MUTAGENIC DISTINCTION BETWEEN THE RECEPTOR-BINDING AND FUSION SUBUNITS OF THE SARS-COV-2 SPIKE GLYCOPROTEIN

ROBERT PENNER

Abstract. We observe that a residue R of the spike glycoprotein of SARS-CoV-2 which has mutated in one or more of the current Variants of Concern or Interest and under Monitoring rarely participates in a backbone hydrogen bond if R lies in the S_1 subunit and usually participates in one if R lies in the S_2 subunit. A possible explanation for this based upon free energy is explored as a potentially general principle in the mutagenesis of viral glycoproteins. This observation could help target future vaccine cargos for the evolving coronavirus as well as more generally.

Introduction

This short note isolates a specific and elementary observation about Protein Data Bank (PDB) [1] files concerning the mutated residues in the current Variants of Concern and of Interest plus the Variants under Monitoring, as per [10], of the SARS-CoV-2 spike glycoprotein S. This observation has not, to our knowledge, appeared in the literature other than in our own earlier work [13] in the context of specific Variants of Concern, and it may be material going forward in designing mRNA or other types of vaccine cargos as the coronavirus continues to evolve. It is anyway worth considering, in this by-now highly studied example, as a potentially more general statement about viral glycoprotein mutagenesis, since we provide a possible explanation for our observation based upon general principles involving free energy.

To state this fact about mutagenic pressure in the spike, recall that as in many examples of viral glycoproteins, in particular commonly with Class I fusion mechanisms [16, 4, 8], S is composed of two subunits S_1 and S_2. S_1 mediates receptor-binding extracellularly, and S_2 mediates fusion within an endosome. One particularity of S is a host-furin mediated cleavage between S_1 and S_2, at residue number 685-686. There is furthermore a second cleavage lying adjacent to the fusion peptide, mediated by host-cathepsin or serine protease, of a C-terminal segment

Date: November 11, 2021.

It is a pleasure to thank Pablo Guardado-Calvo for critical comments.

S_2' of S_2, at residue number 815-816. See [2] for more information on cleavage in the SARS-CoV-2 spike.

In any protein, hydrogen bonds form between backbone Nitrogen atoms N_i-H_i and Oxygen atoms O_j=C_j in the peptide unit, and these are called Backbone Hydrogen Bonds (or BHBs). (To be precise: A DSSP [9] hydrogen bond is accepted as a BHB provided that furthermore the distance between H_i and O_j is less than 2.7 Å and $\angle NHO$ and $\angle COH$ each exceed 90°.) A protein residue R_i itself is said to *participate* in a BHB if either the nearby Nitrogen N_i-H_i donates to or the nearby Oxygen O_i=C_i accepts a BHB. (Again to be precise, if at least two monomers of the trimeric spike participate, then the residue itself participates.) On average for all proteins, roughly 70-80% of all residues participate in a BHB [7].

Here is the main easily confirmed empirical observation of this paper, which is quantified in Table 1 and subsequently discussed:

> A residue R of the SARS-CoV-2 spike glycoprotein S which has mutated in one or more of the current Variants of Concern or Interest or under Monitoring [10] (cf. Table 1 for these mutagenic residue numbers), rarely participates in a BHB if R lies in S_1 and usually participates in a BHB if R lies in S_2.

A general possible explanation for this involves the free energy of structural details stabilized by BHBs. Namely, viral glycoproteins which mediate receptor binding and membrane fusion are by their very nature metastable. It follows that successful viral mutation can neither increase free energy by so much as to disturb stability of the molecule nor decrease it by so much as to interrupt near-instability, for otherwise the molecule will respectively either explode or fail to reconform and function correctly. The minimal way to avoid this twofold constraint is to mutate residues that do not participate in BHBs at all, and that is precisely what we find in S_1 before mutation.

We shall discuss S_2 subsequently, only after including certain salient definitions, facts, and data, and note in this Introduction simply that the existence of BHBs and their free energies are obviously functions of pH. This alone might account for differences between S_1 and S_2, since the endocytotic pathway is highly acidifying.

