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

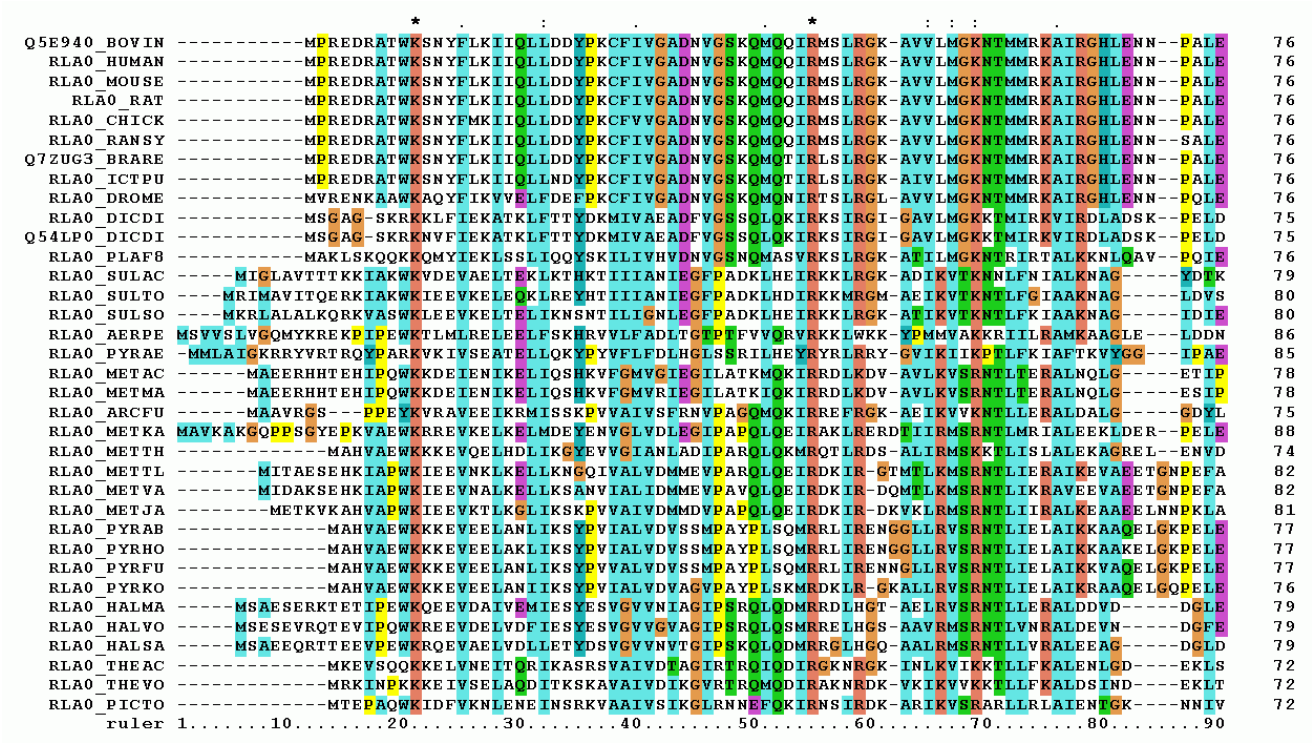
Training the HMM
--train training.fasta : a set of high confidence transcripts
--cd_hit_est : path to CD-hit-est tool, a clustering tool to produce a non-redundant protein set
-G: genetic code

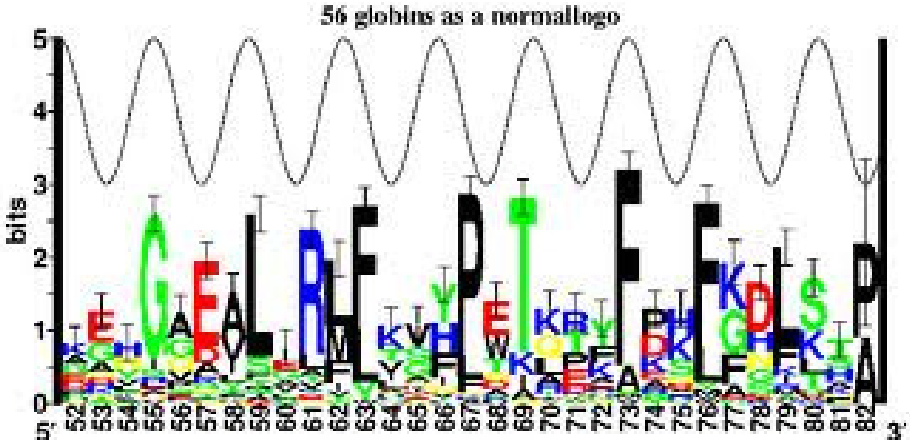
Output:
--retain_long_orfs 900: all ORF longer than 900 nt will be retained
-m 100: minimum protein length

The ORF prediction can refined by homology protein search

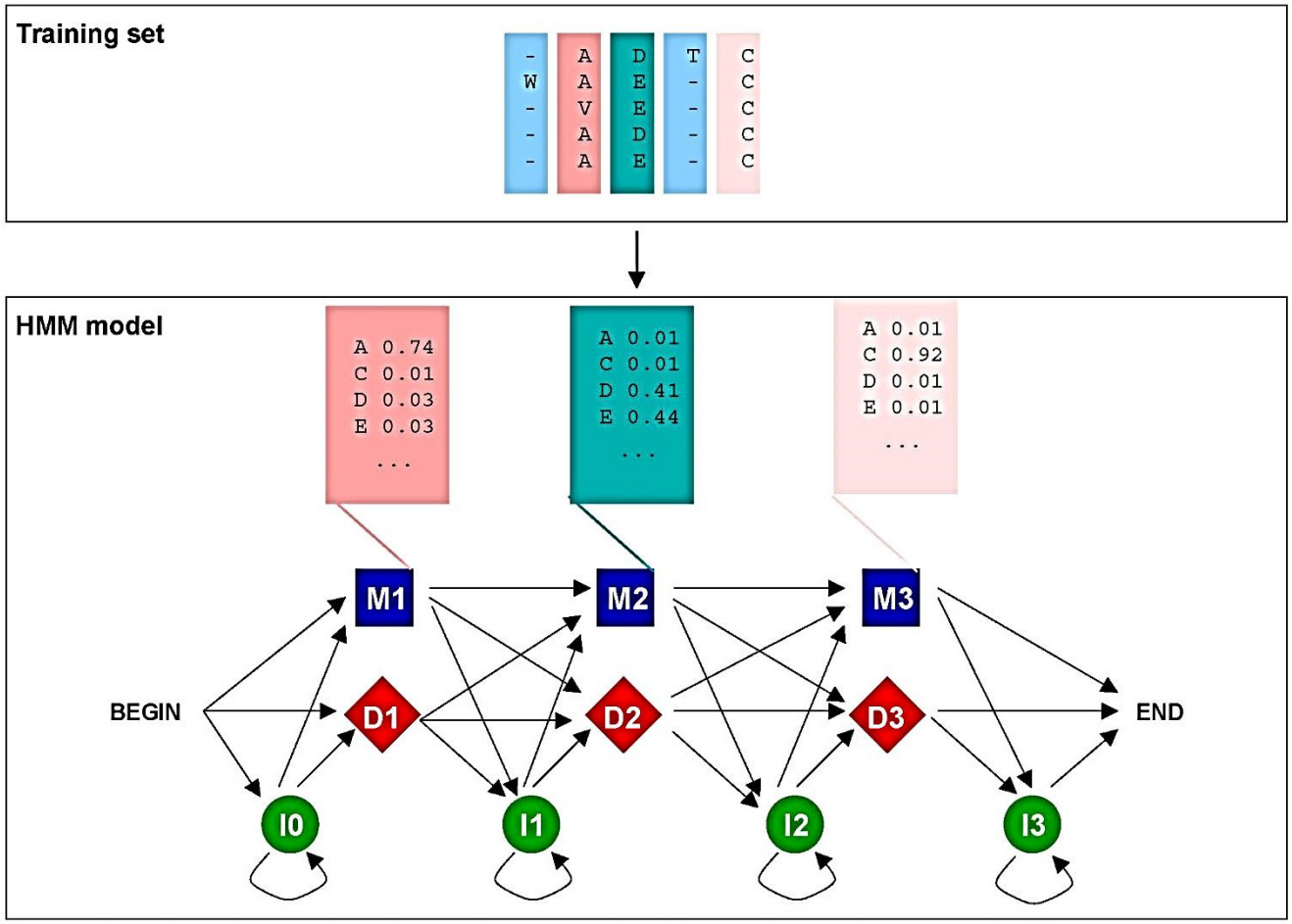
```
blastp -query transdecoder_dir/longest_orfs.pep -db
