INTRODUCCIÓN A LA BIOINFORMÁTICA

POSTGRADO Y FORMACIÓN PROFESORADO

TRABAJO FINAL

<u>UNED</u>

PRÁCTICA O LECCIÓN GUIADA

UTILIDADES DE LA BIOINFORMÁTICA

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

Santiago Royuela Samit <u>Mayo de 2022</u>

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 físicoquímica 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: <u>https://www.ncbi.nlm.nih.gov/</u>) 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.

| Search NCBI | covid 19 | | X Se | arch | |
|--|--|--|-------------------|--|---|
| esults found in 25 databases | | | | | |
| TAXONOMY Severe acute Severe acute of Severe acute Taxonomy ID: 2097 NCBI SARS-COM | e respiratory syn espiratory syndrom e respiratory syndro 049 I-2 resources | Was t drome coronavirus 2 e coronavirus 2 is a below-specie ome-related coronavirus | es classification | NCBI Virus The most up-to-de SARS-CoV-2 nucle protein sequences | ate set of eotide and s |
| NCBI Virus Browse and download | | | | | |
| Literature | | Genes | | Proteins | |
| Bookshelf | 4,404 | Gene | 433 | Conserved Domains | 0 |
| MeSH | 144 | GEO DataSets | 5,546 | Identical Protein Groups | 0 |
| NLM Catalog | 1,375 | GEO Profiles | 0 | Protein | 551,767 |
| PubMed | 255,633 | HomoloGene | 0 | Protein Family Models | 0 |
| PubMed Central | 328,540 | PopSet | 86 | Structure | 3 |
| Genomes | | Clinical | | A CONTRACTOR | The second se |
| Assembly | 0 | ClinicalTrials.gov | 8,828 | N. Carlos | |
| BioCollections | 0 | ClinVar | 12 | J. S. | |
| BioProject | 417 | dbGaP | 0 | | × |
| BioSample | 4,869,249 | dbSNP | 0 | | |
| Genome | 0 | dbVar | 9,335 | SARS-CoV-2 protein struct | ures |
| Nucleotide | 53,083 | GTR | 99 | View 3D structures and conserv novel coronavirus proteins, inclu | ed domains of Iding (S)pike. |
| SRA | 3,933,529 | MedGen | 106 | (E)nvelope), (M)embrane, and (N | v)ucleocapsid |
| Taxonomy | 1 | OMIM | 5 | | |
| PubChem | | | | | |
| BioAssays | 610 | | | | |
| Compounds | 1,668 | | | | |

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

Vemos que si clicamos en <u>TAXONOMY</u> nos llevará a una página donde aparece un resumen de la taxonomía del virus, su genoma e identificaciones pertinentes que

Severe acute respiratory syndrome coronavirus 2 🔅

Severe acute respiratory syndrome coronavirus 2 is a below-species classification of Severe acute respiratory syndrome-related coronavirus

2,278

| Browse taxonomy | | |
|-----------------------------|---|--|
| Current scientific name | Severe acute respiratory syndrome coronavirus 2 | |
| Acronym | SARS-CoV-2 | |
| Genome type | ssRNA(+) | |
| NCBI Taxonomy ID | 2697049 | |
| For more details see NCBI T | axonomy | |

Genome Browse all genomes in NCBI Virus

Pathways

Substances

Reference genome ASM985889v3 Jan 13, 2020 Referg GCF_009858895.2

| Download | |
|----------------|---------|
| Genome size | 29.9 kb |
| Viral segments | 1 |
| Genes | 11 |

Annotation from NCBI RefSea



External links Encyclopedia of Life Wikipedia pasaremos a analizar.

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 | enome | | | | | | | | |
|--|---------|--|--|--|--|--|--|--|--|
| owse all genomes in NCBI Virus | | | | | | | | | |
| Reference genome ASM985889v3 Jan 13, 2020 RefSeq GCF_009858895.2 | | | | | | | | | |
| GCF_009858895.2 | | | | | | | | | |
| Genome size | 29.9 kb | | | | | | | | |
| Viral segments | 1 | | | | | | | | |

Genes

Annotation from NCBI RefSeq



11

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

| 29.9 kb |
|-----------------|
| 1 |
| 1 |
| 37.5 |
| 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 "<u>View</u> <u>all genes</u>".

