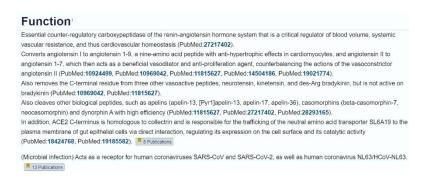
Taller 7- Uniprot

Grupo 1:

- Ana Sofia Villa Benavides 201923361
- Ronald Yesid Diaz Pardo 202111309
- Gabriela Moreno Prieto 201713956
- 1. Utilice UniProtKB para estudiar la proteína ACE2 de humanos, altamente relacionada con la pandemia COVID-19.



a. Acerca de su función:



i. ¿Cuál es su rol principal?

El rol principal de la proteína ACE2 es actuar como una carboxipeptidasa reguladora clave en el sistema renina-angiotensina, controlando el volumen sanguíneo, la resistencia vascular y la homeostasis cardiovascular. En relación al COVID-19, ACE2 actúa como el receptor para el virus SARS-CoV-2, permitiendo su entrada a las células humanas.

ii. ¿Posee actividad catalítica?

Catalytic activity	
Rhea 26554 t3	angiotensin II + H2O = angiotensin-(1-7) + L-phenylalanine
Rhea 63532 t3	angiotensin I + H2O = angiotensin-(1-9) + L-leucine
Rhea 63536 t3	bradykinin(1-8) + H2O = bradykinin(1-7) + L-phenylalanine A 2 Publications This reaction proceeds in the forward direction. A 2 Publications View Rhea reaction ✓
Rhea 63540 t3	H2O + neurotensin = L-leucine + neurotensin-(1-12) Tipublication This reaction proceeds in the forward direction. Tipublication View Rhea reaction >
Rhea 63572 년	H2O + neurotensin-(1-8) = L-arginine + neurotensin-(1-7) Tpublication This reaction proceeds in the forward direction. Tpublication View Rhea reaction >
Rhea 63544 t3	H2O + kinetensin = kinetensin-(1-8) + L-leucine Tpublication This reaction proceeds in the forward direction. This reaction proceeds in the forward direction.

Si posee actividad catalítica en 12 reacciones diferentes.

iii. ¿En qué vías metabólicas interviene?



De acuerdo con la sección de enzimas y vías metabólicas, en específico en relación con reactome estas son las vías

Metabolismo de hormonas peptídicas				
Sistema renina-angiotensina				
Regulación de la presión arterial por señales				
químicas				
Infección por SARS-CoV-2				
Vías de señalización de neuropéptidos				
Vías antihipertensivas				
Vías de vasodilatación				
Homeostasis cardiovascular				

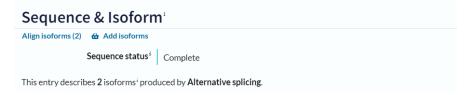
iv. ¿Con qué proteínas interacciona?

Interaction¹ Subunitⁱ Homodimer (PubMed:32132184). Interacts with the catalytically active form of TMPRSS2 (PubMed:21068237). Interacts with SLC6A19: this interaction is essential for expression and function of SLC6A19 in intestine (By similarity) Interacts with ITGA5:ITGB1 (PubMed:15276642, PubMed:33102950). Probably interacts (via endocytic sorting signal motif) with AP2M1; the interaction is inhibited by phosphorylation of Tyr-781 (PubMed:33436498). Interacts (via PDZ-binding motif) with NHERF1 (via PDZ domains); the interaction may enhance ACE2 membrane residence (PubMed:34189428). 🔼 By Similarity (Microbial infection) Interacts with SARS coronavirus/SARS-CoV spike protein. 4 Publications (Microbial infection) Interacts with SARS coronavirus-2/SARS-CoV-2 spike protein (via RBD domain). 10 Publications (Microbial infection) Interacts with human coronavirus NL63 spike protein. It Publication (Microbial infection) Interacts with human coronavirus NL63/HCoV-NL63 spike glycoprotein. 1 Publication Processed angiotensin-converting enzyme 2 (Microbial infection) Interacts with SARS coronavirus-2/SARS-CoV-2 spike protein; the interaction is increased by AVP/Arg-vasopressin with which they may form a complex. 🔼 1 Publication Binary interactions Q9BYF1 has binary interactions with 26 proteins

ACE2 interactúa directamente con 26 proteínas, esto lo pudimos observar en la sección de interacción, particularmente en *binary interaction*. De estas podemos destacar principalmente, que ACE2 interactúa con la proteína de spike NL63 del coronavirus SARS-2/SARS-CoV-2. A continuación, está un pantallazo de la tabla con las 26 proteínas reportadas, aquí se puede ver también información sobre el número de experimentos que respaldan esta interacción.

