

## Review of exercise 1

```
tblastn -num_threads 2 -db contig -query DH10B.fasta -out blastout.xls  
-evalue 1e-10 -outfmt "6 qseqid sseqid qstart qend sstart send length  
nident pident evalue"
```

### Other options:

-max\_target\_seqs : maximum number of targets to report

-perc\_identity: percentage identity cutoff

-task blastn-short : for short queries

### Output format:

-outfmt "6 col1 col2 col3 ..."

std: commonly used 16 columns

stitle: description line of the target sequence

## Translate RNA sequence to Protein sequence

1. 6-frame translation.
2. ORF detection tool to find the correct frame
  - ORF length
  - HMM model trained on a set of genuine proteins (from the same species to capture the correct codon bias signal)
  - BLAST or PFAM scan

## TransDecoder – an ORF finder tool of the Trinity package

```
TransDecoder -t transcript.fasta
```

Other parameters:

-S: only analyze top strand

### Training the HMM

--train training.fasta : a set of high confidence transcripts

--cd\_hit\_est : path to CD-hit-est tool, a clustering tool to produce a non-redundant protein set

-G: genetic code

### Output:

--retain\_long\_orfs 900: all ORF longer than 900 nt will be retained

-m 100: minimum protein length

## The ORF prediction can refined by homology protein search

```
blastp -query transdecoder_dir/longest_orfs.pep -db  
uniprot_sprot.fasta -max_target_seqs 1 -outfmt 6 -evaluate 1e-5 -  
num_threads 10 > blastp.outfmt6
```

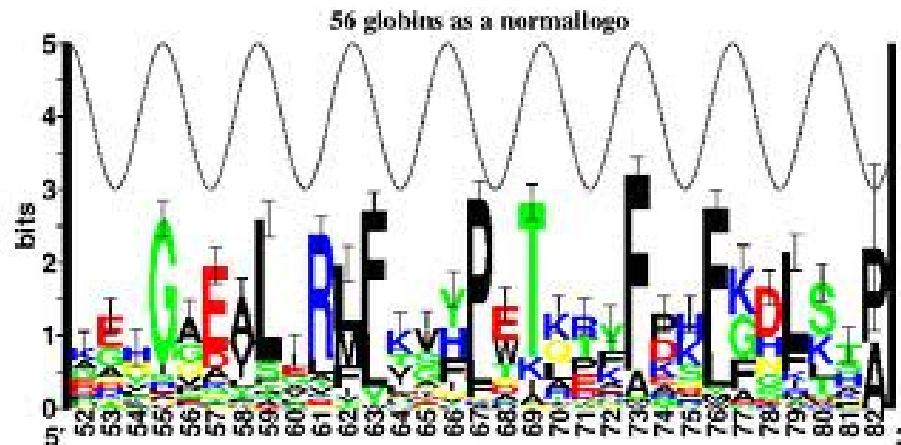
```
hmmsearch --cpu 8 --domtblout pfam.domtblout /path/to/Pfam-A.hmm  
transdecoder_dir/longest_orfs.pep
```

```
TransDecoder.Predict -t target_transcripts.fasta --  
retain_pfam_hits pfam.domtblout --retain_blastp_hits  
blastp.outfmt6
```

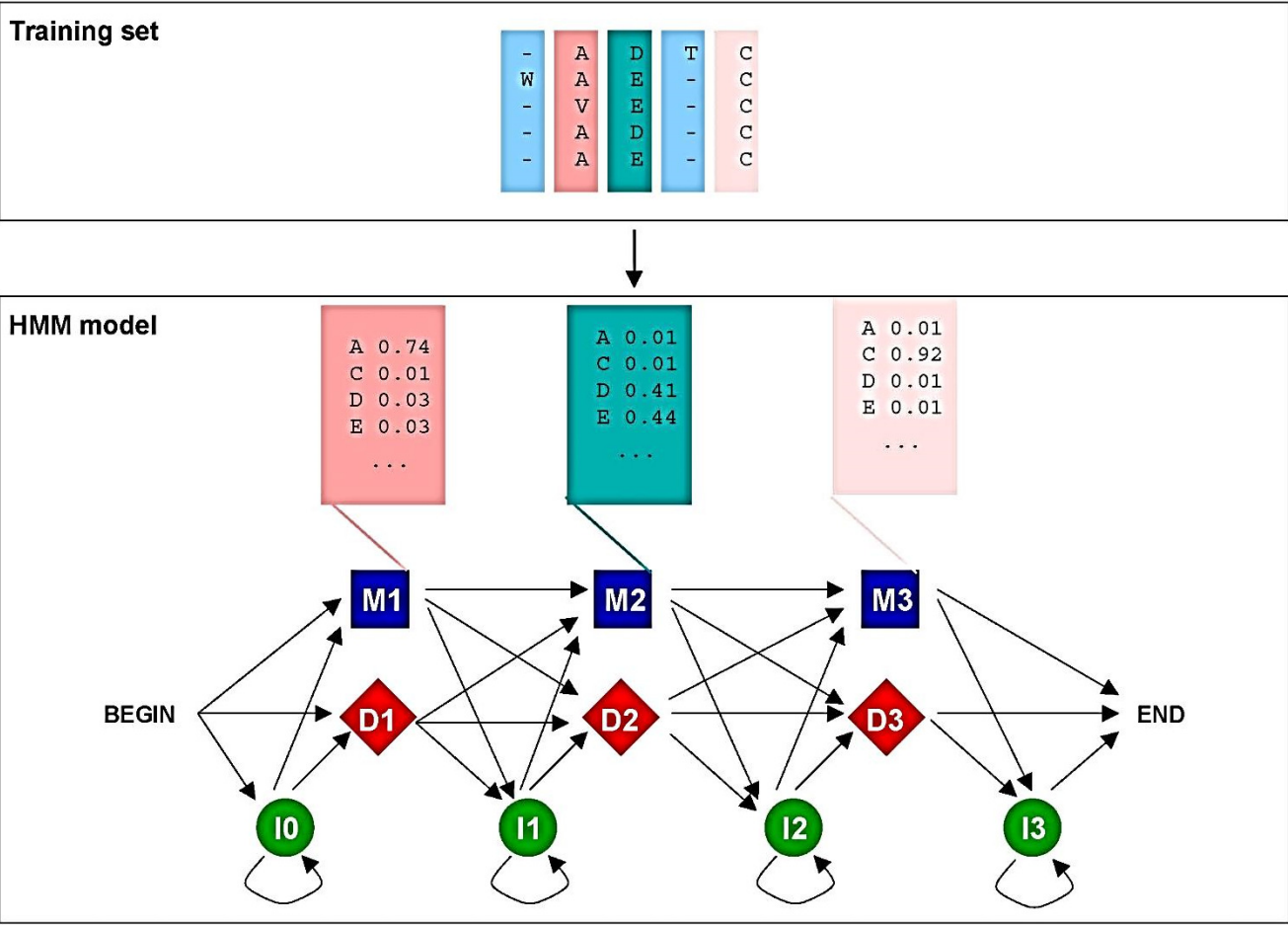
# HMMs are trained from a multiple sequence alignment

