2nd International *Blastocystis* Conference
9-12th October 2018
Bogotá, Colombia

Workshop session 7: *Blastocystis* comparative genomics and evolution
Andrew J. Roger: Andrew.Roger[at]Dal.Ca

Comparative genomics is a fast-moving field that is heavily dependent on technical advances in DNA sequencing technology and bioinformatics method development. The kind and quality of genome and transcriptome sequence data and genome assemblies/annotations available for *Blastocystis* subtypes in public databases is rapidly changing and will continue to do so.

Three reasonable quality draft genomes with annotated genomes and predicted gene sets are published and publicly available. These include:
***Blastocystis* sp. ST7- Singapore isolate B (genome contigs, EST data and predicted genes)**
***Blastocystis* sp. ST4-WR1 isolate (genome contigs and predicted genes)**
***Blastocystis* sp. ST1- NandII isolate (genome contigs and predicted genes)**

However there are also unpublished draft genome sequences (with no gene predictions) for:
***Blastocystis* sp. ST2 (Flemming isolate)**
***Blastocystis* sp. ST3 (ZGR isolate)**
***Blastocystis* sp. ST6 (SSI:754 isolate)**
***Blastocystis* sp. ST8 (Dmp/08-128 isolate)**
***Blastocystis* sp. ST9 (F5323 isolate)**

In this workshop I will cover a number of topics related to the bioinformatic analysis of *Blastocystis* gene, genome and predicted proteome data.

Teaching outcomes:

After this short workshop you should be able to:

- Understand how to use the **NCBI website** to download gene or genome sequences for *Blastocystis* subtypes and do **BLAST** searches for homologs in *Blastocystis* genomes
- View and extract information from an annotated genome contig in **Artemis** (or Integrative Genomics Viewer)
- Navigate the use of a number of tools (**eggNOG-mapper**, **Interproscan**, etc) to help with functional annotation of a set of proteins of interest by searching orthologous groups and profile HMMs
- Use tools to help you predict the subcellular localizations of proteins of interest within *Blastocystis* cells
- Use tools (MAFFT) to do multiple alignments and investigate the phylogenies of particular genes
- Find more information about particular classes of proteins such as proteases and carbohydrate active enzymes (CAZy)

More reading on bioinformatics/genomics theory and methods

- The BLAST Sequence analysis tool: https://www.ncbi.nlm.nih.gov/books/NBK153387/
- What is a hidden Markov Model (HMM)? https://www.nature.com/articles/nbt1004-1315
- Profile HMMs for protein families: http://www.biology.wustl.edu/gcg/hmmanalysis.html

Part 1 – Browsing genome assemblies downloaded from databases

Tools/databases used:
NCBI (http://www.ncbi.nlm.nih.gov)
Artemis (https://www.sanger.ac.uk/science/tools/artemis)

Introduction
When genomes are sequenced and published, they are usually deposited in a database such as GenBank at NCBI (or EMBL-EBI or DDBJ). The files that are deposited end up in GenBank format (.gbff files) as a Biosample within a Bioproject. They are also usually accessible from the "Genome" database. Annotated genomes (e.g. ST7, ST4, and ST1 above) have associated information about locations of genes and other features in the sequence. If the genomes are just contigs and are not annotated with predicted genes etc. (like ST2, 3, 6, 8 and 9), then these files are just nucleotide sequences only with no 'Features' mapped in the GenBank file. Here you will investigate these files in GenBank in a genome viewer/editor **Artemis.**

Steps to follow:

1. Go to NCBI (http://www.ncbi.nlm.nih.gov) and select "Genome" from the search menu.
2. Type in **Blastocystis** and hit search. You will see two entries, one for **Blastocystis hominis** and one for **Blastocystis**.
3. Click on each of them and look carefully at the information on the pages.
4. Ask yourself what genome assemblies are present and for what subtypes?
5. Go to the *Blastocystis* sp. ATCC 50177/NandII assembly and look in the INDSC column.

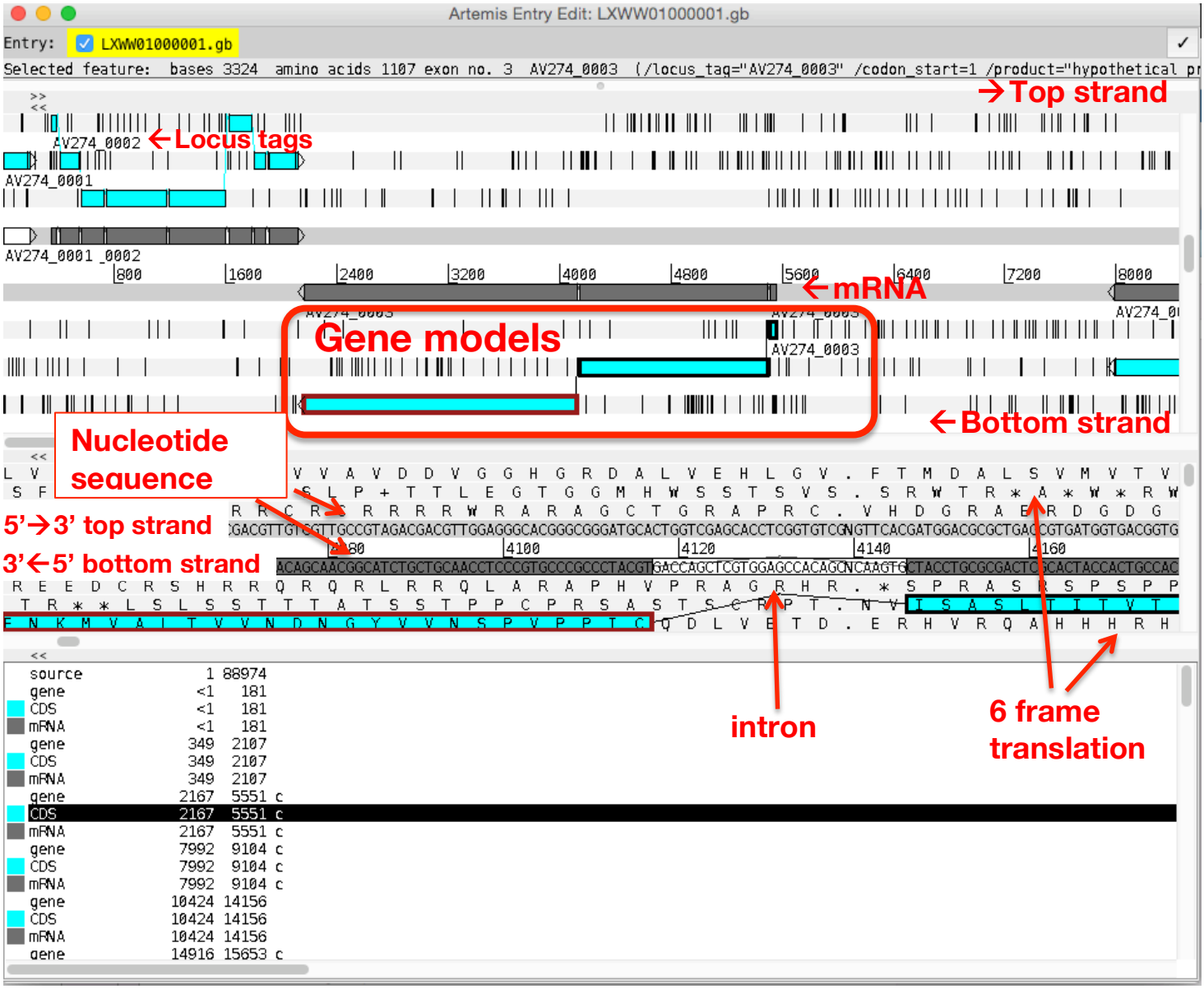
INDSC stands for International Nucleotide Sequence Database Collaboration, a long-standing collaboration between NCBI, DDBJ and EMBL-EBI databases. Accession numbers shown in this column are used by all three databases.

