Supplementary Material for Predicting the Emergence of SARS-CoV-2 Clades

Siddharth Jain[1], Xiongye Xiao[2], Paul Bogdan[2,*], and Jehoshua Bruck[1,*]

[1]California Institute of Technology, Electrical Engineering, Pasadena, 91125, USA
[2]University of Southern California , Electrical and Computer Engineering, Los Angeles, 90007, USA
[*]pbogdan@usc.edu, bruck@caltech.edu

ABSTRACT

Evolution is a process of change where mutations in the viral RNA are selected based on their fitness for replication and survival. Given that current phylogenetic analysis of SARS-CoV-2 identifies new viral clades after they exhibit evolutionary selections, one wonders whether we can identify the viral selection and predict the emergence of new viral clades? Inspired by the Kolmogorov complexity concept, we propose a generative complexity (algorithmic) framework capable to analyze the viral RNA sequences by mapping the multiscale nucleotide dependencies onto a state machine, where states represent subsequences of nucleotides and state-transitions probabilities encode the higher order interactions between these states. We apply computational learning and classification techniques to identify the active state-transitions and use those as features in clade classifiers to decipher the transient mutations (still evolving within a clade) and stable mutations (typical to a clade). As opposed to current analysis tools that rely on the edit distance between sequences and require sequence alignment, our method is computationally local, does not require sequence alignment and is robust to random errors (substitution, insertions and deletions). Relying on GISAID viral sequence database, we demonstrate that our method can predict clade emergence, potentially aiding with the design of medications and vaccines.

