Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Computers in Biology and Medicine 122 (2020) 103849

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: http://www.elsevier.com/locate/compbiomed

Bioinformatics studies on a function of the SARS-CoV-2 spike glycoprotein as the binding of host sialic acid glycans

B. Robson

Ingine Inc. Cleveland Ohio USA and the Dirac Foundation, Oxfordshire, UK

ARTICLE INFO

Keywords:
Coronavirus
SARS-CoV-2
COVID-19
Bioinformatics
Spike glycoprotein
Spike protein
S1-NTD
Sialic acid
Sugar binding
Prediction of sialic acid binding

ABSTRACT

SARS-CoV and SARS-CoV-2 do not appear to have functions of a hemagglutinin and neuraminidase. This is a mystery, because sugar binding activities appear essential to many other viruses including influenza and even most other coronaviruses in order to bind to and escape from the glycans (sugars, oligosaccharides or polysaccharides) characteristic of cell surfaces and saliva and mucin. The S1 N terminal Domains (S1-NTD) of the spike protein, largely responsible for the bulk of the characteristic knobs at the end of the spikes of SARS-CoV and SARS-CoV-2, are here predicted to be "hiding" sites for recognizing and binding glycans containing sialic acid. This may be important for infection and the ability of the virus to locate ACE2 as its known main host cell surface receptor, and if so it becomes a pharmaceutical target. It might even open up the possibility of an alternative receptor to ACE2. The prediction method developed, which uses amino acid residue sequence alone to predict domains or proteins that bind to sialic acids, is naïve, and will be advanced in future work. Nonetheless, it was surprising that such a very simple approach was so useful, and it can easily be reproduced in a very few lines of computer program to help make quick comparisons between SARS-CoV-2 sequences and to consider the effects of viral mutations.

1. Introduction and review

1.1. Background

As far as is known to history, no coronavirus [1] has been as disturbing to humanity as the human pandemic of the 2019 novel coronavirus [2] now known as SARS-CoV-2. In quick response to determination of the final version of the RNA sequence of the Wuhan seafood market isolate, the present author examined functional sites of SARS-CoV-2 that are highly conserved across the coronaviruses [3,4], and which thus likely to exhibit escape mutation that can quickly undo the good work of the developers of vaccines and therapeutic agents. So far, the published papers have concerned the spike glycoprotein [3–5]. Exploration of known and newly found proteolytic cleavage sites in the spike glycoprotein of SARS-CoV-2 that are well conserved is a popular area of inquiry for SARS researchers because such sites can interact with human host airway proteases that could be the target for protease inhibitors as potential drugs (e.g. Ref. [6]). The difficulty is that inhibiting the action of human proteins could have undesirable effects on the host [6], so parallel work on other kinds of functional sites in the coronavirus proteins is of great importance.

The present paper explores another potential functional site in the spike protein, but this one is a different because the site is not a proteolytic cleavage site, and it is not well conserved, except, it is argued, for a characteristic composition of particular amino acid residues. Expressed another way, there can exist certain subsequences of a protein sequence that *are* well conserved, but only in respect to some pattern or property that is less obvious than the order of amino acids. Finding them (or as is more correctly stated, predicting them) may therefore require a more subtle and, in the present case, novel bioinformatics tool, compared with the standard bioinformatics tools which were essential in the preceding papers [3–5]. Comparisons with other proteins as described below suggest that the subsequence of interest in this paper could have a crucial function, and a high degree of conservation is, even by itself, also a clue as having a role important to the virus [5]. Hence such a site may represent a potential therapeutic target, perhaps as well as representing a synthetic vaccine target. However, until very recently, that crucial function did not even seem to be possessed by SARS-CoV and SARS-CoV-2, and the details have yet to be elucidated.

E-mail address: barryrobson@ingine.com.

https://doi.org/10.1016/j.compbiomed.2020.103849
Received 7 May 2020; Received in revised form 4 June 2020; Accepted 4 June 2020
Available online 8 June 2020

1.2. Binding sialic acid glycans - a traditional picture from the influenza virus

The particular virus function that is considered in the present paper is non-covalent binding to the sialic acid glycans, i.e. oligosaccharides or polysaccharides that contain sialic acid residues. They are sometimes called sialylated glycans. Interest in this binding arose as follows. It seems unlikely (although of course possible) that functions important for many different kinds of virus are of little importance to others, especially if they have a common lifestyle such as infection of the respiratory system or alimentary tract, typically reflected by common symptoms. If such functions are absent, it begs the question of how the virus copes. Though glycan binding of SARS-CoV and SARS-CoV-2 seems absent, diminished, or relatively neglected in the literature (see Section 1.5), many coronaviruses such as human coronavirus OC43 and bovine coronavirus appear to recognize sialic acid as a receptor. However, most biology students are more familiar with the hemagglutinin and neuraminidase of influenza, the H and N in, for example H1N1 (the numbers such as 1 being based on immunological typing of these proteins), that bind to glycans, (sugar chains, oligosaccharides or polysaccharides) at cell surfaces notably those chemically bound to membrane proteins, hence called glycoproteins, of host cells. The surfaces of many animal and all vertebrate cells are dressed with a dense and complex array of glycans primarily containing sialic acids, attached to proteins and lipids at the cell surface. Such glycans also occur to a lesser extent in other organisms, ranging from fungi to yeasts and bacteria, and they are present at the surface of many viruses derived from animal hosts. Glycans can contain several kinds of sugar, including notably sialic acid, glucose, mannose, fucose, N-Acetylglucosamine, and N-Acetylgalactosamine. The standard emotive picture is that the influenza hemagglutinin binds the cell surface glycan molecules to first locate the lung cell surface, and that the neuraminidase has a later role, to enable many thousands (perhaps hundreds of thousands of) "baby viruses", i.e. the newly formed virions, to cut their way out the protective layer of glycans when emerging from the cell. More correctly stated, when the replicated viruses bud from the host cells, they remain attached to the host-cell surface by binding between hemagglutinin and the "tips" of the glycan chains, and the neuraminidase is used to sever that link by breaking certain links between the component sugar residues (see below). Recent work has supported this long standing picture for influenza viruses, but also answers affirmatively to the question that must have arisen in many student's minds, i.e. that the neuraminidase must also be important for the virus to cut its way into the cell in the first place [7]. Any such description of entry does not, however, quite fit in with the above "more correctly stated" model for final release of the virion progeny, because it is not obvious why the incoming infecting virus should bind to the cell surface and then be made to disengage. Nonetheless, many viruses appear to need and do have an enzyme to achieve similar results, even if that enzyme is not of neuraminidase type and dissociates the virus from the cell in other ways: e.g. see Ref. [8] and discussion below. It does seem reasonable that all such similar results must provide assistance in the mobility of virus particles through the respiratory tract mucus, but a fuller picture should perhaps include the notion of "decoy" glycan molecules [9] as discussed below.

1.3. The great diversity of sialic acid glycans

The seeming challenge for research into non-covalent binding to glycans is that they are diverse molecules, and that remains true even among the sialic acid glycans. At first consideration, this would seem to suggest that the prediction of glycan binding sites will be difficult. Critical features that distinguish glycans of interest in the present paper are the terminal sugar residues on the oligosaccharide chains at the distal ends (i.e. the "tips"), where sialic acids are important, and the amino acid residue attachment site (*N*- or. *O*-) to the membrane protein at the base. The N (nitrogen) protein attachment point attachment is asparagine and the O (oxygen) protein attachment point is usually serine or threonine, but sometimes tyrosine, or occasionally other amino acids which are hydoxylated as a post-translational modification. Covalent connections of that kind are those to the host cell surface proteins but, as indicated above, it is the non-covalent binding of a virus to them that is of interest here. Covalent binding is controlled by host enzymes and so would appear, again at first consideration, to be more specific. Enveloped viruses such as SARS-CoV-2 also have their own bound sialic acid glycans (the spike protein is usually referred to as the spike glycoprotein), but these are of less direct interest here although they can clearly influence binding of a virus to various receptors.

Despite their diversity, and perhaps because of it (i.e. because that diversity implies more information content) sialic acid glycans of host cells are key molecular recognition features not only for entry of viruses such as influenza, but also in embryonic development, neurodevelopment, reprogramming, and oncogenesis. Correctly speaking, even sialic acid itself is diverse. It is a generic term for a family of derivatives of the nine-carbon sugar neuraminic acid. The sialic acid family includes some 43 derivatives of neuraminic acid, but these acids rarely appear free in nature. Members include *N*-acetylneuraminic acid, 2-keto-3-deoxy-D-*glycero*-D-*galacto*-nonulosonic acid, 5,7-diamino-3,5,7,9-tetra-deoxy-D-*glycero*-D-*galacto*- nonulosonic acid, and 5,7-diamino-3,5,7,9-tetra-deoxy-L-*glycero*-L-*manno*-nonulosonic acid. If the term "sialic acid" is used unqualified, it usually refers to the representative member of this group, *N*-acetylneuraminic acid. The variability of glycans is not random but reflects their modes of synthesis. In eukaryotes generally, a typical N-linked glycan has an initial core that consists of 14 residues (3 glucose, 9 mannose, and 2 *N*-acetylglucosamine). This preassembled glycan is usually transferred by a glycosyltransferase oligosaccharyltransferase to a nascent peptide chain within the reticular lumen. This initial core 14-sugar unit is assembled in the cytoplasm and endoplasmic reticulum and other sugars may be added later. In contrast, O-linked glycans are assembled one sugar at a time at the outset on proteins in the Golgi apparatus.

There are some specific features of medical interest as relevant to the human host of viruses (but by no means unique to humans). The lung epithelial glycans are typical by having sialic acids as the distal residues, and it is these that the influenza neuraminidase cleaves away. Most soluble secreted proteins are also similarly decorated with such glycans. That includes the proteins that make up saliva and mucus in the airway, and are in general important for viral infection. Both N- and O- and glycosphingolipid-glycans are found in human lungs, and they include large and complex-type N-glycans with linear poly-N-acetyllactosamine $[\text{-3Gal}\beta1\text{–}4\text{GlcNAc}\beta1\text{-}]_n$ extensions, which are predominantly terminated in α2,3-linked sialic acid. In contrast, the smaller N-glycans lack poly-N-acetyllactosamine but are enriched in α2,6-linked sialic acids. There are also large glycosphingolipid glycans, which also consists of poly-N-acetyllactosamine, usually terminating in α2,3-linked sialic acid. While it is commonly maintained that viruses such as influenza virus bind to the sialylated glycans, and this is assumed in the present paper, some care is required, because there are also non-sialylated glycans in human lungs on which viral binding could occur.

1.4. Most coronaviruses have "receptor destroying activity"

While it is binding of viruses to sialic acid glycan that is of interest here, it should be kept in mind that it is often associated with a catalytic activity in which the sialic acid glycan is the substrate. In influenza these two aspects are particularly distinct by being on separate surface proteins, the hemagglutinin and neuraminidase respectively, but that separation is not true of all viruses. Not surprisingly, as noted above, most coronaviruses of the coronaviridae family also have capabilities to bind to sialic acid glycans, but they also have the ability to cleave the glycans, which is often described by authors as a "receptor destroying activity". Like influenza C viruses, purified bovine coronavirus preparations have an esterase activity which inactivates O-acetylsialic acid-containing

receptors on erythrocytes; diisopropyl fluorophosphate completely inhibited this receptor-destroying activity suggesting that the viral enzyme is in this case a serine esterase [8]. This is believed to facilitate the spread of virus infection by removing receptor determinants from the surface of infected cells (see discussion below) and prevent the formation of virus aggregates. Another coronavirus, porcine transmissible gastroenteritis virus (TGEV) recognizes N-glycolylneuraminic acid. Nor does it depend on the sialic acid binding activity for infection of cultured cells, but interaction with sialoglycoconjugates appears to help the virus to pass through the sialic acid-rich mucus layer in the epithelium of the small intestine. Hemagglutinin-esterases are a family of viral envelope glycoproteins that mediate reversible attachment to *O*-acetylated sialic acids. These too are said to be receptor-destroying, but the enzymic activity reaction is in this case not a cleavage in the sense used concerning neuraminidase, but rather a change in molecular recognition by removal from of the acetyl group from the C9 position of the above acetylated neuraminic acid residues. The other and probably major reason for researchers thinking of these actions as "receptor destroying" is because the picture is a little more complex than just allowing entry and exit from the host cell. Because many viruses attain host cell specificity by being selective for particular types of sialic acid, these may occur as decoys to the virus on off-target host cells and on free molecules in the extracellular environment. To prevent irreversible binding to these decoys, many viruses including many coronaviruses have receptor-destroying enzymes that are therefore interesting targets for antiviral intervention, exemplified by the influenza A virus neuraminidase [9].

1.5. Where are these important functions in SARS-CoV and SARS-CoV-2?

Even though such glycan binding domains and enzymes as neuraminidases are found in many coronaviruses, there seems to be no such enzymes in SARS-CoV and SARS-CoV-2. Viruses of the lower respiratory tract, such as influenza virus, respiratory syncytial virus, and SARS-related coronaviruses, are generally considered as having key differences that require different therapeutics [10] even though relatively little is lost in considering already approved drugs for one of such viruses against the other (e.g. Refs. [11]). Typically, the apparent absence of glycan binding and enzymic sites in SARS-CoV and SARS-CoV-2 has been dismissed as due to the fact that the virus enters on ACE2, i.e. angiotensin converting enzyme type 2 (e.g. see Ref. [5] for discussion), not on a glycoprotein. This does not, however, escape from the intuitively important need for preliminary binding, cell entry and exit through the glycan layer, and probably the decoy-related function discussed above. There appears to be growing evidence of significant lectin-binding capability. Lectins are the carbohydrate-binding proteins that are highly specific for sugar groups of other molecules. Activation of C-type lectin receptor and other similar receptors contributes to pro-inflammatory response to many coronavirus infections. There also are studies over several years that locate glycan binding and even related catalytic activities in the spike glycoprotein. It has been noted that E3 protein of bovine coronavirus is a receptor-destroying enzyme with acetylesterase activity [8], and the 3D structure of coronavirus hemagglutinin-esterase offered insight into coronavirus and influenza virus evolution, with implications for drug and antibody discovery [9].

