Abstract
A safe and effective vaccine is urgently needed to bring the current SARS-CoV-2 pandemic under control. The spike protein (SP) of SARS-CoV-2 represents the principal target for most vaccines currently under development. Despite the presence of a CoV proof-reading function in viral replication, SP protein from SARS-CoV still extensively mutates, which might have an impact on current and future vaccine development. Here, we present analysis of more than 1600 SP unique variants suggesting that vaccine candidates based on the Wuhan-Hu-1 reference strain would be effective against most of currently circulated SARS-CoV-2 viruses, but that further monitoring of the evolution of this virus is important for identification of other mutations, which could affect the effectiveness of vaccines.

Keywords
COVID-19, SARS-CoV-2, mutations, vaccine effectiveness
Introduction
The current Coronavirus Disease 2019 (COVID-19) pandemic, caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), represents an unprecedented health threat resulting in over 1.1 million deaths globally (www.worldometers.info/coronavirus) and significant economic damage. There is a consensus that development of an effective and safe vaccine represents the best strategy to control the COVID-19 pandemic, and consequently, this endeavor remains a top public health priority. Researchers and drug companies around the world are working hard to develop a COVID-19 vaccine, and currently 282 vaccine candidates are in clinical trials or will enter clinical investigation soon (clinicaltrial.gov). The spike glycoprotein (SP) of SARS-CoV-2 is the principal target for most these vaccine candidates. Despite the impact of a CoV proof-reading function in viral replication, SARS-CoV still extensively mutates, which might have an impact on current and future vaccine development. In the GISAID database, more than 1,600 SARS-CoV-2 strains with unique combinations of mutations in SP are deposited. In order to fully understand the impact of the mutations on the SP properties (pathogenesis, virulence, transmissibility, immunogenicity), advanced laboratory studies are required. These investigations take time and are done in close collaboration between different research groups, institutions, and academia. However, some important information about the biological effects of mutations could and should also be obtained by the in silico analysis of SP, which can be performed quickly and with a minimal experimental foreknowledge at the time of analysis.

Previously, a novel bioinformatics approach, which is based on electronic biology, was used in the assessment of the effect of mutations on the vaccine effectiveness (VE) for seasonal flu vaccines. Most recently, this approach allowed successful prediction of VE for two successive flu seasons. Here, we used this bioinformatics approach for the analysis of the impact of mutations in SP from SARS-CoV-2 virus on the effectiveness of the COVID-19 vaccines and/or vaccine candidates. The presented results showed that VE of vaccine candidates that are based on SP from the reference SARS-CoV-2 strain YP_009724390 remains between 76% and 94%, despite an extensive mutation rate in this important and likely protective antigen. Our analysis also suggests that further close monitoring of SP mutations might be essential in order to be able to respond to changes that may influence VE in the future.

Methods
Viruses
We analyzed the non-redundant subunit 1 of S proteins (SP1) from human SARS-CoV-2 viruses deposited in the GISAID (https://www.epicov.org/epi3/cfrontend#18c7c) from January to October, 2020, and SP1 from viruses isolated from mink that are presented in the same database. Prototype virus Wuhan-HU-1 (YP_009724390) is used in the analysis as the wild type (WT) and as the vaccine virus.

Informational spectrum method (ISM)
The ISM is the virtual spectroscopy method for the analysis of the protein biological properties. This bioinformatics approach encompasses three basic steps: (i) the representation of the primary structure of the protein as the numerical sequence by the assignment to each amino acid of the corresponding value of the electron-ion interaction potential (EIIP), (ii) the transformation of the obtained numerical sequence into the informational spectrum (IS), and (iii) the calculation of the cross-spectrum (CS) between interacting proteins.

The EIIP is the physical parameter that determines the long-range interactions of biological molecules (interactions at distances >5Å). This molecular descriptor is defined by the following equation:

$$W = 0.25 \frac{Z^* \sin(1.04 \pi Z^*)}{2\pi}$$

where $Z^*$ is the average quasivalence number (AQVN):

$$Z^* = \frac{1}{N} \sum_{i=1}^{N} n_i Z_i$$

where $N$ is the total number of atoms, $n_i$ is the number of the $i$-th component, $Z_i$ is the valence number of the $i$-th atomic component in the molecule and $m$ is the number of components. The EIIP values calculated according to the Equation (1) are given in Rydbergs (Ry).

