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 119 (2020) 103670

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

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

B. Robson [a,b,*]

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

ARTICLE INFO

Keywords:
2019-nCoV coronavirus
Wuhan seafood market coronavirus
Bioinformatics
Synthetic vaccine
Peptidomimetic
Retroinverso
Q-UEL language
Automatic browser

ABSTRACT

This paper concerns study of the genome of the Wuhan Seafood Market isolate believed to represent the causative agent of the disease COVID-19. This is to find a short section or sections of viral protein sequence suitable for preliminary design proposal for a peptide synthetic vaccine and a peptidomimetic therapeutic, and to explore some design possibilities. The project was originally directed towards a use case for the Q-UEL language and its implementation in a knowledge management and automated inference system for medicine called the BioIngine, but focus here remains mostly on the virus itself. However, using Q-UEL systems to access relevant and emerging literature, and to interact with standard publically available bioinformatics tools on the Internet, did help quickly identify sequences of amino acids that are well conserved across many coronaviruses including 2019-nCoV. KRSFIEDLLFNKV was found to be particularly well conserved in this study and corresponds to the region around one of the known cleavage sites of the SARS virus that are believed to be required for virus activation for cell entry. This sequence motif and surrounding variations formed the basis for proposing a specific synthetic vaccine epitope and peptidomimetic agent. The work can, nonetheless, be described in traditional bioinformatics terms, and readily reproduced by others, albeit with the caveat that new data and research into 2019-nCoV is emerging and evolving at an explosive pace. Preliminary studies using molecular modeling and docking, and in that context the potential value of certain known herbal extracts, are also described.

1. Introduction and brief review of coronaviruses and uses of synthetic peptides

1.1. Coronaviruses

Coronaviruses are viruses of both medical and veterinary importance [1]. They include transmissible gastroenteritis virus (TGEV), porcine epidemic diarrhea virus (PEDV), and the human coronaviruses severe acute respiratory syndrome coronavirus (SARS-CoV) responsible for the epidemic in 2003, and Middle East respiratory syndrome coronavirus (MERS-CoV). In the past ten years, many new coronaviruses have been identified. They infect a wide range of hosts from mammals to birds and closely related coronaviruses have been identified in distantly related animals suggesting recent interspecies jumps [1]. Like influenza viruses that have similar epidemic properties, they are single stranded RNA viruses with a lipoprotein envelope, and enter the host cells by a Class I fusion protein, i.e. one that does not require any other viral surface proteins for fusion. Viral fusion proteins are potential therapeutic and vaccine targets, and the above is indicative of the possibility that the kind of considerations discussed in the present paper can be extended to broader range of viruses.

1.2. Purposes of the present paper

The present series of studies (e.g. Ref. [2]) concerns studies on the genome of the Wuhan seafood market isolate [3] believed responsible for the current pandemic. It is believed that this is the same virus as that also called SARS-CoV-2 or COVID-19 virus after the disease now designated COVID-19. The nomenclature is apparently still not fully settled and the preceding choice of name 2019-nCoV remains popular, so it is used throughout below. See Section 1.3 for a brief account concerning the evolving name of this virus and concerns over strains of it. As

* Ingine Inc., Cleveland, Ohio
E-mail address: barryrobson@ingine.com.

https://doi.org/10.1016/j.compbiomed.2020.103670
Received 2 February 2020; Received in revised form 17 February 2020; Accepted 17 February 2020
Available online 26 February 2020

far as matters relating to findings regarding the genomics of 2019-nCoV and the design of synthetic vaccines, diagnostics and peptidomimetic drugs are concerned, ref [2] may be regarded as an early preprint to the present paper. There is discussion here regarding particular World Wide Web access methodologies used to facilitate the study (See Section 2 below), but the virus remains the priority and further and new work on the virus itself is included. However, although not absolutely essential, using a "robot" autobrowser to surf the internet to capture knowledge and keep up with developments proved very valuable in this case. That is because the situation concerning important scientific data about 2019-nCoV is changing rapidly, as one may expect in an epidemic (now technically a pandemic at least by traditional criteria) of such concern, but seemingly more than in previous epidemics at such an early stage.

Some keys examples of rapid progress, and the timeline of some particularly relevant events in the last phase of this study, are as follows. The above paper [2] was posted on 30th January 2020. The first version of the present paper was submitted to the journal on 2nd February 2020. After the completion of the study and the first version of the present paper, a bat virus with 97.41% identity of the spike protein) that is of particular interest in the present paper (as described in Section 1.4 below). This recent bat coronavirus sequence was entered into Genbank as entry QHR63300.1. In addition, two days before the time of writing the present paper, on the 5th February 2020, entry 6LU7 (DOI: 10.2210/pdb6LU7/pdb) appeared in the protein data bank (PDB), being the three dimensional structure of the crystal structure of the 2019-nCoV main protease that is encoded within the viral genome, in complex with an inhibitor. As may be expected, papers considering modeling of various viral proteins and interaction are appearing almost by the minute. Using the automated surfing methods described in the present paper, examples appearing include a Lawrence Livermore National Laboratory announcement about researchers developing a preliminary set of predictive 3D protein structures of 2019-nCoV to aid research efforts to combat the disease, as discussed in Section 4.1.1. The above protease is a potential target, but not the subject of interest in the present study.

Because at the time of this study and the preparation of this paper it was unclear, and perhaps even discouraged, that the virus of interest should be considered as a strain of SARS (see Section 1.3), much below, and of ref [2] as an early preprint of it, follows the notion of using bat and particularly human SARS as a reference model. This is after ref [2] identified these viruses as closely related to 2019-nCoV. The important assumption is that the molecular mechanism of infection is very similar. With 2019-nCoV now seen as a type or strain of SARS, this assumption seems not only justified but also rather trivial. However, it will still be emphasized to some extent because it is, of course, only the classification nomenclature that has changed, not the molecular relationships between sequences and structures described. Moreover, as discussed in this paper, it is not just 2019-nCoV itself or viruses very similar to it that help design means of combatting the virus, but actually those with more weakly matching genomes that are most informative in these kind of studies, which might be described as "comparative bioinformatics". Indeed, it is conserved surface regions that are of particular interest, and hence, should it emerge that the viral strain or nomenclature being considered is incorrect, the conclusions of a paper like this one would likely remain essentially the same, providing there is still sequence similarity for the surface protein of interest. 2019-nCoV continues to be of great concern due to its high rate of infectivity and protracted time to show symptoms. For that reason analysis is done in a way that relates to the design of diagnostic, vaccines, and therapeutic agents. These are peptides, i.e. molecules composed of a few (say, 10–25) amino acid resides, that relate to actual parts of the amino acid sequence of a whole viral protein: see Section 1.5. These sections are continuous sections; the significance, limitations, and remedies for this are discussed in Discussion and Conclusions Section 5. Potential candidate segments of sequence converge in Results Section 4 to one particularly favored section of sequence of 13 amino acid residues, and modifications of it as "peptide designs". Some observations regarding work identified as relating to smaller rigid compounds as potential drugs are also discussed, along with some preliminary conformational and docking studies.

1.3. The coronavirus genome of interest, variations, and design issues

Despite the above comments, it remains that in the early stages of a pandemic, one needs to ensure that one is addressing the virus described, and indeed be cautious as to what the phrase "the same virus" might mean in practice. In that regard, as well as a matter of literature searching, it has helped neither researchers nor news organizations (and meeting and conference organizers) that the name for the virus and disease has constantly changed over the past few weeks, and indeed days. The WHO decided to prefer the name "2019-nCoV" hours before release of the previous report by the present author in January 2020 [2], although the final decision on the virus' official name still awaited the International Committee on Taxonomy of Viruses. At the time of final writing of this text, the WHO named the *disease* COVID-19, so one can reasonably speak of the "COVID-10" or "Covid-19 virus". The above previous report by the author [2] was early in pointing out that the (Wuhan seafood market isolate) virus surface protein of particular interest below is closely related to that of human and bat SARS, and at the time of rewriting the present paper, the Coronavirus Study Group (CSG) of the International Committee on Taxonomy of Viruses decided that 2019-nCoV was a variant of the coronavirus that caused an outbreak of severe acute respiratory syndrome (SARS) in 2002–03. However the new name SARS-CoV-2 is not universally popular, and the nomenclature 2019-nCoV virus or just 2019-nCoV is used throughout this text. Chinese researchers and government have worked industriously to publish their determination of the early detected Wuhan virus on several web sites. At the time of writing, there were some concerns expressed, often outside the mainstream peer reviewed scientific literature, that what is seen as the current 2019-nCoV pandemic is not due to the same virus as that specified as "Wuhan seafood market pneumonia virus isolate" (e.g. Ref. [3]), i.e. GenBank entry MN908947, which is used here. On January 17th' 2019, MN908947.3 replaced MN908947.2, and probably represents an adequate stable description of the sequence for research into that strain isolate [3]. The entry describes an RNA virus with an RNA sequence of 29033 bases. Despite the above concerns, it is believed that MN908947.3, available at the time of the present study, is essentially the Wuhan seafood market isolate and 2019-nCoV except for any genome changes due to accepted mutations, as is typically the case in the course of a viral epidemic. RNA viruses usually held to have much higher mutation rates than DNA viruses.

The above considerations impact a central theme of this paper. It is parts of the surface of a virus that normally interacts with the host cells for entry and which are vulnerable to the host immune system. Unfortunately, it is also exposed regions on surface proteins that accept mutations more readily, except at important interaction sites. If compounds designed as weapons against the virus are based on sections of amino acid sequence that can readily change, then they will quickly become useless ("escape by mutation"). The strain is thus again important but, as noted above, it is no less the *variations* between virus genomes, taking into account the relationships of the genome of interest with those of both closely related and distantly related viruses that are particularly informative for the present kind of study. This is also important because direct information about the structure, function and action of the proteins encoded by the genome is not, at time of writing, available for 2019-nCoV, and one seeks to make reasonable extrapolations from what is known for other related viruses.

1.4. Coronavirus spike protein as therapeutic target

More specifically focus is on the Class I fusion protein of the coronaviruses which is a glycoprotein known as the spike protein (S) that

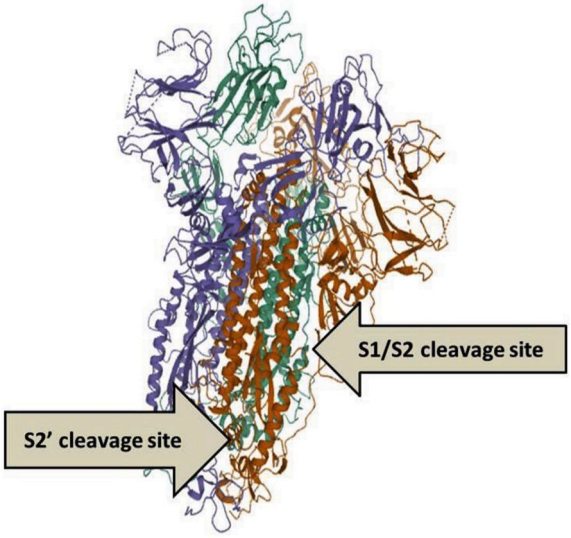


Fig. 1. SARS spike glycoprotein showing approximate position of the two cleavage sites known for that virus. They are assumed to be the same cleavage points with high degree of homology in the Wuhan Seafood Market Coronavirus.

protrudes extensively from the virus envelope surface. It is responsible for binding to the receptor on the host cell as well as mediating the fusion of host and viral membranes [4]. S, most frequently referred to as the "spike protein" or "spike glycoprotein" below, is synthesized as a single-chain precursor of approximately 1300 amino acids and forms a trimer of 3 S proteins on folding. The trimeric SARS coronavirus (SARS-CoV) spike glycoprotein consists of three S1–S2 heterodimers and binds the cellular receptor angiotensin-converting enzyme 2 (ACE2). It mediates fusion of the viral and cellular membranes through a pre-to postfusion conformation transition. Airway protease cleavage site along the amino acid sequence of SARS-CoV S glycoprotein have been identified [5]. Reading from the N-terminus of S1, the important functional elements of the spike including protease cleavage sites as described from SARS studies [5,6] are the S1 N-terminal domain (S1-NTD), the S1 C-terminal domain (S1-CTD), the S1/S2 as the first protease cleavage site as a loop between a pleated sheet and a-helix, the fusion peptide (FP) associated with a highly disordered loop between two a-helices which contains the second cleavage site S2', and a heptad repeat (HR). Fig. 1 shows the approximate positions of the cleavage points in the SARS S1 spike protein within the trimer, superimposed on the Protein data bank entry 5XLR. The heptad repeat has so far attracted most interest as the research lead for a therapeutic, preventative agent. Membrane fusion and cell entry is mediated through extensive post-receptor-binding structural rearrangements in which two heptad repeat (HR) regions play a key role: HR1, positioned C-terminal of a hydrophobic fusion peptide, and HR2 or, located N-terminal of the transmembrane domain. Central to the fusion process is the formation of the stable six-helix bundle in which HR2 regions assemble into the grooves formed by the trimeric HR1 coiled coil.

Some care is needed in deducing mechanisms of activation and cell entry, and hence potential therapies, on the basis of using different types of virus as reference models. This is because there are four main types of coronavirus with differences in spike proteins and it is known that here are some differences in activation and lung cell entry. Renaming the believed causative agent of COVID-19 to indicate it as a type of SARS does not alleviate the concern that different strains use different mechanisms: very few point mutations might change the modes of cell entry. The entry types with represented examples are as follows [4]. Genus *Alphacoronavirus*: human coronavirus NL63 (HCoV-NL63), porcine transmissible gastroenteritis coronavirus (TGEV), porcine epidemic diarrhea coronavirus (PEDV), and porcine respiratory coronavirus (PRCV). Genus *Betacoronavirus*: severe acute respiratory syndrome coronavirus (SARS-CoV), Middle East respiratory syndrome coronavirus (MERS-CoV), bat coronavirus HKU4, mouse hepatitis coronavirus (MHV), bovine coronavirus (BCoV), and human coronavirus OC43. Genus *Gammacoronavirus* avian infectious bronchitis coronavirus (IBV). Genus *Deltacoronavirus*: porcine deltacoronavirus (PdCV).

