

# Transcript Functional Annotation

GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC  
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA  
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCCTGGTCCCT  
TGGAGGCATGCAGTTCAGCAGACAGTGAAGTCAAGCCATCCACCCAACATGCGGAACGTGTC  
TCTTCTGCAGGTCCCAGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG  
TCTGGA TCGAC  
TCTCC TCCA  
AAAGAC CCTGG  
GGCTTC CCTAA  
TGACCT TGCTG  
GAAAA CAGCC  
TTGTC TCCA  
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG  
ATGTGGTTTTTGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA  
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA  
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG  
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT  
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG  
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT  
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC  
AGGCCATTGCCGAACGCCTGGGCTGCACCCTACCCAGCTGGCCATAGCCTGGTGCCTGA  
GGAATGAGGGTGTGTCAGCTCCGTGCTTCTGGGTGCTTCCAATGCAGAACAACCTTATGGAGA

Can we gather hints of biological function  
from sequence?

# Methods used to predict function from sequence

- Sequence homology

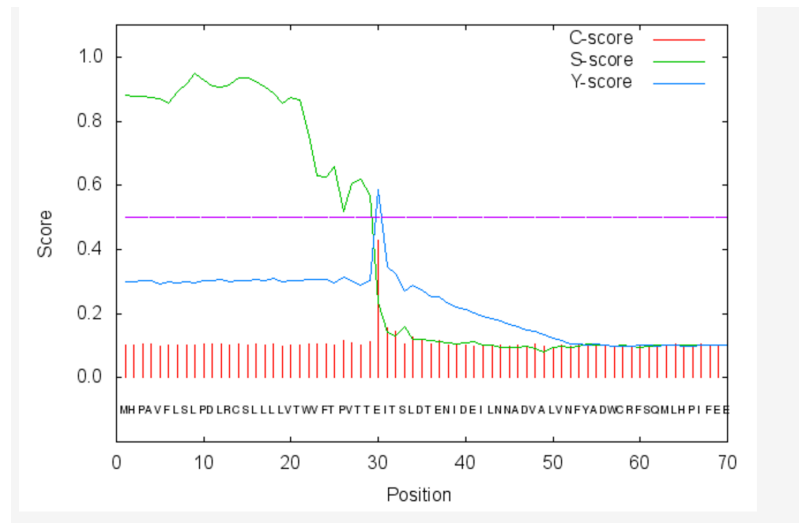
Searching protein database for sequence similarity

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

- 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



## 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...](#)

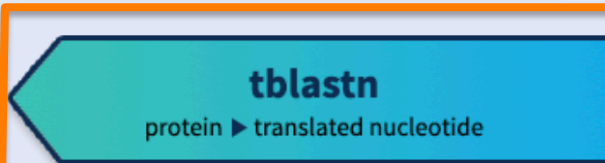
## Web BLAST



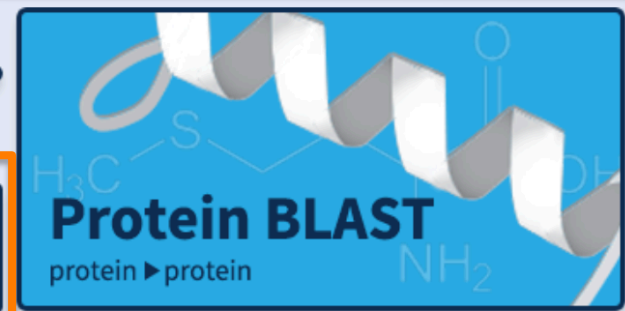
**Nucleotide BLAST**  
nucleotide ► nucleotide



**blastx**  
translated nucleotide ► protein



**tblastn**  
protein ► translated nucleotide



**Protein BLAST**  
protein ► protein

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

Secure | <https://www.uniprot.org>

UniProtKB  Advanced Search

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 (557,275)  
Manually annotated and reviewed.

TrEMBL (114,759,640)  
Automatically

(as of May, 2018)

**UniRef**  
Sequence clusters

**UniParc**  
Sequence archive

**Proteomes**

**Supporting data**

Literature citations  
Cross-ref. databases

Taxonomy  
Diseases

Subcellular locations  
Keywords

News

[Forthcoming changes](#)  
Planned changes for UniProt

[UniProt release 2018\\_04](#)  
The Matrix (enzymes) Reloaded | Cross-references to GlyConnect

[UniProt release 2018\\_03](#)  
Ama-(not a)-toxin: a cap on death | Cross-references to VGNC

[News archive](#)

## Getting started

### Text search

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

### BLAST

Find regions of similarity between your sequences



## UniProt data

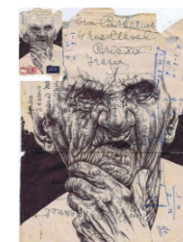
### Download latest release

Get the UniProt data

### Statistics

View Swiss-Prot and TrEMBL statistics

## Protein spotlight



### Giving In To Time

May 2018

Time runs its treacherous fingers along everything. The smoothed edges of a pebble. The polished wood of a staircase. The worn

# Example of a Swiss-Prot Record

www.uniprot.org/uniprot/Q9H479

UniProtKB  Advanced

BLAST Align Retrieve/ID mapping Peptide search Help Contact

## UniProtKB - Q9H479 (FN3K\_HUMAN)

Display

Entry Publications Feature viewer Feature table

None

Function

Names & Taxonomy

Subcell. location

Pathol./Biotech

PTM / Processing

Expression

Interaction

Structure

Family & Domains

Sequence

Cross-references

Entry information

Miscellaneous

**Protein** | **Fructosamine-3-kinase**

**Gene** | **FN3K**

**Organism** | *Homo sapiens (Human)*

**Status** |  - Annotation score: ●●●●○ - Experimental evidence at protein level<sup>i</sup>

### Function<sup>i</sup>

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 function<sup>i</sup>

- fructosamine-3-kinase activity
- kinase activity

Complete GO annotation...

#### GO - Biological process<sup>i</sup>

- epithelial cell differentiation
- fructosamine metabolic process
- fructoselysine metabolic process
- post-translational protein modification

Complete GO annotation...