1. Materials and Methods

As is customary, we record mutations relative to an original Wuhan genome called Wuhan-Hu-1 and its corresponding spike protein (UniProt

Code P0DTC2) by considering only structure files with resolution below some bound, for our purposes resolution at most 3.0 Å, neither cleaved nor bound to antibody or receptor, and computed via cryo-electron microscopy. These 15 exemplar structures 6VXX 6X29 6X7 6XLU 6XM0 6XM3 6XM4 6ZB5 6ZGE 7A4N 7AD1 7DDD 7DF3 7DWY 7JWY for S from the PDB depend upon various techniques of stabilizing S in its prefusion conformation. The molecules are therefore not truly identical, hence the utility of taking consensus and average data across the collection of PDB files, as we shall do.

Some of the previous considerations can be calibrated by employing a new concept and quantity in structural biology, the so-called backbone free energy (BFE) from [11], which can be computed from a PDB file, to be called simply a *structure*. Roughly, the BFE of a structure stabilized by a BHB is computed from geometry by comparing the planes containing the peptide units of the donor and accepter of the BHB, and applying the Pohl-Finkelstein quasi Boltzmann Ansatz [15, 5, 6]. See [11] for the details of this application in the general case of protein backbones, [12] for application to coronavirus spikes, and [13] for the SARS-CoV-2 spike S in particular.

There is a trichotomy of possibilities for a residue R in a specific structure: R may be modeled in the structure and participate in a BHB or not, and in this latter case we say R is *absent*, but R may also simply be *missing* from the PDB file. (As before, these properties of residues are taken as consensus data from the three monomers.) R can be missing for a number of simple reasons: the protein may be disordered at R; the experiment may be inaccurate or problematic at R; the data and its refinement may not model R to within reasonable parameters; or R may be C-terminal or N-terminal to the synthesized sub-peptide of the protein S.

The average of resolutions in our collection of structures is 2.77 Å, of Clashscores is 4.19, and of percentage Ramachandran outliers is 0.1, so these are all high-quality experimental structures. As argued in [13], it follows that the first among the possibilities for R missing is the most likely, so one might *conflate missing with disordered* for high quality structures within PDB-range. The consensus range of our collection of structures for S is comprised of residue numbers 27 to 1147.

A residue that is missing or absent is said to be *unbonded* and is *bonded* otherwise. If a residue R_i is bonded, then it participates in a BHB so there is either a BHB with donor N_i-H_i or one with acceptor O_i=C_i or both, and the BFE of the residue R_i is defined to be the maximum of the BFEs of these one or two BHBs, first averaged over

the two or three monomers. If a residue is unbonded, then its BFE is undefined.

The basic fact, established in [11] for viral glycoproteins, is that residues of large BFE target locations of large conformational change in the backbone, in particular typically including the fusion peptide. Specifically to give a quantitative sense to what follows, the range of BFE values is -2.9 to +6.85 kcal/mole with approximate 50th, 90th, and 99th percentile cutoffs respectively given by 1.4, 4.6, and 6.6. The validated hypothesis is that if the BFE of a residue lies in the 90th percentile, i.e., is at least 4.6 kcal/mole, then within one residue of it along the backbone, the sum of the two adjacent backbone conformational angles changes by at least 180 degrees. The converse does not hold.

2. Results

As argued before in order to preserve metastability of the molecule, the change in BFE of a mutation must be more or less conserved by the BFE across the spike, which is depicted in Figure 1. Higher BFE is evidently concentrated in S_1 compared to S_2. The several regions of meaningful negative BFE are illustrated in the figure by the intersections of the plot with the grey horizontal line, which correspond to nearly ideal α helices. Notice that each cleavage site is surrounded by a region of high BFE, and likewise for the two ends of HR1.

However as depicted in Figure 2, the single mutation D614G, which quickly globally overtook Wuhan-Hu-1 as the predominant strain, alters BFE along the entire backbone by as much at 5.10 kcal/mole at residue 134, whereas by only 0.14 kcal/mole at residue 614 itself. Thus, a single local change of primary structure can engender a long-range change of BFE across the whole spike glycoprotein.

