UNIVERSIDAD AUTÓNOMA DE SAN LUIS POTOSÍ

FACULTAD DE CIENCIAS QUÍMICAS

Programa de Posgrado en Ciencias Farmacobiológicas

“**Predicción y evaluación de epítopes de células T potencialmente inmunogénicos contra el coronavirus SARS-CoV-2 mediante estudios de vacunología reversa”**

Tesis para obtener el grado de

Doctor en Ciencias Farmacobiológicas

Presenta:

Aguayo Martinez Elsa Yamelie

Directora de Tesis:

Dra. Diana Patricia Portales Pérez

Codirector de Tesis:

Edgar E. Lara Ramírez

SAN LUIS POTOSÍ, S.L.P Diciembre 2024

Este proyecto fue realizado en Laboratorio de Inmunología y Biología Celular y Molecular, Facultad de Ciencias Químicas adscrito a la Universidad Autónoma de San Luis Potosí, San Luis Potosí, México y la “Unidad de Investigación Biomédica IMSS Zacatecas”, en el periodo comprendido entre agosto 2020 y agosto 2024.

El programa de Doctorado en Ciencias Farmacobiológicas de la Universidad Autónoma de San Luis Potosí pertenece al Programa Nacional de Posgrados de Calidad (PNPC) del CONAHCyT, registro 003383, en el nivel 291236.

Numero de registro de beca otorgada por CONAHCyT/CVU: 934987

Los datos del trabajo titulado “Predicción y evaluación de epítopos de células T potencialmente inmunogénicos contra el coronavirus SARS-CoV-2 mediante estudios de vacunología reversa” se encuentran bajo el resguardo de la Facultad de Ciencias Químicas y pertenecen a la Universidad de San Luis Potosí.

UNIVERSIDAD AUTÓNOMA DE SAN LUIS POTOSÍ
Facultad de Ciencias Químicas
Centro de Investigación y Estudios de Posgrado
Posgrado en Ciencias Farmacobiológicas
Programa de Doctorado
Formato D13

Aprobación de Tema de Tesis

San Luis Potosí SLP a 22 de noviembre del 2024

Comité Académico

La presente es para que quede asentado que el tema de Tesis de Doctorado: "Predicción y evaluación de epítopos de células T potencialmente inmunogénicos contra el coronavirus SARS-CoV-2 mediante estudios de vacunología reversa" del estudiante: Elsa Yamelie Aguayo Martinez que se llevará a cabo en el laboratorio de Inmunología y Biología Celular y Molecular, Facultad de Ciencias Químicas, Universidad Autónoma de San Luis Potosí, San Luis Potosí, México y la Unidad de Investigación Biomédica IMSS- Zacatecas, es APROBADO.

Sin más por el momento, quedo de Uds.

A T E N T A M E N T E

Dr. Sergio Zarazúa Guzmán
Coordinador del Posgrado

UNIVERSIDAD AUTÓNOMA DE SAN LUIS POTOSÍ

FACULTAD DE CIENCIAS QUÍMICAS

Programa de Posgrado en Ciencias Farmacobiológicas

"**Predicción y evaluación de epítopos de células T potencialmente inmunogénicos contra el coronavirus SARS-CoV-2 mediante estudios de vacunología reversa"**

Tesis para obtener el grado de Doctor en Ciencias Farmacobiológicas

Presenta:

Aguayo Martinez Elsa Yamelie

Presidente: Dr. Sergio Zarazúa Guzmán

Secretario: Dr. Fidel Martinez Gutiérrez

Vocal: Dra. Diana Patricia Portales Pérez

Vocal: Dr. Edgar E. Lara Ramírez

Vocal: Dr. Julio E. Castañeda Delgado

Diciembre 2024

INTEGRANTES DEL COMITÉ TUTORIAL ACADÉMICO

Dra. Diana Patricia Portales Pérez. Adscrita al Laboratorio de Inmunología y Biología Celular y Molecular, Facultad de Ciencias Químicas, Universidad Autónoma de San Luis Potosí, San Luis Potosí, México.

Dr. Edgar E. Lara Ramírez. Adscrito al Laboratorio de Biotecnología Farmacéutica, Centro de Biotecnología Genómica, Instituto Politécnico Nacional, Reynosa, Tamaulipas, México

Dr. Fidel Martínez Gutiérrez. Adscrito al Laboratorio de Antimicrobianos, Biopelículas y Microbiota, Centro de Investigación en Ciencias de la Salud y Biomedicina (CICSaB), Facultad de Ciencias Químicas, Universidad Autónoma de San Luis Potosí, San Luis Potosí, México.

