

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io

Supported by



Creative Commons

This page is available in the following languages:

Afrikaans Български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски (latinica) Sotho svenska
中文 華語 (台灣) isiZulu



Attribution-Share Alike 2.5 Canada

You are free:



to **Share** — to copy, distribute and transmit the work



to **Remix** — to adapt the work



Under the following conditions:



Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).



Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.

- For any reuse or distribution, you must make clear to others the licence terms of this work.
- Any of the above conditions can be waived if you get permission from the copyright holder.
- The author's moral rights are retained in this licence.

[Disclaimer](#)

Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
[English](#) [French](#)

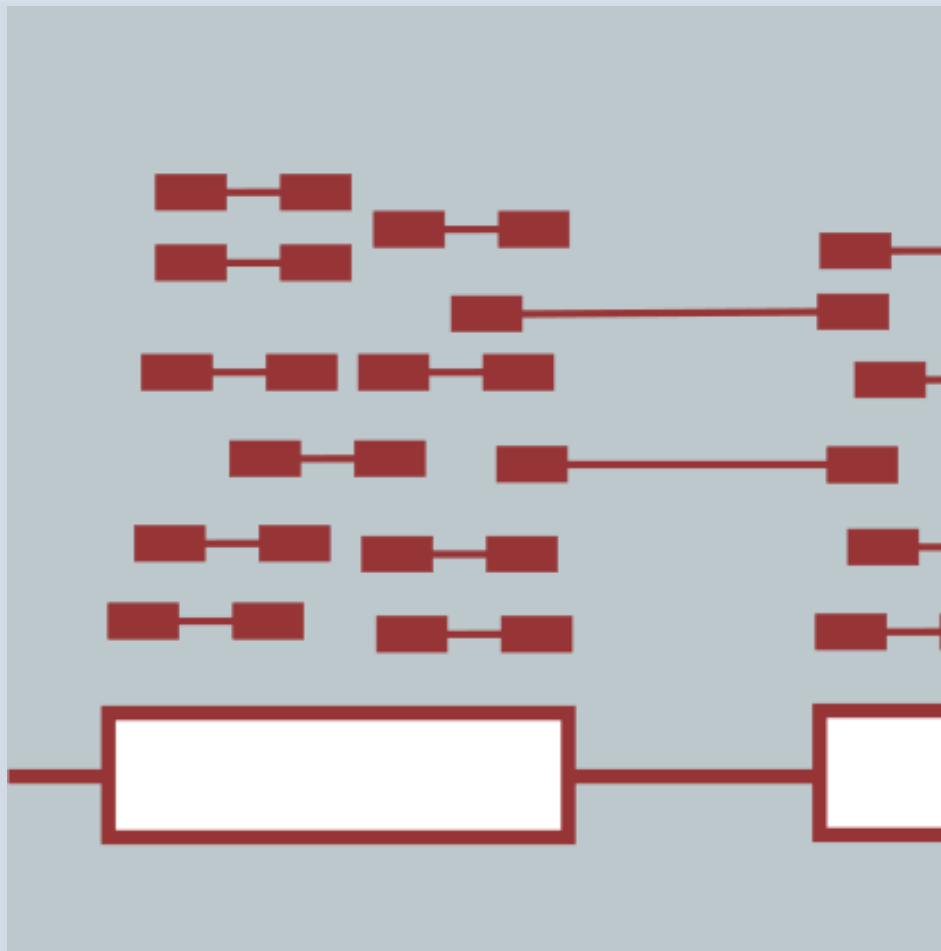
[Learn how to distribute your work using this licence](#)

Functional Annotation and Analysis of Transcripts

Brian Haas

Informatics for RNA-Seq Analysis

June 11-13, 2019



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

GGAGCTGGAGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGGCTGGGCCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGGCCCTGGTTTGTAGTCTCTGAGTGTGCA
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCCGTGGTCCT
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCAGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATTCGAC
TCTCCCTCCCA
AAAGACCTGG
GGCTTCCTAA
TGACCTTGCTG
GAAAACAGCC
TTGTCATTCCA
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG
ATGTGGTTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC

Can we gather hints of biological function
from sequence?

Methods used to predict function from sequence

- Sequence homology

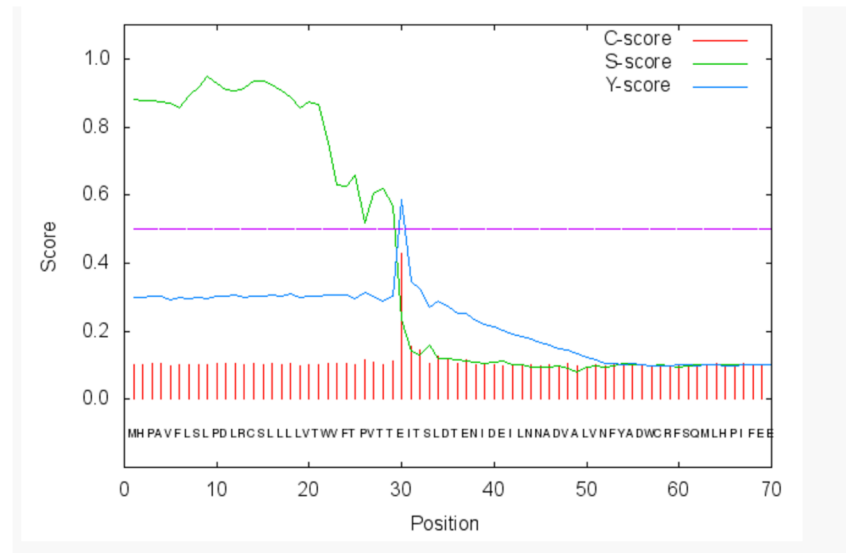
Searching protein database for sequence similarity

```
Query  THVHRPYNEHKSLSGTARYMSINTHLGREQSRDDLESMDGHVFMFLRGSLPW--QGLKA
       T   P + K   GT Y S + HLG   RR DLE +G   L   LPW Q L A
Database Match  TGDFKP-DPKMHNGTIEYTSRDAHLG-VPTRRADLEILGYNLIIEWLGAELPWVTQKLLA
```