Q5E940_BOVIN	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_HUMAN	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_MOUSE	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_RAT	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_CHICK	-----MPREDRATWKSNYFMKIIQLDDYPKCFVVGADNVGSKOMQOIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_RANSY	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE	76
Q7ZUG3_BRARE	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_ICTPU	-----MPREDRATWKSNYFLKIIQLDDYPKCFIVGADNVGSKOMQOIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE	76
RLA0_DROME	-----MVRENKAANKAQYFIKVVYELDFPKCFIVGADNVGSKOMONIRTSIRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE	76
RLA0_DICDI	-----MSGAG-SKRKKLFIEKATKLFETTVDKMIVAEADFGSSQLOKIRKSIRGI-GAVLMGKNTMIRKVIRDLADSK--PELD	75
Q54LP0_DICDI	-----MSGAG-SKRKNVFIEKATKLFETTVDKMIVAEADFGSSQLOKIRKSIRGI-GAVLMGKNTMIRKVIRDLADSK--PELD	75
RLA0_PLAF8	-----MAKLSKQKKQMYIEKLSSLQQYSKILIVHVDNVGSKOMASVRSRSGK-ALILMGKNTIRRTALKKNLQAV--PQIE	76
RLA0_SULAC	-----MIGLAVTTTKTAKWKVDEVAELTEKRLKTKHTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLNFNIALKNAG----YDLK	79
RLA0_SULTO	-----MRIMAVITQERKIATKWKIEEVKELEOKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG----LDVS	80
RLA0_SULSO	-----MKRLALALKQRKVASWKLIEYKELTELKNSNTLILGNLECFPADKLHEIRKKLRGK-ADIKVTKNTLFGIAAKNAG----IDIE	80
RLA0_AERPE	MSVVSIVGQMYKREKIPENKTLMLRELEELFSKIRVILFADITGIPTFVYVVRVKKLWKK-YPMMVAKKRILLRAMKAGLE--LDDN	86
RLA0_PYRAE	MMLATGKRRYVRTROYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEVRYRLRRY-GVIKIIKPTLFGIAFTKVVYGG--TPAE	85
RLA0_METAC	-----MAEERHHTETIPQWKKDEIENIKELIQSHKVFQMGVIEGILATKMOKIRRDLDKV-AVLKVSRLNTEALNQLG----ETIP	78
RLA0_METMA	-----MAEERHHTETIPQWKKDEIENIKELIQSHKVFQMGVIEGILATKMOKIRRDLDKV-AVLKVSRLNTEALNQLG----ESIP	78
RLA0_ARCFU	-----MAAVRGS--PPEYKVRAVEEIKRMISSEKVVVAIVSFRNVVAGQMOKIRREFRGK-AEIKVVKNTLLEALDNLG----GDL	75
RLA0_METKA	MAVKAKGQPPSGYEPKVAEWKRREYKELKELMDEYENVGLVDLEGIPAPQLOEIRAKLRERDTIIRMSRNTLMRIAILEKLDER--PELE	88
RLA0_METTH	-----MAHVAEWKKKEVEEELANLKSYPVIALVDVSSMPAYPLSQMRRLIRENGCLLRVSRNTLIELAIKKAAQELGKPELE	74
RLA0_METTL	MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLOEIRDKIR-ETMLKMSRNTLIELAIKKAAQELGKPELE	82
RLA0_METVA	MIDAKSEHKIAPWKIEEVNKLKELLKNSANVIALIDMMEVAVQLOEIRDKIR-DQMLKMSRNTLIELAIKKAAQELGKPELE	82
RLA0_METJA	-----METKVAHVADPKKIEEVKTLKGLIKSKPVVAIVDMMDVPAQLOEIRDKIR-DKVKLRMSRNTLIELAIKKAAQELGKPELE	81
RLA0_PYRAB	-----MAHVAEWKKKEVEEELANLKSYPVIALVDVSSMPAYPLSQMRRLIRENGCLLRVSRNTLIELAIKKAAQELGKPELE	77
RLA0_PYRHO	-----MAHVAEWKKKEVEEELAKLKSYPVIALVDVSSMPAYPLSQMRRLIRENGCLLRVSRNTLIELAIKKAAQELGKPELE	77
RLA0_PYRFU	-----MAHVAEWKKKEVEEELANLKSYPVIALVDVSSMPAYPLSQMRRLIRENGCLLRVSRNTLIELAIKKAAQELGKPELE	77
RLA0_PYRKO	-----MAHVAEWKKKEVEEELANLKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKKAAQELGQPELE	76
RLA0_HALMA	MSAESEKRTETIPENKQEEVDIVEMIESVESVGVVNIAGIPSRLODMRRDLHGT-AELRVSRNTLIEALDDVD--DGLE	79
RLA0_HALVO	MSESEVRQTEVIPQWKREVDDELVDLIESVESVGVVNVAGIPSRLODMRRDLHGS-AAVRMSRNTLVNRALEVN--DGF	79
RLA0_HALSA	MSAEEQRTTEVIPENKQEEVDDELVDLIESVESVGVVNVAGIPSRLODMRRDLHGS-AAVRMSRNTLVNRALEVN--DGLD	79
RLA0_THEAC	-----MKEVSQKKELVNEITQRIKASRSVAIVDLAGIRTRQIODIRGNRGG-INLKVIKKTLFLKALENLGD--EKLS	72
RLA0_THEVO	-----MRKINPKKKEIVSELAQDITKSKAVAVDIDKGVYTRMODIRAKNRDK-VKIKVVKTLFLKALDLSND--EKLT	72
RLA0_PICTO	-----MTEPAQWKIDFVKNLENEINSRKVAALVSKLGRNNEFOKIRNSIRDK-ARIKVSARLLRLALENTGK--NNIV	72

