## Concise mathematical statement of problem

I have 3 non-negative, bounded, real-valued functions $A(x), B(x), C(x)$ of a common discrete variable $x = 1, ..., 1000$ or so. Each of $A, B, C$ can furthermore be in one of two configurations called *up* (u) and *down* (d); this combinatorial data is fixed once and for all for a given triple $A, B, C$ for all $x$. (It is an over-simplification to report just the 2-states u and d since there are necessarily intermediate states though experimentalists do often so characterize it, and in many cases, the configuration is indeed definitive.) In fact, $A, B, C$ are *a priori* unlabelled–indistinguishable–and are labelled just so as to express the data.

There is a collection $A_n(x), B_n(x), C_n(x)$ of such data and in addition a 3-letter word in the alphabet {u,d} giving the up/down configuration of the $n$th dataset $A_n, B_n, C_n$ in this order, for $n = 1, ..., N$. There are several possible ways to group the data, and depending upon this, $N$ ranges from 1 to 40 or so.

I would like to find a method that determines which values of $x$ especially correspond to either up or down. There is clearly insufficient data for machine-learning tools. A further detail is that there is a fixed universal value $Y$, and I am only interested in finding inputs $x$ whose output is at least $Y$. For example for the word u-u-d if $A(x_0)$ and $B(x_0)$ are roughly equal and small and $C(x_0) > Y$ is large, one might think that down correlates with a large value at this $x_0$. This is just a toy example. The signal for up/down may well involve multiple values of $x$.

Just to give some actual data (from the virus that causes COVID-19), Figure 1 illustrates the three functions $A, B, C$ (here called *chains*) for three examples with configurations d-d-d (called 6VXX), d-u-d (called 6VYB) and u-d-d (called 6VSB) given in the order $A - B - C$.

There is actually another detail: $A, B, C$ are multivalued functions, but typically with only a few inputs with multiple outputs, which may vary with $n = 1, ..., N$, that are significant in the sense that the cutoff value $Y$ is achieved. Perhaps let us just ignore this aspect for the moment...though I suspect these few significant branch points are important.

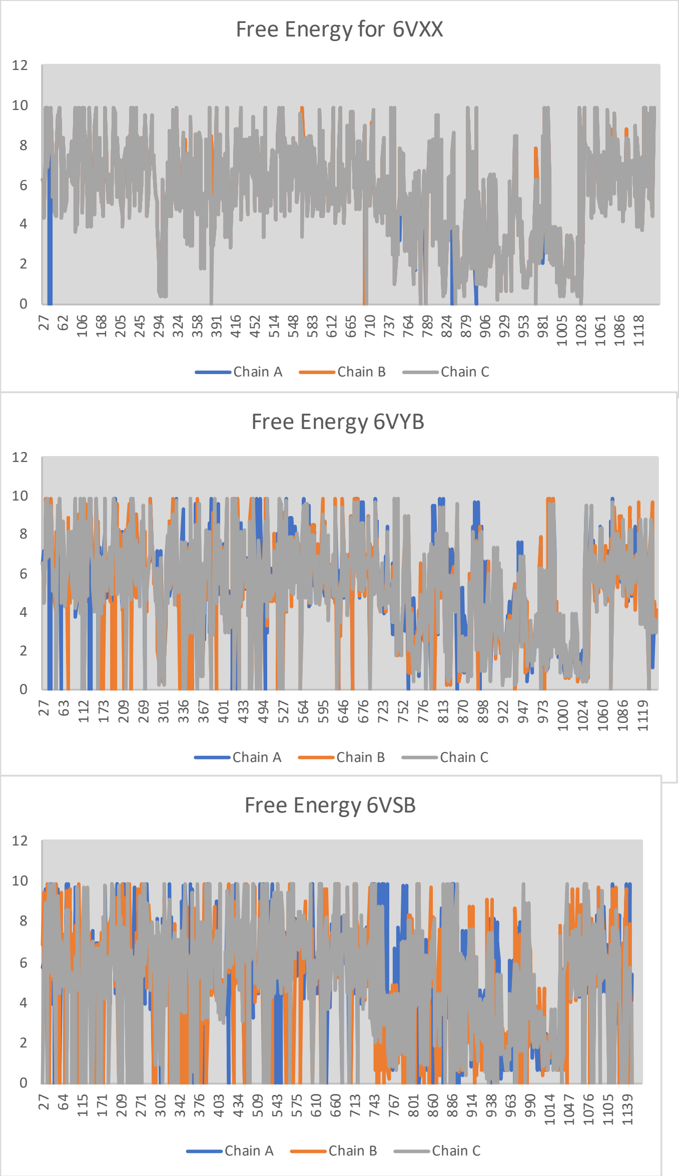

Figure 1: Plot of normalized free energy $A(x), B(x), C(x)$ for the three chains $A, B, C$ as functions of residue number $x$ for SARS CoV-2 spike glycoprotein from the PDB files 6VXX (d-d-d), 6VYB (d-u-d), and 6VSB (u-d-d).

