SARS-CoV-2 origin: an affair of codons?

Antonio R. Romeu[1] and Enric Ollé[2]

[1]: Chemist. Professor of Biochemistry and Molecular Biology at the Rovira i Virgili University. Tarragona. Spain. Corresponding author. Email: antonioramon.romeu@iubilo.urv.cat

[2]: Veterinarian, Biochemist. Associate Professor of the Department of Biochemistry and Biotechnology of the Rovira i Virgili University. Tarragona. Spain. Email: enric.olle@urv.cat

Abstract

The furin cleavage site, with an arginine doublet (RR), is one of the clues of the SARS-CoV-2 origin. This furin-RR is encoded by the CGG-CGG sequence. Because arginine can be encoded by six codons, in a previous work we found that in SARS-CoV-2, CGG was the minority arginine codon (3%). Also, analyzing the RR doublet from a large sample of furin cleavage sites of several kinds of viruses, we found that none of them were encoded by CGG-CGG. Here, we come back to the core of the matter, but from the perspective that in the human genome, in contrast, CGG is the majoroty arginine codon (21%). Here, we highlighted that the 6 arginine codons provide genetic markers to a traceability on the RR origin in the furin site, as well as, to weigh the probability of the theories about the origin of the virus.

Key words

SARS-CoV-2 origin, Furin Cleavage Site, Arginine Codon Usage, Bioinformatics.

As it is known (1,2,3) and we have also addressed (4), the origin of SARS-CoV-2 could be reduced to the origin or acquisition of the furin cleavage site in its S protein. This was a gain of function: there had been an insertion in the S gene that had caused the S protein to gain more capacity for human infection. SARS-CoV-2 belongs to the group of Betacoronaviruses, Sarbecoviruses (Lineage B). It was surprising that it was the only Sarbecovirus species with such furin site. At present, still nobody knows how and when it got to the virus. this is a key point in the controversy over the origin of the virus pandemic.

Before going ahead, some basic concepts of biology need to be absolutely clear. Based on the universal genetic code and in the protein sinthesis, when the sequence of a gene is read in frames of three, the combination of the 4 letters taken three by three, results in the existence of 64 triplets or codons. Since there are only 20 protein amino acid, there are more than enough codons to go around, allowing some amino acids to be specified by more than one codon. The arginine (whose symbol is R), can be encoded by any of the 6 triplets: AGG, AGA, CGA, CGC, CGG, CGT.

In SARS-CoV-2, the furin site is characterized by the insertion of a 4 amino acid sequence (PRRA), which corresponds to the insertion of 12 nucleotides (3 x 4). In SARS-CoV-2, the RR doublet of the furin site is encoded by the CGG-CGG codons. It is worth bearing in mind that furin cleavage site, with RR doublet, is common in the world of viruses (5) (including Coronaviruses, but excluding Betacoronaviruses). Recombination with other viruses is the most plausible explanation for the acquisition of this site in SARS-CoV-2 (6). However, to analyze the likelihood of such virus recombination, we screened the databases, and analyzing the RR doublet from a large sample of furin cleavage sites of several kinds of viruses. We found that there were no RR doublets encoded by the CGG-CGG codons (4). We observed that AGA triplet was the majority codon involved in these viral RR doublets. In all genetic recombination, there is always a donor and an acceptor of genetic material; and the donor code is passed to the acceptor. If the SARS-CoV-2 has acquired the furin site by recombination with another virus, at the moment, from the information available in the genomic databases, we can't know what it may have been.

With these results, we were interested in determining the arginine codon usage in SARS-CoV-2. Studying the composition of all its proteins, we found (4): AGG (13%), AGA (45%), CGA (5%), CGC (10%), CGG (3%), CGT (24%). The AGA triplet was the majority, and interestingly, CGG was the minority. In the specific case of S protein, of the 42 arginines it has, 20 are encoded by AGA, and only 2 by CGG (4). It was surprising that CGG-CGG codons, were those that encoded the RR doublet of the SARS-CoV-2 furin site.