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.

| Download 🗸 | Select columns | | Rows per page 20 - 1-11 of 11 < | > |
|------------|----------------|--------------------------------------|---------------------------------|--------|
| Sene ID | Symbol | Gene name | Gene type Transcripts | Action |
| 3740568 | S | surface glycoprotein | protein-coding | : |
| 13740569 | ORF3a | ORF3a protein | protein-coding | : |
| 13740570 | E | envelope protein | protein-coding | : |
| 13740571 | м | membrane glycoprotein | protein-coding | : |
| 13740572 | ORF6 | ORF6 protein | protein-coding | : |
| 13740573 | ORF7a | ORF7a protein | protein-coding | : |
| 3740574 | ORF7b | ORF7b | protein-coding | : |
| 13740575 | Ν | nucleocapsid phosphoprotein | protein-coding | : |
| 13740576 | ORF10 | ORF10 protein | protein-coding | : |
| 13740577 | ORF8 | ORF8 protein | protein-coding | : |
| 13740578 | ORF1ab | ORF1a polyprotein;ORF1ab polyprotein | protein-coding | : |

Si clicamos en cada uno de los **11 genes** podemos acceder a un navegador genómico dichas para ver 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

| | | J | |
|----------|--------|---------------|--|
| Cono ID | Symbol | Gene name | Download |
| | Symbol | Gene name | Download a data package for gene (GeneID: |
| 13740568 | S | surface glyco | 43740568) Jin |
| 13740569 | ORE3a | ORE3a protei | Select file types - estimated size 1 Mb |
| 10740000 | | | ✓ Gene sequences (FASTA) |
| 13740570 | E | envelope pro | Transcript sequences (FASTA) |
| 13740571 | М | membrane g | Protein sequences (FASTA) |
| 13740572 | ORF6 | ORF6 protein | Your selected data will be downloaded as a ZIP archive |
| 13740573 | ORF7a | ORF7a prote | Name your file |
| | | | ncbi_dataset.zip |
| 13740575 | N | nucleocapsic | |
| 127/0576 | ODE10 | ODE10 proto | |

proteína para analizarlo con el programa SnapGen Viewer que veremos más R 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**.