## Narrative description of biological content of problem

One indispensable step in the successful infection of a host cell by a virus is receptor binding, whereby the viral particle recognizes and attaches to the host cell receptor. For viruses such as coronaviruses which are enveloped in a lipid bilayer, this is accomplished by structures called *viral glycoproteins* which cover the virus surface like a somewhat dense forest. An approximate scale and shape to keep in mind for the coronaviruses considered here is a roughly spherical enveloped viral particle of approximately 130nm diameter with its forest of viral glycoproteins standing at 15nm or so tall above the viral surface. The glycoprotein of a coronavirus is furthermore covered by a complicated linked collection of glycans presumably to shield it from immune system recognition, and its comparatively thin spots have been proposed as vaccine or drug targets.

For coronaviruses such as SARS CoV-1, SARS CoV-2 and MERS, respectively causing the diseases SARS (from 2003-2004), COVID-19 and MERS (2012-present), this viral glycoprotein is called the *spike*, which we imagine for a moment as a palm tree with its roots inside the viral particle, its branchless trunk or *stalk* extending distally from the virus surface and its canopy at maximum distance in a mixed metaphor now imagined as its head. Like the hell-dog Cerberus, the spike actually has three *heads* atop its stalk. The host cell receptor for SARS CoV-1 and SARS CoV-2 is ACE2 and for MERS is DPP4. In the human body, epithelial cells in both the respiratory and digestive systems are rich in ACE2 while DPP4 is ubiquitous.

Each head lies in a distinct *chain*, in this case a linear monomer comprised of roughly 1200 *residues* that occur in a linear order along what is called the protein *backbone*. The ordered set of residues in each of the three chains is identical for a fixed virus though it can vary from strain to strain especially in its head. The chains and their constituent heads are labelled $A, B, C$ though they are in fact indistinguishable.

The three heads oscillate independently between an *up* configuration (u) leaning away and splayed out from the stalk and a *down* configuration (d). (Up/down are also sometimes called open/closed and standing/lying.) In the up configuration and only in this state, each head exposes its receptor binding domain which can bind to the host cell receptor. Also in the up state is each head fully exposed to the immune system, and for SARS CoV-1 and MERS, human antibodies which defeat the viral onslaught attach to the exposed up-configured head; for MERS, an engineered antibody G4 attaches differently on the virus membrane-proximal region. The part of the head that is exposed when it is down is highly disordered providing a defense against immune system attack as well as presumably ameliorating receptor binding when it is up. Thus, each head varies between an offensive and vulnerable up and an inert and protected down state. Incapacitating this oscillation presumably either leaves the virus vulnerable to the immune system or inactivates its binding capacity.

Clearly the head must visit transitional states between up and down. For much of the data, this issue is moot since the configuration is definitively up or down as often reported by the experimentalist. However for some data, one or more heads may be intermediary. For now, this is qualitatively reported in the three-letter word of up/downs by a letter u$'$ (and d$'$ respectively) indicating a transitional state that is more up than down (and more down than up). It may be useful to quantify this.

The next indispensable step of coronavirus attack is fusion of the virus membrane with either the membrane of the host cell itself or via so-called endocytosis, namely uptake by the cell and subsequent fusion with the membrane of a vesicle surrounding it; there is recent evidence of the latter pathway. In either case, this depends upon a so-called fusion peptide and finally allows entry of the viral genome into the host cell cytoplasm for replication there in the case of coronaviruses. The presumptive fusion peptide for coronaviruses lies centrally between the heads and is covered and protected from immune system recognition when the heads are down.

Although the precise mechanism is unknown for coronaviruses, once the head binds to the host cell receptor, parts of the three heads are cleaved away followed by large conformational changes exposing the fusion-ready peptide. For SARS CoV-2 there is actually a further cleavage similar to MERS but not shared by SARS Cov-1. In any case, the entire spike reconforms to produce what is called a six-helix-bundle that is characteristic of so-called Class I fusion glycoproteins.

