The Protein Journal (2021) 40:328–340
https://doi.org/10.1007/s10930-021-09988-3

In Silico Study of Mutational Stability of SARS-CoV-2 Proteins

Dwaipayan Chaudhuri[1] · Satyabrata Majumder[1] · Joyeeta Datta[1] · Kalyan Giri[1]

Accepted: 12 April 2021 / Published online: 22 April 2021

Abstract
Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), an enveloped RNA virus transmits by droplet infection thus affects the respiratory system. Different genomes have been reported globally for SARS-CoV-2 with moderate level of mutation which makes it harder to combat the virus. Mutational profiling and the relevant evolutionary aspect of coronavirus proteins namely spike glycoprotein, membrane protein, envelope protein, nucleoprotein, ORF1ab, ORF3a, ORF6, ORF7a, ORF7b and ORF8 were studied by in silico experiments. Clustering of the protein sequences and calculation of residue relative abundance were done to get an idea about the protein conservancy as well as finding out some representative sequences for phylogenetic and ancestral reconstruction. By mutational profiling and mutation analysis, the effect of mutations on the protein stability and their functional implication were studied. This study indicates the mutational effect on the proteins and their relevance in evolution, which directs us towards a better understanding of these variations and diversification of SARS-CoV-2 for useful future therapeutic study and thus aid in designing therapeutic agents keeping the highly variable regions in mind.

Keywords SARS-CoV-2 proteins · Mutation stability · Clustering · Shannon entropy · Mutation profiling

1 Introduction

Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) is the seventh coronavirus which is known to infect humans. Previously reported SARS-CoV (2002–03) and MERS-CoV (2012) also lead to severe respiratory illness whereas HKU1, NL63, OC43 and 229E are associated with relatively mild diseases [1]. The virus is a member of the betacoronavirus family, which is confirmed by the fact that it has 79.6% identity to SARS-CoV [2]. Phylogenetic studies have also confirmed that the virus belongs to the betacoronavirus family and is closer to the SARS-like coronavirus in bat [3]. The virus shows a 4% sequence variability when compared with its closest relative, bat SARS-related corona virus RaTG13 and the variability reaches to almost 17% in the neutral sites. The development of new variations in the virus which has given rise to the more prevalent L type and less found S type. This is not only due to recombination but can also be attributed to mutation and natural selection [4].

The virus has approximately 30 kb long positive sense single stranded RNA genome with 38% GC content. The virus particle has a spherical shape with some polymorphism, the diameter of which range from 60 to 140 nm with distinctive spikes about 8 to 12 nm in length [5]. The virus consists of a 7096 amino acid long replicase polyprotein 1ab which is cleaved to form several proteins which play a very important role in virus replication. The components are host translation inhibitor nsp1, RNA dependent RNA polymerase, helicase, protease, endonuclease, exonuclease, methyltransferase and many others. This polyprotein is synthesized in such a manner that during translation there is a ribosomal frameshift to the − 1 frame which makes this polyprotein different in comparison to polyprotein 1a which is translated from the same RNA region. The spike glycoprotein is one of the most important proteins on the surface of the virus which interacts with the host ACE2 receptor for entry into the cell [6]. The protein encoded by ORF7a is a non-structural protein which plays a very important role in viral replication in cell culture by modulating the host G0/G1 transition checkpoint. The protein encoded by the ORF3a forms a homotetrameric K^+ ion channels and can also modulate the release of virus particles. The membrane protein and envelope protein both constitute the virion outer envelope.

✉ Kalyan Giri
kalyan.dbs@presiuniv.ac.in

[1] Department of Life Sciences, Presidency University, 86/1 College Street, Kolkata 700073, India

The nucleoprotein packages the viral RNA genome to form the helical ribonucleocapsid and also plays a very important role in viral assembly through its interactions with the viral genome and membrane protein. The envelope protein besides its role in forming the viral envelope also acts as a viroporin to form pentameric protein-lipid pores that allow ion transport. The proteins encoded by ORF6, ORF7b and ORF8 may function as a determinant of virus virulence, integral component of viral membrane and a relay point for host-viral interaction, respectively [7].

In this study almost all of the proteins belonging to the virus were considered and those were characterized using several in silico methods. This has been done to create a mutation profile and thus analyze all the mutations with respect to a parent ancestral strain from where the virus has diversified. These findings provide idea about the impact of different mutations on the proteins in terms of their stability and overall influence aiding in the study of evolution of this virus.

1.1 Methodology

1.1.1 Sequence Retrieval

The protein sequences of ORF1ab, ORF3a, ORF6, ORF7a, ORF7b, ORF8, spike glycoprotein, nucleocapsid, envelope protein and membrane protein were retrieved from the NCBI virus page. Globally 1740 (one from Africa, 163 from Asia, 28 from Europe, 1542 from North America, one from Oceania and four from South America) complete genomes have been sequenced with the human host till 9th May, 2020. Amino acid sequences of these ten proteins belong to the genomes mentioned above.

1.1.2 Clustering and Shannon Entropy Calculation

Sequences were clustered using CD-HIT (CDHIT: https://pubmed.ncbi.nlm.nih.gov/16731699/) at a threshold of 100% identity to determine the total number of different unique sequences that are present for each protein [8]. This also helps to identify what could be the native residue, at a particular position which has undergone mutation over the course of time as well as provides the data of relative abundance of different amino acids at each position of a protein. This also aids in getting idea to gain more information on the possible ancestral strain of the virus which is later done by phylogenetic reconstruction of the closest ancestral sequences. Shannon entropy (H) is used to estimate the diversity among protein sequences by the equation:

$$H = -\sum_{i=1}^{M} P_i \log_2 P_i$$

(P_i = probability of i, where i refers to the amino acid type and M refers to the number of amino acid type) [9]. H ranges from 0 (only one residue is present at that position) to 4.322 (all 20 residues are equally represented in that position). High entropic value at a particular position means higher variability of amino acid at that position and vice versa. When Shannon entropy is applied to a multiple sequence alignment (MSA), it computes the entropy (variability) for every position. In statistical mechanics, entropy is a measure of the number of ways a system can be arranged or number of states. In MSA, the entropy value for a particular position indicates the number of amino acids can be present in that position (more like informational entropy). Thus, for DNA or protein sequence it can compute the relative variation (with respect to nucleotide or amino acid, respectively) present in different sequences of a set as well as provide some quantitative data to guess the nucleotide or amino acid which is present with higher percentage for a particular sequence set [10]. Here, Los Alamos Shannon entropy tools were used for the calculation Fig. 1. Execution of these two processes were used to map the relative abundance of the residues in the proteins along with some representative mutant sequences.

