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 125 (2020) 103963

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

The use of knowledge management tools in viroinformatics. Example study of a highly conserved sequence motif in Nsp3 of SARS-CoV-2 as a therapeutic target

B. Robson [a,b]

[a] Ingine Inc., Cleveland, OH, USA
[b] The Dirac Foundation, Oxfordshire, UK

ARTICLE INFO

Keywords:
Knowledge management
Coronavirus
SARS-CoV-2
COVID-19
Mutations
Conservation
Bioinformatics
Therapeutic
X domain
Macro domain

ABSTRACT

Knowledge management tools that assist in systematic review and exploration of scientific knowledge generally are of obvious potential importance in evidence based medicine in general, but also to the design of therapeutics based on the protein subsequences and fold motifs of virus proteins as considered here. Rapid access to bundles (clusters) of related elements of knowledge gathered from diverse sources on the Internet and from growing knowledge repositories seem particularly helpful when exploring less obvious therapeutic targets in viruses (for which knowledge new to the researcher is important), and when using the following concept. Subsequences of amino acid residue sequences of proteins that are conserved across strains and species are (a) more likely to be important targets and (b) less likely to exhibit escape mutations that would make them resistant to vaccines and therapeutic agents. However, the terms "conserved" and even "highly conserved" used by authors are matters of degree, depending on how distant from SARS-CoV-2 they wished to go in comparing other sequences. The binding site to the human ACE2 protein as virus receptor and human antibody CR3022 binding site on the spike glycoprotein are rather variable by the criteria used in the present and preceding studies. To look for more strongly conserved targets, open reading frames of SARS-CoV-2 were examined for extremely highly conserved regions, meaning recognizable across many viruses and organisms. Most prominent is a motif found in SARS-CoV-2 non-structural protein 3 (Nsp3). It relates to a fold called type called the macro domain and has remarkably wide distribution across organisms including humans with significant homologies involving three especially conserved subsequences (a) VVVNAANVYLKHGGGVAGALNK, (b) LHVVGPNVNKG, and (c) PLLSAGIFG. Careful study of the variations of these and of the more variable sequences between and around them might provide a finer "scalpel" to ensure inhibition of a vital function of the virus without impairing the functions of related host macro domains.

1. Introduction

1.1. Background

In late 2019 a strain of coronavirus [1], originally frequently called the Wuhan Seafood market isolate [2], was obtained from patients associated with the location of that name. The final confirmed genomic sequence MN908947.3 entered on the GenBank data base on 23rd January 2020 is held by many workers essentially to define the virus now known as SARS-CoV-2. A rapid response was made by the present author using bioinformatics analysis for design of synthetic vaccines and peptidomimetic therapeutics [3–6], the kind of study that is reasonably called "viroinformatics". This quick response, in which the research and writing of the first publically released papers took less than a week [3,4], was facilitated by knowledge gathering and processing tools with an Artificial Intelligence flavor [4]. These are methods of interacting with the World Wide Web. They are in a continuing research and development phase, and some of the development and testing is described in this paper. However, the focus is still on finding targets for attack against SARS-CoV-2.

1.2. The importance of conserved subsequences

Subsequences of virus proteins that are highly conserved are of

E-mail address: barryrobson@ingine.com.

https://doi.org/10.1016/j.compbiomed.2020.103963
Received 6 July 2020; Received in revised form 7 August 2020; Accepted 7 August 2020
Available online 13 August 2020

particular interest, for two reasons. First, they are likely to be involved in roles or functions that must be important to the virus for some reason, and so represent an "Achilles heel" worth examining as a vaccine and therapeutic target [5]. Second, in the design of vaccines and diagnostics, one should not, ideally, target a site that is highly variable, i.e. subject to escape mutations and appearance of drug resistance, otherwise a great deal of research time and effort can be wasted that could have been spent on solutions with more permanent effect. For present purposes, subsequences of amino acid sequence that are recognizable as related across many virus species, and even beyond, are of interest. A virus species is considered as class of viruses that comprises several strains, constitutes a replicating lineage, and occupies a particular ecological niche. For example, the KRSFIEDLLFNKV motif of the spike protein in the vicinity of the S2' entry activation cleavage site was of interest [4] because it seemed unique in the spike protein by having a broad conservation across all coronavirus species examined, even appearing to have traces in other nidoviruses [5]. A browser search for more recent closely related work, using as query the string KRSFIEDLLFNKV alone, revealed many hits (1780 Google hits on 5th August 2020) that did not show prior to the above publications [3–5] (a tiny few that did show earlier are substrings of longer sections of related sequences for other coronaviruses typically described around 2012). Inspection of current hits shows that industry appreciates the importance of such well-conserved sites. Organizations include synthetic chemistry companies subsequently selling the above peptide for research, and usually cite the above papers [3–5], while market prediction reports extract from the text of these papers to use it as the example of a substantial SARS-CoV-2 biopharmaceutical market, or even a more general peptide market. Of course, it is not the only specific site that is considered to be an interesting potential SARS-CoV-2 target, and the region that binds to the angiotensin converting enzyme type 2 (ACE2, that the virus uses as its main entry receptor) indirectly gets far more browser hits on the query string ACE2. It is certainly feasible as a target *a priori*, but the sequences involved are much more variable [4,5]. Approaches based on inhibiting ACE2, or for example antibodies directed against the spike glycoprotein [1] may buy valuable time until more enduring solutions are found, but much is to be potentially gained by parallel studies focused on relatively invariant regions of the genome as the permanent solution. The most recent paper [6] in this series nonetheless made the point that a region of the protein sequence can have conserved properties of functional importance that are not necessarily apparent in a specific order of amino acid residues along the primary sequence of a protein. The motif mainly discussed in the present paper has aspects of both: the overall fold is well conserved but short sections of its sequence are reasonably well conserved well beyond the virus kingdom.

1.3. Virus evolution and vaccine and drug resistance

Attention must be paid to avoiding escape mutations and emergent drug resistance because under the selective pressure of vaccines and therapeutics, virus evolution takes place at a rate many orders of magnitude faster than long term evolutionary distances between viruses might imply. At the time of writing this in early 2020, it was understood that there were already roughly 50 isolates around December 2019 to January 2020 with approximately, in the present author's estimation, 2.6 accepted base mutations and 1.4 missense mutations per isolate genome that result in amino acid residue changes. This provides some (albeit extremely rough) indication of the variations, i.e. 1-2 amino acid changes, that we can expect somewhere in the 29 proteins in the proteins of SARS-CoV-2 over the stage of evolution of the virus and number of people infected in early 2020. However, even though coronaviruses do not accept mutations particularly quickly compared with many other RNA viruses (e.g. Refs. [7,8]), the probability of new accepted mutations emerging is properly determined as proportional to the rate of mutation per virus times the number of viruses of that kind in the world. At the time of writing there are about 2 million COVID-19 cases worldwide, with some claims that this may represent only some 6% of infected people. Back-of-the-envelope calculation by the present author, considering rates of progeny production and shedding, viability and survival, suggests that the number of actual SARS-CoV-2 genome copies in the world could be around 10^{21}, each of roughly 30,000 RNA bases. That is 60,000 bits, i.e. possibly some 10^{26}–10^{27} bits of viral computational power worldwide working by natural selection to ensure survival of the specie. Be that as it may, it will undoubtedly be an astronomic number, and it would be wise to look far afield in coronavirus relationships to ensure that a subsequence of the proteins is well conserved. *A priori*, any motif of interest could even be found well beyond the virus kingdom, and study of it can indicate the kind of variations that are possible in the rapid evolution under the pressure from vaccines and therapeutics. Many workers consider that viruses and cellular organisms coexisted since earliest life, and exchanged genes. In general, it is the evolution of the gene that matters, not the organism [9]. By computing the rate of accepted mutations in certain genes, analyses of sequences relationships can determine the time that has elapsed since common ancestors and of the first emergence of a major group such as the coronaviruses [10]. It was once commonly held that the most recent common ancestor of coronaviruses existed around 10,000 years ago, but this is a relatively young age compared with the evolutionary history of their presumed natural hosts, which started to diversify tens of millions of years ago. By taking account of variation in the strength of natural selection over time, it has been found that the time to the most recent ancestor common for all coronaviruses is likely many millions of years, much longer than has previously been believed [10].

1.4. Previous computational work

Examples of related efforts specifically using computers directed against SARS-CoV and/or SARS-CoV-2 are diverse. Examples include means of monitoring and displaying the pandemic in real time (e.g. Ref. [11]), which helped the author understand emergent strains, and displaying the *interactome* of all the interactions between SARS-CoV-2 proteins and RNA, and human proteins [12], which helped interpret analyses by bioinformatics. While huge benefit for therapeutic research for COVID-19 can be gained by considering what was learned from the SARS outbreaks [13], *ab intio* studies of the three dimensional structures and functions of targets in SARS-CoV and SARS-CoV-2, i.e. using computational chemistry, are also taking a honorable place alongside experiment in the discovery of therapeutics [14], and approaches of this kind continue to play some role in the present project [4–6]. This involves modeling protein interactions with candidate drugs on a computer, a practice that was arguably not taken too seriously until the design of protease inhibitors in that way early in the AIDS epidemic. As for many viruses including HIV and coronaviruses, a protease target arises because the virus initially makes a large *polyprotein* from which two or more viral proteins are obtained by proteolysis by virus proteases. HIV protease has continued to be a popular model [15], and the experience naturally inspired research into therapeutics against the coronavirus of the earlier SARS outbreaks [16]. Such enzymes are by no means the only targets of interest. Another popular target during the earlier SARS outbreaks was the RNA preplicase [17].

1.5. Use of knowledge management tools

As described in Ref. [4], the current COVID-19 project was aided by algorithms of an Artificial Intelligence flavor; they are represented by a collection of modules known as the BioIngine (e.g. Refs. [18–26]). Of particular interest in the present work was the modules for automatic interaction with the World Wide Web to extract knowledge from natural language text, structured data, plus automated interaction with webpages for publically available tools. The underlying theory of knowledge representation and use in inference is based on Dirac's notation and algebra and has been developed by the author over several years, e.g. see

Refs. [18–21] and references therein. These efforts culminated in the Hyberbolic Dirac Net (HDN) as a network of probabilistic knowledge for inference, and the associated Q-UEL (Quantum Universal Exchange) language. Thus far, the use cases have been largely in the area of healthcare and standard clinical practice, public health and socioeconomic analysis, but a recent extension to bioinformatics and genomics [25] has been timely in providing a basis for the current SARS-CoV-2 project.

Because the importance of this technology in the context of virology is the ability to make rapid response to new viral epidemics [3–5], it is should be stated that the present paper was originally the third in the series by the present author responding to the appearance of the final version of the genome of the Wuhan Seafood Market isolate on GenBank. The author had significant early experience regarding rapid response to HIV, Bovine Spongiform Encephalopathy, and veterinary viruses, but quickly needed to obtain knowledge concerning SARS-like coronaviruses, aided by the above technology. The research was done in January and in February 2020, with the first recorded draft document on 1st March, and a version focusing on the standard bioinformatics rather than automated knowledge gathering was submitted to the present journal on the 23rd April. As demonstrating an early response to a new epidemic using bioinformatics, this was therefore reasonably successful, especially when considered along with the other papers in the series [3–6]. Unfortunately, however, the opportunity was also taken to extend earlier work on the automatic generation of Systematic Reviews [23] to automated construction of scientific papers, using the first version of this present paper, in a kind of blind "Turing test", with some minimal human intervention to ensure avoidance of plagiarism and appropriate accreditation throughout. Unfortunately, as a kind of "Grand Challenge" demonstration of Artificial Intelligence, that part of the project was evidently premature and over-ambitious: the reviewers objected that multiple things were discussed in the paper, looked more like a review, that the manuscript lacked adequate flow and structure, and required a substantial revision from scratch. Nonetheless, the future in which a "robot" may write an accepted not-too-short scientific paper that is accepted as-is may not be too far off, and the following theory, methods, and results examples should be taken of some indication of how that might be achieved. To some extent the following text does follow the appearance of knowledge extracted in order to support this idea.

2. Theory

2.1. Canonical representations of elements of knowledge

A conceptual theory behind having a computer automatically write a scientific paper (or at least a report) is usefully based on the idea of having a machine play the role of a student responding to a question in a science or medical examination paper, when this is the more traditional kind where the answer required is in the form of an essay. So far, success has been confined to the more modern form of multiple-choice examination for which there is a set of prewritten candidate answers [22]. The latter is the approach built on here, but of course both kinds of exam require a student to have some repository of knowledge. The elements of knowledge gathered in the present project and previous work [18–24] take a canonical form based on the *Dirac notation* with its *braket* <A|B> and *bra-operator*-ket <A | **R** |B>. When appearing in Q-UEL, it is called a "tag" by analogy with XML, and Q-UEL could be considered an extension of XML for probabilistic semantics. In most cases, the Q-UEL tags are to be used in reasoning can be seen as elements of knowledge or definitions, or facts and data concerning, say, a species, or a patient or the health of a country or US state. Examples of virological interest, as Q-UEL's somewhat XML-like tags, are as follows.

< SARS-CoV-2 | **'is the causative agent of'** | COVID-19 >
< 'Epstein-Barr virus' | **'may cause'** | fatigue >
< 'Epstein-Barr virus' | **causes** | increased 'serum aspartate aminotransferase activity' >
< 'Epstein-Barr virus' | **'does not cause'** | increased 'serum bilirubin' >
< 'Epstein-Barr virus' | **'does not cause'** | increased 'serum alkaline phosphatase' activity >

Note that relationships can be negative, which is important in weighing the balance of evidence for and against something when using tags like those above in automated reasoning: see below and discussion of module MARPLE in Methods Section 3.4.

For practical medical applications where quantum mechanical waves or other waves are not being considered, the imaginary part is based not on $\boldsymbol{i}$ such that $\boldsymbol{ii} = -1$, but on $\boldsymbol{h}$, also rediscovered in different guises by Dirac, such that $\boldsymbol{hh} = +1$. In theoretical physics derived from Dirac's system one uses the *spinor projectors*, in the above written $\mathfrak{u} = ½ (1+\boldsymbol{h})$ and written $\mathfrak{u}^* = ½ (1-\boldsymbol{h})$ and it is readily shown that $\mathfrak{u}\mathfrak{u} = \mathfrak{u}$, $\mathfrak{u}^*\mathfrak{u}^* = \mathfrak{u}^*$, $\mathfrak{u}\mathfrak{u}^* = \mathfrak{u}^*\mathfrak{u} = 0$, and $\mathfrak{u} + \mathfrak{u}^* = 1$. Note that an asterisk as a post-appended superscript represent complex conjugation, i.e. it changes the sign of the imaginary part, equivalent to changing $+\boldsymbol{h}$ to $-\boldsymbol{h}$ and *vice versa*. This leads to a formal grammar and algebra as a probabilistic semantic theory (e.g. Refs. [18–25]). In brief, a construction like < type 2 diabetes | **causes** | obesity > has the value $\mathfrak{u}$ P("type 2 diabetes causes obesity") + $\mathfrak{u}^*$P("obesity causes type 2 diabetes"), i.e. in general,

$$<A \mid \mathbf{R} \mid B> = \mathfrak{u}\ P(\text{“A R B”}) + \mathfrak{u}^*\ P(\text{“B R A”}) = \{0.85, 0.12\} \quad (1)$$

From the perspective of semantic theory, this is a semantic triple (ST), e.g. as in subject-verb-object. From a mathematical perspective, the above is a Dirac *bra-operator-ket* for the purely $\boldsymbol{h}$-complex case, since the P( ) are purely real values. The $\boldsymbol{h}$-complex value encodes the *probability dual*, {P("A R B"), P("B R A")}, so that in this case P("A R B"), = 0.85 for P("type 2 diabetes causes obesity"), and P("B R A") = 0.12 for P ("obesity causes type 2 diabetes".). There is the even more basic *braket* <X|Y>, considered even basic because from this bra row vectors < A| and ket column vectors |B> can be built from several such, and so can matrices as operators **R,** and also the bra-operator-ket in Eqn. (1) itself. Dirac's system is amazingly self-defining in a circular kind of way, so for present purposes, it is sufficient to think of <A|B> as the special case of <A | **if** | B>, i.e. when **R** is specifically the conditional relationship **if**. Relators **R** are most often Hermitian, as in quantum mechanics, such that <A | **R** | B > = (<B | **R** | A>)* = <B | **R*** | A>. Here **R*** naturally represents active-passive inversion of a verb, as in conversion of **causes** to '**is caused by**', or other relationships, e.g. (arguably) **if** to **then**. The information in the present study is either derived from medical text that is assumed to be authoritative, or represents an assertion awaiting refutation, so $<A \mid \mathbf{R} \mid B> = \mathfrak{u}1 + \mathfrak{u}^*1 = 1$. That is, unless a specific degree of truth, range of applicability, or degree of belief or reliability is clearly stated in the text, which is uncommon. A default scale based on "few", "most etc. does exist, but the ultimate default is 1 for several important reasons [26].

For the algorithms used in the present study, it has been convenient to define a *linear semantic multiple* (LSM), so-called to distinguish it from the semantic triple ST.

$$< A \mid \mathbf{R} \mid B \mid \mathbf{S} \mid C \mid \mathbf{T} \mid D \ldots. > = < A \mid \mathbf{R} \mid B > < B \mid \mathbf{S} \mid C > <C \mid \mathbf{T} \mid D> \ldots. \quad (2)$$

Eqn. (2) implies a the use of a logical operator **and** between bra-

opertaor-kets, but properly by default it described as **rand** which means that it assumes independence (random association). In the case of the probabilities being default 1 throughout, the result is of course trivially 1, but there are other operators, even many that are still logical operators, that will have different consequences. Most importantly, there is in our approach also a *context-dependent* functional operator **cand.** It multiplies probabilities form the **rand** result (typically 1 in the present study) by the extent of relationship in the characters words etc. in the readily decomposed into STs, here < A | **R** | B> and <A | **S**| C>. BSMs be seen in the so-called Q-UEL XTRACT tags discussed later below, and these tags are currently the only representations of branch knowledge structures in our system that are in a single tag. XTRACT tags are particular kinds of BSM in which an algorithm XTRACTOR as sought to process source text and maximize the linear LSM form, for example as follows.

<Q-UEL-XTRACT-Marple41Cstep2
"(virions?) [0https://en.wikipedia.org/wiki/Electron_microscopy] |**^have**| `a fringe |**of**| large **&or** bulbous ( ^surface?) projections (creating) |**as**| `an image reminiscent |**of**| `a crown |**of**| `the solar corona [0https://en.wikipedia.org/wiki/Solar_corona]: (fringe?) morphology [0https://en.wikipedia.org/wiki/Morphology_(biology)]; (image?) **|^is ^created by|** `the viral spike |**as**| S peplomers [0https://en.wikipedia.org/wiki/Peplomer] |**^are**| proteins [0https://en.wikipedia.org/wiki/Proteins] **|^populate|** 'the (^surface?) |**of**| `the virus; proteins |**^determine|** host tropism [0https://en.wikipedia.org/wiki/Host_tropism]"
(source:='https://en.wikipedia.org/wiki/Coronavirus' time:='Fri Jan 31 18:43:03 2020' extract:=450) Q-UEL-XTRACT-Marple41Cstep2>

(see Ref [22]) in the two bra-operator-kets being multiplied (on a scale 0…1, and often less than 1).

$$< A' |\mathbf{R}'|B' |\mathbf{S}'|C' |\mathbf{T}'|D' \ldots. > = < A |\mathbf{R}| B > \textbf{cand} < B |\mathbf{S}| C > \textbf{cand} <C |\mathbf{T}| D> \ldots. \quad (3)$$

Importantly, in actual Q-UEL applications, the tags representing STs on the right of the equation need not be STs: they can themselves be LSMs, or even approximations of LSMs as discussed shortly below. In simple forms of automated reasoning, A and A′ are often the same, and similarly for **R** and **R**′, B and B′ and so on, but this is not necessarily so here. This is not too important here because what matters is that the right hand side of Eqn. (3): it is a kind of chain or sometimes network of reasoning that links a block of text like a student exam question, to one or more statements that act like the candidate multiple choice answers in an exam. The left-hand side of the equation might be thought of as a pared-down representation of the exam question (albeit that it would be an ideal one from the student's point of view because it would be somewhat highly informative in pointing to the answer), and the right most term on the right-hand side could be considered as a representation of a candidate answer. Because the **cand** operator is symmetrical (i.e. x **cand** $y = y$ **cand** x) the result is a purely real probability when all the tags involved are purely real, and this is of course so when they the default value 1, as is normally so in the present case. In general, however, it should be thought of as the first value of the probability dual of Eqn. (3). In the applications in the present paper, this would be seen as an estimate P(answer[i] | question) where answer[i] is the ith candidate answer in a multiple-choice examination.

In the case of the module MARPLE that was primarily used to gather knowledge from the Internet in the present project [22], the algorithm is fairly tolerant of the kinds of tag used as knowledge elements, and of the quality of the way in which knowledge is represented in them, as follows. In automated analyses of natural language text to extract knowledge, parsing of sentences and reduction to a linear form is the main action, but reduction to a purely linear form is not always possible for any graph representation of knowledge. Like the parsed structure of a sentence as commonly perceived, there can be branch points, and these are represented in Q-UEL tags by semicolons. Such entities are reasonably called Branched Semantic Multiples (BSMs). Evidently, BSMs < A | **R** |B; A |**S**| C> or just < A |**R** |B; |**S**| C> if A is the "root", can be still be

Some clarification is beneficial at this point. Although all Q-UEL tags are designed to be readable by humans for reasons stated in Discussion Section 5.2, XTRACT tags are not usually seen by the user: they are intended to be used in automated inference. They can be seen as ungrammatical or stilted when presented directly to the human user. Nonetheless, direct access by the user, which was often the case in the present project, is still useful. One may also see how in principle these can be used directly to help write a scientific paper, by usually being automatically reorganized as much as possible into linear form they avoid plagiarism, oblige some rewriting (automatic or otherwise) to restore good grammar, and because they are insistent regarding sources as provenance. The XTRACT tags represents a large number of knowledge elements retained in a Knowledge Representation Store (KRS) that can be queried, and they not only provide latest information but also a time stamped record of early forms of the source, which is particularly true of Wikipedia entries for the regularly undated entries regarding COVID-19. Related to that is a more theoretical consideration: some Q-UEL applications can be as tolerant as the human reader, albeit in a simple way. In the module MARPLE discussed below, XTRACT tags are treated as in Eqn. (3), i.e. as if the components were LSMs, even if the XTRACTs do contain branched relationships. Since there may be chains in parallel of be different Q-UEL tags that relate, for example, a question to a possible answer, there may even be a fairly large complex network reminiscent of a Feynman path integral in quantum mechanics [22]. The approximation that the above implies emerges as found not to be serious in practice, but it should be technically called a use of "presyllogistic" logic or reasoning to distinguish it from a more logically rigorous (but more time consuming) method. It has been considered plausible that medical students sitting a medical licensing exam often take the same approach mentally [22].

3. Methods

3.1. The general approach

The most unusual method used in the present study is the repurposing of the multiple choice exam mode [22] as a way of obtaining a *bundle* of knowledge elements which are selected as related but which do

not necessarily come from the same source webpages. These direct the research and help considerably in writing up the report or paper. It quickly gathers elements of knowledge that are related in several ways,

(a) as the result of an optional logical query that first selects a subset of tags in the KRS for consideration, or a subdirectory of the KRS, e.g. infectious diseases,
(b) by a search automatically initiated first in the KRS and then using the World Wide Web, in order to form a chain of tags between question and each candidate answer based purely on similarity in their text content (Eqn. (3)),
(c) by containing links to the original source webpage and a pointer to position,
(d) by containing links that were embedded in the source text (when present), associated primary source references (when present),
(e) by same or similar data (time stamp) when rapidly changing information is involved, and
(f) by the action of dictionaries of words and phrases that control the "flavor" of data examined on the Internet, i.e. normally ensuring that it is authoritative medical text, but which could be adjusted to obtain popular, political, or other content.

Note that though in the current work the search starts with Wikipedia, it may automatically surf beyond it. The general approach and strategy used here, is described in Results Section 4.1, but in more methodological detail it is as follows.