uniprot_sprot.fasta -max_target_seqs 1 -outfmt 6 -evalue 1e-5 -
num_threads 10 > blastp.outfmt6

hmmscan --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



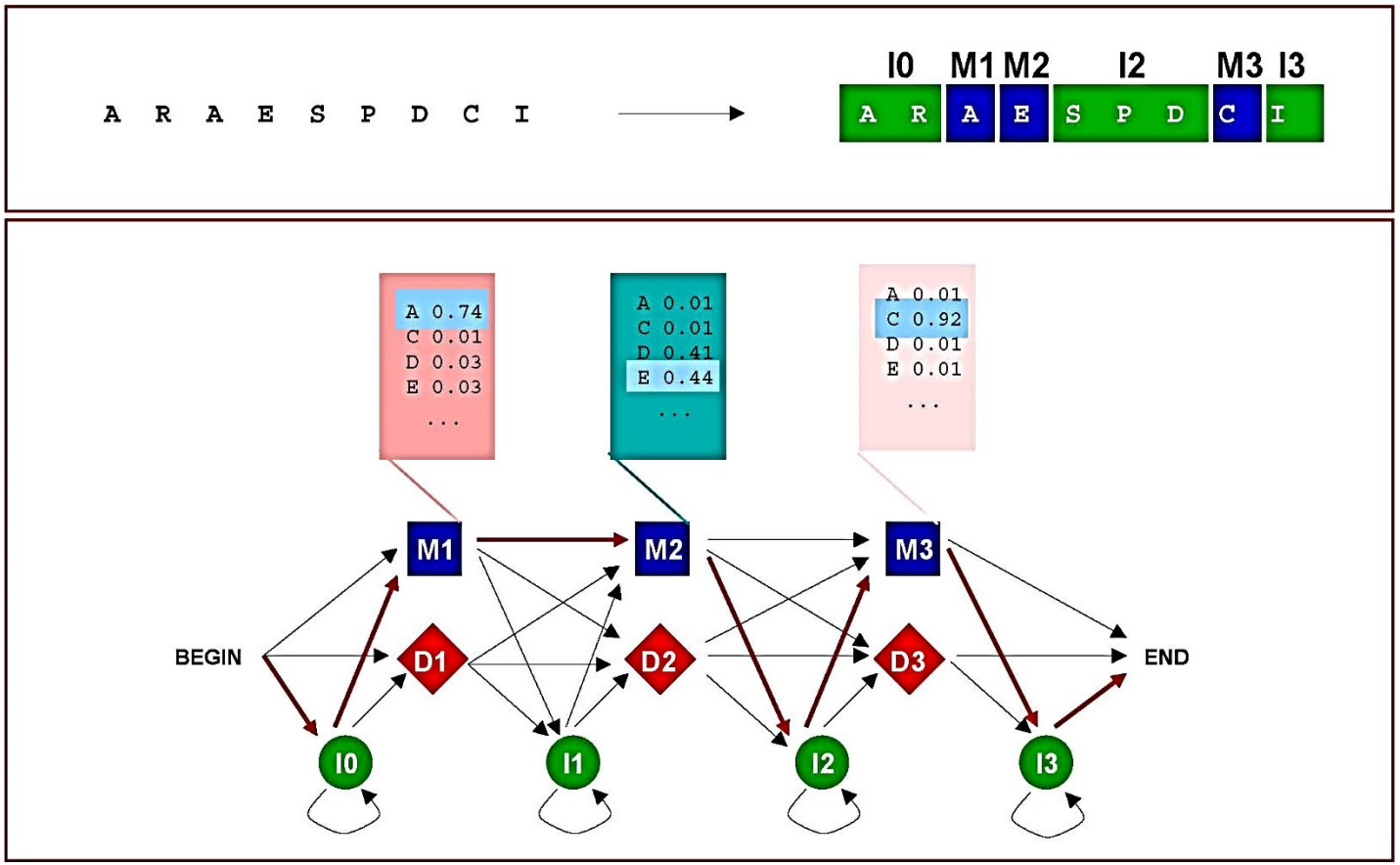


Hidden Markov Model (HMM) is more general than PSSM



Pagni, et al. EMBnet Course 2003

Match a sequence to a model
Application: Function Prediction



Pagni, et al. EMBnet Course 2003

```
>unknown_protein
MALLYRRMSMLLNIILAYIFLCAICVQGSVKQEWAEIGKNVSLECASENEAVAWKLGNQTINKNHTRYKI
RTEPLKSNDDGSENNDSQDFIKYKNVLALLDVNIKDSGNYTCTAQTGQNHSTEFQVRPYLPSKVLQSTPD