These reconformations are driven neither by chemical reaction nor by dephosphorylization but rather by hydrogen bonding. Indeed, the spike glycoprotein should be thought of as a mechanical device primed for reconformation which is triggered by suitable stimuli such as host receptor binding. In order to accomplish such hydrogen bond-based reconformation, certain geometric motifs of spike are especially unstable or in other words harbor reservoirs of high *free energy* compensated by other regions of low free energy in order to stabilize the overall protein structure. The free energy of a residue is defined to be the maximum of the free energies of its one or two adjacent hydrogen bonds if either exists and is undefined otherwise, and a residue is called *exotic* if the free energy of any adjacent bond lies in the top tenth percentile.

By combining essentially geometric methods with a Boltzmann-like Ansatz, the free energy of residues can be estimated for each chain as explained in

[Penner; Backbone free energy estimator applied to viral glycoproteins; next issue of Journal of Computational Biology; IHES preprint M/20/07].

The input to the method is the so-called *PDB file* which gives the spatial locations of each of the constituent atoms in a protein. (PDB=Protein Data Bank=the repository of all known 3d protein structures, which can be accessed by typing RCSB into your browser.)

For a given PDB file of the spike glycoprotein, which captures spike in a specific conformation, this provides three functions $A(x), B(x), C(x)$ giving the free energy at the residue $x$ for each chain $A, B, C$. Because hydrogen bonds can be bifurcated, these are in fact multivalued functions, but in practice exotic bifurcated bonds are rather rare. For the spikes of the viruses SARS CoV-1 and CoV-2, there are exotic bifurcated bonds, and for MERS there is even an exotic trifurcated bond in the data. (Here we distinguish between the PDB data and the realistic physics.) In addition for each PDB file, a three-letter word in the alphabet $\{$u,d$\}$ is determined expressing the respective up/down conformation of the heads of chains $A, B, C$ sometimes with u/d replaced by u$'$/d$'$ for transitional states as discussed before.

It is not necessarily the case that the sought-after exotic regions that correlate with up/down are nearby the head, neither nearby in 3-space nor nearby in residue. Furthermore, the best-possible solution would be one that lies on the stalk in a region not only conserved from strain to strain, but even better if it were conserved across different human coronaviruses.

As to the scale of the normalized free energy of a residue, which is denoted $\Pi$, so $A(x) = \Pi(x)$ for residue $x$ in chain $A$ for instance, we have $0 \leq \Pi \leq 9.85$, and the respective 90th, 95th, 99th and 100th percentile cutoffs for $\Pi$-values of hydrogen bonds are given by $\Pi = 7.5, 8.5, 9.5$ and $9.85$. The units in which $\Pi$ is measured are $kT_C$, where $k$ is the Boltzmann constant and $T_C$ is an effective temperature, the *conformational temperature* roughly equal to the protein melting point, which is typically near 350 degrees Kelvin. Insofar as $T_C$ varies from one protein to another, this must be taken into account when comparing these free energies across different proteins. A natural shift based on the generally accepted free energy of -2 kcal/mole for a turn interior to an ideal right alpha helix puts the zero-point absolute free energy at roughly $\Pi = 2.9$ but again this depends upon the protein itself through $T_C$.

Table 1 contains all the files for the spike of SARS CoV-1, SARS CoV-2, MERS and a handful of other human coronaviruses in the PDB at this moment together with their other pertinent information.

## Explanation of Data

Let us next describe the data at hand. Fix a PDB file, say 6VSB for SARS CoV-2. The data for this is in the folder francois_SARS-CoV-2_20200323 lying in the folder 6vsb, which in turn contains two files relevant to the current discussion, namely, two comma-separated value files:

FILE1=6vsb_merged_stress_density_full_graph.csv

FILE2=6vsb_merged_stress_density.csv

which are described here.