(i) As a general preparation for the project in the manner of a review [23], the user writes a short paragraph in English describing the project and content of particular interest. This is seen as analogous to a question in a computer-based multiple choice examination.
(ii) The user then enters a list of key words and phrases of interest which he or she wishes the system to address. This is done by the algorithm first querying a large KRS and then by automatically browsing the World Wide Web. This list of points of interest is analogous to the list of candidate answers in the multiple choice examination. The relevant knowledge found is that represented by the tag or tags that form the path (or paths) between the question and each candidate answer, essentially as represented by Eqn. (3), noting the comment beneath it that the tags to the right and representing the path may themselves be LSMs or even approximations of such. A candidate answer need not appear directly in the question, in which case more than one tag separates question and an answer. However, the construction of paths more than three tags long is computationally intensive and is abandoned.
(iii) The Internet is accessed to find proteins that contain subsequences of amino acid residues that significantly match, and the protein and the species to which that protein belongs are noted. Of particular interest are those that match many diverse species. Also predictions by the Hyperbolic Dirac Net (HDN) learn to identify related proteins, and related domains as modular parts of proteins, by sequence [18,24].
(iv) Steps (i) and (ii) are repeated using proteins found above, e.g. ADP-ribose-1″-phosphatase in the present study. This can also lead to discovery in the literature of a more general notion such as a particular class of domain, e.g. the X domain and macro domain as highlighted in the present study.
(v) Highly conserved subsequences as potential functional motifs are studied in the context of the fold motif in which they occur, if a

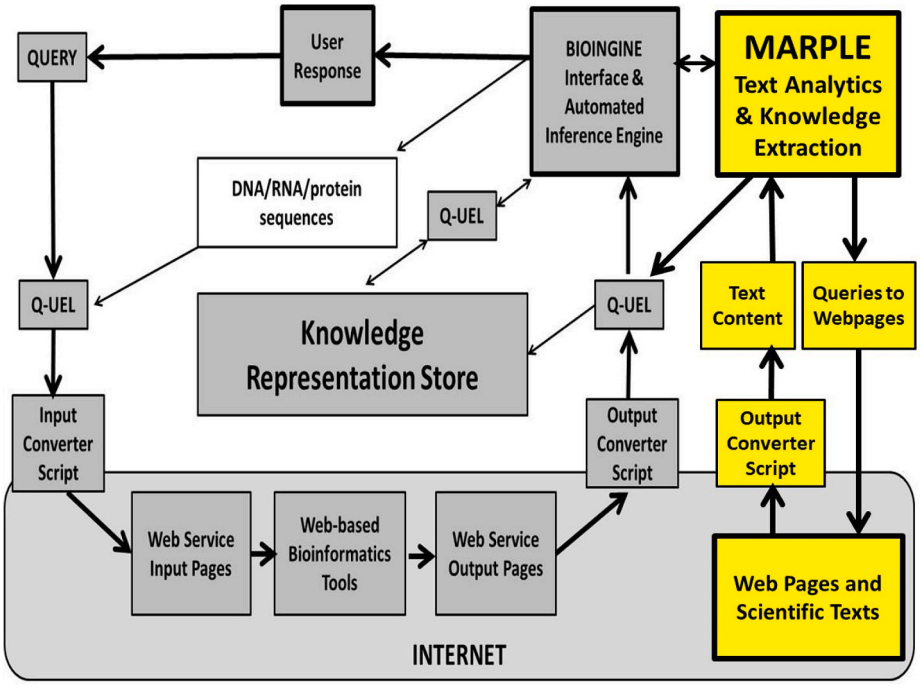


Fig. 1. Flowchart of the BioIngine when used in Virus Bionformatics.

relationship between subsequence and a particular fold motif is detected. One reason for this is that functional motifs that occur both in humans and SARS-CoV-2 need to be considered together in order to ensure that any therapeutic agents design to work against SARS-CoV-2 do not have an undesirable effect on humans.

(vi) Structural bioinformatics and computational chemistry tools are accessed to help design ligands that bind well to the SARS-CoV-2 motif as potential antagonists and hence a candidate for drug development.

In practice, there are many departures from the above particularly because the user will often wish to launch further queries against the KRS and World Wide Web in order to uncover further relationships and "drill down" to investigate certain aspects more deeply. In consequence it is best to consider the Q-UEL systems as composed of modules connected by Q-UEL but which can be invoked by the user as tools when needed. Fig. 1 shows the workflow as implemented in the BioIngine system. The BioIngine interface in the rightmost grey box in the upper right corner interacts with the human researcher. It provides decision support through automated inference tools [18–20] that use knowledge stored in the KRS as well as knowledge currently being returned from the Internet. The Q-UEL language is the architectural glue of the BioIngine, i.e. the main means of communications between modules. There is occasional use of comma separated value files with first row as metadata (column headers) and natural language text when these are steps in process that are useful in their own right, e.g. when reordering records in a structured for auditing in the former case, and the above "question and candidate answers" paradigm in the second.

The Q-UEL language is a means of computation for inference [18–20] as well as for interoperability [21], as well as the canonical form for storing knowledge in the KRS. This knowledge has been derived from natural language text analytics [22,23], and represents the activity in the yellow area of Fig. 1. In comparison with previous studies, relatively little use was made of structured data as a source of knowledge [24,25] with the important exception of biosequences, which are a special case discussed next in Section 3.2).

3.2. Standard tools and data sources used

"Bioinformatics knowledge" is derived from analysis of the content of, and relationships between, DNA, RNA, or protein sequences. As indicated in the grey area of Fig. 1, such knowledge can captured in Q-UEL canonical form by the use of converters that interact with webpages for publically available bioinformatics tools on the Internet [25]. This tends to be efficient and is important for the author and collaborators in order to capture appropriate knowledge in the KRS, but of course the results obtained can be reproduced by a researcher interacting in the standard way at the websites. A particularly important standard tool used was BLAST, in the present study mainly at the NIH National Library of Medicine site [27]. The common tools of bioinformatics have been described by Lesk [28]. Chemoinformatics (chemical informatics) [29] is also relevant for the design of therapeutic drugs against COVID-19, the principal tools of interest being drug screening *in silico*, involving docking of candidate drugs to protein targets followed by high-grade molecular dynamics simulations to determine free energy of binding. The ZINC database [30] is an example of a data base of a large collection of ligands, and is popular with researchers carrying out such work. All these have been used in the current SARS-CoV-2 project [4,5]. Some Q-UEL tags capturing, for example, protein sequence information can of course be regarded as source data and input for further analysis, e.g. as follows.

```
<Q-UEL-ORF-PROTEIN:=(application:='Perl version v5.16.3':= 'GenBank query',
tagtime(gmt):= 'Fri Gen 12 15:41:23 2020' source:=('GenBank
entry':=https://www.ncbi.nlm.nih.gov/nuccore/MN908947.3,
process:='GenBank query':= https://www.ncbi.nlm.nih.gov/genbank/,
definition:='Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome.
', accession:=MN908947, version:=MN908947.3))
ORF:='orflab'
product:= 'orflab polyprotein'
'protein id':='QHD43415.1'
sequence:= 'IUPAC 1 letter aa code':=
'MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQHLKDGTCGLVEVE
:
[omitted for brevity]
:
GQINDMILSLLSKGRLIIRENNRVVISSDVLVNN'
'size class':='protein':=full:=7097
|'has a well conserved subsequence as':='transformed by':=(converter:=BLASTPXTRACTION-
exptal4, 'using':=BLASTp:=
'https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LO
C=blasthome:=('Database':='Non-redundant protein sequences (nr)', 'Organism':= 'Viruses
(taxid:10239)','exclude:= '+', top:=100)) |
'conserved sequence':= 'IUPAC 1 letter aa code':=VVVNAANVYLKHGGGVAGALNK'
Q-UEL-ORF-PROTEIN>
```

As exemplified above, it is not unusual to include useful detail such as most conserved sequence of significant length to avoid repeating computations on the sequence in the future. The above, along with tags such as the following, exemplify the relatively simple forms of reasoning used in the present study, i.e. that the sequence starting VVVNAAN is a potential target for therapeutic design.

```
< Q-UEL-ORF-PROTEIN:=
 source:=('GenBank entry':=https://www.ncbi.nlm.nih.gov/nuccore/MN908947.3)
ORF:='orf1ab'
product:= 'orf1ab polyprotein'
'protein id':='QHD43415.1'
| 'is coded in genome of' |
pathogen:=Viruses:=Riboviria:=Orthornavirae:=Pisuviricota:=
Pisoniviricetes:=Nidovirales:=Cornidovirineae:=Coronaviridae:=Orthocoronaviri
nae:=Betacoronavirus:=Sarbecovirus:=SARS-CoV-2
Q-UEL-ORF-PROTEIN>

<Q-UEL-ORF-PROTEIN
pathogen:=SARS-CoV-2
|'has possible target site at'|
'conserved sequence':= 'IUPAC 1 letter aa code':=VVVNAANVYLKHGGGVAGALNK'
Q-UEL-ORF-PROTEIN>
```

3.3. Extraction of knowledge from the internet by XTRACTOR

The present study relied heavily on the ability of XTRACTOR to autosurf (automatically browse) the World Wide Web and gather knowledge in a canonical form (Q-UEL tags) that can be understood directly by the user. Natural language text on webpages is automatically reparsed to have LSM form (Eqns. (2) and (3)) wherever possible, i.e. the form that many modules can readily use in inference (see Theory Sec. However, they are typically directly readable by the human user as seen throughout this paper, and the "A.I. flavor" of the current approach is more in regard to the means by which they are gathered, as discussed in Section 3.4.

Wikipedia was mostly accessed, although the autosurfing procedure can move on to other kinds of site. Although a secondary source in the sense of being encyclopedic review of primary sources, Wikipedia can contain authors' interpretations, novel material and writing style of literary merit, so it is good practice to also reference the source of the Wikipedia text.

<Q-UEL-marple39
"`one billion `common colds [0https://en.wikipedia.org/wiki/Common_colds] |**^occur**| `each year |**In**| `the US
[9https://web.archive.org/web/20081001232444/http://www3.niaid.nih.gov/topics/commonCold/"
(source:='https://en.wikipedia.org/wiki/Lung diseases' time:='Mon May 25 14:35:18 2020' extract:=42)
Q-UEL-marple39 >

<Q-UEL-marple39
"|In| `the UK |**with _value ^approximately by**| 1 |**with _value in (every)?** | 7 individuals |**^are ^affected by**| 'some
(^form)? |**of**| chronic lung _disease 'most (^commonly) |**by**| chronic obstructive pulmonary _disease |**^includes|**

[0https://en.wikipedia.org/wiki/Chronic_obstructive_pulmonary_disease] asthma |**^includes|**
[0https://en.wikipedia.org/wiki/Asthma] &and chronic bronchitis |**^includes|**
[0https://en.wikipedia.org/wiki/Bronchitis#Chronic_bronchitis] &and emphysema |**^includes|**
[0https://en.wikipedia.org/wiki/Pneumatosis#Lungs] **|^includes|**
[11https://www.blf.org.uk/support-for-you/copd/what-is-copd"
(source:='https://en.wikipedia.org/wiki/Lung diseases' time:='Mon May 25 14:35:18 2020' extract:=43)
Q-UEL-marple39>

<Q-UEL-XTRACT-marple42 "respiratory disease) |**(with _value) ^support ^accounted for**| 93.3% |**of**|
ICU utilization |**in**| ^United ^States|[0https://www.hcup-us.ahrq.gov/reports/statbriefs/sb185-Hospital-Intensive-Care-Units-2011.jsp] |**In**| 2011"
| 'was extracted from' |
source:='https://en.wikipedia.org/wiki/Lung diseases' time:='Sat May 23 14:25:13 2020' extract:=46 Q-UEL-XTRACT-marple42>

<Q-UEL-Marple41 "`the GOR method |**^is ^based on**| probability
[0https://en.wikipedia.org/wiki/Probability] parameters **|Like**| Chou-Fasman; probability parameters
|^derived from| empirical studies |**of**| (known) protein tertiary _structures
[0https://en.wikipedia.org/wiki/Tertiary_structure] |**^solved by**| X-ray crystallography
[0https://en.wikipedia.org/wiki/X-ray_crystallography]" (source:='https://en.wikipedia.org/wiki/GOR method'
time:='Fri Jun 5 13:10:39 2020' extract:=60) Q-UEL-Marple41>

<Q-UEL-Marple41 "(`the GOR method |**^unlike**| Chou-Fasman; `the GOR method **|^takes into ^account not only**| 'the propensities **|of**| individual amino acids
[0https://en.wikipedia.org/wiki/Amino_acid] |**to ^form**| particular secondary structures; (However) **|but also** | `the conditional probability [0https://en.wikipedia.org/wiki/Conditional_probability] |**of**| `the amino `acid **|to ^form**| `a secondary _structure |**^given**| `its immediate neighbors |**^have already ^formed|**
_structure" (source:='https://en.wikipedia.org/wiki/GOR method' time:='Fri Jun 5 13:10:39 2020'
extract:=61) Q-UEL-Marple41>

The relator '**was extracted from**' is formally required by the Q-UEL specification's but in practice is usually omitted for brevity, because many such tags will be in the KRS. Refs. [31–27] are examples of primary sources providing XTRACTs of experimental findings directly relevant to the present study. XTRACT examples more explicitly appropriate to methodology are as follows.

<Q-UEL-Marple41 "`the GOR method |**^is ^based on**| probability [0https://en.wikipedia.org/wiki/Probability] parameters |**Like**| Chou-Fasman; probability parameters |**^derived from**| empirical studies |**of**| (known) protein tertiary _structures [0https://en.wikipedia.org/wiki/Tertiary_structure] |**^solved by**| X-ray crystallography [0https://en.wikipedia.org/wiki/X-ray_crystallography]" (source:='https://en.wikipedia.org/wiki/GOR method' time:='Fri Jun 5 13:10:39 2020' extract:=60) Q-UEL-Marple41>

<Q-UEL-Marple41 "(`the GOR method |**^unlike**| Chou-Fasman; `the GOR method |**^takes into ^account not only**| 'the propensities |**of**| individual amino acids [0https://en.wikipedia.org/wiki/Amino_acid] |**to ^form**| particular secondary structures; (However) |**but also** | `the conditional probability [0https://en.wikipedia.org/wiki/Conditional_probability] |**of**| `the amino `acid |**to ^form**| `a secondary _structure |**^given**| `its immediate neighbors |**^have already ^formed**| _structure" (source:='https://en.wikipedia.org/wiki/GOR method' time:='Fri Jun 5 13:10:39 2020' extract:=61) Q-UEL-Marple41>

3.4. MARPLE

The yellow area of Fig. 1 also represents the method that has the most "A.I. flavor" in the present paper. The gathering of bundles of related elements of knowledge and reasoning with them is primarily in regard to displaying contents of chains of XTRACT and other Q-UEL tags that link questions to candidate answers using the multiple-choice exam paradigm. Otherwise, any A.I. flavor does not from the use of the use of those Q-UEL tags for inference [26], but from the method of controlling the autosurfing the Internet that generates them, in order to obtain bundles of related knowledge as described at the beginning of Section 3.1. The process, which uses the paradigm of a multiple choice examination for medical students [22], does not require a user to see XTRACT tags, but the main purpose in the present case is to ensure that the automated browsing, via the in-text-links and links to references, stays on the track of relevant information. In effect, the candidate answers can be considered as queries, with the autosurfing kept on track (i.e. kept relevant) by the question and by so-called BUZWORDS and BADWORDS files. BUZWORDS consisted of hundreds words, phrases, medical terms and Greek and Latin roots statistically associated with authoritative medicine, while BADWORDs contains hundreds of roots words and phrases such as log in, news, report, music, flight, hotel, quote, awesome, league, victory etc. statistically associated with non-medical or less serious medical texts. If the score based on the latter exceeds that of the former, the webpage is abandoned.

An example question is as follows. "For the past week, an 18-year-old man has had fever, sore throat, and malaise with bilaterally enlarged tonsils, tonsillar exudate, diffuse cervical lymphadenopathy, and splenomegaly. There is lymphocytosis with atypical lymphocytes. The patient tests positive for heterophil antibodies." Given an extensive list of pathogens as candidate answers, MARPLE correctly gave the highest probability, 10.79%, to Epstein-Barr virus, significantly above the next possible answer *Streptococcus pyogenes* at 7.4%. In this case, the correct answer did have direct strong "pro-clues" and several incorrect answers had strong "anti-clues" in the KRS.

```
STRONG PRO-CLUE FOR < Streptococcus pyogenes > FROM KNOWLEDGE < Streptococcus pyogenes
| causes | sore throat >
STRONG PRO-CLUE FOR < EpsteinBarr virus > FROM KNOWLEDGE < Epstein-Barr virus | causes
| sore throat >
STRONG PRO-CLUE FOR < EpsteinBarr virus > FROM KNOWLEDGE < Epstein-Barr virus | causes
| lymphocytes | with | atypical features >
STRONG PRO-CLUE FOR < EpsteinBarr virus > FROM KNOWLEDGE < Epstein-Barr virus | causes
| cervical lymphadenopathy >
STRONG PRO-CLUE FOR < EpsteinBarr virus > FROM KNOWLEDGE < Epstein-Barr virus | causes
| splenomegaly >
STRONG ANTI-CLUE FOR < Streptococcus pneumoniae > FROM KNOWLEDGE < Streptococcus
pneumoniae | does not usually cause | sore throat >

Definitions and facts found for case 12.
A. Adenovirus in medical special service coded CPT 90476 Adenovirus vaccine, type 4,
live, for oral use in group Therapy - vaccination - other vaccine.
E. Chlamydia psittaci in risk factor coded DXX A70 Chlamydia psittaci infections in
group Empyema, bronchiectasis, Pneumonias.
T. Streptococcus pneumoniae in Streptococcus pneumoniae may cause mallaise.
U. Streptococcus pyogenes in Streptococcus pyogenes causes sore throat.
U. Streptococcus pyogenes in pharyngeal redness indicates Streptococcus pyogenes.
Found 5.
```

It is not always the case that clues are so direct. There can several pairs of interrelated tags that can give a link between the question and a candidate answer, and many can all be used at the same time, so the level for reporting clues is set to report only strong clues (in principle, medium and weak clues can be reported). MARPLE can be repurposed without modification for other applications as described in Results Section 4.1. It also served the manner of a simple symptoms checker. In the present study this was a useful means of soliciting information.

The final probability weighting of the candidate "answers" was similar in this case at 24% for (A) and 19% for the rest, but this does not have any quantitative significance in the present context and simply suggests that the process of reasoning found each query highly relevant to the "question" part. This kind of weighting was considered less important in the present context and the Internet can be set to be queried by the candidate "answers" even if the question is blank. Searching the established KRS for matches was previously the original default if there is no question.

```
CASE 1:
A patient reports that she "may have coronavirus". Symptoms seem severe. What is the
checklist of recent and present symptoms?
.....................................................................
A. (P= 4.16%) Cough or wheezing
B. (P= 3.56%) Chills
C. (P= 4.16%) Fever or raised temperature
D. (P= 4.16%) Shortness of breath or difficulty breathing
E. (P= 3.68%) Fatigue
F. (P= 3.57%) Aches and pains
G. (P= 3.62%) Headache
H. (P= 3.57%) Loss of taste
[other symptoms omitted for brevity]
.

Definitions and facts found for case 1.
A. Cough, wheezing, gasping associated with throat, airway and lung infections.
A. COVID-19 is consistent with cough
B. Chills in complication coded DX 7806 means "fever" in group Fever &amp;amp; Chills.
C. fever in complication coded DX 7806 means fever in group Fever &amp;amp;
D. Shortness of breath in risk factor coded DXX R0602 Shortness of breath in group
"Other Cardiovascular Disease"
D. COV-19 is associated with shortness of breath
E. fatigue in disease diagnosis coded DX 78079 means "Other malaise and fatigue"
in group "Sxs - malaise / fatigue"
F. pain in disease diagnosis coded DX 78096 means "Generalized pain" in group
"Sxs - symptoms -other - other"
G. headache in disease diagnosis coded DX 7840 Headache in group "NS - headache"
H. COVID-19 is consistent with loss of taste
H. COVID-19 is not consistent with heightened taste
[definitions and facts information omitted for brevity]
```

In response to the above question 3 as query, the following steps were reported, and indicate the methods of natural language text processing. Importantly, these steps each display information that is, to varying extents, directly useful to the user, meaning in this case the SARS-CoV-2 project. Recall that it is the HTML of the web page that is being analyzed, from which text is extracted in a form which XTRACTOR deduces is the essential text that the user wishes to see.

4. Results

4.1. Preparative studies on coronaviruses and respiratory diseases

Much of the information about SARS-CoV and SARS-CoV-2 was gathered in the manner of semi-automated Systematic Review [23], using MARPLE with the XTRACTOR module in the mode that addresses both the KRS and autosurfs the World Wide Web. An example at the start of the present study in January 2020, the query as "exam question" using MARPLE was as follows.

Case 3.
Investigation into the Wuhan seafood market coronavirus, bioinformatics analysis of the spike protein
(A) Wuhan coronavirus
(B) SARS
(C) Coronaviruses
(D) Epidemiology of coronaviruses
(E) Lung diseases

Table 1
Open reading frames of SARS-CoV-2 genbank entry MN908947.3Original Wuhan seafood market isolate.

Open Reading Frame	Amino acid residue sequence. Only the beginning of ORF1 and ORF2 are shown.	Cover% Identity% Range with other coronaviruses	Dominant Non-human hosts reported	Non-virus matches
ORF1ab polyprotein QHD43415.1	MESLVPGFNEKTHVQLSL PVLQVRDVLVRGFGD SVEEVLSEARQ….	(100%,99.97%) to (100%,86.10%)	bat	Broad animal distribution. e.g. Human RNA polymerase (1076–1173) Cover 129 Identities 41(32%) Positives 63(48%) Gaps 16(12%) ADP-ribose glycohydrolases, Superfamily I DNA and/or RNA helicases - replication, recombination and repair.
ORF2 S spike glycoprotein QII57278.1	MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRS….	(100%,99.92%), (94%,99.59%) to (100%,75.88%)	bat, pig	None significant overall outside the nidoviruses, but see Ref [3] for detailed comparisons. Does contain some viral binding motifs such as PPxY near the N-terminus (see underlined, Col. 2, left).
ORF3a protein QHD43417.1	MDLFMRIFTIGTVTLKQGEIKDATPSDFVRA TATIPIQASLPFGWLIVGVALLA VFQSASKIITLKKRWQLALSKGVH FVCNLLLLFVTVYSHLLLVAAGLEA PFLYLYALVYFLQSINFVRIIMRL WLCWKCRSKNPLLYDANYFLC WHTNCYDYCIPYNSVTSSIVITSG DGTTSPISEHDYQIGGYTEKWESG VKDCVVLHSYFTSDYYQLYST QLSTDTGVEHVTFFIYNKIV DEPEEHVQIHTIDGSSGVV NPVMEPIYDEPTTTTSVPL	(100%,99.64%) to (82%,26.99%) (72%,71.91%), (58%,25.54%)	bat	Weak matches from 28% cover 100% match down to 10% cover 41% match - bacteria, archaea, flatworms
ORF4 structural protein E protein envelope protein, QHD43418.1	MYSFVSEETGTLIVNSVLLFLA FVVFLLVTLAILTALRLCAYCCNIVNVSL VKPSFYVYSRVKNLNSSRVPDLLV	(100%,98.67%) to (97%,31.88%), (86%,28.79%) (73%,34.55%)	bat, ferret, mink	Weak matches around cover 25% matches 68% to cover 54% matches 48%, bacteria, trichomonads, fungi, flatworms, nematodes. Tapeworm *Echinococcus granulosus* of several animals notably dogs, Identities: 19/55(35%) Positives: 33/55(60%) Gaps: 0/55(0%) (host, e.g. contaminants?)
ORF5 M membrane glycoprotein HD43419.1	MADSNGTITVEELKKLLEQW NLVIGFLFLTWICLLQFAYAN RNRFLYIIKLIFLWLLWPVTLA CFVLAAVYRINWITGGIAIAMACL VGLMWLSYFIASFRLFARTRSMW SFNPETNILLNVPLHGTILTRPLLESEL VIGAVILRGHLRIAGHHLGRCDIKDLPK EITVATSRTLSYYKLGASQRVAGD SGFAAYSRYRIGNYKLNTD HSSSSDNIALLVQ	(100%,99.95%) to (98%,39.27%) (95%,42.06), (92%,41.95%), (90%,42.36%) (93%,33.17%)	bat, pig, camel, hedgehog	Weak matches around cover 89% matches e.g. 42%"Unknown E. coli protein" GenBank WP_148724442.1 4-205 Cover 203 residues Identity 87 (43%) Positives 124(61%) (host, e.g. contaminants?).
ORF6 protein QHD43420.1	MFHLVDFQVTIAEILLIIMRTFKVSIWNLD YIINLIIKNLSKSLTENKYSQLDEEQPMEID	(100%,98.36%) (100%,73.77%) (100%,67.21%) (100%,51.82%) to (81%,68.00%) (65%,55.00%) (100%,47.54%)	bat, civet	None reported. Salmonella enterica match withdrawn by submitter as contaminant.
ORF7a protein QHD43421.1	MKIILFLALITLATCELYHYQECVRGTTV LLKEPCSSGTYEGNSPFHPLADNKFAL TCFSTQFAFACPDGVKHVYQLRARSV SPKLFIRQEEVQELYSPIFLIVAAIVFITLCFTLKRKTE	(100%,99.17%) (100%,84.43%) to (98%,55.37%) (97%,23.20%) (62%92.74% (23%,100%)	bat, civet	No significant matches reported.
ORF8 protein QHD43422.1	MKFLVFLGIITTVAAFHQECSLQSCTQHQPY VVDDPCPIHFYSKWYIRVGARKSAPLIELC VDEAGSKSPIQYIDIGNYTVSCLPFTINC QEPKLGSLVVRCSFYEDFLEYHDVRVVLDFI	(100%,99.17%) (100%,94.21%) to (100%,49.59%) (98%,30.65%) (99%,28.00%) (97%, 27.64%) (43%, 35.19%)	bat, civet bat	Weak matches around cover 43% matches 35%, many with bacterial nitrogenase cofactor biosynthesis. Also Galleria mellonella, moth of the family Pyralida Identities: 25/86(29%) Positives: 39/86(45%) Gaps: 6/86(6%).

