International Journal of Coronaviruses
Open Access - Peer-Reviewed - ISSN 2692-1537

Portico Archived | Crossref DOI | CC BY 4.0 Licensed

Research Article

An Algorithm to Predict the Possible SARS-CoV-2 Mutations

Raúl Isea [1*]

[1] Fundación Instituto de Estudios IDEA, Hoyo de la Puerta, Baruta, Venezuela.
*** Correspondence:** Raúl Isea, Fundación Instituto de Estudios IDEA, Hoyo de la Puerta, Baruta, Venezuela

Published citation: Raúl Isea (2021) **An Algorithm to Predict the Possible SARS-CoV-2 Mutations**. *International Journal of Coronaviruses* - 3(1):1-7. https://doi.org/10.14302/issn.2692-1537.ijcv-21-3804

Article DOI: 10.14302/issn.2692-1537.ijcv-21-3804

Abstract

An algorithm to determine the possible mutations that can occur in the S protein responsible of the Covid-19 in humans is designed. To do that, nine tridimensional sequences available in the Protein Data Bank similar to the initial strain sequenced in Wuhan (December 2019) are identified. The conditions driving this potential mutation are: (1) an accumulated number of mutations greater than (or equal to) 5 in each position; (2), a cumulative value of the different variations of Gibbs free energy less than -2.0 Kcal/mol; and (3), a squared fluctuation greater than 1.6 Å obtained according to calculations for normal mode analysis based on anisotropic network models (ANM) after averaging the first 20 vibration modes. The result is that 491 positions can mutate, while 424 positions did not provide any mutation. Finally, the results reveal that there are mutations that cannot be predicted, so more studies are needed to determine why they are present in the human population.

Keywords:

Covid-19; SARS-CoV-19; Mutation; ΔΔG; Gibbs; ANM; Python

Article Information	
Academic Editor	Amin Ataie, Babol university of Medical science, department of Pharmacology and toxicology, Iran.
Received Date	5 April 2021
Accepted Date	13 April 2021
Published Date	17 April 2021
Online	17 April 2021
Publication	Volume 3; Issue 1; Pages 1–7
Copyright	© 2021 Raúl Isea

Article Information (continued)

License	This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. https://creativecommons.org/licenses/by/4.0/

Introduction

At the end of December 2019, the first episodes of Covid-19 were registered from patients from the Huanan Seafood Wholesale Market in the city of Wuhan (China) who presented a new atypical pneumonia, fever, cough, and in the most severe cases, dyspnea and bilateral lung infiltration. In view of this, on December 31, the Wuhan Municipal Health Commission reported the incident to the World Health Organization (WHO).

The genome of the virus was made public on January 2020 [1]. This study allowed the International Committee on Taxonomy of Viruses to rename it as SARS-CoV-2. From there, the virus began to spread to other cities in China, and later on to other countries in the world. In view of the high number of infections, the WHO determined a new Covid-19 pandemic on March 11.

The genome data was published in 2020 [1]. It was observed that it is a new betacoronavirus. It is 79% and 50% identical with respect to SARS-CoV and MERS-CoV, respectively. In other words, Covid-19 seems to be more related to the episode registered in 2002 rather than the incident that occurred in 2012. This observation is being investigated in more detail and will be published in a future work.

In this paper, there is a focus on the S protein because it is involved in the process of entering the virus into the receptor, and therefore it is a target for the design of possible vaccines against Covid-19. Recall that the S protein of SARS-CoV-2 consists of approximately 1,273 amino acids (aa), slightly higher than that found in SARS-CoV (1,255 aa).

Up to date, the emergency use of the vaccines Sputnik V from Russia, Sinopharm from China, among others, have been authorized. They are based on the analysis of the sequences that occurred in the initial studies of the virus, but a range of mutations can occur such that they are inefficient over time. Hence the need to predict possible mutations that can modify the effectiveness of vaccines [2,3].

The first mutations registered in Brazil were found in a patient in Rio of Janeiro infected in October 2020 [4]. This mutation (E484S) spread to multiple countries including USA, Singapore, Argentina, Denmark, Ireland, England, Canada, etc.; but it was not found in Africa (remember that this variant is also known as 484 K.V2).