- Sequence composition

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

- Neural Networks
- Hidden Markov Models



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

Browser address bar: <https://www.uniprot.org>

UniProtKB Advanced

BLAST Align Retrieve/ID mapping Peptide search Help Contact

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 (560,292)
Manually annotated and reviewed.
Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (158,257,522)
Automatically annotated and not reviewed.
Records that await full manual

(as of June, 2019)

UniRef
Sequence clusters

UniParc
Sequence archive

Proteomes

Supporting data

- Literature citations
- Cross-ref. databases
- Taxonomy
- Diseases
- Subcellular locations
- Keywords

News

Forthcoming changes
There are currently no changes planned

UniProt release 2019_05
Love's Labour (nearly) Lost

UniProt release 2019_04
A pox on your messenger | Removal of the cross-references to HOVERGEN, ProteinModelPortal and UniGene

[News archive](#)

YouTube

Text search
Our basic text search allows you to search all the

UniProt data

Download latest release
Get the UniProt data

Protein spotlight

Twisting Fate
May 2019

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 levelⁱ

Functionⁱ

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

GO - Molecular functionⁱ

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

Complete GO annotation...

GO - Biological processⁱ

- 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...

Keywordsⁱ

Molecular, Kinase, Transferase

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

Gene Ontology: a structured relational vocabulary for describing biological functions

www.ebi.ac.uk/QuickGO/GTerm?id=GO:0030387#te...

QuickGO Search! Web Services Dataset Term Basket: 0

Term Information Ancestor Chart Child Terms Protein Annotation Co-occurring Terms Change Log

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

```
graph BT; F3K[fructosamine-3-kinase activity] --> KA[kinase activity]; KA --> P[phosphorylation]; P --> TPCM[transferase activity, transferring phosphorus-c]; P --> PM[phosphorus metabolic process]; TPCM --> T[transferase activity]; PM --> CMP[cellular metabolic process]; T --> CA[catalytic activity]; CMP --> CP[cellular process]; CMP --> MP[metabolic process]; CA --> MF[molecular function]; CP --> BP[biological process]; MP --> BP;
```

A	Is a	B
A	Part of	B
A	Regulates	B
A	Positively regulates	B
A	Negatively regulates	B
A	Occurs in	B
A	Capable of	B
A	Capable of part of	B

QuickGO - <http://www.ebi.ac.uk/QuickGO>

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?

```
GGAGCTGGAGGCCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGGCTGGGCCCCTCCC  
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGGCCCTGGTTTGTAGTCTCTGAGTGTGCA  
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCCTGGTCCCT  
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC  
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG  
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC  
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA  
AAAGACAGCTCCAGTTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG  
GGCTTGGAACATGGGTGACCTTCGGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA  
TGACCTTGGCCTACGATAATGGCATCAACCTGTTTCGATACGGCGGAGGTCTACGCTGCTG  
GAAAAGCTGAAGTGGTATTAGGGAACATCATTAAAGAAGAAGGGATGGAGACGGTCCAGCC  
TTGTCATCACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA  
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG  
ATGTGGTTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA  
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA  
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG  
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT  
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG  
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT  
TGAAGGACAAGATCCTGAGTGAGGAGGGTTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC
```


Is there an ORF for a potential Coding Region?

```
GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGGCTGGGCCCCTCCC  
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGGCCCTGGTTTGTAGTCTCTGAGTGTGCA  
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCCTGGTTCCT  
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC  
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG  
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC  
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA  
AAAGACAGCTCCAGTTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG  
GGCTTGGAACATGGGTGACCTTCGGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA  
TGACCTTGGCCTACGATAATGGCATCAACCTGTTTCGATACGGCGGAGGTCTACGCTGCTG  
GAAAAGCTGAAGTGGTATTAGGGAACATCATTAAAGAAGAAGGGATGGAGACGGTCCAGCC  
TTGTCATCACCACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA  
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG  
ATGTGGTTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA  
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA  
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG  
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT  
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG  
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT  
TGAAGGACAAGATCCTGAGTGAGGAGGGTTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC
```

Is there an ORF for a potential Coding Region?

GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGGCTGGGCCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGGCCCTGGTTTGTAGTCTCTGAGTGTGCA
GTTGCTGCAC**ATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCCTGGTTCCT**
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA
AAAGACAGCTCCAGTTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG
GGCTTGGAACATGGGTGACCTTCGGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA
TGACCTTGGCCTACGATAATGGCATCAACCTGTTTCGATACGGCGGAGGTCTACGCTGCTG
GAAAAGCTGAAGTGGTATTAGGGAACATCATTAAAGAAGAAGGGATGGAGACGGTCCAGCC
TTGTCATCACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG
ATGTGGTTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT
TGAAGGACAAGATCCTGAGTGAGGAGGGTTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC

Find all ORFs using ORFfinder

← → ↻ Secure <https://www.ncbi.nlm.nih.gov/orffinder/> 

NCBI Resources How To Sign in to NCBI

ORFfinder


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



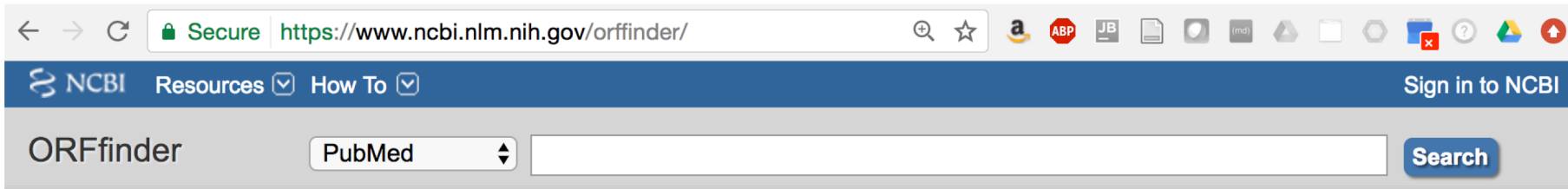
Enter Query Sequence

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

```
GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTAGTCTCTGAGTGTGCA
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCTGTTGGGCCCGTGGTCCT
TGGAGGCATGCAGTTCAGCAGACAGTGACTCAGCCATCCACCCAACATGCGGAACGTGTC
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA
AAAGACAGCTCCAGTTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG
GGCTTGGAAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA
TGACCTTGGCCTACGATAATGGCATCAACCTGTTTCGATACGGCGGAGGTCTACGCTGCTG
```

 **From:** **To:**

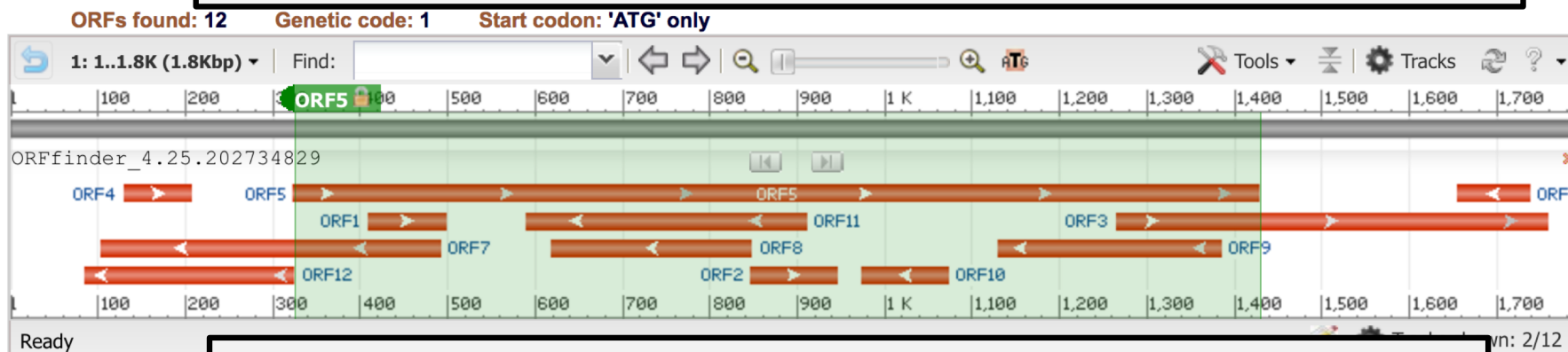
ORFfinder finds all open reading frames and provides translations



Open Reading Frame Viewer

Sequence

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



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

Add six-frame translation track

ORF5 (367 aa)

Display ORF as...

Mark

```
>lcl|ORF5
MYPESTTGSPARLSLRQTGSPGMIYSTRYGSPKRQLQFYR
NLGKSGLRVSLGLGTWVTFGGQITDEMAEHLMTLAYDNG
INLFDTAEVYAAGKAEVVLGNIKKGWRRSSLVITTKIF
WGGKAETERGLSRKHIIIEGLKASLERLQLEYVDVVFANRP
DPNTPMEETVRAMTHVINQGMAMYWGTSRWSSMEIMEAYS
VARQFNLIIPPICEQAEYHMFQREKVEVQLPELPHKIGVGA
MTWSPLACGIVSGKYDSGIPPYSRASLKG YQWLKDKILSE
EGRRQQA KLKELQAI AERLGCTLPQLAIWCLRN EGVSSV
LLGASNAEQLMENIGAIQVLPKLS SIVHEIDSILGNKPY
SKKDYRS
```

Mark subset...