(*continued on next page*)

Table 1 (*continued*)

Open Reading Frame	Amino acid residue sequence. Only the beginning of ORF1 and ORF2 are shown.	Cover% Identity% Range with other coronaviruses	Dominant Non-human hosts reported	Non-virus matches
ORF 9 nucleocapsid phospho-protein QHD43423.2	MSDNGPQNQRNAPRITFGGPSDSTGL PNNTASWFTALTQHGKEDLKFPRGQ GVPINTNSSPDDQIGYYRRATRRIRGG DGKMKDLSPRWYFYYLGTGPEAGLP YGANKDGIIWVATEGALNTPKDHIG TRNPANNAAIVLQLPQGTTLPKGFY AEGSRGGSQASSRSSSRSRNSSRN STPGSSRGTSPARMAGNGGDAAL ALLLLDRLNQLESKMSGKGQQQQ GQTVTKKSAAEASKKPRQKRTAT KAYNVTQAFGRRGPEQTQGNF GDQELIRQGTDYKHWPQIAQFAP SASAFFGMSRIGMEVTPSGT WLTYTGAIKLDDKDPNFKDQVILLNKHID AYKTFPPTEPKKDKKKKADETQALP QRQKKQQTVTLLPAADLDDFSKQLQQSMSSADSTQ	(89%,100%) to (46%,92.59%), (77%,47.53%)		Matches from cover 87% matches 50% with E. coli, especially nucleoprotein. Apolygus lucorum, a species of true bug in the Miridae family Identities: 19/55(35%) Positives: 29/55(52%) Gaps: 0/55(0%).
ORF10QHI42199.1	MGYINVFAFPFTIYSLLLCRMN SRNYIAQVDVVNFNLT	(86%,100%) to (81%100%)	(human only reported)	No significant matches reported.

Doing (A).
Relevance score was 1.12 and https://en.wikipedia.org/wiki/Wuhan coronavirus does not contain mention of “Wuhan seafood market coronavirus” per se. Xtract Maker call 1. <markers> <non-qualifying pronouns> <nouns> <picturable nouns> <adjectives> <compound 4 word prepositions> <compound 3 word prepositions> <compound 2 word prepositions> <prepositions> <adverbs> <regular verbs> <irregular verbs> has been built.
CROSS REFS (IN-TEXT LINKS FOUND):
[0] noref
[1] 0https://en.wikipedia.org/w/index.php?title=Wuhan_coronavirus&redirect=no
[2] 0https://en.wikipedia.org/wiki/2019%E2%80%9320_Wuhan_coronavirus_outbreak
[3]0https://en.wikipedia.org/wiki/Talk:Novel_coronavirus_(2019-nCoV) #Requested_move_22_January_2020
[4] 0https://en.wikipedia.org/wiki/Portal:Current_events
[5] 0https://en.wikipedia.org/wiki/2019%E2%80%9320_Wuhan_coronavirus_outbreak
[6] 0https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources#Breaking_news
[7] 0https://en.wikipedia.org/wiki/Wikipedia:Risk_disclaimer
[8] 0https://en.wikipedia.org/wiki/Talk:Novel_coronavirus_(2019-nCoV)
[9] 0https://en.wikipedia.org/wiki/Help:Maintenance_template_removal
[10] 0https://en.wikipedia.org/wiki/Virion
[11] 0https://en.wikipedia.org/wiki/Virus_classification
[12] 0https://en.wikipedia.org/wiki/Virus
[13] 0https://en.wikipedia.org/wiki/Synonym_(taxonomy)
[14] 0https://en.wikipedia.org/wiki/Simplified_Chinese_characters
[15] 0https://en.wikipedia.org/wiki/Wuhan coronavirus#cite_note-Brain-num-9
[16] 0https://en.wikipedia.org/wiki/Wuhan coronavirus#cite_note-Brain-num-6
[17] 0https://en.wikipedia.org/wiki/Wuhan coronavirus#cite_note-Brain-num-2
[18] 0https://en.wikipedia.org/wiki/Traditional_Chinese_characters
[19] 0https://en.wikipedia.org/wiki/Wuhan coronavirus#cite_note-Brain-num-2
[20] 0https://en.wikipedia.org/wiki/Wuhan coronavirus#cite_note-Brain-num-0
[21] 0https://en.wikipedia.org/wiki/Wuhan coronavirus#cite_note-Brain-num-2
[22] 0https://en.wikipedia.org/wiki/Wuhan
[23] 0https://en.wikipedia.org/wiki/Hubei
[24] 0https://en.wikipedia.org/wiki/China
[25] 0https://en.wikipedia.org/wiki/2019%E2%80%9320_Wuhan_coronavirus_outbreak
[26] 0https://en.wikipedia.org/wiki/Genomic_sequencing
[27] 0https://en.wikipedia.org/wiki/Positive-sense_single-stranded_RNA_virus
[28] 0https://en.wikipedia.org/wiki/Coronavirus
[29] 0https://en.wikipedia.org/wiki/Huanan_Seafood_Wholesale_Market
[30] 0https://en.wikipedia.org/wiki/Zoonosis

There is a great deal of content in web pages like those of Wikipedia that is not relevant and currently changes in this and in the mode of presentation typically require corresponding changes to the converters in Fig. 1. The converters are fairly smart and tolerant, but not indefinitely so, and sometimes have to be modified to keep up with webpage and html styles. As of January 2020, an example of text identified as relevant was as follows.

Virus:
Novel coronavirus (2019-nCoV)
Wuhan, China, the epicenter of the only recorded outbreak
Synonyms [#13]
Wuhan coronavirus[1] (simplified Chinese [#14]: #### [#15] #### [#16]#### [#17]; traditional Chinese [#18]: #### [#19] #### [#20] ### [#21])
Wuhan seafood market pneumonia virus[2]
The novel coronavirus (2019-nCoV),[3][4] informally known as the Wuhan coronavirus,[1] is a contagious virus that causes respiratory infection and has shown evidence of human-to-human transmission, first identified by authorities in Wuhan [#22], Hubei [#23], China [#24], as the cause of the ongoing 2019#### Wuhan coronavirus outbreak [#25].[5] Genomic sequencing [#26] has shown that it is a positive-sense, single-stranded RNA [#27] coronavirus [#28].[6][7][8]
Because many early cases were linked to a large seafood and animal market [#29], the virus is thought to have originated in animals [#30], but this has not been confirmed.[9] Comparisons of the genetic sequences of this virus and other virus samples have shown similarities to SARS-CoV [#31] (79.5%)[10] and bat coronaviruses (96%),[10] indicating that an origin in bats is likely.[11][12][13]

PROCESSING
Question/Case 3 (A) Wuhan coronavirus.
Link https://en.wikipedia.org/wiki/Special:MyTalk has relevance=52%
Preliminary identification of <subject clause| and |relator clause|.

---

and so on. XTRACT tags are themselves bundles of elements of knowledge. Recall that to facilitate extraction of semantic triples, XTRACTS are if necessary reparsed into *linear semantic multiples* (LSMs) as much as possible, and to do this, semantic triples such as < late 2017 Chinese scientists |**ˆtraced**| the virus > are identified first, and then assemble into the LSMs. For example, <late 2017 Chinese scientists |**ˆtraced**| the virus>, <the virus |**through**| the intermediary>, and <the intermediary |**of**| civets > were generated from a subsequent web page on SARS (see example below), automatically loaded via a link in the preceding web page, and would if occurring alone become, at a near final stage, the LSM < late 2017 Chinese scientists |**ˆtrac**ed| the virus | t**hrough**| the intermediary |**of**| civets >. In practice, as shown below, the final LSM was even longer by joining more semantic triples. In general the assembly need not come from content of one webpage; that is the case in the present study, although separate sentences with common content from a paragraph can be joined (or sentences split). Some branches in XTRACTs are inevitable in some cases, and a semicolon '; ' is used to indicate a break in the linear relation, and hence a branch point. The presumed noun phrase following the semicolon is typically the first (subject) noun phrase encountered. Some indication of the initial "raw" processing into what XTRACTOR considered semantic triplets can be deduced by examining the following; evidently XTRACTOR is imperfect in some cases. In many cases these triplets can be used as knowledge elements. Any that are not usefully informative are not likely to do any harm except to take up unnecessary storage.

<Severe acute respiratory syndrome |**as**| SARS>
<SARS |**^is**| a viral respiratory disease [8]>
<a viral respiratory disease [8] |**of**| zoonotic [9] origin>
<zoonotic [9] origin |**^caused by**| the SARS coronavirus [10]>
<the SARS coronavirus [10] |**as**| SARSXhyphenXCoV>
<SARSXhyphenXCoV |Between| November 2002 and July 2003 >
<November 2002 and July 2003 an outbreak |**(^outbreak)? of**| SARS>
<SARS |**in**| southern China [11]>
<southern China [11] |**^caused**| an eventual 8098 cases>
<an eventual 8098 cases |**with value ^resulting in**| 774 deaths reported>
<774 deaths reported |**with value in**| 17 countries&XhashX911&XhashX93>
<17 countries&XhashX911&XhashX93 |with| the majority>
<the majority |**of**| cases>
<cases |**in**| mainland China and Hong Kong&XhashX912&XhashX93>
<China and Hong Kong&XhashX912&XhashX93 |**with value as**| 96XpcentX fatality [12] rate>
<96XpcentX fatality [12] rate |**according to**| the World Health Organization [13]>
<the World Health Organization [13] |**as**| WHO &XhashX912&XhashX93 No cases>
<WHO &XhashX912&XhashX93 No cases |**of**| SARS>
<SARS |**^have ^been ^reported**| worldwide>
<worldwide |**with value since**| 2004&XhashX913&XhashX93>
<2004&XhashX913&XhashX93 |**In**| late 2017 Chinese scientists>
<late 2017 Chinese scientists |**^traced**| the virus>
<the virus |**through**| the intermediary>
<the intermediary |**of**| civets [14]>
<civets [14] |**to**| cave-dwelling horseshoe bats>
<horseshoe bats [15] |**in**| Yunnan [16] province&XhashX914&XhashX93>

and so on. There are readily processed to bullet points.

- <Severe acute respiratory syndrome |**as|** SARS |**^is**| a viral respiratory disease [8] |**of**| zoonotic [9] origin |**^caused by**| the SARS coronavirus [10 |**as**| SARS-CoV>
- < SAR-CoV |**Between**| November 2002 and July 2003 an outbreak |**(^outbreak)? of**| SARS |**in**| southern China [11]>
- < SAR-CoV |**^caused**| an eventual 8098 cases |**with value ^resulting in**| 774 deaths reported |**with value in**| 17 countries&XhashX911&XhashX93 |**with**| the majority |**of**| cases> |**in**| mainland China and Hong Kong&XhashX912&XhashX93 |**with value as**| 96% fatality [12] rate>
- <96XpcentX fatality [12] rate |**according to**| the World Health Organization [13] |**as**| WHO &XhashX912&XhashX93 >
- < No cases |**of**| SARS |**^have ^been ^reported**| worldwide |**with value since|** 2004&XhashX913&XhashX93>
- < late 2017 Chinese scientists |**^traced**| the virus SAR-CoV |**through**| the intermediary |**of**| civets [14]] |**to|** cave-dwelling horseshoe bats [15] |**in**| Yunnan [16] province&XhashX914&XhashX93>

These are also linked together in a subsequent step before tag tidying as discussed next. Automatic attempts to correct and to tidy are made in the final stages of forming an XTRACT tag. This includes a step in which links in text (starting '[0) and links in cited references (starting '[1', '[2' etc.) are directly expressed in the XTRACT tag, so that they may be read by human eye as well as the computer. In some cases sentences are split up into entries in separate tags, and sometimes XTRACTOR makes the judgement that they sufficiently intertwined to represent one knowledge element. For example:

<Q-UEL-XTRACT-Marple41W. "Severe acute respiratory syndrome |**as**| SARS |**^is**| `a viral respiratory _disease [0https://en.wikipedia.org/wiki/Respiratory_disease] |**of**| zoonotic [0https://en.wikipedia.org/wiki/Zoonotic] origin |^caused by| `the SARS coronavirus [0https://en.wikipedia.org/wiki/SARS_coronavirus] |**as**| SARS-CoV |Between| November 2002 {AND} July 2003 'an (^outbreak) |**of**| SARS |**in**| southern China [0https://en.wikipedia.org/wiki/Southern_China] |**^caused**| `an eventual 8098 cases |**with _value ^resulting in**| 774 _deaths reported |**with _value in**| 17 countries [1https://www.sciencedirect.com/science/article/abs/pii/S0277953606004060?via%3Dihub] |**with**| `the majority |**of**| cases |**in**| mainland| China {AND} Hong Kong [2https://www.who.int/csr/sars/country/table2004_04_21/en/] |**with _value as**| 96/ fatality [0https://en.wikipedia.org/wiki/Case_fatality] _rate |**according to**| `the World Health _Organization [0https://en.wikipedia.org/wiki/World_Health_Organization] |**as**| (WHO) [2https://www.who.int/csr/sars/country/table2004_04_21/en/] |**as**| `No cases |**of**| SARS |**^have ^been ^reported|** worldwide |**with _value since**| [3https://www.nhs.uk/conditions/sars/; |**In**| late 2017 Chinese scientists |**^traced**| `the virus |**through**| `the intermediary |**of**| civets [0https://en.wikipedia.org/wiki/Civets] |**to** | cave-dwelling horseshoe bats [0https://en.wikipedia.org/wiki/Horseshoe_bat]"
| 'was extracted from' |
source:='https://en.wikipedia.org/w/index.php?title=SARS&redirect=no' time:='Fri Jan 31 15:20:26 2020' extract:=69 Q-UEL-XTRACT-Marple41W>

The Wuhan seafood market Wikipedia entry was continuously being updated on at very least a daily basis, emphasizing the value of Q-UEL XTRACT tags including a time stamp as well as other provenance. Future mining of essentially the same knowledge elements will thus differ and with the time stamps represent an analyzable chronological development. No direct reference to "spike protein" was found at the time of this study in a query to Wikipedia (whether automated as above or manual), and MARPLE received a message *"The page 'Spike protein' does not exist. You can ask for it to be created, but consider checking the search results below to see whether the topic is already covered"*. However, other tags already created that mention both spike protein and virus and coronavirus can heal the broken path through the series of links.

4.2. External sites described as well conserved in the literature

The virus surface sites considered as well conserved in the SARS-CoV-2 spike glycoprotein by the present author have been described in Refs [2–5], although Ref [5] was concerned with the need for sialic acid glycan functional "hemagglutinin-like" site that is determined by conserved properties of the subsequence rather than specific order of amino acids, and appears at different sites in different coronaviruses. Ref [4] has already addressed the type 2 angiotensin converting enzyme (ACE2) which is a popular target for researchers and considered the site of SARS-CoV-2 binding to it to be relatively variable. An example of a Q-UEL XTRACT tag that stimulated these studies was as follows.

Q-UEL-Marple41W2
"`the spike |**as**| S |**of**| envelope |**as|** E |**of**| membrane |**as**| M {AND} nucleocapsid |**as**| N |**In**| `the specific case |**of**| `the SARS coronavirus | **as ^see below|**
[0https://en.wikipedia.org/wiki/Coronavirus#Severe_acute_respiratory_syndrome] **|^defined ^receptor-binding|** domain |**on**| S mediates |**as**| `the attachment |**of**| `the virus |**to**| `its cellular receptor |^angiotensin-converting| enzyme 2 [0https://en.wikipedia.org/wiki/Coronavirus#cite_note-Brain-num-2] |**as**| ACE2 [8 https://www.semanticscholar.org/paper/Structure-of-SARS-coronavirus-spike-domain-with-Li-Li/bbedaafec1ea70e9ae405d1f2ac4c143951630bc] coronaviruses |**as ^specifically by**| `the members |**of|** Betacoronavirus [0https://en.wikipedia.org/wiki/Betacoronavirus] subgroup |**as**| `A |**also ^have**| `a spike-like protein |**^called**| hemagglutinin esterase [0https://en.wikipedia.org/wiki/Hemagglutinin_esterase] |as| (?_he) [4NOLINKREF de Groot RJ, Baric R, Enjuanes L, Gorbalenya AE, Holmes KV, Perlman S, Poon L, Rottier PJ, Talbot PJ, Woo PC, Ziebuhr J (2011). 'Family Coronaviridae'. In AMQ King, E Lefkowitz, MJ Adams, EB Carstens (eds.). Ninth Report of the International Committee on Taxonomy of Viruses. Elsevier, Oxford. pp. 806–828] "
(source:='https://en.wikipedia.org/wiki/Coronavirus' time:='Fri Jan 31 17:36:06 2020' extract:=453)
Q-UEL-Marple41W2 >

Table 2
Summary of VVVNAAN Domain Core BLASTp Matches on the Animal kingdom.

Species	Group	Score
Acropora millepora	stony corals	79.0
Orbicella faveolata	stony corals	74.7
Stylophora pistillata	stony corals	73.2
Pocillopora damicornis	stony corals	72.4
Branchiostoma belcheri	lancelets	72.0
Actinia tenebrosa	sea anemones	70.9
Mizuhopecten yessoensis	bivalves	70.9
Apaloderma vittatum	birds	70.5
Crassostrea virginica	bivalves	70.1
Terrapene carolina triunguis	turtles	68.9
Chrysemys picta bellii	turtles	68.9
Platysternon megacephalum	turtles	68.9
Protobothrops mucrosquamatus	snakes	68.6
Gouania willdenowi	bony fishes	68.2
Saccoglossus kowalevskii	hemichordates	68.2
Tetraodon nigroviridis	bony fishes	68.2
Sphaeramia orbicularis	bony fishes	68.2
Thamnophis sirtalis	snakes	68.2
Python bivittatus	snakes	67.8
Rhinatrema bivittatum	caecilians	67.8
Egretta garzetta	birds	67.4
Chelonia mydas	turtles	67.4
Lingula anatina	brachiopods	67.4
Lottia gigantea	gastropods	65.1
Exaiptasia pallida	sea anemones	63.2

Following identification of publications that relate to escape form vaccine and therapeutic agents (e.g. Refs. [31–34]), there were searches with MARPLE queries to find what other authors considered as *well conserved* subsequences of SARS-CoV-2 spike glycoprotein. A paper finding a neutralizing antibody (CR3022) against Sars-CoV-2 obtained from an earlier SARS-CoV patient was entitled "A highly conserved cryptic epitope in the receptor-binding domains of SARS-CoV-2 and SARS-CoV" [35] and is of considerable interest. However, this was not found to be "highly conserved" in the sense used in the present paper. Analysis of the Protein Data Bank 6W41Crystal structure of human antibody CR3022 in complex with a SARS-CoV-2 domain by the present author showed that the following spike protein subsequences, and especially those in italics, had significant interactions with the antibody.

RGDEVRQIAPGQTGKIADYN,

LDSKVGGNYNYLYRLFRK,

EIYQAGSTPCNGVEGFNCYFPL*QSYGFQPTK*GV

If these CR3022 binding regions are to continue to serve as epitopes without escape mutations, one would hope to see them as not differing by non-conservative mutations in early SARS isolates. For example, we would hope that SARS isolates from Hong Kong patients have similar subsequences, and they certainly are [36] when compared with the large range of the less close variants [4] and of more distant coronaviruses [5]. Nonetheless, one sees differences within those regions that would not normally be considered as conservative substitutions: serine (S) to phenylalanine (F) (small polar to large non-polar), threonine (T) to methionine (M) (small polar to large non-polar), and histidine (H) to asparagine (N) (large partially charged to small neutral polar). An example in the above region of interest is the following BLASTp comparison of SARS-CoV-2 with a Hong Kong SARS-CoV isolate Genbank ABD72970.1.Opportunities for escape mutation should, nonetheless, be considered in the light of comparisons between more distantly related coronaviruses, because any pattern of changes in a single coronavirus strain is evidently preserving the virus protein structure and functions. Unlike the KRSFIEDLLFNKV motif, which is easily recognizable in the coronaviruses of birds and reptiles and perhaps even as remnants in fish coronaviruses [3], one does not have to go too far from the Wuhan seafood market isolate MN908947.3 to other strains and species to find that the above CR3022 antibody binding regions start varying considerably. Different coronavirus strains even in the same host species can differ drastically in these regions, as for the human common cold coronaviruses, where parts of the corresponding subsections are arguably just recognizable due to the help of an overall spike protein alignment using Clustal Omega at https://www.ebi.ac.uk/Tools/msa/clustalo/. However, only a phenylalanine (F), cysteine/cystine (C), and threonine (T) are conserved. For example, this shows up well in the Wuhan seafood market isolate MN908947.3 when aligned examples the two dominant strains of human common cold coronavirus NP_073551.1 and AIV41987.1. The differences between coronavirus long established in hosts increases dramatically with taxonomic difference between host species, for example, in a full Clustal Omega alignment of MN908947.3 spike protein sequences with avian spike proteins such as KX266757, KC119407.1, KM454473, NC016991, and NC016993. The corresponding regions found by that alignment method are unrecognizable as related to the antibody binding subsequences. Only two cystines (C) and a tyrosine (Y) are conserved in the antibody binding regions. Recall that under selective pressure of vaccines and therapeutics, virus evolution precedes many orders of magnitude faster than the above evolutionary distances might imply, not least because there is a substantial yield of virus particles per cell [37] and a large global prevalence of SARS-CoV-2 as discussed in Section 1.1. Consequently, the above notion of "highly conserved" site as for the CR3022 antibody binding does not seem likely to be conserved sufficiently to provide a long term solution, in the author's opinion.

4.3. Extended search for conserved subsequences in other SARS-CoV-2 proteins

The above was extended to broader exploration of conserved regions of the SARS-CoV-2 using BLASTp [27,28] whether used interactively at a website [28] or accessed via a utility such as the BioIngine. See Table 1. In the last column the non-virus matches of all the ORF amino acid residue sequences are shown (the ORF 1 and 2 sequences are shown partially only because of their considerable length). This line of investigation supported not only the idea that ORF 6 and 10 have no obvious matches so far outside moderately related coronaviruses, but also (as yet) with no other sequences in living organisms. A this initial semi-automated level, all ORFs we were treated as one protein although the product of ORF 1 is a polyprotein from which several proteins are derived from the ORF 1 protein product by proteolytic cleavage. It is the matches of non-viral proteins with the ORF1 polyprotein that are the most persuasive, and one group is of particular interest for exploration the present paper. Other interesting matches with products of the polyprotein will be discussed elsewhere.

