Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2021 doi:10.20944/preprints202110.0080.v1

SARS-CoV-2 and MERS-CoV Share the Furin Site CGG-CGG Genetic Footprint

Antonio R. Romeu[1]

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

Abstract

At present, the polybasic furin cleavage site on the spike glycoprotein of the SARS-CoV-2 is still a missing link. Remarkably, the two arginine residues of this site are encoded by the CGG arginine codon, which is rare in Betacoronavirus proteins. Arginine dimers are common at viral furin sites, but are not CGG-CGG encoded. The question is: Is that genetic footprint, encoding arginine pairs, unique to the SARS-CoV-2? To address the issue, using Perl scripts, here I dissect in detail the NCBI Virus database in order to report the arginine dimers that exist in Betacoronavirus proteins. As main result, a set of Middle East respiratory syndrome-related coronavirus (MERS-CoV) (isolates: camel/Nigeria/NVx/2016, host: *Camelus dromedarius*) have the CGG-CGG encoded arginine pair in the spike protein polybasic furin cleavage site. In addition, CGG-CGG encoded arginine pairs were also found in the orf1ab polyprotein from HKU9 and HKU14 Betacoronavirus, as well as, in the nucleocapsid phosphoprotein from few SARS-CoV-2 isolates. To quantify the presence probability of CGG-CGG arginine-arginine in Betacoronavirus, a First-Order Markov Chain was defined. It is highly unlikely to find it in betacoronaviruses wildlife, but it is there. Collectively, results shed light on recombination as origin of the virus CGG-CGG arginine dimer in the S1/S2 cleavage site.

Key words

SARS-CoV-2, MERS-CoV, Arginine Dimer, Polybasic Furin Cleavage Site, Aginine Codon Usage, Markov Model, Bioinformatics

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2021 doi:10.20944/preprints202110.0080.v1

First of all, the structure and availability of the NCBI Virus database information (1), that makes this work possible, must be appreciated. Arginine is a polar and non-hydrophobic amino acid, with a positive charged group a physiological pH. Arginine participates in the binding of negatively charged substrates and/or protein actives sites (2). Consistently, arginine is involved in viral polybasic proteolytic cleavage sites, even as a dimers, as recognition motif of the ubiquitously expressed furin serine protease (3,4).

A notable characteristic of the SARS-CoV-2, that distinguishes from the rest of Sarbecovirus, is the acquisition of a polybasic furin cleavage site (PRRAR) at the S1-S2 boundary of the S protein (5). It greatly mediates the fusion of human cell and viral membranes, and the rapid human-to-human virus transmission (5-7). That acquisition was achieved through the insertion of four amino acids (PRRA) in the S protein. However, this site is common in viral proteins, such as the hemagglutinin (H5) protein of the avian influenza viruses (3) or the S protein of some of the seventh coronavirus known to infect humans (5): HCoV-HKU1 (RRKRR-756, coordinate based on S protein), HCoV-OC43 (RRSRR-764) and MERS-CoV (MLKRR-700).

Another notable characteristic of the SARS-CoV-2 is the CGG-CGG coding sequence of the arginine dimer in that polybasic furin cleavage site. In the genetic code, arginine is encoded by six codons AGA, AGG, CGC, CGA, CGG and CGT codons. CGG is a minority arginine codon in SARS-CoV-2 (8). Consistently, CGG-CGG encoded arginine dimers at viral polybasic furin cleavage sites have not been found (9). In this sense, SARS-CoV-2 has the most extreme CpG deficiency in all known Betacoronavirus genomes, probably to avoid the human antiviral defence, mediated by the zinc finger antiviral protein (ZAP) (10). On the other hand, the other thirteen SARS-CoV-2 proteome arginine dimers, which are strictly conserved in the closest Sarbecovirus strains, are not CGG-CGG encoded either (11).

