Ankit Jambusaria

Candidate for Bioinformatics Engineer, Algorithm

**Invitae Bioinformatics Exercise B.3**

The program can be executed using the following command:

python3 translate\_transcript\_to\_genomic\_coords.py --genome-mapping-file input\_file1.txt --transcript-processing-file input\_file2.txt --output output.txt

translate\_transcript\_to\_genomic\_coords.py is a script that reads in a genome mapping file and a transcript processing file which are given as command line arguments and translates input transcript coordinates to genomic coordinates using the position and CIGAR string. The (0-based) position is input as CHR1:3, and the CIGAR string is 8M7D6M2I2M11D7M.

**Assumptions:**

* The transcript is always mapped from genomic 5’ to 3’.
* The transcript and genomic coordinates are 0-based.
* The transcript processing file contains transcript ids where each transcript maps to a unique location on at most one chromosome.

|  |  |
| --- | --- |
| **Strengths** | **Weaknesses** |
| The user-friendly software allows for input files to be given as command line arguments. | Unit tests are needed to test small components of the software |
| Tests have been written into the program to check for the format of input files. | Functional tests are needed to ensure that components of the software operate correctly with each other |
| Tests have been written to verify the format of CIGAR strings | Calculating and storing the coordinate mappings for entire alignments could cause memory issues when processing a large number of queries. |
| The software generates the coordinate mappings between transcript coordinates and genomic coordinates across all transcript ids and stores it in the memory. This allows for quick retrieval of the genomic mapping for any transcript coordinate without the need to repeatedly process the CIGAR string. |  |
| The software has a structure which enables development of additional functionality and features. |  |