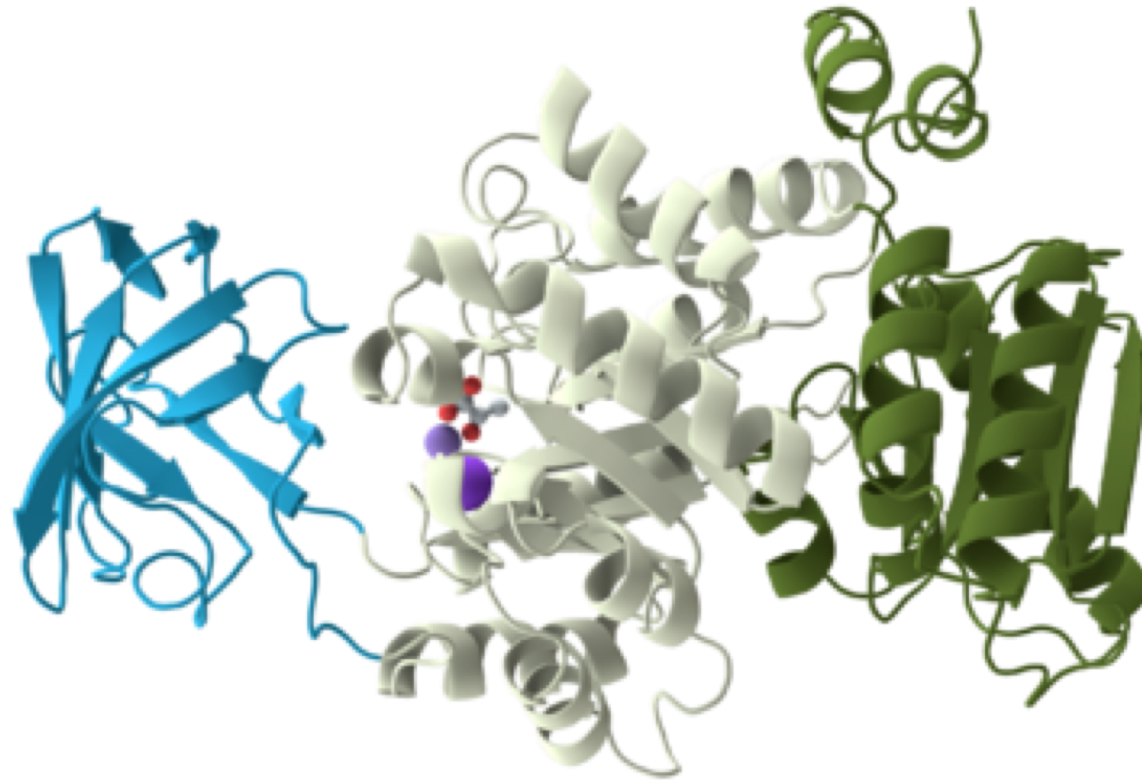
Marked: 0

Download marked set

as Protein FA

Label	Strand	Frame	Start	Stop	Length (nt aa)
ORF5	+	3	324	1427	1104 367
ORF3	+	1	1264	1758	495 165
ORF7	-	1	492	103	390 129
ORF11	-	3	910	590	321 107
ORF9	-	3	1384	1130	255 85
ORF12	-	3	325	86	240 79
ORF8	-	2	848	618	231 77

Can we recognize functional domains in putative coding regions?



Hints at substrate binding or catalytic activity

DNA, RNA, calcium,
phosphate, etc.

Glycosylase, 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...](#)

QUICK LINKS

- [SEQUENCE SEARCH](#)
- [VIEW A PFAM ENTRY](#)
- [VIEW A CLAN](#)
- [VIEW A SEQUENCE](#)
- [VIEW A STRUCTURE](#)
- [KEYWORD SEARCH](#)
- [JUMP TO](#)

ANALYZE YOUR PROTEIN SEQUENCE FOR PFAM MATCHES

Paste your protein sequence here to find matching Pfam entries.

```
METGGRARTGTPQPAAPGVWRARPAGGGGGGASSWLLDGNWLLCYGFLY
LALYAQVQSQKPCERTGSCFSGRCVNSTCLCDPGWVGDCQHCQGRFKLT
EPSGYLTDGPINYYKYTKCTWLIIEGYPNAVLRFRNFATECSWDHMYVY
DGDSIYAPLIAVLSGLVPEIRGNETVPEVTTSGYALLHFFSDAAYNLT
GFNIFYSINCPNNCSGHGKCTTSVSVPSQVYCECDKYWKGEACDIPYCK
ANCGSPDHGYCDLTGEKLCVCNDSWQGPDCSLNVPSTESYWILPNVKPFS
PSVGRASHKAVLHGKFMWVIGGYTFNYSSFQMLVNLNLESSIWNVGTPSR
GPLQRYGHSLALYQENIFMYGGRIETNDGNVTDLWVFNHSQSWSTKTP
TVLGHGQQYAVEGHSAHIMELDSRDVVMIIIFGYSAYIGYTSSIQEYHIS
SNTWLVPETKGAIVQGGYGHSTSVYDEITKSIYVHGGYKALPGNKYGLVDD
LYKYEVTKTWTILKESGFARYLHSAVLINGAMLIFGGNTHNDTSLNGA
KCFSAFLAYDIACDEWKILPKPNLHRDVRNRFHSAVINGSMYIFGGFS
SVLLNDILVYKPPNCKAFRDEELCKNAGPGIKCVWNKNHCSWESGNTNN
ILRAKCPPKTAASDDRCYRYADCASCANTNGCQWCDDKCCISANSNCM
SVKNYTKCHVRNEQICNKLTSCKSCSLNLCQWDQRQEQCALPAHLCGE
GWSHIGDACLRNVSSRENYDNAKLYCYNLSGNLASLTSKEVEFVLDEIQ
KYTQKQVSPWGLRKINISYWGVEDMSPFTNTTLQWLPGEPNDSGFCAYL
ERAAVAGLKANPCTSMANGLVCEKPVVSPNQARPCKKPCSLRTSCSNCT
SNGMECMWCSSTKRCDVSNAYIIFPYGQCLEWQTATCSPQNCGLRTCG
QCLEQPGCGWCNDPSNTGRGHCIEGSSRGPMKLGIMHHSEMVLDTNCPK
EKNYEWSFIQCPACQCNHSTCINNVCQCKNLTGKQCQDCMPGYG
PTNGGQCTACTCSGHANICHLHTGKCFCTTKGIKGDQCQLCDSERNRYGN
PLRGTCYYSLLIDYQFTFSLQEDDRHHTAINFIANPEQSNNLDISINA
SNNFNLNITWSVSGTAGTISGEETSIVSKNNIKEYRDSFSYEFNFRSNP
NITFYVYVSNFSWPIKIQIAFSQHNTIMDLVQFFVTFSCFLSLLVAAV
VWKIKQTCWASRRRQQLLRERQMQASRPFASVDVALEVGAEQTEFLRGL
EGAPKPIAIEPCAGNRAAVLTVFLCLPRGSSGAPPPGQSGLAIASALIDI
SQQKASDSKDKTSGVRNRKHLSTRQGTCV
```

Go
Example

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



[HOME](#) | [SEARCH](#) | [BROWSE](#) | [FTP](#) | [HELP](#) | [ABOUT](#)



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.

