ABSTRACT

SARS-CoV-2 is an enveloped positive-sense single-stranded RNA coronavirus that causes COVID-19, of which the current outbreak has resulted in a high number of cases and fatalities throughout the world, even when vaccine doses are being administered. The aim of this work was to scan the SARS-CoV-2 genome in search for therapeutic targets. We found a sequence in the 5'UTR (NC\_045512:74-130), consisting of a typical heptamer next to a structured region that may cause ribosomal frameshifting. The potential biological value of this region is relevant through its low similarity with other viruses, including coronaviruses related to SARS-CoV, and its high sequence conservation within multiple SARS-CoV-2 isolates. We have predicted the secondary structure of the region by means of different bioinformatic tools. We have suggested a most probable secondary structure to proceed with a 3D reconstruction of the structured segment. Finally, we carried out virtual docking on the 3D structure to look for a binding site and then for drug ligands from a database of lead compounds. Several molecules that could be probably administered as oral drugs show promising binding affinity within the structured region, and so it could be possible to interfere its potential regulatory role.

Keywords: SARS-CoV-2; frameshifting; 5'UTR; pseudoknot; molecular docking.

RESUMEN

El SARS-CoV-2 es un coronavirus de ARN monocatenario de sentido positivo envuelto que causa COVID-19, del cual el brote actual ha provocado una gran cantidad de casos y muertes en todo el mundo, incluso cuando se están administrando dosis de vacunas. En este trabajo hemos escaneado el genoma del SARS-CoV-2 en busca de dianas terapéuticas. Encontramos una secuencia en el 5'UTR (NC\_045512:74-130), que consiste en un heptámero típico junto a una región estructurada que puede causar cambios en la pauta de lectura. El valor biológico potencial de esta región es relevante debido a su baja similitud con otros virus, incluidos los coronavirus relacionados con el SARS-CoV, y su alta conservación de secuencia dentro de múltiples aislados de SARS-CoV-2. Hemos predicho la estructura secundaria de la región mediante diferentes herramientas bioinformáticas. Hemos sugerido una estructura secundaria más probable para así proceder al acoplamiento virtual en la estructura 3D para buscar un sitio de unión y luego ligandos de fármacos. Hemos encontrado varias moléculas que probablemente podrían administrarse como fármacos orales muestran una afinidad de unión prometedora dentro de la región estructurada, por lo que es posible que interfieran en su posible función reguladora de la replicación viral.

Palabras clave: SARS-CoV-2, cambios en la pauta de lectura, 5'UTR, pseudonudo, acoplamiento virtual.

INTRODUCTION

On March 11th, the World Health Organization (WHO) declared COVID-19 a clinical pandemic (primarily pneumonia and gastroenteritis) caused by the SARS-CoV-2 virus. As of end October 2021 the pandemic outbreak has caused almost five million deaths worldwide, although almost 7 million peo-
ple are vaccinated. SARS-CoV-2 belongs to the *Coronaviridae* family and is related to SARS-CoV and Middle East Respiratory Syndrome (MERS)-CoV (79% and 50% genomic similarity, respectively). SARS-CoV caused an epidemic outbreak in 2003 and MERS caused an outbreak in 2012 [1]. Those three viruses belong to the *Betacoronavirus* genus. Coronavirus causes zoonotic infections, so they may spill over from a host species to a different one through small changes in their genome. SARS-CoV-2 demonstrated a high genetic similarity (more than 85%) to a virus group known as SARS related coronavirus (SARSr-CoV), which are isolated from animal hosts, including *Hipposideros* bats and pangolins (*Manis javanica*). These species seem to be candidates as intermediate hosts for SARS-CoV-2 [2,3].

These viruses have a positively translated single strand RNA genome and they use programmed −1 ribosomal frameshifting (−1 PRF) to direct the synthesis of immediate early proteins that prepare the infected cell for takeover by the virus. Frameshifting is a smart mechanism for the translation of a genome sequence into two different proteins by moving the translation frame one position in the union between RNA and the ribosome [4]. A typical frameshifting signal has two essential elements: a characteristic heptanucleotide called the ‘slippery’ sequence, at which the ribosome-bound tRNAs slip into the −1 frame, and an adjacent mRNA secondary structure that stimulates this slippage process. The intermediate sequence between these two elements also has a typical size of less than twelve nucleotides. Often the secondary structure is more complex than a simple stem-loop between palindromic sequences, expanding into pseudoknots [5]. In terms of structure, a pseudoknot forms upon the base-pairing of a single-stranded region of RNA in the loop of a hairpin to a stretch of complementary nucleotides elsewhere in the RNA chain.