The location of any sialic acid glycan binding region of SARS-CoV-2 is, *a priori* unclear, although intuitively (a) it would likely be associated with the cap or knob at the outer end of the spike protein, or (b) at least not involve exactly the same domain as is required for other important functions. Although throughout the coronaviruses various external proteins and domains can recognize either protein or sugar receptors or both, the majority of such studies like those above implicate the S1 region in their spike glycoproteins, but as discussed in the present paper, there are other potential sugar binding sites that are still within the spike protein. Overall, the SARS-CoV-2 spike glycoprotein has 1273 amino acid residues and until early 2020 understanding of structure was heavily based on SARS-CoV spike glycoprotein (1255 amino acids) with 20–27% amino acid residues similarity among non-SARS coronaviruses. Most of the spike protein appears to be involved in the specific stages of cell entry. The spike glycoprotein of SARS-CoV and SARS-CoV-2 is translated as a large polypeptide that is later cleaved to S1 and S2 sites. After binding to the main receptor that that is held to be primarily ACE2, the host proteases activate the virus by cleaving first at the S1/S2 boundary (i.e. S1/S2 site) and then within S2, i.e. at the S2' site. The spike of similar coronavirus have long been considered as being in two main states (i) the pre-fusion form (the form of the mature virion) and (ii) post-fusion form, the form after membrane fusion has been completed). More detailed studies have split the latter into a pre-hairpin intermediate state, and post-fusion hairpin state. Somewhat like in all virus Class I fusion proteins, the S2 protein contains two heptad repeat regions (HRs) of which one (HR2) is located close to the transmembrane anchor. Membrane fusion occurs when there is a conformational change in the HRs to form a fusion core. The HRs of the protein fold into a coiled-coil structure, known as the "fusogenic state". As virus and target cell membranes fuse, the coiled coil regions (called heptad repeats) become a trimer-of-hairpins structure. The S2' cleavage site appears particularly important by being well conserved [3–5] and proteolysis by cathepsin appears sufficient to expose the fusion peptide of S2 and activate fusion within the host cell endosome. In general, S2' is now considered as the key viral fusion peptide which is unmasked following S2 cleavage. Subsequently, S1 dissociates from S2, allowing S2 to transition to the post-fusion structure.

The following locations in the sequences of amino acid residues apply specifically to SARS-CoV-2. These vary somewhat with author, and the following are used here. The signal peptide (SP) comprises residues 1–19. On the inside of the lipid membrane, the carboxyl terminus (C-terminus) is comprised of the transmembrane region (TM) comprising residues 1214–1236 and the cytoplasmic tail (CT) residues1237-1273. The extracellular domain of the spike glycoprotein is comprised of N-terminal domain (S1-NTD) comprising residues 20–286, and is of particular interest here. The host cell receptor binding domain (RBD) comprising residues 319–541. In summary the key regions are as follows.

SP 1-19
S1-NTD 20–286
RBD 319-541
S2 686-1213
TM 1214-1236
CT 1237-1273

Fig. 1 shows the external part 20–1213 of the spike glycoprotein of SARS-CoV-2 in the closed state prior to ACE2 binding, with S1-NTD domain (the "ears", dark blue) of interest here, RBD (at tip, subdomains light blue, blue-green), S2 (subdomains, orange, green, yellow). The orange-white, green-white and yellow-white helical structures are the α-helices of the trimer that form the neck associated with S2, and the red-white helical structures are the start of the transmembrane α-helices TM.

1.6. The function of S1-NTD

In at least some coronaviruses, S1-NTD is known to be involved in binding host proteins or glycans, but coronaviruses show great diversity in their binding which presumably underlies their ability to jump between very different host species. While the role of S1-NTD Compared with the current reasonably detailed knowledge of the remarkable mechanism of cell entry involving ACE2 and changes to the spike protein on cleavage, the specific function of S1-NTD of SARS-CoV-2 has not been elucidated (at least, not by the time of writing in April 2020). As noted above, S1 in SARS-CoV-2 is now well known to have a region which is the receptor binding domain to human ACE2 but also, significantly for

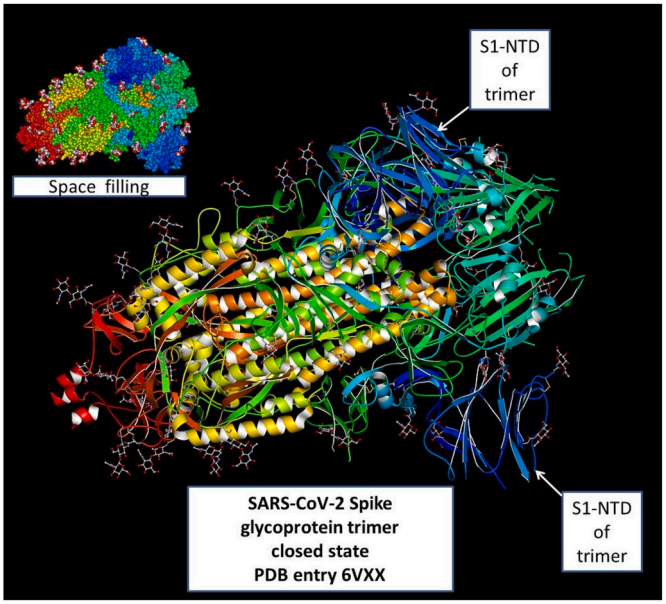

Fig. 1. Spike protein of SARS-CoV-2 PDB entry 6VVX, showing S1-NTD domain (dark blue). See text in regard to the significance of the other colors.

what follows in the text below, SARS-like coronaviruses can bind CLEC4M/DC-SIGNR C-type lectin domains on host cells. See Ref. [12] for review of the diverse receptor recognition mechanisms of coronaviruses up to 2015, which represented the body of understanding until the COV-19 pandemic. Bovine coronavirus is an example of a coronavirus for which it seems clear that S1-NTD has an established glycan-binding function. Although the structure of a sugar-bound Bovine CoV S1-NTD was not available, some conclusions could be reached by researchers using structure-guided mutagenesis and comparisons with different coronaviruses. As well as evidence in 2008 linking hemagglutinin-esterase to the S1 domain of at least some coronaviruses [9], Zhang and Yap [13] had reported in 2004 a rational 3D model for S1 domain of SARS-CoV spike protein by fold recognition and molecular modeling techniques, and there they noted a suggestive structure similarity between S1 protein and influenza virus neuraminidase [14]. This opened up the possibility for those authors that existing anti-influenza virus inhibitors and anti-neuraminidase antibody could be used as a starting point for designing anti-SARS drugs, vaccines and antibodies [14].

Based on such observations and discussion so far, it is therefore reasonable to propose that S1-NTD could be important in the binding of certain alternative host cell surface receptors, or perhaps which aid in targeting the virus to ACE2, and so might provide a helpful therapeutic target (as well as candidate antigenic site for synthetic vaccine design). Nonetheless, such functions if present in SARS-CoV-2 could, *a priori*, reside in other domains at the virus surface. The challenge for research here is that the substantial knowledge concerning such matters in well-studied coronaviruses is not readily transferable. Variously throughout the coronaviruses, S1-NTD, CTD and S2 regions can recognize either protein or sugar receptors or both in various cases, and very similar coronavirus spike protein domains within the same genus may recognize different host cell receptors, while many very different coronaviruses may recognize the same host cell receptor. The studies mentioned above also suggested that at least some coronavirus S1-NTDs are evolutionarily related to human galectins, the term typically used for the lectins as carbohydrate-binding proteins that are specifically involved in inflammation, immune responses, cell migration, autophagy and signaling; however the viral domains derived from them have diverged with specificities for different sugar receptors [12]. Further review is given throughout Results Section 4 where appropriate. Another challenge discussed in Results Section 4 is that key regions for which an experimental 3D structure would help resolve the matter are disordered, and hence invisible, in current available spike protein structures.

2. Theory

Less familiar theoretical principles do arise in the knowledge gathering, inference, and prediction methods developed by the author and colleagues; they are used in the present COV-19 project as described in Ref. [4]. They also include approaches to facilitate interaction with the standard bioinformatics web tools which are stated below in Methods Section 3. However, while tools of these various kinds do speed and facilitate a project of this nature, the usual methods of investigating literature and accessing bioinformatics data bases and tool are sufficient for reproducibility of the work described in this paper, at least by researchers reasonably familiar with bioinformatics and protein structure. An algorithm for predicting the domains and proteins involved in sialic acid glycan binding is developed in the course of the project described in Results Section 4, but this is primarily of a highly empirical nature. Future work to advance this algorithm on a sounder theoretical basis is underway. See brief discussion on future work in Discussion Section 5.1, where some of the above-mentioned theoretical approaches of the author and colleagues, also to be used in the development of the algorithm, are cited.

3. Methods

The overall approach comprised the following steps.

i. Automated gathering of information from the World Wide Web regarding hemagglutinins, neuraminidases, and sugar binding proteins, particularly but not solely of viruses, using “autosurfing”, natural language processing, and knowledge extraction techniques [4]. Note that it is in particular non-covalent sialic acid glycan binding sites that are being explored, not for example asparagine or serine or threonine sites to which glycans are *covalently* linked. This knowledge gathering approach, as described in Refs. [4], is not essential for reproducing the present work or for carrying out comparable studies, but it does greatly accelerate research and preparation of the scientific paper, allowing fast responses to a new epidemic [3–5].

ii. Attempted discovery of continuous short sequences of amino acid residues (potential “sequence motifs”) with patterns and amino acid content common to SARS-CoV-2 spike protein amino acid sequences and hemagglutinins and neuraminidases, particularly those of influenza viruses. Also, more generally, in preparative work, comparison of the spike protein sequence of the spike protein with sialic acid glycan binding proteins and other sugar binding proteins, or domains of them. Protein sequences or parts of them used as input for any part of the study were obtained from GenBank https://www.ncbi.nlm.nih.gov/genbank/. The standard method of bioinformatics used for detecting in large protein sequence databases any amino acid residue sequences similar to those of an input sequence was primarily BLASTp at https://blast.ncbi.nlm.nih.gov/Blast.cgi. The standard tool for a more formal and typically multi-sequence alignment was Clustal Omega at https://www.ebi.ac.uk/Tools/msa/clustalo/. These tools can be automatically accessed by the present author’s methods [4] but again that is not essential for reproducibility of the present work.

iii. Examination of patterns in potential or known short subsequences in small proteins or domains known to have a function involving non-covalent sialic acid binding and, in absence of any clear patterns, study of the amino acid content of the subsequences. This established a preliminary sialic acid glycan binding score (SABS) for the twenty naturally occurring amino acid residues. However, the short subsequences identified were to

be considered as "signals" or "fingerprints" for sialic acid glycan binding domains as a whole. That is, direct contact with sialic acid was not necessarily at (or solely at) the specific short subsequence, for reasons discussed in Results Section 4. This, plus a sequence rather than three dimensional structure perspective, and a specific focus on binding sialic acid glycans rather than sugars in general, resulted in a substantial difference in scores from another major method of predicting sugar binding regions of proteins also discussed later below.

iv. Development of the above as an algorithm SABR-P for identifying potential small proteins or domains of proteins that non-covalently bind sialic acid glycans, by predictions on a test data set of protein sequences. As noted above the subsequences predicted are taken to indicate the glycan binding domain as a whole, not necessarily the sialic acid sites *per se*, but they may be. This approach involved noting true positive and negative predictions and false positives and negative predictions so as to optimize sensitivity and specificity. This was done specifically in regard to non-covalent binding of sialic acid glycans. In other words, it was done so as to distinguished sialic acid glycan binding domains not only from those domains known not to bind sugars but also from those that bind sugars and glycans that do not contain sialic acid.

v. Examination of the three dimensional structures of regions of the regions of the SARS-CoV-2 spike protein predicted as binding sialic acid glycans to propose and locate a sialic acid binding function of SARS-CoV-2 (possibly but not necessarily associated with some kind of enzymic activity).

Results and discussion in the present paper used the same amino acid codes as the above tools and data bank use, i.e. the IUPAC (International Union of Pure and Applied Chemistry) one letter amino acid codes, given in Table 1 below. For completeness, conservative replacements in column 3 of Table 1 are given. They relate largely to substitutions that can usefully be made in the design of synthetic peptides [4,5]. This is an application which is not specifically discussed in the present paper but which could be a basis for design of synthetic vaccines and preventative or therapeutic agents [4,5], in this case targeted at sialic acid glycan binding site of a virus. As discussed in the sequence of steps for the methodology above, the sialic acid glycan binding motifs are taken to be indicators of the sialic acid binding domain and not necessarily of the target sites *per se*, but they may be, and often are, potential target sites. The list of conservative replacements also remains useful for considering substitutions that are conservative in maintaining similar amino acid properties when detecting and comparing related sequences.

Table 1
One letter amino acid codes and sialic acid site binding region measures discussed in the text.

One letter code	Amino acid	Conservative replacements	Preliminary sialic acid binding amino acid score SABS (see Results)	SABR-P prediction method refined parameters (see Results)
A	alanine	A, E, S, T	1	1
C	cysteine/ cystine	S, T, V	1	1
D	aspartic acid	E	1	1
E	glutamic acid	A, D	0	0
F	phenylalanine	M, W, Y	1	2
G	glycine	N, P	1	1
H	histidine	K, R	1	2
I	isoleucine	L, V	0	0
K	lysine	H, R	0	0
L	leucine	I, V	0	0
M	methionine	F, W, Y	0	0
N	asparagine	G, D, Q	1	1
P	proline	G	0	0
Q	glutamine	N, E	0	0
R	arginine	H, K	0	0
S	serine	A, T	1	1
T	threonine	A, I, S	1	1
V	valine	A, I, L	0	0
W	tryptophan	F, M, Y	2	4
Y	tyrosine	F. M, W	1	2

Since for both reasons they be useful in deeper consideration of many results in the present paper, some comment may be useful to researchers less familiar with bioinformatics. See Ref. [4] for a further account. Note that the work of considering what is a conservative replacement is done automatically by the standard bioinformatics tools used. The replacements in Table 1 are consistent with the conservative replacement rules implied by the tables of weights implemented automatically in BLASTp and Clustal Omega mentioned above, which are discussed at those sites. However, the original intent as an application to peptide design means that in Table 1 there is a degree of asymmetry based on the author's experience in peptide design [4] because one is going from a natural protein state to less natural one without evolution making compensatory changes in the rest of the protein or system. For example, empirical studies show that serine (S) can be replaced by alanine (A) or threonine (T) but it is frequently important that a replacement to threonine should be isoleucine (I) in order to retain stability of a β-pleated sheet in which they occur. Strictly speaking, these are just fairly crude rules-of-thumb: the best replacements are dependent on more specific circumstances and detailed conformational and binding calculations. The assignment in Table 1 are not seen as controversial because apart from the asymmetry they relate to the "interchangability" or "alternative rule" of amino acid residues by many authors that are intended as universal, i.e. intended to apply to all proteins. This is because they relate to similarity of amino acid residues in terms of physicochemical, conformational, as well as biological properties of many sequences that are at least universal to, say, vertebrates. However, they are historically more directly empirically based on well-known studies probabilities of amino acid differences found by comparing amino acid residue sequences amongst fairly related proteins from a wide range of sets of different proteins, such that one is comparing sequences of hemoglobins, or of lysozymes, or of cytochromes C, and so on.

