Genomic Signal Processing

Motivation

We hear about outbreaks of deadly diseases in different parts of the world all the time. Examples include Birds Flu, Ebola and MERS Coronavirus. It is interesting to think about what makes such viruses so unique and one way to investigate this is through genomic signal processing. The genome is the blueprint of life that contains all the information needed for all living beings to perform all their functions. Any genome consists of a unique long sequence of only four different proteins linked together. Such proteins are termed A, C, T and G. The genome of Coronavirus for example is given as [1]:

```
GATTTAAGTGAATAGCTTGGCTATCTCACTTCCCCTCGTTCTCTTGCAGAACTTTGATTTTAAC
GAACTTAAATAAAAGCCCTGTTGTTTAGCGTATCGTTGCACTTGTCTGGTGGGATTGTGG …
```

In genomic signal processing, we give each protein a number and consider the above sequence as an array of numbers (or practically a genomic “signal”) that can be processed using the available signal processing tools such as the Fourier transform.

Design Problem

Design a methodology to compare the Ebola virus to the MERS Coronavirus.

Design Input

- Genome sequences of both Ebola virus [2] and the MERS Coronavirus [3].
- Matlab code to converts genome sequence to a numerical genomic signal.

Design Output

- A report and documented Matlab code for a method to characterize and differentiate between the two viruses by their sequence variations, Fourier transform variations, spectrogram, etc.

Design Evaluation Criteria

- Qualitatively by comparing the spectra of their genomic signals.
- Quantitatively by stating the features where they exhibit largest differences.

References

[1] http://www.ncbi.nlm.nih.gov/nuccore/667489388?report=fasta&to=30119

[2] http://hgdownload.cse.ucsc.edu/goldenPath/eboVir3/bigZips/

[3] http://www.ncbi.nlm.nih.gov/genome/viruses/variation/MERS/