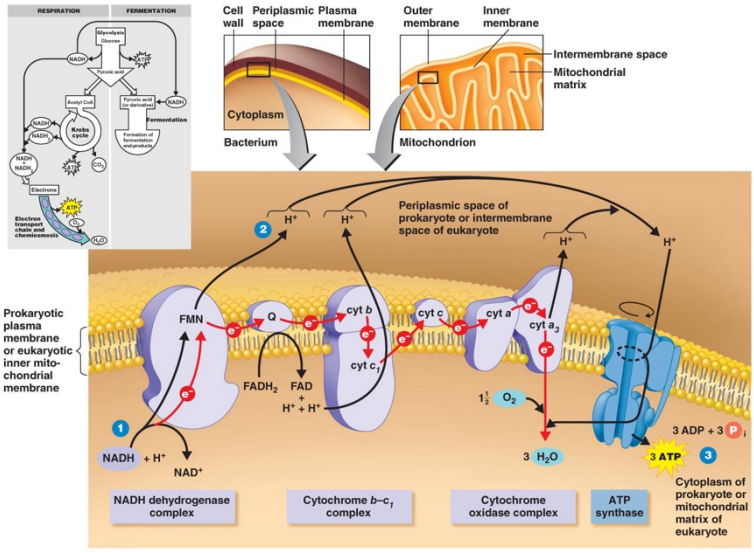
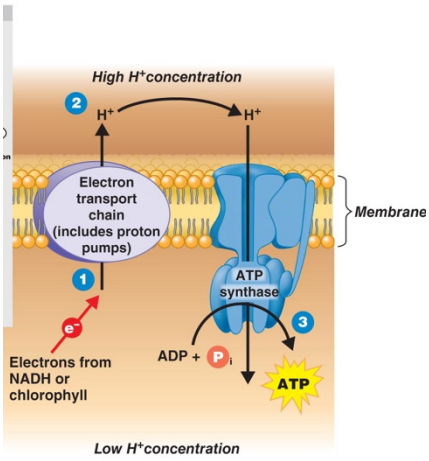
Phylogeny of COI exercise

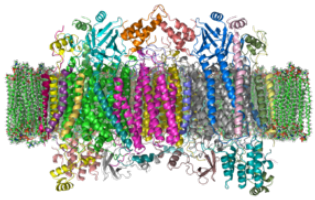
Electron transport chain



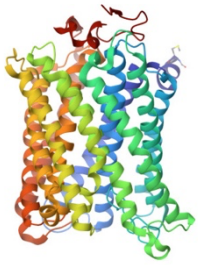
- Occurs in the plasma membrane of **prokaryotes**
- Occurs in the inner mitochondrial membrane of **eukaryotes**
- Series of carrier molecules are oxidized and reduced as electrons are passed down the chain
- Electrons (from NADH) pass down the electron transport chain while protons (H+) are pumped across the membrane
- This makes a proton gradient (proton motive force)
- Protons in higher concentration on one side of the membrane diffuse through ATP synthase
- *ATP synthase* uses the passage of H+ to make ATP – meaning it uses the electrochemical proton gradient across the membrane to synthesise ATP



Cytochrome oxidase complex -> Also referred to as "Complex IV" is last enzyme of the electron transport chain and passes the electron to O2 which generates water.



- **Cytochrome c oxidase I** is the main subunit of the Cytochrome oxidase complex
- Can have multiple names: COX1, COI, CO1
- Mitochondria encoded protein in eukaryotes



To do: Reconstruct the **Cytochrome c oxidase I** phylogenetic tree from multiple organisms.

The "COI_NCBI_aa.fasta" file contains the amino acid sequences of **Cytochrome c oxidase I** from various organisms.

1) Open the "COI_NCBI_aa.fasta" file and take a look at it…

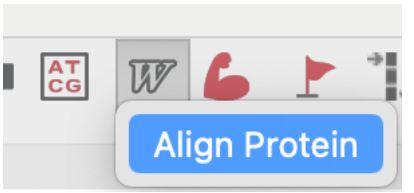
What do the fasta header names tell you about the sequences?