The second variant in Brazil has the following mutations: L18F, T20N, P26S, D138Y, R190S, K417T, N501Y, D614G, H655Y, and V1176F in the S protein [5], and was designated as B.1.1.248. It is interesting to see that the variant formed by K417T, E484K, and N501Y was independently named 28-AM-I, although both variants have also been assigned as P.1 variant as part of the B.1.1.28 nomenclature.

In view of this, nine three-dimensional (3D) structures of the S protein were selected to avoid any ambiguity in the results. Subsequently, an exploration of all the possible mutations that occur in each of the positions that make up the S protein was carried out, using the calculations of Gibbs free energy ($\Delta\Delta G$) [6] as described in the next section. Finally, three scenarios will be handled in order to determine if a mutation in such a protein is possible.

Methodology

From one of the sequences published after the incident in the Wuhan market in December 2019, nine structural sequences of the S protein were selected from the Protein Data Bank (www.pdb.org), using the tools from the NCBI portal (blast.ncbi.nlm.nih.gov/). These sequences should provide a similarity higher than 98% with respect to the Wuhan sequence.

To consider a mutation as valid, two out of these three conditions should be met: (1) positions where at least five or more possible mutations (the half of the selected sequences) can occur; (2) amino acids that present a quadratic fluctuation equal to or greater than 1.6 Å obtained from an anisotropy network model calculations; and (3), the accumulation of the variation Gibbs energy is less than -2.0 Kcalc/mol.

The first mutations registered after the initial outbreak in Wuhan are analyzed, *i.e.:* Y28, A67, N74, W152, Y200, R273, F275, L276, E298, K300, T302, G485, A570, D614, A653, L752, P812, I818, G838, F1103, and V1104 [7-11]. In addition, to corroborate whether the two variants registered in Brazil (until February 2021) could be predicted with the results of the paper, the E484K mutation was initially registered in a patient infected in Rio of Janeiro in October 2020 [3], and later a pool of mutations that occurred in the second Brazilian variant (also known as P.1) are present, which are: L18F, T20N, P26S, D138Y, R190S, K417T, N501Y, D614G, H655Y, and V1176F [3].

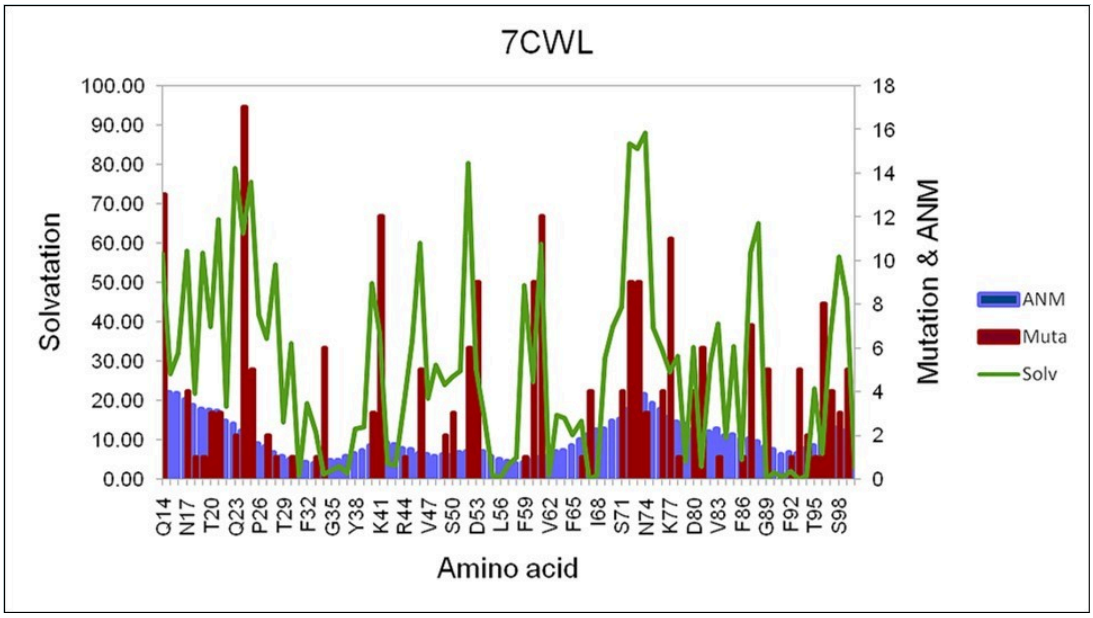

Figure 1. Selected region of the 7CWL sequence between Q14 until S98 of the S protein. The degree of exposure to the solvent is depicted in green color, the fluctuation due to the square displacement is depicted in light blue, and the number of mutations in red color.