As is to some extent customary in the field, *three letter* codes (such as GLY for glycine) are used for the amino acids in the molecular graphics figures; these codes are fairly self-evident at least in the direction of deducing the full name of the amino acid being represented. There was also use of data and the associated graphics tools in the Protein Data Bank (PDB) at https://www.rcsb.org/and in Japan https://pdbj.org/ which was used for Fig. 1.

Energy calculations by the author's own KRUNCH and by a commercial Sculpt protein modeling package were used in Ref. [4], but were not required for the present study, with the exception that some calculations by these tools were used to obtain earlier unpublished results on sugar binding to amino acid residue sidechains. This provided a check on the preliminary sialic acid binding capability shown in column 4 of Table 1, used initially in the present paper. KRUNCH is a molecular mechanics modeling package that essentially functions like many standard molecular modeling packages. There is the arguable exception that it gives much more attention than usual to novel algorithms for navigating through multiple energy minima and discovering new conformers, but that capability did not appear to be too important in the present study. For the much greater part, however, these binding assessments were based on the amino acid residues observed by the author in sequences involved in sugar binding sites in proteins (found by visual examination of binding sites of entries in the PDB) and similar qualitative observations by other authors. That is, they also reflect rather general opinion of what amino acids are involved in sugar binding and in its most general formulation this intuitively comprises aromatic residues, and hydrogen bonding residues to interact with the sugar hydroxyl groups. At the outset, as a starting point only, column 4 of Table 1 of these preliminary sialic acid binding amino acid scores (SABS) are

really assignments that are qualitative, using 0 for not often present in sialic acid glycan binding sites and 1 for often present,. However, tryptophan was assigned a double score of 2 reflecting its larger size and double ring. How reliable these assignments are in regard to sialic acid glycan binding is what is assessed on a more objective basis by the prediction method developed in this paper, including a degree of recalibration. The marginally modified parameters are also shown in the last column Table 1 for convenience of comparison. While as a methodological strategy it was tempting to start from an alternative more objective and established approach discussed in Results Section 4, or at least to use it as a starting point or as an important "gold standard" for comparison, it has substantially different aims.

4. Results

4.1. Putative sugar binding sites in the SARS-CoV-2 spike glycoprotein S1-NTD domain deduced on the basis of alignments

As a first step in an investigation making use of bioinformatics, a common strategy is the use of protein sequence alignments so that a functional part that is known in one protein might suggest an analogous functional part in another. As discussed here, this approach has proven somewhat less successful for sialic acid glycan binding sites than for other functions considered elsewhere (e.g. Refs. [3–5]), and necessitated development of alternative techniques of which discussion occupies the major part of this paper, but the results are consistent and to some degree suportive. As the basis of discussion and starting point for the present study, the following is a SARS-CoV-2 spike protein sequence in which the S1-NTD is shown in italics. The short section of sequence underline emerges below as of particular, but by no means sole, interest.

More subtle computational techniques based on protein conformation have in this project suggested relationships between SARS-CoV and hemagglutinins and neuraminidases of other viruses, but they essentially support prior work. Recall that Zhang and Yap [13] noted a suggestive structure similarity between SARS-CoV S1 and influenza virus neuraminidase. Using various computational techniques they compared the 3D structure of SARS-CoV S1 Protein Data Bank 3D structures 1INY (neuraminidase from influenza A virus complexed with a sialic acid phosphonate analogue inhibitor) and 1B9T (neuraminidase from influenza B virus with novel aromatic inhibitors). This observation does not necessarily suggest by itself a common function, because many protein folding patterns are used by nature for diverse purposes, but armed with that information a more careful use of sequence alignment is helpful. The following part of a Clustal omega sequence alignment done in the present study also includes the hemagglutinin esterase sites in E3 of the bovine coronavirus [8] (E3-CoV-Q14EB1.1 below) and that studied by Zeng et al. [9] (Hem-est-CoV-3CL5 below). Substrate binding residues in deduced for SARS-CoV S1 by Zhan and Yap [13] are shown by #.

```
SARScoV2-S-MN908947.3      -FEYVSQPFLMD----LEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSA   222
SARScoV-S-NP_828851.1      -FEYISDAFSLD----VSEKSGNFKHLREFVFKNKDGFLYVYKGYQPIDVVRDLPSGFNT   215
Hem-est-CoV-3CL5           NYSYMDLNPALCDSGKISSKAGN-SIFRSFHFTD----FYNYTGEGQ-------------   77
E3-CoV-Q14EB1.1            NYGYLDIHPSLCNNGKISSSAGD-SIFKSYHFTR----FYNYTGEGD-------------   89
                            : *:.    :     :. . *: . ::.: *.     :  *.
                                                ######
SARScoV2-S-MN908947.3      LEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPR-------TF   275
SARScoV-S-NP_828851.1      LKPIFKLPLGINITNFRAILTAFS------PAQDIWGTSAAAYFVGYLKPT-------TF   262
Hem-est-CoV-3CL5           ---QIIFYEGVNFTPYHAF------KCTTSGSNDIWMQNKGLFYTQVYKNMAVYRSLTFV   128
E3-CoV-Q14EB1.1            ---QIIFYEGVNFNPHHRF------KCFFNGSNDVWIFNKVRFYRALYSNMALFRYLTFV   140
                               . :  *:*:. .: :            :.. *  .   ::    .         .
                                                            ####

SARScoV2-S-MN908947.3      LLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNL   335
SARScoV-S-NP_828851.1      MLKYDENGTITDAVDCSQNPLAELKCSVKSFEIDKGIYQTSNFRVVPSGDVVRFPNITNL   322
Hem-est-CoV-3CL5           NVPYVYNGSAQSTALCKSGSLV-------------------------------------   150
E3-CoV-Q14EB1.1            DILYNFSFSI-KANICNSNILS-------------------------------------   161
                            : *  . :  .:  *  . *
```

The situation is of course not clear when the degree of amino acid residue match between sequences is poor because similarities can arise by chance, but aligning more sequences can be insightful, and in this case including influenza A and B sequences seemed particularly helpful. Although still barely conserved compared with the motifs in the current project (published [4,5], submitted, or under investigation) the most plausible match for further study is at SSSGWTAGAAAYY of SARScoV2-S-MN908947.3, because others would include at least one extensive deletion areas from one protein. Including influenza A and B sequences interferes with much of the alignment but it does preserve the above match.

```
>6VSB_1|Chains A,B,C|SARS-CoV-2 spike glycoprotein|Severe acute respiratory syndrome
coronavirus 2 (2697049)
MFVFLVLLPLVSSQCVNLT - signal peptide
TRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWI
FGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLRE
FVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLL
KYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVA
DYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNY
NYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVK
NKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVP
VAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPGSASSVASQSIIAYTMSLGAENSVAY
SNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPP
IKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLA
GTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQ
LSSNFGAISSVLNDILSRLDPPEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGKGYHLMS
FPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTV
YDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQGSGYIPEAPRDGQAY
VRKDGEWVLLSTFLGRSLEVLFQGPGHHHHHHHHSAWSHPQFEKGGGSGGGGSGGSAWSHPQFEK
```

```
FluB-N-AGA18961.1          LSCPGS-T-------------F-QKALLISPHRFGETKGNS----------ACGPKECKH   119
FluB-N-pdb1B9T             LSCQGS-T-------------F-QKALLISPHRFGEIKGNSAPLIIREPFVACGPKECRH   53
FluA-N-ACQ76385.2          ---------------------------YSKDNSIRIGSKGDVFVIREPFISCSPLECRT   131
FluA-N-pdb1INY             ---------------------------YGKDNAVRIGEDSDVLVTREPYVSCDPDECRF   50
SARScoV2-S-MN908947.3      FRVYS-SANNCT---------FEYVSQPFLMD----LEGKQGNFKNLREFVFKNIDG-YF   201*
SARScoV-S-NP_828851.1      TMIFD-NAFNCT---------FEYISDAFSLD----VSEKSGNFKHLREFVFKNKDG-FL   194*
Hem-est-CoV-3CL5           --LFGDARSDCNHVVNTNPRNYSYMDLNPALCDSGKISSKAGN-SIFRSFHFTD-----F   69*
E3-CoV-Q14EB1.1            --LFGDSRSDCNHINNLKIKNYGYLDIHPSLCNNGKISSSAGD-SIFKSYHFTR-----F   81*
                                                                .    #####

FluB-N-AGA18961.1          FALTHYAAQPGGYYNGTREDRNKLRHLISVKLGKIPTVENSIFHMAA------WSGSACH   173
FluB-N-pdb1B9T             FALTHYAAQPGGYYNGTRKDRNKLRHLVSVKLGKIPTVENSIFHMAA------WSGSACH   107
FluA-N-ACQ76385.2          FFLTQGALLNDKHSNGTIKDRSPYRTLMSCPIGEVPSPYNSRFESVA------WSASACH   185
FluA-N-pdb1INY             YALSQGTTIRGKHSNGTIHDRSQYRALISWPLSSPPTVYNSRVECIG------WSSTSCH   104
SARScoV2-S-MN908947.3      KIYSKHTP--INLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWT   259*
SARScoV-S-NP_828851.1      YVYKGYQP--IDVVRDLPSGFNTLKPIFKLPLGINITNFRAILTAFS------PAQDIWG   246*
Hem-est-CoV-3CL5           YNYTGEGQ-----------------QIIFYEGVNFTPYHAF------KCTTSGSNDIWM   105*
E3-CoV-Q14EB1.1            YNYTGEGD-----------------QIIFYEGVNFNPHHRF------KCFFNGSNDVWI   117*
                              .                        .    .   .  .           :

FluB-N-AGA18961.1          DGREWTYIGVDGPD------------------------------S---------------   188
FluB-N-pdb1B9T             DGREWTYIGVDGPD------------------------------N---------------   122
FluA-N-ACQ76385.2          DGINWLTIGISGPD------------------------------N---------------   200
FluA-N-pdb1INY             DGKTRMSICISGPN------------------------------N---------------   119
SARScoV2-S-MN908947.3      AGAAAYYVGYLQPR-------TFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGI   312
SARScoV-S-NP_828851.1      TSAAAYFVGYLKPT-------TFMLKYDENGTITDAVDCSQNPLAELKCSVKSFEIDKGI   299
Hem-est-CoV-3CL5           QNKGLFYTQVYKNMAVYRSLTFVNVPYVYNGSAQSTALCKSGSLV---------------   150
E3-CoV-Q14EB1.1            FNKVRFYRALYSNMALFRYLTFVDILYNFSFSI-KANICNSNILS---------------   161
                              .
```

This preserved match region also places the notable tryptophan (W) within short distance of alignment in the second subsequence corresponding to the Zhan-Yap binding site, but in procedures of this kind it is important to establish that such features are not simply artifacts of the particular sequences included. The essential features of the above alignments at the binding site are not perturbed by removing the influenza neuraminidases which provided a basis for the Zhan-Yap catalytic regions, as follows.

should probably be distinguished. As discussed later below, charged amino acid residues glutamate (E), arginine (R), and lysine (K) also frequently make intimate contact with sialic acids but that is in three dimensions, not together in a subsequence. A likely relevant observation was that the first set of amino acid residues (the set containing aspartate) and binding sialic acids tended to occur in a subsequence that adopted a local loop conformation, while the second set (that containing glutamate) were frequently associated with α-helices and particularly their termini. However, this was an empirical and qualitative observation regarding a tendency, and a more objective quantification of the importance of the aspartate set is the purpose of the prediction algo-

```
SARScoV2-S- MN908947.3     -FEYVSQPFLMD----LEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSA   222
SARScoV-S-NP_828851.1      -FEYISDAFSLD----VSEKSGNFKHLREFVFKNKDGFLYVYKGYQPIDVVRDLPSGFNT   215
Hem-est-CoV-3CL5           NYSYMDLNPALCDSGKISSKAGN-SIFRSFHFTD----FYNYTGEGQ-------------   77
E3-CoV-Q14EB1.1            NYGYLDIHPSLCNNGKISSSAGD-SIFKSYHFTR----FYNYTGEGD-------------   89
                             : *:.    :     :. . *: . ::.: *.     :  *.
                                                 ######
SARScoV2-S-MN908947.3      LEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPR-------TF   275
SARScoV-S-NP_828851.1      LKPIFKLPLGINITNFRAILTAFS------PAQDIWGTSAAAYFVGYLKPT-------TF   262
Hem-est-CoV-3CL5           ---QIIFYEGVNFTPYHAF------KCTTSGSNDIWMQNKGLFYTQVYKNMAVYRSLTFV   128
E3-CoV-Q14EB1.1            ---QIIFYEGVNFNPHHRF------KCFFNGSNDVWIFNKVRFYRALYSNMALFRYLTFV   140
                               . :  *:*:. .: :            :.. *  .   ::    .          .
                                                           ####
```

4.2. Sialic acid binding residues

Alignment studies like those above are helpful in forming initial hypotheses, but they are clearly not convincing by themselves when the degrees of match between aligned residues is weak; however, it is possible that the actual composition in terms of amino acids in the subsequence and potential motif of interest is more important than their actual order in that subsequence. Based on many studies exemplified by the above, it was noted that tryptophan is a recurrent, albeit not absolutely essential, feature of sugar binding including sialic acids. References to that also recur throughout the biochemical literature. For example, see e.g. Ref. [14], and also Fig. 2 shows the tryptophan interaction with and sialic acid in the Influenza Virus B neuraminidase (PDB Entry 2BAT). As might be expected by their similar aromatic character, the alternative amino acid residues as sugar binders, and residues frequently supporting tryptophan in the binding site, tend to be aromatic sidechains, notably tyrosine (Y), sometimes phenylalanine (F) and also histidine (H). A preliminary survey of sequence motif patterns in sugar binding proteins suggests that invariant amino acid residues across a family of proteins tend be one or more of the above residues supported by negatively charged aspartate (D), asparagine (N), serine (S), threonine (T) glycine (G) and sometimes alanine (A) that provide the hydrogen bonding. However, particularly in regard to the non-aromatic residues, the binding of acidic and non-acidic sugars rithm developed below. Three dimensional considerations, however, give insight and sometimes explain why influences can be somewhat indirect. Hydrogen bonding that occurs between the hydroxyl groups of carbohydrate ligands and polar amino acid residues at the binding site is typically supported by water-mediated hydrogen bonding networks in

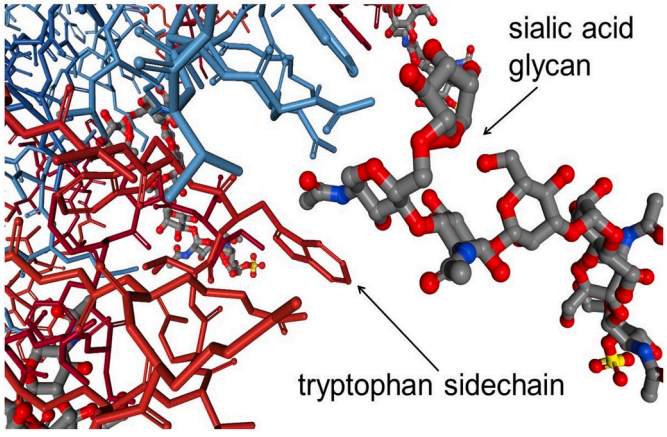

Fig. 2. The influenza virus B neuraminidase tryptophan interaction with and sialic Acid(PDB entry 2BAT).

which serine and threonine are fairly commonly involved. Nonetheless, the most outstanding feature of carbohydrate binding sites from a three dimensional perspective would appear to be the position and orientation of tryptophan (W), tyrosine (Y), and/or phenylalanine (F), which usually provide a hydrophobic plate for close interaction with the planar face of sugar rings, an interaction resembling hydrophobic stacking interactions, as in Fig. 2. The importance of these and to some extent of histidine (H) in a sequence motif seems reasonable.