ruler 1.....10.....20.....30.....40.....50.....60.....70.....80.....90

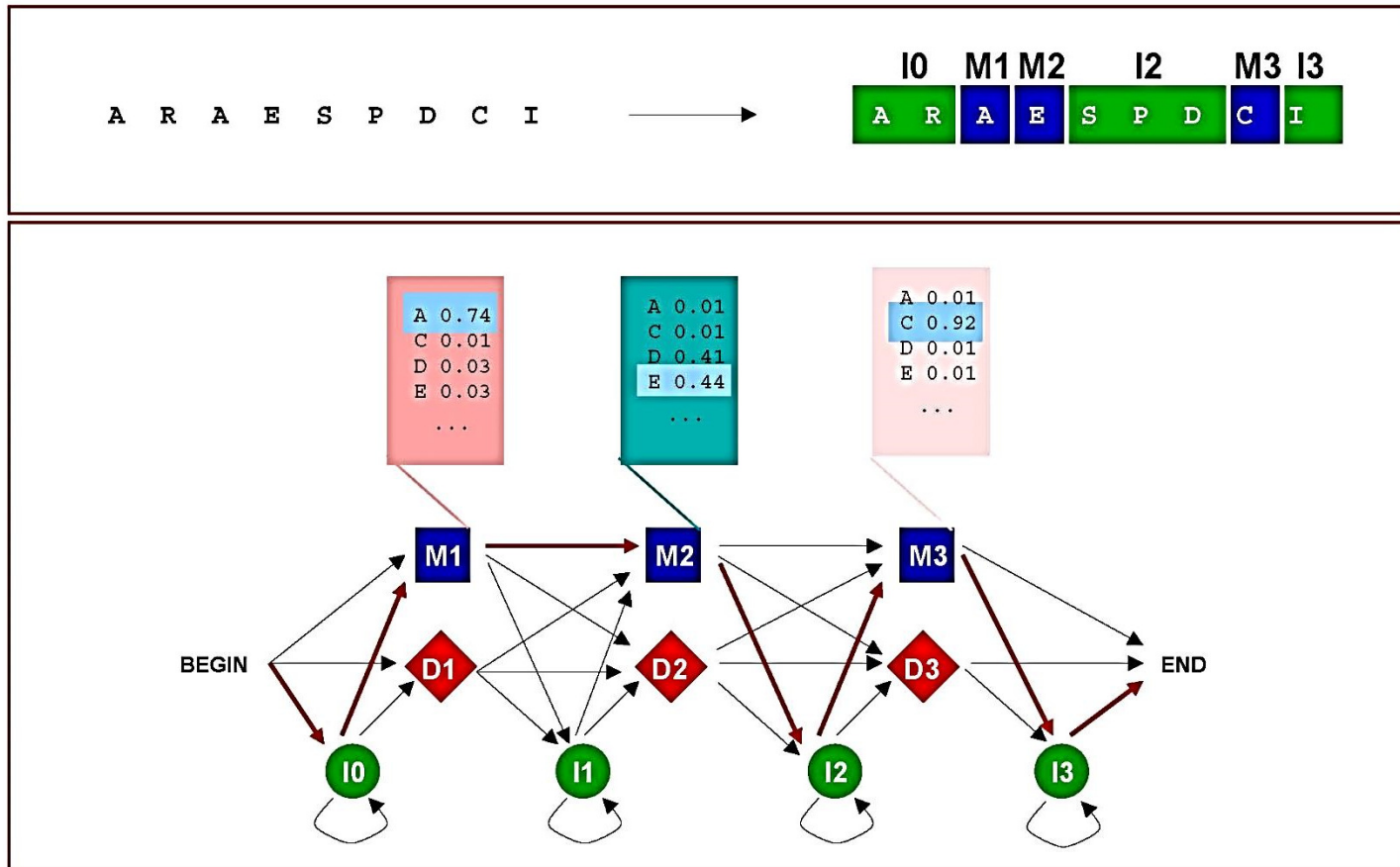


# Hidden Markov Model (HMM) is more general than PSSM



# Match a sequence to a model

## Application: Function Prediction



```

>unknown_protein
MALLYRRMSMLLNIIILAYIFLCAICVQGSVKQEWAEIGKNVSLCASENEAVAWKLGNTQINKNHTRYKI
RTEPLKSNDDGSENNDSQDFIKYKNVLALLDVNIKDSGNYTCTAQTGQNHSTEFQVRPYLPSKVLQSTPD
RIKRKIKQDVMLYCLIEMPQNETTNRNLKWLKDGSGQFEFLDTFSSISKLNTHLNFLEFTEVYKKENG
TYKCTVFDDTGLEITSKEITL FVMEVQVSIDFAKAVGANKIYLNWTVNDGNDPIQKFFITLQEAGTPTF
TYHKDFINGSHTSYILDHFKPNTTYFLRIVGKNSIGNGQPTQYPQGITTLSYDPIFIPKVETTGSTASTI
TIGWNPPLIDYIQYYELIVSESGEVPKVEEAIYQQNSRNLPMYFDKLTATDYEFVRVACSDLTKT
CGPWSENVNGTTMDGVATKPTNLSIQCHHDNVTRGNSIAINWDVPKTPNGKVVSYLIIHLGNPMSTVDRE
MWGPKIRRIDEPHHKTLYESVSPNTNYTVTVSAITRHKKNGEPATGSCMLPVPSTDAIGRTMWSKVNLD
KYVLKLYLPKISERNPICCYRLYLVRINNDNKELPDPEKLNIAIYQEVHSDNVTRSSAYIAEMISSKYF
RPEIFLGDEKRFSENNDIIRDNDIICRKCLEGTPLRKP EIIHIPPQGSLSNSDSELPILSEKDNLIKGA
NLTEHALKILESKLRDRNAVTSDENPILSAVNPVPLHDSSRDVFDGEIDINSNYTGFL EIIIVRDRNNA
LMAYSKYFDIITPATEAEPIQSLNMDYYLSIGVKAGAVLLGVILVFLVWVFFHKKTKNELQGEDTLTL
RDSLRLALFGRRNHHCHEITTCNKKCFDAGRTURLDLFNAYKRNHKTQVYCFLEVEMLDNEFCDDTIAN
SDLKENACKNRYPI
EQHLEIIVMLTNL
RRQITQYHYLTWK
SVSIYNTVCDLRH
EKLLATADEISKS
QDPLENTIGDFWR
TNCKIDDTLKVTQ
VAMCILVQHLRLE

```

# PFAM

a pre-constructed HMM model database  
for protein function domain prediction

**Sequence search results**

[Show](#) the detailed description of this results page.

