Introduction week

1. *Plain text*

A. Download the three files sent to you the email from the instructor. Save them in a Folder with your name on the Desktop.

B. Open the .docx file with Word, the .rft file with Wordpad and the .txt file with Wordpad. At first glance, do these file look similar?

C. Open all three files with Notepad++. At first glance, do these files look similar?

D. For most applications in bioinformatics, you need to use PLAIN TEXT format (.txt). It is recommended that you use Wordpad or Notepad++ for this class. Verify that you save your files as .txt and not as .rtf and .docx. Never use .docx for sequence or data files.

2. *Jalview sequence viewer*

A. The sequences in the .txt file are in FASTA format, but for this exercise the names are replaced with a simple letter code.

```
>SEQUENCE_NAME
MDQLGEEDPDPSLSPPLSQETFEEIWALKMISPFMNTEQLMASTQPIAFPEEGASAMEQNTYPLLDPHVSTTPLAPLAEG
YLNGGDFVIPPANHYLANPGVMLQLAEDYPPGDSGVFPPTEDYPGCYGFNLEFEQSGTAKSVTYTYSPVLNKLFCQMGKT
```

Save the plain text file as dataset1_your_name.fa

B. Use Mozilla Firefox to open Jalview webstart (http://www.jalview.org/). If it asks if you want to update Java, say Later.

Jalview opens a few example windows upon starting up. When all are open and nothing seems to happen, close the "internal" windows, but keep the main Jalview window open.

Under File, open dataset1_your_name.fa

This opens up and it is all a bunch of letters in black and white. What are the letters? Color by Taylor and it is a bit easier to look at. How are sequences organized? What is the main difference between the n-term and the c-term? Why?

C. In Jalview, Select all.

Go to Web Service, Alignment: Muscle with default

This builds a multiple sequence alignment (MSA), aiming to line up evolutionary related (homologous) sites into columns. When no homology is found a gap ('dash') is inserted.

When the alignment is done, color by Taylor.

D. How has the alignment step changed what you see on your screen?

Can you find sites that are 100% conserved (not changing from one sequence to another)? Can you find non-conserved sites?

What does Conservation and Consensus show?

E. Mouse over the alignment. If you place the cursor over a residue, a number occurs in the lower left. What does this number mean?

E. Save your multiple sequence alignment as dataset1_your_name_muscle.fa

3. *BLAST - what is your mystery sequence?*

You have been assigned a sequence. Copy that sequence only and paste into the BLAST window at NCBI:

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

Make no changes and press BLAST towards the lower end of the screen.

A. What is the protein (accession number)?
B. What species is it from (taxonomic AND common name)?
C. What is the e-value?
D. How many sequences are 100% identical?
E. How similar is your sequence to the human sequence?
F. What is your poorest hit (based on the default sorting)? (Accession number, % sequence identity to your protein, e-value)?

4. Substitution matrices and k-mers

A. Which are the 3-letter k-mers for the following sequence: GSRPTYKLF

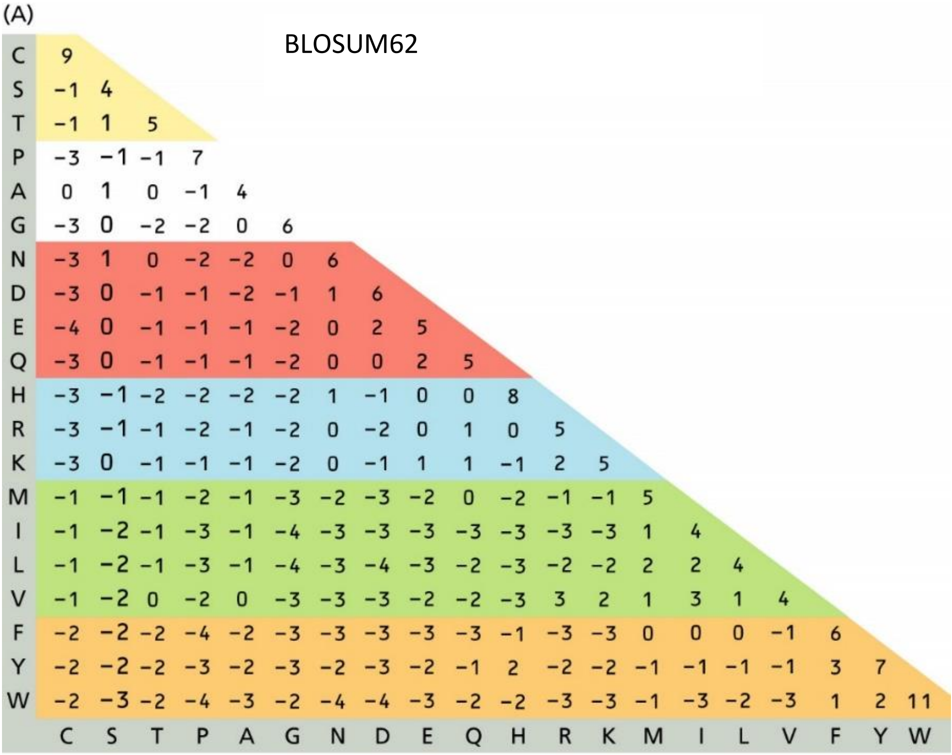
B. Calculate the score and rank the following k-mers by score.

```
1.      SPL                 4. SPL
        +||                    |
        TPL                    AWK

2.      SPL                 5. SPL
        ||+                    | |
        SPI                    SWL

3.      SPL           Use the scores from
        | |           BLOSUM62
        SAL
```

3. How are the amino acids colored and arranged in the matrix?



(A)