Along with the occasional appearance of cysteine (C), sometimes as a serine (S) and particularly a threonine (T) substitution, the residues mentioned above can be used as the basis of a preliminary and essentially qualitative model for assessment of sugar binding as given column 4 of Table 1. Often a valine (V) substitution was seen in potential glycan binding sites, although this was not significantly supported by the optimization of the predictive technique described below. However, glycans containing sialic acids appear to bind somewhat differently to other sugars. Taking this as a hypothesis and focusing on these, protein sites for binding them may have a variety of affinities for different subtypes. For example, all influenza A virus strains critically depend on sialic acid to bind to host cells and the different forms of sialic acids all show different affinities that change with influenza A virus variety, important because it determines which species can be infected. There has also been very relevant work that can complicate the details of what can be meant by, for example, "sugar binding motif". Indeed, Zhang and Yap [13] themselves noted that tryptophan was frequently involved in strong protein fold interactions that stabilized the sugar binding domain fold, yet lacked any direct interactions with sugars. In the influenza neuraminidase they comprised three pairs of main chain-side chain interactions: tryptophan (W) 171 (donor) and phenylalanine (F) 179 (acceptor), alanine (A) 210 (donor) and phenylalanine (F) 29 (acceptor), leucine(L) 209 (donor) and tryptophan (W) 171 (acceptor). For example, tryptophan accepted one N–H···π bond from Leucine (L) 209 and donates one N–H···π bond to phenylalanine 179. It is possible that the aromatic residues could be induced to be more exposed in certain types of binding, but the above interactions appeared to stabilize the structure of S1 fold motif while reducing the active site cavity for ligand binding [13].

4.3. More detailed analysis of the importance of tryptophan

Features of sialic acid glycan binding regions represented by a run of amino acid residues have diverse sequence patterns, but evidently the simple presence of the above residues in a section of sequence is itself a strong signal feature; this seems particularly so for tryptophan. To examine this further, many sequence alignments were examined that relate to role of the aromatic amino acids and the contribution of tryptophan to known or suspected sialic acid glycan binding, and these were investigated in a more quantitative way. For example, the following represent a summary of influenza subsequences that contain tryptophan. With it is associated the preliminary, essentially qualitative sialic acid binding residue score (SAB-S) discussed above and based on observation of multiple sugar binding sites and literature survey, to each of the amino acids mentioned above. At this stage, there is no significant algorithm except that the sum of scores over the residues in the short sequence is divided by the number of residues (mostly 16 or 17) so that it expressed on an averaged, per residue, basis for that sequence. Recall that it is qualitative that the amino acid residues are given a score of 1 (and 0 otherwise), except that tryptophan is given an extra weight of 2. In many cases the sidechain is involved, especially for the aromatic residues, but this not obligatory: it could be a backbone interaction (as is necessary for glycine (G) that lacks a sidechain). This is the "qualitative" model that was shown in Table 1, which will be extended to the SABR-P method later below. There is in the following an attempt at local alignment to highlight similar features because there is often some indication that a motif is reused even within a protein, although that assertion is not required for present purposes. The sum of score is divided by the sequence length that excludes relative deletions '-'.

```
FHMAAW-SGSACHDGRE - 1NSB_1 neuraminidase Influ. B subseq. 1, QRBS 0.88
SKIGRWYSRTMSKTERM - 1NSB_1 neuraminidase Influ. B subseq. 2, QRBS 0.53
YDGDPW-TDSDALAHSG - 1NSB_1 neuraminidase Influ. B subseq. 3, QRBS 0.94
DGGKTWH-SAATAIYCA - 1NSB_1 neuraminidase Influ. B subseq. 4, QRBS 0.94

QVCIAW-SSSSCHDDKA - 2BAT_1 neuraminidase Influ. A subseq. 1, QRBS 0.81
QGVKGWAFDNGNKDLRS - 2BAT_1 neuraminidase Influ. A subseq. 2, QRBS 0.71
KVIGGW--STPNSKSQI - 2BAT_1 neuraminidase Influ. A subseq. 3, QRBS 0.60
QETRVWWTSNSIVVFCV - 2BAT_1 neuraminidase Influ. A subseq. 4, QRBS 0.65

RALISWPLSSPP-DDKT - 7NN9_1 neuraminidase Influ. A subseq. 1, QRBS 0.56
VECIGWSSTSCHDGKTR - 7NN9_1 neuraminidase Influ. A alt.align.1B, QRBS 0.76
DGVNTWLGRTISIASRS - 7NN9_1 neuraminidase Influ. A subseq. 2, QRBS 0.71
VLNTDW-SGYSGSFTQG - 7NN9_1 neuraminidase Influ. A subseq. 3, QRBS 0.875
KEDKVWWTSNSIVSMCV - 7NN9_1 neuraminidase Influ. A subseq. 4, QRBS 0.65
```

Many hemaglutanin and neuraminidase matches were found with SSSGWTAGAAAYY using BLASTp, the top 100 varying from 64% match and 63% identities that preserve the tryptophan, such as hemagglutinin Influenza A virus, GenBank AXB35920.1 SSGW–GAVN, and neuraminidase, Influenza A virus AFK13818.1 SGW——AAY, and neuraminidase, Influenza A virus ANZ90284.1, SSAWSASA. Looking at the larger sequence context, 95% of these scored over 0.75 on the above system. However, *nucleocapsid* proteins of Influenza A virus such as sequence QIQ4588 with SG-TAGAA and AAZ08011.1 as STSG-AAGAA also appear in the top 100 matches along with the above and with comparable scores, but with one obvious and significant difference, that the tryptophan (W) is missing.

The possible importance of tryptophan in binding in SARS-CoV-2 is exemplified in the following sequence of the SARS-CoV-2 S1-NTD section of the spike glycoprotein (including the signal peptide, in order to conserve standard numbering). The sequence is from the S1-NTD of SARS-CoV-2 GenBank entry MN908947.3, along with a brief description of accessibility in PDB entry 6VVX, which is for the spike protein in the closed state.

```
MFVFLVLLPLVSSQCVNLT - signal peptide
                   TRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFF**SNVTWFHAIHV** 60-70 W exposed
SGTNGTKRFDNPVLPFNDGVYFASTEKS**NIIRGWIFGTTLD**SKTQSLLIVNNATNVVIKVCEFQFCNDPF 112-124 cleft
LGVYYHK**NNKSWMESEFR**VYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPI 147-160 disord.
NLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPG**DSSSGWTAGAAAY**YVGYLQPRTFLLKYN 253-265 disord.
ENGTIT
```

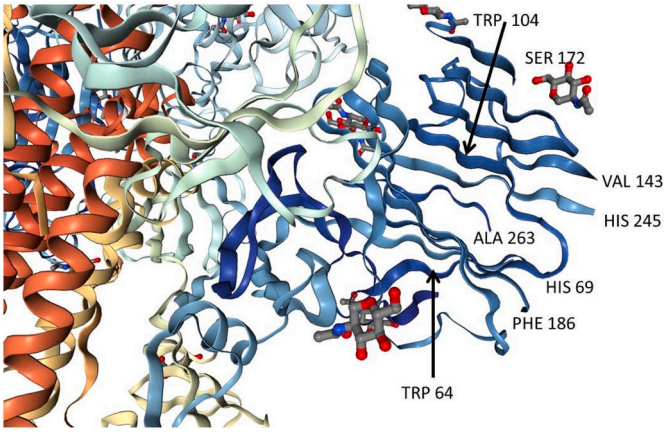

Fig. 3. S1-NTD (PDB 6VXX) Showing Residues at Boundaries of Invisible (Disordered) Segments and the Location of the Two Visible Tryptophan (TRP) Residues. These are the last visible residues of the non-disordered region bounding a disordered region, and this figure shows how the missing sections result in "cut ends" in displays of the reported three dimensional structure.

Here, "disord." means a disordered (flexible) loop: the tryptophan (W) and adjacent residues are not seen in the experimental 3D structure determination 6VVX. This is also true of PDB entry 6VSB the spike protein in the open state, 6VYB and other SARS-CoV-2 accessible to the author at the time of writing. See Fig. 3. The fact that disordered suggest an open loop like structure that be may be capable of binding to, and perhaps adjusting to, disordered targets, and certainly does not prohibit functional significance.

The following summary Table 2 has comments on the status of exposure of the tryptophan and surrounding residues, and on the involvement in binding based on alignments with the Zhang-Yap analysis.

4.4. Development of a simple algorithm for prediction of sialic acid binding domains

Evidently the above is not perfect as a basis for a prediction method, and the next step in this study required more detailed considerations. it also required a more comprehensive analysis of predictive capability that allows for false positives and false negatives. The method developed here takes account of the above results in the light of several pieces of experimental and theoretical evidence, which is usefully briefly reviewed at this point in the narrative. Several kinds of evidence were taken into account in development of SABR-P, the Sialic Acid Binding Region Prediction. The emerging principles used, not all of which are immediately intuitive, are argued as follows so that they may help researchers develop improved versions.

(i) *The method specifically concerns predicting regions that interact with glycans containing sialic acid residues.* The development of the method started with the qualitative sialic acid binding region SABS score in Table 1 and used in preliminary studies above. Other prediction methods such as that discussed below also address sugars or oligoscharrides in general. Recall, however, that even the sialic acid components comprise a fairly diverse family of sugar types (see Introduction Section 1.3). Also, while described as a Sialic Acid Binding Region Prediction, and this is believed to be the naturally emerging emphasis, it is formally the binding to glycans that contain sialic acids that matters.

(ii) *The emphasis is also on binding, not catalysis.* The method is not specifically concerned with catalytic amino acid residues involved in neurominidase action (nor any other activity of cleaving modifying the glycan to reduce virus binding, e.g. esterases, deacetylases). While evidence of neuraminidase activity in SARS-CoV-2 would be very important, the current evidence is that focus should be on sialic acid binding. This is because of lack of any strong evidence for such enzymic activities at the time of writing, and also because neuraminidase inhibitors approved as dugs, Oseltamivir (Tamiflu), and Zanamivir (Relenza) have been tested on SARS-CoV-2 *in vitro* and were not found effective [15]. It is nonetheless the binding of the appropriate glycans containing sialic acid that may still be important for SARS-CoV and SARS-CoV-2 infection *in vivo*, as it is for other viruses [16]. The fact that other viruses, including other coronaviruses, require that function, followed by any demonstration that SARS-CoV and SARS-CoV-2 still possess that function, would suggest that it could still be a target for pharmaceutical drug development, at least as a preventative. This argument is not unique in its general form. For example, Fantani and colleagues believe that the SARS-CoV-2 also uses sialic acids linked to host cell surface gangliosides to somehow facilitate host cell entry and so could represent a therepeutic target [17].

(iii) *The method seeks to predict whole domains and proteins that bind sialic acid, not specific short sequence motifs.* That is the case even though the predictions assign a sialic acid binding score to each residue in the domain or protein, which can of course still be inspected as of potential involvement in direct binding. The threshold and scale of the SABR-P prediction method is set such that any residue (in practice it is almost always a continuous run of residues) with a sialic acid binding score of more than 100 signals that the whole domain or protein, whichever was

Table 2
Sialic acid binding scores and surface exposure of tryptophan -containing subsequences.