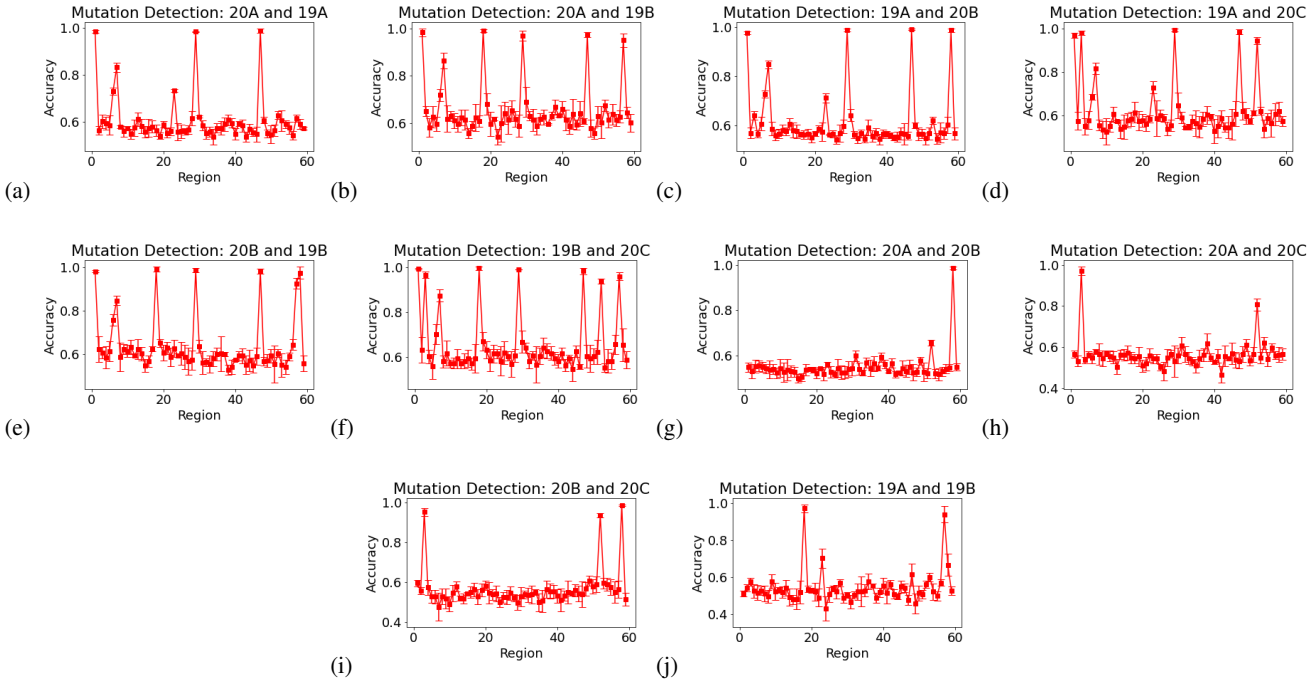

Figure S1. Mutation detection with region length 500. SARS-CoV-2 RNA is divided into continuous non-overlapping regions of length 500. Spikes represent the regions where mutations differentiating clade pairs 20A and 19A (a), 20A and 19B (b), 19A and 20B (c), 19A and 20C (d), 20B and 19B (e), 19B and 20C (f), 20A and 20B (g), 20A and 20C (h), 20B and 20C (i), 19A and 19B (j) are detected.

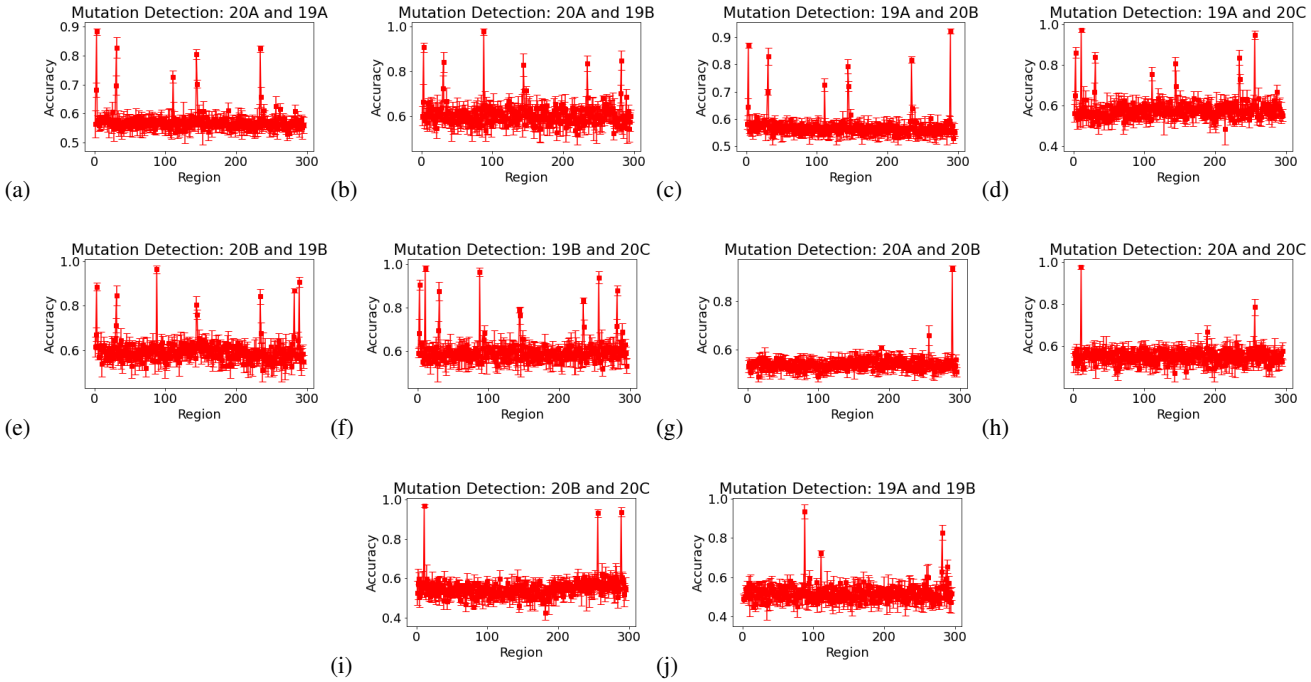

Figure S2. Mutation detection with region length 100. SARS-CoV-2 RNA is divided into continuous non-overlapping regions of length 100. Spikes represent the regions where mutations differentiating clade pairs 20A and 19A (a), 20A and 19B (b), 19A and 20B (c), 19A and 20C (d), 20B and 19B (e), 19B and 20C (f), 20A and 20B (g), 20A and 20C (h), 20B and 20C (i), 19A and 19B (j) are detected.

