Sequence bioinformatics, computer exercise X

University of Potsdam, WS 2023/2024, Stefanie Hartmann

Give it all you've got! And since the mere presence of smartphones reduces available cognitive capacity (`https://www.journals.uchicago.edu/doi/full/10.1086/691462`), I suggest you put your cell phone out of sight and into your bag for the duration of this computer lab.

Multiple sequence alignments

Objective

After completing this exercise, you will be familiar with the use and output of commonly used alignment software. Today you will use the default settings of most programs, but feel free to experiment with the different parameters, as they can have large effects on the final result.

1 Alignment viewing

☞ Although it is possible to view the plain-text version of an alignments in the terminal using `less`, reading and interpreting alignments is much (!) easier with a dedicated alignment viewer.

- on the computer pool machines, I recommend using an online alignment viewer, e.g., `https://alignmentviewer.org/` or `https://www.ebi.ac.uk/Tools/msa/mview/`. Note: you can also install one of the viewers listed below, but they'll be gone the the next time you log in.
- on your own computer, I recommend downloading and installing JalView (`http://www.jalview.org/getdown/release/`). If for some reason this won't work on your machine, you can also try AliView (`http://www.ormbunkar.se/aliview/`) or the simple ClustalX Viewer (`http://www.clustal.org/download/current/`).

2 Logging into the cluster and preparing for the analyses

☞ Log into the university's computing cluster as described before and change to your directory for this class on the /work partition. Create a new directory for today's exercise (e.g., lab10) and change into it.

☞ You'll use four different MSA programs today, all of which are globally installed. Load the corresponding modules:

- `module load bio/Clustal-Omega/1.2.4-foss-2018b`
- `module load bio/MAFFT/7.453-GCC-9.3.0-with-extensions`
- `module load bio/PRANK/170427-GCC-9.3.0`
- `module load bio/TCoffee/13.45.0-GCCcore-10.2.0`

☞ Start an interactive session: `srun --cpus-per-task=1 --mem=5G -p teaching --pty bash`
Please coordinate with your neighbors so not everyone starts the interactive session at the same time.

3 The data

For today's computer lab, you'll work with a set of sequences that belong to the toc75 gene family. In theory, you could use the database sequences from last week that you considered to be homologous to the query sequence. In practice, I'm providing a slightly modified set of sequences: I've removed some isoforms and a pseudogene, and I've added a few other sequences not present in last week's database file.

☞ The sequences are available in the shared folder /home/group/seqBioinf/ as tocs.fasta.gz. Copy them your own directory, then extract and inspect the file. Sequence IDs have the same species' prefixes as before.

☞ How many sequences are there? Do the sequences all have approximately the same length? And why might this information be useful to have before computing a multiple sequence alignment?

4 Computing the alignments

Since you will compute alignments of the same sequences with different alignment programs, you need to keep track of which software generated which output file. I recommend coding this information within the name of the output file.

4.1 ClustalO

4.1.1 Computing the alignment

Clustal-Omega (or ClustalO) is a recent implementation (`https://doi.org/10.1038/msb.2011.75`) of progressive sequence alignment that is designed for large data sets.

☞ `clustalo -h` gives you information about the usage and parameters of ClustalO.

☞ To just output the alignment, you'll only need to use the parameters -i and -o. If you'd like to also output the distance matrix and guide tree, you'll need three additional parameters: --distmat-out=, --full, and --guidetree-out= . Compute the alignment but note that all but the parameter '--full' require a file name; use the extension ".fasta" for the name of the output alignment. As just mentioned, I recommend naming the output files such that you'll know that they were computed using ClustalO.

☞ If you've computed the matrix and guidetree, you can view them with `less`. For the matrix you'll need to make your terminal window wide (and/or use the -S option). The guide tree is in newick format (`https://evolution.genetics.washington.edu/phylip/newicktree.html`) and a little difficult to read for humans. You can view it using the retree program of the PHYLIP package:

1. load the module using `module load bio/PHYLIP/3.697-GCC-9.3.0`
2. open the program by typing `retree`
3. type Y and hit enter; type the name of the guidetree file and hit enter
4. you'll only see the top half of the tree. Type j and hit enter, and you'll see the bottom half
5. type x and hit enter, type n and hit enter. Now you've got your terminal back...
   (Admittedly, this was a little cumbersome. In two weeks you'll compute phylogenies and use a different viewer for visualizing trees in newick format.)

