Contents lists available at ScienceDirect

Journal of Virological Methods

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

A novel method for conserved sequence extraction with prospective mutation prediction for SARS-CoV-2 PCR primer design

Saygın Hüseyin Portakal [a], Beyza Kanat [a], Murat Sayan [b,c], Burak Berber [d], Osman Doluca [a,*]

[a] Izmir University of Economics, Faculty of Engineering, Department of Biomedical Engineering, Izmir, Turkey
[b] Kocaeli University, Faculty of Medicine, Clinical Laboratory, PCR Unit, Kocaeli, Turkey
[c] Turkish Republic of Northern Cyprus
[d] Eskisehir Technical University, Faculty of Science, Department of Biology, Eskisehir, Turkey

ARTICLE INFO

Keywords:
Conserved sequence
COVID-19
SARS-CoV-2
Evolution
Probe design
Primer design
Phylogenetics
Mutation prediction

ABSTRACT

While the whole genomic sequence of SARS-CoV-2 had been revealed, it was also demonstrated that the genome of SARS-CoV-2 exhibits identity with the genome of SARS-CoV and MERS-CoV with ratios of 80 % and 50 % respectively. In the light of SARS-CoV-2 infection and mortality data, diagnosis and treatment of COVID-19 came into prominence around the world. As such many RT-PCR kits have been developed by biotechnology scientists. However viruses are fast mutating organisms and in order to increase accuracy, feasibility in long term and avoid the off target results of RT-PCR assays, regions of viral genome with low mutation rate and designing of primers targeting these regions are quite important. In this scope, we are presenting a novel algorithm that could be used for finding low mutation rate regions of SARS-CoV-2 and primers that were designed according to findings from our algorithm in this study.

1. Introduction

Since recently, our world has been combating an outbreak initiated with the observation of a simple viral-pneumonia at Wuhan city of China in late December of 2019 (Peng et al., 2020). The analysis demonstrates that this outbreak is caused by a novel coronavirus (nCov) belonging to the *Coronaviridae* family (Kooraki et al., 2020). While the World Health Organization (WHO) named the disease caused by nCov as COVID-19, the International Committee on Taxonomy of Viruses (ICTV) named the nCov as Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) (Ather et al., 2020; Wu et al., 2020). The previous outbreaks that are severe acute respiratory syndrome (SARS) and the Middle East respiratory syndrome (MERS) originated by coronaviruses were defined as great threats now that they exhibit high death rate (Xie and Chen, 2020). During the writing of this paper, 5,807,015 coronavirus cases have been detected with 357,800 deaths around the world. This data indicates that SARS-CoV-2 is causing quite a large amount of death with a higher infection rate despite having a lower mortality rate which was discovered as 5 % according to SARS-CoV (Jiang et al., 2020).

Coronaviruses are sort of RNA viruses that possess single stranded RNA molecules as the genetic material with a length of approximately 26−32 kb (Mousavizadeh and Ghasemi, 2020). There are four subgroups of the *Coronaviridae* that are alpha (α), beta (β), gamma (γ), and delta (δ) (Shereen et al., 2020). In the genomic scope, the structural proteins of viruses are encoded by four specific genes that are called spike (S), membrane (M), envelope (E), and nucleocapsid (N) (Grifoni et al., 2020). It is revealed that the receptor binding domain of spike of SARS-CoV-2 recognizes and uses angiotensin-converting enzyme 2 (ACE2) in infecting host cells just as SARS-CoV (Q. Wang et al., 2020). Furthermore, it is also discovered that SARS-CoV-2 binds to ACE2 with more than 10 fold affinity according to SARS-CoV due to the homologous recombination observed on spike glycoprotein (Y. Yang et al., 2020). The SARS-CoV-2 shows higher infection rate in comparison to previous SARS-CoV and threatens human health around the world is highlighted by many researches. Under the light of this information the mutation analysis on coronavirus genomes has come into prominence in order to both detect SARS-CoV-2 infection and develop treatment approaches targeting the genome of the virus.

As the most widely used diagnostic approach, RT-PCR utilizes the RNA isolated from upper and lower respiratory specimens that is reverse transcribed to cDNA and subsequently amplified. During the amplification method, the primers amplify a target region and probe anneals to a selected target sequence located between the forward and reverse primers. Often the diagnostic kits are designed with an accompanying

[*] Corresponding author.
E-mail address: osman.doluca@ieu.edu.tr (O. Doluca).

https://doi.org/10.1016/j.jviromet.2021.114146
Received 10 July 2020; Received in revised form 29 March 2021; Accepted 30 March 2021
Available online 1 April 2021

probe which is degraded during the extension step of the PCR cycle, due to the 5' nuclease activity of *Taq* polymerase, causing the reporter dye to separate from the quencher. A wide range of targets may be used for RT-PCR based diagnosis, but *RdRp*, E and N genes are the most common targets used in many diagnostic kits. However like the rest of the viral genome, these targets are also prone to mutations resulting in lower efficiency in primer binding and even complete fall of the primers and probes from respective target regions.

Considering that the viruses are fast mutating organisms, the high mutation rate of viral genomes provides a high survival rate to viruses along various conditions (Sanjuán et al., 2010) along with a high chance of any PCR primer to lose their specificity. This represents an important problem for the scientists who work on designing detection methods for the viral genomes (Peck and Lauring, 2018). As such, the identification of conserved sequences is a significant approach to overcome this limitation. Many algorithms identifying conserved sequences of viral genomes were developed and published in literature over the years. For instance, Sadeque and his colleagues developed an algorithm which is called as JaPaFi in 2010 in order to detect conserved regions of DNA molecules including poxvirus promoter elements (Sadeque et al., 2010). In addition, Upton and his colleagues developed another algorithm which is called as POCs, in order to detect conserved families of poxviruses genes (Ehlers et al., 2002). However, all these algorithms require the existence of sequencing data to detect mutations that have occurred and do not make estimation of future possible mutations. Since the sequencing is not a routine work in many laboratories, novel mutations are often discovered too late and false negative results may be reported after PCR applications until the mutation is identified, often requiring complete redesign of the primer/probe sets.

