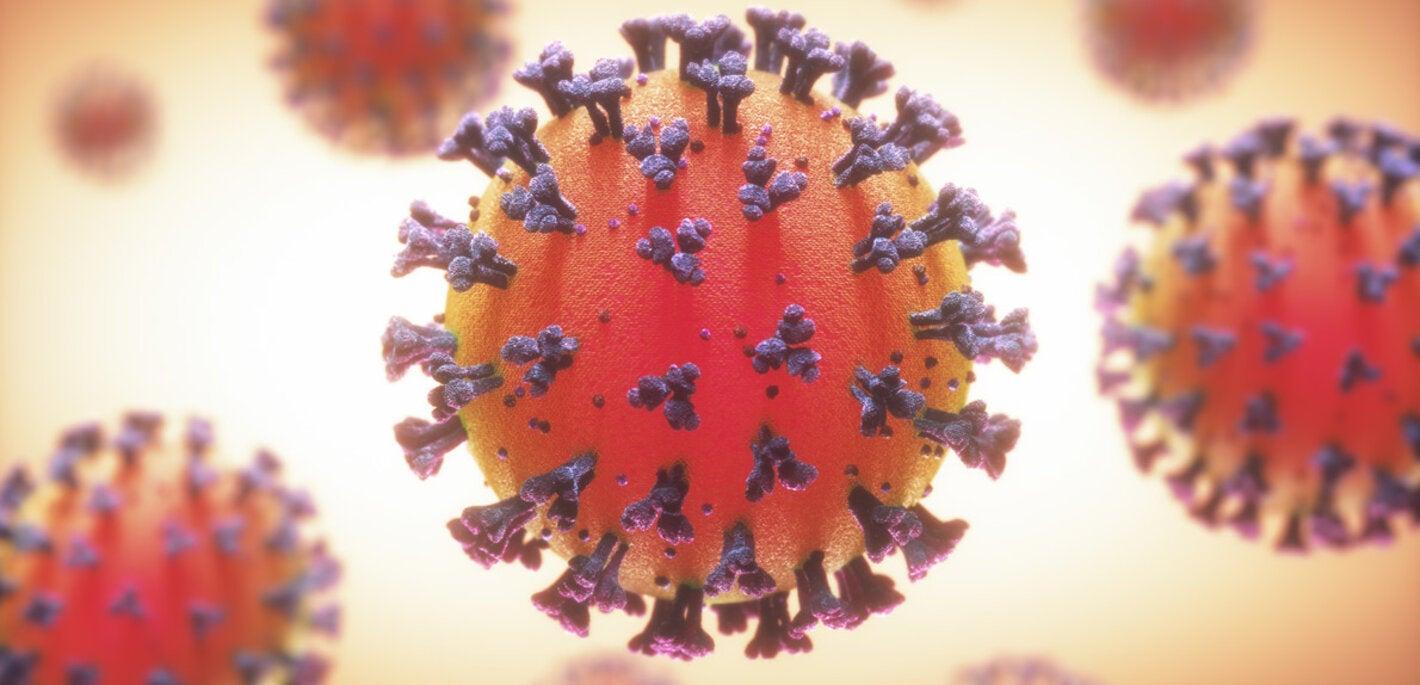


Genética

Año: 2021

Trabajo Integrador Final

“Herramienta para el estudio de la evolución del virus del SARS-COV-2”



Salim Taleb, Nasim A; Tolley, Rocio M

Docentes: Schierloh, Pablo; Machtey, Matias.