Select v	ENTRY 1 Q9BYF1 →	ENTRY 2	NUMBER OF EXPERIMENTS	INTACT
BINARY	Q9BYF1	ACE2 Q9BYF1	8	EBI-7730807, EBI-7730807 ☑
BINARY	Q9BYF1	AGT PRO_0000032457 P01019	2	EBI-7730807, EBI-25493366 13
BINARY	Q9BYF1	AGT PRO_0000032458 P01019	5	EBI-7730807, EBI-6622938 ☑
BINARY	Q9BYF1	AP2M1 Q96CW1	2	EBI-7730807, EBI-297683 🖸
BINARY	Q9BYF1	APP PRO_0000000092 P05067	3	EBI-7730807, EBI-821758 🗗
BINARY	Q9BYF1	CLEC4M Q9H2X3	3	EBI-7730807, EBI-1391211 ☑
BINARY	Q9BYF1	DEFA5 PRO_0000417390 Q01523	2	EBI-7730807, EBI-25634589 🗗
BINARY	Q9BYF1	GRM2 Q14416	2	EBI-7730807, EBI-10232876 🗹
BINARY	Q9BYF1	HSPA5 P11021	6	EBI-7730807, EBI-354921 월
BINARY	Q9BYF1	ITGB1 P05556	4	EBI-7730807, EBI-703066 🖸
BINARY	Q9BYF1	KDM1A 060341	19	EBI-7730807, EBI-710124 [2
BINARY	Q9BYF1	KNG1 PRO_0000006688 P01042	5	EBI-7730807, EBI-6623273 🖸
XENO	Q9BYF1	Kpna2 P52293	4	EBI-7730807, EBI-3043908 🖸
BINARY	Q9BYF1	NHERF1 014745	7	EBI-7730807, EBI-349787 ☐
BINARY	Q9BYF1	NTS PRO_0000019524 P30990	2	EBI-7730807, EBI-6655817 🖸
BINARY	Q9BYF1	PDZK1 Q5T2W1	3	EBI-7730807, EBI-349819 🖸
XENO	Q9BYF1	S A0A6B9WHD3	2	EBI-7730807, EBI-25496413 [2
XENO	Q9BYF1	S A0A6G6A1M4	4	EBI-7730807, EBI-26997256 🖸
XENO	Q9BYF1	S A0A6M3G9R1	4	EBI-7730807, EBI-26997195 🖾
XENO	Q9BYF1	S P0DTC2	252	EBI-7730807, EBI-25474821 13
XENO	Q9BYF1	S PRO_0000449647 P0DTC2	2	EBI-7730807, EBI-25490323 13
XENO	Q9BYF1	S P59594	57	EBI-7730807, EBI-15582614 [3
XENO	Q9BYF1	S PRO_0000037209 P59594	3	EBI-7730807, EBI-25475261 12
XENO	Q9BYF1	S Q5GDB5	2	EBI-7730807, EBI-25566398 🖸
XENO	Q9BYF1	S Q6Q1S2	9	EBI-7730807, EBI-15814420 [™]
BINARY	Q9BYF1	SFTPD P35247	2	EBI-7730807, EBI-11316157 년
BINARY	Q9BYF1	SHANK1 Q9Y566	2	EBI-7730807, EBI-3442234 [3
BINARY	Q9BYF1	SLC6A19 Q695T7	4	EBI-7730807, EBI-25475705 ₺
BINARY	Q9BYF1	SNX27 Q96L92	2	EBI-7730807, EBI-2514865 🗹
BINARY	Q9BYF1	TMPRSS2 O15393	3	EBI-7730807, EBI-12549863 [™]
BINARY	Q9BYF1	VIM P08670	4	EBI-7730807, EBI-353844 13

- b. Acerca de su secuencia y estructura:
 - i. ¿Cuántas isoformas se han identificado?