Blastocystis sp. ATCC 50177/Nand II
Submitter: Dalhousie University
Environment: OptimumTemperature: C, Habitat:HostAssociated

Loc	Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	Other RNA	Gene	Pseudogene
		master WGS	-	LXWW00000000.1	16.47	53.0	6,544	18	6,617	55
	Un	-	-	-	16.47	53.0	6,544	18	6,611	49

6. **Click on** the accession number there below INSDC.
7. This is the GenBank entry for the whole genome shotgun (WGS) sequencing project. **At the bottom of the entry** you will see the list of the contigs: **LXWW01000001-LXWW01000580. Click on those.**
8. This will bring up the **'Sequence Set browser'** that shows each of the 580 contigs from the assembly. **Click on the 'GenBank' link for the first contig: LXWW01000001**
9. Scroll up and down to see the format of the file. The 'FEATURES' section provides the information about the locations of the genes, putative mRNA start and stop sites, intron positions, annotation information, the inferred amino acid sequences of proteins. At the bottom you will see 'ORIGIN' with the entire nucleotide sequence of the whole contig.
10. Go to the top right hand side and click on 'Send to:' and choose 'Complete record' and 'File'. Select **GenBank (full)** and **save the file** to your working directory with the name **LXWW01000001.gb**
11. Start **Artemis (double click)** and click '**OK'** on the first dialogue box.
12. Under **'File'** choose open **Project Manager**. Then click on the green '**+**' sign to create a new project (name it whatever you want)
13. Select the **LXWW01000001.gb** file you just saved in your working directory and click **'OPEN'**

14. You should see something like this:



15. There are 3 different fields of view. The top shows the gene models (blue) the mRNAs (dark grey) and the stop codons (black ticks). The middle shows the nucleotide sequence with the predicted amino acid sequence above and below (top strand is 5'→3' and bottom strand is 3'→5'). The bottom panel shows all the features of the contig as indicated in the GenBank file.
16. Double click on one of the gene models in the top panel. You'll notice that that area is highlighted below and you can even see the introns in the sequence (usually about ~30bp in length which is characteristic for *Blastocystis*)
17. If you 'Right click' on the gene model a menu will come up with all sorts of choices. Choose '**View**' and '**Amino acids of selection'**. You can output it in FASTA or other format. This is the predicted protein for that locus (the locus tags are listed in the top panel).
18. If you want to search for some locus tag or a sequence motif or something, use the **Goto menu→ Navigator** to help you find things.
19. Go to **Select → All CDS Features**. Then under '**File' → Write → Amino Acids of Selected Features**. Save the file as **'contig1_proteins.fasta'**
20. **Quit Artemis.**

Part 2 – Functional annotation of genes

Tools/databases used:
NCBI BLAST/Conserved domains: http://www.ncbi.nlm.nih.gov
Interproscan: https://www.ebi.ac.uk/interpro/
eggNOG-mapper/eggNOG database: http://eggnogdb.embl.de/#/app/home

Introduction:
Predicting putative functions for protein-coding genes is complicated and error-prone. Various methods include basing a prediction on one or several of the following:

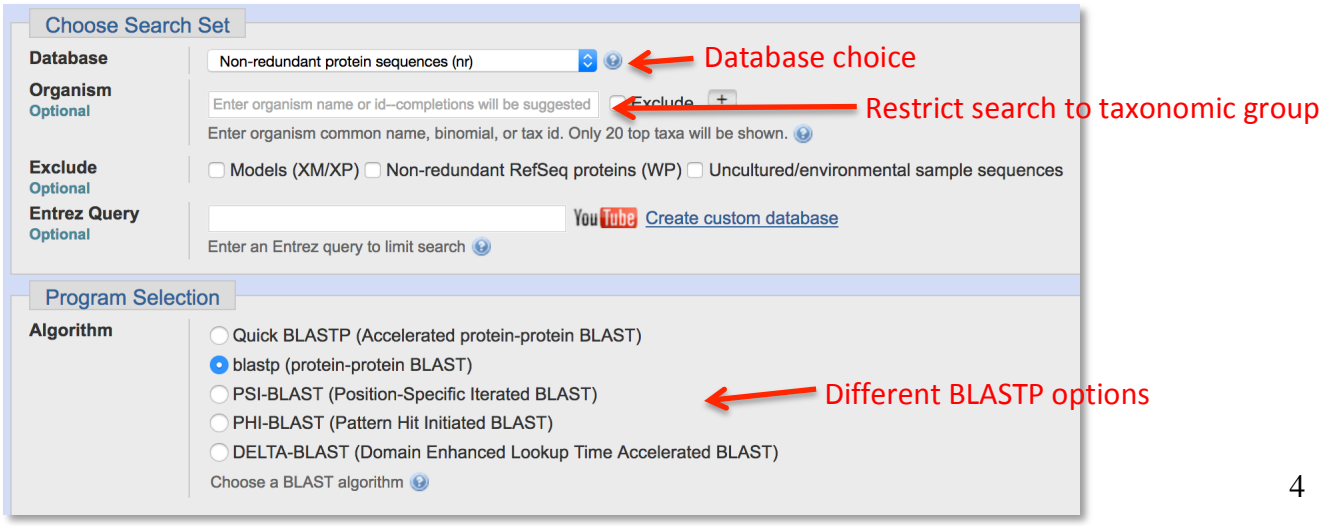
- the functional annotation of the 'top hits' from a BLASTP search against a protein database (e.g. databases such as: nr, uniref, UniProtKB/Swiss-Prot etc.).
- the functional annotation of the best 'match' by searches using hidden Markov models (HMM) based on alignments orthologous groups in databases such as eggNOG, KOG, Panther etc. (e.g. eggNOG-mapper, PANNZER2, Panther)
- annotating the conserved domains *within* the protein using domain-based search tools (usually HMM-based) and domain databases such as Interproscan, PFAM, SMART etc.

Finding best hits and conserved domains with BLAST

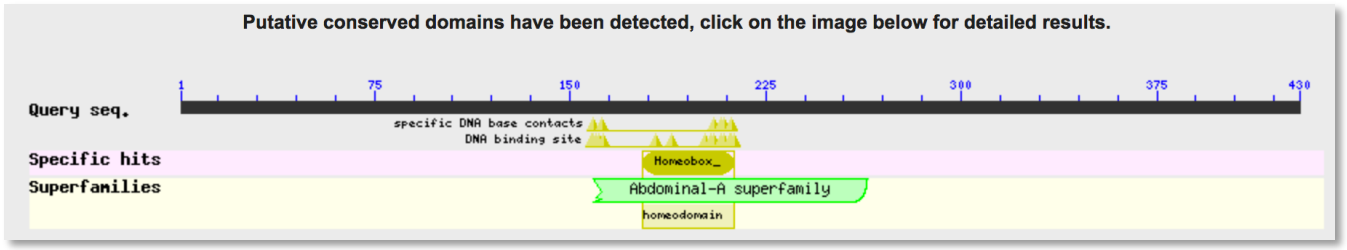
1. **Open** the **contig1_proteins.fasta** file in Notepad (or open it on this **weblink**) and scroll down until find the protein with the name **AV274_0014**. It should have the words 'homeobox protein' in its name. **Copy its FASTA entry.** You can also find this FASTA of this protein at this **weblink**
2. Go to NCBI page (http://www.ncbi.nlm.nih.gov) and click BLAST on the right hand side menu.

