INTRODUCCIÓN A LA BIOINFORMÁTICA

POSTGRADO Y FORMACIÓN PROFESORADO

TRABAJO FINAL

UNED

PRÁCTICA O LECCIÓN GUIADA

UTILIDADES DE LA BIOINFORMÁTICA

EL COVID 19, UN ANÁLISIS DEL GENOMA Y HERRAMIENTAS DE BIOINFORMÁTICA

Santiago Royuela Samit

Mayo de 2022

INTRODUCCIÓN

En la siguiente práctica o lección guiada vamos a introducir al alumno en los conceptos básicos de la biología molecular y la computación, centrándonos en el código genético universal codificado en los ADN's y ARN's, y atendiendo a los distintos recursos informáticos a los que podemos acceder en la actualidad, dando a conocer el gran potencial de esta nueva ciencia que emerge desde la biología, la matemática, la físico-química y la computación. De esta manera, se le abre al alumno una puerta para darle a conocer los diferentes recursos que existen en la red para el análisis de datos biológicos, genéticos, proteómicos, etc, todo ello mediante una práctica guiada en la que verá el potencial de diferentes recursos y en donde se enfatizarán aspectos teóricos.

Para centrar al alumno, se le guiará a través del genoma del conocido virus **COVID-19**, despertando el interés de la bioinformática aprovechando la ingente información e impacto de esta pandemia acontecida, así como las nuevas técnicas de inmunización basadas en **vacunas de ARNm**. De esta manera, se trata de una práctica guiada o magistral del profesor al alumno profano con algo de conocimiento, mediante la cual, a la vez que se le muestran herramientas informáticas accesibles en red, se le explican conceptos propios de la biología molecular y de la bioinformática. Durante la práctica o lección guiada, el profesor podrá o deberá ir explicando conceptos como los de transcripción, traducción, código genético universal, gen, ADN, ARN, ARNm, splicing alternativo, ORF, CDS, intrones y exones, regiones UTR, codones, pautas de lectura de codones en una secuencia, así como los conceptos de biología o físico-químicos pertinentes en cada momento etc.

ACCEDIENTO A LA INFORMACIÓN DEL VIRUS DEL COVID-19

Cuando escuchamos hablar de la **pandemia del virus del Covid-19** y no somos expertos en la materia, podemos proceder a realizar un estudio con base científica para conocer muchos aspectos a tener en cuenta ante esta alarmante situación. Si lo que pretendemos es conocer dicho virus, sus variantes y vacunas posibles, recomendaremos acudir a la página web de **National Center for Biotechnology Information** (NCBI: https://www.ncbi.nlm.nih.gov/) para comenzar a buscar información acerca de este virus.

Accediendo a la página realizaremos una búsqueda seleccionando "**All DataBase**" introduciendo la palabra "**COVID 19**" y nos redirigirá a la página con los siguientes resultados de búsqueda que mostramos en la figura 1.

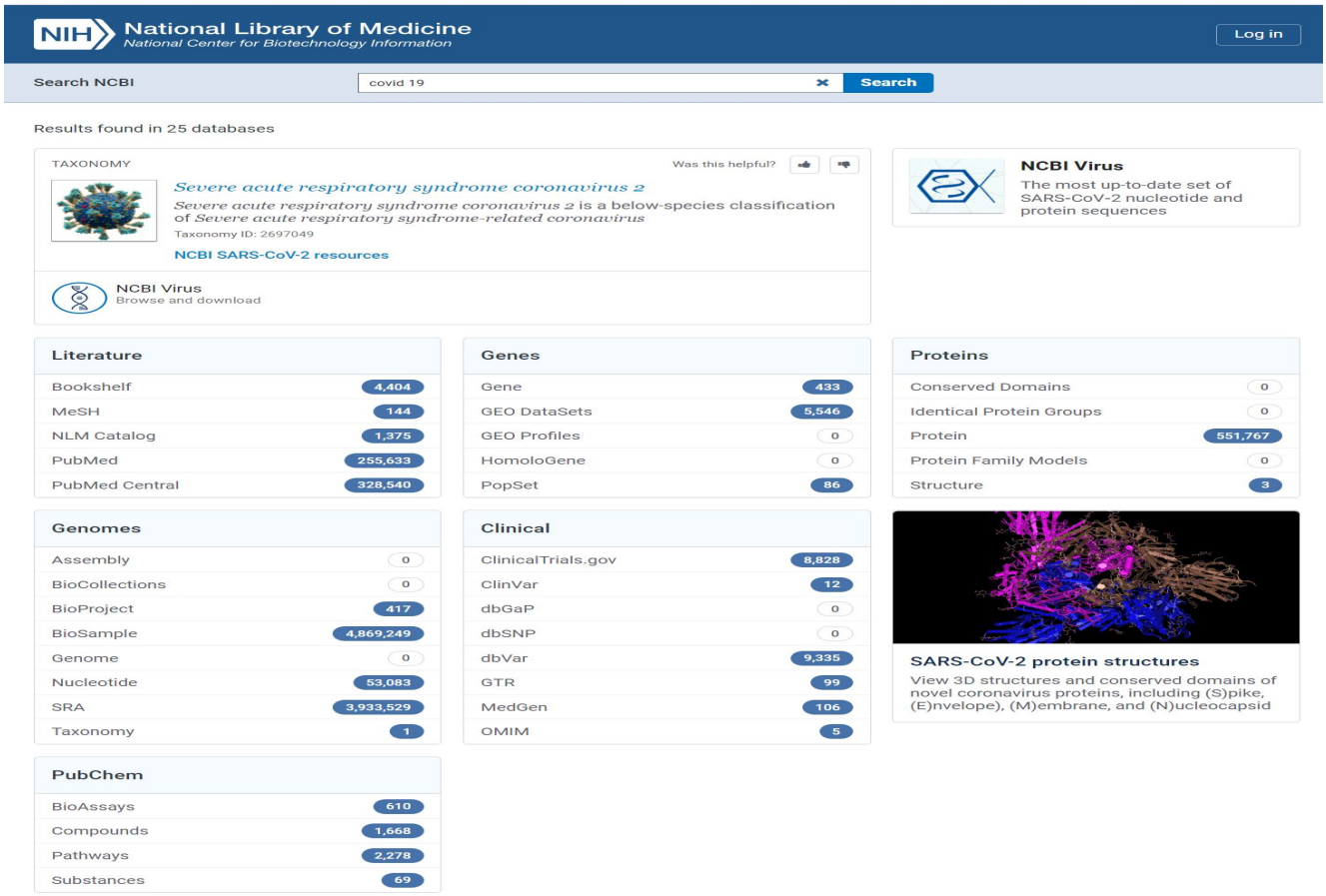
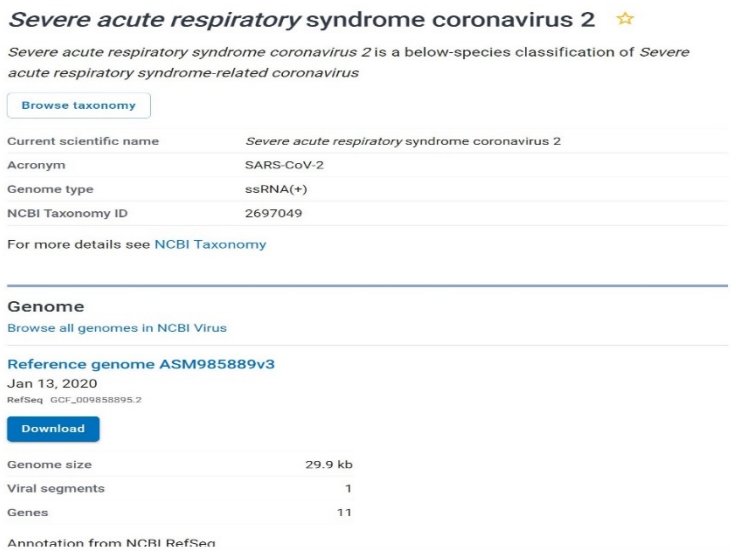
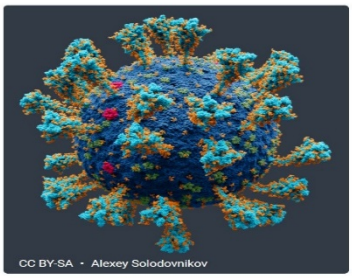


Ilustración 1. Resultado de la búsqueda en el NCBI por la palabra "Covid 19".

Vemos que si clicamos en **TAXONOMY** nos llevará a una página donde aparece un resumen de la taxonomía del virus, su genoma e identificaciones pertinentes que pasaremos a analizar.