The numerical sequence, representing the primary structure of protein, is transformed into the IS by the discrete Fourier transformation

$$X(n) = \sum_{m=1}^{N} x(m) e^{-\text{i}2\pi m (n-1)/N}, \quad n = 1, 2, ..., N/2$$

where $x(m)$ is the $m$-th member of a given numerical series, $N$ is the total number of points in this series, and $X(n)$ are discrete Fourier transformation coefficients. In this way, the information defined by the sequence of amino acids is represented as the series of frequencies and their amplitudes. The frequencies in IS correspond to the distribution of the structural motifs with defined physicochemical properties determining a biological characteristics of a protein. When comparing proteins, which share the same biological or biochemical function, the ISM technique allows detection of code/frequency pairs that are specific for their common biological properties, or which correlate with their specific interactions.

Phylogenetic analysis
The algorithm of the ISM-based phylogenetic analysis, which was used for the assessment of the biological effect of mutations in SP1 proteins from SARS-CoV-2, was previously described in detail. Figure 1 provides the schematic presentation of this algorithm. Here, we used an ISM distance measure $d$ defined on the ratio between specific frequencies $F(0.257)$ and $F(0.479)$ in IS of SP1 protein, which characterize its interaction with host factors.

Results and discussion
Figure 2 presents the ISM-based phylogenetic tree (the high resolution tree is given in Extended data: Figure IS) calculated for 1643 nonredundant SP1 proteins from SARS-CoV-2
Figure 1. The schematic representation of the ISM-based phylogenetic algorithm ISTREE.

Figure 2. The ISM-based phylogenetic tree of the non-redundant SP1 proteins from SARS-CoV-2 viruses deposited in GISAID from January to October 2020.
viruses collected from January to October, 2020. Analyzed SP1s in this tree are grouped into four separated clusters. The largest cluster A, encompassing 78% of analyzed proteins, contains two sub-clusters, which includes the Wuhan-Hu-1 reference sequence for the Spike protein that is also the basis for different vaccine candidates (sub-cluster A2) and viruses with SP1 with the most abundant mutation D614G (sub-cluster A1). According to the IS concept, proteins that are grouped in the same cluster in the ISM-base phylogenetic tree have similar interacting and immunological profiles. This indicates that a vaccine based on SP from the Wuhan-Hu-1 reference strain will be effective against viruses that are grouped together with this virus in the sub-cluster A2. It was suggested that a single vaccine candidate based on the reference SARS-CoV-2 strain would likely match most of currently circulating variants, including the mutant D614G. Recently, it also showed that sera from D614-infected hamsters exhibit modestly higher neutralization titers against G614 virus than against D614 virus, indicating that this mutation may not reduce the ability of vaccines in clinical trials to protect against COVID-19. This suggests that viruses which are grouped together with the mutant D614G in the sub-cluster A1 would also be responsive to the vaccine candidates that are based on the Wuhan-Hu-1 reference strain. In contrast to the cluster A, viruses that are grouped into other three small clusters B, C and D do not perfectly match the vaccine virus and might be resistant to the vaccine.

In Table 1, the distribution by months of non-redundant SP1 from potentially vaccine responsive SARS-CoV-2 is reported. These data suggest that VE of the vaccine candidates, estimated on the base of the IS parameters would be between 76% and 95%. This estimation of VE is limited with the assumption that viruses from the cluster A dominate among currently circulating SARS-CoV-2. Despite this limitation, the predicted VE is in accord with the results of the recently completed Phase-3 clinical trial of the candidate COVID-19 vaccine. This study, which enrolled 43,538 subjects, showed that the tested vaccine candidate to be more than 90% effective in preventing COVID-19 in participants.

