BMI/CS 576 Fall 2015
Homework #1

Prof. Colin Dewey

Due Thursday, October 1st, 2015 by 11:59pm

The goal of this assignment is to become more familiar with the algorithms for sequence assembly.

To turn in your assignment, copy all relevant files to the directory:
`/u/medinfo/handin/bmi576/hw1/USERNAME`
where `USERNAME` is your account name for the BMI network. Please note the homework policies posted at `http://www.biostat.wisc.edu/bmi576/hw.html`

1. (60 points) Write a program, `greedy_assemble`, that takes as input a set of read strings and uses the greedy fragment assembly algorithm to output a single superstring that contains all reads as substrings. You must use the graph-based (Hamiltonian path) version of the algorithm. We will assume that (1) we are assembling a single-stranded sequence and (2) that no read is a substring of any other read.

   Your program should be run via a wrapper shell script `greedy_assemble.sh` that is given a single command line argument specifying the name of a file containing the sequence reads, one read per line. The superstring should be printed to the standard output stream. A template wrapper script is available from:
   `http://www.biostat.wisc.edu/bmi576/hw/hw1/greedy_assemble.sh`

   For sanity checking (but not sufficient for complete testing), an example set of reads and the superstring produced from them by the greedy algorithm are posted at:
   `http://www.biostat.wisc.edu/bmi576/hw/hw1/`

   Here is an example of running the program, with the example set of reads:
   ```
   $ ./greedy_assemble.sh test.reads
   the_quick_brown_fox_jumps_over_the_lazy_dog
   ```

   For the purpose of making this algorithm deterministic, we must establish tiebreaking criteria for edges in the overlap graph that have the same weight. For two edges with the same weight, we will first choose the edge whose source node read is first in lexicographical order. If the source nodes are identical, then we choose the edge whose target node read is first in lexicographical order. For example, if $e_1 = ATCGGA \rightarrow$

$GGAT$ and $e_2 = ATCGGA \rightarrow GGAA$, we will attempt to use edge e_2 first because $GGAA < GGAT$ according to lexicographical order.

2. (10 points) Use your `greedy_assemble` program to assemble a small subset of the reads used to assemble the genome of an isolate of the Ebola virus, which caused a major epidemic in West Africa last year. Once correctly assembled, these reads form a short segment of the genome of this virus. To allow your assembler to succeed, the reads have been cleaned of errors and have have been oriented so that they all come from the same strand of the genome. Once you have assembled the genomic segment, use the BLAST web service to search the NCBI database of proteins with your assembled sequence. Which gene is contained within this genomic segment? Reads and BLAST instructions are provided at: `http://www.biostat.wisc.edu/bmi576/hw/hw1/`

3. (20 points) For the following strings, (i) give the $k = 3$ spectrum for the string, (ii) draw the SBH graph for the spectrum, (iii) give one Eulerian path and its corresponding string for the SBH graph, and (iv) show whether or not there exists an Eulerian path in the graph that corresponds to the original string.

   (a) `TACCGGACTTAGG`

   (b) `TATCGGATCGTTA`

4. (10 points) How would the SBH graph for a circular genome be different from that of a linear genome? Assume that there are no repetitive k-mers in the genomes.