In this scope, the discovery of regions that did not only record less observed mutations, but also that are likely to resist potential new mutations which are yet to occur, is important, especially in pandemic situations. As such, we have developed an algorithm and a process to identify regions that are less likely to undergo mutations for viral genomes, in particular, of the SARS-CoV-2. Briefly, a consensus sequence was generated using the SARS-CoV-2 genome and calculated a *mutability score* for each nucleotide depending on the impact of the change on the coded protein. The exchanging frequency of the nucleotides was revealed by the nucleotide substitution model using Mega X. Afterwards, using BLAST + and RefSeq databases, the open reading frames and coding sequences of consensus sequences have been identified. Using aforementioned data and the BLOSUM100 matrix we predicted the impact of all amino acid changes. Iterating through different possible mutations a *mutability score* was calculated for the protein-coding sequences and filtered to yield a low mutable sequence. Using Primer3 software, primers targeting the conserved regions were detected and the efficiency of the algorithm was analyzed statistically by comparing with new mutations. The results demonstrated that our developed algorithm provides conserved regions with low mutability, eliminating up to 78 % of new mutations, in order to guide design appropriate primers to utilize in PCR for detection of SARS-CoV-2.

2. Algorithm

The algorithm requires multiple sequence alignment of the available target genome or gene sequences as an input. In this study all multiple sequence alignments were performed by and downloaded from NCBI Virus (Virus Variation Resource - improved response to emergent viral outbreaks).

The sequence alignment was used for establishing a nucleotide substitution model using MEGA X software and the default parameters (Kumar et al., 2018). The substitution model with the lowest BIC score was selected to be used for further analysis. For complete COVID-19 sequences obtained until 14/04/2020, the general time-reversible model with a proportion of invariable sites and rate of variation across sites (GTR + G+I) showed the lowest BIC score (Miura, 1986; Shoemaker and Fitch, 1989; Z. Yang, 1994).

In parallel, the multiple sequence alignment was used to generate a consensus sequence. Simply, any nucleotide where a substitution had occurred at least a given minimum percent identity was replaced by "N" to generate the consensus sequence. The purpose of creating a consensus sequence is to preliminarily eliminate any nucleotide position with an already recorded variation. In this study a minimum percent identity was chosen as 95 %. This is an arbitrarily chosen threshold and implies that any nucleotide would be replaced by "N" only if at least 5 % of the given sequences are different at that position. Otherwise, the most abundant nucleotide will remain in the consensus sequence. The users may alter this threshold depending on the requirements of the situation. However, behind the choice of this value lies that similarity between the homologous proteins of SARS-CoV-2 and SARS-CoV do not exceed ~96 % while similarity varied dramatically between 32 % and 90 % for most homologous pairs (Cagliani et al., 2020).

The open reading frames along the consensus sequence were found using the tblastn tool of BLAST + software (Camacho et al., 2009) and protein-coding sequences from RefSeq database (**NC_045512.2**). In order to take amino acid changes into account the BLOSUM100 matrix is chosen since it is a better representative for closely related proteins, and we are interested in immediate changes (Henikoff and Henikoff, 1992).

For each nucleotide inside the open reading frames, its substitution may result in a change in amino acid sequence. Such a change can alter the structure and activity of the protein, and needs to be tolerated structurally and functionally for the virus to survive and proliferate. Thus the change in the nucleotide sequences may have varying amounts of impact on the protein. As protein substitution matrices, such as BLOSUM100, represent the recorded frequency of mutations of known proteins that survived evolutionary processes, it also implies how well an amino acid substitution may be tolerated. Thus there is a relation between the position and type of nucleotide substitutions and tolerability of the resulting amino acid changes. This can also be referred to as the mutability of the nucleotide. The algorithm assumes any deleterious mutations would not be tolerated and survive, thus these mutations are ignored.

To quantify the mutability for each nucleotide of the consensus sequence, the frequency of a particular nucleotide substitution was multiplied by the corresponding value in the BLOSUM100 matrix that matches to the substitution from the wild type amino acid at the codon that the nucleotide is found in, into the amino acid resulting due to the nucleotide change. The products resulting from different substitutions from the same nucleotide were then summed and recorded into an array with a size of consensus sequence, the *mutability profile*. This profile contains a *mutability score* for each nucleotide, given that the nucleotide is inside an open reading frame and its codon does not already contain an observed mutation. Otherwise, any nucleotides outside the open reading frames or inside a codon with a known mutation do not receive a *mutability score* and corresponding position in the *mutability profile* was left empty. This profile represents how tolerable a mutation at such a position would be by its corresponding protein (Fig. 1).

The Eq. (1) given below summarizes the calculation of mutability score (*MS*) at a given position *i* in the mutability profile. Unless the position is conserved (*Con*) and located inside a protein-coding sequence (*CDS*), the score is *Null* indicating there is no pressure against protein-level mutations. Otherwise, the mutability score is calculated by the sum of multiplication of nucleotide sequence frequency ($F[n_0,n]$) between the wild type nucleotide (n_0) and a nucleotide from the nucleotide list N={A,T,C,G} (n), and amino acid substitution score ($S[a_0,a']$) between the wild type amino acid coded by the codon corresponding to the nucleotide n_0 at position i (a_0) and the amino acid coded by n (a') whether n equal to n_0 or not. F is obtained from optimal the nucleotide substitution frequency model and S is an amino acid substitution score matrix, such as BLOSUM100.