BLASTN is for nucleotide sequence searches against nucleotide databases,
BLASTX does 6 frames translations of a nucleotide sequence and search against amino acid databases
BLASTP is for amino acid sequence searches against amino acid databases
tBLASTn is for amino acid sequence searches against translated nucleotide databases (all six frames)

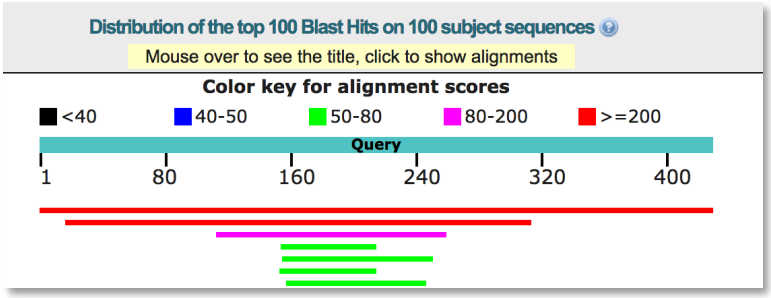
3. Choose **BLASTP** and **paste** the **AV274_0014** sequence into the window.
4. Choose the database you want to use:
   - **nr** non-redundant collection of almost all protein sequences in GenBank
   - **uniref** - comprehensive and non-redundant set of sequences from major research organisms that is well annotated
   - **UniProtKB/Swissprot** - very clearly and carefully annotated protein database
5. Note a number of other options including i) restricting the search to particular taxonomic groups, and iii) different versions of BLASTP for different settings



6. Choose the default and click '**Submit'.**
7. If the protein contains domains of interest (i.e. conserved protein domains that are found in databases), then it will be shown as follows. Later we will click on this (not now).



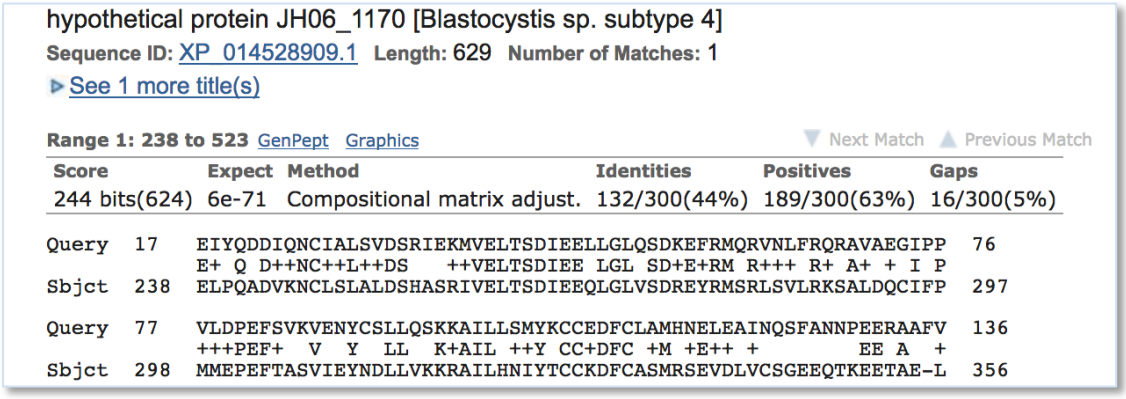
8. Below this you should see all the matching sequences in the database coming up as colored lines (colors are coded according to score – high scores means your sequence matches database sequence better)



9. Scroll down and you can see the summary of the search. How many hits are there to any *Blastocystis* subtype? What subtypes have homologs?

Note: you will not find homologs from Blastocystis ST2, 3, 6, 8 and 9 amongst the hits because these genomes are not annotated and so don't have 'proteins' in the protein database. If you want to know if a homologous gene is present in those genomes you should go back to the BLAST front page (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and use ***tblastn*** *and select the '***Whole-genome shotgun contigs (wgs)'** *database to search (****tblastn*** *will translate those genomes in all 6 frames and BLAST your protein sequence against these translations)*

10. Keep scrolling down and you will see the actual BLAST alignments.



hypothetical protein JH06_1170 [Blastocystis sp. subtype 4]

Sequence ID: XP_014528909.1 **Length:** 629 **Number of Matches:** 1

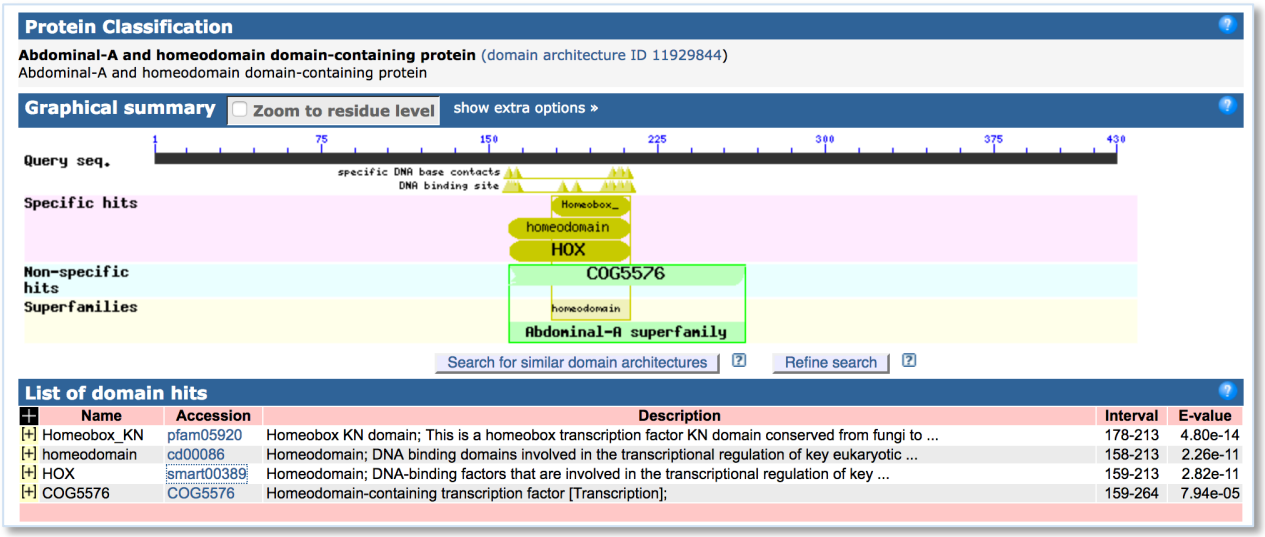
See 1 more title(s)