BLOSUM62

C	9																			
S	−1	4																		
T	−1	1	5																	
P	−3	−1	−1	7																
A	0	1	0	−1	4															
G	−3	0	−2	−2	0	6														
N	−3	1	0	−2	−2	0	6													
D	−3	0	−1	−1	−2	−1	1	6												
E	−4	0	−1	−1	−1	−2	0	2	5											
Q	−3	0	−1	−1	−1	−2	0	0	2	5										
H	−3	−1	−2	−2	−2	−2	1	−1	0	0	8									
R	−3	−1	−1	−2	−1	−2	0	−2	0	1	0	5								
K	−3	0	−1	−1	−1	−2	0	−1	1	1	−1	2	5							
M	−1	−1	−1	−2	−1	−3	−2	−3	−2	0	−2	−1	−1	5						
I	−1	−2	−1	−3	−1	−4	−3	−3	−3	−3	−3	−3	−3	1	4					
L	−1	−2	−1	−3	−1	−4	−3	−4	−3	−2	−3	−2	−2	2	2	4				
V	−1	−2	0	−2	0	−3	−3	−3	−2	−2	−3	3	2	1	3	1	4			
F	−2	−2	−2	−4	−2	−3	−3	−3	−3	−3	−1	−3	−3	0	0	0	−1	6		
Y	−2	−2	−2	−3	−2	−3	−2	−3	−2	−1	2	−2	−2	−1	−1	−1	−1	3	7	
W	−2	−3	−2	−4	−3	−2	−4	−4	−3	−2	−2	−3	−3	−1	−3	−2	−3	1	2	11
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W

5. Run BLAST with the same sequence against different databases (https://www.be-md.ncbi.nlm.nih.gov/blast/Blast.cgi) Use default BLAST parameters.

```
>AAA60082.1 phenylalanine hydroxylase [Homo sapiens]
MSTAVLENPGLGRKLSDFGQETSYIEDNCNQNGAISLIFSLKEEVGALAKVLRLFEENDVNLTHIESRPS
RLKKDEYEFFTHLDKRSLPALTNIIKILRHDIGATVHELSRDKKKDTVPWFPRTIQELDRFANQILSYGA
ELDADHPGFKDPVYRARRKQFADIAYNYRHGQPIPRVEYMEEEKKTWGTVFKTLKSLYKTHACYEYNHIF
PLLEKYCGFHEDNIPQLEDVSQFLQTCTGFRLRPVAGLLSSRDFLGGLAFRVFHCTQYIRHGSKPMYTPE
PDICHELLGHVPLFSDRSFAQFSQEIGLASLGAPDEYIEKLATIYWFTVEFGLCKQGDSIKAYGAGLLSS
FGELQYCLSEKPKLLPLELEKTAIQNYTVTEFQPLYYVAESFNDAKEKVRNFAATIPRPFSVRYDPYTQR
IEVLDNTQQLKILADSINSEIGILCSALQKIK
```

A. Run blastp with the sequence above against the nr database.
   i. How many sequences are in the nr database?
   ii. What taxonomics group do you get hits in?
   iii. What is the query cover, e-value, and sequence id for the first and last hits?

B. Run blastp with the sequence above against the refseq database.
   i. How many sequences are in the refseq database?
   ii. What taxonomic group do you get hits in?
   iii. What is the query cover, e-value, and sequence id for the first and last hits?

C. Run blastp with the sequence above against the refseq database but specify the following organisms: *Homo sapiens, Mus musculus, Monodelphis domestica, Gallus gallus, Danio rerio, Drosophila melanogaster, Dictyostelium discoideum, Zea mays*

   i. Which species do you get hits in?

   ii. What is the query cover, e-value, and sequence id for the first and last hits?

   iii. Which species do you get “true” hits in? (here let “true” mean e-value < 1e-20 and > 25% query cover, but this is set on a case-by-case basis)

   iv. What are common names for the true hit proteins? How are Drosophila’s hits names different?

   v. The hit set has isoforms. Some are listed to be isoforms, but to be sure, we need to check the GeneID. As an example, we will look at the hits in *Drosophila melanogaster*.

      - Click in the box for every Drosophila hit to select them. Show GenPept for all Drosophila hits.

      - Scroll down the GenPept to GeneID.

      - How many different Drosophila genes did you get?

      - Which of the Drosophila isoforms are from the same gene?

   vi. Download and save all true hits as FASTA complete sequence.

6. Multiple sequence alignments

A. Open Jalview (http://www.jalview.org/, Launch Jalview Desktop, Open with Web Start)

B. Place all sequences you downloaded last time into Jalview (there are some different ways, you can e.g. open the file or paste the sequences in FASTA format into the textbox.).

C. Select all sequences

D. Align with Tcoffee

E. Save in fasta format, name the file: tcoffee.mfa

F. Do not close the unaligned sequence window, but do close the tcoffee.mfa file (but remember where it is saved, you will need it again)

G. Repeat steps C (starting from the unaligned sequences) to F 3 times with three different multiple sequence alignment methods: muscle, mafft, and clustal. In the end, you should have save 4 files, named tcoffee.mfa, muscle.mfa, mafft.mfa, and clustal.mfa.