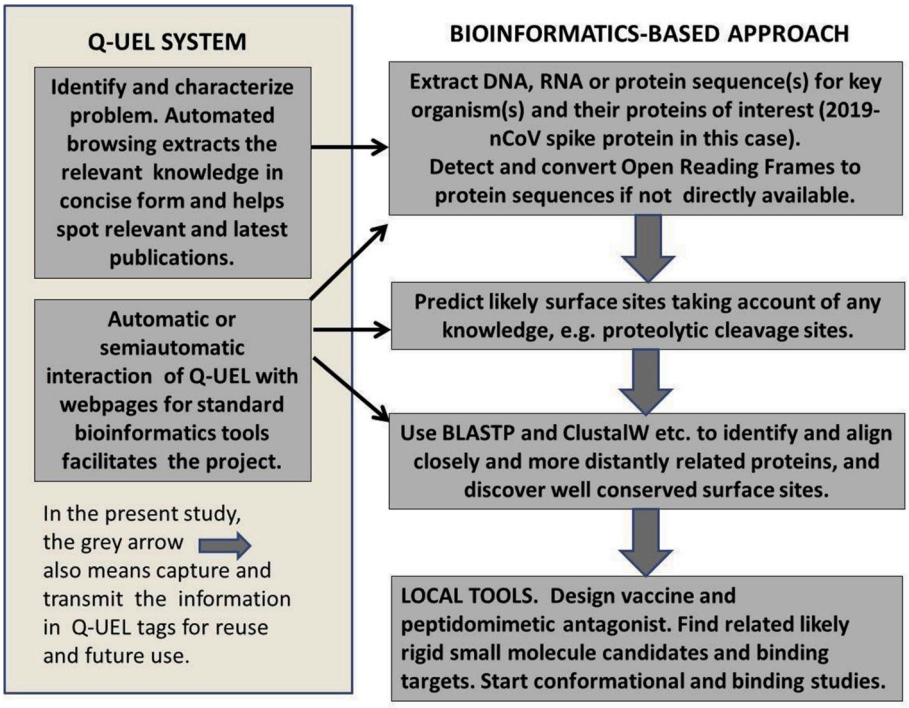


Fig. 2. The Main Workflow in this Study.
The bioinformatics-based approach can also be reproduced by standard tools, but is facilitated by Q-UEL tools, especially with the explosive growth in studies regarding 2019-nCoV.

The alphacoronavirus HCoV-NL63 and the betacoronavirus SARS-CoV recognize a zinc peptidase angiotensin-converting enzyme 2 (ACE2) for target cell binding but alphacoronaviruses such as TGEV, PEDV, and PRCV recognize a zinc peptidase called aminopeptidase N (APN). Also for example, SARS-CoV and betacoronaviruses can recognize different receptors: MERS-CoV and HKU4 recognize a serine peptidase, dipeptidyl peptidase 4 (DPP4). It is known from studies on the SARS coronavirus that activation by a trypsin-like protease naturally present in the lung appears essential to activate the virus for cell binding and entry by cleaving S at two sites [5,6]. The three dimensional structure of the original human SARS spike protein is known by high resolution electron microscopy [7], though again very few point mutations might change the entry mechanism.

1.5. Peptides as the bases of potential therapeutics and vaccines. Previous work

If an Achilles' heel of a coronavirus resides in certain parts of the spike protein as discussed above, the natural choices of preliminary weapons are synthetic peptides (e.g. Refs. [8–10]). This choice reflects a history for synthetic (or otherwise biotechnologically produced) peptides that has developed rapidly since the 1980s, seen as the basis of therapeutics [11,12] and vaccines [11,13]. Related work by other researchers are exemplified by Refs. [8–13], and other efforts are cited at appropriate points below, and in regard to pathogens more generally in Section 5.6. Like many of those efforts, the author and colleagues also had significant success in the past in designing synthetic peptides that at least in some useful way helped combat disease and epidemics. These included some 50 reported studies including HIV that proved diagnostic value in the laboratory (e.g. Refs. [14–16]), and BSE (Bovine spongiform encephalopathy, Mad Cow Disease) [17,18] which led to the early useful diagnostic marketed worldwide by Abbott Laboratories [19]. Other example studies have included immunotherapeutic strategies, and putative vaccines against bacterial pathogens, Ebola and flaviviruses: see Ref. [20] for discussion and a general introductory review of this kind of approach.

These projects resulted in a body of experience and design rules-of-thumb which the author and collaborators are seeking to capture by the methods described in Section 4. It is to be understood that peptides proposed as B-epitopes for the construction of synthetic peptides rarely work well alone (recall that immunogenicity is the ability of a particular substance to provoke an immune response in the body, and that an epitope or antigenic determinant is the small part of an antigen that is recognized and distinguished by the immune system, specifically by antibodies, B cells, or T cells). Ensuring choice of segment of sequence for synthesis and presence of amino acids with at least some reasonable level of immunogenicity is just a small (but important) first step of the vaccine construct.[1] None of such studies by the present author and collaborators have previously addressed a coronavirus in any significant way, but several laboratories skilled in coronavirus science have used peptides based on the heptad repeat HR; they have been considered as antiviral drugs to prohibit virus cell entry, and there has been significant success in the laboratory for some coronaviruses and host species (e.g. Refs. [7,8]). However, to the author's knowledge that has not been a significant success toward a medical product based on the HR. One reason may be that coronaviruses escape from HR2-derived Peptide Entry Inhibition by mutations. This is particularly indicated in the HR1 domain of the spike fusion protein, suggesting that only limited options exist for escape from the inhibitory effect of the HR2 peptide [9,10]. An alternative is the two cleavage sites S1/S2 and S2′ required for activation for cell entry. Since these sites are exposed to the solvent [7] they are plausible targets for several preventative, therapeutic, and diagnostic purposes. The sites may also represent epitopes serving as either the basis for design of a synthetic peptide vaccine, when linked to a carrier protein or a deigned cloned protein. Even if they fail to implant immune memory, they may also form the basis for production antibodies as "sera" for passive immunity and for usein diagnostic kits and biosensors. Importantly there is also the possibility that, as in the case of peptide analogues of the heptad repeats, they might serve as the basis for design for inhibitors of cell entry, not by binding to the coronavirus, but to the trypsin-like proteases of the lung in a preventative manner, prior to virus exposure.

Very early in the above peptide studies by the author and colleagues it was discovered that the bioinformatics and peptide and protein modelling steps required in each investigation were capable of a high degree of automation [21–25]. The PROMETHEUS expert system based on the GLOBAL polymorphic programing language also developed by the author and colleagues [26,27] played a major role in many of the later peptide studies including the BSE diagnostic. Some of the general ideas from a user perspective are used in the present studies, but they employed the more recent Q-UEL language [28]. This language is significantly different from GLOBAL and other languages, being a radical extension of XML to probabilistic semantics with the tags having algebraic force [28], but it had already been seen to provide a useful platform for studies of epidemics and design of peptides [20].

2. Theory

The theory behind Q-UEL has been described and developed in several essentially mathematical papers. Refs [28–36] provide a more relevant and applied mathematical account, in conjunction with algorithm and software development. Ref [28] provided the first detailed description of the Q-UEL language as a means of interacting with the World Wide Web and potentially proving the basis of a "Thinking Web" for medicine, primarily by rendering the emerging Semantic Web as more fundamentally probabilistic [28]. The underlying theory of probabilistic semantics is somewhat irrelevant in the present case because there is no great need to consider probabilities associated with elements of knowledge about the bioinformatics of coronaviruses that Q-UEL's XML-like tags express. Note, however, that probabilities do become more relevant in clinical genomics involving patient DNA sequences [36], and a similar case can be made that homologies as extents of match between viral protein sequences, and uses of automated probabilistic reasoning, will be useful in the future.

3. Methods

3.1. Q-UEL tools

Q-UEL is in part concerned with automation of computer and Internet-centric research tasks. Mainly, in the present study this involved automatic surfing of the World Wide Web to speed access to information about coronaviruses, and to capture that in canonical form, i.e. as Q-UE's XML-like "tags" as statements of knowledge. A major

[1] In brief summary of the previous work above and current conventional thinking, the peptides seen as epitopes are generally envisaged as needing to be attached to immunogenic proteins safe to the patent or receiving test organism, but not so familiar to the receiving immune systems as to cause significant cross reactions (e.g. inducing autoimmune disease). Also while B epitopes from exposed regions of a pathogen are key to the antibody response, T epitopes from anywhere in the pathogen proteins may be needed for T-system memory and cellular responses, and the overall structure may benefit from linkage to molecular adjuvants such as muramyl dipeptide, and possibly to other stimulating and anti-suppressor peptides. In short, the peptide of initial interest may only be a small part of what would look like a complex molecular 'Swiss Army knife". Nonetheless, these additions are to varying extents already understood support technologies (even if not as yet favored for human use – see discussion later below). The various components can be seen as interchangeable and/or "reusable" parts, the essence of a "cartridge" approach to synthetic vaccines intended to facilitate a quicker response to new infectious diseases. Among them the B-epitope from some part of the virus surface remains key to immune system recognition.

Table 1
One letter amino acid codes used in the text.

One letter code	Amino acid	Conservative replacements
A	alanine	A, E, S, T
C	cysteine/cystine	S, T, V
D	aspartic acid	E
E	glutamic acid	A, D
F	phenylalanine	M, W, Y
G	glycine	N, P
H	histidine	K, R
I	isoleucine	L, V
K	Lysine	H, R
L	leucine	I, V
M	methionine	F, W, Y
N	asparagine	G, D, Q
P	proline	G
Q	glutamine	N, E
R	arginine	H, K
S	serine	A, T
T	threonine	A, I, S
V	valine	A, I, L
W	tryptophan	F, M, Y
Y	tyrosine	F, M, W

difference from previous genomics use cases for Q-UEL [36] is that the notion of types of coronavirus and strain replaced the notion of human patient. The specific software of the decision support system interacting with the Internet was the more recent BioIngine implementation of Q-UEL and associated software applications [29–35]. These applications generate and use Q-UEL tags to extract, communicate knowledge and draw conclusions from it by automated inference. The interaction with source text and bioinformatics tools on the World Wide Web is in keeping with the originally intended spirit of facilitating a "Thinking Web" for the Internet [28]. However, the Q-UEL tags here served more in the role of an architectural principle. See Fig. 2. Most of the "Bioinformatics Based Approach" on the right hand side of Fig. 2 could be carried out in the usual way of interacting manually with the web sites, and sometimes were. Q-UEL there has the role of facilitating but not of being essential for the study, except in the sense that this paper may, arguably, may have been written very differently and much later in the face of explosive development of worldwide research into 2109-nCoV.

3.2. Expertise capture

One of several motivations [36] for developing an integrated genomics and bioinformatics approach based on Q-UEL was that the popular highly integrated Biology Workbench at the San Diego Supercomputer Center is no longer available at the time of writing [37]. However, some standard web tools used prominently in present work on the coronavirus spike sequence are already in a suite that at least integrates aspects concerned with a particular tool, and the web pages are sufficiently easy to use in the normal manual way. Nonetheless, an automated approach is valuable for speedier research, and for capturing knowledge from the Internet in Q-UEL tags. As well as capturing expertise of the human user regarding the workflow, it leads to full reproducibility. For the most part, output from BLASTP as used at https://blast.ncbi.nlm.nih.gov/Blast.cgi [38] is still expressed below in that form familiar to bioinformatics researchers, rather than as it appeared in Q-UEL tags. Also see Ref. [36] for the formats for Q-UEL representation of bioinformatics information of this kind.

3.3. Use of MARPLE/HDNstudent with XTRACTOR

Previously, a number of ways of gathering information from the Internet into Q-UEL tag form have been explored. For example, a Q-UEL-CTRACT tag in Ref. [28] was generated by interacting with the HTML on the webpage for Google queries and the resulting list of hits. In the present paper, extensive use was made of MARPLE/HDNstudent with XTRACTOR [30,31]. This provides convenient ways of making search queries for initiating "autosurfing" of the Internet. In the present case, there were some 170,108 Q-UEL tags already representing statements of medical knowledge in the KRS (albeit several tags have been found that essentially duplicate the same knowledge). However, these are not as yet exhaustive of medical knowledge and focused on a number of specialized medical areas. While many related to viral infections, there was no detail at the time of the study regarding 2019-nCoV structure. Hence the gathering of new knowledge from the Internet was important. In the course of the recent phase of study, a site particularly frequently "hit" was https://en.wikipedia.org/wiki/Novel_coronavirus_(2019-nCoV). Its content changed significantly since first appearance. It remains that a good secondary "scientific news" source concerning the course of the epidemic is that Wikipedia entry because it appears to regularly updated.

3.4. Amino acid residue conventions and conservative replacements

Table 1 shows the one letter amino acid code used for peptide and protein primary structure throughout the study and in the following text, and also shows a consensus over conservative substitutions in the course of evolution (as accepted mutations), with common secondary structure propensities, and as alternatives in peptide design, i.e. amino acid residues with similar properties [11]. With a focus on design, the relationships are not intended to be symmetric, e.g. starting from E gives one more option than going from D to E, and actual choices in simulation studies can change depending on details of interactions.

4. Results

4.1. Brief comments of use of Q-UEL and related Software.Applications

XTRACTOR was accessed in several ways, and one generated tags such as the following which describes how Lawrence Livermore National Laboratory (LLNL) researchers "have developed a preliminary set of predictive 3D protein structures of the virus to aid research efforts to combat the disease". This and the following example are new tag specifications developed for the present project.

```
< Q-UEL-XTRACTOR-WEBPAGE:= tagtime(gmt):= 'Sat Feb 8 6:22:48 2020'
protein structure prediction coronavirus
 | 'was found at' |
URL:=https://www.llnl.gov/news/lawrence-livermore-researchers-release-3d-protein-structure-predictions-
novel-coronavirus
Q-UEL-XTRACTOR-WEBPAGE>
```

Note that the query is "protein structure prediction coronavirus" but the words are also legally "nominal categorical" data types and legal attributes in the above tag construction. Another example with a status report was as follows.

< Q-UEL-XTRACTOR-WEBPAGE:= tagtime(gmt):= 'Mon Feb 10 14:59:05 2020'
2019-nCoV spike protein
| 'was found at' |
URL:=https://www.biorxiv.org/content/10.1101/2020.01.30.927871v1
status:=('reprint detected', 'withdraw detected')
Q-UEL-XTRACTOR-WEBPAGE>

In contrast, XTRACTOR as it is used by MARPLE [36] has the big advantage of capturing, parsing, and standardizing content. In the present case, it was essentially repurposed. A number of examination-like questions were set, with the "question "replaced by a brief project description and "answers" replaced by queries, e.g. Case 1. Investigation into the Wuhan seafood market coronavirus, bioinformatics analysis of the spike protein. Answers. (A) Wuhan coronavirus, (B) SARS, (C) Coronaviruses, (D) Spike protein, (E) Lung diseases. The final weighting of the candidate "answers" was similar in this case at 24% for (A) and 19% for the rest and simply suggests that the process of reasoning found each query highly relevant to the "question" part. In its process of scoring the World Wide Web for answers to the queries, text is captured, parsed and reorganized into XTRACT tags that for future use in automated inference are readily decomposed into semantic triples and linear semantic multiples [36]. However, in the present project they are primarily of value for collecting knowledge for review by the researcher, as part of an automated systematic review process [33], and to time stamp the tags along with provenance information. This is very useful here because research and even these sites, are changing. Some examples are as follows. Note that the contraction into basic relationships, with reparsing and annotation into canonical form to make use easy for computers, sometimes gives a rather stilted "bad English" presentation. Automatic extraction of relevant text prior to this processing, with references, is also accessible for easy human reading.