Subsequence	Protein	GenBank or PDB entry	Score	Experimental observation or exposure suggested by structural analogy
FFSNVTWFHAIHVSGTN	SARS-CoV-2 S1-NTD	MN908947.3	0.88	Tryptophan exposed
FYSNVTGFHTIHTFGNP	SARS-CoV S1-NTD	NP_828851.1	0.82	Site aligning with above, no tryptophan, exposed
SNIIRGWIFGTTLDSKT	SARS-CoV S1-NTD	MN908947.3	0.71	Example of buried Tryptophan, but possibly accessible at base of cleft
HDGGKTWHSAATAIYCA	neuraminidase Influenza B	PDB 1NSB	0.94	Tryptophan exposed
EGKQGNFKNLREFVFKN	SARS-CoV-2 S1-NTD	MN908947.3	0.47	Aligns with a Zhang-Yap binding site
SEKSGNFKHLREFVFKN	SARS-CoV S1-NTD	NP_828851.1	0.53	Aligns with a Zhang-Yap binding site
KAGNSIFRSFHFTDFYN	Hemagglutinin esterase	PDB 3CL5	0.88	Aligns with a Zhang-Yap binding site
SAGDSIFKSYHFTRFYN	CoV - E3	PDB 4EB1.1	0.82	Aligns with a Zhang-Yap binding site
GDSSSGWTAGAAAYYVG	SARS-CoV-2	MN908947.3	0.94	Aligns with a Zhang-Yap binding site
SPAQDIWGTSAAAYFVG	SARS-CoB	NP_828851.1	0.82	Aligns with a Zhang-Yap binding site
SGSNDIWMQNKGLFYTQ	Hemagglutinin esterase	PDB 3CL5	0.71	Aligns with Zhang-Yap binding site
NGSNDVWIFNKVRFYRA	CoV - E3	Q14EB1.1	0.71	Aligns with Zhang-Yap binding site

provided as the whole sequence in input, is a binding module for glycans containing sialic acid. There are several reasons for using this as a marker of a domain or protein of interest. One is that the binding is to the whole glycan as an oligosaccharide or polysaccharide, not necessarily the sialic acid components *per se*, and therefore the ligand is a large structure that could involve several binding sites. Recall that tryptophan was found to be frequently involved in strong protein fold interactions that stabilized the sugar binding domain fold, yet were lacking any direct interactions with sugars and even buried [13]. There are at least somewhere between 10 and 100 amino acid residue sequence motifs known to be associated with sugar binding in general, and they fall into some 7 fold motifs [18]. Also, as the function of sialic acid recognition may be to guide the virus to and across the host cell surface [17], sites on the virus that enable this motility and hunt out points for catalysis or simply the strong binding appear to be as important as key strong binding and catalytic sites themselves (e.g. Ref. [18]). In addition, predictions of localized regions of proteins as sugar binding sites are naturally not as good as those based on predicting that a whole domain or protein is involved in sugar binding, but such a prediction remains valuable and indeed well suited to the present study. In a particularly detailed investigation by Taroni et al., analysis of the characteristic properties of sugar binding sites was performed on a set of 19 sugar binding proteins [19]. Their prediction was optimized on a training set of 19 non-homologous carbohydrate binding structures and tested on a test set of 40 protein–carbohydrate complexes. The thoroughness of that study and the inclusion of many kinds of information would make it a reasonable choice of "gold standard" for comparison, except that the aims were substantially different, and the overall accuracy of prediction achieved was only 65%. It is true nonetheless that results were very good for carbohydrate-binding enzymes as opposed to lectins, with a rate of success of 87%. This emphasis on enzyme catalytic sites again argues for avoiding the use of this kind of approach in the present paper, as covered by point (i) above. In summary, it may be that a large domain or protein that has strong signals for sialic acid binding may be indicative of a sialic acid binding and/or catalytic function even if a specific short subsection of the sequence predicted as binding sialic acids is not in the position for which a strong binding or catalysis has been observed.

(iv) *The prediction is nonetheless based on sequences and scores for residues in continuous subsequences.* Only the sequence is considered in the method, not residues brought together by the folding of the protein in space. Most prediction methods for predicting sugar binding such as the example of Taroni et al. [19] discussed above are not based on sequence and the sugar-binding propensity of amino acids in short segments alone, but on regions on a protein surface that require account of 3D structural information, analogous to "discontinuous" epitopes or "discontinuous determinants" of a pathogen protein in the study of immune response and in synthetic vaccine design. This kind of approach was abandoned in the present study not just because it took away the emphasis on the binding functions of whole domains or proteins but most importantly because many of the sites of particular interest are in disordered regions, or otherwise invisible in the available experimental 3D structure, as discussed later below. It might even be that a degree of disorder is important for binding some sugars. Also and not least, 3D structural information is not always available.

(v) *Glycophilicty parameters for amino acid residues or sequence patterns as obtained by other workers were not used.* This is particularly because, the aims, i.e. what things are wanted to be known, are usually very different in these methods. Although Taroni et al. showed that out of 6 partially parameters partially dependent on 3D information the sugar binding propensities of certain amino acids was prominent in having discriminatory power, this was in regard to the tendency for being in putative a sugar binding patch compared with protein surfaces in general (requiring ordered conformation and experimental information about it). At the same time, it still concerns a specific region of a kind rather than predictions of domain or proteins as sugar binding as a whole. Perhaps most importantly these studies were concerned with binding sugars in general (not specifically sialic acids). Not surprisingly, therefore, the amino acid residue propensities have no significant overall correlation with the propensities in Table 1 used and developed in the present paper. Notably, alanine (A), serine (S), threonine (T) and cysteine (C) have a propensity against sugar binding in their approach. There is nonetheless the significant exception that tryptophan (W) is the strongest in both that and the present study.

(vi) *A simple predictive model was developed based on optimized parameters.* If a simpler method works as well as a more rigorous approach, it can sometimes provide insight and help build even better rigorous approaches. The method used in the present case was initially based on the GOR method [20] for protein secondary structure prediction in which sialic acid glycan binding state of residues replaced α-helix, β-sheet, and coil (or loop) states of residues. However, it was found that the results were essentially reproduced by a simpler model. This is primarily because the directional effects (in terms of N-terminal direction or C-terminal direction along the amino acid residue sequence) was found to be equivalent (i.e., symmetrical) and to persist for some 8 residues in both directions, very like the influence of alanine on α-helix formation in the GOR method [20]. Consequently the final method used in the present study consisted of just two changes to that used in previous Results Section 4.3 above, the second being described in point (vii) below. The qualitative scores as parameters to types of amino acid residue as shown in Table 1 were based on visual observations by the present author and a survey of sites in the literature, and this was improved by optimization on essentially the same set of proteins examined. The 20 amino acids were initially assigned the qualitative parameters of Table 1, then these 20 parameters were optimized to optimize sialic acid glycan binding predictions as positive for 20 sialic acid glycan binding proteins, and negative for 10 proteins not considered as binding sugars, and 10 proteins such as lectins that bind other sugars that do not contain sialic acids. This gave as before glycine (G), alanine (A) aspartate (D), asparagine (N), serine (S), threonine (T), cystine/cysteine (C) each with a score of 1, and the rest assigned 0, except for tyrosine (Y phenylalanine (F) and histidine (H) that were now assigned larger parameter values of 2 and tryptophan (W) that was assigned a larger parameter value of 4.

(vii) *Parameters describing directional influences of amino acid residues were modeled in a simple way.* In part this is a response to the above point that the directional effects (in terms of N-terminal direction or C-terminal direction along the amino acid residue sequence) are symmetrical, but more specifically the essential features of the interactions between neighboring residues can be deduced from parameters like those in Table 1 in much the same way as the shape of an equilateral triangle standing on its base can be deduced from its height. As well as using new parameters, the initial score for a particular residue was computed in the same way as that for the whole segment of 17 residues in Section 4.3, but was specifically considered as associated with the central residue, and the sum was retained as that sum rather than the overall score being averaged by the number of residues in the segment. Deletions to consider alignments of segments were not applied in the SABR-P method. The above sum associated with each central, i.e. 9th, residue, was then averaged over the corresponding sums from the residues up to 8 away on the N-

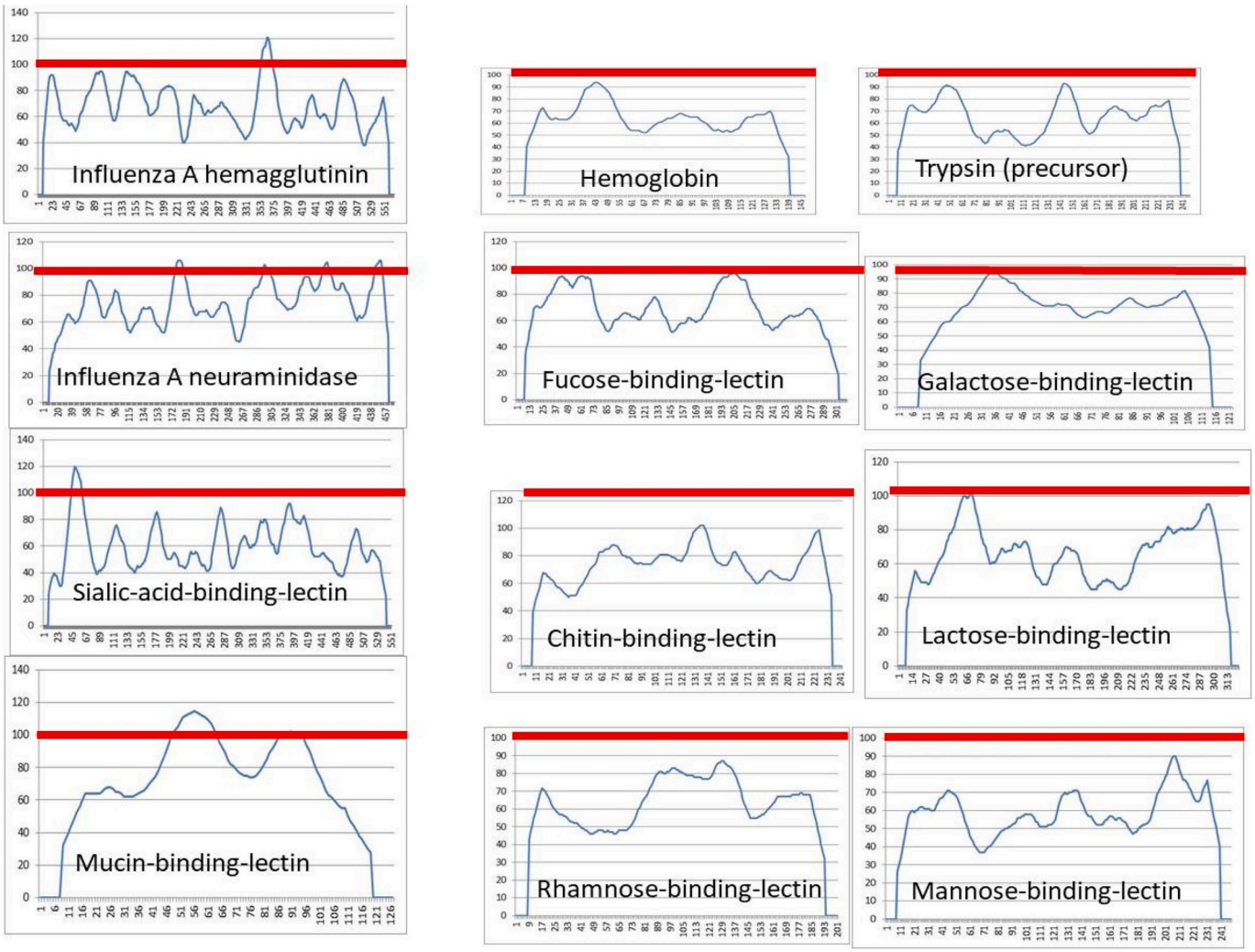

Fig. 4. Examples of Prediction of Sialic Acid Binding Sites by SABR-P.
In each case, the abscissa (x axis) is the distance along the sequence (residue number) and the ordinate (y axis) is the predicted sialic acid glycan binding propensity. Scores above a threshold of 100 (red line) for any residues are taken as a prediction that the domain or protein binds sialic acid glycans.

terminal direction up to 8 residues in the C-terminal direction. This was done for every residue in the input sequence. It formally models parameters describing the directional characteristics of different types of amino acid residue, albeit in a highly empirical way. However, it is simply analogous to a smoothing of the propensity plot for residues along the overall input sequence.

(viii) *Plot scaling was applied to produce a convenient decision threshold.* An appropriate height scale of the plot, in which a residue with a score of over 100 would be considered as the whole input sequence indicative of a sialic acid residue binding domain or protein, was obtained by an empirical pseudo-normalization consisting of dividing each residue score by 300. Whenever the final results by SABR-P are expressed as a percentage-like score by multiplying all residue scores on the plot by 100, this is of course equivalent to dividing by 3. On that percentage basis, there seems no benefit it describing scores more accurately than to the nearest integer.

In summary, the essential features of the simple resulting algorithm are as follows.

(1) *Residue Binding Parameter Assignment.* Examine every residue in the sequence and assign it the parameter 0,1,2,3, or 4 of the SABR-P prediction method refined parameter (last column of Table 1).

(2) *Basic Motif Score.* For each residue number i, obtain a score by summing over the run of 17 residues of which i is the central (9th) residue and assign the resulting sum (the score) to each residue i (providing that a residue numbers i-8, i-7 etc. up to i+8 lies within the sequence).

(3) *Smoothed Motif Score.* Smooth the implied plot of scores to model a directional information effect characteristic of secondary structure prediction [20] by repeating step (2) with the resulting scores, but do not add the basic motif score for each residue i to itself. More specifically stated, examine the above basic motif score for each ith residue from i = 1 to end of sequence in turn, and add it to the score of the residue at i-8, i-7, etc up to i-1, and to the score of the residue i+1, i+2, etc. up to i+8 (providing that the residue numbers i-8 etc. lie within the sequence).

(4) *Normalized Score as SABR-P propensity.* Examine each smoothed score resulting from the above for each residue i, and divide by 300. This is the current value of the optimized normalization parameter that sets the threshold as 100 after multiplying by 100, above which sialic acid glycan binding is predicted.

(5) *Reporting.* The normalized score (SABR-P propensity) is plotted as a function of residue number from 1 to end of the sequence, i.e. as

a "SABR-P propensity plot". The actual predictions are also written out separately as text and these report only the amino acid residue (as G, A, V, etc.) along with the score, for those residues for which the normalized scores exceed 100. Particularly because of step (3), these will tend to occur conveniently in runs of approximately 8–18 contiguous residues, but sometimes shorter. In the present study the normalized score was formalized as a simple measure by rounding to the nearest integer prior to reporting, though this made no significant difference in the present paper.

This simple approach will be better developed as a GOR-like method, but for present purposes it suffices to give an estimate of the likely sialic acid glycan binding domains on the spike protein and proceed with further investigations and results described below. At the above optimized threshold, this the example SABR-P propensity plots shown in Fig. 4. The initial studies using a random selection (more correctly stated, an arbitrary selection) of proteins except that a predominance of sialic acid or sialic acid glycan binding proteins were sought, the method initially gave 18 true positives out of 20 proteins representing group A, i.e. those that are known to bind sialic acids or glycan molecules containing them, 10 true negatives for prediction of the above same kind of sialic acid binding out of group B, i.e. 10 proteins believed *not* to bind any kinds of sugars significantly, and 6 true negatives for the 10 proteins comprising group C, i.e. proteins that bind sugars but not those containing sialic acids, and so representing the biggest challenge. This initial result represented 85% accuracy, 82% sensitivity, and 80% specificity, a reasonable preliminary result for such a simple model. Note that there were initially no false negatives for those proteins not believed to bind any kinds of sugars, so a specific search for at least one case of a false negative (discussed below), which is strictly speaking a bias, dropped the prediction quality to 83% accuracy, sensitivity 82%, and 76% specificity. A recent number of further studies exemplified in the discussion below were also, strictly speaking, a bias including comparisons between weakly homologous proteins, extended the sample to 80 proteins and reproduced the original quality to within the nearest percentage: 85% accuracy, 82% sensitivity, and 80% specificity.