In the other ORFs, i.e. not ORF 6 and 10, there are often matches with gene sequences from bacteria and parasites, some of which that could, in principle, could be due to contamination by viruses, notably in the gut. Some submitters have withdrawn entries matching viruses on that interpretation. Also some cellular pathogens and parasites could acquire a protein or domain by gene transfer. However, contamination and even gene transfer after early life on Earth seems unlikely to be the explanation for ORF 3a non-virus matches, which include many thermophilic bacteria species, e.g. from thermal vents and hots springs, and archae. The archae can be found on skin, but they are not generally considered as pathogens, although some share some characteristics with known pathogens that could imply the potential to cause disease.

4.4. Brief studies on the SARS-CoV-2 small open reading frames (SORFs)

In principle, a virus can match a host protein not just by amino acid sequence but by some kind of related function. Small open reading frames (SORFs) for ribosomal synthesis of small peptides [25] are of relatively recent interest in molecular biology, and SORFs 6 and 10 of SARS-CoV-2 (again, see Table 1) contained some features, not least small size, reminiscent of the content of some of the 85 Q-UEL

knowledge tags describing known and predicted human mitochondrial mini-proteins that serve cytoplasmic signaling functions [25]. There is also a potential functional connection because these signaling processes are also involved in the innate immune response to viral infection, and SARS-CoV-2 products might have evolved to interfere with these signals to the benefit of the virus. One such KRS entry is as follows, which relates to viruses via the KRS tags < humanin | is | a mitochondrial derived peptide > and < mitochondrial derived peptides | 'may be involved in' | viral infection >. It is a typical example of a Q-UEL knowledge tag for genomics and bioinformatics tag.

sense of a RNA-dependent-RNA-polymerase activity does not necessarily apply to all protein products coded in ORF1. Rather, ORF1 is considered by researchers as not only concerned with proteins with such activity but also with those proteins that indirectly support replication, or use some similar recognition functions, use similar catalytic chemistry to do similar tasks with different purposes. Consistent with the entry for this ORF in rightmost column of Table 1, replication and transcription was found in the literature detected by MARPLE with XTRACTOR to be a dominant theme of ORF1 for the 30-kb plus-strand SARS-CoV-2 RNA genome and these are elaborate processes. It takes place at cytoplasmic membranes and involves continuous and discontinuous RNA synthesis

<Q-UEL-ORF-PROTEIN:=(application:='Perl version v5.16.3':=FASTAtoTags.txt, tagtime(gmt):='Thu Mar 8 13:46:53 2018' source:=(patient:=75229, process:=interpretation:='ORFfinder+BLAST':='https://www.ncbi.nlm.nih.gov/orffinder/') 'for patient':=75229 'putative mitochondrial protein':='best match':='ORF8:2953:3054 SHLP6 (Small Humanin-like Peptides)' **| is |** sequence:='IUPAC 1 letter aa code':='MLESMSTMGFTTSMLDQDIPMVQPLLKVRLFND' 'size class':='peptide-miniprotein (21-35 aa)':=33 Q-UEL-ORF-PROTEIN>

There was some initial interest in relation to a purine binding motif, a superfamily of actual or putative helicases of bacteria, yeast, insects, mammals, pox and herpesviruses, three groups of positive strand RNA viruses, and significantly a mitochondrial product. However, no significant sequence matches have been found, yet, between the SARS-CoV2 genome and the human mitochondrial genome.

4.5. The subsequence VVVNAANVYLKHGGGVAGALNK

In the studies for Table 1, the subsequence of any significant size that was most conserved, in the sense of the variety of not only virus but also cellular species matched, was SARS-CoV-2 ORF1 subsequence VVVNAANVYLKHGGGVAGALNK. Similar subsequences starting VIVNAAN… were a common variant. I-for-V represents a very conservative substitution as both are β-branched aliphatic residues differing by a single CH_2 group. In bacteria the top 100 matches varied from 100% cover 86% residue match and 90% cover 85% residue match down to 100% cover 78% match and 77% cover 88% match. By "cover" is meant that the subsequence match would not have matches at one or both ends of the VVVNAAN… subsequence, i.e. it is "only a piece" of the query sequence. Thermophilic bacteria are prominent in the best matches, some of which are at least as good as those found in the coronaviruses themselves. All these strong matches are not, however, identities. As one might expect, viruses most closely related to SARS-CoV-2 had best matches, but they still showed some variation. A match with an original SARS-CoV coronavirus sequence (SARS coronavirus TJF Genbank entry AAT76146.1) described as "replicase 1 ab" was as follows.

by the viral replicase, a large protein complex expressed by the 20-kb replicase gene, notably two thirds of the genome. This complex, non-structural and used inside the host cell, is currently believed to have 16 viral origin subunits (Nsps 1–16) and can make use of several hijacked cellular proteins. The Nsps have multiple enzymatic functions, including protease, polymerase, helicase, and RNase activities. The RNA-dependent RNA polymerase, RNA helicase, and protease activities are common to RNA viruses. In addition, originally based on sequence analysis, the coronavirus replicase is believed to employ a variety of other RNA and nucleotide processing enzymes, some of which are absent or rare in other RNA viruses. For the majority of these proteins, MARPLE with XTRACTOR obtained few hits and apparently the available functional information is still limited.

4.6. Subsequence VVVNAANVYLKHGGGVAGALNK is not part of the RNA-dependent RNA polymerase from SARS-CoV-2

Very frequently and prior to COVID-19 emergence, coronavirus proteins with sequences homologous to the above have been considered as directly representing part of the replicase mechanism. A key point as that the above subsequence is *not* found to relate to RNA-dependent RNA polymerase for genome replication, nor indeed to any currently popular target for of therapeutic drugs against CoV-SARs-2, despite the fact that its high degree of conservation suggests a function important to the virus. The subsequence and the region of sequence following it, all of which is described later below as the *VVVNAAN core*, is not the same as that of the RNA-dependent RNA polymerase from SARS-CoV-2 as in, for example, in entries 6M71, 7BTF 7BW4 and 7BV1 in the protein Data Bank https://www.rcsb.org. Nor does it relate to any of part of them.

```
Replicase 1ab [SARS coronavirus TJF]
Sequence ID: AAT76146.1 Length: 7073 Number of Matches: 1
Range 1: 1034 to 1147
Identities: 91/114(80%) Positives:91/114(80%) Gaps:1/114(0%)
Query  1     VVVNAANVYLKHGGGVAGALNKATNNAMQVESDDYIATNGPLKVGGSCVLSGHNLAKHCL  60
             V+VNAAN++LKHGGGVAGALNKATN AMQ ESDDYI  NGPL VGGSC+LSGHNLAK CL
Sbjct  1034  VIVNAANIHLKHGGGVAGALNKATNGAMQKESDDYIKLNGPLTVGGSCLLSGHNLAKKCL  1093

Query  61    HVVGPNVNKGEDIQLLKSAY-NFNQHEVLLAPLLSAGIFGADPIHSLRVCVDTV  113
             HVVGPN+N GEDIQLLK+AY NFN  ++LLAPLLSAGIFGA P+ SL+VCV TV
Sbjct  1094  HVVGPNLNAGEDIQLLKAAYENFNSQDILLAPLLSAGIFGAKPLQSLQVCVQTV  1147
```

Note, however, that "replicase 1 ab" above clearly applies to the overall polyprotein reading frame residues 1 to 7073 as the above states, and therefore essentially to the whole "orf1ab polyprotein" (protein ID QHD43415.1) of 7096 residues for SARS-CoV-2 Wuhan Seafood isolate GenBank MN908947. Referring it to having a replicase function in the

The alignment given by Claustal Omega at https://www.ebi.ac.uk/Tools/msa/clustalo for these polymerases and VVVNAANVYLKHGGGVAGALNK and the region of sequence following, is shown below. It is not significant, though one may note that some weak match between the polymerases and the overlapping part VYLKHGGGVAGALNKATNNAM from SARS-CoV-2 ORF1 is suggestive.

```
VVVNAANcore       ------VVVNAANVYLKHGGGVAGALNKATNNAMQVESDDYIAT--NGPLKVGGSCVLSG 52
7BV1_1|Chain      AQVLSEMVMCGGSLYVKPGGTSSGDATTAYANSVFNICQAV-TANVNALLST------DG 713
7BTF_1|Chain      AQVLSEMVMCGGSLYVKPGGTSSGDATTAYANSVFNICQAV-TANVNALLST------DG 712
6M71_1|Chain      AQVLSEMVMCGGSLYVKPGGTSSGDATTAYANSVFNICQAV-TANVNALLST------DG 712
7BW4_1|Chain      AQVLSEMVMCGGSLYVKPGGTSSGDATTAYANSVFNICQAV-TANVNALLST------DG 703
7BTF_2|Chain      ------------------------------------------------------------ 0
6M71_2|Chain      ------------------------------------------------------------ 0
7BW4_3|Chain      ------------------------------------------------------------ 0
7BV1_3|Chain      ------------------------------------------------------------ 0
7BTF_3|Chains     --------------SLPSYAAFATAQEAYEQAVANGDSEVVLKKLKKSLNV-AKSEFDR 51
6M71_3|Chains     --------------SLPSYAAFATAQEAYEQAVANGDSEVVLKKLKKSLNV-AKSEFDR 51
7BW4_2|Chains     --------------SLPSYAAFATAQEAYEQAVANGDSEVVLKKLKKSLNV-AKSEFDR 51
7BV1_2|Chains     --------------SLPSYAAFATAQEAYEQAVANGDSEVVLKKLKKSLNV-AKSEFDR 52

VVVNAANcore       HNLAKHCLHVVGP----N--VNKGEDIQLLKSAY-NFNQHEVLLAPLLSAGIFGADPIHS 105
7BV1_1|Chain      NKIADKYVRNLQHRLYECLYRNRDVDTDFVNEFYAYLRKH-------FSMMILSDDA--- 763
7BTF_1|Chain      NKIADKYVRNLQHRLYECLYRNRDVDTDFVNEFYAYLRKH-------FSMMILSDDA--- 762
6M71_1|Chain      NKIADKYVRNLQHRLYECLYRNRDVDTDFVNEFYAYLRKH-------FSMMILSDDA--- 762
7BW4_1|Chain      NKIADKYVRNLQHRLYECLYRNRDVDTDFVNEFYAYLRKH-------FSMMILSDDA--- 753
7BTF_2|Chain      ------------------------------------------------------------ 0
6M71_2|Chain      ------------------------------------------------------------ 0
7BW4_3|Chain      ------------------------------------------------------------ 0
7BV1_3|Chain      ------------------------------------------------------------ 0
7BTF_3|Chains     DAAMQRKLEKMADQAMTQMYKQARSEDKRAK---------------------------- 82
6M71_3|Chains     DAAMQRKLEKMADQAMTQMYKQARSEDKRAK---------------------------- 82
7BW4_2|Chains     DAAMQRKLEKMADQAMTQMYKQARSEDKRAK---------------------------- 82
7BV1_2|Chains     DAAMQRKLEKMADQAMTQMYKQARSEDKRAK---------------------------- 83
```

The subsequence VVVNAANVYLKHGGGVAGALNK (but again commonly VIVNAAN….) resides in nsp3 (nonstructural protein 3), the largest subunit of the so-called "replicase" ORF1while in contrast the RNA -dependent RNA polymerase (which is made up from nsp 7, 8, and 12). A quick means of verification consists of entering "SARS-CoV-2 nsp3" at NCBI site https://www.ncbi.nlm.nih.gov/protein gives many hits at the top of which is the link to https://www.ncbi.nlm.nih.gov/protein/YP_009742610.1, and the FASTA file https://www.ncbi.nlm.nih.gov/protein/YP_009742610.1?report = fasta has a sequences of 1945 residues which contains the subsequence VVVNAANVYLKHGGGVAGALNK. Nsp3 itself contains several conserved domains, including an N-terminal domain enriched in Glu and Asp residues ("acidic domain"), one or two papain-like proteases (PL1pro and PL2pro), a domain originally called the X domain well conserved in the Togaviridae, Coronaviridae, and Hepeviridae, and a C-terminal conserved domain ("Y domain") containing putative transmembrane and metal ion-binding domains. The above subsequence of interest resides in the region originally called the X-domain, but see Section 4.7.

4.7. *Subsequence VVVNAANVYLKHGGGVAGALNK relates to a macro domain*

Q-UEL tags for interacting with standard bioinformatics tools such as BLASTp and Clustal Omega have been discussed in Ref. [25]. A new experimental one reporting drew attention to the macro domain in the course of the study, e.g. as follows.

```
<Q-UEL-BIOINFORMATICS:=( application:='Perl version v5.16.3':=BLASTtestScript.txt, time:='Thu May 27 14:32:49
2020')
tool:=BLASTp:= https://blast.ncbi.nlm.nih.gov/Blast.cgi,
'source header check':='!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"',
html:=http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd,
Query:=' VVVNAANVYLKHGGGVAGALNK'
'input parameters':=(standard, 'job title':='VVVNAAN domain core', database:='Non redundant protein sequences
(nr)', organism:=exclude:= 'viruses (taxid:10239)', 'max target sequences':=100)
| generated |
'commonly recurrent strings':=('macro':=63, 'domain:=63, 'domain-containing':=58, 'TPA':=37, 'ribose':=13,
'ADP':=13, 'ATP':=12, AAA family':=12).
'query' hits:=('macro':=63, 'AMP':=0, 'phosphat':=0,'polymerase':=0)
Q-UEL-BIOINFORMATICS>
```

Q-UEL tags of that kind triggered XTRACT tags related to these findings which were of particular interest in the current analysis, e.g. as follows.

```
<Q-UEL-marple39 "`the Macro domain &or A1pp domain |^is| `a module |In| molecular biology &or, |as| `a module
|with _value of about| 180 amino acids [0https://en.wikipedia.org/wiki/Amino_acid] |^can ^bind| ADP-ribose
[0https://en.wikipedia.org/wiki/ADP-ribose], `an NAD
[0https://en.wikipedia.org/wiki/Nicotinamide_adenine_dinucleotide] metabolite
[0https://en.wikipedia.org/wiki/Metabolite] |^related| ligands [0https://en.wikipedia.org/wiki/Ligands]"

(source:='https://en.wikipedia.org/wiki/macro domain' time:='Tue May 26 10:43:36 2020' extract:=62) Q-UEL-
marple39>
```

The XTRACT tags generated for this Wikipedia entry as well as other sites indicate the functions of the macro domain are various (as described fairly extensively below) but the general feature is considered to be the binding of ADP-ribose.

<Q-UEL-marple39 "(token subject) [0https://en.wikipedia.org/wiki/Binding_(molecular)] |to| ADP-ribose (Binding) [0https://en.wikipedia.org/wiki/Binding_(molecular)] (token object), ADP-ribose |**^can ^be**| `either covalent [0https://en.wikipedia.org/wiki/Covalent] {OR} non-covalent [0https://en.wikipedia.org/wiki/Non-covalent]:[1https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1594587] |in| certain cases (?_it) |^is ^believed to| bind non-covalently [2https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7094737] |while in| `other cases | **such as**| Aprataxin [0https://en.wikipedia.org/wiki/Aprataxin] (?_it) |**^appears to**| bind |**as**| `both |**^non-covalently through**| `a zinc _finger [0https://en.wikipedia.org/wiki/Zinc_finger] motif |**^covalently through**| `a separate region |**of**| `the protein [0https://en.wikipedia.org/wiki/Protein] [3https://doi.org/10.1038%2Fnature06420]" (source:='https://en.wikipedia.org/wiki/macro domain' time:='Tue May 26 10:43:36 2020' extract:=63) Q-UEL-marple39>

As the XTRACTS from this source go on to emphasize, the macro domain is an ancient and highly evolutionarily conserved protein domain, widely distributed throughout all kingdoms of life.

<Q-UEL-marple39 " Macro **|(is) as**| `the C-terminal domain |of| mammalian [0https://en.wikipedia.org/wiki/Mammalia] core histone macro-H2 |**^has been renamed**| Macro |**as it is**| the C-terminal domain of mammalian core histone macro-H2A [0https://en.wikipedia.org/wiki/H2AFY][5https://doi.org/10.1016%2Fs0968-0004%2801%2901787-x][6https://www.sciencedirect.com/science/article/abs/pii/S002228360300473X?via%3Dihub]; Macro domain proteins [0https://en.wikipedia.org/wiki/Protein] |**^can ^be ^found in**| eukaryotes [0https://en.wikipedia.org/wiki/Eukaryote] |**in as ^mostly by**| pathogenic bacteria [0https://en.wikipedia.org/wiki/Pathogenic_bacteria] |**in**| archaea [0https://en.wikipedia.org/wiki/Archaea] |**in**| ssRNA ^viruses [0https://en.wikipedia.org/wiki/SsRNA] (token object)" (source:='https://en.wikipedia.org/wiki/macro domain' time:='Tue May 26 10:43:36 2020' extract:=66) Q-UEL-marple39>

That includes, as BLASTp indicated above, the coronaviruses.

<Q-UEL-marple39 "(token subject) [0https://en.wikipedia.org/wiki/SsRNA] | `**such as|** coronaviruses| [0https://en.wikipedia.org/wiki/Coronavirus] &and Rubella [0https://en.wikipedia.org/wiki/Rubella_virus] **&and** Hepatitis E viruses [0https://en.wikipedia.org/wiki/Hepatitis_E_virus] (token object)" (source:='https://en.wikipedia.org/wiki/macro domain' time:='Tue May 26 10:43:36 2020' extract:=67) Q-UEL-marple39>

4.8. Subsequence VVVNAANVYLKHGGGVAGALNK in the context of the macro domain

The subsequence VVVNAANVYLKHGGGVAGALNK (or again commonly VIVNAAN…) as the most prominent match is not the same as the macro domain, but it is part of it. This is the match that BLASTp tends to find, because the region outside that subsequence is much more variable across many organisms. The extent of the macro domain varies somewhat with different authors and is often seen as up to 190 amino acid residues in length, but is typically considered as around 165 amino acid residues. It is more specifically the motif VVVNAANVYLKHGGG-VAGALNK of 22 amino acid residues that is of primary interest, but almost as importantly, its extension as the region VVVNAAN… LRVCVDT comprises 133 amino acid residues. This subsequence is called the "VVVNAAN domain core" in the present project. It is the region involved in many prominent BLASTp matches in the present project, of which there are examples given later below. The following shows the sequence (indicated below as SCV2) in SARS-CoV-2 that would be considered as the macro domain, and the VVVNAAAN… …LRVCVDT sequence is underlined and in bold. The experimental secondary structure of the domain for SARS-CoV2 is shown for Protein Data Bank entry 6W6y. Here this is not intended to be a study of secondary structure variation in the macro domain, nor of the accuracy of the secondary structure prediction (also shown), but one might (arguably) expect that such an important ancient and widely spread domain has strong tendency for secondary structure prediction by the HDN [24] based on local effects. The essential features predicted are reproducible by GOR4 publically available at https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_gor4.html, also developed by the present author and colleagues, and also access able via the Q-UEL system. The main difference is that the HDN predicts in two directions of conditionality, but only prediction of secondary structure given sequence is important in the present case.

However, the particular way that the HDN version is implemented means that it dramatically improves as it starts to learn the relationships between amino acid sequence and secondary structure in folding motifs of domains that it has encountered before [24]. In that implementation, proteins are successively added to the training set but have their structure predicted (and compared with the observed structure) immediately prior to inclusion in that training set. Inspection of descriptions and/or structure of those proteins responsible for peaks of performance [18] allows one to identify the domain fold, which was seen to be a macro domain in the present case. It remains that the prediction itself largely reflect the innate conformational tendencies of amino acid residues and local interactions in the "secondary structure" of the sequence, not the impact of interactions in the three dimensional "tertiary structure". In the following prediction of secondary structure for the SARS-CoV-2 macro domain of current interest, there are seen to be significant discrepancies between the predicted secondary structure PRED and the observed secondary structure of SARS-CoV-2 macro domain in the

Protein Data Bank entry 6W6Y. Nonetheless, in the VVVNAAN domain core region itself, and particularly for the well conserved segment VVVNAANVYLKHGGGVAGALNK the prediction is for the most part reasonable. This suggests that the local tendencies are significantly retained in the final observed conformation. That is, the tertiary interactions do not so much compete with and override local effects, so that the VVVNAAN core is likely to have significant conformational stability.

```
              1         2         3         4         5         6         7
SCV2 IEVNSFSGYLKLTDNVYIKNADIVEEAKKVKPTVVVNAANVYLKHGGGVAGALNKATNNAMQVESDDYIA
6W6Y ccccccccEEccccEEEEEcHHHHHHHHHHccEEEEEccccccccHHHHHHHHHHHcHHHHHHHHHHH
PRED ccccccccEEEcccEEEccHHHHHHHHcccccEEEEEccEEEEEccccHHHHHHHHHccHHHHcccEEE

              8         9         0         1         2         3         4
SCV2 TNGPLKVGGSCVLSGHNLAKHCLHVVGPNVNKGEDIQLLKSAYENFNQHEVLLAPLLSAGIFGADPIHSL
6W6Y HHHccccccEEEEEccccccEEEEEEccHHHHHcHHHHHHHHHHHHHHHHcEEEEccccHHHHHcHHHHH
PRED ccccEEEccEEEEEcccccccccEEEEcccccccccHHHHHHHHHHHHHcHHHHHHHHHHcccccccccEE

              5         6
SCV2 RVCVDTVRTNVYLAVFDKNLYDKLV
6W6Y HHHHHHHccEEEEEEcHHHHHHHHH
PRED EEcEEccccccEEEEEccccccEEc
```

The inner region of this VVVNAAN domain core also tends to be more variable than the ends when comparing sequences, and a finer analysis is given shortly below. It remains that, in talking about properties and functional roles rather than matches and homologies, what is said for the VVVNAAN domain core is valid for the macro domain, and *vice versa*. For comparison with the matches in the present paper and in relation to a known three dimensional protein structure, an example of variation of the above subsequence in SARS coronaviruses is as follows.

```
Chain A, Replicase Polyprotein 1ab (pplab) (orflab) [Severe acute respiratory
syndrome-related coronavirus]
Sequence ID: 2FAV_A Length: 180 Number of Matches: 1
Range 1: 41 to 154 Identities:91/114(80%) Positives: 101/114(88%)  Gaps: 0/114(0%)
Query  1    VVVNAANVYLKHGGGVAGALNKATNNAMQVESDDYIATNGPLKVGGSCVLSGHNLAKHCL  60
            V+VNAAN++LKHGGGVAGALNKATN A Q ESDDYI  NGPL VGGSC+LSGHNLAK CL
Sbjct  41   VIVNAANIHLKHGGGVAGALNKATNGAXQKESDDYIKLNGPLTVGGSCLLSGHNLAKKCL  100

Query  61   HVVGPNVNKGEDIQLLKSAYENFNQHEVLLAPLLSAGIFGADPIHSLRVCVDTV  114
            HVVGPN+N GEDIQLLK+AYENFN  ++LLAPLLSAGIFGA P+ SL+VCV TV
Sbjct  101  HVVGPNLNAGEDIQLLKAAYENFNSQDILLAPLLSAGIFGAKPLQSLQVCVQTV  154
```

2FAV is the Protein Data Bank entry for crystal structure of SARS macro domain discussed further in the following Section 4.10. In between the two subsections underlined in the above alignment, the sequences vary significantly more, and in roughly the same manner as the regions of the macro domain that extend outside the VVVNAAN domain core. In broader studies involving many such comparisons between coronaviruses, however (as well as with a much broader range of organism's as discussed later below), subsequence LHVVGPNVNKG is seen as moderately well conserved. Three sequences are therefore are notable for their persistence and called by the present author "VVVNAAN domain core subsequences (a), (b) and (c)" as follows.

(a) VVVNAANVYLKHGGGVAGALNK,
(b) LHVVGPNVNKG
(c) PLLSAGIFG

It is the first subsequence (a) that shows reasonable agreement with secondary structure prediction, and is thus a more strongly locally determined structure, and possibly a strong early forming nucleating structure during folding. A scan against known sequence motifs in Prosite https://prosite.expasy.org/scanprosite/recognized (a) as a signature of the macro domain, but initially not subsequence (b) or (c). Consistent with observations above, it is a motif common in thermophilic bacteria, for example IVNAANAYLRHGGGVAGA in *Thermotogae bacterium*. A comment on variation is inevitably biased by availability of sequence information, but on the whole it seems reasonable to say that the second half of subsequence (a), notably HGGGVAGA, is more conserved than the first. Initially this did not produce any significant hits on Prosite even with the sensitivity increased to allow greater flexibility in the match. Use of Prosite is discussed later below in regard to macro domain function (Section 4.0). Conservation of the second half with variation in the first half seems particularly noticeable in extremophiles, for example there is a match with IVNPANAYLRHGGGVAGAL in GenBank entry WP_169700310.1, describing a previously unknown mesophilic, anaerobic, rod-coccoid-shaped bacterium, having a sheath-like outer structure and isolated from a water sample collected in the area of an underground gas storage aquifer.