H. Your next task is to compare the alignments built with different methods. You can compare with your neighbors, displaying one each, but also, do the following:

I. Open the .mfa files, one at the time in a plain text editor, such as Notepad ++.

J. For tcoffee.mfa, replace '>' with '>tcoffee_' (use Find & Replace), save as tcoffee2.mfa

K. For mafft.mfa, replace '>' with '>mafft_' (use Find & Replace), save as mafft2.mfa

L. For muscle.mfa, replace '>' with '>muscle_' (use Find & Replace), save as muscle2.mfa

M. For clustal.mfa, replace '>' with '>clustal_' (use Find & Replace), save as clustal2.mfa

N. Make a plain text file that contains all content from tcoffee2.mfa, muscle2.mfa, mafft2.mfa, and clustal2.mfa. Call it MSA_compare.mfa

O. Open MSA_compare.mfa in Jalview.

P. Color by Taylor.

Q. Under View, select the Overview window.

R. Are the different alignment methods producing different alignments? What seem to agree and disagree? Do you notice any patterns?

7. Brief Introduction to BASH (credits: Christian Balbin)

The terminal is used to control the computer in a non-graphical manner. Many data science applications do not have a GUI (Graphical User Interface). As you become familiar with BASH you will find removing the complexity of a GUI will actually increase productivity and reduce tediousness in many applications.

A. Open up Bash on Ubuntu on Windows. If on Mac, open up Terminal

B. What do you see?

C. The terminal automatically opens up in your home directory. This is the root directory.

D. Type `ls` (list) and hit enter.

  a. What is the command `ls` used for? What did the terminal print to the screen?

  b. Type `mkdir Desktop`

  c. Type `ls` what has changed?

  d. Type `pwd` to see your print working directory (this is where your are)

E. Type `cd Desktop` and hit enter. (Change Directory)

  a. What directory (called folder in the GUI environment) are you currently working in?

F. `ls` in this directory, did the output change?

G. Let's keep our files neat. Type `mkdir bioinformatics` in your current directory (Desktop).

H. `ls` again in your current directory

  a. What's new?

I. Now let's enter our newly created directory.

J. Let's make a text file using a command line text editor. Type `nano` and press enter to use the GNU nano text editor.

K. Type "`hello world`" into the text editor.

L. Press control + x

M. Press y indicating you want to save your file

N. Name your file `mydoc.txt`, press enter

O. `ls` again in your current directory. Do you see your newly created file?

P. Let's make a backup of this file and store it in our home directory

Q. Use the command `cp` (copy) to make a copy of the file

R. The first argument is the path to the file you want to copy, the second argument is path where you want the copy to be saved. Name the backup `mydoc.txt.bkp`
(`cp mydoc.txt ~/mydoc.txt.bkp`)

Hint: `~/` stands for your home directory, `../` stands for one directory before the current working directory, `../../` stands for two directories above the current working directory etc. Knowing this, what are two possibilities for the second argument of the `cp` command in order to back up our file to your home directory?

a. cd to your home directory. Do you see your backup?

S. Change back into your bioinformatics directory.

T. The echo command will route any input to standard out, many programs we will work will direct their output to standard out

U. Type `echo hello world`, press enter

V. Type `echo this will overwrite text > mydoc.txt`

W. Hint: when providing the path to `mydoc.txt` simply *type m and hit tab for auto completion*

X. View the contents of mydoc.txt without opening it by using `cat mydoc.txt`

Y. What do you notice?

Z. Type `echo this will append text >> mydoc.txt`

AA. `cat` the file, what do you notice?

BB. let's mv (move) our backup back into our bioinformatics folder. The syntax is similar to the copy command,

CC. the first argument is the path to the file you want to move. The second argument is where you want to move it.

DD. How can you move your backup into your bioinformatics folder without leaving the bioinformatics directory?

EE. Hint: Bash interprets . as your current directory

FF. Hint: Remember the ~/ or ../ syntax from earlier)

GG. There is no rename command in Linux. You simply mv a file from one name to the another

HH. Rename the mydoc.txt.bkp to mydoc.txt using mv
Note: let's not *delete* anything yet, we'll get to that later.

8. Building a dataset using local blast and the command line

You will rebuild the dataset using the same species as previously but excluding *Zea mays* since there was not a true hit in this species. You will use a small database that only contains Uniprot canonical reference proteomes for the following species: *Homo sapiens, Mus musculus, Monodelphis domestica, Gallus gallus, Danio rerio, Drosophila melanogaster,* and *Dictyostelium discoideum.*

UniProt mission is *"*to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information."

Canonical means that there is only one isoform per gene so you don't have to consider isoforms.

Reference proteome means that these proteomes cover well-studied organisms but it is important to note that the annotations can be done both manually and algorithmically. Even if these are considered high-quality data, be alert to potential errors.

1. We need to find the Uniprot acession number for the query used in exercise 5 (AAA60082.1) by performing a BLAST at **https://www.be-md.ncbi.nlm.nih.gov/blast/Blast.cgi**
   a. Specify the UniprotKB database and Homo sapiens
   b. One hit is 100% identical (P00439.1) and the top of GenPept looks like this:

```
LOCUS       PH4H_HUMAN               452 aa            linear   PRI 30-AUG-2017
DEFINITION  RecName: Full=Phenylalanine-4-hydroxylase; Short=PAH; AltName:
            Full=Phe-4-monooxygenase.
ACCESSION   P00439
VERSION     P00439.1
DBSOURCE    UniProtKB: locus PH4H_HUMAN, accession P00439;
```

The accession number is P00439

2. Open up Bash on Ubuntu on your desktop.

3. Find the query sequence by extracting the sequence with accession P00439 from the local database that is located at: ____

   (exact location of the database will be given in the classroom)

   Use the following command:

```
  blastdbcmd -db /mnt/c/Deployment/BSC4434/combined.fa -entry
P00439 > P00439.fa
```

   `blastdbcmd` lets us communicate with the database and this grabs sequence P00439 from the database and creates a new fasta file called P00439.fa. Use `cat P00439.fa` to see what it is in it.

