

Assignment 2

Topics: DNA sequencing data handling and evolutionary analysis

Due date: 17 September 2025 at 11:59pm

Rules:

- You can work in group, but write your own answers
- You can use AI to help, but don't abuse it. Credit AI when used
- The objective of the assignment is to provide you with experience. Explain your work and observations. Don't just paste a screenshot of the result.
- You can contact me to ask for clarification

Credit: GPT-5 was used to aid the design of the assignment

Caution: Analyses in this assignment can take >30 minutes

Part A. Hands-On Mini-Pipeline (Galaxy Platform)

We will use the **Galaxy server** (<https://usegalaxy.eu> or <https://usegalaxy.org>), which allows running sequencing pipelines through a graphical web interface.

1. Go to <https://usegalaxy.eu> or <https://usegalaxy.org>. Create a free account.
2. Download DNA sequencing data for yeast *S. cerevisiae* from **Sequence Read Archive** accession number **SRR15616231** (900 Mb)

Sequence Read Archive Search Run Browser Analyses Study Provisio

Run Browser > SRR15616231

DNA-seq of *Saccharomyces cerevisiae* (SRR15616231)

Metadata Analysis Reads Data access FASTA/FASTQ download

Download for Experiment SRX11913458

<input type="checkbox"/> Accession	Total Bases	Spots	
		Total	Filtered
<input type="checkbox"/> SRR15616231	2.4Gbases	8.0M	

Filter Runs

Search by sub-sequence,



[What can the filter be applic](#)

3. Upload the data onto your **Galaxy** account. Select **S. cerevisiae Apr 2011** as the associated genome.


Download from web or upload from disk


Regular Composite Collection Rule-based


You added 1 file(s) to the queue. Add more files or click 'Start' to proceed.


Name	Size	Type	Genome	Settings	Star
 SRR15616231.fastq.c	873.7 MB	Auto-de...	unspecified (?)		0%
<div><div>cerevis</div><div><div><div>. cerevisiae Apr. 2011 (SacCer_Apr2011/sacCer3)</div><div>. cerevisiae June 2008 (SGD/sacCer2)</div><div>. cerevisiae Oct. 2003 (SGD/sacCer1)</div></div></div></div>					


4. Perform quality check using **FastQC**. You can search for the tool on Galaxy. The tool will automatically select your FASTQ file as the input. Click **Execute** without changing any parameter.

 **Galaxy Pasteur**

Tools 

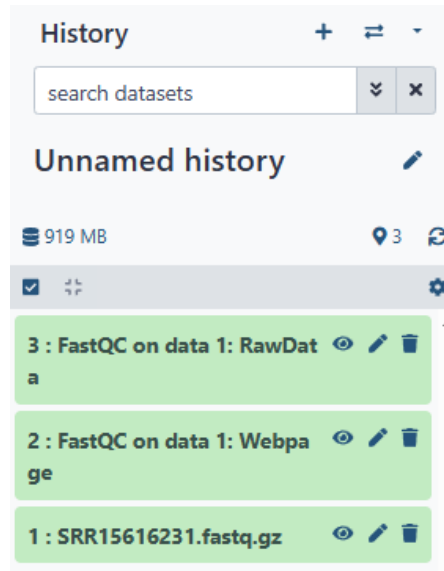
fastqc 

 Upload Data

 Show Sections

FastQC Read Quality reports

5. Progress of your analysis will be shown in the **History** bar. Once an analysis is done, it will turn green. Here, look for **FastQC on data 1: Webpage**. Click on it and save the **FastQC result in .html format**. Open the .html file on your computer and summarize the quality check result.
 - How many reads are present, and how long?
 - What is the GC%? Does it match genomic GC% for *S. cerevisiae*?
 - Is there any issue with base quality, duplication, or adapter sequence?



6. Perform read trimming using **Trimmomatic**. Select your FASTQ file as the input.
7. Perform sequence alignment using **Bowtie2**. Select the output of **Trimmomatic** as the input. Select **scerevisiae** as the reference genome. A **BAM** file should be produced.
8. *[Trick: You can queue up the analysis while **Bowtie2** is running]* Perform variant calling using **FreeBayes**. Select the output of **Bowtie2** as the input. Select **scerevisiae** as the reference genome. Save the resulting **VCF** file to your computer and inspect it.
 - Filter variants based on QUAL of 30 or above. How many variants passed?
 - Examine the **FORMAT** column. What are the details reported in this VCF?