FILE1 gives data for the three chains of the spike glycoprotein, which are denoted A,B,C for 6vsb and in general with notation as in the Table. The first column gives the residue number, which is common to the three chains. Starting with chain A, columns 2,3,4,5 give the free energies $\Pi$ of the backbone hydrogen bonds (BHBs) corresponding to the residue-row in non-increasing order of magnitude. Then columns 6,7,8,9 have entries a/d/x for acceptor/donor/absent for the respective BHBs indicated in columns 2,3,4,5. Then comes the analogous data for chain B in columns 10-17 and chain C in columns 18-25. Only the 4 highest energy BHBs adjacent to any residue are tabulated. Bifurcated and higher-order BHBs are evident from the appearance of more than one a or d in a single row for a single chain. For example for 6vsb at residue 120, there are three adjacent BHBs: a bifurcated donor and a singleton acceptor, but none of these are exotic, and so neither is residue 120, since the free energies are all less than $\Pi = 7.5$.

FILE2 presents all data for all chains in a PDB file in a different format as next described; spike glycoproteins in complex with antibodies/receptors or with duplicate molecules can have many chains as indicated in the Table. The first column merely enumerates the rows in the file, and there is one row in FILE2 for each BHB in the PDB file. Please see reference above for details on the determination of BHBs from PDB file. The next two columns give the donor of the hydrogen bond, first the chain and then the residue-1. (The -1 is because the numeration is that of peptide groups in the computer code producing this file.) The next two rows give the acceptor, first the chain and then the residue itself (with no -1). The columns for cluster and stress should be ignored. The last -log Density column gives the normalized free energy $\Pi$ of the BHB.

(The so-called B-factors for the donor and acceptor residues are respectively given in columns 8 and 9 and can be ignored at first glance. The B-factor gives a measure of the disorder of the protein at the residue...or equally might reflect experimental error of measurement there. B-factors less than 60 are acceptably small and greater than 100 indicates error or disorder.)

For example, row 3 of FILE2 indicates a BHB with donor given by chain A/residue 44 and acceptor given by chain C/residue 565 with normalized free energy $\Pi = 9.47$, a highly exotic and hence significant BHB (with rather large B-factors however).

Table 1: **HCoV spike glycoprotein up/down conformations**

| PDB | Res.(Å) | Chain/Conf. | Other Chains | Comments |
|---|---|---|---|---|
| | | | **SARS CoV-2** | |
| 6VXX | 2.8 | A-B-C/d-d-d | none | 2P and C-terminal foldon |
| 6VYB | 3.2 | A-B-C/d-u-d | none | 2P and C-terminal foldon |
| 6VSB | 3.5 | A-B-C/u-d′-d′ | none | 2P and C-terminal foldon |
| | | | **SARS Cov-1** | |
| 5WRG | 4.3 | A-B-C/d-d-d | none | Arg667 mutation to Ala |
| 5XLR | 3.8 | A-B-C/d-d-d | none | Arg667 mutation to Ala |
| 5X58 | 3.2 | A-B-C/d-d-d | none | S2-cleavage site mutation |
| 5X5B | 3.7 | A-B-C/u-d-d | none | S2-cleavage site mutation |
| 6CRV | 3.2 | A-B-C/u-u-u | none | trypsin stabilized 2P, C3 symmetry |
| 6CRW | 3.9 | A-B-C/d-u-d | none | trypsin stabilized 2P |
| 6CRX | 3.9 | A-B-C/u-u-d | none | trypsin stabilized 2P |
| 6CRZ | 3.3 | A-B-C/u′-u-d′ | none | trypsin stabilized 2P, claims C3 symmetry |
| 6CS0 | 3.8 | A-B-C/d′-u-d′ | none | trypsin cleaved |
| 6CS1 | 4.6 | A-B-C/u-u-d′ | none | trypsin cleaved |
| 6CS2 | 4.4 | A-B-C/u′-u-u′ | D=ACE2 | trypsin cleaved, B receptor bound |
| 6ACC | 3.6 | A-B-C/d-d-d | none | +ACE2, receptor bound but removed<br>trypsin cleaved, low pH |
| 6ACD | 3.9 | A-B-C/d-d-u | none | +ACE2, receptor bound but removed<br>trypsin cleaved, low pH |
| 6ACG | 5.4 | A-B-C/d-d-u | D=ACE2 | C receptor bound<br>trypsin cleaved, low pH |
| 6ACJ | 4.2 | A-B-C/d-d-u | D=ACE2 | C receptor bound<br>trypsin cleaved, low pH |
| 6ACK | 4.5 | A-B-C/d-d-u | D=ACE2 | C receptor bound<br>trypsin cleaved, low pH |
| 6NB6 | 4.2 | A-B-C/d-u-u′ | H,I=$S230_h$<br>L,M=$S230_\ell$ | B,C antibody bound 2P |
| 6NB7 | 4.5 | A-B-C/u′-u′-u′ | D,G,H=$S230_h$<br>E,I,L=$S230_\ell$ | A,B,C antibody bound 2P |