Results

The amino acid sequence of the S protein selected in the work corresponds to one of the first episodes registered in Wuhan (China) in December 2019, whose NCBI ID was MN908947. Thus, nine of their sequences were randomly selected from the result obtained with the Blastp program, which are deposited in the Protein Data Bank (PDB).

The nine sequences selected in humans correspond to the following PDB identifiers: 7JJI, 6VSB, 7KDI, 7KDJ, 6ZOW, 6XCM, 7CWL, 7K8S, and 7C2L. The next step was to calculate for each one of them, all the possible mutations that can occur from the calculation of the variation of Gibbs free energy after using the PoPMuSiC program.

Figure 1 shows the number of mutations, the degree of exposure to the solvent, and the quadratic fluctuation in the region between Q14 and S98 in the 7CWL sequence. The figure shows how the number of mutations, the degree of exposure to the solvent, and the quadratic fluctuation are correlated throughout this region.

Figure 2 shows the result of the evaluation of the possible mutations that may occur at position 200 of the S protein found in seven sequences selected in the work. This figure verifies the excellent agreement of the results obtained between all of them, as well as the average value depicted in blue (noted as <>).

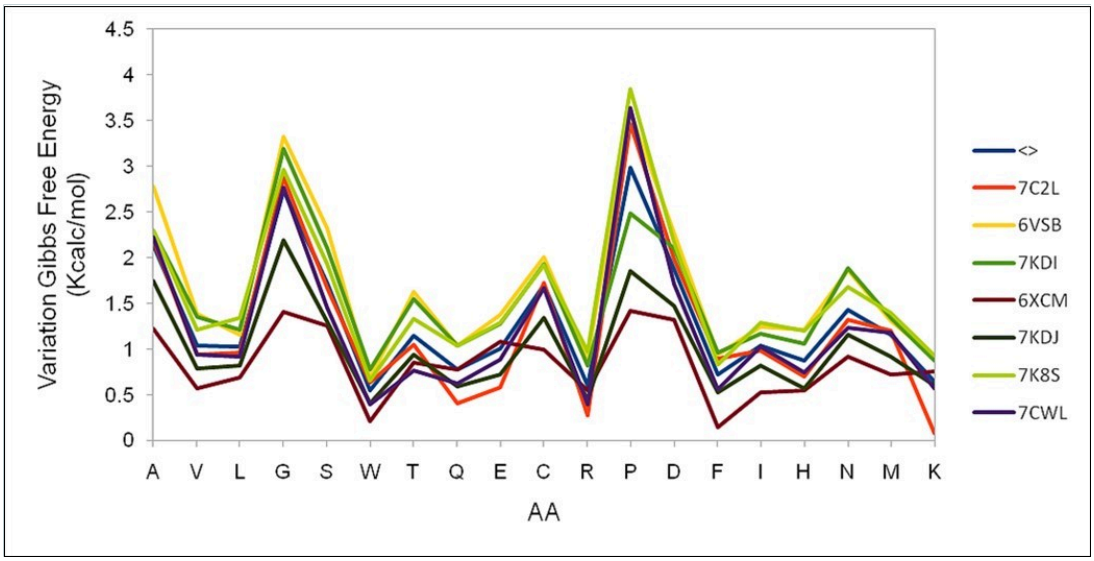

Figure 2. Different results for the variation of the Gibbs free energy at position 200 of the S protein obtained in seven different sequences are shown. The average value is depicted in blue.

Figure 3 shows the possible mutations in a small region between the positions Q14 and I68 in five different sequences. The distribution is not uniform as in the previous case, so the accumulated value of the mutations is considered instead of the average value.

