SUBSTITUTION MODELS AND MEGA (COOL) PHYLOGENETICS

AARHUS UNIVERSITY
DEPARTMENT OF MOLECULAR BIOLOGY AND GENETICS

TA
29 AUGUST 2023

BJARKE MEYER PEDERSEN
PH.D STUDENT

DISTANCES BETWEEN SEQUENCES

AGATGCTAGCATCGACTAGCATCAGCTGACCCCGCGCGCGCAT

AGATAATAGCATCGACTAGCATCAGCGGACCCCGCGCGCGCAC

How far are these a part?

P-DISTANCE

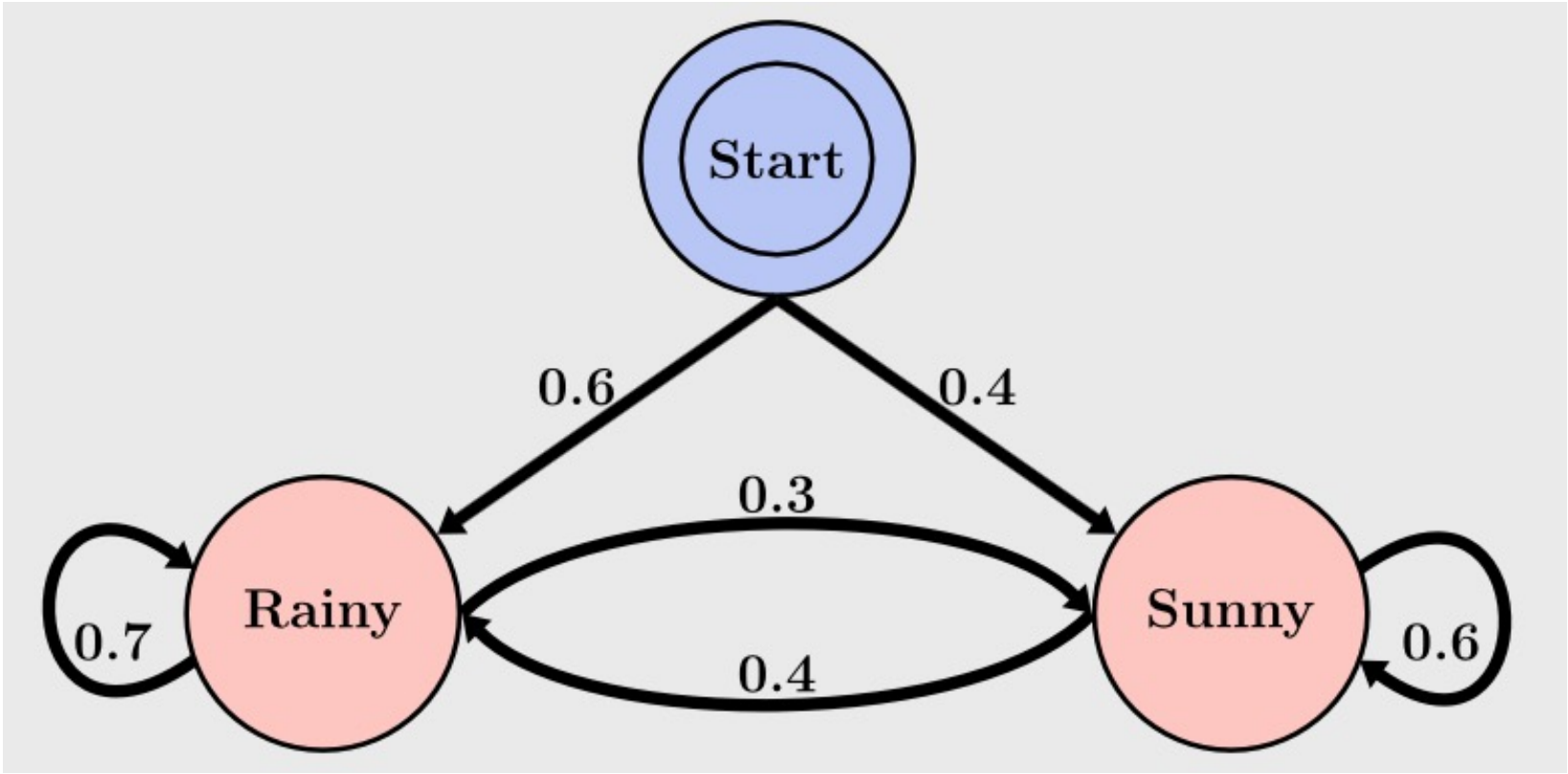
The proportion of differences between the two sequences

Differences / Total number of nucleotides compared

The sequences before were 43 long, and there were 4 differences

What would be their p-distance?
When is this appropriate? When is it not?