1.1.3 Phylogenetic Analysis

All the clustered representatives were phylogenetically analyzed using the Dayhoff model of amino acid substitution and 1000 bootstraps in the IQTREE server (http://iqtree.cibiv.univie.ac.at/) [11]. Approximate Bayes' test was also done for the same. The final trees were visualized in iTOL (https://itol.embl.de/) to get a phylogram [12] Fig. 2. So by this analysis, it was possible to map the sequence which is closest to and have arisen from the root itself as well as the diversified sequences. By ancestral sequence reconstruction from these in-group phylogenetic trees, the root sequence i.e. the first node sequence was reconstructed and compared with existing sequences. Phylogenetic analysis (using outgroup protein sequences) was performed to detect the true root of the sequences, where the Bat Coronavirus sequences served as outgroups. Due to change in sequence size owing to evolution, the size of the ancestral sequences could be different in certain cases and these changed length sequences were considered for further analysis involving stability upon mutation. Multiple sequence alignment was done using MultAlin (http://www.sacs.ucsf.edu/cgi-bin/multalin.py) to compare the Wuhan reference strain with the most recent ancestor reconstructed from ingroup and outgroup phylogenetic trees.

1.1.4 Mutation Profile

The mutational profile of the proteins were generated using EVmutation (EVcouplings) web server (https://evcouplings.org) [13, 14]. Presently the unique mutation profiling algorithm used in this server, accepts data

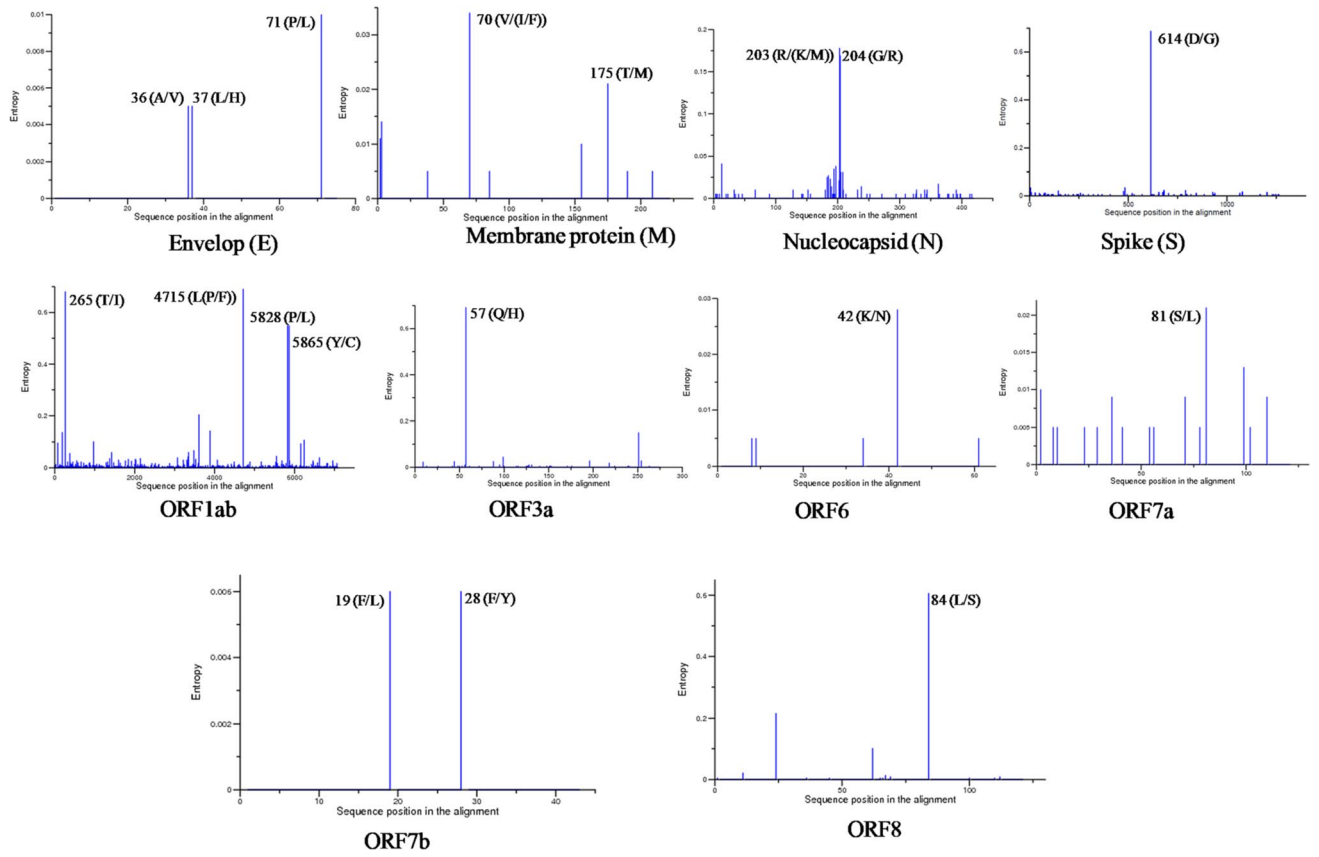

Fig. 1 Shannon entropy calculation of the aligned protein sequences of SARS-CoV-2. The bars show the frequency of variation of amino acids at that position. Lower entropic frequency means the presence of highly conserved amino acid residue at that position

