

# BMI/CS 576 Fall 2009

## Homework #1

Prof. Colin Dewey

Due Wednesday, September 30th, 2009 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. You must submit a file named `README` to this directory, which gives directions on how to compile (if necessary) and run your programs. For each question below, the `README` file should list the files relevant to that question (e.g., code, other files with written answers). Please note the homework policies posted at <http://www.biostat.wisc.edu/bmi576/hw.html>

1. Write a program, `GreedyAssemble`, 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. Depending on the language you use, your program should be run from the command line with one of the following commands:

```
GreedyAssemble readsfile
java GreedyAssemble readsfile
python GreedyAssemble.py readsfile
perl GreedyAssemble.pl readsfile
```

where `readsfile` is the name of a file containing sequence reads, one read per line. The superstring should be printed to the standard output stream.

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.

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/hw1/>

2. The H1N1 virus genome consists of eight RNA segments (separate molecules). A set of (simulated) sequencing reads from one of these segments is provided on the course website. Use your GreedyAssemble program to assemble these reads into the full segment (if your program is correct, it should also assemble the segment correctly). Once you have assembled the segment, use the BLAST web service to search the NCBI database of nucleotide sequences with your assembled sequence. To which segment *number* does this sequence correspond? Reads and BLAST instructions are provided at: <http://www.biostat.wisc.edu/bmi576/hw1/>
3. 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. 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.