CONTINUOUS-TIME MARKOV CHAINS



BACK TO THE DISTANCES AND SEQUENCES

AGATGCTAGCATCGACTAGCATCAGCTGACCCCCGCGCGCGCAT

AGATAATAGCATCGACTAGCATCAGCGGACCCCCGCGCGCGCAC

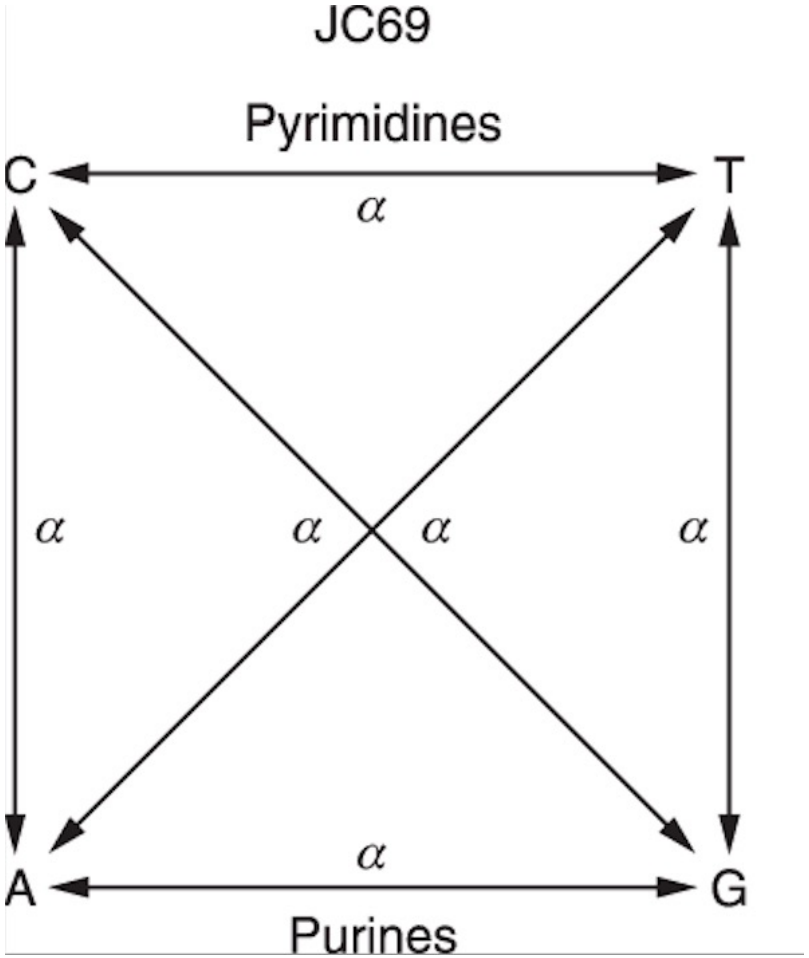
We want to quantify the unobserved, which is difficult.

We can do so by treating this as a Markov chain, but it will require us to assume somethings:

- Each site in the sequence are independent

Not always true but we will assume more unbelievable stuff later so never mind.

JUKES CANTOR (69)



$$P(t) = \{p_{ij}(t)\} \approx I + Qt = \begin{bmatrix} 1-3\lambda t & \lambda t & \lambda t & \lambda t \\ \lambda t & 1-3\lambda t & \lambda t & \lambda t \\ \lambda t & \lambda t & 1-3\lambda t & \lambda t \\ \lambda t & \lambda t & \lambda t & 1-3\lambda t \end{bmatrix}$$

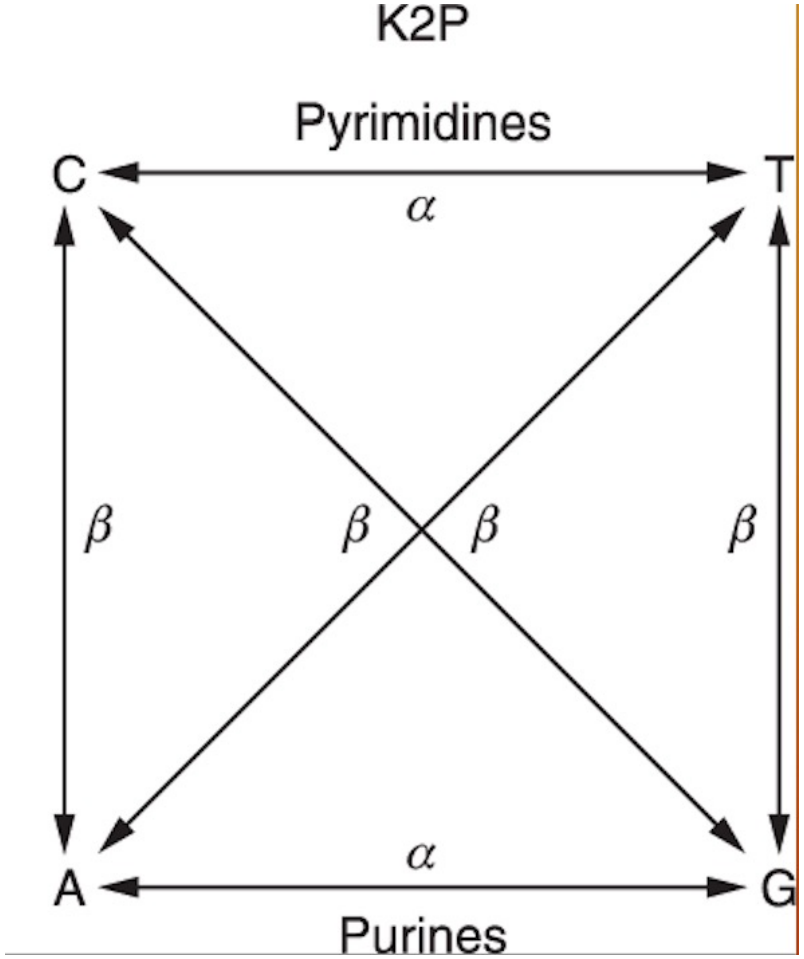
$$P(t) = e^{Qt} = \begin{bmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{bmatrix}, \text{ with } \begin{cases} p_0(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}, \\ p_1(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}. \end{cases} \quad (1.4)$$

$$\hat{d} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\hat{p}\right), \quad (1.7)$$

What would the distance be between the two seqs now?

