SARS-Cov-2 Mutational Networks

Bio-Inspired Computing, CSE 598
Assignment 3, Spring Semester 2020

1 Introduction

In this assignment, we will learn about the SARS-Cov-2 virus and study its mutational network. You are invited to work in pairs for this project, but you are not required to.

DUE DATE: April 22, 2020 3:05 pm

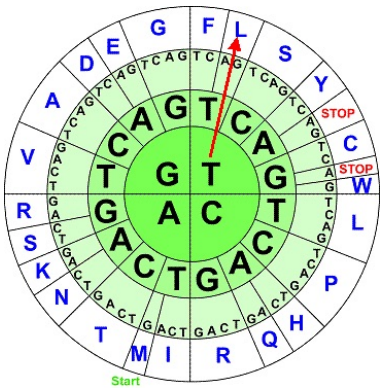
2 Introduction

The SARS-Cov-2 virus is responsible for the respiratory disease known as COVID-19. SARS-Cov-2 is an RNA virus and part of the group of coronaviruses, which cause respiratory tract infections, including the common cold and more serious diseases such as SARS, MERS, and now COVID-19. Scientists would like to know how the SARS-Cov-2 virus (or SARS-2 for short) evolved from other coronaviruses and how likely it is to evolve new forms that might be more or less lethal. An important distinction to keep in mind is 'genotype vs. phenotype,' by which we mean that genetic mutations do not always lead to phenotypic changes in the virus. Scientists know a great deal about base genetic mutation rates and how to use those to calculate the probability of certain new sequences emerging, but the calculations need to account for *neutral* mutations, which here we will define as nucleotide sequences that have the same phenotype as the original.

Some relevant references are posted with this assignment (Wan et al. 2020; Wagner 2012). In addition, Trevor Bedford's Twitter thread is fascinating and easy to read: `https://twitter.com/trvrb/status/1242628550563250176`.

1. Assuming that SARS-2 is 30,000 nucleotides long, how many RNA strands of the same length are exactly 1 mutation away?

2. Approximately what fraction of the possible RNA strands are neutral (have the same phenotype as the original)? Nucleotide substitutions that encode the same amino acid are neutral, and those that encode different amino acids are not neutral. Does your answer change if you consider that some substitutions have more impact on phenotype than others? Use the following codon wheel as needed in your calculations (taken from `http://en.bioinformatica-na-escola.org/activities/vision/research3/genetic-code.html`):



3. We will focus on the mutations that encode the SARS "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.) Here they are:

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

   Black letters indicate the reference sequence, SARS2002; blue indicate mutations in the Civit; green indicate mutations in the bat; and red indicate mutations in SARS-2. Note: SARS-2 also has two of the bat mutations, suggesting that it emerged from bats.

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

4. Calculate the number of possible genomes 1, 2, 3, all the way up to 15 mutations away from an original length 15 genome. Remember not to double-count reverse mutations, i.e., if C mutates into G and then mutates back to C. Show these in a figure. Consider using a log scale.

5. 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 occurs on average once per 7 days. Ignore neutral networks for the 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.

6. Now, incorporate our assumption about the neutral network: Any silent mutation (nucleotide mutation that doesn't change the amino acid) is neutral in any place in the genome. For the 15 nucleotides relevant to the viral variants mentioned above, assume that some small b% of non-neutral mutations (that change the amino acid) are neutral or beneficial and will persist in the population. Assume that (1 - b)% of non-neutral mutations are deleterious and will not persist. Note that there are 3 different amino acids in positions 1 and 2, so in those cases three out of the twenty amino acids (15%) could reasonably be assumed to be beneficial or neutral. Comment on which of the observed variants provide a neutral path from SARS-2 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 additional assumptions you make.

3 Reporting results

Please hand in a 3-4 page report using the ACM format, written in appropriate academic style with proper citations. Send a .pdf of your assignment to me directly on the Due Date and submit a .pdf of the writeup and a .zip file with your code and instructions for running it through Canvas. Your paper should describe the following:

1. The problem you are trying to solve (in your own words), including important terms, concepts and definitions.
2. How you set out to solve the problem, e.g., which method you used to generate negative detectors, which matching algorithm you used, etc.
3. For each question, a separate section or subsection that contains:
   (a) The method you used to answer the question, how you generated each figure and any associated calculations
   (b) Figures with captions that are self-explanatory, have labeled axes, legends, etc.
   (c) A short narrative summarize the results in the figure, interpreting them, and noting any relevant shortcomings, innovations, or ideas for how to improve it.
4. Discussion of what you learned and how relevant you think it is or is not to understanding the evolution of SARS-2.
5. A pointer to your code with instructions for how to run it.