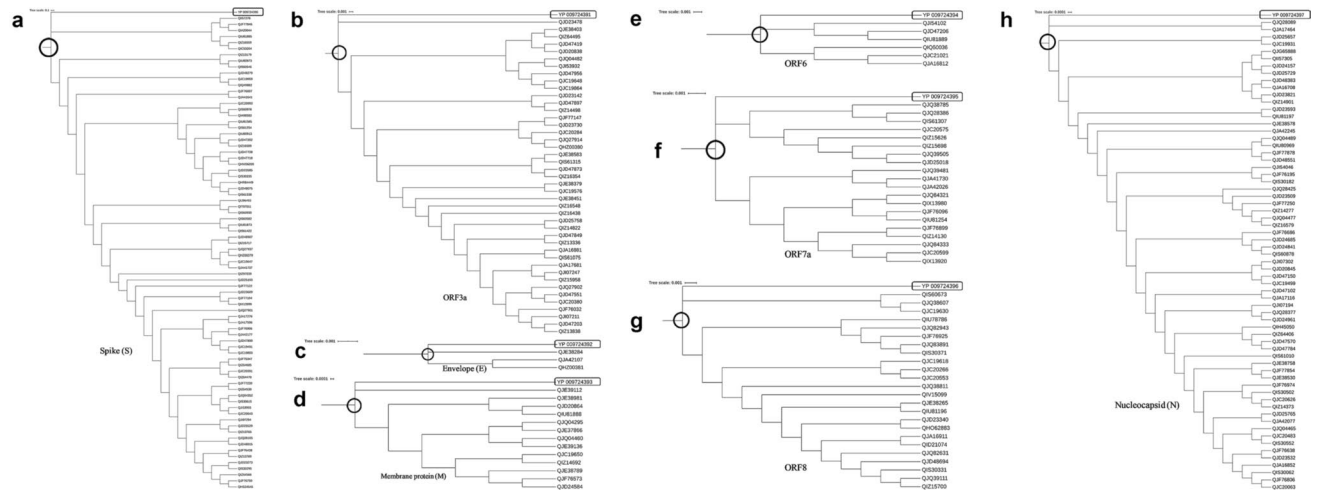

Fig. 2 Phylogenetic trees of eight SARS-Cov-2 proteins: **a** spike, **b** ORF3a, **c** envelope, **d** membrane protein, **e** ORF6, **f** ORF7a, **g** ORF8, **h** nucleocapsid, are mentioned in the text in Newick format. Representative sequences derived from clustering data were used for phylogenetic tree construction. Results show that the initial sequences (YP009724390, YP009724391, YP009724392, YP009724393, YP009724394, YP009724395, YP009724396, YP009724397) belong to the same strain for all proteins. The circle in each cases indicates the starting point for diverging out of phylogenetic tree. The rectangular boxes denote the representative sequence (ancestral sequence)

belonging to the range of 50–500 amino acids. Thus, ORF1ab, spike protein and ORF7b were not considered for being out of acceptance range, while the rest seven proteins abiding by the requisites of the server were considered for further analysis being in the acceptable region. In case of polyprotein ORF1ab, any analysis other than the stability of the protein would not be accurate as it gets cleaved into different individual proteins with independent functions in many cases inside the host cell. In each of these profiles the ancestral sequence reconstructed from the ingroup phylogenetic trees was taken as the starting point and effect of all mutations mapped using two separate algorithms: the independent model and the epistatic model. The independent model Fig. 3 considers only local effect of each mutation, having an independent effect on the overall characteristics of the protein while the epistatic model Fig. 4 considers coupling of both local and global effect of mutation and calculates the change in energy using an energy function on the basis of statistical physics. Both these models and their respective algorithms are part of the EVmutation server.

1.1.5 Mutation Analysis

The iMutant 2.0 server (https://folding.biofold.org/i-mutant/i-mutant2.0.html) was used to predict the difference of change in Gibbs free energy between wild type protein (in this case is the ancestral reconstructed strain from the outgroup trees so that the true root is used for analysis) and the mutated sequence to predict the effect of mutation on stability [15–17]. The ΔG of formation of both wildtype and mutant proteins is the change in energy of the same from unfolded to folded state (i.e. ΔG = energy of folded state—energy of unfolded state) and ΔΔG is the difference between ΔG of formation of wild type and ΔG of formation of mutant protein with both values being negative as folding of protein leads to decrease in energy ($\Delta\Delta G = \Delta G_f^{WT} - \Delta G_f^{Mutated}$). Thus a negative value indicates the wild type value is more negative than mutant implying that the wildtype is more stable and mutation leads to decrease of stability. On the other hand a positive value indicates that the wild type value is less negative than the mutant protein implying the increase of stability upon mutation. For sequence based prediction, the server