At this stage of the present project the predictive method was only required to give approximate results as guidelines as to the regions the spike protein that might be further examined for sialic acid glycan binding capability, but the method appears promising and justifies some further comments and some further exploratory investigation, as follows.

4.5. Predictions of group a proteins

In group A, proteins being tested are believed to be able to accommodate sialic acid, sialic acid glycan, and related compounds by noncovalent binding. Predicted correctly, these would represent the true positives, but predicted incorrectly, they would represent false negatives. Such proteins include various hemagglutinins such as that of influenza A at GenBank CAA24291.1.

```
M,17,105
I,18,106
D,19,107
G,20,108
W,21,109
Y,22,108
G,23,106
F,24,104
R,25,101
```

It also included various neuraminidases such as influenza A in GenBank entry AAL60438.1 with the following predicted subsequence.

```
A,178,103
W,179,106
S,180,106
A,181,107
S,182,107
A,183,106
C,184,106
H,185,105
D,186,104
G,187,102
W,296,103
H,297,103
G,298,102
S,299,102
N,300,102
R,301,101
P,302,101
W,303,102
W,380,101
```

Another example is human neuraminidase (GenBank entry CAB41449.1).

```
H,200,102
D,201,104
H,202,106
G,203,107
R,204,107
T,205,107
W,206,107
A,207,105
R,208,101
H,297,102
P,298,102
T,299,102
H,300,102
```

Generally arbitrarily selected neuraminidases and hemagglutinins are true positives but the above kind of pattern, a segment of approximately 9–23 residues with a score exceeding 100 and centered on tryptophan with the peak score, is not universal. Sometimes it is another hydrophobic residue. In a "haemaglutinin repeat-containing protein" of a yet to be fully classified micororganism *Candidatus Kentron* a tryptophan (W) with a score of 118 is slightly displaced from the central peak of 119 associated with a run of three glycines (G). It is two glutamate residues (E), negatively charged that are in the vicinity in the sequence that appear to perturb the usual central role of tryptophan. Similarly, a rat neuraminidase is of some interest in that the predicted sequence SLDHGHTW surrounds the glycine (G) with peak score 106 and terminates at tryptophan (W) with a score of 102. The tryptophan is, however, immediately followed by a glutamate (E).

The significance of the above comments is as follows. Of the amino acids with charged sidechains, only aspartate (D) and histidine (H) appear to favor such binding in the present author's analysis, while glutamate (E) arginine (R), lysine (K) have zero valued parameters and so are contraindications of sialic acid or sialic acid glycan binding. However, the latter three charged amino acid residue can sometimes be found in scialic acid binding sites and in sites predicted as such by the present method. Indeed, they are often dominant features of residues directly interacting with sialic acid. In the sialic acid binding site of the globular head region of the Newcastle disease virus haemagluttinin a glutamate (E), three arginine residues (R) and a lysine (K) are intimate contact with sialic acid ligand. A difference is that these residues make intimate contact in space and are not together in a one or very few subsequences. That is, they do not necessarily constitute a sequence motif.

The actual extent of any counter-predictive impact of the above charged residues other than aspartate (D) and histidine (H), when they together in a subsequence of amino acids examined by the present algorithm, can be seen in false negatives. The neuraminidase of the plant *Striga asiatica* (Asian witchweed), was one of the false negatives: the

analogous region to many of the above, at least in the sense of surrounding the only tryptophan (W) and with the peak value, is EGAVDRW**R**GEANF where the tryptophan (W) has only a peak value of 88, and has an arginine (R), positively charged, on both sides. The other false negative in the original study was a bacterial (Chryseobacterium) haemagluttinin with the peak value of 93 at a tryptophan, but with two lysine residues (K), also positively charged sidechains, in the vicinity. Later studies also noted that N-acetyl neuraminic acid synthetase NeuB of *E. coli*, which has 4 tryptophan residues (W), but did not exceed a score of 95 which was the score for a phenylalanine (F), and there is a single lysine near the peak value at FN of 95 in subsequence FNLYK. This enzyme catalyzes the condensation of phosphoenolpyruvate and N-acetylmannosamine) to form N-acetyl neuraminic acid, so it is possible that the binding to N-acetyl neuraminic acid is much weaker in order to release the product.

Nonetheless, closer examination shows that the contraindicative effect of glutamate (E), arginine (R) and lysine (K is not strong in the context of the algorithm and typical subsequences, so failure to predict is in practice more an effect of the subsequence as a whole). In the above and similar cases, substituting the lysine or arginine by, say, a serine (S) in "computer experiments" does not typically increase the score of any residue to more than 100. For example, in the case of the witchweed neuraminidase, changing RWR to SWS raised the peak value at the tryptophan to 99, a significant increase but still not exceeding 100. In some cases a value exceeding 100 can of course be attained. A false negative also found in later studies investigating these issues was the ox neuraminidase, with the subsequence DDHGVSWRYGGGVS containing the tryptophan (W) with peak value 96. Changing to an serine (S) the arginine (R) adjacent to the tryptophan (W) did have an effect, albeit that the only change was having the tryptophan as the only residue predicted, with a marginal score of 101.

4.6. Predictions of group B proteins

Two kinds of potential true negative or false negative were distinguished in the study. The group that appeared to do particularly well at predicting those proteins that do not binding sialic acid or sialic acid glycan was, perhaps not surprisingly, group B, i.e. those proteins not expected to bind any kind of sugars. Out of the original sample of 10 proteins not expected to bind any kinds of sugars significantly, and that also confirmed that expectation, i.e. true negatives as far as predicting sialic acid binding is concerned, two (i) hemoglobin and (ii) trypsin precursor for which the prediction plots for which are shown in Fig. 4. The others not shown are mostly quite large proteins containing more than 5 tryptophan residues (W) sites, for which the method still correctly predicted as non-sialic-acid binders, and so worthy of some comments. They included (iii) human ubiquitin C of 685 residues and no tryptophans (W) and no residue score exceeding 56, in contrast to (iv) human progesterone receptor of 933 residues of which 6 were tryptophan but none of which exceeded 100 (one had the highest score of 95). The remaining true negative cases are (v) fatty acid oxidation complex subunit alpha FadB of *Acinetobacter calcoaceticus* of a substantial 717 residues, (vi) the mitochondrial NADH-ubiquinone oxidoreductase 75 kDa subunit of the camel, which comprised a substantial 733 residues but did notably not exceed a score of 77 for any residue, (vii) human cytochrome C with a maximum score of 87 for the second tyrosine (Y) in TGQAPGYS**Y**TATAANKN, and (viii) alcohol dehydrogenase (human, 1A) that did not exceed a score of 81 for any residue despite the "concern" that ethanol having basic sugar-like features and so could, *a priori*, be marginal. Perhaps unfairly included as rather small, (ix) proinsulin nonetheless does contain a tryptophan (W) which correctly did not exceed 100 and indeed only had a score of 72 in LLALLAL**W**GPDPAAA; the phenylalanine (F) in GPDPAAA**F**VNQHLCG had a highest score of 81 in the sequence. In the initial study, the case most closely approaching a false positive in this group was (x) human prothrombin with a substantial number of 622 residues, there was only one residue, glycine (G), that reached a score of 100, and a residue score should exceed 100 to classify the whole domain or protein as sialic acid binding. It is possibly best declared as an example of a marginal case. At the outset false positives were expected to appear in this non-sugar-binding group groups as the sample is increased, not least because of the preliminary nature of the method. Human angiotensin converting enzyme

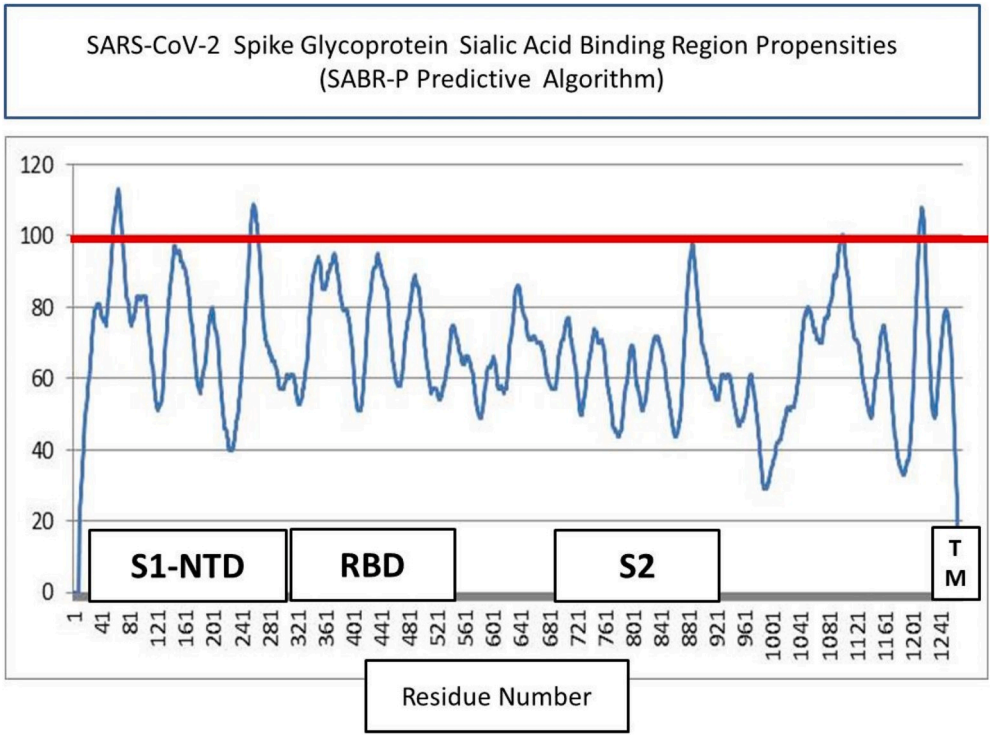

Fig. 5. Prediction of sialic acid glycan binding regions applied to the SARS-CoV-2 spike glycoprotein sequence.

type 2 (ACE2) was the first exception found and it is a significant exception in that it had two substantial regions 198–276 and 599–610 both exceeding scores of 100 throughout and peaking at substantial scores of 112.

4.7. Predictions of group C proteins

This group is interesting in that persistently predicting these as false positives might incline a researcher to abandon specifically predicting sialic acid glycan binding and instead consider their method as better positioned to predict sugar binding in general, but this turned out not to be the case. For Group C, i.e. those sugar-binding proteins that are believed to bind sugars but not to bind sialic acids or glycans containing them, the prediction plots for the 6 true negatives are shown to the lower right in Fig. 4. Human α-amylase, not shown, was an interesting false positive with subsequence of residues all only marginally exceeding 100 YSG**W**DFWGEGW but containing three tryptophans (W) of which the first had the highest score of 103. The human lysozyme precursor was also a false positive, albeit only marginally so, with one sequence KWESGYNTRA exceeding 100 and with the peak at tyrosine (Y) with a score of 103. In later studies examining the effect of considering weakly homologous sequences, the significantly different hen lysozyme precursor is interesting as still being a false positive but representing and even more marginal case, having one short subsequence with residues exceeding 100, namely WV, with tryptophan (W) with final score 102 followed by valine (V) with final score 101. While there are significant amounts of sialic acids in hen egg white, the author is not aware of any extract of hen egg white lysozyme that has these bound to the protein. Evidently comparisons of homologous proteins might be helpful to improve prediction power. Another false positive is the human ATP-dependent translocase ABCB1 isoform 2 which has a ribose binding site has a considerable size of 1280 residues yet with just one subsequence, SYALAF**W**YGMMYFSYAGCF, exceeding 100, and has the peak of 107 at the tryptophan (W). For such reasons, one may suspect that a fairer set of criteria for assessing predictive performance would be an average per residue basis. One of the more recent studies included the SARS-CoV-2 polyprotein which contains proteins with known RNA and ribose binding functions. Out of 7095 residues, 184 residues exceeded scores of 100. The proteins and domains there, however, include several of which all possible functions are not yet known. Evidently, those regions with scores over 100, by virtue of being in proteins of SARS-CoV-2, are certainly worthy of future further examination in this project. However, this was considered beyond present scope for the present paper, as the focus is on the spike glycoprotein.

The more recent studies of sugar binding proteins that did not include suspected sialic acid or sialic acid glycan binding regions included sugar isomerases, such as the L-rhamnose isomerase of *Rubinisphaera brasiliensis*, which with 423 residues containing 9 tryptophans (W) was a true negative. They also included a comparison between sugar transporter proteins, such as the UDP-galactose transporter of Drosophila melanogaster had 357 residues including 4 tryptophans (W), the PTS fructose transporter subunit EIIC of *Aeromonas hydrophila* with 589 residues containing 7 tryptophans, and a facilitated glucose transporter member 1, which despite having 492 residues containing 6 tryptophans (W), all of which were true positives by not predicting any sialic acid or sialic acid glycan ability. However, some sugar transporters were false positives. For example, the arabinose transporter of *E. coli* had the subsequence F**W**LYTA which exceeded 100 with the tryptophan scoring 104.

4.8. Predictions of none-covalent sialic acid glycan binding in SARS-CoV-2 spike glycoprotein

Armed with the above predictions and insights, one may make better informed judgements as to whether a domain for non-covalently binding host sialic acid glcycans may exist in the SARS-CoV-2 spike protein See Fig. 5. The above results suggesting the ability to distinguish between three classes of protein in relation to sialic acid glycan binding is perhaps surprising, not least because non-sialic sugars such as fucose and mannose can occur in cell surface glycans (including sialic acid glycans) as indicated in Introduction Section 1.3. This is discussed in Discussion Section 5.2. At this stage, only the ability to show propensities in different regions of the SARS-CoV-2 (GenBank entry MN908947.3) spike glycoprotein is required.