Subscript $h, \ell$ respectively denote heavy, light and 2P denotes 2 proline point mutation to stabilize prefusion

Table 1: (continued) **HCoV spike glycoprotein up/down conformations**

| PDB | Res.(Å) | Chain/Conf. | Other Chains | Comments |
|---|---|---|---|---|
| | | | **MERS** | |
| 5W9H | 4.0 | p-q-r/u-d-d | B,E,H=$G4_h$; C,F,I=$G4_\ell$<br>A,D,G=spike | antibody bound 2P |
| 5W9I | 3.6 | B-J-F/u-u-u | C,G,K=$G4_h$; D,H,L=$G4_\ell$<br>A,E,I=spike | antibody bound 2P |
| 5W9J | 4.8 | J-K-L/d-d-d | E,B,H=$G4_h$; F,C,I=$G4_\ell$<br>D,A,G=spike | antibody bound 2P |
| 5W9K | 4.6 | J-K-L/u-d-u | B,E,H=$G4_h$; C,F,I=$G4_\ell$<br>A,D,G=spike | antibody bound 2P |
| 5W9L | 4.8 | B-C-J/u-d-d | E,H=$G4_h$; F,I=$G4_\ell$<br>A,D,G=spike | antibody bound 2P |
| 5W9M | 4.7 | E-F-J/u-d-d | B,H=$G4_h$; C,I=$G4_\ell$<br>A,D,G=spike | antibody bound 2P |
| 5W9N | 5.0 | H-I-J/u-d-d | B,E=$G4_h$; C,F=$G4_\ell$<br>A,D,G=spike | antibody bound 2P |
| 5W9O | 4.5 | J-K-L/u-d-d | B,E,H=$G4_h$; C,F,I=$G4_\ell$<br>A,D,G=spike | antibody bound 2P |
| 5W9P | 4.0 | A-C-I/u-u-u | F,D,K=$G4_h$; G,E,L=$G4_\ell$<br>J,B,H=spike | antibody bound 2P |
| 6Q04 | 2.5 | A-B-C/d-d-d | none | glycan 5-N-acetyl neuraminic acid bound |
| 6Q05 | 2.8 | A-B-C/d-d-d | none | glycan sialyl-lewisX bound |
| 6Q06 | 2.7 | A-B-C/d-d-d | none | glycan 2,3-sialyl-N-acetyl-lactosamine |
| 6Q07 | 2.9 | A-B-C/d-d-d | none | glycan 2,6-sialyl-N-acetyl-lactosamine |
| 6PZ8 | 4.2 | B-J-F/u-u-u | H,C,D=$G2_h$; L,G,K=$G2_\ell$<br>A,E,I=spike | antibody bound |
| 5X59 | 3.7 | A-B-C/u-u-u | none | S2-cleavage site mutation |
| 5X5C | 4.1 | A-B-C/u-d-u | none | S2-cleavage site mutation |
| 5X5F | 4.2 | A-B-C/d-d-u | none | S2-cleavage site mutation |
| 6NB3 | 3.5 | A-B-C/d-d-u | L,E=$LCA60_\ell$; H,D=$LCA60_h$ | A,C antibody bound, 2P |
| 6NB4 | 3.6 | A-B-C/u-d-u | H=$LCA60_\ell$; L=$LCA60_h$ | A antibody bound, 2P |
| | | | **Other** | |
| 6U7H | 3.1 | A-B-V/d-d-d | none | HCoV-229E, 2P |
| 6OHW | 2.9 | A-B-C/d-d-d | none | HCoV-OC43, 2P |
| 6NZK | 2.8 | A-B-C/d-d-d | none | HCoV-OC43, 2P<br>receptor 9-*O*-Ac-Me-Sia bound |
| 5SZS | 3.4 | A-B-C/d-d-d | none | HCoV-NL63, extensive glycan shield |
| 5I08 | 4.0 | A-B-C/d-d-d | none | HCoV-HKU1(isolate N5)<br>mutated furin-cleavage site |

Subscript $h, \ell$ respectively denote heavy, light and 2P denotes 2 proline point mutation to stabilize prefusion. For 5W9H-5W9P, antibody G4 dimerizes spikes, and only Chain/Conf. spike is resolved in PDB file.