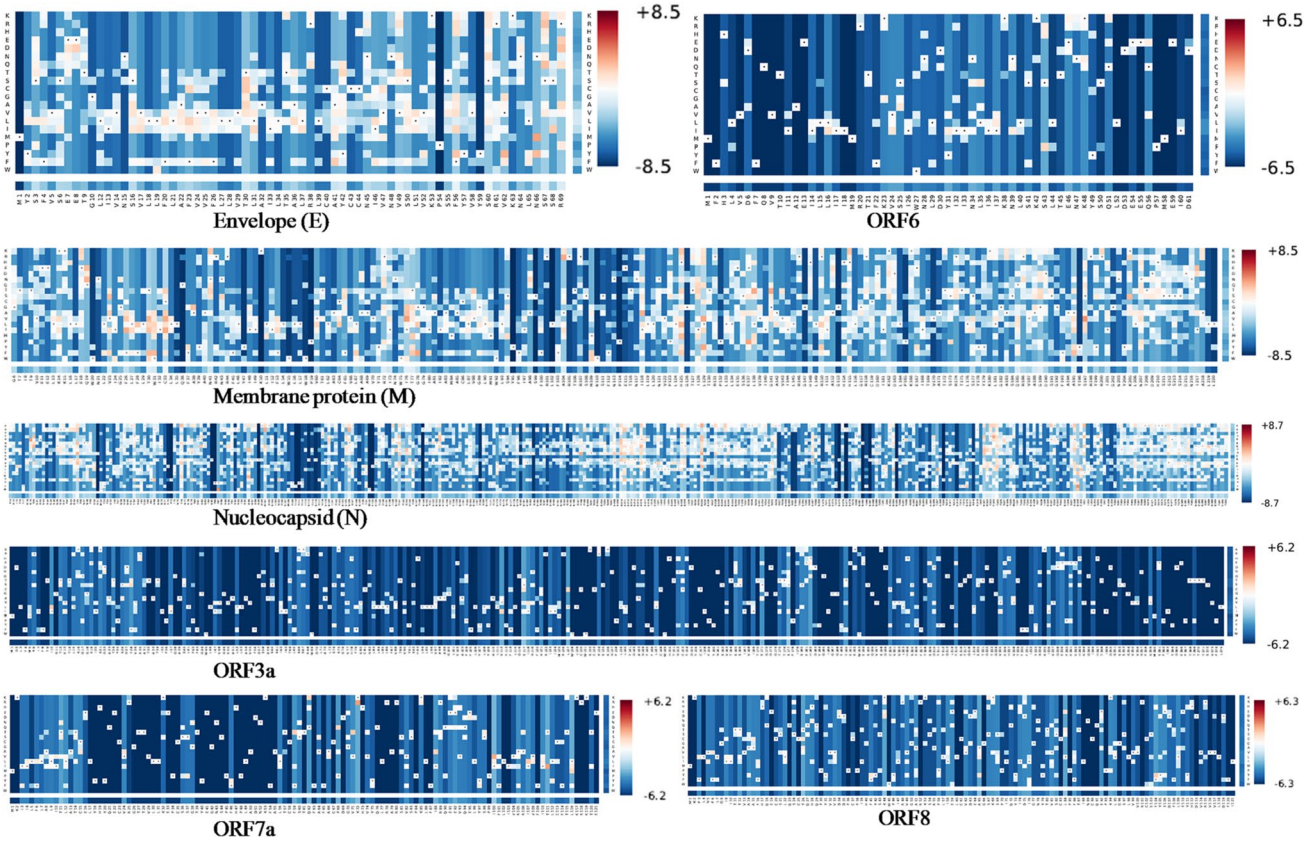

Fig. 3 Independent model of mutational profiling using EVmutation server. X-axis of the graph represents the amino acid sequences with corresponding position of the mentioned protein and Y axis represents the amino acid (KRHEDNQTSCGAVLIMPYFW) substitution. The colour gradient from higher intensity of blue towards red refers to the lowering of harmful effect due to mutations on the protein

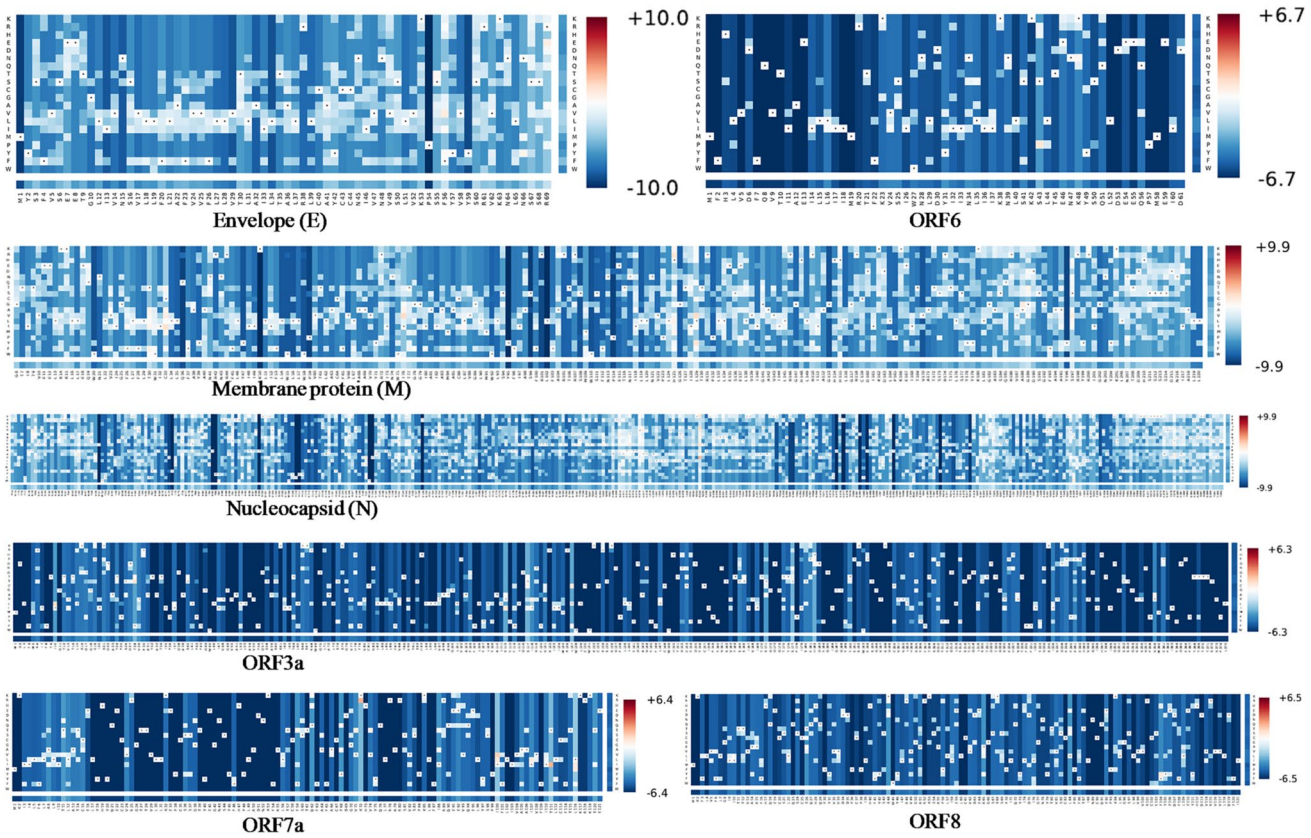

Fig. 4 Epistatic model, considering both local and global effect of mutation using EVmutation server. X-axis of the graph represents the amino acid sequences with corresponding position of the mentioned protein and Y axis represents the amino acid (KRHEDNQTSCGAVLIMPYFW) substitution. The colour gradient from higher intensity of blue towards red refers to the lowering of harmful effect due to mutations on the protein

has shown an accurate result in 77% of the cases using the dataset belonging to Protherm, the database for all experimentally determined thermodynamic parameters of wild type and mutant proteins. According to the server ΔΔG is the difference of unfolding Gibbs free energy between the mutated and wild type of the protein which is a positive value as proper folding of the protein leads to decrease of energy. So when the difference between these ΔG are considered for wild type and mutant species, a negative value is considered to be destabilizing as the difference between the energy of unfolded and folded state gets reduced in the mutant strain thus making it smaller than that for wild type sequence. So a positive difference indicates that the ΔG of mutant sequence is greater (greater negative value) than the wild type sequence and thus the mutation has a stabilizing effect on the protein. All mutations were performed at 25 °C and pH 7. The temperature was selected such that it is close to average temperature of the world and pH such that it is neutral and does not influence any particular type of mutation. Both these factors affect the nature of mutation and its effect on stability where changing either of values can affect the nature of results. All mutations were performed in two sets at 25 °C, pH 7 and 37 °C, pH 7.4. The temperatures were selected such that it is close to average temperature of the world and the conventional temperature at which enzymes get activated. Similarly, neutral pH as well as human blood pH were selected so that none of them influence any particular type of mutation. The results were compared by performing a paired T-test (significance level 0.05) to investigate statistical significance between the two sets. Paired T-test is performed between two similar sets at two different scenarios and gives result of significant difference between the two.

