Due Date: Thursday 28 March, 4pm

Please submit a **PDF** by email to: teaching@biods.org with subject “341 Assignment 1”. Make sure that your name and ID number are on the PDF called: YourName_assignment1.pdf

What you hand in should be all your own work.

Please note the following

- There are 22 marks in this assignment but the **maximum** you can get is **10**.
- Marks outside the set $\{0, 10\}$ are very uncommon for question 3.
- Unless we receive a correct solution to question 3, the solution will not be made available within this course.

Questions

1. (a) Assume that a gene is a word over the alphabet $\{A, T, C, G\}$. Assume also that a protein is unambiguously defined by a finite set of genes and different sets of genes define different proteins. Is the set of all proteins finite or infinite? If infinite, is the set countable or uncountable? Justify your answers formally (with proofs). (**3 marks**)

$\{A, T, C, G\}^*$ is the set of genes. Let P be the set of proteins. Then we can write P is a subset of the power set of $\{A, T, C, G\}^*$ such that all sets in P are finite:

$$P = \{X | X \subset \{A, T, C, G\}^*, |X| < \mathbb{N}\} \subset \mathcal{P}(\{A, T, C, G\}^*)$$

The set of genes $\{A, T, C, G\}^*$ is countable, because following bijection exists:

$$f : \{A, T, C, G\}^* \to \mathbb{N} \text{ with } f(g) = \sum_{i=0}^{n-1} 4^i g_{n-i},$$

where $g_1 g_2 \dots g_n$ represents a gene g :

$$g_i = \begin{cases} 1 & \text{if } g \text{ has an } A \text{ at position } i \\ 2 & \text{if } g \text{ has a } T \text{ at position } i \\ 3 & \text{if } g \text{ has a } C \text{ at position } i \\ 4 & \text{if } g \text{ has a } G \text{ at position } i \end{cases}$$

f maps all elements of $\{A, C, G, T\}^*$ bijectively to $\mathbb{N}$ (for brevity, we skip proving this):

$$\begin{array}{lll} \lambda \to 0 & AA \to 5 & AAA \to 21 \\ A \to 1 & AC \to 6 & AAC \to 22 \\ C \to 2 & AG \to 7 & AAG \to 23 \\ G \to 3 & AT \to 8 & AAT \to 24 \\ T \to 4 & CA \to 9 & ACA \to 25 \\ & \dots & \dots \end{array}$$

f gives us an order of the genes. We can now define a bijection h from the set of proteins to the natural numbers. Therefore, we order assign all proteins a natural number:

The empty set is assigned to 0 and $\{A\}$ to 1. Now we define the following assignments of proteins to natural numbers, that is our new bijection h, recursively: For each following gene g in the list determined by f, we add the protein $\{g\}$ that only consists of this gene and assign it to the next natural number k (that is not covered by h already). Then, for all l proteins we already listed before, we add the union of this protein with $\{g\}$ to our list of proteins. These new proteins then are assigned the numbers from $k+1$ to $k+l$. Below one can see an illustration of this recursively defined function h:

$$\begin{array}{ccc}
\emptyset \to 0 & \{AA\} \to 17 & \{AA, AC\} \to 49 \\
\{A\} \to 1 & \{A, AA\} \to 18 & \{A, AA, AC\} \to 50 \\
\{C\} \to 2 & \{C, AA\} \to 19 & \{A, C, AA, AC\} \to 51 \\
\{A, C\} \to 3 & \{A, C, AA\} \to 20 & \dots \\
\{G\} \to 4 & \dots & \{AG\} \to 65 \\
\{A, G\} \to 5 & \{AC\} \to 33 & \{A, AG\} \to 66 \\
\{C, G\} \to 6 & \{A, AC\} \to 34 & \{C, AG\} \to 67 \\
\{A, C, G\} \to 7 & \dots & \{A, C, AG\} \to 68 \\
\dots & \dots & \dots
\end{array}$$

Since this function is a bijection from the set of proteins to $\mathbb{N}$, the set of proteins is countably infinite.

(b) Consider F, the set of all functions $f : \{0,1,2\}^* \to \{0,1,2\}^*$ mapping from ternary numbers to ternary numbers. Is F countable or uncountable? Justify your answer formally. (**3 marks**)

Solution:

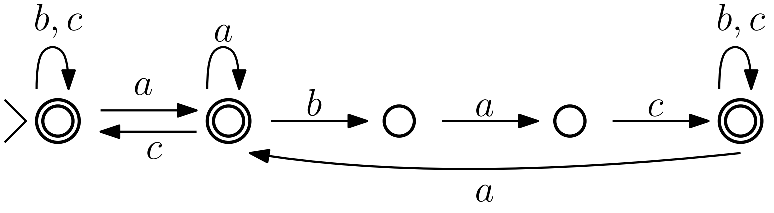
F is uncountable.

For proving that F is uncountable we will use Cantor's diagonal argument. For contradiction, we assume that F is countable. Then we can write the elements of F as list $f_0, f_1, f_2, f_3, \dots$. Let f^* be a function from $\{0,1,2\}^*$ to $\{0,1,2\}^*$ with $f^*(t) = f_t(t) + 1$ where we use the addition defined for ternary numbers. According to the definition of f^* it differs from every f_t of our list. Therefore, we found an element of F that is not covered by the list. This is a contradiction to our assumption, that F is countable. Therefore, F is uncountable.