4.2. *Simple epitope prediction combined with deduction of S1/S2 and S2' sites in 2019-nCoV*

The following shows the amino acid sequence (in standard IUPAC one letter code: see Table 1) for 2019-nCoV spike glycoprotein QHD43416.1, deduced from the Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome, GenBank: MN908947.3. Technically speaking, the existence and nature of the spike (S) glycoprotein of 2019-nCoV is theoretical, being the computed translation of a predicted ORF (open reading frame), as far as in known to the author at the present time. However, comparisons with SARS would make any alternative interpretations unlikely even prior to the recent naming as a SARS type. The very recent close match with a SARS virus entry in the databases makes the assumptions used in the present paper much more plausible, however. As described in Ref. [36] a Q-UEL tag obtaining an ORF (open reading frame) from GenBank is only slightly different from that in Ref. [36] which was a specification suited to human mitochondrial genomics.

<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>

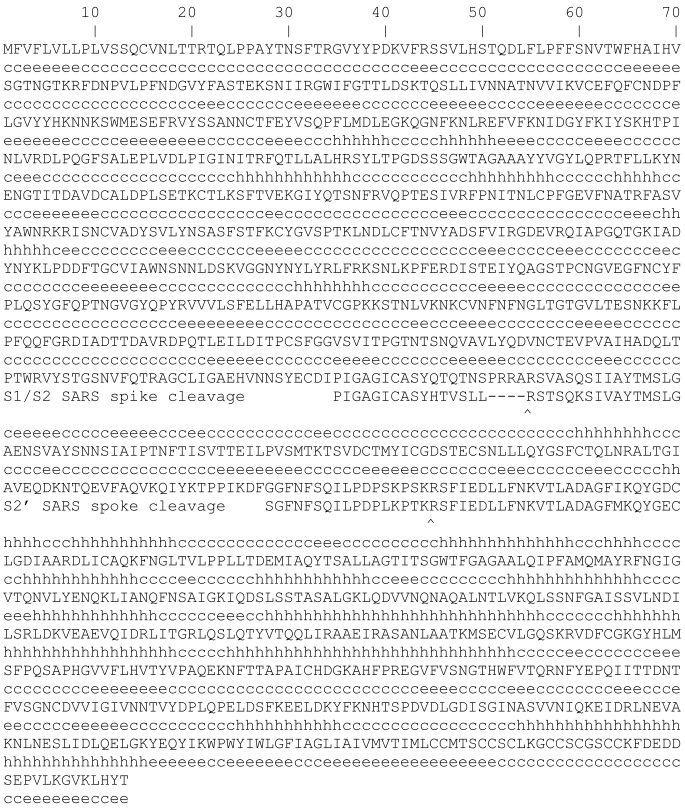
<Q-UEL-Marple41W2 "`the spike |as| S |of| envelope |as| E |of| membrane |as| M {AND} nucleocapsid |as| N |In| `the specific case |of| `the SARS coronavirus | as ^see below| [0https://en.wikipedia.org/wiki/Coronavirus#Severe_acute_respiratory_syndrome] |^defined ^receptor-binding| domain |on| S mediates |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]"
| 'was extracted from' |
source:='https://en.wikipedia.org/wiki/Coronavirus' time:='Fri Jan 31 17:36:06 2020' extract:=453 Q-UEL-Marple41W2>

```
<Q-UEL-ORF-PROTEIN:=(application:='Perl version v5.16.3':= 'GenBank query',
tagtime(gmt):= 'Sun Feb  2 10:58:29 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:=( '21563..25384:= 'S', 'codon start':=1)
product:= 'surface glycoprotein'
'protein id':=QHD43416.1.'
| is |
sequence:= 'IUPAC 1 letter aa code':=
'MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVSGTNGTKRFDNPVLP
FNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYV
SQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTPGDSSSG
WTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFN
ATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDF
TGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSF
ELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPG
TNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSV
ASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVE
QDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLT
VLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASAL
GKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSE
CVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITT
DNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQEL
GKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT'

'size class':='peptide-protein (21-35 aa)':=35
Q-UEL-ORF-PROTEIN>
```

There is a similar Q-UEL-ORF-AA-2RY-STRUCTURE tag that also carries a secondary structure prediction, usually in terms of h (α-helix), e (extended chain, β-pleated sheet) or c (coil or loop) as the conformational state of each amino acid residue. Such tags, along with QUEL-ORF-AA-CONSENSUS-SEQUENCE tags, help ratify that the translation is plausibly a real one and give insight into structure and function. Consensus tags show sequences of amino acid residues in the translated ORF that are usually associated with posttranslational modifications, binding and catalytic sites etc.,. For example, in the above putative 2019-nCoV S protein there are 20 occurrences of N (asparagine) followed by a residue that is not P (proline) which is followed in turn by S (serine) or T (threonine) at which glycosylation is likely to occur, followed by a residue that is not P (proline). This is typical of an extracellular protein. The secondary structure prediction extracted is shown below along with the known cleavage sites from the SARS spike protein are shown short sections of SARS sequence at which cleavage occurs, with the arginine R at which the cut occurs in indicated by the character ^. Underneath is shown (as strings of characters, h, e, c) which are the secondary structure prediction using the GOR method [39], in Version IV available at the Rhone-Alpes Protein Institute of Biology and Chemistry website [40]. There is not a Q-UEL tag as yet that combines such cleavage site sequences with a larger amino acid sequence, although alignment tags and similar BLAST result tags can be used to play a similar role, and if cleavage site rules for a protease are included in the dictionary of consensus sequence, they would at least show the amino acid residue at the point of the cut (e.g. K lysine or R arginine in the case of trypsin) or its immediate vicinity.

The purpose is here was not to predict secondary structure in the context of the folded protein structure (the "native conformation") which known experimentally for SARS [4,7] and likely to be similar to the 2019-nCoV glycoprotein. Indeed, secondary structure could ultimately be deduced that way in fairly high resolution, e.g. terms of dihedral angles expressed in terms of 9 x 9 ranges (e.g. A38G18T38T29…) [11], although doubtless there will then ultimately be found to be significant differences between 2019-nCoV and SARS at that level of resolution, especially considering that the difference in sequence is significant. Rather the purpose is otherwise, and twofold. The first purpose is simply based on the empirical finding of the author that residues predicting as coil or loop c tend to form a good basis for peptides chosen in the design of diagnostics and vaccines (see discussion later below), irrespective of whether the prediction is correct and irrespective of rules for water-liking character, exposure, or intrinsic immunogenicity (discussed later below), albeit that they naturally overlap. The second is to give the peptide designer some idea of the likely most populated conformation of residues as h, e, and c (and for some more detailed kinds of prediction, types of helix, sheet, turn or bend, and so on). The latter is in regard to the conformational behavior of peptides that are copied or modeled from local regions of the sequence and are of some 20 to 25 amino acid residues in length or less. When synthesized as peptides, they likely exist in solution as an open dynamic, random configuration that may be envisaged a statistical mechanical average over many overall conformations of similar energy, corresponding more to the above prediction. The prediction used here deliberately does not include the strong influence on and perturbation of intrinsically preferred residue conformation that occurs when the amino acid residue must fit into the rest of the folded structure, i.e. the so-called tertiary interactions. GOR method predictions that also make use of alignment of sequences from many homologous proteins were not used, because they are deliberately designed to be influenced by the conformational preferences of residues in compact folded motifs. Similarly the intrinsic Q-UEL learning approach applied to predicting secondary structure [34] was not used. Fig. 2 in that paper illustrates the point. It shows that after *circa* 150 arbitrarily selected proteins are brought into the training set., the predictions begin to change and improve dramatically as more and more homologous proteins of similar folded structure are encountered and the coverage of sampled proteins spans the commoner protein families Obvious pre-known homologues were discarded in that study [34], but the above concern still applies.



These two sites are prominent but not alone amongst the cleavage points that researchers have discovered in the SARS spike protein. MARPLE with XTRACTOR was able to extract from the World Wide Web a number of sequences that were cleaved by various proteases, including refs [5, 6]. Ref [5] also noted 11 potential sites for trypsin, plasmin and Transmembrane protease TMPRSS11a. Others extracted from Websites included also the following

Trypsin: S1/S2 HTVSLLRSTSQKSIVAYTMSL, S2' LPDPLKPTKRSFIEDLLFNKV
Cathepsin: S1/S2 HTVSLLRSTSQKSIVAYTMSL
Elastase: S2' LPDPLKPTKRSFIEDLLFNKV
Plasmin: S1/S2 HTVSLLRSTSQKSIVAYTMSL, S2' LPDPLKPTKRSFIEDLLFNKV
Transmembrane protease/serine TMPRSS subfamily:
TMPRSS1: S1/S2 HTVSLLRSTSQKSIVAYTMSL,
TMPRSS2: Multiple sites
TMPRSS11a: S1/S2 HTVSLLRSTSQKSIVAYTMSL, S2' LPDPLKPTKRSFIEDLLFNKV

In general the sequence in the S2′ region LKPTKRSFIEDLLFNKVTLADAGFMK was already looking of particular interest because of the variety of proteins that can cleave within those regions and potentially activate the virus for cell entry, including trypsin, elastase, plasmin, TMPRSS1, TMPRSS2, and TMPR11a. There are diverse types of proteases present in the human lung include various serine, cysteinyl, aspartyl and metalloproteases. These can for the most part function extracellularly as well intracellularly to perform a variety of normal functions such as neutrophil chemotaxis, tissue remodeling, mucin expression, and not least viral and bacterial destruction, a function which could *a priori* be involved in coronavirus spike activation for any of the above proteins, albeit likely to different extents. However, the above region is by no means the only region in which proteolysis could occur, as exemplified above.

4.3. S protein homologies and variations in the S1/S2 and S2' sites

As of 1/29/2020, the closest homology (sequence match) of the whole 2019-nCoV sequence (other than with its own entry at 100%) was 81% with the spike protein of the Bat SARS-like coronavirus, sequence ID GenBank entry AVP78031.1, with 100% coverage. Sequences of the same name as entry that matched closely were AVP78042.1 at 80.32% and ATO98205.1 at 77.07%. All the top matches are bat host species. Three matches at 83–84% were noted in early studies but they were not listed as matches recently (the reason is unclear). The remaining closely matching sequences out of the top 100 have coverage 98%–100% and matches are in the range 77.04%–77.38%. That is however somewhat deceptive. For example, by searching specifically for spike glycoprotein sequences related to host as pig *Sus scrofa* one obtains a variety of matches with only *circa* 30%–31% homology and 56% coverage which is typically regarded as very weak. A common rule of thumb is that more than 30% identity over the entire lengths is required for significance, because higher identities are seen by chance in short alignments. However, as is not infrequently the case in homology studies, the assumption of meaningful homology and alignment is strengthened by known common function and suspected common ancestry of the proteins and by persuasive matches of sections, often of known functional significance.

As might be expected from the above considerations, searching not

on the whole sequence but only on the two sections around the two arginine cleavage sites reveals a broader variety of coronaviruses with host species, human, bat, pig, and civet. This group of species has a relatively close set of matches in the regions of S1/S2 and S2′ sites and this is of interest because it illustrates the kind of acceptable mutations, and in synthetic peptides, substitutions, of single amino acid residues that can occur near the cleavage points, i.e. the kinds of variations and alternatives that may be of interest in peptide design uncluttered by too many variations. All of human, bat, pig, and civet host species were encountered when searching with the S1/S2 or "PIGAG"-containing motif sequence, called PIGAG for short. That is, it matches with coronavirus isolates from human, bat, pig, and civet. Some arbitrary examples selected to show patterns in variation are as follows. These are summaries of alignments made as discussed in Section 3.4. While showing significant variation in the case of pig, becoming a PLGDG motif, the essential sequence features and the 2019-nCoV cleaved arginine R is still retained.

```
PIGAGICASYQTQTNSPRRARSVASQ-SIIAYTMSLG  - 2019-nCoV  spike glycoprotein
PIGAGICASYHTVSLL----RSTSQK-SIVAYTMSLG  - SARS-CoV spike glycoprotein
PIGAGICASYHTASVL----RSTGQ-KSIVAYTMSLI  - [recombinant CoV] ID:ACJ60703.1
PIGAGICAKYGISSNT--- RLRSNSQSIVAYTMSLG  - [SARS-like CoV BatCoV/BB9904]
PIGAGICASYHTASTL----RSVGQK-SIVAYTMSLG  - [Bat CoV] ID:ARI44799.1
PIGAGICASYHTVSSL----RSTSQK-SIVAYTMSLG  - [SARS CoV civet014] ID:AAU04661.1
PIGAGICASYHTASIL----RSTSQK-AIVAYTMSLG  - [Bat SARS-like CoV] ID:AVP78031.1
PIGAGICASYHTVSLL----RSTSQK-SIVAYTMSLG  - [SARS CoV BJ302] ID: AAR07628.1
PIGAGICASYHTASLL----RSTGQK-SIVAYTMSLG  - [BtRf-BetaCoV/HeN2013] ID:AIA62339.
PLGAGFCADLGRGVSV--R-RIAFER-HDTTYVAPVI  - [bat CoV HKU2-related ID:QHA24703.1
PLGDGFCADLGSGVVV--R-RMTFEK-HDTTYVAPVT  - [Swine diarrhea CoV] ID:AVM80484.1
PLGDGFCADLLGNVAV--R-RMTFEK-HDTTYVAPVT  - [Swine enteric alphaCoV] ID:QEH62669.1
PLGDGFCADLKGTVAV--R-RMGFEA-HDTTYVANVI  - [BtRf-AlphaCoV/YN2012] ID:YP_009200735.1
```