Table 3
Summary of VVVNAAN domain core BLASTp matches on the vertebrates only.

Species	Group	Score
Apaloderma vittatum	birds	70.5
Terrapene carolina triunguis	turtles	68.9
Chrysemys picta bellii	turtles	68.9
Platysternon megacephalum	turtles	68.9
Protobothrops mucrosquamatus	snakes	68.6
Thamnophis sirtalis	snakes	68.2
Thamnophis sirtalis	snakes	68.2
Tetraodon nigroviridis	bony fishes	68.2
Sphaeramia orbicularis	bony fishes	68.2
Gouania willdenowi	bony fishes	68.2
Python bivittatus	snakes	67.8
Rhinatrema bivittatum	caecilians	67.8
Chelonia mydas	turtles	67.4
Egretta garzetta	birds	67.4
Notechis scutatus	snakes	67.0
Ophiophagus hannah	snakes	67.0
Notechis scutatus	snakes	67.0
Oryzias melastigma	bony fishes	67.0
Chelonoidis abingdonii	turtles	66.6
Thamnophis elegans	snakes	66.2
Chanos chanos	bony fishes	66.2
Catharus ustulatus	birds	66.2
Erythrura gouldiae	birds	66.2
Aotus nancymaae	primates	66.2
Saguinus labiatus	primates	65.9
Anolis carolinensis	lizards	65.9
Carassius auratus	bony fishes	65.9
Lates calcarifer	bony fishes	65.9
Gambusia affinis	bony fishes	65.5
Acipenser ruthenus	bony fishes	65.1
Oreochromis niloticus	bony fishes	65.1
Latimeria chalumnae	coelacanths	65.1
Cavia porcellus	rodents	65.1
Lithobates catesbeianus	frogs & toads	64.7
Labrus bergylta	bony fishes	64.3
Myripristis murdjan	bony fishes	64.3
Takifugu rubripes	bony fishes	64.3
Kryptolebias marmoratus	bony fishes	63.9
Esox lucius	bony fishes	63.9
Electrophorus electricus	bony fishes	63.9
Parambassis ranga	bony fishes	63.9
Kryptolebias marmoratus	bony fishes	63.9
Amazona aestiva	birds	63.9
Xiphophorus couchianus	bony fishes	63.5
Austrofundulus limnaeus	bony fishes	63.5
Callipepla squamata	birds	63.5
Nestor notabilis	birds	63.5
Serinus canaria	birds	63.5
Pseudonaja textilis	snakes	63.5
Paroedura picta	lizards	63.5
Podarcis muralis	lizards	63.5
Pogona vitticeps	lizards	63.2
Gavia stellate	birds	63.2
Myotis davidii	bats	63.2
Strigops habroptila	birds	62.8
Sorex araneus	insectivores	62.8
Cebus capucinus imitator	primates	62.8

4.9. Subsequence VVVNAANVYLKHGGGVAGALNK relates to an ADP-ribose-1″-phosphatase

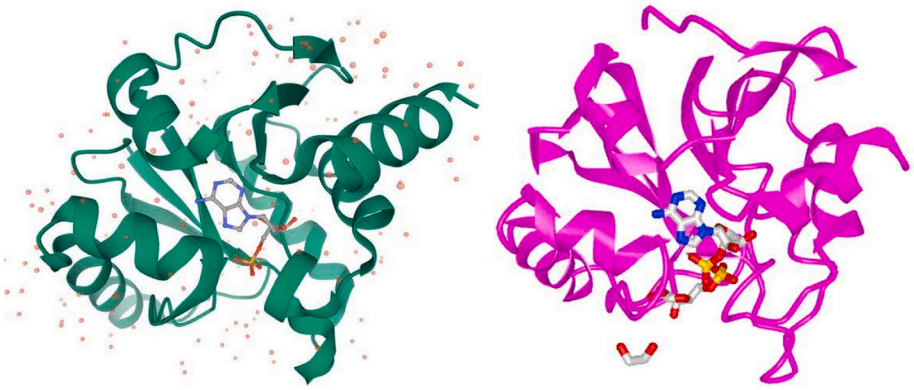
The experimental three dimensional structure 2FAV used in the alignment above is a crystal structure of the SARS macro domain in complex with ADP-ribose, and the depositors considered this as likely to be the structural basis for ADP-ribose and polyADP-ribose binding by viral macro domains in general. Originally on the basis of homology, researchers found that an important function was, at least in some cases, conversion of ADP-ribose-1″-monophosphate to ADP-ribose, and this relates to the original identification of the X domain (the "conceptual ancestor" of the macro domain as discussed above). It is known that in SARS-CoV ADP-ribose-1″-phosphatase is responsible for ADP-ribose-1″-phosphate dephosphorylation involving a conserved domain of nsP3 [38], in contrast to the RNA-dependent RNA polymerase (which recall is made up from nsP 7, 8), and 12, and that the SARS-unique domain (SUD) of SARS coronavirus contains two macro domains that bind G-quadruplexes, i.e. unusual nucleic-acid structures formed by consecutive guanosine nucleotides, where four strands of nucleic acid are forming a superhelix [39]. It is now appreciated that macro domain functions can include a variety of ribose-phosphate-related binding and catalytic activities, including putative sequence-specific endoribonuclease, 3′-to-5′ exoribonuclease, 2′-O-ribose methyltransferase, as well as ADP ribose 1″-phosphatase (or phosphohydrolase) and in some coronaviruses, cyclic phosphodiesterase activities [40]. Many of these can also exhibit a variety of more general nucleic acid and nucleotide functions, but while adenosine phosphate binding is the dominant theme, the domain can bind a variety of unrelated ligands [41] which emphasizes its potential interest as a therapeutic target.

The ability to use the VVVNAANVYLKHGGGVAGALNK motif as a predictor of the macro domain or specifically for ADP-ribose-1″-phosphatase has not yet been quantified in detail, and any result would be somewhat arbitrary because it depends, of course, on the sequences available for examination. However, it is clear that not all SARS-CoV-2 proteins associated with replicase activity or called "polymerase proteins", that must inevitably have nucleotide binding functions, relate to the VVVNAAN domain core *per se*. Consequently, if used as a predictor of such activity the prediction would have many true positives but many false negatives. That is, it is a sensitive but not specific test. The SARS-CoV-2 subsequence of interest here is that of Nsp3, and so should also be distinguished from the Nsp12 polymerase for which the structure, bound to Nsp7 and Nsp8 co-factors, has been determined [17]. Protein Data Bank entry 6W4B entered in March 2020, Nsp9 RNA binding protein is described as a replicase protein and believed to mediate viral replication and virulence (it is interesting as having an unexpected a peptide-binding site that needs to be understood to understand Nsp9 function). These proteins are best not considered as "sister" entities to Nsp3. For example, the Nsp9 sequence is aligned by Clustal Omega with the above VVVNAAN… sequence as follows, and despite a few tentative hints at common sequence features, there is only 20% identity. That would usually be considered within the range that is not likely to be significant, i.e. consistent with random match [5].

```
VVVNAAN...      ---VVVNAANVYLKHGGGVAGALNKATNNAMQVESDDYIATNGPLKVGGSCVLSG-----    52
Nsp9            SNAMNNELSPVALRQMSCAAGTTQTAC------TDDNALAYY-NTTKGGRFVLALLSDLQ    53
                   :  : : * *:: . .**: :.*        .*: :*     . **  **:

VVVNAAN...      -----------------HNLAKHCLHVVGPNVNKGEDIQLL--KSAYENFNQHEVLLAPL    93
Nsp9            DLKWARFPKSDGTGTIYTELEPPCRFVTD--TPKGPKVKYLYFIKGLNNLNRGMVL-GSL    110
                              :*   * .*..  . ** .:: *   .. :*:*:  ** . *

VVVNAAN...      LSAGIFGADPIHSLRVCVDTV   114
Nsp9            A----------ATVRLQ----   117
                         ::*:
```



6W6Y
Crystal Structure of ADP ribose phosphatase of NSP3 from SARS CoV-2 in complex with AMP

6W02
Crystal Structure of ADP ribose phosphatase of NSP3 from SARS CoV-2 in the complex with ADP ribose

Fig. 2. Experimental structures of SARS-CoV-2 ADP-ribose-phosphatase.

```
WLVNASNVDHRPGGGLCHAFYQ - Hepatitis E virus X domain
VVVNAANVYLKHGGGVAGALNK - SARS-CoV-2 VVVNAAN domain core segment(a)

LHVVGPNVNKG - Hepatitis E virus X domain
LHAVAPDYRLE - SARS-CoV-2 VVVNAAN domain core segment(b)

PLLSAGIFG - Hepatitis E virus X domain
PLLGTGIYQ - SARS-CoV-2 VVVNAAN domain core segment(c)
```

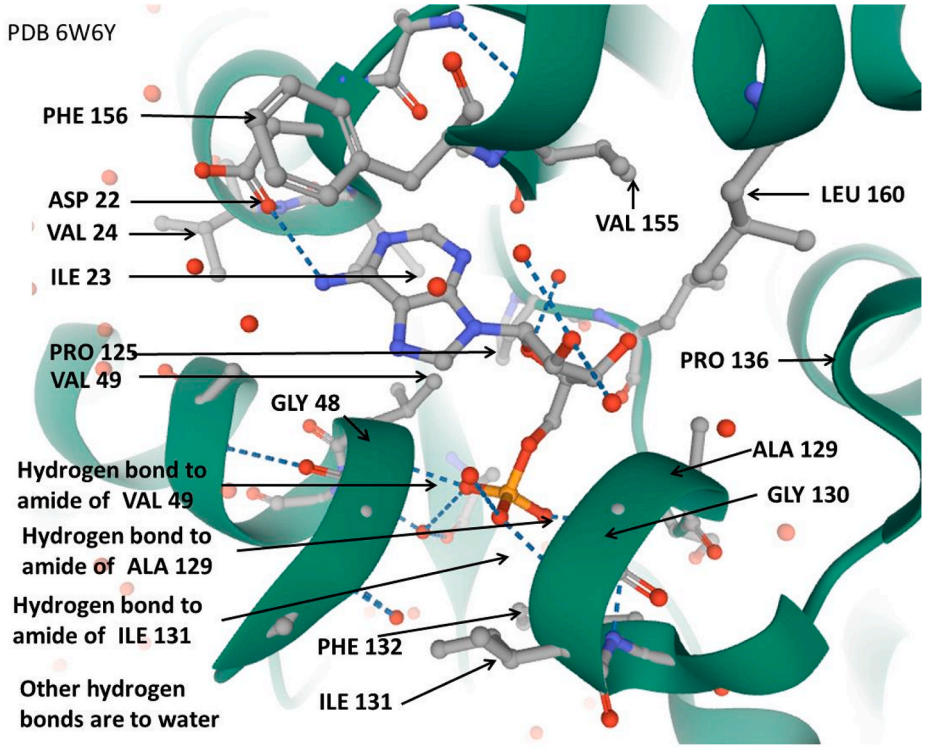


Fig. 3. SARS-CoV-2 VVVNAAN domain core binding site 1.
Adenosine Monophosphate Interactions with ADP ribose phosphatase of SARS-CoV-2 Nsp3 in Complex with Adenosine Monophosphate in PDB Entry 6W6Y.

A general feature of matches with VVVNAANVYLKHGGGVAGALNK is an involvement with molecules containing purine and/or phosphate moieties. Other purine binding motifs have been found in a superfamily across many organisms, but they seem distinct from the locus of the above motif. Of several recurrent themes found, the patterns GxGKS/T or G/AxxxxGKS/T (where x is any amino acid) associated with phosphate binding are prominent; they are well-known and considered definitive of two classes of helicase-like domain. However, in the present study no obvious homologues of these patterns are found within the section of sequence VVVNAAN…VCVDTV matching sequences AAT76146.1 and 2FAV_A discussed above, nor in the examples of matches discussed elsewhere in this paper. There was occasionally found some indication of weaker homology with the above and particularly synthetic hexapeptide SGAGKT shown to bind inorganic phosphate strongly [41], but these not consistently conserved in all matches and may be an impression gained from the examples accessed. A domain considered as an X domain from a conformational perspective is found in Hepatitis E virus where the subsequence spans PDGSKVFAGSLF… VPIGSFDAWER, but it has no obvious relation to the segment VVVNAAN *per se*. However, there is a possible weak homology between the X domain and the VVVNAAN domain core on alignment especially if emphasis is on the three well conserved segments discussed above.

4.10. VVVNAAN domain core in multicellular organisms

Examining the occurrence of the macro domain in higher organisms may give clues as to origins, functional importance, and the risks to a host of thereeputic treatment intended to inhibit the virus. As seen above in Table 1, there are homologies of coronavirus proteins with proteins of cellular microorganisms, notably the archaea. These may reflect an ancient association, and the interest in more complex organisms is more in relation to their involvement as relatively modern coronavirus hosts and the possibility of more recent gene transfer. Microorganisms are of course not necessarily excluded from that host role, although viruses of the archaea so far all have double-stranded DNA genomes. References to macro domains tend to highlight vertebrates:-

<Q-UEL-marple39 "`the domain |^occurs in| eg |In| vertebrates [0https://en.wikipedia.org/wiki/Vertebrates]" (source:='https://en.wikipedia.org/wiki/macro domain' time:='Tue May 26 10:43:36 2020' extract:=68) Q-UEL-marple39>

However, they are of course found more generally, and it cannot even be said that the sequence matches of interest in the present study are always significantly less obvious in the invertebrate case, as discussed shortly below. Recall three subsequences from Section 4.7 that were of particular interest as well conserved: (a) VVVNAAN-VYLKHGGGVAGALNK, (b) LHVVGPNVNKG, and (c) PLLSAGIFG. Similar sequences are found across the animal kingdom with match to (a), (b) and (c) entered together and in the above order, or simply with the whole subsequence matching replicase entry AAT76146.1 and 2FAV_A in Section 4.2 (i.e. VVVNAAN… VCVDTV) are found to be widespread. This might be expected from the dominant match found by subsequence (a), primarily directly or indirectly concerned with nucleic acid replication, recombination and repair (but also frequently

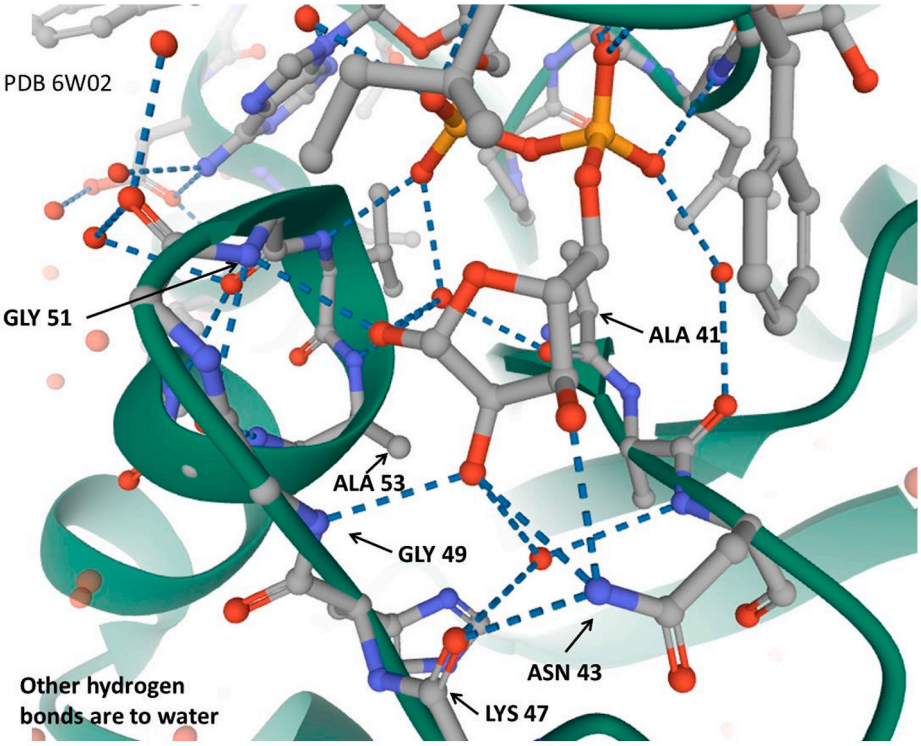


Fig. 4. SARS-CoV-2 VVVNAAN domain core binding site 2.
Ribose Phosphate Interactions in ADP ribose phosphatase of SARS-CoV-2 Nsp3 in Complex with ADP ribose PDB Entry 6W02.

involving small nucleotides). BLASTp searches on each one separately were particularly insightful.

Subsequence (a) entered alone into BLASTp was found in many animal polymerases and mono-ADP-ribosyltransferases. It is this subsequence that seems most definitive of the function of this domain, because matches of this kind are not significant when (b) and (c) are entered as queries into BLASTp separately or together. For the above reasons, the overall sequence VVVNAAN… VCVDTV is conveniently called the *VVVNAAN domain core motif*.

Subsequence (b) is found with 78% identity in mammalian vomeronasal type-1 receptors (e.g. in the rat, GenBank XP_032765894.1). This kind of matching protein has already been noted in Ref. [5] in relation to the SARS-CoV-2 spike glycoprotein where FNCTWP is suspiciously a subsequence in the mammalian vomeronasal type-2 receptor 1 on sensory cells within the main nasal chamber that detects heavy moisture-borne odor particles. There were also comparable levels of match with diverse proteins such as thioredoxin reductases, sugar transporters, acyl-coenzyme A thioesterases and hydroxysteroid dehydrogenases, toll-like receptors, and flottillins (possibly involved in vesicular trafficking and signal transduction).

Subsequence (c) is, in contrast, found with a number of matches different to those of both the above, but surprisingly prominent is a variety of solute carrying proteins. This family is diverse and transports both charged and uncharged organic molecules as well as inorganic ions and ammonia gas. If there are any hints to be found of a meaningful match in the attempted Nsp9 alignment earlier above, they end around the beginning of subsequence (c).

One difference in functions of the macro domain in *invertebrates* relates to the fact that NK cells, antibodies, and cytotoxic T cells appear to be considered as lacking in such organisms. This is discussed later below, but the appearance of the VVVNAAN domain core in descendants of primitive invertebrates such as coelenterates certainly seems unlikely to have any connection to the interferon system discussed below and as currently understood. However, an important component of a pathway activated by interferon in mammals, the enzyme 2′,5′-oligo A synthetase, has been reported in sponges. *Acropora millepora* is a species of branching stony coral native to the western Indo-Pacific where it is found in shallow water from the east coast of Africa to the coasts of Japan and Australia. The protein in the GenBank data base matched is stated as involving an "uncharacterized protein" of 138 residues, but the entry also describes it as a "macro domain, a high-affinity ADP-ribose binding module found in a variety of proteins as a stand-alone domain or in combination with other domains like in histone macroH2A …".

interferon mechanisms are very old, if not the interferon system itself. There is some debate as to when sponges, animals belonging to the phylum Porifera, first emerged. Some authors consider that emergence was not until the Cambrian period, between 541 million and 485 million years ago, but some workers put it as early as Precambrian times, some 760 million years ago. A taxonomic tree for all the species examined would be very rich and complex, but for present purposes Table 2 (and Table 3 later below) suffices to give some indication based on BLASTp score. Indeed both give a somewhat confused impression and it seems that as measured by sequence similarity there is rather little trend in terms of taxonomic distance between animal species. Recall that these are subsequence of proteins in the organisms stated, not necessarily coronavirus hosts. Nonetheless, relationships may be more to do with gene jumping between host and virus. The main point here is the widespread appearance, and hence apparent great antiquity, of the gene.

As discussed below, any differences between vertebrates and invertebrates might be due to differences in immunity and notably also innate cellular immunity. It thus seems less surprising that, in the present study, there are as yet no obvious plant protein matches to the above VVVNAAN domain core (the section VVVNAAN… VCVDTV used in matches above). The antiviral defenses are of course different in plants. However, there are analogous systems and there are remnants as possible matches between other parts of the SARS-CoV-2 polyprotein with helicases in leguminous plants including *Arachis ipaensis* is a herb in the Faboideae family, Lupinus albus, the white lupin or field lupine, Vigna angularis, the adzuki bean, Cajanus cajan, the pigeon pea, in genomes the emphases on which presumably reflects interest in plant food products. These are matches of the order of 28% identities in *circa* 275 residues and subsequences GDPAQLPA and marginally ITRAK may be worth investigating, but are beyond present scope. However, these matches might have something to do with nitrogen fixing bacteria hosted by these leguminous plants.

The persistence or otherwise of the domain core in humans and other vertebrates is an important consideration for the development of antiviral therapeutic agents for human and veterinary medicine, as well as in regard to both wild and domesticated animals as coronavirus hosts. Here are some odd fluctuations to an expected evolutionary trend, which may be indicative of gene transfer but which seem to defy any unifying theme or hypothesis even on a gene transfer basis. One may compare the SARS-CoV-2 sponge match with some significantly weaker vertebrate matches, e.g. genbank entry xp_030436584.1, identities:45/125(36%) positives:64/125(51%) gaps: 16/125(12%), in the same region of a gene of *Gopherus evgoodei*, the Sinaloan Thornscrub Tortoise, a relatively

```
LOC114959026 [Acropora millepora]
Sequence ID: XP_029192739.1 Length: 238Number of Matches: 1
Range 1: 67 to 174
Identities; 48/108(44%) Positives 63/108(58%) Gaps:10/108

Query  2    VVNAANVYLKHGGGVAGALNKATNNAMQVESDDYIATNGPLKVGGSCVLSGHNLA-KHCL  60
            +VNAAN +L+HGGGVAGA+ K   N +Q +SD+++A +GP+K GG  V    NL     +
Sbjct  67   IVNAANSWLRHGGGVAGAIVKKGGNQIQADSDEFVAKHGPVKTGGIAVTEAGNLPCSIII  126

Query  61   HVVGPNV---NKGEDIQLLKSAYNF------NQHEVLLAPLLSAGIFG  99
            H VGP      KGED  L  + YN        Q   L AP +S+GIFG
Sbjct  127  HAVGPVWEGGQKGEDKCLRDAMYNSLVECRKRQLVSLAAPAISSGIFG  174
```

The above sponge match has 44% identities, 58% positives with the SARS-CoV-2 motif. That the above match is to a poly ADP-ribose polymerases is significant because this has been long suggested as a major function of the domain in vertebrates.