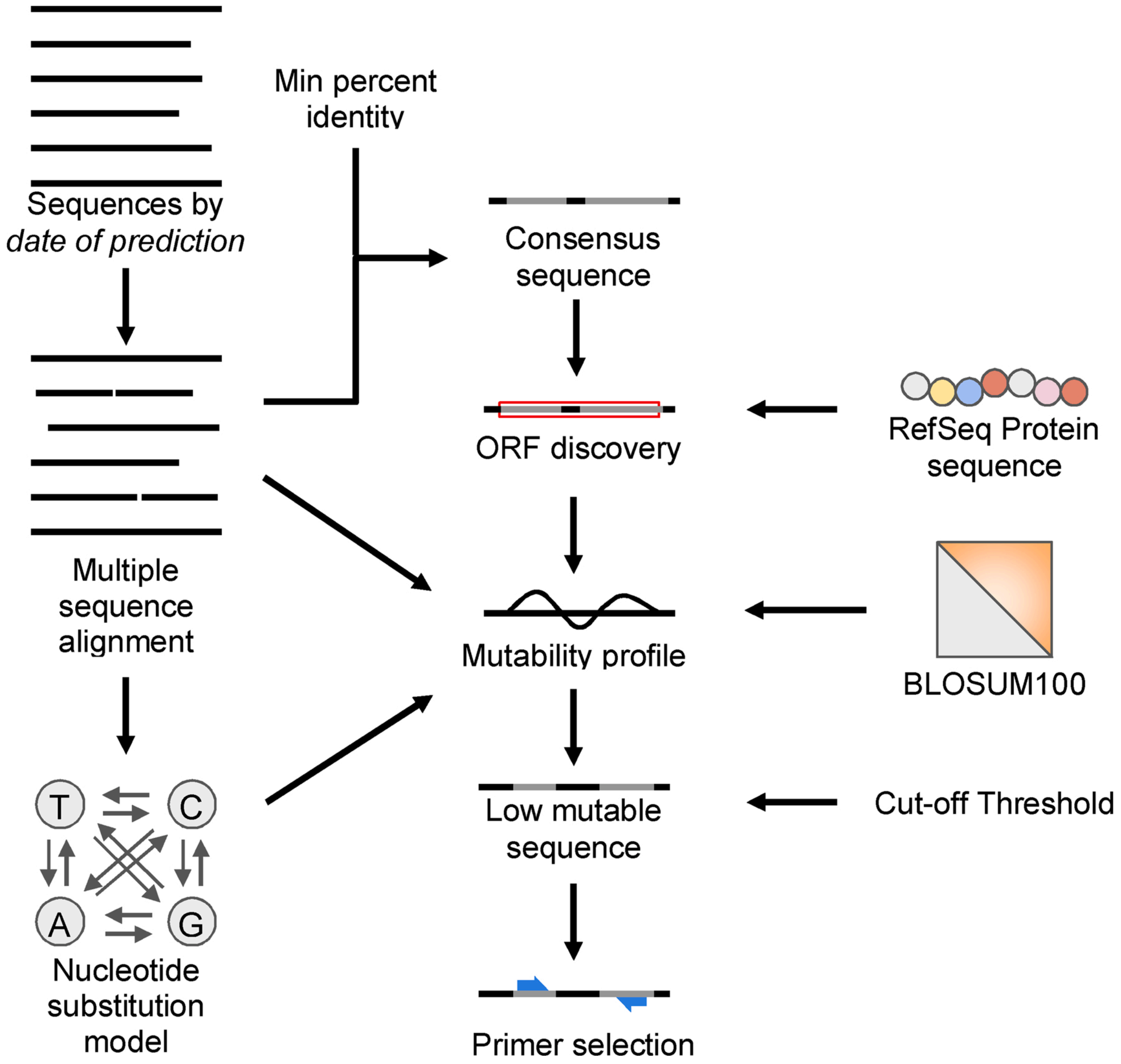

Fig. 1. The flowchart for the primer selection process using BLOSUM100 amino acid substitution matrix.

$$MS_i = \begin{cases} \sum_{n \in N} F[n_0, n] \times S[a_0, a'] & \textit{if i in CDS and Con} \\ \textit{Null} & \textit{if i not in CDS and Con} \end{cases} \quad (1)$$

Briefly, MS_i represents the possibility of mutation at a particular position. The rationale behind the equation is that, for each position this possibility is contributed by the chance of a mutation to occur in form of $F[n_0,n]$ and the chance of that mutation to be conserved as a codon $S[a_0, a']$. However, this predicts only evolutionary pressure on the protein, not any secondary nucleic acid topologies, thus we would have to ignore any mutation outside the CDS.

Mutations inside codons, especially if located at the end of a codon, may be silent and do not impact the resulting protein. However, any mutation along a primer binding site may have varying degrees of impact on its ability to bind depending on its position. For that reason, the PCR primers should be chosen from low mutable regions. This will also lower the chance of the primers becoming ineffective due to novel mutations. While no region is completely resistant to mutations, low mutability regions can be extracted by determining a cut-off threshold. By replacing nucleotides above the cut-off with "N", a new sequence is generated containing only less mutable non-N residues. The extracted sequence can then be used for primer design as the target sequence using different primer-design tools, ie. primer3 (Kõressaar et al., 2018).

3. Materials and methods

All sequences were downloaded from NCBI Virus that are released prior to April 14th and contain a complete genome. The list of all accession numbers are available in the supplementary info. Multiple sequence alignments were done by NCBI Virus Align tool and estimation of a nucleotide substitution model was performed by Mega X using Neighbor-joining method with maximum likelihood as statistical method and the substitution type set to nucleotide. The best substitution model was selected based on the lowest BIC score. The rest of the analysis was performed using Python code written in Python 3.8. All codes are available at https://github.com/odoluca/low_mutable_sequence_extraction. For the discovery of open reading frames RefSeq protein sequences were used with an accession number of NC_045512.

3.1. Stretch analysis

Stretch analysis simply looks for continuous stretches of non-N residues inside the low mutability regions for usable primer binding targets. Since the low mutability regions are determined by both the *mutability profile* and cut-off, we have compared different cut-off thresholds to observe the change in the number of continuous non-N stretches with lengths between 18 and 24 nt versus cut-off threshold.

