Directory Parser & DNA Transcription Tool

Author: Haegi Oh (pronounced “Hayjee Oh”)

Introduction

It is undeniable that as every year passes, the volume of big data increases. Learning how to analyze big data is important because it provides crucial information for companies, researchers, and influential decision-makers (Cai, L. and Zhu, Y., 2015). Especially when it comes to biology, datasets - such as Genbank files from NCBI - can be extremely long.

In a real-world setting, it is important for one to be able to analyze large datasets like Genbank files to extract information of biological importance, such as FASTA sequences, so that they can be used for processes like transcription – an important step for building proteins.

While there are many useful tools out there to help aid specific needs, this research project will consist of three scripts that can parse a directory containing Genbank values, convert Genbank sequences to FASTA format, and transcribe these sequences to RNA.

Materials

- Bash shell-scripting in Unix environment
- Python3
- 40-50 truncated Genbank files with “.gbk” extension as input data
  - Inside directory called ”genomes”
  - **Note: only “bacillus” files contain sequences**

Methods

- Files
  - **directory.sh** - parse directory to copy specific files of one’s choice into new folder
  - **fasta.sh** – convert sequences from Genbank file of choice to FASTA format ***(file name and “ > output.txt” are typed into command line)***
  - **transcription.py** – transcribe DNA sequence to RNA
- Techniques
  - For loops
  - Grep
  - Regular expressions
  - Functions
- Usage
  - Use Terminal
  - **Note: Please have genomes folder and scripts all in *same working directory, including those that are copied into new_folder***

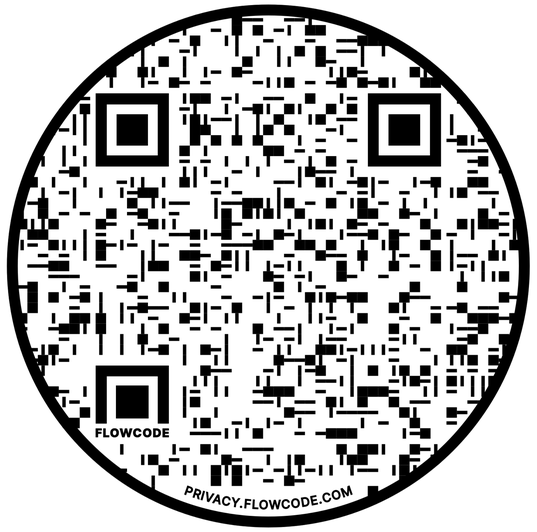
Results

directory.sh → new_folder → NC_004567_head.gbk

NC _004567_head.gbk → fasta.sh → output.txt

output.txt → transcription.py → DNA to RNA

Visual Example (QR Code)



Conclusion

- The scripts, when used in conjunction, are useful for analyzing large amounts of Genbank files to:
  - find specific files of interest and copy them into new directory
  - Extract DNA sequence from file of interest from new directory and turn it into FASTA format
  - Take this FASTA sequence and transcribe it into RNA

  Ultimately one could could use this result for doing more research on building proteins, or on RNA’s function as a blueprint for proteins.

Works Cited

- Cai, L. and Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal*, 14, p.2. DOI: http://doi.org/10.5334/dsj-2015-002
- Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. *Nucleic acids research*, *44*(D1), D67–D72. https://doi.org/10.1093/nar/gkv1276
- Toal, Ray. (n.d.). Introduction to Bash. Retrieved December 09, 2020, from https://cs.lmu.edu/~ray/notes/bash/