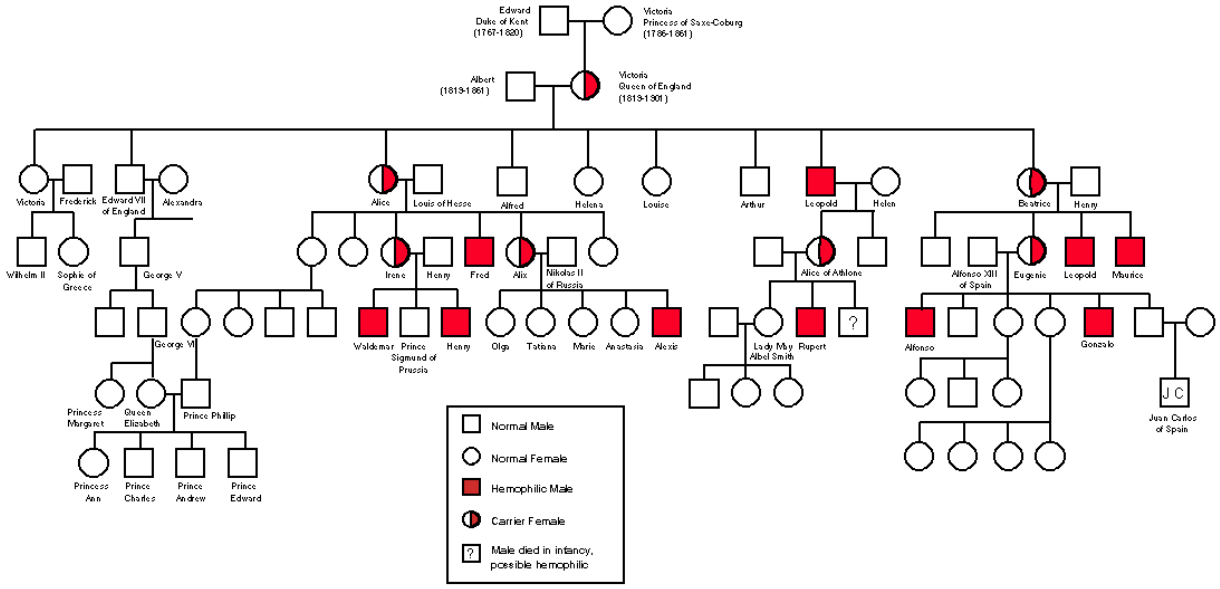
Running Exercise III

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



In January 1918, the Romanov family was executed during the uprising of the Bolshevik revolution. There are several accounts on what happened, but it is believed that angry mob shot all members of the family, 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 clear 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 are the real Anastasia. In 1932, DNA evidence (assume they existed back then) and testimonies were collected from most of the people involved.

Testimonies:

Anastasia1. 33 years old. She has blue eyes and "orange" hair. She possess an ivory hair brush 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 was able to point the investigators to hidden places in the palace unknown to anyone except the royal family. She fully complied with the genetic analysis.

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 hair brush. She presented her childhood picture which is 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 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 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 can prove that she is not his real mother, but no one did such 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 be a heavy drinker. The farmer admitted to have 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 permission of the mother (the Farmer's daughter) and after the birth certificate documents were provided since she was curious to know what diseases he has.

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. Read about 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. Alternatively, you may use online tools such as those listed here: http://www.ebi.ac.uk/Tools/psa/. Note, sequence alignment tools require that your sequences should be submitted in particular formats, such as FASTA format.
4. In analyzing the genetic evidence please note that there is 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.
7. Recommended reading: Ancient DNA: the first three decades – papers from a conference on ancient DNA. This paper is strongly recommended for this exercise. http://rstb.royalsocietypublishing.org/content/370/1660/20130371. Other papers from the meeting are available here: http://rstb.royalsocietypublishing.org/content/370/1660.

Goal. Given a very limited budget for DNA analyses and access to the 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 analysis **only** to the information and data provided in this exercise.

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

1. Do 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 (everyone)

Deadline. Submit your findings by Sunday 25th at 17:00.

Bonus. The first group to submit their materials with the correct answer will receive three extra points for their final grade (for each member). Note, multiple submissions will disqualify you from the bonus.

It is OK to: help each other out by reviewing each other's code, test them, and debug them. Remember that this exercise prepares you for the test, so code independently as much as possible. Work together on the paper.

It is not OK to: communicate with members of other groups.

Part one (Student I)

Goal. We want to calculate how well two DNA sequences are aligned with each other. For that, we need to write a multiple sequence aligner. It is similar to the pairwise sequence aligner that you wrote in Running Exercise II. Multiple sequence aligner works on files with multiple sequences and outputs 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.

Write a program that takes the fasta file and a file with the weights and outputs the scores.

The program should be run as MSA.py fasta_file weight_parameters output_file

Data. Data for the assignment can be downloaded from Canvas. The name of the archive file is *GeneticData.txt*. You should create two fasta files, one for mtDNA and one for the Y chromosome.

Output. You should output the MAS results to a file that looks like this:

```
SampleA SampleB IdentityScore Score
A1 A2 55.2% 35
A1 A3 51.3% 17
.
.
```

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 score of the alignment by summing over the individual alignment scores.

Part two (Student II)

Goal. Build a similarity matrix. Read the MSA file. Convert the two measures (identity score and alignment score) to two tables and print them to two tab-delimited files.

```
   A1 A2 A3 A4 A5
A1 0 2 4 1 3
A2 1 . . . .
A3 3 . . . .
A4 3 . . . .
A5 5 . . . .
```

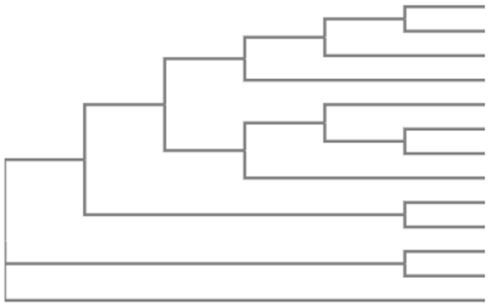
Write another python script that reads an individual name, reads the table, and outputs the most similar individual.

Part three (Student III)

Goal. Calculate a hierarchical clustering of the samples. Read the MSA file. Calculate the hierarchical clustering by importing the following modules:

```
from scipy.cluster.hierarchy import dendrogram, linkage
from matplotlib import pyplot as plt
```

Print the hierarchical clustering (see example below) based on the results of the table. Write the sample names and the score on the right.



Submission. The results of your analysis should be summarized in an informal **Google Colab** report of 1,300-1,500 words. The information should include a description of your analyses, a description of your python commands, and your consideration of the results to address the question. If you find it difficult to write text in Google Colab, you may attach the Google Colab file as an appendix and write the report in Word. Summarize your results and write your conclusions. Your report should reveal who is Anastasia, what happened to her, and the methods that you used to infer it.

Good luck!