External links
Encyclopedia of Life
Wikipedia

Ilustración 2. Identificación del virus SARS-CoV-2 y acceso a su genoma en el NCBI. Podemos ver el tamaño de la secuencia del genoma, los segmentos virales y los 11 genes que posee el virus.

Genome

Browse all genomes in NCBI Virus

Reference genome ASM985889v3

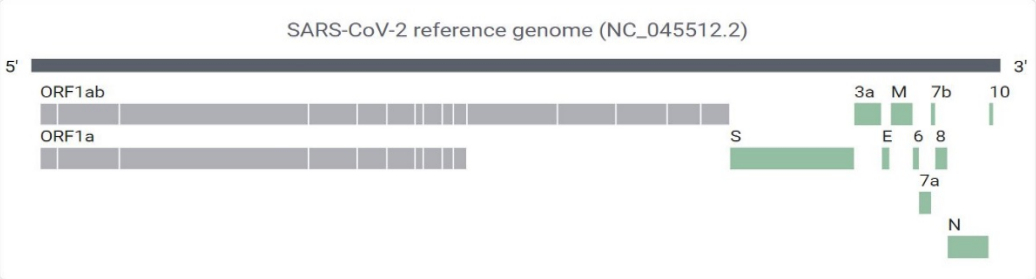
Jan 13, 2020

RefSeq GCF_009858895.2

Download

Genome size	29.9 kb
Viral segments	1
Genes	11

Annotation from NCBI RefSeq



External links

Encyclopedia of Life

Wikipedia

Ilustración 3. Visión esquemática del genoma del Virus SARS-CoV-2 y sus genes. Es importante el número de referencia de la secuencia o la referencia del genoma de GenBank, que en este caso es NC_045512.2

Assembly statistics

These statistics describe the nuclear genome of the reference sequence, GCF_009858895.2

Genome size	29.9 kb
Number of chromosomes	1
Number of scaffolds	1
GC percent	37.5
Assembly level	Complete Genome

Annotation details

Provider	
Name	NCBI RefSeq
Date	Jan 13, 2020
Genes	11
Protein-coding	11
Non-coding	0

View all genes (includes updated and unannotated genes)

Chromosomes

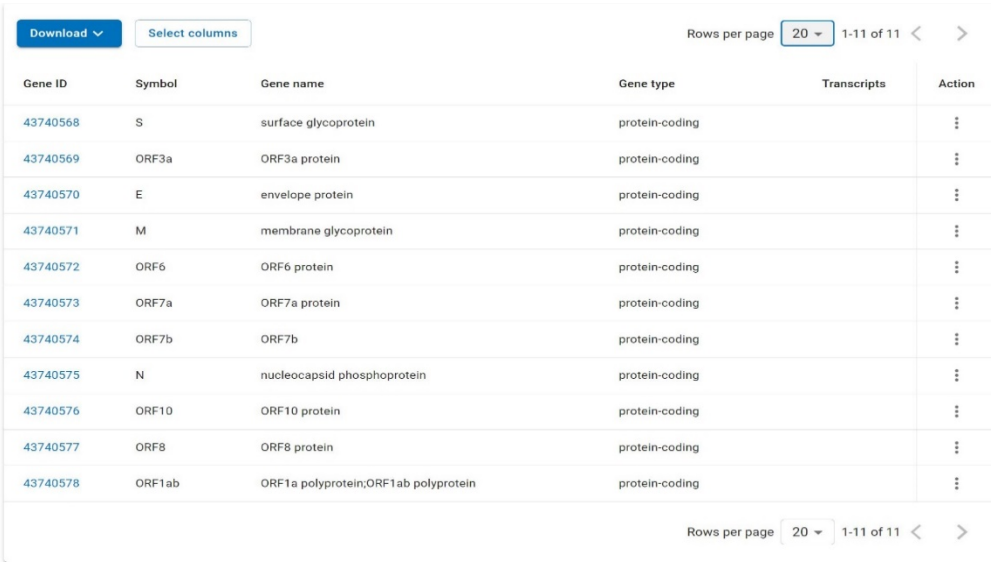
Chromosome	GenBank	RefSeq	Size (bp)	GC content (%)
ANONYMOUS	MN908947.3	NC_045512.2	29.903	37,5

Si clicamos en "**Reference genome**" nos llevará a una página en donde aparece una descripción más detallada del genoma del virus y podremos acceder a un enlace "**View all genes**".

Ilustración 4. Descripción estadística del genoma del SARS-CoV-2

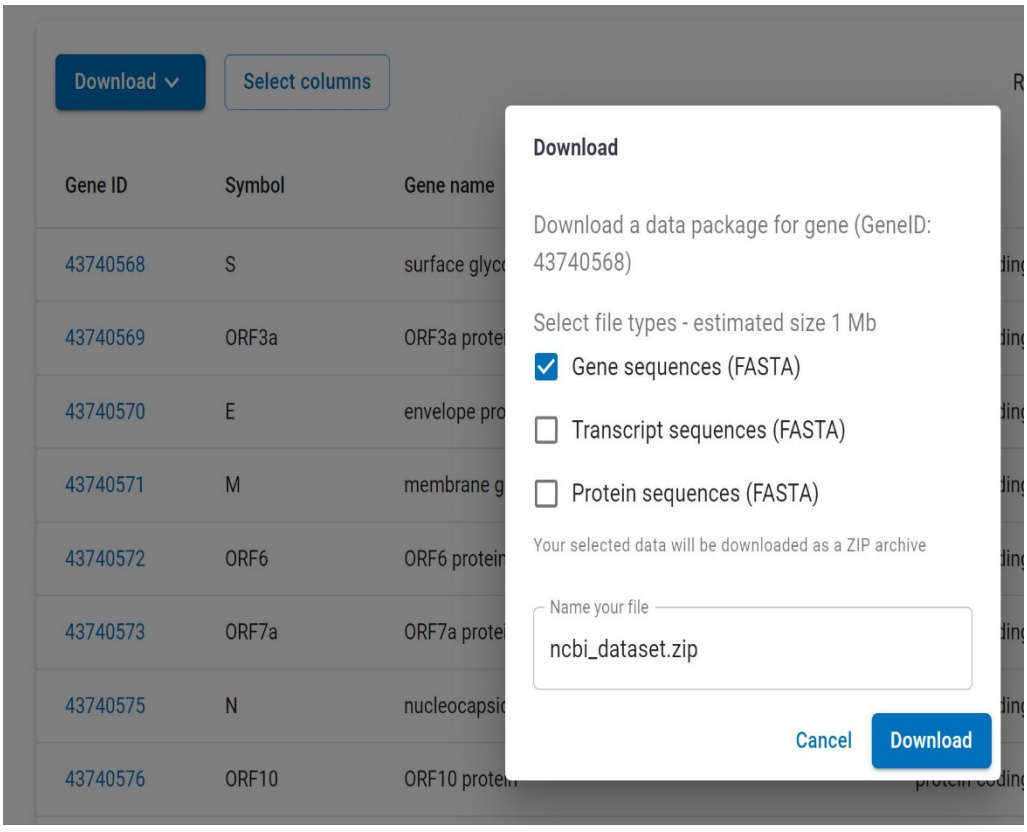
Vemos el tamaño del genoma y su porcentaje de nucleótidos GC. También podemos ver que hay 11 genes, indicando su identificación, símbolo y nombre. Vemos que los 11 genes de SARS-CoV-2 codifican para proteínas:

Ilustración 5. Los 11 genes del virus de SARS-CoV-2 y las proteínas víricas para las que codifican.



Si clicamos en cada uno de los **11 genes** podemos acceder a un navegador genómico para ver dichas regiones en el genoma, así como sus transcritos y productos. A estas alturas, ya hemos identificado que el **COVID-19** (SARS-CoV-2) es un **virus de ARN** con un **genoma** de **tamaño 29.9 kb**. Observamos que de 11 genes salen 28 productos proteícos.

Ilustración 6. Podemos acceder al fichero FASTA del gen que codifica a la proteína S, así como su transcrito y proteína para analizarlo con el programa SnapGen Viewer que veremos más adelante.



Si descargamos los ficheros **FASTA** del gen, transcrito y proteínas podemos ejecutarlos en **SnapGene Viewer** y ver la secuencia de **aa** del transcrito.

Las herramientas que vamos a utilizar son la base de datos del **NCBI**, el buscador **ORF FINDER** y el programa de descarga gratuíta **SnapGene Viewer**.