#### Keywords<sup>i</sup>

Molecular Kinase Transferase

## Gene Ontology (GO):

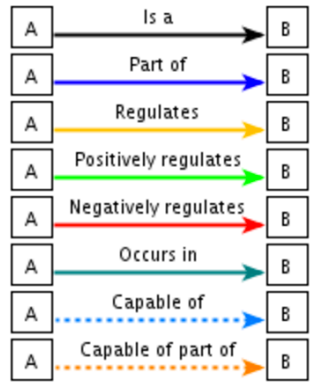
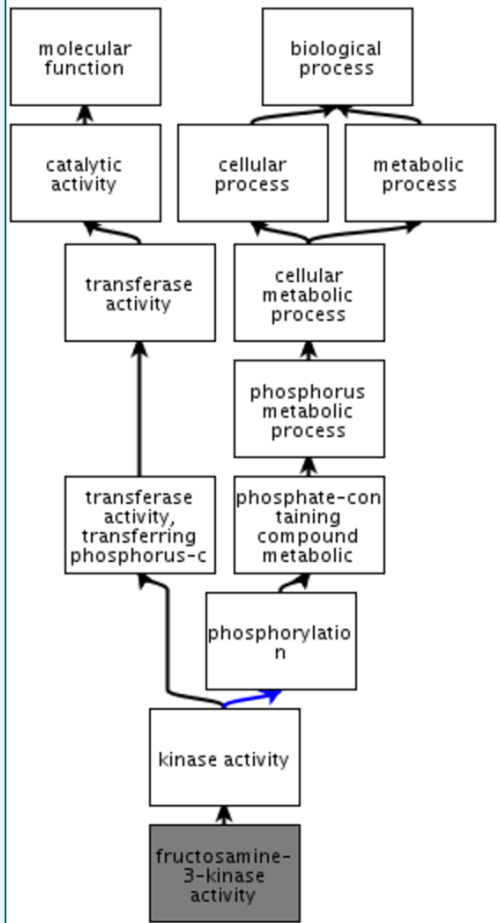
Structured vocabulary for defining molecular functions, biological processes, and cellular components.

# Gene Ontology: a structured relational vocabulary for describing biological functions

**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.



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
<b>green marbles</b>	$k$	$K - k$	$K$
<b>red marbles</b>	$n - k$	$N + k - n - K$	$N - K$
<b>total</b>	$n$	$N - n$	$N$

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?

GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC  
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA  
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCCTGGTCCCT  
TGGAGGCATGCAGTTCAGCAGACAGTGAATCAGCCATCCACCCAACATGCGGAACGTGTC  
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCTCCCTCTGTGAAAAGGCACGCTGATCTG  
TCTGGATAAGTGTGGCCGGCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC  
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA  
AAAGACAGCTCCAGTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG  
GGCTTGGAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA  
TGACCTTGGCCTACGATAATGGCATCAACCTGTTTCGATACGGCGGAGGTCTACGCTGCTG  
GAAAAGCTGAAGTGGTATTAGGGAAACATCATTAAGAAGAAGGGATGGAGACGGTCCAGCC  
TTGTCATCACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA  
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG  
ATGTGGTTTTTGGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA  
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA  
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG  
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT  
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG  
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT  
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCCGCCAGCAGGCCAAGCTGAAGGAACTGC  
AGGCCATTGCCGAACGCCTGGGCTGCACCCTACCCAGCTGGCCATAGCCTGGTGCCTGA  
GGAATGAGGGTGTGTCAGCTCCGTGCTTCTGGGGTGGCTTCCAATGCAGAACAACCTTATGGAGA



# Is there an ORF for a potential Coding Region?

GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCCTCCC  
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA  
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCCTGGTCCCT  
TGGAGGCATGCAGTTCAGCAGACAGTGAATCAGCCATCCACCCAACATGCGGAACGTGTC  
TCTTCTGCAGGTCCCAGTCCACAGCAGGATTCACCCCTCTGTGAAAAGGCACGCTGATCTG  
TCTGGATAAGTGTGGCCGGCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC  
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA  
AAAGACAGCTCCAGTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG  
GGCTTGGAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA  
TGACCTTGGCCTACGATAATGGCATCAACCTGTTTCGATACGGCGGAGGTCTACGCTGCTG  
GAAAAGCTGAAGTGGTATTAGGGAAACATCATTAAGAAGAAGGGATGGAGACGGTCCAGCC  
TTGTCATCACCAACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA  
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG  
ATGTGGTTTTTGGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA  
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA  
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG  
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT  
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG  
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT  
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCCGCCAGCAGGCCAAGCTGAAGGAACTGC  
AGGCCATTGCCGAACGCCTGGGCTGCACCCTACCCAGCTGGCCATAGCCTGGTGCCTGA  
GGAATGAGGGTGTGTCAGCTCCGTGCTTCTGGGTGCTTCCAATGCAGAACAACCTTATGGAGA