N=43
Differences=4

KIMURA (80) AKA K80 AKA AKA K2P



$$Q = \begin{bmatrix} -(\alpha+2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha+2\beta) & \beta & \beta \\ \beta & \beta & -(\alpha+2\beta) & \alpha \\ \beta & \beta & \alpha & -(\alpha+2\beta) \end{bmatrix}.$$

$$P(t) = e^{Qt} = \begin{bmatrix} p_0(t) & p_1(t) & p_2(t) & p_2(t) \\ p_1(t) & p_0(t) & p_2(t) & p_2(t) \\ p_2(t) & p_2(t) & p_0(t) & p_1(t) \\ p_2(t) & p_2(t) & p_1(t) & p_0(t) \end{bmatrix}, \quad (1.10)$$

where the three distinct elements of the matrix are

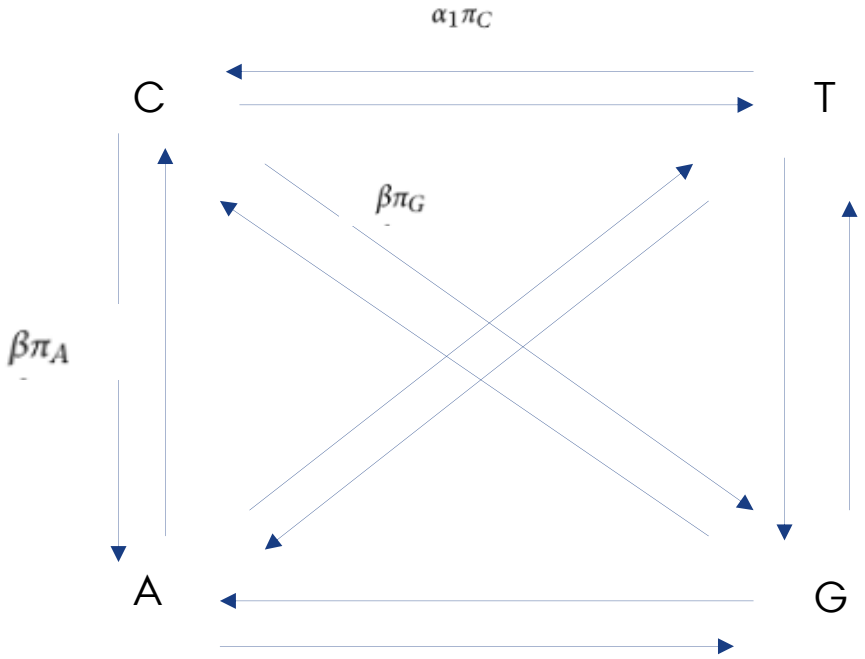
$$\begin{aligned} p_0(t) &= \tfrac{1}{4} + \tfrac{1}{4}e^{-4\beta t} + \tfrac{1}{2}e^{-2(\alpha+\beta)t} = \tfrac{1}{4} + \tfrac{1}{4}e^{-4d/(\kappa+2)} + \tfrac{1}{2}e^{-2d(\kappa+1)/(\kappa+2)}, \\ p_1(t) &= \tfrac{1}{4} + \tfrac{1}{4}e^{-4\beta t} - \tfrac{1}{2}e^{-2(\alpha+\beta)t} = \tfrac{1}{4} + \tfrac{1}{4}e^{-4d/(\kappa+2)} - \tfrac{1}{2}e^{-2d(\kappa+1)/(\kappa+2)}, \\ p_2(t) &= \tfrac{1}{4} - \tfrac{1}{4}e^{-4\beta t} = \tfrac{1}{4} - \tfrac{1}{4}e^{-4d/(\kappa+2)} \end{aligned} \quad (1.11)$$

$$\begin{aligned} \hat{d} &= -\tfrac{1}{2}\log(1-2S-V) - \tfrac{1}{4}\log(1-2V), \\ \hat{\kappa} &= \frac{2\times\log(1-2S-V)}{\log(1-2V)} - 1 \end{aligned} \quad (1.12)$$

What would the distance be between the two seqs now?

N=43
Differences=4 (S=2, V=2)
S = transitions
V= transvertions

TN93



$$Q = \begin{bmatrix} -(\alpha_1\pi_C + \beta\pi_R) & \alpha_1\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha_1\pi_T & -(\alpha_1\pi_T + \beta\pi_R) & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & -(\alpha_2\pi_G + \beta\pi_Y) & \alpha_2\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha_2\pi_A & -(\alpha_2\pi_A + \beta\pi_Y) \end{bmatrix}. \quad (1.16)$$

$$\hat{d} = \frac{2\pi_T\pi_C}{\pi_Y}(a_1 - \pi_R b) + \frac{2\pi_A\pi_G}{\pi_R}(a_2 - \pi_Y b) + 2\pi_Y\pi_R b,$$

What would the distance be between the two seqs now?

If you want you can do this one at home..