4. To run BLAST, we need to determine which algorithm to run. In this case the query is protein and the database is protein so blastp seems appropriate. Use the following command:

```
blastp -db /mnt/c/Deployment/BSC4434/combined.fa -query P00439.fa
```

This will write your results to the screen and it looks similar to the results from the NCBI blast server. At the end of the results, you will have information about the database. How many sequences are in the database?
Also, what matrix and which gap penalties were used?

Let's send the results to a file instead. Use the following command:

```
blastp -db /mnt/c/Deployment/BSC4434/combined.fa -query P00439.fa >
blastp.out
```

If we use `cat blastp.out` we can see what is in the file, but since this file is long try the following command instead:

```
       head -100 blastp.out
```

This will show the first 100 lines in the file and this is where we find the list with hits and the corresponding e-values. Typically, we would look for huge jump in e-value to determine which sequences to include (the good hits) and exclude (the poor hits) from the blast hits. From your blastp.out, what e-values are bordering the jump in e-value?

Next, you want to rerun blastp, **specifying an e-value cutoff (X)** that will include all "good" hits but exclude the "poor" hits. Use the following command, but replace X with an e-value.

```
blastp -db /mnt/c/Deployment/BSC4434/combined.fa -query P00439.fa -
evalue X > blastp_evalue.out
```

Look at the top 100 lines in your output to see if you captured what you meant to capture with the specified e-value. If not, redo until you do.

When you have only the output you desire (the good hits), run blastp again but this time specify a different way to view the output (remember to **replace X with your e-value cutoff**):

```
blastp -db /mnt/c/Deployment/BSC4434/combined.fa -query P00439.fa -
evalue X -outfmt "6 sacc qstart qend sstart send evalue bitscore score
length qcovs pident stitle" > blastp_table.out
```

What does this output show? What is listed in the different columns?

Column 1 has the accession codes for the sequences we have identified. We need to have a file with only the accession codes, listed in one column. Modify the command above to give only the column you are interested in. Call the new file `blastp_accessions.out.`

Next, we need to extract the sequences corresponding to the accessions from the database, so we need to communicate with the database again:

```
blastdbcmd -db /mnt/c/Deployment/BSC4434/combined.fa -entry_batch
blastp_accessions.out > blastp_results.fa
```

What do you find if you `cat blastp_results.fa`

This file should contain 27 sequences and while we should be able to count to 27, it is tricky to count on the screen, and whatif you have 300 sequences? Instead, let the computer count for you. By searching for patterns using regular expressions, we can search through the file and count the number of times we have a new sequence. In the command below, grep is tool that lets you search for a pattern. -c instructs grep to count how many times the 'pattern' occurs. What pattern (it can be as short as one character) could you use to count how many sequences are in blastp_results.fa? Try it:

```
grep -c 'pattern' blastp_results.fa
```

(e.g. if you pattern is ## you would type: `grep -c '##' blastp_results.fa`)

Confirm that you have 27 sequences in total.

Also, use grep to confirm that you have:

4 Homo sapiens, 4 Mus musculus, 4 Monodelphis domestica, 5 Gallus gallus, 6 Danio rerio, 3 Drosophila melanogaster, and 1 Dictyostelium sequence in the dataset.

Next, copy blastp_results.fa to Documents?

```
cp blastp_results.fa /mnt/c/Users/vh133/Documents/my_blast_results.fa
```

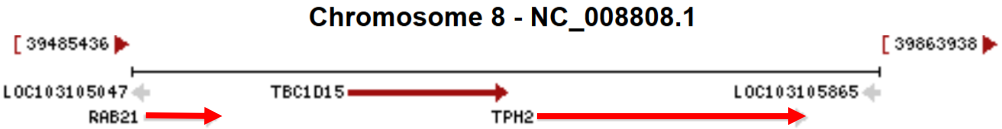
Last, go to Windows and copy my_blast_results.fa to your USB drive!

9. Building trees from protein sequence

Part I

(Please complete 9A before class on Monday Oct 2. Remember to bring the files to class)

A. Use the sequence data in blastp_results.fa that was generated using local blast. However, there is one sequence (F7GI81_MONDO) in this file that has a faulty annotation. This entry contains sequence from 3 consecutive genes (red arrows) on chromosome 8 in the Monodelphis genome:



We only have a hit for the TPH2 region and will simply replace it for now and discuss what has happened with this sequence later. Make a new file that has the same sequences as blastp_results.fa except that it does not have F7GI81_MONDO and instead has the sequence below (use only the high-lighted accession code to search for this sequence in NCBI).

```
>XP_001363075.1 PREDICTED: tryptophan 5-hydroxylase 2 isoform X1 [Monodelphis
domestica]
MQPAMMMFSSKYWARRGLSLDSAVPEEHQLLSSLTFNRANDEKKGNKGSGKNEVAGESGKTAVVFSLKNE
VGELVKALRLFQEKHVNMVHIESRKSRRRNSEVEIFVDCDCNKKEFNELIQLLKFQTTIVTLNPPENIWT
EEEDLEGVLWFPRKIAELDKCSHRVLMYGSELDADHPGFKDNVYRQRRKYFVDVAMSYKYGQPIPRVEYT
EEETRTWGVVFRELTKLYPTHACREYLKNLPLLTKYCGYREDNVPQLEDVSIFLKERSGFTVRPVAGYLS
PRDFLAGLAYRVFHCTQYVRHGSDPLYTPEPDTCHELLGHVPLLADPKFAQFSQEIGLASLGASDEDVQK
LATCYFFTIEFGLCKQEGQLRAYGAGLLSSIGELKHALSDKACVKAFDPKTTCLQECLITTFQEAYFVSE
SFEEAKEKMRDFAKSITRPFSVYFNPYTQSIEILKDTRSIENVVQDLRSDLNTVCDALSKMNKYLGI
```

