BLAST

- makeblastdb -in Ecoli.faa -dbtype prot -out ECDB to make databases
- blastp -query query.fasta -db swissprot -outfmt "6 qseqid sseqid pident length qstart qend sstart send evalue stitle" -out blast.out
- awk '{print $2}' blast.out | sort | uniq -c | awk '$1>1'   to know which gave multiple hits
- grep “Campylobacter jejuni” blast.out to get species specific hits
- blastn -query query.fasta -db nt -remote -out results_online.txt -outfmt "6 qcovs pident evalue sacc staxid stitle” for remote database searches
- wget "https://rest.uniprot.org/uniprotkb/P11926.fasta" to get full length sequence
- awk '$3>50 && $3<=60' blast.out | head -n 5 > 50to60_top5.out
- awk '$3>60 && $3<=70' blast.out | head -n 5 > 60to70_top5.out
- awk '$3>70 && $3<=80' blast.out | head -n 5 > 70to80_top5.out
- awk '$3>80 && $3<=90' blast.out | head -n 5 > 80to90_top5.out
- awk '$3>90' blast.out | head -n 5 > above90_top5.out
- To get representative sequences
  (head -n 1 50to60_top5.out; head -n 1 60to70_top5.out; head -n 1 70to80_top5.out; head -n 1 80to90_top5.out; head -n 1 above90_top5.out) > filtered.out
- If you want to get only one top hit in the target database for each query protein, modify the command as:
  blastp -db ECDB -query StaphAur.faa -evalue 1e-04 -outfmt 6 -max_target_seqs 1 -out Sprot-to-ECDB_Max1
- psiblast -query HBB.fasta -db HP -num_iterations=4 -evalue=0.01 -out HBB-to-HP_psiblast-4 -out_pssm=PSSM4
- download.sh

  ```
  for id in `cut -c11-18 forMSA.out`
  do
  wget https://rest.uniprot.org/uniprotkb/$id.fasta
  done
  ```
- tblastn -query query.fasta -db nt -remote -outfmt 6 -out tblastn.out

PAIRWISE ALIGNMENT

- needle -asequence reference.fasta -bsequence sequence1.fasta -gapopen 10 -gapextend 0.5 -outfile alignment1.txt
- needle -asequence reference.fasta -bsequence sequence1.fasta -gapopen 8 -gapextend 0.2 -datafile EBLOSUM80 -outfile alignment1_custom.txt
- water -asequence reference.fasta -bsequence sequence1.fasta -gapopen 10 -gapextend 0.5 -oufile local_alignment1.txt

MSA, PHYLOGENY & MEME

- clustalw -INFILE=sequences.fasta -OUTFILE=clustalw_alignment.aln -OUTPUT=PIR
- clustalo --infile=seq.fasta --outfile=clustalo.aln --wrap=60 –outfmt=clu
- t_coffee -seq RbcL_Proteins.fasta -output=clustalw_aln -outfile=tcoffee_alignment.aln
- T-Coffee is often more reliable than ClustalW or Clustal Omega because it uses a consistency-based scoring approach, which integrates information from multiple pairwise alignments to guide the final multiple sequence alignment. This method ensures that aligned residues are consistent across all sequence comparisons, reducing errors introduced by the progressive alignment strategies used in ClustalW and Omega. As a result, T-Coffee produces more accurate alignments, especially for distantly related sequences or those with low similarity, where traditional methods may struggle.
- Mustang download in legacy pdb format
- awk '($1~"^ATOM$" && $5~"^B$")' 8wix.pdb > 8wix_B.pdb
- awk '($1~"^ATOM$" && $5~"^A$")' 1bom.pdb > 1bom_A.pdb
- mustang -i 8wix_A.pdb 1bom_A.pdb
- Type mega on commandline → mega opens up → edit/build alignment → restrieve sequence from file → ok → select file → ctrl + A → alignment → align by → save in .mas format → exit → phylogeny → NJ tree → bootstrap 500 → substitution model → p distance → ok
- ITOL
- meme forMSA.fasta -o motif -nmotifs 5 -minw 6 -maxw 15

HMM

- **grep NAME Pfam-A.hmm | wc -l**
- to know the number of entries in pfam databse
-
- **hmmstat pfam-A.hmm**
- provides a quick statistical summary of one or more HMM profiles, reporting key information such as model name, length, number of sequences used to build it, and its information content—helpful for evaluating the quality and characteristics of HMMs.
- grep NAME Pfam-A.hmm | grep -i globin
-
- **hmmfetch Pfam-A.hmm Globin > globin.hmm**
- Pfam-A.hmm has hmm profile of 24000 families from this we are taking hmm profile of globin family in globin.hmm
-
- input to build hmm profile is msa
- hmmbuild <outputfile_name> <inputfile_name>
- **hmmbuild test.hmm test.clustal**
-
- **hmmemit -o emitted.fa test.hmm**
- generate a representative sequence of that family of protien
- the sequence does not really exist it is the representative sequence that fits the profile
- if you do hmmemit again it will give some other sequence
- it is like random number generator
- **hmmemit -n 5 -o emitted.fa test.hmm** to emit 5 sequences
-
- took last sequence from globins45.fa and saved in hbb2.fa
-
- search sequence against hmmfile
- **hmmpress Pfam-A.hmm** to make databse in proper format so that we can use hmmscan
- **hmmscan Pfam-A.hmm hbb2.fa** gives to what family the sequence might be belonging to it gave 2 peptidase or globin… lower evalue for globin hence it might be belonging to globin family
- try the same with sequence e emitted previously
-
- to align sequences using hmm profile
- **clustalo -i seq.txt -o seq.clustal** Do MSA
- **hmmbuild test.hmm test.clustal** Build HMM profile

- **hmmalign seq.hmm addseq.txt > align.aln**
-
- jackhmmer is equivalent of psiblast
- Finding the homologues
- **hmmbuild msa.hmm msa_clustal.aln**
- **hmmsearch -o hmm.ecoli.txt --tblout hmm.tab.txt msa.hmm Ecoli_Proteins.faa**
-
- **hmmsearch -o hmm.human.txt --tblout hmm.tab.txt1 msa.hmm HSA_Proteins.fasta**
-
-
- **hmmsearch query is hmm profile and database is fasta database it is like blast**
- **hmmscan query is fasta and we are searching against hmm profile**

- **hmmalign basically aligns the 2 sequences based on hmm profile so inputs for hmmalign would be a hmm profile (we made for msa) and a file containing both sequence that we want to align and then run the command** hmmalign msa1profile.hmm seq1seq2.fasta > aligned.fasta

- Go to https://www.rcsb.org/ → search for → click → left top (biological unit and asymmetric unit)
- go to https://www.ebi.ac.uk/thornton-srv/databases/pdbsum/ → enter pdb code → find → go to ligands → list of interactions
- Go to https://web.expasy.org/compute_pi/ → add uniprot id → click here to compute molwt/pI → submit
- Go to ncbi blast → tblastn → enter query sequence → search click on top hit (evalue 0, query covergae 100, percent identity 100) → genbank
- go to https://www.ebi.ac.uk/interpro/ → enter the sequence → search → click on sequence name