3.2. Primer3 analysis

Since there is more to primer design than only right length, we decided to use low mutability regions in primer picking by Primer3 software. The change in statistical analysis obtained from online primer3 web server versus different cut-off thresholds were recorded. Default parameters set by the web server were used.

3.3. Z score calculation

To test the success of the predictions, we compared a set of newly observed mutations that were not included in the prediction dataset to randomly generated mutations. The testing procedure was as follows. Aligned sequences were separated into two groups according to a picked release date, which will be referred to as the *date of prediction*. The sequences released prior to any chosen *date of prediction* were taken and used to generate a *mutability profile* as described above. The rest of the sequences released *post-date*, were used to obtain consensus sequence using a minimum percent identity of 100 % to identify any new mutation. The positions of mutations were recorded in a *post-date mutation profile*. Using the *post-date mutation profile*, 300 *random mutation profiles* were generated by shuffling mutations only between positions where a corresponding *mutability score* exists in the *mutability profile*. The purpose of using mutations with only a corresponding *mutability score* is to avoid a bias due to high mutation count in the sections of the sequence with already observed mutations. We also made sure the number of mutations *post-date mutation profile* and *random mutation profiles* remained equal (Fig. 2).

A *mutability score sum* was obtained for each *random mutation profile* as well as the *post-date mutation profile*, by adding *mutability scores* of corresponding mutations in these profiles. A normal distribution was obtained using the mean and standard deviations of the *mutability score* sums of *random mutation profiles*. A z-score and a p-value was obtained for the *post-date mutation profile* with respect to the distribution using z test.

4. Results

A multiple sequence alignment of all complete nucleotide sequences available by 04.14.2020 at NCBI Virus belonging to Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, taxid: 2697049) was obtained using NCBI align tool. The multiple sequence alignment was then downloaded and used to extract a consensus sequence using custom python code and estimate a nucleotide substitution model using MEGA-X software as described in the algorithm section.

The protein RefSeq sequence (**NC_045512**) was then aligned on the consensus sequence using the offline Blast tool to discover the open reading frames. Using the nucleotide substitution model, open reading frames, and BLOSUM100 amino acid substitution matrix, a *mutability score* was calculated for each nucleotide position and their distribution was plotted. (Fig. 3)

It is important to note that the choice of the amino acid substitution matrix has a significant impact on the score. Positive mutabilities indicate a higher chance of mutation in the future, while lower scores indicate a higher chance of future conservation. The *mutability scores* are calculated only for nucleotides that are conserved and nucleotides inside an open reading frame.

4.1. Stretch analysis

A cut-off threshold was then selected within the range of the distribution. The purpose of the cut-off is to mark nucleotides that are likely to mutate in the future with “N” and to have these positions be disregarded by the primer design tools. However choosing a low threshold would eliminate continuous stretches of conserved sequences required for a primer binding site. Then we have analyzed the number of possible primer binding sites in comparison to varying cut-off threshold values. (Fig. 4) This analysis showed that no continuous stretch of low mutable nucleotides would be available below the cut-off of 0.6. The number of such putative primer targets rise with increasing cut-off. It is important to make the choice of the cut-off to be done in the light of the distribution or such stretch analysis.

Similarly, the use of primer3 tool with low mutable sequences

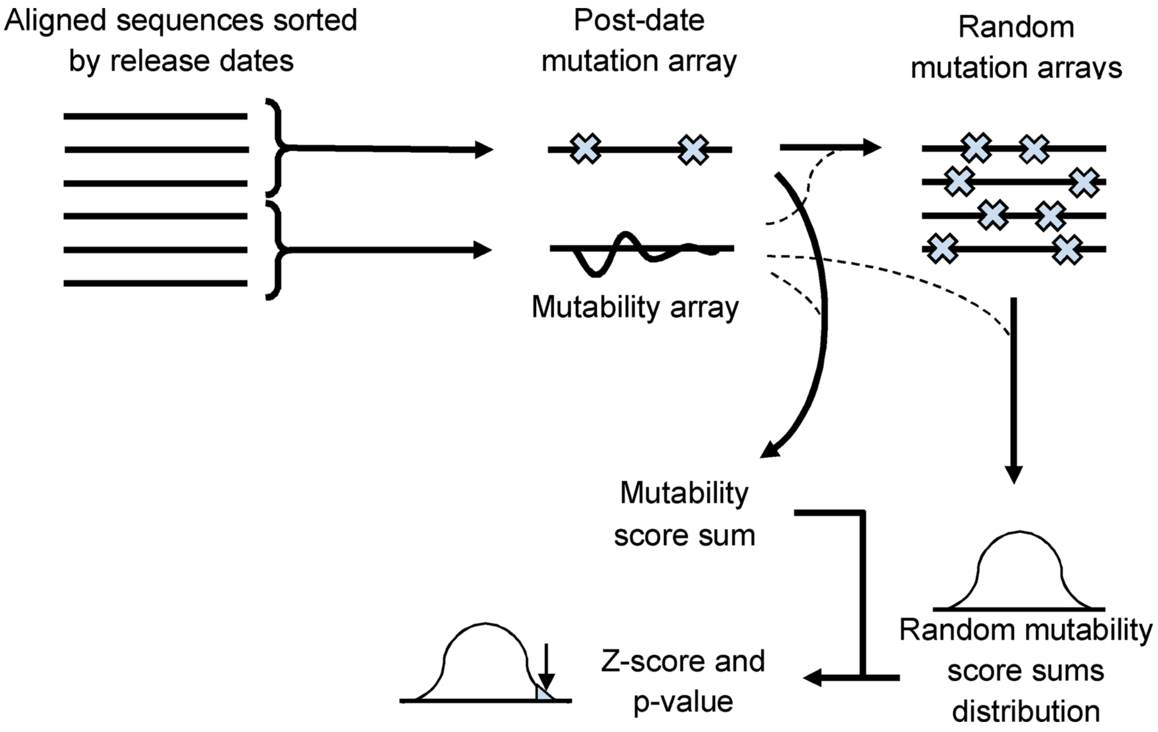

Fig. 2. Evaluation flowchart for evaluation of the algorithm.

