Origen del SARS-CoV-2 desde una perspectiva Bioinformática

Origin of SARS-CoV-2 from Bioinformatics perspective

Raúl Isea[1]
Fundación IDEA, Hoyo de la Puerta, Caracas, Venezuela[1]
raul.isea@gmail.com[1]

Fecha de recepción: 16/04/2021
Fecha de aceptación: 05/05/2021
Pág: 2 – 12

Resumen

El SARS-CoV-2 es el tercer Coronavirus altamente patogénico que afronta la humanidad tras declararlo como una pandemia el 11 de marzo de 2020. Se ha planteado que el origen del SARS-CoV-2 es producto de una evolución natural del virus, aunque algunos trabajos sostienen la hipótesis de que fue creado en un laboratorio. Por esa razón, se realiza un estudio estadístico y bioinformático con énfasis en un análisis multivariable por componentes principales en diecinueve secuencias estructurales. Tras los resultados obtenidos, se puede inferir que el origen de la infección en humanos está estrechamente vinculado con los murciélagos en vez de los pangolines, y además se descarta la idea que fue creado en un laboratorio.

Palabras clave: Covid-19, SARS-CoV-2, Evolución, Origen, Laboratorio.

Abstract

SARS-CoV-2 is the third highly pathogenic Coronavirus facing humanity after declaring it a pandemic on March 11, 2020. It has been suggested that the origin of SARS-CoV-2 is the product of a natural evolution of the virus, although some papers support the hypothesis that it was created in a laboratory. For this reason, a statistical and bioinformatic study is carried out with emphasis on a multivariate analysis by principal components in nineteen virus sequences. It could be that the origin of the infection in humans is more closely related to bats instead of pangolins according to evolution studios, and it can also be discarded the idea that it was created in a laboratory.

Key words: Covid-19, SARS-CoV-2, Evolution, Origin, Laboratory.

ISSN: 2244-7423

Introducción

El 31 de diciembre de 2019, las autoridades de la Comisión de Salud de China informaron a la Organización Mundial de la Salud (OMS) un incidente acontecido con 27 personas que presentaban una neumonía de etiología desconocida en la ciudad de Wuhan (China).

Posteriormente las autoridades del Centro de Control de Enfermedades (CDC) de China señalaron el 7 de enero de 2020 que dicho incidente se trataba de un nuevo Coronavirus, tras haberlo secuenciando unos días antes [Wu et al., 2020]. Al incrementarse el número de personas contagiadas fuera del área de Wuhan se confirmó la transmisión entre humano y humano, y el 20 de enero de 2020 la OMS informó que se había vuelto un problema de salud pública de carácter internacional.

El Comité Internacional de Taxonomía denominó al nuevo Coronavirus SARS-CoV-2, y el 11 de marzo de 2020 fue declarado como una pandemia. Lamentablemente se han registrado más de 136 millones de personas contagiadas en más de 210 países para el 12 de abril de 2021 de acuerdo a los datos obtenidos en la Universidad Johns Hopkins, disponible gratuitamente en coronavirus.ujh.edu.

El secuenciamiento del genoma reveló su gran tamaño (aproximadamente 30 mil nucleótidos) con respecto a otros virus del tipo ARN, como por ejemplo, los 10 mil nucleótidos del SIDA o los 19 mil del Ébola. Asimismo, permitió determinar que el SARS-CoV-2 es un Betacoronavirus, perteneciente a la subfamilia *Coronavirinae*, del orden Nidovirales [Wu et al., 2020].

En paralelo se determinó que el genoma del SARS-CoV-2 es un 79 % y 50 % idéntico a los genomas del SARS-CoV y MERS-CoV, respectivamente; es decir, el SARS-CoV-2 pareciera estar más relacionado con los episodios registrados en 2002 en vez del incidente ocurrido en 2012 [Zhou et al., 2020], dando una brecha de tiempo para crear diversas especulaciones.