TN93, THE COMPLICATED CASE

Talk to your group for 2 minutes:

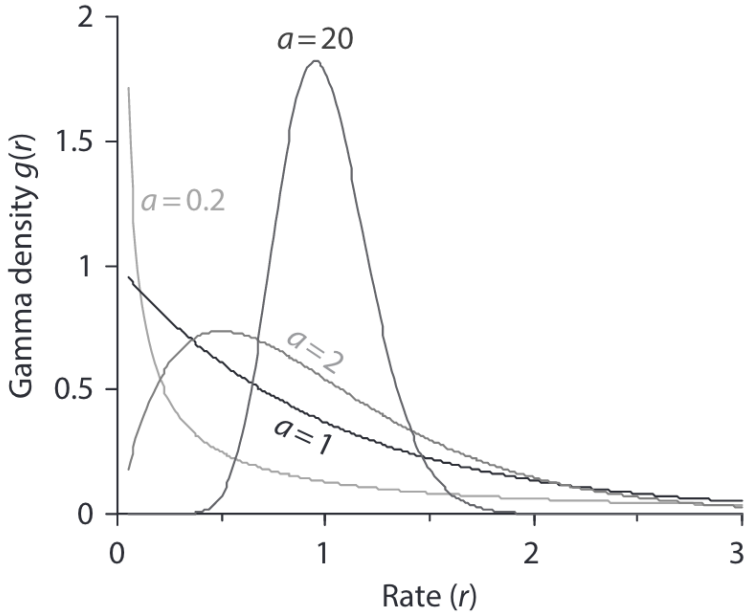
What would we have to change to make TN93 -> JC69?

What would we have to change to make TN93 -> K2P

(Fun) fact this is a nested model, which they'll cover in Datascience at some point.

EXTENTIONS OF THE MODELS

Not all sites are equally likely to change, we can adjust for this using a gamma distribution.



MAXIMUM LIKELIHOOD ESTIMATION (MLE)

Estimate the parameters for your model by picking the parameters that maximize the likelihood of seeing the data

It can get awefully complicated, but it is indeed very handy

EXERCISES

Now you have ~45 minutes to make the exercises for today

Ask questions if you have any

EVALUATION OF THE EXERCISES

The Sequence of the Human Genome

J. CRAIG VENTER, MARK D. ADAMS, EUGENE W. MYERS, PETER W. LI, RICHARD J. MURAL, GRANGER G. SUTTON, HAMILTON O. SMITH, MARK YANDELL, CHERYL A. EVANS, [...], AND XIAOHONG ZHU +264 authors Authors Info & Affiliations

SCIENCE · 16 Feb 2001 · Vol 291, Issue 5507 · pp. 1304-1351 · DOI: 10.1126/science.1058040

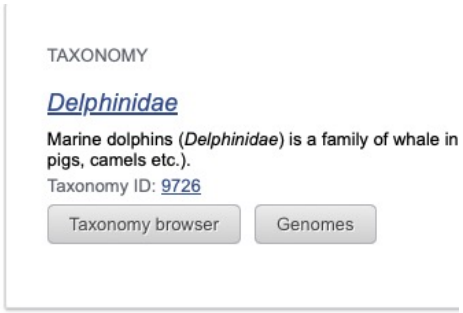
1. Understanding PubMed and GenBank

GenBank is a useful database that contains DNA, RNA and protein sequences publicly available that we will access through MEGAX to download sequences of interest. You can read more about it here.

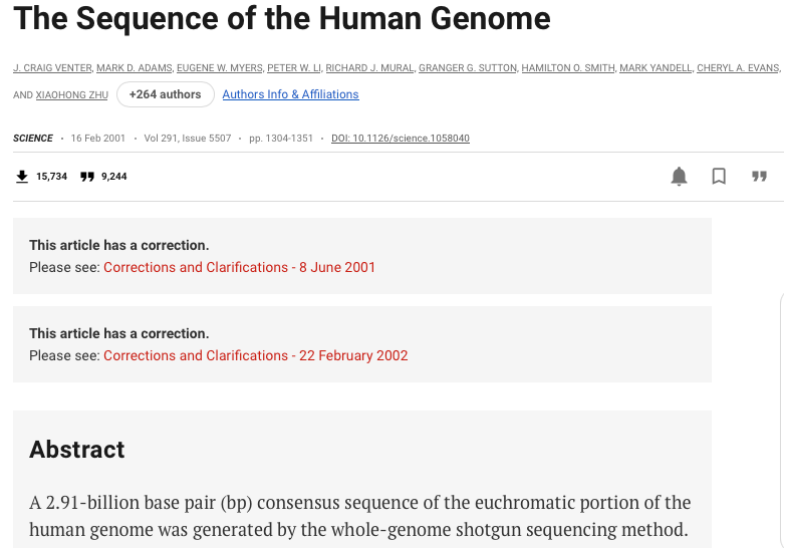
In order to use GenBank efficiently, as other databases such as PubMed used to search for papers, it is a good idea to use specific search fields. These can be specified as e.g. [Author], [pdat] (publication time) and [Title] and use logical operators to combine search terms, e.g. AND, OR, NOT. For example, try to search "Zhao[Author] AND Wu F[Author] AND Yu[Author] AND coronavirus[Title] AND 2020[pdat]" on PubMed. Which paper comes up?