A set of bioinformatic tools has already been developed to predict these structures [6]. The mechanism of action of pseudoknots is not completely understood; some authors suggest that it appears to be linked to the helicase activity of the ribosome. When pseudoknots are located in coding regions, they modulate the elongation and termination steps of translation: the ribosome is able to switch from the zero reading frame to the −1 frame and translation continues in the new frame. When pseudoknots are in non-coding regions, they act on the regulation of the initiation of protein synthesis and on template recognition by the viral replicase guiding viral replication and packaging [7].

All coronaviruses have been reported to utilize programmed −1 ribosomal frameshifting to control the expression of their proteins. In 2005, Plant et al. [8] identified a threestemmed mRNA pseudoknot inducing an efficient −1 ribosomal frameshift in the SARS-CoV genome. By this mechanism, the virus may produce a fusion protein that overlaps the regions ORF1a and ORF1b. This element encodes an ORF1ab polyprotein involved in ablating the host cellular innate immune response. Mutations affecting this structure decreased the rates of −1 PRF and had deleterious effects on the virus propagation. Recently, Kelly et al. [9] described the same pseudoknot in SARS-CoV-2, and they demonstrated frameshifting. This area is highly conserved between SARS-CoV and SARS-CoV-2, as there is only one single nucleotide difference, a C to A substitution at position 13,533 bp.

The frameshifting regions could be used as a target to fight viral infection [9]. Starting with early studies, point mutations at the slippery sequence have proved to have an important effect on viral replication [8]; thus, they can be also interesting points in the engineering of an attenuated virus for vaccine development. The inhibition of these regions by peptide antisense oligomers was studied by Neuman et al. [10]. After several passages in cell culture, virions escape the inhibition of replication but show attenuated forms. Rangan et al. [11], described highly structured areas of RNA that might be less accessible to complementory oligomers, but these convoluted areas would provide small binding sites for conventional drug molecules; therefore, a combination of scanning for structure and sequence conservation may be appropriate to find therapeutic targets. Previous studies using *in silico* methods found drug-like molecules that would inhibit SARS-CoV replication by action on the frameshifting region at the overlap between ORF1a and ORF1b [12]. The same molecule has been shown to affect replication in SARS-CoV-2 [9].

In this work, we scanned the SARS-CoV-2 genome to seek for novel likely critical areas for virus replication focusing on frameshifting predictors. We explored the likely biological relevance of this feature through the study of sequence conservation and its suitability as a potential drug target by the analysis of the structural properties and the drug docking prediction.

**MATERIAL AND METHODS**

**Genomewide frameshifting signal search.** A prediction of the relevant sequence and structures in the viral reference genome for SARS-CoV-2 (NC 045512) [13] was performed using the KnotInFrame tool [14]. The output determined the sequence and position of slippery sequences and nearby pseudoknots, since both criteria are needed to predict frameshifting. Our focus on a particular region was established by combining KnotInFrame output with biological knowledge. We focused on previously undescribed frameshifting regions and the likely regulatory roles of UTR regions. Once a sequence of interest met these criteria, an inspection of the predicted secondary structure was achieved with additional tools ipknot [15] and RNAfold from Vienna Suite [16]. The secondary structure for the segment of interest in dot bracket notation was chosen from the inspection of the overall conformation of the 5’UTR and assuring to include the slippery region. The likelihood of the secondary structure was assessed by computing the minimum free energy (MFE) of a large number of random sequences of SARS-CoV-2 of the same length as the sequence of interest into mFold, in order to obtain an empirical distribution of MFE and so assess how dominant the proposed structure would be [17].