Recordemos que el SARS-CoV se originó en la provincia de Guandong (China) donde se confirmaron 8.098 casos y 744 fallecidos en 2002, mientras que el MERS-CoV se rastreó hasta Arabia Saudita (en 2012), registrando 2494 casos contabilizados hasta noviembre de 2018 [Zhou et al., 2020].

Gracias a estudios genéticos, se determinó que la glicoproteína de espícula (S) es la responsable en el proceso de entrada del virus en el receptor, es decir, es la vía por la cual se absorbe el virus en el epitelio respiratorio [Matheson y Lehner, 2020], y por ello es un blanco en el diseño de vacunas. Se debe tener presente que la glicoproteína S del SARS-CoV-2 consta de 1,273 aminoácidos (AA), ligeramente superior a la del SARS-CoV (1,255 AA). La

glicoproteína S presenta dos subunidades conocidas denotadas como S1 y S2. La subunidad S1 es la que se une a los receptores de la enzima convertidora de angiotensina 2 (ACE2), mientras que la subunidad S2 determina la fusión, permitiendo la entrada al virus por endocitosis [Matheson y Lehner, 2020].

Origen del SARS-CoV-2

La literatura científica ha sostenido que el origen del SARS-CoV-2 fue producto de un proceso de evolución natural [Andersen et al., 2020], [Zhou et al., 2020], [Wu et al., 2020], [Latinne et al., 2020], [Baruah et al., 2020]; aunque algunos trabajos sostienen la idea de que fue creado en un laboratorio [Latham y Wilson, 2020] [Segreto et al., 2021] [Yan et al., 2020] (así como las citas encontradas en dichos trabajos). Los que defienden un origen natural sostienen que el SARS-CoV-2 fue producto de un evento zoonótico que traspasó las barreras de las especies [Andersen et al., 2020] [Zhou et al., 2020] [Ge et al., 2013], proviniendo de los murciélagos (más específicamente de la secuencia identificada como RaTG13, detalles en [Rahalkar y Bahulikar, 2020a]). No obstante, es importante indicar que esta secuencia (RaTG13) data de 2013, y su genoma fue publicado en 2020, sin que los autores indicaran el lugar de procedencia de la misma.

Otros han asociado el origen del virus con las secuencias de los pangolines [Zhang et al., 2020] [Liu et al., 2020], aunque se ha sospechado que fueron el paso intermedio de contagio entre los murciélagos y los humanos [Isea, 2021]. Sin embargo, en [Latham y Wilson, 2020] alegan que las secuencias en los animales antes mencionados son un fraude, y que el SARS-CoV-2 fue producto de un laboratorio empleando para ello la secuencia de los virus ZC45/ZXC21 [Latham y Wilson, 2020] [Yan et al., 2020].

En paralelo, se ha indicado que el SARS-CoV-2 presenta una alta similitud con un dominio de una polimerasa BCoV/4991 aislado de unos mineros que estaban en una mina de la provincia de Yunnan (China) en 2012, donde lamentablemente fallecieron tres de ellos cuando limpiaban los excrementos de los murciélagos [Rahalkar y Bahulikar, 2020b].

Es importante destacar que la mayoría de los trabajos antes indicados en esta sección están basados en estudios de filogenia molecular (así como las referencias citadas en dichos trabajos), donde el consenso es que el origen fueron los murciélagos tras analizar los árboles filogenéticos, y por esa razón, este tipo de estudio se descarta en este trabajo.

A raíz de ello, se van a seleccionar varias secuencias correspondientes al SARS-CoV-2 detectadas en humanos, así como las secuencias provenientes de murciélagos, de SARS-CoV, MERS-CoV y pangolines para determinar si existen diferencias o similitudes entre ellas, empleando para ello varias herramientas básicas de Bioinformática como se indicarán a continuación.

Metodología

A partir de una secuencia de una glicoproteína de espícula del SARS-CoV-2 proveniente de los primeros episodios registrados en diciembre de 2019, se van a seleccionar diversas secuencias de los individuos humanos que fueron contagiados por dicho virus, las cuales deben estar depositadas en la base de datos de proteínas abreviada como PDB (acceso gratuito, disponible en `www.pdb.org`). Dicha selección fue posible tras utilizar el programa Blast [Johnson et al., 2008] disponible en `blast.ncbi.nlm.nih.gov`, restringiendo la búsqueda al Protein Data Bank (PDB). Asimismo, seleccionamos otras secuencias específicas correspondientes a murciélagos, pangolines, SARS-CoV y MERS-CoV.