The S2′ or "FIEDLL"-containing motif sequence, called FIEDLL for short, also matches with coronavirus isolates from human, bat and pig. Examples are as follows. While showing significant variation in the case of pig, the KRSFIEDLL sequence is highly conserved and hence the cleaved arginine R is still retained.

```
FGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC  - 2019-nCoV CoV spike glycoprotein
FGGFNFSQILPDPLKPTKRSFIEDLLFNKVTLADAGFMKQYGEC  - [Bat SARS-like CoV] ID:ATO98218.1
FGGFNFSQILPDPLKPTKRSFIEDLLFNKVTLADAGFMKQYGEC  - [SARS CoV BJ302] ID: AAR07628.1
FGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC  - [BtRf-BetaCoV/HeN2013] ID:AIA62339.1
YSGFNFSQILPDPLKPTKRSFIEDLLFNKVTLADAGFMKQYGEC  - SARS-CoV spike glycoprotein
YSGFNFSQILPDPLKPTKRSFIEDLLFNKVTLADAGFMKQYGEC  - [recombinant  CoV] ID:ACJ60703.1
YSGFNFSQILPDPLKPTKRSFIEDLLFNKVTLADAGFMKQYGEC  - [SARS CoV BJ182-] IDACB69883.1
FGGFNFSQILPDPLKPTKRSFIEDLLFNKVTLADAGFMKQYGEC  - [SARS CoV C028]ID:AAV98001.1
FGGFNFSQILPDPSKPTKRSFIEDLLFNKVTLADAGFMKQYGDC  - [Bat SARS HKU3-13] ID:ADE34823.1
FGGFNFSQILPDPLKPTKRSFIEDLLYNKVTLADAGFMKQYADC  - [BtRf-BetaCoV/HeN2013]ID:AIA62339.1
VLGVSVYDPASGRVVQ-KRSFIEDLLFNKVVTNGLGTVDEDYKRC - [Pig epi. diarrhea v.] ID:ALB35880.1
VLGVSVYDPASGRVVQ-KRSFIEDLLFNKVVTNGLGTVDEDYKRC - [spike Sus scofa] ID: QGV12780.1
VLGVSVYDPASGRVVQ-KRSFIEDLLFNKVVTNGLGTVDEDYKRC - [Sus scrofa]ID:QGV12779.1
```

Note that the section KRSFIEDLLFNKV is highly conserved overall apart from a tyrosine (Y) replacement for the second phenylalanine (F) in one case, and this pattern is seen over a broader set of alignments. Again, isolates of coronavirus from pig were a notable distinct but related group. VLGVSVYDPASGRVVQ and FNKVVTNGLGTVDEDYKRC around (bounding) KRSFIEDLL are sufficient markers by close match for identifying pig diarrhea epidemic coronaviruses. Amongst any spike glycoprotein sequences examined, NVL was a common triplet occurring triplet often occurring twice in many other spike glycoproteins, e.g., bat [BtRf-BetaCoV/HeN2013] AIA62339.1, which also have the FIEDLL motif, but unless it shortly preceded the FIEDLL motif as in NVLGVSVYDPAS GRVVQKRSFIEDLLNKVVT [Pig epi. diarrhea v.] ID: ALB35880.1 the internal match with the FIEDLL motif is not significant and the NVL and following sequence usually carry no cleavage arginine R at the corresponding point.

4.4. Absence of a clear "PIGAG" (S1/S2 related) site in some coronaviruses

A variety of standard bioinformatics formats have been explored for Q-UEL attributes that express alignment of biosequences, finally settling on a few user-selected options that are most popular with the author and his collaborators [36]. It is a "movable attribute" (meaning that it can optionally appear in many kinds of tags) because in genomics, sequences are usually compared with a universally agreed reference standard in order to help the original alignment, detect possible errors, and highlight variations. In the present study, they often appeared in Q-UEL-ORF--PROTEIN tags concerning an ORF for the spike protein of a coronavirus that is *not* 2019-nCoV and the sequence that is analogous to the reference sequence in Ref. [36] *is* 2019-nCoV. Evidently, this choice of tag vehicle is somewhat arbitrary and the "reference standard" is only appropriate for a specific virus project of the present kind. Also, use of BLASTP [38] dominated the present project (in contrast with the previous genomic study that also made extensive use of alignment technology related to CLUSTALW [36]). Pairwise alignments are already naturally generated by BLASTP, so that format was used. It comprises use of blanks for non-matches and a plus '+' for a conservative difference, i.e. the amino acids matching are not identical, but a similar type of amino acid. The default notion of "similar" in is not, however, the same as in Table 1 above. An example attribute concerning a section of sequence as follows (see Ref. [36] for extension to span a whole gene or protein).

```
Alignment:=(process:=BLASTP:=https://blast.ncbi.nlm.nih.gov/Blast.cgi:=
(
'AIA62339-S',642-1003,          'ICASYHTAS----LLRSTGQKSIVAYTMSL'
'BLASTP match code',973-1003,   'ICASY T +      RS   +SI+AYTMSL'
'MN908947-S ',973-1003,         'ICASYQTQTNSPRRARSVASQSIIAYTMSL'
)
```

Below, however, matches are reported in the more familiar output from BLASTP as it actually appeared at https://blast.ncbi.nlm.nih.gov/Blast.cgi, with the "Query" being usually MN908947 S protein (note, now at the top), and the 'Subjct' being one of the proteins found to match with it (note, now at the bottom).

```
Query  521   PATVCGPKKSTNLVKNKCVNFNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQ  580
             P  + G    + +  + C  +   G  G GV+T +N  FL    +  D      A ++
Sbjct  632   PKPLEGVTDVSFMTLDVCTKYTIYGFKGEGVITLTNSSFLAGVYYTSDSGQLL-AFKNVT  690

Query  581   TLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGS  640
             +  +  +TPCSF            S Q A +  D+      V + +   + T+
Sbjct  691   SGAVYSVTPCSF------------SEQAAYVDDDI------VGVISSLSSSTFNSTRELP  732

Query  641   NVFQTRAGCLIGAEHVNNSYECDIPI----GAGICASYQTQTNSPRRARSVASQSIIAYT  696
               F           H N+   C  P+      G+C S      S     S + Q  IA T
Sbjct  733   GFFY----------HSNDGSNCTEPVLVYSNIGVCKS-----GSIGYVPSQSGQVKIAPT  777

Query  697   MSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQY  756
             ++            +I+IPTNF++S+ TE L +  T  SVDC  Y+C  ++ C  LL QY
Sbjct  778   VT-----------GNISIPTNFSMSIRTEYLQLYNTPVSVDCATYVCNGNSRCKQLLTQY  826

Query  757   GSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGG--FNFSQILP----DPS  810
              + C  +  AL   A  +      +    ++  +   I  F G  +NF+ +L     DP+
Sbjct  827   TAACKTIESALQLSARLESVEVNSMLTISEEALQLATISSFNGDGYNFTNVLGVSVYDPA  886

Query  811   KP---SKRSFIEDLLFNKVTLADAGFIKQ-YGDCLGDIAARDLICAQKFNGLTVLPPLLT  866
                    KRSFIEDLLFNKV     G + + Y  C     +  DL+CAQ ++G+ VLP ++
Sbjct  887   SGRVVQKRSFIEDLLFNKVVTNGLGTVDEDYKRCSNGRSVADLVCAQYYSGVMVLPGVVD  946

Query  867   DEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQ  926
              E +  Y+++L+ G +  G+T    AA  +PF+  +  R N + +  +VL  NQ+L+A
Sbjct  947   AEKLHMYSASLVGGMVLGGFT----AAAALPFSYAVQARLNYLALQTDVLQRNQQLLAES  1002

Query  927   FNSAIGKIQDS--------------LSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGA  972
             FNSAIG I  +               L++ A AL K+Q+VVN    AL  L  QL  NF A
Sbjct  1003  FNSAIGNITSAFESVKEAISQTSKGLNTVAHALTKVQEVVNSQGAALTQLTVQLQHNFQA  1062

Query  973   ISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSEC  1032
             ISS ++DI SRLD + A+VQ+DRLITGRL +L  +V Q L +  E++AS  LA  K++EC
Sbjct  1063  ISSSIDDIYSRLDILSADVQVDRLITGRLSALNAFVAQTLTKYTEVQASRKLAQQKVNEC  1122

Query  1033  VLGQSKRVDFC-GKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAIC-HDGKAHFP  1090
             V  QS+R  FC G G H+ S  Q+AP G++FLH   VP     +      +C +D  A
Sbjct  1123  VKSQSQRYGFCGGDGEHIFSLVQAAPQGLLFLHTVLVPGDFVDVIAIAGLCVNDEIALTL  1182

Query  1091  REGVFV--------SNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVV-IGIVNNTVYDPL  1141
             RE   V           T +FV+ R  +EP+  T  +     +C V  + +  + + D +
Sbjct  1183  REHGLVLFTHELQNHTATEYFVSSRRMFEPRKPTVSDFVQIESCVVTYVNLTRDQLPDVI  1242

Query  1142  QPELDSFK--EELDKYFKNHTSPDVDL-----------GDISGINASVVNIQKEIDRLNE  1188
                +D  K  +E+     N T P + L           G+I+ +     +++   + L
Sbjct  1243  PDYIDVNKTLDEILASLPNRTGPSLPLDVFNATYLNLTGEIADLEQRSESLRNTTEELQS  1302

Query  1189  VAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIM  1233
             +  N+N +L+DL+ L + E YIKWPW++WL     LI +V + +
Sbjct  1303  LIYNINNTLVDLEWLNRVETYIKWPWWVWLIIFIVLIFVVSLLVF  1347
```

As noted above, by searching specifically for spike glycoprotein sequences related to host as pig *Sus scrofa* one obtains many matches with only *circa* 30%–31% homology and 56% coverage with weak but persuasive matches of sections. The interest point alluded to above is that while the so-called FIEDLL motif clearly evident and retains the cleavage arginine R, though the PIGAG motif is essentially missing, here more correctly meaning that it is highly modified with the cleavage arginine R missing. The alignment in the region of the two cleavage sites is as follows, with the extent of the cleavage segments as mostly discussed above underlined. "Query" indicates the 2019-nCoV sequence and the Sbjct a pig virus sequence such as GenBank AZL47249.1 also typical of a large number of pig diarrhea coronaviruses differing by just one or two amino acid residue differences in the spike protein, and which did not align well particularly on the N-terminal side of the FIEDL motif.

Compare for example variations in the "PIGAG" region of S glycoprotein [SARS coronavirus BJ302] ID: AAR07628 with respect to the 2019-nCoV spike glycoprotein where there are significant variations in the vicinity but in which the PIGAG motif is preserved, and the cleavage arginine is found.

```
Query  564   QFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVA  623
             QFGRD++D TD+VRDP+T EILDI+P SFGGVSVITPGTN S++VAVLYQDVNCT+V  A
Sbjct  550   QFGRDVSDFTDSVRDPKTSEILDISPRSFGGVSVITPGTNASSEVAVLYQDVNCTDVSTA  609

Query  624   IHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRR  683
             IHADQLTP WR+YSTG+NVFQT+AGCLIGAEHV+ SYECDIPIGAGICASY T +
Sbjct  610   IHADQLTPAWRIYSTGNNVFQTQAGCLIGAEHVDTSYECDIPIGAGICASYHTVS----L  665

Query  684   ARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYIC  743
              RS + +SI+AYTMSLGA++S+AYSNN+IAIPTNF+IS+TTE++PVSM KTSVDC MYIC
Sbjct  666   LRSTSQKSIVAYTMSLGADSSIAYSNNTIAIPTNFSISITTEVMPVSMAKTSVDCNMYIC  725
```

and similarly in glycoprotein [BtRf-BetaCoV/HeN2013] Sequence ID: AIA62339.

```
Query  550   GVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVA  609
             GVLT+S+K F  FQQFGRD +D TD+VRDPQTL ILDI+PCSFGGVSVITPGTNTS+ VA
Sbjct  522   GVLTDSSKTFQSFQQFGRDASDFTDSVRDPQTLRILDISPCSFGGVSVITPGTNTSSAVA  581

Query  610   VLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAG  669
             VLYQDVNCT+VP  +HADQL P+WRVY+TGS VFQT+AGCLIGAEHVN SY+CDIPIGAG
Sbjct  582   VLYQDVNCTDVPTTLHADQLAPSWRVYTTGSYVFQTQAGCLIGAEHVNASYQCDIPIGAG  641

Query  670   ICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPV  729
             ICASY T +      RS   +SI+AYTMSLGAENSVAY+NNSIAIPTNF+ISVTTE++PV
Sbjct  642   ICASYHTAS----LLRSTGQKSIVAYTMSLGAENSVAYANNSIAIPTNFSISVTTEVMPV  697
```

In consequence of the loss of the "PIGAG" site and cleavage arginine in some cases, the focus is on the S2' section DPSKPSK**R**SFIEDLLFNKV (the arginine constituting the cleavage point is underlined). The "PIGAG" site would not be so suitable for designs based on the assumption of binding to a protease, and in general would seem to promise an increased ability for the virus to escape, by mutation, any successful vaccine or therapeutic based on that region. Admittedly, the vicinity of the "FIEDLL" site can show great variation, but KRSFIEDLLFNKV tends to be well preserved, e.g. as in alignment with Pig S protein ID:QGV12784.1.

```
Query  807   P----DPSKPS-----KRSFIEDLLFNKV-TLADAGFI-KQYGDCL-G
                DP   S     KRSFIEDLLFNKV T    G     Y  C
Sbjct  879   GVSVYDPA--SDRVVQKRSFIEDLLFNKVVT-NGLGTVDEDYKRCSNG
```

4.5. Design of synthetic peptides as a basis for diagnostics and vaccines

Peptides that mimic part of a protein of protein sequence in terms of amino acid sequence and, by intent, the biological function or an inhibitor of it, are often referred to as *peptidomimetics*. Some authors use the term more specifically for analogues that are not dependent on an immune response, but act as an antagonist to, in this case, binding or cleavage. From the point of synthesis of a peptide as a plausible analogue of an immunogenic part ("epitope") of a protein for development of diagnostics and the peptide of interest is:

(NH_3^+)-GPSK**R**SFIEDLLFNKVTLAC-(COO^-)

The rationale is that the section KRSFIEDLLFNKV is exposed as associated with S2′ at the surface but highly conserved as shown in the second (i.e. "FIEDLL") alignment in Section 4.3. To bring it to the length of *circa* 20 amino acid residues considered in the author's experience (and frequently by other workers) as most suitable as a basis of a B-epitope for vaccine and diagnostic design, it was noted that the preceding serine (S) and following leucine (L) and alanine (A) are found in 2019-nCoV, well conserved in human host (AAR07628.1) and bat host species, and represent fairly conservative substitutions in the next most related coronaviruses, the pig host species. The remaining principles used in this design are summarized at the end of Section 4.6 alongside the similar discussion for a peptidomimetic, because the method of design (as is typical for epitopes in synthetic peptide vaccines) is based on rules-of-thumb as discussed in that Section, and are similar in the two cases. The choice is supported by the fact that it corresponds to an exposed loop in the SARS coronavirus, which will be important for raising antibodies. The ends of the peptides will have amino and carboxyl groups (shown above), which are linked to the rest of the protein sequence by neutral peptide links. The ends of the peptide are thus selected on the rationale that (a) the exposed N-terminal NH_3^+ -glycine (G) mimics (albeit imperfectly) the positively charged lysine (K) in the KPS triplet and (b) cysteine-(COO^-) provides a means for a Cys-S-S-Cys disulfide link to a carrier protein and the exposed carboxy terminus mimics (again albeit imperfectly) the aspartate (D) in the LAD triplet. A carrier protein is required to be added for antibody production and T-system memory, because short peptides have limited antigenicity. Keyhole limpet haemocyanin is often used, at least for raising antibodies, and for early stage vaccine studies, using laboratory animals. It is derived from the limpet, a gastropod; it is phylogenetically distant from mammalian proteins, thus reducing false positives in immunologically-based research techniques in mammalian model organisms, and clinically avoiding autoimmune effects. Some further conventional wisdom for selection of amino acid residues to synthesize as peptides and used as the basis of diagnostics and vaccines is as follows. Concerns are for raising antibodies (B immune system), implanting immune system memory in the case of vaccines), and specificity, and avoiding unwanted cross reactions with other proteins with similar epitopes to the target protein of interest. The peptide length should in general best be 10–20 amino acids of which many but not all are significantly polar (hydrophilic, water-liking) (residues in the set [STDENQKRHYP]). The length is thus considered appropriate at 20 amino acid residues.

