CS423/523 Complex Adaptive Systems

Assignment 2: Analyze viral evolution

Due Tuesday, March 28 (before midnight, 11:59 pm)

Introduction to the assignment

Viruses are replicating information encoded in minimal amounts of matter (minimal compared to other replicating living systems). This leads to a hypothesis: protecting people from viruses is largely an information problem. This leads to the speculation that computer scientists have unique skills to protect people from viruses. Your job in this assignment is to understand viral evolution in terms of the constraints on replicating information. You will need to know a little biology (just enough to be dangerous), but you should approach the problems from a computational perspective -- what information and computations can tell us about the past and future of SARS-CoV-2 evolution?

Background:

You will need to understand the Wagner paper on neutral networks and the Smith paper on antigenic maps. You will also build on 2 papers by Jesse Bloom, a computational immunologist who has created two calculators: one estimates the antigenic escape of mutations at different locations of the receptor binding domain (RBD) of the spike protein of SARS-CoV-2 which is roughly 200 base pairs. This is the region of the viral spike protein with the most neutralizing antibody activity for several reasons: 1) it has to be exposed (perhaps only briefly due to the infamous furin cleavage site) when the virus attaches to the ACE2 receptor in cells of the respiratory tract and 2) it is essential to the virus that it successfully bind to the receptor, so the virus is constrained in how much that site can mutate.Therefore, as an exposed, relatively constrained location, it is a logical place for antibodies to bind.

Experiments verify that many antibodies neutralize this site. The Bloom calculator (https://jbloomlab.github.io/SARS2_RBD_Ab_escape_maps/escape-calc/) shows how mutation to new *amino acids* in over 30 sites affect the binding of the population of antibodies most commonly generated in people in response to the virus. What has surprised many scientists is how quickly SARS-CoV-2 has been able to find so many mutations that successfully bind to the ACE2 receptor but escape antibodies.

The second Bloom calculator estimates the fitness impact of mutations at various locations (https://jbloomlab.github.io/SARS2-mut-fitness/S.html). Fitness impact is estimated (imperfectly) as the difference between the expected percentage of mutations of a certain type and the observed percentage of mutations of a given type in circulating genomes. It is known that different viruses have different likelihoods of different nucleotide mutations. For example in SARS-CoV-2 C → T mutations are most common. (T is the nucleotide in DNA corresponding to U in the codon table for RNA, and you can see that C →U mutations are often neutral, also

called synonymous or silent). The expectation is defined from the data on all circulating genomes as the overall probability for each possible mutation. Mutations that are selected against are observed less frequently at particular locations; mutations that are selected for are observed more frequently than chance at particular locations.

The take away message here is this: in addition to the synonymous mutations that are neutral because they do not change the encoded amino acid (i.e. mutating ACC -> ACA is neutral because it still encodes Threonine), there are other mutations that are a) *antigenically neutral* because they do not change antibody binding or b) functionally neutral because they do not change fitness. You will consider these three types of neutral mutation for different parts of this assignment.

Summary of the assignment:

The assignment has 3 parts. First, you will calculate some properties of relevant segments of the virus and its neutral network. Second, you will create several neutral networks and create an antigenic map of variants accessible by the amino acid neutral network and the antigenic neutral network. Third, you will make some predictions about what is possible and/or likely for viral evolution. What you investigate here will be largely up to you, but interesting avenues include analyzing the plausibility of natural spillover vs lab leak origins, or predicting how far in the future escape from vaccines or prior infection will occur. You look at how long it should have (in theory) taken to evolve escape antibodies in prior variants and/or future variants. You could ask about the probability of beneficial or detrimental variants (from the perspective of humans, other animals in “spillback” or the virus). You can focus on theoretical arguments (what if experiments with evolution over neutral networks) or empirical realities (analyze data from past evolution) or a combination. This last part should include stating hypotheses and testing them with a set of simulations or calculations and should have at least some evidence from a published paper (either data or interpretations from biologists).

Part 1 - Calculations

The full viral genome is just under 30,000 nucleotides and therefore codes for approximately 10,000 amino acids
The spike protein RBD is approximately 600 nucleotides and 200 amino acids
(i.e., shown in comparison to other related coronaviruses here: https://www.nature.com/articles/s41423-020-0400-4)

You can use these approximations of length of find a paper that gives more definitive numbers for the following.

Start with the original Wuhan strain.

QA1) How many genomes are 1 basepair mutation away from the original strain? First consider neutral and non-nuetral mutations in base pairs (simple bit flips, but with 4 options, noting these are also called synonymous and nonsynomous because here we are considering only the nucleotide mutations that still code for the same amino acid.)

QA2)How many of these are synonymous mutations and how many are nonsynonomous and *might* have some affect on fitness?

Create a neutral network of all genomes with 3 mutations.
QA3) How many genomes are in that neutral network (don't count reversions back to previous genomes as new)?
QA4) How many genomes are 1 mutation away from that neutral network?
QA5) How many of these genomes (the ones 1 mutation away from the 3-step-neutral network) are synonymous and how many non-synonomous.
QA6) Calculate how fast synonymous and non-synonymous mutations are accumulated by generating n-step-neutral networks and looking 1 mutation from that neutral network.
QA7) Explain in 1-2 paragraphs what exploration at the edges of neutral networks mean for viral evolution.

Now consider mutations that are *antigenically* neutral. For this assignment you should establish a threshold in the Bloom Escape Calculator to determine when something is antigenically neutral. Further, only consider the antigenically neutral mutations in the RBD of the spike protein.