RIKRKIKQDVMLYCLIEMYPQNETTNRNLKWLKDGSQFEFLDTFSSISKLNDTHLNFTLEFTEVYKKENG
TYKCTVFDDTGLEITSKEITLFVMEVPQVSIDFAKAVGANKIYLNWTVNDGNDPIQKFFITLQEAGTPTF
TYHKDFINGSHTSYILDHFKPNTTYFLRIVGKNSIGNGQPTQYPQGITTLSYDPIFIPKVETTGSTASTI
TIGWNPPPPDLIDYIQYYELIVSESGEVPKVIEEAIYQQNSRNLPYMFDKLKTATDYEFRVRACSDLTKT
CGPWSENVNGTTMDGVATKPTNLSIQCHHDNVTRGNSIAINWDVPKTPNGKVVSYLIHLLGNPMSTVDRE
MWGPKIRRIDEPHHKTLYESVSPNTNYTVTVSAITRHKKNGEPATGSCLMPVSTPDAIGRTMWSKVNLDS
KYVLKLYLPKISERNGPICCYRLYLVRINNDNKELPDPEKLNIATYQEVHSDNVTRSSAYIAEMISSKYF
RPEIFLGDEKRFSENNDIIRDNDEICRKCLEGTPFLRKPEIIHIPPQGSLSNSDSELPILSEKDNLIKGA
NLTEHALKILESKLRDKRNAVTSDENPILSAVNPNVPLHDSSRDVFDGEIDINSNYTGFLEIIVRDRNNA