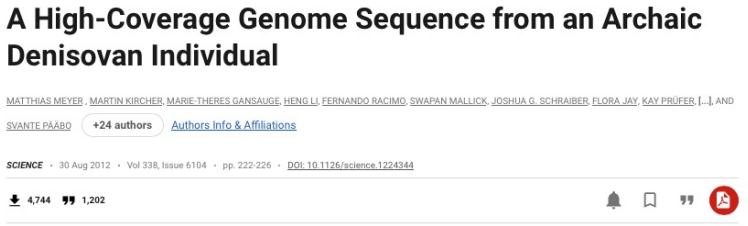
By searching papers in PubMed and GenBank (and Google for sure), answer the following questions:

1. Find and download the paper of the first sequence of the human genome by the International Human Genome Sequencing Consortium or the one assembled by Craig Venter et al in 2001. Where was it published? How can the sequence be obtained?
2. Find the paper on the high coverage archaic Denisovan sequence published by Svante Pääbo's group published in 2012 and answer the same questions. Who were the Denisovans?
3. How many sequences from the dolphin (you might want to use the latin name *Delphinidae*) can you find? You can choose any other animal!
4. Find the Taxonomic position of Dolphins and find the number of DNA sequences deposited in GenBank for some of the dolphin species
5. Find sequences from FOXP2. What can you learn about this protein?



Items: 1 to 20 of 557525





Taxonomy Browser

Selected taxa

Delphinidae (marine dolphins) Enter one or more taxonomic names

Taxonomic name	Genomes
Eukaryota (eukaryotes)	34,056
Metazoa (animals)	12,392
Chordata (chordates)	6,426
Mammalia (mammals)	2,820
Artiodactyla (even-toed ungulates)	415
Delphinidae (marine dolphins)	26
Cephalorhynchus	0
Delphinus	3
Feresa	0
Grampus	2
Lagenodelphis	0
Lagenorhynchus	3
Lissodelphis	0
Orcaella	0
Orcinus	3
Peponocephala	0
Globicephala (pilot whales)	1
Pseudorca	0
Sotalia	0
Sousa	2
Stenella	2
Steno	1
Tursiops	9

Example 2.1:

Launch the *Alignment Explorer* by selecting the ***Align | Edit/Build Alignment*** on the launch bar of the main *MEGA* window.

Select ***Create New Alignment*** and click ***Ok***. A dialog will appear asking "Are you building a DNA or Protein sequence alignment?" Click the button labeled "**DNA**".

From the *Alignment Explorer* main menu, select ***Data | Open | Retrieve sequences from File***. Select the "hsp20.fas" file from the MEG/Examples directory.

Aligning Sequences by *ClustalW*

You can create a multiple sequence alignment in *MEGA* using either the *ClustalW* or Muscle algorithms. Here we align a set of sequences using the *ClustalW* option.

Example 2.2:

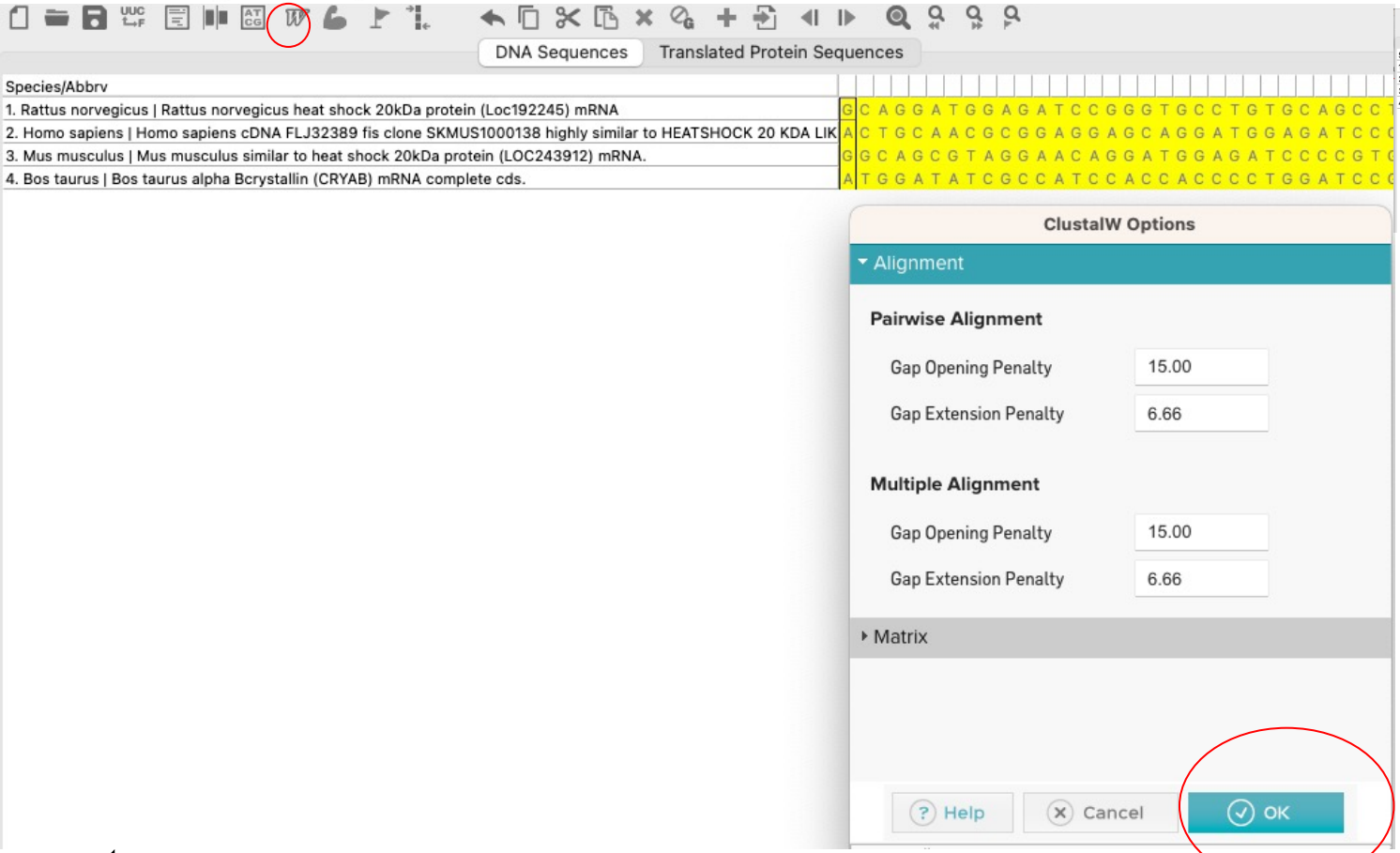
Open the alignment file (using the instructions above) hsp20.fas.