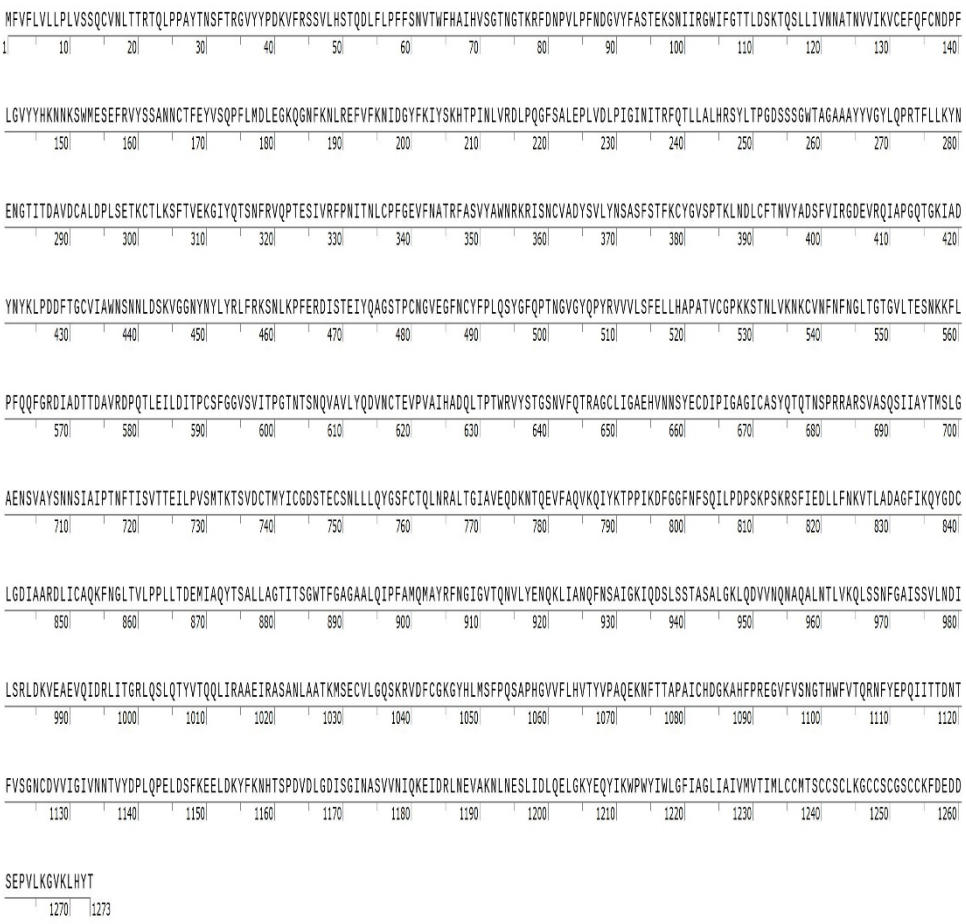


Ilustración 7. Secuencia de aa en código de 1 letra para la proteína S en SnapGene Viewer. El programa numera la secuencia desde el 1 en adelante, no teniendo en cuenta en este caso la secuencia total del genoma, pues hemos descargado solo la traducción de la proteína S desde el NCBI en la página comentada.

Podemos acceder a las propiedades de la secuencia de aa de la proteína desde SnapGene Viewer donde nos indicará el número de tipos de aa que aparecen, su porcentaje, así como datos físico-químicos de la proteína S

	Whole Protein
Length	1273 aa
▶ Molecular Weight	141.178,79 Da
▶ Extinction Coefficient (280 nm)	146.460 M^{-1} cm^{-1}
▶ Absorbance (280 nm, 0,1%)	1,04
Isoelectric Point (pI)	6,10
Charge at pH 7,0	-26,80

Amino Acid			Number	Percent
A	Ala	Alanine	79	6,21
C	Cys	Cysteine	40	3,14
D	Asp	Aspartic Acid	62	4,87
E	Glu	Glutamic Acid	48	3,77
F	Phe	Phenylalanine	77	6,05
G	Gly	Glycine	82	6,44
H	His	Histidine	17	1,34
I	Ile	Isoleucine	76	5,97
K	Lys	Lysine	61	4,79
L	Leu	Leucine	108	8,48
M	Met	Methionine	14	1,10
N	Asn	Asparagine	88	6,91
P	Pro	Proline	58	4,56
Q	Gln	Glutamine	62	4,87
R	Arg	Arginine	42	3,30
S	Ser	Serine	99	7,78
T	Thr	Threonine	97	7,62
V	Val	Valine	97	7,62
W	Trp	Tryptophan	12	0,94
Y	Tyr	Tyrosine	54	4,24

Ilustración 8. Propiedades de la proteína S en SanpGene Viewer. Podemos ver el número de cada tipo de aa y su porcentaje en la secuencia, así como sus propiedades físico-químicas.

También podemos ir a los detalles de los genes y accederemos a una página con un **Summary**, un **Genomic Context** y a **Genomic regions, transcripts, and products y otros de mucho interés.**

Podemos obtener mucha información de dicho gen que codifica para la famosa proteína S, así como al transcrito y productos finales.

Con el fichero **FASTA** de la secuencia genómica de la proteína S descargado podemos ejecutarlo en el programa **SnapGene Viewer**, donde podremos acceder a la secuencia y seleccionar el marco o pauta de lectura 1, que es el que empleará el ribosoma para la traducción de la proteína en cuestión.

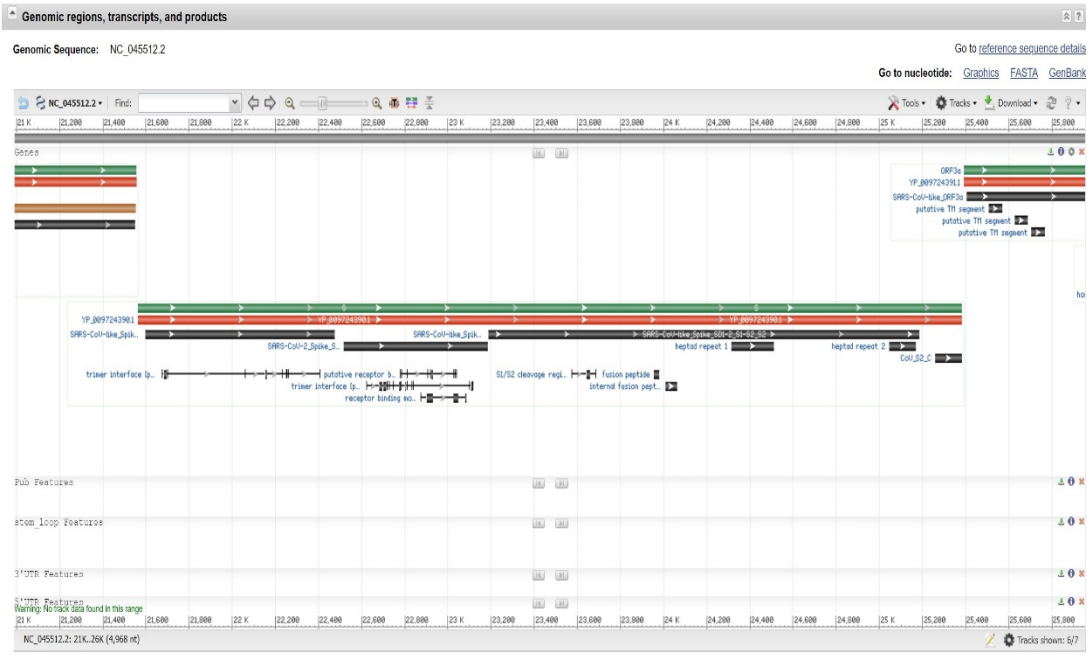


Ilustración 9. Región Genómica de la proteína S, su ORF, CDS, Transcritos y Productos. Podemos seleccionar cada unos de ellos y acceder a información descriptiva del mismo, así como a enlaces de la biblioteca del NCBI para obtener más información de cada transcrito o producto.

Desde la página anterior,**https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_009858895.2/** podemos acceder al fichero de **GenBank** del: **Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome** NCBI Reference Sequence: NC_045512.2. Desde aquí podemos acceder al fichero en formato **FASTA** para obtener la **secuenciación del genoma** a la que se hace referencia, donde indica la fecha, laboratorio, autores y métodos empleados, así como otros datos de interés para el investigador. Vemos que, a pesar de ser un **virus de ARN**, la notación en su secuenciación se hace en base al código de "letras" de nucleótidos de ADN ("A","T","C","G"), donde aparece la **Timina** en vez del **Uracilo** propio de los **ARN's**.