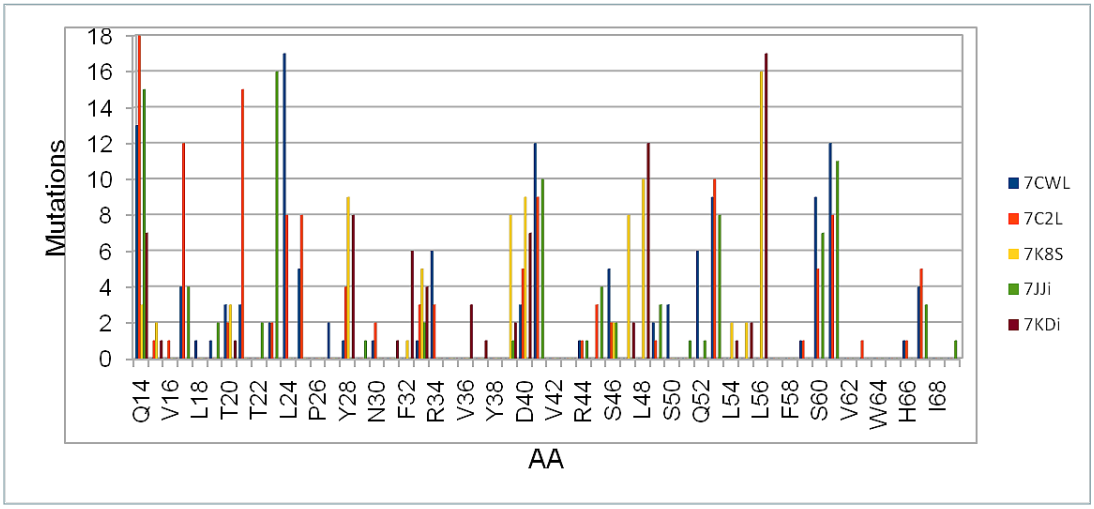

Figure 3. Cumulative number of mutations obtained in the region between Q14 and I68 of the S protein obtained from five sequences considered in the work.

It should be noted that a total of 177,463 different Gibbs free energy calculations must be analyzed in the nine sequences selected in the work, so it was necessary to implement small scripts in the Python programming language to analyze those results.

The calculations revealed that 424 positions do not present any mutation in the nine sequences selected (indicated in Table 1). Table 2 shows the positions that meet two out of the three conditions proposed in this paper.

Table 1. The 424 amino acid positions where no mutation was detected:

P26, S31, F32, G35, V36, Y37, Y38, V42, F43, V47, L48, L54, F55, L56, P57, F58, V62, W64, V70, L84, V90, Y91, I100, I101, W104, I105, F106, L110, L117, L118, I119, V120, V126, V127, I128, V130, C131, F133, Y144, Y145, S151, S155, V159, Y160, C166, P174, F175, G184, F186, F192, V193, F194, I197, G199, Y200, F201, I203, Y204, L216, F220, L223, P225, I235, T236, F238, L241, G252, D253, S254, S255, W258, A260, Y265, Y266, V267, Y269, L270, F275, L276, Y279, G283, T284, I285, V289, D294, L296, L303, S305, F306, G311, I312, Y313, S316, F318, I326, F329, C336, P337, V341, F342, F347, S349, W353, I358, N360, C361, Y365, L368, F374, T376, F377, C379, G381, V382, S383, L387, L390, C391, F392, T393, V395, F400, V401, I402, D405, V407, Q409, I410, P412, G413, T415, G416, I418, Y421, Y423, L425, F429, C432, V433, W436, G447, N448, Y451, Y453, L455, F456, L461, P463, F464, S469, I472, Y473, C480, G482, V483, N487, C488, L492, F497, P507, Y508, V510, V511, V512, L513, T523, V524, C525, L533, V534, N536, C538, V539, F541, F543 ,N544, G545, L546, V551, L552, F559, P561, F562, F565, V576, P579, I584, L585, I587, P589, C590, F592, V595, V597, I598, V610, L611, Y612, V615, W633, V635, Y636, S637, V642, F643, G648, C649, L650, I651, C662, I664, P665, I666, G667, A668, G669, I670, S673, Y674, I692, Y695, M697, L699, Y707, N710, I712, A713, I714, T716, N717, F718, I720, V722, T723, I726, L727, V729, M731, V736, D737, C738, M740, Y741, I742, S746, C749, L752, L753, Y756, G757, S758, F759, Q762, L763, A766, E773, Q774, K776, Q779, V781, F782, V785, I788, Y789, F797, G798, G799, F802, S803, I805, L806, P807, R815, S816, I818, L821, L822, F833, I834, C840, L841, G842, D843, I844, A845, A846, L849, I850, C851, F855, G857, L858, V860, L861, P863, L864, L865, E868, M869, I870, A871, Q872, Y873, L877, I882, G885, W886, F888, A890, L894, I896, P897, M900, M902, F906, G910, V911, L916, Y917, N919, I923, F927, I934, L948, V951, Q954, Q957, T961, L962, L966, S968, F970, G971, S974, S975, I980, L984, A989, I993, R995, L996, L1001, Q1002, L1004, Q1005, T1006, Y1007, V1008, Q1010, Q1011, L1012, I1013, R1014, A1015, I1018, R1019, A1025, A1026, C1032, V1033, L1034, G1035, R1039, F1042, C1043, G1044, G1046, Y1047, H1048, L1049, P1057, V1060, V1061, F1062, L1063, V1065, Y1067, V1068, F1075, I1081, C1082, G1085, F1089, P1090, G1093, V1094, F1095, V1096, G1099, W1102, F1103, V1104, F1109, Y1110, P1112, I1115, T1116, F1121, V1122, G1124, G1131, V1137, P1140, L1141