Range 1: 238 to 523 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
244 bits(624)	6e-71	Compositional matrix adjust.	132/300(44%)	189/300(63%)	16/300(5%)

```
Query  17   EIYQDDIQNCIALSVDSRIEKMVELTSDIEELLGLQSDKEFRMQRVNLFRQRAVAEGIPP  76
            E+ Q D++NC++L++DS   ++VELTSDIEE LGL SD+E+RM R+++ R+ A+ + I P
Sbjct  238  ELPQADVKNCLSLALDSHASRIVELTSDIEEQLGLVSDREYRMSRLSVLRKSALDQCIFP  297

Query  77   VLDPEFSVKVENYCSLLQSKKAILLSMYKCCEDFCLAMHNELEAINQSFANNPEERAAFV  136
            +++PEF+  V  Y  LL  K+AIL ++Y CC+DFC +M +E++ +         EE A  +
Sbjct  298  MMEPEFTASVIEYNDLLVKKRAILHNIYTCCKDFCASMRSEVDLVCSGEEQTKEETAE-L  356
```

11. The Query is your sequence of interest and the 'Sbjct' is the database sequence aligned to it. Note how divergent the homolog of ST4 is from your ST1 (NandII) query sequence. They are only 44% identical!
12. Scroll back up to the 'Conserved domain' graphic at the top. Click on it and you should see something like this:



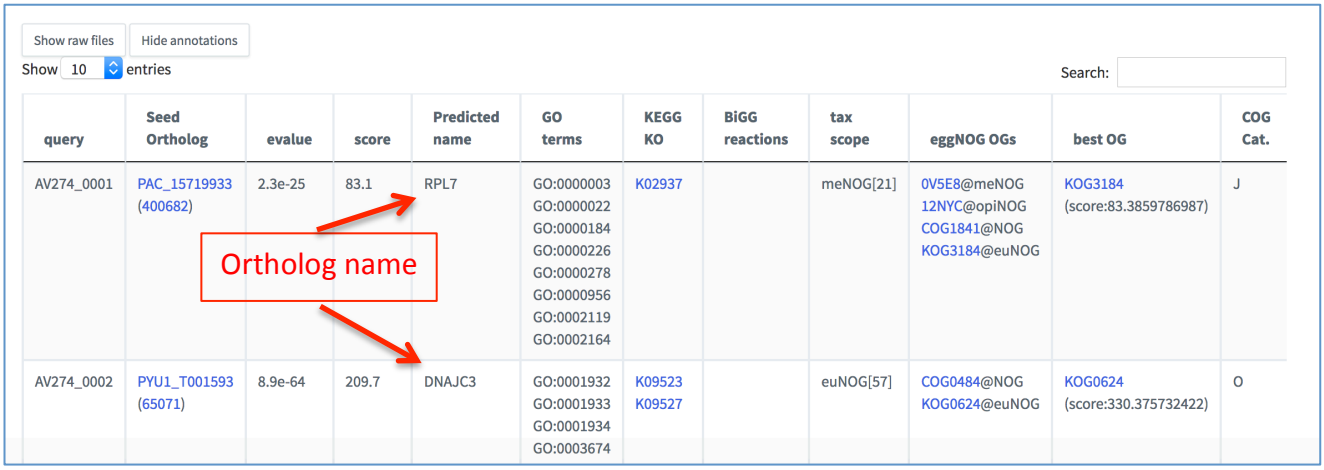
13. This is a summary of all the hits to conserved domain databases and specific information about specific functional residues in the protein (inferred from homologs)
14. Click 'Zoom to residue level' and you can see in detail all of the annotated features in your query sequence including DNA binding motifs etc.
15. Mouse over each of the features. A box should come up explaining more about that feature. If you click on various domain **Accessions** you will get an idea of the function of this domain.

Finding conserved domain structure and functional sites with Interproscan

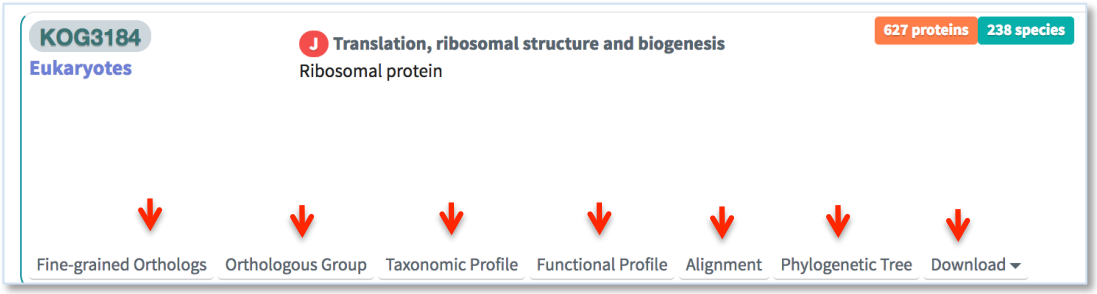
1. Go to **Interproscan** http://www.ebi.ac.uk/interpro/search/sequence-search and note the box where you could **paste the same sequence** (**weblink**).
2. I've already completed the search; the result is here: https://www.ebi.ac.uk/interpro/sequencesearch/iprscan5-S20181009-170444-0466-5488606-p2m
3. The output will show predicted protein family, domains, other predicted features (such as disordered regions) and annotated residues. Click on them and learn about the regions at your leisure.

EggNOG functional annotation steps to follow:

1. Go to the EggNOG database: http://eggnogdb.embl.de/#/app/home and click on **eggNOG-mapper** on the top left
2. Select **Browse** and upload your **contig1_proteins.fasta** file.
3. **Do NOT** hit **Run** as you will not have time to get the results.
4. Go to this web address for the results from contig1 proteins: http://eggnogdb.embl.de/#/app/emapper?jobname=MM_wWTmMX
5. Click on '**Explore annotations**'. You should see something like this:



6. **Click on the KEGG KO**. This brings you to the **Kyoto Encyclopedia of Genes and Genomes (KEGG)** annotation of that 'orthologous group' of proteins. KEGG is a database that associates molecular functions with proteins and biochemical pathways
7. You will see **GO terms**. These are 'Gene ontology' numbers that refer to general classes of molecular functions. More can be found about GO here: http://www.geneontology.org/page/introduction-go-resource
8. **Click on the best orthologous group** (**best OG**). This gives you the putative annotation of your protein.
9. Note that you can access all kinds of information about this OG at the bottom of this entry:



16. A text file of annotations for all of the proteins (for which eggNOGs were found) can be downloaded from the original results page by clicking on **'Download annotations'**

Notes:

- You can do a similar kind of analysis to eggNOG-mapper (i.e. annotate a large number of proteins in FASTA format) using **PANNZER2:** http://ekhidna2.biocenter.helsinki.fi/sanspanz/ .
- Individual protein functions can also be investigated in the well annotated **Panther** database**:** http://www.pantherdb.org/tools/hmmScoreForm.jsp

Part 3 – Prediction of subcellular localization

Background reading on the secretory pathway: "*Overview of the Secretory Pathway*": https://www.ncbi.nlm.nih.gov/books/NBK21471/

Tools used:

TargetP: http://www.cbs.dtu.dk/services/SignalP/
TMHMM: http://www.cbs.dtu.dk/services/TMHMM-2.0/
BUSCA: http://busca.biocomp.unibo.it/

Introduction:

Signal or targeting peptides can be at the N-terminus (signal peptides for: secretory pathway, endomembrane system, and targeting peptides for mitochondria, chloroplasts), within the sequence (e.g., nuclear localization signals) or C-terminus (e.g. PTS1 for peroxisomes and ER-retention signals). The properties of target peptides vary depending on the organelle or membrane to which they are targeted. Bioinformatic prediction tools tend to work best if they have been 'trained' on experimental localization data for a large set of proteins from the organism of interest. Currently there is not enough experimental data from *Blastocystis* to train such methods, so we must rely on predictions from other organisms that will necessarily be less accurate.

