UTM UNIVERSITI TEKNOLOGI MALAYSIA
SCHOOL OF COMPUTING
Faculty of Engineering

Semester II 2020/2021

Subject : Bioinformatics I (SCSB2103)
Section : 01 – Dr Haslina Hashim
Topic : Lab 03 – Pairwise Sequence Alignment

Name : Gui Yu Xuan
Matric ID : A20EC0039

1) Perform pairwise alignment at the NCBI BLAST website.

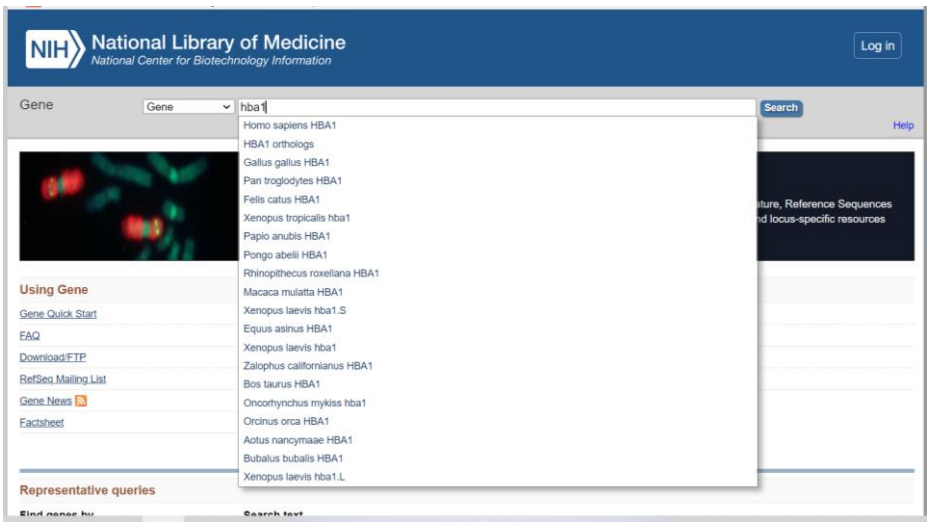
i) Obtain the human HBA and HBB protein sequences.

ii) Perform pairwise alignment at the NCBI BLAST website.

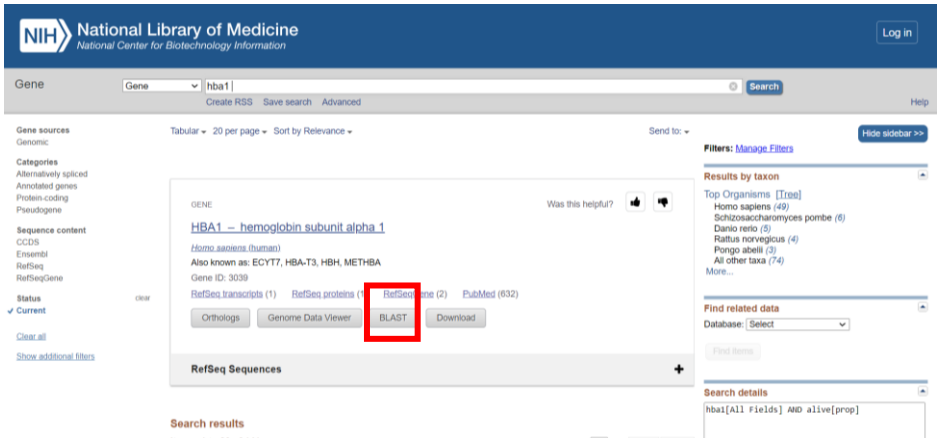
iii) Then use a comparison tool from the EBI website.

iv) Vary the scoring matrix (e.g., try different PAM and BLOSUM matrices) and record the effects on the score, the number of gaps, the percent identity, and the length of the aligned region.

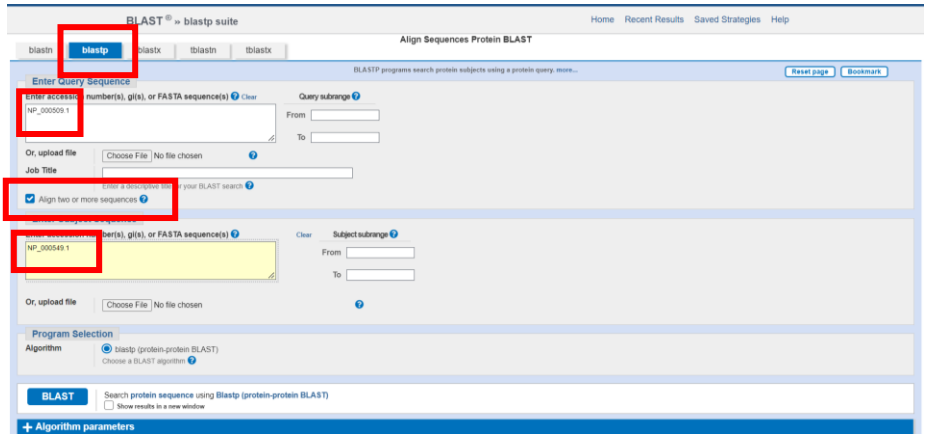
v) For the NCBI BLASTP program note that the output of a pairwise alignment includes a dot matrix view.