**Conservation of the sequence of interest.** Sequence conservation was assessed for the sequence of interest deter-
mined in the previous step as a reliable trait of biological relevance. The conservation of the sequence of interest was evaluated in two steps. First, the conservation between SARS-CoV-2 and other human and animal hosted coronavirus genomes was studied by the computation of a cladogram and by the search for the alignment of the sequence of interest against a comprehensive viral database. A total of 21 high quality genomes from coronavirus hosted in humans and other species were selected based on subjective criteria regarding variability and relevant facts to build a cladogram. The genomes were downloaded from GenBank and aligned with Clustal Omega [18] using the default parameters. The cladogram was constructed using a maximum likelihood estimate with FastTree [19], under a GTR model of nucleotide evolution. The package ggtree [20] was used in R [21] to generate the graphic of the cladogram and the multiple sequence alignment (MSA). In addition to this alignment of the SARS-CoV-2 and another 20 coronavirus genomes, the sequence of interest was examined by ViroBLAST [22]. This tool provides a blastn [23] alignment with a comprehensive database of all types of virus, so that we would assess any casual homology with any other virus. Secondly, we evaluated the conservation of the sequence of interest within SARS-CoV-2 isolates from different geographic locations since the onset of the pandemic. We took advantage of the fast contribution of genomes into the GISAID database. We filtered the genomes in the database in order to retain only high quality records (length greater than 29,000 nt and with a low number of undetermined positions). The number of variant site strains was assessed by blastn [23], making the distinction of variants at the whole 5’UTR region (1-265 nucleotide positions); and the number of variants at the position of the sequence of interest. Further individual inspections of mismatched genomes were conducted to ensure whether the variation was not due to technical sequencing reasons.

**Prediction of 3D structure and molecular docking.** Upon consideration of different alternatives, the structure of the sequence of interest in dot-bracket notation and the underlying nucleotide sequence were imported into Rnacomposer [24] to obtain a 3D structure prediction in .pdb format. The file in .pdb format was used as input for the virtual scan for active sites. This task was carried out using Autodock tools suite [25]. This suite comprises the AutoGrid and Autoligand tools for the search of active sites in a molecular 3D structure. A combination of manual selection of the region of interest and automatic search space by the tools was used to obtain the coordinates and dimensions of a putative active site. These data were used as inputs for the molecular docking by the Autodock Vina tool. The virtual docking tested the binding of a set of molecules specially selected for drug screening; the NCBI maximum diversity set II. The affinity of molecules to bind the active site was assessed by the minimum free energy in Kcal/mol.

**RESULTS**

**Frameshifting prediction.** We used KnotInframe program [14] to detect cis-acting signals, the nucleotide sequence and the position of the heptameric slippery sites and near pseudoknots as prediction for −1 PRF. A list of genomic regions of SARS-CoV-2 (NC 045512.2) where a frameshifting signal was predicted by the KnotInFrame program is shown in Table 1. The stability of every predicted structure is also indicated by the MFE value, on the rightmost column. A more negative value of MFE represents a more stable and likely to be a functional structure. This value is mainly dependent on the length of the sequence. The pseudoknot we propose associated with the pattern sequence at position 76 (UUUAAAA) that was identified as −1 PRF, is ranked fourth and exhibit the lowest MFE

### Table 1

| Slippery sequence | Slippery pos. | Pseudoknot start | Pseudoknot end | Length | Deltarel | MFE  |
|-------------------|---------------|------------------|----------------|--------|----------|------|
| TTTAAAC           | 13462         | 13469            | 13549          | 80     | 0.126    | -34.80 |
| GGGTTTA           | 4261          | 4268             | 4328           | 60     | 0.092    | -15.60 |
| AAATTITG          | 6071          | 6078             | 6158           | 80     | 0.076    | -16.10 |
| TTTAAAA         * | 76            | 83               | 123            | 40     | 0.070    | -14.00 |
| GGGITTTI          | 13348         | 13355            | 13475          | 120    | 0.051    | -34.90 |
| GGTTTTIT          | 8183          | 8190             | 8270           | 80     | 0.049    | -15.10 |
| TTTAAAT           | 4264          | 4271             | 4331           | 60     | 0.047    | -15.60 |
| CCAAAA            | 20646         | 20655            | 20773          | 120    | 0.046    | -29.00 |
| TTTAAAA           | 6514          | 6521             | 6621           | 100    | 0.038    | -19.40 |
| TTTAAC            | 20817         | 20824            | 20924          | 100    | 0.035    | -30.20 |
| TTTTTIT           | 11076         | 11085            | 11183          | 100    | 0.035    | -19.60 |

*The line in bold face was chosen as the sequence of interest. MFE: minimum free energy*
different programs in dot bracket notation. The IPknot tool [15] has been used in two fashions, firstly by the only input of the sequence of interest and secondly, by the input of the whole 5'UTR region, and then cutting out the prediction for positions 7:130. In both cases, IPknot predicts a knotted structure just downstream of the slippery region (the pattern of opening and closing brackets do not match, meaning that stem-loops bind to outer regions). Interestingly, the prediction we obtained using IPknot fully agrees with the prediction obtained recently in that genomic area using the Rosetta tool [11].