| MFV | FLV | LLPLV | SSQC | VNLTT | RTQLPF | PAYTNS | SFTRG | VYYPDK | VFRSS | VLHST | QDLFI | LPFFSI | NVTWFH | AIHVS | GTNG | TKRFDI | VPVLP | FNDGVY | FAST | EKSNII | RGWIF | GTTLE | DSKTQS | LLIV | NNATNV | VIKVCE | FQFCNDPF |
|------|-------|-------|------|--------|--------|--------|--------|--------|--------|-------|--------|--------|--------|---------|---------|--------|---------|--------|--------|---------|--------|--------|--------|--------|---------|--------|-------------|
| 1 | 1 | 10 | | 20 | | 30 | | 40 | | 50 | L. | 60 | | 70 | | 80 | | 90 | 1 | 100 | | 110 | | 120 | | 130 | 1 40 |
| LGV | YYH | KNNKS | WMES | EFRVYS | SSANNO | TFEY | /SQPF | LMDLEG | KQGNF | KNLRE | FVFKI | NIDGY | FKIYSK | (HTPI) | ILVRD | LPQGF | SALEP | LVDLPI | GINI | TRFQTI | LALHR | SYLT | GDSSS | GWTA | GAAAYY | VGYLQF | RTFLLKYN |
| | 1 | 150 | 1 | 160 | 1 | 170 | T | 180 | | 190 | ļ | 200 | ļ | 210 | 1 | 220 | | 230 | | 240 | | 250 | 1 | 260 | Į. | 270 | 280 |
| ENG | TIT | DAVDC | ALDP | LSETKO | CTLKSF | TVEK | GIYQT | SNFRVQ | PTESI | VRFPN | ITNL | CPFGE | VFNATF | FASV | (AWNR | KRISN | CVADY | SVLYNS | ASFS | TFKCYG | SVSPTK | LNDLO | CFTNVY | ADSF | VIRGDE | VRQIAF | ∙GQTGKIAD |
| | | 290 | | 300 | | 310 | | 320 | I | 330 | | 340 | | 350 | | 360 | | 370 | | 380 | | 390 | 1 | 400 | | 410 | 420 |
| YNY | KLP | DDFTG | CVIA | WNSNNL | LDSKVG | GNYNY | (LYRLI | FRKSNL | KPFER | DISTE | IYQA | GSTPCI | NGVEGF | NCYFF | PLQSY | GFQPTI | VGVGY | QPYRVV | VLSFI | ELLHAF | PATVCO | PKKST | FNLVKN | IKCVN | FNFNGL | TGTGVL | .TESNKKFL |
| | 1 | 430 | | 440 | | 450 | | 460 | | 470 | | 480 | | 490 | | 500 | | 510 | | 520 | | 530 | 1 | 540 | | 550 | 560 |
| PFQ | QFG | RDIAD | TTDA | VRDPQT | TLEILD | ITPCS | SFGGV | SVITPG | TNTSN | QVAVL | YQDVI | NCTEVI | PVAIHA | DQLTF | TWRV | YSTGSI | VVFQT | RAGCLI | GAEH | VNNSYE | CDIPI | GAGI | CASYQT | QTNSI | PRRARS | VASQSI | IAYTMSLG |
| | ł | 570 | ļ | 580 | l | 590 | 1 | 600 | 1 | 610 | - L | 620 | | 630 | 1 | 640 | j | 650 | | 660 | I | 670 | i | 680 | Į, | 690 | 700 |
| AEN | SVA | YSNNS | IAIP | TNFTIS | SVTTEI | LPVSM | (TKTS) | VDCTMY | ICGDS | TECSN | ILLLQ | YGSFC | TQLNRA | LTGIA | VEQD | KNTQE | /FAQV | KQIYKT | PPIK | DFGGFN | FSQIL | .PDPSH | (PSKRS | FIED | LLFNKV | TLADAG | FIKQYGDC |
| | 1 | 710 | ļ | 720 | 1 | 730 | 1 | 740 | Ţ | 750 | ļ | 760 | | 770 | Ī | 780 | 1 | 790 | 1 | 800 | | 810 | 1 | 820 | [| 830 | 840 |
| LGD | IAA | RDLIC | AOKF | NGLTVI | LPPLLT | DEMIA | AOYTS | ALLAGT | ITSGW | TFGAG | AALO | IPFAM | OMAYRF | NGIG | /TONV | LYENOI | KLIAN | OFNSAI | GKIO | DSLSST | ASALO | KLOD | /VNONA | (OALN' | TLVKOL | SSNFGA | ISSVLNDI |
| | 1 | 850 | | 860 | 1 | 870 | | 880 | 1 | 890 | | 900 | | 910 | 1 | 920 | | 930 | | 940 | 1 | 950 | 1 | 960 | 1 | 970 | 980 |
| I SR | I D K | VFAFV | OTDR | LITGRI | 05101 | YVTOC |)I TRA | AFTRAS | ANI AA | TKMSF | CVI GI | OSKRVI | DECGKO | SYNT MS | SEPOS. | APHGV | /FI HV | TYVPAC | FKNF | TTAPAI | CHDGK | AHEPE | REGVEN | SNGT | HWEVTO | RNEYEF | OTITIONT |
| | 1 | 990 | 1 | 1000 | 1 | 1010 | 1 | 1020 | 1 | 1030 | Ţ | 1040 | 1 | 1050 | 1 | 1060 | 1 | 1070 | T | 1080 | 1 | 1090 | 1 | 1100 | 1 | 1110 | 1120 |
| FVS | GNC | DVVTG | TVNN | туулрі | OPEL | ISEKEE | | EKNHTS | וחעח | GDISC | TNAS | /VNTO | KETNRI | NEVA | (NI NE) | | 1 FI GK | VENVIK | .MDMA. | TWI GE1 | AGI TA | TVMVI | ттинсо | NTSC | rsni kg | 201211 | CCKEDEDD |
| | 1 | 1130 | 1 | 1140 | -41-55 | 1150 | | 1160 | 1 | 1170 | | 1180 | | 1190 | | 1200 | (| 1210 | 1 | 1220 | | 1230 | 1 | 1240 | | 1250 | 1260 |
| | | | | | | | | | | | | | | | | | | | | | | | | | | | |