Se han identificado 2 isoformas.

ii. ¿Cuántos sitios activos posee?



Posee 2 sitios activos, uno en la posición 375 y otro en la 505. Ambos están descritos como aceptores de protones.

iii. Entre las posibles variantes, ¿Hay alguna/s que inhiban la interacción con la proteína Spike de SARS-CoV-2? Si lo hay, cite la bibliografía correspondiente



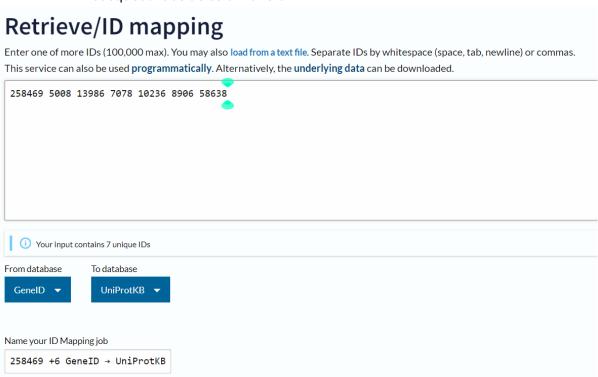
De todas las posibles variantes recopilamos las que mencionaban inhibición de la interacción con la proteína spike. Encontramos que hay algunas variantes que ligeramente inhiben, otras que lo hacen fuertemente y una sola que indica inhibir sin indicar la fuerza de la inhibición. Todas las variantes de este tipo están soportadas por un mismo artículo de Li et al., (2005). Este se llama" Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2". Sin embargo, ninguna de estas variantes hace referencia al SARS-CoV-2 en particular si no al SARS-CoV.

Hay otras variantes que si mencionan puntualmente al SARS-CoV-2, sin embargo, estas no inhiben y en cambio pueden aumentar ligeramente la interacción. Estas variables todas están reportadas por otro artículo: "Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2" (Chan et al., 2020).



- 2. Dada la siguiente lista de identificadores Entrez ID utilice la herramienta de mapeo de UniProt: 258469 5008 13986 7078 10236 8906 58638
 - a. Identifique los Entrez ID que no fueron mapeados a proteínas en UniProt. ¿Cuántos son?

Para realizar la búsqueda de estos IDs se ha usado la herramienta de ID Mapping de Uniprot provista en el siguiente link. Allí se ingresaron los Ids y se realizó inicialmente la búsqueda con los valores por defecto de UnitProt, en este caso por defecto la base de datos escogida es UniProtKB AC/ID y UniprotKB, no obstante, dentro de esta base de datos no se hallaron resultados que coincidan con lo que se busca, así se cambió la base de datos 1 a la de Genome Annotation Databases /GeneID, esto debido a que se identifica que estos Ids hacen referencias a IDs provistos en NCBI, específicamente Gene Id, por lo cual hacer la búsqueda en esta base de datos nos dan resultados adecuados, dejando la base de datos #2 como UniProtKB. De acuerdo con lo mencionado la query de búsqueda luce de esta manera:



Finalmente, como resultado de esta búsqueda se encontró que en total 2 IDs no fueron mapeados: 58638 y 13986, y de los que si fueron mapeados se obtuvieron 13 resultados en total, el acceso a este trabajo puedo ser obtenido en este <u>link</u>.

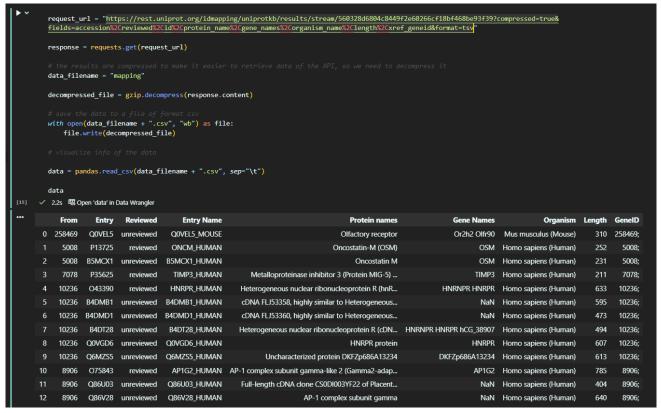
- b. Genere un archivo (.csv o .tab) que contenga una tabla de todas las proteínas mapeadas con las siguientes columnas:
 - i. UniProt ID, UniProt name, Protein name, Gene name, Organism, Entrez ID.

