**Hamilton’s challenge**

In 1990’s when Tim Hamilton invented short-read sequencing, he had to convince Craig Venter that it works. He worked all night creating a simulation where he simulated the human DNA (at this point, it wasn’t sequenced yet), chopped it to very small, overlapping segments (=reads), and then aligned them. He showed that the alignment matched the simulated human genome. In this assignment, you are asked to replicate Hamilton’s work.

Specific tasks

1. Simulate a random DNA sequence (start with something small like *n*=100, but demonstrate your solution on something large like *n*=100,000; *n* should be a parameter in the program).

For example (with a sequence of *n*=18): ACGATCAGACTAGCTACG

1. Chop it to *n*x30 reads of length l=200 bases (*l* should also be a parameter in the program, again, start with something small. If you use *n*=100, then *l* can be 20).

For example (the above sequence of *n*=18 will have reads of *l*=4): ACGA, CGAT, GATC…

You can assume no errors in this procedure.

1. Provided a file with these reads sorted at random, align those reads based on their sequence similarity. Remember, you cannot use the DNA sequence that you simulated – it is unknown. Rely only on the similarity of the sequences.

For example (with reads of *l*=4):

ACGA,

 CGAT,

GATC…

Notice the vertical alignment; you can be certain that there is a CGA in the original sequence.

1. Predict the DNA sequence and calculate the accuracy of your prediction by comparing it with your simulated sequence.

For example, original sequence: ACGATCAGACTAGCTACG

Reconstructed sequence: ACGATCAGACTAGCTAC**T**



Almost! 94.44% accuracy!

1. Write a short summary of how the results change with reads of different lengths.