SEPVLKGVKLHYT

1270 1273

| | Leng | gth | | Whole Protein 1273 aa | | | | | | | |
|---|------|----------|----------------------|--|---------|--|--|--|--|--|--|
| × | Mole | ecular V | Weight | 141.178 | 3,79 Da | | | | | | |
| | Exti | nction | Coefficient (280 nm) | 146.460 M ⁻¹ cm ⁻¹ | | | | | | | |
| | Abse | orbanc | e (280 nm, 0.1%) | 1.04 | | | | | | | |
| | Isoe | lectric | Point (pI) | 6,1 | 0 | | | | | | |
| | Cha | rge at | pH 7,0 ~ | -26 | ,80 | | | | | | |
| | Ami | no Acio | 1 | Number | Percent | | | | | | |
| | А | Ala | Alanine | 79 | 6,21 | | | | | | |
| | С | 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 | | | | | | |
| | н | His | Histidine | 17 | 1,34 | | | | | | |
| | I | Ile | Isoleucine | 76 | 5,97 | | | | | | |
| | к | Lys | Lysine | 61 | 4,79 | | | | | | |
| | L | Leu | Leucine | 108 | 8,48 | | | | | | |
| | М | Met | Methionine | 14 | 1,10 | | | | | | |
| | N | Asn | Asparagine | 88 | 6,91 | | | | | | |
| | Р | Pro | Proline | 58 | 4,56 | | | | | | |
| | Q | Gln | Glutamine | 62 | 4,87 | | | | | | |
| | R | Arg | Arginine | 42 | 3,30 | | | | | | |
| | S | Ser | Serine | 99 | 7,78 | | | | | | |
| | т | Thr | Threonine | 97 7,62 | | | | | | | |
| | V | Val | Valine | 97 | 7,62 | | | | | | |
| | vv | Trp | Tryptophan | 12 | 0,94 | | | | | | |
| | Y | Tyr | Tyrosine | 54 4,24 | | | | | | | |
| | | | | | | | | | | | |

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

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 físicoquímicas.

También podemos ir alos detalles de los genes yaccederemos a una páginacon un Summary, unGenomic Contexty aGenomic regions,transcripts, and products yotros 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,<u>https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF 009858895.2/</u> 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. <u>https://www.ncbi.nlm.nih.gov/orffinder/</u>

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.



| Open Reading Frame Viewer | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|--|---|--------|---|-------|---|---------|---------------------------------|---------------------------------------|-------------------------------|-------------------------|-------------------------------|-------------------------|--|--|-----------------|--------------|---|----------------------------------|
| Sequence | | | | | | | | | | | | | | | | | | | | | | |
| | 0 . 7 | | | | | | | | | | | | | | | | | | | Q.r. | a. Ara | - A 0 - |
| | | IR K | 19.K 1 | RK (11 | K 112 K | Ita K | 0953 | 115 K | 16 K | 17 K | 118 K | 19 K | 120 K | 121 K | 122 K | 23 K | 124 K | 125 K | DE K | 10 I27 K | ols • Paci | S* 10 7* |
| P* P* P* P* | | | | | · | | UNU 2 W | | | | | | | | | | | | | | | |
| (U) ORFfinder_5.13.04317166 | | _ | | | | _ | - | | | | | | | 00000 | | | _ | | | _ | 00531 | 00> |
| 06F12 06F32 06F32 06F42 06F33 06F47 0FF36 06F155 0FF14 06F13 06 | ORF158 ORF58 ORF49 OF | 0RF116 01 8F117 0RF11 0RF53 0RF11 9RF149 0RF54 | ORF147 ORF5 RF55 0RF56 ORF114 0RF113 ORF148 0 | 7 ORF1 ORF111 ORF ORF146 ORF146 ORF145 ORF112 ORF | 0RF143 92 0RF58 0 0RF14 0RF142 F144 0R | 0RF139 | 0RF91 0RF11 0RF90 0RF90 0RF12 0RF12 2 | 0RF13 | 0RF18 0 0RF188 0RF188 0RF17 0RF15 | ORF186 | ORF133 ORF ORF21 ORF21 ORF88 | 97 1F132 22 0 0RF23 0 0RF | 0RF24 0RF24 0RF25 86 | ORF83 ORF83 ORF25 | 8F63 🔽 08F 08F64 12 📕 0 | 0RF68 ORF68 ORF67 | RF130 OF ORF104 ORF70 ORF70 ORF129 | 08F71 08F7 08F4 128 00 08F72 08F73 00F73 | 0RF127 0RF5 | 38 08F7 08F7 | 0RF31 0RF79 25 0RF78 0RF99 0RF124 0RF124 0RF124 | ORF98 ORF32 ORF81 ORF81 |
| 0F96 0F97 | 0 | RF52 CRF93 | | | | | 0010 | | ORF136 | 89 0RF1 | 85 | | | | | UKPOP | • | ORF183 | ORF74 ORF29 ORF | 0RF7 | 16 | |
| 1 1K 2K 3K 4K 5K | 6 K 7 K | 8 K | 9 K 1 | ек рі | К 12 К | 13 K | 14 K | 15 K | 16 K | 17 K | 18 K | 19 K | 20 K | 21 K | 22 K | 23 K | 24 K | 25 K | 26 K | 27 K | 28 K | 29,98 |
| 1: 130K (29,903 nt) | | | | | | | | | | | | | | | | | | | | | 🕺 🌻 Trac | ks shown: 2/13 |
| ORF2 (2595 as) Display ORF as Mark | Mark subset Label St ORF9 ORF2 ORF26 ORF26 | Marked: 0 rand Fran + + + + | Download m ie Start 2 266 1 13768 2 21536 2 28274 | arked set 3 Stop 13483 21555 25384 29533 | as Protein FAST Length (nt aa) 13218 4405 7788 2595 3849 1282 1260 419 | A ~ | | | | | | | | | | | | | | | | |
| VSTCSTHTINGPHQCLLSSTAAT6GATV/TGTSKEVCGAHMILKEVVSOV ENPHLM6007PKCDR#MPMHLRIMASLVLARKHTTCCSLSHRFYRLANEC | ORF75 | + : | 3 26523 | 20220 | 669 222 | | | | | | | | | | | | | | | | | |
| OPE2 Norked est (0) | ORF77 | + | 3 27894 | 28259 | 366 121 | | | | | | | | | | | | | | | | | |
| SmartBLAST SmartBLAST best hit titles 9 | ORF7 | + | 1 27394 | 27759 | 366 121 | | | | | | | | | | | | | | | | | |
| BLAST | ORF150 | - | 3 6489 | 6187 | 303 100 | | | | | | | | | | | | | | | | | |
| BLAST | ORF78 | + | 3 28284 | 28577 | 294 97 | | | | | | | | | | | | | | | | | |
| BLAST Database: Non-redundant protein sequences (nr) Y | | | 20102 | 00400 | 004107 | | | | | | | | | | | | | | | | | |