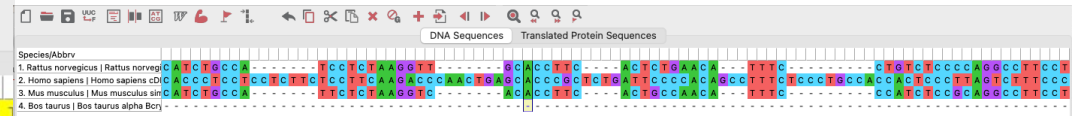
Select the ***Edit | Select All*** menu command to select all sites for every sequence in the data set.

Select ***Alignment | Align by ClustalW*** from the main menu to align the selected sequences data using the *ClustalW* algorithm. Click the "***Ok***" button to accept the default settings for *ClustalW*.

Once the alignment is complete, save the current *alignment session* by selecting **Data | *Save Session*** from the main menu. Give the file an appropriate name, such as "hsp20_Test.mas". This will allow the current *alignment session* to be restored for future editing.

Exit the *Alignment Explorer* by selecting ***Data | Exit AIn Explorer*** from the main menu.





Aligning Sequences Using Muscle

Here we describe how to create a multiple sequence alignment using the Muscle option.

Example 2.3:

Starting from the main *MEGA* window, select ***Align | Edit/Build Alignment*** from the launch bar. Select ***Create a new alignment*** and then select ***DNA***.

From the *Alignment Explorer* window, select ***Data | Open | Retrieve sequences from a file*** and select the "Chloroplast_Martin.meg" file from the MEGA/Examples directory.

On the *Alignment Explorer* main menu, select ***Edit | Select All***.

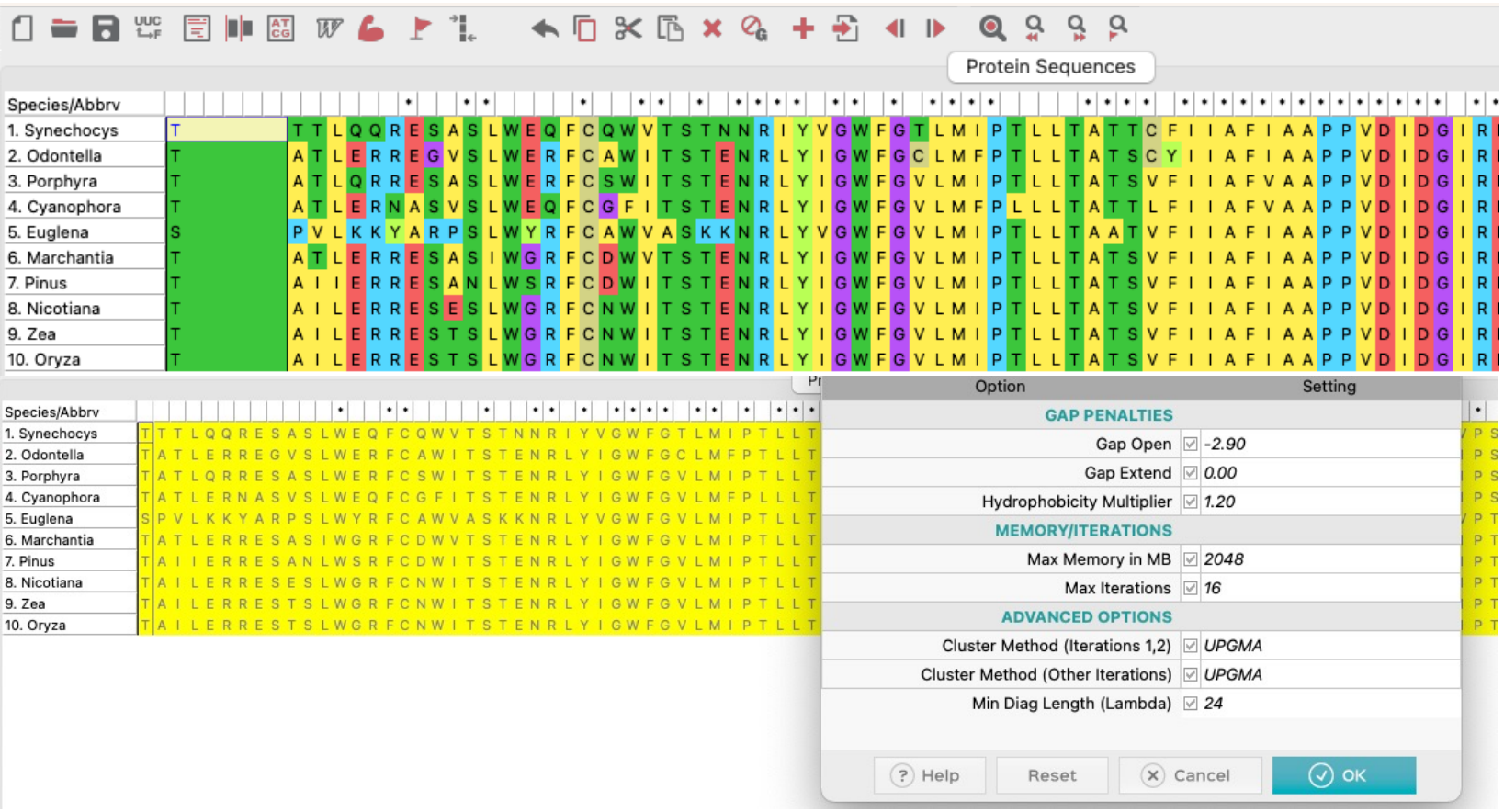
On the *Alignment Explorer* launch bar, you will find an icon that looks like a flexing arm. Click on it and select ***Align DNA***.