Ahora, para probar herramientas de bioinformática y aprender conceptos, al margen de que tengamos los genes identificados del **virus SARS-CoV-2**, pasaremos a buscar los **ORFs** copiando la secuencia del fichero **FASTA** y entrando en la página del **NCBI** (*National Center for Biotechnology Information* del *National Institutes of Health*, que es una institución pública de los EEUU). En esta página hay una herramienta de análisis. **https://www.ncbi.nlm.nih.gov/orffinder/**

ORF FINDER es un programa *on line* para encontrar **marcos abiertos de lectura** en secuencias de **DNA**, o **RNA** en nuestro caso. Al ejecutar el programa nos aparecen **159 marcos de lectura abiertos**. Hemos de tener en cuenta que **ORF finder** solo busca **marcos de lectura abiertos –**un concepto teórico que no necesariamente ha de corresponder a un **CDS-**, determinados por los **codones** de **inicio** y **stop**, sin tener en cuenta las regiones promotoras, **UTR's**, etc, es así que salen más posibles proteínas que las que realmente se dan en la descripción del virus en **GenBank**.

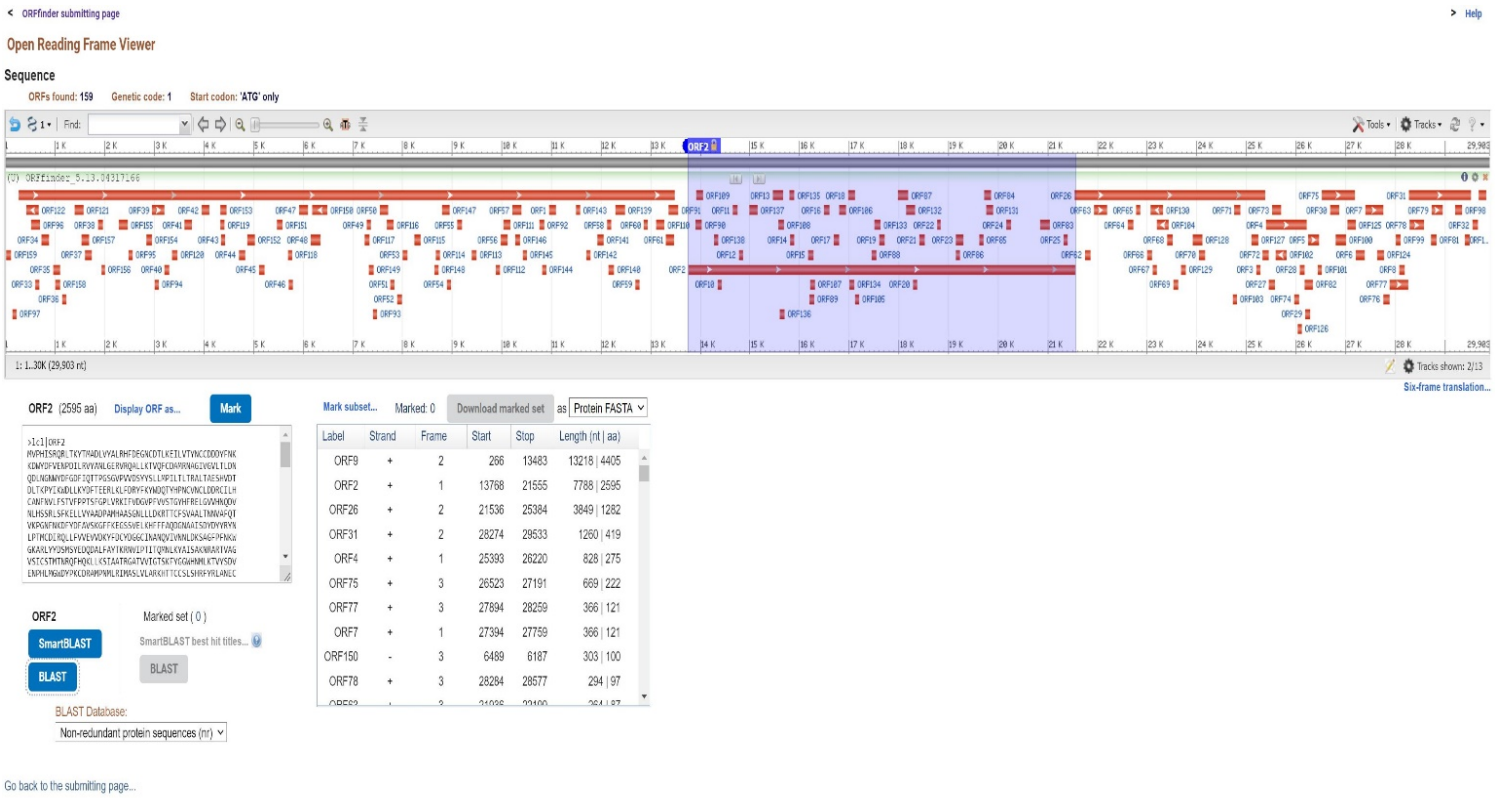


Ilustración 10. Marcos Abiertos de Lectura en el genoma del SARS-CoV-2

Sabemos que La **región de codificación** de un gen, también conocida como **CDS** (*Coding Sequence*), es esa porción del ADN de un gen o bien ARN que codifica la proteína. La región generalmente comienza en el extremo 5' por un codón de inicio y termina en el extremo 3' con un codón de terminación. Analizando el fichero de **Gen Bank NCBI**, Reference Sequence: **NC_045512.2** podemos extraer la información real del virus secuenciado y sus proteínas o productos génicos funcionales que se indican en los **FEATURES**. Podemos resumirlos para el caso del **SARS-CoV-2** concreto secuenciado por este equipo y laboratorio que se indica. Pasamos a resumir cierta información de los **FEATURES** como los **11 genes** y sus **CD's** correspondientes, así como sus localizaciones y rangos en la secuencia de la cadena de **ARN** que, por convenio, se escribe en el sentido **5'->3'** (aquí no adjuntamos la secuencia de la cadena, pues es muy larga):

```
FEATURES             Location/Qualifiers
     source          1..29903
                     /organism="Severe acute respiratory syndrome
coronavirus 2"
                     /mol_type="genomic RNA"
                     /isolate="Wuhan-Hu-1"
                     /host="Homo sapiens"
                     /db_xref="taxon:2697049"
                     /country="China"
                     /collection_date="Dec-2019"
```

Ilustración 11.Parte del FEATURES del fichero donde indica cuándo y dónde se secuenció, así como el organismo y su taxón de referencia.

Del fichero **GenBank** referente a la referencia indicada del virus **SARS-CoV-2** podemos extraer la siguiente información que resumimos sobre sus **genes** y los

respectivos **CDS's**, así como sus **traducciones a aa**, que no los adjuntaremos por ahorrar espacio.

- Región **5' UTR** 1..265
- **Gene:** 266..21555 /gene="ORF1ab" con la Regiones Codificadoras:
  - CDS: join(266..13468,13468..21555)
  - CDS: 266..13483
- **Gene:** 21563..25384 /gene="S" /locus_tag="GU280_gp02" /gene_synonym="spike glycoprotein"
  - CDS: 21563..25384
- **Gene:** 25393..26220 /gene="ORF3a" /locus_tag="GU280_gp03" /db_xref="GeneID:43740569"
  - CDS: 25393..26220
- **Gene:** 26245..26472 /gene="E" /locus_tag="GU280_gp04" /db_xref="GeneID:43740570"
  - CDS: 26245..26472
- **Gene:** 26523..27191 /gene="M" /locus_tag="GU280_gp05" /db_xref="GeneID:43740571"
  - CDS: 26523..27191
- **Gene:** 27202..27387 /gene="ORF6" /locus_tag="GU280_gp06" /db_xref="GeneID:43740572"
  - CDS: 27202..27387
- **Gene:** 27394..27759 /gene="ORF7a" /locus_tag="GU280_gp07" /db_xref="GeneID:43740573"
  - CDS: 27394..27759
- **Gene:** 27756..27887/gene="ORF7b" /locus_tag="GU280_gp08" /db_xref="GeneID:43740574"
  - CDS: 27756..27887
- **Gene:** 27894..28259 /gene="ORF8" /locus_tag="GU280_gp09" /db_xref="GeneID:43740577"
  - CDS: 27894..28259
- **Gene:** 28274..29533 /gene="N" /locus_tag="GU280_gp10" /db_xref="GeneID:43740575"
  - CDS: 28274..29533
- **Gene:** 29558..29674 /gene="ORF10" /locus_tag="GU280_gp11" /db_xref="GeneID:43740576"
  - CDS: 29558..29674
- Región **3'UTR**: 29675..29903

Cabe señalar que estos **CD's** indicados en el fichero de **GenBank** del virus son los **marcos de lectura abiertos** que **realmente se expresarán en el virus**, siendo que los localizados por **ORF Finder** anteriormente lo son desde un aspecto teórico, siendo la experimentación la que nos indicará cuáles de ellos verdaderamente son codificadores. En los **FEATURES,** en cada **CDS** aparece la **traducción a proteínas** o **producto génico** que se expresa realmente en el virus. Para probar la eficacia de **ORF FINDER** intentaremos localizar el **ORF** correspondiente a la famosa "**proteína S**" del virus, y la compararemos con la traducción que aparece en el fichero de **GenBank** que hemos consultado.