Tackling the issue of avoiding potential immunological cross-reactivity with other human proteins, the closest match of the appropriate corresponding non-terminal sequence KGPSKRSFIEDLLFNKVTLACD with human genome protein products in GenBank is as follows.

```
Query  6     RSFIEDLLFNKVTL  19
             RSF E L FNK TL
Sbjct  119   RSF-EGLIFNKYTL  131
```

The match at 56% identity with 77% coverage is with tumor protein D55 isoform 2 [*Homo sapiens*], ID: NP_001001874.2, and similarly with Tumor protein D52-like 3 [*Homo sapiens*] ID: AAH33792.1. Next match is in regard to neprilsyn entries at only 56% match and 55% coverage. These matches are unlikely to be of concern.

Conventional wisdom in designing peptides as the basis of diagnostics and vaccines is also that one should avoid pairs or triplets etc. of proline (P) or serine (S) and RG pairs that are in common motifs with special functional roles. There is a pair PS in the candidate peptide but that doublet has not been a concern in the author's experience. Asparagine (N) and serine (S) or threonine (T) should be avoided in patterns NXS or NXT (where X is not a proline) because the S or T is likely to be glycosylated in extracellular proteins, unless the designer is going to arrange for glycosylation of the synthetic peptide. That is not an issue in any significant coronavirus matches here. Cysteine C should be avoided (perhaps replaced by serine S) not least because of the formation of unwanted Cys-Cys S–S disulfide bonds, although at either end of the peptide a Cys (C) already present or added can provide the link to a carrier protein that is essential for B immune system antibody response and T system memory. In general, it is information-rich content for binding in terms of polar and nonpolar (hydrophobic, water hating) that matters. Conventional wisdom also says avoid extensive stretches of strongly nonpolar amino acids (residues VLIFW), although extensive hydrophobicity is also considered the characteristic feature of immunogenic MHC Class I T cell epitopes, and some epitopes can play a dual B and T epitope role. Patterns of hydrophobic residues interspersed by hydrophilic residues are considered by many as essential for immunogenicity, and in the author's experience residues with sidechains that are

partly polar and partly nonpolar, notably tyrosine Y and histidine H, seem particularly helpful in terms of immunogenicity. Researchers also sometimes avoid N-terminal asparagine N and glutamine Q, and C-terminal proline P and glycine G. These are not considered to be an issue as regards the immunogenic purposes of the above candidate.

4.6. Some rule-of-thumb principles for design of inhibitors

Simple rules-of-thumb for design of peptides, which for the most part can certainly be automated, are based on the appropriate section or sections of sequence in the protein of the pathogen that one wishes to "attack", and its variations in strongly and weakly related viruses. They take less consideration of the human host proteins that cleave it by proteolysis or bind to it. However, knowledge of the human protein involved would certainly be useful for some features of design, not least because one might use it as an experimental and computational model. One might expect amino acid sequences in peptide and protein substrates cleaved by airway proteases to relate to those cleaved in the S glycoprotein. Unfortunately, this is not obviously apparent at the sequence level, as further discussed in Section 4.7 below. Type II transmembrane serine proteases are likely targets for cleavage activation of 2019-nCoV based on studies of SARS-CoV, but they still require clarification as to the natural sequences typically cleaved, and as discussed in Section 4.7, the relation between KRSFIEDLLFNKV from 2019-nCoV and a model substrate sequence for type II transmembrane serine proteases, AEAALRKLLEVA, is probably not significant. Trypsin cleaves peptides on the C-terminal side of lysine (K) and arginine (R) amino acid residues. If a proline residue (P) is on either side, cleavage is impeded. The zinc-based angiotensin converting enzyme ACE can cleave at cleavage sites (L|F)(H|R|K)L (where the vertical bar '|' separates alternatives), and converts Angiotensin I DRVYIHPFHL to angiotensin II DRVYIHPF. While there are certain suggestive stretches of sequence in coronaviruses that weakly resemble angiotensin, e.g. RVVHPALHL in AVP78050.1 and ATYYHPNIRVGVFHD in ADV71747.1, none correspond to or lie near to the trypsin cleavage points in the SARS S glycoprotein. At best they only imaginatively match the S2′ segment DPSKPSKRSFIEDLL, though the DP (aspartate-proline) in that is suggestive of some features of ACE inhibitors [41]. There is no obvious relation of the S protein cleavage sites with angiotensinogen.

A major difficulty with peptides as pharmaceutical agents is that they lack biostability. They are readily cleaved by proteases, in part a protective and disposal mechanism of the body (albeit one that is used to advantage by the coronaviruses). Also, by being susceptible to partial proteolysis they can, if large enough or by binding to endogenous proteins, cause unwanted immune effects. Judicious use of D (dextro) amino acids, mirror images of their natural L (laevo) counterparts, enhances biostability because they cannot be recognized by common proteases of the body, Consequently peptides containing them will not be degraded easily and have a longer-lasting effect as both vaccines and inhibitor drugs [42–44]. Inserting just a few D-amino acids amongst many L-amino acid residues requires some difficult design and much trial and error, although an exception is the C-terminal conjugation of D-amino acids or glycosides to L-peptides that usually provides a simple way to gain long-term biostability. Because of its relative simplicity and the theoretical basis, the *retroinverso* approach is regaining some popularity [45]. The retroinverso analog of a natural polypeptide is entirely composed of D-amino acids but can sometimes mimic the structure and function of the natural peptide. These analogues have peptide sequence in reverse direction with respect to natural peptide and also have chirality of amino acid inverted from L to D. In other words, the sequence is synthesized backwards, last residue first, and D amino acids are used instead of L. The overall effect is that leaves the sidechains on the original side of the peptide backbone, the one disadvantage being the amide N–H and carbonyl C═O groups are interchanged in the backbone. The problem is probably not so much that a different conformation is achieved in general, since the sidechains are potentially

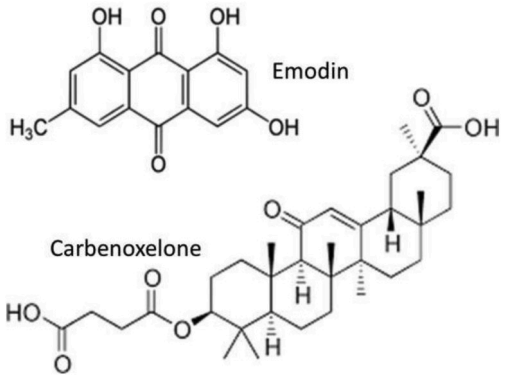


Fig. 3. Emodin, found elsewhere to be an inhibitor of SARS-CoV entry, has some binding features involving the ketone group that are akin to carbenoxelone and to a number of molecules that may be considered as substructures of carbenoxelone's steroid-like ring.

correctly placed and there is some induction by interactions with the receiving protein binding site (in the present case, the lung proteases) just as for a surface loop in the originating protein (in the present case, the spike glycoprotein). Rather, it is that if one or more N–H···O═C (i.e. amide group…carbonyl group) hydrogen bonds are formed between the backbone of the loop on the originating protein and the binding site, they will now be unstable NH···HN or C═O···O═C interactions, unless accommodated correctly by conformational change or intervening water. The latter is not unreasonable, since simulations of peptides in water have shown that one or two water molecules can often intervene even between correctly paired NH and O═C groups in the manner N–H···O–H···O═C. [46], typically by pointing water hydrogen atoms at the oxygen lone pair electron clouds or positions in which they would occur if represented [47]. More easy to handle is the problem that the terminal amino (NH_3^+) and carboxyl (COO^-) groups are also interchanged. As well as being retroinverso, there should be modifications of the sequence to accommodate these termini. Consequently the original sequence below is replaced by the one below that to emulate a lysine (K) at the new N-terminus and an aspartate (D) at the C-terminus.

Original L-Mimetic. (NH_3^+)-GPSK**R**SFIEDLLFNKVTLAC-(COO^-)

retroinverso mimetic (NH_3^+)-dextro-[GNFLLDEIFSRKSRKSPC]-(COO^-)

Neither of the first L-peptide or following D-peptide corresponds simply to the peptide that would be synthesized as the original well conserved KRSFIEDLLFNKV sequence. The steps in defining the *details* in these sequences are drawn from personal experience and common known practices [14–21], various published sources (e.g. Refs. [13, 49–52]), experimental studies of binding and cleavage studies specifically on SARS-CoV and other relevant systems (e.g. Refs. [53–56]), related SARS-CoV studies on smaller ligands (e.g. Refs. [53–56]) discussed later below, and general studies of ligand binding to proteins when hydrophobic groups or hydrogen bonding groups are removed or inserted (e.g. Ref. [57]). Fortunately (not least because capturing expertise for automation is of interest), it possible to write down a core set of rules that have endured, and which explain with the choices of peptides above, once the section of sequence of interest has been determined by bioinformatics methods. These rules are innately logical and extensible to other studies of this general kind. The one debatable exception to this not seen above but seen in the original proposal for the retroinverso peptide [2] is the proline (P) that substituted for arginine (R) in the end section RKSPC of the retroinverso peptidomimetic shown above. This is discussed in Discussion and Conclusions Section 5.2. The primary rules of thumb used are as follows. First recall the sequence in the Wuhan seafood market isolate in this region as follows.

FGGFNFSQILPDPSKPS(KRSFIEDLLFNKV)TLADAGFIKQYGDC

To compare, an alternative to the part in brackets is RRSFIDELAFGRG: it is an example of a tentative *specific* design "hunch" based on a section of a human semaphorin (GenBank NP_001243276.1) produced in response to lung diseases, a suggestive biologically-related match with conservative replacements but *not* one suggested by the *general* rules and probably coincidental. In the first peptide suitable for as a candidate for a peptide vaccine, the principles, once given the core KRSFIEDLLFNKV sequence motif, are as follows. This is to add a C-terminal cysteine (C) as a linkage (other chemical linkages are certainly possible) to the carrier, connected by a moderately flexible TLA (threonine-leucine-alanine) arm. This is found at the same locus in the conservatively retained in other coronaviruses. The sequence is specifically after LA because the -(COO^-) terminus of the protein will mimic the aspartate (D) found in that position in the original sequence (but not, accordingly, in the peptide modeling the epitope). At the C-terminus is also a similarity in that a cystine S–S bridge residue is found on the right (C-terminal side) of this section in the natural sequence, albeit that this is probably not a significant consideration because the naturally occurring C is 11 residues way. Similarly at the N-terminus, the truncation is immediately following an lysine (K) because an amino terminal (NH_3^+)-mimics it. However, a glycine (G) is included to give flexibility and length consistent with mimicking the lysine sidechain. For the retroinverso mimetic, after writing the sequence backward and thinking in terms of D-amino acids, the use of end charges as charged sidechain mimics is reversed, or more correctly stated, the roles of positively and negatively charged sidechains used there are switched. The C terminal linker is sited where an aspartate (D) is in DPSKP…so that the -(COO^-) terminal will mimic it, while the end toward the N-terminus the sequence is truncated at NFLLD… so that the charged N-terminus can mimic the lysine (K) in KPS….. Again here, a glycine is inserted to give appropriate flexibility and length so that the lysine sidechain can be better mimicked.

4.7. Binding and activation considerations for simulation studies

Simulation and molecular modeling studies of a potential agonist go beyond rules-of-thumb and clearly benefit from knowledge of the three dimensional structures, in atomic detail, of the proteins with which they interact. If the automatic vigilance methods of the author and collaborators (and indeed more standard use of the Internet by the author) have failed in this respect, then they will need to be improved. As the situation appears at this time, major simulation studies in this project will be delayed until an experimental structure of the 2019-nCoV spike protein is forthcoming which, at current rate of research on 2019-nCoV, will doubtless be soon. The experimental three dimensional structure of 2019-nCoV spike protein itself is not known at the time of writing, and as far as the author is aware at this time, those SARS CoV spke protein structures that are known are only *circa* 75%–81% homologous with SARS spike proteins. GenBank entry QHR63300.1 for a bat coronavirus spike protein entered on the 27 January 2020 after the present study is 97.4% homologous with SARS proteins, but again the three dimensional structure is not known, or at least not accessible to the present author. Also, protein modeling and binding studies are hampered in that the region around the S2' site of SARS-CoV is partly disordered in conformation, and require a longer period of study.

The human proteins responsible for binding and for cleavage of 2019-nCoV are also not completely clear, though by now several laboratories may have a clearer picture. The recent notion that 2019-nCoV is a kind of SARS will doubtless promote that avenue of enquiry. Of course, a number of studies concerning human protein binding and activation of SARS-CoV have already, some time ago, been carried out by several groups (e.g. Refs. [53–55]). While angiotensin converting enzyme type 2 (ACE2) is responsible for binding SARS, it is a type II transmembrane serine protease (TMPRSS2) that appears to be responsible for the activation cleavage for SARS-CoV [55]. There are several structures known on the Protein Data Bank for proteases of this general kind, including entry 20Q5 for DESC1 and 3T2N for hepsin. Several groups have examined the specificity cleavage specificity analysis of six Type II Transmembrane Serine Proteases (e.g. Ref. [56]). As noted above, there are no obvious sequence relationships between KRSFIEDLLFNKV and type II serine proteases, angiotensinogen, angiotensin 1, or various substrates on which enzymes of potential interest may act, or AEAALRKLLEVA which has been used for exploring active site docking [56]. This was deduced, in the manner of design, from studies of the cleavage specificity analysis of six type II transmembrane serine proteases using PICS (Proteomic Identification of protease Cleavage Sites) with proteome-derived peptide libraries [56]. There is no significant relationship except for the hint of reversal of charges (e.g. ED for RK) between KRSFIEDLLFNKV and the AEAALRKLLEVA peptide. A number of agents that can block binding or activation have been studied [57–59] but there appears to be a number of different mechanisms (e.g. Ref. [59]).