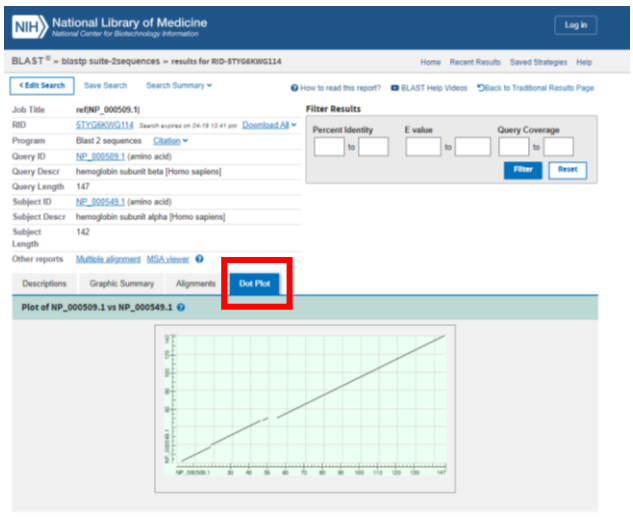
First, go to NCBI websites and type hba1 at the search bar.



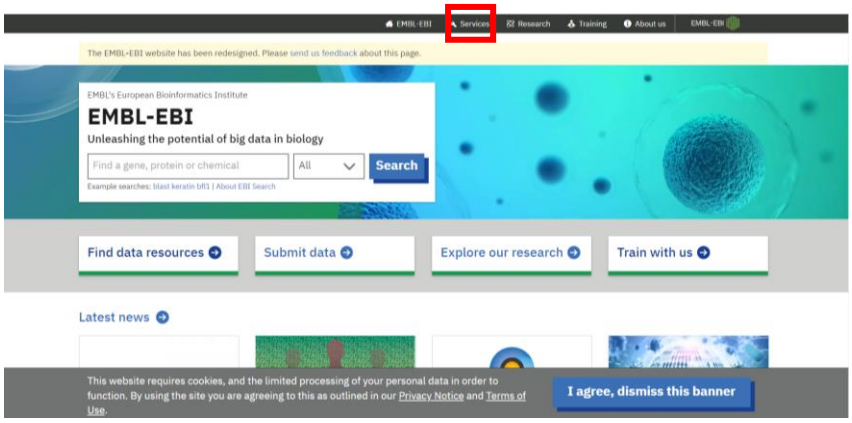
Then, choose the BLAST and click on it.



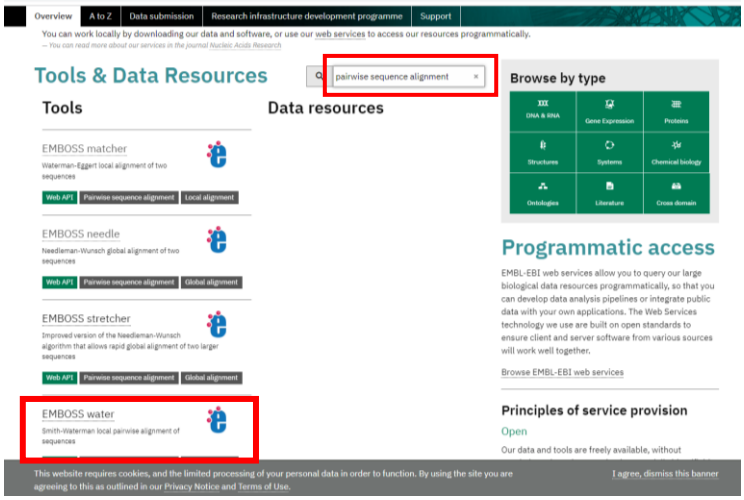
Then, choose the blastp section and enter the information and click on the BLAST.