Para obtener los datos en un formato de .csv o .tab se identificó que UniProt no ofrece la posibilidad de descargar los datos en estos formatos, es por ello por lo que se ha desarrollado una herramienta de acceso programático que accede al API de UniProt con la URL generada de una búsqueda con las tablas requeridas:

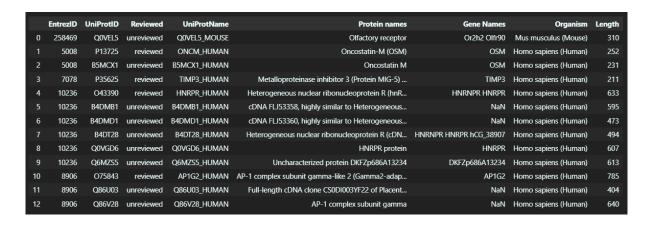
URL de request: https://rest.uniprot.org/idmapping/uniprotkb/results/stream/560328d6804c844 9f2e60266cf18bf468be93f39?compressed=true&fields=accession%2Creviewe d%2Cid%2Cprotein_name%2Cgene_names%2Corganism_name%2Clength%2Cxref_geneid&format=tsv

Tener en cuenta que este acceso programático se está haciendo con el valor de parámetro de compressed = True, por lo que los datos están comprimidos y requieren ser descomprimidos al obtenerlos para poder visualizarlos correctamente y no en lenguaje binario.

Código de acceso programático con archivo resultante mapping.csv visualizado por medio de Pandas:

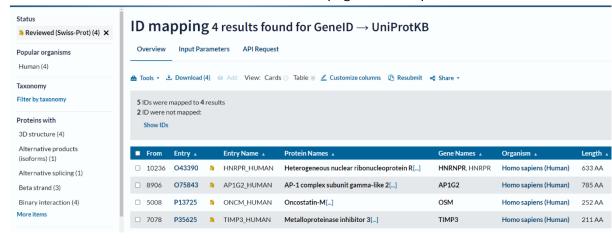


No obstante, se requiere de un refinamiento de las columnas para adecuar los nombres de estas las columnas que se piden, por lo cual se obtiene la siguiente tabla luego del refinamiento de acuerdo con las columnas de datos que se solicitan:

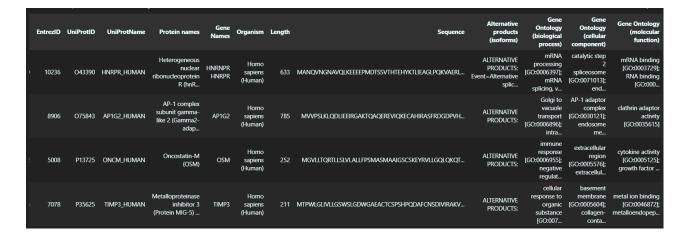


- c. Genere un archivo (.csv o .tab) que contenga una tabla sólo de las proteínas curadas con la siguiente información:
 - UniProt ID, UniProt name, Protein name, Gene name, Organism, Entrez ID, Sequence, Length, Alternative products, Anotaciones en GO-BP, GO-CC, GO-MF.

Para descargar los datos que corresponden a proteínas curadas se ha seleccionado la sección de Reviewed dentro de la página de búsqueda:



Aquí dentro, se extrajo la URL de API para realizar la búsqueda con las columnas seleccionadas, por lo cual se obtuvo el siguiente resultado luego de filtrar:



El acceso a las herramientas programáticas usadas en el código de cuaderno Jupyter y a los archivos resultantes .csv puede ser encontrado en nuestro repositorio GitHub en este link.

- 3. Análisis a partir de la secuencia de AA provista,
 - a. Realice un BLAST para determinar las proteínas similares a la secuencia objetivo.