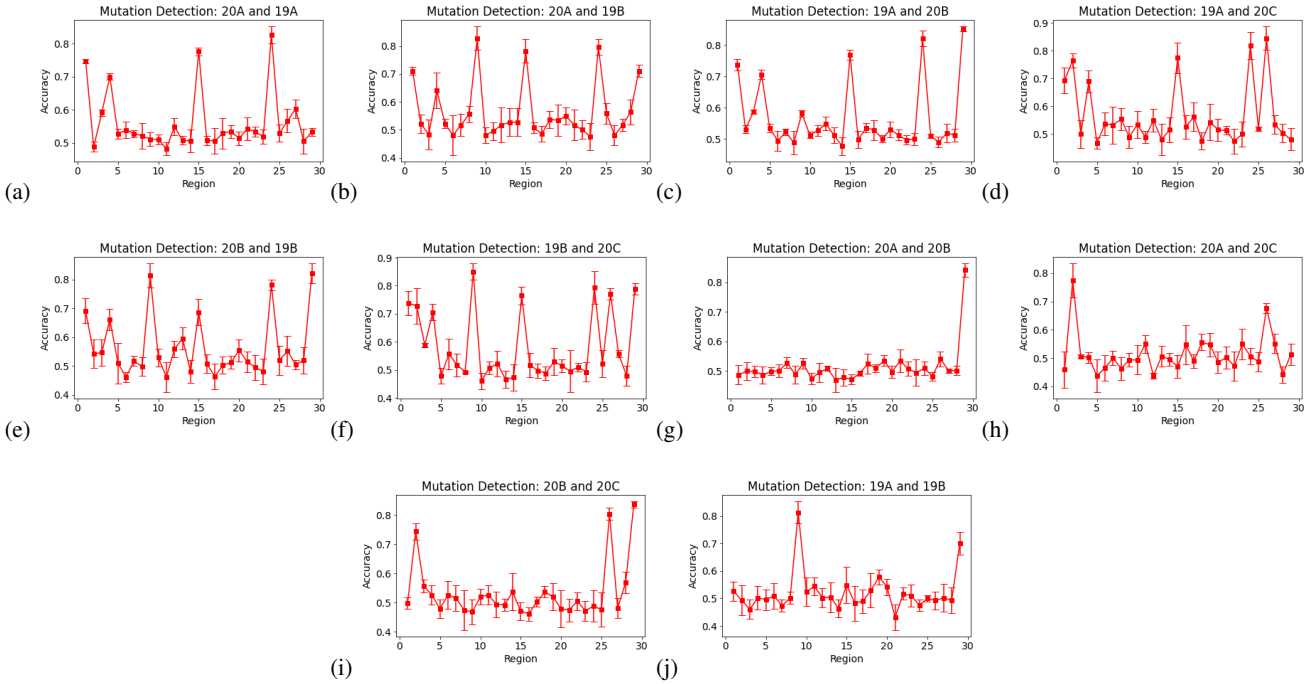

Figure S3. Mutation detection in the presence of synthetically added noise. We synthetically added 10% noise in the SARS-CoV-2 viral RNA sequences by adding substitution, insertion and deletion mutations. We show here that the mutations differentiating clades can still be detected, meaning that the detection using state transition probability matrix is robust to noise.