Table 1 presents findings and data about the mutating residue numbers M common to one or more of the variants under consideration here. Specifically the table summarizes BFEs and numbers of absent, missing, and unbonded residues in each of the molecules S, S_1 and S_2 as well as in their respective intersections $\bar{S} = S \cap M, \bar{S}_1 = S_1 \cap M$ and $\bar{S}_2 = S_2 \cap M$ with the mutagenic residues M under consideration.

Several trends present themselves:

- S_1 is more disorganized than S_2 (i.e., # missing is larger);
- there are more loops in S_1 than S_2 (i.e., # absent is larger);
- the BFE of S_1 is larger than S_2;
- the same three assertions hold for $\bar{S}_1$ and $\bar{S}_2$.

Moreover, this table quantifies our main finding that

Summary Statistics of Variant Mutagenic Residues

mol	#res	avg #missing	avg #absent	avg #unbonded	avg BFE
S	1121	1.39	4.91	0.35	2.41
$\bar{\mathrm{S}}$	56	4.71	5.34	0.61	2.47
S_1	655	1.85	5.52	0.42	3.16
$\bar{\mathrm{S}}_1$	43	6.14	5.69	0.74	2.80
S_2	466	0.73	4.05	0.25	1.60
$\bar{\mathrm{S}}_2$	13	0.00	4.15	0.15	2.14

Table 1. Summary statistics for the PDB-covered submolecule S of the SARS-CoV-2 spike glycoprotein (residues 27 to 1147) and its sub-molecules S_1 (residues 27 to 681) and S_2 (residues 682 to 1147). Averages are per residue in each molecule and over monomers in the 15 structures in the third and fourth columns. A residue is *missing* if it is not modeled in the PDB file (interpreted as disorganized), it is *absent* if it occurs in the PDB file but does not participate in either nearby backbone hydrogen bond (along the backbone), and it is *unbonded* if it is either missing or absent in at least 10 of the 15 PDB files in the database for that residue. In each case, the bar over the molecule denotes the subset of mutated residues M of Wuhan-Hu-1 among the Variants of Concern or Interest and those under Monitoring, namely, residues (5 9 12 18-20 26) 52 67 69-70 75-76 80 95 136 138 144-145 152 156-158 190 215 243-244 246-253 346 417 449 452 478 484 490 501 570 614 641 655 677 679 681 701 716 796 859 888 899 950 982 1027 1071 1092 1101 1118 (1176), where the residues in parentheses are outside PDB-coverage and are not reflected in this table. A residue contributes to the average free energy BFE only if it is bonded.

- the residues mutating in the variants under consideration are likely to be unbonded in S_1 and bonded in S_2,

as well as the observation that

- a greater ratio $\frac{43}{655} \approx 0.066$ of residues in S_1 are mutating in the variants under consideration than $\frac{13}{460} \approx 0.028$ in S_2.

Note that the missing and absent columns in the table come directly from the PDB and DSSP with no provisos (other than those conventions in parentheses in the text). The unbonded column depends upon a cutoff 10, namely, it is either absent or missing in at least 10 of the 15 structures. The last BFE column depends not only on the cutoff 10 but also on our theory of BHBs. All of the preceding trends are invariant under changing this cutoff by unity, with this data not presented.

3. Conclusions

We find that mutagenic pressure on S_1 exceeds that on S_2, as expected based on function and location of both subunits, and that the former is more disorganized and with a lower percentage of bonded residues than the latter. These findings are consistent with the general trend that B-factors [3] in the receptor-binding subunit usually

exceed those in the fusion subunit of a viral glycoprotein, at least in the prefusion conformation.

It is argued that mutation of unbonded residues avoids the twofold constraint on BFE imposed by metastability of the viral glycoprotein, thus explaining the tendency of mutating residues in S_1 to be unbonded.