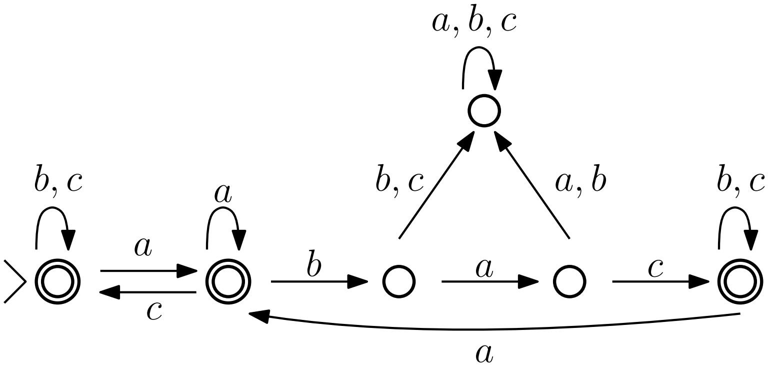
2. Design an automaton for each of the following languages over the alphabet $\Sigma = \{a, b, c\}$. If your automaton is non-deterministic, provide a DFA for the same language. Explain why your DFA recognises the same language. If you decide to use the automaton determinisation algorithm, list all steps of the algorithm you make. (**2 marks each**):

(a) The set of strings in which every ab is immediately followed by a ac.

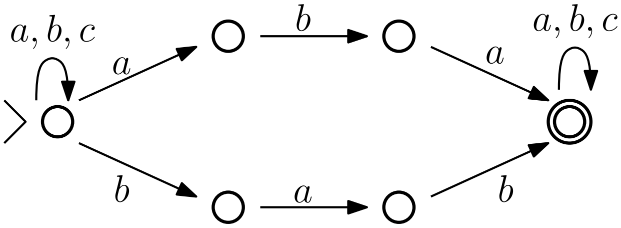
NFA:



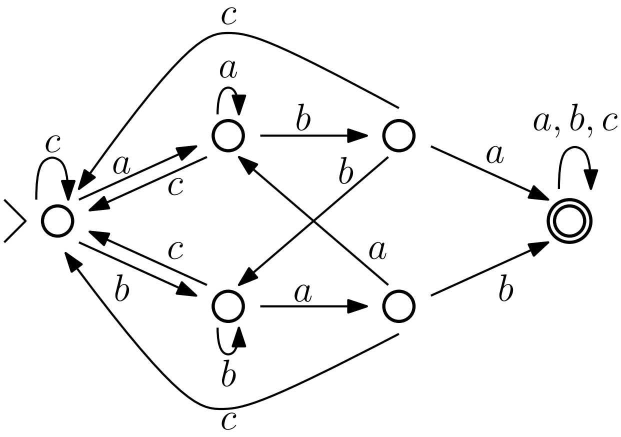
DFA:



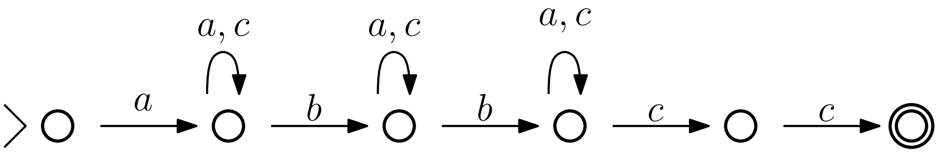
(b) The set of strings that contain *aba* or *bab* (or both) as a substring.
NFA:



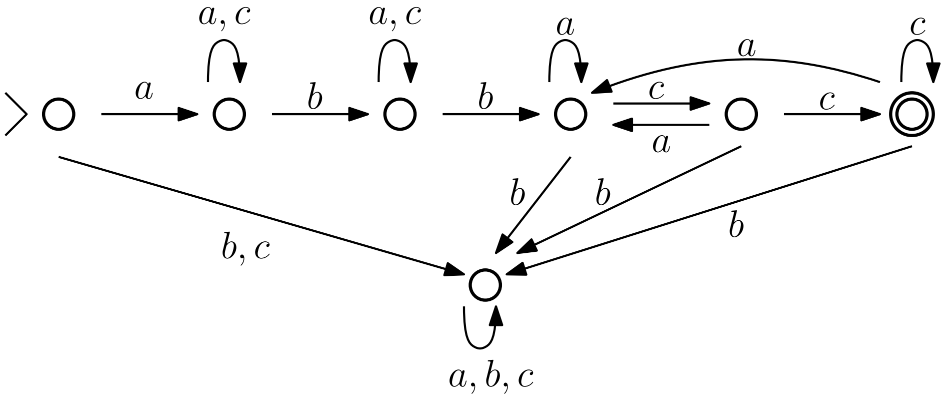
DFA:



(c) The set of strings that begin with *a*, contain exactly two *b*'s and end with *cc*.
NFA:



DFA:



3. Prove that the Pumping Lemma is not a sufficient condition for a language to be automatic, that is, the Pumping Lemma is not an "if and only if" statement. (**10 marks**)