La traducción del **CDS** que da lugar a la **proteína S** produce una secuencia, según el fichero **Gen Bank**, que viene dada por los aminoácidos:

/translation="MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVS
GTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKS
WMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPI
GINITRFQTLLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKGI
YQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCF
TNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIY
QAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGV
LTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLTPT
WRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSI
AIPTNFTISVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIK
DFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSAL
LAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQ
ALNTLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKR
VDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTD
NTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL
QELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT"

Sabemos, por **GenBank**, que dicha proteína viene codificada en la secuencia del genoma en la región **CDS: 21563..25384**. Ello nos indica que **ORF Finder** busca

marcos abiertos de lectura, pero que no contempla el hecho de los **promotores** que son necesarios para el inicio de la traducción y que vendrán a ser los **CDS's** que realmente se traducen a **aminoácidos**. Ahora, al conocer donde es el inicio de la **proteína "S"** en la secuencia, procederemos a ponerlo en el buscador de **ORF FINDER**, situándonos en el nucleótido en cuestión donde comienza el codón de inicio de dicha proteína.

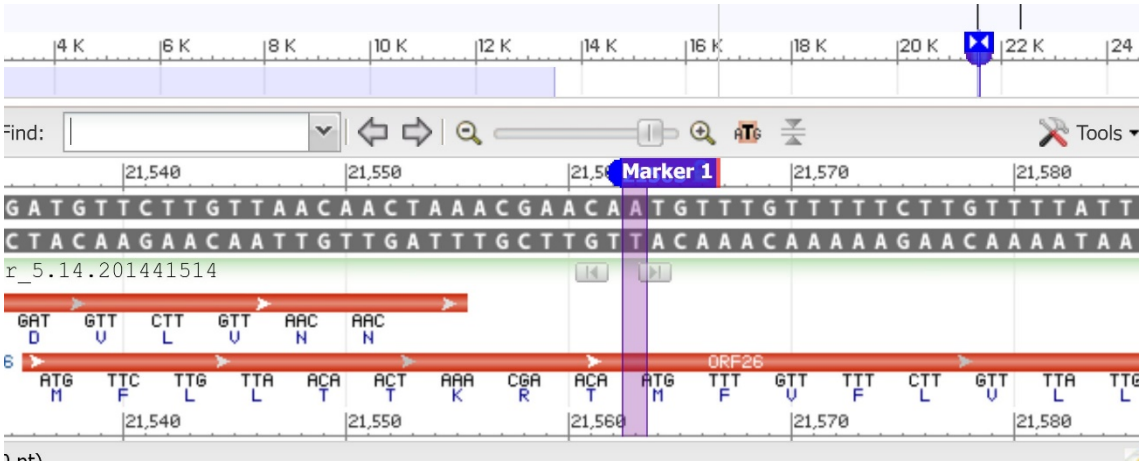


Ilustración 12. Inicio de la proteína S en la secuencia del genoma en la etiqueta Marker 1. Podemos ver los 8 primeros aa de la proteína que se corresponden con los indicados en GenBank. Vemos que el inicio de la proteína S está dentro de un ORF ya iniciado anteriormente en la secuencia por ORF FINDER.

Desde **ORF FINDER** podemos poner un marcador de inicio donde indica el CDS de la **proteína S** en el fichero **Gen Bank**, así como otro marcador en su fin: **21563..25384**. Vemos que el **CDS** de **la proteína S** se encuentra dentro de un **ORF** identificado por **ORF FINDER**, pero no en el inicio de éste a causa de la **pauta de lectura**. En ese rango de la secuencia que hemos marcado, podemos observar como **ORF FINDER** nos muestra otros marcos de lectura abiertos, algunos en la **hebra codificante** con una u otra pauta de lectura posible, otros lo mismo, pero en la hipotética **hebra complementaria** que, en nuestro caso, al tratarse un **virus monocatenario de ARN**, no existe, pero que **ORF FINDER** sí contempla en este caso.

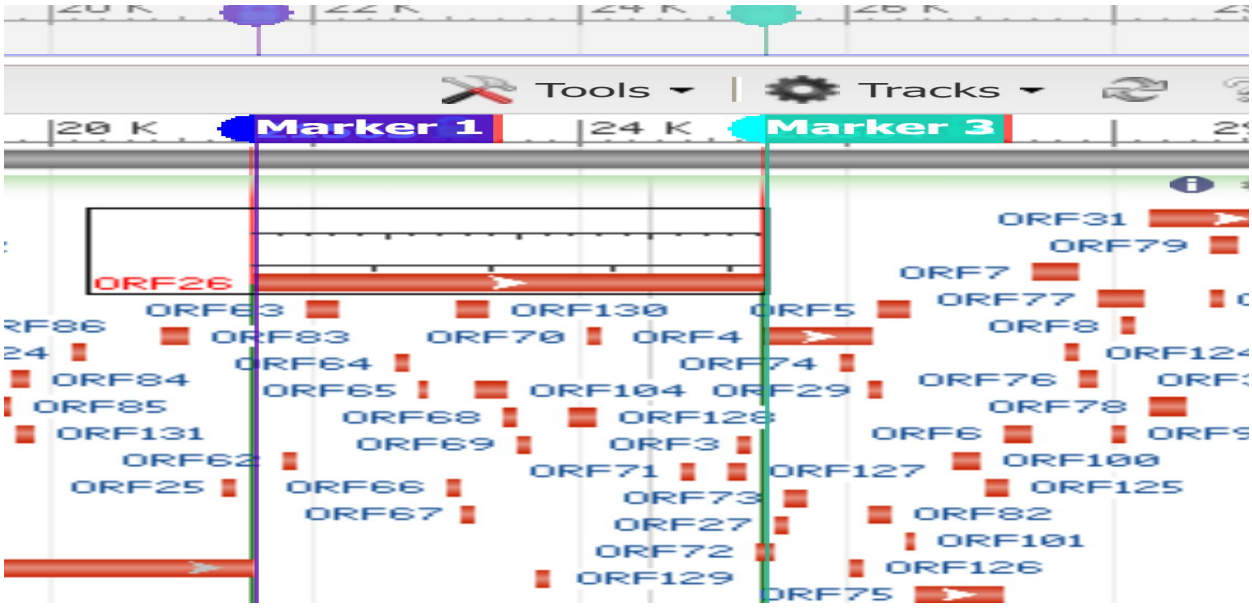


Ilustración 13.Dos marcas en el inicio y fin del CDS correspondiente a la proteína S dentro del ORF26 identificado por el programa. Mediante el buscador del navegador podemos poner marcas en la secuencia o seleccionar rangos dentro de esta para realizar búsquedas en Gen Bank del NCBI.

SnapGen Viewer

Podemos utilizar el programa gratuíto **SnapGene Viewer** y cargar la secuencia **FASTA** del virus para obtener un gráfico de su **genoma**, con los **11 marcos de lectura** correspondientes a los **11 genes** con sus respectivos **CDS**, indicados en **color naranja** y en el sentido de lectura de la hebra de **ARN**, de **5'->3'**, siendo que también aparecen otros 4 **CDS's**, en color verde y sentido opuesto, que el programa calcula al interpretar la lectura de la "teórica e inexistente, en este caso, **hebra molde**" –pues nuestro **virus** es de **ARN monocatenario positivo**, solo que **SnapGene Viewer** nos calcula la lectura de la hebra complementaria, hipotética en este caso-. También podemos observar en la secuencia dónde actúan las diferentes **enzimas.**

