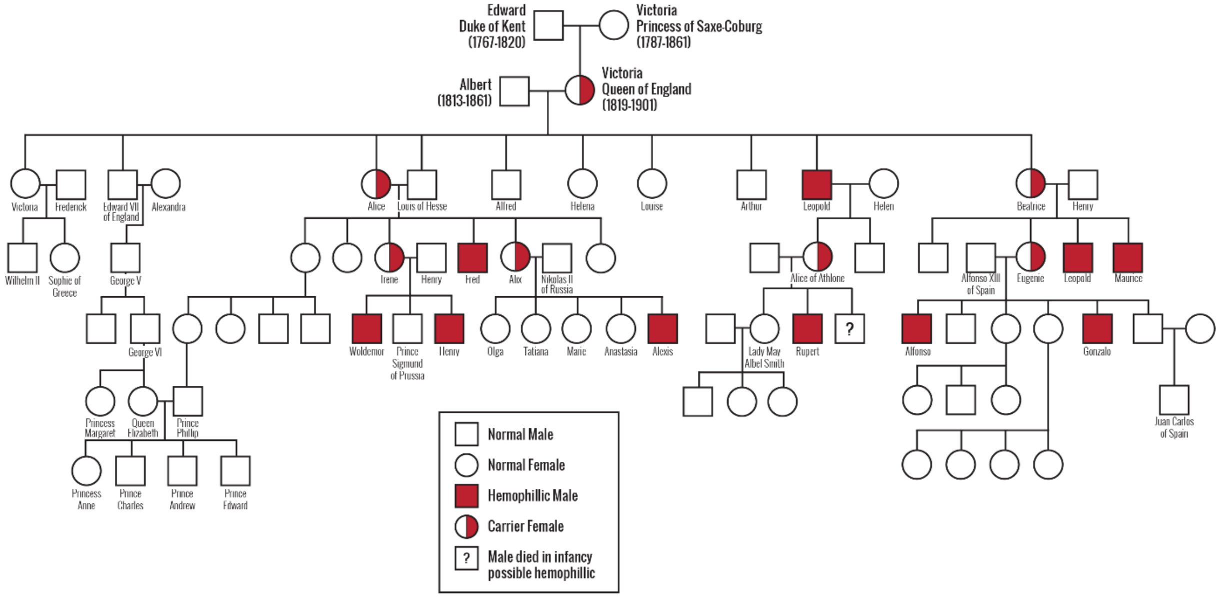
Running Exercise III (2025)

In 1894, Nicholas II Romanov married Alexandra, daughter of Princess Alice, daughter of Queen Victoria. The couple had five children, as shown in the family tree below. Their son Alexei had Hemophilia. This is the known family tree.



Source: "Hemophilia" by Aronova-Tiuntseva and Herreid

In January 1918, the Romanov family was executed during the uprising of the Bolshevik revolution. There are several accounts of what happened, but it is believed that an angry mob shot all family members, burnt the bodies, and disposed of them in the forest. Some ten years later, the bodies were discovered by an amateur archaeologist. It quickly became apparent that one body that of Anastasia may be missing.

Anastasia was the youngest daughter of Tsar Nicholas II. She was delivered by Grigori Rasputin, a Russian peasant and a questionable doctor who spent much time drinking. He reported that Anastasia was born in 1901. Family portraits indicate that Anastasia had blue eyes and strawberry-blonde hair. Unfortunately, much of the DNA evidence was unclear, mainly due to the poor condition of the findings. By 1931, five women claimed that they were the real Anastasia. In 1932, DNA evidence (assume they existed back then) and testimonies were collected from most people involved.

Testimonies:

Anastasia1. 33 years old. She has blue eyes and orange-like hair. She possesses an ivory hairbrush with the name "Anastasia Romanov." Your expert historian confirmed the authenticity of the object. She fully complied with the genetic analysis.

Anastasia2. 31 years old. She has blue eyes and blonde hair. She has a son with hemophilia in support of her claimed royal origin. She fully complied with the genetic analysis.

Anastasia3. 30 years old. She has blue eyes and "yellow" hair. She has very detailed memories of her family and the palace. She could point the investigators to hidden places in the palace unknown to anyone except the royal family. She refused to comply with the genetic analysis. However, her husband (an unemployed actor) provided their son's DNA.