Go back to the submitting page

< ORFfinder submitting page

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: <u>NC 045512.2</u> 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 | 129903 |
| | <pre>/organism="Severe acute respiratory syndrome</pre> |
| 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.



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.



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

jauguu cuuguuaaca acuaaacgaa caauguuugu uuuucuuguu uuauugccac uagucucuag



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.

21.600

caauguuugu uuuucuuguu uuauugccac uagucucuag 21.600



caacucagga cuuguucuua ccuuucuuuu ccaauguuac

21.750

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: <u>S surface glycoprotein [Severe acute</u> <u>respiratory syndrome coronavirus 2] Gene ID: 43740568, updated on 7-May-2022</u> (<u>https://www.ncbi.nlm.nih.gov/gene/43740568</u>). 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 <u>**BLAST Protein: YP 009724390.1**</u>, 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**.

| De | scriptions | Graphic Summary | Alignments | Taxon | omy | | | | | | | | | | | |
|----|----------------|----------------------------------|------------------------|-------------|-----------|------------------|---------|----------------|--------------|----------------|----------------|------------|--------------------|-------------|---------------|-----|
| Se | quences pro | oducing significant al | ignments | | | | | Downlo | ad 〜 | s | elect o | olumn | s ~ S | how | 100 🗸 | 0 |
| | select all 1 | 00 sequences selected | | | | Gen | Pept | Graphics | Distan | ce tree | of res | ults N | <u>lultiple al</u> | ignme | nt MSA View | ver |
| | | Description | | | | Scientifi | c Name | | Max Score | Total Score | Query Cover | E value | Per. Ident | Acc. Len | Accession | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | 5.2] | Severe ad | cute respiratory | syndror | ne coronaviru. | . 2557 | 2557 | 99% | 0.0 | 100.00% | 1282 | BCN86353.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe ad | oute respiratory | syndron | ne coronaviru. | . 2555 | 2555 | 99% | 0.0 | 100.00% | 1273 | QIZ15717.1 | |
| ~ | Chain A. Spike | glycoprotein [Severe acute res | piratory syndrome core | onavirus 2] | Severe a | cute respiratory | syndron | ne coronaviru. | . 2554 | 2554 | 99% | 0.0 | 100.00% | 1310 | <u>6XR8 A</u> | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe a | cute respiratory | syndron | ne coronaviru. | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QOF12329.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe ad | cute respiratory | syndror | ne coronaviru. | 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QWC77885.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe a | cute respiratory | syndror | ne coronaviru. | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QRN63738.1 | |
| | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe a | cute respiratory | syndror | ne coronaviru. | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QMT96172.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe a | cute respiratory | syndror | ne coronaviru | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QIU80973.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe ad | cute respiratory | syndror | ne coronaviru | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QOF15989.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | 3.2] | Severe a | cute respiratory | syndron | ne coronaviru. | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QMT94564.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe ad | cute respiratory | syndron | ne coronaviru. | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QKU28906.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>[]</u> | Severe a | cute respiratory | syndror | ne coronaviru. | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QOU86714.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe a | cute respiratory | syndror | ne coronaviru. | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QZJ49063.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe ad | cute respiratory | syndror | ne coronaviru. | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QKV35819.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe a | cute respiratory | syndror | ne coronaviru. | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QIA98583.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe a | cute respiratory | syndror | ne coronaviru. | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QIZ14569.1 | |
| | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe a | oute respiratory | syndror | ne coronaviru. | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QIZ16559.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe a | cute respiratory | syndror | ne coronaviru. | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QIU81873.2 | |
| | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe a | cute respiratory | syndror | ne coronaviru. | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QOU93902.1 | |
| | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe a | cute respiratory | syndror | ne coronaviru. | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QOF10625.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe a | cute respiratory | syndror | ne coronaviru. | . 2553 | 2553 | 99% | 0.0 | 100.00% | 1261 | QXF49728.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe a | cute respiratory | syndror | ne coronaviru | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QIU81885.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe a | cute respiratory | syndror | ne coronaviru. | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QWC76579.1 | |
| | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | 5 2] | Severe a | cute respiratory | syndron | ne coronaviru | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QKV06859.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe a | cute respiratory | syndror | ne coronaviru | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QOF08429.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe a | cute respiratory | syndror | ne coronaviru | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QIS61254.1 | |
| ~ | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>s 2]</u> | Severe ad | cute respiratory | syndror | ne coronaviru | . 2553 | 2553 | 99% | 0.0 | 99.92% | 1273 | QNV50022.1 | |
| | surface glycop | rotein [Severe acute respiratory | syndrome coronavirus | <u>[2]</u> | Severe a | cute respiratory | syndron | ne coronaviru. | . 2553 | 2553 | 99% | 0.0 | 100.00% | 1273 | YP 009724390 | 1.1 |