QB1-7 (multipart) Repeat questions QA1-7, this time calculating how many amino acid changes are posssible that are *antigenically* neutral (ie they do not affect the non-neutral binding sites of prior antibodies.)

QC) Assume 0.001% of all neutral mutations compensate for some deleterious effect of 1 mutation that escapes antibodies. What are the chances of antibody escape that has such negligible fitness?

Part 2 - Build a neutral network (of your choice) and create an antigenic map from variants you create from it. Traverse this neutral network (perhaps with a random walk) to create new variants that are 1 (or more) mutations away from neutral network. Create an antigenic map of these new variants you have imagined based on the Bloom Escape Calculator.

Discuss how likely any of these variants are based on different assumptions you can make about finding an epistatic change that will compensate for deleterious effects of escaping antibody? For this, use the Bloom Fitness Calculator to estimate or extrapolate how likely it is that one or some of your variants could find epistatic mutations that compensate for deleterious effects of their antibody escape. You can speculate here, but back up your speculations with a mathematical analysis.

Translation table 1 [edit]

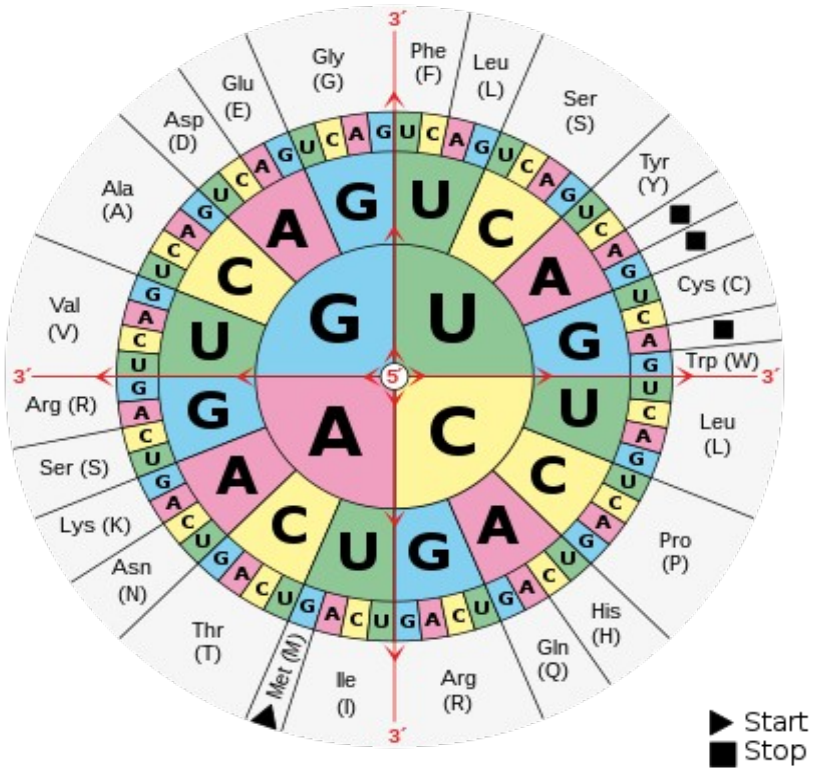
Standard RNA codon table [edit]

Amino-acid biochemical properties	Nonpolar †	Polar †	Basic ‡	Acidic ↓

Termination: stop codon *	Initiation: possible start codon →

Standard genetic code[1][10]

1st base	2nd base U		2nd base C		2nd base A		2nd base G		3rd base
U	UUU	(Phe/F) Phenylalanine †	UCU	(Ser/S) Serine †	UAU	(Tyr/Y) Tyrosine †	UGU	(Cys/C) Cysteine †	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L) Leucine †	UCA		UAA	Stop (*Ochre*) *[note 2]	UGA	Stop (*Opal*) *[note 2]	A
	UUG →		UCG		UAG	Stop (*Amber*) *[note 2]	UGG	(Trp/W) Tryptophan †	G
C	CUU		CCU	(Pro/P) Proline †	CAU	(His/H) Histidine ‡	CGU	(Arg/R) Arginine ‡	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	(Gln/Q) Glutamine †	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	(Ile/I) Isoleucine †	ACU	(Thr/T) Threonine †	AAU	(Asn/N) Asparagine †	AGU	(Ser/S) Serine †	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K) Lysine ‡	AGA	(Arg/R) Arginine ‡	A
	AUG →	(Met/M) Methionine †	ACG		AAG		AGG		G
G	GUU	(Val/V) Valine †	GCU	(Ala/A) Alanine †	GAU	(Asp/D) Aspartic acid ↓	GGU	(Gly/G) Glycine †	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E) Glutamic acid ↓	GGA		A
	GUG →		GCG		GAG		GGG		G



Part 3 Expand your analysis in some meaningful way. Extend some analysis you started in part 1 or 2, or if you're feeling bold, develop a new analysis. Your new approach can be interesting mathematically, because its exploring a concept in evolution or because it is grounded in published literature, or a combination. DISCUSS - what does this mean for SARS-CoV2 past and future evolution? If you choose to model origins, think about what it could mean - could it have evolved from known (or unknown) viruses in bats or other mammals? Is there some evidence that the virus evolved with a large jump, rather than what could have occurred over a neutral network? Propose the data that would be needed to support or refute a spillover hypothesis.