- **TargetP** is a leading neural-network-based method for testing for the presence of localization signals for secretory pathway, mitochondria and chloroplasts
- **TMHMM** is a hidden Markov Model-based method for determining if proteins have transmembrane regions
- **BUSCA** is a server that uses a large number of different prediction tools to come to a consensus localization prediction

Steps to follow to run TargetP (targeting peptide) prediction:

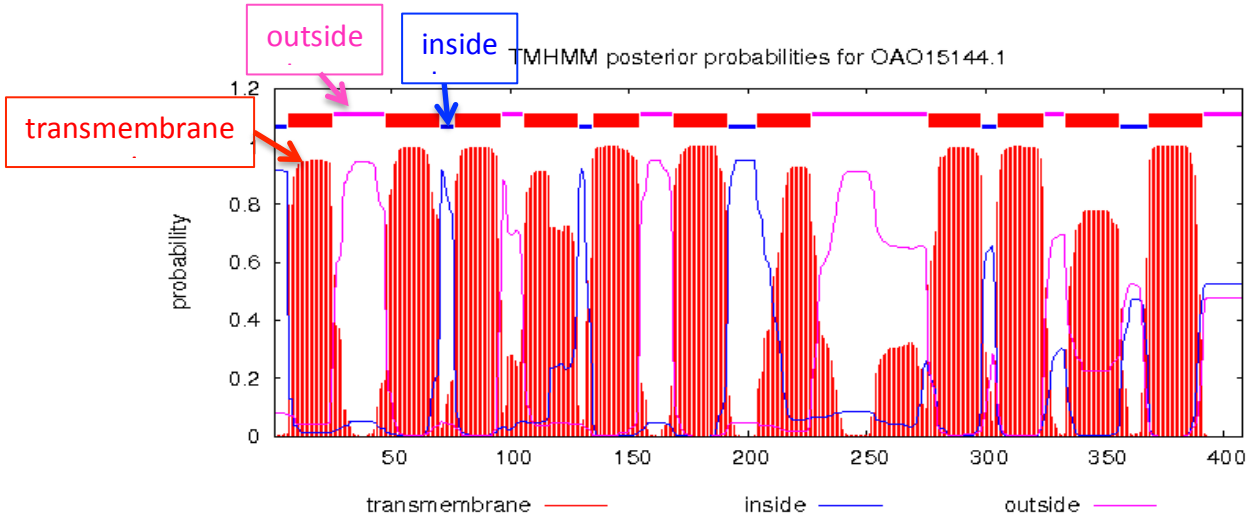
1. Go to the TargetP server in Denmark: http://www.cbs.dtu.dk/services/TargetP/
2. Note all the various options you can select. Under '**Performance scope'** click the '**Perform cleavage site predictions'** option
3. Open a new browser tab and open this **weblink** containing two proteins: i) a putative fucose permease protein and ii) the alpha subunit of succinyl-CoA synthetase from *Blastocystis* sp. ST1 (NandII)
4. Copy and paste the sequences into the box in the TargetP box and click '**Submit**'
5. The result should look like this:

```
### targetp v1.1 prediction results ###################################
Number of query sequences:  2
Cleavage site predictions included.
Using NON-PLANT networks.

Name                   Len            mTP     SP  other  Loc  RC  TPlen
--------------------------------------------------------------------------
OAO15144.1             408          0.010  0.948  0.118   S    1     26
ABY62723.1             318          0.871  0.026  0.149   M    2     18
--------------------------------------------------------------------------
cutoff                              0.000  0.000  0.000

Explain the output.  Go back.
```

6. The first protein is predicted to have a signal peptide (SP) for the secretory pathway (S) and the second one has a mitochondrial targeting peptide (mTP) for mitochondria (M). Each occurs at the N-terminus of the protein of interest. Note the prediction of the length of the peptides (under **TPlen**).
7. Click on the '**Explain**' button to understand the various abbreviations

TMHMM (transmembrane regions) prediction, steps to follow:

1. Go to the TMHMM server: http://www.cbs.dtu.dk/services/TMHMM-2.0/
2. Copy the FASTA sequences of the same two proteins as above (**weblink**).
3. Copy and paste the sequences into the box and click '**Submit**'
4. The **first** result should look like this:



5. The protein clearly has 11 transmembrane segments indicating it is an integral membrane protein. If it is on the cell surface then 'inside' refers to the cytoplasm and outside to the outside of the cell.
6. The second protein (succinyl-CoA synthetase, alpha subunit) has no predicted transmembrane domains.

BUSCA subcellular localization prediction, steps to follow:

1. Go to the BUSCA website in Italy: http://busca.biocomp.unibo.it/
2. Choose the taxonomic origin of your sequences from the pull-down menu: '**Eukarya – Other - 9 compartments**'
3. Paste the sequences of the same two sequences (**weblink**) into the box
4. Click '**Start prediction'.** This might take a little while so you can continue with the next exercise and leave the webpage open. The result should look like this:

Protein Accession/ID	GO-id	GO-term	Score	Alternative Localization	Features
OAO15144.1	GO:0012505	C:endomembrane system	0.67	GO:0005886 - C:plasma membrane (score=0.47)	Transmembrane Alpha Helix
ABY62723.1	GO:0005739	C:mitochondrion	0.82	-	Mitochondrial Transit Peptide

5. Note for the putative fucose permease (OAO15144.1) there are two possible localizations: endomembrane system (better score) and plasma membrane (not quite as good score).
6. Optional (if time permits): You can try this tool with this **weblink** to 5 cysteine proteases obtained from a text search of the Protein database at NCBI

Notes

Other useful tools include localization predictions of a new 'deep learning' algorithm DeepLoc: http://www.cbs.dtu.dk/services/DeepLoc/ , the tool Mitofates (for mitochondrial prediction): http://mitf.cbrc.jp/MitoFates/cgi-bin/top.cgi and Phobius (for signal peptide prediction): http://phobius.sbc.su.se/.

Part 4 – Alignment and phylogeny of a protein family:

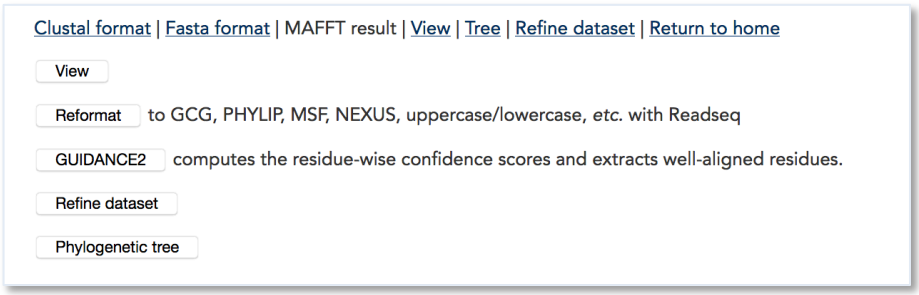
Tools used:
MAFFT version 7 (online): https://mafft.cbrc.jp/alignment/server/
Phylo.io (online): http://phylo.io/

Introduction:
An important part of comparative genomics is inferring the phylogenies of gene families in the genomes of interest. To do this you first need to retrieve close homologs of the proteins using the top BLAST hits or based on the orthologous groups (e.g. eggNOGs) that you have found are the best match. All aligned sequences (in FASTA format) can then be aligned together using a multiple alignment program. There are many different multiple alignment tools, but a popular tool is MAFFT. One the alignment is estimated, it should be viewed using a viewer like Seaview or an online tool. Then a phylogeny can be estimated. Here we use a 'quick and dirty' pairwise distance matrix method method to get a rough picture of relationships. To get a more robust phylogeny, a maximum likelihood or Bayesian analysis should be conducted.

