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.

Journal of Immunological Methods 481–482 (2020) 112787

Contents lists available at ScienceDirect

Journal of Immunological Methods

journal homepage: www.elsevier.com/locate/jim

Technical Note

In the search of potential epitopes for Wuhan seafood market pneumonia virus using high order nullomers

Daniele Santoni[a,*], Davide Vergni[b]

[a] Institute for System Analysis and Computer Science "Antonio Ruberti", National Research Council of Italy, Via dei Taurini 19, 00185 Rome, Italy
[b] Institute for applied mathematics "Mauro Picone", National Research Council of Italy, Via dei Taurini 19, 00185 Rome, Italy

ARTICLE INFO

Keywords:
Nullomers
Peptide-HLA
Immunoinformatics
Viral genomes
SARS-CoV-2
Self/Non-Self

ABSTRACT

Alarms periodically emerge for viral pneumonia infections due to coronavirus. In all cases, these are zoonoses passing the barrier between species and infect humans. The legitimate concern of the international community is due to the fact that the new identified coronavirus, named SARS-CoV-2 (previously called 2019-nCoV), has a quite high mortality rate, around 2%, and a strong ability to spread, with an estimated reproduction number higher than 2. Even though all countries are doing their utmost to stop the pandemic, the only reliable solution to tackle the infection is the rapid development of a vaccine. For this purpose, the means of bioinformatics, applied in the context of reverse-vaccinology paradigm, can be of fundamental help to select the most promising peptides able to trigger an effective immune response. In this short report, using the concept of nullomer and introducing a distance from human self, we provide a list of peptides that could deserve experimental investigation in the view of a potential vaccine for SARS-CoV-2.

1. Introduction

Coronaviruses belong to the family of Coronaviridae in the order of Nidovirales. For humans they can be responsible for mild respiratory infections (like cold) and also for much more serious diseases (like Severe acute respiratory syndrome –SARS or Middle East respiratory syndrome – MERS). Recently, in December 2019 an alarm emerged in the city of Wuhan, China, for a new viral pneumonia infection due to a novel coronavirus, falling within the genus Betacoronavirus, that was traced to a seafood wholesale market and named as SARS-CoV-2. Preliminary analysis revealed that different SARS-CoV-2 sequenced genomic strains exhibit more than 99.98% sequence identity, and they are closely related (88% identity) to bat-derived SARS-like coronaviruses, while they are more distant from human SARS-CoV (about 79%) and human MERS-CoV (about 50%) (Lu et al., 2020).

According to current statistics the new coronavirus seems to be less lethal than previous SARS or MERS outbreaks with a percentage of deaths, so far, around 2%, but it spreads out very quickly and in a couple of months the epidemic has already infected and killed more patients than previous coronavirus infections of SARS (2003) and MERS (2012). By now (middle of February 2020) there are more than 2000 dead and more than seventy thousand infected in China. In this view the rapid development of a vaccine is a focal issue to avoid pandemics. According to the new paradigm of reverse vaccinology (Rappuoli, 2001) identification of epitopes, able to trigger an immune response against pathogens, is the fundamental step to develop effective new vaccine.

In this work we applied an original computational methodology together with reliable and effective bioinformatics tools in order to select promising viral peptides able to trigger an effective immune response. In particular, the core of the proposed methodology is based on the selection of viral peptides that are more than three mutation steps far from human self. In other terms we selected those *high order nullomers*, i.e. viral peptides that are absent in human (simple *nullomers* (Hampikian and Andersen, 2007)), such that all the peptides obtained by mutating any three aminoacids of them are still absent. As highlighted in (Santoni, 2018) it has been shown that high order nullomers have a higher probability to bind HLA than expected. Moreover, in our opinion, the identification of farthest-from-human peptides could be interesting in the view of vaccine design both for cross-reactivity reasons, since we expect on average an high number of potential antibodies able to recognize them, and for avoiding autoimmunity risks, because of the very low similarity with human self.

