Introduction to Unix shell - EXERCISES

- 2016/17 Part II BBS Bioinformatics
- 16 Jan 2017, 15:00-17:00
- Bioinformatics Training Room, Craik-Marshall Building, Downing Site
- Alexey Morgunov

Contents

Download and unpack the `text files` and the `human reference genome annotation`.

Exercises - Shakespeare

1. Check you have the `shakespeare.txt` file (all works of Shakespeare as text). Process it to output a list of words with frequency counts. Be careful not to count capitalised and non-capitalised words separately, and take care of the apostrophe!
2. Check how many times some country names are mentioned by Shakespeare. What are the most common words to co-occur in the same line with a country name? Filter your output for words shorter than four letters.
3. Find the most common bigrams Shakespeare uses. Trigrams?

Exercises - Human genome

1. How many genes are there in the `reference genome`? Don't forget to unpack the file.
2. How many transcripts does your favourite gene have, e.g. ENSG00000001461?
3. How many exons?
4. Produce a tab separated file with these columns: transcriptID, exon_number, exon_length.
5. Which exon is the longest?

Exercises - PDB

1. Extract the protein sequence from the PDB file `1A8Q.pdb` (ATOM instances).
2. Chech if the sequence from ATOM instances matches the one in SEQRES.

Exercises - More sed

1. Write a script that would combine the split lines in `split_lines.txt` using `sed`.

License

Many of the shell scripting exercises are taken from Linux Shell Scripting Tutorial (LSST) v2.0 under a CC-BY-NC-SA license.

This material is released under a CC-BY-NC-SA license 

.