Alignment and phylogeny of a protein family in *Blastocystis*, steps to follow:

1. Copy the fucose permease sequence from the file of two proteins above (or **weblink**) and copy and paste it into the NCBI BLASTP window: https://blast.ncbi.nlm.nih.gov/Blast.cgi
2. Click **BLAST**
3. When the results return, scroll down to the list of best hit and click on '**Select: All**' and then '**Download**'. This would save the file in FASTA (Complete sequences) format in your working directory. This contains 100 of the best hitting sequences.

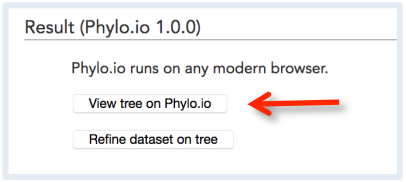
4. Note that you can actually get a view of a multiple alignment and distance tree from the above links as well. Feel free to check that out. We will use a different method.
5. **Go to the online MAFFT multiple alignment server**: https://mafft.cbrc.jp/alignment/server/
6. Open this **weblink** in a new tab of your web browser. This is the same file as above but with only 50 best hit sequences saved.
7. **Copy and paste it into the MAFFT server window**. Scroll down and note the various options for the multiple alignments. Just use the defaults.
8. Rather than waiting for the results of submitting this run, just go to the following URL where it has already finished: https://mafft.cbrc.jp/alignment/server/spool/_out.181010085518363SqoGkIB3NyHsRK2TEAjD9lsfnormal.html
9. Scroll up and down. You can see the fucose permease that you used to BLAST is at the top and is aligned to homologs



10. The various options include viewing the alignment in **Fasta format**, alignment '**View**', phylogenetic analysis ('**Tree**') etc.
11. Click on '**View**' and select **MSA viewer** (in a new window). This allows you to inspect your multiple alignment and choose sequences to remove etc.
12. Go back the original results window and **click on 'Tree'**
13. There are multiple options for making the tree. **Choose the default options** (faster) and click '**Go!**'

*Note: Ideally you **should** allow for different rates of evolution at different sites ('**Heterogeneity among sites'**) and you **should** select '**Estimate'** for the alpha shape parameter (this parameter governs the how variable the rates are at different sites – you **should** estimate it from your alignment). You **should** also conduct **Bootstrap** analysis to determine how robust the estimated groupings in your tree are (high bootstrap values → more robust). All of that takes much longer.*

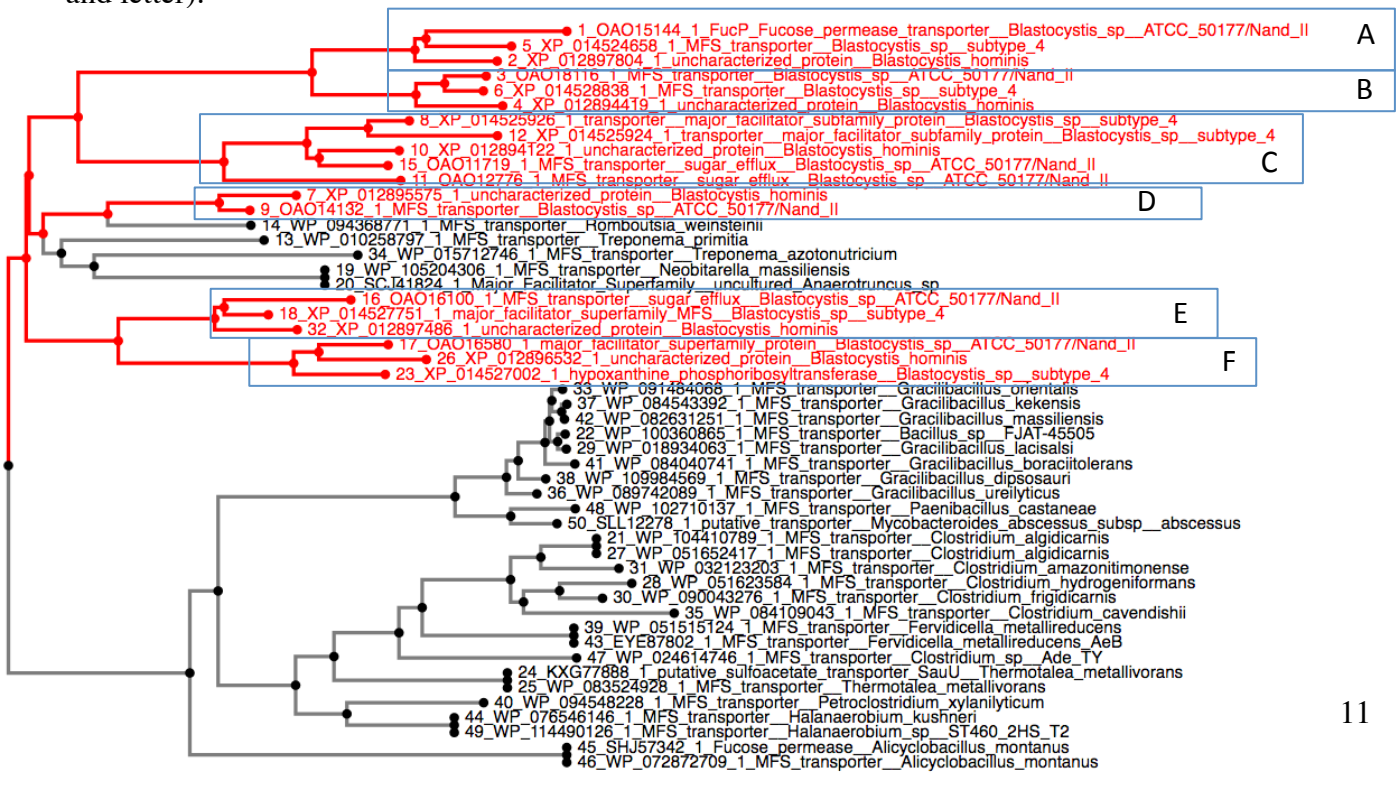
14. You will see:



15. **Click on 'View tree on Phylo.io'**

Note that you can use this phylogenetic tree viewer online at http://phylo.io/ for any tree that you generate with any phylogenetic program that outputs 'Newick' format tree files (including IQ-TREE, RAxML, MrBayes etc.)

