Minitask: Pairwise Sequence Alignment via Compression Algorithms

Markus Frohmann k12005604, Janis Traweger k1627710

January 2024

1 Task Preparation

To get MT genomes for various organisms, we use the NIH Nucleotide Search and search for "complete mitochondrial genome", and additionally filter for "Mitochondrion" (using the filter "Genetic compartments" on the left). Then, we select mammals from the search results with the structure as follows: complete mitochondrial genome.

We select the following mammals for our comparisons:

- Felis catus (Cat)
- Bos taurus (Cattle)
- Mus musculus (Mouse; strain Balb/cJ)
- Rattus norvegicus (Rat; wild-caught)
- Canis lupus (Wolf)
- Ovis aries (Sheep)

2 Pairwise Sequence Alignment

To get a viable distance measure between aforementioned genome sequences we employed the use of compression algorithms proposed in the lecture. Given two sequences A and B we compress them and their concatenation. Then we apply the proposed formula:

$$d(A,B) = 1 - \frac{c(A) + c(B) - c(AB)}{max(c(A), c(B))}$$

where c(A) is the Compression of sequence A. In theory a perfect compression algorithm will break a sequence down to its bare minimum and decoding its pattern. As there are different compression algorithms commonly used we compared gzip and 7zip to see if there happens to be a noticeable difference

between them. In general 7zip compression results are more compact (ie.: the resulting compression file is smaller) [1] meaning this algorithm should be better for the task at hand. As both algorithms are lossless no information gets lost, and so the better compression algorithm will "find" more patterns/generalise better.

3 Results

Our analysis of the results was executed in a jupyter notebook (contained in the "Code" - folder). We plotted the distance between each sequence in a heatmap (Figure 1) and notice that, while the distance is generally smaller for the 7zip-algorithm, the resulting differences in the distances between the sequences are more pronounced.

In general we concluded that 7zip algorithm seems to be better for the task as there should not be high a distance (like 0.9 in the gzip) between any of the genome sequences as they have a similar task. The smallest difference in the mitochondrial genome appears to be between sheep/cattle and rat/mouse. As the former are domesticated herbivores and the latter have apparent physical similarity between the two, the algorithm/distance formula seems to work as intended. At the end we also employed a neighbor-joining tree algorithm to construct a phylogenetic tree (Figure 2). When we compare the trees with a neighbor-joining tree created from the clustal omega algorithm (Figure 3) we confirm that the compression distance metric seems to do a similar job.

References

[1] https://www.infiniroot.com/blog/1289/performance-comparison-different-compression-methods-mysqldump

Figure 1: Heatmap of distance between MT sequences

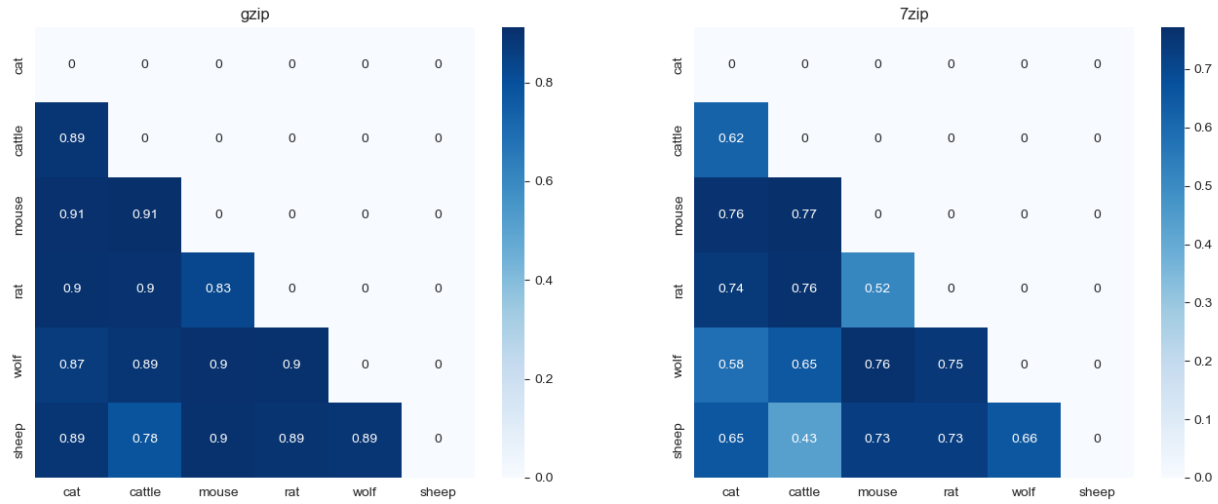

Figure 2: Phylogenetic tree

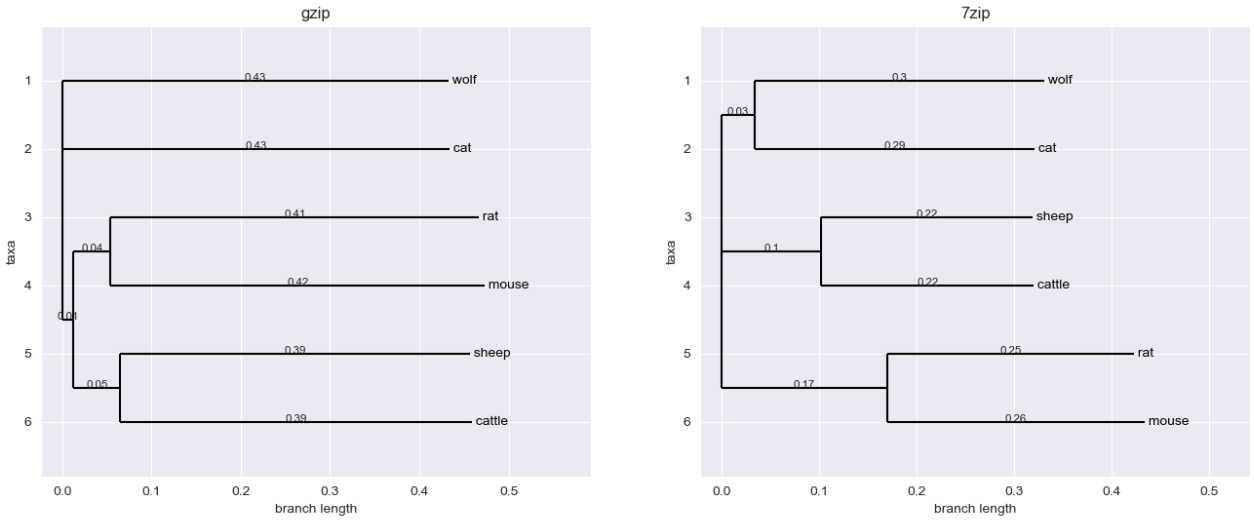

Figure 3: Phylogenetic tree