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 be 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
Match a sequence to a model
Application: Function Prediction

Pagni, et al. EMBnet Course 2003
>unknown_protein
MALLYRMSMLNIIAYIFLCAICVQGSKYQEWAIKGKNSLECASNEAAMWKLGRNQTIKNKHTRYKIRTEPLKSNDNSEEENSDQDFIKYKVNVLALDLVNIKDSGNYTCAQTGQHSTEFQVRPYLPSKVLQSTPD
RIKRKIKQDMYCLIMYPQNETTRNLKWKDGQSEFEJLTFSISIKLNDLHNLTLEFTSTEYVQENG
TYKCTVFDDGTIEISITFLMVEMVPQVSIDFAKAVGANLITYLWINTVNGNPIQFFITLQAGTFPT
TYHKDFINGSHTYLDHKFNPNTTYFLRIVGKNSINGQPTQYQGQITLSDYDFIPFIEPKVETGTASTI
TIGWNPPIPDPDDIOITQQYQLIVSESEGGEPKVIIEAAYQYQSNRLNPMYDKLXATDYEYFRVACSDLPT
CGPWSEVNYITTMDGVTTPNLSDIQCHHNVTRGNSTAIWQNDPVTPGKRVSVYILHLGNPMSVTDRE
MWGPKIRIREPEHHTLYIESVSPNTNTVTMTAITEHKKNGEPATGSLMPVSDAPIAGTRMWSKVNLDS
KYVLKLYLPKISENGPICYRLVLYRINNKELDPPEKLNIATYGEVHSNVTSSAYIAEMISSKYF
RPEIFLGDEKFRSENNDIODRTNECMCLETGFPLRKPEIIHPPQGSLSNSDELPLESEKEOVLKGA
NLTEHAIKILESDKRNVTDEPILS4AVNPVPHDSRSDFEGIEIDINSNYGFLEIVRDRNNA
LMAYSKYFDIIIPATEAEPIGQSLNMDYLYLSIGKAVAGVLGIVLWFFHHKTKELNLQGQDTTLT
RLDSLRALGRHKNHWSCEGTGEGIDGAGIQPNIKLRNMWGAHKOGVTKSWLWMKRVNDTTEQNT
SDLKENACKNYRPFGRQITQYLTVKNQETTEDEVDLMLQPHDPNEQSFENLPYRTNLQKPRNLN
QRQPLTIGDFWMNQERDLRAEKLADEISKEQKQLERQQLQDYQQRKLLQVAMCILVQHLRL

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

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
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

TMHMM

Others

Gene ontology

Various protein features, eg. Transmembrane, signal peptide
Gene Ontology
### Gene Ontology

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Type</th>
<th>GO Identifier</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRMZM2G035341</td>
<td>molecular_function</td>
<td>GO:0008270</td>
<td>zinc ion binding</td>
</tr>
<tr>
<td>GRMZM2G035341</td>
<td>molecular_function</td>
<td>GO:0046872</td>
<td>metal ion binding</td>
</tr>
<tr>
<td>GRMZM2G035341</td>
<td>cellular_component</td>
<td>GO:0005622</td>
<td>intracellular</td>
</tr>
<tr>
<td>GRMZM2G035341</td>
<td>cellular_component</td>
<td>GO:0019005</td>
<td>SCF ubiquitin ligase complex</td>
</tr>
<tr>
<td>GRMZM2G035341</td>
<td>biological_process</td>
<td>GO:0009733</td>
<td>response to auxin</td>
</tr>
<tr>
<td>GRMZM2G047813</td>
<td>molecular_function</td>
<td>GO:0003677</td>
<td>DNA binding</td>
</tr>
<tr>
<td>GRMZM2G047813</td>
<td>cellular_component</td>
<td>GO:0005634</td>
<td>nucleus</td>
</tr>
<tr>
<td>GRMZM2G047813</td>
<td>cellular_component</td>
<td>GO:0005694</td>
<td>chromosome</td>
</tr>
<tr>
<td>GRMZM2G047813</td>
<td>biological_process</td>
<td>GO:0006259</td>
<td>DNA metabolic process</td>
</tr>
<tr>
<td>GRMZM2G047813</td>
<td>biological_process</td>
<td>GO:0034641</td>
<td>cellular nitrogen compound metabolic process</td>
</tr>
<tr>
<td>GO ID</td>
<td>GO Category</td>
<td>Annotation</td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>------------------------------------</td>
<td>-----------------------------</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G035341</td>
<td>molecular_function</td>
<td>GO:0008270 zinc ion binding</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0046872 metal ion binding</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cellular_component</td>
<td>GO:0005622 intracellular</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0019005 SCF ubiquitin ligase complex</td>
<td></td>
</tr>
<tr>
<td></td>
<td>biological_process</td>
<td>GO:0009733 response to auxin</td>
<td></td>
</tr>
<tr>
<td>GRMZM2G047813</td>
<td>molecular_function</td>
<td>GO:003677 DNA binding</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cellular_component</td>
<td>GO:005634 nucleus</td>
<td></td>
</tr>
<tr>
<td></td>
<td>biological_process</td>
<td>GO:005694 chromosome</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:006259 DNA metabolic process</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>GO:0034641 cellular nitrogen compound metabolic process</td>
<td></td>
</tr>
</tbody>
</table>
The Necessity for GO Slim
The Necessity for GO Slim

