Minimizer counting for GWAS and Genome Size Estimation

- [x] ~~Check if Jellyfish works with k-mer size sequences in FASTA / FASTQ [Kushagra]~~
- [ ] Skim through MSP Kmer counter / KMC3 as prospective code to build up on
- [ ] [Big] Build a code to read FASTA/FASTQ and generate the minimizer(w,k) in O(n) time [Souvadra]
- [ ] Parallelize the [Big] code

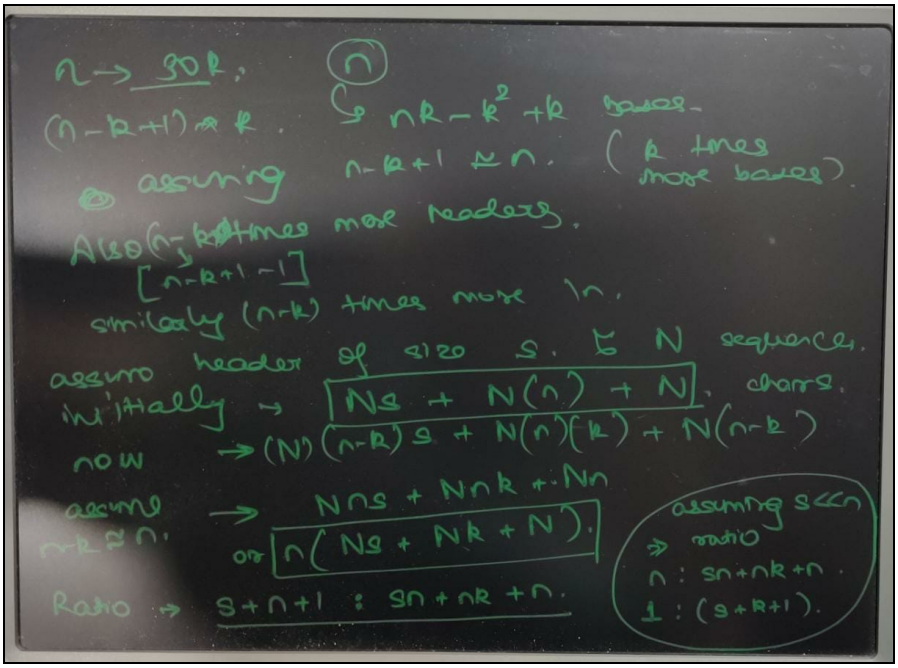
TASK 1 (Check if Jellyfish works with kmer size sequences) [KUSHAGRA]

Used 5098 SARS-CoV-2 sequences (~30k bases each). File size was 156 mb. On running Jellyfish count on this, got the sequences_counts.jf binary file (1.9 mb).

Command with execution details:

Created the kmers individually (separate sequences) called kmer.fasta using k=21. The size of this file was 5.1 gb (~33 times the original). File size scales up due to repetition in bases being stored with an approximate factor of 1: (s+k+1). Rough calculation below:

N is number of sequences, n is length of a sequence, k is kmer size chosen, s is number of characters in headers.



On running Jellyfish count on this:

```
kushagra@kushagra:~/Documents/IISc_Internship/Jellyfish/Time_Comp/5098_seq$ /usr/bin/time --verbose jellyfish count -m 21 -s 100M -t 4 -C kmer.fasta
        Command being timed: "jellyfish count -m 21 -s 100M -t 4 -C kmer.fasta"
        User time (seconds): 66.14
        System time (seconds): 2.58
        Percent of CPU this job got: 201%
        Elapsed (wall clock) time (h:mm:ss or m:ss): 0:34.02
        Average shared text size (kbytes): 0
        Average unshared data size (kbytes): 0
        Average stack size (kbytes): 0
        Average total size (kbytes): 0
        Maximum resident set size (kbytes): 516420
        Average resident set size (kbytes): 0
        Major (requiring I/O) page faults: 31
        Minor (reclaiming a frame) page faults: 127429
        Voluntary context switches: 56411
        Involuntary context switches: 401
        Swaps: 0
        File system inputs: 9918360
        File system outputs: 3624
        Socket messages sent: 0
        Socket messages received: 0
        Signals delivered: 0
        Page size (bytes): 4096
        Exit status: 0
```

Took nearly 4x-5x user time. Average resident size remained similar though.

To verify converted both the binaries to dumps.

```
/usr/bin/time --verbose jellyfish dump sequences_counts.jf > sequences_counts_dumps.fa
```

Then checked the md5sum for both the dumps.

```
kushagra@kushagra:~/Documents/IISc_Internship/Jellyfish/Time_Comp/5098_seq$ md5sum kmer_counts_dumps.fa
f0b518e4492021d512482d11b82ecdd9  kmer_counts_dumps.fa
kushagra@kushagra:~/Documents/IISc_Internship/Jellyfish/Time_Comp/5098_seq$ md5sum sequences_counts_dumps.fa
f0b518e4492021d512482d11b82ecdd9  sequences_counts_dumps.fa
```