Similarly, the secondary structure of the whole 5'UTR regions was also obtained by IPknot tool and the graphical representation of that secondary structure using the VARNA software [26] is shown in Figure 2. Upon the inspection of the secondary structure, a sequence of interest spanning from position 74:130 was selected and it is shown in Figure 1 along with the secondary structure as predicted by the selected predicted frameshifting region clearly falls into the 5'UTR of NC_045512.2 reference genome [14], which spans from 1 to 265 nt as the first start codon for the coding sequence is at 266 position in SARS-CoV-2. However, if −1 PRF occurs within the 5'UTR region, probably at the U nucleotide at position 95 nt, then there is an upstream AUG codon at position 107 that can act as start codon and viral translation might be altered. The sequence of interest spanning from position 74:130 was selected and it is shown in Figure 1 along with the secondary structure as predicted by
structure as indicated by its computed MFE, we analyzed 1,509 random sequences from NC_45512.2 of the same length as the sequence of interest. Their MFE values were computed by mFold [17] to obtain an empirical distribution. The predicted value for the sequence of interest (-11 Kcal/mol) was ranked 1,509 random sequences from NC_45512.2 of the same length as the sequence of interest. Their MFE values were computed by mFold [17] to obtain an empirical distribution. The predicted value for the sequence of interest (-11 Kcal/mol) was ranked

slippery sequence and neighbouring structured segment. The sequence (positions 74—130) was also used to proceed with the analysis so that it includes the slippery region and the structure of the stem and loop and the pseudoknot.

In order to test the probability of the predicted secondary structure as indicated by its computed MFE, we analyzed 1,509 random sequences from NC_45512.2 of the same length as the sequence of interest. Their MFE values were computed by mFold [17] to obtain an empirical distribution. The predicted value for the sequence of interest (-11 Kcal/mol) was ranked
Putative Secondary Structure at 5'UTR as a Potential Antiviral Target against SARS-CoV-2

E. Garcia-Moran, et al.
Rev Esp Quimioter 2022;35(2): 204-212

Pseudoknot and thus, along with the immediate slippery sequence form a frameshifting signal.

Genomic similarity. The conservation of the 74--130 region among Coronaviridae family is shown in Figure 4. Interestingly, while this region was identical in all the isolates from SARS-CoV-2 including isolates from human patients from distant geographical localizations (MT370831, New York; and

| Host                    | GenBank accession no. | Date (year) | Score | Identities (Query length) | Percentage | Expect |
|-------------------------|-----------------------|-------------|-------|----------------------------|------------|--------|
| Rhinolophus pusillus    | JX93987.1             | 2011        | 86.0  | 52/55 (57)                 | 95         | 1e-15  |
| Rhinolophus sinicus     | KJ473814.1            | 2013        | 86.0  | 53/57 (57)                 | 93         | 1e-15  |
| Rhinolophus sinicus     | MG772933.1            | 2017        | 86.0  | 53/57 (57)                 | 93         | 1e-15  |
| Rhinolophus sinicus     | MG772934.1            | 2017        | 86.0  | 53/57 (57)                 | 93         | 1e-15  |
| Mus musculus            | HQ890526.1            | 2008        | 80.6  | 52/57 (57)                 | 91         | 5e-14  |
| Mus musculus            | HQ890527.1            | 2008        | 80.6  | 52/57 (57)                 | 91         | 5e-14  |
| Mus musculus            | HQ890528.1            | 2008        | 80.6  | 52/57 (57)                 | 91         | 5e-14  |
| Mus musculus            | HQ890529.1            | 2008        | 80.6  | 52/57 (57)                 | 91         | 5e-14  |
| Mus musculus            | HQ890530.1            | 2008        | 80.6  | 52/57 (57)                 | 91         | 5e-14  |
| Mus musculus            | HQ890531.1            | 2008        | 80.6  | 52/57 (57)                 | 91         | 5e-14  |

within empirical distribution of the MFE values. The histogram and frequency curve of this distribution is shown in Figure 3. The vertical line is set at -11 Kcal/mol. Clearly, few random sequences show this value. This value was at the top 5% of the negative endo of the distribution. This reveals that the predicted structure is fairly stable in relation to other segments of NC\_45512, and supports that this sequence may occur in the predicted form of a stem-loop with outer bindings to form a pseudoknot and thus, along with the immediate slippery sequence form a frameshifting signal.

