Introduction to Genomics in Python - Exercises

Kevin Rue-Albrecht

Oxford Biomedical Data Science Training Programme

2020-05-06 (updated: 2020-05-06)

Yesterday ... ♫

- Write an algorithm to compute GC content
    - https://en.wikipedia.org/wiki/GC-content
    - http://rosalind.info/problems/gc
- Test data:
    - https://www.ensembl.org/Homo_sapiens/Gene/Sequence?db=core;g=ENSG00000172216
- Supply input (FASTA) and output (TXT) file names on the command line
    - Use the `argparse` module: https://docs.python.org/3/library/argparse.html
- Pseudocode

```
# import module
# initialise the command line parser
# add arguments to the command line parser
# parse the command line arguments
# check the validity of the arguments parsed
# use those arguments in your program
```

Genomic file format conversion exercise

- Copy the SAM file from shared/week2/ERR...
- Write a Python script to convert the SAM file to a BED file
    - Supply the SAM file name on the command line using –i or --input
    - Supply the BED file name on the command line using –o or –output
    - Format your output using F-strings
    - Provide a command line argument to pad the intervals in the bed file
    - Output a file with the coordinates of the sequenced fragments

Comparing Genomic Intervals

- Download 2 BED files from ENCODE
  - Different tracks for the same cell line
  - Same track for different cell lines
- Write a Python script to count the number of intervals that overlap between the two files
- Both files should be provided as command line arguments

Advanced

- Calculate the number of overlapping bases