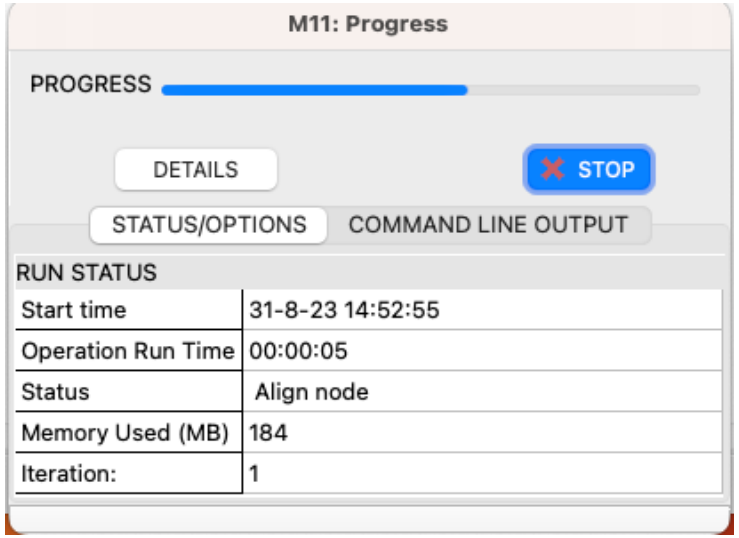
Near the bottom of the *MUSCLE - AppLink* window, you will see a row called *Alignment Info*. You can read information about the Muscle program.

Click on the ***Compute*** button (accept the default settings). A *Progress* window will keep you informed of Muscle alignment status. In this window, you can click on the *Command Line Output* tab to see the command-line parameters which were passed to the Muscle program. Note: The analysis may complete so fast, that you won't be able to click on this tab or read it. The information in this tab isn't essential, it's just interesting.

When the Muscle program has finished, the aligned sequences will be passed back to *MEGA* and displayed in the *Alignment Explorer* window.

Close the *Alignment Explorer* by selecting ***Data | Exit AIn Explorer***. Select ***No*** when asked if you would like to save the current *alignment session* to file.





Obtaining Sequence Data from the Internet (GenBank)

Using *MEGA's* integrated web browser you can fetch GenBank sequence data from the *NCBI* website if you have an active internet connection.

Example 2.4:

From the main *MEGA* window, select ***Align | Edit/Build Alignment*** from the main menu.

When prompted, select ***Create New Alignment*** and click ***ok***. Select ***DNA***

Activate *MEGA's* integrated browser by selecting ***Web | Query Genbank*** from the main menu.

When the *NCBI: Nucleotide* site is loaded, enter **CFS** as a search term into the search box at the top of the screen. Press the ***Search*** button.

When the search results are displayed, check the box next to any item(s) you wish to import into *MEGA*.

If you have checked more than one box: locate the ***Display Settings*** dropdown (located near the top left hand side of the page directly under the tab headings). Change the value to ***FASTA (Text)*** and click the ***Apply*** button. This will output all the sequences you selected as a text in the FASTA format.

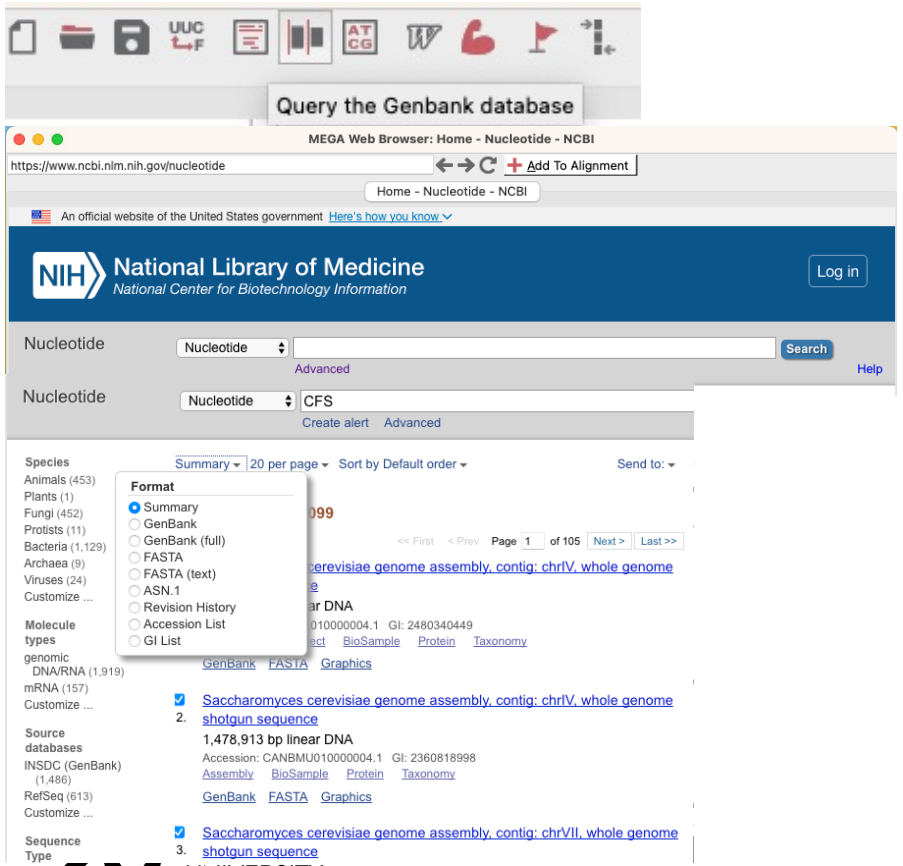
Press the ***Add to Alignment*** button (with the red **+** sign) located above the web address bar. This will import the sequences into the *Alignment Explorer*.