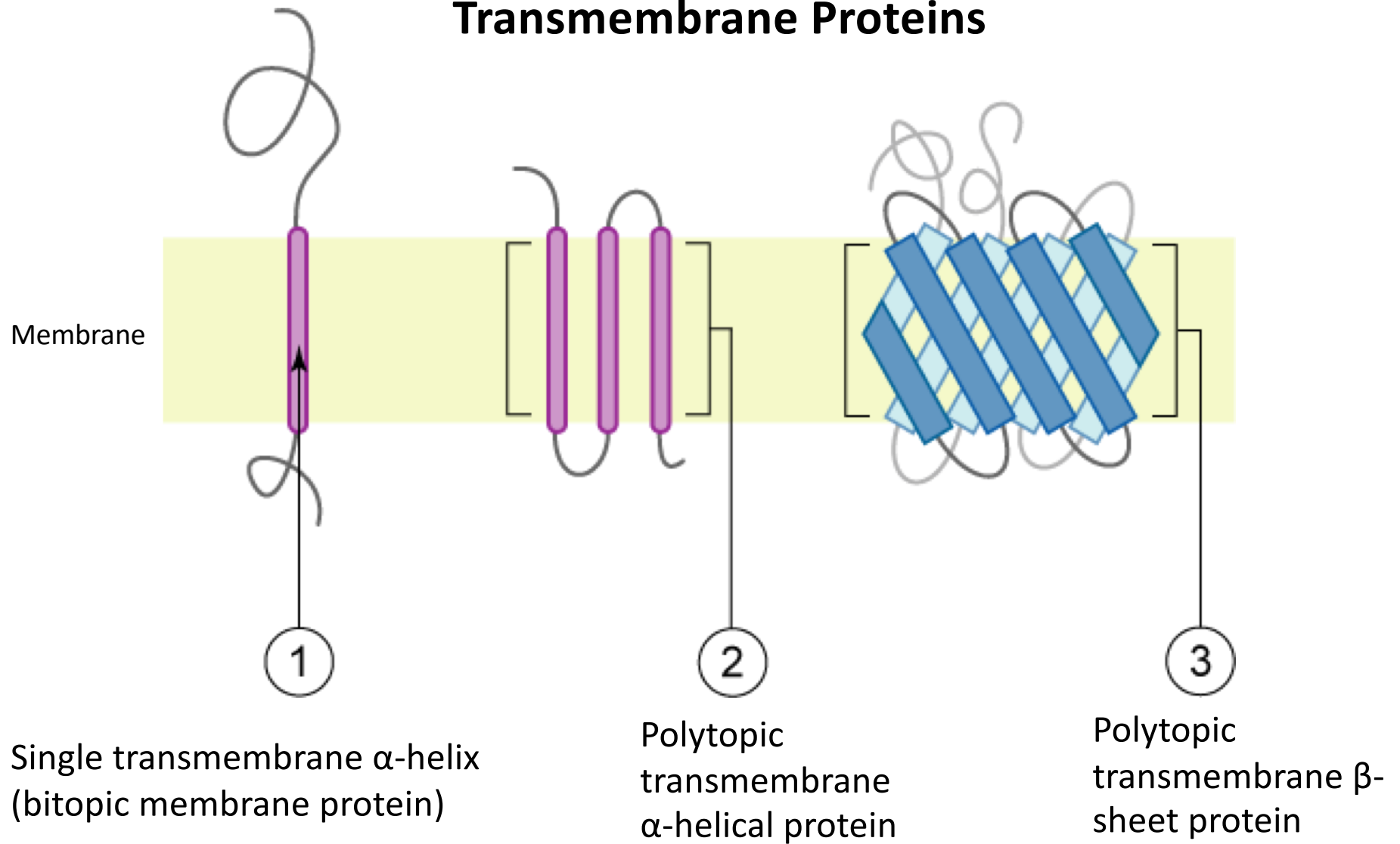
Significant Pfam-A Matches

[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

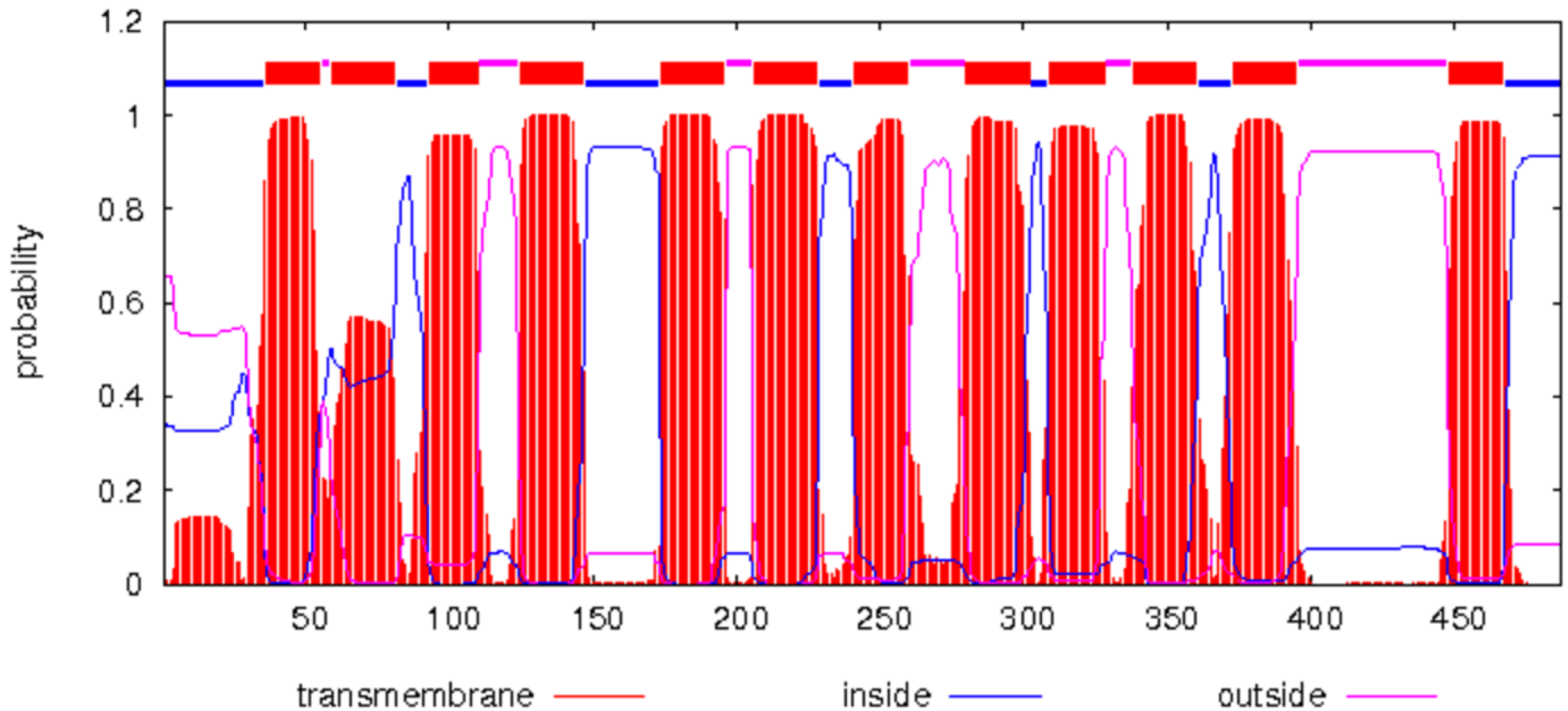
Comments or questions on the site? Send a mail to pfam-help@ebi.ac.uk.
European Molecular Biology Laboratory