4.1.2 Viewing the alignment

☞ Use `less` to view the alignment you just computed. The alignment is also in fasta format. (Note: if unaligned and aligned sequences can both be in fasta format, how do you know by looking at the data with `less` whether the sequences are aligned or unaligned?)

To see the aligned columns underneath each other and color-coded, you'll have to view the file with a dedicated alignment viewer, on your local machine.

☞ Use the `scp` command as described before to copy the newly computed alignment file to the computer pool machine or to your own computer. View it with an online or locally installed MSA viewer as described above. Explore the viewing options (which are different for different viewers!): font size, color-coding of amino acids, showing conserved columns underneath the alignment, etc.

☞ Take your time getting familiar with the alignment. Scan the alignment columns and evaluate their level of conservation and gappiness.

☞ How long is the alignment? How conserved are the different regions? Is the level of conservation similar throughout the entire alignment length? Are there long gaps or many short indels?

4.2 MAFFT

4.2.1 Computing the alignment

MAFFT can compute multiple sequence alignments using different methods, e.g., optimized for speed or optimized for accuracy. `mafft -h` will provide a quick overview of the different methods, and much more information about the different ways of running MAFFT is available at `https://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html`

☞ Select whichever option seems more appropriate. If you're unsure, I suggest using a setting that is optimized for accuracy, is recommended for smaller alignments (200 sequences or fewer), and uses an iterative refinement method that incorporates local pairwise alignment information: `mafft --maxiterate 1000 --localpair tocs.fasta > aln.mafft.fasta`

4.2.2 Viewing the alignment

☞ The only output generated is the alignment. Use `scp` to copy it to your local computer and inspect it with an alignment viewer.

☞ Answer the same questions for this alignment that you answered for the ClustalO alignment. Compare the two alignments.

4.3 T-Coffee

4.3.1 Computing the alignment

TCoffee is a fairly complex software with lots of run modes and parameters. Type `t_coffee` to see a rather long list of parameters and programs it can work with. The full documentation is available online at `https://tcoffee.readthedocs.io/en/latest/tcoffee_main_documentation.html`.

☞ Compute a multiple sequence alignment using the command `t_coffee tocs.fasta -output fasta,html`. This will take approximately 5 minutes to complete. Do not enter another command while the program is running!

☞ Once TCoffee is done, it will output the alignment to the terminal, and it will also provide some information about the run and its output files. Type `ls` to see that three files have been generated: The guide tree ('tocs.dnd'), the alignment in fasta format ('tocs.fasta_aln'), and a file in html format ('tocs.html').

☞ Rename the files so you'll know they have been generated using the TCoffee program. Use ".fasta" as the file extension for the alignment file.

☞ If you like, view the guide tree using the retree program as before.

4.3.2 Viewing the alignment

☞ Use `scp` to copy the alignment file and the html file to your local computer and inspect it.

☞ Inspect the alignment using JalView. Answer the same questions for this alignment that you answered for the ClustalO and MAFFT alignments. Compare the three alignments.

☞ On your local computer, use the internet broswer of your choice to open the html file. You'll see the same alignment as with the MSA viewer, but TCoffee consistency scores can be seen as well.

> **From the manual:** The CORE index is an estimation of the consistency between your alignment and the computed library. The higher the consistency, the better the alignment. The score reported with every T-Coffee alignment is the consistency score (depending on the version it can be normalized to 100 or 1000). If you want to go further and estimate the local consistency (known as the CORE index), an html file is automatically created each time you run T-Coffee; it is colored version of your alignment where residues are colored according to their consistency score, from blue (low consistency) to red (high consistency). It is not fool-proof but in principle you can expect positions with a score above 6 to be correctly aligned.