Table 2. The 491that canoccur a mutation (the mutations observed in the Brazilian variants are highlighted in bold for quick visualization).

Q14, N17, **T20**, R21, Q23, L24, P25, A27, **Y28**, R34, D40, K41, R44, S45, S46, Q52, D53, S60, N61, H66, A67, I68, H69, G72, T73, N74, K77, R78, D80, N81, P82, P85, N87, D88, G89, A93, S94, T95, E96, K97, N99, R102, G103, T108, D111, S112, K113, Q115, S116, N121, N122, A123, T124, N125, K129, E132, Q134, C136, N137, **D138**, P139, F140, L141, G142, V143, K147, N148, K150, M153, E154, E156, R158, S162, A163, N164, N165, E169, S172, Q173, L176, M177, D178, L179, E180, K182, Q183, K187, N188, **R190**, E191, K195, N196, D198, K202, S205, K206, H207, T208, P209, I210, N211, L212, V213, R214, D215, Q218, S221, E224, L226, D228, P230, I231, N234, R237, Q239, T240, L242, A243, L244, H245, R246, Y248, L249, S256, T259, G261, A262, A263, A264, G268,Q271,**R273**,K278,E281,N282, D287, D290, A292, **K300**, C301, T302, K304, E309, K310, Q314, N317, Q321, T323, E324, S325, R328, P330, N331, T333, N334, L335, F338, G339, E340, N343, T345, R346, N354, R355, K356, R357, V362, A363, D364, V367, N370, S371, A372, S373, S375, K378, T385, K386, N388, N394, A397, D398, S399, R403, G404, E406, Q414, **K417**, D420, N422, K424, D427, D428, T430, G431, A435, N439, N440, L441, D442, S443, K444, V445, G446, Y449, N450, R454, R457, K458, N460, K462, E465, R466, I468, T470, E471, Q474, A475, G476, S477, T478, P479, E484, G485, F486, Y489, F490, P491, Q493, S494, G496, Q498, T500, **N501**, G502, V503, G504, S514, E516, L517, L518, H519, A520, G526, P527, K528, K529, S530, T531, N532, K535, K537, N540, N542, G548, E554, S555, N556, K557, K558, Q563, Q564, G566, R567, D568, I569, D571, T572, T573, D574, A575, R577, Q580, L582, E583, D586, S591, G594, S596, T599, G601, N603, N606, Q607, A609, Q613, **D614**, N616, C617, T618, E619, N641, Q644, R646, A653, E654, **H655**, V656, N657, N658, S659, Y660, E661, D663, A672, Q675, T676, Q677, S689, Q690, S691, I693, A694, T696, A701, E702, N703, S704, V705, A706, N709, S711, T724, E725, K733, T734, S735, D745, T747, S750, N751, L754, C760, T761, N764, R765, T768, G769, A771, V772, D775, N777, T778, E780, A783, K786, Q787, K790, T791, P793, I794, K795, N801, Q804, D808, P809, K811, P812, K814, N824, K825, T827, A831, K835, Q836, G838, D839, Q853, K854, N856, P862, T866, D867, T874, A879, G880, T881, T883, S884, G891, A892, A893, Q895, A899, R905, N907, G908, I909, Q913, N914, E918, K921, L922, A924, N925, N928, S929, I931, G932, K933, Q935, D936, S937, S939, S940, T941, A942, S943, L945, G946, Q949, D950, N953, N955, L959, N960, K964, Q965, N969, V976, N978, D979, K986, V987, E988, E990, V991, Q992, D994, T998, G999, R1000, S1003, A1020, S1021, A1022, K1028, S1030, Q1036, K1038, K1045, S1051, P1053, Q1054, S1055, G1059, H1064, T1066, Q1071, E1072, K1073, N1074, T1077, A1078, P1079, A1080, H1083, D1084, K1086, A1087, H1088, R1091, E1092, S1097, N1098, T1100, H1101, T1105, Q1106, E1111, Q1113, I1114, T1117, D1118, T1120, S1123, N1125, D1127, V1128, I1130, I1132, V1133, N1134, N1135, T1136, Q1142, P1143, E1144, L1145, D1146, S1147