Posteriormente se realizó un alineamiento múltiple de proteínas con el programa Cobalto, de acceso gratuito y disponible en `www.ncbi.nlm.nih.gov/tools/cobalt/cobalt.cgi?CMD=Web`, para poder realizar las comparaciones de similitud e identidad entre las secuencias [Papadopoulos y Agarwala, 2007].

El próximo paso fue determinar la frecuencia en los aminoácidos que componen las secuencias seleccionadas para detectar posibles diferencias entre especies, o apreciar cualquier anormalidad entre ellas, gracias al uso de varios algoritmos escritos en Python donde se empleó Biopython (una revisión completa está en el libro publicado por Rocha y Ferreira, 2018). En paralelo, se determinó el porcentaje de identidad y similitud entre dichas secuencias empleando para ello una matriz del tipo Blosum62 gracias a los algoritmos desarrollados en este lenguaje de programación [Rocha y Ferreira, 2018].

Finalmente, se realizó un cálculo de análisis multivariable producto de la covarianza obtenida tras las similitudes obtenidas entre pares de secuencias. Los detalles técnicos están publicados en el trabajo de [Tharwat, 2016]. Para visualizar dicho resultado, se realiza una representación gráfica 3D de las tres primeras componentes principales donde se agregaron manualmente unas líneas punteadas para ayudar a visualizar los resultados. La revisión de la metodología computacional ha sido validada y publicada en la literatura científica por [Konishi et al., 2019].

Resultados

La secuencia correspondiente a uno de los primeros episodios registrados en Wuhan (China) en diciembre de 2019 fue MN908947. A partir de dicha secuencia, y como se indicó en la sección anterior, se realizó un Blast de proteínas para seleccionar aquellas estructuras que presentan mayor similitud con ella, seleccionando diecinueve secuencias como están indicadas en la Tabla 1, es decir, doce secuencias que corresponden a contagios en humanos, cuyos identificadores son: 6XCM, 6ZP1, 6VSB, 7KDI, 7C2L, 6ZOW, 7K8S, 7KDJ, 7JJI, 7CWL, 6XS6 y 6ZB4, mientras que únicamente seleccionamos una secuencia correspondiente a un pangolín (identificador 7BBH), y otra correspondiente al SARS-CoV (5X58). Asimismo, fueron

obtenidas dos secuencias correspondientes a murciélagos (6ZGF y 7CN4), y tres secuencias que corresponden a MERS-CoV (5W9H, 5X59 y 6NB3).

A modo de ejemplo, se muestra en la figura 1 una pequeña sección del alineamiento obtenido con las secuencias utilizadas en el trabajo, obtenida con el programa Cobalto, donde se aprecian pequeñas diferencias entre ellas.

En la figura 2 se aprecia el porcentaje de identidad y dentro de un paréntesis el grado de similitud entre pares de secuencias, donde la secuencia 6XCM presenta una menor identidad y similaridad con respecto al resto de las secuencias en humanos, en vista del alto número de gaps que posee la misma tras el alineamiento obtenido con Cobalto.

Figura 1: Una corta región del resultado del alineamiento de algunas de las secuencias estudiadas en el trabajo.

Cuando se comparan las similitudes de las secuencias que infectan a los humanos con respecto a SARS-CoV y MERS-CoV, se obtienen un 74 % y 29 % de identidad, respectivamente; es decir, las secuencias de SARS-CoV-2 efectivamente están más relacionadas con el episodio de 2002 que con respecto al episodio acontecido en 2012 (MERS-CoV). Por otra parte, se evidencia la baja similitud existente cuando entre la secuencia del pangolín con respecto de los murciélagos.