Part B. Literature Analysis (GWAS/QTL/Phylogenetics)

Study this landmark GWAS study of age-related macular degeneration, and eye disease, from 2005, <https://pmc.ncbi.nlm.nih.gov/articles/PMC1512523/>, and answer the following questions:

- What was the biological question?
- How many samples and SNPs were considered?
- How were the SNPs filtered?
- Was correction for multiple testing performed? Which techniques?
- How was the GWAS result narrowed down to pinpoint key genes?

Part C. Experimental Design utilizing Genotype-Phenotype analysis

Suppose you are studying **drug resistance in a yeast population (1,000 strains)**. For each strain, you recorded the highest concentration of the drug at which the yeast survived. From these results, you want to identify genetic variants responsible for the resistance.

Propose an experimental design and computational analysis using techniques we have learned so far in this course to pinpoint candidate variants as specific as possible.

- What samples would you sequence?
 - Will you sequence the whole genome, or only some genomic elements?
 - Describe the sequencing data processing steps
 - What analysis method(s) would you apply?
 - How would you test that the identified variants are associated with resistance?
-

Part D. Critiquing LLM Responses

Do this after finishing Part C.

1. Feed the question from Part D into an LLM/AI of your choice:

"Design a DNA sequencing-based experiment and computational analysis to identify genetic variants associated with antifungal drug resistance in yeast. For each strain, the highest concentration of the drug at which the yeast survived was recorded."

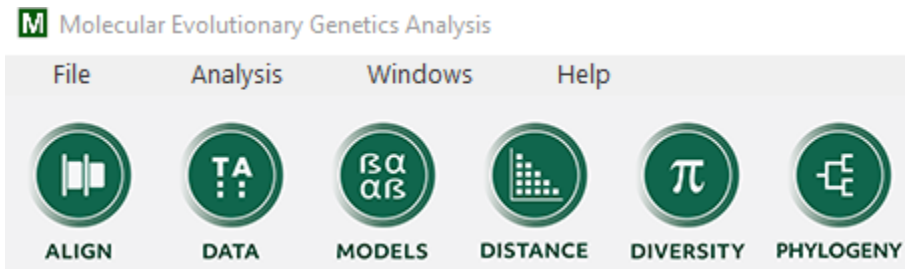
2. What was the LLM's response?
 3. Critique the response:
 - Which parts of the response are scientifically sound and useful?
 - Which parts of the response are vague, inaccurate, or incomplete?
 - Compare your own design from **Part C** with LLM's response.
 - Reflect: What risks are there in relying solely on LLMs for scientific planning?
-

Part E. Phylogenetics reconstruction

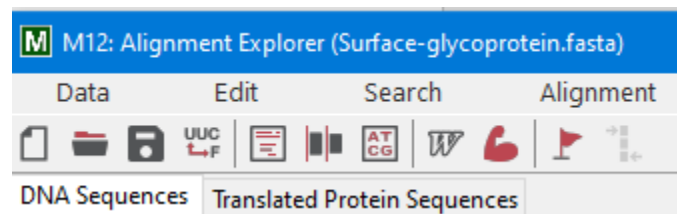
We will use MEGA (<https://www.megasoftware.net/>) which is a graphical interface for phylogenetics analysis. Download and install version X, 11, or 12.

The FASTA file containing nucleotide coding sequences of selected surface glycoproteins is provided at <https://github.com/cmb-chula/comp-biol-3000788/blob/main/data/Surface-glycoprotein.fasta>.

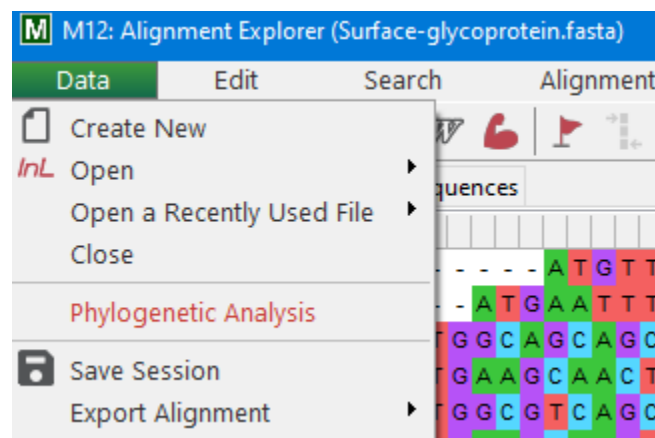
Here is a quick guide to MEGA's GUI:



