Bioinformatics, computer lab VI
University of Potsdam WS 2019/2020, S. Hartmann

Multiple sequence alignments

Objective

The goal of this exercise is to compute, read, and evaluate a multiple sequence alignment of seven globin protein sequences from a variety of organisms. Specifically, you'll work with four hemoglobin sequences from mammals, one globin from a fish, one myoglobin from a mammal, and one leghemoglobin from a leguminous plant.

The data

☞ Download the file `globins.fasta` from Moodle, this file contains seven globin sequences. Look at it with `less` in a Linux terminal or with a text-only editor: do they all have approximately the same length? Can you find all sequences that are listed above?

1 Computing the alignment

You will use the Clustal algorithm to align the rather conserved globin protein sequences. For relatively easy alignments, the Clustal programs are a good choice. However, as discussed in class, other alignment software may be a better choice for alignments with many and/or very long and/or very divergent sequences.

For the command-line version, the general format of the program call is as follows:

```
clustalw inputfile -OUTFILE=outputfile
```

Of course, you need to substitute 'inputfile' with the name of the globin file and 'outputfile' with an appropriate name for the output file. I suggest using the extension ".aln" for the output file.

☞ Execute the command to compute the alignment.

☞ Look at the output that is generated in the terminal: which part of the output was generated during which of the steps of the progressive alignment algorithm?

☞ Use `ls` to see which files were generated.

1.1 Viewing the guide tree

One of the generated files is the guide tree that was discussed in class. It has the file ending .dnd. You can look at the file using `less`, but the guide tree is written in newick format, which is a little difficult to read for the human eye.

If you want to see a graphical representation of this tree, use the program `retree` from the phylip software package. Type `phylip retree` to start this program. This will show you the short `retree` menu: The first option is for viewing a tree from a file, which is what you want to do. So type `y` to accept all defaults, then hit Return. You will then be asked which tree you want to view. Type the

name of the .dnd file and hit Return. (The numbers on the internal nodes are simply IDs and not relevant here at all.)

To exit the retree program, type `q` (for 'quit'), then `n` because you don't want so save anything now.

2 Viewing the multiple sequence alignment

☞ Use `less` to view the alignment file that ClustalW generated. The alignment is broken up into three "chunks" that are displayed underneath each other.

☞ It's just a black and white representation, but does it seem like a good alignment? Why or why not? And can you guess what the symbols " * ", " : ", and " . " below the alignment columns mean?

☞ For an enhanced viewing experience you can use the program ClustalX or Jalview to view the alignment. Both of them can color DNA and amino acids for easier viewing, and both of them are installed on the machines in the computer pools.

In the terminal, type `jalview` to open its graphical interface. The program starts, and a News window is briefly displayed and then disappears. Under the **File** menu, select **Input Alignment** → **From File** and select the alignment file you just generated. Make sure the "Files of Type" option at the bottom of the window is set to "All Files" or to "Clustal (.aln)".

You can change multiple aspects of the alignment display, for example:

- the font size (under "Format" → "Font...")
- to see the entire alignment, resize the jalview window and under "Format", click on the "Wrap" option
- to color amino acids by physico-chemical properties or similarity, select one of the option under "Color". Try out a few different of the options and select one that allows you to easily see where alignment regions are more or less conserved.

3 Reading and evaluating the alignment

☞ Last week you used BLAST to identify a plant leghemoglobin sequence as potentially homologous to the human hemoglobin. The pairwise alignment, however, covered only the last third of both sequences. How conserved are the alignment columns in the first third of the alignment? the last two thirds? overall? between the different types of globins? ...?

☞ How many completely conserved amino acids does this alignment contain? (Hint: look at the information about "Conservation" below the alignment) Are these all in one region, or are they spread out? What might be the biological significance of these?

☞ Can you tell from looking at the alignment which two sequences are the most similar? Which are the least similar sequences? – Probably not. You can get a sense for how similar or distant the sequences in the alignment are, but you cannot determine the precise relationships between these sequences without computing a phylogenetic tree. We will cover phylogenetic inference in more detail next week.