Save as blastp_results_MD_replaced.fa

You will use this dataset to build phylogenetic trees with PHYML. PHYML uses as input a multiple sequence alignment in PHYLIP format. The sequences in blastp_results_MD_replaced are in fasta format, but that can easily be converted with Jalview. However, saving as PHYLIP truncates the names so you need to shorten the names to 9 characters or less. Rename the sequences in a logical manner. This is best done in the FASTA document in a simple text editor. For instance:

```
>sp|P18459|TY3H_DROME Tyrosine 3-monooxygenase OS=Drosophila melanogaster GN=ple …

>tr|Q9PU40|Q9PU40_CHICK Tyrosine hydroxylase OS=Gallus gallus GN=tyrosine hydroxy…
```

Can be renamed as shown in **BOLD** (also note the space between the new and old name)**:**

>**TH_DM** sp|P18459|TY3H_DROME Tyrosine 3-monooxygenase OS=Drosophila melanogaster GN=…

>**TH_GG** tr|Q9PU40|Q9PU40_CHICK Tyrosine hydroxylase OS=Gallus gallus GN=tyrosine hydro…

As you can see, for Drosophila melanogaster, the abbreviated species name is the first letter of ***D**rosophila* and ***m**elanogaster* and for chicken (CHICK), the abbreviated species name is the first letter of ***G**allus* and ***g**allus*. The species in the dataset are:

Homo sapiens HUMAN [HS]	Mus musculus MOUSE [MM]
Monodelphis domestica MONDO [MD]	Gallus gallus CHICK [GG]
Danio rerio DANRE [DR]	Drosophila melanogaster DROME [DM]
Dictyostelium discoideum DICDI [DD]	

Use the abbreviation within the brackets.

To rename the protein names, use the following:
Phenylalanine hydroxylase or phenylalanine monooxygenase [PAH]
Tyrosine hydrolase or tyrosine monooxygenase [TH]
Tryptophan hydroxylase or tryptophan monooxygenase [TPH]
If e.g. Tryptophan hydroxylase 1 [TPH1] or tyrosine hydroxylase-like [THL]
Henna [HN]
Uncharacterized protein [UP]

Note: All names must be unique! You may need to include TPH1, TPH2, TH2, UP1, UP2, etc. Double check that all names are unique before you open this file in Jalview!!!

B. The space after the new and the old name is critically important. When you open this FASTA file in Jalview, it will recognize the new part as the name but it will keep the old name as information.

Make your multiple sequence alignment with Muscle. When you save your alignment, save it twice: 1 as FASTA and 2 as PHYLIP. The FASTA file will have the code for your renamed sequences so you know what name corresponds to which accession number.

C. According to the Phylip alignment, how many sequences are in your dataset and how long is the alignment? (it should be 27 species and 649 characters)

D. Build a tree with Phyml (http://www.atgc-montpellier.fr/phyml/): upload the phylip file, specify what type of data you have, use the smart model test determined by AIC, and build the tree with SH-like support. Name your analysis as you please and specify your email address. Submit. Check that the tree is running. If not, trouble shoot. The tree should take about 10-15 min or less.

E. Run Phyml (http://www.atgc-montpellier.fr/phyml-sms/) again with everything the same except build it with 100 bootstraps. This tree will not be done today but you should have it before Monday. Bring it to class.

[While the trees are running…

10. Species trees, not gene or protein trees

Use NIH taxonomy – Common Tree – to generate a species tree for the species from 5C (except *Zea mays*): *Homo sapiens, Mus musculus, Monodelphis domestica, Gallus gallus, Danio rerio, Drosophila melanogaster, Dictyostelium discoideum*

A. Which species is the outgroup?

B. Save the tree as Phylip tree. (This format is also called Newick.)

C. Open your tree with FigTree (http://tree.bio.ed.ac.uk/software/figtree/): the program must be downloaded, but does not need to be installed, just unzip and click on the .exe file.
(You can save FigTree on your usb for easy access in the future)

D. Generate a nice colorful graphic of your tree, rooted with the outgroup, in FigTree.
Save the graphic as a png (not a screenshot...)

Return to your protein trees]

F. Open the folder with files that you got from PhyML for the two different trees. Which file is different? Which files have your SH-like and your bootstrap tree?

G. Open the SH-like tree in Figtree.

i. Since this tree has node support values, FigTree will ask if you want to call them ‘label’ say yes or rename them as you wish.

ii. Root the tree based on the outgroup

iii. show the node label support.

iv. Make a nice graphical representation,
- make sure that the names are readable,
- the scale bar is shown and its legend is readable
- show the labels at the nodes
- be creative (you can color by clade, by branch and by taxa)
- save the graphic

v. Reroot the tree at mid-point, show the node label support, and save as a graphic (png)

Do the tree tell the same story for the two different roots?

vi. Repeat for the bootstrap tree

11. Building trees from protein sequence

Part II

Interpreting the phylogenies How are the sequences grouped? (by species name or by major protein type: PAH, TH, TPH)

H. Are the nodes well supported (node support > 0.75 (or 75 if 100 bootstraps))? Any exceptions?

I. Compare the sequence trees rooted with the outgroup to the species tree from NIH.

  i. Can you find Gene Duplication (GD) and Gene Loss (GL) events?

  ii. Where (in which protein type and where in the species tree) do your trees indicate that GDs have occurred?

  iii. Where (in which protein type and where in the species tree) do your tree indicate that GLs have occurred?

  iv. Ignoring small differences in branch lengths, do all your trees tell the same story?