# Is there an ORF for a potential Coding Region?

GGAGCTGGAGGCCCCAGGCAACTACACCGTCCACGTACCCAGAGGGGCTGGGCCCTCCC  
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTGTTAGTCTCTGAGTGTGCA  
GTTGCTGCAC**ATGGGGCCCTGGCGCTTGCTGCACCAACTTCCTGTTGGGCCCGTGGTCCT**  
**TGGAGGCATGCAGTTCAGCAGACAGTGA**CTCAGCCATCCACCCAACATGCGGAACGTGTC  
TCTTCTGCAGGTCCCGGTCCACAGCAGGATTCCCCCTCTGTGAAAAGGCACGCTGATCTG  
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC  
TCTCCCTGCGGCAGACAGGCTCCCCCGGGATGATCTACAGTACTCGTTATGGGAGTCCCA  
AAAGACAGCTCCAGTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG  
GGCTTGGAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA  
TGACCTTGGCCTACGATAATGGCATCAACCTGTTTCGATACGGCGGAGGTCTACGCTGCTG  
GAAAAGCTGAAGTGGTATTAGGGAACATCATTAAGAAGAAGGGATGGAGACGGTCCAGCC  
TTGTCATCACCAACCAAGATCTTCTGGGGTGGAAAAGCGGAGACTGAGAGAGGCCTTTCCA  
GGAAGCACATAATTGAAGGACTGAAAGCGTCCCTGGAGCGGCTGCAGCTGGAGTACGTGG  
ATGTGGTTTTTGGCCAACCGCCCAGACCCCAACACGCCCATGGAAGAGACCGTGCGGGCCA  
TGACCCATGTCATCAACCAGGGGATGGCCATGTACTGGGGCACATCACGCTGGAGCTCCA  
TGGAGATCATGGAGGCCTACTCGGTGGCTCGGCAGTTCAACCTGATCCCGCCCATCTGCG  
AGCAAGCGGAATATCACATGTTCCAGAGGGAGAAGGTGGAGGTCCAGCTGCCAGAGCTGT  
TCCACAAGATAGGAGTAGGTGCCATGACCTGGTCCCCTCTGGCGTGCGGCATCGTCTCAG  
GGAAGTATGACAGCGGGATCCCACCCTACTCCAGAGCCTCCCTGAAGGGCTACCAGTGGT  
TGAAGGACAAGATCCTGAGTGAGGAGGGTCGCCGCCAGCAGGCCAAGCTGAAGGAACTGC  
AGGCCATTGCCGAACGCCTGGGCTGCACCCTACCCAGCTGGCCATAGCCTGGTGCCTGA  
GGAATGAGGGTGTGTCAGCTCCGTGCTTCTGGGGTGGCTTCCAATGCAGAACAACCTTATGGAGA

# Find all ORFs using ORFfinder

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

NCBI Resources How To Sign in to NCBI

ORFfinder PubMed  Search

## 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  
ACCAGAGACCACGCCCTGGTGTGCCTTAGGGGCCCTGGTTTTGTTAGTCTCTGAGTGTGCA  
GTTGCTGCACATGGGGCCCTGGCGCTTGCTGCACCAACTTCTGTGGGCCCGTGGTCCT  
TGGAGGCATGCAGTTCAGCAGACAGTGAAGTCCAGCCATCCACCAACATGCGGAACGTGTC  
TCTTCTGCAGGTCCCGGTCCACAGCAGGATCCCCCTCTGTGAAAAGGCACGCTGATCTG  
TCTGGATAAGTGTGGCCGGCCCCATGTATCCGGAATCAACCACGGGGTCCCCAGCTCGAC  
TCTCCCTGCGGCAGACAGGCTCCCCCGGATGATCTACAGTACTCGTTATGGGAGTCCCA  
AAAGACAGCTCCAGTTTTACAGGAATCTGGGCAAATCTGGCCTTCGGGTCTCCTGCCTGG  
GGCTTGAACATGGGTGACCTTCGGGGGCCAGATCACGGATGAGATGGCAGAGCACCTAA  
TGACCTTGGCCTACGATAATGGCATCAACCTGTTTCGATACGGCGGAGGTCTACGCTGCTG
```

From:  To:

# ORFfinder finds all open reading frames and provides translations

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

NCBI Resources How To Sign in to NCBI

ORFfinder PubMed  Search

## Open Reading Frame Viewer

Sequence

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

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



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

Add six-frame translation track

ORF5 (367 aa)

Display ORF as...

Mark

Mark subset...

Marked: 0

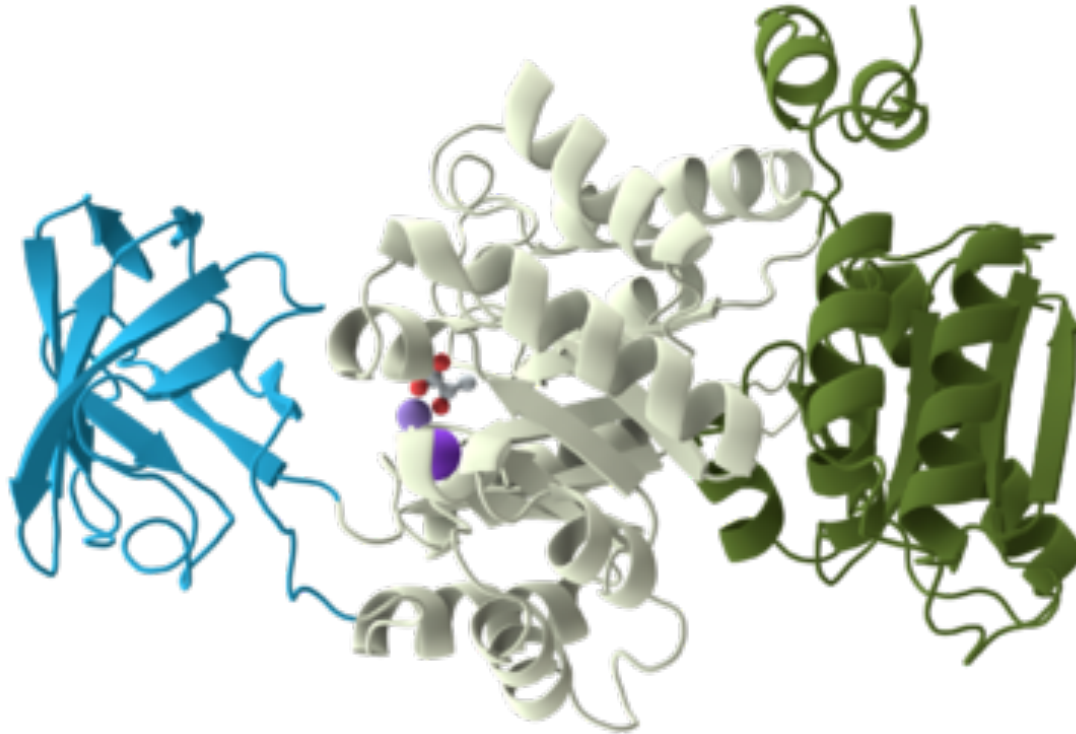
Download marked set

as Protein FA

