Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 11 May 2020
Applications Note

Sequence analysis

Preliminary RNA secondary structural analysis of the SARS-CoV-2 spike glycoprotein coding region at the inter- and intraspecies levels

Lawrence Chillrud [1,*] and Itsik Pe’er [1]

[1] Department of Computer Science, Columbia University, New York, 10027, USA

* To whom correspondence should be addressed.

Abstract

Motivation: By May 10, 2020, a total of 4,088,393 cases of COVID-19, caused by SARS-CoV-2, had been confirmed worldwide, causing nearly 300,000 deaths and significantly impacting daily life for millions (JHU, 2020). A large body of research has identified the spike (S) glycoprotein region in SARS-CoV-2 as an area of interest due to its crucial role in allowing the virus to gain entry to and subsequently infect host cells (Walls *et al.*, 2020). Furthermore, the SARS-CoV-2 S gene seems to have diverged from the S genes of other related beta coronaviruses, making it a genetically unique region that warrants close attention (Wen *et al.*, 2020). As such, this study performs a preliminary RNA secondary structural analysis of the S gene coding region across inter- and intraspecies datasets, in the hope of identifying conserved secondary structures that could be targeted when developing a vaccine.
Results: This study identified 15 potentially conserved structures within the S gene at the interspecies level, and 43 at the intraspecies level. Each structure was also given an associated likelihood. The largest conserved structure identified, comprised of 259 nucleotides (6.8% of the S gene), was found at the intraspecies level, located at positions 2761-3019 within the 3822-nucleotide-long S gene. Examining a subset of the interspecies dataset (using only SARS-CoV-2 and RaTG13, the closest related beta coronavirus) yielded the same conserved structure highlighted by the analysis, this time slightly shorter at 215 nucleotides in length, located at positions 2805-3019.
Availability: https://github.com/lawrence-chillrud/SARS-CoV-2-Seq-Analysis.
Contact: lgc2139@columbia.edu

1 Introduction

The pandemic caused by SARS-CoV-2 has brought global economies, daily routines, and human lives to an abrupt halt as of May, 2020. A large body of research has identified the spike (S) glycoprotein region in SARS-CoV-2 as an area of interest due to its crucial role in allowing the virus to gain entry to and subsequently infect host cells (Walls *et al.*, 2020). Furthermore, the SARS-CoV-2 S gene seems to have diverged from the S genes of other related beta coronaviruses (Wen *et al.*, 2020). RNA secondary structures are known to play a critical role in the viral life cycle, even influencing the infectivity and lethality of a virus, and underscoring the famous dogma, “structure determines function,” (Andrews *et al.*, 2019). As such, this study performs a preliminary RNA secondary structural analysis of the S gene coding region across inter- and intraspecies datasets in the hope of identifying conserved secondary structures that could be targeted when developing a vaccine.

2 Methods

The interspecies dataset is made up of five isolates (including SARS-CoV-2), all of which belong to the beta coronavirus genus. These isolates were chosen after reviewing the relevant literature discussing the phylogeny of the novel coronavirus (Tang *et al.*, 2020). The five isolates, along with their reference genome’s GenBank accession number and their percent identity to SARS-CoV-2, can be found in supplementary Table 1. The complete reference S gene for each interspecies isolate was downloaded in FASTA format from the NCBI database (Sayers *et al.*, 2010).

The intraspecies dataset consists of the SARS-CoV-2 reference genome, along with 413 other isolates of SARS-CoV-2 that were arbitrarily chosen from the isolates available in the NCBI database at the time. The S genes of these 414 isolates were scraped and downloaded in GenBank format from the NCBI database by a custom Python script.

Multiple sequence alignments (MSAs) were then obtained for each of the two datasets with the Clustal Omega algorithm, which uses seeded guide trees and HMM profile-profile techniques to generate the alignments along with other useful metadata such as percent identity matrices (Madeira *et al.*, 2019). With the percent identity matrix for the intraspecies dataset, it was found that many of the intraspecies isolates shared 100.00% identity with one another. Therefore, a very basic dimensionality reduction was performed on the intraspecies data; those isolates that had 100.00% overlap in their S gene were deemed redundant and were subsequently removed, save for one representative. This reduced the number of contributing interspecies isolates from 414 to 65.

RNA secondary structure prediction was then performed on the SARS-CoV-2 reference S gene by the program RNAfold from the ViennaRNA 2.0 package (Stadler *et al.*, 2011). RNAfold calculates minimum free energy secondary structures of RNA sequences, reporting the resulting structures in the common dot-bracket notation.