Dr. Sergio Zarazúa Guzmán. Adscrito al Laboratorio de Neurotoxicología, Facultad de Ciencias Químicas, Universidad Autónoma de San Luis Potosí, San Luis Potosí, México

Dr. Julio Enrique Castañeda Delgado. Adscrito a la Unidad de Investigación Biomédica de Zacatecas, Instituto Mexicano Del Seguro Social, Zacatecas, México; Cátedras-CONACYT, Unidad de Investigación Biomédica de Zacatecas, Instituto Mexicano Del Seguro Social, Zacatecas, México.

UNIVERSIDAD AUTÓNOMA DE SAN LUIS POTOSÍ
Facultad de Ciencias Químicas
Centro de Investigación y Estudios de Posgrado
Posgrado en Ciencias Farmacobiológicas
Programa de Doctorado
Formato D5

Carta Cesión de Derechos

San Luis Potosí SLP a 22 de noviembre del 2024

En la ciudad de San Luis Potosí el día 22 del mes de Noviembre del año 2024, el que suscribe Elsa Yamelie Aguayo Martinez Alumno(a) del programa de posgrado Doctorado en Ciencias Farmacobiológicas adscrito a la Facultad de Ciencias Químicas manifiesta que es autora intelectual del presente trabajo terminal, realizado bajo la dirección de: la Dra. Diana Patricia Portales Pérez y el Dr. Edgar E. Lara Ramírez y cede los derechos del trabajo titulado "Predicción y evaluación de epítopos de células T potencialmente inmunogénicos contra el coronavirus SARS-CoV-2 mediante estudios de vacunología reversa" a la Universidad Autónoma de San Luis Potosí, para su difusión con fines académicos y de investigación.

Los usuarios de la información no deben reproducir de forma total o parcial texto, gráficas, imágenes o cualquier contenido del trabajo si el permiso expreso del o los autores. Éste, puede ser obtenido directamente con el autor o autores escribiendo a la siguiente dirección *eyamelie.amtz@gmail.com; doc_lara_ram@hotmail.com; dportale@uaslp.mx*. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

_____

Elsa Yamelie Aguayo Martinez

UNIVERSIDAD AUTÓNOMA DE SAN LUIS POTOSÍ
Facultad de Ciencias Químicas
Centro de Investigación y Estudios de Posgrado
Posgrado en Ciencias Farmacobiológicas
Programa de Doctorado

Formato D28

Carta de Análisis de Similitud

San Luis Potosí SLP a 22 de noviembre de 2024

L.B. María Zita Acosta Nava
Biblioteca de Posgrado FCQ

Asunto: Reporte de porcentaje de similitud de tesis de grado

Por este medio me permito informarle el porcentaje de similitud obtenido mediante Ithenticate para la tesis titulada "**Predicción y evaluación de epítopos de células T potencialmente inmunogénicos contra el coronavirus SARS-CoV-2 mediante estudios de vacunología reversa"** presentada por el autor Elsa Yamelie Aguayo Martinez. La tesis es requisito para obtener el grado de Doctorado en el Posgrado en Ciencias Farmacobiológicas. El análisis reveló un porcentaje de similitud de Porcentaje de Similitud 11 % excluyendo referencias y metodología.

Agradezco sinceramente su valioso tiempo y dedicación para llevar a cabo una exhaustiva revisión de la tesis. Quedo a su disposición para cualquier consulta o inquietud que pueda surgir en el proceso.

Sin más por el momento, le envío un cordial saludo.

A T E N T A M E N T E

_____

Coordinador Académico del Posgrado
Dr. Sergio Zarazúa Guzmán

El conocimiento es un viaje continuo, y cada descubrimiento es un paso más hacia la comprensión de lo desconocido.

AGRADECIMIENTOS

Quiero expresar mi más profundo agradecimiento a la Universidad Autónoma de San Luis Potosí (UASLP), institución que ha sido fundamental en mi formación académica y profesional. A lo largo de este camino, encontré un espacio de crecimiento, aprendizaje y desarrollo personal, por lo que siempre estaré orgulloso de ser parte de esta comunidad universitaria.