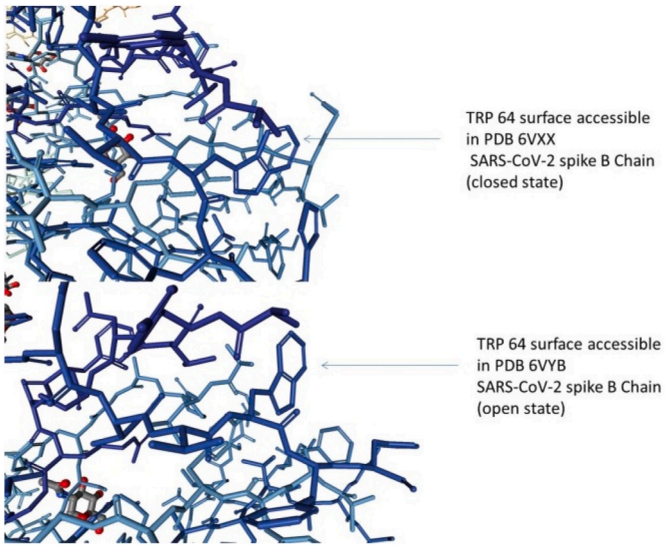

Fig. 6. The Tryptophan sidechain in FFSNVTWFHAIHV 58–70 of Spike Glycoprotein is Exposed in a Site that Has all the Appearance of a Sialic Acid Glycan binding site.

4.9. S1-NCD motif FFSNVTWFHAIHV

In this stage of the study the predicted regions of the SARS-CoV spike protein were examined in more detail. While as discussed above tryptophan can be involved in the fundamental structure of a sialic acid binding domain, an exposed tryptophan like that binding a sialic acid glycan in Fig. 2 would be particularly persuasive. The first predicted sialic acid binding motif is segment FFSNVTWFHAIHV 58–70 with SABR-P score ranging from 101 for valine (V) to 113 for tryptophan (W) and for phenylalanine (F) is visible in PDB entry 6VVX (spike closed state) and 6VYB (spike open state).

```
F,58,102
F,59,105
S,60,107
N,61,108
V,62,109
T,63,111
W,64,113
F,65,113
H,66,111
A,67,109
I,68,106
H,69,104
V,70,101
```

As shown in Fig. 6, the tryptophan sidechain in FFSNVTWFHAIHV as residues 58–70 of the sars-cov-3 spike glycoprotein is exposed in a site that has all the appearance of a sialic acid glycan binding site, comparable with the influenza virus B neuraminidase (PDB entry 2BAT) tryptophan site known to have interaction with and sialic acid, that was shown in Fig. 2.

It is useful to see how recurrent a potential more universal motif may be, to give insight, to avoid cross reactions of vaccine in human and veterinary patients as hosts, to detect an underlying common function that one might not wish to inhibit in the host with an anti-viral therapeutic, and so on. A BLAST search on non-viruses picks this subsequence

up as FFSNVTNIAWIHAI of *Parasteatoda tepidariorum*, the common house spider, and related sequences, a zinc figure domain because it contains FYVE the zinc finger motif, but more abundantly it picks up sugar-binding proteins such as glycosyl transferases such as FFSPVWARTPNVTWFH-HV of actinobacteria, Ribulose-phosphate 3-epimerase of animals, a-amylase of the Chitinophagia ("chitin eating") bacteria, C-type lectin 37Db-like of *Drosophila hydei*, and related sugar binding proteins, all varying around 100% cover and 55% match, 92% cover 71% match, 76% cover 70% match, and so on.

4.10. S1-NCD motif FFSNVTWFHAIHV

The second predicted sialic acid binding subsequence SSSGWTAGAAA has been of interest in the preceding Sections 4.1-4.4, where it seemed a likely sugar binding site based on circumstantial evidence such as homologies to neurmaindiases and glycan esterases. Unfortunately, it lies in the range that generally disordered in experimental 3D structure, e.g. residues 246–262 in VVX (spike closed state) and 243–262 in VYB (spike open state). It also has modest maximum scores of 109, but the above considerations do not prohibit its potential importance.

```
S,254,101
S,255,103
S,256,106
G,257,107
W,258,109
T,259,109
A,260,108
G,261,106
A,262,104
A,263,103
A,264,101
```

Again, as for the previous subsequence, it is useful to see how recurrent a potential more universal motif may be. BLASTp searches of non-viruses find subsequences in sugar binding proteins such as SSAGWTAGAA of microbacteria (90% cover 90% match), but compared with FFSNVTWFHAIHV discussed above, searche generate a lot of closer matches (many 100% matches with the top 99 at 100% cover 82% match or better) which, however, involve an even more diverse set of proteins. At first examination most appear less directly relevant to sugar, but many of these merit more examination. Details of each case are beyond present scope, but for example a particularly recurrent match example is SSSGWTAGA or similar sequences of proteins that contain the twin-arginine translocation pathway signal of most bacteria and archaea. For example such as *Ruegeria marisrubri*, of the *Rhodobacteraceae* has the above matching subsequence. The twin-arginine translocation pathway transports folded proteins across the cytoplasmic membrane of these microorganisms. The proteins are targeted by signal peptides containing a conserved twin-arginine motif, and the literature does not always mention any N-glycosylation and indeed there may not in every case be any directly relevant evidence of such. However, there is certainly known involvement in the production of secretory and extracellular N-linked glycoproteins in bacteria such as *Escherichia coli*.

4.11. S2 Motif HWKWPWYIWL

The third predicted subsequence is HWKWPWYIWL 1102–1218 with a peak score of 108 relates to IKWPWYIWL in the original Wuhan seafood market isolate (GenBank MN908947.3) and lies at the boundary between the C-terminal end of S2 and the transmembrane part 1214–1273. Its 3D structure is not, to the present author's knowledge, available, as it lies in residues 1147 onward are generally excluded from experimental structure (e.g. VVX and VYB).

```
H,1101,100
W,1102,100
K,1211,100
W,1212,104
P,1213,106
W,1214,108
Y,1215,107
I,1216,106
W,1217,104
L,1218,100
```

Again, matches with non-virus or host proteins may be of interest as biologically and medically important. BLASTp searches on non-viruses with the sequence as query inappropriately pick up a lot of coronaviruses by accidently relating to the host name, but this is indicative of a strong recurrence of the motif across coronaviruses. Otherwise, there is a diverse set of matches especially but not solely with bacterial proteins, and unfortunately most are described as hypothetical proteins for which the function is typically unclear. Of those that are named, there are some indications of involvements with sugar binding in many cases. Many are ribosome proteins or phosphatases relevant directly or indirectly to RNA or ribose binding. The PAP2 superfamily is characterized by a core consisting of a 5-helical bundle and includes functions involving glycerol phosphates but also sugars such as in the case of Glucose-6-phosphatase. This subsequence HWKWPWYIWL also matches sequences in proteins with an SPFH domain which is implicated in regulating targeted protein turnover in stomatins and other membrane-associated proteins. HWKWPWYIWL has of course a notable tryptophan (W) repeat that hints a special role of its own, certainly worthy of analysis but outside present scope. In general, however, it does appear that across many proteins it has other functions than sugar binding (perhaps diverse or multiple functions), although sugar binding is not excluded.

5. Discussion

5.1. Significance of the present work

The significance and innovation of the present work is that it proposes a sialic acid glycan binding function for the SARS-CoV-2 spike protein that has been largely neglected by other workers, apparently on the rationale that ACE-2 binding is the important first step in cell entry. Sites involved in the characteristic cap or knob of the spike protein appear partially persuasive in the light of their role as binding to host cells. There is a further possible site towards the base of the external part of the spike protein, which seems less likely by virtue of its position and weaker prediction. Interaction with sialic acid glycan with or without associated catalytic activity would be consistent with such functions observed in many respiratory and alimentary tract viruses, and not least in many or most other coronaviruses, and so such a function must be important to these viruses. On these grounds, it may be a target for therapeutic agents against SARS-CoV-2, particularly perhaps preventatives as well as means of impeding spread from lung cell to lung cell, and an exposed target for antibodies raised by synthetic vaccines. Although other authors have recently touched on such a glycan binding ability in SARS (as discussed in this paper above and particularly below), it has not been to the present author's knowledge analyzed in comparable detail and do not appear to relate to the same site. Nor do they propose a general prediction method for sialic acid glycan binding as described in the present paper. Of course, in the present paper this is still a prediction and not an experimental result, but it will hopefully encourage experimental researchers to investigate the glycan binding properties of SARS-CoV-2 more extensively. A further innovative feature

is that predictive method, which is expected to be worthy of investigation for the proteins of other viruses and even of other organisms. Like many predictive methods in bioinformatics it is not perfect, i.e. there are false positives and false negatives in prediction, so it is actually conceivable that the method is useful even if it is not correct in the particular case of SARS-CoV-2. In that sense, it may emerge as the more important contribution.

5.2. The quality of predictions by the current SABR-P algorithm and future work

The current SABR-P predictive algorithm is naïve and it is not expected that it will resemble closely the final refined form of the algorithm, which will based on more rigorously on principles closer to those of the GOR method [20], the Hyperbolic Dirac Net [21–23], the association Q-UEL language [24], and the BionIngine implementation including its new algorithms [25–28]. The impression of good performance for the current SABR-P method largely arises from the fact that it is only required to predict the sialic acid glycan binding properties of whole domains or proteins, not highly localized subsequences or surface patches. In essence, the method is really doing little more than capture and quantify in an algorithm the visual inspection of sugar binding domains and proteins and the observations of other workers as discussed above. However, the method was only required to help explore potential non-covalent sialic acid glycan binding sites in the spike glycoprotein, and in that regard it has proven adequate and valuable for present purposes. It also suggests a more refined approach may perform well because false positives and false negatives were mainly just over the boundary and just under it respectively. Resolution should be increased.

5.3. Distinguishing proteins binding different glycans and sugars

The current predictions also indicate a research direction in which to explore. The parameters for the general sugar binding capacity of amino acids residues are very different to those used by other workers and here the focus has been on sialic glycans versus other saccharide-based molecules. In this, perhaps the most surprising finding of all is the apparent ability of the method to distinguish between sialic acid containing glycans and other sugars in the case of lectins. This is because non-sialic sugars such as mannose and fucose can occur in sialic acid glycans, and prediction results hint that there is likely to be some distinguishing feature for a majority of cases that makes a specific recognition. In this respect it would seem initially of concern for the sensibleness of the predictions that, for example, mannose-binding lectin binds to a range of sugars that also include N-acetyl-D-glucosamine, N-acetyl-mannosamine, fucose and glucose. It is therefore possible that research in this direction will not be so profitable because the above distinguishing behavior of the algorithm might be to some extent coincidental. Be that as it may, the predictions are remarkably much better than expected, and should certainly be challenged by researchers in order to improve such methods.

5.4. Future challenges

Specifically, a larger sample may require a threshold adjustment or corresponding rescaling, perhaps resulting in a deterioration of performance particularly in regard to distinguishing lectins. Nonetheless, it is noteworthy that ultimately human glycan binding proteins have to overcome the same problem as the above kind of prediction algorithm. While this broad range of sugar recognition by the mannose-binding lection permits that lectin to interact with a wide selection of pathogens (viruses, bacteria, yeasts, fungi and protozoa) decorated with such sugars, there must be some kind of distinguishing aspect such that is not decoyed by the sialic acid glycans of the human host. A more mundane problem in extending the study is that the correct state as sialic acid glycan binding, other sugar binding, or not binding any kind of sugar, may be uncertain or a matter of degree. Further studies at time of writing suggested only about 70% for each of accuracy, sensitivity, and specificity, but this larger set is, as yet, of dubious quality for the purpose. Some proteins were believed, rather than known, to bind sialic acid glycans, binding might be weak or less specific or of multiple types, or the domain or approximate location of the binding site can be unclear. Related to that is a difficulty that the performance of any prediction method of this kind is defensible, and possibly unfairly defensible, in regard to false positives: it may be that experiment shows that a particular virus predicted to bind sialic acid glycans does not specifically do so, but perhaps it once did, in evolutionary terms. This is particularly relevant in regard to studying coronaviruses because, as discussed above, many coronaviruses certainly do bind sialic acid glycans. Of course, the prediction method would then still be subject to the criticism that it insufficiently sophisticated to manage the impact of small changes. For purely theoretical methods, that may be an issue for some time: in the present author's experience even simulations of binding of sugars to proteins in atomic detail tend to be difficult in view of the complex role of water molecules. For example, water molecules commonly represent protein-to-sugar bridges as discussed in this paper.

5.5. Biological implications

Potential biological implications arguably support the above prediction for SARS-CoV-2 spike protein. That is, the story "makes sense". Although the involvement of SARS-CoV and SARS-CoV-2 with sialic acid glycans has been rather neglected in the literature (but see below), such involvement represents a prominent and well known feature in the life history of influenza and other viruses, and appears no less important in the life of many other coronaviruses. Admittedly, HIV and many other enveloped viruses do not encode hemagglutinin for sialic acid binding. Instead, they interact using N-terminal sialic acid bound to envelope-associated proteins, like gp120 on HIV-1. However, the mode of infection is different. The SARS-CoV-2 coronavirus cannot jump in a magical way from contaminated surfaces to the lung, and it is doubtful that infective small loads of virus rely on chance to travel from the infecting person to the lung cell ACE2 receptor of the next human host. It is as yet unclear how many virus particles of SARS-CoV-2 are needed for infection, but the virus is clearly very contagious, and this may be because rather few particles are needed for infection.

In any event, the virus has to survive, and ideally even benefit for its survival, stages in a complex journey in mucus of a sneeze or on hands, face, eye, nose, or mouth, and in the various stages of the airway. Initial cell entry points are unlikely to be only the lung epithelium. SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes [29]. Viral mechanisms relating to these various surfaces could be fairly sophisticated. In biology in general, flexibility in carbohydrate recognition contributes to the targeting efficiency of carbohydrate-active enzymes in environments where there is diverse range of saccharides [18]. In a virus, more than one saccharide-binding site or multiple sugar binding sites in a protein could act to increase or decrease the overall affinity and increase or decrease virus mobility at different locations, while conformational changes could make available some sites and not others could regulate the extent of movement of the virus. Some binding sites have evolved to distinguish not just the sugar residue components but several types of monosaccharide or glycosidic bond linkage.

Once having reached the vicinity of a cell with an ACE2 receptor, the virus still needs to recognize the cell surface and raft across the cell surface to reach the ACE2 receptor. Fantani et al. [17] argued that a new type of ganglioside-binding domain exists at the tip of the N-terminal domain of the SARS-CoV-2 S protein, and that the subsequence 111–158, conserved among clinical isolates, may improve attachment of the virus to lipid rafts and facilitate contact with the ACE-2 receptor. This study also showed that, in the presence of CLQ or its more active derivative, hydroxychloroquine, the spike protein is no longer able to

bind gangliosides [17]. The present study does not support (nor necessarily refute) their conclusions in terms of such specific details, but the general argument concerning guidance to the ACE-2 receptor is compatible. Very recently Milanetti and colleagues have made available a preprint [30] that is tune with such ideas, and specifically states that binding sialic acids provides a second means of entry, other than ACE2. Rather like the approach of Thornton et al. [19], this is based more on interactions between surfaces of molecules in three dimensional space.