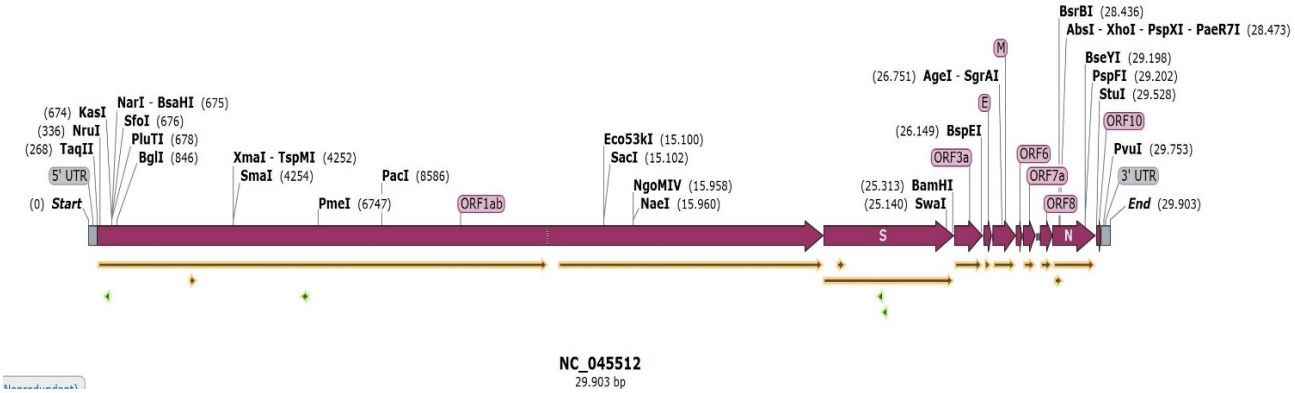
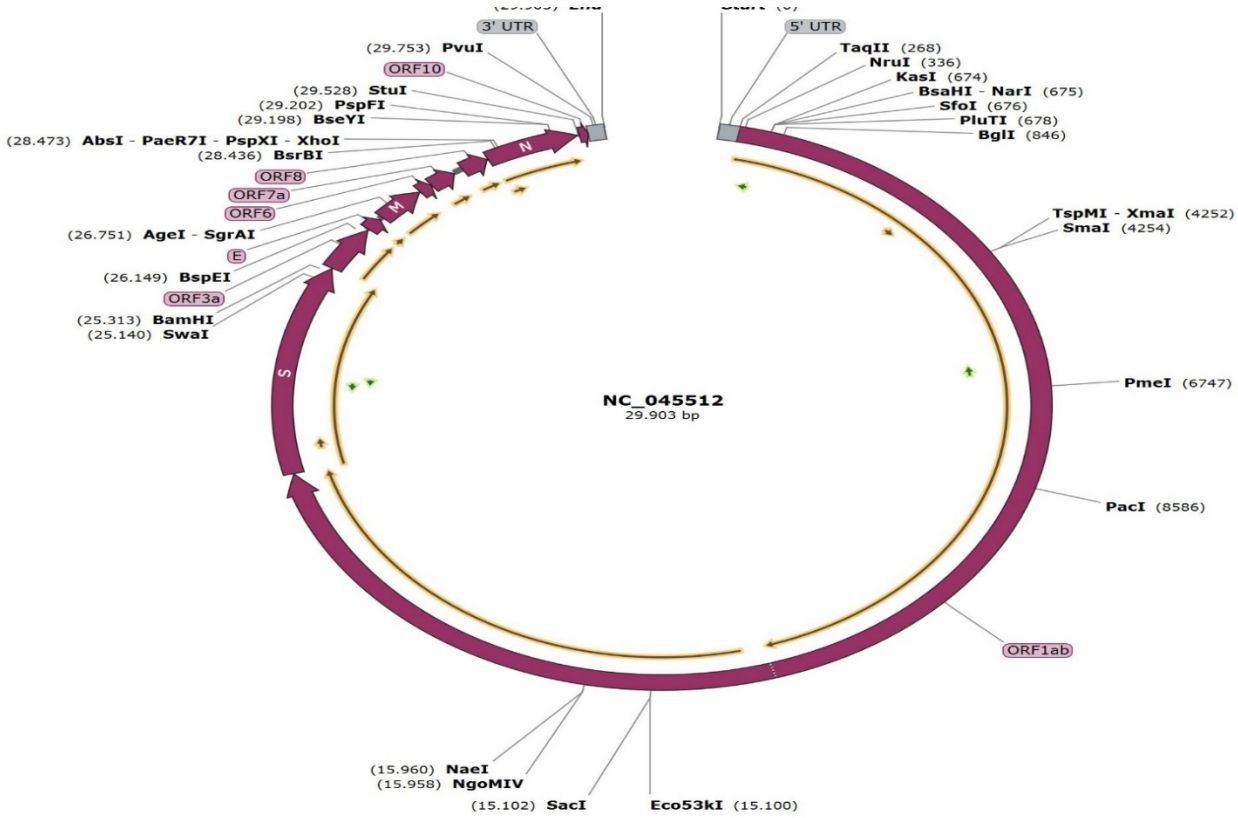


Ilustración 15. Esquema del genoma del SARS-CoV-2 indicando sus genes en naranja y las enzimas asociadas a las regiones del genoma.

Ilustración 14. Mapa del genoma del SARS-CoV-2 con sus CDS, UTR'S y productos génicos asociados.

De otro lado, podemos ver en la secuencia del genoma, mediante **SnapGen Viewer**, el



porcentaje de nucleótidos C, G, que sirve de base para el análisis de un gen por las propiedades fisicoquímicas de estos dos nucleótidos.

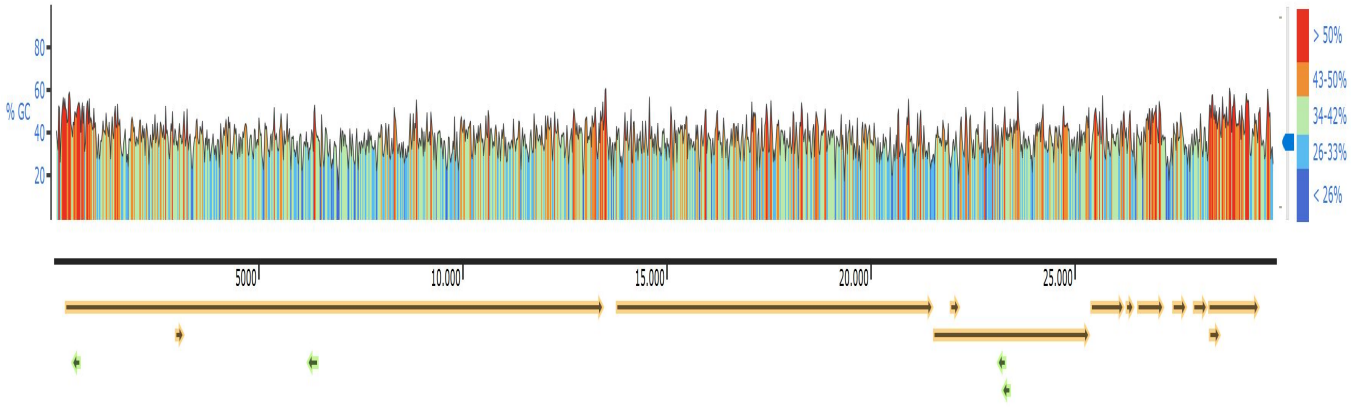


Ilustración 16. Contenido en porcentajes de bases C/G en la cadena de ARN del genoma del virus SARS-CoV-2. Observamos que en los extremos de la secuencia del genoma de ARN aumenta el porcentaje de bases de desoxirribonucleótidos C/G.

También, desde el **NCBI** podemos descargar el fichero **FASTA** del genoma de **SARS-CoV-2** secuenciado en cuestión y abrirlo con **SnapGene Viewer.**

Hemos de tener presente que **SnapGene Viewer** permite carga de la página del **NCBI** los genomas y sus **FEATURES** si le indicamos la referencia o número de la **secuencia NCBI** (en nuestro caso, para el **SARS-CoV-2** secuenciado según la referencia que el **NCBI Virus** nos proporciona, enlace proporcionado en la búsqueda primera por "**Covid 19**" que hemos realizado al principio):

NC_045512
Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome (Wu,F., et al.)
Attributes
Nuc Completeness: complete
Length: 29903
Mol Type: RNA
Host: Homo sapiens
Geo Location: China
Collection Date: 2019-12

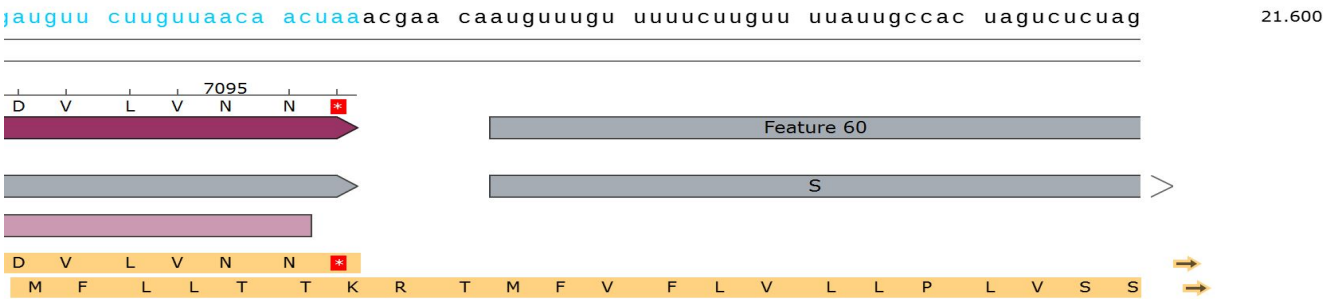


Ilustración 17. Observamos donde comienza el inicio de la codificación de la proteína S, procedente de un ORF ya abierto anteriormente en la secuencia. SnapGene Viewer nos muestra los posibles marcos de lectura abiertos, pero en este caso identifica el CDS de la proteína S, que será la codificación que finalmente se traducirá o expresará genéticamente.