```
>lcl|ORF5
MYPESTTGSPARLSLRQTGSPGMIYSTRYGSPKRQLQFYR
NLGKSGLRVSLGLGTWVTFGGQITDEMAEHLMTLAYDNG
INLFDTAEVYAAGKAEVVLGNIIKKKGWRRSSLVITTKIF
WGGKAETERGLSRKHIEGLKASLERLQLEYVDVVFANRP
DPNTPMEETVRAMTHVINQGMAMYWGTSRWSSMEIMEAYS
VARQFNLIPIPCQAEYHMFQREKVEVQLPELPHKIGVGA
MTWSPLACGIVSGKYDSGIPPYSRASLKGQWLKDKILSE
EGRRQQAKLQELQAIERLGTLPQLAIAWCLRNQGVSSV
LLGASNAEQLMENIGAIQVLPKLSISSIVHEIDSILGNKPY
SKKDYRS
```

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
ORF8	-	2	848	618	231   7

# 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.xfam.org



HOME | SEARCH | BROWSE | FTP | HELP | ABOUT



## 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.

```
METGGRARTGTPQPAAPGVWRARPAGGGGGGASSWLLDGNWLLCYGLFLY
LALYAQVSQSKPCERTGSCFSGRCVNSTCLCDPGWVGDCQHCQGRFKLT
EPSGYLTDGPINIKYKTKCTWLIEGYPNAVLRFRNFHATECSWDHMYVY
DGSYIAPLIAVL SGLIVPEIRGNETVPEVVTTSYALLHFFSDAAYNLT
GFNIFYSINSCPNNGSGHGKCTTSVSVSPQVYCECDKYWKGEACDIPYCK
ANCGSPDHGVCYDLTGEKLCVCNDSWQGPDCSLNVPSTESYWILPNVKPFS
PSVGRASHKAVLHGKFMWVIGGYTFNYSSFQMVLYNLESIWNVGTPSR
GPLQRYGHSLALYQENIFMYGGRIETNDGNVDELWVFNHISQSWSTKTP
TVLGHGQQYAVEGHS AHIMELDSRDVVMIIIFGYS AIYGYTSSIQEYHIS
SNTWLV PETKGAIVQGGYGHSTSVYDEITKSIYVHGGYKALPGNKYGLVDD
LYKYE VNTKTWILKESGFARYLHSAV LINGAMLIFGGNTHNDTSLN SGA
KCF SADFLAYDIACDEWKILPKPNLHRDVNRFHGS AVVINGSMYIFGGFS
SVLLNDILVYKPPNCKAFRDEELCKNAGPGIKCVWNKNHCESWESGNTNN
ILRAKCPPKTAASDDRCRYAD CASCTANTNGCQWCDKCKICISANSNC SM
SVKNYTKCHVRNEQICNKLTSCKSCSLNLCQWDQRQQEQCALPAHL CGE
GWSHIGDA CLR VNSSRENYDNAKLYCYNLSGNLASLTT SKEVEFLDEIQ
KYTQQKVSPWGLR KINISYWG WEDMSPFTNTTLQWLPGEPNDSGFCAYL
ERA AVAGLKANPCTSMANGLVCEKPVVSPNQNARPCKKPCSLRTSCSNCT
SNGMECMWCSSTKRCVDSNAYIIFPYGQCLEWQTATCSPQNC SGLRTCG
QCLEQPGCGWCNDPSNTGRGHCIEGSSRGPMLIGMHSEMVLDTNLC PK
EKNYEW SFIQCPACQCNGHSTCINNVC EQCKNLTGKQCQDCMPGY YGD
PTNGGQCTACTCSGHANICHLHTGKCFCTTKGIKGDQCQLCDSENRYVGN
PLRGTCYYSLLIDYQFTFSLQEDDRHHTAINFIANPEQSNK NLDISINA
SNNFNLNITWSVGSTAGTISGEETSIVSKNNIKEYRDSFSYKFNFRSNP
NITFYVYVS NFSWPIKIQIAFSQHNTIMDLVQFFVTF FSCFLSLLVA AV
VWKIKQTCWASRRRREQLLRERQQMASRPFASVDVALEVGAEQTEFLRGPL
EGAPKPIAIEPCAGNRAAVLTVFLCLPRGSSGAPPGQSGLAIASALIDI
SQQKASDSKDKTSGVRNRKHLSTRQGTCV
```

Go  
Example

This search will use 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.

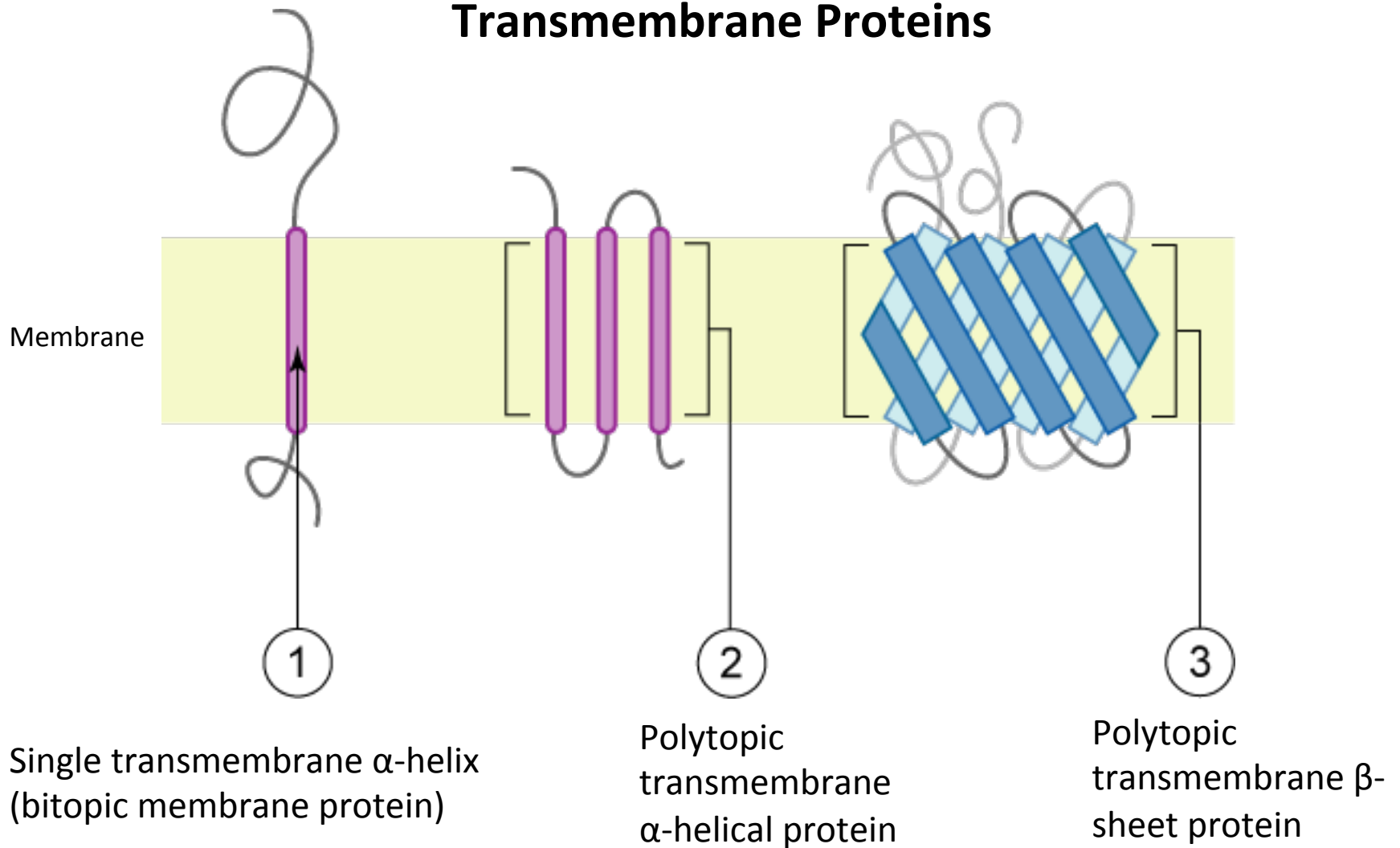
## Significant Pfam-A Matches

[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					
<a href="#">CUB</a>	CUB domain	Domain	<a href="#">CL0164</a>	93	206	93	206	1	110	110	42.2	7.7e-11	n/a	<a href="#">Show</a>
<a href="#">EGF_2</a>	EGF-like domain	Domain	<a href="#">CL0001</a>	249	280	249	280	1	32	32	22.5	0.0001	n/a	<a href="#">Show</a>
<a href="#">Kelch_5</a>	Kelch motif	Repeat	<a href="#">CL0186</a>	351	393	352	392	<b>2</b>	<b>41</b>	42	33.7	2.2e-08	n/a	<a href="#">Show</a>