Transmembrane Proteins



Trans-membrane Domains via TmHMM

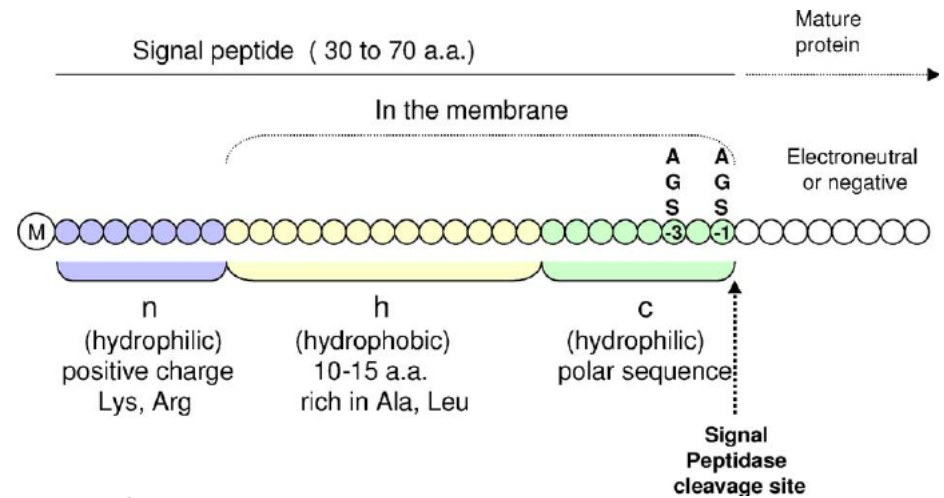
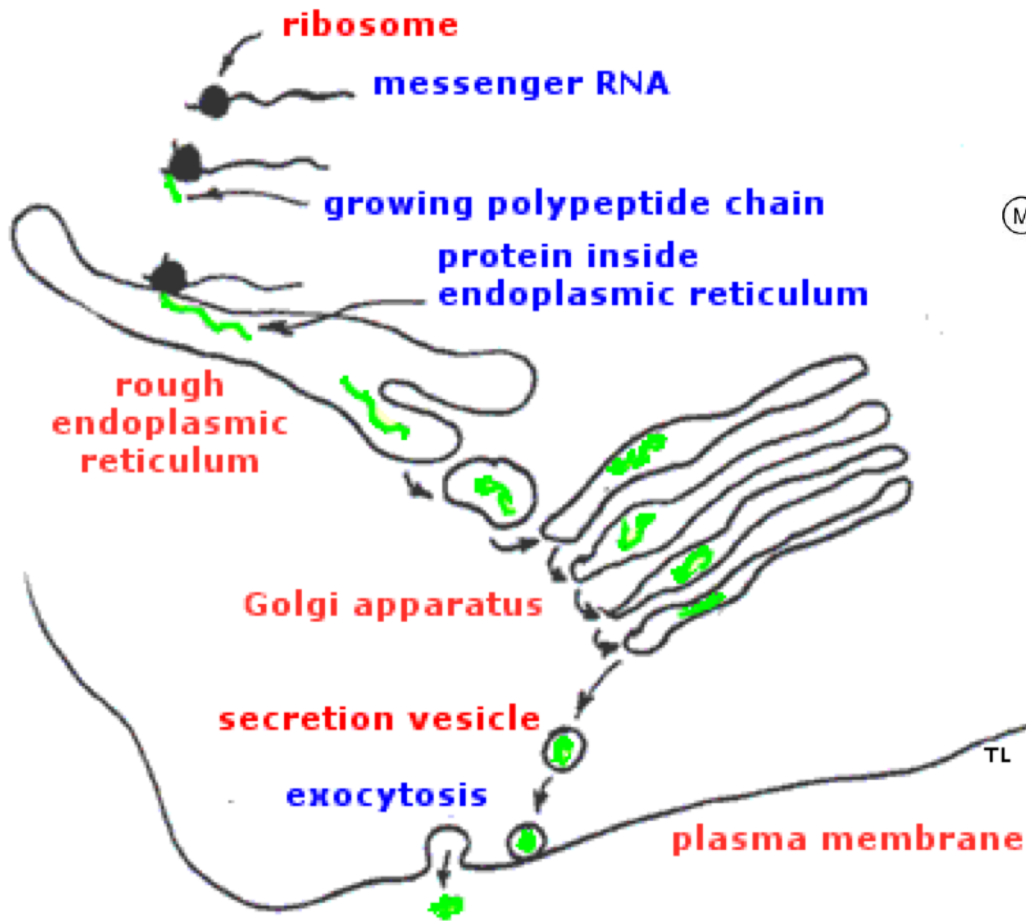
TMHMM posterior probabilities for WEBSEQUENCE



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/ ☆

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

[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](#).