Table 3 shows some of the mutations identified in the scientific literature, where the amino acid and position are indicated in the first column (AA) as well as the cumulative number of predicted mutations (Mut). The next column shows the accumulated value of the Gibbs free energy variations obtained by adding all the different contributions of the possible mutations (an example of this calculation is shown below), and the fourth column shows the mean square fluctuation obtained with an ANM methodology.

It is not possible to predict some mutations ie., Y200, F275, L276, L752, I818, F1103, and V1104 (Table 3). In fact, when checking the results in positions F275, L276, and I818, do not present any mutation. However, other mutations were predicted, such as Y28 or D614. The R273 and K300 positions are determined by a high number of mutations.

In order to understand the accumulated value of the Gibbs free energy variations (ΔΔG accum), the results obtained at position D614 are selected. This position has been found to be mutated to a GLY (G). The results obtained in each of the nine sequences are as follows:

6VSB: -1.04, CYS, PHE, GLY, HIS, ASN

6XCM: -2.11, CYS, PHE, GLY, HIS, ASN, TYR

7K8S: -1.35, CYS, GLY, HIS, ASN

7C2L: -0.58, CYS, GLY, HIS, ASN, VAL

7KDJ: GLY, 0,

7KDI: GLY, 0,

7JJI: -2.61, CYS, PHE, LEU, TRP, TYR

6ZOW: -1.48, CYS, PHE, GLY, HIS, ASN, PRO, THR, VAL, TYR

7CWL: -2.86, CYS, PHE, GLY, HIS, LEU, MET, ASN, TYR

Table 3. Mutations observed in the S protein (AA) as well as the accumulated mutations obtained (Mut). The cumulative value of the different variations of Gibbs free energy at that position (ΔΔG accum), and the quadratic displacement obtained with ANM.

AA	Mut	ΔΔG acumu.	<ANM>
Y28	14	-1,58	1,8
A67	16	-2,96	3,2
N74	14	-2,29	4,7
W152	1	-0,04	4,8
Y200	0	0,00	3,1
R273	44	-6,29	1,4
F275	0	0,00	1,1
L276	0	0,00	1,1
E298	14	-1,50	1,0
K300	63	-31,64	1,1
T302	38	-6,62	1,2
G485	6	-0,91	23,2
A570	2	-0,27	4,8
D614	42	-12,03	1,7
A653	36	-10,00	1,3
L752	0	0,00	6,5
P812	6	-0,37	2,3
I818	0	0,00	1,1
G838	15	-5,18	3,5
F1103	0	0,00	3,4
V1104	0	0,00	3,1

Where the 6VSB sequence predicted five possible mutations (Cys, Phe, Gly, His, and Asn) such that the total sum of the five variations of the Gibbs free energy is equal to -1.04 Kcal/mol. The 7K8S sequence predicts four mutations and the cumulative Gibbs energy variation is -1.35 Kcalc/mol, and so on. Therefore, it can be verified that in position 614, 42 mutations present in 7 sequences are predicted, while two of them do not predict any mutation (7KDI and 7KDJ).

When reviewing the different mutations that are predicted in each of the seven sequences, it can be seen that the most frequent amino acid is a Gly (G), occurring 8 times. Cys (C) also appears 7 times, among others. Hence, an accumulated Gibbs energy of -12.03 Kcalc/mol was found (result of the sum of -1.04, -2.11, -1.35, - 0.58, -2.61, -1.48, and -2.86).