<a href="#">Kelch_4</a>	Galactose oxidase, central domain	Repeat	<a href="#">CL0186</a>	466	518	468	514	<b>3</b>	<b>44</b>	49	20.6	0.0003	n/a	<a href="#">Show</a>
<a href="#">Kelch_1</a>	Kelch motif	Repeat	<a href="#">CL0186</a>	520	574	520	573	1	<b>45</b>	46	20.0	0.00033	n/a	<a href="#">Show</a>
<a href="#">Kelch_5</a>	Kelch motif	Repeat	<a href="#">CL0186</a>	579	614	581	613	<b>5</b>	<b>40</b>	42	25.3	9.7e-06	n/a	<a href="#">Show</a>
<a href="#">Lectin_C</a>	Lectin C-type domain	Domain	<a href="#">CL0056</a>	765	874	766	874	<b>2</b>	108	108	70.2	2e-19	n/a	<a href="#">Show</a>
<a href="#">PSI</a>	Plexin repeat	Family	<a href="#">CL0630</a>	889	939	890	938	<b>2</b>	<b>50</b>	51	27.8	2.5e-06	n/a	<a href="#">Show</a>
<a href="#">PSI</a>	Plexin repeat	Family	<a href="#">CL0630</a>	942	1012	942	1012	1	51	51	50.0	2.9e-13	n/a	<a href="#">Show</a>

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

# Transmembrane Proteins

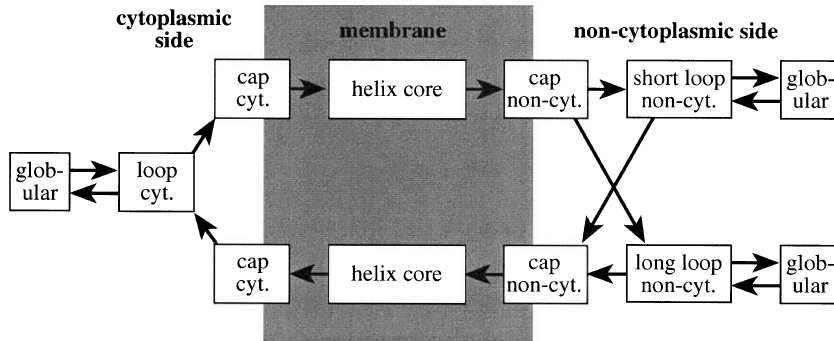




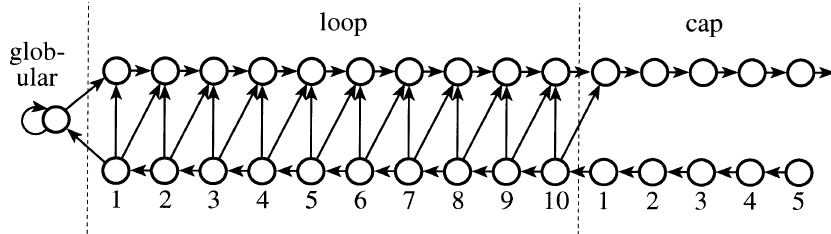
# Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes

Anders Krogh<sup>1\*</sup>, Björn Larsson<sup>1</sup>, Gunnar von Heijne<sup>2</sup> and Erik L. L. Sonnhammer<sup>3</sup>

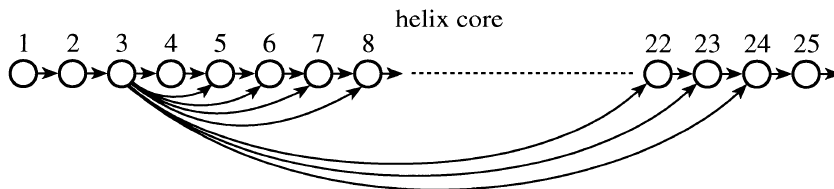
(a)



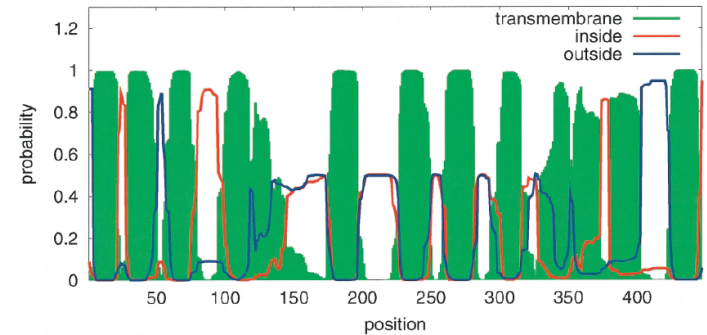
(b)



(c)



**Figure 1.** The layout of the hidden Markov model. (a) The overall layout. Each box corresponds to one or more states in the HMM. Parts of the model with the same text are tied, i.e. their parameters are the same. Cyt. represents the cytoplasmic side of the membrane and non-cyt. the other side. (b) The detailed structure of the inside and outside loop models and helix cap models. (c) The structure of the model for the helix core modelling lengths between five and 25, which translates to helices between 15 and 35 when the caps are included.



**Figure 2.** Posterior probabilities for a single sequence. The posterior probability for transmembrane helix, inside, or outside displayed for the gluconate permease 3 from *E. coli* (SWISS-PROT entry Gntp\_ECOLI), for which the structure is unknown. Some parts of the protein are relatively certain, whereas other parts are less certain.

# Using TMHMM to identify putative transmembrane proteins

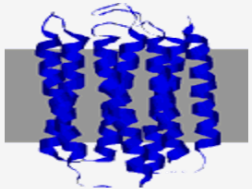
← → ↻ ⓘ www.cbs.dtu.dk/services/TMHMM/ ☆ amazon ABP JB [document icon] [mail icon] [share icon] [help icon] [refresh icon]

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

[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)

No file chosen

OR by pasting sequence(s) in **FASTA** format:

```
MEILCEDNTSLSSIPNSLMQVDGDSGLYRNDNFNSRDANSSDASNWTIDGENRTNLSFEG
YLPPTCLSILHLQEKNWSALLTAVVILTIAGNILVIMAVSLEKKLQATNYFLMSLAIADMLL
GFLVMPVSMILTILYGYRWPLPSKLCVWYLDVLFSTASIMHLCAISLDRYVAIQNPIHHSR
FNSRTKAFLKIIAVWTISVGVSMPIPVFGLQDDSKVFKQGSCLLADDNFVLIQSVFAFFIPLTI
MVITYFLTIKSLQKEATLCVSDLSRAKSLASFSFL
```

**Output format:**

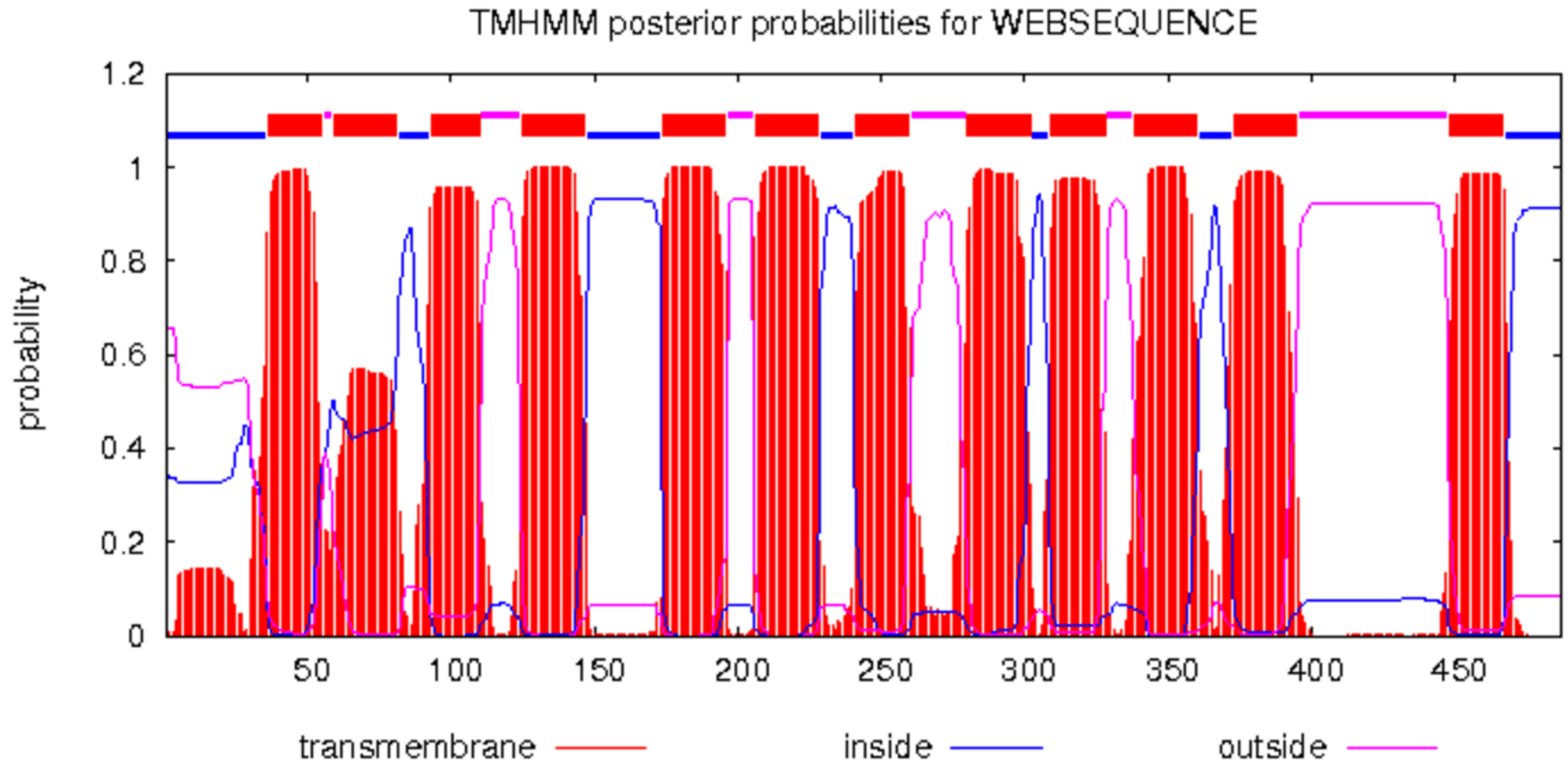
- Extensive, with graphics
- Extensive, no graphics
- One line per protein

**Other options:**

- Use old model (version 1)

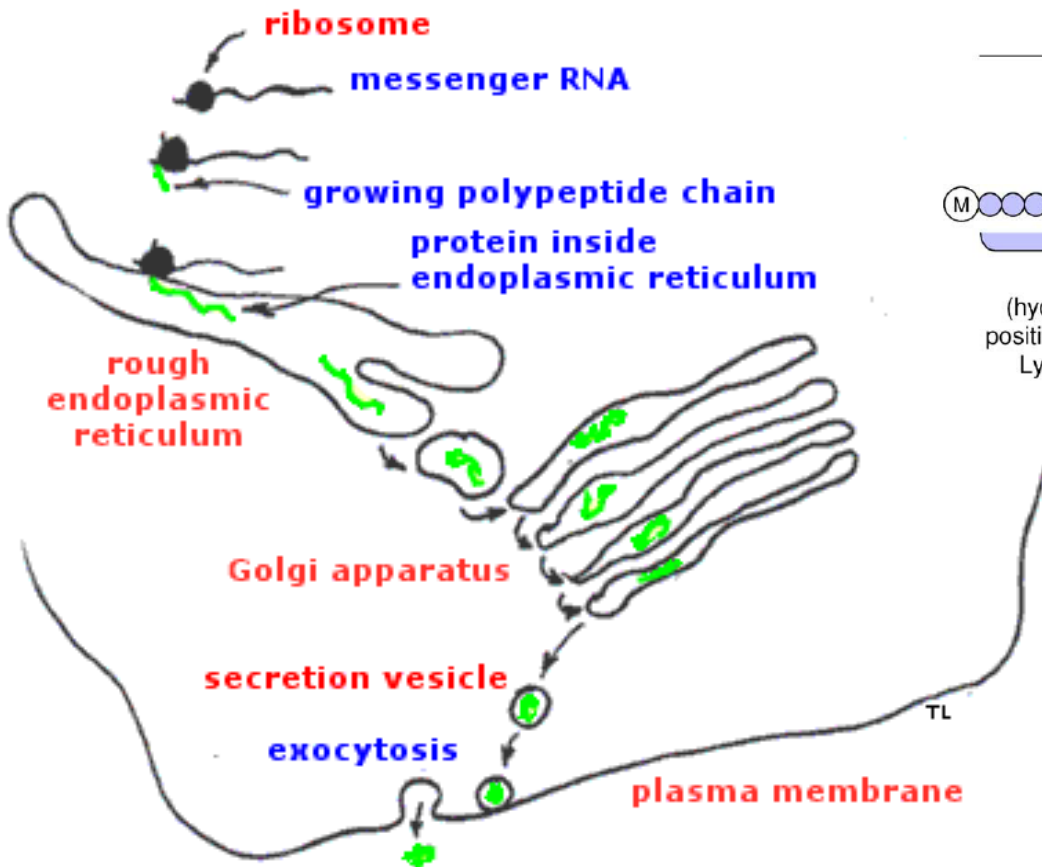
CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS ■ TECHNICAL UNIVERSITY OF DENMARK DTU