Las proteínas similares son las denominadas: Glicoproteína de espiga, glicoproteína S, E2, proteína peplómero, escindida en proteína de espiga S1, proteína de espiga S2, proteína de espiga S2'



b. Identifique a qué proteína es más probable que pertenezca la secuencia provista.

La proteína con la que tiene mayor identidad es la denominada P0DTC2 con una identidad del 98.9%



c. Determine si las diferencias entre la secuencia provista y la proteína más similar ocurren en features de interés

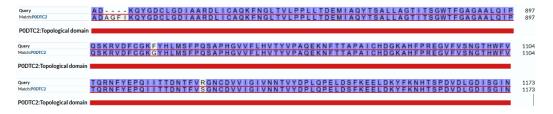
Al resaltar el BLAST de las dos secuencias por dominio se observa una zona donde se cambian 3 aminoácidos. Un aminoácido cerca a los 296 y dos aminoácidos seguidos cerca de los 518. Cuando las secuencias de proteínas tienen diferencias en la zona de un dominio, esto puede tener implicaciones importantes, ya que los dominios proteicos son las regiones estructurales y funcionales clave de una proteína.



También, al resaltar el BLAST por mutagénesis se observa una zona cerca a los 720 aminoácidos donde hay una deleción de 4 aminoácidos. Al revisar las anotaciones de esta zona se observa que está asociado a: la perdida completa de la O-glicosilación por parte del huésped, pérdida completa de la escisión por la furina del huésped, mejora parcial de la escisión por la furina del huésped y aproximadamente un 50% más de eficiencia de entrada en células Vero E6, mientras que un 30% menos de eficiencia de entrada en células BHK que expresan hACE2.



Por último, al resaltar el BLAST por topological domain está asociado a la parte extracelular y se observa que va desde el aminoácido 12 hasta el 1211. Donde están incluidas las diferencias ya mencionadas y otras adicionales como la deleción de 4 aminoácidos cerca de los 897 y el cambio de dos aminoácidos por los 1104 y 1173 aminoácidos. Estos cambios pueden tener cambios significativos en la proteína como: la interacción con otras moléculas o señales y cambiar la orientación de la proteína dentro de la membrana para su correcta localización.



MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLP
FFSNVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKT
QSLLIVNNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYV
SQPFLMDLEGKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVD
LPIGINITRFQTLHALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTI
TDAVDCALDPLSETKCTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVF
NATRFASVYAWNRKRISNCVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADS
FVIRGDEVRQIAPGQTGKIADYNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLF
RKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLYRYGFQPTNGVGYQPYRVVVL
SFELLHAPATVCGPKKSTNLVKNKCLNFNFNGLTGTGVLTESNKKFLPFQQFGRDIA
DTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHAD
QLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASYQTQTNSRSVA

SQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGD
STECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNF
SQILPDPSKPSKRSFIEDLLFNKVTLADKQYGDCLGDIAARDLICAQKFNGLTVLPP
LLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQ
KLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVL
NDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVL
GQSKRVDFCGKFYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPR
EGVFVSNGTHWFVTQRNFYEPQIITTDNTFVRGNCDVVIGIVNNTVYDPLQPELDSF
KEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGK
YEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCS

Referencias

Li, W., Zhang, C., Sui, J., Kuhn, J. H., Moore, M. J., Luo, S., Wong, S. K., Huang, I. C., Xu, K., Vasilieva, N., Murakami, A., He, Y., Marasco, W. A., Guan, Y., Choe, H., & Farzan, M. (2005). Receptor and viral determinants of SARS-coronavirus adaptation to human ACE2. *The EMBO journal*, *24*(8), 1634–1643. https://doi.org/10.1038/sj.emboj.7600640

Chan, K. K., Dorosky, D., Sharma, P., Abbasi, S. A., Dye, J. M., Kranz, D. M., Herbert, A. S., & Procko, E. (2020). Engineering human ACE2 to optimize binding to the spike protein of SARS coronavirus 2. *Science (New York, N.Y.)*, 369(6508), 1261–1265. https://doi.org/10.1126/science.abc0870