In the absence of any direct indication of relatively recent gene transfer by virus, the gene associated with the VVVNAAN core motif would seem to very old, consistent with discussions of the macro domain. At least the *foundations* of the interferons and viral anti-

confined species, being found in the relatively arid part of the Tropical Deciduous Forest of western Mexico and relatively desert-like Sonoran Foothills. The entry describes it as a mono-ADP-ribosyltransferase. It might hint at the SARS-CoV-2 sponge case being a gene transfer, except that there are many stronger matches more comparable with the sponge match or stronger that are found in many other reptiles. For example *Protobothrops mucrosquamatus*, a mono-ADP-ribosyl transferase of a venomous pit viper species endemic to Asia, Genbank XP_029140551.1 identities: 46/110(42%) positives: 58/110(52%) gaps: 11/110(10%). As

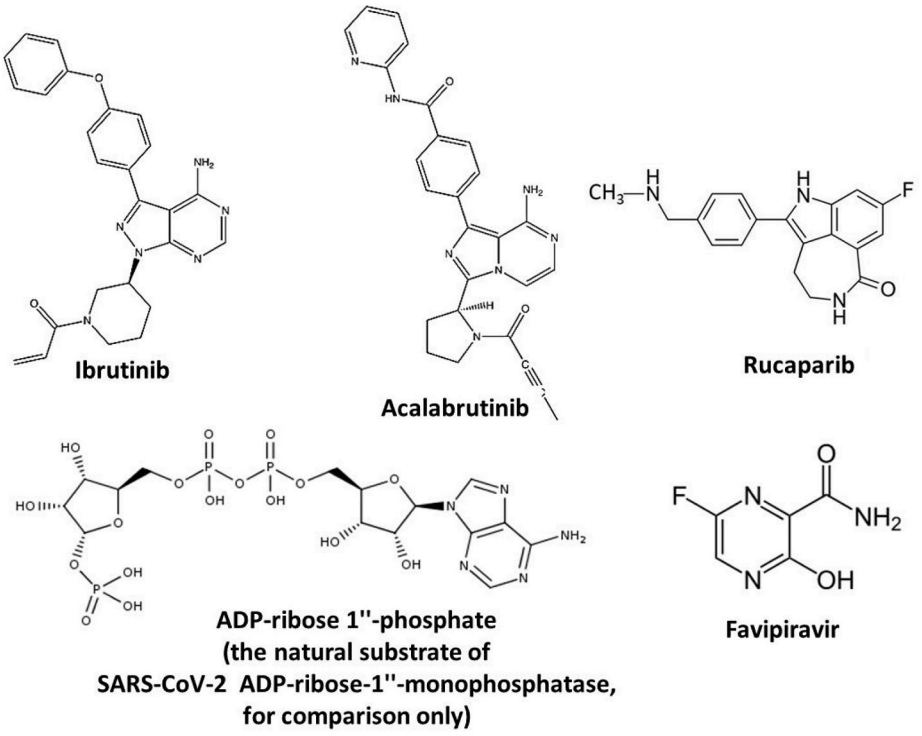


Fig. 5. Medium Binders to ADP ribose phosphatase of SARS-CoV-2 Nsp3 (estimated binding free Energy) −9 to −11 Kcal/Mole).

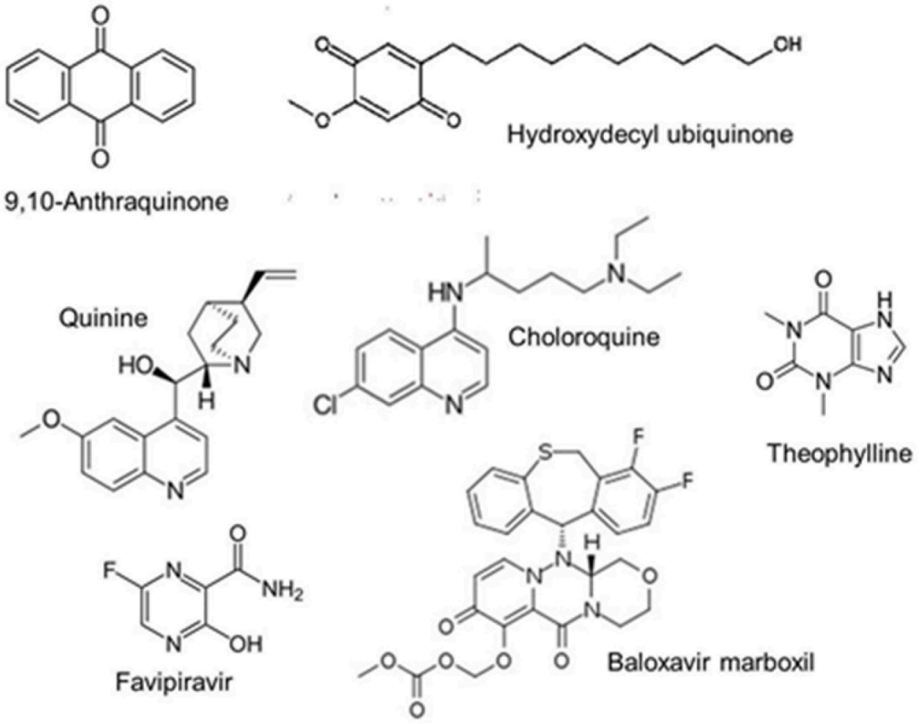


Fig. 6. Weak-to-medium Binders ADP ribose phosphatase of SARS-CoV-2 Nsp3 (estimated binding free Energy) −6 to −8 Kcal/Mole).

may be expected, the ubiquitous nature of this domain includes the birds. For example, *Apaloderma vittatum* the Bar Tailed Trogon, has a poly ADP-polymerase GenBank entry XP_009869054.1, identitie49/125 (39%), positives 68/125(54%) gaps: 17/125(13%). Matches with polymerases and mono-ADP-ribosyltransferases certainly also include the fishes. *Gouania willdenowi* is the blunt-snouted clingfish, a species of clingfish found along the Mediterranean Sea coasts from Syria to Spain, GenBank entry XP_028330112.1, identities 47/122(39%), positives 69/122(56%), gaps14/122(11%).

Although it is just one domain in an animal protein that is usually matched by BLASTp, putting the above human matching sequence into BLASTp generates two or more domains coexisting in many proteins concerned with nucleic acids in many organisms. An example with three is *Chrysemys picta bellii* is the painted turtle, GenBank entry

XP_023960593.1 which has two in a poly ADP-ribose polymerase. It is the most widespread native turtle of North America. It lives in slower moving fresh waters from southern Canada to northern Mexico, and from the Atlantic to the Pacific. A further reason for mentioning this match is that the first of the three is at an extremely interesting 96% match level with the SARS-CoV-2 subsequence of interest. There were no XTRACT tags obtained indicating that give an account of humans eating this specific species of turtle, although there were indications (e.g. at https://www.nwf.org/Magazines/National-Wildlife/2006/Asias-Turtle-Tragedy) that a variety of turtles are eaten are ancient symbols of longevity.

<Q-UEL-marple39 " `The domain |**^was ^described ^originally in association with**| `the ADP-ribose1-phosphate |**as**| Appr-1-P (-processing) activity |**as**| A1pp |**of**| `the yeast [0https://en.wikipedia.org/wiki/Saccharomyces_cerevisiae] YBR022W protein |^called| A1pp [4https://doi.org/10.1126%2Fscience.286.5442.1153] ; `the domain |**^has ^been ^renamed**| Macro" (source:='https://en.wikipedia.org/wiki/macro domain' time:='Tue May 26 10:43:36 2020' extract:=64) Q-UEL-marple39>

Pinning down the importance to SARS-CoV and SARS-CoV-2 of the functional and biological role of the macro domain has not been very easy because of a variety of ribose-phosphate-related activities detected within the macro domains themselves [40]. These include putative sequence-specific endoribonuclease, 3′-to-5′ exoribonuclease, 2′-O-ribose methyltransferase, ADP ribose 1″-phosphatase and in some coronaviruses, cyclic phosphodiesterase activities [41]. Ribose is a common theme, along with phosphate binding features [42], but macro domains can not only recognize ADP-ribose both in its free and protein linked

poly [ADP-ribose] polymerase 14-like [Chrysemys picta bellii]
Sequence ID: XP_023960593.1 **Length: 1823Number of Matches: 3**
Range 1: 833 to 95
Identitities: 115/120(96%)
Positives: 116/120(96%)
Gaps: 0/120(0%)

```
Query  1    VVVNASNEDLKHIGGLAEALLKAAGPELQTECDHIVRKRGPLQPGHAVITDAGNLPCKQV  60
            VVVNASNEDLKHIGGLAEALLKAAGPELQTECDHIVRKRGPLQPG AVITDAGNLPCKQV
Sbjct  833  VVVNASNEDLKHIGGLAEALLKAAGPELQTECDHIVRKRGPLQPGRAVITDAGNLPCKQV  892

Query  61   IHAVGPRWRDHEPGKCVHLLKRAIKESLHLAETFNHHSIAIPAISSGIFGFPLKLCAQSI  120
            IHAVGPRWRDHEP KCV LLKRAIKESL LAET+NHHSIAIPAISSGIFGFPLKLCAQSI
Sbjct  893  IHAVGPRWRDHEPEKCVRLLKRAIKESLQLAETYNHHSIAIPAISSGIFGFPLKLCAQSI  952


Range 2: 1049 to 1165
Identities: 44/119(37%)
Positives: 66/119(55%)
Gaps: 3/119(2%)
Query  1     VVVNASNEDLK-HIGGLAEALLKAAGPELQTECDHIVRKRGPLQPGHAVITDAGNLPCKQ  59
              V+V++  +DL+  +G L+++LL+ AGP LQ E +   +++ P Q G    T   NL C
Sbjct  1049  VIVSSVGQDLRLGVGPLSQSLLQKAGPTLQLEFNEESQRQVPTQ-GSVFHTSGCNLACSF  1107

Query  60    VIHAVGPRWRDHEPGKCVHLLKRAIKESLHLAETFNHHSIAIPAISSGIFGFPLKLCAQ  118
              + HAV P W D   G  +  L+  +KE L   E  +  SI  PAI +G FGFP  + A+
Sbjct  1108  LFHAVVPVW-DQGRGGAMTNLEDIVKECLKKTEELSLRSITFPAIGTGGFGFPKPIVAK  1165


Range 3: 1257 to 1351
Identities: 34/111(31%)
Positives: 55/111(49%)
Gaps: 17/111(15%)
Query  1     VVVNASNEDLKHIGGLAEALLKAAGPELQTECDHIVRKRGPLQPGHAVI-TDAGNLPCKQ  59
              V+VN SN       G+ +A+++AAGP+++ EC+ +       LQP    I T  G L C +
Sbjct  1257  VIVNISNSSFNAKSGVFKAVMEAAGPQVKLECNMLA-----LQPHSGFITTQGGKLMCNK  1311

Query  60    VIHAVGPRWRDHEPGKCVHLLKRAIKESLHLAETFNHHSIAIPAISSGIFG  110
              +IH +      H+       +K  + + L   E   + S+A PAI +G  G
Sbjct  1312  IIHLI------HQKD-----VKAQVSKVLQECELRKYTSVAFPAIGTGQAG  1351
```

Again as measured by sequence similarity there is rather little trend in terms of taxonomic distance and relationships may be more to do with gene jumping between host and virus.

4.11. Other macro domain functions of potential importance to SARS-CoV-2

By "function" is here meant something closer to binding specificity and any associated catalytic mechanism; involvement in the biology of infection requires consideration of several functional activities and is discussed later below. Originally, several research groups considered activity related to ADP-ribose-1′ monophosphate to be, in effect, a form, and related ligands such as O-acyl-ADP-ribose, but also ligands unrelated to ADP ribose. As far as nucleoside phosphates are concerned, the binding activity does not appear to extend beyond adenine recognition [43]. Nonetheless, many plus-strand RNA viruses can also bind poly(A) or poly(G) by using a similar enzyme. Overall, it is ADP-ribose 1″-phosphatase (ADRP) activity that has in particular been indicated, normally described as highly specific for ADP-ribose. ADP-ribose 1″-phosphate differs from ADP-ribose-phosphate by the addition of the further phosphate to the ribose sugar. It appears that many phosphatases can attack both forms albeit at different rates. ADP ribose and ADP-ribose 1″-phosphate can interconvert. ADP-ribose 1″-phosphate phosphatase catalyzes the following reaction. It is an equilibrium

reaction that lies to the right and the amount of phosphate released is proportional to the amount of the substrate added. In ADP ribose the ring is predominantly in open form favored in alkaline conditions.

ADP-ribose 1″-phosphate + H_2O ⇌ ADP-ribose + phosphate

In bacterially cloned forms of human coronavirus 229E (HCoV-229E), SARS coronavirus X domains were shown to dephosphorylate the substrate of the side product of cellular tRNA splicing, thus converting it to ADP-ribose, in a highly specific manner. XTRACT tags highlighted that in cellular organisms, ADP-ribose 1″-phosphate phosphatases are normally seen as working in concert with 2′,3′-cyclic-nucleotide 3′-phosphodiesterase in the breakdown of adenosine diphosphate ribose 1″,2″-cyclic phosphate, a by-product of tRNA splicing.

BLASTp continued to be an important tool for exploring function. Several BLASTp matches to the VVVNAAN domain core pattern (a) relate directly or indirectly to proteins of that system. For example, an interesting one is to the interferon α/β receptor, particularly noted in the type 2 interferon α/b receptor of certain reptiles. *Platysternon megacephalum* is the big-headed turtle is a species of turtle from Southeast Asia and southern China.

domain", in the motif pattern match data bases PROSITE PATTERN, PROSITE PROFILE, NCBI-CDD and Pfam that can be accessed at https://www.genome.jp/tools/motif/. The corresponding interferon receptor alignment, does seem to be a better known pattern, motif NCBI-CDD ID: 239235. Here it is described as cd02907, Macro_Af1521_BAL_like, Macro domain, Af1521-and BAL-like family. It also appear as smart00506, A1pp, Appr-1″-p processing enzyme, cd02908, Macro_Appr_pase_like, Macro domain, Appr-1″-pase_like family, PRK00431, ADP-ribose-binding protein, and pfam01661.

4.12. Indirect evidence from bioinformatics that the primary function of the VVVNAAN domain core could be as part of a poly-ADP-ribose polymerase

When using IHAVGPRWRDHEP that aligned with the second or (b) LHVVGPNVNKG, the result was similar to that above. BLASTp matches were dominated at highest levels of match by poly-ADP-ribose polymerases type 14 and mono-ADP-ribosyltransferase of turtles and birds, but also O-acetyl-ADP-ribose deacetylase of many bacteria, particularly the fermicutes, at 80–85% match. In searching the above data bases for

```
interferon alpha/beta receptor 2 [Platysternon megacephalum]
GenBank  ID: TFK07665.1 Length: 1692 Number of Matches: 1

Identities: 46/125(37%) Postives: 65/125(52%) Gaps: 17/125(13%)
Query  1    VVVNAANVYLKHGGGVAGALNKATNNAMQVESDDYIATNGPLKVGGSCVLSGHNL-AKHC  59
            VVVNA+N  LKH GG+A AL KA    +Q E D  +   GPL+ G + +    NL  K
Sbjct  854  VVVNASNEDLKHIGGLAEALLKAAGPELQTECDHIVRKRGPLQPGHAVITDAGNLPCKQV  913

Query  60   LHVVGP---NVNKGEDIQLLKSAY--------NFNQHEVLLAPLLSAGIFGADPIHSLRV  108
            +H VGP   +   G+ + LLK A          FN H + + P +S+GIFG      L++
Sbjct  914  IHAVGPRWRDHEPGKCVHLLKRAIKESLHLAETFNHHSIAI-PAISSGIFG----FPLKL  968

Query  109  CVDTV  113
            C  ++
Sbjct  969  CAQSI  973
```

Interferons are common feature of the vertebrates, and generally considered as lacking in invertebrates, at least as a major integrated cellular defense system. However, matches to interferon were the exception rather than the rule.

Another approach to exploring domain functions is similar but focuses on component parts and consider whether any parts of them correspond to known motifs of *known* function. The subsequences of the VVVNAAN domain, i.e. (a) VVVNAANVYLK HGGGVAGALNK, (b) LHVVGPNVNKG, and (c) PLLSAGIFG, separately produced many matches in the animal kingdom. Interestingly, the counterparts to the above three subsequences as found in the above interferon receptor generally appear to be better known as motifs than do the VVVNAAN domain core sequence. A BLASTp search was performed using only VVVNASNEDLKHIGGLAEALLK which was the *interferon receptor alignment* to the (a) subsequence VVVNAANVYLKHGGGVAGALNK from the VVVNAAN domain core of SARS-CoV-2. As discussed above, Prosite recognizes it, but as macro domain motif without comment on function. VNAANVYLKHGGGVAGALNK similarly matches "PF01661, macro known motifs, the SARS-CoV-2 subsequence (b) in this case produced no matches with known motifs at all, while for the corresponding interferon receptor subsequence IHAVGPRWRDHEP, results were very similar to those for the corresponding interferon receptor matches obtained in relation to subsequence (a).

Perhaps more surprisingly, for the above interferon receptor subsequence PAISSGIFG that aligned with the third or (c) PLLSAGIFG of SARS-CoV-2, the results also resembled those corresponding to (a), producing extensively poly-ADP-ribose polymerases and mono-ADP-ribosyltransferase type 14, albeit with a greater variety of vertebrate species, and with 100% coverage and 100% match. PLLSAGIFG did not correspond significantly to entry on the motif data basses. However, the analogous interferon receptor subsequence PAISSGIFG is a known motif NCBI-CDD ID: 239235 that can, as above, be accessed at https://www.genome.jp/tools/motif/. Here it is described as cd02907, Macro_Af1521_BAL_like, Macro domain, Af1521-and BAL-like family, consistent with the results for subsequence (a). In contrast the match with human interferons itself are not significant, e.g. the following has only 15% identity spanning the 113 residue VVVNAAN domain core.

```
VVVNAANcore      ------VVVNAANVYLKHGGGVAGALNKATNNAMQVESDDYIATNGPLK-VGGSCVLSGH   53
AAA36123.1       MAFVLSLLMALVLVSYGPGGSLGCDLSQ--NH-VLVGRKNLRLLDEMRRLSPHFCLQDRK   57
                       :::  . *    **.:.  *.:  *: : *  .:    :   :     *: . :

VVVNAANcore      NLAKHCLHVVGPNVNKGEDIQLL----KSAY-NFNQH-------EVLLAPLLSAGIFGAD   101
AAA36123.1       DFALPQEMVEGGQLQEAQAISVLHEMLQQSFNLFHTEHSSAAWDTTLLEPCR-TGL--HQ   114
                 ::*     * * ::::.: *.:*    :.::  *: .        .** *   :*:   :

VVVNAANcore      PIHSLRVCVDTV-------------------------------------------------   113

AAA36123.1       QLDNLDACLGQVMGEEDSALGRTGPTLALKRYFQGIHVYLKEKGYSDCAWETVRLEIMRS   174
                  :..* .*:. *

VVVNAANcore      ----------------------  113
AAA36123.1       FSSLISLQERLRMMDGDLSSP  195
```

The role of ADP-ribose-1″-phosphatase activity was in earlier studies considered not to be of particularly great biological importance, at least in studies *in vitro* [42]. Guided by the crystal structure of AF1521, an X domain homologue from Archaeoglobus fulgidus a sulfur-metabolizing microorganism; known *Archaeoglobales* are anaerobes, most of which are hyperthermophiles), potential active-site residues of the HCoV-229E X domain were targeted by site-directed mutagenesis. It was deduced that the HCoV-229E replicase polyprotein residues, Asn 1302, Asn 1305, His 1310, Gly 1312, and Gly 1313 are important at the active site [42]. That characterization of an ADRP-deficient HCoV-229E mutant revealed no significant effects on viral RNA synthesis and virus titer, and no reversion to the wild-type sequence was observed when the mutant virus was passaged in cell culture, thus seemed surprising. The authors [42] thus concluded the conserved X domain activity *in vitro* was *dispensable*, and that coronavirus replicase polyproteins have evolved to include nonessential functions. They cautioned, however, that the biological significance of the novel enzymatic activity *in vivo* remains to be investigated. This caution was judicious. It is highly unlikely that such a protein so well conserved in the coronaviruses, as in mammals, has no useful function, even if the main functional role is not exactly that first considered. The activities were investigated in more detail in viruses including coronaviruses, toroviruses, alphaviruses, hepatitis E (e.g. Refs. [43–47]). It was early considered that ADP-ribose 1″-phosphatase itself should be distinguished from related enzymes that attack ADP-ribose to produce AMP and ribose 5-phosphate, as well as several other enzymes involving substrates with adenine, ribose, and phosphate components, including importantly poly-ADP-ribose polymerase. However, many cases these enzymes have also been shown be or have macro domains.

As described below in Section 4.14, there are clearly many ways in which macro domains could potentially be involved the cell's defenses against viruses, but it remains that some macro domains certainly recognize poly ADP-ribose as a ligand. Poly-ADP-ribosylation is a common post-translational modification and an immediate DNA-damage-dependent post-translational modification of histones and other nuclear proteins. Apart from BLASTp matches with TFK07665.1 of the turtle P. Megacephalum itself, the top 100 matches of the VVVNAAN domain core are with sequences described as poly ADP-ribose polymerases type 14 and mono-ADP-ribosyltransferase from 96% match down to 73%, closest matches being with other turtles down to approximately 85%, and the rest primarily birds with a few reptiles. The discovery that mammalian macro domain proteins enzymatically remove ADP-ribose from proteins stimulated many studies of macro domains defined these domains as de-ADP-ribosylating enzymes, which indicates that these viruses have evolved to counteract antiviral ADP-ribosylation, likely mediated by poly-ADP-ribose polymerases. ADP-ribose is covalently attached to target proteins by poly-ADP-ribose polymerases using nicotinamide adenine dinucleotide (NAD+) as a substrate. ADP-ribosylation can alter enzyme activity, protein–protein interactions, and protein stability. Several type of poly-ADP-ribose polymerase are induced by interferon and are known to have antiviral properties, implicating ADP-ribosylation in the host defense response. Viruses can counter this innate immune response by interfering with PARP-mediated antiviral defenses and inhibiting cytokine production and the inflammatory response (see later below).

4.13. Experimental evidence that functions of the VVVNAAAN domain core are important to SARS-CoV-2 survival

Recall that ADP-ribose 1″-phosphatase activity is not, at the time of writing, considered by many authors as the biological main role of the coronavirus enzyme. Viral macro domains have relatively poor ADP-ribose 1″-phosphohydrolase activities, but efficiently bind free poly-ADP-ribose *in vitro*, whether it is free or bound to proteins [43]. It is now appreciated that coronaviruses with mutations in the macro domain are also highly attenuated pathogens in mammal host studies. A 2011 study [44] found that genetically engineered mutants of SARS-CoV and human coronavirus 229E (HCoV-229E) expressing ADRP-deficient macro domains displayed an increased sensitivity to the antiviral effect of α-*interferon*, compared with their wild-type counterparts. This data suggested that ADRP activities may well have a role in viral escape from the innate immune responses of the host [44]. Stated precisely for coronaviruses, in the words of those authors, the ADP-ribose-1″-monophosphatase domains of SARS-coronavirus and human coronavirus 229E mediate resistance to antiviral interferon responses [44]. In 2019, researchers [49] identified the ADP-ribosyltransferase poly-ADP-ribose polymerase family member 11 (PARP11) as a potent regulator of antiviral efficacy. It did not restrict type 1 interferon (IFN–I) production induced by certain viruses, but it did reduce signal activation by IFN-I. PARP11 mono-ADP-ribosylates the ubiquitin E3 ligase β-transducin repeat-containing protein (β-TrCP) and this promotes interferon α and β receptor subunit 1 (IFNAR1) ubiquitination and degradation. Ubiquitantion is the addition of ubiquitin proteins to proteins that can mark them for degradation via the proteasome, alter their cellular location, affect their activity, and promote or prevent protein interactions. It is part of a system "intended" to prevent or limit virus infection of cells but it is also an example of a host cell systems that viruses can sometimes hijack for their own "purposes". PARP11 expression is *upregulated* by a variety of virus infections, promoting ADP-ribosylation-mediated viral evasion. The researchers considered [46] ADP-ribosylation inhibitors and found that *rucaparib*, a first-in-class drug targeting the DNA repair enzyme poly-ADP ribose polymerase-1 (PARP1), can push PARP11 to stabilize IFNAR1. Rucaparib-rendered mice become more resistant to viral infection [45]. Ubiquitin E3 ligase β-transducin repeat-containing protein regulates its substrates and may itself provide a "druggable target for improving IFN antiviral efficacy" [45].

4.14. How important is poly-ADP-ribose polymerase activity compared with other possible functions of the VVVNAAN domain core?