4.8. Preliminary conformational studies on the proposed peptides

The conformational flexibility of the epitope and retroinverso agonist proposed above is also a challenge. That flexibility is an advantage for pharmaceutical action *in vivo*, because although it increases an entropy cost of binding, it is more likely to adjust to fit a binding site, and even fit different relevant sites, in the manner of a locksmith's "skeleton key". In that sense, peptidomimetics are tolerant of design choice: if the intended fit is not perfect, the molecule can adjust. Computationally, however, it increases a well-known problem called the multiple minima problem [11]. With 18 D-amino acid residues, dextro-[GNFLLDEIFSRKSPKSPC] with just 10 distinct conformers there are 10^{18} possibilities. Even if one considers just 5 conformational states per amino acid residue (a gross underestimate), there would be $5^{18} = 3{,}814{,}697{,}265{,}625$ distinct conformers to energy-minimize or subject to molecular dynamics to consider in the free peptide alone, often separated by significant energy barriers. To obtain deeper insight, studies have been performed exploring the intrinsic conformational preferences of KRSFIEDLLFNKV and dextro-[GNFLLDEIFSRKSPK SPC], i.e. with no receptor present, in which the most stable top hundred might be identified.

The software used was the KRUNCH suite illustrated in the appendix to Ref. [48], using a model of "pseudoatoms" (artificial interaction centers) representing first shell of water. The approach, described in more detail in Section 4.8, is one of molecular mechanics using energy minimization capable of exploring multiple minima, boosted by various algorithms to help explore the conformational energy space as a whole. However, many conformers of comparable low energy are found, simply supporting the idea that these candidate peptides in Section 4.6 are conformationally flexible. Starting from an extended structure, essentially e for each residue but specifically backbone dihedral angles $\Phi = -90^\circ$, $\Psi = +150^\circ$, there is a preferred overall secondary structure in reasonable accord with the predictions made on the 2019-nCoV S protein made above using GOR IV for the segment PSKRSFIEDLLFNKVT, and the D-amino acid sequence in reverse is consistent with the mirror image conformation of the reverse sequence. In effect, that simply means that an α-helical conformation (and its mirror image in the D-amino acid peptide case) did not appear. There was departure from local deep minima. Notably EDLLFNK tended occasionally to adopt angles around $\Phi = -60^\circ$ $\Psi = -50^\circ$ characteristic of α-helix, but they returned to the extended conformer. Very preliminary binding studies similar to those of Barr'e et al., but using 2019-nCoV spike protein and the retorinverso peptide suggest various modes of binding varying between *circa* −11 and −16 kcal/mol but these numbers should not be taken too seriously. Only relative values are meaningful but they have been scaled ("adjusted") by correction factors established in previous higher grade molecular dynamics calculations [48]. In any event the above calculations do not include a potentially strong entropy contribution for each conformer (and ideally for the overall solute-solvent

system).

4.9. Preliminary studies on the proposed peptides and smaller ligands using a model pharmacophore for 2019-nCoV antagonism

The system used to explore 2019-nCoV to host cell binding, activation and cell entry, i.e. the "binding model", was chosen on the basis of inhibitors common to, or similar between, (a) SARS binding, activation and entry and (b) and ligands including inhibitors of an enzyme of medical importance, as discussed here below. This model system was initially and in part simply a practice "set up" for relevant studies, suggested by certain relationships between the actions of small molecules revealed by auto-surfing as described below. The choice is undoubtedly likely to be of concern because it involves a human enzyme unlikely to be involved in 2019-nCoV entry. Nonetheless, it proved insightful, and consider that, until relatively recently in the history of pharmacology, ignorance of the atomic details of a protein to which a ligand (as the proposed bioactive small molecule) binds, has been the norm. Pharmacologists have used evidence and hunches to deduce a pharmacophore, i.e. *an abstract description of molecular features that are necessary for molecular recognition of the ligand by the protein*. The rational for the strategy used in the present study is that one can make model binding sites and draw some useful conclusions, but using the actual experimental structure of a protein known to bind similar ligands is more likely to give realistic insight. A justification is that if the pharmacophore in the target choice has no clear features to refute that choice, which the present author calls the "target refutation principle", then it is at least a worthy first choice as a "straw man", i.e. a conjecture for criticism and further debate. After all, any preliminary pharmacophore model might be later refined by adding an "except when" condition to its description.

The choice of the target protein as model was governed by several considerations discussed below, mostly revealed by the auto-surfing approach, and motivated by the above problems of conformational flexibility for peptide ligands. Other investigations by the author and collaborators have focused on rigid, steroid-like scaffolds for binding groups, and also on molecules roughly resembling steroid-like "pieces". In one such study [48], using as the lead a known inhibitor and a steroid-like plant derivative called carbenoxolone, a list of such "pieces" and related molecules screened for binding more at least a superficial resemblance to emodin (1,3,8-trihydroxy-6-methylanthraquinone). This compound came to attention as important in the present study because the "auto-surfing" found that emodin *has already been shown to be an inhibitor* of SARS-CoV entry [57,58]. See Fig. 3. Thus by analogy, it is a putative antagonist for 2019-nCoV entry. Even more importantly, it was found that several studies (e.g. Ref. [60] have shown that emodin, carbenoxelone (again, see Fig. 3), and other molecules with some similarity can inhibit 11-beta-hydroxysteroid dehydrogenase type 1 in rodents and humans. The experimental structure of this enzyme is available with bound carbenoxolone (Protein Data Bank entry 2BEL; the structure also includes the NADP cofactor). In a series of anthraquinone compounds shown to inhibit the steroid-processing enzyme 11-beta-hydroxysteroid dehydrogenase type 1 of both humans and mice, emodin was identified as the most potent selected [60].

The KRUNCH method of molecular modeling used [48,62,63] is not commercially available although the molecular dynamics, docking and other parts of the suit as a whole are essentially the same as standard available versions, and have in the past sometimes been replaced by the standard available versions [48]. The core approach, which is of less usual character, is predominantly concerned with conformational space, not phase space based on molecular dynamics with a dimension of simulated time. This core approach is a descendant of earlier developments in protein and peptide modeling [11] that did not simulate Newtonian dynamics but rather they applied energy minimization techniques, and in particular those called Simplex methods [11]. Because they do not depend on continuous derivatives these methods can navigate a potential energy surface pitted with multiple small energy minima [11,63]. This approach is then extended in pursuit of identifying the global minimum or minima by a Globex method that looks for trends in deep minima on the larger scale [11], and then finally embellished by a variety of techniques that explore conformational space efficiently [63]. Roughly speaking, however, history of a simulation gives a similar impression to molecular dynamics, albeit run at high temperatures with periods of cooling or annealing.

Initial binding studies using 11-beta-hydroxysteroid dehydrogenase type 1 as "receptor" and activator of the virus are similar to those used in Ref. [48]. Early alternative choices of binding protein as model included growth hormone secretagogue receptor type 1 GenPept accession NP_796304 because of some evidence in the literature of the action of anthraquinone, emoghrelin derivatives, and indanthrone, amongst others, which have some of the features of emodin, and because of an initial suspicion that a sequence segment in growth hormone receptor might contribute to binding and had some features relating to the current peptide of interest, but none of the alternatives were ultimately seen as persuasive as the steroid dehydrogenase choice. Preliminary studies in the present project also suggested possible similarities of binding to the active site of steroid dehydrogenase between (a) the peptide epitope proposed above, (b) the retroinverso peptidomimetic, (c) a compound carbenoxelone already studied somewhat intensively by the author and collaborators [48], and (d) emodin. See below. Although initial results for binding of the peptides −8 to −10 kcal/mol initially seemed promising, emodin and carbenoxelone bound significantly lower as discussed below, and the peptide binding differed by showing multiple similar energy binding modes.

Although the dehydrogenase is a very different model system to any directly relating to 2019-nCoV, there were some interesting observations regarding emodin binding that might be extensible to more relevant simulations in a manner consistent with the above "target refutation principle". Using the software described in Ref. [62] and the appendix to Ref. [48], initial studies comprised simply superimposing the smaller analogues into the steroid core of carbenoxolone using the position of the analogous ketone O11 atom on C11 of the steroid-like framework of carbenoxelone and serine 170 side chain oxygen atom as a pivot, removing the carbenoxolone, and initially locally minimizing the energy of the new enzyme-ligand system. Conformational changes of the enzyme on replacing by emodin depended critically on choice of dialectric constant, values less than 20 gave less than 1.9 Å rms. In practice, one conventional rule of thumb is that each aromatic ring can contribute up to about −1.7 kcal/mol in going from an aqueous to a non-polar environment. But although carbenoxolone is steroid-like, and so expected to fit in very hydrophobic pockets, there is ample evidence that there is often hydrogen bonding occurring between steroids and their receptors [61]. In the present system, however, despite significant and differing conformational adjustments by the protein to accommodate these two ligands, there is a negative electrostatic tension in the region of the ketone group because due to close approach of the serine 170 side chain oxygen atom the phenolic oxygen of Tyr 183 and also by oxygen O7N in the NADP cofactor. These electrostatic tensions would seem to provide evidence against the choice of the target protein according to the above target refutation principle. However, it is to be recalled that there is catalysis to consider (including a serine protease type of mechanism) for the proteolysis that is believed to be required to activate the virus for entry. Also, this is at a cut at the lysine residue following an arginine, both carrying a positively charged sidechain. Consequently, relevance cannot be dismissed. Also of course, binding can be favorable overall. Indeed, despite the electrostatic tension, binding energies of −11 and −16 kcal/mol (again adjusted against previous high grade calculations as in Section 4.7) were obtained for emodin and carbenoxelone respectively which seem reasonable indicators of significant binding. Again, however, such values should not be taken as predictions of experimental binding free energy because the above comments on relative values and entropy still apply for the overall

solute-solvent system even though the ligands themselves are much more rigid. Other compounds containing one or few steroid-like rings such as (2,3)-dithio-6-hydroxy-8-carboxy-(1,7,9)-azanaphthalene showed as binding similarly in some binding simulation methods but not in others, and when binding showed multiple binding modes.

5. Discussion and conclusions

5.1. The proposed L-peptide

Recall that it is the L-amino sequence GPSKRSFIEDLLFNKVTLAC that was proposed as an B-epitope to be synthesized with L amino acids and attached to a carrier, for purposes of raising antibodies as a diagnostics, or as a potential vaccine. In the letter case it might require an additional peptide as a T-epitope (for immune system memory), which might come from anywhere in the S protein or even elsewhere in virus, and perhaps other agents such as molecular adjuvant. These are known considerations in the state of the art [50,51], although vaccines based on this synthetic approach are still relatively disfavored for human use despite being in increased use in veterinary medicine. The reason may in part be the historical difficulty in synthesizing longer peptides without side reactions and accumulative errors in the insertions of the amino acids, but now whole proteins can be synthesized, even out of D-amino acids (in which case they fold in mirror image of the native form) [43]. It is increasingly held that the synthetic peptide vaccine approach can sometimes be more effective than traditional vaccine methods [50], and it has been steadily increasing in popularity [51]. The main problem to overcome for vaccine and diagnostic design, with some analogous issues for therapeutic peptidomimetics, seems to be that in many pathogens, the B epitope antigenic determinants for antibody production are discontinuous sections of sequence, and using just one of them typically delivers poor results [51]. However, since sections can be linked chemically with runs of connecting amino acids such as glycine, this is, strictly speaking, just a further layer of design, especially if the three dimensional structure of the protein is known. For the immediate future, that will probably still require a significant amount of experimental trial and error to refine the design, but it is certainly facilitated if one can first obtain activity from studies on different continuous epitopes, ultimately linking the peptides together as above.

5.2. The proposed D-peptide

Recall also that is was D-amino acid sequence GNFLLDEIFSRKSRKSPC that was proposed for synthesis from D-amino acids to act as an antagonist to viral entry. The reason for the retroinverso approach is that an L-peptide would be susceptible to proteolysis from a variety of human proteins. *In silico* (computer simulations) make it possible to use both an L-peptide and D-peptide, as standard molecular mechanics or dynamics will not emulate catalysis (and it would take a remarkably sophisticated and fast quantum mechanical calculation to do so), and the L-peptide results may also be helpful in the design process. *In vivo*, a D-peptide would have a much longer half-life to work either at the human protein required for binding, most often believed to be the angiotensin converting enzyme type 2 (ACE2) as responsible for binding SARS-CoV to lung cells, or as an inhibitor preventing proteolysis required for activation of the S spike protein. At the time of writing, it is the type II transmembrane serine protease TMPRSS2 that is favored as responsible for the activation cleavage. The author favors the inhibition of proteolysis. If the peptide is correct, at least we know form the SARS spike protein experiments that the lysine (K) in the terminal segment PKSPC corresponds to the proteolytic cleavage point in the segment PSKRS. There must be a sufficient degree of binding of an inhibitor to the active (i.e. catalytic) site and many competitive inhibitors work by binding reversibly to it. It is a common view even to the point of view of some authors defining competitive inhibition that way. However, the main binding and recognition site may not be the same thing, and there are also mechanisms by which an enzyme may bind either the inhibitor or the substrate but never both at the same time. Allosteric interactions allow competitive, non-competitive, or uncompetitive inhibition, and the peptide binding at another site could act to enhance enzyme activity (and so act as an agonist to viral entry). Certainly a practical advantage of focusing on the protease in the laboratory is that its inhibition can initially be tested without having to use virus or virus protein. Strictly speaking it is the relative strengths of binding virus and peptide that matter, but this is not generally seen as a serious obstacle in practice.

The original proposal for the retroinverso compound [2] suggested a proline (P) to substitute for arginine (R) in the end section RKSPC. This modification had somewhat the status of a "hunch" and is at best a "weak rule" in the sense of Section 5.3 below. Having been demoted in the present paper it would seem hardly worthy of mention here except that proline has played a recurrent high profile role in the history of design of peptidomimetics, and the modification might still be worth considering if binding of the recommended retroinverso peptide is negligible or weak. Briefly, the various ideas used in the president study involved considering the advantages of a proline residue lacking an NH group that in the retroinverso peptide was suspected to involve an unfavorable N–H.H–N interaction, and a different flexibility at that critical point in which peptide bond cis-trans isomerization replaced extensive N-Cα rotation. There was also a concern that residual L or specific D proteolysis or racemization to L can still occur at some points in a D-amino acid residue sequence. However, the former stereochemical considerations apply to the steroid processing enzyme as the binding and inhibitor model, not necessarily to the lung proteases, and even in considering the former model there are also some objections, discussion of which is beyond current scope.