Although potentially vaccine non-responsive viruses in the clusters B, C and D represent only 22% of all 1,643 viruses with non-redundant SP1, these viruses could negatively affect VE. Mass vaccination, which is expected in the near future, creates an immune pressure in the population that would lead to increase in mutations in SP1 as well as in natural selection and preservation of escape mutants. Moreover, this fraction of mutant viruses that can escape vaccine protection and be efficiently transmitted could significantly increase among circulating SARS-CoV-2 once vaccination starts. For this reason, real-time monitoring of virus evolution with different methods, including the ISM-based phylogenetic tool for detection of “functional” changes in the SP1 of SARS-CoV-2 is needed.

In September 2020 in Denmark, 12 human cases of COVID-19 have been identified with SARS-CoV-2 unique variants associated with farmed minks. Viruses isolated from these patients had a combination of mutations, or changes that have not been previously observed and the implications of the identified changes in this variant are not yet well understood. Preliminary findings indicate that this particular mink-associated variant identified in both minks and the 12 human cases could negatively affect VE because mutations in these viruses moderately decreased sensitivity to neutralizing antibodies. Danish authorities have undertaken different actions to limit the further spread of this variant of the virus among mink and human populations, including culling of all farmed mink in Denmark (more than 17 million).

In order to assess the effect of these mutations on VE we calculated the ISM-phylogenetic tree for 5 non-redundant SP1 sequences from SARS-CoV-2 viruses from mink deposited in GISAID (Figure 3). As presented, four of five viruses are grouped with the vaccine virus suggesting that these variants would not significantly affect VE of the COVID-19 vaccines that are based on the SP from the Wuhan-Hu-1 reference strain. The virus with the combination of mutations G261D and Y453F is out of this cluster and it could be potentially resistant to these vaccines. This suggests that further spread of this variant should be carefully monitored. Of note is that the virus with single mutation Y453F is grouped with the vaccine virus indicating that this variant will not significantly affect VE although this mutation is located in RBD. Contrary, this mutation in combination with G261D could decrease responsiveness to vaccine. This result also points out that the assessment of biological effect of mutations should be included in analysis all mutations that are present in protein.

Conclusions
The presented analysis of SP1 proteins, performed by the electronic biology tool, suggest (i) that vaccine candidates based on the Wuhan-Hu-1 reference strain would be effective against most of currently circulated SARS-CoV-2 variants, and ii) further

<table>
<thead>
<tr>
<th>Month</th>
<th>Number of viruses</th>
<th>Vaccine responsive viruses [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>All nonredundant SP1</td>
<td>1643</td>
<td>78</td>
</tr>
<tr>
<td>February</td>
<td>42</td>
<td>95</td>
</tr>
<tr>
<td>March</td>
<td>431</td>
<td>86</td>
</tr>
<tr>
<td>April</td>
<td>525</td>
<td>83</td>
</tr>
<tr>
<td>May</td>
<td>360</td>
<td>86</td>
</tr>
<tr>
<td>June</td>
<td>370</td>
<td>76</td>
</tr>
<tr>
<td>July</td>
<td>322</td>
<td>78</td>
</tr>
<tr>
<td>August</td>
<td>329</td>
<td>87</td>
</tr>
<tr>
<td>September</td>
<td>333</td>
<td>92</td>
</tr>
<tr>
<td>October</td>
<td>121</td>
<td>76</td>
</tr>
</tbody>
</table>
ISM-based monitoring of the evolution of SARS-CoV-2 is important for identification of other mutations, which could affect the effectiveness of vaccines against this virus. Therefore, the scientific community needs to be proactive in submitting genetic sequences into databases such as GISAID. For example, we have States such as Wisconsin, South and North Dakota with low mortality rate 0.8, 1.2, 1.2, respectively (GISAID accessed on November 39, 2020), but no ability to check genetic sequences of viruses from those States as none were deposited. It is very important to support genetic data collection and to share that data in real time in order to organize successful monitoring using different approaches including the one published in this paper.

Data availability
Sequence data of the viruses were obtained from the GISAID Database. To access the database each individual user should complete the “Registration Form For Individual Users”. This form, together with detailed instructions, are available on the website. After submission of the Registration form, the user will receive a password. There are no other restrictions for access to GISAID. Conditions of access to, and use of, the GISAID Database and Data are defined by Terms of Use.