The DUET web server (http://biosig.unimelb.edu.au/duet/stability) was used to study the effect of mutation on protein stability at the structural level [18]. It takes into account a statistical potential energy and graph based signatures to represent the 3D protein environment to predict the effect of the mutation on protein stability. The structures for the same was constructed using homology modelling taking 7jtl.pdb as template for ORF8, 6xdc.pdb for ORF3a, 6m3m.pdb for nucleocapsid and 7cn8.pdb for spike. These structures were used for study of mutations' effect on stability.

1.1.6 Functional Effect of Mutations

The functional impact of mutations i.e. whether the mutation is tolerated by the protein or not is predicted by the SIFT web server (http://sift-dna.org) which makes use of similar protein to evaluate the effect of amino acid substitution based on sequence homology [19] thus taking into consideration the evolutionary conservancy of residues where it is thought that functionally important residues are those which remain conserved throughout the course of evolution. Thus only proteins which have a minimum number of sequences can be analyzed using the same. Due to lack of acceptable number of sequences available, ORF8 and ORF7b could not be analyzed to take a look at the functional relevance of the mutations taking the Wuhan reference strain (NC_045512) as reference for evaluating the tolerance of the mutation.

2 Result

2.1 Identification of Unique Sequences of SARS-CoV-2 Proteins Sequences

Protein sequences with complete information regarding their mutational changes, are only considered in this study for all SARS-CoV-2 proteins except the polyprotein ORF1ab where all sequences were taken into consideration. This polyprotein gives rise to all the functional and nonstructural proteins of the virus like protease, exonucleases, endonucleases, helicase, polymerase etc. and loss of information is for just a part of the protein and not all the protein. So for this protein only the unknown positions designated as X positions were discarded and not the whole sequences. The number of different unique sequences for the various proteins are as follows: 630 for ORF1ab (including the incomplete sequences), 79 for spike glycoprotein, four for Envelope, 14 for membrane protein, 66 for Nucleoprotein, 44 for ORF3a, seven for ORF6, 18 for ORF7a, 25 for ORF8 and three for ORF7b. All these sequences differ from each other in at least one position. At 90% identity it has been seen that these clusters coalesce to a single cluster, reiterating the fact that they do not have too many mutations amongst them and these mutations can only be identified at a very fine resolution.

2.2 Calculation of Relative Abundance of Amino Acids of SARS-CoV-2 Proteins

Shannon entropy was calculated for the protein sequences to determine the relative conservancy and abundance of each amino acid at each position Fig. 1. The relative abundance data provides us with the information of what all mutations are found at each respective position for the proteins. The Shannon entropy data is useful in knowing the conservancy of an amino acid in a particular position which may have functional implications as it is the functionally important residues which remain evolutionarily conserved.

2.3 Phylogenetic Analysis of the Representative Sequences of SARS-CoV-2 Proteins

The representative sequences from clustered data were used for phylogenetic tree construction using the IQTree software package where the Dayhoff model was used for the same along with Bayes' test and 1000 bootstraps. The reconstructed sequence at the root for the reconstructed trees considering ingroup sequences, was again seen to be the sequence of the Wuhan 2019 strain. Phylogenetic trees Fig. 2 were built independently for all the proteins separately except for ORF7b, because it has only three different sequences those differed from the root sequence by one amino acid. The result showed that the initial sequence or the ancestral origin is the same for all of the proteins (YP009724390, YP009724391, YP009724392, YP009724393, YP009724394, YP009724395, YP009724396, YP009724397). While constructing the ancestral sequences i.e. the closest ancestor from which the SARS-CoV-2 sequences emerged considering the closest relative i.e. Bat Coronavirus as an outgroup it was seen that the reconstructed sequences were highly similar to the Wuhan strain of the virus. The sequences were absolutely identical for proteins M, S and ORF8 while for E there is one substitution and one addition, which has been removed in the course of evolution. Similarly in case of ORF3a there is one substitution while for ORF6 and N there is one substitution and two amino acids addition for both proteins. These additional amino acids indicate that evolution has shortened the sequence lengths in these proteins. From the data of ingroup sequences, it can be seen that the Wuhan strain is actually the ancestral strain, while taking in consideration the outgroup sequences it can be seen that the Wuhan strain is very close to the ancestral strain which in terms of substitution is one amino acid apart in few cases while in others they are absolutely identical.

2.4 Mutation Profiling of SARS-CoV-2 Proteins Using Independent and Epistatic Models

The resulting data of all proteins here, follows a trend of having an overall impact of the mutations on the phenotype of the protein and not just on their stability. The effect of the mutations on the stability is mostly additive because, according to the independent model, it considers the effects of each mutation to be independent of the other mutations and thus shows additive nature being mutually exclusive events. But, in some cases their effects may not be independent of each other and influence each other as well (epistatic mutation).

Table 1 Number of mutations and mutational tolerance in SARS-CoV-2 proteins

Name of the protein		Stabilizing mutation		Destabilizing mutation		Total no of mutation	Mutations tolerated		% of destabilized mutation	% of tolerating mutation
		Mild	Strong	Mild	Strong		Yes	No		
Envelope (E)		0	1	1	1	3	1	2	66.7	33.3
Membrane protein (M)		0	0	7	4	11	7	3	100	63.6
Nucleocapsid (N)		17	3	34	10	64	33	31	68.75	51.6
Spike (S)		10	1	36	24	71	54	17	84.5	76.05
ORF1ab	NSP1	3	0	8	7	18	3	15	78.94	55.9
	NSP2	19	3	32	26	80	55	25		
	NSP3	26	4	62	55	147	105	23		
	NSP4	9	2	15	11	37	14	12		
	NSP5	7	0	13	4	23	12	7		
	NSP6	2	0	6	11	18	12	5		
	NSP7	1	0	3	1	5	0	10		
	NSP8	5	1	14	4	24	14	4		
	NSP9	1	0	5	3	9	5	2		
	NSP10	0	0	2	2	4	2	32		
	NSP12(RdRP)	7	0	23	17	47	15	24		
	NSP13	11	1	19	18	49	25	17		
	NSP14	6	0	17	16	39	22	12		
	NSP15	6	1	8	12	27	15	13		
	NSP16	1	0	13	8	22	9	17		
	Total	104	12	240	195	551	308	218		
ORF3a		6	1	17	17	41	6	35	82.9	14.6
ORF6		2	0	0	3	5	1	4	60	20
ORF7a		2	1	6	7	16	2	14	81.25	14.3
ORF7b		0	0	1	1	2	Not applicable		100	-
ORF8		2	0	5	8	15			86.67	-
Total		162		617		779	414	338	79.20	53.14