With the data now displayed in the *Alignment Explorer*, you can close the *Web Browser* window.

Align the new data using the steps detailed in the previous examples.

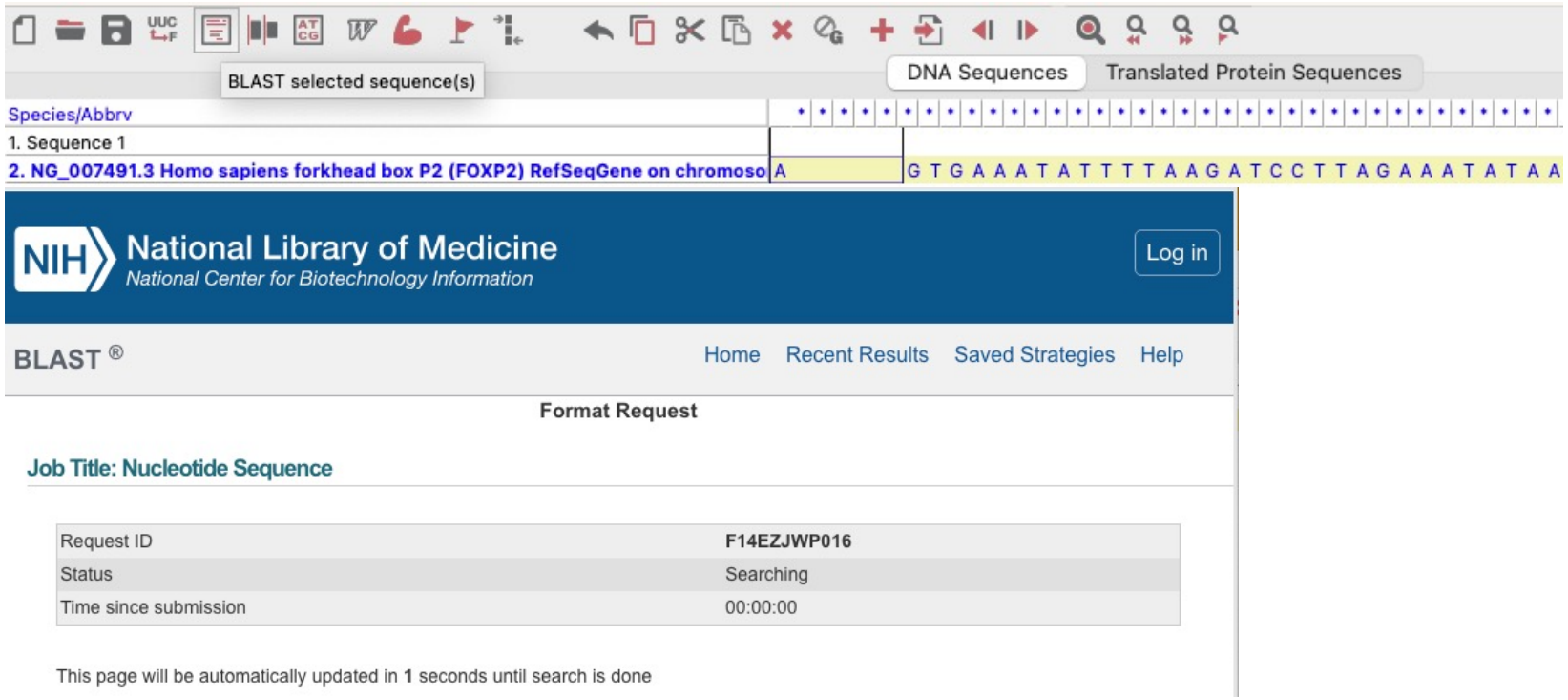
Close the *Alignment Explorer* window by clicking ***Data | Exit Aln Explorer***. Select ***No*** when asked if you would like the save the current *alignment session* to file.

Note: We have aligned some sequences and they are now ready to be analyzed. Whenever you need to edit/change your sequence data, you will need to open it in the *Alignment Editor* and edit or align it there. Then export it to the *MEGA* format and open the resulting file.



BLAST

BLAST (basic local alignment search tool) is an algorithm and program for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA and/or RNA sequences. A BLAST search enables a researcher to compare a subject protein or nucleotide sequence (called a query) with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold. Wikipedia.



EVALUATION

JOIN THE BARC FACEBOOK TODAY

AARHUS
UNIVERSITY