Dylan Shiramizu
3/13/2020
BME 160

FastQ Translator

Abstract:

FastQ formatting has become the most commonly used way of sharing sequencing data. Especially when it comes to storing biological sequences and their respective quality scores. Although there hasn't been one uniform way of formatting it, they all have their own unique niches to make them useful. Using the three different FASTQ standards, Sanger and SOXELA/Illumina, the FastQ Translator is able to convert quality scores between specified FASTQ Format file types. These quality scores can be used to determine sequence quality or to remove (end clip) or take out low-quality sequences. Quality scores are a useful tool when mapping/looking for genes. The scores can be seen as the error probability of each base. The higher the Q score, the higher chances that the base call will be correct. The Phred quality scores became the standard for DNA sequencing. A majority of the DNA sequenced during the Human Genome Project was processed by Phred. There are also different applications/software packages that use the quality scores and sequences to achieve different things.

Introduction To Problem:

Before the FastQ format became the default format, there was the FASTA format. The FastA Format allowed for sequences to be long without line wrapping. The

FastQ format, on the other hand, is an extension of FastA. The FastQ format gives the ability to store sequences with their corresponding quality score. So, FastQ is very useful. But, Since there have been many versions and changes to the way FASTQ formatting is read and done, a translator is required so that sequences and their quality scores are usable across all platforms. The problem is that since there are many ways of expressing data and there are many ways to encode quality mapping, so it should be able to switch from one standard to another. I.e. SOLEXA -> Sanger translation.

Description of Solution:

A solution to this is to make a program that translates the quality scores from one mapping format to another. FastQTranslator is a program that takes an input file, either in FastA or FastQ format, and translates it from four possible PHRED mappings, PHRED33,PHRED64, PHRED64(B offset), or PHRED64(SOLEXA) to either PHRED33 or PHRED64. THe program knows what the user wants through command line inputs. The program then looks at what PHRED it was given and what it needs to translate it to. It then finds a Qscore from Q(0,40) or Q(2,40) or Q(-5,40) depending on if the inputs were PHRED33 or PHRED64, PHRED64 with B off set or PHRED64 SOLEXA, respectively. Then it adjusts the quality scores to the desired mapping format. This new list of quality scores is used with the ASCII table to create the new quality mapping. The program prints to an outfile where each of the valid sequences entered will have new quality mapping to the desired mapping format. This program is helpful when switching from one PHRED mapping to another. This would be helpful if someone were to get a

file from NCBI and wanted to be able to use it for a different mapping than the file came with without having to remap the genes and give it a new quality mapping that fits your needs.

Discussion/ Evaluation:

The program takes in arguments from the command line and individually looks at every sequence. In certain parts, they might be reading them again redundantly but it runs and gives the correct desired output. Also, the management and storage of objects that accumulate throughout the program could be neater. The program will only translate sequences input with the correct format. Meaning that the sequence exists, the length of the sequence and quality mapping are equal, the ID of the sequence and the quality mapping are equal, the ID's start with the proper character ('>' for fastA, '@' and '+' for fastQ). Also, with parsing through FastQ, the quality lines are allowed to contain the @ character. This meant that if the program looked for @'s to find IDs it could accidentally pick up a quality line instead of an actual ID line. So, when the input file was a fastQ file the fastAreader just reads the file straight down as header, sequence, header2, qualityline. The program removes sequences that aren't valid while also raising an error to stderr. Also, it could have many more implementations such as telling the user whats wrong with it, check for only unique sequences or print out the list of quality scores.

Conclusion:

For future work regarding this program, it could use a lot of cleaning up. The formatting gets a little wild and some functions could be condensed but it runs. Thinking back at the program it can definitely be done in a much much nicer way than my implementation. The program can get really messy if you don't keep track of variables or name objects without thinking. I created a new class to be able to call from the main function to translate the sequences. This was a challenge but I'd be open to learn more and discover my future opportunities for research/work.

References

Peter J. A. Cock, Christopher J. Fields, Naohisa Goto, Michael L. Heuer, Peter M. Rice, The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants, *Nucleic Acids Research*, Volume 38, Issue 6, 1 April 2010, Pages 1767–1771, https://doi.org/10.1093/nar/gkp1137

"Phred Quality Score." *Wikipedia*, Wikimedia Foundation, 13 Dec. 2019, en.wikipedia.org/wiki/Phred_quality_score.

"FASTQ Format." *Wikipedia*, Wikimedia Foundation, 16 Dec. 2019, en.wikipedia.org/wiki/FASTQ_format#Quality.