Those peptides (named third order nullomers, according to (Vergni and Santoni, 2016), or $W4$, as specified in the following) were further processed in order to select the ones with the highest likelihood to be exposed on cell surface, according to three selection steps: i) the probability to be the product of proteasome cleavage, ii) the probability

* Corresponding author.
E-mail address: daniele.santoni@iasi.cnr.it (D. Santoni).

https://doi.org/10.1016/j.jim.2020.112787
Received 20 February 2020; Accepted 22 March 2020
Available online 23 April 2020

to be carried out by TAP-transport complex and iii) the probability to strongly bind at least one HLA among the 89 considered in this study. Finally we identified a minimal set of 9 peptides of interest that could deserve further experimental investigation. Previous studies involving nullomers have been proposed for designing novel therapeutical strategies Silva et al. (2015); Alileche and Hampikian (2017), and the use of high order nullomers instead of simple nullomers certainly represents a step forward in such a context.

2. Materials and methods

2.1. Proteomes

The *Homo sapiens*' proteome (GRCh38) was downloaded from Ensembl site (http://ftp.ensembl.org/pub/current_fasta/homo_sapiens/pep/).

Wuhan-Hu-1 (MN908947) was taken as a reference for SARS-CoV-2 from ncbi site (https://www.ncbi.nlm.nih.gov/nuccore/MN908947.3/).

Ad hoc scripts were designed in order to extract all possible 9-mers from the sequence of Human proteome and for SARS-CoV-2. It is worth noting that, in this work, only contiguous peptides of size 9 were considered. This choice does not significantly affect the results of the work since the largest part of peptides presented on cell surface by MHC class I complex are 9-mers contiguous in sequence. Therefore, hereafter the term peptide will indicate 9-mers contiguous peptides and, without ambiguity, we will use *HSA* and SCV_2 to indicate Human and SARS-CoV-2 peptides, respectively.

2.2. Peptide classes: distance from human self

Given two peptides $p \equiv (p_1,p_2,\cdots,p_9) \in SCV_2$ and $q \equiv (q_1,q_2,\cdots,q_9) \in HSA$, we computed the distance between them by counting the number of positions in which the aminoacids p_i and q_i are different $D(p,q) = \Sigma_{i=1}{}^{9} B(i)$, where $B(i) = 1$ if $p_i \neq q_i$, otherwise is zero. By using the distance D it is possible to compute the minimal number of mutations that are needed to transform a peptide $p \in SCV_2$ into a human peptide

$$M(p) = \min_{q \in HSA} \{D(p,q)\} \tag{1}$$

By definition if $M(p)$ is equal to m no human 9-mer can be obtain from p by mutating a number of aminocids smaller than m. According to the distance M we defined different disjoint subsets of SCV_2, starting from the common class, C, as:

$$C = \{p \in SCV_2 \mid M(p) = 0\} \tag{2}$$

i.e., the subset of SCV_2 in common with *HSA*, to the first class of absent peptides, W_1, as

$$W_1 = \{p \in SCV_2 \mid M(p) = 1\} \tag{3}$$

containing 9-mers peptides of SARS-CoV-2 that are not present in human but that can be changed in at least one human peptide with a single mutation step, and generalizing to the m-th class of absent peptides, W_m, as

$$W_m = \{p \in SCV_2 \mid M(p) = m\}. \tag{4}$$

It follows that a peptide belonging to class W_m needs m mutation steps (not less than m) in order to be changed in at least one human peptide. In the following we will also refer to peptides belonging to set W_1 as simple nullomers, i.e., absent words of the set *HSA*, while to peptides belonging to class W_i (with $i > 1$) as high order nullomers. Unlike Vergni and Santoni (2016) in which nullomers and high order nullomers are gathered in non-disjoint sets, in this work we preferred, for exposition clarity, to use disjoint classes of nullomers, W_i.

2.3. Prediction softwares

2.3.1. Peptide-MHC class I interaction: NetMHC

The software NetMHC (version 4.0) (Lundegaard et al., 2008) has been used to predict the interaction of peptides with MHC class I complex in terms of binding probability score, taking into account 89 different HLAs (38 class A, 39 class B 10 class C and 2 class E). According to NetMHC software a peptide is predicted to strongly bind (SB) a given HLA when the related interaction score is smaller or equal to 0.5, while for a predicted weak bind (WB) the score falls in the range from 0.5 to 1.5.

2.3.2. TAP-transport and proteasome cleavage: NetCTL

The software NetCTL (version 1.2) (Larsen et al., 2007) has been used to predict proteasome cleavage probability (indicated as CLE) and transport scores related to the transporter associated with antigen processing (indicated as TAP).