As was already mentioned, the different pH of activation for the two subunits S_1 and S_2 may explain the opposite trend in the latter that mutating residues tend to be bonded, since the prefusion stabilized spike structures may better reflect the actual geometry and consequent BHBs of S_1 compared to S_2. Another related possibility is that pre-cleavage, S_1 sits on top of S_2 as a kind of cap, thereby sterically constraining the latter, so the active geometry of S_2 is displayed only post-cleavage and in the course of post-fusion reconformation.

In any case, the findings on S_1 suggest a strategy for anticipating residues primed for mutation therein. However going forward, it is the residues that are unbonded for the currently mutated variants, rather than for Wuhan-Hu-1, that should be considered as likely future candidates.

These considerations may be of utility in predicting future variants, not only of the SARS-CoV-2 spike, but more generally for any Class I fusion viral glycoprotein, and this may be of substance in general for vaccine design.

References

[1] Berman, H.M., Westbrook, J., Feng, Z., et al. 2000. The Protein Data Bank. *Nucleic Acids Research* 28, 235–242. Available online: `http://www.rcsb.org/pdb/`.

[2] Bollavaram, K., et al. 2021, Multiple sites on SARS-CoV-2 spike protein are susceptible to proteolysis by cathepsins B, K, L, S, and V, *Protein Science* **30**(6), 131–1143.

[3] Carugo, O. 2018. How large B-factors can be in protein crystal structures? *BMC Bioinformatics*, **19**, 61.

[4] Chernomordik, L.V., and Kozlov, M.M. 2009. Mechanics of membrane fusion, *Nature Structural and Molecular Biology*, **15**, 675–683.

[5] Finkelstein, A.V., et al. 1995, Boltzmann-like statistics of protein architectures: Origins and consequences. In B.B. Biswas and S. Roy, eds., Proteins: Structure Function, and Engineering. *Subcellular Biochemistry* **24**, 1–26, Springer, Boston, MA.

[6] Finkelstein, A.V., et al. 1995 Why do protein architectures have Boltzmann-like statistics? *Proteins* **23**, 142–150, Springer-Nature.

[7] Finkelstein, A.V., and Ptitsyn, O. 2016. Protein Physics, A Course of Lectures. 2nd edition. Academic Press, London, UK.

[8] Harrison, S.C. 2008. Viral membrane fusion, *Nature Structural and Molecular Biology* **15**, 690–698.

[9] Kabsch, W. and Sander, C. 1983, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers* **22**, 257–637. Available online: `https://swift.cmbi.umcn.nl/gv/dssp/`.

[10] Mullen, J.L., Center for Viral Systems Biology, et al. 2020, outbreak.info. Available online: `https://outbreak.info/`.

[11] Penner, R. 2020 Backbone Free Energy Estimator Applied to Viral Glycoproteins, *Journal of Computational Biology*, **27**, 10, 1495–1508, Liebert.

[12] Penner, R. 2020 Conserved High Free Energy Sites in Human Coronavirus Spike Glycoprotein Backbones, *Journal of Computational Biology*, **27**, 11, 1622–1630, Liebert.

[13] Penner, R. 2021, Antiviral Resistance against Viral Mutation: Praxis and Policy for SARS-CoV-2, *Computational and Mathematical Biophysics* **9**(1), 81–89.

[14] Penner, R., et al. 2014, Hydrogen bond rotations as a uniform structural tool for analyzing protein architecture, *Nature Communications*, **5**, 5803, Springer-Nature. Available online: `https://bion-server.au.dk/hbonds/`

[15] Pohl, F.M., 1971 Empirical protein energy maps, *Nature New Biololgy* **234**, 277–279, Springer-Nature.

[16] White, J.M., Delos, S.E., Brecher, M., et al. 2008. Structures and mechanisms of viral membrane fusion proteins: multiple variations on a common theme, *Crit Rev Biochem Mol Biol* **43**, 189–219.