[FAQ](#)[Article abstracts](#)[Instructions](#)[Output format](#)[Performance](#)[Data](#)

SUBMISSION

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

```
MHPAVFLSLPDLRCSLLLLLVTVVFTPVVTEITSLDTENIDEILNNADVALVNFYADWCRFSQMLHPFEEASDVIKEEFPNENQVVFARVDCDQHSDIAQRYRISKYPTLKLFRNGMMM  
KREYRGQRSVKALADYIRQQKSDPIQEIRDLAEITLDRSKRNIIGYFEQKSDNYRVFERVANILHDDCAFLSAFGDVSKPERYSGDNIYKPPGHSAAPDMVYLGAMTFDVTYNWIQ  
DKCVPLVREITFENGEELTEEGLPFLILFHMKEDTESLEIFQNEVARQLISEKGTINFLHADCDKFRHPLLHIQKTPADCPVIAIDSRHMYVFGDFKDVLPKGLKQFVFDLHSGKLHREF  
HHGPDPTDAPGEQAQDVASSPPESSFQKLAPSEYRYTLLRDRDEL
```

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

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:

D-cutoff for SignalP-noTM networks

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

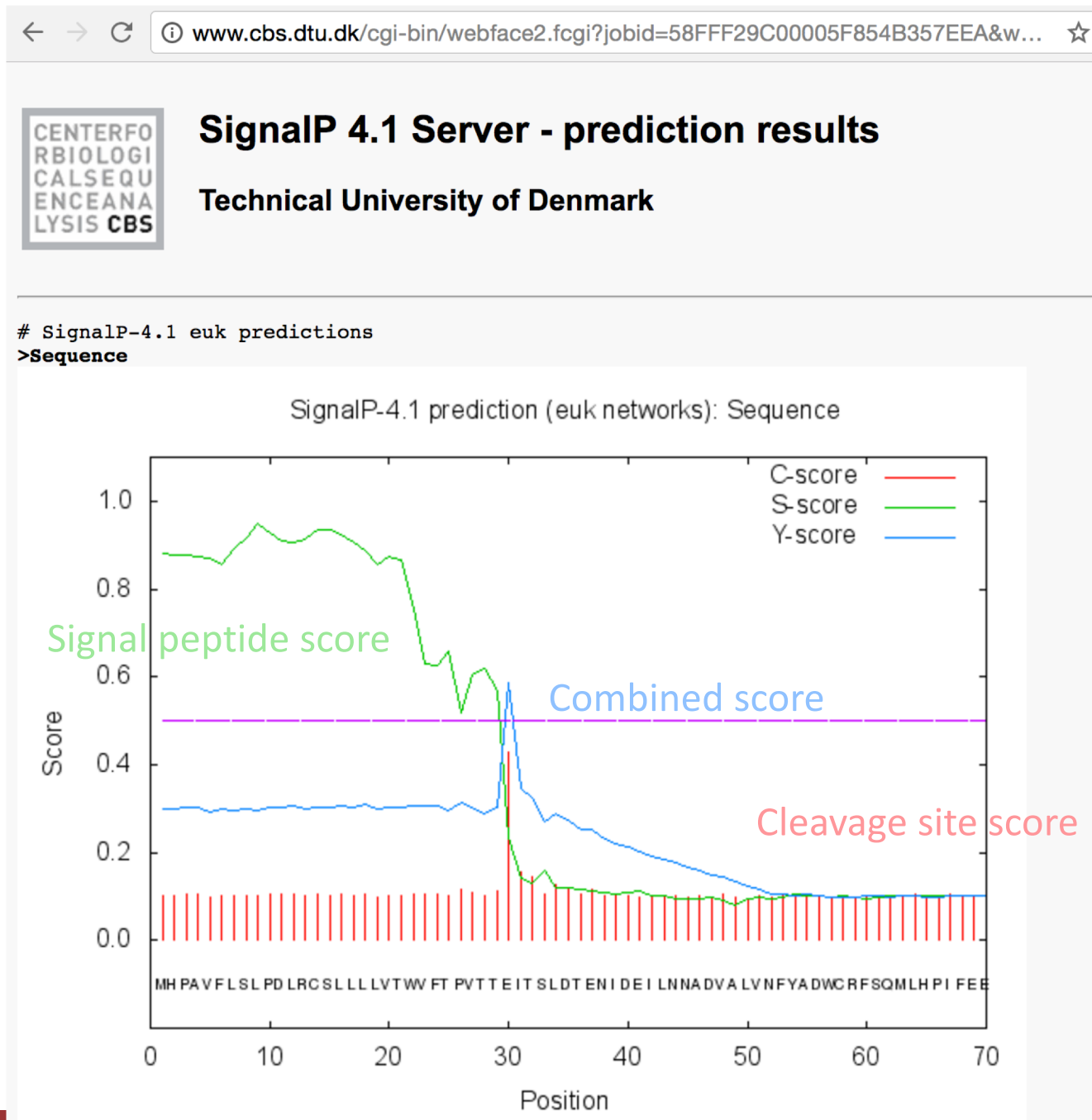
Input sequences do not include TM regions

Positional limits ([explain](#))

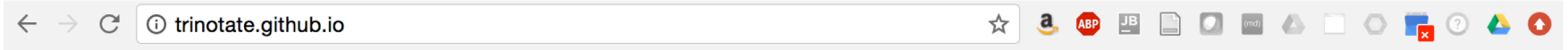
Minimal predicted signal peptide length. *Default: 10*

N-terminal truncation of input sequence (0 means no truncation).
Default: Truncate sequence to a length of 70 aa

Example SignalP predicted signal peptide



Transcriptome-scale functional annotation using Trinotate



Trinotate: Transcriptome Functional Annotation and Analysis

Trinotate

TransDecoder



TMHMM

SignalP



Pfam



eggNOG
version 3.0



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

There's no substitute for experimentally validating protein functions



Transcriptome Assembly is Just the End of the Beginning...