4.4 Prank

4.4.1 Computing the alignment

☞ The fourth and last MSA program you'll use is Prank (http://wasabiapp.org/software/prank/). To find out about its usage, type `prank -help`.

☞ The minimum set of parameters you need to specify are -d= and -o= (followed by the input and output file name, respectively). Use this to compute an alignment. Again, this will take about 5 minutes to complete. Do not enter another command while the program is still running.

From the terminal output you'll see that Prank also does iterative refinement. By default, Prank computes five iterations, but this number can be changed.

☞ Once you have the terminal prompt back, check that the alignment has been computed.

4.4.2 Viewing the alignment

☞ Use `scp` to copy the alignment to your local computer. Inspect it and compare it to the other three alignments.

5 Comparing the alignments

5.1 Comparison by manual inspection

Some MSA viewers (e.g., JalView) allow you to open and view more than one alignment, but you'll need to adjust the window of each alignment, and you'll also need to decrease the font size of the amino acids. This way, you can compare alignment regions between the different results. If you color-code the amino acids (e.g., using the ClustalX, Zappo, or Taylor color scheme, `https://www.jalview.org/help/html/colourSchemes/index.html`), you'll be able to more easily compare aligned regions.

☞ Are there broad alignment regions (e.g., the beginning or middle or end) that are aligned very differently by the different programs? Or more or less the same?

☞ Can any of these differences be explained by the algorithm of the alignment programs?

☞ Discuss the quality of your multiple sequence alignments. Which alignments do you think are good, which one do you think is the the *best* – and what are the criteria for this decision?

☞ Is one of the multiple sequence alignments you computed particularly bad? If so: which alignment, and by which criteria? How would you improve alignment quality?

☞ Do any of the sequences appear misaligned in one or more of the alignments? Are there sequences that, in your opinion, shouldn't even be in the alignment(s)?

☞ In lecture I discussed the problems of aligning multi-domain sequences. Do you think all of these sequences contain the same functional domain(s)? Are they globally homologous? And, therefore, can and should this set of sequences be globally aligned?

5.2 Meta-alignment

The questions in the previous section might not be so easy to answer, especially if you don't yet have a lot of experience with multiple sequence alignments. Even with more experience, coming up with the answers will require some serious alignment gazing. Alternatively, you can use the M-Coffee approach, which computes a single multiple sequence alignment that is a consensus alignment from two or more input files. Residues are color-coded according to their agreement with the different input alignments:

☞ Use the *M-Coffee server* at `http://tcoffee.crg.cat/apps/tcoffee/do:mcoffee` to newly generate alternative alignments that are then combined into one multiple sequence alignment. Direct your browser to the M-Coffee server, upload the unaligned sequences. Under "Show more options" select the MSA programs to be used: Preferably select those that you just used here (although not all of them might be available).

☞ In which regions do the different alignment programs agree, which alignment regions are more problematic? Which sequences are more difficult to align?

5.3 Automated comparison of alignment quality

In bioinformatic pipelines, manual inspection of tens or hundreds (or even thousands) of multiple sequence alignments is not feasible, and the *automated* comparison of alignment quality is often the only option.

☞ Suggest (not use, not implement, just describe) an automated approach that you would use to compare the quality of alignments – i.e., without opening them in an alignment viewer and looking at them. This approach can be an existing one, or it can be of your own design. Briefly justify why you think this is a good way of automated alignment comparison. And can you think of potential problems of this approach?

6 Exiting the terminal

Log out of the interactive session with `exit`.

Then log out of the current session to the cluster, type `exit` again.

Finally, to cleanly close the terminal window, type `exit` yet again.

Do not close a terminal window with the mouse, especially not if an ssh session is still active.

7 Summary

☞ Summarize what you have learned about the computation of multiple sequence alignments from this set of alignments.

☞ Summarize what you have learned about the biology/sequences/evolution of the toc75 gene family from today's exercises. No additional information needs to be looked up, and speculative answers are again encouraged.

➜ Don't forget to take your flash drive with you when you leave!