Now the question: is the a CGG-CGG encoded arginine dimer unique to the SARS-CoV-2 polybasic furin cleavage site? Based on the NCBI Virus database as a source of information, through a bioinformatics approach and using Perl scripts, all current Betacoronavirus arginine dimers and their coding regions are here reported. Full results are available in a Google Drive Folder. Interestingly, arginine dimers were widely distributed in Betacoronavirus proteins, about 30% of them contained one or more of the amino acid pair. These proteins were mostly members of the non-structural orf1ab-polyportein complex, and also in the structural S protein and nucleocapsid phosphoprotein. As regards the arginine codon usage focused on the Betacoronsvirus arginine dimers, AGA was the majority (about 50%), followed by CGT (about 24%). CGG was minoritary (about 5%).

As main result, Table 1 summarizes the Betacoronavirus arginine dimers, that were encoded by CGG-CGG. However, the most remarkable discovery was the CGG-CGG arginine dimer of the polybasic furin cleavage site of S protein from a set of MERS-CoVs. In addition, the Betacoronavirus species MERS-CoV, *Rousettus* and *Eidolon helvum* bat coronavirus HKU9 and Rabbit coronavirus HKU14 had CGG-CGG encoded arginine dimers in their orf1ab-polyprotein. Within the SARS-CoV-2 species, only two SARS-CoV-2 isolates from North America showed a CGG-CGG arginine dimer in the orf1ab-polyprotein, and few SARS-CoV-2 isolates, also from North America, showed the first (out of four) nucleocapsid phosphoprotein arginine dimer encoded by CGG-CGG.

To quantify the probability of the presence of the CGG-CGG genetic footprint encoding arginine pairs in Betacoronavirus, a First-Order Markov Chain was defined, based on the corresponding arginine codons. The states of the Markov model were the arginine codons themselves. Markov nodel allowed to determine the probability of the second arginine codon depending on the previous codon.

Since arginine has six codons, in an arginine dimer there are 36 (6 x 6) chances of finding an codon pair (like a roll of two dice: 36 possible outcomes). In this sense, the frequencies of the arginine codons pairs could be understood as the probability that the second codon depends on the presence of first codon. By normalizing these frequencies, the stochastic matrix of the Markov chain could be created, whose elements are the transition probability between codons (states). As an example, Table 2 shows a stochastic Markov matrix, based on the arginine dimers found in a recent Asia Betacoronavirus protein sample. In this sample, if the first codon was AGG or CGA, the second was most likely AGA. If the first codon was CGG, the second was most likely CGT.

The elements on the main diagonal mean the probability that the second codon was the same as the first. The significant presence of two arginine condons in a row occurred in AGA-AGA.

In this work, 9,017,367 Betacoronavirus protein sequences, and their coding regions, have been analysed. The sample had 14,993,143 arginine dimers (RR). It is a large sample which grow day by day. Accordingly, the present results (available in a Google Drive Folder) are also updated. Excluding the arginine dimers of the SARS-CoV-2 furin site (P**RR**AR), which are encoded by CGG-CGG, the results shown here highlight that this genetic footprint encoding arginine dimers are highly unlikely in betacoronaviruses wildlife, but they are there. Just because that presence, recombination probably operated into the origin of the virus CGG-CGG arginine dimer in the S1/S2 cleavage site. Recombination is the common method of viruses picking up new skills (12-15).

Full updated results:

https://drive.google.com/drive/folders/1Dp04BHDyMay1sB0GX0O0IFzfZTp_VrBu?usp=sharing

Acknowledgements

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

Competing interest declaration

Author declare that he has no conflicts of interest.

References

1. National Center for Biotechnology Information (NCBI). NCBI Virus database. Accessed September 23, 2021. https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/. Eneida L Hatcher, Sergey A Zhdanov, Yiming Bao, Olga Blinkova, Eric P Nawrocki, Yuri Ostapchuck, Alejandro A Schäffer, J Rodney Brister. Virus Variation Resource-improved response to emergent viral outbreaks. Nucleic Acids Res. 45(D1):D482-D490, 2017. PMID: 27899678 doi: 10.1093/nar/gkw1065.