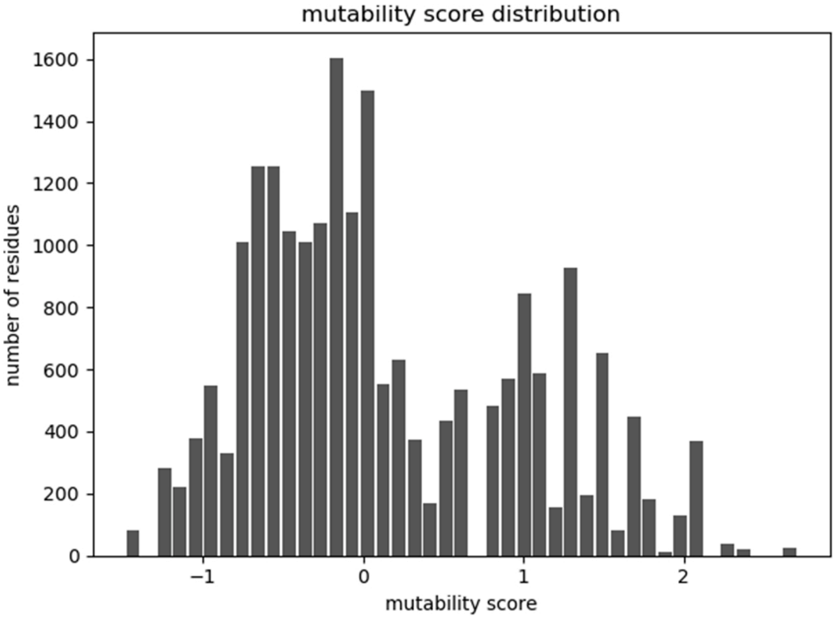

Fig. 3. Distribution of mutability scores calculated from COVID-19 sequences, using GTR + G+I and BLOSUM100 matrix.

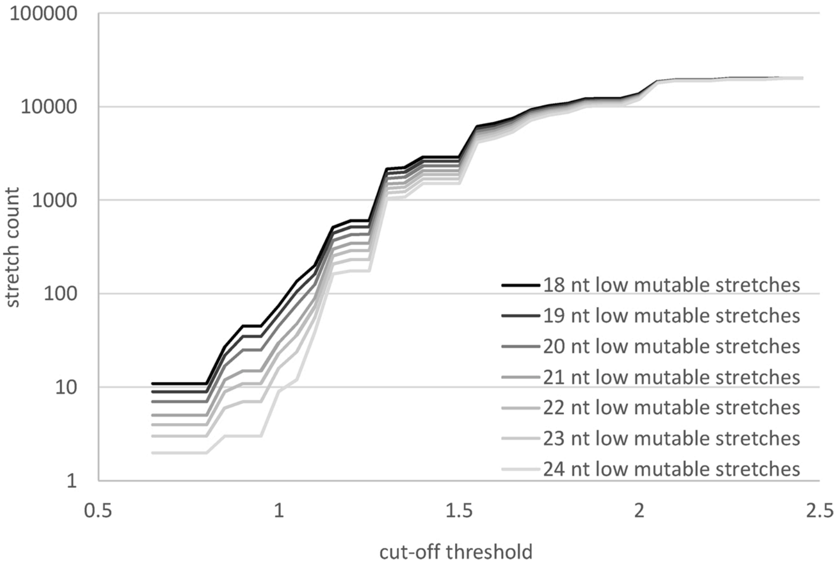

Fig. 4. Number of putative primer binding regions with varying sizes containing only non-N residues (solid grayscale lines) with varying cut-off threshold values.

obtained with varying cut-off thresholds resulted in putative primers in numbers that are correlated with stretch analysis. A cut-off threshold of at least 0.6 was required to obtain any putative primer binding site, eliminating the top 30.7 % of the mutable nucleotides. In addition the primer3 tool managed to pick a potential primer pair with a cut-off threshold as low as 0.9, indicating the applicability of the algorithm eliminating the top 22 % of mutable nucleotides. (Fig. 5)

4.2. *Z score calculation*

To test the success of the algorithm we decided to use it for a set of sequences made available by a previous date, also referred as the *date of prediction*. Conserved sequences and then the *mutability profiles* were determined for different dates of prediction. Comparing it to the sequences made available after those dates, *post-date observed mutations* were determined. For all *post-date mutations*, the sum and the mean of *mutability scores* of corresponding mutation sites were calculated. Similarly *mutability score sums* were also calculated for randomly generated 300 consensus sequences with random mutations. A Z-score and a p-value were calculated using the *mutability score sum* obtained by the *post-date observed mutations* and the distribution of the *sum of the mutability scores* of the mutation sites of the randomly generated sequences. It was assured that the number of random mutations were equal to *post-date observed mutations*. A Z-score above 3 indicates a *mutability score sum* above 99.7 % of the randomly generated mutations and any predicted *mutability profile* with a *mutability score sum* above 3 indicates statistically significant difference. (Figure S2)

Accordingly, starting from the sequences obtained at the first days of the pandemic, reliable *mutability profiles* were obtained. The gradual decrease through April is a result of a low number of samples after the *date of prediction* since the increasing p-value indicates a decreased significance of the Z-score as well. However, undeniably, the emergence of new epidemic centers also increases the variation in the Covid-19 genome, and subsequently, results in a lower prediction power. The