Extended data
This project contains the following extended data:
- Figure 1S

Data are available under the terms of the Creative Commons Zero “No rights reserved” data waiver (CC0 1.0 Public domain dedication).

Acknowledgement
This work was supported by the research grant from Agricultural Department of North Dakota, USA, to Biomed Protection ND (Grant Number 20-342A2).

References
PubMed Abstract | Publisher Full Text | Free Full Text

http://www.doi.org/10.7910/DVN/JLXWWP

PubMed Abstract | Publisher Full Text | Free Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

PubMed Abstract | Publisher Full Text | Free Full Text

Publisher Full Text

PubMed Abstract | Publisher Full Text
The article by Veljkovic et al. describes the application of an *in silico* method previously developed by the group to monitor the "functional" evolution of viral sequences based on the informational spectrum method (ISM). This method is based on the primary amino acid sequences of proteins, from which informational spectra are being derived that ultimately inform about protein-protein interactions. This method allows a rapid assessment of functional alterations in evolving proteins like changes in antigenicity or receptor usage and has been validated in previous studies of the Veljkovic group. A key study was the validation of mutations in the hemagglutinin sequences of H5N1 Influenza viruses predicted by ISM to alter receptor specificity in functional in vitro studies by Schmier et al. (2015), which should be mentioned in the introduction.1

Here this method is applied to analyze more than 1600 non-redundant sequences of the spike glycoprotein of the SARS-CoV-2 actually circulating and deposited in the GISAID database over time to derive their immunological profiles and to detect potential escape from the actual vaccine strains. The derived phylogenetic tree uncovers 2 major clusters corresponding to the Wuhan-Hu1 reference group and a derived D614G mutant covered by the actual vaccines and 3 minor clusters representing 22% of sequences, of which vaccine coverage is not known. These results nicely show how potential candidates for vaccine escape mutants can rapidly be identified by the ISM method, which can then be proven experimentally. The authors also identified by ISM a subgroup of mink SARS-CoV-2 sequences transmitted to humans in Denmark, which represent outliers compared to the actual vaccine strains and should thus be watched carefully in the future. Overall, this paper nicely underlines the value of the ISM method, which can rapidly be performed on a huge set of related sequences to identify potentially "dangerous" outliers, which may be functionally altered in terms of virulence or antigenic escape. This sets the ground to focus experimentally on these outlier strains for functional implications of the predicted mutations, thus limiting long-lasting experimental work to essential strains. As such, the manuscript potentially contributes essentially to the control of the most acute viral health problem we are actually facing.

The paper is generally well written, minor linguistic issues should be corrected in the manuscript.

Minor issues:

- **Abstract**: "....most vaccines currently in clinical applications or under development"; "Here
we present computational analysis of more than...”; “effective against most of currently circulating SARS-CoV-2 viruses. However, further monitoring...”.

- **Introduction**: “…in silico analysis of SP, which can be performed much more rapidly and with minimal experimental knowledge at the time of analysis. Furthermore, in a proof of principle study, it was shown that HA mutations predicted in silico to alter avian receptor specificity of H5N1 Influenza A viruses indeed result in enhanced specificity for human receptors (Schmier *et al.*, Scientific Reports 2015).

- **Results**: In Fig. 2 it should be indicated, what percentage of sequences has been assigned to the different groups (A1, A2, B, C, D). Also in Table 1, it should be indicated in an additional column, what percentage of sequences belongs to the minor groups (B, C, D) over the months to detect a potential increase over time.

**References**


---

**Is the work clearly and accurately presented and does it cite the current literature?**

Yes

**Is the study design appropriate and is the work technically sound?**

Yes

**Are sufficient details of methods and analysis provided to allow replication by others?**

Yes

**If applicable, is the statistical analysis and its interpretation appropriate?**

Not applicable

**Are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions drawn adequately supported by the results?**

Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Virology, HIV-1, Influenza Viruses, vaccine development, virus entry, antigenicity

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com