Finally, it is verified that the mutations registered in the new Brazil variants also appear in the results of this work, which are T20N, D138Y, R190S, K417T, N501Y, D614G, and H655Y. The L18 and P26 positions do not count on predicted mutations, and unfortunately the position V1176 was not present in the sequences.

Conclusion

This work determines the different positions where a mutation can occur in the S protein in order to explain the different variants that are occurring in SARS-CoV-2. It is interesting to note that it is possible to actually predict those observed in the new variant of Brazil, but it was not possible to explain some of the mutations detected at the beginning of the contagion by Covid-19 (L18, P26, Y200, F275, L276, L752, I818, F1103 and V1104).

Acknowledgment

I'd like to acknowledge Rafael Mayo-Garcia for his comments on this manuscript.

References

1. Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., et al (2020). A new coronavirus associated with human respiratory disease in China. *Nature* Vol. 579(7798), 265-269. https://doi.org/10.1038/s41586-020-2008-3
2. Fontanet, A., Autran, B., Lina, B., Kieny, M.P., Karim, S.S.A., and Sridharj, D. (2021). *Lancet.* Vol 397(10278): 952–954.
3. dos Santos, W. G. (2021). Impact of virus genetic variability and host immunity for the success of COVID-19 vaccines. *Biomed Pharmacother.* Vol. 136: 111272. https://doi.org/10.1016/j.biopha.2021.111272
4. Voloch C.M., da Silva F Jr R., de Almeida L.G.P., et al. Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. 2020. *medRxiv* 2020.12.23.20248598; https://doi.org/10.1101/2020.12.23.20248598
5. Toovey O.T.R. , Harvey K.N., Bird P.W., and Tang J. W-T (2021) Introduction of Brazilian SARS-CoV-2 484K.V2 related variants into the UK. *J Infect.* 2021 Feb 3:S0163-4453(21)00047-5. doi: 10.1016/j.jinf.2021.01.025.
6. Dehouck, Y., Kwasigroch, J.M., Gillis, D., and Rooman, M (2011). PoPMuSiC 2.0: a web server for the estimation of protein stability of protein changes upon mutation and sequence optimality. *BMC Bioinformatics*, 12: 151. https://doi.org/10.1186/1471-2105-12-151
7. Durmaz, B., Abdulmajed, O., and Durmaz R (2020). Mutations Observed in the SARS-CoV-2 Spike Glycoprotein and Their Effects in the Interaction of Virus with ACE-2 Receptor. *Medeni Med J.* Vol. 35(3): 253–260. https://doi.org/10.5222/mmj.2020.98048
8. Chand, G.B., Banerjee, A., and Azada, G.K (2020). Identification of twenty-five mutations in surface glycoprotein (Spike) of SARS-CoV-2 among Indian isolates and their impact on protein dynamics. *Gene Rep.* 2020 Dec; 21: 100891. https://doi.org/10.1016/j.genrep.2020.100891
9. Chen, R.W.J., Gao, K., Hozumi, Y., Yin, C., and Wei, G-W. (2021). Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. *Commun Biol* 4, 228.
10. Teng, S., Sobitan, A., Rhoades, R., Liu, D., and Tang, Q. (2021). Systemic effects of missense mutations on SARS-CoV-2 spike glycoprotein stability and receptor-binding affinity, *Briefings in Bioinformatics*, Vol. 22(2): 1239–1253.
11. Singh, P.K., Kulsum, U., Rufai, S.B., Mudliar, S.R., and Singh, S (2020). Mutations in SARS-CoV-2 Leading to Antigenic Variations in Spike Protein: A Challenge in Vaccine Development. *J Lab Physicians.* Vol. 12(2): 154–160 https://doi.org/10.1055/s-0040-1715790

Publisher's Note

All claims, opinions, data, and conclusions expressed in this article are solely those of the author(s) and contributor(s), and do not necessarily represent those of Open Access Pub, the journal, its editors, or reviewers. Open Access Pub does not guarantee or endorse any product, method, instruction, or claim discussed in the article, and disclaims responsibility for any injury to people or property resulting from ideas, methods, instructions, or products referred to in the content.