6. Conclusion

The results and conclusions of this study are speculative in the sense that they are applications of computers (using the techniques of bioinformatics and a new predictive method), and hence they are essentially theoretical. Their role has been to highlight the likelihood that the SARS-CoV-2 spike has a biological function of binding host cell sialic acid glycans (and probably across cells surfaces by that means, as discussed below). In particular, a domain in the cap or knob of the SARS-CoV-2 spike, which has so far been somewhat neglected, is involved in the non-covalent binding of host sialic acid glycans. It is perhaps curious that subsequences found as conserved by use of bioinformatics tools such BLASTp and Clustal Omega (also used here as described in Methods Section 4.1), or detected as a known or new functional motif, often seem in the literature to be considered as having the status of experiment or observation, while consideration of more complex patterns with more sequence options tend to be treated as theory and prediction. This caution is justified in the present study because further study and confirmation is required along the lines discussed in Discussion section 5 above. above. To the extent that it is a prediction, it is a prediction for SARS-CoV-2 made in advance of experiment in order to provide an objective and fair test of the methodology and it is hoped that it will stimulate experimental study in this area whether the experiments confirm or refute that prediction. Either result would likely be of ultimate medical importance. This is essentially typical of the more interesting roles of computers in biomedical research, although the general infrastructure and support that they provide for more routine tasks is of course of great importance.

The present paper possibly still stands as the first reported attempt to establish means of making use of sequence motifs that could be recognized between strains, albeit that the order and to some extent precise nature of the amino acid residues appears less important, or perhaps more subtle, than has been considered in previous papers in this series [4,5], which is why it required the development of the predictive technique (SABR-P). This will be advanced in future work, but the present method already helps to make quick comparisons between SARS-CoV-2 sequences and to consider the effects of viral mutations. However, it was surprising that a very simple approach was so useful, and it can easily be reproduced in a very few lines of computer program. The important consequence of the present study, however, is that there are already a variety of inhibitors of sialic acid binding that may serve as anti-viral agents, and this will be examined elsewhere.

Declaration of competing interest

This project used some of the knowledge gathering methods described in preceding papers although the work is described entirely in terms of standard bioinformatics tools. These knowledge gathering methods are used, amongst many others in an integrated way, in the algorithms and internal architectural features of the BioIngine.com, a distributed system developed by Ingine Inc. Cleveland, Ohio, for the mining of, and inference from, Very Big Data for commercial purposes.

References

[1] P.S. Masters, The molecular biology of coronaviruses, Adv. Virus Res. 66 (2006) 193–292.

[2] R. Lu, X. Zhao, J. Li, P. Niu, B. Yang, H. Wu, W. Wang, H. Song, B. Huang, N. Zhu, Y. Bi, X. Ma, F. Zhan, L. Wang, T. Hu, H. Zhou, Z. Hu, W. Zhou, L. Zhao, J. Chen, Y. Meng, J. Wang, Y. Lin, J. Yuan, Z. Xie, J. Ma, W.J. Liu, D. Wang, W. Xu, E. C. Holmes, G.F. Gao, G. Wu, W. Chen, W. Shi, W. Tan, Genomic characterization and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding, Published online January 29, 2020, www.thelancet.com, 2020, https://doi.org/10.1016/S0140-6736(20)30251-30258.

[3] B. Robson, Preliminary Bioinformatics Studies on the Design of Synthetic Vaccines and Preventative Peptidomimetic Antagonists against the Wuhan Seafood Market Coronavirus. Possible Importance of the KRSFIEDLLFNKV Motif, Circulated and Published on ResearchGate, 2020, https://doi.org/10.13140/RG.2.2.18275.09761.

[4] B. Robson, Computers and viral diseases. Preliminary bioinformatics studies on the design of a synthetic vaccine and a preventative peptidomimetic antagonist against the SARS-CoV-2 (2019-nCoV, COVID-19) coronavirus, Comput. Biol. Med. (2020), 103670 published online 26 February 2020.

[5] B. Robson, COVID-19 coronavirus spike protein analysis for synthetic vaccines, a peptidomimetic antagonist, and therapeutic drugs, and analysis of a proposed achilles' heel conserved region to minimize probability of escape mutations and drug resistance, Comput. Biol. Med. 121 (2020), 103749, https://doi.org/10.1016/j.compbiomed.2020.103749. Published online June 2020, In press.

[6] B. Coutard, C. Valle, X. de Lamballerie, B. Canard, N.G. Seidah, E. Decroly, The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade, Antivir. Res. 176 (2020), 104742.

[7] M.N. Matrosovich, Y.T. Matrosovich, T. Gray, N.A. Roberts, H.-D. Klenk, Neuraminidase is important for the initiation of influenza virus infection in human airway epithelium, J. Virol. 78 (22) (2004) 12665–12667.

[8] R. Vlasak, W. Luytjes, J. Leider, W. Spaan, P. Palese, The E3 protein of bovine coronavirus is a receptor-destroying enzyme with acetylesterase activity, J. Virol. 62 (12) (1988) 4686–4690.

[9] Q. Zeng, M.A. Langereis, A.L.W. van Vliet, E.G. Huizinga, R.J. de Groot, Structure of coronavirus hemagglutinin-esterase offers insight into corona and influenza virus evolution, Proc. Natl. Acad. Sci. Unit. States Am. 105 (26) (2008) 9065–9069.

[10] M.A. Behhzadi, V. Leyva-Grado, Overview of current therapeutics and novel candidates against influenza, respiratory syncytial virus, and Middle East respiratory syndrome coronavirus infections, Front. Microbiol. 10 (1327) (2019), https://doi.org/10.3389/fmicb.2019.01327. https://www.frontiersin.org/articles/10.3389/fmicb.2019.01327/ful.

[11] H. Lu, Drug treatment options for the 2019-new coronavirus (2019-nCoV), Biosci. Trends 14 (1) (2020) 69–71, https://doi.org/10.5582/bst.2020.01020, 2020.

[12] Fang Li, Receptor recognition mechanisms of coronaviruses: a decade of structural studies, J. Virol. 89 (4) (2015).

[13] X.W. Zhang, Y.L. Yap, The 3D structure analysis of SARS-CoV S1 protein reveals a link to influenza virus neuraminidase and implications for drug and antibody discovery, THEOCHEM 681 (1) (2004) 137–141.

[14] A. Seno N. Jimbo, I. Matsumoto, Tryptophan residues and the sugar binding site of potato lectin, J. Biochem. 95 (1) (1984) 267–275, 1984.

[15] L.C. Emily, T.E.E. Ooi, C.-Y. Lin, H.C. Tan, A.E. Ling, B. Lim, L.W. Stanton, Inhibition of SARS coronavirus infection in vitro with clinically approved antiviral drugs, Emerg. Infect. Dis. 10 (4) (2004) 581–586, https://doi.org/10.3201/eid1004.030458.

[16] C. Schwegmann-Wessels, G. Herrler, Sialic acids as receptor determinants for coronaviruses, Glycoconj. J. 23 (51–8) (2006).

[17] j. Fantini, C. Di Scala, H. Chahinian, N. Yah, Structural and molecular modelling studies reveal a new mechanism of action of chloroquine and hydroxychloroquine against SARS-CoV-2 infection, Int. J. Antimicrob. Agents (2020), https://doi.org/10.1016/j.ijantimicag.2020.105960 [Epub ahead of print].

[18] A.L. van Bueren, E. Ficko-Blean, Carbohydrate-binding modules, CAZYpedia. https://www.cazypedia.org/index.php/Carbohydrate-binding_modules. (Accessed 1 May 2020).

[19] C. Taroni, S. Jones, J.M. Thornton, Analysis and prediction of carbohydrate binding sites, Protein Eng. Des. Sel. 13 (2) (2000) 89–98, https://doi.org/10.1093/protein/13.2.89.

[20] J. Garnier, J.F. Gibrat, B. Robson, GOR method for predicting protein secondary structure from amino acid sequence, Methods Enzymol. 266 (1996) 540–553.

[21] B. Robson, Hyperbolic Dirac Nets for medical decision support, Theor. Methods Compar. Bayes Net Comp. Biol. Med. 51 (2014) 183–197.

[22] S. Deckelman, B. Robson, Split-complex numbers and Dirac bra-kets, Commun. Inf. Syst. 14 (3) (2015) 135–149.

[23] B. Robson, Bidirectional General Graphs for inference. Principles and implications for medicine, Comput. Biol. Med. 10 (2019) 382–399.

[24] B. Robson, T. Caruso, U.G.J. Balis, Suggestions for a web based universal exchange and inference language for medicine, Comput. Biol. Med. 1 (12) (2013) 2297–2310, 43.

[25] B. Robson, S. Boray, Implementation of a web based universal exchange and inference language for medicine. Sparse data, probabilities and inference in data mining of clinical data repositories, Comput. Biol. Med. 66 (2015) 82–102.

[26] B. Robson, Studies in using a universal exchange and inference language for evidence based medicine, Semi Automat. Learn. Reason. PICO Method. Syst. Rev. Environ. Epidemiol. Comput. Biol. Med. 79 (2016) 299–323.

[27] B. Robson, S. Boray, Studies in the extensively automatic construction of large odds-based inference networks from structured data. Examples from medical, bioinformatics, and health insurance Claims data, Comput. Biol. Med. 95 (2018) 147–166.

[28] B. Robson, Extension of the quantum universal exchange language to precision medicine and drug lead discovery. Preliminary example studies using the

mitochondrial genome, Comput. Biol. Med. 117 (2020), 103621 printed online Feb 2020, In press.
[29] W. Sungnak, Ni Huang, Christophe Bécavin, Marijn Berg, Rachel Queen, Monika Litvinukova, Carlos Talavera-López, Henrike Maatz, Daniel Reichart, Fotios Sampaziotis, Kaylee B. Worlock, Masahiro Yoshida, L. Josephine, Barnes, HCA Lung Biological Network,1., SARS-CoV-2 entry factors are highly expressed in nasal epithelial cells together with innate immune genes, Nat. Med. (2020), https://doi.org/10.1038/s41591-020-0868-6.
[30] E. Milanetti, M. Miotto, L. Di Rienzo, M. Monti, G. Gosti, G. Ruocco, In-Silico Evidence for Two Receptors Based Strategy of SARS-CoV-2, BioRyX, 2010, https://doi.org/10.1101/2020.03.24.006197 preprint (not peer reviwed), https://www.biorxiv.org/content/10.1101/2020.03.24.006197v1.

Barry Robson BSc(Hons) PhD DSc, Professor Emeritus Epidemiology Biostatistics & Evidence Based Medicine is a US and UK citizen. He was five years as Chief Scientific Officer IBM Global Healthcare, Pharmaceutical, and Life Sciences and, prior to that, six years as the Strategic Advisor at IBM Global Research Headquarters (T. J. Watson Research Centre). For most of those 11 years he held the prestigious title of IBM Distinguished Engineer. According to Barry's two page biography written by journalist Brendan Horton in Nature (389,418–420, 1997), Barry was a pioneer in bioinformatics, protein modeling, and computer-aided drug design, interests that he actively continues today. He is the recipient of several honours including the Asklepios Award for Outstanding Vision in Science and Technology at the Future of Health Technology Congress at M.I.T. in 2002. He has helped start up several other companies or divisions in the UK and USA. Barry continues as CEO of The Dirac Foundation in the UK, and Distinguished Scientist (Admin.) at the University of Wisconsin-Stout Department of Mathematics, Statistics, and Computer Science. He is also cofounder of Ingine Inc., Cleveland Ohio USA, a medical A.I. company. While continuing to work for, and then collaborate with IBM, he was also University Research Director and Professor of Epidemiology Biostatistics and Evidence Based Medicine at St. Matthew's University School of Medicine which he helped established in its earlier days in the Cayman Islands. Barry also holds a Harvard-Macy Certificate in the Business of Medical Education. Immediately prior to joining IBM in 1998 he was hired as Principal Scientist at MDL Information Systems in California to help put together the technology for the multimillion sale of a bioinformatics system to the holding company forming Craig Venter's Celera Genomics that produced the first draft of the human genome. Prior to that, he was CSO of Gryphon Sciences (later Gryphon Pharmaceuticals) in South San Francisco, California, a bio-nanotechnology ultrastructural chemistry start-up largely held and then acquired by SmithKline Beecham. Before moving to the US, Barry was the scientific founder of Proteus International plc in the UK, designing and leading the development of the PROMETHEUS Expert System and its underlying GLOBAL Expert System, bioinformatics and simulation language for drug, vaccine, and diagnostic discovery. It sold for the equivalent of $9.4 million to the pharmaceutical industry in the mid-1990s. At Proteus, he also led the team that used the above Expert System to invent and patent several diagnostics and vaccines including the Mad Cow disease diagnostic subsequently marketed worldwide by Abbott. He has over 300 scientific publications in Nature, Science, J. Mol. Biol. Biochemical J., including some 50 patents and two books: "The Engines of Hippocrates. From the Dawn of Medicine to Medical and Pharmaceutical Informatics" Robson and Baek, 2009, Wiley, 600 pages)" and "Introduction to Proteins and Protein Engineering" (B. Robson and J. Garnier, 1984, 1988, Elsevier, 700 pages). He has contributed to several reports to governments (EU, US, Denmark) including Panels of the National Innovation Initiative for "Innovate America" published by The Council on Competitiveness, Washington D.C. (2004) as a whitepaper to the President of the United States. He was also an advisor in relation to a major scientific computer-aided drug design collaboration and network for Peter Feinstein Consultants between work between US scientists and the Russian Science City Arzamas. For five years, Barry was a Nature "News and Views" Correspondent on biomolecules. He was Visiting Scholar Stanford University School of Medicine 1997–1998, Professorial Lecturer Mount Sinai NYC during part of his period at IBM Corporation, and held visiting positions and professorships in INRA and U. Paris-Sud France, and a Technical University of Copenhagen under Sir Rodney Cotterill, as well a postdoctoral position at Oxford (Wolfson College) under Sir David Phillips while Reader in Biochemistry at the University of Manchester.