Carrera: Lic. en Bioinformática

# Introducción al problema

El alto grado de infecciosidad del SARS-COV-2 ha llevado a su amplia distribución y elevado número de casos a nivel mundial de la enfermedad, lo que implica una proliferación desmedida del virus [[1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7307108/)]. En las sucesivas replicaciones, pueden darse mutaciones que alteren aspectos del mismo, llevando a cambios que, mientras que en su mayoría resultan perjudiciales para la viabilidad del virón, pueden conducir a variantes más infecciosas o letales [[1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7307108/),[2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7418704/)]. Si suficientes mutaciones logran acumularse en ciertas secuencias codificantes, en particular de la glicoproteína de superficie S, conocida como la proteína ‘spike’, es posible que surjan variantes que evadan los efectos de las vacunas desarrolladas hasta el momento [[2](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7418704/),[3](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0250780)].

Por lo tanto, resulta de interés monitorear las mutaciones encontradas en las secuencias del virus, y analizar si existe cierta tendencia a presentar mutaciones en regiones específicas, o si puede demostrarse un impacto en las presiones selectivas del virus al encontrarse gran parte de la población vacunada.

# Objetivo

Desarrollar una herramienta capaz de analizar distintas secuencias de ADN o proteínas y cuantificar las variaciones que existen con respecto a una secuencia inicial, también identificar las regiones donde se producen la mayor cantidad de mutaciones en la secuencia(hotspots), mostrando los resultados gráficamente.

# Desarrollo

La herramienta fue programada en el lenguaje C + + en el entorno de desarrollo de Qt y toda su interfaz fue desarrollada mediante las herramientas que este provee. Su funcionamiento consiste en la lectura de secuencias en un archivo obtenido mediante la base de datos NCBI [[4](https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=SARS-CoV-2,%20taxid:2697049)], en el cual éstas ya se encuentran alineadas. Después de esta lectura se realiza un procesamiento en función de lo que desee el usuario:

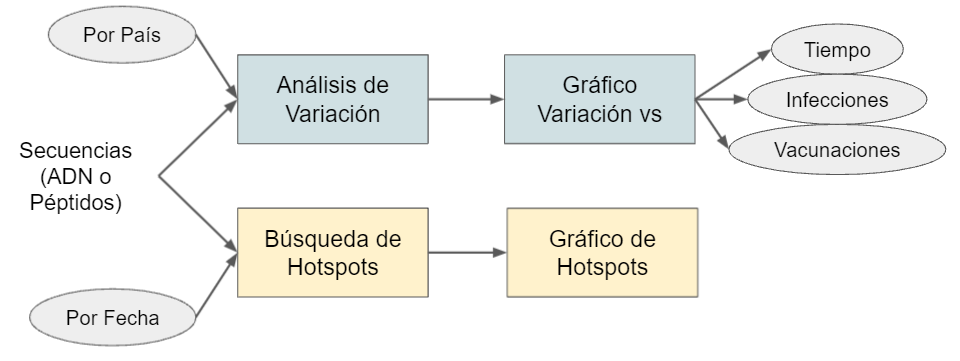


Figura 1: Cuadro conceptual del funcionamiento de la herramienta.

El ingreso puede ser mediante una secuencia de ADN o proteínas, y en el caso del análisis de variación puede realizarse en un determinado país, o en la búsqueda de hotspots, en un determinado rango de fecha.

Los resultados del análisis de variación se muestran en forma de gráfica con ejes, permitiendo elegir qué mostrar en el eje de las abscisas para permitir comparación con múltiples parámetros,

Los resultados obtenidos de la búsqueda de hotspots se muestran mediante una gráfica unidimensional donde se puede apreciar la variabilidad en las distintas regiones.

.

A continuación se muestra en detalle una imagen tomada de la interfaz inicial de la herramienta, indicando con números las distintas funcionalidades:

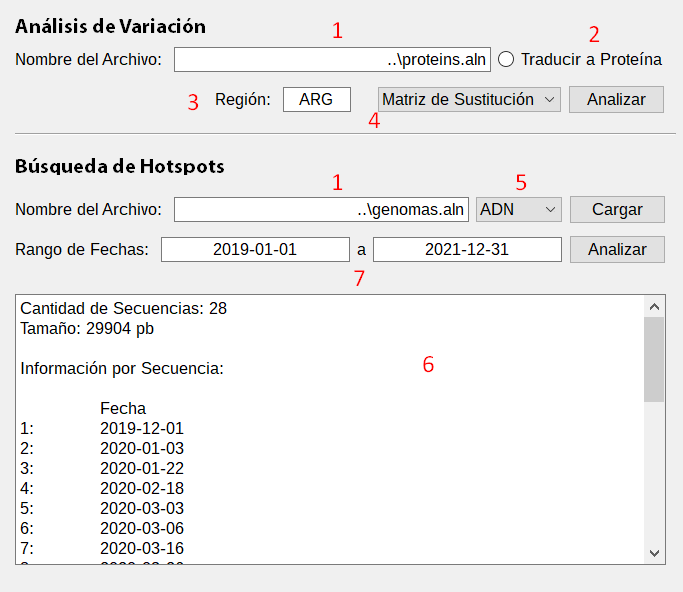


Figura 2: Interfaz principal de la herramienta.

**Funcionalidades:**

1. Líneas de texto en las cuales se indica el nombre y dirección del archivo a partir del cual se cargarán las secuencias para cada tipo de procesamiento.
2. Se da como opción la posibilidad de ingresar secuencias de ADN para el análisis de variación, y que sus regiones codificantes sean traducidas a proteínas.
3. Es posible elegir la región de la cual tomar los valores de cantidad de infecciones y vacunaciones, provenientes de los datos de la base de datos Our World in Data.
4. Se da la opción de seleccionar una matriz de sustitución previamente cargada a la herramienta. Por defecto se emplea un cálculo simplificado que cuantifica la cantidad de puntos en los cuales la secuencia varía de la de referencia.
5. Es posible seleccionar que la búsqueda de hotspots se realice en secuencias de ADN o de péptidos.
6. Presionar el botón de ‘Cargar’ muestra datos generales de las secuencias incluidas en el archivo, incluyendo la cantidad de secuencias en el alineamiento, su longitud, y la fecha correspondiente a cada una.
7. Es posible seleccionar un rango de fechas a comparar en la búsqueda de hotspots.

# Resultados obtenidos y esperados

### Análisis de la Variación

Se programó un módulo en el cual se calcula un grado de variación de las secuencias con respecto a la más antigua (de fecha menos reciente) dentro del conjunto de secuencias en el alineamiento empleado. La variación puede calcularse aplicando distintas matrices que se seleccionan en la interfaz principal de la herramienta, para así aplicar un valor distinto dependiendo de la naturaleza de la variación. El gráfico resultante se construye a partir del valor calculado de variación sobre el eje vertical, y el eje horizontal puede ser modificado para reportar el índice (en base al orden de las secuencias), la fecha asociada a cada secuencia, la cantidad de infecciones correspondientes a la fecha de la secuencia, y la cantidad de vacunaciones correspondientes a la fecha de la secuencia.

