HPBBM. Exercises R. Permafrost and stop codons programming practice.

Ramon Diaz-Uriarte*

2016-11-17 (Release: Rev: fe4c981)

1 Is there a procaryote stop codon in here?

Someone in your lab has a set of many sequences (here, 5000) of DNA from a bunch of samples collected from permafrost areas. There are some hypothesis about the types of organisms in those areas but, to make the story short, the idea is that as part of your research you want to separate those sequences between those that have the procaryote stop codons and those that don't and check:

- Are there differences in the length of the two types of sequences?
- Are there differences in the frequency of T between the two types of sequences?

All the files are in the directory `permafrost` (part of the compressed file `permafrost.zip`). What do we want to do?

- Get an idea of the length of the sequences and the frequency of Ts (number of Ts divided by length of sequence) for those sequences that have and do not have prokariotic with stop codons.

How exactly will we do this? As follows:

1. Read all files. In fact, **any** number of files, where each file is a sequence.
2. Write functions that, for any sequence, will return:
   - Its length (the number of characters)
   - The frequency of Ts (number of Ts divided by length of the sequence)
3. Then, keeping separately sequences that contain the stop codons from those that do not have stop codons, use `summary` to provide, well, a summary of
   - The length of the sequences.
   - Frequency of T (the total number of Ts divided by the length of the sequence)

These are some hints:

- We want to be able to read any number of files.
- We know the stop codons of prokaryotes are `stopCodons <- c("TAA", "TAG", "TGA")`. A good thing of that is that if I change my mind later, I can search for whatever.
- When reading the files, take a look at function `scan`.

*Dept. of Biochemistry, Universidad Autónoma de Madrid, Spain, http://ligarto.org/rdiaz, rdiaz02@gmail.com

- When checking if a sequence has a stop codon, any of the functions from the `grep` family can be useful. Use the help.
- You can really ace this exercise if, instead of using `sapply` you think about `vapply`. Again, the help is your friend.

To give you an example, these are the frequencies of Ts for the sequences with stop codon I get (the names you use might be different, of course)

```
summary(sapply(seqsStop, fnucl))

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.05556 0.23163 0.25000 0.25118 0.27037 0.46154
```

(you need to obtain also the lengths and similar statistics for those without stop codons, of course).

We are done!