Xiheng He

Lisanne Friedrich

Exercises for Algorithmic Bioinformatics II

Assignment 8

Xiheng He

December 2021

Exercise 4 (ClustalO & ClustalW, 10P):

Read the following publication on the multiple sequence alignment algorithm Clustal Omega:

Sievers, Fabian, and Desmond G. Higgins. *Clustal Omega for making accurate alignments of many protein sequences.* Protein Science 27.1 (2018): 135-145.

Apply both algorithms (ClustalO + ClustalW) to the sequences below. Visualize and discuss the resulting MSA appropriately:

```
Human: TAVCMECFREAAYTKRLGTEKEVEVIGGADKYHSVCRLCYFKKA
Danio Rerio: NAVCMQCFKEAAYTKRLGAEKEVEVIGGSDKYHAVCRCCY
Tryp Brucei: SAVCMECHNRKASFTYRTVKSDERKLVGGSDMYMSVCRSCYETK
Mus Musc: TAVCMECFREAAYTKRLGLEKEVEVIGGADKYHSVCRLCYFKKS
Vac Virus: TAVCMKCFKEASFSKRLGEETEIEIIGGNDMYQSVCRKCY
Leishm Major: TAVCMMCHEQPACFTRRTVNVEQQELIGGADMYIATCRECYSKQ
Thermot Marit: AVCHRCGEYNATLTLKVAGGEEEIDVGGQEKYIAVCRDCY
Human M163: TAVCMECFREAAYSKRLGTEKEVEVIGGADKYHSVCRLCYFKKA
```

For extra points: compare and discuss the ClustalO/ClustalW MSA with the result of your center-star implementation!

- In a nutshell:

  ClustalW is a matrix-based algorithm where Clustal Omega is consistency-based. They both use pairwise alignments

like EMBOSS and LALIGN, but ClustalW use similarity for MSA where Clustal Omega uses seeded guide trees and a new HMM engine that focuses on two profiles to generate MSA.

- Algorithm:
  - ClustalW:

    ClustalW uses progressive alignment methods as stated above. In these, the sequences with the best alignment score are aligned first, then progressively more distant groups of sequences are aligned. This heuristic approach is necessary due to the time and memory demand of finding the global optimal solution. The first step to the algorithm is computing a rough distance matrix between each pair of sequences, also known as pairwise sequence alignment. The next step is a neighbor-joining method that uses midpoint rooting to create an overall guide tree. The process it uses to do this is shown in the detailed diagram for the method to the right. The guide tree is then used as a rough template to generate a global alignment.
  - ClustalO:

    Clustal Omega has five main steps in order to generate the multiple sequence alignment. The first is producing a pairwise alignment using the k-tuple method, also known as the word method. This, in summary, is a heuristic method that isn't guaranteed to find an optimal alignment solution, but is significantly more efficient than the dynamic programming method of alignment. After that, the sequences are clustered using the modified mBed method. The mBed method calculates pairwise distance using sequence embedding. This step is followed by the k-means clustering method. Next, the guide tree is constructed using the UPGMA method. This is shown as multiple guide tree steps leading into one final guide tree construction because of the way the UPGMA algorithm works. At each step, (each diamond in the flowchart) the nearest two clusters are combined and is repeated until the final tree can be assessed. In the final step, the multiple sequence alignment is produced using HHAlign package from the HH-Suite, which uses two profile HMM's. A profile HMM is a linear state machine consisting of a series of nodes, each of which corresponds roughly to a position (column) in the alignment from which it was built.

- Time complexity:

  ClustalW : $\Omega(N^2)$, ClustalO : $\Omega(L^N)$ (L is the length of the alignment, N is the number of sequences)

```
Thermot         -AVCHRCGEYNATLTLKVAGGEEEIDVGGQEKYIAVCRDCY----   40
Vac             TAVCMKCFK-EASFSKRLGEETEIEIIGGNDMYQSVCRKCY----   40
Danio           NAVCMQCFK-EAAYTKRLGAEKEVEVIGGSDKYHAVCRCCY----   40
Mus             TAVCMECFR-EAAYTKRLGLEKEVEVIGGADKYHSVCRLCYFKKS   44
Human           TAVCMECFR-EAAYTKRLGTEKEVEVIGGADKYHSVCRLCYFKKA   44
Human_M163      TAVCMECFR-EAAYSKRLGTEKEVEVIGGADKYHSVCRLCYFKKA   44
Tryp            SAVCMECHNRKASFTYRTVKSDERKLVGGSDMYMSVCRSCYETK-   44
Leishm          TAVCMMCHEQPACFTRRTVNVEQQELIGGADMYIATCRECYSKQ-   44
                 ***  * .  *  : :     :   :** : * :.** **
```

Figure 1: mutiple sequence alignment

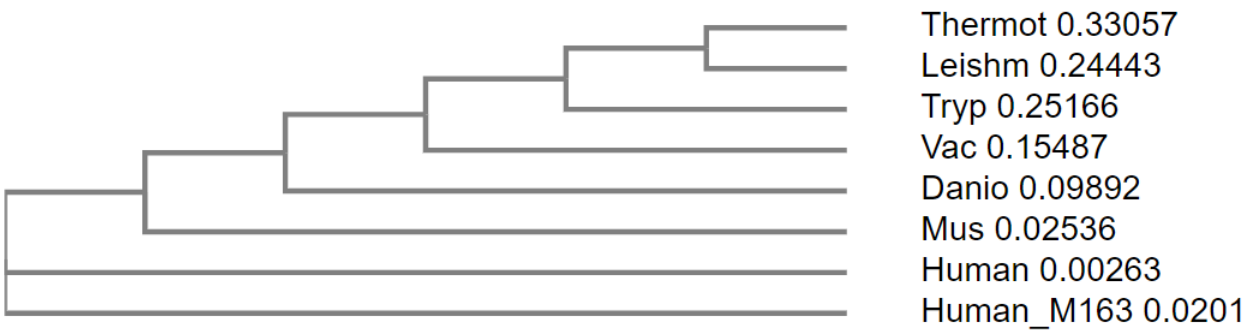

Figure 2: phylogenetic tree