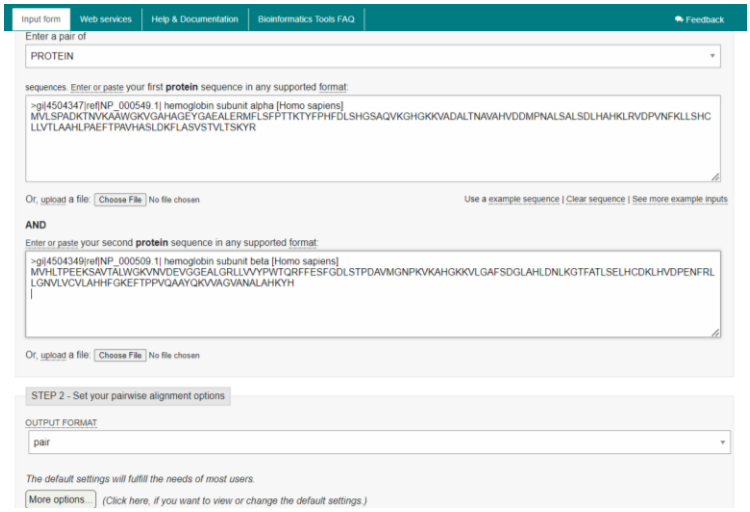
After that, select the dot plot to see the graph and the results.



Next, go to EBI website and click on services.



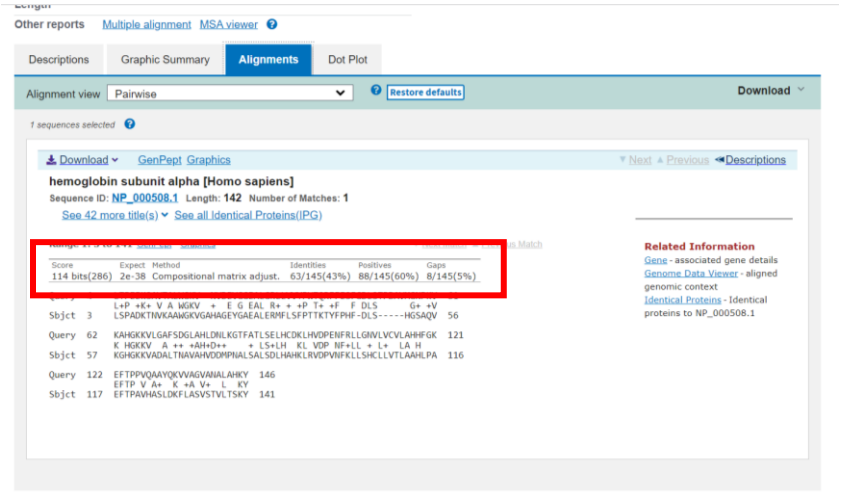
At the search bar type pairwise sequence alignment and choose EMBOSS water.



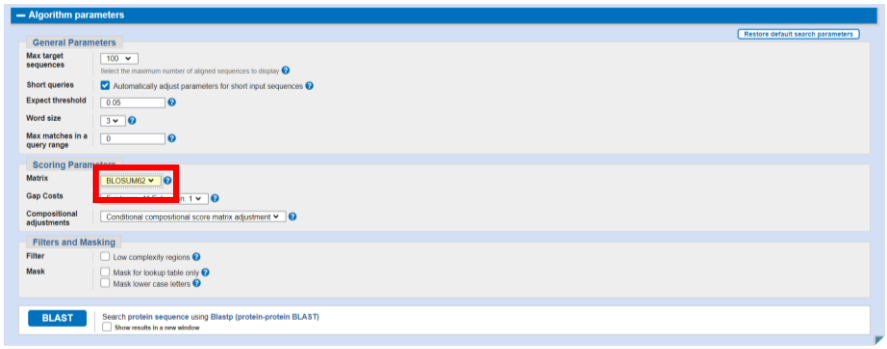
Then, enter the information as shown above, and click submit button. The output will be showed as below.

```
########################################
# Program: water
# Rundate: Mon 18 Apr 2022 06:52:55
# Commandline: water
#    -auto
```

```
#    -stdout
#    -asequence emboss_water-I20220418-065251-0610-14348719-p2m.asequence
#    -bsequence emboss_water-I20220418-065251-0610-14348719-p2m.bsequence
#    -datafile EBLOSUM62
#    -gapopen 10.0
#    -gapextend 0.5
#    -aformat3 pair
#    -sprotein1
#    -sprotein2
# Align_format: pair
# Report_file: stdout
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: NP_000549.1
# 2: NP_000509.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 145
# Identity:      63/145 (43.4%)
# Similarity:    88/145 (60.7%)
# Gaps:           8/145 ( 5.5%)
# Score: 293.5
#
#
#=======================================

NP_000549.1        3 LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-     50
                     |:|.:|:.|.|.||||  :..|.|.|||.|:.:.:|.|:.:|..| |||
NP_000509.1        4 LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLST     51

NP_000549.1       51 ----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDP     96
                         .|:.:||.|||||..|.:::.:||:|::....:.||:||..||.|||
NP_000509.1       52 PDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDP    101

NP_000549.1       97 VNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKY    141
                     .||:||.:.|:..||.|...||||.|.|:..|.:|.|:..|..||
NP_000509.1      102 ENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKY    146


#---------------------------------------
#---------------------------------------
```



From the above results, you can see the result for BLOSUM 62 scoring metrics. Then, at the top left, click on the edit search button.



