lab5_questions_and_maj_answers

Course: Genomics & Bioinformatics

Lab: 5 - Alignments of mitochondrial sequence

PART 2

Author: Mohammed Al-Jaff

Part 2: Pairwise Alignment

Question 1. To which species do the different sequences belong?

Ans:

```
seq1 - Acinonyx jubatus (cheetah)
seq2 - Caenorhabditis elegans (worm/nematode),
seq3 - Drosophila simulans (fly)
seq4 - Pongo abelii (orangutan)
seq5 - Allomyces macrogynus (fungus)
seq6 - Monosiga brevicollis (flagellated eukaryote)
seq7 - Mus musculus domesticus (mouse)
seq8 - Homo sapiens (human)
seq9 - Drosophila melanogaster (fruit fly)
seq10 - Saccharomyces cerevisiae (fungus)
seq11 - Lynx canadensis (lynx )
seq12 - Caenorhabditis remanei (nematode)
seq13 - Fritillaria persica (flowering plant)
seq14 - Phodopus sungorus (mouse)
```

Question 2. There one sequence that should stand out from the others. Write down the name of the corresponding fasta file, the species name, and the type of sequence. How did you identify it?

Ans: All search was done with nucleotide blast with the whole sequence provided as blast query. In each case, we found exact and identical hits meaning that these sequences are not 'undocumented'.

The outlier looks to be sequence 13, which turns out to be the the mitochondrial genome of the flowering plant *Fritillaria persica*. Why? Because the rest of the seqeucens seem to come in groups of at least two organisms corresponding to the same clad, eg (humans and urangutag), (lynx, cheetah), (bakers yeast, some other fungus). Only this one is a plant among all the given sequences. Not only that, what is truly distinguishing of this sequences is that it is a chloroplast genomes and not a mitochondria genomes.

Question 3. What is the common name of this family of species? Which other species from the dataset belong to that same family? Perform the same thing (and answer the second part of the question) with the species from Seq14.

Ans:

According to wikipedia, the family that Pongo abelii (seq 4) belongs to is called Hominidae (primates) containing the genera that itself contains species such as humans (genera homo), chimpanzees (genera pan) and gorillas (genera gorilla), in addition to the urangotangs.

As for Phodopus sungorus (seq14), we find that this mouse species belongs to the 'Cricetidae' family of rodents which seem to be the label for rodents found in the 'new world'/americas which includes all the other rodents such as hamsters and rats and mice, also from the new world. An interesting thing here is that we also happen to have an 'old world' rodent representative from Mus musculus domesticus (seq 7) which belong to the 'sibling' family Muridae within the same superfamily Muroidea. '

Question 4. What do you learn from this table? What can you expect from the alignment? (make at least one hypothesis)

Based off of figure 3:

Hypothesis: Just based off of figure 3, one of the fungi species [seq5 Allomyces macrogynus] will have much better global alignment to the human sequences [seq 8] compared to the other fungi representative seq10 - Saccharomyces cerevisiae

further more:

Hypothesis: Our fungi sequences [seq5 - Allomyces macrogynus and seq10 - Saccharomyces cerevisiae] should end up having a relatively better alignment to each other than to any other sequence when genes relating to translations are considered.

Question 5. Write down the command.

Ans:

```
command=
"/sw/apps/bioinfo/MAFFT/7.407/snowy/bin/mafft"  --retree 1 --clustalout --reorder "all_14_mts.fasta" > "lab5_all_14.clustal"
```

For the lrRNA case

```
"/sw/apps/bioinfo/MAFFT/7.407/snowy/bin/mafft"  --retree 1 --clustalout --reorder "all_10_lrRNA.fasta" > "multi_align_10_lrRNA.clustal"
```

Question 6. Which version of clustalw did you load?

Ans:

Using

```
module spider clustalw
```

I only found i version of clustalw: 2.1. That was the version loaded and used.

Question 7. Normally one of the sequences should stand out. Which one?

Ans: From the alignment visualisation, the A. jubatus sequence is the one that stands out by it being much longer than the other. It looks like a good chunk of 'extra' DNA has been included in the sequences for its lrRNA sequences.

See accompanying clustal file.

Question 8. Show your new alignment to a teaching assistant. If you cannot show it, submit the corresponding alignment file (.clustal).

Ans: See accompanying cluster file.

Question 9. Do you think that it was meaningful to align these 13 mitochondrial genomes? Would you remove some if you were to do it again? Which?

What do you see? Does it match your expectations after filling the feature table?

Ans:

Two things jumped out when viewing the multiple alignment of the whole mitochondrial genomes (plus the chloroplast genome of the flowering plant F. persica). 1) being the influence of the sequences for F. persica almost seeming to distort the whole alignment by its long length and its dissimilarity [by visual inspection] from the rest of the sequences. Although this should have been expected given that its a chloroplast genome and that chloroplast are evolutionary distinct from each other [if i remember correctly, the current consensus is that they originate from distinct endosymbiotic events between a pro-eukaryote and prokaryotes.]. And second: If i assume correctly that the sequences are stacked on top of each other relative to their distance/similarity to one another, they the stacking showed unexpected grouping between what i believe are actual closely related species. Examples of this is that the fruit fly being group nearer to the microb *M. brevicollis* then to the expected *D. stimulans.* I guess this stems from the combination of us aligning the whole genomes (plus a chloroplast genome) and not only coding genes/exomes that might have a higher degree of mutational constraint.

Regarding the choice of genomes we have aligned: I don't think I'm opposed to the choice of genomes selected here, even the flowering plast chloroplast because i can well imagine a situation where you would like to study the relative phylogenetic placing of all the included organisms and have F. persica as an outgrip representative sequence. But in that case maybe add a bit more plant sequences. What i might find problematic in problem 2d is that we choice to align whole genomes and not only focus on a concatenated set of core/highly conserved regions of the mitochondria, which would then give us less risk of having to deal with aligning divergent regions which in turn might detrimentally influence our final alignment.