To download premade GO Slim:

Create your own GO Slim:
http://oboedit.org/docs/html/Created_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](http://www.blast2go.com)

- **Interproscan**
  Function prediction mostly based on PFAM, PSI-BLAST and other motif scanning tools.
  [http://www.ebi.ac.uk/interpro/](http://www.ebi.ac.uk/interpro/)

- **Trinotate**
  Function prediction based on BLAST, PFAM, SignalP, TMHMM, RNAmermer
  [http://trinotate.github.io/](http://trinotate.github.io/)
<table>
<thead>
<tr>
<th>Program Name</th>
<th>Description</th>
<th>Abbreviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>BlastProDom</td>
<td>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.</td>
<td>ProDom</td>
</tr>
<tr>
<td>FPrintScan</td>
<td>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.</td>
<td>PRINTS</td>
</tr>
<tr>
<td>HMMPIR</td>
<td>Scans the hidden markov models (HMMs) that are present in the PIR Protein Sequence Database (PSD) of functionally annotated protein sequences, PIR-PSD.</td>
<td>PIRSF</td>
</tr>
<tr>
<td>HMMPfam</td>
<td>Scans the hidden markov models (HMMs) that are present in the PFAM Protein families database.</td>
<td>PfamA</td>
</tr>
<tr>
<td>HMMSmart</td>
<td>Scans the hidden markov models (HMMs) that are present in the SMART domain/domain families database.</td>
<td>SMART</td>
</tr>
<tr>
<td>HMMTigr</td>
<td>Scans the hidden markov models (HMMs) that are present in the TIGRFAMs protein families database.</td>
<td>TIGRFAM</td>
</tr>
<tr>
<td>ProfileScan</td>
<td>Scans against PROSITE profiles. These profiles are based on weight matrices and are more sensitive for the detection of divergent protein families.</td>
<td>PrositeProfiles</td>
</tr>
<tr>
<td>HAMAP</td>
<td>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.</td>
<td>HAMAP</td>
</tr>
<tr>
<td>PatternScan</td>
<td>PatternScan is a new version of the PROSITE pattern search software which uses new code developed by the PROSITE team.</td>
<td>PrositePatterns</td>
</tr>
<tr>
<td>SuperFamily</td>
<td>SUPERFAMILY is a library of profile hidden Markov models that represent all proteins of known structure.</td>
<td>SuperFamily</td>
</tr>
<tr>
<td>SignalPHMM</td>
<td></td>
<td>SignalP</td>
</tr>
<tr>
<td>TMHMM</td>
<td></td>
<td>TMHMM</td>
</tr>
<tr>
<td>HMMMPanther</td>
<td></td>
<td>Panther</td>
</tr>
<tr>
<td>Gene3D</td>
<td></td>
<td>Gene3d</td>
</tr>
<tr>
<td>Phobius</td>
<td></td>
<td>Phobius</td>
</tr>
<tr>
<td>Coils</td>
<td></td>
<td>Coils</td>
</tr>
</tbody>
</table>
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

• **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


   Software page: http://cbsu.tc.cornell.edu/lab/labsoftware.aspx
   BLAST2GO page: http://cbsu.tc.cornell.edu/lab/doc/instruction_blast2go.htm