We found **7** Pfam-A matches to your search sequence (**all** significant). You did not choose to search for Pfam-B matches.

[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="#">Ig_2</a>	Immunoglobulin domain	Domain	<a href="#">CL0011</a>	24	127	35	126	<b>11</b>	<b>78</b>	80	27.0	3.5e-06	n/a	<a href="#">Show</a>
<a href="#">Ig_2</a>	Immunoglobulin domain	Domain	<a href="#">CL0011</a>	132	233	135	233	<b>4</b>	80	80	19.8	0.00063	n/a	<a href="#">Show</a>
<a href="#">fn3</a>	Fibronectin type III domain	Domain	<a href="#">CL0159</a>	237	321	244	320	<b>8</b>	<b>84</b>	85	39.3	5.2e-10	n/a	<a href="#">Show</a>
<a href="#">fn3</a>	Fibronectin type III domain	Domain	<a href="#">CL0159</a>	333	425	340	425	<b>6</b>	85	85	40.9	1.6e-10	n/a	<a href="#">Show</a>
<a href="#">fn3</a>	Fibronectin type III domain	Domain	<a href="#">CL0159</a>	439	534	452	532	<b>11</b>	<b>83</b>	85	27.3	2.8e-06	n/a	<a href="#">Show</a>
<a href="#">Y_phosphatase</a>	Protein-tyrosine phosphatase	Domain	<a href="#">CL0031</a>	916	1154	916	1153	1	<b>234</b>	235	283.6	9.6e-85	1096,1096	<a href="#">Show</a>
<a href="#">Y_phosphatase</a>	Protein-tyrosine phosphatase	Domain	<a href="#">CL0031</a>	1212	1448	1212	1447	1	<b>234</b>	235	211.8	8.5e-63	1390,1390	<a href="#">Show</a>

Comments or questions on the site? Send a mail to [pfam-help@sanger.ac.uk](mailto:pfam-help@sanger.ac.uk). Our [cookie policy](#).

The Wellcome Trust

<http://pfam.sanger.ac.uk/>



# Other prediction tools

## SignalP: Predicting Signal Peptide

The screenshot shows the SignalP 4.0 Server web interface. At the top, there is a navigation menu with various categories: EVENTS, NEWS, RESEARCH GROUPS, CBS PREDICTION SERVERS, CBS DATA SETS, PUBLICATIONS, EDUCATION, STAFF, CONTACT, ABOUT CBS, INTERNAL, CBS BIOINFORMATICS TOOLS, CBS COURSES, and OTHER BIOINFORMATICS LINKS. Below the menu, the breadcrumb path is "CBS >> CBS Prediction Servers >> SignalP". The main heading is "SignalP 4.0 Server". A paragraph describes the server's function: "SignalP 4.0 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." Below this, there is a link to the "version history" and a note about the ePrint of the SignalP 4.0 paper. A horizontal navigation bar contains links for "Background", "Article abstracts", "Instructions", "Output format", and "Data". The "SUBMISSION" section includes a text area for pasting a single amino acid sequence or several sequences in FASTA format. Below the text area is a "Choose File" button with the text "No file chosen". There are three columns of options: "Organism group" (Eukaryotes, Gram-negative bacteria, Gram-positive bacteria), "Method" (Input sequences may include TM regions, Input sequences do not include TM regions), and "Graphics" (No graphics, PNG (inline), PNG (inline) and EPS (as links)). There are also "Output format" (Standard, Short (no graphics), Long, All - SignalP-noTM and SignalP-TM output (no graphics)), "Optional - User defined D-cutoff values" (D-cutoff for SignalP-noTM networks, D-cutoff for SignalP-TM networks), and "Truncate sequence" (Default: Truncate sequence to a length of 70 aa - 0 means no truncation). At the bottom, there are "Submit" and "Clear fields" buttons.

CBS >> CBS Prediction Servers >> SignalP

### SignalP 4.0 Server

SignalP 4.0 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:** ePrint of the SignalP 4.0 paper is available, see [Citations](#).

[Background](#) [Article abstracts](#) [Instructions](#) [Output format](#) [Data](#)

#### SUBMISSION

Paste a single amino acid sequence or several sequences in [FASTA](#) format into the field below:

Submit a file in [FASTA](#) format directly from your local disk:  
 No file chosen

**Organism group**

- Eukaryotes
- Gram-negative bacteria
- Gram-positive bacteria

**Method**

- Input sequences may include TM regions
- Input sequences do not include TM regions

**Graphics**

- No graphics
- PNG (inline)
- PNG (inline) and EPS (as links)