2. Michael J. Harms, Jamie L. Schlessman, Gloria R. Sue, and Bertrand García-Moreno. Arginine residues at internal positions in a protein are always charged. Proc. Natl. Acad. Sci. U S A. 108(47):18954-18959, 2011. PMID: 22080604. doi:10.1073/pnas.1104808108.

3. 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.

4. Imène Kara, Marjorie Poggi, Bernadette Bonardo, Roland Govers, Jean-François Landrier, Sun Tian, Ingo Leibiger, Robert Day, John W M Creemers, Franck Peiretti. The Paired Basic Amino Acid-cleaving Enzyme 4 (PACE4) Is Involved in the Maturation of Insulin Receptor Isoform B. J. Biol. Chem. 290:2812-2821, 2015. PMID: 25527501. doi: 10.1074/jbc.M114.592543.

5. Kristian G. Andersen, Andrew Rambaut, W Ian Lipkin, Edward C Holmes, Robert F Garry. The proximal origin of SARS-CoV-2. Nat. Med. 26(4):450-452, 2020. PMID: 32284615. doi: 10.1038/s41591-020-0820-9.

6. Philip V'kovski, Annika Kratzel, Silvio Steiner, Hanspeter Stalder, Volker Thiel. Coronavirus biology and replication: implications for SARS-CoV-2. Nat. Rev. Microbiol. 19(3):155-170, 2021. PMID: 33116300. doi: 10.1038/s41579-020-00468-6.

7. Markus Hoffmann, Hannah Kleine-Weber, Stefan Pöhlmann. Multibasic Cleavage Site in the Spike Protein of SARS-CoV-2 Is Essential for Infection of Human Lung Cells. Mol. Cell 78(4):779-784, 2020. PMID: 32362314. doi: 10.1016/j.molcel.2020.04.022.

8. Xuhua Xia. Domains and Functions of Spike Protein in Sars-Cov-2 in the Context of Vaccine Design. Viruses. 13(1):109, 2021. PMID: 33466921. doi: 10.3390/v13010109.

9. Antonio R. Romeu, Enric Ollé. SARS-CoV-2 and the Secret of the Furin Site. Preprints 2021, 2021020264. doi: 10.20944/preprints202102.0264.v1.

10. Shuai Xia, Qiaoshuai Lan, Shan Su, Xinling Wang, Wei Xu, Zezhong Liu, Yun Zhu, Qian Wang, Lu Lu, Shibo Jiang. The role of furin cleavage site in SARS-CoV-2 spike protein-mediated membrane fusion in the presence or absence of trypsin. Signal Transduct. Target Ther. 5(1):92, 2020. PMID: 32532959. doi: 10.1038/s41392-020-0184-0.

11. Antonio R. Romeu, Enric Ollé. The SARS-CoV-2 arginine dimers. Research Square Preprints 2021. doi:10.21203/rs.3.rs-770380/v1.

12. Simon-Loriere, E. & Holmes, E.C. Why do RNA viruses recombine? *Nat. Rev. Microbiol.* **9**, 617-26 (2011). PMID: 21725337. doi: 10.1038/nrmicro2614.

13. Zhou, H. et al. Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses. *Cell* **184**, 4380-4391.e14 (2021). PMID: 34147139. doi: 10.1016/j.cell.2021.06.008.

14. Li, L.-L. et al. A novel SARS-CoV-2 related coronavirus with complex recombination isolated from bats in Yunnan province, China. *Emerg. Microbes Infect.* **10**, 1683-1690 (2021). PMID: 34348599. doi: 10.1080/22221751.2021.1964925.