2) Open the MEGA software. Here are links to user manuals:
Mega11:
https://www.megasoftware.net/web_help_11/index.htm#t=First_Time_User.htm
Mega12:
https://www.megasoftware.net/web_help_12/index.htm#t=First_Time_User.htm

3) Open the "COI_NCBI_aa.fasta" dataset (Align option)– take a bit of time to look at the sequences
   - What do you think the colour code represents? (consider your lecture on biomolecules)

Amino Acid	Abbreviations	
	3-letter	1-letter
Glycine	Gly	G
Alanine	Ala	A
Valine	Val	V
Leucine	Leu	L
Isoleucine	Ile	I
Proline	Pro	P
Methionine	Met	M
Phenylalanine	Phe	F
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Serine	Ser	S
Threonine	Thr	T
Asparagine	Asn	N
Glutamine	Gln	Q
Cysteine	Cys	C
Aspartic Acid	Asp	D
Glutamic Acid	Glu	E
Lysine	Lys	K
Arginine	Arg	R
Histidine	His	H

4) First you will need to align the amino acids
   - arranging the sequences to identify regions of similarity.
   - Use Clustal W - a general purpose multiple alignment program for DNA or proteins
   - Use default values

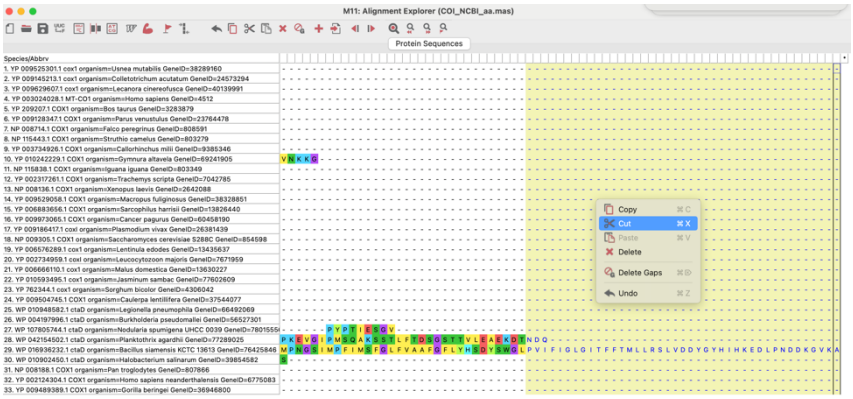


   - Can you identify portions of the proteins that are conserved and ones that are not?

   - Can you identify potential insertions or deletions in the sequences of these organism?

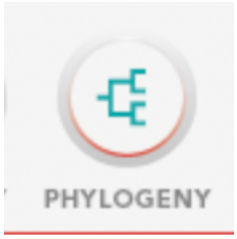
5) Optional: you can manually edit the alignment

- Trim the ends that may not be useful
- Add or remove spaces

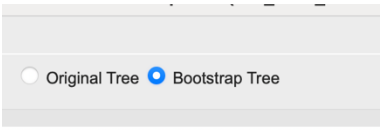


6) Save the alignment as a .mas file

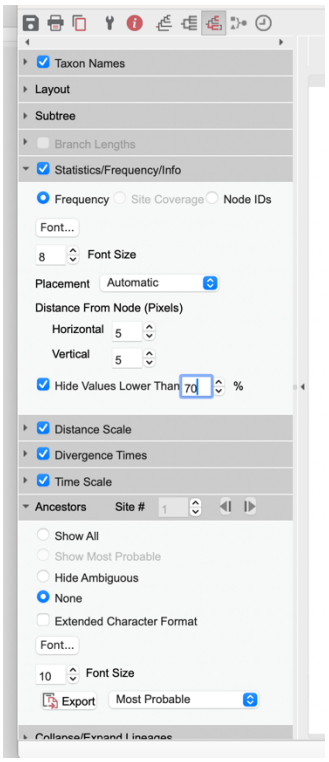
7) Construct a Maximum Parsimony tree (use the .mas file you created)



- Use the Bootstrap method to test how reliable your phylogeny is. To do this you need to select "Test of Phylogeny" in the options **before** running the analysis.
- The program might prompt you to calculate branch length distances. Don't do this as it takes too long.
- After the tree is made, you mind need to specifically select the bootstrap tree.



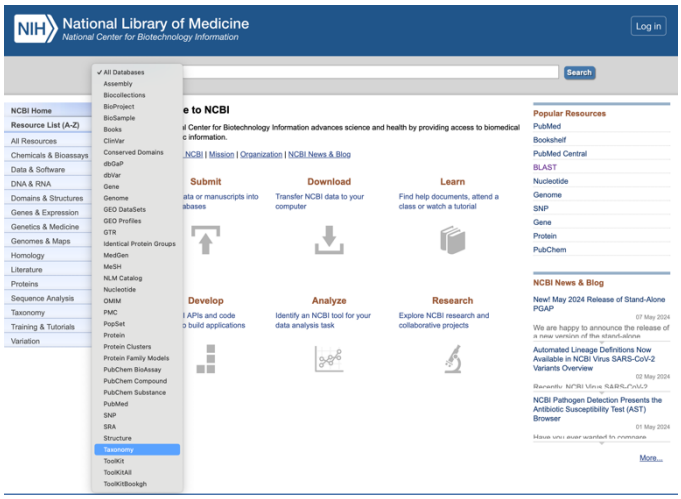
8) You can make cut-off values for the % confidence of trees that recover each phylogenetic relationship

