Project2: Covid-19 evolution and spread

Electronic turnin (code and pdf of your paper) due **Monday March 30 (11:59pm)**
Hard copy of your paper due **Tuesday March 31 (at the start of class)**

Your report should be up to 6 pages. The report will contain

- an **Abstract** that summarizes the purpose and main findings of your paper
- an **Introduction** that introduces important terms, concepts and definitions
- a **Methods & Results** section that contains
    - Figures with captions that are self-explanatory
    - Explanation of your Methods explaining how you generated each figure and any associated calculations
    - Results text that summarizes each figure that explains and interprets that figure
- a **Discussion & Conclusions** section that interprets and summarizes overall findings
- a **References** section that lists your citations.
- a **Contributions Statement** that describes each team members contribution
- **Code turnin:** You will turn in code on Learn. You do NOT need to print out code.

Follow the IEEE format for references (included in the LaTeX download). You can use Overleaf to collaboratively write your papers in LaTex. Check your grammar and spelling (Grammarly is free). Avoid long, run on sentences. Make this paper easy to read. See resources to be posted on Learn for scientific paper writing. A grading rubric will also be posted online.

Your goals for Project 2 are to explain how concepts from complex adaptive systems can aid understanding of the evolution and spread of SARS-CoV-2 (Covid-19). You will address two questions, and build and analyze a model of evolutionary change and/or epidemic spread.

- **Part 1: How can neutral networks help us to understand the evolution of covid-19?**
- **Part 2: How can cellular automata help us to understand epidemic spread of Covid-19?**
- **Part 3: Build a model that gives new insight into the evolution and/or spread of Covid-19**

<u>The Details</u>

- **Abstract (~1/2 page)**
    - State the purpose of the paper and motivates why it is important
    - Summarize key results
    - Use proper grammar and spelling in clear and easily understood text. No convoluted or run-on sentences.
    - Note: References are not needed in the abstract.
- **Introduction (!~1 page)**
  Define the following terms and explain how they relate to the evolution and spread of covid-19.
    - Top down vs bottom up causality
    - Logistic map (and chaotic population dynamics)
    - Frozen accidents
    - Measures of complexity
    - Neutral networks
    - Epidemic spread
    - Include 1 -2 paragraphs that summarize the purpose, methods and approach of your paper.
    - Include citations of relevant references. Uses proper grammar and spelling in clear and easily understood text. No convoluted or run-on sentences.

- **Methods & Results**

(Note, relevant references will be posted on Learn)

Part 1: How can neutral networks help us to understand the past and potential future evolution of covid-19?

Your explanation should include answers to the following questions:

1a. Assuming the covid-19 RNA is 30,000 nucleotides long, how many RNA strands of the same length are exactly 1 mutation away?

Approximately what fraction of the possible RNA strands have the same phenotype as the original? (Remember, there are 64 codons that redundantly encode 20 amino acids, see the wheel below). Use this definition: Nucleotide substitutions that encode the same amino acid are *neutral*, and those that encode different amino acids are *not neutral*.

1b. Now focus on the mutations that encode the "Spike protein" that enables binding and entry into mammalian cells and likely elicits an immune response (Li, 2016). Wan et al. (2020) show sequences of beta-coronaviruses, the group of coronaviruses that infect bats, and cause SARS, MERS and covid-19. They show 5 amino acids (encoded by 15 nucleotides) that are functionally different and vary between human SARS-2002 and covid-19. (4 of those 6 are also different in bat-SARS-2013 and 2 are different in civet-SARS-2002.)

RYYYLNYNYTTY
RYYYLNYKYTSY
RYYSFNYNYTNY
NYYLFNYQQTNY (assume the first N mutation is not important)

Note: the 6 black letters represent amino acids that don't change (YY NY T-Y) and are here to show alignment. The relevant amino acids

SARS-2002	R	Y	L	N	Y	T
Civet-2002	R	Y	L	K	Y	S
Bat-2013	R	S	F	N	Y	N
Covid-2019	N	L	F	Q	Q	N
position	1	2	3	4	5	6

Calculate the total number of possible combinations of nucleotides and the number of possible amino acids from these 15 nucleotides and 5 amino acids.