From here, a custom algorithm was written in Python to extract potentially conserved secondary structures given. The algorithm begins by lining the MSA up with the dot-bracket notation encoding the secondary structure of the SARS-CoV-2 spike glycoprotein. Each single site in the MSA is then parsed as a vector, $\vec{v}_i \in \mathbb{R}^d$, $\forall i \in \{1, 2, \ldots, length(\text{MSA})\}$, where $d = 5$ for the interspecies case, and $d = 65$ for the intraspecies case (due to the dimensionality reduction).

Next, each vector encoding a single site in the MSA then undergoes a collapsation, $C(\vec{v}_i) : \mathbb{R}^d \mapsto \mathbb{R}^d$, in which the only information kept is whether or not an element in the vector matches the element corresponding to the SARS-CoV-2 sequence, as well as how many different elements are present in the vector. The collapsed expression of a vector, $\vec{v}_i$, is denoted as $\vec{c}_i$. This kind of collapsation can be seen in supplementary Fig. 1: the top of the figure shows an example of the initial MSA for the interspecies case, the middle shows the corresponding dot-bracket notation, and the bottom shows the “collapsed” version of the same MSA.

At this stage in the algorithm, the procedure is to step through the “collapsed” MSA one site (i.e. vector) at a time, all the while checking the corresponding dot-bracket notated secondary structure per step. Basic bracket matching / balancing is done via a stack, with which an open bracket is matched with its corresponding closed bracket. The idea is that if there are long stretches in which an open bracket’s associated $\vec{c}_i$ perfectly matches the corresponding closed bracket’s associated $\vec{c}_j$, then the region those two brackets span is flagged as a potentially conserved secondary structure (provided there are also perfectly matching vectors for every pair of brackets inside that span). Essentially, if $\vec{c}_i = \vec{c}_j$ and $i \neq j$, and the bracket at position i, denoted b_i, is balanced by b_j, and this holds true for all pairs of balanced brackets in between the span of i and j, then the region $[i, j]$ is flagged as a potentially conserved structure.

Returning to supplementary Fig. 1 as an example, and the described algorithm would flag sites $[1, 7]$ as making up a potentially conserved secondary structure. This is because every pair of brackets $\{(1, 7), (2, 6), (3, 5)\}$ has matching collapsed vectors (expressed as the brightly colored columns in the figure).

Finally, the validity of the flagged sequence is evaluated. Working under an i.i.d. assumption, each collapsed vector, $\vec{c}_i$, is given an associated probability, $\Pr[c_i]$, obtained by taking the raw count of the given collapsed vector over the entirety of the MSA, and dividing that count by the number of sites total in the MSA (i.e. $length(\text{MSA})$). The likelihood of seeing a given conserved region $[i, j]$, $\Pr[i, j]$, is then given by equation (1).

$$\Pr[i, j] = \sum_{i}^{j} \log\Big(\Pr[c_i]\Big). \tag{1}$$

This log likelihood is then compared against the log likelihood of seeing a random sequence of collapsed vectors of the same size. Again, via the i.i.d. assumption, this is easily obtained: a mean probability, μ, and variance, σ^2, were calculated from the distribution of probabilities made earlier for each $\vec{c}_i$. A Gaussian distribution is parameterized, $\mathcal{N}(\mu, \sigma^2)$, from which $(j - i + 1)$-many independent draws can be made. These draws’ log probabilities are then summed, giving the basis for comparison. We call a potentially conserved secondary structure identified by our algorithm “likelier than not” if its log probability is higher than that of a random sequence of the same size generated in this way.

3 Results

This study identified 15 potentially conserved structures within the S gene at the interspecies level, and 43 at the intraspecies level. The largest conserved structure identified came from the intraspecies dataset, and was 259 nucleotides long (6.8% of the S gene). It was located at positions 2761-3019 within the 3822-nucleotide-long S gene and deemed “likelier than not” with a log likelihood of -95, while a random sequence of the same size had one of -132. When the interspecies dataset was relaxed to only contain SARS-CoV-2 and RaTG13, it also highlighted this region as a potentially conserved secondary structure, only slightly shorter: 215 nucleotides long, found at positions 2805-3019. It was also deemed “likelier than not”, with a log likelihood of -44, while a random sequence of the same size had one of -60. The common structure can be seen in supplementary Fig. 2. Unsurprisingly, 54 potentially conserved structures were identified with this relaxed interspecies analysis, and on average produced conserved regions that were much more “likely” and much longer than those structures identified using the full interspecies dataset.