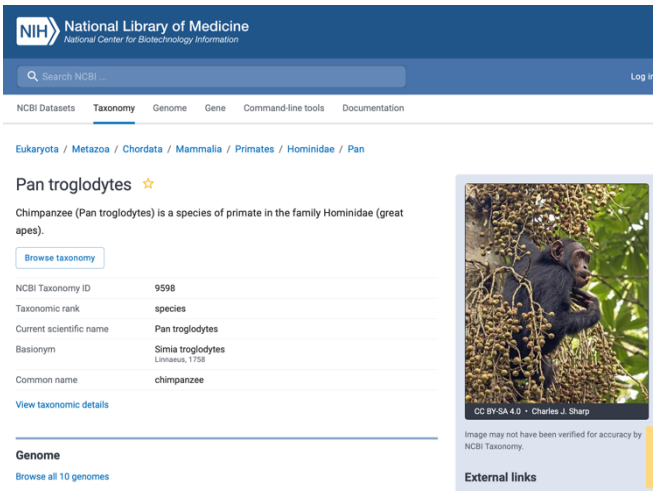


9) Explore the tree:

- Which taxa seem related?
- Do you see any patterns that make sense, for example are organism with vertebrates more closely related to each other? What are their sister taxa?
- Are there any polytomy groups (consider bootstrap values when you think about this)
- You can look up taxonomic identities and classification of these organism on NCBI
- You can also google
- Are there prokaryotes in the tree? Why do you think this is? What function does this protein serve in prokaryotes?

- Why is ***Cytochrome c oxidase I*** useful for reconstructing phylogenies across the tree of life? Check out ***Castresanal et al.*** paper (PDF on Moodle)





10) Try using other methods for constructing trees and testing phylogeny
    - Some may be computationally long, don't focus on these (trial and error)
    - Do you get the same results?

11) Using NCBI add 5 (or more!) organisms of your choice to the fasta file and see how the relationship changes between the taxa.
    - There are multiple ways of getting the amino acid sequence of the protein you are interested in, for example you can use BLAST to look for similar sequences as in your fasta file – you can choose which organisms you want to find related species of.
    - Or you can search in the Gene or Protein database of NCBI (either should allow you to get the AA sequence) using the protein / gene name.

12) You also could have reconstructed the phylogenetic trees with the nucleotide sequence of genes. You have the fasta file of the nucleotide sequence available to you as well "COI_NCBI_na.fasta".
    - Try aligning and reconstructing the tree with this (this will take longer, more computational demand) – I have provided the aligned file .mas if your laptop is struggling.
    - Did you have any trouble?
    - How does the alignment look?
    - Do you need to manually edit it more than the AA one?
    - Does the tree look similar?

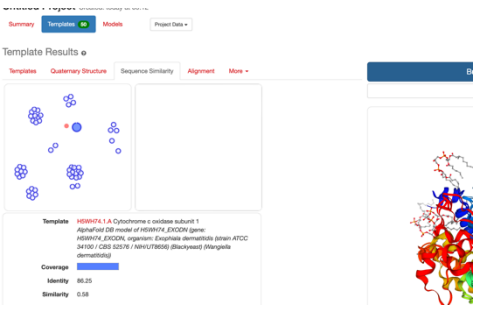
- Why do you think there is discrepancy?
- Which do you think is better?

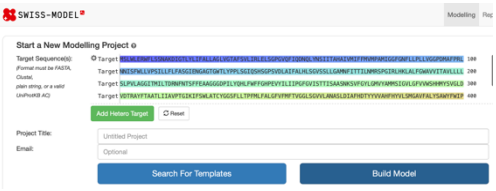
13) Lets explore the 3D structure of this protein.

