PREPRINT

April 10th, 2020: 1-5

Email: ymorales@una.ac.cr

Coronavirus 2: Analysis of Regularity of Complete Genome (SARS-CoV-2/Hu/DP/Kng/19-020 RNA)

Yuri Morales-López
Universidad Nacional, Costa Rica

Abstract

This paper shows a technique for the detection of regularities in the complete genome of Coronavirus 2. For this, the concept of intervals in the RNA chain and their comparison within the known sequence is used. The *Maximum Regularity Index* was applied. The results show that areas of high regularity are identified that coincide with ORF10, Gene N, nucleocapsid phosphoprotein and others. The detection of regularities can enhance the creation of biochemical or genetic techniques to deactivate certain functions of Coronavirus 2 or control its effects.

Keywords: Coronavirus 2; SARS-CoV-2; Complete Genome; Regularity Analysis; RNA; index of maximum regularity; Covid-19

INTRODUCTION

Coronavirus 2 has a sequence of 29902 nt (April 10th, 2020) (National Center for Biotechnology Information [NCBI], 2020). Its full lineage is: Viruses; *Riboviria; Nidovirals; Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus; Severe acute respiratory syndrome-related coronavirus* and its genbank acronym is *SARS-CoV2*.

A method has been applied in which chains of RNA of a certain length are taken and compared with chains of the same length within the complete genome. In previous work this technique has been applied to other microorganisms (Ugalde, Morales, & Láscaris-Comneno, 2010; Morales López, Ugalde león, & Láscaris-Comneno, 2010; Láscaris-Comneno-Slepuin, Ugalde-León, & Morales-López, 2011)

METODOLOGY

The index of maximum regularity $i_{max,r}$ was defined in Láscaris-Comneno, Skliar, & Medina (1999) and in other papers such as Láscaris-Comneno-Slepuin, Ugalde-León, & Morales-López (2011) has been widely described.

In summary, a string or interval l_c is defined. It considers the new sequences that are generated $l_c - 1$ when a shift is made, one nucleotide at a time. Each ordered pair is part of the set $B \times B$, where $B = \{A, G, T, C\}$. This procedure is repeated considering the original sequence together with the one generated by reversing its order.

Denoted by $n_{A,A,d}$, $n_{A,G,d}$, $n_{A,C,d}$, $n_{A,T,d}$, $n_{G,A,d}$, $n_{G,A,d}$, $n_{G,G,d}$, $n_{G,C,d}$, $n_{G,T,d}$, $n_{C,A,d}$, $n_{C,G,d}$, $n_{C,C,d}$, $n_{C,T,d}$, $n_{T,A,d}$, $n_{T,C,d}$, $n_{T,G,d}$y $n_{T,T,d}$ the number of ordered pairs actually counted. The $i_{max,r}$ is defined as:

$$i_{max,r} = \frac{max\{\, D_1, \cdots, D_{l_c-1}, D_0^*, \cdots, D_{l_c-1}^*\}}{\left(l_c - \frac{l_c}{16}\right)^2 + 15\left(\frac{l_c}{16}\right)^2}$$