The ability of both viral and human cytoplasmic proteins to carry out similar reactions and to do the reverse reactions, and the fact that both forward and backward seem capable of being processes in both in attack and defense, often results in authors speaking of a virus as upsetting a delicate balance of cellular response rather than committing to a specific biological role. This may well be justified, but despite that, the role of the above SARS-CoV-2 macro domain as enzyme must be important to the virus because of the high conservation of the VVVNAAN core sequence. If its binding function can be specifically inhibited without serious detriment to the host cells, then it could be the basis of a therapeutic agent. To help in such studies, several three dimensional protein structures are available. In 2006, Egloff et al. [43] determined the crystal structure of the SARS-CoV domain at 1.8-Å resolution in complex with ADP-ribose. In addition, Protein Data Bank entry 6W02, a crystal structure of ADP ribose phosphatase of NSP3 from SARS CoV-2 in the complex with ADP ribose had been submitted by Michalska and colleagues although an assisted journal publication is not yet available at the time of writing. Other Protein Data Bank entries described as SARS-CoV-2 ADP-ribose-phosphatases include 6W67, which are examined in some detail later below alongside 6W02, 6WEN, 6WCF, and 6VXS. Nonetheless, it remains that the above structures are described as ADP ribose phosphatases, and it does not automatically follow that the role of the macro domain with the VVVNAAN domain core is more specifically as a poly-ADP-ribose polymerase because phosphatase activity of various kinds figures prominently in the defense of cells against virus infection. There are many other possibilities, so the proper answer to the question posed by the title of this subsection 4.14 is probably that it is currently unclear. The complex subject of the *interactome* between virus and host cell proteins has many unresolved aspects but is gaining increased understanding under the pressure of the COVID-19 pandemic. Viruses evade the interferon system by partially blocking interferon

synthesis or interferon action, and the number and variety of these across the virus kingdom is indicative of the balance in the long coexistence of viruses and vertebrates [37]. Several virus proteins appear to interact with proteins involved in the innate immune pathways, notably the interferon signaling pathways involving type I interferon (IRF-1) production, interferon type III (IRF-3) activation, triggering the inflammatory response known as NF-kappa B.

4.15. Examples of XTRACT tags in gathering knowledge about the interactome

Development of a connection graph of the interactome for SARS-CoV-2 is extremely valuable [4], but in the work of the present kind one also often needs more details and the relation to other relevant things, needs to catch latest information, and to catch interpretations, opinions and theories even in the popular and professional media, that can generate ideas and hypotheses. XTRACT tags are already subgraphs that can be pieced together build an interactome of current scientific knowledge and opinion. In the present case, the knowledge obtained was more straightforward, but was instructive in appraising the significance of the above macro domain in the interactome. Examples are as follows.

Protein phosphorylation and dephosphorylation play important roles in innate immune responses to RNA viruses by regulating the activation and deactivation of multiple RLR-mediated signaling components, such as those known as RIG-I, VISA, TRAF3, TBK1 and IRF3. Some subgraphs of the interactome are traditionally well known as pathways. MARPLE's autosurfing commonly led to the web page at https://en.wikipedia.org/wiki/JAK-STAT_signaling_pathway. The so-called JAK-STAT pathways, i.e. pathways involving the Janus kinase (JAK)-signal transducer and activator of transcription (STAT). Some of the same genes can also be induced directly by viruses and double-stranded RNA produced during virus infection. A kinase domain appears important for JAK activity, since it allows JAK proteins such as JAK1, JAK2, JAK3, TYR2 of this pathway to phosphorylate (add phosphate groups to) proteins. Macro domain elements such as the VVVNAAN motif were not found in human versions of these proteins in the present study.

The above account so far concerns protein phosphorylation, but there are frequent adjacent references and links to RNA recognition.

<Q-UEL-Marple41 "`Another cellular enzyme **&and** RNAse L [0https://en.wikipedia.org/wiki/RNAse_L] **|also ^induced by| interferon |^action- destroys**| RNA |**within**| `the cells **|to**| further |**^reduce**| protein synthesis |**of**| `both viral **&and** host genes, protein synthesis **|^impairs**| `both virus replication |**^Inhibited**| protein synthesis, |as| `both virus replication |**^infected**| host cells" (source:='https://en.wikipedia.org/wiki/interferon' time:='Fri May 29 14:33:46 2020' extract:=125) Q-UEL-Marple41>

<Q-UEL-Marple41 "(interferon) |**^is ^activated by**| double-stranded RNA [0https://en.wikipedia.org/wiki/Double-stranded_RNA] |**as**| dsRNA **|^introduced to**| `the cells |**by**| `a viral infection" (source:='https://en.wikipedia.org/wiki/Protein kinase R' time:='Fri May 29 15:05:34 2020' extract:=679) Q-UEL-marple41>

This links back to https://en.wikipedia.org/wiki/Protein kinase R'.

<Q-UEL-marple41 "PKR |**^is ^induced by**| interferon [0https://en.wikipedia.org/wiki/Interferon] **|in**| `a latent (state)" (source:='https://en.wikipedia.org/wiki/Protein kinase R' time:='Fri May 29 15:05:34 2020' extract:=682) Q-UEL-marple41>

Nonetheless, despite a high number of hits relating to protein phosphorylation, there are also ample links relating this to RNA recognition. Weights can be assigned be the system based on recognition of key words and phrases on webpages such as protein phosphorylation and RNA simply by searching the XTRACTS generated. Overall the autosurfing produced proportional weights 0.6:0.4 for protein phosphorylation in proportion to RNA interactions.

<Q-UEL-Marple41 "Response |**to**| interferon; cells |**^produce**| large _amounts |**In**| response; large _amounts **|of**| `an enzyme [0https://en.wikipedia.org/wiki/Enzyme] **|^known as**| protein kinase R [0https://en.wikipedia.org/wiki/Protein_kinase_R] |as| PKR: (This) enzyme **|^phosphorylates|** [0https://en.wikipedia.org/wiki/Phosphorylation] |**as**| `a protein |**^known as**| eIF-2 [0https://en.wikipedia.org/wiki/interferon#cite_note-Brain-num-2] **|in**| response |**to**| `new viral infections" (source:='https://en.wikipedia.org/wiki/interferon' time:='Fri May 29 14:33:46 2020' extract:=123) Q-UEL-Marple41>

<Q-UEL-Marple41 "`the |**^phosphorylated**| eIF-2 |**^forms**| `an inactive `complex |**with**| `another protein |**^called|** eIF2B [0https://en.wikipedia.org/wiki/EIF2B] |**to ^reduce**| protein synthesis |**within**| `the cell" (source:='https://en.wikipedia.org/wiki/interferon' time:='Fri May 29 14:33:46 2020' extract:=124) Q-UEL-Marple41>

<Q-UEL-marple41 "PKR |**^can ^also ^be ^activated by**| `the protein PACT [0https://en.wikipedia.org/wiki/PRKRA] {OR} |**by**| heparin; PKR |**^contains**| `an N-terminal [0https://en.wikipedia.org/wiki/N-terminal] dsRNA (binding) domain |**as**| dsRBD {AND} |**as**| `a C-terminal [0https://en.wikipedia.org/wiki/C-terminal] kinase [0https://en.wikipedia.org/wiki/Kinase] domain |**^gives**| (?_it) pro-apoptotic [0https://en.wikipedia.org/wiki/Pro-apoptotic] |**as^cell-killing**| functions" (source:='https://en.wikipedia.org/wiki/Protein kinase R' time:='Fri May 29 15:05:34 2020' extract:=680) Q-UEL-marple41>

<Q-UEL-marple41 "`The dsRBD |^consists of| `two tandem (^copies)? | **of**| 'a (^conserved)? (^double)? stranded RNA (binding) motif **&and** dsRBM1 **&and** dsRBM2" (source:='https://en.wikipedia.org/wiki/Protein kinase R' time:='Fri May 29 15:05:34 2020' extract:=681) Q-UEL-marple41>

In considering design of inhibitors of macro domain function it was important to keep a link to the bioinformatics findings. Notably, the double-stranded RNA-binding protein in humans, e.g. GenBank NP_001157854.1, does not appear to contain a VVVNAAN or similar motif. Expression of type I and III interferons is induced in virtually all cell types upon recognition of viral molecular recognition patterns, especially on nucleic acids, by cytoplasmic and endosomal receptors.

Several families of RNA viruses have macro domain enzymes that remove ADP-ribose from proteins and so counter innate immune responses to virus infection. The human TARG1/C6orf130, MacroD1, and MacroD2 proteins reverse ADP-ribosylation by acting on ADP-ribosylated substrates through the hydrolytic activity of their macro domains. Despite other possible functions of macro domains in the interactome involving both protein phosporylation and ADP ribosylation, the term repeatedly encountered in relation to macro domains was PARPs, which as was indicated earlier above is simply the acronym for poly-ADP-ribose polymerases, e.g.

(H), and two glycine residues (G) that appear important in interactions with a ligand. As noted above, various experimental structures specifically for ADRPs and related enzymes have been known for some time. For example, the structures of ADP-ribose-1″-monophosphatase from yeast and its complex with ADP-ribose were determined to 1.9 A and 2.05 A, respectively in 2005 [46], and has been carefully analyzed.

The structure of the 284-amino acid protein shows a two-domain architecture consisting of a three-layer α−β−α sandwich N-terminal domain joined to a small C-terminal α-helical domain. Loop-region residues asparagine (N) at residue 80, aspartate (D) 90, and histidine (H) 145 may form a catalytic triad. The structure in complex with ADP-ribose revealed an active-site water molecule well positioned for nucleophilic attack on the terminal phosphate group. Fig. 2 shows two Protein Data Bank entries for the corresponding SARS-CoV-2 ADP-ribose-phosphatase, 6W6Y and 6W02. The left (green) structure is in complex with AMP, the right (magenta) is the structure similarly oriented with ADP-ribose. These contain the VVVNAAN domain core motif. The sequences given for SARS-CoV-2 ADRP in the Protein Data Bank as related entry 6WCF, and 6WO2, are as follows.

```
>6WCF_1|Chain A|SARS-COV-2 ADP RIBOSE PHOSPHATASE (ADRP)|Severe acute
respiratory syndrome coronavirus 2 (2697049)
GEVNSFSGYLKLTDNVYIKNADIVEEAKKVKPTVVVNAANVYLKHGGGVAGALNKATNNAMQVESDDYIATNGPLKV
GGSCVLSGHNLAKHCLHVVGPNVNKGEDIQLLKSAYENFNQHEVLLAPLLSAGIFGADPIHSLRVCVDTVRTNVYLA
VFDKNLYDKLVSSFLE

>6W02_1|Chains A,B|ADP ribose phosphatase|Severe acute respiratory syndrome
coronavirus 2 (2697049)
SNAGEVNSFSGYLKLTDNVYIKNADIVEEAKKVKPTVVVNAANVYLKHGGGVAGALNKATNNAMQVESDDYIATNGP
LKVGGSCVLSGHNLAKHCLHVVGPNVNKGEDIQLLKSAYENFNQHEVLLAPLLSAGIFGADPIHSLRVCVDTVRTNV
YLAVFDKNLYDKLVSSFLE
```

<Q-UEL-marple39 "(protein)? |**^predicted**| poly-ADP-ribose polymerases [0https://en.wikipedia.org/wiki/Poly(adp-ribose)_polymerases] |**as**| PARPs {AND} B aggressive lymphoma [0https://en.wikipedia.org/wiki/B-cell_lymphoma] |**as**| BAL |**as**| protein" (source:='https://en.wikipedia.org/wiki/macro domain' time:='Tue May 26 10:43:36 2020' extract:=69) Q-UEL-marple39>

In addition this explanation came from several other sources, e.g. GenBank entry XP_029192739.1 discussed above, where the annotation included the text describing macro domains as found in a variety of proteins as a stand-alone domain or in combination with other domains like in histone macroH2A and some PARPs.

4.16. Mechanistic features of the VVVNAAN domain core

The 3D structure of the macro domain target comprises a mixed alpha/beta fold of a mixed beta sheet sandwiched between four helices, and the ligand-binding pocket lies within the fold. There is also an asparagine-rich (N, i.e. Asn) commonly associated with catalytic sites of macro domains. Many have asparagine residues (N), a histidine residue

It is useful to split the ADRP sequence into four sections (which are actually contiguous). Analysis of the experimental three-dimensional structure by the present author considers the regions underlined and in bold font to be segments of sequence that contain residues involved at the active site.

(a) GEVNSFSGYLKLTDNVYIKNA**DIVE**EAKKVKPT
(b) ***VVVNAANVYLKHGGGVA****GALNK*ATNNAMQVESDDYIATNGPLKVGGSCVLSGHNLAKHC
(c) ***LHVVGPNVNK***GEDIQLLKSAYENFNQHE***VLLAPLLSAGIFG***ADPIHSLRVCVDTV
(d) RT**NVYLAVFDKNL**YDKLVSSFLE

Only (b) and (c) correspond to the so-called VVVNAAN domain, i.e. the subsections of SARS-CoV-2 ADRP that are found to detectably match ADRPs and related proteins animal in all the BLASTP matches discussed above. Recall the conserved regions from the many BLASTp matches

addressed VVVNAANVYLKHGGGVAGALNK, LHVVGPNVNKG, and PLLSAGIFG. These are reproduced in italics in the four contiguous sequences (a)-(c) above. Sections (b) and (c) make up the main features of the binding site or at least residues likely to significantly influence it, and (a) and (c) appear for the most part, peripheral (although some residues make import contributions).

Overall, it would seem unreasonable to speak of the binding site as split into an adenosine phosphate domain and a separate ribose binding domain. Sections (b) and (c), the so-called VVVNAAN domain core, represent a "domain core" to the eye. By "domain" is in this case, however, meant a single reasonably compact fold, a "mini-protein", and

SARS-CoV-2 ORF1 polyprotein and full human polymerase sequence is shown below. Only the (single) region that matches is shown. It will be assumed that the interactions of the substrate will be similar in the human polymerase. Residues likely to have direct or indirect influence on substrate binding are underlined. Those with more direct interactions and corresponding to Figs. 3 and 4 are underlined and in bold, and above is written the symbol @ for the involved in the adenine monophosphate component of the substrate, and those where above is written the symbol # are more involved in the phosphate ribose component.

```
                                                  #   #
                                @@@             # #   #  @ @@@
SARSCoV2ORF1      ------------DIVEEAKKVKPTVVVNAANVYLKHGGGVAGALNKATNNAMQVESDDY  1090
Human_PARP        FSRTVLAPGVVLIVQQGDLARLPVDVVVNASNEDLKHYGGLAAALSKAAGPELQADCDQI  569
                                :  ::   *****:*  *** **:*.**.**:.  :*.:.*:

SARSCoV2ORF1      IATNGPLKVGGSCVLSGH-NLAKHCLHVVGPNVNKGEDIQL---LKS-------AYENFN  1139
Human_PARP        VKREGRLLPGNATISKAGKLPYHHVIHAVGPRWSGYEAPRCVYLLRRAVQLSLCLAEKY-  628
                  :  :* *  *.: : ..     :* :*.***. .  *  :    *:          *::
                        @   @@@@   @                 @@   @
SARSCoV2ORF1      QHEVLLAPLLSAGIFGADPIHSLRVCVDTVRTNVYLAVFDKNLYDKLVSSFLEMKSEKQV  1199
Human_PARP        KYRSIAIPAISSGVFGFPLGRCVETIVSAIKENFQFKKDGHCL----KEIYLVDVSEKTV  684
                  ::. :  * :*:*:**    :.:.. *.::: *. :   .: *    . :*   *** *
```

it remains that there are two discernable binding sites in the VVVNAAN domain core. They are usefully examined in the Adenosine Monophosphate (AMP) interactions in the substrate complex in Protein Data Bank (PDB Entry 6W6Y) as in Fig. 3, and the ribose phosphate interactions in adenosine diphosphate ribose complex in PDB entry 6W02 (Fig. 4) as details are then arguably slightly clearer, because it is of particular interest to focus on the smaller AMP structure as a basis for developing candidate drugs, but focusing on either whole structure gives similar conclusions, and binding studies could be carried out on 6W6Y or 6W02. Note that the residue numbering of each of these PDB entries as used in Figs. 3 and 4 differs slightly but is used here for reproducibility in reference to the source PDB entries.

4.17. The need to avoid drug interactions with the human polymerase

To design therapeutic agents that antagonize the virus without impairing functions in the human host will require care because the VVVNAAN domain core persists as a recognizable match to a human polymerase, as follows.

In designing antagonists to inhibit these regions from binding to a *protein receptor*, a retro-inverso approach might be attempted as a first step in developing smaller drug liagnads. Here the section of sequence is written backward and synthesized with D-amino acids [4], but in this case they represent recognition sites with adenine nucleotides, not a protein, and design of an organic ligand is indicated. What the above indicates is, for example, that a ligand is required that binds to SARS-CoV-2 macro domain sequence VVVNAANVYLKHGGGVA and not to VVVNASNEDLKHYGGLA, its human polymerase counterpart, and at the same time to bind to

LHVVGPNVNKGEDIQLLKSAYENFNQHEVLLAPLLSAGIFG

and not to

IHAVGPRWSGYEAPRCVYLLRRAVQLSLCLAEKYKYRSIAIPAISSGVFG

its human polymerase counterpart. In the first case the need to interact with VYLKHG rather than EDLKHYG is indicated. In the second case, the need to interact with NVNKGEDIQ rather than RWSGYEAPRCVY is indicated. In addition recall that residues DIVE interact with the adenine, but it shortly precedes the VVVNAAN motif. It is replaced by VQQG in the human polymerase. However, not all these residues in all these subsequences interact intimately with the substrate, as a consequence of how their structure is folded in space.

```
poly (ADP-ribose) polymerase family, member 14, isoform CRA_a [Homo sapiens]
GenBank ID: EAW79467.1 Length: 1424 Number of Matches: 1
Range 1: 535 to 654
Identities: 40/124(32%) Positives:61/124(49%) Gaps: 14/124(11%)

Query  1056  VVVNAANVYLKHGGGVAGALNKATNNAMQVESDDYIATNGPLKVGGSCVLSGHNLA-KHC  1114
             VVVNA+N  LKH GG+A AL+KA    +Q + D  +   G L  G + +     L   H
Sbjct  535   VVVNASNEDLKHYGGLAAALSKAAGPELQADCDQIVKREGRLLPGNATISKAGKLPYHHV  594

Query  1115  LHVVGPNVNKGED---IQLLKSAYE------NFNQHEVLLAPLLSAGIFGADPIHSLRVC  1165
             +H VGP  +  E    + LL+ A +          ++  +  P +S+G+FG       L  C
Sbjct  595   IHAVGPRWSGYEAPRCVYLLRRAVQLSLCLAEKYKYRSIAIPAISSGVFG----FPLGRC  650

Query  1166  VDTV  1169
             V+T+
Sbjct  651   VETI  654
```

The query is the SAR-CoV-2 VVVNAAN domain core sequence and subject (Sbjct) is the human polymerase. The residues at, and in the vicinity of, the binding site residues of SARS-CoV-2 ADP-ribose-phosphatase, of those that can be considered as continuous peptides, are shown underlined and in bold. The Clustal Omega alignment for the full

One may focus on short subsequences where (a) all or almost all the residues interact with the substrate, (b) would make key differences by involving binding an inhibitor to the SARS-CoV-2 ADPR rather than the human polymerase, and (c) take account of the more detailed

orientations of sidechains to see where the differences in amino acids residues would reasonably have strongest effect. Given these requirements, the choices arguably reduce just to very few of which two are of particular interest as follows. A drug would need to fit the SARCoV-2 aspartate-isoleucine-valine (DIV) sequence, but not the human polymerase valine-glutamine-glutamine sequence (VQQ), and the valine-phenylalanine-aspartate-lysine (VFDK) sequence, not the lysine-aspartate-glycine-histidine (KDGH) sequence. As peptides they would for example be lysine-valine-isoleucine (KVI) and phenylalanine-valine-lysine-aspartate (FVKD), with some allowance for one to three residues on either side to enhance binding (steric hindrances in the binding site permitting), and consideration given to rendering them as retroinverso peptidomimetics [4,5]. There is also the concern that there are other proteins to which binding should be avoided, and indeed this is ultimately the basis of most adverse reactions to pharmaceutical drugs, but poly-ADP-ribose polymerase was the closest human match found in this case.

4.18. Preliminary binding studies in silico

The preceding papers [4,5] reported preliminary binding studies for ligands that were potential drug candidates to a different target, and these methods were used here in preliminary studies to assess the challenges for the phosphatase. Except for addition of tools that hope to achieve better exploration of conformational space, the methods essentially follow standard approaches to estimating binding strength, more detailed docking and estimation of binding energy. Results are particularly preliminary in the present case because of the involvement of water molecules in the binding, and the phosphate charge. Multiple local minima for binding modes were found and no strong binders have as yet found in the preliminary study (as classified in the preceding papers [4,5]), suggesting that he results are as yet far from complete. The study addressed an arbitrary selection of compounds that are already available and basically simply reflected reflect placing the adenine or adenine-like component of the molecule in the same spatial location and orientation as the adenine ligand in the above experimental structures (Figs. 5 and 6).

5. Discussion and conclusions

5.1. The use of knowledge-gathering techniques and augmented reasoning

Abandoning the original, rather demanding, intention to have the present paper written largely automatically by the system has created less need for the tentative further new steps in technology that the present author has recently explored. This new development is not abandoned, but COVID-19, not algorithmic development, has priority where other routes are possible. The simple use of MARPLE, repurposed from its original multiple-choice examination role (and knowledge testing and curation roles), has been useful and sufficient for managing knowledge and its extraction from Internet text, in the present study. No recoding (reprogramming) was required. As has been emphasized above and in the previous papers on SARS-CoV-2, none of Q-UEL-related applications (modules) used in the project [3–6] are essential for *reproducing* the work in this project. Bioinformatics tools available on the Internet can be access by the researcher in the usual way, rather than though the Q-UEL system. Nonetheless, Q-UEL applications can facilitate a rapid response to a new kind of pandemic, as was particularly obvious in Refs [3,4] where the research and writing of the papers took approximately one week following appearance of the final version of the Wuhan seafood market isolate on GenBank. See also Section 5.4 below.

Despite the less ambitious approach, the knowledge-gathering techniques as used here essentially represent an application to virology of our algorithms for a semi-automated approach to systematic review [23] which includes gathering information at the start of a new project, or even in deciding upon a particular project under the threat of a new disease. Although some detail was given in Ref. [4], the subsequent papers focused on bioinformatics and results. In contrast, in the present paper, examples have been given much more extensively. The approach does not appear to have close counterparts in other published work except those of the present author and colleagues (e.g. Ref. [22]) which as noted above previously had a different purpose (to test and curate gathered knowledge for automated medical reasoning). Obviously, it is ultimately akin to a browser such as Google, but it works in a different way (e.g. it does not index web pages) and importantly it is an automated one that runs in the background and gathers knowledge while other work is being done at the computer. It is certainly somewhat unusual as applied in the current paper because the queries take the form of an examination question with candidate answers to focus the search on the topics of interest. By accumulating knowledge in canonical form also usable in inference [21–26], it also has obvious affinities to projects such as the Semantic Web and even earlier Expert Systems as discussed in Refs [21–26], but it is not a static repository of knowledge. Rather, it is a combination of previously acquired and fresh knowledge. The Knowledge Representation Store KRS is searched first to attempt to resolve queries in the format of an examination question, and then new knowledge is added by automated surfing.

The molecular and biomedical findings remain of interest and, as noted above, have the priority. In some ways, the current content is a little less pressing. Unlike subsequences of interest in the previous papers, the macro domain turns out to be already a known motif. Nonetheless, as far as SARS-CoV-2 and even the previous SARS coronaviruses are concerned, it has been “an orphan target” looking for a home, and even a fuller understanding. The function and biological roles of the VVVNAAN Domain Core of the Nsp3 macro domain in SARS-CoV-2 and other coronaviruses, and the extent of its importance to survival and perpetuation of the virus, have been less clear. To the author’s knowledge at the time of writing, *detailed* consideration has not been given to the motif as a therapeutic target in SARS-CoV-2. Some subsequent developments are however discussed later below.

5.2. Direct use of XTRACT tags

Some further clarification may be made in regard to the comments made at the end of Theory Section 2. In the present study, a great deal of use was made of querying the knowledge data base for relevant Q-UEL tags, including XTRACTS. Although all Q-UEL tags are designed to be readable by eye for ease of maintenance and to provide medical data if part of the IT infrastructure collapsed in a disaster, XTRACT tags were not intended to be normally seen because reading of these by eye, with their distorted grammatical structure that facilitates decomposition into semantic triples such as subject-verb-object facilities interpretation by computer. The canonical form enabling that is doubtless not to everyone’s taste when reading directly. There are tools to facilitate reading, such as expressing the intermediate steps as bullet points, or simply displaying segments of extracted text as originally written, a trivial option because it is the first step in the production of XTRACT tags. For the user experienced with XTRACT tags, however, it provides links of several kinds (in-text-links, references in the reference list, and number and time stamp of the extract) for the human user, or the automated system, to follow through. These also allow the user to consider the changes in the text and knowledge represented in rapid change COVID-19 webpages, comparing entries.