Since each species has its own codon usage. Regarding the amino acid arginine in Homo sapiens, the frequency of use of triplets is (7): AGG (20%), AGA (20%), CGA (11%), CGC (19%), CGG (21%), CGT (9%). This pattern of the human genome contrats with that of the virus. In the human species, the CGG triplet is the majority (21%) and in the virus it is the minority (3%). It is surprising that in SARS-CoV-2, the furin site RR doublet uses the arginine majority codon of the human genome.

Circumstance like this fuels the great dilemma on conflicting theories about the origin of the virus: natural or laboratory (biotechnological).

The theory that supports a natural origin of SARS-CoV-2, should consider that two independent events have occurred in time: (i) the insertion by mutation or recombination of a 12 nucleotides sequence, encoding the furin site, in a strategic site of the S gene; and (ii) such inserted sequence must contain the majority human arginine codon: CGG. Both events already have a low probability, however, since they are independent and had had to be simultaneous, the probability would be even lower.

The theory that supports a laboratory origin of SARS-CoV-2, should contemplate that the insertion of the furin site in the virus genome had been in a controlled way. Considering the current applications of genetic engineering, genes of certain human proteins are now inserted (cloned) into microorganism genomes suitable for their commercial manufacture. A typical example is the insulin production obtained from the yeast *Sacharomyces cerevisae*. In these circunstances, human genes are biotechnology optimized with the majority codons of the recipient microorganism. In the case of the SARS-CoV-2 furin cleavage site, it gives the impression that a 12-nucleotide sequence had been cloned into the genome of a given virus originating the SARS-CoV-2. Of course, this cloning sign can only be glimpsed since arginine has 6 triplets in the universal genetic code. In the impossible case that arginine had 2 codons, such as lysine (by the way, lysine is also a positive

amino acid and present in viral furin sites), that traceability does not it would make sense and, there would be no doubt about the natural origin of the virus.

This is the state of the art on the origin of SARS-CoV-2, from the genetic perspective. The issue must also be approached with a forensic genetic mindset. Results based on genetic markers are always expressed in stochastic or probabilistic terms. In genetics, there are no absolute certainties, but there are evidences so highly probable that allow sentencing of guilt or the determination of paternity and/or maternity relationships. Thus, under the umbrella of the sequence analyses, the origin of SARS-CoV-2 cannot be proved as a mathematical theorem. However, between two theories, which have the same consequences, the simplest explanation is usually the most probable: Occam's razor.

Acknowledgements

This work has not been awarded grants by any research-supporting institution.

Competing interest declaration

All authors declare that they have no conflicts of interest.

References

1. Britt Glaunsinger. Coronavirus biology. The second lecture in the COVID-19, SARS-CoV-2 and the Pandemic Series. University of California, Berkeley. 2020. Accessed June 2, 2021. https://www.youtube.com/watch?v=r2mOU2qOCYs.

2. Nicholas Wade. Origin of Covid — Following the Clues. Accessed June 2, 2021. https://nicholaswade.medium.com/origin-of-covid-following-the-clues-6f03564c038.

3.Kristian G Andersen, Andrew Rambaut, W Ian Lipkin, Edward C Holmes, Robert F Garry. The proximal origin of SARS-CoV-2. Nat. Med. 26:450-452, 2020. PMID: 32284615. doi: 10.1038/s41591-020-0820-9.

4. Romeu, A.R.; Ollé, E. SARS-CoV-2 and the Secret of the Furin Site. Preprints 2021, 2021020264 (doi: 10.20944/preprints202102.0264.v1). Accessed June 1, 2021..

5. Elisabeth Braun, Daniel Sauter. Furin-mediated protein processing in infectious diseases and cancer. Clin. Transl. Immunol. E1073, 2019. PMID: 31406574. doi.org/10.1002/cti2.1073.

6. William R Gallaher. A palindromic RNA sequence as a common breakpoint contributor to copy-choice recombination in SARS-COV-2. Arch. Virol. 165:2341-2348, 2020. PMID: 32737584. doi: 10.1007/s00705-020-04750-z.

7. GenScript Codon Usage Frequency Table(chart) Tool. Accessed June 2, 2021. https://www.genscript.com/tools/codon-frequency-table.