lab4_questions_and_maj_answers

Course: Genomics & Bioinformatics

Lab: Lab4 - Annotate your mitochondrial genome

Author: Mohammed Al-Jaff

Step 1: Perform a first annotation with GeSeq

Question 1. How many of these features are found by GeSeq?

Ans: No features of the types 'intron' or 'exon' were seen. In total: 29 features were found. But only 7 tRNA genes were found (Gly, His, Ile, Met, Cys, Ala, Gln)

Step 2: Complete the annotation of tRNAs with tRNAscan

Question 2. How many tRNAs were identified? Were they already identified by GeSeq? Do the coordinates differ between the two softwares? If yes, write the two sets of coordinates down. Which do you think are the correct coordinates?

Ans: tRNAScan detected all 22 putative tRNA genes in the Congo mitochondrial genome given. As mentioned above, only qe

Predicted tRNA gene	tRNAScan coordinates	GeSeq endpoints	Disagreement
Gly (G)[TCC]	4106-4173	4106-4173	-
His (H)[GTG]	6253-6320	6253-6290	Yes
Ile (I)[GAT]	14846-14914	14846-14914	-
Met (M)[CAT]	14985-15052	14985-15052	-
Cys (C)[GCA]	16407-16341	16407-16359	Yes
Ala (A)[TGC]	16236 -16168	16235-16168	Yes
Gln (Q)[TTG]	14983 -14912	14983-14929	Yes

Step 3: Use web-based blast to confirm features boundaries

Question 3. *Is the gene complete and in one exon? If not, how does it look like? About the alignment: what is the first amino acid position of the reference ('Sbjct') which align well to your mitochondrial genome?*

Ans: For the COX2 gene in humans and its match in out Pongo mitochondria, the BLASTX search indeed resulted in "complete" single exon alignment where the first amino acid position is 1 from our subject/reference.

Further more: The match and alignment resulted in an ridiculously small E-value (good) and roughly 95 precent sequence identity. In short, the alignment was very good from the start to end positions of the alignment (apart from the expected divergence of some aa) As for the range, the alignment agreed with the start position of 1682 but the disagreed with the end position. The end position from the blast alignment being 2359 and for the GeSeq prediction being 2365 (ie two codon difference). As the instructions mention, this might be explained by "there might be a difference due to incorporating or not the stop codon" between GeSeq and BLAST

Question 4. *Which gene did you choose? Does it have one or several exons? Do the boundaries match with the GeSeq boundaries? If not, list the GeSeq and the blast boundaries. Does the alignment start at position 1 of the subject? If not, make an hypothesis concerning the location of the start codon in your own genome (in terms of protein sequences).*

Ans: I chose the gene ND1 from the coding sequence of the H. Sapiens reference: ND1 (subunit 1 of NADH dehydrogenase).

```
    - GeSeq boundaries:            13892..14845
    - BLASTX alignment boundaries: 13956..14843
```

Blastx and GeSeq boundaries did disagree and the subject startposition began at "23", meaning that BLAST somehow was not able to align tehe 22 amino acids reference protein.

To check which boundary made more sense, I translated the region inside the boundary from GeSeq into its protein sequence. Lo and behold, the sequence looked very similar in a.a. sequence composition and order to the first 22

```
>first 22 aa of Pongo ND1:
MPMINLLLLIMSILIAMAFLML
----------------------
MPMANLLLLIVPILIAMAFLML
>first 22 aa of Human reference ND1:
```

This makes me more likely to trust that the GeSeq boundary is indeed the correct one and that the BLAST alignment and boundaries are not correct.

Step 4: Use web-based blast to improve the annotation of rRNA

Question 5. Do the coordinates differ between GeSeq output and what blastn suggests?

Ans: My GeSeq files shows that the Pongo long rRNA region is split into 3 different features. Where the first two are rRNA2 fragments. The boundaries being:

```
RNR2-fragment: 12253..12650
RNR2-fragment: 12653..12965
RNR2:          13008..13686
```

What is interesting is that, the blast alignment between the Pongo mitochondria sequence with the human reference rRNA resulted in the alignment across and including all 4 fragments found in Pongo. The "subject" start position was at 1 and the dot product showed "perfect linearity". This makes me simply think that GeSeq didnt manage to detect those 3 regions as a single one. Also, 2 of the fragments are sub-500 in length indicating that GeSeq has fallen victim to it's known inadequces according to *"f GeSeq predicted the genes to be in several pieces; or if the genes predicted by GeSeq were very short (less than 500 base pairs would be short)."*

```
Alignment boundary from blast : 12323..13808
```

Due to the very good blast alignment, i changed the GeSeq boundaries to the ones given it instead and removed the 3 fragment parts as they were included.

Step 5: Draw a visual representation of your annotated mitochondria and identify unannotated regions

Question 6. Now, look at the output. Do you see regions devoid of annotated features? See the example below (Figure 1)

Ans: No regions of devoid annotation.... See attached pdf of OGDraw output.

Step 6: Look for open reading frames with ORF finder

Question 7 (alternative 2). What are the genes identified by ORFfinder? Do the results match with your previous results, e.g. from GeSeq? (e.g. in terms of boundaries)

Ans: ORF-Finder found 11 ORFs.

ORFs	**ORF Boundaries**	**SmartBlast Hit**	**GeSeq comparision**
ORF1	1..1551	cytochrome b [Pongo abelii]	COX1 1..1536

ORF11	complement (8262..8786)	NADH dehydrogenase subunit 6 [Pongo abelii]	ND6 (-)(8262..8786)
ORF4	8860..10050	cytochrome b [Pongo abelii]	CYTB 8860..10000

In terms of gene "identification", ORF+Smart blast resulted in similar genes as GeSeq, where here to we had some minor slight differences between their respective boundaries.

Step 7: A bit of programming

Task: Submit python script.

Ans: See attached script

```
q8_script.py
```