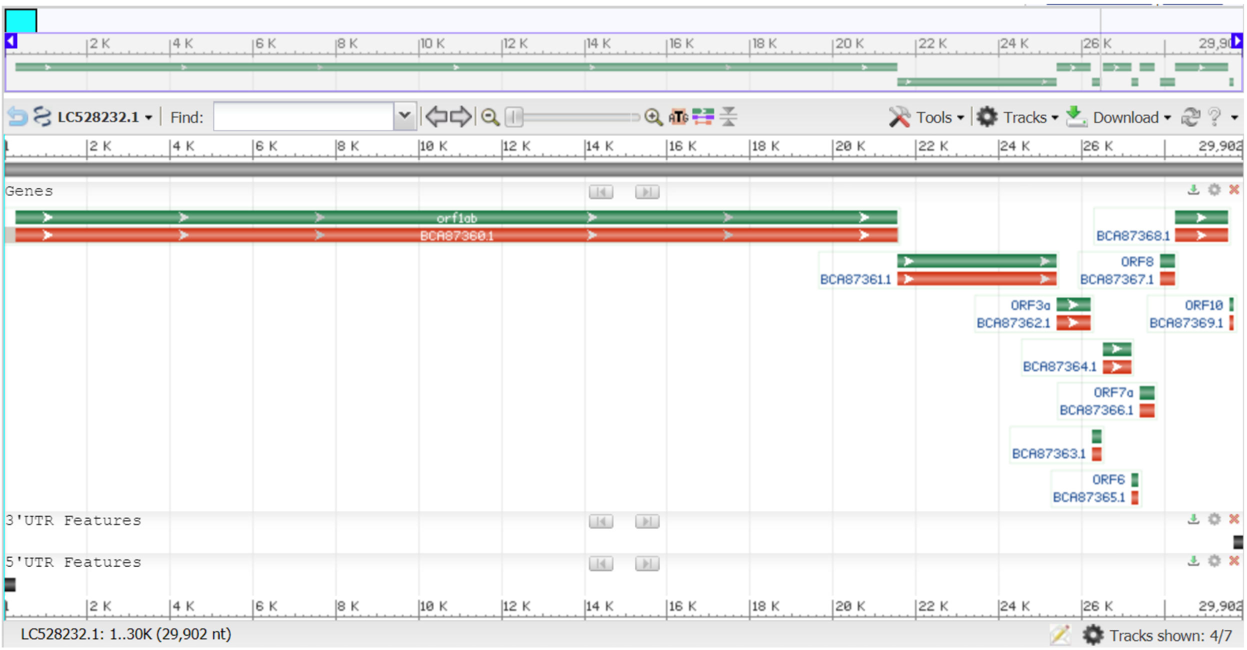
Where $D^*{}_d = \left(n_{A,A,d} - \frac{l_c}{16}\right)^2 + \left(n_{A,G,d} - \frac{l_c}{16}\right)^2 + \cdots + \left(n_{T,T,d} - \frac{l_c}{16}\right)^2$

RESULTS

The $i_{max,r}$ was applied to the complete genome of Coronavirus 2. Below are the results of coincidence with the genes already discovered by other techniques (NCBI, 2020) and possible areas of study for possible deepening.

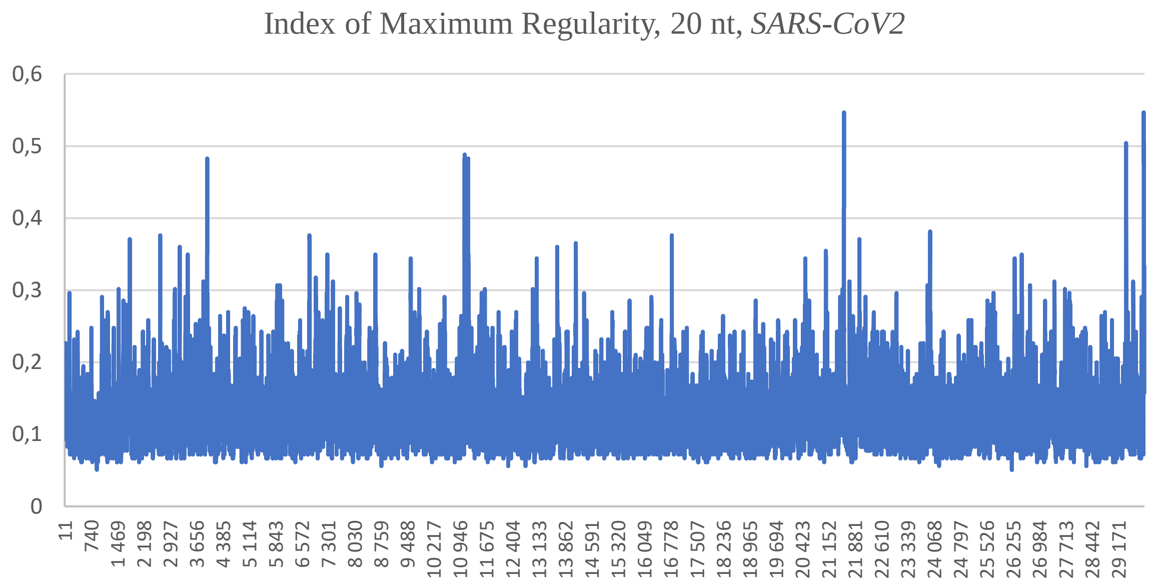
For a visualization on the components of Coronavirus 2, the configuration of the NCBI (2020) is shown.