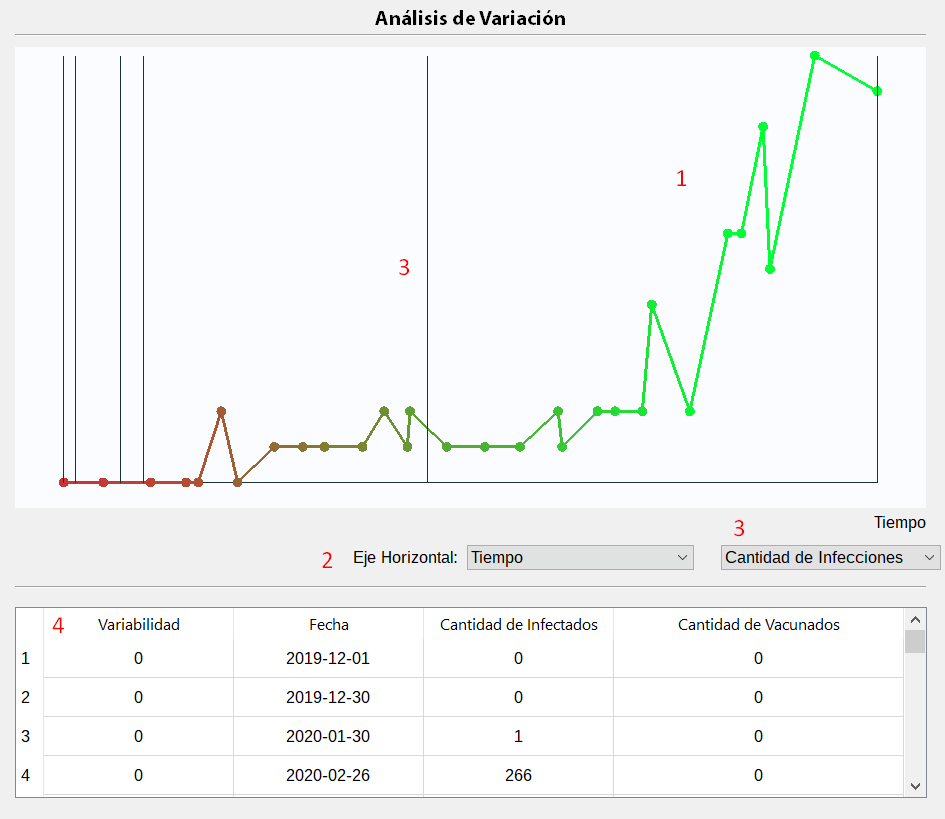


Figura 3: Interfaz correspondiente a los resultados del análisis de variación.

**Funcionalidades:**

1. Gráfico de variación en función del eje seleccionado. Cada punto se corresponde con una de las secuencias incluídas en el alineamiento analizado.
2. Es posible seleccionar uno de cuatro ejes: por índice, por tiempo (en días), cantidad de infecciones en el país seleccionado en la interfaz principal, y cantidad de vacunaciones en el país seleccionado en la interfaz principal.
3. Para el eje de tiempo, se posibilita seleccionar una marca vertical que indique las fechas en las cuales se dio un aumento en un factor 100 de la cantidad de infecciones o de vacunaciones.
4. Se muestran datos asociados por cada secuencia: el valor de variabilidad con respecto a la primer secuencia, la fecha, las infecciones totales para la fecha, y las vacunaciones totales para la fecha.

### Búsqueda de *Hotspots*

Se desarrolló un módulo que compara las secuencias en el alineamiento y grafica su longitud, marcando con colores diferentes los sitios en los cuales se detecta una mayor variabilidad entre las distintas secuencias.

A continuación se muestra una captura de la salida de la herramienta, indicando con números en rojo las distintas funcionalidades.