5.3. Use of bioinformatics and selection of sequences for synthesis

There has been an argument amongst pharmaceutical scientists that, especially at early stages of a study, bioinformatics combined with hunches of the experienced pharmaceutical synthetic chemistry can often provide more reliable guidelines than, for example computational ligand-protein binding simulations that are hampered by the complexity of conformation space [63]. At least, for some decades investigation by bioinformatics has been considered as an important early step, certainly in the views of the present author and collaborators [64–66]. The main role of the project described in the present paper was initially to describe bioinformatics strategies, rather than binding studies, that have worked in other cases as discussed briefly shortly below, and to explore automation. Compared with the tasks of automating protein modeling and drug design in which the present author has been involved (e.g. Refs. [21–27,48]) it seems clear that the above relatively simple series of bioinformatics steps can readily be automated. The main tasks are for this kind of work are (a) detecting those sequences that are conserved in a surface protein across a viral (and potentially bacterial or parasitic) group of organisms, strain and likely to be exposed to the exterior (prediction of coil C helps here as discussed in the text), and (b) programming the set of rules for the detailed choice of amino acid residue sequence and changes to it. Section 4.6 gave a set of rules for the present study that are readily automatable, and those rules are essentially the ones that have stood the test of time and appear logically justified. For design of synthetic vaccines, such automation has already been extensively implemented as prediction software by many workers [49]. All these are essentially tools for predictions, and they are rules-of-thumb particularly in the sense that they may not work in very case, and may needed to be adjusted on a trial-and-error basis. There are also "weak" rules for which evidence for general applicability is questionable, but which could work in a few special cases. The only "rule" in the original specification of Section 4.6 that on further consideration lacked generality, is that for proline discussed above in Section 5.3.

5.4. Computational calculations and binding simulations

Currently, computational chemistry, protein modeling, and binding simulations are popular tools. In practice, however, they must be used with caution because of complexity of the conformational space and, related to that, strong entropic contributions, including of course from the surrounding solvent. This is probably the main reason why IBM's very high performance computer Blue Gene, originally developed on the premise by some workers (including the present author) that it might be able to predict at least the three dimensional structure of some smaller proteins by folding simulation lasting for about a year of computer time [63], did not succeed. It did however achieve many insights into protein modeling as discussed in Ref. [63], and went on to have other applications, including drug design based on reading all US patents at the time, and evolving designs by DOCK and molecular dynamics simulations [48]. The scaling "adjustment" factors use in the molecular mechanics KRUNCH approach in Section 4.8 also depended on these more sophisticated, high grade simulations [48]. Other sources of error includes also (typically) include neglect of quantum mechanical contributions, neglecting quantization of vibrations and with modeling based primarily with atoms representing centers of interaction, essentially meaning limitations in potential functions, which in turn may largely mean neglect of changes in interactions and interaction centers in the environment of other electrical fields.

From the preliminary conformation and binding studies carried out in the present study, one may conclude that these approaches are, for the proposed peptides, challenging by involving many degrees of freedom, both in the conformational and general sense. Simulation methods need to be enhanced by methods that can assess protein and protein-ligand interaction model (e.g. Ref. [62]), and heuristics for overcoming the multiple minimum problem (e.g. Ref. [63]), but so far these particular cited approaches have not provided a quick solution on a standard personal computer..For pharmaceutical reasons too, moving toward smaller non-peptide analogues may be the best direction. It is probably best to await detailed experimental structures of viral and human proteins involved in activation, and their interaction. The electrostatic tension that applied in the above choice of pharmacophore may not make it an optimal model system, but "computer experiments" and incidental observations can provide useful clues and insights. In earlier studies on related compounds including carbenoxelone, converting the ketone group = O to a thioketone = S derivative in most of the compounds studied appeared to enhance binding at first by relaxing the electrostatic tension despite increasing increased the van der Waal's repulsion, but it was subsequently appreciated that a stronger partial charge is required for the thioketone sulfur on a benzene ring because chemically this is strongly electronegative, carrying closer to a full electron unit of charge. The C═S bond length of thiobenzophenone is 1.63 Å thioformaldehyde, but due to steric interactions, the phenyl groups are not coplanar and the dihedral angle SC-CC is 36°. Nonetheless, the binding energy of the emodin thioketone fell by 2 kcal/mol, and this is being investigated. However, the thioketone group is not an optimal choice to explore and might well cause oligomerization in practice, but studies like that above do support the idea that other replacements to the corresponding ketone group in multiple ring substructures of the steroid-like structure might reasonably be explored in pursuit of potent analogues. Of course, the appropriate pharmacophore could be one with hydrogen bonding requirements not met with the above model.

5.5. Effectiveness of peptide based approaches in general

Many principles similar to those of epitope selection apply to therapeutic pepidomimetic design, at least for the relatively simple step of proposing a retroinverso candidate. It is a worthy early strategy, if only on the grounds that it shows suggestive wanted activity, or it does not, and one might well consider it unfortunate if a great deal of computational medicinal chemistry and laboratory synthesis and testing is done, only to find later that using the retroinverso approach would have created a successful first candidate at the outset. A limitation of the approach is that one must be aware of coincidental matches that are not truly significant unless one can see that the proteins being compared are essentially of the same function or family, with the order of similar sections preserved, and that the correspondences make sense in the light of background biological knowledge. Even armed with such knowledge, one must be wary of making too big a jump. A curious example from the present study is as follows. Because in the author's experience a simple reversal of the peptide sequence (i.e. using L-amino acids) can sometimes have required activity, it was interesting to also perform a BLAST search on that reverse sequence. One of the closest matches is with a close match of the above reverse sequence with a sequence in a bacterial enzyme that interconverts D and L alanine: bifunctional UDP-N-acetylmuramoyl-tripeptide:D-alanyl-D-alanine ligase, also known as alanine racemase. More specifically, NFLLDEIFSRKSPKS finds a similar section NFLLNEEFSVIKSPKS in the racemases ID WP_090409885.1 and WP_158209963.1. This is an interesting coincidence, but it does not (as yet) stand up to investigation. This is primarily because amino acid residues from a variety of parts of the racemase appear to interact with itssubstrate, and because the sequence is not significantly preserved in a large number of both related and diverse racemases.

5.6. Future therapeutic aspects and studies

In the ongoing fight against infectious disease there is accelerating progress in the identification of relevant peptides and, where appropriate, design of smaller drug molecules based on them should be applied (e.g. Refs. [67–72]) Such progress provides many additional tools and strategies that could be used. In some cases, one might find that repurposing existing drugs or herbal extracts with suspected genuine efficacy may help in the challenge. As a kind of "stop press", note that at the time of this study it seemed plausible that angiotensin converting enzyme inhibitors could do the same job as a petidomimetic. This was based on certain molecular similarities, but recent studies have highlighted that customary ACE inhibitors are likely to have a counterproductive effect, although an ACE 2 inhibitor might well work usefully. It is likely that the concentrations required may have an excessive effect on blood pressure, but an aerosol preparation might be useful. However, much research on this is needed. However, a peptidomimetic based on the above motif perhaps also delivered as an aerosol, could perhaps be more specific and have less side effects. Design or discovery of more rigid molecules with similar van der Waals's and electrostatic surfaces would be the more traditional pharmaceutical choice.

In the 2011 fictional movie "Contagion" [73] a herbal plant supplement claimed to cure a fictional viral infection played the role of the villain compared with established mainstream methods of vaccine and drug design; however, the villainy was not in the herbal origins but rather the fraudulent experiment in the movie that claimed to support its efficacy. For example, medicinal plants are often found to have an ACE inhibitor action (e.g. ref. [74]). Many compounds studied in works referenced in Section 4.7 were plant extracts, derivatives, or analogues. Carbenoxelone is a derivative of a product from the licorice plant already proven in clinical trials and once marketed as an anti-inflammatory treatment for stomach ulcers, and is steroid-like. Emodin (6-methyl-1,3,8-trihydroxyanthraquinone) can be isolated from rhubarb, buckthorn, and Japanese knotweed and is produced by many species of fungi. The preliminary comparisons with the retroinverso peptide, carbenoxelone and emodin could, of course, emerge as a coincidental similarity of the same kind as that between the reverse sequence used in the retroinverso petidomimetic and a sequence in alanine racemease discussed above.

A future Thinking Web WW4 beyond the current World Wide Web and emerging Semantic Web [28] remains an important goal to facilitate

the response to old and emerging diseases. It seems clear, nonetheless, that rapid access to the emerging literature and the bioinformatics tools available on the Internet, guided by a human researcher, will for some time yet remain important weapons in the battle against viruses and other pathogens. Q-UEL tools, or other similar approaches, can help there.

Declaration of competing interest

This paper is provided to the community to promote the more general applications of the thinking of Professor Paul A. M. Dirac in human and animal medicine in accordance with the charter of The Dirac Foundation, to emphasize the advantages and simplicity of the basic form of the Hyperbolic Dirac Net, to encourage its use, and to propose at least some of the principles of the associated Q-UEL, a universal exchange language for medicine, as a basis for a standard for interoperability. These mathematical and engineering principles are used, amongst many others in an integrated way, in the algorithms and internal architectural features of the BioIngine.com, a distributed system developed by Ingine Inc. Cleveland, Ohio, for the mining of, and inference from, Very Big Data for commercial purposes.

References

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

[2] B. Robson, Preliminary Bioinformatics Studies on the Design of Synthetic Vaccines and Preventative Peptidomimetic Antagonists against the Wuhan Seafood Market Coronavirus. Possible Importance of the KRSFIEDLLFNKV Motif, Circulated and Published on ResearchGate, 2010, https://doi.org/10.13140/RG.2.2.18275.09761.

[3] F. Wu, S. Zhao, B. Yu, Y.-M. Chen, W. Wang, Y. Hu, Z.-G. Song, Z.-W. Tao, J.-H. Tian, Y.-Y. Pei, M.L. Yuan, Y.-L. Zhang, F.-H. Dai, Y. Liu, Q.-M. Wang, J.-J. Zheng, L. Xu, E.C. Holmes, Y.-Z. Zhang, A Novel Coronavirus Associated with a Respiratory Disease in Wuhan of Hubei Province, China, , Wuhan Seafood Market Pneumonia Virus Isolate Wuhan-Hu-1, Complete Genome, 2019. GenBank submission MN908947.3, https://www.ncbi.nlm.nih.gov/nuccore/MN908947. (Accessed 26 January 2020).

[4] F. Li, Structure, function, and evolution of coronavirus spike proteins, Annu. Rev. Virol. 3 (1) (2016) 237–261.

[5] Y.W. Kam, Y. Okumura, H. Kido, L.F.P. Ng, R. Bruzzone, R. Altmeyer, Cleavage of the SARS coronavirus spike glycoprotein by airway proteases enhances virus entry into human bronchial epithelial cells in vitro published, PloS One (2009), https://doi.org/10.1371/journal.pone.0007870. November 17. (Accessed 26 January 2020).

[6] Belouzard, S., Chu, V. C. and Whittaker, G. R., Activation of the SARS coronavirus spike protein via sequential proteolytic cleavage at two distinct sites, Proc. Natl. Acad. Sci., , 106(14, 5871-5876; https://doi.org/10.1073/pnas.0809524106 (last accessed 1/26/2020).

[7] M. Gui, W. Song, H. Zhou, J. Xu, S. Chen, Y. Xiang, X. Wang, Entity 1 containing Chain A, B, C SARS-CoV spike glycoprotein, Cell Res. 27 (2017) 119–129.

[8] 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.

[9] 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.

[10] J.B. Berend, J.W.A. Rossen, W. Bartelink, C. Zuurveen, A.M. de Hann, Duquerroy, C.A.B. 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.

[11] B. Robson, J. Garnier, Introduction to Proteins and Protein Engineering, second ed., Elsevier Press, 1998.

[12] S. Sachdeva, Peptides as 'drugs': the journey so far, Int. J. Pept. Res. Therapeut. 23 (2017) 49, https://doi.org/10.1007/s10989-016-9534-8 (2017).

[13] W. Li, M.D. Joshi, S. Singhania, K.H. Rasey, A.K. Murthy, Peptide vaccine: progress and challenges, Vaccines (Basel) 2 (3) (2014) 515–536.

[14] B. Robson, R.V. Fishleigh, C.A. Morrison, Prediction of HIV vaccine, Nature 4 (325) (1987) 395.

[15] Fishleigh, R. V. and Robson, B, Synthetic Peptides Related to HIV-Env Proteins, patent Patent: EP00371046A1, (1990).

[16] R.V. Fishleigh, B. Robson, R. Aston, Synthetic Polypeptides Derived from the HIV Envelope Glycoprotein.Patent : EU0636145, 1995.

[17] Fishleigh, R. V. and Robson, B, Fragments of Prion Proteins, patent EP00636145A1, (1995),.

[18] Fishleigh, R. V. and Robson, B, and P. Mee, Fragments of Prion Proteins (1998) patent US05773572 (1998).

[19] Citywire. https://citywire.co.uk/new-model-adviser/news/protherics-mad-cow-test-goes-international/a220861. (Accessed 28 January 2020).

[20] B. Robson, From Zika to Flu and Back Again, 2016, https://doi.org/10.13140/RG.2.1.5000.6808. CAVIRC (Report of the Caribbean Anti-Virus Informatics Research Center), https://www.researchgate.net/publication/296667599_From_Zika_to_Flu_and_Back_Again_CAVIRC_Caribbean_Anti-Virus_Informatics_Research_Center.

[21] B. Robson, Computer aided peptide and protein engineering, in: Applied Biotechnology, Proceedings of Biotech 86' Europe, Held in London, 1, 1986, pp. B9–B14.

[22] B. Robson, D.J. Ward, A. Marsden, The EPSITRON concept of peptide and protein engineering. Applications of computer-aided molecular design, Chem. Des. Autom. News 1 (7) (1986) 9–11.

[23] B. Robson, E. Platt, A. Marsden, P. Millard, An expert system for protein engineering. Its application in the study of chloramphenicol acetyltransferase and avian pancreatic polypeptide, J. Mol. Graph. 5 (1987) 8–17 (1987).

[24] R.V. Fishleigh, B. Robson, J. Garnier, P.W. Finn, Studies on rationales for an expert system approach to the analysis of protein sequence data - preliminary analysis of the human epidermal growth factor receptor, FEBS Lett. 2 (4) (1987) 219–225.

[25] J. Garnier, J.F. Gibrat, J. Levin, B. Robson, Modélisation des polypeptides: application aux oligopeptides vaccinants" INRA/EUC Rapport FRT - 85 T 0606, 1988.

[26] J. Ball, R.V. Fishleigh, P.J. Greaney, A. Marsden, E. Platt, L.J. Pool, B. Robson, A polymorphic programming environment for the chemical pharmaceutical and biotechnology industries, in: D. Bawden, E.M. Mitchell (Eds.), Chemical Information Systems - beyond the Structure Diagrams, Ellis Horwood Press, 1990, pp. 107–123.

[27] B. Robson, E. Platt, J. Li, Computer aided design of biomolecules: the big hammer approach, in: David L. Beveridge, Richard Lavery (Eds.), Theoretical Biochemistry and Molecular Biophysics 2 Proteins, Adenine Press, 1992, pp. 207–222.