Under the algorithm parameters, trying to change the matrix under scoring parameters. Then, you may get the results like the table below.

Matrix	E-value	Identities	Positives	Gaps
BLOSUM 62	2e-38	63/145 (43%)	88/145 (60%)	8/145 (5%)
BLOSUM 90	4e-37	63/145 (43%)	81/145 (55%)	8/145 (5%)
PAM 30	4e-34	65/149 (44%)	74/149 (49%)	9/149 (6%)
PAM 250	2e-38	63/146 (43%)	109/146 (74%)	8/146 (5%)

The results of these scoring metrics are almost the same. When two proteins have large different or only some parts of the proteins are well aligned, then a large differences will been seen easily.

2) Perform pairwise alignment at the UCSC website.

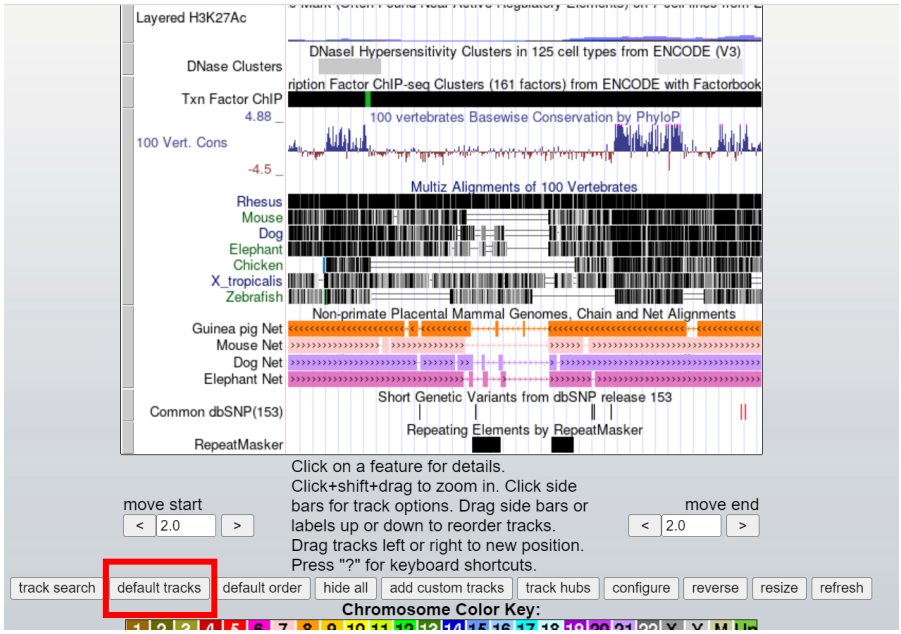
i) Go to http://genome.ucsc.edu. Follow the link to the genome browser, select the human genome hg19 build, and enter a query of hbb. This should direct you to chr11:5,246,696-5,248,301 (a region of 1,606 base pairs encompassing the beta globin gene, HBB.

ii) Click the box to set the view to default tracks.

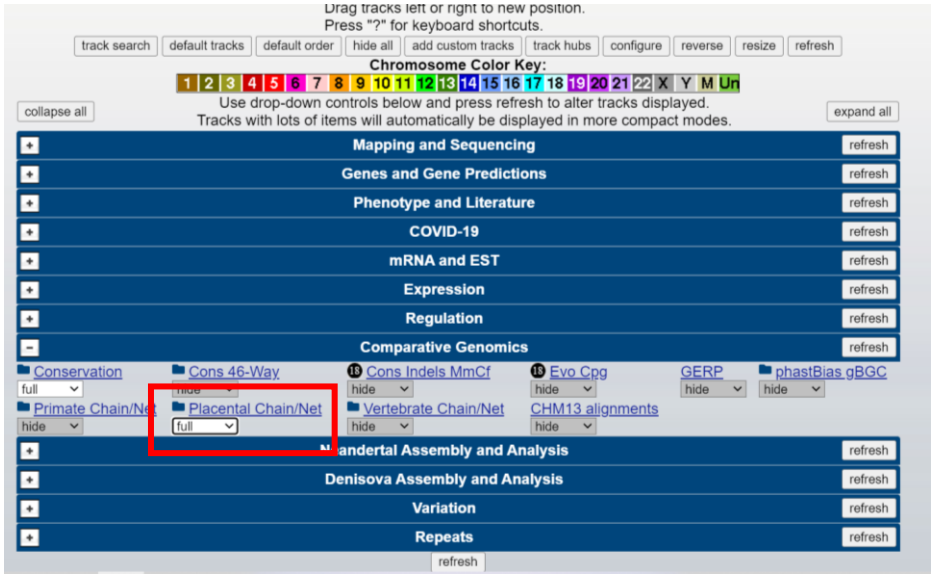
iii) Under Comparative Genomics" select Placental Chain/Net and set the display to full. By clicking the Placental Chain/Net header you can view a series of options. Set Chains to full view and Nets to full view. Set the species to horse (deselect other species). Click submit.