J. Compare the sequence trees rooted midpoint to the species tree from NIH.

  i. Can you find Gene Duplication (GD) and Gene Loss (GL) events?

  ii. Where (in which protein type and where in the species tree) do your trees indicate that GDs have occurred?

  iii. Where (in which protein type and where in the species tree) do your tree indicate that GLs have occurred?

  iv. Ignoring small difference in branch lengths, do all your trees tell the same story?

K. Do the trees rooted with an outgroup and the trees rooted mid-point tell the same story?

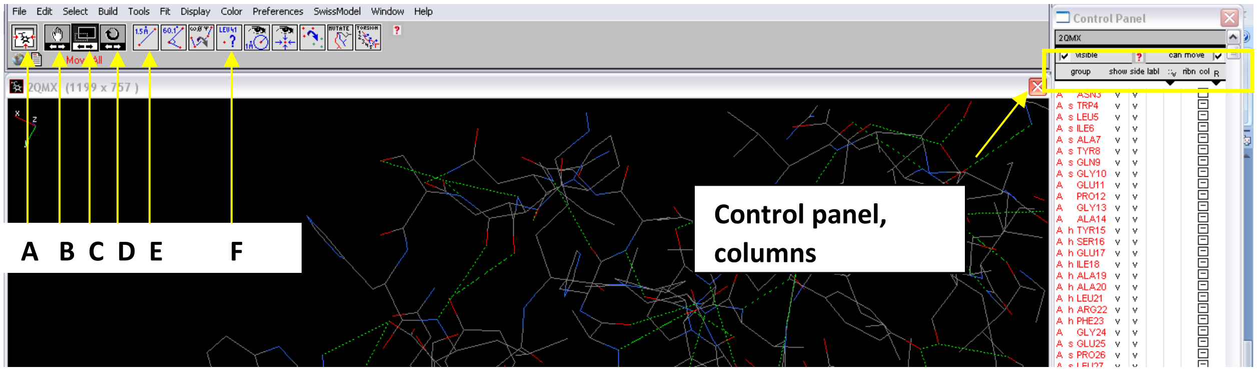
L. What is HN_DM?

M. For the proteins names UP, what do you think their functional annotation could be?

SAVE ALL FILES, KEEP THEM ORGANIZED – WE WILL USE THEM AGAIN!!!

Swiss PDB viewer quick tutorial

Download and install Swiss-PDB viewer from http://spdbv.vital-it.ch/.



A – Centers all the molecules displayed in the viewer

B – Allows you to move the molecules, but does not rotate them

C – Zoom in and out

D – Rotate

E – Allows you measure a distance between two atoms picked on the screen

F – Allows you to label a residue that is picked on the screen.

NOTE: *Questions and answers in the tutorial are for your own use.*

1. Download PDB file 2QMX from www.pdb.org. Open in SwissPDBviewer.

2. Under Window, choose Control Panel. The control panel opens up on the right side of the screen. In the control panel, the columns from left to right are

   i) Group; the chain (e.g. A or B if given in the PDB file), the secondary structure (s – beta strand and h for helix), the residues and their numbering. Chains, secondary structures, and residues can be selected by clicking on them. Holding the shift button down lets you select a residue range by clicking on the first and the last in the range. Holding the Ctrl button down lets you select all you click on. A residue can be deselected by "reselecting" it. Explore how the shift and ctrl button lets you select and deselect in various ways.

   ii) show; left clicking in the column lets you add one residues backbone at the time, while right clicking shows all backbones. If you click on some of the residue names in the group column they will be selected. To show only the backbones of the selected residues, click on the word show at the top of the column

   iii) side, similar to show but acts of sidechain instead of the backbone

   iv) label, allows you to label the residues in the viewer

   v) v acts on surfaces (but this functions is not so good in this program so we will not use it today)

vi) ribn, similar to show and side, but shows the ribbon representation instead

vii) col, shows how the residues are colored, and

viii) the last column shows if the coloring acts on e.g. backbone (B), sidechain (S), or ribbon (R). To change between the different, click on the black arrow and a pull down menu will allow you select another option.

Explore different possibilities. When you proceed, leave the ribbon representation displayed in the viewer.

3. Explore the different coloring options, note how the first option in this vertical menu allows you to choose if you want the coloring command to act on e.g. backbone, side chain, or ribbon. End with coloring by chain acting on ribbon. How many chains do you have in 2QMX? *The answer is 2 chains.*

4. Select all residues in chain A. Open the Ramachandran plot under Window. If there are any outliers, which are they? *Glycines – we will talk about the Ramachandran plot.*

5. Represent the entire protein as backbone and sidechain (right clicking in these columns in the control panel will do the trick), hide the ribbon (again, right click in the R column).