Genomic similarity. The conservation of the 74--130 region among Coronaviridae family is shown in Figure 4. Interestingly, while this region was identical in all the isolates from SARS-CoV-2 including isolates from human patients from distant geographical localizations (MT370831, New York; and
Putative Secondary Structure at 5'UTR as a Potential Antiviral Target against SARS-CoV-2

E. Garcia-Moran, et al.

Rev Esp Quimioter 2022;35(2): 204-212

Table 3: Results of docking of lead compounds from NCI diversity set II against the predicted active site in the sequence of interest.

| NSC id   | pubChem id | MFE*   | Molecular formula    | H bond donors | H bond acceptors | Active torsions | Mol weight |
|----------|------------|--------|----------------------|---------------|------------------|-----------------|------------|
| 293778   | 325266     | -12.2  | C40H32N2O4           | 0             | 5                | 2               | 594.7      |
| 308835   | 328761     | -11.1  | C39H32N2O4           | 0             | 4                | 0               | 484.6      |
| 61610    | 247228     | -11.1  | C34H24N6O2           | 4             | 4                | 4               | 548.6      |
| 37641    | 235856     | -11    | C29H13F6O6           | 2             | 7                | 4               | 496.6      |
| 319990   | 330740     | -10.7  | C23H18N6O2S2         | 4             | 6                | 4               | 474.6      |
| 93354    | 261360     | -10.6  | C28H33N02S2          | 1             | 4                | 1               | 447.6      |
| 122819   | 452548     | -10.5  | C32H22O23S           | 3             | 14               | 7               | 656.7      |
| 37553    | 235811     | -10.5  | C30H28N4O2           | 2             | 2                | 2               | 476.6      |
| 293781   | 235856     | -10.5  | C29H36F6O6           | 2             | 7                | 4               | 496.6      |

*Values of predicted MFE in Kcal/mol. MFE: minimum free energy

DISCUSSION

This study has revealed a previously unnoticed feature in the SARS-CoV-2 genome, which is likely to play a biological role on account of the remarkable conservation of its sequence and stability of the structure. The close occurrence of the slippery sequence and a likely stable pseudoknot suggests that this may be an area of frameshifting, in addition to the previously described overlapping region of ORF1a and ORF1b, where frameshifting has been proven for SARS-CoV [8] and also present in SARS-CoV-2 [9]. We focused on a different region, previously unnoticed in the 5'UTR. The fact that no protein may be linked with the sequence may argue against frameshifting, as may that of the overlap between ORF1a and ORF1b. Supporting the role of 5'UTR, Zhu et al. [28] demonstrated that different natural deletions in the 5'UTR of FMDV (foot-and-mouth disease virus) markedly affected the pathogenicity and species tropism of the virus. Frameshifting linked with 5'UTR has been described in HIV-1 [29], and in this case the structure next to the slippery sequence is a stem and loop, without additional pseudoknotting.

Another important endeavour of this work is to consider this RNA structured area as a useful target for feasible drug
Putative Secondary Structure at 5’UTR as a Potential Antiviral Target against SARS-CoV-2