The intensity of different shades of blue color represent the extent of damaging influence of the mutations created upon the overall protein Figs. 3 and 4. The changes from blue to white exemplify the decreasing amount of damaging influence while the change from white to red marks the amount of beneficial influence of the mutations on the protein. Thus by comparing the results of both the models of any particular mutation on each of the proteins, it was seen that the epistatic model Fig. 4 gives a wider range of energy than the independent model Fig. 3. This indicates that the mutation which in some cases appear to be less damaging or even beneficial to some extent, actually has far-reaching effects beyond their additive nature when the change of interactions with surrounding residues is taken into consideration. Thus a mutation which is independently more or less neutral has a deeper effect which cannot be understood by looking at the mutation alone but needs to be considered as a collective. The evolutionary fitness data of a particular mutation can also be predicted from this analysis due to the fact that the model compares the effects of the mutations as a whole which would help to determine the overall stability of the proteins and their phenotypic consequence, thus provide knowledge about the possible evolutionary pathways the proteins have undergone or will undergo. The stability data also provide information about how much the protein is evolutionarily fit as in order to evolve the protein has to cross the evolutionary fitness barrier which is again governed by the stability of the protein. Thus both the models suggest the impact of mutations on the fitness of the proteins in the evolutionary timescale but the data obtained by employing the epistatic model has much wider and in depth implications by taking into consideration both local and global effects of the mutation and is thus more robust of the two models used in this study.

2.5 Mutation Stability Analysis of SARS-CoV-2 Proteins

Each mutation was considered separately as single point mutation for most of the sequences Table 1. In most cases it has been seen that these mutations show additive effect upon protein stability according to the independent model of mutation. Thus, these single point mutations can be extrapolated to multiple mutations in cases where applicable. When a single residue has been mutated to multiple residues Table 2, they have been considered as separate mutations. In Table 2, the difference of energy occurred as a result of the mutations has been recorded and those data imply the stability of the proteins on the basis of change in Gibbs free energy, and the changes of amino acids from wild type to mutated strains. The data tabulated in Table 1, classified as four categories with respect to $\Delta\Delta G$ values as mild stabilizing (0–1), strong stabilizing (greater than 1), mild destabilizing (0 to − 1) and strong destabilizing (less than − 1) mutations. Table 1 data shows 617 destabilizing mutations and 162 stabilizing mutations i.e. destabilizing mutations are 79.20% of the total mutations (779). So, it can be seen that the destabilizing mutations dominate over the stabilizing ones and thus leads to a decrease in stability of the proteins in turn promoting further evolution. Both $\Delta\Delta G$ values (25 °C and 37 °C) were compared using two tailed paired T-test (at 0.05 significant level). For all eight proteins, (except ORF 7b and envelope due to low sample size) results indicate non-significant difference between the two groups Table 3.

The effect of mutation on stability were mostly in agreement to the data generated based on sequence analysis. For Orf8, it was seen that 80% of the mutations were destabilizing at the structure level, for nucleocapsid this percentage was 77.7%, 75% for ORF3a and 61% for spike. It was seen that for a majority of the proteins, full structure could not be constructed and thus the effect of all the mutations could not be studied at the structural level due to the constraint of non-availability of the region in the respective structures. This makes the predictions at the sequence level essential for these mutations as no structural data is available for all the proteins.

2.6 Effect of Mutations on Protein Function

Analyzing the tolerance of mutation on the individual proteins taking the Wuhan reference strain protein sequence as the reference it was seen that the overall tolerance level varied based on proteins Table 1. In some proteins like spike glycoprotein (S), the majority of mutations are tolerant which makes the virus infective in spite of such high mutations in the protein. This data proves that though the protein may lose out on stability due to the mutations, its function is not compromised. On the other hand, there are proteins like ORF3a and ORF7a where most of the mutations are non-tolerant as their functions get compromised which points to the fact that there is little acceptance to the change in the particular proteins. In the RNA directed RNA polymerase (RdRP) the prevalence of non-tolerating mutations show that the protein needs specific residues to carry out its function and mutations compromise the function in most cases and may also play a role in decreasing infectivity of the virus with time. Thus it has been seen that there is no uniform rule for the functional impact of the mutations on the protein. The impact of mutations vary from protein to protein and is based mainly on their evolutionary conservancy, which directly points to the functional impact of the proteins in the virus lifecycle as it is the functionally important residues which have been seen to be more resilient to mutations (to preserve their function) and highly conserved across evolution. Conversely, there are proteins which needs

Table 2 The Gibbs free energy changes of four SARS-CoV-2 proteins due to the corresponding mutational changes

Name of the protein	Amino acid position	$AA_{wild\ type}$	AA_{Mutant}	ΔΔG at 25ºC	ΔΔG at 37ºC
Membrane protein (M)	2	A	V	− 0.39	− 0.29
			S	− 0.46	− 0.45
	3	D	G	− 0.49	− 0.22
	38	A	S	− 0.29	− 0.27
	70	V	I	− 1.27	− 1.11
			S	− 3.66	− 3.36
	85	A	S	− 0.76	− 0.74
	155	H	Y	− 0.22	− 0.21
	175	T	M	− 0.59	− 0.57
	190	D	N	− 1.85	− 1.65
	209	D	Y	− 1.85	− 1.73
Envelope (E)	36	A	V	+ 1.47	1.59
	37	L	H	− 1.76	− 1.74
	71	L	P	− 1.34	− 1.27
ORF6	8	Q	H	− 1.9	− 1.87
	9	V	F	− 2.36	− 2.43
	34	S	N	0.85	1
	42	K	N	− 1.41	− 1.29
	61	D	Y	0.97	1.07
ORF7b	19	F	L	− 1.72	− 1.63
	28	F	Y	− 0.89	− 0.86

to evolve with time to maintain its infectivity and thus needs to be more flexible with respect to mutations.

2.7 Discussion

In this short span of time SARS-CoV-2 has mutated itself which has led to its strain diversification. Here, first using the clustering the different types of sequence were identified which helps to know the relative percentages of the different sequences in existence. The less amount of divergence per residue indicates the fact that the virus has not originated a very long time ago as with passage of time the amount of divergence goes on increasing as can be seen in the case of HIV-1 [20].

The presence of one phenotypic mutation in most of the cases leads to the fact that the virus is relatively new. The mutation profile indicates the overall influence of all the mutations possible in these proteins. Most of these mutations have not been reported till now and many may never be reported as well due to their destabilizing effects on the virus. From the overall view of the effects of all of these mutations on these proteins it can be said that majority of the mutations are destabilizing in nature to the respective proteins which in turn acts as an incentive for their further evolution so that they can survive to attain a higher stability and not be rejected during natural selection [21].