Anastasia4. 32 years old. She has blue eyes and strawberry-blonde hair. She claims that Anastasia3's dad was the architect who designed the palace and that Anastasia1 was her childhood friend who stole her hairbrush. She presented her childhood picture, which is nearly identical to the official picture. She has a son named Alexei II with hemophilia. She refused to submit to a full genetic analysis, claiming the evidence is very clear, but she agreed to submit her son's DNA to a full genetic analysis.

Anastasia5. 35 years old. She explained her age in that Rasputin erred in filling her birth certificate because he was drunk. She claims that Anastasia2 adopted her son (Anastasia2's son) from a nearby orphanage and that nobody wanted him because of his disease. She said that a simple DNA test could prove that she is not his birth mother, but no one did such a test. She has brown eyes and brown hair. She claims that the family portrait painters painted her differently to flatter her and raise the popularity of the family among the public. She complied with the genetic analysis.

Farmer. 52 years old. Admitted to being a heavy drinker. The farmer admitted to having witnessed the murder. He said each one of the Romanov family was shot once. He was paid 50 Rubles to burn and dispose of all the bodies. The farmer pointed the investigators to the bodies from which the DNA was extracted, and the identification of 6 members was confirmed. When the investigators reached the alleged Anastasia's grave, the farmer was so drunk that he tripped and fell on the single bone that remained of the body.

Farmer's only daughter. 36 years old. She has green eyes and black hair. She followed her dad wherever he went and said very little. The farmer said she lost the ability to speak after the birth of her son. The researchers noted the similarity of her son to Rasputin. The mentioning of that name made the farmer's daughter scream in agony.

Farmer's grandson. A 1-year-old. Genetic data were collected with the mother's permission (the farmer's daughter). She produced his birth certificate because she was curious to know what diseases he had.

Grigori Rasputin. Unknown age. The color of the eyes could not be determined. His hair was blond but was clearly dyed. He was too drunk to say anything but agreed to submit his genetic data to full genetic testing.

Tips

1. Remember that you are geneticists, not detectives.
2. Start by plotting the family tree.
3. Your data are aligned, but you will need to evaluate the alignment. For that, read about the pairwise global alignment of DNA sequences (for example, http://a-little-book-of-r-for-bioinformatics.readthedocs.org/en/latest/src/chapter4.html). Read also about multiple sequence alignment (MSA). You can adopt any reasonable scoring matrix and calculate the similarity between the sequences.
4. In analyzing the genetic evidence, note a 10% genotyping error rate, on average.
5. Due to your low budget, you could afford only low-coverage sequencing, so some nucleotides could not be determined with any reliability and are marked in (?).
6. Assume there is only one type of hemophilia, which is a recessive X-linked disorder.

General goal. Given a minimal budget for DNA analyses and access to testimonies, the purpose of this running exercise is to carry out a forensic investigation that will answer the questions.

- Which of the women involved, if any, is Anastasia Romanov?
- What happened to Anastasia Romanov?

Limit your analyses **only** to the information and data provided in this exercise.

To answer the above questions, write a Python code that will:

1. Evaluate the multiple sequence alignments of DNA sequences (student I).
2. Build a similarity matrix (student II).
3. Calculate a hierarchical clustering of the samples (student III).

Finally, write an essay that summarizes your methods and findings and answers the questions above (all students).

The essay. Write a formal essay (submit it as a Google Notebook file). Your essay should be structured like an academic paper with an introduction, methods, results, and discussions. There should be no references. Figures should be numbered and include legends. The essay should include a description of your analyses, a description of your Python commands, and your consideration of the results to address the questions. Your report should reveal who Anastasia is, what happened to her, and the methods you used to infer it. The results of your analysis should be summarized in an informal **Google Colab** essay of 1,300-1,500 words (ipynb).

Your essay in your Google Colab report should include: introduction, methods, results [google Colab code for all the students], figures\tables with legend, and discussion.