LMAYSKYFDIITPATEAEPIQSLNNMDYYLSIGVKAGAVLLGVILVFIVLWVFHHKKTKNELQGEDTLTL
RDSLRALFGRRNHNHGHEIISGNHKGEDACDIHDLDLENAYKNPHKDTDYGFLDEYEMLDNPEGDDTIKN
SDLKENACKNRYP
EQHLEIIVMLTNL
RRQITQYHYLTWK
SVSIYNTVCDLRH
EKLLATADEISKS
QDPLENTIGDFWR
TNCKIDDTLKVTQ
VAMCILVQHLRLE
```

PFAM

a pre-constructed HMM model database

for protein function domain prediction

wellcome trust sanger institute

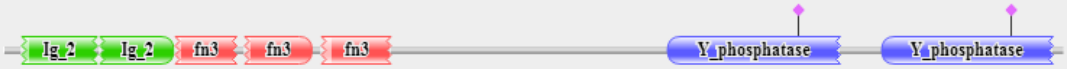
HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

Pfam keyword search Go

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

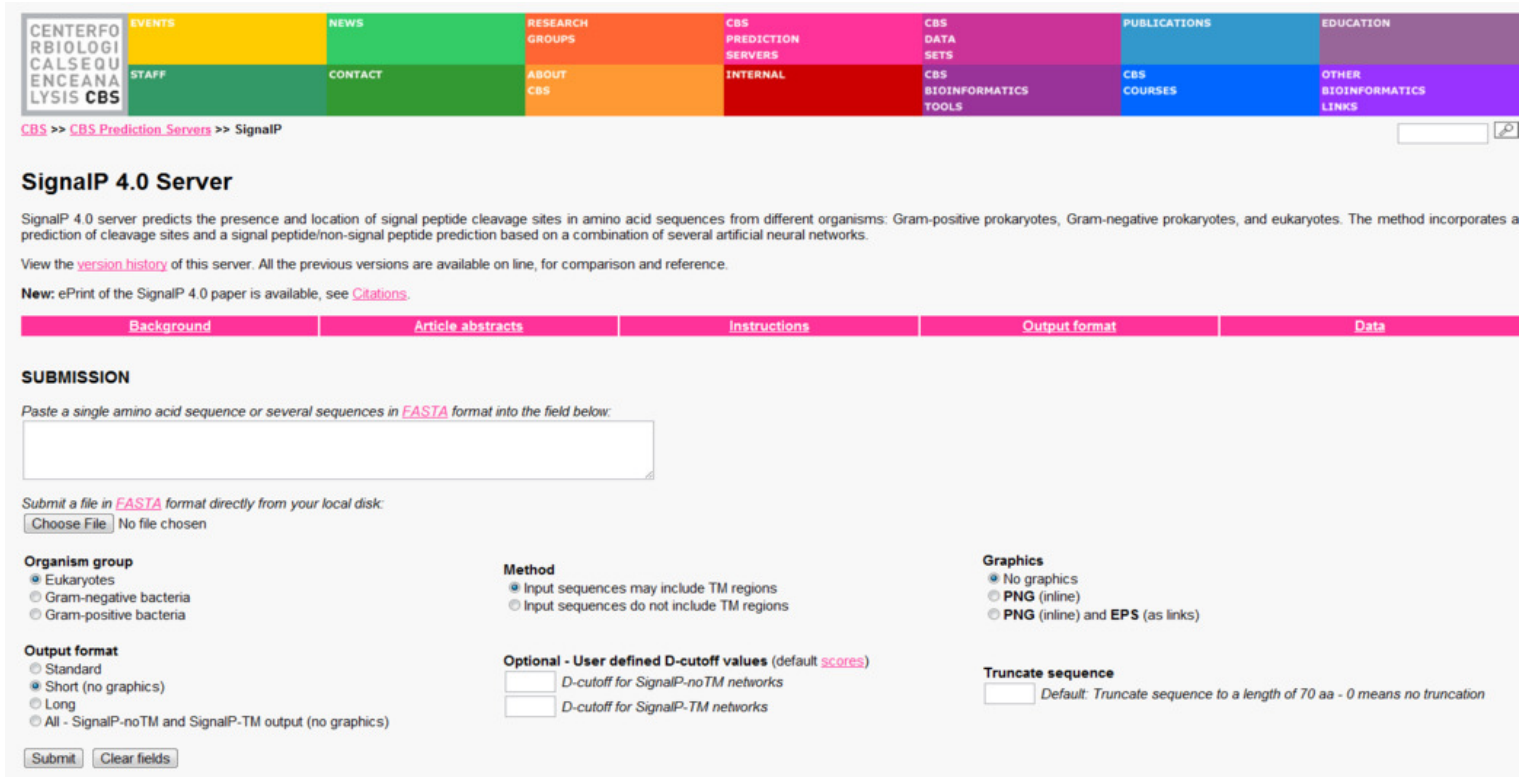