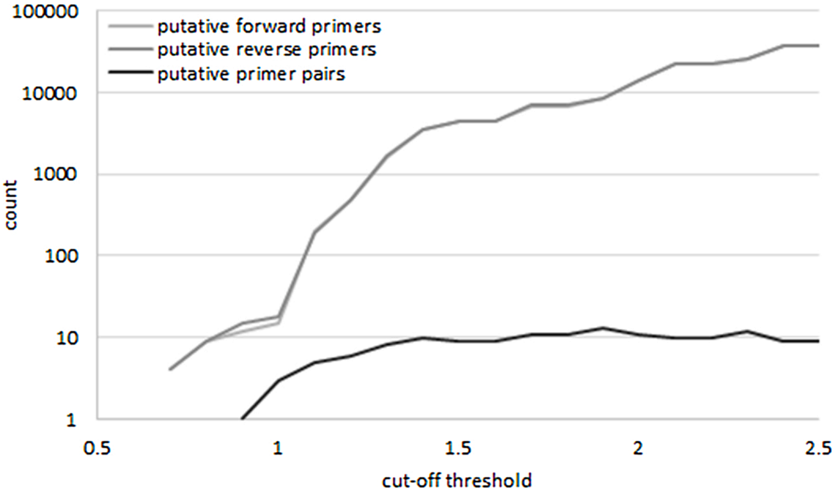

Fig. 5. The number of putative forward (light gray), reverse (gray) primers and primer pairs discovered by primer3 tool for low mutable sequences obtained using different cut-off thresholds.

analysis with newly generated sequence data should improve the predictability (Fig. 6).

A more tangible evaluation of the algorithms may be considered using the low mutable sequence. The number of new mutations that were predicted and the number of missed mutations were considered for a specific cut-off. When a cut-off score of 1.0 was used, throughout the pandemic, consistently around 60 % of new mutations were predicted and filtered out by the algorithm. More importantly, the prediction power rose to 78 % by the end of the tested period, indicating that more recent mutations are more likely to be predicted. (Fig. 7)

Another test is also performed using a random number generator instead of BLOSUM100 matrix. The p-values showed no significant difference from the random *mutability score sums*. (Figure S1) The fact that a random matrix does not yield a statistically significant result, indicates that the substitution matrix in an important element of this algorithm. However, when it comes to choice of substitution matrix, one might argue that different viruses DNA or RNA may have different evolutionary speeds and the choice of matrix may have significant impact. In order to demonstrate this, we have repeated the Z-score calculations using various PAM and BLOSUM matrices. Despite the fact that a random matrix does not yield any significant difference, any of the PAM or BLOSUM matrices labels statistically significant number of putative mutations (Figure S3). This indicates that different viruses with different molecular clocks may be applied without overworking on the choice of a substitution matrix.

To demonstrate how to evaluate already designed primers and probes we have chosen the primer probe pairs made available by the Central for Disease Control (CDC) for research purposes. (Table S1) These include four primers and two probes (excluding the chemically modified versions of the two probes). Then we have calculated the average mutability score for each primer or probe by finding the mean of the mutability scores of the nucleotides that these oligonucleotides are designed to bind on the viral genome. To compare, we have also calculated the average mutability score of all putative primers of same length that could bind to the viral genome. Any putative primer that contains a nucleotide position that does not have a mutability score due to absence of coding sequence or known mutations are discarded. It should be noted that these scores are obtained by averaging small sets of mutability scores of nucleotides found in Fig. 3. The average mutability scores for putative primers remained around 0.133 for BLOSUM100. When compared, it was apparent that N1-For and N2-Probe have an average mutability score above the average of the putative primers indicating that these sequences are relatively more prone to mutations, and there is space for improvement. It was encouraging that the rest

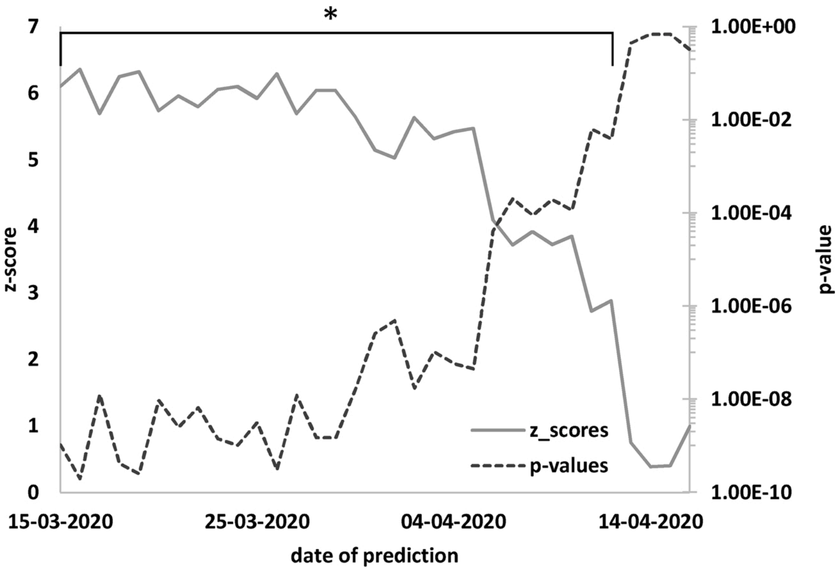

Fig. 6. The Z-scores (solid line) and associated p-values (dashed line) of the sum of mutability scores of post-date mutations with respect to random mutations. The Z-scores indicate the prediction power, as the degree of the difference of the mutability score sum of post-date mutations, from a random pool of mutability score sums obtained from randomized mutations. The asterisk indicates when the sum of mutability scores of mutations observed after the corresponding date of prediction is significantly different from sequences with randomized mutations with a p-value below 0.01.