5.3. Truly “very conserved” sites

As emphasized from the outset (Introduction Section 1.2), it is important to target molecular recognition sites of SARS-CoV-19 that are highly conserved sequences of amino residues across many strains and species. These are unlikely to mutate readily and escape from the actions of vaccines and therapeutics. Otherwise, a synthetic peptide vaccine or novel therapeutic antiviral drug that still inevitably requires taken many

person-hours to develop could be useless in weeks or months. That is not only because of the rapid rates of mutation of RNA viruses [7,8] but also the huge number of people currently infected with COVID-19, representing an astronomic number of virus particles all responding rapidly to Darwinian natural selection. At the time of writing, SARS-CoV has unfortunately shown the expected capability to escape [33,34]. Also, as discussed in this paper, neutralizing antibodies recently found against COVID19 that are claimed to bind a "highly conserved" site [35] are binding a site that is not so well conserved by the present author's rather stringent criteria of "conserved". Temporary solutions could of course "hold the fort" against the attack of the virus for a useful period. The same problem could be so for new therapeutics. A compromise solution may be possible: in such cases it may be particulars wise to attack COVID-19 at several sites even if each site is susceptible to accepted mutation, since the probability of an escape mutation is then dramatically decreased.

Fortunately, there are sites and patterns of amino acid residues in proteins inside and at the surface of the virus that are conserved to a peculiar degree, and many in non-structural proteins coded by the viral genome that are not incorporated into the virus particle. The VVVNAAN domain core described in the present paper is the most conserved in all the SARS-CoV-2 proteins as judged by the reasonable criteria used here. It contains much of the active site of an enzyme that (as hinted above but even right from the outset), seems to have been somewhat of an "orphan", somewhat in the backstreets of coronavirus research. This is firstly because the virus can replicate in cells in the "test-tube" with that gene eliminated, secondly because it seemed to have very weak enzymic activity, and thirdly it seemed to possess an enzymic activity of no obvious great importance. This is the ADP-ribose 1″-phosphatase, ARDP discussed in this paper. However, the fact that an amino acid sequence appears to have been preserved by evolution for millions of years is a powerful clue in bioinformatics.

5.3. The cellular function of the VVVNAAN domain core of SARS-CoV-2

Some further comment on the biology, discovered later in the project by MARPLE, may be helpful. In many ways the widespread nature of the macro domain and its ADP-ribose-phosphate-related functions across the living world is a confusing feature because both virus and host are armed with what looks like similar weapons. ADP-ribosylation is a reversible post-translational modification that occurs in animals, plants, and bacteria. The ADP-ribosyl transferases in all cases appear to add poly-ADP-ribose and mono-ADP-ribose to proteins. In the human genome, there are 15 genes coding at least for the former, and these are the PARPs (PARP1–PARP16). Not all have the macro domain. There has been an evolutionary conservation in parts of the sequences of these PARPS as well as divergence among primates and non-mammal species of specific regions of PARP9, 14 and 15 and it is these that are the PARPs with macro domains. PARP14 has been directly implicated in the induction of interferon in mouse and human cells, indicating a critical role in the regulation of innate immunity. The activity attributed to ARDP has been recently found to play a key role in a pathway protecting the virus against the ancient innate immune system inside our cells, involving interferons and ubiquitin proteins [44–46], but the balance is subtle. It is known that the ADP-ribose-1″-phosphatase domains of SARS-coronavirus and human coronavirus 229E mediate resistance to antiviral interferon responses [44], while ADP-ribosyltransferase PARP11 might modulate that attack by mono-ADP-ribosylating the ubiquitin E3 ligase β-TrCP [45]. Overall, it is increasing the level of β-interferon that seems to be particularly important for the defense of mammalian cells against viruses. In the current COVID-19 pandemic, there is news that deficiency in β-interferon production by the lung could explain the enhanced susceptibility of these at-risk patient groups to developing severe lower respiratory tract (lung) disease during respiratory viral infections [47].

The essential features of the current (2020) understanding are described by Fehr and colleagues [48,49]. The role as a eukaryote Poly-ADP-ribose polymerase (PARP) is the likely actual role, not as a replicase but as involved in a number of cellular processes such as DNA repair and stability, and programmed cell death. However, ADP-ribose phosphate monomers could play a role in innate immunity which is essentially the same as the corresponding polymers. As noted above, not all PARPS have a macro domain but the PARPs in general are involved in ADP-ribosylation as a widely distributed post-translational modification. The ADP-ribose moiety is transferred from NAD^+ to amino acid residues of target proteins, leading to either mono-ADP-ribosylation or poly-ADP-ribosylation. This post-translational modification regulates a number of biochemical processes. Poly-ADP-ribosylation was initially seen as an immediate DNA-damage-dependent post-translational modification of histones and other nuclear proteins, with PARPs signaling the presence of DNA damage by adding ADP-ribose units to DNA, histones, and other DNA repair enzymes, and facilitating repair of DNA strand breaks. It is nonetheless now fairly well accepted that several of the mammalian mono-ADP-ribosylating PARPs are powerful antiviral proteins that are able to inhibit viruses of many kinds. Recent reports are making it clear that in many cases ADP-ribosylation is more general still, maintaining protein homeostasis by an interwoven set of processes that regulate the levels and stability of proteins in cells. Rack, Perina, and Ahel give a good account of the functions of macro domains in general [50], while Poltronieri gives a convenient editorial summary of the PARPS as a whole [51], in which there has been some rapid evolution, particularly among those without macro domains, suggestive to many authors for a role of ADP-ribosylation in host-virus conflicts. PARP13 restricts the replication of several families of viruses and shows sites of positive selection in the PARP catalytic domain, an apparent target for genetic conflicts with viruses. The PARP13 zinc finger directly binds to viral RNA and this solicits the exosome for the specific degradation of the viral RNA. PARP13 (ZAP), and PARP4 are involved and PARP1, 7, 10, and 12 have specifically been shown to play roles in repressing viral replication.

5.4. More recent developments

It is well known that research into SARS-like coronaviruses and their potential therapeutic targets has progressed rapidly in the present pandemic. As a kind of "stop press", some additional comments may be made. Recent papers essentially support both the above findings and the degree of caution shown regarding a detailed interpretation over what role the Nsp3 macro domain does have, in defending SARS-CoV-2 against the host cell. Since preparation of this manuscript several preprints have indicated that the viral macro domain counters host antiviral ADP-ribosylation and specifically that it removes ADP-ribose from proteins, e.g. Ref [52]. However, strictly speaking that reference appears to be a proposal and is not specifically focused on SARS-CoV-2. Removing ADP-ribose appears the simplest and natural interpretation of biological role *a priori*, but the issue remains that the role of the viral macro domain may be more subtle. ADP-ribosylation might activate or inactivate a protein. A viral enzyme could (in principle) ADP-ribosylate a host protein to activate or to deactivate it. In a recent paper [53] it is confirmed that the SARS-CoV-2 macro domain binds ADP-ribose, and that this is a step needed to justify screens for potential antivirals that bind in place of ADP-ribose, but that paper also explicitly states that more work needs to be done because the biological role for ADP-ribose binding is not completely understood. What seems clear is that this macro domain is important to the virus, and so a point of vulnerability: blocking its action is likely to be important for a therapeutic solution, although the following consideration remains.

After the initial preparation of the present paper, Webb and Saard [54] noted the sequence homology between human PARP14 and the SARS-CoV-2 ADP ribose 1′-phosphatase, and similarly noted its potential importance and potential as a "druggable target". This highlights that the traditional use of bioinformatics can, as already emphasized

above, obtain similar discoveries and confirm present results without the knowledge gathering and related Q-UEL methods. However, it is noticeable that the electronic publication date of their paper (11th June 2020) is still a relatively late date following the publication of the Wuhan Seafood market isolate in GenBank in January [2] as discussed in Refs. [3,4]. It seems surprising that there have been no extensive publications on this topic specific to SARS-CoV-2 prior to June. It could be argued that because the macro domain was known to occur in coronaviruses and is widely spread across so many species well beyond viruses, that he discovery of it in SARS-CoV-2 is less profound (i.e. not unexpected), but lack of publications does not seem consistent with the opportunities of it as a therapeutic target (which Webb and Saard emphasize) and the pressing importance of COVID-19. In contrast a different version of the present paper highlighting the homology and then entitled "Example Studies of Well Conserved Regions in SARS-CoV-2 Proteins as Targets with Lower Risk of Escape Mutations. Examples of … and a VVVNAAN Domain Core of the Nsp3 Macro Domain" was submitted to the present journal on 23rd April. Although it has been extensively rewritten with new emphasis and material, that is at least evidence that the present methods can help achieve a rapid response to a new epidemic.

The ability to extract knowledge rather than simply discover papers was the original aim, but the ability also to do the latter can of course be tested by more standard browsing methods such as Google, with which it can of course be usefully combined. The approach did fail to discover the paper by Webb and Saard when applied to help in the later rewriting, and this omission is insightful regarding possible improvements. It seems an unfortunate target to miss because "ADP ribose 1′-phosphatase" appears in their title. The paper was not well cited at that time so there were less opportunities for searches to thread through to that paper, and it is noticeable that focus and definitions can evolve at early stages of a research topic, which can create challenges for automated surfing of this kind. Triggering queries were at that stage on a poly-ADP-ribose polymerase activity as the most interesting-looking activity of the same enzyme at that time, the 1″ phosphatase, and particularly on NSp3 and macro domains. These did not appear in the summary and main body of the text of the paper, and there is no mention of subsequences VVVNAAN and VIVNAAN etc. characteristic of the virus macro domains, in the *text*. They do clearly appear in a *figure*, but such content is not accessible for standard text analytics, which makes a case for bioinformatics output being rendered as text in publications. That all said, the present paper is arguably more extensive in its observations of homology and analysis: as stated earlier above, PARP9, 14 and 15 contain macro domains and PARP14 is particularly interesting because it has been directly implicated in the regulation of innate immunity. In summary, despite some limitations that are arguably not too surprising, the approach remains promising for facilitating and accelerating research.

5.5. The need for a finer scalpel

Virologists long experienced in coronavirus studies may well be able to add much more insight to the above observations. Human expertise is important, and COVID-19 is rapidly driving great understanding of the above processes. Insight evolves daily. However, few researchers would deny that the balance between viral infection and the host cell's defenses is a fine one: some potential therapeutic compounds might be rather blunt instruments. Published works suggest that ADP-ribosylation is involved in distinguishing and appropriately partitioning "good" and "bad" or damaged proteins, including targeting viral proteins for ubiquitination and degradation. Findings indicate that the VVVNAAN domain core in the macro domain of the coronavirus is required by the virus to prevent PARP-mediated inhibition of coronavirus replication, but the macro domain is also used directly in the enhancement of interferon production. In general, the macro domains in human proteins regulate a wide variety of cell processes, and are involved in repairing damage to the genetic material, signal transduction, the immune response as a whole, cancer, defects in development, and neurodegenerative disease. However, recognizing the difference between good and bad is not perfect, and viruses are notorious for being able to hijack or "twist" the cell's defenses to their own advantage. A role for PARP favoring viral infection was originally suggested by findings that benzamide and benzopyrone analogues that inhibit PARP diminish retroviral infection, but this was not seen in all studies and depends on the virus and strain. Mutations in the virus have a lot of scope for changing the balance of power and the specific mechanisms used in the interactome, and those can change with virus strains. Only relatively recently was it shown that PARP inhibition enhanced replication and inhibited interferon production in primary macrophages infected with macro-domain-mutant (but not wild-type) coronavirus [49]. Knockdown of two abundantly expressed PARPs, PARP12 and PARP14, led to increased replication of mutant but had no effect on the wild-type virus. At the time of the present paper going to press, many papers by many groups regarding SARS-CoV-2 mutations and interactions with host cell proteins look likely to be publically available soon.

Consequently, the SARS-CoV-2 macro domain looks likely to be well recognized as a potent target to counter COV-19, but care in molecular design will be required to avoid interfering with the important functions of human macro domain proteins. A finer "scalpel" is required to avoid unwanted action on human relatives. Continued careful analysis of the conserved VVVNAAN domain core motif and the variations of the regions of amino acid residue sequence of the conserved three subsequences (a) VVVNAANVYLKHGGGVAGALNK, (b) LHVVGPNVNKG, and (c) PLLSAGIFG, as well as of the more variable regions between and around them, well might hopefully provide that.

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 in January 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. June 2020.

[6] B. Robson, Bioinformatics studies on a function of the SARS-CoV-2 spike glycoprotein as the binding of host sialic acid glycans, Comput. Biol. Med. 122 (July 2020) 103849, 2020.

[7] R. Sanjuán, M. Nebot, N. Chirico, L.M. Mansky, R. Belshaw, Viral mutation rates, J. Virol. 84 (19) (2010) 9733–9748, https://doi.org/10.1128/JVI.00694-10.

[8] K.M. Peck, A.S. Laurin, Complexities of viral mutation rates, J. Virol. 92 (14) (2018) 1031–1037.

[9] R. Dawkins, The Selfish Gene, Oxford University Press, 1976.

[10] J.O. Wertheim, D. Chu, J.S.M. Peiris, S.L.K. Pond, L.L.M. Poon, A case for the ancient origin of coronaviruses, J. Virol. 87 (12) (2013) 7039–7045.

[11] E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, Lancet Infect. Dis. (2020).

[12] David E. Gordon, Gwendolyn M. Jang, Mehdi Bouhaddou, Jiewei Xu, Kirsten Obernier, Matthew J. O'Meara, Jeffrey Z. Guo, Danielle L. Swaney, Tia A. Tummino, Huettenhain Ruth, Robyn M. Kaake, Alicia L. Richards, Beril Tutuncuoglu, Helene Foussard, N.J.K. Jyoti Batra, A SARS-CoV-2-Human Protein-Protein Interaction Map Reveals Drug Targets and Potential Drug Repurposing, 2020.

[13] J.S. Morse, T. Lalonde, S. Xu, W.R. Liu, Learning from the past: possible urgent prevention and treatment options for severe acute respiratory infections caused by 2019-nCoV, Chembiochem 21 (2020) 730–738.

[14] G. Li, E. De Clercq, Therapeutic options for the 2019 novel coronavirus (2019-nCoV), Nat. Rev. Drug Discov. 19 (2020) 149–150.

[15] M. Adachi, T. Ohhara, K. Kurihara, T. Tamada, E. Honjo, N. Okazaki, S. Arai, Y. Shoyama, K. Kimura, H. Matsumura, S. Sugiyama, H. Adachi, K. Takano, Y. Mori, K. Hidaka, T. Kimura, Y. Hayashi, Y. Kiso, R. Kuroki, Structure of HIV-1 protease in complex with potent inhibitor KNI-272 determined by high-resolution X-ray and neutron crystallography, Proc. Natl. Acad. Sci. Unit. States Am. 106 (12) (2009) 4641–4646, https://doi.org/10.1073/pnas.0809400106.
[16] Y.M. Báez-Santos, S.E. St John, A.D. Mesecar, The SARS-coronavirus papain-like protease: structure, function and inhibition by designed antiviral compounds, Antivir. Res. 115 (2015) 21–38.
[17] R.N. Kirchdoerfer, A.B. Ward, Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors, Nat. Commun. 10 (2019) 2342.
[18] B. Robson, Hyperbolic Dirac nets for medical decision support. Theory, methods, and comparison with Bayes nets, Comput. Biol. Med. 51 (2014) 183–197.
[19] S. Deckelman, B. Robson, Split-complex numbers and Dirac bra-kets, Commun. Inf. Syst. 14 (3) (2015) 135–149.
[20] B. Robson, Bidirectional general graphs for inference. Principles and implications for medicine, Comput. Biol. Med. 108 (2019) 382–399.
[21] B. Robson, T. Caruso, U.G.J. Balis, Suggestions for a web based universal exchange and inference language for medicine", 43, Comput. Biol. Med. 1 (12) (2013) 2297–2331.
[22] B. Robson, B, S. Boray, Data mining to build a knowledge representation Store for clinical decision support. Studies on curation and validation based on machine performance in multiple choice medical licensing examinations, Comput. Biol. Med. 73 (2015) 71–93.
[23] B. Robson, Studies in using a universal exchange and inference language for evidence based medicine. Semi-Automated Learning and Reasoning for PICO Methodology, Systematic Review, and Environmental Epidemiology", Comput. Biol. Med. 79 (2016) 299–323.
[24] 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.
[25] 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) in press.
[26] B. Robson, POPPER a simple programming language for probabilistic semantic inference in medicine, Comput. Biol. Med. 56 (2014) 107–123.
[27] National institutes of Health, National Library of Medicine, Blast. https://blast.ncbi.nlm.nih.gov/Blast.cgi.
[28] A.M. Lesk, Introduction to Bioinformatics, fifth ed., Oxford University Press, 2019.
[29] T. Engel, J. Gasteiger (Eds.), Applied Chemoinformatics: Achievements and Future Opportunities, Wiley-VCH, 2018.
[30] ZINC15 database (last acessed March 9 2020), https://zinc.docking.org/.
[31] I.j. Liu, W.T. Tsai, L.E. Hsieh, L.L. Chueh, Peptides corresponding to the predicted heptad repeat 2 domain of the Feline coronavirus spike protein are potent inhibitors of viral infection, PloS One 8 (12) (2013), e82081.
[32] D. Forni, G. Filippi, R. Cagliani, L. De Gioia, U. Pozzoli, N. Al-Daghri, M. Clerici, Manuela Sironi, The heptad repeat region is a major selection target in MERS-CoV and related coronaviruses, Sci. Rep. 5 (2015) 4480.
[33] J.B. Berend, J.W.A. Rossen, W. Bartelink, C. Zuurveen, A.M. de Hann, C.A. B. Duquerroy Boucher, P.J. Rottier, Coronavirus escape from heptad repeat 2 (HR2)-Derived peptide entry inhibition as a result of mutations in the HR1 domain of the spike Fusion protein, J. Virol. (2008) 2580–2585. March.
[34] B. Rockx, E. Donaldson, M. Frieman, T. Sheahan, D. Corti, A. Lanzavecchia, R. S. Baric, Escape from human monoclonal antibody neutralization affects in vitro and in vivo Fitness of severe acute respiratory syndrome coronavirus, JID (J. Infect. Dis.) 201 (6) (2010) 946–955, https://doi.org/10.1086/651022.
[35] M. Yuan, N.C. Wu, X. Zhu, C.-C.D. Lee, T.Y. Ray, R.T.Y. So, H. Huibin Lv, C.K. P. Mok, I.A. Wilson, A highly conserved cryptic epitope in the receptor-binding domains of SARS-CoV-2 and SARS-CoV, Science (03 Apr 2020) eabb7269, https://doi.org/10.1126/science.abb7269.
[36] J.W. Tang, J.L. Cheung, I.M. Chu, M. Ip, M. Hui, M. Peiris, P.K. Chan, Characterizing 56 complete SARS-CoV S-gene sequences from Hong Kong, J. Clin. Virol. 38 (1) (2007) 19–26.
[37] F.S. Cohen, How viruses invade cells, Biophys. J. 110 (5) (2016) 1028–1032, https://doi.org/10.1016/j.bpj.2016.02.006.
[38] K.S. Saikatendu, J.S. Joseph, V. Subramanian, T. Clayton, M. Mark Griffith, K. Moy, J. Velasquez, B.W. Neuman, M.J. Buchmeier, R.C. Raymond C Stevens, P. Peter Kuhn, Structural basis of severe acute respiratory syndrome coronavirus ADP-ribose-1''-phosphate dephosphorylation by a conserved domain of nsP3, Structure 13 (11) (2005) 1665-1675, https://doi.org/10.1016/j.str.2005.07.022.
[39] J. Tan, C. Vonrhein, O.S. Smart, G. Bricogne, M. Bollati, Y. Kusov, G. Hansen, J. R. Mesters, C.l. Schmidt, Guido Hansen, R. Hilgenfeld, The SARS-unique domain (SUD) of SARS coronavirus contains two macrodomains that bind G-quadruplexes, PLoS Pathog. 5 (5) (2009), e1000428, https://doi.org/10.1371/journal.ppat.1000428.
[40] L. Mousavizadeha, S. Ghasemi, Genotype and phenotype of COVID-19: their roles in pathogenesis, J. Microbiol. Immunol. Infect. (2020), https://doi.org/10.1016/j.jmii.2020.03.022 [Epub Ahead of print].
[41] A. Bianchi, C. Giorgi, P. Ruzza, C. Toniolo, E.J. Milner-White, A synthetic peptide designed to resemble a proteinaceous P-loop nest is shown to bind inorganic phosphate, Proteins 80 (2020) 1418–1424, https://doi.org/10.1002/prot.24038.
[42] A. Putics, W. Filipowicz, J. Hall, A.E. Gorbalenya, J. Ziebuhr, ADP-ribose-1"-monophosphatase: a conserved coronavirus enzyme that is dispensable for viral replication in tissue culture, J. Virol. 279 (20) (2005) 12721–12731.
[43] M.-P. Egloff, H. Malet, K. Putics, M. Heinonen, H. Dutartre, A. Frangeul, A. Gruez, V. Campanacci, C. Cambillau, J. Ziebuhr, T. Ahola, B. Canard, Structural and functional basis for ADP-ribose and poly(ADP-ribose) binding by viral macro domains, J. Virol. 80 (17) (2006) 8493–8502, https://doi.org/10.1128/JVI.00713-06.
[44] T. Kuri, K.K. Eriksson, A. Putics, R. Züst, E.J. Snijder, A.D. Davidson, S.G. Siddell, V. Thiel, J. Ziebuhr, F. Weber, The ADP-ribose-1''-monophosphatase domains of SARS-coronavirus and human coronavirus 229E mediate resistance to antiviral interferon responses, J. Gen. Virol. 92 (2011) 1899–1905, 08.2011.
[45] T. Guo, Y. Zuo, L. Qian, J. Liu, Y. Yuan, K. Xu, Y. Miao, Q. Feng, X. Chen, L. Jin, L. Zhang, C. Dong, S. Xiong, H. Zheng, ADP-ribosyltransferase PARP11 modulates the interferon antiviral response by mono-ADP-ribosylating the ubiquitin E3 ligase β-TrCP, Nat. Micobiol. 4 (11) (2019) 1872–1884.
[46] D. Kumaran, S. Eswaramoorthy, F.W. Studier, S. Swaminathan, Structure and mechanism of ADP-ribose-1''-monophosphatase (Appr-1''-pase), a ubiquitous cellular processing enzyme, Protein Sci. 14 (3) (2005) 719–726.
[47] R. Staines, Duo of UK firms announce Coronavirus therapy trials, Pharmaphorum, https://pharmaphorum.com/news/duo-of-uk-firms-announce-coronavirus-therapy-trials/.
[48] A.R. Fehr, S.A. Singh, C.M. Kerr, S. Mukai, H. Higashi, M. Aikawa, The impact of PARPs and ADP-ribosylation on inflammation and host–pathogen interactions, Genes Dev. 34 (2020) 341–359.
[49] M.E. Grunewald, Y. Chen, C. Kiny, T. Maejima, R. Lease, D. Feraris, M. Aikawa, C. S. Sullivan, S. Perlman, A.R. Fehr, The coronavirus macrodomain is required to prevent PARP-mediated inhibition of virus replication and enhancement of IFN expression, PLOS Pathgens (2019), https://doi.org/10.1371/journal.ppat.1007756. May, https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1007756.
[50] J.G. Rack, D. Perina, I. Ahel, Macrodomains: structure, function, evolution, and catalytic activities, Annu. Rev. Biochem. 2 (85) (2016) 431–454, https://doi.org/10.1146/annurev-biochem-060815-014935.
[51] P. Poltronieri, ADP-Ribosylation Reactions in Animals, plants, and bacteria, Eidtorial, Challenges 8 (2017) 14.
[52] Y.M.O. Alhammad, A.R. Fehr, The viral macrodomain counters host antiviral ADP-ribosylation, Viruses 12 (2020) 384, https://doi.org/10.3390/v12040384, 2020.
[53] D.N. Frick, R.S. Virdi, N. Dahal, N.R. Silvaggi, Molecular basis for ADP-ribose binding to the Mac1 domain of SARS-CoV-2 nsp3, Epub. June, Biochemistry (Am. Chem. Soc.) (2020), https://doi.org/10.1021/acs.biochem.0c00309, 2020.
[54] T.E. Webb, R. Saard, Sequence homology between human PARP14 and the SARS-CoV-2 ADP ribose 1'-phosphatase, Immunol. Lett. 38–39 (2020).