iv) The display now shows human/horse chained alignments and and alignment nets.

First, go to UCSC Genome Browser. Then, move the mouse over the Genomes and select Human GPCh37/hg19.

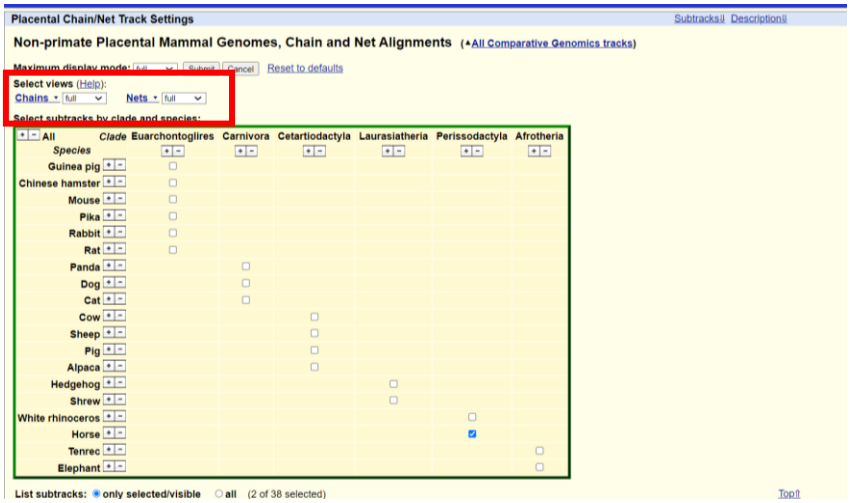


Then, select the default tracks.

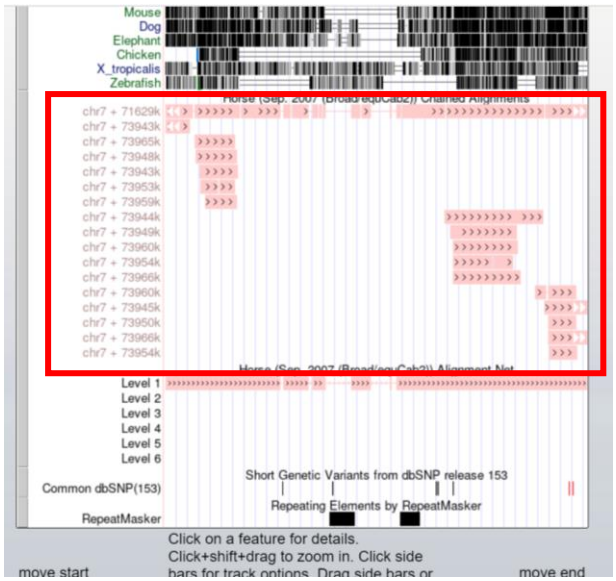


Under the Comparative Genomics, change the Placental Chain/Net to full and click on refresh.

After that, click on the Placental Chain/Net header.



Then, change the Chains and Nets to full. Select the horse species and deselect other species.



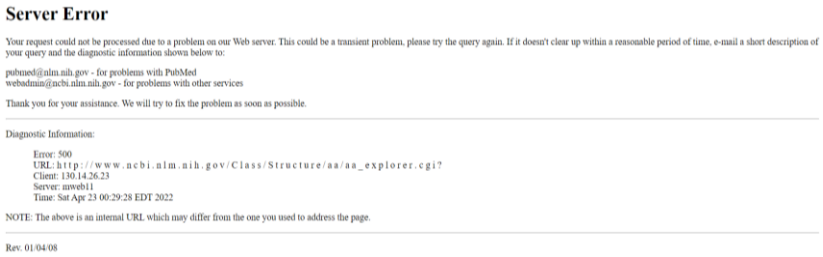
You will be direct to this page. Select the elements inside the box.

Select the DNA sequence alignment. You will see the results as the below picture.

