CSC 314, Bioinformatics Lab #7:
Sequence-Based Pathogen identification

As a result of genomic sequencing, it has become possible to rapidly identify infectious diseases in an individual, which is particularly powerful in situations where standard diagnostic tools fail. In one example, a 14-year old boy was put into a medically induced coma after falling critically ill with a disease that could not be identified. However, by sequencing DNA found in blood samples from the child, researchers were able to identify the presence of the pathogen *Leptospira santarosai*. The appropriate treatment (which turned out to be the antibiotic penicillin) was then administered, and the boy's life was saved.

Press release: https://www.ucsf.edu/news/2014/06/114946/faster-dna-sleuthing-saves-critically-ill-boy
Article on diagnosis: http://www.nejm.org/doi/full/10.1056/NEJMoa1401268
Article on sequencing pipeline: http://genome.cshlp.org/content/early/2014/05/16/gr.171934.113

The purpose of this lab is to carry out an analysis in order to identify what pathogen may have infected a patient who cannot be diagnosed by traditional means. For simplicity, we will limit ourselves to a small number of potential pathogens (3 rather than 1000s), and a small number of sequencing data (105 fragments of length 50 bp rather than trillions of segments of length ~ 200 bp). The following files are provided:

1. samples.txt – the sequence data from the patient
2. human.txt – the human reference, consisting of a fragment of chromosome 15
3. ecoli.txt – subset of the *E.coli* genome (Sakai strain)
4. raoultella.txt – subset of the *Raoultella ornithinolytica* genome
5. zika.txt – subset of the Zika virus genome

E.coli are a species of bacteria, most of which are harmless but can cause food poisoning symptoms. The Sakai strain caused a minor outbreak of food poisoning in Japanese school children in 1996. *Raoultella ornithinolytica* is a bacteria that may cause skin lesions and flu-like symptoms, and is difficult to treat, though is rare in humans. The Zika virus, which is spread through mosquito bites, can cause fever, rash, joint pain, and conjunctivitis (pink eye).

Assignment: Complete the python notebook to identify the pathogen which may be infecting the patient, using the provided data files.