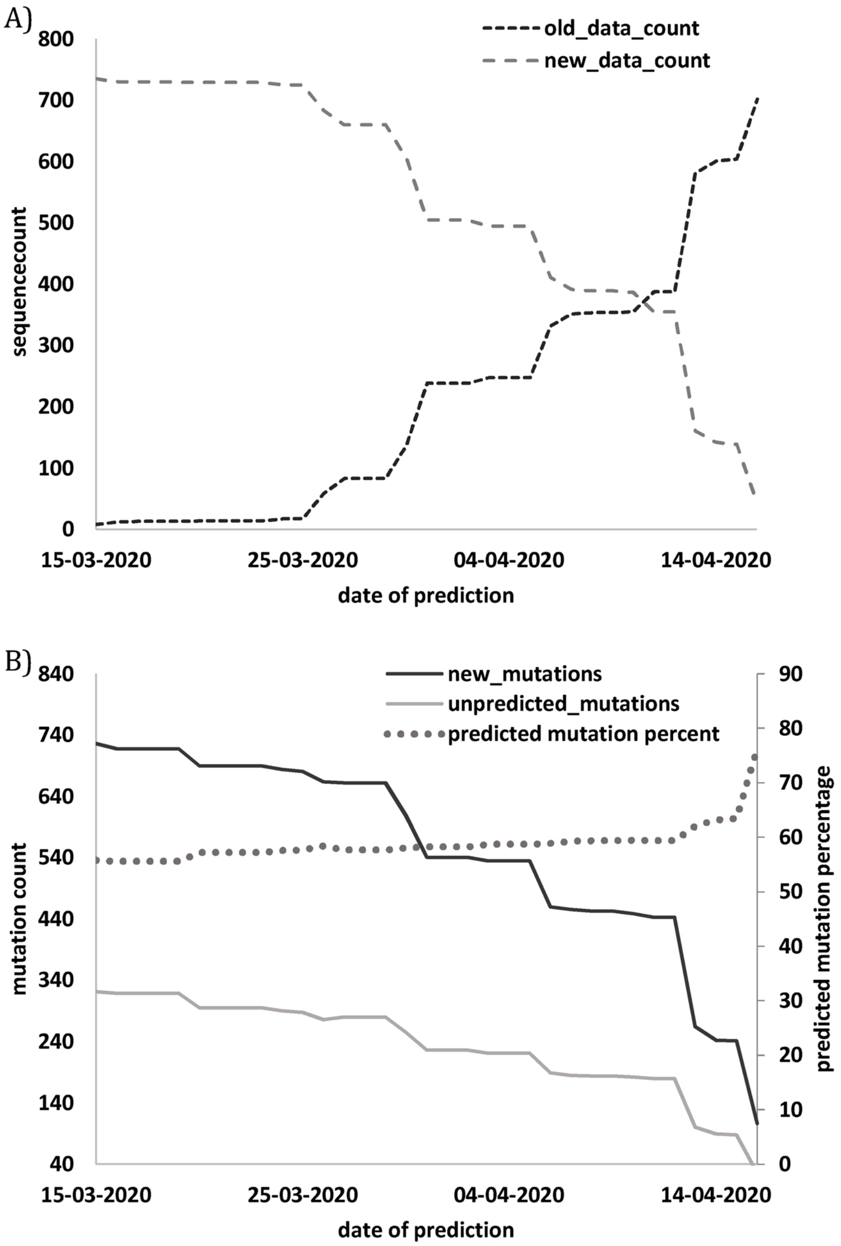

Fig. 7. The number of sequences before and after any prediction date (A, dark and light dashed lines), the number of all and unpredicted mutations in the post-date sequences when a cut-off of 1.0 was used (B, dark and light solid lines) and the percentage of predicted mutations versus all new mutations (B, dotted line).

were above or equal in mutability to the average. Especially N1-Probe showed an average mutability score of 0.040 indicating a better design. However, it should be noted that there is more to primer design than only immutability, such as avoiding primer-dimer formation, optimum GC content, specificity etc.

5. Discussion

Especially considering emerging pandemic situations, the ability to have high prediction power starting from the beginning of pandemic is extremely important to be able to avoid false positives due to unforeseen mutations. Here we introduce a process for the extraction of a low mutable sequence to be used in the design primer and probes for PCR applications. Our algorithm shines especially with its ability to filter out mutations that are yet to be observed even at early stages of the pandemic. When observed forthcoming mutations are compared to randomized mutations the mutability profiles showed significantly higher values throughout the tested period. (Fig. 6) Accordingly, the process showed that low mutable sequences obtained even at early stages of a pandemic may be adequate to avoid approximately 60 % of new mutations to be observed in the upcoming period. (Fig. 7B)

It should be noted that this rate can be improved even more by using a lower cut-off threshold (ie. 0.8), however also decreasing the number of putative primer binding sites as well. The primer3 and stretch analysis showed that possible continuous non-N residues required for designing primers may be obtained with cut-off thresholds (as low as 0.6) while eliminating the nucleotides with the highest mutability scores (up to 30.7 %). (Figs. 4 and 5)

We hope that the algorithm introduced here, will not only help design improved primers and probes for SARS-CoV-2 but also prepare diagnostic kit manufacturers for future epidemics. Especially considering the lack of sequencing data at the early stages of an epidemic, the detection primers may often be required to be modified due to novel mutations. Our algorithm may decrease the frequency of such modifications as shown. Additionally, in the case of the already available primers the algorithm also provides a method for mutation variation analysis as part of the risk assessment for the manufactured kits. However, there is potential for further improvement. It might be possible to improve the output using alternative amino acid substitution matrices. Another possible route may be using algorithms that predict impact of mutations, such as I-Mutant, (Capriotti et al., 2005) instead of an amino acid substitution matrix.

Authorship contributions

O. Doluca: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing – Original Draft, Visualization, Supervision, Project administration

H.S. Portakal: Conceptualization, Investigation, Data curation, Formal analysis, Writing – Original Draft

B. Kanat: Conceptualization, Investigation, Data curation, Formal analysis, Writing – Original Draft

B. Berber: Investigation, Resources, Methodology, Supervision

M. Sayan: Investigation, Resources, Writing – Review & Editing

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

I would like to thank and acknowledge the Scientific and Technological Research Council of Turkey (TUBITAK) for the support in this and other endeavors. (Project no: 216S854)

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jviromet.2021.114146.