In this study, through mutation analysis, it has been seen that the predominant of the mutations are destabilizing in nature ranging from 60 to 100% [22]. Thus the mutations lead to a decrease in stability of the proteins by both structural and functional aspects. It has been seen in several studies that the increase in protein core size leads to an increase in the mutational robustness of the proteins thus making them more susceptible to the mutations which occur as a result of polymerase errors, host encoded mutation rate modifiers besides the fact that the virus contains a single stranded RNA genome [23]. It is seen here that larger proteins which are not disordered also undergo the same as is evident from 71 mutation sites in Spike protein (1273 amino acids) to just two mutation sites in protein ORF7b (43 amino acids). Thus it can be said that larger protein cores can accumulate more mutations leading to more mutational robustness. In both eukaryotes and prokaryotes it has been seen that the median of the mutations lie in the destabilizing range, the number being further more for archaea. Thus larger protein cores accumulate more destabilizing mutations which also reflects the strong stability determined constraints on proteins. Studies have shown that highly stable proteins have slower evolutionary rates as stability retards structure evolution owing to the formidable fitness valley barriers. Evolution of new structure occurs by passing through a series of unstable intermediates. This formation of intermediates is hindered by the high fitness barrier. This can also be proved by the

Table 3 Statistically significant difference between ΔΔG data set of eight SARS CoV-2 proteins at 25 °C and 37 °C (except Envelope and ORF7b due to low sample size)

Protein	T value	p value	Statistical significance
Membrane (M)	0.26132	0.796515	Non-significant
Nucleocapsid (N)	0.60242	0.547981	Non-significant
Spike (S)	0.43077	0.66731	Non-significant
ORF1ab			
NSP1	0.24095	0.81104	Non-significant
NSP2	0.50945	0.61115	Non-significant
NSP3	0.66357	0.507488	Non-significant
NSP4	0.14688	0.883634	Non-significant
Protease	0.43088	0.668568	Non-significant
NSP6	0.06324	0.949926	Non-significant
NSP7	0.11161	0.913886	Non-significant
NSP8	0.35656	0.723051	Non-significant
NSP9	0.18162	0.85816	Non-significant
NSP10	0.29084	0.780965	Non-significant
RdRp	0.34436	0.731363	Non-significant
NSP13	0.2186	0.827428	Non-significant
NSP14	0.43463	0.665065	Non-significant
NSP15	0.18243	0.855952	Non-significant
NSP16	0.37798	0.707346	Non-significant
ORF3a	0.14018	0.888874	Non-significant
ORF6	0.06503	0.949744	Non-significant
ORF7a	0.1184	0.906516	Non-significant
ORF8	0.14751	0.883788	Non-significant

fact that the abundant proteins which have evolved towards greater stability have slow evolution rates [20]. Compactness of the protein leads to stability in the protein. This compactness can be described using contact density in the protein which may thus act as a proxy for thermodynamic stability which has been shown to be negatively correlated with the evolutionary rate.

For a protein to evolve, a threshold level of stability is needed. If the protein exceeds this threshold by a considerable margin it can accumulate more of these destabilizing mutations and lead to a higher evolutionary rate. But if the protein is very much close to the threshold then too much of these mutations will lead to it falling below the threshold and thus lead to loss of structure and function of the protein and eventually being removed by natural selection. This is where the role of the less prevalent stabilizing mutations come in which may play an enabling role so that to increase the breadth of these destabilizing mutations that can be accommodated [24]. These destabilizing mutations lead to structural disorder which allows the protein to occupy multiple conformations and by more evolutionary refinements can lead to specialization and further stabilization. Thus the fitness and functional advantage is given priority by selection even compromising the structural stability of the protein [25].

An example for the same can be considered in the form of HIV-1. In spite of using combinatorial drugs, the problem with HIV-1 is that they frequently become drug-resistant. Thus evolutionary trace back of the same was done to arrive at evolutionary primitive sequences. When these sequences were compared with drug resistant strains for amino acid replacements it was seen that the predominant of these substitutions were destabilizing in nature. The overall evolutionary history of HIV-1 has shown a tendency for decrease in stability [26]. When the non-drug resistant strains were considered it was seen that the specific accessory mutations spotted were mostly stabilizing in nature and thus stabilized the proteins. But overall the same story was seen where the proteins in the course of evolution show a decrease in stability. It has been known for a long time that HIV-1 has a very high rate of mutation which renders the drug ineffective and a possible reason for the high rate could be this accumulation of destabilizing mutations which lower the fitness barrier. The same results in case of SARS-CoV-2 could indicate the same where the high prevalence of destabilizing mutations can be one of the primary reasons for the huge variation and strain diversification of the virus.

Based on the functional implications of the mutations, it can be seen that the impact varies according to the importance of the protein in the virus and its role in the virus lifecycle. Some mutations may lead to change in the phenotype of the protein leading to various drugs becoming ineffective in clinical trial as well as increasing the viral survival capability as can be seen in D614G mutation of spike protein [27]. It has also been seen that mutation lead to greater infectivity of the virus [28].

Proteins involved in the basic function of the virus have been primarily seen to be resilient to mutations and a majority of the mutations cannot be functionally tolerated by the proteins, this data points that the residues which are involved in mutation have a chance of affecting the active site and thus lead to a reduced efficacy in the proteins. On the other hand of the spectra, there are proteins which are involved in virus infectivity and thus to maintain the same they need to be flexible and open to varied types of mutation but still maintain the infectivity otherwise the virus would lose infectivity very easily and would perish even before entering the host cells. Proteins like the spike glycoprotein(S) which have a whole plethora of mutations which in spite of adversely affecting the protein stability have a positive role in function and thus the mutations play very important in evolution by promoting evolution in both ways.

Thus it is the mutations' effect on protein stability which reduces the fitness barrier and also on protein function that promotes the role of the mutations in the ability to evolve the virus such that it is causing a wide spread pandemic by adapting to new hosts and environments through increasing its genetic diversity in a very short period of time [23]. The data can be validated with the data obtained from SARS-CoV-1 where a specific selection for some sequences have occurred due to availability of divergence time (supplementary data).

