Exercise 1

- What is the computational *complexity* of distance matrix calculations: with increasing number of proteins, how will the run time increase? Assuming that the distance matrix for 1000 proteins takes 1s to compute, how long will it (roughly) take for 10.000, 100.000, 1.000.000, 10^9 proteins?
    - Pairwise distance calculations square quadratically ($O(n^2)$ in computer science lingo), i.e. if you have 10 times more sequences, overall runtime will be 100 times slower (and so on). Our examples would thus take 100s, 10.000s, 1.000.000s and 1.000.000.000.000s, respectively.
- As said before, the Jaccard index is quite simple, and probably provides a very inaccurate estimate of the evolutionary distance between two sequences. In which ways could the distance calculation be modified to provide a more realistic estimate of the "true" underlying evolutionary processes? Which parameters could have the strongest influences on the accuracy of any distance estimate?
    - Jaccard assumes that transition probabilities between all amino acids are equal, which is not realistic. Depending on structural and biochemical similarities, some amino acid transitions are known to be more likely than others. More complex methods account for this effect by employing special transition matrices to weight mismatches. Other factors include for instance multiple substitutions at the same site and dependencies between sites (e.g. insertions/deletions, selective pressure affecting multiple co-localized sites), which are accounted for by modeling approaches.
- Protein sequences and nucleotide sequences (DNA/RNA) provide similar yet distinct types of information. What are the most striking differences between the two with regard to the calculation of evolutionary distances? For which kinds of problems or studies would you prefer the one or the other, and why?
    - Amino acid sequences are evolutionarily more conserved than nucleotide sequences: synonymous changes in DNA/RNA space do not lead to changes in protein space (due to wobble base pairs). Thus, protein information allows us to look further back in time (and are what is typically used for questions on non-recent evolution). DNA/RNA, on the other hand, has a higher short-term resolution and can thus be used to distinguish for instance between recent viral strains or between human individuals (molecular fingerprinting in forensics).

Exercise 3

Take a bit of time to inspect the tree:

- How many major clades can you see? Which clade features, on average, the most distantly related members?
    - The tree has 3 major clades: Two small ones (genus names starting with Fuso/Fuso/Rick and Camp/Halo/Thio/Legio, respectively) and one big ones (all remaining leaves). The highest evolutionary divergence (= most distantly related members) is found in the Camp/Halo/Thio/Legio clade.
- What information do the taxonomic labels provide with regards to taxonomic ranks?
    - The first two words in each label refer to genus and species identifiers. Some labels have additional information on subspecies and type/culture strains (provided after the species).

Exercise 4

- Can you identify clear differences between the trees? From an evolutionary perspective, what do these differences imply?
    - Firstly, the UPGMA tree is ultrametric (i.e. all leaves have the same distance to the root) while the NJ tree is not. This is because the UPGMA algorithm assumes a uniform molecular clock across the whole tree: all taxa are expected to evolve/diverge at the same speed. This assumption is, however, known to be violated regularly (some microbial groups evolve faster than others, for instance via hypermutability traits). NJ tress may reflect differences in evolutionary divergence more accurately by resolving branches at varying depths. As a second difference, the NJ tree shows only two clades instead of three: the Fuso/Fuso/Rick clade is now a deep-branching part of the larger clade from before. This suggests a considerably closer evolutionary relatedness between these clades than predicted by UPGMA.
- What is the most important algorithmic difference between NJ and UPGMA?
    - NJ does not only consider the raw similarity between two sequences when deciding which leaves to connect next, but also uses their distance to all other sequences. Thus, if two sequences are distant from each other, but even more distant from all other sequences (relatively speaking), they may still form clearly detectable clades. In contrast, UPGMA tends to depict members of such clades as only distantly related to each other, which can complicate interpretation. As another difference, the uniform molecular clock assumption made by UPGMA dictates a natural root for the tree. However, NJ does not make that assumption and thereby produces unrooted trees by default. In practice, midpoint rooting is often applied to the resulting trees, which places the root such that the distance of all leaves to the root is somewhat similar. This accounts for the fact that divergence is expected to be broadly consistent across the tree of life, but allows for individual fast-evolving outlier leaves or clades.