NATURE PROTOCOLS | PROTOCOL

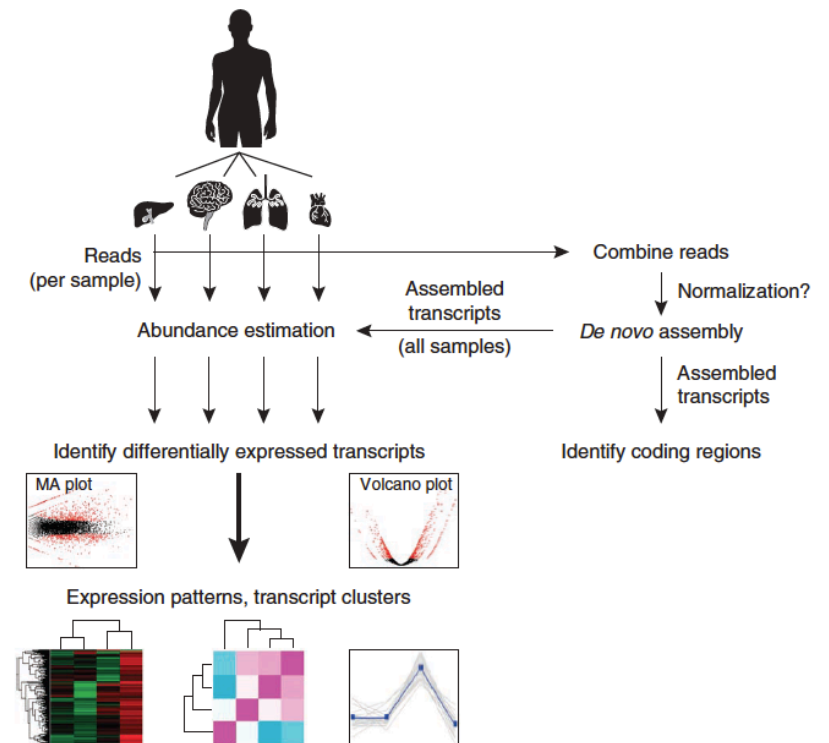
De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis

Brian J Haas, Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, Matthias Lieber, Matthew D MacManes, Michael Ott, Joshua Orvis, Nathalie Pochet, Francesco Strozzi, Nathan Weeks, Rick Westerman, Thomas William, Colin N Dewey, Robert Henschel, Richard D LeDuc, Nir Friedman & Aviv Regev

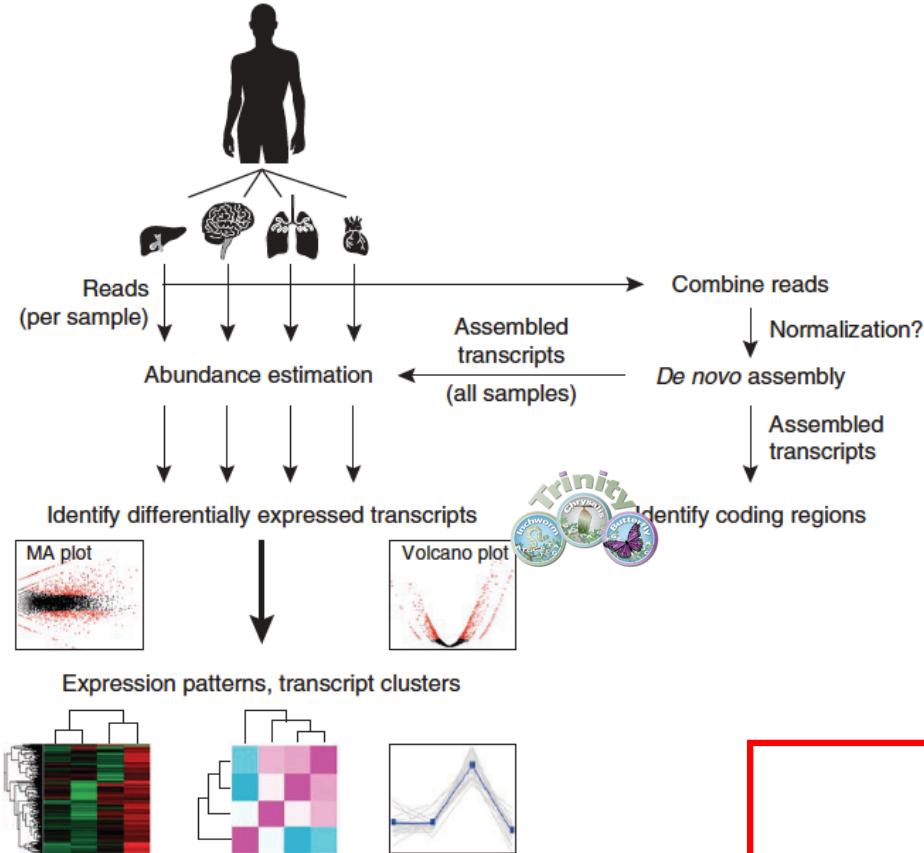
Affiliations | Contributions | Corresponding authors

Nature Protocols 8, 1494–1512 (2013) | doi:10.1038/nprot.2013.084

Published online 11 July 2013



Trinity Framework for De novo Transcriptome Assembly and Analysis



Bioconductor,
& Trinity

Trinotate

TrinotateWeb using **canvaXpress**

The screenshot displays three main plots: a volcano plot on the left, a scatter plot in the middle, and a heatmap on the right. A legend indicates significant up-regulated (red) and down-regulated (blue) genes.

Trinotate Functional Annotation Lab



We are on a Coffee Break & Networking Session

Workshop Sponsors:



Canadian Centre for
Computational
Genomics

