CS4120/4121/5120/5121—Spring 2016
Homework 1
Lexical Analysis
Due: Friday, February 5, 11:59PM

0 Updates

- In Problem 1. the alphabet is `A`, `C`, `G`, `U` (RNA encoding), not `A`, `C`, `G`, `T` (DNA encoding).
- Fixed typo in Problem 3.: the *file* is denoted by a letter and precedes the *rank*, which is denoted by a number.

1 Instructions

1.1 Partners

You may work alone or with *one* partner on this assignment. But remember that the course staff is happy to help with problems you run into. Use Piazza for questions, attend office hours, or set up meetings with any course staff member for help.

1.2 Homework structure

There are two parts of the homework. The first part is required of all students. The second part is required of students taking CS5120, but those enrolled in CS4120 are welcome to try it for good **HARMA**.

1.3 Tips

You may find the Dot and Graphviz packages helpful for generating drawings of DFAs and other graphs. You can get these packages for multiple OSes from the Graphviz download page. An example of a DFA drawn using Graphviz may be useful.

2 Problems

1. Design of finite automata

Living cells on earth encode their genetic information in a chemical code made from repetitions of four basic compounds abbreviated as `A`, `C`, `G`, and `U`. These constituents are arranged in long linear sequences of triplets, e.g., `GCU` and `UGA`, referred to as *codons*.

The meaning of a short subsequence of a genetic encoding is often not immediately clear, since there are different *reading frames* in which to interpret the information. For instance, the partial sequence `CUUCUCCGCAAUUUAC` might specify any of the following codons, where question marks indicate missing data:

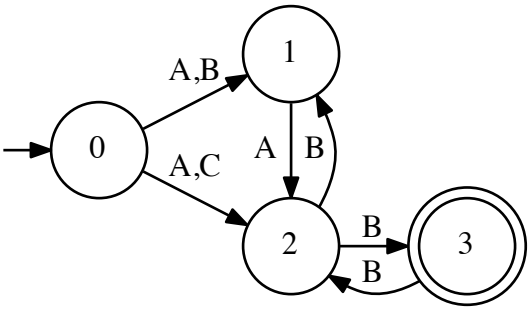
- `CUU CUC CGC AAU UUA C??`
- `?CU UCU CCG CAA UUU AC?`
- `??C UUC UCC GCA AUU UAC`

Certain codons have special meanings that resolve the reading frame ambiguity. The *start codon* `GUG` denotes the beginning of a gene-encoding region. One of the three *termination codons* `UAG`, `UAA`, and `UGA` marks the end of such a region.

A commonly asked question is whether a partial sequence acquired from a biological sample contains a complete gene-coding region under some reading frame. Design a nondeterministic finite automaton that accepts such a subsequence. Ensure that the start and termination codons are correctly aligned.

2. Determinization of automata

Construct a deterministic version of the following nondeterministic finite automaton. Make sure to indicate the initial and terminal states. Label each DFA state with the set of NFA states to which it corresponds.



3. Regular expressions

A popular way of encoding chess games is the *Portable Game Notation* (PGN), which is essentially a numbered sequence of moves in *algebraic notation*[1] with optional comments between moves. A (simplified) format description of PGN is given by the following set of rules:

1. **Game**: A game consists of a sequence of turns and ends with one of these three result identifiers: `1-0`, `0-1`, or `1/2-1/2`.
2. **Turns**: Each turn begins with the turn number in decimal, followed by a period. The turn number cannot begin with `0`. Following the period are one or two moves, one move per player. A move can be a normal move or castling.
3. **Normal moves**: A normal move is given by the piece that moved, an optional `x` if there was a capture, and the coordinate of the destination square of the move. Each coordinate is a letter for the *file* (i.e., the column), followed by a digit for the *rank* (i.e., the row). Pieces are denoted as `K` for king, `Q` for queen, `R` for rook, `B` for bishop, `N` for knight, and the empty string for pawns.

[1] Not to be confused with algebraic data types. See Wikipedia for more information.

4. **Disambiguating pieces**: In rare occasions, two or more identical pieces can move to the same destination. In this case, one or both of the file and the rank indicating the departure square is inserted after the piece identifier.
5. **Castling**: Kingside castling is denoted by `O-O`, and queenside castling is indicated by `O-O-O`.
6. **Check and checkmate**: A plus sign + is appended to a checking move. A hash sign # is appended to a checkmating move.
7. **Comments**: An optional comment may be inserted after a move. They contain arbitrary descriptions between curly braces. Comments do not nest.

For instance, here is a PGN encoding of a game between Robert Fischer and Boris Spassky:

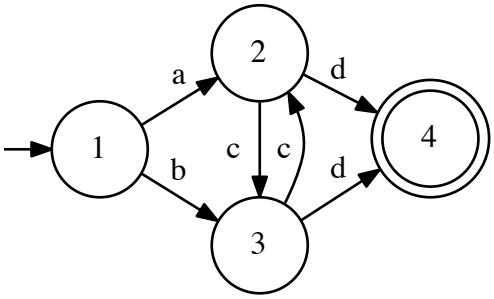
```
1. e4 e5 2. Nf3 Nc6 3. Bb5 {This opening is called the Ruy Lopez.} a6
4. Ba4 Nf6 5. O-O Be7 6. Re1 b5 7. Bb3 d6 8. c3 O-O 9. h3 Nb8 10. d4
Nbd7 11. c4 c6 12. cxb5 axb5 13. Nc3 Bb7 14. Bg5 b4 15. Nb1 h6 16. Bh4
c5 17. dxe5 Nxe4 18. Bxe7 Qxe7 19. exd6 Qf6 20. Nbd2 Nxd6 21. Nc4 Nxc4
22. Bxc4 Nb6 23. Ne5 Rae8 24. Bxf7+ Rxf7 25. Nxf7 Rxe1+ 26. Qxe1 Kxf7
27. Qe3 Qg5 28. Qxg5 hxg5 29. b3 Ke6 30. a3 Kd6 31. axb4 cxb4 32. Ra5
Nd5 33. f3 Bc8 34. Kf2 Bf5 35. Ra7 g6 36. Ra6+ Kc5 37. Ke1 Nf4 38. g3
Nxh3 39. Kd2 Kb5 40. Rd6 Kc5 41. Ra6 Nf2 42. g4 Bd3 43. Re6 1/2-1/2
```

Write a regular expression that describes a game of chess in the PGN format as closely as possible using the regular expression syntax discussed in class. Spaces and newlines can be ignored. To avoid writing down one giant regular expression, you can define parts of it as smaller, named regular expressions. (Be careful to ensure that your language is still regular.)

3 Problem for CS5120

4. DFA simplification

Simplify the following DFA using the method presented in class. Show the minimized version, as well as any intermediate steps you took.



4 Submission

Submit your solution as a PDF file on CMS. This file should contain your name, your NetID, all known issues you have with your solution, and the names of anyone you have discussed the homework with.