Submission

1. Submit all the required scripts, images, and output files.

2. Each person is required to upload their own part of the project files individually, not collectively. This helps us keep track of individual contributions. **Clearly mark your own part**.
3. Write your essay in your Google Colab Jupyter Notebook. This ensures easy access and readability for review purposes.
4. Download your Jupyter Notebook(s) and send the downloaded file(s) rather than sharing a link to a Google Colab notebook. This allows us to access your work without any access restrictions. **One student from each group should submit it**.
5. Whether you have a single Jupyter Notebook or multiple files, please organise them in a folder. After organizing, zip the folder containing all your files. Uploading a zipped folder keeps all your work together and prevents any missing files during the submission process.
6. We should be able to run all your code through the Jupyter Notebook.

Deadline. Upload your findings to Canvas by 30/10 at 15:00. Late submissions will be penalized heavily.

Bonus 1. Groups that submit all their materials by 30/10 at 13:00 will receive 1 extra point for their final grade (maximum of 25) for each member. Note: multiple submissions will disqualify you from the bonus.

Bonus 2. Groups that submit all their materials on time with the correct answer will receive 5 extra points for their final grade (maximum of 25) for each member. Note: multiple submissions will disqualify you from the bonus.

It is OK to: help your group members by reviewing their code, testing it, and debugging it. Remember that this exercise prepares you for the test. Code independently as much as possible. Work together on the paper.

It is not OK to: communicate with members of other groups, write code for others, or dictate the code for them. Be a good friend and help your friend UNDERSTAND how to code.

Grade. 50% of your grade will be based on your code, and 50% will be a group grade based on your essay.

Part one (Student I)

Goal. Calculate how well the DNA sequences provided to you are aligned with each other. It is expected that the real Anastasia's DNA sequence would be more similar to the DNA of her family.

1. The parser

Write a *parser* that reads the GeneticData – [your group #].txt (for simplicity, *GeneticData.txt*) file and outputs two fasta files, one for mtDNA and one for the Y chromosome.

Data. The file *GeneticData.txt* contains all the DNA information available.

Execution. The program should run as:

```
python FastaParser.py text_file output_fasta_file
```

2. Evaluates multiple aligned sequences

Write a program that reads an aligned fasta file (you need to test that it is aligned), evaluates the alignment, and outputs two measures of the alignment quality: the identity score and the alignment score of all possible pairs. Allow the user to determine the weight matrix. For your testing, you may adopt any reasonable weight matrix. Your script should take the fasta file and the weights file and output the scores.

Aligned sequences allow us to calculate the **percent identity** (or identity score) as follows: 100 * identical nucleotides/total nucleotides. For example, if there are eight different positions and 31 identical positions, we can say that the two sequences are 100*31/39 = 79% identical.

You can also calculate the alignment score by summing over the individual alignment scores.

Execution. The program should run as

```
python ExamineMSA.py fasta_file weight_parameters output_file
```

You should output your results to a file that looks like this:

```
SampleA | SampleB | Comparable nucleotides | uncertain nucleotides | Total nucleotides|
IdentityScore | Score
A1 A2 12 6 18 55.2% 35
A1 A3 13 5 18.3% 17
.
.
```

Part two (Student II)

Goal. Identify the most genetically similar individuals

1. From similarity to genetic distances

Write a script that builds a similarity matrix based on the scores between all the individuals. For this, you will need to read the two files that contain the identity and the alignment scores (one file for each parental chromosome), convert these two measures to similarity (genetic distances) between the individuals (read up on the subject and choose a reasonable way to do so). Print the similarity (genetic distances) between all the individuals into two tab-delimited files.

Execution. The program should run as:

```
python build_similarity_matrix.py input_file output_file
```

Your output files will look like this (A1-A5 are example individuals). The cells include the distances.

```
   A1 A2 A3 A4 A5
A1  . . . .
A2  . . . .
A3  . . . .
A4  . . . .
A5  . . . .
```

2. Printing the genetic similarity in files

Write a Python script that takes an input file with a similarity matrix (as in the above) and either ALL (for all people) or an individual name. *ALL* would print the similarity (you choose which measure) between all the individuals (all against all comparisons). *An individual's name* would print the most similar individual/s to that individual (one against all).