Comments or questions on the site? Send a mail to **pfam-help@sanger.ac.uk**. Our **cookie policy**.

The Wellcome Trust

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

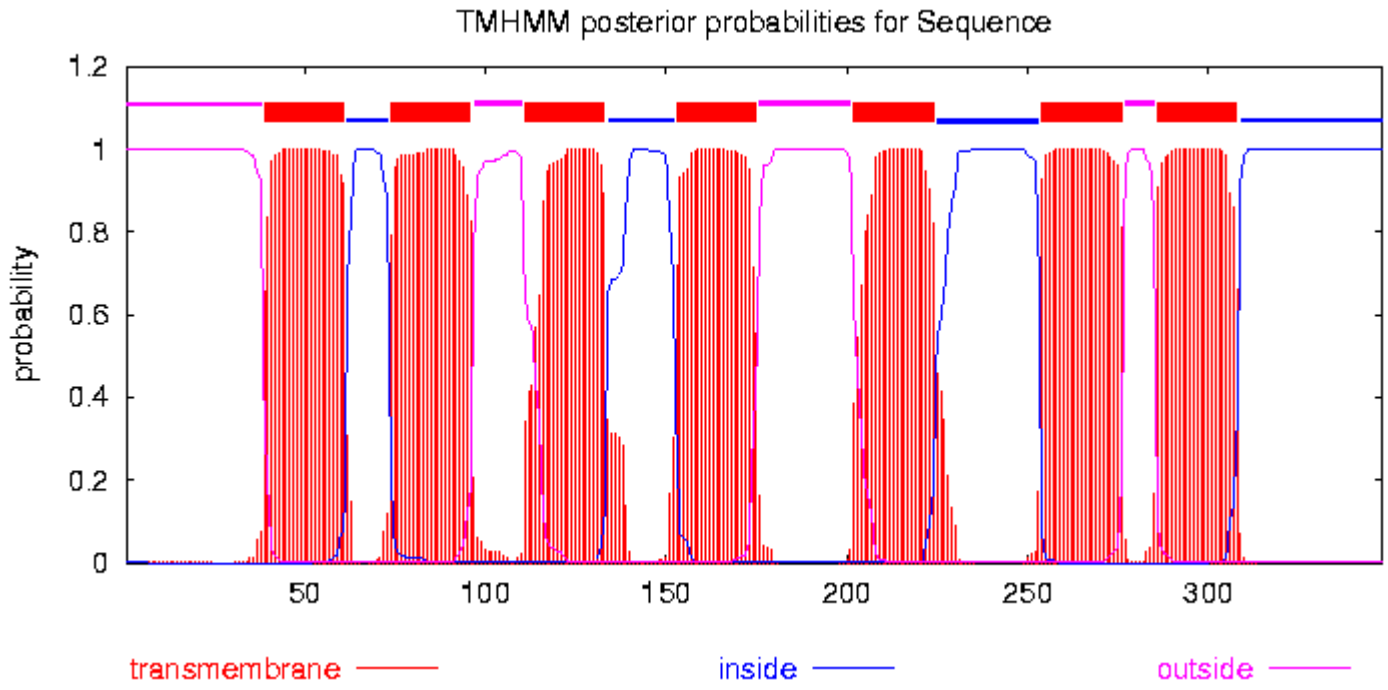
Other prediction tools

SignalP: Predicting Signal Peptide



Other prediction tools

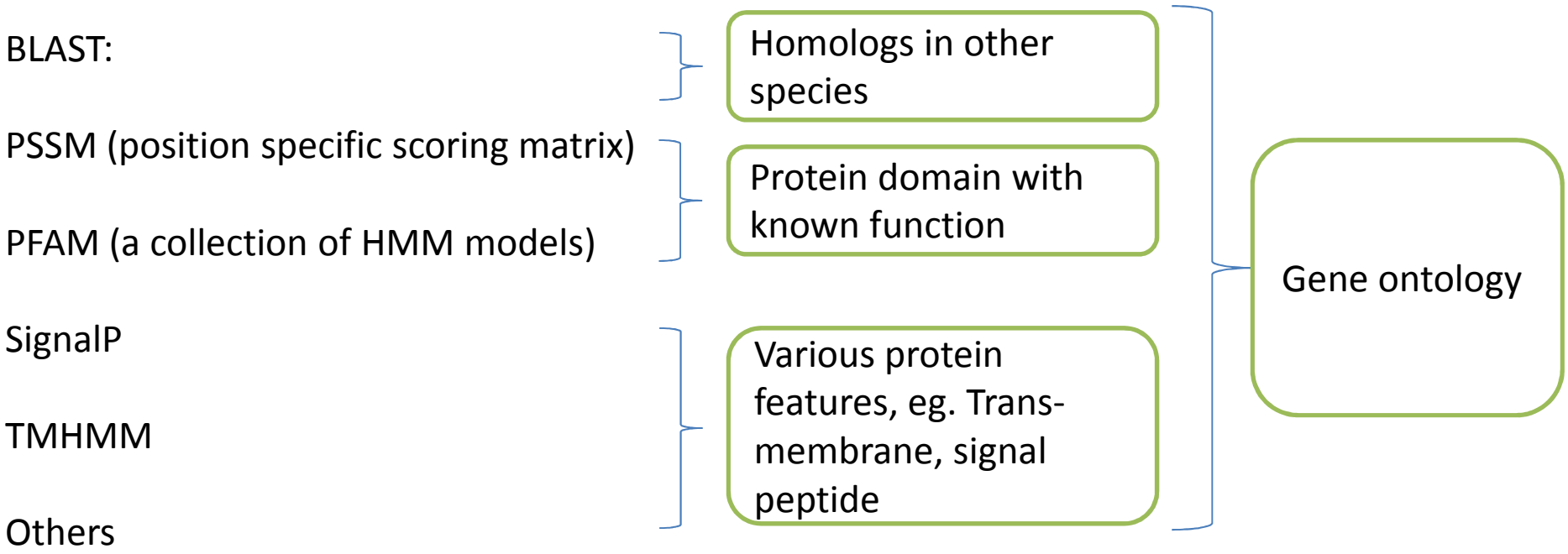
TMHMM: trans-membrane proteins



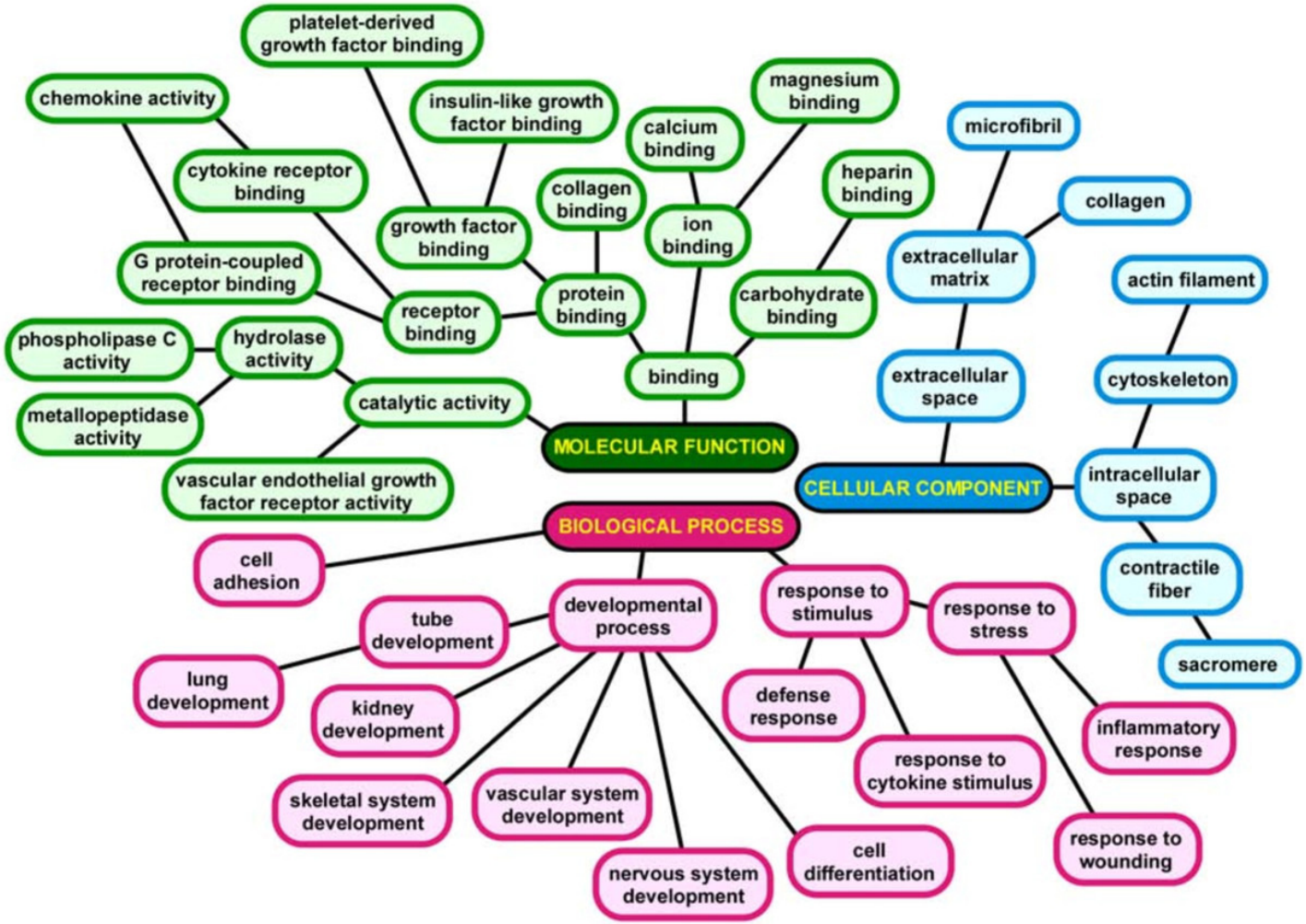
Summary

RNA -> Protein:

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



Gene Ontology

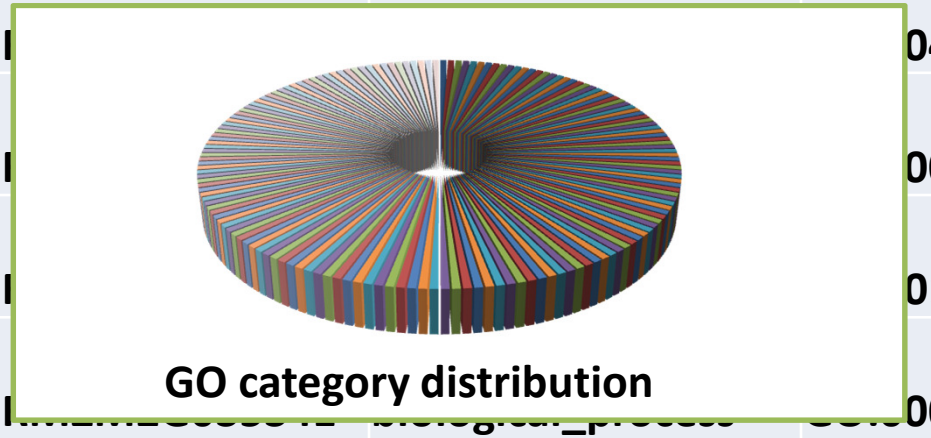


Gene Ontology

