Odds and Ends

Last day of class is usually a mix of trying to find gaps or fill in holes.

I discussed phylogenetic tree building.

```
git clone https://github.com/biodataprog/GEN220_2025_classexamples.git
cd GEN220_2025_classexamples/Trees
module load muscle
module load fasttree
module load iqtree
module load trimal
module load clipkit

# build an alignment of sequences already identified as homologs
# previously I had started with MET12 (S. cerevisiae) enzyme
more  MET12.fa # single sequence
ls -l MET12.hit_seqs.fasta # the collection of homologs for MET12 in a few yeast fungi
# denovo multiple alignment - writes in multi-fasta format
muscle -align MET12.hit_seqs.fasta -output MET12.hit_seqs.fasaln

# trim sequences - using automated parameters - see http://trimal.cgenomics.org/trimal for
trimal -automated1 -in MET12.hit_seqs.fasaln -out MET12.hit_seqs.mfa.trim
# this is an alternative alignment trimmer
clipkit MET12.hit_seqs.fasaln
# build a tree w fastree (FastTreeMP uses multiple processors, FastTree uses 1 processor on
FastTreeMP <  MET12.hit_seqs.fasaln >  MET12.hit_seqs.tre

# build a tree with IQ-TREE2 - ultrafast bootstrap and first determine optimal number of pr
iqtree3 -s MET12.hit_seqs.fasaln -nt 2 -bb 1000 -alrt 1000
```

Some links * Muscle - Multiple alignment tool * TrimAl - alignment trimming tool * clipkit - alignment trimming tool * HMMER - HMMER - Hidden Markov Model for biosequence analyses. * FastTree - Fast Phylogenetic Tree construction * IQ-TREE - Phylogenetic Tree construction * RAxML; a tutorial * iTOL - Tree visualization (web-based) tool * FigTree - Tree visualization (can run on HPCC if you have X11 enabled: `module load figtree; figtree`) * ggtree - R package for Tree rendering

```
module load hmmer

# build an HMM from a multiple alignment
hmmbuild MET12.hmm MET12.hit_seqs.fasaln
```

This is a little circular I am searching the HMM back against the original sequences, but if you wanted to instead search this HMM against a database of proteins (eg swissprot or your collection of proteins from species)

```
module load hmmer
```

```
module load samtools
# domtbl is the result file which has columns of data that are parseable instead of more co
hmmsearch -E 1e-3 --domtblout MET12.search.domtbl MET12.hmm DATABASE > MET12.search.hmmsearc
#. you could can download a database like swissprot
#curl -O https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/comple
#gunzip uniprot_sprot.fasta.gz
DB=uniprot_sprot.fasta
# or one on the cluster
cp /srv/projects/db/Swissprot/2023_03/uniprot_sprot.fasta $DB
hmmsearch -E 1e-3 --domtblout MET12.search.domtbl MET12.hmm $DB > MET12.search.hmmsearch

# retrieve these hits
grep -v ^# MET12.search.domtbl | awk '{print $1}' | samtools faidx -r - uniprot_sprot.fasta

# to align a set of proteins back to an HMM (which is instead of doing a denovo multiple al
hmmalign MET12.hmm MET12.sprot_hits.fasta  > MET12.sprot_hits.stk
# convert the stockholm format to multifasta (afa)
esl-reformat afa MET12.sprot_hits.stk > MET12.sprot_hits.fasaln
# convert the stockholm format to clustal
esl-reformat clustal MET12.sprot_hits.stk >  MET12.sprot_hits.clustalw

# rebuild a tree
clipkit MET12.sprot_hits.fasaln
# build a tree w fastree (FastTreeMP uses multiple processors, FastTree uses 1 processor on
FastTreeMP <  MET12.sprot_hits.fasaln.clipkit >  MET12.sprot_hits.tre
# read this tree
cat MET12.sprot_hits.tre
# try copy and pasting this and opening in iTOL
# https://itol.embl.de/
```