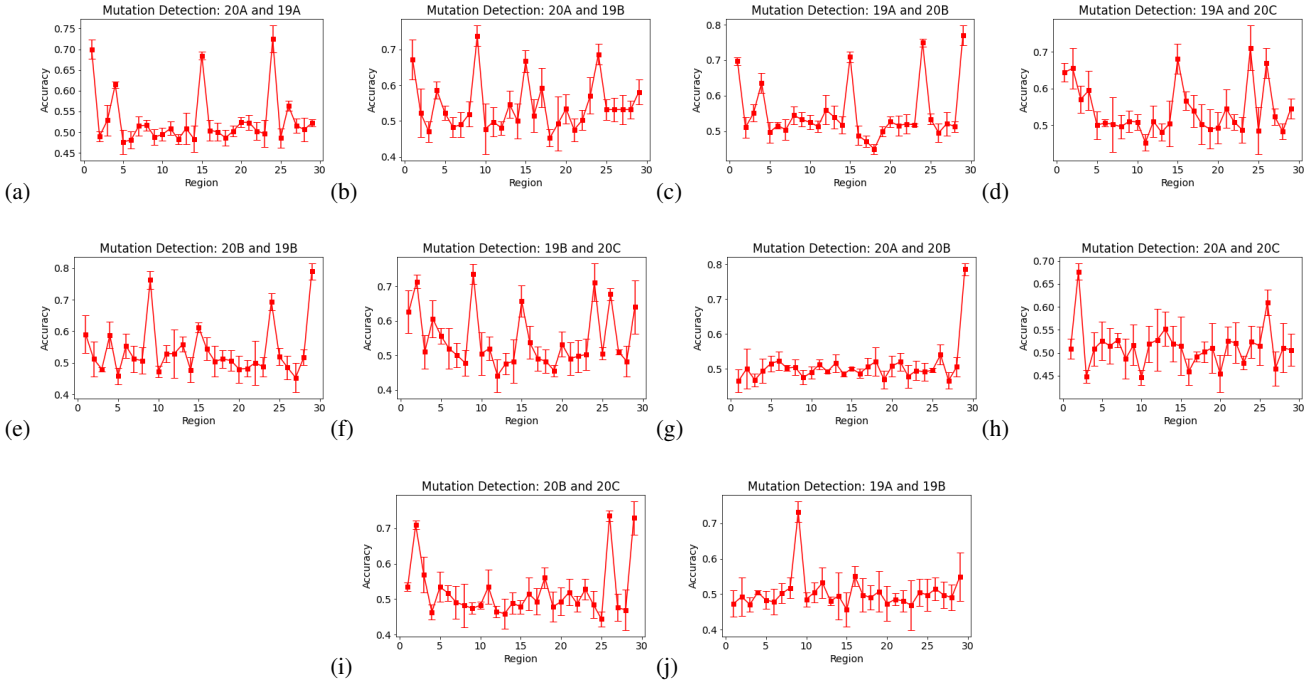

Figure S4. Mutation detection in the presence of synthetically added noise. We synthetically added 15% noise in the SARS-CoV-2 viral RNA sequences by adding substitution, insertion and deletion mutations. We show here that the mutations differentiating clades can still be detected, meaning that the detection using state transition probability matrix is robust to noise.