Institut des Hautes Études Scientifiques, 35 route des Chartres, Le Bois Marie, 91440 Bures-sur-Yvette, France, and Mathematics Department, UCLA, Los Angeles, CA 90095, USA

E-mail address: `rpenner@ihes.fr`

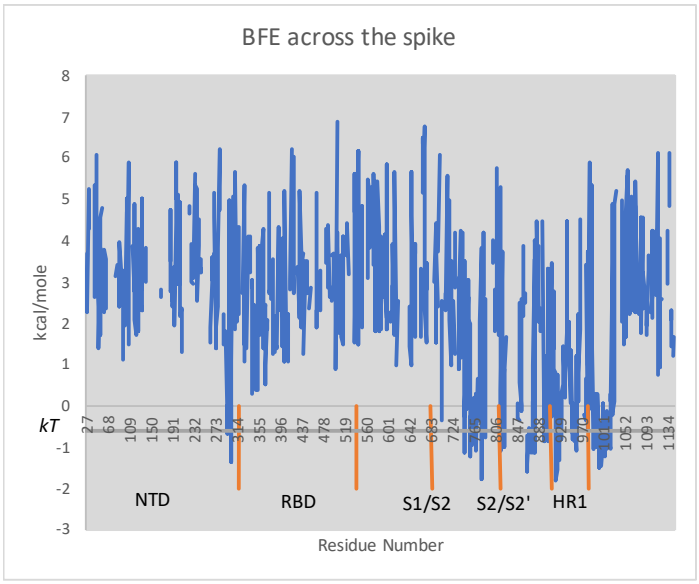

Figure 1. Plot of BFE by residue across the PDB-covered spike, residues 27-1147 of S. Illustrated in orange are the respective residue ranges for the N-Terminal Domain, the Receptor Binding Domain, the S_1/S_2 cleavage, the S_2/S_2' cleavage, and the first Heptad Repeat Domain. The grey horizontal line indicates one "heat quantum" $kT \approx 0.6$ kcal/mole. One confirms by comparison with the structure itself that the intersections of BFE with this line correspond to α helices, and in fact ones that are especially near ideal α helices according to considerations of free energy.

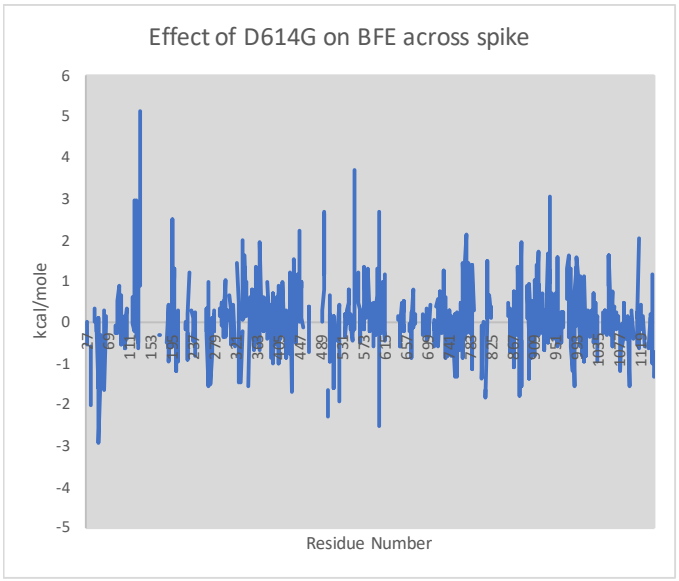

Figure 2. Comparison of BFE across the spike from the single mutation D614G. The BFE of Wuhan-Hu-1 is computed at each residue as the average of PDB structures 7KDG and 7KDH, which are stabilized in the prefusion conformation by mutations R682G, R683G, R685G plus the 2P mutation given by K986P and V987P; the BFE of the D614G mutation is computed as the average of structures 7KDK and 7KDL, analogously stabilized but also with the D614G mutation. In each case, missing or absent residues give null. Plotted is the difference of the former minus the latter. The Wuhan-Hu-1 BFE at residue 614 itself is 2.26 kcal/mol compared to 2.12 kcal/mole for D614G, but despite this near equality at residue 614, the BFE across the entire spike is altered.