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: <u>BCN86353.1</u> Length: 1282 Number of Matches: 1

Range 1: 21 to 1282 GenPept Graphics

Vext Match 🔺 Pi

| Score 2557 b | oits(662 | Expect 7) 0.0 | Method Compositional | matrix adjust. | Identities 1262/1262(100 | Positives 9%) 1262/1 | 262(100 | Gaps %) 0/12 |
|-----------------|----------|------------------|--|--|----------------------------------|--------------------------|----------------|-----------------|
| Query | 12 | SSOCVNI | TTRTOL PPAYTN | SETRGVYYPDKV | FRSSVLHSTODLFL | | ATHVS | 71 |
| Sbjct | 21 | SSQCVNI | TTRTQLPPAYTN | SFTRGVYYPDKV | FRSSVLHSTQDLFL | PFFSNVTWFF | AIHVS | 80 |
| Query | 72 | GTNGTK | REDNPVLPENDGV | YFASTEKSNIIR YFASTEKSNIIR | GWIFGTTLDSKTQS | | VIKVC | 131 |
| sbjct | 81 | GTNGTK | REDNPVLPENDGV | YFASTEKSNIIR | GWIFGTTLDSKTQS | LLIVNNATN | VIKVC | 140 |
| Query | 132 | EFQECNE | OPFLGVYYHKNNK | SWMESEFRVYSS SWMESEFRVYSS | ANNCTFEYVSQPFL ANNCTFEYVSOPFL | MDLEGKOGNE | KNLRE | 191 |
| Sbjct | 141 | EFQFCN | PFLGVYYHKNNK | SWMESEFRVYSS | ANNCTFEYVSQPFL | MDLEGKQGNF | KNLRE | 200 |
| Query | 192 | FVFKNI | OGYFKIYSKHTPI OGYFKIYSKHTPI | NLVRDLPQGFSA | LEPLVDLPIGINIT | REQTLLALHE | SYLTP | 251 |
| sbjct | 201 | FVFKNI | OGYFKIYSKHTPI | NLVRDLPQGFSA | LEPLVDLPIGINIT | REQTLIALHE | SYLTP | 260 |
| Query | 252 | GDSSSGL | TAGAAAYYVGYL | QPRTFLLKYNEN OPRTFLLKYNEN | GTITDAVDCALDPL GTITDAVDCALDPL | SETKCTLKSF | TVEKG | 311 |
| bjct | 261 | GDSSSGL | TAGAAAYYVGYL | <i><u>QPRTFLLKYNEN</u></i> | GTITDAVDCALDPL | SETKCTLKSF | TVEKG | 320 |
| Query | 312 | IYOTSNE | RVQPTESIVRFP | NITNLCPFGEVF NITNLCPFGEVF | NATREASVYAWNRK | RISNCVADYS | VLYNS VLYNS | 371 |
| bjct | 321 | IYQTSNI | RVQPTESIVRFP | NITNLCPFGEVF | NATRFASVYAWNRK | RISNCVADYS | VLYNS | 380 |
| Query | 372 | ASESTE | | CFTNVYADSFVI CFTNVYADSFVI | RGDEVRQIAPGQTG RGDEVROIAPGOTG | KIADYNYKLF | PDDFTG | 431 |
| bjct | 381 | ASESTE | CYGVSPTKLNDL | CFTNVYADSFVI | RGDEVRQIAPGQTG | KIADYNYKLF | PDDFTG | 440 |
| Query | 432 | CVIAWNS | SNNLDSKVGGNYN | YLYRLFRKSNLK | PFERDISTEIYQAG | STPCNGVEGE | NCYEP | 491 |
| bjct | 441 | CVIAWNS | SNNLDSKVGGNYN | YLYRLFRKSNLK | PFERDISTEIYQAG | STPCNGVEGF | NCYFP | 500 |
| Query | 492 | LOSYGE | 2 PTNGVGYQPYRV 2 PTNGVGYOPYRV | VVLSFELLHAPA | | KCVNENENGL | TGTGV | 551 |
| bjct | 501 | LQSYGFO | PTNGVGYQPYRV | VVLSFELLHAPA | TVCGPKKSTNLVKN | KCVNFNFNGL | TGTGV | 560 |
| Query | 552 | | CELPEQQEGRDIA | DTTDAVRDPQTL DTTDAVRDPOTL | EILDITPCSFGGVS EILDITPCSFGGVS | VITPGTNTS | | 611 |
| bjct | 561 | LTESNK | -LPFQQFGRDIADTTDAVRDPQTLEILDITPC | EILDITPCSFGGVS | VITPGTNTSN | IQVAVL | 620 | |