**Output format**

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

**Optional - User defined D-cutoff values** (default [scores](#))

D-cutoff for SignalP-noTM networks

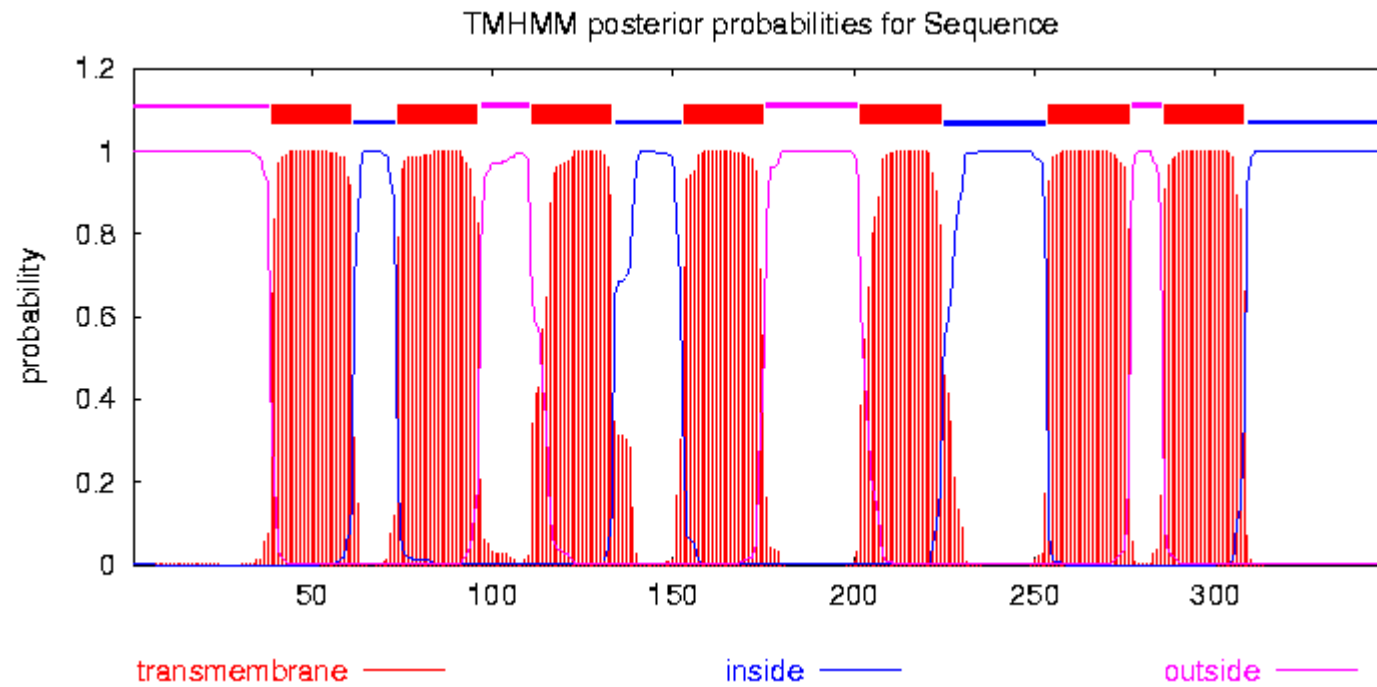
D-cutoff for SignalP-TM networks

**Truncate sequence**

Default: Truncate sequence to a length of 70 aa - 0 means no truncation

## Other prediction tools

### TMHMM: trans-membrane proteins



# Summary

RNA -> Protein:

1. 6-frame/3-frame translation translation
2. ORF detection tool, e.g. TransDecoder

BLAST:

Homologs in other species

PSSM (position specific scoring matrix)

Protein domain with known function

PFAM (a collection of HMM models)

SignalP

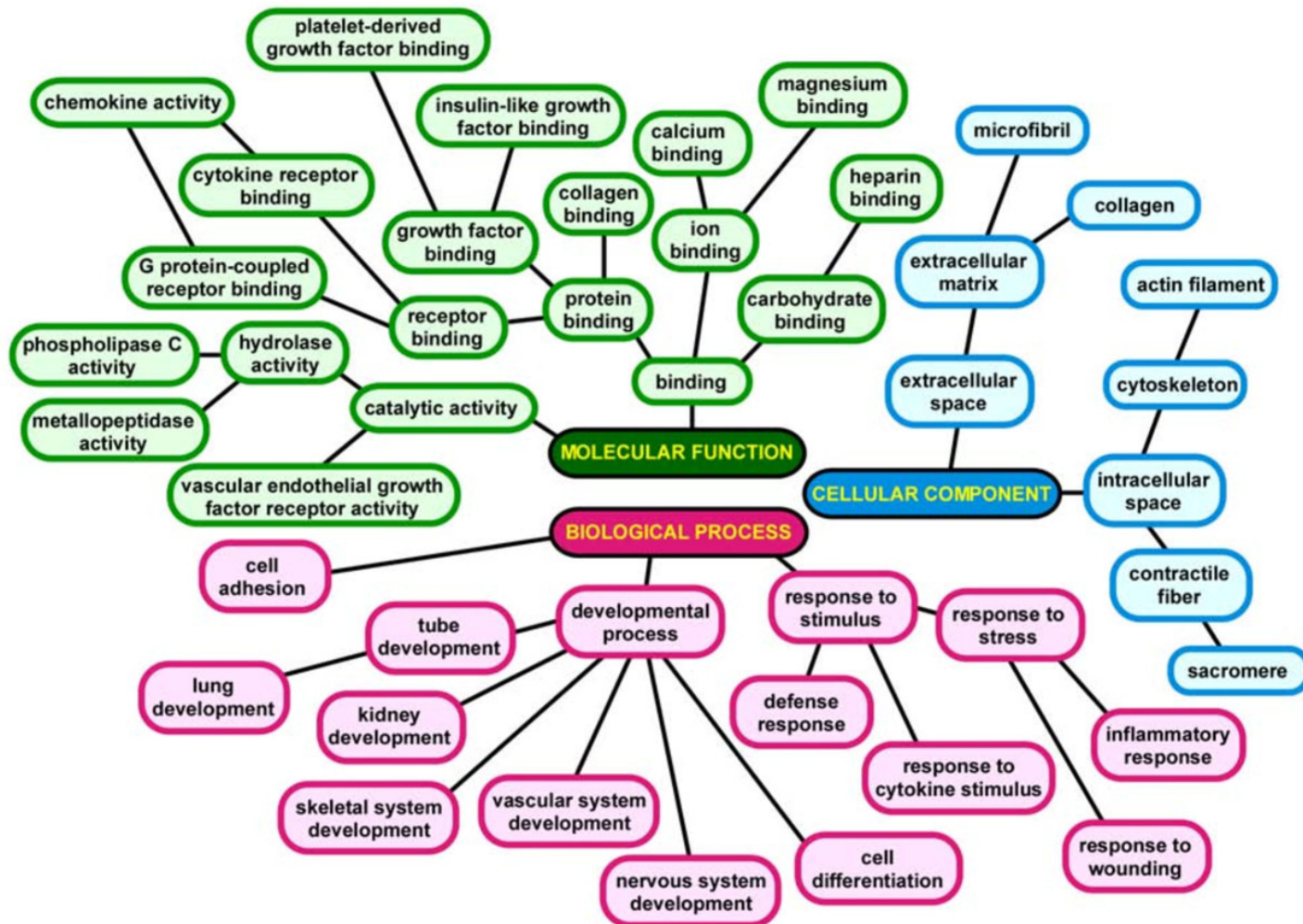
Various protein features, eg. Trans-membrane, signal peptide