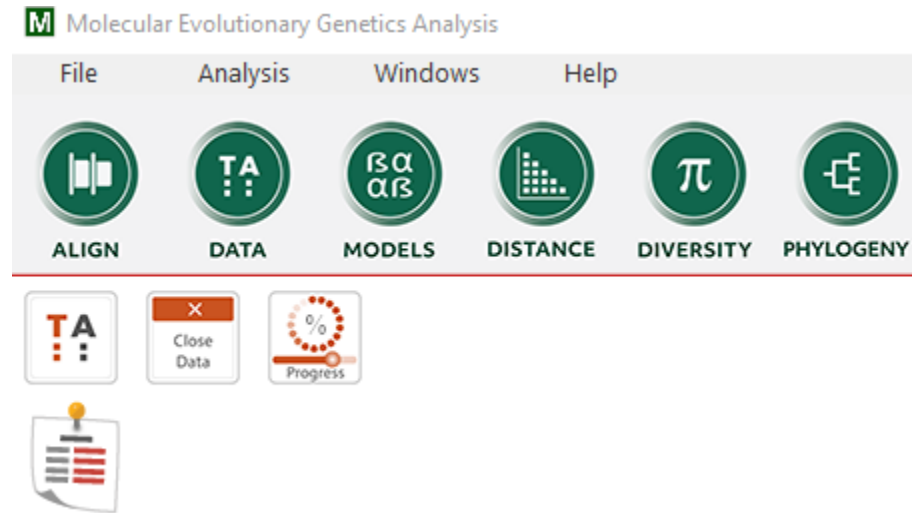
New sequence data can be imported through **DATA** module. Multiple sequence alignment can be performed within that module via the **Alignment** menu.



To make the alignment result available to other modules, you need to save the result via the **Data** menu and load them in other modules. Alternatively, you can select **Phylogenetic Analysis** from the **Data** menu. This will save and load the alignment result for you into the main GUI.



You will be able to tell that the results are available to other modules within MEGA if additional icons show up in the main GUI.



Then, you can use the **MODELS** module to test multiple nucleotide/protein substitution models or the **PHYLOGENY** module to build a phylogenetic tree.

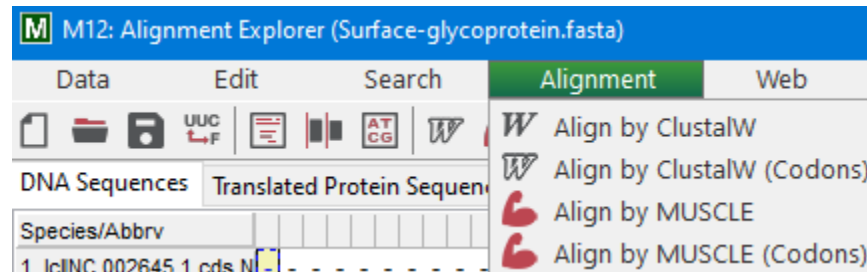
For the sake of simplicity, **we will not adjust any parameters in MEGA analysis unless instructed to do so.**

1. Inspect the header of the first sequence

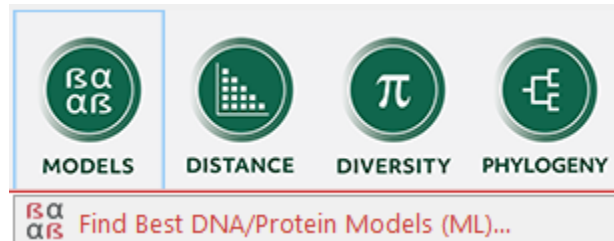
```
>Ic|NC_002645.1_cds_NP_073551.1_1 [gene=S] [locus_tag=HCoV229Egp2]
[db_xref=GeneID:918758] [protein=surface glycoprotein] [protein_id=NP_073551.1]
[location=20570..24091] [gbkey=CDS]
```

What does each piece of information inside [...] mean?