    - Under Tools, select Compute H-bonds. All the H-bonds between polar atoms off opposite polarity within 3.3 Å of each other will be shown as green dotted lines.
    - Zooming in and rotating can help you find what you are looking for, but for viewing H-bonds it helps to be specific. For instance, if you want to show all the residues within 6 Å off a ligand, select the ligand (called A PHE303 located at the end of the residues in the control panel), Go to Select, click on Neighbors of selected aa and choose the option Display only groups that are within 6 Å.
    - Center your atoms using the center button. (upper left)
    - Make sure the coloring acts on backbone and sidechain. Color by CPK. Then click on the square in the control panel after the ligand (A PHE303). This will allow you to change the color of the ligand only, select purple. Now, you should see the ligand in purple and the residues with 6 Å of it in CPK. The ligand forms 5 H-bonds to other residues, which are they? *2 H-bonds to the side chain of D224, 2 H-bonds to the backbone of L225, and one H-bond to the backbone of L211.*

6. Color by chain. Is the ligand located at an interface between two chains? *Yes*

7. *Explore a protein structure of your choice from www.pdb.org using SwissPDBviewer to learn to maneuver selecting, displaying, coloring, etc.*

12. Homology modeling

1. Create an account on SwissModeller (https://swissmodel.expasy.org/)

2. Build a model for TPH2_GG from the dataset you used to build the phylogenetic trees. However, in the interest of time, use the accession code Q6PKI7 to grab the FASTA file from NCBI (https://www.ncbi.nlm.nih.gov/). You do not need to do a BLAST search.

3. Use the SwissModel server to identify a template that has Tryptophan as ligand. If more than one such template is found, use the first one.

4. Check the box for the identified template only (you need to recheck the top hit on the list as it is selected by default). When the Build Model button says Build Model 1, click the button.

5. While the model is being built, what is the PDB id and resolution of the template? What protein's structure is represented in that PDB file?

6. When your model has completed, do the GMQE and QMEAN scores indicate that you have a pretty good or a not so good model?

7. What is the sequence identity between the target and the template and what region of the target sequence was modelled?

8. Use the different way to color the alignment and the 3D model in the window to answer the following questions?
   a. Which part of the model have poorer QMEAN4 score?

   b. What is the secondary structure in the N-term and in the C-term, respectively?

   c. Are the polar residues primarily found on the outskirts of the structure or in the middle of the structure (roughly speaking)?

9. What information is found in the Modelling log?

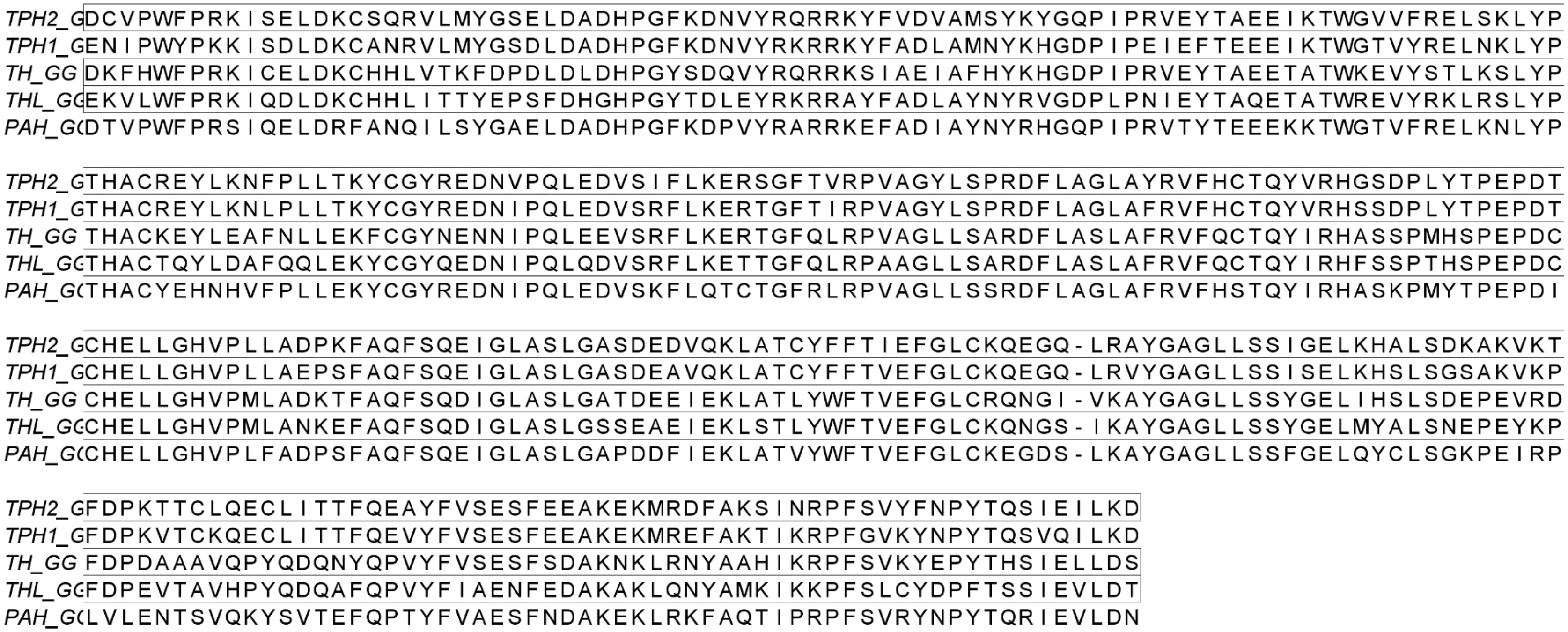
10. What information is found the Model report (save the report as PDF)?