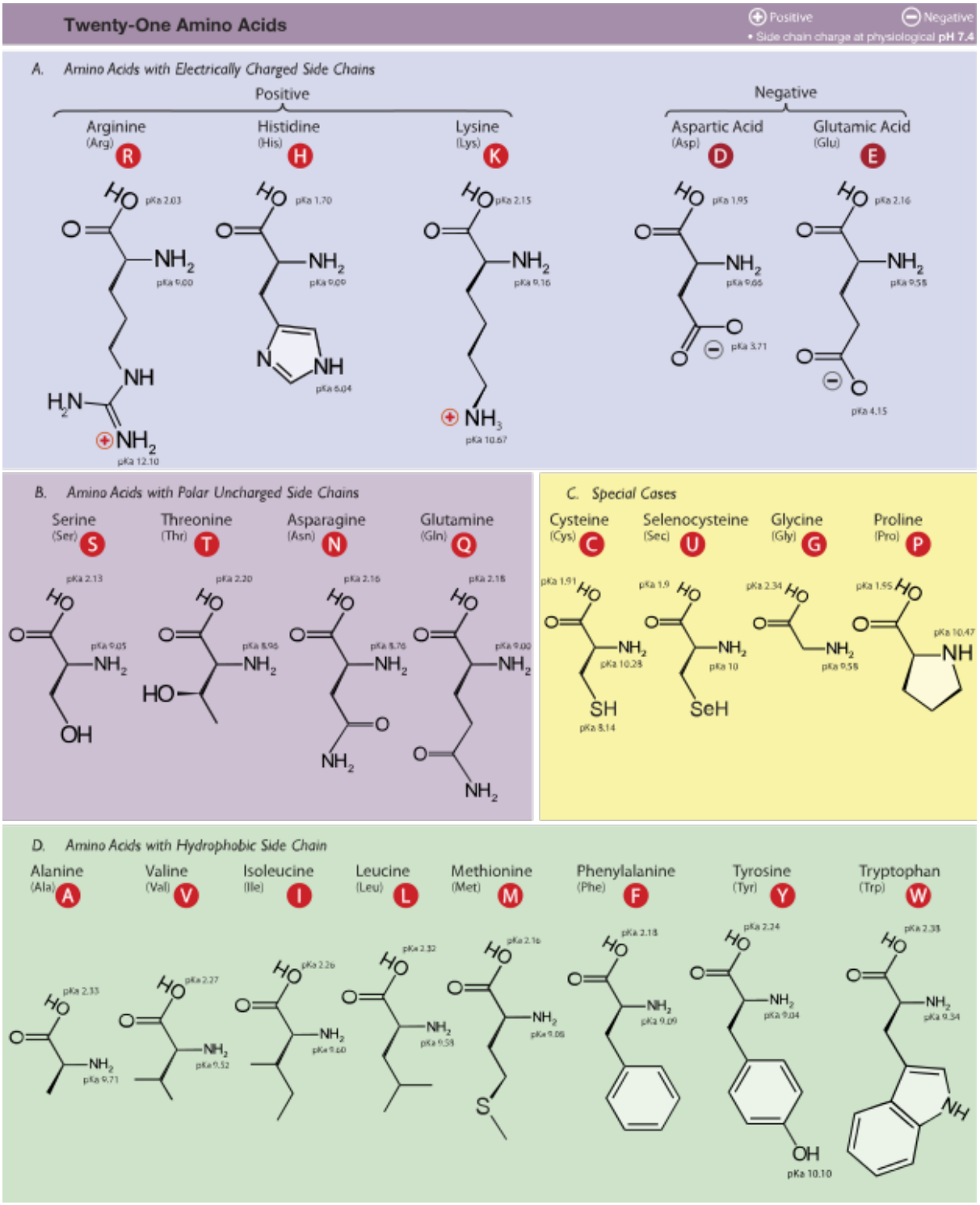
- Go on https://swissmodel.expasy.org
- This is an online resource for modelling 3D structure of AA sequences

14) Try modelling the 3D structure of a few of your proteins (take divergent ones for example)

- How is their 3D structure different?
- Explore the information that the site can provide.
- Look at related Templates (verified models of proteins)






Twenty-One Amino Acids
Positive
Negative
• Side chain charge at physiological pH 7.4
A. Amino Acids with Electrically Charged Side Chains
Positive
Arginine (Arg) R
Histidine (His) H
Lysine (Lys) K
Negative
Aspartic Acid (Asp) D
Glutamic Acid (Glu) E
B. Amino Acids with Polar Uncharged Side Chains
Serine (Ser) S
Threonine (Thr) T
Asparagine (Asn) N
Glutamine (Gln) Q
C. Special Cases
Cysteine (Cys) C
Selenocysteine (Sec) U
Glycine (Gly) G
Proline (Pro) P
D. Amino Acids with Hydrophobic Side Chain
Alanine (Ala) A
Valine (Val) V
Isoleucine (Ile) I
Leucine (Leu) L
Methionine (Met) M
Phenylalanine (Phe) F
Tyrosine (Tyr) Y
Tryptophan (Trp) W