3. Results and discussion

The total number of different 9-mers occurring as contiguous substrings in the human proteome is 11,224,527, while the total number of unique contiguous 9-mers of SARS-CoV-2 is 9591. Table 1 shows the partition of SARS-CoV-2 peptides with respect to the human self depending on their distance from human peptides, as reported in Section 2.2.

In the following we will focus on the 27 peptides belonging to the class W4, the farthest from human self. 25 out 27 peptides are shared in all the 39 so far available strains of SARS-CoV-2 while the two remaining are missing in just one strain (IMRLWLCWK and MRLWLCWKC are missing in the strain MT039890). This is not surprising because currently available strains exhibit more than 99.98% sequence identity. It is worth to note that no peptide in SCV_2 turns out to be W5, in other words four steps of mutations are sufficient to obtain from every peptide in SCV_2 at least one peptide in *HSA*.

Then, starting from the peptides in the W4 set, we applied further selection steps according to their probability to be the product of proteasome cleavage, to their propensity to be transported on cell surface by TAP complex and to the probability score of binding MHC class I complex. W4 peptides showing at least one predicted HLA strong bind (19 out of 27) are reported in Table 2 together with the following features: the number of HLAs that each peptide strongly and weakly bind, respectively, the ORF in which each peptide occurs (for ORF1 it is also mentioned the nsp nonstructural proteins of orf1a/b according to (Chan et al., 2020)), the position of each peptide with respect to the correspondent ORF, the proteasome cleavage and TAP-transport predicted scores.

The rows related to peptides showing a significant cleavage (greater than 0.4) and TAP scores (greater than 0.5) - 9 peptides - are highlighted in bold. It is worth noting that almost all those bolded peptides have a large number of Strong/Weak Bind HLAs. In particular *YVMHANYIF* occurring in ORF1 (nsp16) is predicted to strongly bind 27 different HLAs and weakly bind 17 ones, it shows a significant probability (0.52) to be cleaved by proteasome complex and a very high TAP score (2.68). *FLCWHTNCY* and *YIKWPWYIW* are predicted to strongly bind 11 and 10 different HLAs, respectively, with very high proteasome cleavage and TAP scores. It is worth discussing also peptide *YYHKNNKSW*, showing 8 Strong Bind HLAs and significant Cleavage and TAP

Table 1
Partition of Wuhan coronavirus 9-mers in classes C, W1, W2, W3 and W4.

	Total	C	W1	W2	W3	W4
Number	9591	2	156	4104	5302	27
Percentage	100	0.02	1.63	42.79	55.28	0.28

Table 2
List of the selected W4 peptides. The first column reports the selected peptides. Second and third columns report the numbers of HLAs the peptide strongly and weakly, respectively, bind. In the fourth column the ORF in which the peptide occur is reported. Fifth column indicates the position of peptide in the correspondent ORF. Proteasome Cleavage and TAP-transport predicted scores are reported in sixth and seventh column, respectively. The peptide YYHKNNKSW, whose aminoacids result to be belong to the solvent-accessible surface area of the spike protein, has been labeled with "*".

Peptide	#SB-HLA	#WB-HLA	ORF(nsp)	Position	CLES-core	TAP-Score
YQCGHYKHI	7	9	ORF1(nsp3)	1831–1839	0.1801	0.7400
CMMCYKRNR	3	2	ORF1(nsp3)	2409–2417	0.0427	1.6400
HNWNCVNCD	1	0	ORF1(nsp3)	2448–2456	0.0541	−1.8390
HIQWMVMFT	1	5	ORF1(nsp4)	3125–3133	0.0236	−0.5460
CISTKHFYW	**3**	**8**	**ORF1(nsp4)**	**3147–3155**	**0.7680**	**0.9770**
YQCAMRPNF	9	15	ORF1(nsp5)	3389–3397	0.0536	2.6460
SWVMRIMTW	3	7	ORF1(nsp6)	3658–3666	**0.6403**	**1.4160**
CKCCYDHVI	2	2	ORF1(nsp13)	5351–5359	0.2798	0.5450
MMGFKMNYQ	2	2	ORF1(nsp14)	5982–5990	0.0269	−0.0460
WHHSIGFDY	**4**	**7**	**ORF1(nsp14)**	**6152–6160**	**0.6175**	**2.9280**
KLMGHFAWW	**8**	**13**	**ORF1(nsp16)**	**6980–6988**	**0.4431**	**1.0240**
YVMHANYIF	**27**	**17**	**ORF1(nsp16)**	**7020–7028**	**0.5218**	**2.6880**
YYHKNNKSW*	**8**	**12**	**ORF2**	**144–152**	**0.9329**	**1.2060**
MQMAYRFNG	3	5	ORF2	900–908	0.0234	−1.0540
YIKWPWYIW	**10**	**14**	**ORF2**	**1209–1217**	**0.8066**	**0.9420**
IMRLWLCWK	**4**	**5**	**ORF3**	**124–132**	**0.8803**	**0.6670**
MRLWLCWKC	1	5	ORF3	125–133	0.0894	0.2950
FLCWHTNCY	**11**	**14**	**ORF3**	**146–154**	**0.9264**	**2.806**
CWHTNCYDY	**3**	**6**	**ORF3**	**148–156**	**0.9621**	**3.2080**

