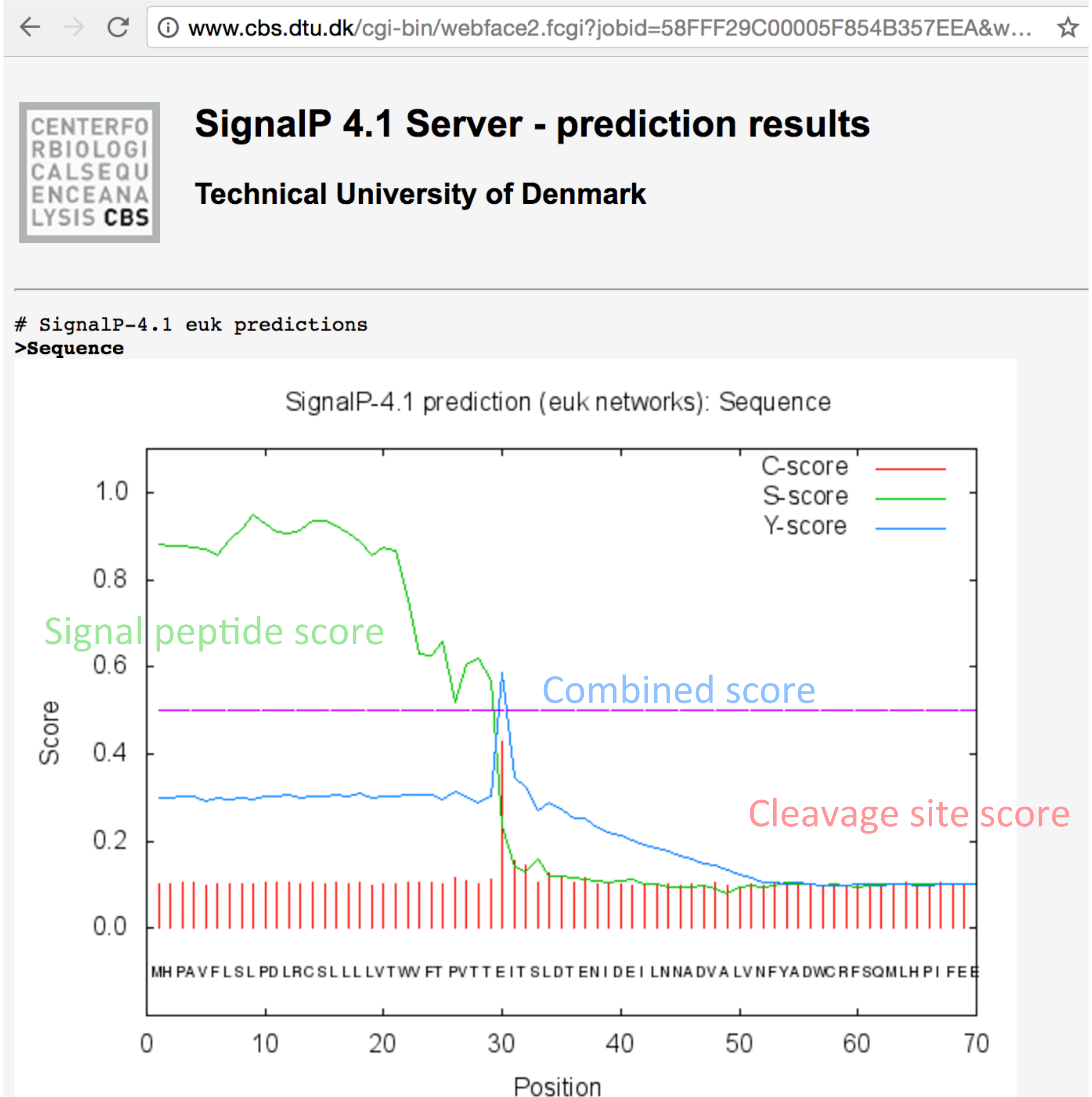
Method (explain)

● Input sequences may include TM regions
○ Input sequences do not include TM regions

Positional limits (explain)

□ Minimal predicted signal peptide length. *Default: 10*
□ N-terminal truncation of input sequence (0 means no truncation).
Default: Truncate sequence to a length of 70 aa

Example SignalP predicted signal peptide



www.cbs.dtu.dk/cgi-bin/webface2.fcgi?jobid=58FFF29C00005F854B357EEA&w...

CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS

SignalP 4.1 Server - prediction results

Technical University of Denmark

```
# SignalP-4.1 euk predictions
>Sequence
```

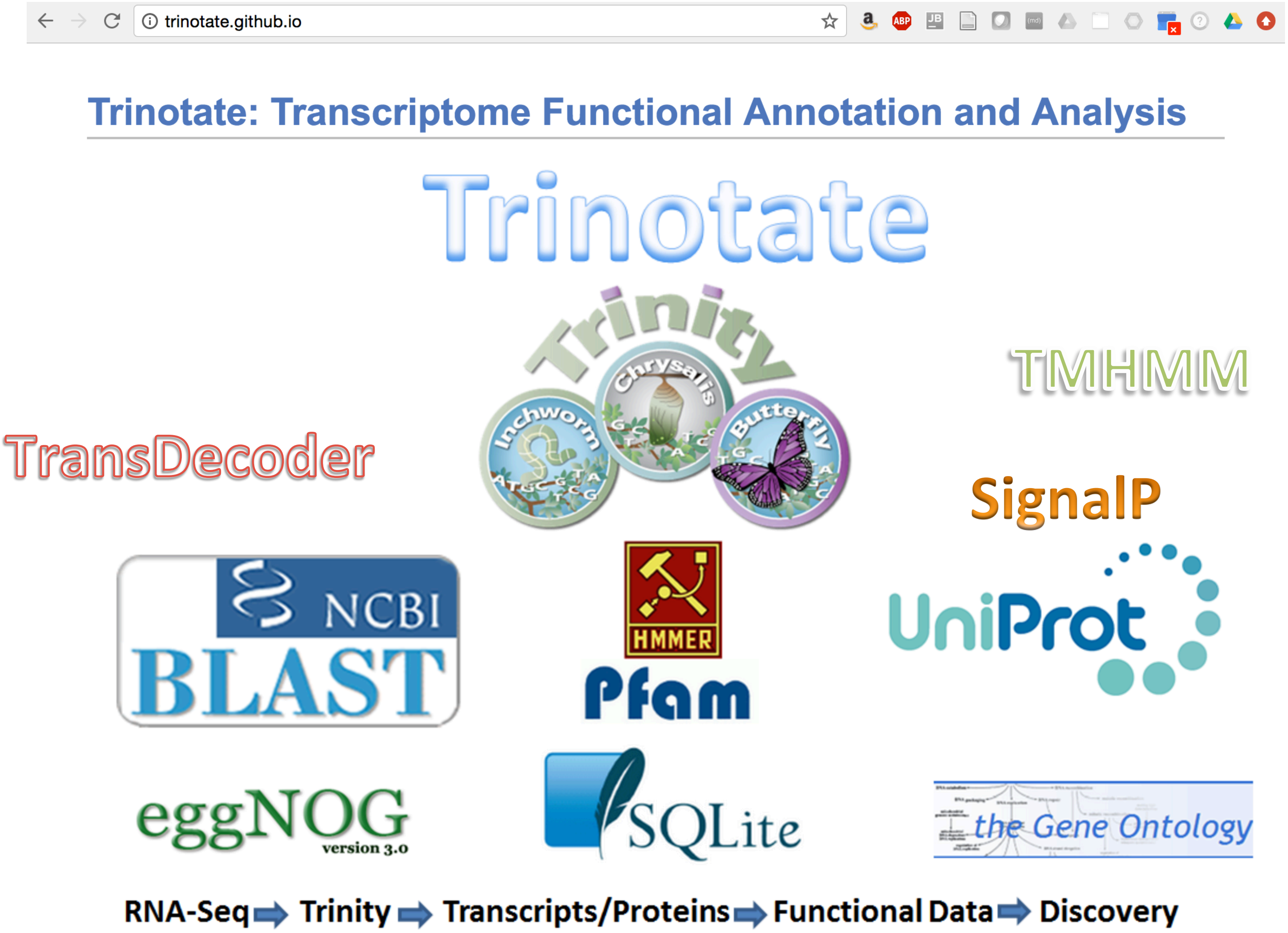
SignalP-4.1 prediction (euk networks): Sequence
C-score
S-score
Y-score
1.0
0.8
0.6
0.4
0.2
0.0
Score
MH PAVFLSLPDLRCSLLLLVTWWFTPVTTEITSLDTENIDEILNNADVALVNFYADWCRFSQMLHPIFEE
0
10
20
30
40
50
60
70
Position

Signal peptide score

Combined score

Cleavage site score

Transcriptome-scale functional annotation using Trinotate



trinotate.github.io

Trinotate: Transcriptome Functional Annotation and Analysis

RNA-Seq → Trinity → Transcripts/Proteins → Functional Data → Discovery

There’s no substitute for experimentally validating protein functions

We are on a Coffee Break & Networking Session