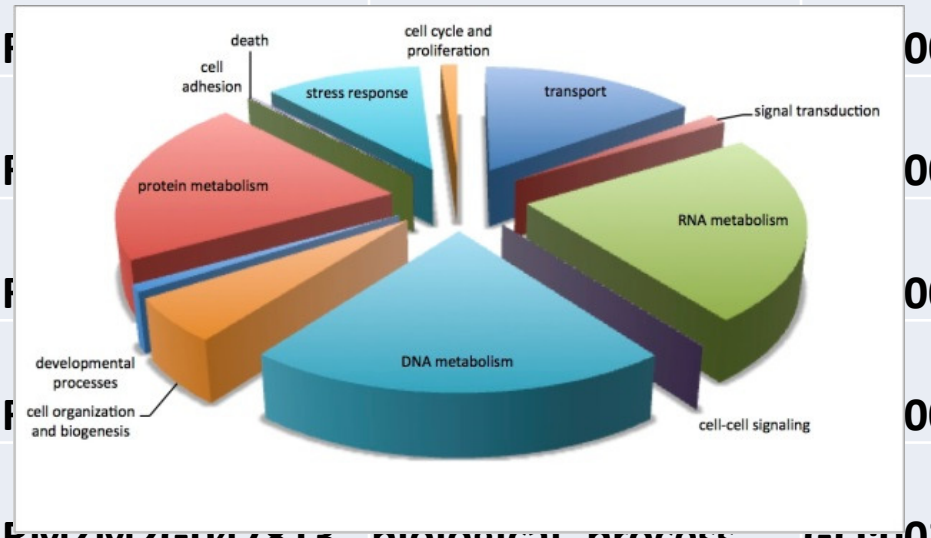
GRMZM2G035341	molecular_function	GO:0008270	zinc ion binding
GRMZM2G035341	molecular_function	GO:0046872	metal ion binding
GRMZM2G035341	cellular_component	GO:0005622	intracellular
GRMZM2G035341	cellular_component	GO:0019005	SCF ubiquitin ligase complex
GRMZM2G035341	biological_process	GO:0009733	response 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

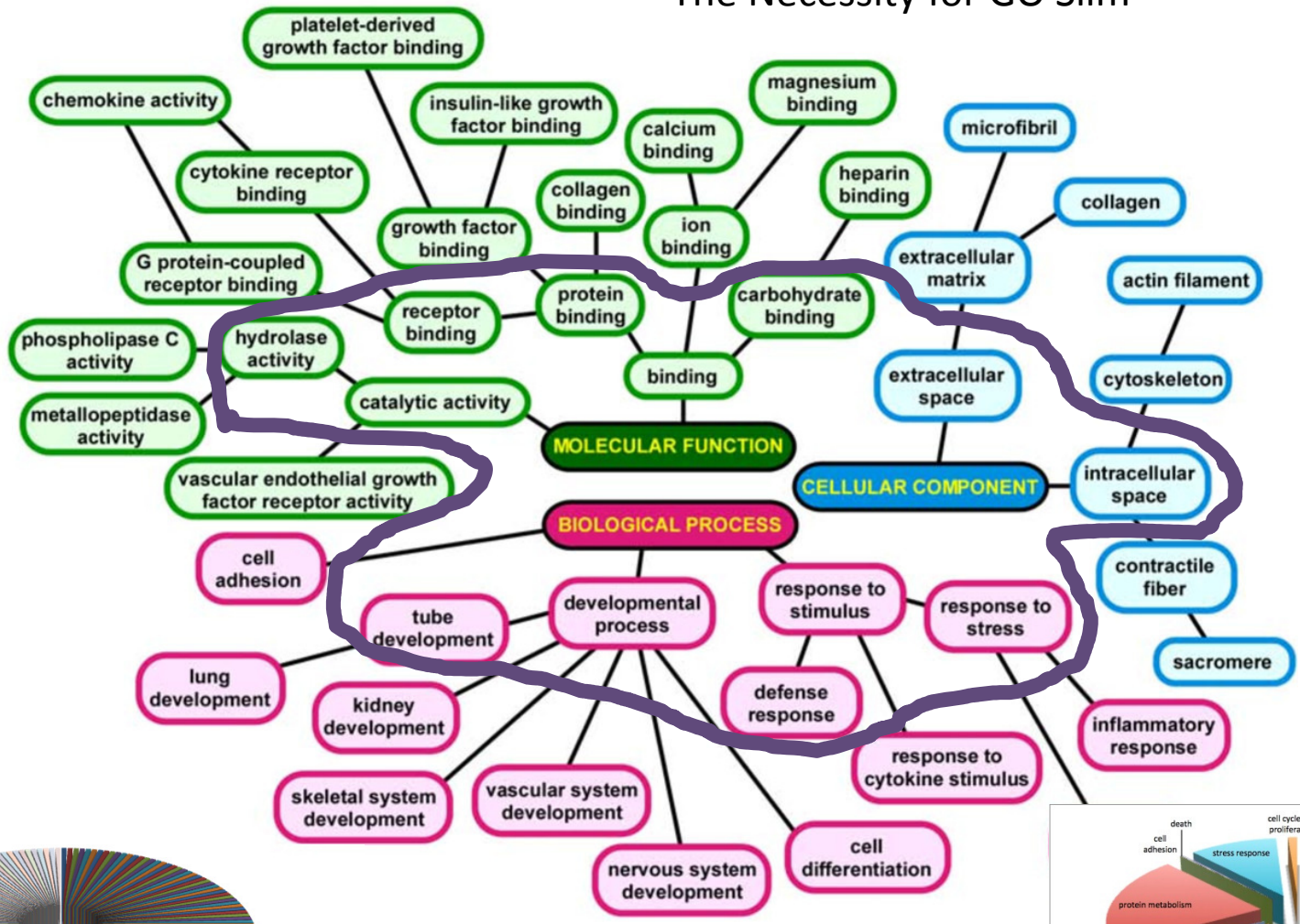
GRMZM2G035341	molecular_function	GO:0008270	zinc ion binding
GI		046872	metal ion binding
GI		005622	intracellular
GI		019005	SCF ubiquitin ligase complex