Extiendo mi gratitud a la Unidad de Investigación Médica del Instituto Mexicano del Seguro Social (IMSS)-Zacatecas, que ha sido mi segundo hogar académico desde la licenciatura hasta este último paso en el doctorado.

Al Consejo Nacional de Humanidades Ciencia y Tecnología (CONAHCYT), quiero expresar mi reconocimiento y agradecimiento por el apoyo financiero y académico brindado durante mis estudios. Su respaldo fue fundamental para que este proyecto pudiera llevarse a cabo, y su contribución al desarrollo de la ciencia en México es invaluable.

Finalmente, agradezco a mi familia, amigos y colegas que estuvieron conmigo en los momentos de esfuerzo, dudas y logros. Cada uno de ustedes dejó una huella en este trabajo, y por ello les dedico este logro con profunda gratitud.

De manera especial, agradezco a mi familia, quienes han sido mi mayor fortaleza y motivación. A mi mamá, cuyo amor incondicional, esfuerzo y sabiduría siempre me han guiado; a mis abuelos, por su apoyo constante, sus palabras alentadoras y por ser un ejemplo de perseverancia y dedicación; a mi papá, por su confianza en mí y su cariño inquebrantable; y a mi hermano, quien con su compañía y ánimo constante hizo más llevaderos los momentos difíciles. Su presencia en mi vida ha sido fundamental para alcanzar este logro.

A mis amigos, les agradezco por cada palabra de aliento, cada vez que me escucharon y estuvieron a mi lado durante este esfuerzo. Sus muestras de apoyo, su paciencia y su confianza en mí me impulsaron a seguir adelante incluso en los momentos más desafiantes.

Finalmente, gracias a todos los que formaron parte de este proceso, directa o indirectamente. Cada uno de ustedes dejó una huella invaluable en este trabajo, y por ello les dedico este logro con profundo aprecio y gratitud. ¡Muchas gracias!

PLOS ONE

RESEARCH ARTICLE

The analysis on the human protein domain targets and host-like interacting motifs for the MERS-CoV and SARS-CoV/CoV-2 infers the molecular mimicry of coronavirus

Yamelie A. Martínez[1,2], Xianwu Guo[3], Diana P. Portales-Pérez[2], Gildardo Rivera[4], Julio E. Castañeda-Delgado[1,5], Carlos A. García-Pérez[6], José A. Enciso-Moreno[1], Edgar E. Lara-Ramírez[1]*

1 Unidad de Investigación Biomédica de Zacatecas, Instituto Mexicano Del Seguro Social, Zacatecas, México, 2 Laboratorio de Inmunología y Biología Celular y Molecular, Facultad de Ciencias Químicas, Universidad Autónoma de San Luis Potosí, San Luis Potosí, México, 3 Laboratorio de Biotecnología Genómica, Centro de Biotecnología Genómica, Instituto Politécnico Nacional, Reynosa, México, 4 Laboratorio de Biotecnología Farmacéutica, Centro de Biotecnología Genómica, Instituto Politécnico Nacional, Reynosa, México, 5 Cátedras-CONACYT, Unidad de Investigación Biomédica de Zacatecas, Instituto Mexicano Del Seguro Social, Zacatecas, México, 6 Information and Communication Technology Department (ICT), Complex Systems, Helmholtz Zentrum München, Neuherberg, Germany

* doc_lara_ram@hotmail.com

“**Predicción y evaluación de epítopos de células T potencialmente inmunogénicos contra el coronavirus SARS-CoV-2 mediante estudios de vacunología reversa”**

Resumen: Los coronavirus: MERS-CoV, SARS-CoV y el SARS-CoV-2 son virus altamente patógenos que causan enfermedades respiratorias graves en humanos. Las interacciones Dominio-motivo entre proteínas son el medio esencial por el cual los virus imitan los procesos biológicos y secuestran la maquinaria de replicación de la célula huésped para sobrevivir. Las interacciones entre proteínas pueden analizarse identificando las expresiones regulares que corresponden con los motivos; en este estudio se realizó minería de datos para mapear interacciones Dominio-motivo en los proteomas de los coronavirus. Se encontró que los motivos expresados en proteínas virales que interactúan con proteínas del huésped se encuentran compartidos y conservados entre los tres coronavirus, indicando que el mimetismo molecular es un mecanismo común en la infección por coronavirus. Además, se realizó un análisis para determinar la ontología de genes, demostrando que los Dominios identificados participan en los procesos relacionados con el uso de fuentes de carbono (proteína N) y regulación de canales de potasio (proteína S). La identificación de estos motivos como potenciales epítopes demostró que en su mayoría estos epítopes se ubican en las proteínas espiga y nucleocápside. Además, estos epítopes se conservan entre los 3 beta coronavirus y son reconocidos por células T CD8+. Este estudio demuestra que la minería de datos enfocada en la identificación de interacciones Dominio-motivo es una estrategia de vacunología reversa interesante para el diseño racional de nuevos tratamientos contra patógenos virales.