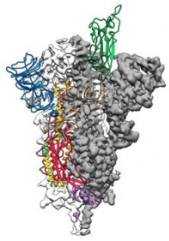
Calculate the number of genomes 1, 2, 3, all the way up to 15 mutations away from a length 15 genome. Show these in some sort of figure. A log scale may be helpful.

Calculate the minimum number of mutations required for the current spike protein to become equivalent to the SARS-2002 spike protein that was much more lethal. Assume that Bedford's blog is correct, and on average 1 of the 30,000 nucleotides mutates per transmission, and that is on average once per 7 days. Ignore neutral networks for a moment, and predict how likely the virus is to generate any one of the 5 mutations required to revert back to the SARS spike protein. Comment on how one

would calculate the likelihood that all of those mutations would happen at once, and how long that might take.

1d. Now make assumptions about the neutral network: any silent mutation (nucleotide change that doesn't change the amino acid) is neutral in any place in the genome. For the 6 nucleotides for which we have seen functional variations, assume that some small *b*% of non-silent mutations (that change the amino acid) are neutral or beneficial and will persist in the population. Assume (1-*b*)% of non-silent mutations are deleterious and will not persist in the population. Note that there are 3 different amino acids in positions 2, 4 and 6, so in those cases 3/20 (15%) of amino acids could reasonably assumed to be beneficial or neutral. Comment on which of the observed variants provide a neutral path from covid-19 back to SARS-2002. Simulate this neutral network and use it to explain the new likelihood and expected time to observe the original SARS-2002 spike-protein. Explain any assumptions you have made.

1e. Use the neutral network from part 1d or another one to hypothesize how neutral networks may affect covid-19 evolution. You could build other relevant neutral networks based on coding regions of proteins from MERS or pangolin-SARS, or the hypothesized 2 strains of covid-19 that may be circulating.

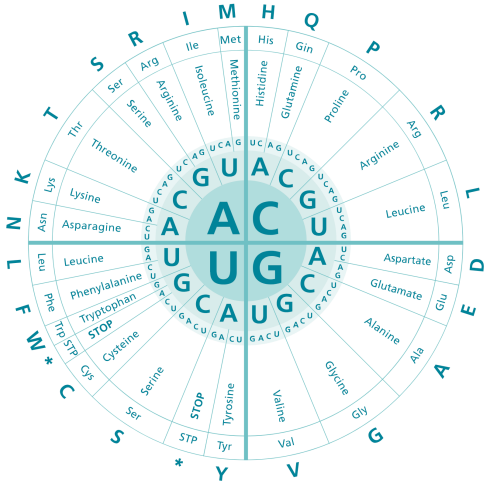


```
Human-SARS-2002   LAWNTRNIDA TSTGNYNYKY RYLRHGKLRP FERDISNVPF SPDGKPCTP-P ALNCYWPLND 480
Civet-SARS-2002   LAWNTRNIDA TSTGNYNYKY RYLRHGKLRP FERDISNVPF SPDGKPCTP-P ALNCYWPLKD 493
Bat-SARS-2013     LAWNTRNIDA TQTGNYNYKY RSLRHGKLRP FERDISNVPF SPDGKPCTP-P AFNCYWPLND 493
2019-nCoV         IAWNSNNLDS KVGGNYNYLY RLFRKSNLKP FERDISTEIY QAGSTPCNGVE GFNCYFPLQS 494
                  :***:.*:*: .  ***** * * :*:.:*:* ******.  : . ...**.    .:***:**:.

Human-SARS-2002   YGFYTTTGIG YQPYRVVVLS FELLNAPATV CGPKL 515
Civet-SARS-2002   YGFYTTSGIG YQPYRVVVLS FELLNAPATV CGPKL 528
Bat-SARS-2013     YGFYITNGIG YQPYRVVVLS FELLNAPATV CGPKL 528
2019-nCoV         YGFQPTNGVG YQPYRVVVLS FELLHAPATV CGPKK 529
                  ***  *.*:* ********** ****:***** ****
```

https://www.livescience.com/coronavirus-spike-protein-structure.html

This is the 3D atomic scale map or molecular structure of the SARS-2-CoV protein "spike" which the virus uses to invade human cells.
(Image: © Jason McLellan/Univ. of Texas at Austin)



Part 2: What can we understand about the spread of Covid-19?