4 Discussion

The fact that both the intraspecies dataset and relaxed interspecies dataset pointed towards the structure located at positions 2805-3019 in the SARS-CoV-2 spike (S) glycoprotein coding region as a potentially conserved site is exciting. It suggests that at both the inter- and intraspecies level, coronaviruses do not like to play with that region. This analysis simply points to that region (among others) as an area for further study: it could be that exploiting the structure found there could aid in the development of a vaccine, or insights could be made into some of the selection processes and pressures SARS-CoV-2 is subject to. Due to time constraints, this study was unable to utilize more robust forms of secondary structure prediction and analysis that exist. A more robust statistical analysis is also warranted.

Acknowledgements

The first author acknowledges the second author in all of their help, guidance, and patience. The first author also acknowledges the European Bioinformatics Institute and the Theoretical Biochemistry Group of the University of Vienna.

References

Andrews, R. J. *et al.* (2019). Mapping the rna structural landscape of viral genomes. *Methods*.

JHU (2020). Johns hopkins university: Coronavirus resource center. *COVID-19 Map*.

Madeira, F. *et al.* (2019). The embl-ebi search and sequence analysis tools apis in 2019. *Nucleic acids research*, **47**(W1), W636–W641.

Sayers, E. W. *et al.* (2010). Database resources of the national center for biotechnology information. *Nucleic acids research*, **39**, D38–D51.

Stadler, P. *et al.* (2011). Viennarna package 2.0. *Algorithms*.

Tang, X. *et al.* (2020). On the origin and continuing evolution of sars-cov-2. *National Science Review*.

Walls, A. C. *et al.* (2020). Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell*.

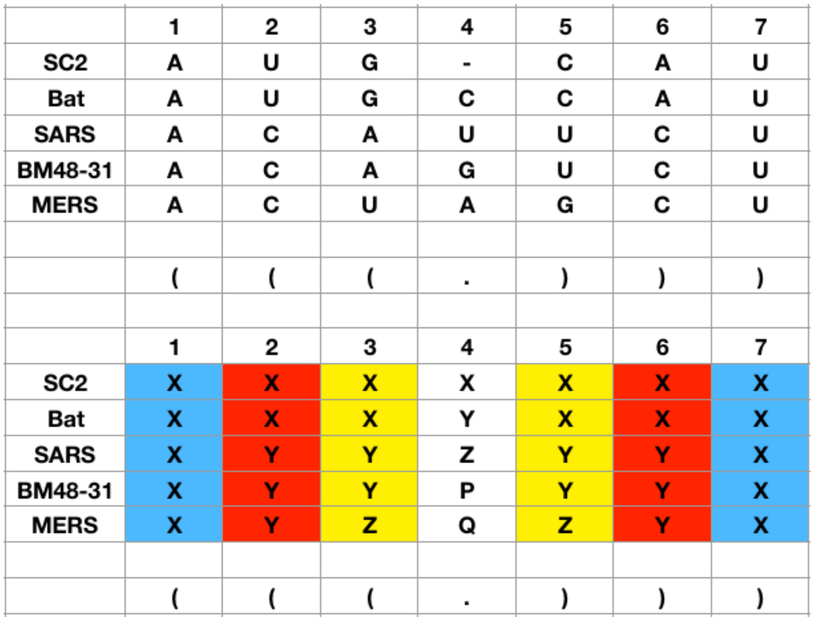
Wen, F. *et al.* (2020). Identification of the hyper-variable genomic hotspot for the novel coronavirus sars-cov-2. *Journal of Infection*.

Supplementary Tables and Figures

Species	Accession	% Identity to SARS-CoV-2*
SARS-CoV-2	NC_045512	100.00%
RaTG13	MN996532	93.15%
SARS-CoV	NC_004718	73.97%
BM48-31	NC_014470	70.40%
MERS-CoV	NC_019843	52.64%