3 Conclusion

COVID-19 caused by SARS-CoV-2 has been made an outbreak since December 2019 and become a worldwide pandemic. Disease specific medication has not yet been discovered and one of the main reason may be the mutations that occur in the protein sequences and leads to different strains of the virus. Investigation and analysis of the mutations in ten SARS-CoV-2 proteins are executed in this study using several in silico methods. Determination of variation in different protein sequences was done by clustering and identification of ancestral sequence done using both Shannon entropy calculation and phylogenetic analysis. By mutational profiling and mutation analysis, the effect of mutations on the protein stability and their functional implication were studied. This may be helpful for future study in identifying proper medication against the virus by characterizing the varying nature of the functionally important residues thus making it effective against a vast range of mutations which are or will be found in the respective proteins. Thus, these

mutations even advocate the need for combinatorial therapy for better efficacy like in the case of HIV-1 where using a single drug leads to development of resistance against the same very quickly, resulting in the need to use combinatorial therapy. Similar cases have also been seen for several respiratory viruses where again combinatorial therapy have given excellent results [29, 30].

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10930-021-09988-3.

Author Contributions Protocol designed and conceptualized by DC, manuscript preparation and data analysis done by SM and JD Project was done under the supervision of K.G. The manuscript was reviewed and approved by all authors.

Funding This study was supported by FRPDF grant of Presidency University.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Corman VM, Muth D, Niemeyer D, Drosten C (2018) Hosts and sources of endemic human coronaviruses. Adv Virus Res 100:163–188. https://doi.org/10.1016/bs.aivir.2018.01.001
2. Zhou P, Yang X, Wang X et al (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 579:270–273. https://doi.org/10.1038/s41586-020-2012-7
3. Xu X, Chen P, Wang J et al (2020) Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. Sci China Life Sci 63(3):457–460. https://doi.org/10.1007/s11427-020-1637-5
4. Tang X, Wu C, Li X, Song Y, Yao X et al (2020) On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev 7(6):1012–1023. https://doi.org/10.1093/nsr/nwaa036
5. Zhu N, Zhang D, Wang W et al (2020) A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 382(8):727–733. https://doi.org/10.1056/NEJMoa2001017
6. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D (2020) Structure, function and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 181(2):281–292. https://doi.org/10.1016/j.cell.2020.02.058
7. UniProt Consortium (2008) The universal protein resource (UniProt). Nucleic Acids Res 36(Database issue):D190–D195. https://doi.org/10.1093/nar/gkm895
8. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13):1658–1659. https://doi.org/10.1093/bioinformatics/btl158
9. Strait BJ, Dewey TG (1996) The Shannon information entropy of protein sequences. Biophys J 71(1):148–55. https://doi.org/10.1016/S0006-3495(96)79210-X
10. Reza FM (1994) An introduction to information theory. Dover Publications, Inc, New York
11. Trifinopoulos J, Nguyen LT, Haeseler AV, Minh BQ (2016) W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res 44(W1):W232–W235. https://doi.org/10.1093/nar/gkw256
12. Letunic I, Bork P (2019) Interactive tree of life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47(W1):W256–W259. https://doi.org/10.1093/nar/gkz239
13. Hopf TA, Green AG, Schubert B et al (2019) The EVcouplings Python framework for coevolutionary sequence analysis. Bioinformatics 35(9):1582–1584. https://doi.org/10.1093/bioinformatics/bty862
14. Rollins N, Brock K, Rollins J, Shen J, Tam A, Shaw A, Bricken T, Luna A, Gauthier N, Hopf T, Sander C, Marks D (2020) SARS-CoV-2 mutation effects and 3D structure prediction from sequence covariation
15. Capriotti E, Fariselli P, Casadio R (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33(Web Server issue):W306–W310
16. Capriotti E, Fariselli P, Calabrese R, Casadio R (2005) Predicting protein stability changes from sequences using support vector machines. Bioinformatics 21(Suppl. 2):ii54–ii58
17. Capriotti E, Calabrese R, Casadio R (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22:2729–2734
18. Pires Douglas E. V, Ascher David B, Blundell Tom L (2014) DUET: a server for predicting effects of mutations on protein stability via an integrated computational approach. Nucleic Acids Res 42(W1):W314–W319. https://doi.org/10.1093/nar/gku411
19. Sim NL, Kumar P, Hu J, Henikoff S, Schneider G, Pauline CN (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Res 40(W1):W452–W457. https://doi.org/10.1093/nar/gks539
20. Serohijos AW, Rimas Z, Shakhnovich EI (2012) Protein biophysics explains why highly abundant proteins evolve slowly. Cell Rep 2(2):249–256. https://doi.org/10.1016/j.celrep.2012.06.022
21. Jay FS (2018) Compensatory mutations and epistasis for protein function. Curr Opin Struct Biol 50:18–25. https://doi.org/10.1016/j.sbi.2017.10.009
22. Faure G, Koonin EV (2015) Universal distribution of mutational effects on protein stability, uncoupling of protein robustness from sequence evolution and distinct evolutionary modes of prokaryotic and eukaryotic proteins. Phys Biol 212(3):035001. https://doi.org/10.1088/1478-3975/12/3/035001
23. Sanjuan R, Domingo-Calap P (2016) Mechanisms of viral mutation. Cell Mol Life Sci 73:4433–4448. https://doi.org/10.1007/s00018-016-2299-6
24. Petrie KL, Palmer ND, Johnson DT et al (2018) Destabilizing mutations encode nongenetic variation that drives evolutionary innovation. Science 359(6383):1542–1545. https://doi.org/10.1126/science.aar1954
25. Gilson AI, Marshall-Christensen A, Choi JM, Shakhnovich EI (2017) The role of evolutionary selection in the dynamics of protein structure evolution. Biophys J 112(7):1350–1365. https://doi.org/10.1016/j.bpj.2017.02.029
26. Olabode AS, Kandathil SM, Lovell SC, Robertson DL (2017) Adaptive HIV-1 evolutionary trajectories are constrained by protein stability. Virus Evol 3(2):vex019. https://doi.org/10.1093/ve/vex019
27. Chen J, Wang R, Wang M, Wei GW (2020) Mutations strengthened SARS-CoV-2 infectivity. J Mol Biol 432(19):5212–5226. https://doi.org/10.1016/j.jmb.2020.07.009
28. Plante JA, Liu Y, Liu J, Xia H et al (2020) Spike mutation D614G alters SARS-CoV-2 fitness. Nature. https://doi.org/10.1038/s41586-020-2895-3
29. Hayden FG (1996) Combination antiviral therapy for respiratory virus infections. Antiviral Res 29(1):45–48. https://doi.org/10.1016/0166-3542(95)00914-0

30. Pirrone V, Thakkar-Rivera N, Jacobson JM, Wigdahl B, Krebs FC (2011) Combinatorial approaches to the prevention and treatment of HIV-1 infection. Antimicrob Agents Chemother 55(5):1831–1842. https://doi.org/10.1128/AAC.00976-10