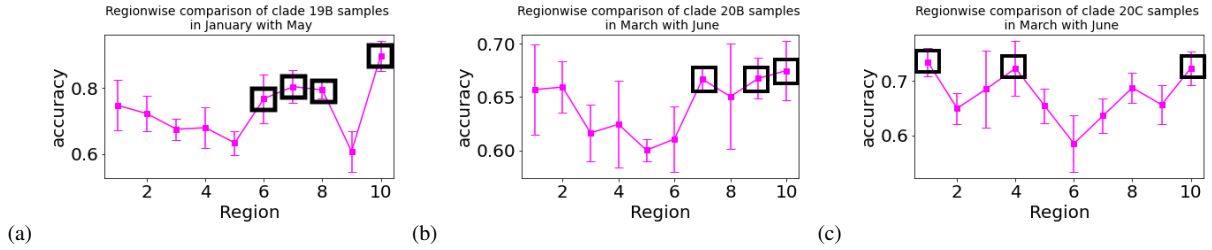

Figure S5. Evolution activity within clades 19B, 20B, 20C. (a) Regions 6 (15000-18000), 7 (18000-21000), 8 (21000-24000) and 10 (27000-30000) are actively mutating within clade 19B and could correspond to the location of mutations in the new clade that may emerge from 19B in the future. (b) Regions 7 (18000-21000), 9 (24000-27000) and 10 (27000-30000) contributes to maximum classification accuracy and hence are the most actively mutating within clade 20B which indicates the possibility of the emergence of a new clade from 20B with mutations in region 7,9 and 10. (c) Regions 1 (0-3000), 4 (9000-12000) and 10 (27000-30000) are the most actively mutating regions within clade 20C which indicates the possibility of emergence of a new clade from 20C with mutations in these regions.

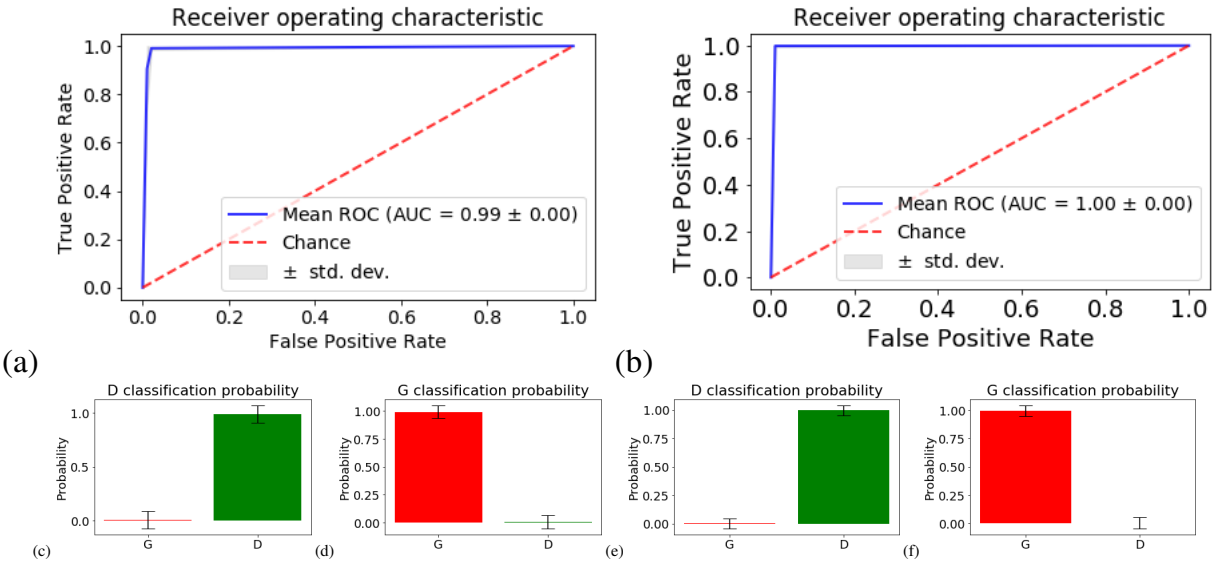

Figure S6. The state machine model can distinguish spike protein mutation sequences from others. We demonstrate here that the sequences with spike protein mutation (D614G) can be distinguished from others using the state machine transformation. Our state machine model accurately captures the higher order differences and classifies sequences with spike protein mutation from the others. (a), (b) ROC curves for the classifier built using the state transition probability as the feature vector to differentiate between sequences with spike mutation from the others. $k = 4, b = 1$ and $k = 5, b = 1$ for (a) and (b) respectively. (c)-(f) show the classification probability on the validation set for sequences with spike mutation and the others using classifier built with state transition probabilities as features.

