Running Exercise I – Hunting malaria genes

RunningExerciseI.tgz – The data file for the assignment.

Background. A common task for bioinformaticians is to merge data from several files into one. You have a fasta file (*malaria.fna*) with multiple DNA sequences. To know which protein sequence those DNA sequences relate to, you used a blastx-run (using the gene sequences as queries and the UniProt database as target). The results are in *malaria.blastx.tab*.

Your task. Insert the information about the protein description (hitDescription) from the blastx results to the fasta file's header and print the output.

The fasta file *malaria.fna* looks like:

```
>1_g    length=1143     scaffold=scaffold00001  strand=-
ATGGATATAAATAAGAAATATTTTGCTCAAGATAAATTGGAACATCACTATCACCA...
>2_g    length=531      scaffold=scaffold00001  strand=+
ATGTTAAAATTTGTAAATATGATTCAGTTATTATCAAGAGGTCACAAACTTGTGGAA...
...
```

Note, that the fields in the id-lines are tab-delimited.

The annotated file *malaria.blastx.tab* looks like:

```
1_g 1143 70 1140 P14223 369 13 369 +1 Fructose-bisphospha...
2_g 531 1 489 Q8IL24 172 1 159 +1 Putative uncharacterize...
...
```

Note, again, that the fields in the lines are tab-delimited.

The output file that your program produces: output.txt should look like:

```
>1_g length=1143 scaffold=scaffold00001 strand=- protein=Fructose...
ATGGATATAAATAAGAAATATTTTGCTCAAGATAAATTGGAACATCACTATCACCA...
>2_g length=531 scaffold=scaffold00001 strand=+ protein=Putative...
ATGTTAAAATTTGTAAATATGATTCAGTTATTATCAAGAGGTCACAAACTTGTGGAA...
...
```

- Note, that the fields in the id lines should be tab-delimited. Use the entire protein description (not included here due to lack of space).
- If the gene doesn't have a blastx hit, which is indicated by null in the blast-file (e.g., 12_g), it *should not be included* in the output file.

Your code should run like this:

```
python malaria.py malaria.fna malaria.blastx.tab output.txt
```

Name the program *malaria.py*. Compress this file with your output file (output.txt). The compressed file will be called *malariaDescription.fna.zip*

Upload *malariaDescription.fna.zip*. to Canvas (Assignments\RE I) by the deadline. Use only materials and modules studied in class (no Pandas).

Last notes:

As in all REs, remember to QC your code.

Deadline: 13/10 12:00 (noon).

Good luck!!!