Tabla 1: Secuencias seleccionadas en el presente estudio, donde se indica el identificador universal en la base de datos PDB, la longitud en aminoácidos de la glicoproteína de espícula, y finalmente de dónde proviene la misma. Todas las secuencias de contagios en humanos son por SARS-CoV-2.

PDB	Longitud	Especie	PDB	Longitud	Especie
7BBH	1248	Pangolín	6XCM	1259	Humano
6ZP1	1251	Humano	6VSB	1288	Humano
7KDI	1288	Humano	7C2L	1283	Humano
6ZOW	1273	Humano	7K8S	1259	Humano
7KDJ	1288	Humano	7JJI	1273	Humano
7CWL	1273	Humano	6XS6	1256	Humano
6ZB4	1259	Humano	6ZGF	1283	Murciélago
7CN4	1267	Murciélago	5X58	1228	SARS-CoV
5W9H	1329	MERS-CoV	5X59	1323	MERS-CoV
6NB3	1359	MERS-CoV			

	6XCM	6ZOW	7K8S	7KDJ	6XS6	6ZB4	6ZGF	7CN4	5X58	5W9H	5X59	6NB3
7BBH	74.31 (76.72)	89.27 (92.40)	89.27 (92.40)	88.93 (91.99)	89.32 (92.47)	(89.02 (92.16)	88.73 (92.05)	88.40 (91.88)	74.11 (82.63)	28.49 (44.54)	28.29 (44.40)	28.51 (44.65)
6XCM		79.69 (79.77)	79.69 (79.77)	79.72 (79.72)	79.80 (79.88)	79.52 (79.60)	78.29 (78.79)	78.21 (78.71)	65.68 (72.71)	24.11 (36.88)	24.18 (37.03)	24.07 (36.79)
6ZOW			100.0 (100.0)	99.42 (99.50)	99.59 (99.59)	99.75 (99.75)	97.19 (97.77)	97.36 (97.77)	74.63 (83.28)	29.04 (45.15)	29.15 (45.24)	29.22 (45.48)
7K8S				99.42 (99.50)	99.59 (99.59)	99.75 (99.75)	97.19 (97.77)	97.36 (97.77)	74.63 (83.28)	29.04 (45.15)	29.15 (45.24)	29.22 (45.48)
7KDJ					99.17 (99.26)	99.50 (99.59)	96.86 (97.36)	96.78 (97.27)	74.55 (83.11)	28.96 (45.10)	29.00 (45.17)	28.92 (45.18)
6XS6						99.50 (99.50)	97.27 (97.85)	97.43 (97.85)	74.65 (83.32)	29.04 (45.15)	29.08 (45.09)	29.22 (45.48)
6ZB4							96.94 (97.52)	97.11 (97.52)	74.71 (83.36)	28.89 (45.07)	29.15 (45.32)	29.07 (45.41)
6ZGF								99.67 (99.83)	75.04 (83.80)	28.97 (44.70)	28.85 (44.64)	28.99 (44.95)
7CN4									75.37 (83.97)	28.97 (44.70)	29.00 (44.71)	29.14 (45.03)
5X58										27.83 (42.68)	28.29 (45.35)	27.99 (42.81)
5W9H											98.45 (98.45)	98.68 (98.84)
5X59												

Figura 2: Porcentaje de identidad, y dentro de un paréntesis el grado de similaridad entre pares de secuencias (más detalles en el texto).