Image 1: Diagram on the structure of Coronavirus 2.



Note: From NCBI (2020).

Graph 1 presents the analysis made for 20nt intervals



In the indicated regions, there are intervals that take higher values (Table 1).

Table 1: Index of Maximum Regularity, 20 nt, *SARS-CoV2 vs NCBI*

Interval (middle point)	$i_{max,r}$	Already classified by NCBI
21578,5	0,546666667	**Gene S (start)**
29875,5	0,546666667	**Gene ORF10**
29876,5	0,546666667	**Gene ORF10**
29877,5	0,546666667	**Gene ORF10**
29390,5	0,504	**Gene N**
29882,5	0,504	**Gene ORF10**
29883,5	0,504	**Gene ORF10**
11077,5	0,488	Undefined. Inside Orf1ab
3952,5	0,482666667	Undefined. Inside Orf1ab
11075,5	0,482666667	Undefined. Inside Orf1ab
11076,5	0,482666667	Undefined. Inside Orf1ab
11078,5	0,482666667	Undefined. Inside Orf1ab
11079,5	0,482666667	Undefined. Inside Orf1ab
11080,5	0,482666667	Undefined. Inside Orf1ab
11180,5	0,482666667	Undefined. Inside Orf1ab
29880,5	0,482666667	**Gene ORF10**
29881,5	0,482666667	**Gene ORF10**

Thus, it is evident that in the areas near the intervals around 21578, 29876, 29390, 3952, 11076 have high values. As consulted in the database, this is positive because **Gene S, Gene N** and **ORF10** were identified. It is important to research in more detail what happens around **11078** interval (20 nt). It should be clarified that **Gene: Orf1ab** is in the range [269, 21 558] (71% of the entire genome) which is extremely wide and the $i_{max,r}$ indicates something of importance around 11000.

The following is the complete genome analysis with 16 nt. (Graph 2 and Table 2)