3) Using the amino acid explorer tool from NCBI

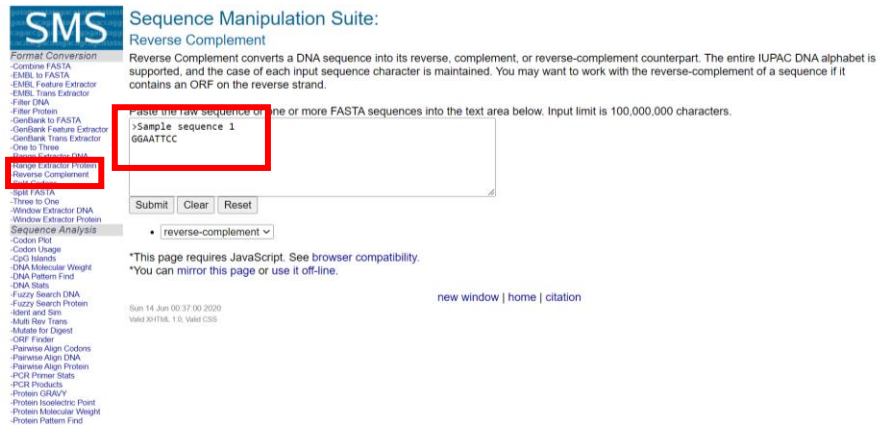
i) Visit http://www.ncbi.nlm.nih.gov/Class/Structure/aa/aa_explorer.cgi Select the Biochemical Properties table. Which amino acid is most abundant? (Is it leucine, at 9.94%?). Use this table to test yourself and make sure you know the one- and three-letter abbreviations for all 20 amino acids, as well as their structures.

ii) Is tyrosine a hydrophobic amino acid? To decide, use the Common Substitutions table. Explore valine (a hydrophobic residue), sort the results by hydrophobicity, and see where tyrosine is located. You can also explore the Structure and Chemistry table.



4) Sequence Manipulation Suite

i) Many tools are available to manipulate sequences. Visit the Sequence Manipulation Suite (http://www.bioinformatics.org/sms2/index.html) to access a large number of tools. What is the reverse complement of the sequence GGAATTCC?



First, go to the website and select Reverse Complement. Then, type the above information. After submit, you will get result as picture below.

5) Perform pairwise alignment using EMBOSS tools via Galaxy and UCSC

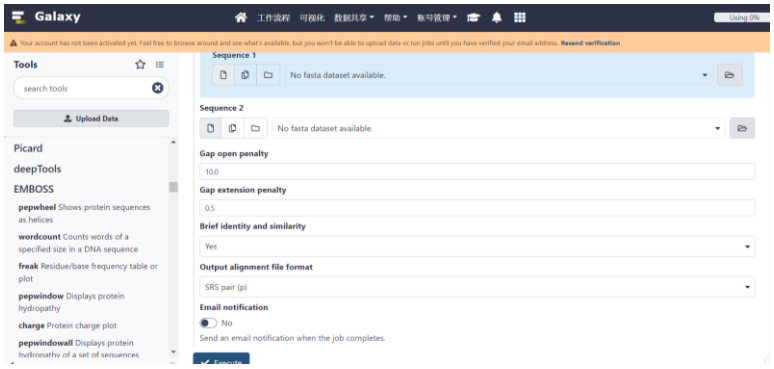
i) In this exercise we will perform global alignment with the EMBOSS package needle and local alignment with the EMBOSS package water.

ii) Both of these are available at the Galaxy public web server (along with over 100 other EMBOSS tools). Box 3.9 introduces EMBOSS and explains how to import beta globin (HBB) and alpha globin (HBA2) proteins from the UCSC Table Browser using Galaxy, and to then align them. This history is saved at https://main.g2.bx.psu.edu/u/pevsner/h/pairwise-alignment-via-ucsc-and-emboss (WebLink 3.14).

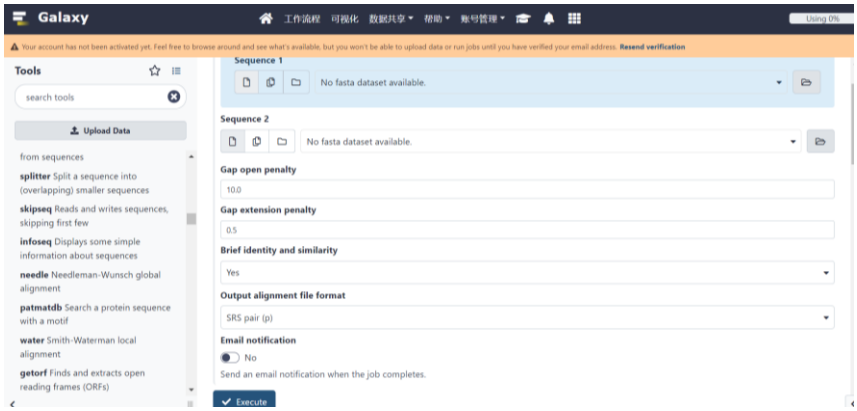
iii) Note that Galaxy is a web-based platform for using hundreds of bioinformatics tools, including next-generation sequence data analysis software.

iv) To use it visit http://usegalaxy.org then go to the public server.