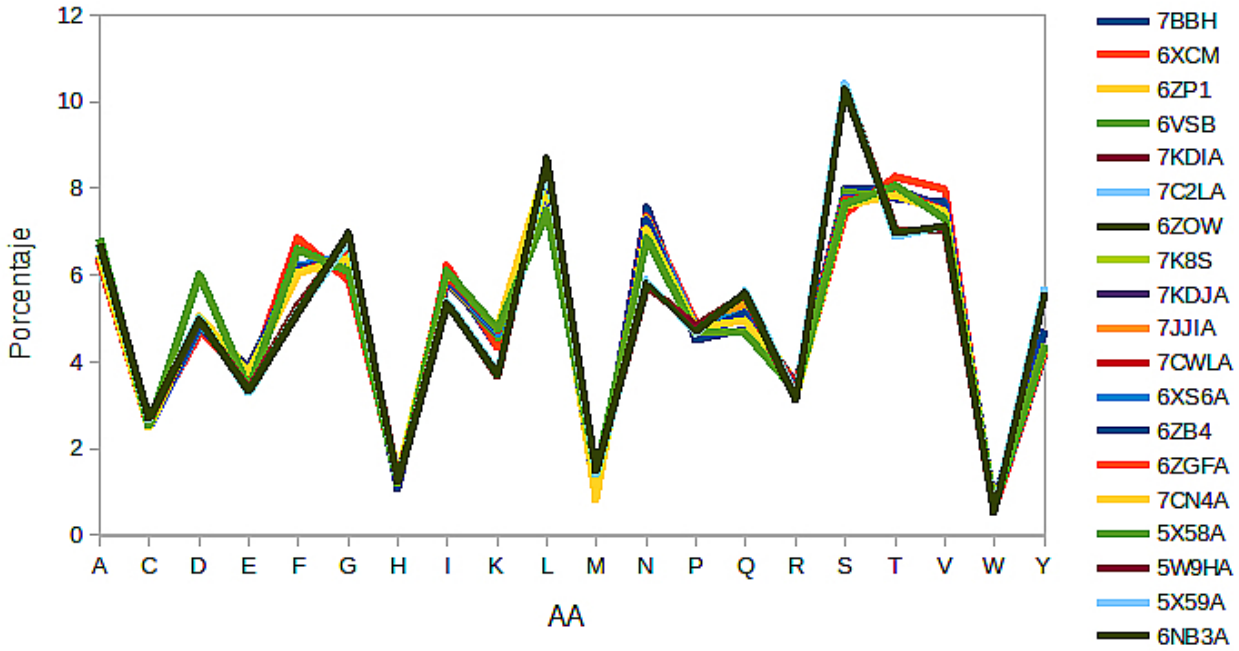

Figura 3: Porcentaje de composición de los aminoácidos (AA) que conforman la glicoproteína de espícula en varias secuencias estudiadas en el trabajo.

En la figura 3 se observa el porcentaje de composición de las secuencias estudiadas en el trabajo. Lamentablemente el trabajo no realiza un estudio detallado de cada una de las secciones (dominios) de las proteínas, donde se detallarían por ejemplo, cómo la secuencia "SPRRA" presente en humanos permite la entrada del virus a la célula [Coutard et al., 2020].

Posteriormente, se empleó un estudio de componentes principales basado en los autovalores obtenidos producto de la covarianza derivada entre pares de secuencias, y de esa manera poder conocer cuán diferentes son las secuencias entre sí. Los resultados obtenidos se pueden visualizar en la figura 4, donde se agregaron manualmente unas líneas punteadas para ayudar a comprender dicho resultado.

Dicha figura representa las tres primeras componentes principales representadas en cada uno de los ejes del gráfico (representado con flechas de color verde). Para simplificar los resultados, se agruparon varias secuencias que están adyacentes entre sí con las etiquetas C1, C2 y C3, donde C1 corresponden a las secuencias del MERS-CoV, mientras que C2 son todas las secuencias en humanos con excepción de 6XCM. C3 son las dos secuencias correspondientes a los murciélagos.