score (0.93 and 1.2 respectively). This peptide occurs in ORF2, coding for the spike protein, that is a surface glycoprotein. Interestingly, computational prediction analysis, performed through NetSurf software ((Klausen et al., 2019)), indicates that the peptide *YYHKNNKSW* is in the solvent-accessible surface area of the spike protein, so it could further considered and investigated as a potential B-cell epitope.

We also analyzed the MHC class I allele distribution in a sample population (121 individuals) of Wuhan, where the first outbreak of epidemics occurred and where the most part of infected subjects are, available from the site http://www.allelefrequencies.net/default.asp. Four peptides out of the selected eight ones (*FLCWHTNCY*, *WHHSIGFDY*, *YIKWPWYIW* and *YVMHANYIF*) are predicted to strongly bind *B*15 HLA family that occurs in around 15% of the individuals recruited in this study, four are predicted to strongly bind *B*40 occurring in 15% of individuals. Three of them (*YVMHANYIF*, *YYHKNNKSW* and *KLMGHFAWW*) are predicted to strongly bind *A*24 occurring in around 17% of individuals.

Finally we performed comparative analysis, looking for the presence of those peptides in close related species such as SARS, MERS or Bat corona viruses. All the eight peptides only occur in the two close related BAT-SARS like viruses and/or in other species of Bat corona viruses. None of them occurs in MERS or human SARS.

In conclusion, in this technical note we provide a list of SARS-CoV-2 peptides potential targets for the immune system, that could be experimentally tested in order to validate their immunogenicity. Those peptides were selected as the farthest, in mutation terms, from the human self in the belief, confirmed by previous bioinformatics analysis (Santoni, 2018), that the higher the distance from human self the higher the probability to bind HLA. Moreover, the selected peptides were also chosen for their high probability to be cleaved by proteasome and TAP transported to the cell surface. Interestingly, one selected peptide, was found to be located in the solvent-accessible surface area of the spike protein.

Acknowledgments

We wish to thank Dr. Massimiliano Adamo for his useful comments and suggestions.

References

Alileche, A., Hampikian, G., 2017. The effect of Nullomer-derived peptides 9R, 9S1R and 124R on the NCI-60 panel and normal cell lines. BMC Cancer 17, 533.
Chan, J.F., et al., 2020. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg Microbes Infect 9 (1), 221–236.
Hampikian, G., Andersen, T., 2007. Absent sequences: nullomers and primes. Pac. Symp. Biocomput. 12, 355–366.
Klausen, M.S., et al., 2019. NetSurfP-2.0: improved prediction of protein structuralfeatures by integrated deep learning. Proteins 87, 520–527.
Larsen, M.V., et al., 2007. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. BMC Bioinformatics. 8, 424.
Lu, R., et al., 2020. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet. https://doi.org/10.1016/S0140-6736(20)30251-8.
Lundegaard, C., et al., 2008. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. Nucleic Acids Res. **1** (**36**) (Web Server issue), W509-512.
Rappuoli, R., 2001. Reverse vaccinology, a genome-based approach to vaccine development. Vaccine. 19, 2688–2691.
Santoni, D., 2018. Viral peptides-MHC interaction: binding probability and distance from human peptides. J. Immunol. Methods 459, 35–43.
Silva, R.M., et al., 2015. Three minimal sequences found in Ebola virus genomes and absent from human DNA. Bioinformatics 31 (15), 2421–2425.
Vergni, D., Santoni, D., 2016. Nullomers and high order Nullomers in genomic sequences. PLoSONE 11 (12), e0164540.