BMIG 5003 Computational Methods for Biomedical Informatics
Horacio Gómez-Acevedo, PhD
Fall 2023
Midterm
Due day: Nov 6th, 12:00 pm CST

Submission:
All your answers should be placed in the Assigned Box Folder. I will accept only Python and text files, NOT jupyter notebooks. Feel free to use any following modules: `numpy, pandas, plotly, scikit-learn, statsmodels, re, and regex`.

1. Develop a program in Python that will read the file `sequences.txt` (provided with this assignment), and perform the following tasks

   (a) It finds the number of occurrences of one or more `N`'s for each line using regular expressions.

   (b) It finds the number of periods in the **whole sequence** (i.e. without break lines), and prints out their type. Namely, it prints out

   ```
   There are x '.'
   There are y '..'
   There are z '...'
   ...
   ```

   (c) It finds out the number of occurrences of `CG`'s in the whole sequence, and prints out the number or such occurrences.

   (d) The program will finally generate a file named `sequences_CG.txt` similar to the `sequences.txt` but have `CG`'s underscored (i.e., if the original file has ACGT the new file will have $\text{A}\underline{\text{CG}}\text{T}$ instead).

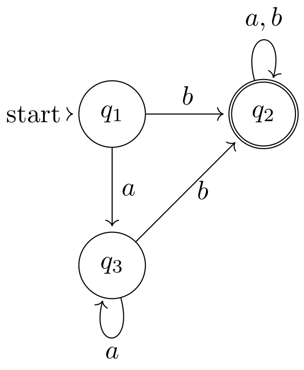
2. Consider the automata M_1 and M_2. A concatenation of those automata $M_1 + M_2$ is defined by the rule: All the accepting strings in M_1 are concatenated to the accepting strings of M_2. Write a Python program that that will implement the automaton $M_1 + M_2$. More specifically, the program should be properly documented that will

   (a) ask the user for a sequence of a's and b's followed by a sequence of 0's and 1's.

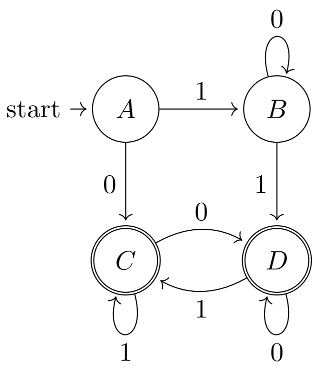
   (b) verify that the input sequence does not contain other characters,

   (c) print out whether the sequence is accepted or rejected, and

   (d) stop when the word 'exit' is typed.



Automaton M_1



Automaton M_2

3. Compose a Python program that asks the user for two non-negative integers (m, n), and a number $p \in (0, 1)$. The program will build a Boolean matrix $A = (a_{rs})$ of type $m \times n$ in which the a_{rs} is T with probability p and F with probability $1 - p$. For reproducibility, set the seed of your random number generator to 500323.

4. Use the `insurance.csv` data set to fit a multi-linear regression model to predict the `charges` based on the variables `age` and `bmi`.

   (a) If you select patients on the northwest region only, what are the values of your β's.

   (b) What is the R^2 of your model?

   (c) If your `bmi` increases by 1 unit while keeping the `age` constant, what would be the expected increase in `charges` according to your model?

   (d) Predict the cost for a person in this region whose age is 45.5 and bmi is 24.9.

   (e) Now consider another region (e.g., northeast) and describe changes in your previous prediction (if any).