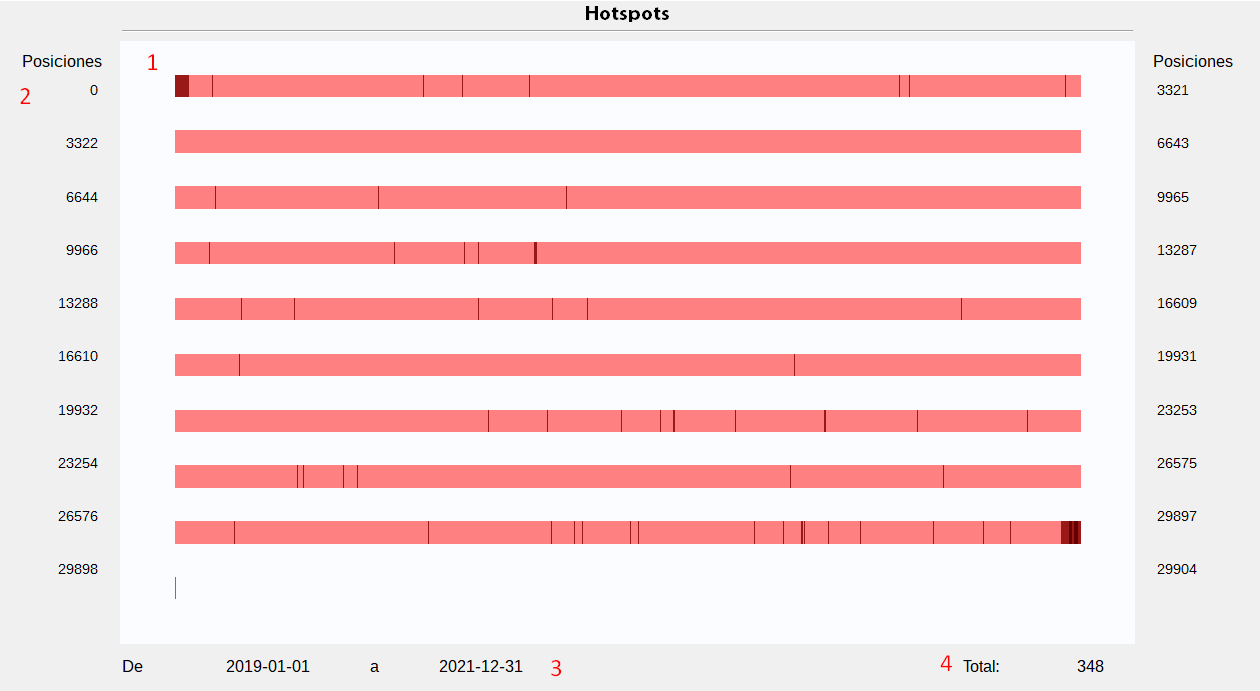


Figura 4: Interfaz correspondiente a los resultados de la búsqueda de hotspots.

**Funcionalidades:**

1. La totalidad de la secuencia se divide en renglones, y se muestran los puntos en los cuales se detectaron cambios de una secuencia a otra con un color más profundo; mientras más profundo, más secuencias en las cuales la posición varió.
2. A cada lado del gráfico se indican los rangos de posiciones ilustrados en cada barra de color.
3. Rango de fechas seleccionado en la interfaz principal.
4. Se muestra el total de posiciones en las cuales se detectó una variación entre las secuencias analizadas.

Cabe recalcar, que puede ocurrir que se detecte una gran variación en los extremos de la secuencia, esto es debido a la alineación de las secuencias donde por lo general se generan secciones al inicio y al final que no contienen nucleótidos.

El programa permite seleccionar un rango de fechas de las secuencias a comparar, para permitir seleccionar un subconjunto de las secuencias incluidas en el alineamiento.

# Ventajas y desventajas

* Fácil uso.
* Rápido funcionamiento.
* Adaptable o mejorable a otros análisis.
* Limitado a archivos de secuencias que sigan el formato designado.
* Depende fuertemente del dataset suministrado para obtener información concisa.

# Aplicación en la práctica

Para el caso del análisis del virus del SARS-COV-2, se puede observar en la siguiente gráfica que la variación entre las secuencias del virus está correlacionado con el aumento en la cantidad de vacunaciones.