E. Garcia-Moran, et al.
Rev Esp Quimioter 2022;35(2): 204-212

1. Coleman CM, Frieman MB. Emergence of the Middle East Respiratory Syndrome coronavirus. PLoS pathogens. 2013;9:e1003595. doi: 10.1371/journal.ppat.1003595.
2. Wang H, Li X, Li T, Zhang S, Wang L, Wu X, Liu J. The genetic sequence, origin, and diagnosis of SARS-CoV-2. Eur. J. Clin. Microbiol. Infect. Dis. 2020; 39(9):1629-35. doi: 10.1007/s10096-020-03899-4.
3. Han GZ. Pangolins Harbor SARS-CoV-2-Related Coronaviruses. Trends Microbiol., 2020,28(7):515-17. doi: 10.1016/j.tim.2020.04.001.
4. Jacks T, Madhani HD, Masiarz FR, Varmus HE. Signals for ribosomal frameshifting in the Rous sarcoma virus gag-pol region. Cell. 1988;55(3):447-58. doi: 10.1016/0092-8674(88)90031-1.
5. Brierley I, Pennell S, Gilbert RJC. Viral RNA pseudoknots, versatile motifs in gene expression and replication. Nat Rev Microbiol. 2007;5(8):598-610. doi: 10.1038/nrmicro1704.
6. Lim CS, Brown CM. Know your enemy, Successful bioinformatics approaches to predict functional RNA structures in viral RNAs. Front Microbiol. 2018;9:2582. doi: 10.3389/fmicb.2017.02582.
7. Fehr AR, Perlman S. Coronaviruses, an overview of their replication and pathogenesis. Methods in molecular biology (Clifton, N.J.). 2015;1282:1-23. doi: 10.1007/978-1-4939-2438-7_1.
8. Plant EP, Perez-Alvarado GC, Jacobs JL, Mukhopadhyay B, Hennig...
M, Dinman JD. A three-stemmed mRNA pseudoknot in the SARS coronavirus frameshift signal. PLOS Biology. 2005;3(6):e172. doi: 10.1371/journal.pbio.0030172.

9. Kelly JA, Olson AN, Neupane K, Munshi S, San Emeterio J, Pollack L, Woodside MT, Dinman JD. Structural and functional conservation of the programmed -1 ribosomal frameshift signal of SARS coronavirus 2 (SARS-CoV-2). J. Biol. Chem. 2020;31;295(31):10741-48. doi: 10.1074/jbc.A120.13449.

10. Neuman BW, Stein DA, Kroeker AD, Churchill MJ, Kim AM, Kuhn P, Dawson P, Moulton HM, Bestwick RK, Versen PL, Buchmeier MJ. Inhibition, escape, and attenuated growth of severe acute respiratory syndrome coronavirus treated with antisense morpholino oligomers. J. Virol. 2005;79(15):9665-76. doi: 10.1128/JVI.79.15.9665-9676.2005.

11. Rangan R, Zhelevud IN, Das R. RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses, a first look. RNA. 2020;26(8):937-59. doi: 10.10121/ma.076141.120.

12. Park H-J, Park S-J. Virtual Screening for RNA-Interacting Small Molecules. In: Dinman J. (eds) Biophysical approaches to translational control of gene expression. Biophysics for the Life Sciences, vol 1. Springer, New York, NY. 2012;1:235-52. doi:10.1007/978-1-4614-3991-2_12.

13. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, Hu Y, Tao Z-W, Tian J-H, Pei Y-Y, Yuan M-L, Zhang Y-L, Dai F-H, Liu Y, Wang Q-M, Zheng J-J, Xu L, Holmes EC, Zhang Y-Z. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265-69. doi:10.1038/s41586-020-2008-3.

14. Theis C, Reeder J, Giegerich R. KnotInframe, prediction of -1 ribosomal frameshift events. Nucleic Acids Res. 2008;36(18):6013-20. doi: 10.1093/nar/gkn578.

15. Sato K, Kato Y, Hamada M, Akutsu T, Asai K. IPknot, fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. Bioinformatics. 2011;27(13):i85-93. doi: 10.1093/bioinformatics/btr215.

16. Lorenz R, Bernhart SH, zu Siederdissen HC, Tafer H, Flamm C, Stadler PF, Hofacker IL. ViennaRNA package 2.0. Algorithms for molecular biology. 2011; 6:26 doi: 10.1186/1748-7188-6-26.

17. Gruber AR, Lorenz R, Bernhart SH, Neubouck R, Hofacker IL. The Vienna RNA website. Nucleic Acids Res. 2008; 36(Web Server issue):W70-74. doi: 10.1093/nar/gkn188.

18. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality multiple protein sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539. doi: 10.1038/msb.2011.75.

19. Price MN, Dehal, PS, Arkin AP. FastTree, computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009;26(7):1641-50. doi: 10.1093/molbev/msp077.

20. Yu G. Using ggtree to visualize data on tree-like structures. Curr Protoc Bioinformatics. 2020;69(1):e96. doi: 10.1002/cpbl96.

21. R Core Team, R, A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2017.