15. Smriti Mallapaty. Closest known relatives of virus behind COVID-19 found in Laos. Nature 597, 603 (2021). PMID: 34561634. doi: https://doi.org/10.1038/d41586-021-02596-2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2021 doi:10.20944/preprints202110.0080.v1

Table 1. CGG-CGG encoded arginine dimer from Betacoronavirus protein sequences. The list is limited to records that exclude those from the SARS-CoV-2 polybasic furin cleavage site (P**RR**AR). The data is grouped by Geographic Regions

Species*	Isolate	Protein**	Acession	Length	Dimer	Protein Position	Coding	Gene/Genome Position	Host
Africa									
MERS-CoV	camel/Nigeria/NV1787/2016	S protein	AVN89398	1353	RR	700	CGGCGG	2100	*Camelus dromedarius*
MERS-CoV	camel/Nigeria/NV1712/2016	S protein	AVN89453	1353	RR	700	CGGCGG	2100	*Camelus dromedarius*
HKU9	PREDICT-GVF-CM-ECO06464	RdRp	ATU79936	112	RR	85	CGGCGG	255	*Rousettus aegyptiacus*
MERS-CoV	camel/Nigeria/NV1673/2016	S protein	AVN89387	1353	RR	700	CGGCGG	2100	*Camelus dromedarius*
MERS-CoV	camel/Nigeria/NV1657/2016	S protein	AVN89442	1353	RR	700	CGGCGG	2100	*Camelus dromedarius*
HKU9	PREDICT-GVF-CM-ECO06646	RdRp	ATU79938	112	RR	85	CGGCGG	255	*Eidolon helvum*
MERS-CoV	camel/Nigeria/NV2020/2016	S protein	AVN89409	1353	RR	700	CGGCGG	2100	*Camelus dromedarius*
MERS-CoV	camel/Nigeria/NV2040/2016	S protein	AVN89420	1353	RR	700	CGGCGG	2100	*Camelus dromedarius*
MERS-CoV	camel/Nigeria/NV1989/2016	S protein	AVN89431	1353	RR	700	CGGCGG	2100	*Camelus dromedarius*
MERS-CoV	camel/Nigeria/NV1405/2016	S protein	AVN89376	1353	RR	700	CGGCGG	2100	*Camelus dromedarius*
Asia									
HKU14		polyprotein	AFE48811	7151	RR	6763	CGGCGG	20289	*Oryctolagus cuniculus*
HKU9	Rousettus spp/Jinghong/2009	ORF1ab	AVP25405	6920	RR	4951	CGGCGG	14853	*Rousettus sp.*
HKU9		ORF1ab	ADM33573	6923	RR	2569	CGGCGG	7707	*Chiroptera*
HKU14		polyprotein	AFE48810	7151	RR	6763	CGGCGG	20289	*Oryctolagus cuniculus*
HKU14		polyprotein	AFE48822	7112	RR	6724	CGGCGG	20172	*Oryctolagus cuniculus*
HKU14		nsp15	YP_009924422	375	RR	286	CGGCGG	858	*Oryctolagus cuniculus*
HKU14		ORF1ab	YP_005454239	7151	RR	6763	CGGCGG	20289	*Oryctolagus cuniculus*
HKU9		ORF1ab	ABN10926	6923	RR	2569	CGGCGG	7707	*Chiroptera*
MERS-CoV		ORF1ab	AHY61336	7179	RR	6174	CGGCGG	18522	*Vespertilio sinensis*
HKU9		ORF1ab	ADM33557	6923	RR	2569	CGGCGG	7707	*Chiroptera*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2021 doi:10.20944/preprints202110.0080.v1

