## Project Proposal

**Team members:** Tanqiu Liu, Wenjie Zhu

**Background:**

(1) CpG island Background

CpG sites are sites in DNA sequence where a cytosine ("C") nucleotide is followed by a guanine ("G") nucleotide[1]. Cytosine in CpG sites are usually methylated by DNA-methyltransferases. Methylated CpG sites affect the expression level of genes they related to and play an important role in gene regulation network in mammalian cells. CpG islands are regions (usually > 200 bp in length) of DNA sequence with high frequency of CpG sites. CpG island is usually associated with gene promoters and therefore becomes an important feature in gene/promoter prediction and epigenetic analysis.

(2) HMM & Viterbi

The Hidden Markov Model (HMM) is a statistical model for sequences of discrete symbols. HMM is defined by Alphabets, Sequence, i-th letter in sequence, and a set of sequence. In HMM the states in the machine are not directly visible but some output at certain states are observable. Each state has a probability distribution over possible output states[2].

Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states. for each intermediate state, until it reaches the end state. At each time only the most likely path leading to each state survives.

**Goal:**

Implement Viterbi algorithm to predict CpG islands within a DNA sequence.

**Data:**

We use human genome assembly hg38 with annotation to build the training sequence data and CpG island label for our project. The dataset can be downloaded from: http://hgdownload.soe.ucsc.edu/downloads.html

CpG annotation:

| bin | chrom | chromStart | chromEnd | name | length | cpgNum | gcNum | perCpg | perGc | obsExp |
|---|---|---|---|---|---|---|---|---|---|---|
| 585 | chr1 | 28735 | 29810 | CpG: 116 | 1075 | 116 | 787 | 21.6 | 73.2 | 0.83 |
| 586 | chr1 | 135124 | 135563 | CpG: 30 | 439 | 30 | 295 | 13.7 | 67.2 | 0.64 |
| 587 | chr1 | 327790 | 328229 | CpG: 29 | 439 | 29 | 295 | 13.2 | 67.2 | 0.62 |
| 588 | chr1 | 437151 | 438164 | CpG: 84 | 1013 | 84 | 734 | 16.6 | 72.5 | 0.64 |
| 588 | chr1 | 449273 | 450544 | CpG: 99 | 1271 | 99 | 777 | 15.6 | 61.1 | 0.84 |
| 589 | chr1 | 533219 | 534114 | CpG: 94 | 895 | 94 | 570 | 21 | 63.7 | 1.04 |
| 589 | chr1 | 544738 | 546649 | CpG: 171 | 1911 | 171 | 1405 | 17.9 | 73.5 | 0.67 |
| 590 | chr1 | 713984 | 714547 | CpG: 60 | 563 | 60 | 385 | 21.3 | 68.4 | 0.92 |

[1] https://en.wikipedia.org/wiki/CpG_site

[2] http://www.cs.ubbcluj.ro/~csatol/mach_learn/bemutato/Mate_Korosi_HMMpres.pdf