TMHMM

Others

Gene ontology

# Gene Ontology

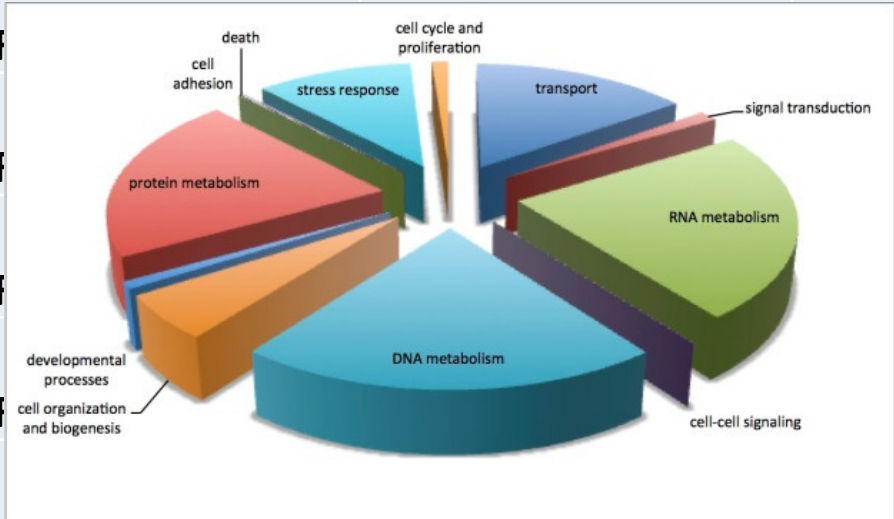
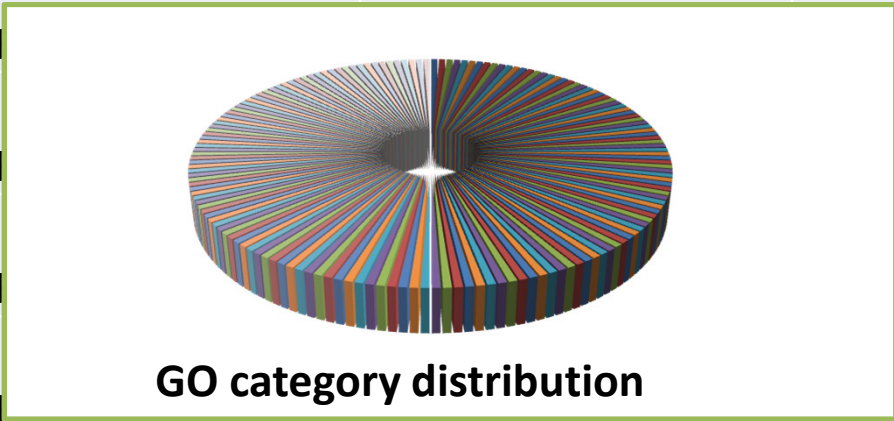


# Gene Ontology

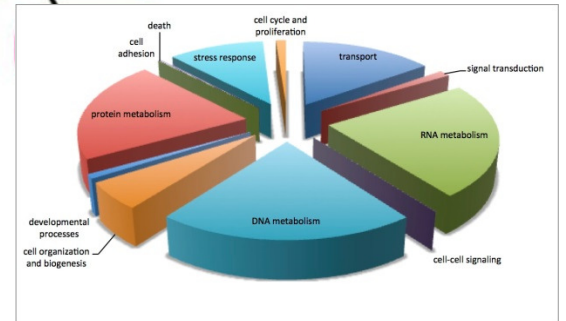
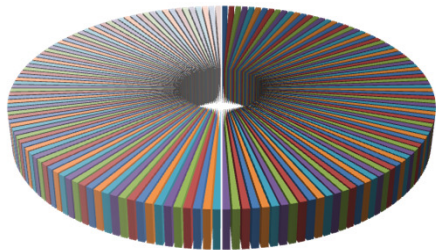
GRMZM2G035341	molecular_function	GO:0008270	zinc ion binding
GRMZM2G035341	molecular_function	GO:0046872	metal ion binding
GRMZM2G035341	cellular_component	GO:0005622	intracellular
GRMZM2G035341	cellular_component	GO:0019005	SCF ubiquitin ligase complex
GRMZM2G035341	biological_process	GO:0009733	response to auxin
GRMZM2G047813	molecular_function	GO:0003677	DNA binding
GRMZM2G047813	cellular_component	GO:0005634	nucleus
GRMZM2G047813	cellular_component	GO:0005694	chromosome
GRMZM2G047813	biological_process	GO:0006259	DNA metabolic process
GRMZM2G047813	biological_process	GO:0034641	cellular nitrogen compound metabolic process

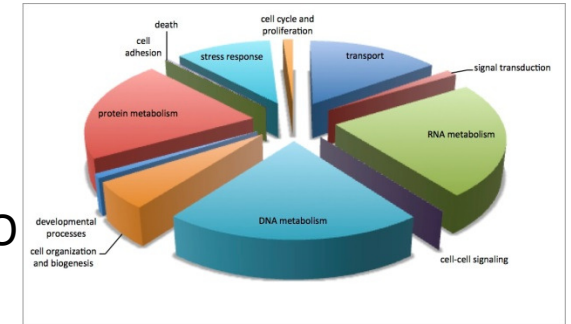
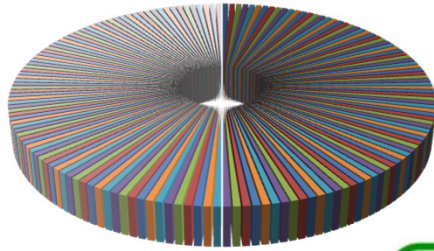
# GO->GOSLIM