| Query | 612 | YODVNC | EVPVAIHADQLT | PTWRVYSTGSNV PTWRVYSTGSNV | FQTRAGCLIGAEHV | NNSYECDIPI | GAGIC | 671 |
| bjct | 621 | YQDVNC | TEVPVAIHADQLT | PTWRVYSTGSNV | FÕTRAGCLIGAEHV | NNSYECDIPI | GAGIC | 680 |
| uery | 672 | ASYQTQ | INSPRRARSVASQ | SIIAYTMSLGAE SIIAYTMSLGAE | NSVAYSNNSIAIPT NSVAYSNNSIAIPT | NFTISVTTEI NFTISVTTEI | LPVSM | 731 |
| bjct | 681 | ASYQTQ | INSPRRARSVASQ | SIIAYTMSLGAE | NSVAYSNNSIAIPT | NFTISVTTEI | LPVSM | 740 |
| Query | 732 | TKTSVDO | TMYICGDSTECS | NLLLQYGSFCTQ NLLLQYGSFCTQ | LNRALTGIAVEQDK | NTQEVFAQVE | QIYKT | 791 |
| bjct | 741 | TKTSVDO | TMYICGDSTECS | NLLLQYGSFCTQ | LNRALTGIAVEQDK | NTQEVFAQVE | QIYKT | 800 |
| Query | 792 | PPIKDFO | GFNFSQILPDPS GFNFSQILPDPS | KPSKRSFIEDLL KPSKRSFIEDLL | FNKVTLADAGFIKQ FNKVTLADAGFIKQ | YGDCLGDIAA YGDCLGDIAA | RDLIC | 851 |
| bjct | 801 | PPIKDFO | GENESQILPDPS | KPSKRSFIEDLL | FNKVTLADAGFIKQ | YGDCLGDIAA | RDLIC | 860 |
| Query | 852 | AQKENG | TVLPPLLTDEMI | AQYTSALLAGTI AQYTSALLAGTI | TSGWTFGAGAALQI TSGWTFGAGAALQI | PFAMQMAYRE | NGIGV | 911 |
| bjct | 861 | AQKENG | TVLPPLLTDEMI | AQYTSALLAGTI | TSGWTFGAGAALQI | PFAMQMAYRF | NGIGV | 920 |
| Query | 912 | | NQKLIANQFNSA | IGKIQDSLSSTA IGKIQDSLSSTA | SALGKLQDVVNQNA SALGKLQDVVNQNA | QALNTLVKQL | SSNFG | 971 |
| bjct | 921 | TQNVLYE | ENQKLIANQFNSA | IGKIQDSLSSTA | SALGKLQDVVNQNA | QALNTLVKQL | SSNFG | 980 |
| Query | 972 | AISSVL | DILSRLDKVEAE | VQIDRLITGRLQ VQIDRLITGRLQ | SLQTYVTQQLIRAA SLQTYVTQQLIRAA | EIRASANLAA | TKMSE | 1031 |
| bjct | 981 | AISSVL | DILSRLDKVEAE | VQIDRLITGRLQ | SLQTYVTQQLIRAA | EIRASANLAA | TKMSE | 1040 |
| Query | 1032 | CVLGQSH | CRVDFCGKGYHLM CRVDFCGKGYHLM | SFPQSAPHGVVF SFPQSAPHGVVF | LHVTYVPAQEKNFT LHVTYVPAQEKNFT | TAPAICHDG | AHEPR | 1091 |
| bjct | 1041 | CVLGQSI | RVDFCGKGYHLM | SFPQSAPHGVVF | LHVTYVPAQEKNFT | TAPAICHDG | CAHEPR | 1100 |
| Query | 1092 | EGVEVS | GTHWFVTQRNFY | EPQIITTDNTFV EPQIITTDNTFV | SGNCDVVIGIVNNT SGNCDVVIGIVNNT | VYDPLQPELD | SFKEE | 1151 |
| bjct | 1101 | EGVEVSI | IGTHWEVTQRNEY | EPQIITTDNTFV | SGNCDVVIGIVNNT | VYDPLQPELD | SFKEE | 1160 |
| Query | 1152 | | HTSPDVDLGDIS | GINASVVNIQKE GINASVVNIQKE | IDRLNEVAKNLNES IDRLNEVAKNLNES | LIDLQELGKY | EQYIK | 1211 |
| bjct | 1161 | LDKYFK | HTSPDVDLGDIS | GINASVVNIQKE | IDRLNEVAKNLNES | LIDLQELGKY | EQYIK | 1220 |
| uery | 1212 | WPWYIW | GFIAGLIAIVMV | TIMLCCMTSCCS | clkgccscgscckF CLKGCCSCGSCCKF | DEDDSEPVLK | | 1271 |
| bjct | 1221 | WPWYIWI | GFIAGLIAIVMV | TIMLCCMTSCCS | CLKGCCSCGSCCKF | DEDDSEPVLK | GVKLH | 1280 |
| uery | 1272 | YT 127 YT | 73 | | | | | |
| Sbict | 1281 | YT 128 | 32 | | | | | |