References

Ather, A., Patel, B., Ruparel, N.B., Diogenes, A., Hargreaves, K.M., 2020. Coronavirus disease 19 (COVID-19): implications for clinical dental care. J. Endod. https://doi.org/10.1016/j.joen.2020.03.008.

Cagliani, R., Forni, D., Clerici, M., Sironi, M., 2020. Coding potential and sequence conservation of SARS-CoV-2 and related animal viruses. Infect. Genet. Evol. 83, 104353 https://doi.org/10.1016/j.meegid.2020.104353.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10 (1 (December)), 1–9.

Capriotti, E., Fariselli, P., Casadio, R., 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res. https://doi.org/10.1093/nar/gki375.

Ehlers, A., Osborne, J., Slack, S., Roper, R.L., Upton, C., 2002. Poxvirus orthologous clusters (POCs). Bioinformatics 18 (11 (November)), 1544–1545.

Grifoni, A., Sidney, J., Zhang, Y., Scheuermann, R.H., Peters, B., Sette, A., 2020. A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. Cell Host Microbe 27 (4 (April)), 671–680 e2.

Henikoff, S., Henikoff, J.G., 1992. Amino acid substitution matrices from protein blocks. Proc. Natl. Acad. Sci. https://doi.org/10.1073/pnas.89.22.10915.

Jiang, S., Hillyer, C., Du, L., 2020. Neutralizing antibodies against SARS-CoV-2 and other human coronaviruses. Trends Immunol. https://doi.org/10.1016/j.it.2020.04.008.

Kooraki, S., Hosseiny, M., Myers, L., Gholamrezanezhad, A., 2020. Coronavirus (COVID-19) outbreak: what the department of radiology should know. J. Am. Coll. Radiol. JACR 17 (4 (April)), 447–451.

Kõressaar, T., Lepamets, M., Kaplinski, L., Raime, K., Andreson, R., Remm, M., 2018. Primer3_masker: integrating masking of template sequence with primer design software. Bioinformatics 34 (11 (June)), 1937–1938.

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol. Biol. Evol. 35 (6 (June)), 1547–1549.

Miura, R.M., 1986. Some Mathematical Questions in Biology: DNA Sequence Analysis. American Mathematical Soc.

Mousavizadeh, L., Ghasemi, S., 2020. Genotype and phenotype of COVID-19: their roles in pathogenesis. J. Microbiol. Immunol. Infect. (March) https://doi.org/10.1016/j.jmii.2020.03.022.

Peck, K.M., Lauring, A.S., 2018. Complexities of viral mutation rates. J. Virol. 92 (14 (July)) https://doi.org/10.1128/JVI.01031-17.

Peng, F., Tu, L., Yan, Y., Hu, P., Wang, R., Hu, Q., Cao, F., et al., 2020. Management and treatment of COVID-19: the Chinese experience. Can. J. Cardiol. (April) https://doi.org/10.1016/j.cjca.2020.04.010.

Sadeque, A., Barsky, M., Marass, F., Kruczkiewicz, P., Upton, C., 2010. JaPaFi: a novel program for the identification of highly conserved DNA sequences. Viruses 2 (9 (September)), 1867–1885.

Sanjuán, R., Nebot, M.R., Chirico, N., Mansky, L.M., Belshaw, R., 2010. Viral mutation rates. J. Virol. 84 (19 (October)), 9733–9748.

Shereen, M.A., Khan, S., Kazmi, A., Bashir, N., Siddique, R., 2020. COVID-19 infection: origin, transmission, and characteristics of human coronaviruses. J. Advert. Res. 24 (July), 91–98.

Shoemaker, J.S., Fitch, W.M., 1989. Evidence from nuclear sequences that invariable sites should Be considered when sequence divergence is calculated. Mol. Biol. Evol. 6 (3 (May)), 270–289.

Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., Lu, G., et al., 2020. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. Cell 181 (4 (May)), 894–904 e9.

Wu, C., Liu, Y., Yang, Y., Zhang, P., Zhong, W., Wang, Y., Wang, Q., et al., 2020. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. Acta Pharm. Sin. B (February). https://doi.org/10.1016/j.apsb.2020.02.008.

Xie, M., Chen, Q., 2020. Insight into 2019 novel coronavirus - an updated interim review and lessons from SARS-CoV and MERS-CoV. Int. J. Infect. Dis. IJID 94 (May), 119–124.

Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39 (3 (September)), 306–314.

Yang, Y., Peng, F., Wang, R., Guan, K., Jiang, T., Xu, G., Sun, J., Chang, C., 2020. The deadly coronaviruses: the 2003 SARS pandemic and the 2020 novel coronavirus epidemic in China. J. Autoimmun. 109 (May), 102434.

Glossary

GTR+G+I: a commonly used nucleotide substitution model that represents frequencies of mutations.

non-N residue: nucleotides whose base is determined and conserved so that there is no ambiguity.

Nucleotide substitution model: Substitution model describes the frequency of transitions in which a sequence of symbols changes into another set of traits or symbols.

Mega X: The Molecular Evolutionary Genetics Analysis (MEGA) software may be a computing platform implementing many analytical methods and tools for phylogenomics and phylomedicine.

BLOSUM100: In bioinformatics, the BLOSUM (BLOcks SUbstitution Matrix) matrix may be a substitution matrix used for sequence alignment of proteins. BLOSUM matrices are accustomed score alignments between evolutionarily divergent protein sequences.

Primer3: It is a web tool that suggests primer and probes for a variety of PCR applications.

BIC score: Bayesian information criterion (BIC) may be a criterion for model selection among a finite set of models. BIC has been widely used for model identification in statistics and regression toward the mean.

tblastn tool: tblastn is an element of the new blast+ package from the NCBI. tblastn compares a protein query sequence against a nucleotide sequence database dynamically translated altogether six reading frames