16. Click 'Render' on the left side and your tree should pop up. You can manipulate the tree and how it is shown using the arrows and the settings.
17. If you want to visualize the Blastocystis homologs, **click on the little search icon on the top right corner** and **type in the word** '**Blastocystis**'. They will show in red like this (without the blue boxes) and letter):



18. You can see from the tree that a number of gene duplications have happened prior to the divergence of the three subtypes generating little 'paralog' trees. The paralogs are labeled A-F. It seems likely there could some functional differences between these 'Major facilitator superfamily' (MFS) proteins → they may be expressed under different conditions, have different affinities for substrates or different substrates.
19. You can go back to the previous menu and choose 'refine dataset on tree'. This allows you to choose subsets of your proteins based on the tree to realign and generate a new phylogeny.

Notes: It would be better to conduct phylogenetic analysis using maximum likelihood (ML) or Bayesian analyses with more sophisticated models of the evolutionary process. A very useful tool for ML analysis is **IQ-TREE** (see www.iqtree.org) that has online servers that you can upload your FASTA file to analyze. Here is an example of a run with IQ-TREE using the same aligned file from MAFFT as above: http://iqtree.cibiv.univie.ac.at/?user=andrewjmroger@gmail.com&jobid=181010020931

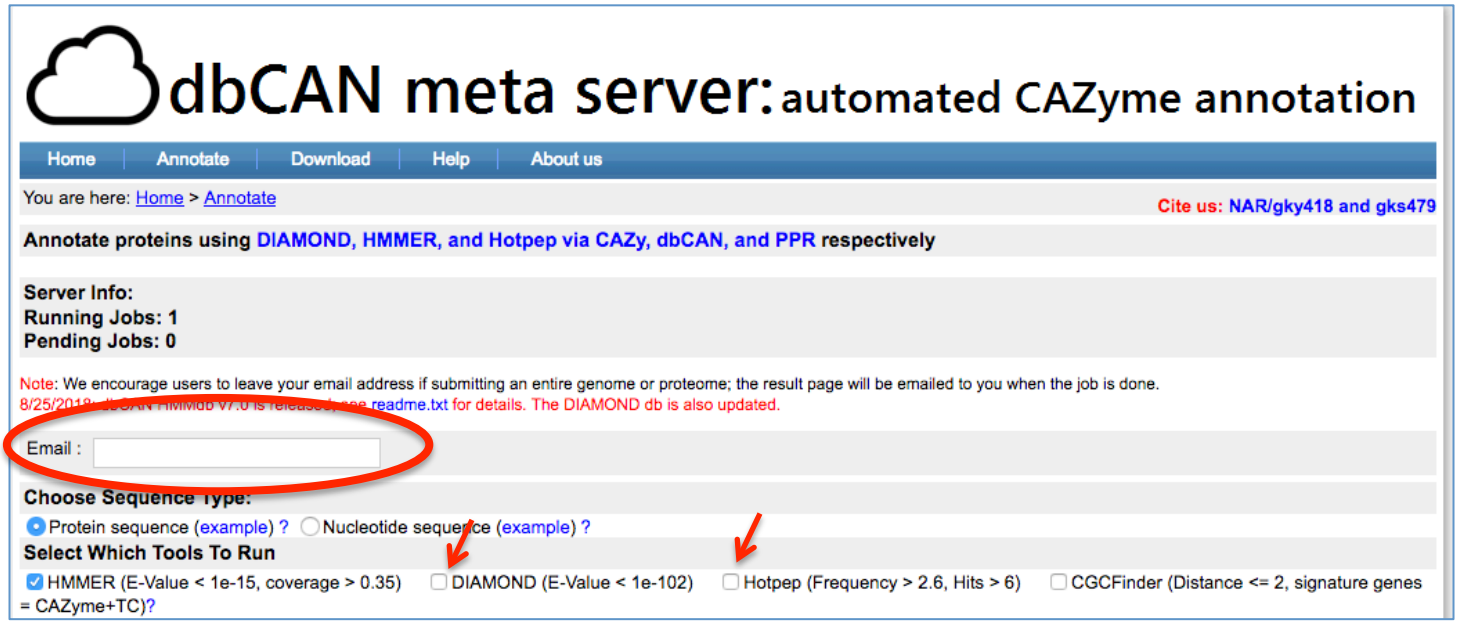
Part 5, Carbohydrate utilizing enzymes (CAZy)

Databases/Tools used:
dbCAN2, a series of tools for searching the CAZy enzymes database:
http://cys.bios.niu.edu/dbCAN2/blast.php

Introduction
There are a large number of enzymes in nature that degrade, modify or create glycosidic bonds, the bonds that link sugar (or polysaccharide) moieties together or to other organic molecules via O, N or S atoms. The CAZy database (www.cazy.org) describes the families of structurally-related catalytic and carbohydrate-binding modules (or functional domains) of enzymes. If a predicted protein is homologous to one of these families or functional modules, it can be useful to know that to predict the function of the protein. Here we use a database searching tool to find CAZy homologs of a Blastocystis ST1 (NandII) protein. Here is a Link to the paper describing the CAZy database: https://academic.oup.com/nar/article/42/D1/D490/1057423

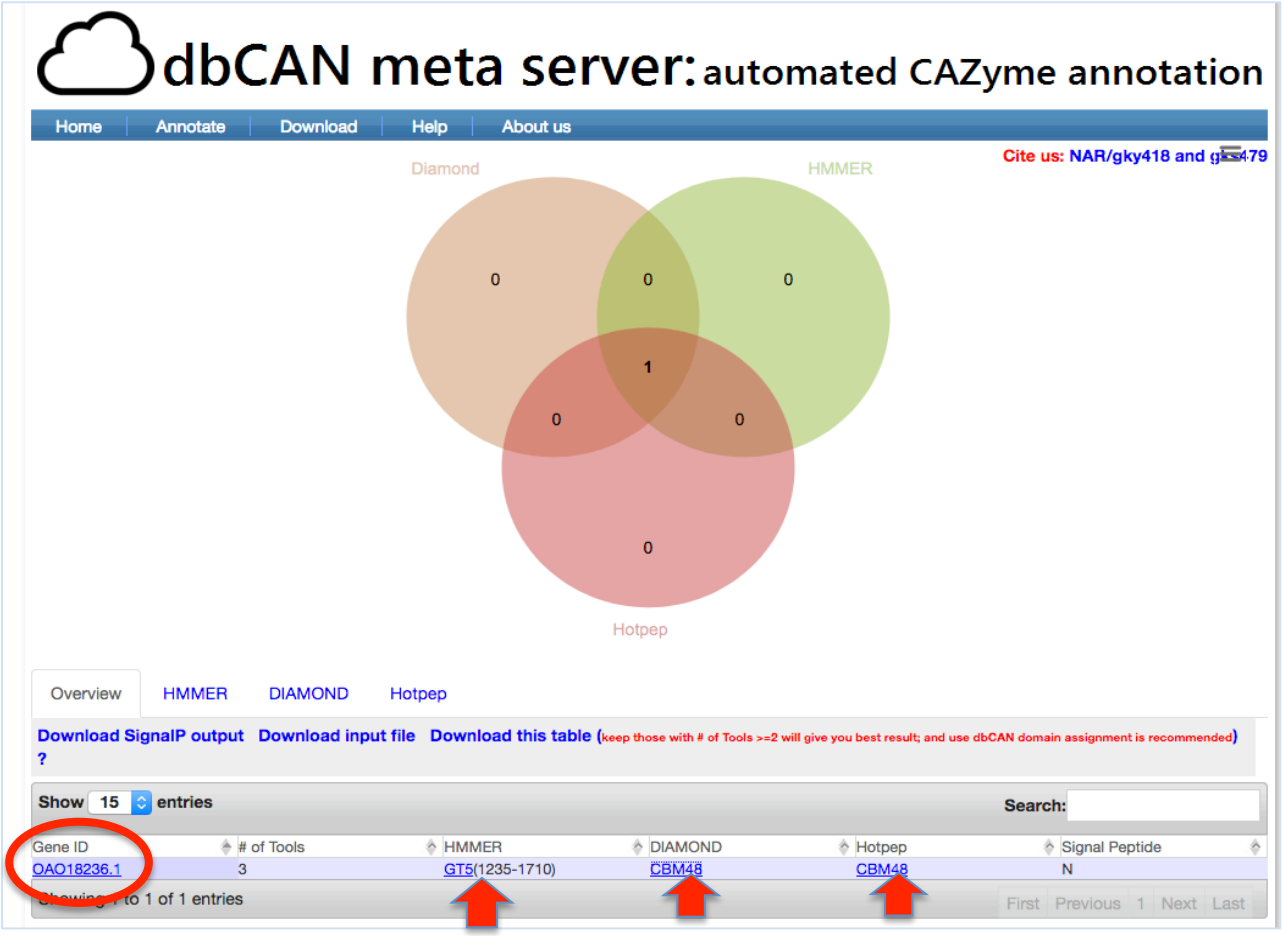
Steps to follow:

1. Go to NCBI (https://www.ncbi.nlm.nih.gov/) and choose 'Protein' from the pull-down menu. Type in '**OAO18236.1**', an accession number of a protein from the first contig of the Blasto NandII assembly.
2. Click on the FASTA link near the top to view the protein in FASTA format. Copy the FASTA entry (header + sequence). This can also be accessed at this **weblink**
3. Go to the dbCAN2 website to search the CAZy database: http://cys.bios.niu.edu/dbCAN2/blast.php
4. Fill in the form with your email. And select the various search tools: HMMER, DIAMOND and Hotpep.



5. Paste the FASTA file in to the box below and click **'Submit'**

6. After it finishes running you should get the following window



Links to the CAZy database for that protein module

7. Click on the **HMMER**, **DIAMOND** and **Hotpep** links. These take you to the CAZy database for the modules identified to give you functional information
8. Click on the Gene ID link. This provides a picture of the best hitting CAZy domains mapped on the sequences (as a line).
9. Using this information plus information from eggNOG-mapper and BLAST, you can make predictions as to the function of the protein.

Part 6, Predicting the type of peptidases/protease you have

Tools/Databases used:
MEROPS: https://www.ebi.ac.uk/merops/
EBI Blast server: https://www.ebi.ac.uk/Tools/sss/ncbiblast/

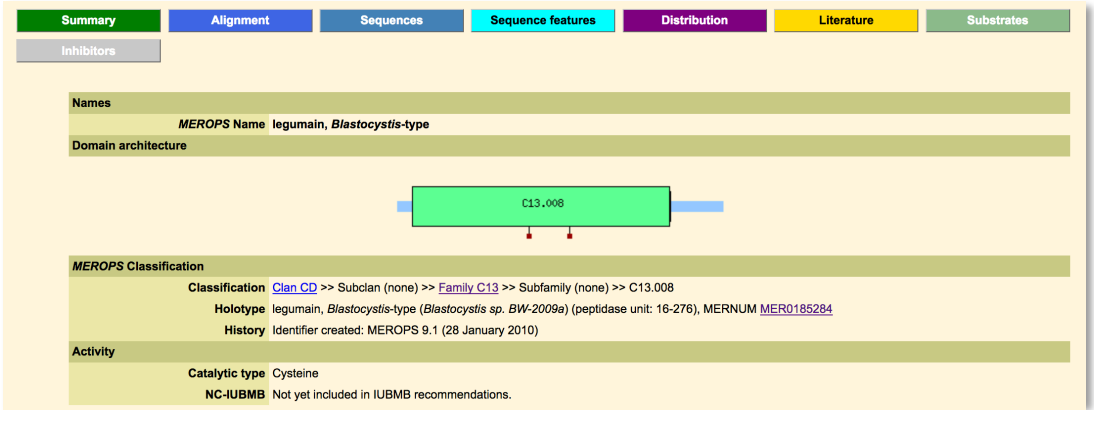
Introduction
Blastocystis subtypes secrete proteases/peptidases that, by degrading various proteins in the host, could be involved in pathogenesis (reviewed in Ajjumpur and Tan (2016) https://www.ncbi.nlm.nih.gov/pubmed/27181702). It is therefore of interest to know what the localization, types and function of the various proteases in the various subtypes. A simple text based search of the 'Protein' database of NCBI using terms such as '**protease and Blastocystis**' or '**peptidase and Blastocystis**' will retrieve the sequences of hundreds of putative examples. **BUSCA** can be used to predict their localization. To try to determine their particular function, the MEROPS Peptidase database is a useful place to start.

Searching the MEROPS database. Steps to follow:

1. Go to the MEROPS database at the EBI: https://www.ebi.ac.uk/merops/
2. Note the resources available on this website. **Click on the 'BLAST MEROPS'** link on the left
3. There are several options to choose from. Select the first option to search the **merops_scan** database. This one is the simplest database to use.
4. This will take you to the **EBI Blast-server** and show you a list of databases to choose to search.
5. **Unselect the UniProt Knowledgebase**
6. Select the **'Other Protein databases'** and then check the **'MEROPS-MPRO (Sequences from the MEROPS scan dataset)'.**
   Open another web browser tab and click on this **weblink** to 5 cysteine proteases from ST1 NandII.
7. Copy these and paste the **first sequence** into the **EBI Blast-server** window and click **'Submit'** at the bottom. The result should look like:

Align.	DB:ID	Source	Length	Score (Bits)	Identities %	Positives %	E()
☑1	MPRO:MER0185284	- legumain, {Blastocystis}-type (Blastocystis sp. BW-2009a) [C13.008]#C13#{peptidase unit: 16-276}~source ACO24555~	261	240.7	43.4	62.5	2.0E-78
☑2	MPRO:MER0004009	- legumain (plant alpha form) (Canavalia ensiformis) [C13.002]#C13#{peptidase unit: 2-277}~source E05717~	276	199.5	39.4	60.3	3.1E-62

8. The top hit is to **'legumain'** of Blastocystis-type. The database ID is on the left and you can click on it. The MEROPS identifier is **C13.008**
9. You can now return to the MEROPS database (https://www.ebi.ac.uk/merops/) and click on **SEARCHES** on the left.
10. Search by 'name' (the first option) using 'legumain' and you get all the types. If you click on the Blastocystis ptype at the bottom right you get specific information about it including domains, inhibitors, substrates and literature:



More bioinformatics tools and databases:
Phobius Subcellular localization prediction: http://phobius.sbc.su.se/
MitoFates Mitochondrial localization: http://mitf.cbrc.jp/MitoFates/cgi-bin/top.cgi
Membrane Transporters at **TransportDB**: http://www.membranetransport.org/transportDB2/index.html
Transporter Classification Database (TCDB): http://www.tcdb.org/
Kinases at **kinase.com**: http://kinase.com/web/current/blast/
MetaCyc Metabolic Pathway Database: https://metacyc.org/