Gorilla Report

Lasse Faurby (faur), Nikolaj Worsøe Larsen (niwl), Philip Flyvholm (phif), Simon Skødt (sijs) and Thomas Rørbech (thwr).

October 3, 2023

Results

Our implementation produces the expected results for all pairs of species. However, it is worth noting that the order of the species comparison may vary from the given `.out` file. Furthermore, our implementation always prioritizes the species with the longest DNA sequence as the first species in the output. As a result, there is a discrepancy in the example of Pig‑‑Deer: 565, where our output is `Deer--Pig: 565`.

We compared the species in `HbB_FASTAs-in.txt` with the `Pig 809283 2PGH`, given by

```
VHLSAEEKEA VLGLWGKVNV DEVGGEALGR LLVVYPWTQR FFESFGDLSN ADAVMGNPKV KAHGKKVLQS FSDGLKHLDN
LKGTFAKLSE LHCDQLHVDP ENFRLLGNVI VVVLARRLGH DFNPDVQAAF QKVVAGVANA LAHKYH
```

The closest species to *Pig* is `Human 2144721 HBHU 4HHB`, with the following optimal alignment (we have inserted spaces for readability):

```
Human--Pig: 646
MVHLTPEEKS AVTALWGKVN VDEVGGEALG RLLVVYPWTQ RFFESFGDLS TPDAVMGNPK VKAHGKKVLG AFSDGLAHLD
NLKGTFATLS ELHCDKLHVD PENFRLLGNV LVCVLAHHFG KEFTPPVQAA YQKVVAGVAN ALAHKYH

 -VHLSAEEKE AVLGLWGKVN VDEVGGEALG RLLVVYPWTQ RFFESFGDLS NADAVMGNPK VKAHGKKVLQ SFSDGLKHLD
NLKGTFAKLS ELHCDQLHVD PENFRLLGNV IVVVLARRLG HDFNPDVQAA FQKVVAGVAN ALAHKYH
```

Implementation details

We have chosen a bottom-up iterative approach for the implementation. This approach computes the results of all sub-problems and stores them in a matrix. The matrix contains the blosum values of the letters in the two DNA strings that need to be compared. Finally, the matrix is used to find the optimal solution through backtracking.

For the comparison of two sequences of length n and m, respectively, our implementation of `solve` uses $O(n \cdot m)$ time. The function `solve` is called $O(k \cdot k)$ times, where k is the number of species that need to be compared. We use $O(n \cdot m)$ space to store the matrix with the results of all sub-problems.