2. Given that these are nucleotide coding sequences, **for the purpose of phylogenetic reconstruction**, what type of multiple sequence alignment would you perform, **nucleotide**, **codon**, or **protein**?
3. Perform the selected alignment in MEGA with **MUSCLE algorithm** and show the screenshot of the result.



4. To perform phylogenetic reconstruction, we need to select an appropriate nucleotide substitution model using a procedure called **nested model testing**. This is done by trying out multiple substitution models, recording the likelihoods, and selecting the most significant model. Do this using the **Find Best DNA/Protein Models (ML)** function in **MODELS** module in MEGA.



This process takes a lot of computing power, and will produce a result like this:

M Results

File Edit View Windows Help

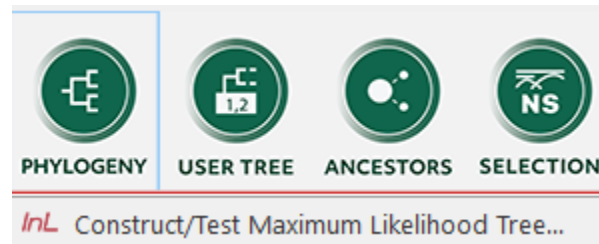
TXT [Icons]

Table. Maximum Likelihood analysis of substitution models

Model	Parameters	BIC	AICc	<i>lnL</i>	(+I)	(+G)	<i>R</i>	<i>f(A)</i>
GTR+G+I	69	192257.965	191599.634	-95730.770	0.01	1.82	1.03	0.314
HKY+G+I	65	192595.824	191975.652	-95922.784	0.01	1.83	1.04	0.314
TN93+G+I	66	192603.417	191973.706	-95920.810	0.01	1.83	1.04	0.314
T92+G+I	63	192702.694	192101.602	-95987.762	0.01	1.83	1.05	0.304
GTR+I	68	194243.826	193595.035	-96729.472	0.01	n/a	0.90	0.314
K2+G+I	62	194437.512	193845.960	-96860.942	0.01	2.06	0.96	0.250
HKY+I	64	194526.294	193915.662	-96893.791	0.01	n/a	0.90	0.314
TN93+I	65	194537.314	193917.143	-96893.530	0.01	n/a	0.90	0.314
T92+I	62	194642.804	194051.252	-96963.588	0.01	n/a	0.91	0.304
JC+G+I	61	195340.901	194758.888	-97318.407	0.01	2.25	0.50	0.250
K2+I	61	196122.428	195540.416	-97709.171	0.01	n/a	0.90	0.250

5. Based on the results from #4, explain what BIC, AICc, and *lnL* mean. How would you select the best substitution model based on these scores? Do they suggest the same substitution model?
6. Let's build a phylogenetic tree using the model **GTR+G+I** with MEGA using the **Maximum Likelihood** method from **PHYLOGENY** module. Adjust the following parameters:
 - a. **Test of Phylogeny:** Standard Bootstrap with 100 Replicates (if your computer is slow, you may reduce the number of Replicates to 20)
 - b. **Model/Method:** According to **GTR+G+I**

c. **Rates among Sites:** According to **GTR+G+I**



M12: Analysis Preferences	
Phylogeny Reconstruction	
Option	Setting
ANALYSIS	
Statistical Method →	<i>Maximum Likelihood</i>
PHYLOGENY TEST	
Test of Phylogeny →	<i>None</i>
SUBSTITUTION MODEL	
Substitutions Type →	<i>Nucleotide</i>
Model/Method →	<i>Tamura-Nei model</i>
RATES AND PATTERNS	
Rates among Sites →	<i>Uniform Rates</i>
DATA SUBSET TO USE	
Gaps/Missing Data →	<i>Use all sites</i>
Select Codon Positions →	<input checked="" type="checkbox"/> 1st <input checked="" type="checkbox"/> 2nd <input checked="" type="checkbox"/> 3rd <input checked="" type="checkbox"/> Noncoding Sites
TREE INFERENCE OPTIONS	
ML Heuristic Method →	<i>Nearest-Neighbor-Interchange (NNI)</i>
Initial Tree for ML →	<i>Make initial tree automatically (Default - NJ/MP)</i>
Branch Swap Filter →	<i>None</i>
SYSTEM RESOURCE USAGE	
Number of Threads →	<i>4</i>

- Discuss the resulting phylogenetic tree. Are you confident in the relationship among the genes as suggested by the tree's topology?