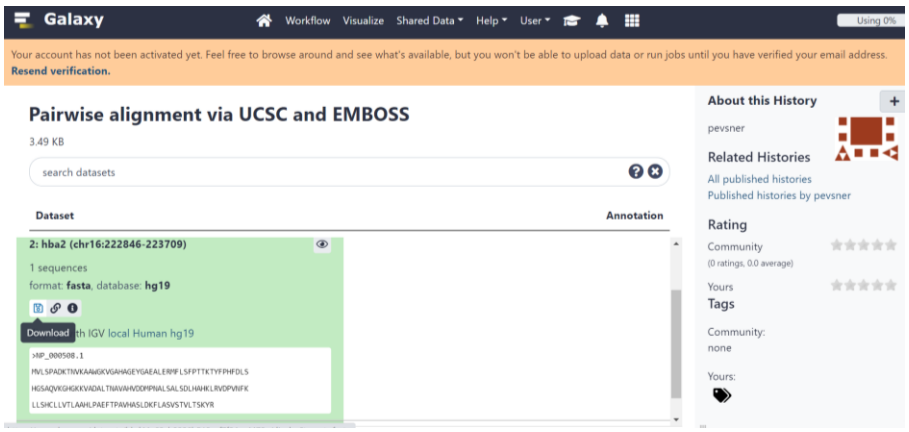
v) Be sure to create a username and log in. This will allow you to continue your work overtime and at different workstations.



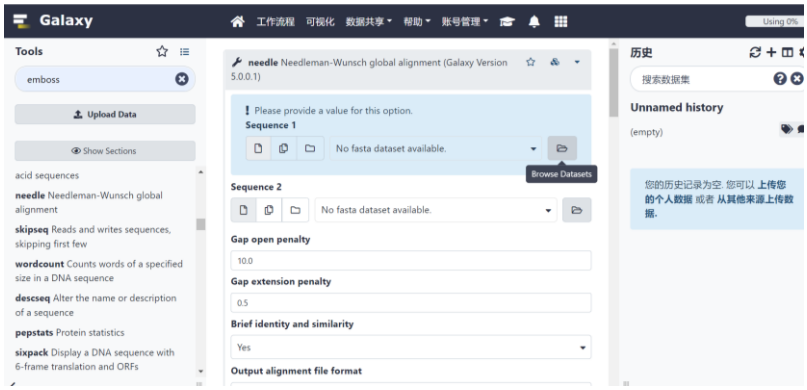
First, go to Galaxy website. At the left-hand side, scroll the scrollbar until you EMBOSS. Then, click on it.



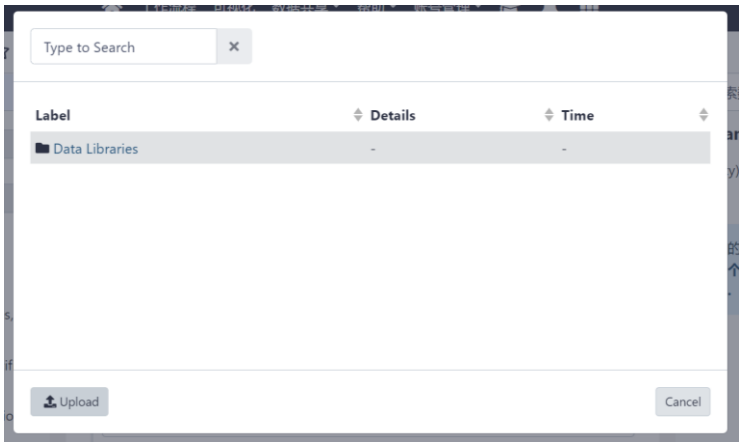
Then, scroll it until you see the needle and water. Click on either one. You will be able to calculate the global alignment by Needleman-Wunsch and local alignment by Smith-Water.