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

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

[30] B. Robson, S. Boray, Interesting things for computer systems to do: keeping and data mining millions of patient records, guiding patients and physicians, and passing medical licensing exams, in: Bioinformatics and Biomedicine (BIBM), Proceedings 2015 IEEE International Conference, IEEE, 2015, pp. 1397–1404.

[31] B. Robson, 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 (2016) 71–93.

[32] B. Robson, S. Boray, Studies of the role of a smart web for precision medicine supported by biobanking, personalized medicine, FTG 13 (4) (2016).

[33] 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.

[34] 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.

[35] B. Robson and S. Boray, Studies in the use of data mining, prediction algorithms, and a universal exchange and inference language in the analysis of socioeconomic health data", Comput. Biol. Med., Sep;112 doi: 10.1016/j.compbiomed.2019.103369. [Epub ahead of print], (2019).

[36] 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 (February 2020), 103621.

[37] The biology Workbench. http://workbench.sdsc.edu/. (Accessed 28 January 2020).

[38] U.S. National Library of Medicine, National center for biotechnology information, national institutes of health, BLASTP SUITE. https://blast.ncbi.nlm.nih.gov/Blast.cgi. (Accessed 28 January 2020).

[39] J. Garnier, B. Robson, The GOR method for predicting secondary structure in proteins, in: G.D. Fasman (Ed.), Prediction of Protein Structure and the Principles of Protein Conformation, Plenum Publishing Corp, 1989, pp. 417–465.

[40] Rhone-alpes Institute of biology and protein chemistry, GOR IV, https://npsa-prabi.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_gor4.html. (Accessed 28 January 2020).

[41] D.P. De Lima, Synthesis of angiotensin-converting enzyme (ACE) inhibitors: an important class of antihypertensive drugs, Quim. Nova 22 (3) (1999).

[42] B. Robson, Beyond proteins, Trends Biotechnol. 17 (8) (1999) 311–315. B. Robson, "Doppelganger Proteins as Drug Leads", B. Robson (1996), Nature Biotechnology, 14, 892-893 (1966).

[43] G.M. Figliozzi, M.A. Siani, .L.E. Canne, B. Robson, R.J. Simon, Chemical synthesis and activity of D, superoxide dismutase, Protein Sci. 5 (suppl. 1) (1996) 72, 15.

[44] B. Robson, Pseudoproteins: Non-protein Protein-like Machines, the Sixth Foresight Conference on Molecular Nanotechnology, 1998. https://foresight.org/Conferences/MNT6/Abstracts/Robson/index.html. (Accessed 7 September 2019).

[45] J. Rai, Peptide and protein mimetics by retro and retroinverso analogs, Chem. Biol. Drug Des. 93 (5) (2019) 724–736.

[46] A.T. Hagler, D.J. Osguthorpe, B. Robson, Monte Carlo simulation of water behaviour around the dipeptide N-acetylalanyl-N'methylamide, Science 208 (1980) 599–601.

[47] B. Robson, Some views of solvation effects in the light of a Monte Carlo simulation, in: F. Franks, F.S. Mathias (Eds.), The Biophysics of Water, John Wiley & Sons Ltd, 1982, pp. 66–67.
[48] B. Robson, R. Dettinger, A. Peters, S.K.P. Boyer, Drug discovery using very large numbers of patents: general strategy with extensive use of match and edit operations, J. Comput. Aided Mol. Des. 25 (5) (2011) 427.
[49] R.E. Soria-Guerr, R. Nieto-Gomez, B.O. Govea-Alonso, S. Rosales-Mendoza, An overview of bioinformatics tools for epitope prediction: implications on vaccine development, J. Biomed. Inf. 53 (2015) 405–414.
[50] D.J. Kao, R.S. Hodges, Advantages of a synthetic peptide immunogen over a protein immunogen in the development of an anti-pilus vaccine for *Pseudomonas aeruginosa*, Chem. Biol. Drug Des. 74 (2009) 33–42.
[51] C.B. Palatnik-de-Sousa1, I.D.S. Soares, D.S. Rosa, Front. Immunol. (18 April 2018), https://doi.org/10.3389/fimmu.2018.00826 (2018).
[52] M.H.V. Van Regenmortel, Synthetic peptide vaccines and the search for neutralization B cell epitopes, *Open Vaccine J.* 2 (2009) 33–44, https://doi.org/10.2174/1875035401002010033.
[53] Y. Wan, j. Shang, R. Graham, R.S. Baric, F. Li, An analysis based on decade-long structural studies of SARS 3, JVI Accepted Manuscript Posted Online 29 January 2020, J. Virol. (2020), https://doi.org/10.1128/JVI.00127-20.
[54] B.A. Katz, C. Luong, J.D. Ho, J.R. Somoza, E. Gjerstad, J. Tang, S.R. Williams, E. Verner, R.L. Mackman, W.B. Young, P.A. Sprengeler, H. Chan, K. Mortara, J. W. Janc, M.E. McGrath, Dissecting and designing inhibitor selectivity determinants at the S1 site using an artificial Ala190 protease (Ala190 uPA), J. Mol. Biol. 344 (2) (2004) 527–547, 19.
[55] M.R. Lennart, M. Reinke, m. Spiegel, t. Plegge, A. Hartleib, I. Nehlmeier, S. Gierer, M. Hoffmann, H. Hofmann-Winkle, M. Winkler, S. Pöhlmann1, Different residues in the SARS-CoV spike protein determine cleavage and activation by the host cell protease TMPRSS2, PloS One 12 (6) (2017), e0179177.
[56] O. Barr'e, A. Dufour, U. Eckhard, R. Kappelhoff, F. Béliveau, R. Leduc, C. M. Overall, Cleavage specificity analysis of six type II transmembrane serine proteases (TTSPs) using PICS with proteome-derived peptide libraries, PloS One 9 (9) (2014), e105984, https://doi.org/10.1371/journal.pone.0105984 (2014).
[57] T.Y. Ho, S.l. Wu, J.C. Chen, C.Y. Hsiang, Emodin blocks the SARS coronavirus spike protein and angiotensin-converting enzyme 2 interaction, Antivir. Res. 74 (2) (2007) 92–101.
[58] S. Schwarz, K. Wang, W. Yu, B. Sun, Schwarz, Emodin inhibits current through SARS-associated coronavirus 3a protein, Antivir. Res. 90 (1) (2011) 64–69.
[59] A.O. Adedeji, W. Severson, C. Jonsson, K. Singh, S.R. Weiss, S.G. Sarafian, Novel inhibitors of severe acute respiratory syndrome coronavirus entry that act by three distinct mechanisms, J. Virol. 87 (14) (2013) 8017–8028.
[60] Y. Feng, S.-L. Huang, W. Dou, S. Zhang, J.-H. Chen, Y. Shen, J.-H. Shen, Y. Leng, Emodin, a natural product, selectively inhibits 11β-hydroxysteroid dehydrogenase type 1 and ameliorates metabolic disorder in diet-induced obese mice, Br. J. Pharmacol. 161 (1) (2010) 113–126.
[61] U. Westphal, Hydrophobicity and hydrophilicity of steroid binding sites, in: Steroid-Protein Interactions II. Monographs on Endocrinology, 27, Springer, 1986.
[62] B. Robson, A. Curioni, T. Mordasini, Studies in the assessment of folding quality for protein modeling and structure prediction, J. Proteome Res. (Am. Chem. Soc.) 1 (2) (2002) 115–133 (2002).
[63] B. Robson, A. Vaithilingham, "Protein folding revisited" pp 161-202 in progress in molecular biology and translational science, in: Molecular Biology of Protein Folding, 84, Elsevier Press/Academic Press, 2008 (2008).
[64] J. Li, B. Robson, Bioinformatics and Computational Chemistry in Molecular Design. Recent Advances and Their Application, Pp 285-307 in Peptide and Protein Drug Analysis, Marcel Dekker, NY, 2000, p. 200.
[65] B. Robson, R. McBurney, The role of information, bioinformatics and genomics, in: R.G. Hill, P. Rang (Eds.), Drug Discovery and Development: Technology in Transition, second ed., Elsevier Press., 2012, pp. 77–94.
[66] D. Smith, Robson, High Throughput Insight: Web-Based Collections of Bioinformatics Tools Catalyzes Scientific Inquiry into Subtle Aspects of Gene Structure and Function, IBC Library Series, 1998.
[67] A. Ali, A. Khan, A.C. Kaushik, et al., Immunoinformatic and systems biology approaches to predict and validate peptide vaccines against Epstein–Barr virus (EBV), Sci. Rep. 9 (2019) 720, https://doi.org/10.1038/s41598-018-37070-z.
[68] S. Kaliamurthi, G. Selvaraj, S. Chinnasamy, Q. Wang, A.S. Nangraj, W.C. Cho, K. Gu, D.-Q. Wei, Exploring the papillomaviral proteome to identify potential candidates for a chimeric vaccine against cervix papilloma using immunomics and computational structural vaccinology, Viruses 11 (2019) 63, https://doi.org/10.3390/v11010063.
[69] A. Mehmood, C.K. Aman, D.-Q. Wei, Prediction and validation of potent peptides against herpes simplex virus type 1 via immunoinformatic and systems biology approach, Chem. Biol. Drug Des. 94 (5) (2019) 1868–1883, https://doi.org/10.1111/cbdd.13602.
[70] Chu, Kaushik, A. C Wang, X. , Wang, W., Zhang, Y., Shan, X Russell, D., Salahub, Xiong, Y., Wei, D-Q, DTI-CDF: Acascade deep forest model towards the prediction of drug-target interactions based on hybrid features, Briefings Bioinf., , bbz152, https://doi.org/10.1093/bib/bbz152.
[71] A.C. Kaushik, A. Mehmood, A.K. Upadhyay, P. Shalinee, S. Srivastava, M. Prayuv, Y. Xiong, X. Dai, D.-Q. Wei, S. Sahi, CytoMegaloVirus infection database: a public omics database for systematic and comparable information of CMV, Interdiscipl. Sci. Comput. Life Sci. (2019, Dec 7), https://doi.org/10.1007/s12539-019-00350-x [Epub ahead of print].
[72] A.C. Kaushik, A. Mehmood, S. Peng, Y.J. Zhang, X. Dai, D.-Q. Wei, A-CaMP: a tool for anti-cancer and antimicrobial peptide generation, J. Biomol. Struct. Dyn. (2020), https://doi.org/10.1080/07391102.2019.1708796 published online, 6 Jan.
[73] Wikipedia, contagion (film). https://en.wikipedia.org/wiki/Contagion (2011_film, last access (2/1/202)).
[74] M.Y. Khan, V. Kumar, Mechanism & inhibition kinetics of bioassay-guided fractions of Indian medicinal plants and foods as ACE inhibitors, J. Tradit. Complementary Med. 9 (1) (2019) 73–78.

Recent Papers

[75] B. Robson, Quantum universal exchange language and hyperbolic Dirac nets for precision medicine and drug design. Proposals with examples from mitochondrial studies, Comput. Biol. Med. 117 (February, 2016), 103621.
[76] B. Robson, S. Boray, Studies in the use of data mining, prediction algorithms, and a universal exchange and inference language in the analysis of socioeconomic health data, Sep;112 in press, Comput. Biol. Med. (2019), https://doi.org/10.1016/j.compbiomed.2019.103369 [Epub ahead of print].
[77] B. Robson, Bidirectional General Graphs for inference. Principles and implications for medicine, Comput. Biol. Med. 10 (2019) 382–399 (2019).
[78] 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.
[79] 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.
[80] B. Robson, S. Boray, Studies of the role of a smart web for precision medicine supported by biobanking, Pers. Med. 13 (2016) 4, https://doi.org/10.2217/pme-2015-0012. Published Online:5 Jul 2016.
[81] B. Robson, 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 (2016) 71–93.
[82] B. Robson, S. Boray, Interesting things for computer systems to do: keeping and data mining millions of patient records, guiding patients and physicians, and passing medical licensing exams, in: Bioinformatics and Biomedicine (BIBM), Proceedings 2015 IEEE International Conference, IEEE, 2015, pp. 1397–1404.
[83] B. Robson, S. Boray, Implementation of a web based universal exchange and inference language for medicine. Sparse data, probabilities and inference in data mining of clinical data repositories, Comput. Biol. Med. 66 (2015) 82–102.
[84] S. Deckelman, B. Robson, Split-complex numbers and Dirac bra-kets, Commun. Inf. Syst. 14 (3) (2015) 135–149.
[85] B. Robson, T. Caruso, U.G.J. Balis, Suggestions for a web based universal exchange and inference language for medicine. Continuity of patient care with PCAST disaggregation, Comput. Biol. Med. 56 (2014) 51–66.
[86] B. Robson, POPPER, a simple programming language for probabilistic semantic inference in medicine, Comput. Biol. Med. 56 (2014) 107–123.
[87] B. Robson, hyperbolic Dirac nets for medical decision support. Theory, methods, and comparison with bayes nets, Comput. Biol. Med. 51 (2014) 183–197.
[88] B. Robson, T. Caruso, U.G.J. Balis, Suggestions for a web based universal exchange and inference language for medicine, 1, Comput. Biol. Med. 43 (12) (2013) 2297–2310, https://doi.org/10.1016/j.compbiomed.2013.09.010. Epub 2013 Sep 20.Also found in preliminary form, with permission of the Editor-in-Chief, at the US Government S&I website:-, http://wiki.siframework.org/file/view/UELRobson102corrections.pdf/451304614/UELRobson102corrections.pdf.
[89] B. Robson, The concept of novel compositions of matter. A theoretical analysis, Intellect. Property Rights 1 (2013) 108, https://doi.org/10.4172/ipr.1000108.
[90] B. Robson, Towards new tools for pharmacoepidemiology, Adv. Pharmacoepidemiol. Drug Saf. 1 (2013) 6, https://doi.org/10.4172/2167-1052.100012.
[91] B. Robson, R. McBurney, The Role of Information, Bioinformatics and Genomics" in Drug Discovery & Development. Technology in Transition, second ed., Churchill Livingstone (Elsevier, 2013, pp. 77–94.
[92] B. Robson, Towards automated reasoning for drug discovery and pharmaceutical business intelligence, 2012, Pharmaceut. Technol. Drug Res. 1 (3) (2012) (27 March 2012) Robson, B., Li, J., Dettinger, R., Peters, A., and Boyer, S.K. (2011), Drug discovery using very large numbers of patents. General strategy with extensive use of match and edit operations. Journal of Computer-Aided Molecular Design 25(5): 427-441 (2011).
[93] B. Robson, U.G.J. Balis, T.P. Caruso, Considerations , for a universal exchange language for healthcare, in: In Proceedings of 2011 IEEE 13th International Conference on E-Health Networking, Applications and Services (Healthcom 2011), IEEE, Columbus, MO, 2011, pp. 173–176, 2011.

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