Graph 2 presents the analysis made for 16 nt intervals

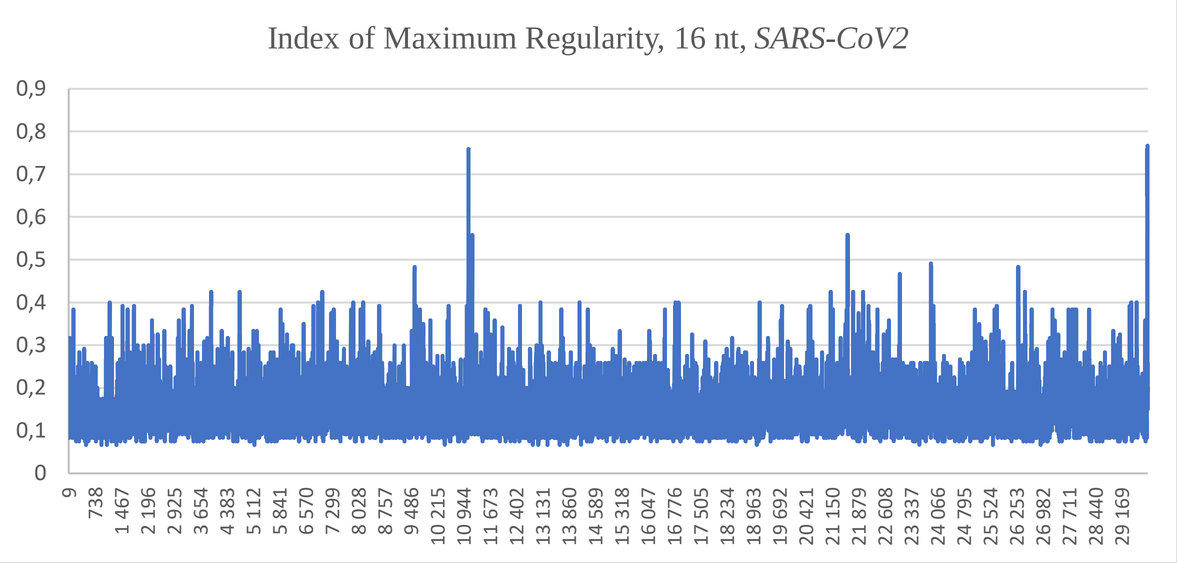


Table 2: Index of Maximum Regularity, 16 nt, *SARS-CoV2 vs NCBI*

Interval (middle point)	$i_{max,r}$	**Already classified by NCBI**
29881,5	0,766666667	**Gene ORF10**
29882,5	0,766666667	**Gene ORF10**
11078,5	0,758333333	Undefined. Inside Orf1ab
11079,5	0,758333333	Undefined. Inside Orf1ab
29878,5	0,758333333	**Gene ORF10**
29879,5	0,758333333	**Gene ORF10**
29880,5	0,65	**Gene ORF10**
29884,5	0,6	**Gene ORF10**
11076,5	0,558333333	Undefined. Inside Orf1ab
11077,5	0,558333333	Undefined. Inside Orf1ab
11182,5	0,558333333	Undefined. Inside Orf1ab
21576,5	0,558333333	**Gene S (start)**
21577,5	0,558333333	**Gene S (start)**
21581,5	0,558333333	**Gene S (start)**
29883,5	0,558333333	**Gene ORF10**
11080,5	0,55	Undefined. Inside Orf1ab
11081,5	0,55	Undefined. Inside Orf1ab
11082,5	0,55	Undefined. Inside Orf1ab
11083,5	0,55	Undefined. Inside Orf1ab
21580,5	0,55	**Gene S (start)**
29875,5	0,55	**Gene ORF10**
29876,5	0,55	**Gene ORF10**
29877,5	0,55	**Gene ORF10**

In this string size, **Gene S (start)** and **Gene ORF10** are identified. And again, the $i_{max,r}$ points to an important zone around the **11000** interval.

When calculating with a length interval of 28 nt, the sectors around **3500** and **11000** appear with high values. Longer chains are not considered because the probability of occurrence decreases with the number of combinations of the four possibilities.

CONCLUSIONS

During the analysis **Gene S, Gene N** and **ORF10** were identified. The interval around **3500** and the interval around **11000**, respectively, appear with high values of regularity and this do not correspond to documented elements (both within Gene: **Orf1ab**). It might be in the interest of giving importance to that region.

The strategy developed here may offer insights into Coronavirus 2 and eventually, with much more research and sophistication, a treatment may be found.

REFERENCES

Láscaris-Comneno, T., Skliar, O., & Medina, V. (1999). Determinación de valores del índice de máxima regularidad correspondientes a diversas secuencias de bases de ADN. Un nuevo método computacional en Genética. *Proceedings IX Congreso Internacional de BioMatemática*, pp. 81-97. Concepción, Chile.

Láscaris-Comneno-Slepuin, T., Ugalde-León, A., & Morales-López, Y. (2011). Análisis de regularidad para el reconocimiento de telómeros en Candida parapsilosis mitochondrion y el cromosoma XVI de Saccharomyces cerevisae. *Revista Tecnología En Marcha*, 24(4), 59-68. https://revistas.tec.ac.cr/index.php/tec_marcha/article/view/157

Morales López, Y., Ugalde León, A., & Láscaris-Comneno Slepuin, tatiana. (2010). Análisis de regularidad de genomas para detección de telómeros y secuencias autónomamente replicativas. *Uniciencia*, *24*(1), 103-110. https://www.revistas.una.ac.cr/index.php/uniciencia/article/view/376

National Center for Biotechnology Information NCBI, (2020). Taxonomy Browser Tool (Gene Bank Database). https://www.ncbi.nlm.nih.gov/nuccore/LC528232.1?report=genbank

Ugalde, A., Morales, Y., & Láscaris-Comneno, T. (2010). Mathematics Applied to the Detection of Genetic Regularities in the Yeast Yarrowia lipolytica. Proceedings of the XVII International Symposium on Mathematical Methods Applied to the Sciences. San José, Costa Rica.