GI		009733	response to auxin
GI		003677	DNA binding
GI		005634	nucleus
GI		005694	chromosome
GI		006259	DNA metabolic process
GRMZM2G047813	biological_process	GO:0034641	cellular nitrogen compound metabolic process

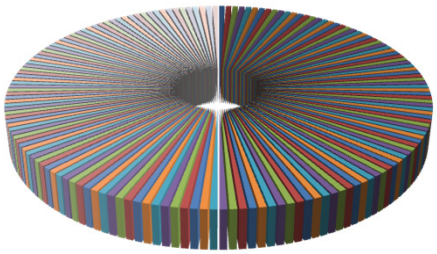


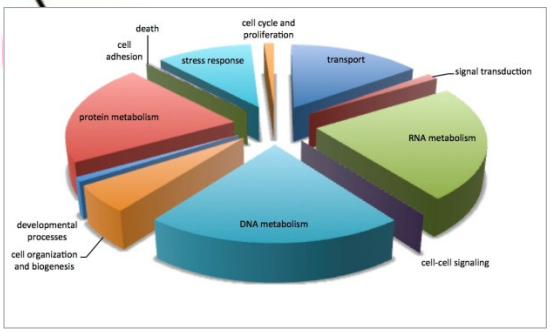
GO category distribution

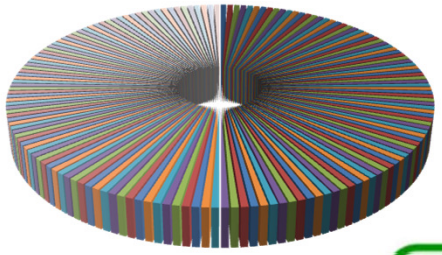


The Necessity for GO Slim

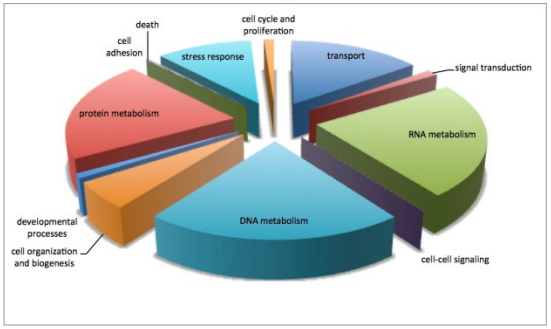


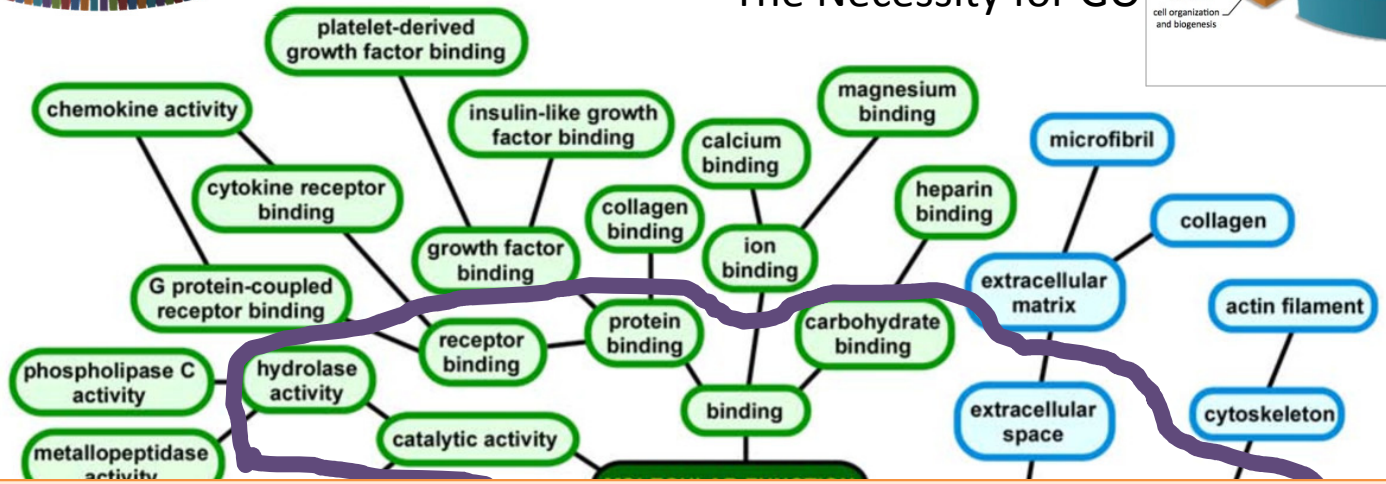






The Necessity for GO





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 scaning tools.