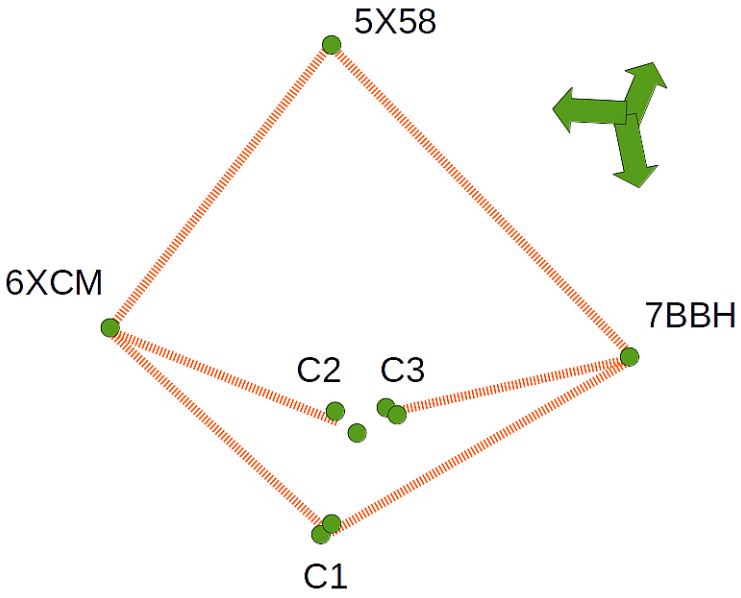

Figura 4: Representación gráfica de las tres primeras componentes principales que representan los ejes de coordenadas de la figura (indicados con unas flechas en color verde en la parte superior derecha). Se han agrupado varias secuencias donde C1 corresponden a MERS-CoV, mientras que C2 a las secuencias infectadas en humanos con SARS-CoV-2, con excepción de 6XCM. Finalmente, en C3 se agrupan las dos secuencias en murciélagos.

La figura 4 destaca que las secuencias del SARS-CoV-2 efectivamente son diferentes a los dos eventos anteriores (SARS-CoV y MERS-CoV). Más aún, en vista que la secuencia 6XCM posee una gran cantidad de gaps debido al alineamiento de las secuencias, no implica que se haya modificado genéticamente la misma, e ilustra lo sensible del método.

Asimismo, la figura nos muestra cuán adyacentes están las secuencias C2 y C3, es decir, no hay grandes diferencias entre las secuencias obtenidas en murciélagos y los casos detectados en humanos, por lo que podríamos afirmar que existe un proceso de evolución entre ellas a raíz de los pequeños cambios observados en dicha figura.

Conclusiones

El trabajo realiza un estudio estadístico y bioinformático del SARS-CoV-2 en diecinueve secuencias de la glicoproteína de espícula S que han sido recopiladas en la base de datos de proteínas – PDB. Al comparar la composición de los aminoácidos entre varias especies, revela diferencias entre ellas acorde a lo señalado en estudios de filogenia molecular.

Finalmente, el análisis de las componentes principales nos señala que efectivamente hay una posible relación evolutiva entre las secuencias de SARS-CoV-2 detectadas en humanos con los

murciélagos, y no con los pangolines, y en vista de todo lo dicho, es posible concluir que la evolución del SARS-CoV-2 es un proceso natural, y no fue creado en el laboratorio.

Agradecimientos

El autor agradece al Comité Editorial así como a los revisores del trabajo por los comentarios realizados en el mismo. Por último, y no por ello menos importante, por las observaciones y la transcripción en LaTeX a Jesús Isea.

Dedicatoria

Este trabajo está dedicado a todas las personas que ayudan incondicionalmente a todos las pacientes que padecen de Covid-19, arriesgando su propia vida por el bien de los demás.

Bibliografía

[Andersen et al., 2020] Andersen, K., Rambaut, A., Lipkin, W., Holmes, E. y Garry, R. (2020). The proximal origin of SARS-CoV-2. *Nature*. Vol. 26(4), 450-452.

[Baruah et al., 2020] Baruah, D., Devi, P. y Sharma, D. (2020). Sequence analysis and structure predition of SARS-CoV-2 accesory proteins 9b and ORF14: evolutionary analysis indicates close relatedness to bat coronavirus. *BioMed Research International*. 2020: 7234961.

[Coutard et al., 2020] Coutard, B., Valle, C., De Lamballerie, X., Canard, B., Seidah, N. y Decroly, E. (2020). The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Research*. 176, 104742.

[Ge et al., 2013] Ge, X., Li, J., Yang, J., Chmura, A., Zhu, G., Epstein, J., Mazet, J. (2013). Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor. *Nature*. Vol. 503, 535-538.

[Isea, 2021] Isea, R. (2021). Analytical solutions of the transmissibility of the SARS-CoV-2 in three interactive populations. *International Journal of Coronavirus*. Vol 2(4), 1-7.