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.

| gi 1958892388 | coprotein [\$ dbj BCN8635 | Severe acut | e respirato | ory syndrom | e coronavi | rus 2] |
|--|---|--|---|--|--|---------|
| d | 50 | 100 | 150 | 200 | 250 | 300 |
| | 353.1 - Find: | 28 | ~ | | 150 | |
| (U) BLAST Re | FVFLVLLPLN nt for group 11 200 sults for: | SSQCVNLTT templates, ref YP_00972 | BLAST Resu | TNSFTRGVYY lts for: ref | PDKV FRSSV YP_00972439 | LHSTQDL |
| BCN86353.1: | | 20 | 30 | 40 | 50 | 60 |
| Contact Co National Center 1 8600 Rockville Pi | MFVFLVLLPLVS LPFFSNVTWFH LDSKTQSLLIVT ANNCTFEYVSQ PQGFSALEPLVI PRTFLLKYNEN VRFPNITNLCPI GVSPTKLNDLC AWNSNNLDSK CYFPLQSYGFQ | SQCVNLTTRTQI IAIHVSGTNGTKI INATNVVIKVCEI IPFLMDLEGKQG GTITDAVDCALD GTITDAVDCALD GEVFNATRFAS FTNVYADSFVIR VGGNYNYLYRLF PTNGVGYQPYR | LPPAYTNSFTRO RFDNPVLPFND FQFCNDPFLGV SNFKNLREFVFK "LLALHRSYLTPO PLSETKCTLKSI VYAWNRKRISN GDEVRQIAPGO FRKSNLKPFERD VV | GVYYPDKVFRSSV GVYFASTEKSNII YYHKNNKSWME NIDGYFKIYSKHT GDSSSGWTAGAA TVEKGIYQTSNF CVADYSVLYNSA QTGKIADYNYKLP ISTEIYQAGSTPO | ALHSTQDLF RGWIFGTT SEFRVYSS TPINLVRDL AYYVGYLQ RVQPTESI SFSTFKCY DDFTGCVI CNGVEGFN | |

Ilustración 22. Hemos seleccionado en el navegador de BLAST de la 1^ª 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

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