Table S1. Highly Likely Transitions

probability ≥ 0.5
$aaagc \to t, aaatc \to a, aacga \to a, aactc \to a, acctc \to a, actat \to t, actca \to a, agcct \to t, ataac \to a, atacg \to t,$ $atcga \to t, attcg \to t, atttc \to a, caacg \to t, caagc \to t, cacag \to a, cagac \to a, cagcg \to t, catac \to a, ccaag \to a,$ $cccat \to t, cccca \to a, cccgc \to a, ccctc \to a, ccgag \to g, ccgcg \to a, ccggg \to t, ccggt \to a, cctcc \to a, cctgc \to t,$ $cctgg \to t, cgaaa \to t, cgact \to a, cgatt \to t, cggac \to a, cggat \to g, cgtag \to t, cgtat \to a, cgtgg \to t, cgtgt \to t,$ $ctaac \to a, ctacc \to a, ctagg \to t, ctcgc \to a, ctctc \to a, gaagc \to t, gaagg \to t, gaccc \to t, gacgg \to t, gactc \to a,$ $gaggc \to t, gagtg \to t, gatgc \to t, gatgg \to t, gatta \to t, gcccg \to t, gcctc \to a, gcgag \to a, gcggc \to a, gcgtc \to a,$ $gctac \to t, gctcc \to a, gctgc \to t, gctgg \to t, ggacc \to t, ggagc \to t, ggatc \to a, ggcag \to a, gggtg \to t, gggtt \to t,$ $ggtga \to t, ggtgc \to t, ggtgg \to t, ggtgt \to t, gtacc \to a, gtatc \to t, gtccc \to t, gtctc \to t, gtgcg \to t, gtggc \to t,$ $gttag \to a, gttct \to t, tactc \to a, tagtc \to t, tatcc \to t, tatcg \to t, tattc \to t, tcacc \to t, tcgag \to g, tcggt \to a,$ $tctcg \to t, tgcag \to a, tgcga \to a, tggac \to a, tggtc \to a, ttagc \to t, ttata \to a, ttatc \to t, ttccc \to t, ttcga \to t,$ $ttcgc \to a, ttgac \to a, ttggt \to g.$
probability ≥ 0.55
$aacga \to a, aactc \to a, agcct \to t, ataac \to a, atacg \to t, attcg \to t, caagc \to t, cagcg \to t, ccaag \to a, cccat \to t,$ $cccca \to a, cccgc \to a, ccgag \to g, ccgcg \to a, ccggg \to t, ccggt \to a, cctcc \to a, cctgc \to t, cctgg \to t, cgaaa \to t,$ $cgact \to a, cggac \to a, cggat \to g, cgtag \to t, ctaac \to a, ctacc \to a, ctcgc \to a, gaagc \to t, gaagg \to t, gaccc \to t,$ $gaggc \to t, gcccg \to t, gcgag \to a, gcggc \to a, gcgtc \to a, gctac \to t, gctcc \to a, gctgc \to t, gctgg \to t, ggagc \to t,$ $ggatc \to a, ggcag \to a, gggtg \to t, ggtgg \to t, ggtgt \to t, gtacc \to a, gtatc \to t, gtccc \to t, gtctc \to t, tatcg \to t,$ $tattc \to t, tcggt \to a, tctcg \to t, tgcga \to a, tggac \to a, ttagc \to t, ttcga \to t, ttgac \to a.$
probability ≥ 0.6
$atacg \to t, caagc \to t, ccgag \to g, ccgcg \to a, ccggg \to t, cctcc \to a, cggac \to a, cggat \to g, cgtag \to t, ctaac \to a,$ $ctacc \to a, gaagg \to t, gaccc \to t, gaggc \to t, gcccg \to t, gcgag \to a, gctcc \to a, gctgg \to t, ggtgt \to t, gtccc \to t,$ $gtctc \to t, tatcg \to t, ttagc \to t.$
probability ≥ 0.65
$ccggg \to t, cggat \to g, gctcc \to a, gtccc \to t, gtctc \to t.$
probability ≥ 0.7
$gtccc \to t.$

Table S2. Mutations differentiating different clades (obtained from nextstrain.org)

Clades	Mutations
19A	Root clade
19B	C8782T, T28144C
20A	C14408T, A23403G
20B	C14408T, A23403G, G28881A, G28882A, G28883C
20C	C14408T, A23403G, C1059T, G25563T