Execution. The program should run as:

```
python most_similar.py input_file All/Name output_file [optional]
```

Output (fake values)

```
SampleA         SampleB         IdentityScore [identity score]    OR Score [alignemnt score]
Princess Irene  Prince Fred     13%     31
Princess Irene  Nicolas II Romanov      15%     9
```

Part three (Student III)

Goal 1. Produce a haplotype file for the mtDNA and Y mutations.

1. Calculate the haplotype map

Write a script that calculates the haplotype map. You were given mtDNA and Y sequence data for multiple individuals. You may assume that they correspond to positions 1..N in each chromosome. Looking at the aligned sequences, for example

AGACCG

AGAC?G

ACAGCG

1. Identify the SNPs (positions 2 and 4) and the minor alleles (C and G, respectively),
2. Calculate the minor allele frequency (1/3 and 1/3, respectively) per SNP.
3. Write a script that produces a haplotype map file that looks like this (one for mtDNA and one for Y)

Chromosome	Position	Alleles	MajorAllele	MinorAllele	MinorFreq
mtDNA	2	A/C	A	C	0.33
mtDNA	4	A/G	A	G	0.33

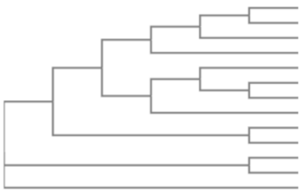
Execution. The program should run as:

```
python CalculateHapmap.py input_file chromosome_name [mtDNA or Y] output_file [optional]
```

Goal 2. Calculate a hierarchical clustering of the samples based on the **genetic distances** (created in part II).

2. Hierarchical clustering

Write a script that calculates the hierarchical clustering (a dendrogram) of files containing pairs of observations and scores. Print the hierarchical clustering or dendrograms in a graphic form (as in the example below):



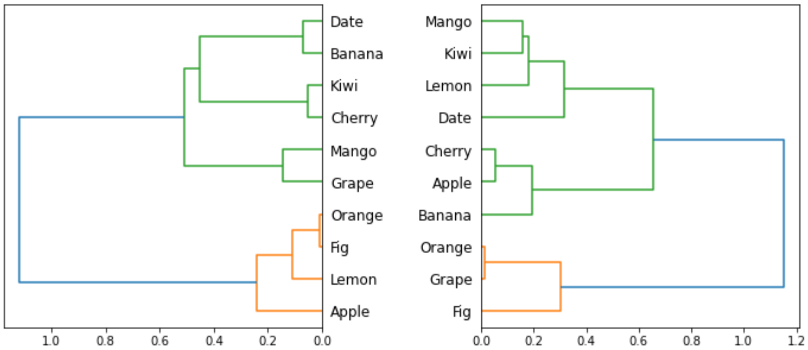
Produce four dendrograms:

1. mtDNA based on alignment score.
2. mtDNA based on identity score.
3. Y based on alignment score.

4. Y based on identity score.

Additional requirements:

1. Show the names of the individuals at the ends of the branches.
2. Color-code the Romanovs in a distinct color from the non-Romanovs.
3. Compare dendrograms #1-2 and #3-4 as in:



Goal 3. Produce a heatmap of the samples based on the genetic distances (created in part II).

Heat map

Write a script that calculates the heat map for a distance matrix. A heatmap is a visual representation of data where colors represent individual values. In genetics, it can show similarities or distances between different samples based on genetic data. For example, if you have genetic distance measurements between various individuals, a heatmap can help you quickly see which individuals are more or less similar (=different) from each other. Use the **genetic distances** between individuals to construct a heatmap. Based on the outcome, decide what happened to Anastasia.

Execution. The program should run as:

```
python PlotDistMatrices.py input_file output_pic1, output_pic2..
```

General notes

You are allowed to use only the modules studied in class. You can use pandas, but you cannot use any built-in modules that will calculate scores for you. That also means you cannot use Biopython for that. Student 3 is allowed to use any external module for the plotting part.

FAQ

- Part 1: Is the FASTA parser supposed to be able to handle multi-line sequences, given that there are none in the genetic file?
    - **Yes**.
- Part 1: Is the weight file a compulsory input, or is it optional?
    - It’s optional, but there should be reasonable default values.
- Part 3: There is little information about input/output files for the heat map.
    - You decide how best to design your program.

Good luck!