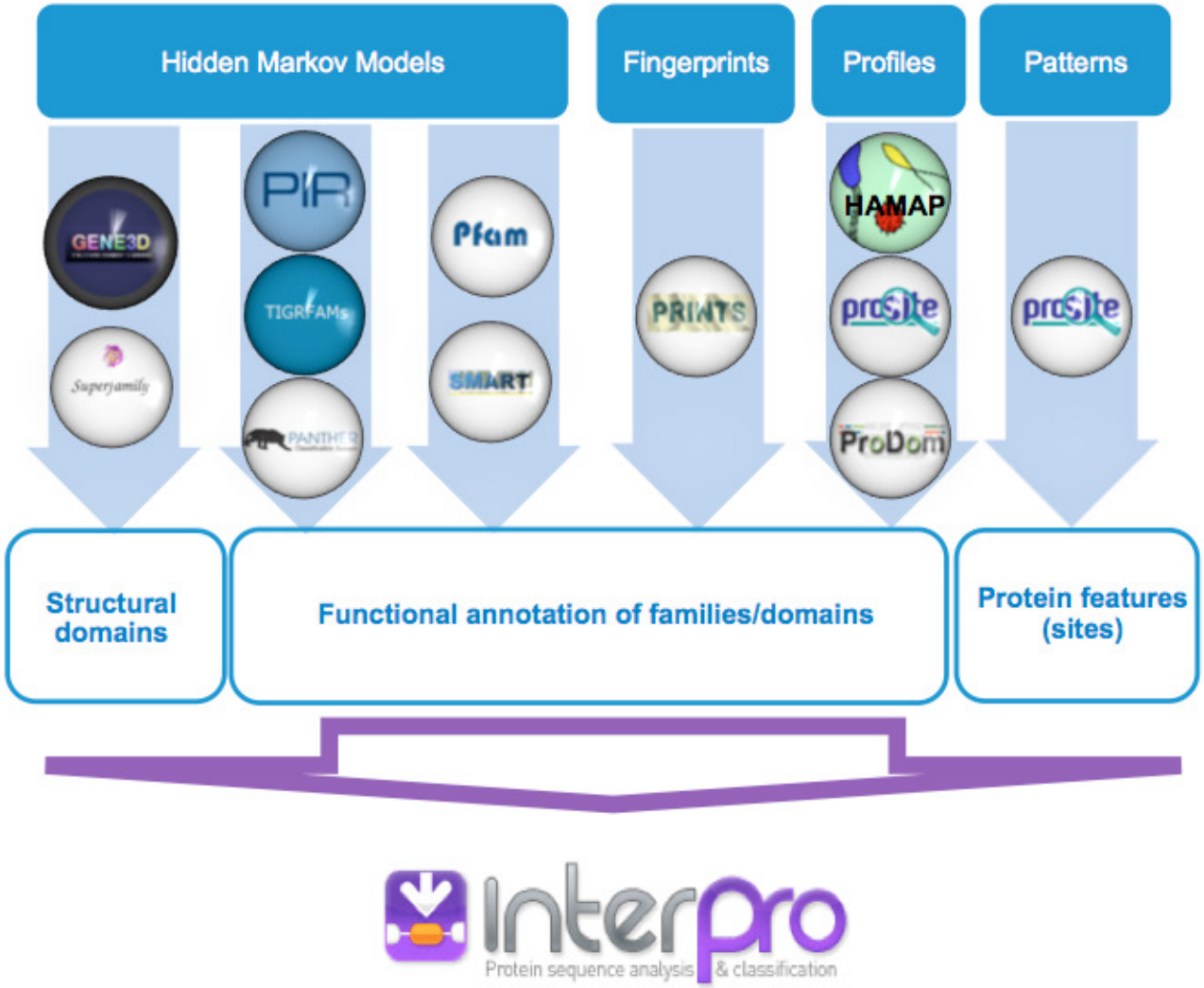
http://www.ebi.ac.uk/interpro/

- **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.

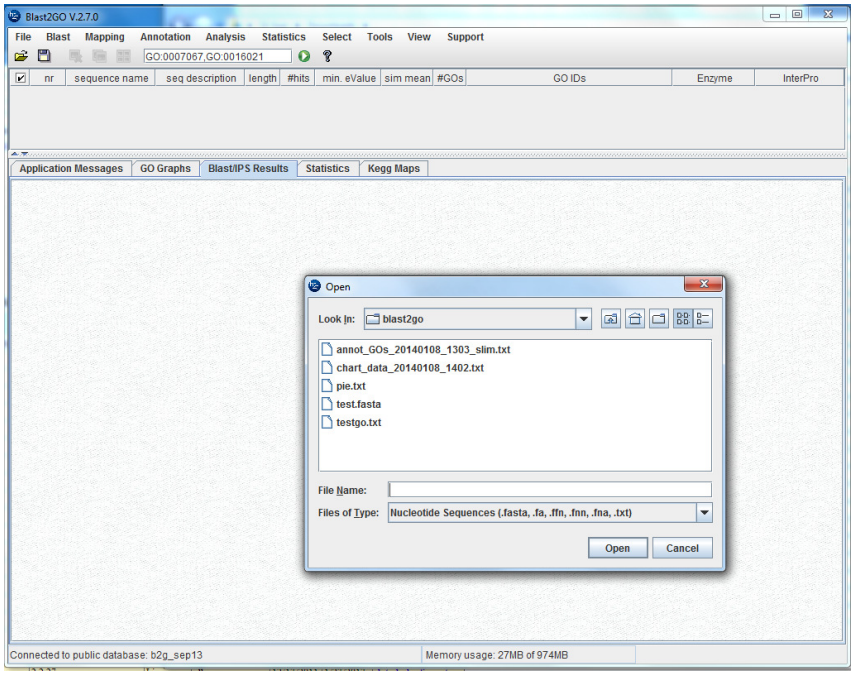
BLAST2GO

- **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