GRMZM2G035341	molecular_function	GO:0008270	zinc ion binding
GI		046872	metal ion binding
GI		005622	intracellular
GI		019005	SCF ubiquitin ligase complex
GI		009733	response to auxin
GRMZM2G035341	biological_process	003677	DNA binding
GRMZM2G035341		005634	nucleus
GRMZM2G035341		005694	chromosome
GRMZM2G035341		006259	DNA metabolic process
GRMZM2G035341			cellular nitrogen compound
GKIVIZM2G047815	biological_process	GO:0034641	metabolic process

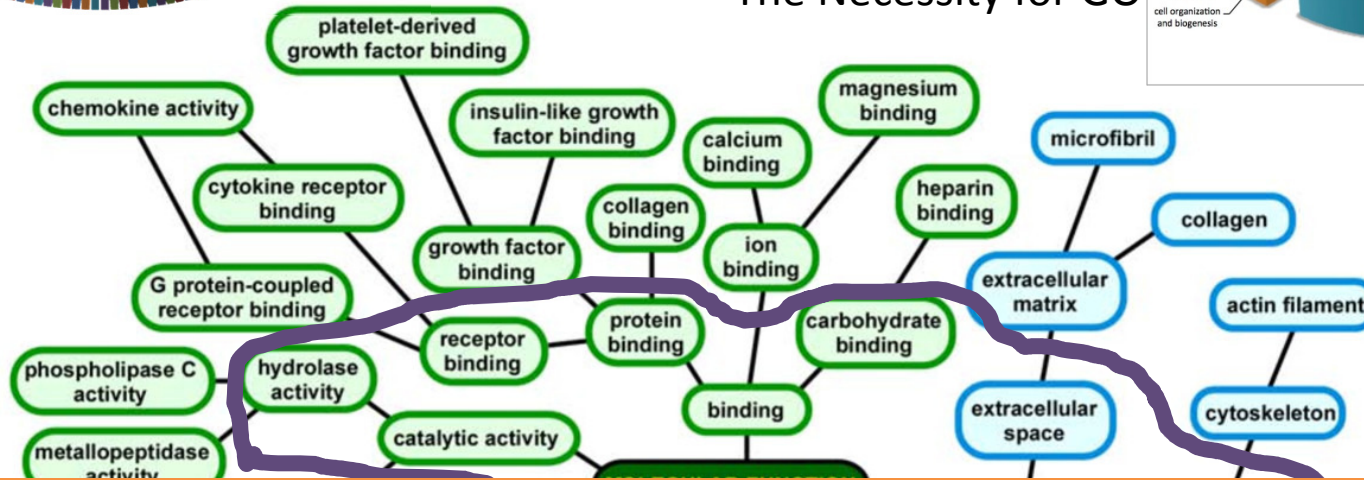


# The Necessity for GO Slim





## The Necessity for GO



## GO Slim

To download premade GO Slim:

<http://www.geneontology.org/GO.slims.shtml>

Create your own GO Slim:

<http://oboedit.org/docs/html/Creating Your Own GO Slim in OBO Edit.htm>



# High throughput gene function prediction

- **BLAST2GO**

Function prediction based on BLAST match to known proteins.

<http://www.blast2go.com>

- **Interproscan**

Function prediction mostly based on PFAM, PSI-BLAST and other motif scanning tools.