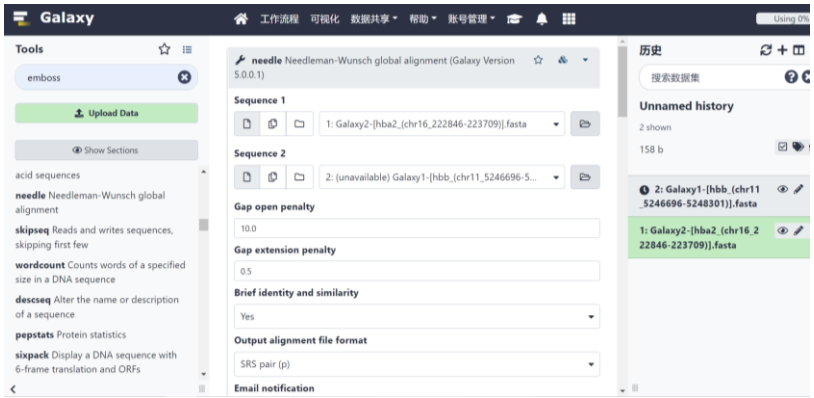
Go to this given link, https://main.g2.bx.psu.edu/u/pevsner/h/pairwise-alignment-via-ucsc-and-emboss. Download the file of the hba2 and hbb.



Then, go back to the Galaxy website and click on the browse datasets icon at the sequence 1 section.



Click on the upload button and upload the file of the hbb. Same step is repeated to upload the file of the hba2 at sequence 2 section.



After two files had been uploaded, scroll down to click on the execute button. Then, you will get the results.

Below is the example of the calculation of Needleman-Wunsch:

```
########################################
# Program: needle
# Rundate: Thu 21 Apr 2022 14:02:26
# Commandline: needle
```

```
#    -asequence /corral4/main/objects/d/3/1/dataset_d319941c-ec6a-41e4-
a112-12f955f600ea.dat
#    -bsequence /corral4/main/objects/0/6/d/dataset_06d70088-a9d7-4c4e-
9d09-d281ec6df3fc.dat
#    -outfile /corral4/main/objects/5/b/4/dataset_5b47c2df-d787-44c3-a3ee-
07e9bfe71137.dat
#    -gapopen 10.0
#    -gapextend 0.5
#    -brief yes
#    -aformat3 srspair
#    -auto
# Align_format: srspair
# Report_file: /corral4/main/objects/5/b/4/dataset_5b47c2df-d787-44c3-a3ee-
07e9bfe71137.dat
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: NP_000508.1
# 2: NP_000509.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 149
# Identity:      65/149 (43.6%)
# Similarity:    90/149 (60.4%)
# Gaps:           9/149 ( 6.0%)
# Score: 292.5
#
#
#=======================================

NP_000508.1        1 MV-LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-D
48
                     || |:|.:|:.|.|.||||  :..|.|.|||.|:.:.:|.|:.:|..| |
NP_000509.1        1 MVHLTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGD
48

NP_000508.1       49 LS-----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLR
93
                     ||     .|:.:||.|||||..|.:::.:||:|::....:.||:||..||.
NP_000509.1       49 LSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLH
98

NP_000508.1       94 VDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR
142
                     |||.||:||.:.|:..||.|...||||.|.|:..|.:|.|:..|..||.
NP_000509.1       99 VDPENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH
147


#---------------------------------------
#---------------------------------------
```

Below is the example of the calculations of the Smith-Water:

```
########################################
# Program: water
# Rundate: Thu 21 Apr 2022 14:03:12
# Commandline: water
```

```
#    -asequence /corral4/main/objects/0/6/d/dataset_06d70088-a9d7-4c4e-
9d09-d281ec6df3fc.dat
#    -bsequence /corral4/main/objects/d/3/1/dataset_d319941c-ec6a-41e4-
a112-12f955f600ea.dat
#    -outfile /corral4/main/objects/f/5/5/dataset_f55d4a6a-e6ac-48ac-a005-
6b42b2534811.dat
#    -gapopen 10.0
#    -gapextend 0.5
#    -brief yes
#    -aformat3 srs
#    -auto
# Align_format: srs
# Report_file: /corral4/main/objects/f/5/5/dataset_f55d4a6a-e6ac-48ac-a005-
6b42b2534811.dat
########################################

#=======================================
#
# Aligned_sequences: 2
# 1: NP_000509.1
# 2: NP_000508.1
# Matrix: EBLOSUM62
# Gap_penalty: 10.0
# Extend_penalty: 0.5
#
# Length: 145
# Identity:      63/145 (43.4%)
# Similarity:    88/145 (60.7%)
# Gaps:           8/145 ( 5.5%)
# Score: 293.5
#
#
#=======================================

NP_000509.1         4 LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLST
51
NP_000508.1         3 LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF-DLS-
50

NP_000509.1        52 PDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDP
101
NP_000508.1        51 ----HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDP
96

NP_000509.1       102 ENFRLLGNVLVCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKY    146
NP_000508.1        97 VNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKY    141


#---------------------------------------
#---------------------------------------
```