HKU14		polyprotein	AFE48800	7151	RR	6763	CGGCGG	20289	*Oryctolagus cuniculus*
HKU9		ORF1ab	ADM33565	6903	RR	2041	CGGCGG	6123	*Chiroptera*
North	America								
SARS-CoV-2	NC-CDC-LC0027271	ORF9	QTG55296	419	RR	41	CGGCGG	123	*Homo sapiens*
SARS-CoV-2	JW0066	ORF9	QXX31736	419	RR	41	CGGCGG	123	*Homo sapiens*
SARS-CoV-2	FL-BPHL-5767	ORF9	QZW21341	419	RR	41	CGGCGG	123	*Homo sapiens*
SARS-CoV-2	NC-CDC-STM-000025458	ORF9	QTC13492	419	RR	41	CGGCGG	123	*Homo sapiens*
SARS-CoV-2	NC-SLPH-0070	ORF9	QTX13191	419	RR	41	CGGCGG	123	*Homo sapiens*
SARS-CoV-2	NC-CDC-STM-000028277	ORF9	QTX74547	419	RR	41	CGGCGG	123	*Homo sapiens*
SARS-CoV-2	MI-CDC-STM-000045686	ORF9	QTW56377	419	RR	41	CGGCGG	123	*Homo sapiens*
SARS-CoV-2	NC-CDC-STM-000032315	ORF9	QTX83255	419	RR	41	CGGCGG	123	*Homo sapiens*
SARS-CoV-2	JW0864	ORF9	QXX38166	419	RR	41	CGGCGG	123	*Homo sapiens*
SARS-CoV-2	NC-CDC-STM-000026863	ORF9	QTC16696	419	RR	41	CGGCGG	123	*Homo sapiens*
SARS-CoV-2	MI-CDC-STM-000046692	ORF9	QTW59960	419	RR	41	CGGCGG	123	*Homo sapiens*
SARS-CoV-2	CA-CDC-ASC210119629	ORF1ab	UBF72283	7096	RR	5767	CGGCGG	17301	*Homo sapiens*
SARS-CoV-2	MD-MDH-4405	ORF1ab	UAB59607	7096	RR	5767	CGGCGG	17301	*Homo sapiens*

* Full name and taxonomic identifier of the Betacoronavirus species: SARS-CoV-2, Severe acute respiratory syndrome coronavirus 2 (taxid:2697049); MERS-CoV, Middle East respiratory syndrome-related coronavirus (taxid:1335626); HKU9, Rousettus bat coronavirus HKU9 (taxid:694006); HKU14, Rabbit coronavirus HKU14 (taxid:1160968).

** Protein name: RdRp, RNA-dependent RNA polymerase; ORF1ab, orf1ab polyprotein; nsp15, non structural protein 15; ORF9, nucleocapsid phosphoprotein.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2021 doi:10.20944/preprints202110.0080.v1

Table 2. Stochastic matrix of the First-Order Markov Chain defined by the arginine codon sequence encoding the arginine dimers of the Asia Betacoronavirus protein sample.

	AGA	AGG	CGA	CGC	CGG	CGT	Sum
AGA	0.29218655	0.18973403	0.11623226	0.01059752	0.00056244	0.39068721	1
AGG	0.80859016	0.00471785	0.07812021	0.00251210	0.08050977	0.02554991	1
CGA	0.81517094	0.00356125	0.01709402	0.05864198	0.00356125	0.10197056	1
CGC	0.21081081	0.35091892	0.04345946	0.07221622	0.01167568	0.31091892	1
CGG	0.08980123	0.00070989	0.10151443	0.01123994	0.00153810	0.79519640	1
CGT	0.39124861	0.27172256	0.00256375	0.02674612	0.30058204	0.00713692	1

The stochastic matrix is a square matrix of transition probability between arginine codons (states). The rows are probabilistic vectors. An element of the matrix means the probability that the second arginine codon would be that of the column if the first is that of the row. Consequently, the sum of the elements of a row is 1. Data used to create this Betacoronavirus-Asia stochastic Markov matrix: Total number of analysed Betacoronavirus protein sequences, 93,977; total number of protein sequences having arginine dimer(s), 34,346 (36.55%); total number of arginine dimers in the sample , 134,249; total number of SARS-CoV-2 (CGG-CGG) polybasic furin cleavage site arginine dimers, 6,859 (5.11%). To avoid distortions in calculations of transition probabilities between arginine codons, the arginine dimers of the SARS-CoV-2 furin site were excluded.