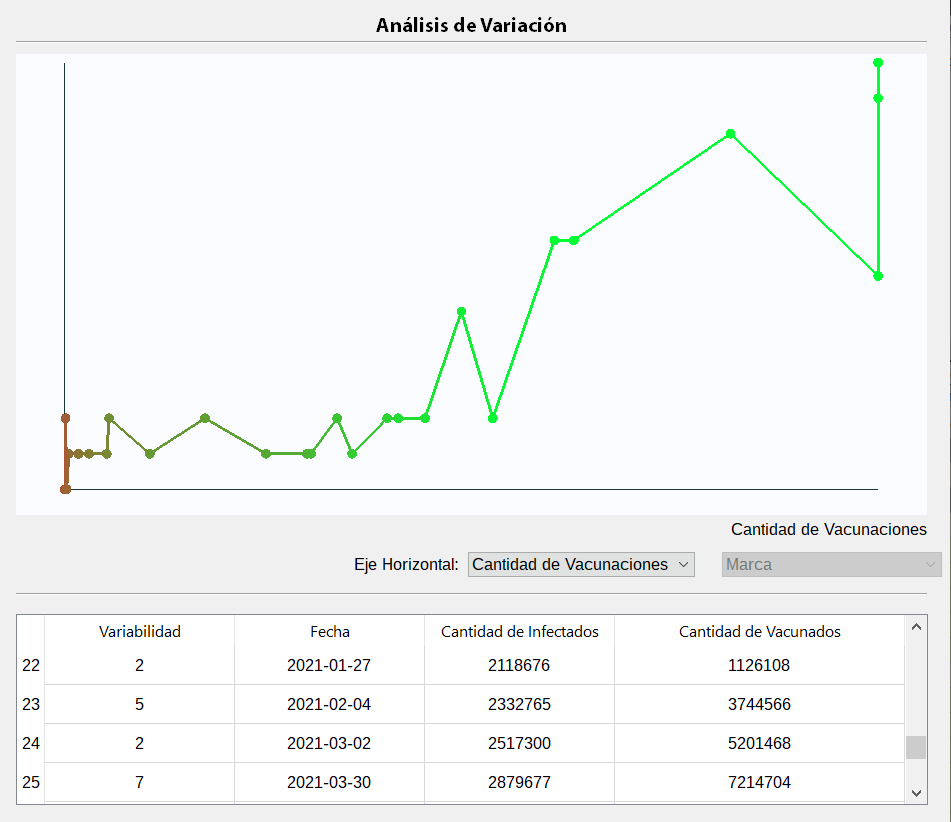


Figura 5: Resultados de la variación para un conjunto de 28 secuencias, representando como eje de abscisas la cantidad de vacunaciones para la fecha asociada a cada secuencia.

# Perspectiva a futuro

Ampliar las funcionalidades disponibles para que el análisis pueda hacerse a nivel proteico teniendo como base la secuencia de ADN, de tal manera que puedan detectarse cambios de nucleótidos que no generen un cambio en el aminoácido producido, también que puedan detectarse las distintas zonas codificantes y no codificantes dentro de la secuencia de ADN provista. Además, cabe recalcar que el uso de la herramienta es extrapolable a cualquier otro organismo, que respeten los formatos de archivos necesarios.

# Referencias

Cagliani, R., Forni, D., Clerici, M., & Sironi, M. (2020). Computational Inference of Selection Underlying the Evolution of the Novel Coronavirus, Severe Acute Respiratory Syndrome Coronavirus 2. Journal of virology, 94(12), e00411-20. <https://doi.org/10.1128/JVI.00411-20>

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K., Dingens, A. S., Navarro, M. J., Bowen, J. E., Tortorici, M. A., Walls, A. C., King, N. P., Veesler, D., & Bloom, J. D. (2020). Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. Cell, 182(5), 1295–1310.e20. <https://doi.org/10.1016/j.cell.2020.08.012>

Van Egeren D, Novokhodko A, Stoddard M, Tran U, Zetter B, et al. (2021) Risk of rapid evolutionary escape from biomedical interventions targeting SARS-CoV-2 spike protein. PLOS ONE 16(4): e0250780. <https://doi.org/10.1371/journal.pone.0250780>

NCBI. (24 de Octubre de 2021). *NCBI*. Obtenido de NCBI: <https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=SARS-CoV-2,%20taxid:2697049>

Global Change Data Lab. (24 de Octubre de 2021). *Our World in Data*. Obtenido de Our World in Data: <https://ourworldindata.org/covid-vaccinations>

The Qt company. (24 de Octubre de 2021). *Qt Documentation*. Obtenido de Qt Documentation: <https://doc.qt.io/qtcreator/>