# Trans-membrane Domains via TmHMM

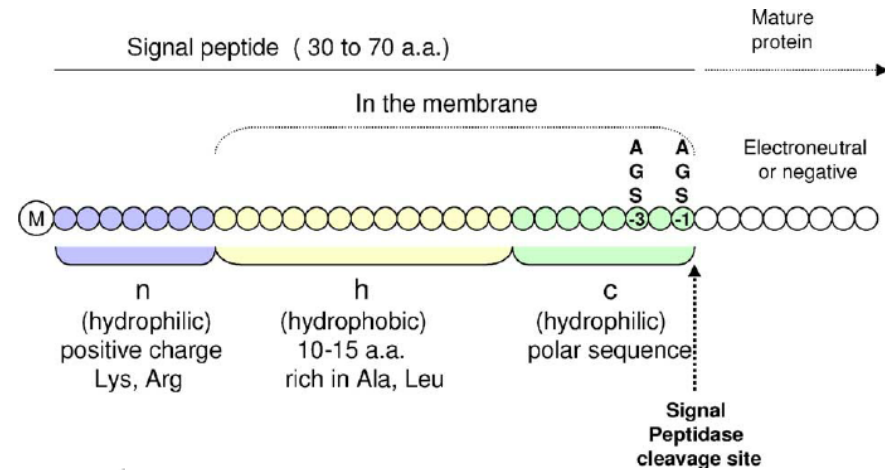


Topology=i36-55o59-81i93-110o125-147i174-196o206-228i241-260o280-302i309-328o338-360i373-395o448-467i

# Predicting Secreted Proteins

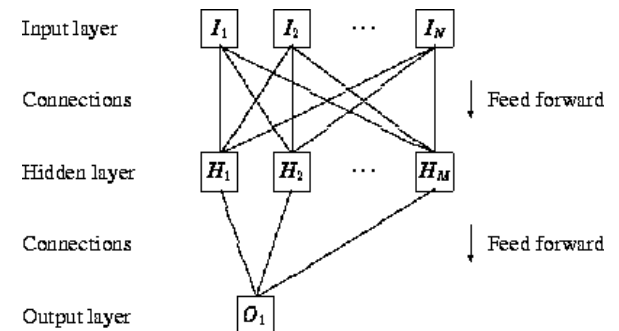


(from: <https://courses.washington.edu/conj/cell/secretion.htm>)



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

## Neural Network Used (in part)



(from <http://www.cbs.dtu.dk/services/SignalP-3.0/background/prediction.php>)

# SignalP: Prediction of N-terminal signal peptides (predict secreted proteins)

www.cbs.dtu.dk/services/SignalP/

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:

```
MHPAVFLSLPDLRCSLLLLVTWVFPVTTEITSLDTENIDEILNADVALVNFYADWCRFSQMLHPIFEEASDVKEEFPNENQVVFARVDCDQHSQDIQRYRISKYPTLKLFRNGMMM  
KREYRGRQSRVKALADYIRQQKSDPIQEIRDLAIEITLDRSKRNIIGYFEQKQSDNYRVFERVANILHDDCAFLSAFGDVSKPERYSGDNIYKPPGHSAPDMVYLGAMTNFDVTVYNIQ  
DKCVPLVREITFENGEELTEGLPFLILFHMKEDTESLEIFQNEVARQLISEKGTINFLHADQDKFRHPLLHIQKTPADCPVIAIDSRHMYVFGDFKDLVLPGLKQFVFDLHSGKLHREF  
HHGPDPTDTAPGEQAQDVASSPPESSFKLAPSEYRYTLLRDRDEL
```

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:
  - 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*

CENTER FOR BIOLOGICAL SEQUENCE ANALYSIS ■ TECHNICAL UNIVERSITY OF DENMARK DTU

# Example SignalP predicted signal peptide

← → ↻ ⓘ www.cbs.dtu.dk/cgi-bin/webface2.fcgi?jobid=58FFF29C00005F854B357EEA&w... ☆

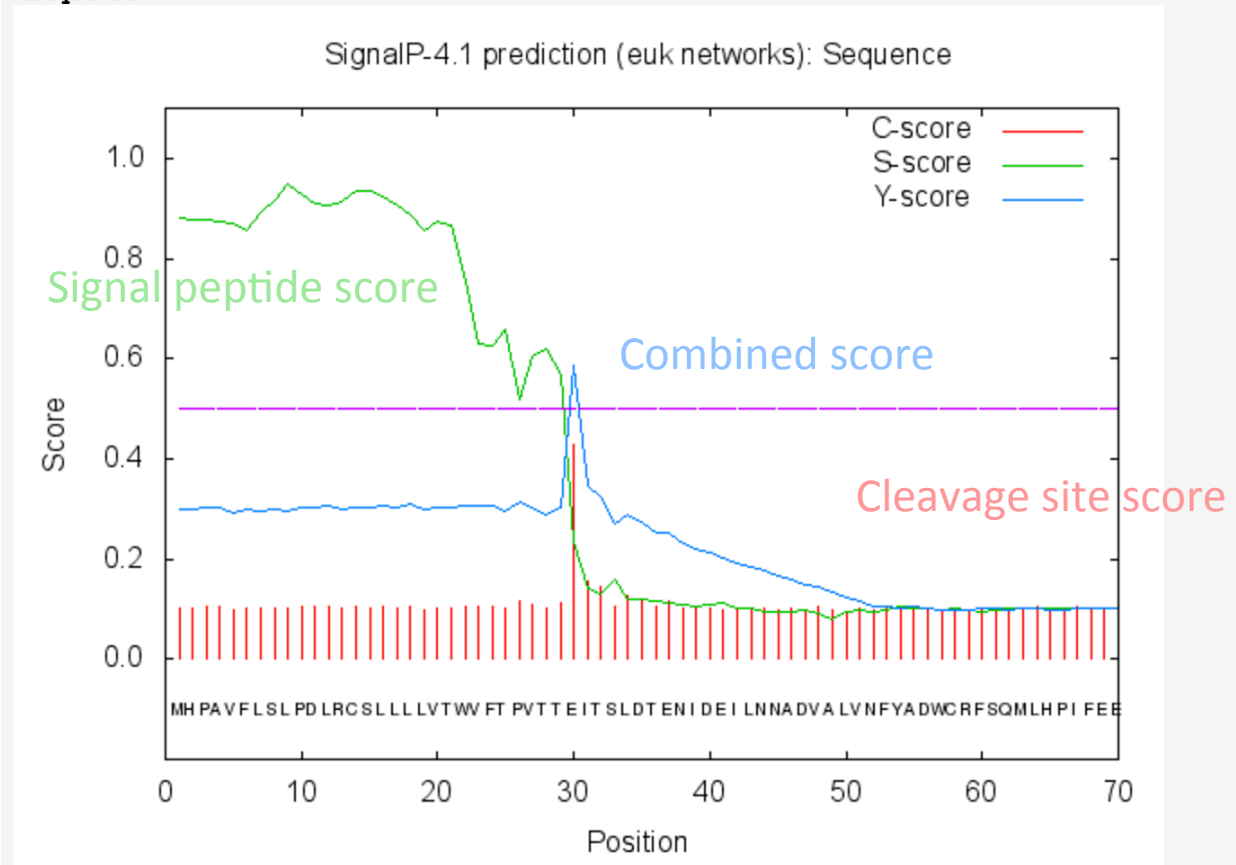


## SignalP 4.1 Server - prediction results

Technical University of Denmark

# SignalP-4.1 euk predictions

>Sequence



# Transcriptome-scale functional annotation using Trinotate



## Trinotate: Transcriptome Functional Annotation and Analysis

# Trinotate

TransDecoder



TMHMM

SignalP



eggNOG  
version 3.0



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

There's no substitute for experimentally validating protein functions





# Practicals

- Trinotate
- TrinotateWeb

# Practicals

- Functional Enrichment Analysis