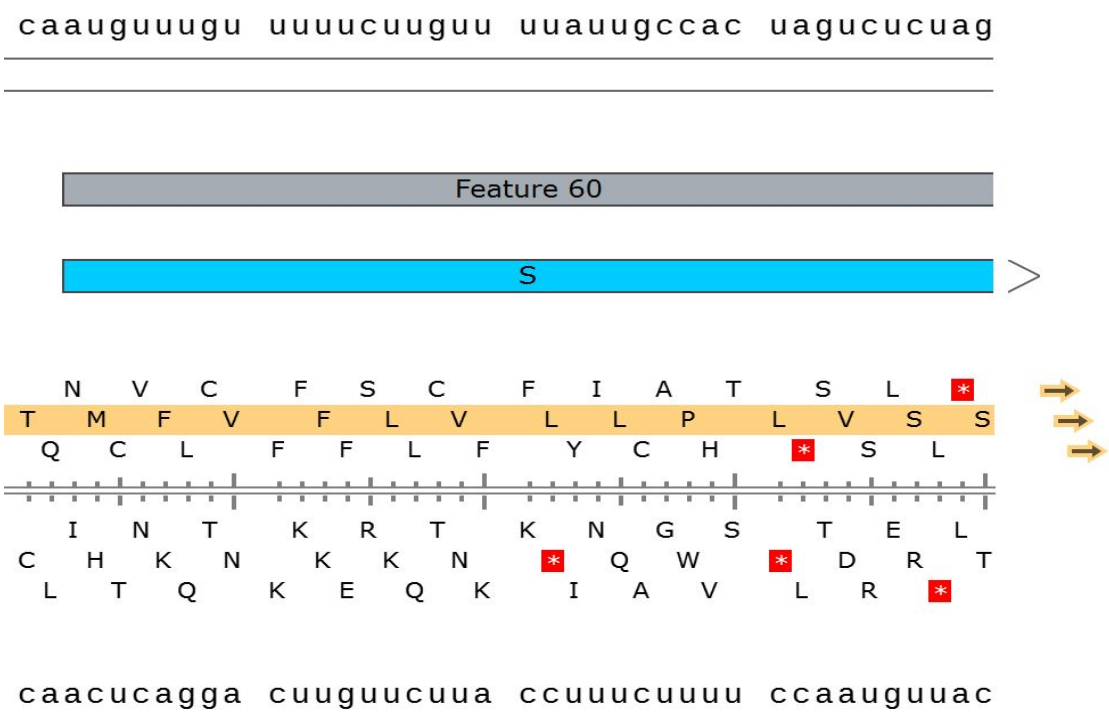


Ilustración 18. Vemos el inicio del gen que codifica para la proteína S en azul.Hay 3 pautas de lectura que se indican, las cuales dan lugar a distintos aa. la nuestra es la marcada en naranja. Vemos que antes de la Metionina que inicia la secuencia de, en el marco abierto le precede una Treonina (T).

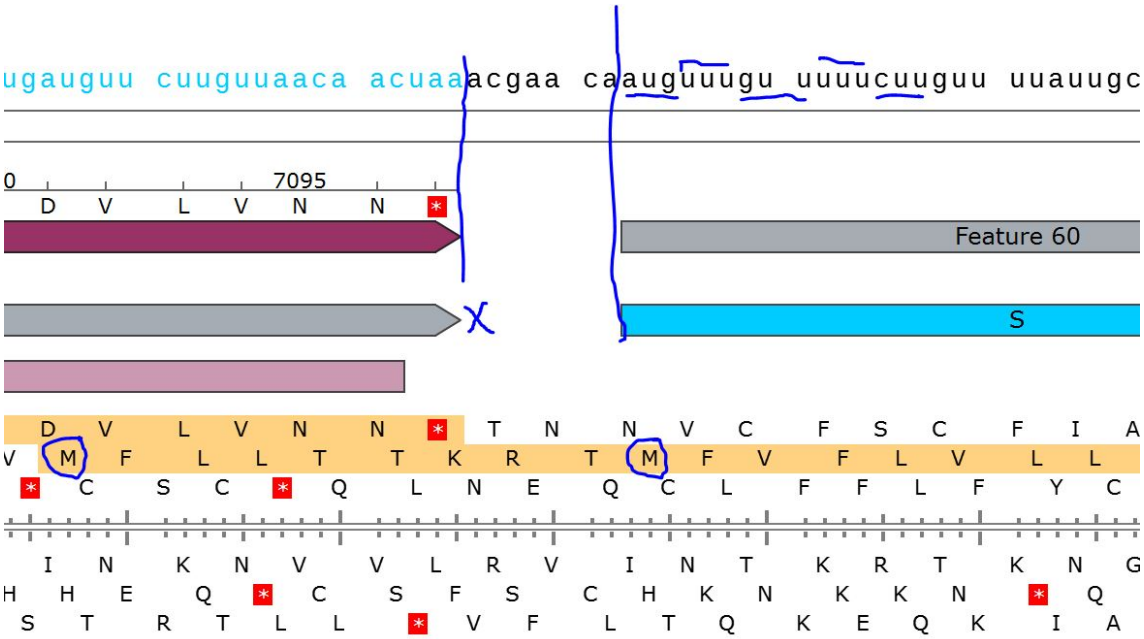


Ilustración 19. Vemos que el CDS que precede a la secuencia del CDS de la proteína S, marcado con una X en en dibujo, tiene una pauta de lectura diferente a la de la codificación de la proteína S, la cual está debajo y cuyo ORF se inicia anteriormente, indicado con la M en un círculo más a la izquierda. Ese ORF ya está abierto, pero en él la cadena ha sido utilizada en una pauta de lectura diferente a la identificada por el programa, la pauta de lectura que produce la traducción que hay encima de ésta. Cuando se inicia la traducción de la proteína S, el ribosoma utilizará esta segunda pauta y no la primera que venía utilizando, siendo que indicamos el primer aa de la proteína S con la M en el segundo círculo de la figura, lugar donde comienza la secuencia de aa de la proteína S.

Ahora podemos volver al **NCBI datasets** del gen que codifica para la **proteína S**, acceder a las regiones genómicas del: *S surface glycoprotein [Severe acute respiratory syndrome coronavirus 2] Gene ID: 43740568, updated on 7-May-2022 (https://www.ncbi.nlm.nih.gov/gene/43740568).* Seleccionamos la banda roja corres`pondiente al **CDS** del **gen S** y accedemos a un menú que nos lleva a la base de datos **BLAST**, en concreto accedemos a **BLAST Protein: YP_009724390.1**, en donde se selecciona la región concreta de la **proteína S**. Nos aparecerá un listado de búsqueda con resultados de secuencias que, por alineamiento, aparecen de mayor a menor identificación con las habidas en la **BB.DD** del **BLAST Protein**.

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments — Download — Select columns — Show 100

select all *100 sequences selected* — GenPept | Graphics | Distance tree of results | Multiple alignment | MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2557	2557	99%	0.0	100.00%	1282	BCN86353.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2555	2555	99%	0.0	100.00%	1273	QIZ15717.1
Chain A, Spike glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2554	2554	99%	0.0	100.00%	1310	6XR8_A
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QOF12329.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QWC77885.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QRN63738.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QMT96172.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QIU80973.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QOF15989.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QMT94564.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QKU28906.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QOU86714.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QZJ49063.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QKV35819.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QIA98583.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QIZ14569.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QIZ16559.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QIU81873.2
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QOU93902.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QOF10625.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	100.00%	1261	QXF49728.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QIU81885.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QWC76579.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QKV06859.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QOF08429.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QIS61254.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	99.92%	1273	QNV50022.1
surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]	Severe acute respiratory syndrome coronaviru…	2553	2553	99%	0.0	100.00%	1273	YP_009724390.1

Ilustración 20. Listado de secuencias encontradas tomando como referencia la proteína S en Blast del NCBI. La mayoría muestrasn un 100% de identidad.