2a: Build a simple 3 state, 2-dimensional CA with SIR dynamics to model epidemic spread. First use deterministic rules and a neighborhood size of 1 (use the Moore neighborhood which includes diagonals so there are 9 neighbors). Explain the rules you chose to determine infection and recovery. Explain how many rules there are, and how many rule tables are possible given the constraint that the order of states must be Susceptible →Infected →Recovered. Recovered cells never change state.
Choose some reasonable size CA and simulate spread assuming an initial configuration in which all cells are in state S, except one cell in state I.

A CA is not a great model of disease spread in a normal population, but is a better (still imperfect) model of a population that does not move (i.e., in a quarantined city, cruise ship or nursing home). Assuming such a scenario, you should model boundaries that do NOT wrap around in a torus. The cells on the edge will have fewer neighbors. Handle that case.

Visualize your CA with states S, I, and R in colors Green, Red, and Blue, respectively. Show a screenshot of your CA at 3-4 time points that illustrate how the disease spreads.

2b. Change your rules from the deterministic ones above so that the probability of transitioning from $P_{S\to I}$ depends on the number of surrounding cells that are in state S, and set $P_{I\to R}$ to be a fixed probability at each time step (this will simulate a Poisson process for recovery).

Add a second variant of the disease with new probabilities $P_{S\to I'}$ and $P_{I'\to R'}$. Show state I' in Orange and R' in Black. Initially set $P_{S\to I'}$ and $P_{I'\to R'}$ to values chosen at uniform random in the interval [0,1]. Use a GA to evolve transition probabilities to keep the 2 variants of the disease in the population at equilibrium (meaning, at the end of the simulation, the number that have recovered from each variant are approximately equal).

Show a screenshot of your CA at 3-4 time points that illustrate how the disease spreads.

Part 3: Choose your own adventure: **Build a model that gives new insight into the evolution and/or spread of Covid-19.** You can expand the work you did on the neutral network in part 1, or the CA (and GA if you'd like) in part 2, or you may choose some other approach. The only requirement is that you use tools or concepts from class and model something relevant to understanding the evolution and/or epidemic spread of covid-19.

Be creative. Cite relevant research where possible. Make "reasonable assumptions" when there are no clear answers, but try to make your analysis relevant to the current pandemic. Do a careful analysis and explain your motivation and results clearly. Provide a clear figure or other succinct presentation of your results, an explanation of your Methods and interpretation of your results.

- Discussion & Conclusions (2 – 4 paragraphs)
    - Summarize the main findings from Parts 1, 2 and 3.
    - Discuss the implications of your analysis for understanding the evolution/spread of covid-19.
    - Refer to your figures and analysis to back up your conclusions.
    - Suggest further steps that would expand your analysis to make other relevant predictions or

explanations

- References
    - List of references cited in the text and any downloaded code used to generate results. Use the references format shown in the IEEE template.
    - A web link is sufficient to cite code. **Failure to cite sources or code will result in a 0 for the entire Project.**
    - Paraphrase and re-interpret any source you cite in your own words. You may also cite videos for this class paper. Remember that Wikipedia, blogs and YouTube videos are generally inappropriate to cite for publications, but you can cite them for this project.
    - You must cite all sources and all code that you use that you did not write. Failure to cite code or sources constitutes academic dishonesty and will result in a 0 on the assignment, a failing grade in the class, and a report to the University.

- Contributions Statement should List the contributions of each team member. An example statement is: *AB wrote the code that generated Figs. 1 and 3 and wrote the associated results paragraphs. XY wrote the code and results for Fig. 2. Both authors collaborated to write the code and results for Fig. 4. AB wrote the Introduction. XY wrote the Discussion and Conclusion. XY formatted the references and edited the final paper. AB&XY consulted with EF to understand how to implement Q and R, but AB & XY wrote their own original code.* Note: the Contributions statement is required to receive a grade.