COMP.4290/5510 HW #2 DUE ON 2/7/22

G-C skew is examined in the textbook.

One can take this further into oligomers (multiple nucleotides).

Consider a genome sequence G, partitioned into 50 equal-length segments. In each segment, the frequency of each k-mer *m* is obtained. Let *m*\* be the reverse complement of *m*.

Now, define

*Si*= *log*(*fm/fm\**),

as a measure of the skew. If frequency is equal to 0, it is set to 1.

Write a program (preferably in Python) that produce a table of Si for each k = 1, 2, and 3. Each k will have 4k rows and 50 columns. For the genome sequence, use E.coli sequence at <https://www.ncbi.nlm.nih.gov/nuccore/U00096.2?report=fasta>.

You are **not** allowed to use any programming packages for programs (in subsequent HWs, as well). For example, you are welcome to use Jupyter Notebook, but your submitted code has to be a Python code that can run on IDLE3. A tutorial on Python is available at <https://stepik.org/course/46745/promo>.

**What to submit:**

Submit a **report** that includes:

1. A brief re-statement of the problem.
2. Results from the program.
3. A discussion on what you found from the results. In particular, try arranging rows of tables such that rows with negative or positive column values in adjacent rows. Can you find any pattern from the tables. A plot of a table can be included, too.
4. In the appendix, program source code. A source code is not a report, and is a supporting document.