Podemos acceder a la información del enlace de la 1ª secuencia de la búsqueda y en ella aparecerá sus descripciones, así como su alineamiento con las secuencias de la BB.DD de BLAST, mostrando las identidades de nucleótidos y posibles gaps en los alineamientos en la búsqueda por comparación de secuencias. Desde este fichero hay un enlace de Graphics donde podemos acceder a un navegador de BLAST sobre la secuencia de aa. En el, en rojo, podemos ver una primera región no alineada, pues el alineamiento correcto comienza en la posición 12, como se indica en el resultado de la búsqueda en la base de datos de BLAST.

Download GenPept Graphics

surface glycoprotein [Severe acute respiratory syndrome coronavirus 2]

Sequence ID: BCN86353.1 Length: 1282 Number of Matches: 1

Range 1: 21 to 1282 GenPept Graphics Next Match

Score	Expect	Method	Identities	Positives	Gaps
2557 bits(6627)	0.0	Compositional matrix adjust.	1262/1262(100%)	1262/1262(100%)	0/12

```
Query   12     SSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVS   71
               SSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVS
Sbjct   21     SSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVS   80

Query   72     GTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVC   131
               GTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVC
Sbjct   81     GTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVC   140

Query   132    EFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLRE   191
               EFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLRE
Sbjct   141    EFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNLRE   200

Query   192    FVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTP   251
               FVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTP
Sbjct   201    FVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTP   260

Query   252    GDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKG   311
               GDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKG
Sbjct   261    GDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVEKG   320

Query   312    IYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNS   371
               IYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNS
Sbjct   321    IYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLYNS   380

Query   372    ASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTG   431
               ASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTG
Sbjct   381    ASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDDFTG   440

Query   432    CVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFP   491
               CVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFP
Sbjct   441    CVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFP   500

Query   492    LQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGV   551
               LQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGV
Sbjct   501    LQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNGLTGTGV   560

Query   552    LTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVL   611
               LTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVL
Sbjct   561    LTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVL   620

Query   612    YQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGIC   671
               YQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGIC
Sbjct   621    YQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGIC   680

Query   672    ASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSM   731
               ASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSM
Sbjct   681    ASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSM   740

Query   732    TKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKT   791
               TKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKT
Sbjct   741    TKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKT   800

Query   792    PPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLIC   851
               PPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLIC
Sbjct   801    PPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLIC   860

Query   852    AQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGV   911
               AQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGV
Sbjct   861    AQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGV   920

Query   912    TQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFG   971
               TQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFG
Sbjct   921    TQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFG   980

Query   972    AISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSE   1031
               AISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSE
Sbjct   981    AISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSE   1040

Query   1032   CVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPR   1091
               CVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPR
Sbjct   1041   CVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPR   1100

Query   1092   EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEE   1151
               EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEE
Sbjct   1101   EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEE   1160

Query   1152   LDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIK   1211
               LDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIK
Sbjct   1161   LDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIK   1220

Query   1212   WPWYIWLGFIAGLIAIVMVTImlccmtsccsclkgccscgscckFDEDDSEPVLKGVKLH   1271
               WPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLH
Sbjct   1221   WPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLH   1280

Query   1272   YT   1273
               YT
Sbjct   1281   YT   1282
```

Ilustración 21. Cadena de aa de la Surface Glycoprotein de la proteína S. La 1ª cadena comienza a alinerse a partir del aa 12 hasta el 1281.

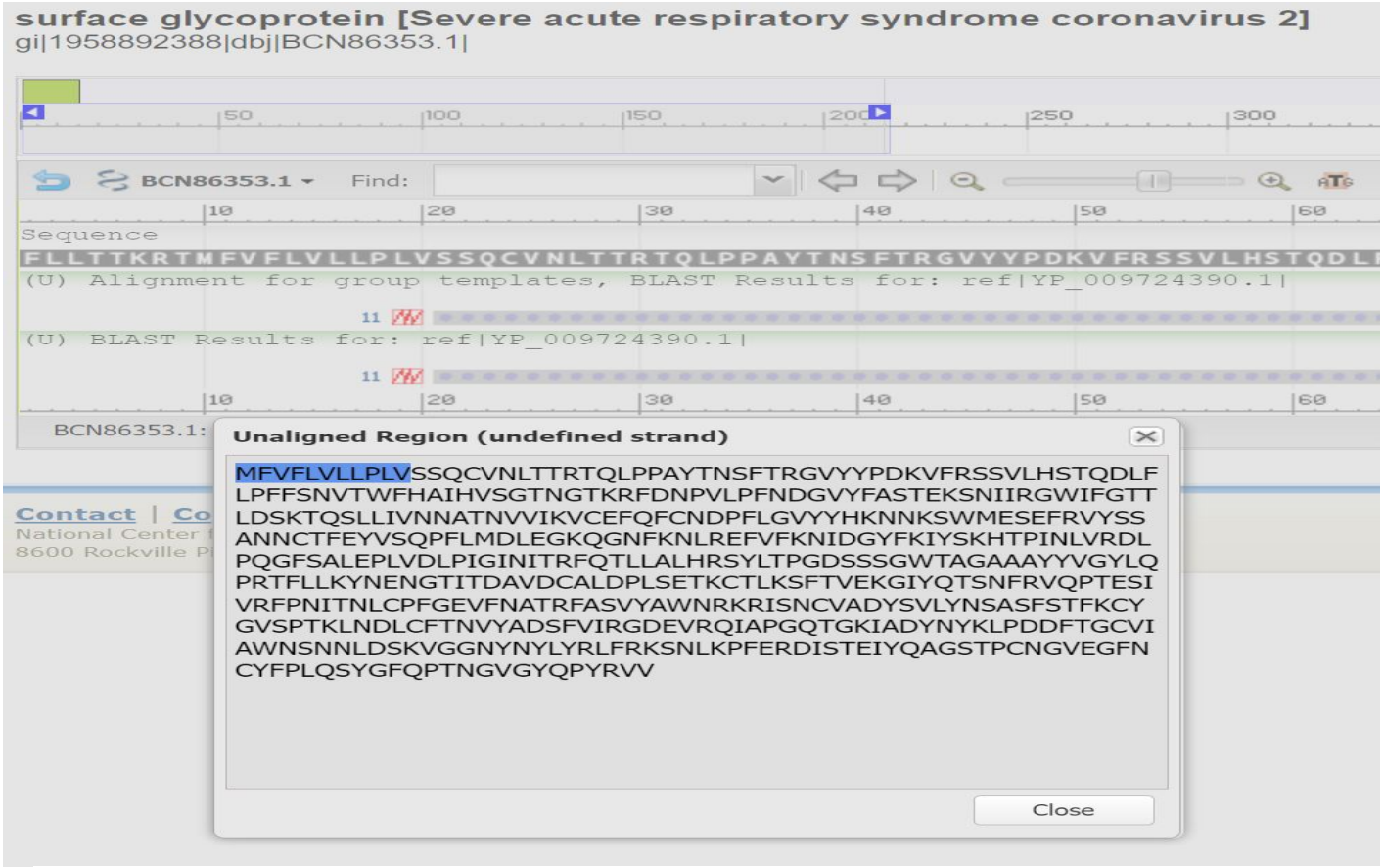


Ilustración 22. Hemos seleccionado en el navegador de BLAST de la 1ª secuencia encontrada la región no alineada, marcada en líneas rojas y descrita en el recuadro donde, en azul, se muestran los aminoácidos que no se han alineado con nuestra secuencia "problema". La cadena empieza a alinearse a partir del aa 12 hasta el 1281.

BIBLIOGRAFÍA SEGUIMIENTO

- https://apps.who.int/iris/bitstream/handle/10665/338892/WHO-2019-nCoV-genomic_sequencing-2021.1-spa.pdf?sequence=1&isAllowed=y
- https://www.ncbi.nlm.nih.gov/sars-cov-2/?utm_source=gquery&utm_medium=referral&utm_campaign=KnownItemSensor:org_genome
- https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Genome&VirusLineage_ss=SARS-CoV-2,%20taxid:2697049
- https://conogasi.org/articulos/gen-desde-el-codigo-genetico-hasta-la-ingenieria-genetica/
- https://www.ncbi.nlm.nih.gov/