**Department of Computing and Information Systems**

**COMP 90016**

Workshop 3

Alignment is one of the core techniques to analyse sequencing data, as it is the first step to comparing sequenced DNA to a reference genome.

We have discussed various techniques to compute alignments in the lectures as well as their theoretical complexities.

In this workshop we are expanding our understanding to practical complexities.

1. Implement the Hamming-distance-aware best-match aligner from the lectures in Python. The reference is going to be the ecoli.fa from assignment 1, which you can either hard-code or pass as a command line argument to the program.   
   The program should compute the match with the lowest Hamming distance in the genome and write this as output to the command line.  
   For simplicity, let us ignore the double-stranded nature of DNA, and only match to the forward strand.
2. Use your program to align a single read to the E. coli reference. You can generate a read any way you wish – between 10 and 50 bases long.   
   Use the unix time command to measure the time it takes to align your read to the reference: i.e. *time python aligner.py ecoli.fa read.fa*
3. Discuss your findings in the group:
   1. Extrapolating this single read runtime, how much would you expect for 1 million reads?
   2. How do these values compare to the theoretical estimates from the lectures?
   3. What if we have to observe both strands of DNA?
   4. How about 100 million reads and a genome of the size of that of humans?
   5. How can we improve the runtime of this simple aligner?