Table 1: The interspecies dataset

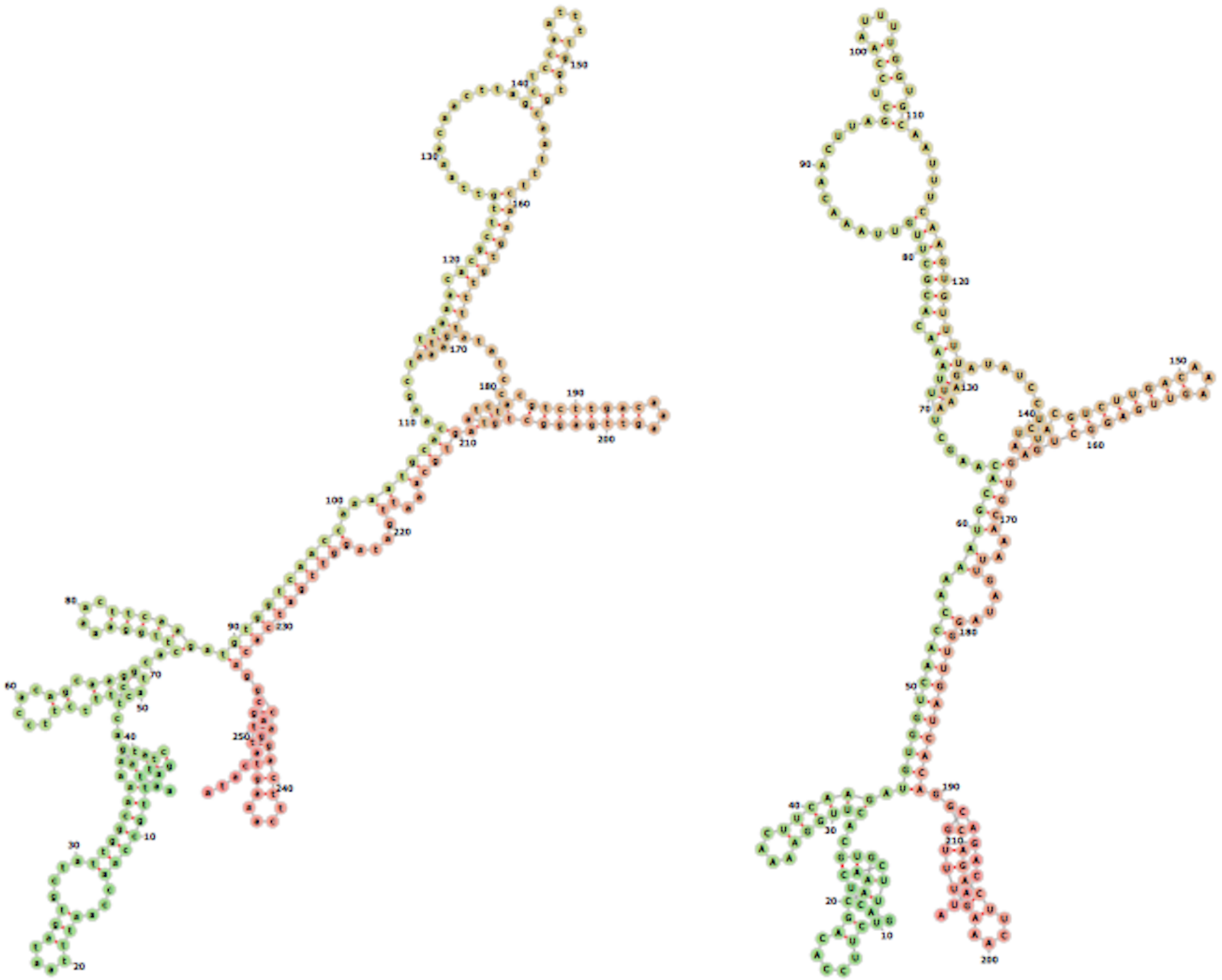
*Note: Percent identity calculated by Clustal Omega



	1	2	3	4	5	6	7
SC2	A	U	G	-	C	A	U
Bat	A	U	G	C	C	A	U
SARS	A	C	A	U	U	C	U
BM48-31	A	C	A	G	U	C	U
MERS	A	C	U	A	G	C	U
	(	(	(	.	)	)	)

	1	2	3	4	5	6	7
SC2	X	X	X	X	X	X	X
Bat	X	X	X	Y	X	X	X
SARS	X	Y	Y	Z	Y	Y	X
BM48-31	X	Y	Y	P	Y	Y	X
MERS	X	Y	Z	Q	Z	Y	X
	(	(	(	.	)	)	)

Figure 1: Visualization of a sample multiple sequence alignment for the interspecies case (top), with its corresponding “collapsed” expression (bottom) utilized in the custom algorithm to identify potentially conserved secondary structures.



(a) Largest potentially conserved secondary structure identified from the intraspecies dataset.

(b) Largest potentially conserved secondary structure identified from the relaxed interspecies dataset.

Figure 2: **(a)** Largest potentially conserved secondary structure identified from the intraspecies dataset. 259 nucleotides in length (6.8% of the S gene). It was located at positions 2761-3019 within the 3822-nucleotide-long S gene and deemed "likelier than not" with a log likelihood of -95, while a random sequence of the same size had one of -132. **(b)** Largest potentially conserved secondary structure identified from the relaxed interspecies dataset (SARS-CoV-2 and RaTG13). 215 nucleotides in length located at positions 2805-3019. Deemed "likelier than not" with a log likelihood of -44, while a random sequence of the same size had one of -60.