11. What information is found in the Model's PDB format (save the PDB format as plain text)?

12. Mark the positions indicated to contact with the Tryptophan ligand on the alignment that only shows the chicken (GG) sequences from the alignment you used to build the tree previously. →

13. Which positions that bind Tryptophan are conserved/not conserved in the other GG sequences?

14. Hypothesize which of the positions involved in coordinating the Tryptophan substrate may be important for the substrate specificity of TPH?



```
TPH2_G DCVPWFPRKISELDKCSQRVLMYGSELDADHPGFKDNVYRQRRKYFVDVAMSYKYGQPIPRVEYTAEEIKTWGVVFRELSKLYP
TPH1_G ENIPWYPKKISDLDKCANRVLMYGSDLDADHPGFKDNVYRKRRKYFADLAMNYKHGDPIPEIEFTEEEIKTWGTVYRELNKLYP
TH_GG  DKFHWFPRKICELDKCHHLVTKFDPDLDLDHPGYSDQVYRQRRKSIAEIAFHYKHGDPIPRVEYTAEETATWKEVYSTLKSLYP
THL_GG EKVLWFPRKIQDLDKCHHLITTYEPSFDHGHPGYTDLEYRKRRAYFADLAYNYRVGDPLPNIEYTAQETATWREVYRKLRSLYP
PAH_GG DTVPWFPRSIQELDRFANQILSYGAELDADHPGFKDPVYRARRKEFADIAYNYRHGQPIPRVTYTEEEKKTWGTVFRELKNLYP

TPH2_G THACREYLKNFPLLTKYCGYREDNVPQLEDVSIFLKERSGFTVRPVAGYLSPRDFLAGLAYRVFHCTQYVRHGSDPLYTPEPDT
TPH1_G THACREYLKNLPLLTKYCGYREDNIPQLEDVSRFLKERTGFTIRPVAGYLSPRDFLAGLAFRVFHCTQYVRHSSDPLYTPEPDT
TH_GG  THACKEYLEAFNLLEKFCGYNENNIPQLEEVSRFLKERTGFQLRPVAGLLSARDFLASLAFRVFQCTQYIRHASSPMHSPEPDC
THL_GG THACTQYLDAFQQLEKYCGYQEDNIPQLQDVSRFLKETTGFQLRPAAGLLSARDFLASLAFRVFQCTQYIRHFSSPTHSPEPDC
PAH_GG THACYEHNHVFPLLEKYCGYREDNIPQLEDVSKFLQTCTGFRLRPVAGLLSSRDFLAGLAFRVFHSTQYIRHASKPMYTPEPDI

TPH2_G CHELLGHVPLLADPKFAQFSQEIGLASLGASDEDVQKLATCYFFTIEFGLCKQEGQ-LRAYGAGLLSSIGELKHALSDKAKVKT
TPH1_G CHELLGHVPLLAEPSFAQFSQEIGLASLGASDEAVQKLATCYFFTVEFGLCKQEGQ-LRVYGAGLLSSISELKHSLSGSAKVKP
TH_GG  CHELLGHVPMLADKTFAQFSQDIGLASLGATDEEIEKLATLYWFTVEFGLCRQNGI-VKAYGAGLLSSYGELIHSLSDEPEVRD
THL_GG CHELLGHVPMLANKEFAQFSQDIGLASLGSSEAEIEKLSTLYWFTVEFGLCKQNGS-IKAYGAGLLSSYGELMYALSNEPEYKP
PAH_GG CHELLGHVPLFADPSFAQFSQEIGLASLGAPDDFIEKLATVYWFTVEFGLCKEGDS-LKAYGAGLLSSFGELQYCLSGKPEIRP

TPH2_G FDPKTTCLQECLITTFQEAYFVSESFEEAKEKMRDFAKSINRPFSVYFNPYTQSIEILKD
TPH1_G FDPKVTCKQECLITTFQEVYFVSESFEEAKEKMREFAKTIKRPFGVKYNPYTQSVQILKD
TH_GG  FDPDAAAVQPYQDQNYQPVYFVSESFSDAKNKLRNYAAHIKRPFSVKYEPYTHSIELLDS
THL_GG FDPEVTAVHPYQDQAFQPVYFIAENFEDAKAKLQNYAMKIKKPFSLCYDPFTSSIEVLDT
PAH_GG LVLENTSVQKYSVTEFQPTYFVAESFNDAKEKLRKFAQTIPRPFSVRYNPYTQRIEVLDN
```

13. Identifying a binding site in SwissPDBviewer

A. Download the PDB file for 3E2T from PDB. 3E2T was the template used for homology modeling. Open 3E2T in Swiss-PDB-viewer

B. Color the TRP ligand (found at the end of the control panel) so that it is easily distinguished from the protein, e.g. in pink.

C. Compute and Display H-bonds.

D. Select and Display the residues that are within 4 Angstrom of the TRP ligand.

E. Color by selection.

F. Open the Alignment window under Wind. What residues are selected? (These are the residues coordinating the ligand.) Mark these onto the alignment (previous page).

G. Are the residues you identified as involved in making the binding site in the template the same as the residues listed by SwissModel for the model?

H. Are these residues in agreement with the residues suggested by PDB to bind TRP in the template: http://www.rcsb.org/pdb/explore/remediatedSequence.do?structureId=3E2T

14. Conservation of the binding site

I. Open the multiple sequence alignment (MSA) you used to build the phylogenetic tree a couple of weeks ago. Arrange the sequences by PAH, TH, TPH1, and TPH2 clade.

J. Use your alignment from the previous page that marks the residues involved in binding TRP to identify the corresponding sites in the MSA.

K. Which positions that bind Tryptophan and are conserved in the TPH1 and TPH2 clades but have another conserved amino acid in one or both of the other two clades can you find?

L. Again, hypothesize which of the positions involved in coordinating the Tryptophan substrate may be important for the substrate specificity of TPH? If you can nominate one amino acid substitution to test its importance for substrate specificity in TPH, which would it be (e.g. A145G means that the Alanine in position 145 is mutated to become Glycine)?