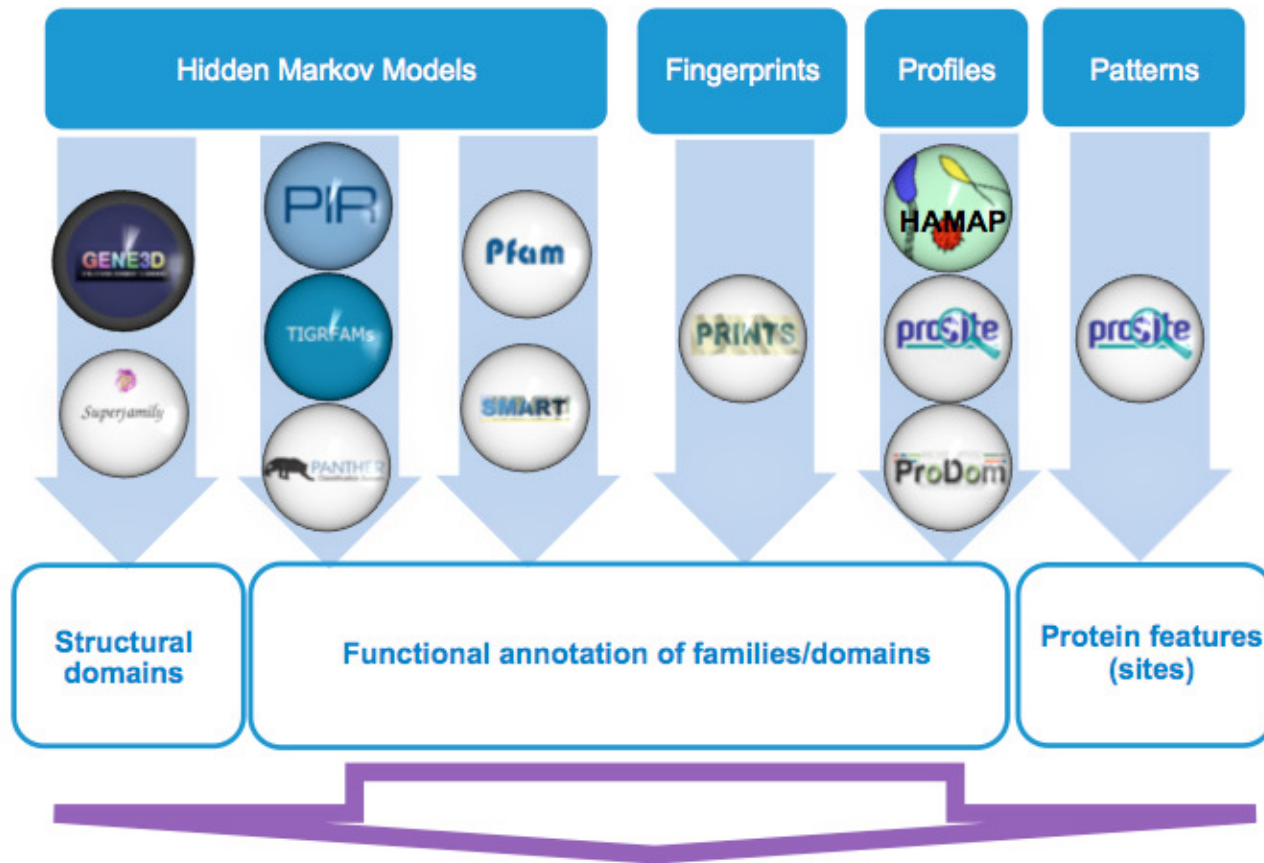
<http://www.ebi.ac.uk/interpro/>

- **Trinotate**

Function prediction based on BLAST, PFAM, SignalP, TMHMM, RNAmmer

<http://trinotate.github.io/>

# InterProScan



# InterProScan

Program Name	Description	Abbreviation
BlastProDom	Scans the families in the ProDom database. ProDom is a comprehensive set of protein domain families automatically generated from the UniProtKB/Swiss-Prot and UniProtKB/TrEMBL sequence databases using psi-blast.	ProDom
FPrintScan	Scans against the fingerprints in the PRINTS database. These fingerprints are groups of motifs that together are more potent than single motifs by making use of the biological context inherent in a multiple motif method.	PRINTS
HMMPiR	Scans the hidden markov models (HMMs) that are present in the PIR Protein Sequence Database (PSD) of functionally annotated protein sequences, PIR-PSD.	PIRSF
HMMPfam	Scans the hidden markov models (HMMs) that are present in the PFAM Protein families database.	PfamA
HMMSmart	Scans the hidden markov models (HMMs) that are present in the SMART domain/domain families database.	SMART
HMMTigr	Scans the hidden markov models (HMMs) that are present in the TIGRFAMs protein families database.	TIGRFAM
ProfileScan	Scans against PROSITE profiles. These profiles are based on weight matrices and are more sensitive for the detection of divergent protein families.	PrositeProfiles
HAMAP	Scans against HAMAP profiles. These profiles are based on weight matrices and are more sensitive for the detection of divergent bacterial, archaeal and plastid-encoded protein families.	HAMAP
PatternScan	PatternScan is a new version of the PROSITE pattern search software which uses new code developed by the PROSITE team.	PrositePatterns
SuperFamily	SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure.	SuperFamily
SignalPHMM		SignalP
TMHMM		TMHMM
HMMPanther		Panther
Gene3D		Gene3d
Phobius		Phobius
Coils		Coils

## Trinotate (Trinity package)

### 1. Predict open-reading-frame (DNA -> protein sequence)

```
TransDecoder -t Trinity.fasta --workdir transDecoder -S
```

### 2. Predict gene function from sequences

#### Trinotate package

- BLAST Uniprot : homologs in known proteins
- PFAM: protein domain)
- SignalP: signal peptide)
- TMHMM: trans-membrane domain
- RNAMMER: rRNA

#### **Output:**

go\_annotations.txt (Gene Ontology annotation file)

Step-by-step guidance: <https://cbsu.tc.cornell.edu/lab/userguide.aspx?a=software&i=143#c>

# BLAST2GO Annotation Steps

Recommended

- **BLAST**: BLAST against “NCBI nr” or Swissprot database;
- **Mapping**: Retrieve GO from annotated homologous genes;
- **Annotation**: Assign GO terms to query sequences.
- **InterProScan (optional)**: Integrate with InterProScan results.

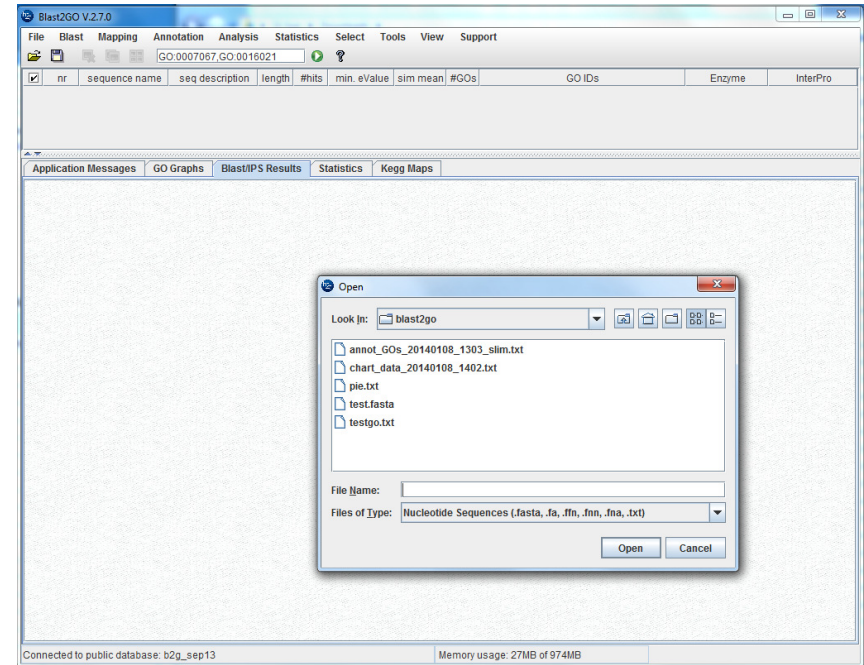
# BLAST2GO

- Do each steps separately.

## BLAST step

Run Command line BLASTX on a BioHPC computer

## BLAST2GO step



BioHPC Lab computer through VNC

# Using BRC Bioinformatics Facility Resource

## 1. Office hour

**1pm to 3pm every Monday, 618 Rhodes Hall**

Signup at: <http://cbsu.tc.cornell.edu/lab/office1.aspx>

## 2. Step-by-step instruction using software on BioHPC computers.

Software page: <http://cbsu.tc.cornell.edu/lab/labsoftware.aspx>

BLAST2GO page: [http://cbsu.tc.cornell.edu/lab/doc/instruction\\_blast2go.htm](http://cbsu.tc.cornell.edu/lab/doc/instruction_blast2go.htm)