[Johnson et al., 2008] Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S. y Madden, T. (2008). NCBI BLAST: a better web interface. *Nucleic Acids Res.* 36, W5-W9.

[Konishi et al., 2019] Konishi, T., Matskukama, S., Fuji, H., Nakamura, D., Satou, N. y Okano, K. (2019). Principal component analysis applied directly to sequence matrix. *Sci Rep.* 9(1), 19297.

[Latham y Wilson, 2020] Latham, J. y Wilson, A. (15 de julio de 2020). A Proposed Origin for the SARS-CoV-2 and the COVID-19 Pandemic. *Printfriendly*. Recuperado de https://www.printfriendly.com/p/g/9jkSGu.

[Latinne et al., 2020] Latinne, A., Hu, B., Olival, K., Zhu, G., Zhang, L. (2020). Origin and cross-species transmission of bat Coronavirus in China. *BioRxiv*. Recuperado de https://doi.org/10.1101/2020.05.31.116061.

[Liu et al., 2020] Liu, P., Jiang, J., Wan, Z., Huan, Y., Li, L., Zhou, J., Wang, Z. (2020). Are the pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? *PloS Pathogens*. 16(5): e1008421.

[Matheson y Lehner, 2020] Matheson, N. y Lehner, P. (2020). How does SARS-CoV-2 cause Covid-19? *Science*. 369(6503), 510-511.

[Papadopoulos y Agarwala, 2007] Papadopoulos, J. y Agarwala, R. (2007). COBALT: constrained-based alignment tool for multiple protein sequences. *Bioinformatics*. 23, 1073-1079.

[Rahalkar y Bahulikar, 2020a] Rahalkar, M. y Bahulikar, R. (2020a). Understanding the origin of BatCoVRaTG13, a virus closest to SARS-CoV-2. *Preprints*. 2020050322. DOI: 10.20944/preprints202005.0322.v1

[Rahalkar y Bahulikar, 2020b] Rahalkar, M. y Bahulikar, R. (2020b). Lethal Pneumonia Cases in Mojiang Miners (2012) and the Mineshaft Could Provide Important Clues to the Origin of SARS-CoV-2. *Front. Public Health*. 8, 581569.

[Rocha y Ferreira, 2018] Rocha, M. y Ferreira, P. (2018). *Bioinformatics Algorithms: Design and Implementation in Python* (1a ed). Academic Press

[Segreto et al., 2021] Segreto, R., Deigin, Y., McCairn, K., Sousa, A., Sirotkin, D., Sirotkin, K., Couey, J., Jones, A. y Zhang, D. (2021). Should we discount the laboratory origin of COVID-19? *Environmental chemistry letters*, 1–15.

[Tharwat, 2016] Tharwat, A. (2016). Principal component analysis – a tutorial. *Int J Appl Patt Rec*. 3(3), 197–240.

[Wu et al., 2020] Wu, F., Zhao, S., Yu, B., Chen, Y., Wang, W., Song, Z., Hu, Y. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*. Vol. 579(7798), 265-269.

[Xiao et al., 2020] Xiao, K., Feng, J., Zhou, N., Zhang, Z., Zou, X., Li, J. (2020). Isolation and characterization of 2019-nCoV-like coronavirus from Malasyan pangolins. *BioRxiv*. Recuperado de https://doi.org/10.1101/2020.02.17.951335.

[Yan et al., 2020] Yan, L., Kang, S., Guan, J. y Hu, S. (2020). Unusual features of the SARS-CoV-2 genome suggesting sophisticated laboratory modification rather than natural evolution and delineation of its probable synthetic route. *Zenodo.org, Preprints*. DOI: 10.5281/zenodo.4028830.

[Zhang et al., 2020] Zhang, T., Wu, Q. y Zhang, Z. (2020). Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. *Curr Biol*. 30(8), 1578.

[Zhou et al., 2020] Zhou, P., Yang, X., Wang, X., Hu, B., Zhang, L., Zhang, W. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2579, 270–273.