Palabras clave: SARS-CoV-2, vacunas, epítope, Coronavirus, interacciones Dominio-motivo

“Prediction and evaluation of potentially immunogenic T cell epitopes against the SARS-CoV- 2 coronavirus using reverse vaccinology studies”

Abstract:

Coronaviruses: MERS-CoV, SARS-CoV, and SARS-CoV-2 are highly pathogenic viruses that cause severe respiratory diseases in humans. Although there are therapeutic strategies for these infections, there is still no specific strategy to eliminate them. Domain-motif interactions between proteins are the essential means by which viruses mimic biological processes and hijack the replication machinery of the host cell to survive. Protein interactions can be analyzed by identifying regular expressions that correspond to motifs; in this study, data mining was conducted to map Domain-motif interactions in the coronavirus proteomes. It was found that motifs expressed in viral proteins interacting with host proteins are shared and conserved among the three coronaviruses, indicating that molecular mimicry is a common mechanism in coronavirus infection. Additionally, an analysis was conducted to determine gene ontology, demonstrating that the identified Domains are involved in processes related to the use of carbon sources (N protein) and regulation of potassium channels (S protein). Identifying these motifs as potential epitopes showed that most of these epitopes are located in the spike and nucleocapsid proteins. Furthermore, these epitopes are conserved among the three beta-coronaviruses and are recognized by CD8+ T cells, as indicated by molecular docking analysis of peptides with recurrent class I HLA receptor haplotypes in SARS-CoV-2 infection. Therefore, this study demonstrates that data mining focused on identifying Domain-motif interactions is an interesting reverse vaccinology strategy for the rational design of new treatments against viral pathogens.

Key words: SARS-CoV-2, vaccine, epitope, Coronavirus, Domain-motif interaction

CONTENIDO

1. INTRODUCCIÓN

El SARS-CoV-2 es el agente etiológico de la enfermedad respiratoria COVID-19. Este virus pertenece a la familia de los virus Coronaviridae, de la cual forman parte también el MERS-CoV y SARS-CoV (Lu et al., 2020), virus causantes de otras enfermedades respiratorias de carácter endémico y pandémico respectivamente; estos tres coronavirus tienen un origen zoonótico y han demostrado ser de relevancia epidemiológica (Arslan, 2021; Forni et al., 2017; Mousavizadeh & Ghasemi, 2021), siendo el SARS-CoV-2 el agente infeccioso de mayor relevancia debido a la pandemia de COVID-19. El surgimiento de la enfermedad de COVID-19 marcó una línea de partida para la propuesta y desarrollo de nuevas estrategias de análisis para identificar y desarrollar alternativas para combatir enfermedades infecciosas, desde el diseño de nuevos tratamientos hasta el desarrollo de vacunas.

El diseño y desarrollo de estos nuevos tratamientos y vacunas toma como base el estudio de las interacciones entre el huésped y el patógeno. El estudio de estas interacciones se enfoca en las interacciones proteína – proteína (IPPs). A inicios de la pandemia se realizaron análisis sobre las IPPs del SARS-CoV-2 para predecir cómo son las interacciones entre este virus y el huésped (Diella et al., 2008; Gordon et al., 2020; Perrin-Cocon et al., 2020). Sin embargo, poco se sabe sobre las interacciones entre Dominio y motivo. Los Dominios son las unidades funcionales de las proteínas, las cuales participan en vías de señalización al interior de la célula (Basu et al., 2009); su longitud es de aproximadamente 200 aminoácidos, y sus patrones de plegamiento son independientes del resto de la proteína(Lin & Zewail, 2012); en contraste, los motivos, son pequeñas secuencias lineales de entre 3 a 15 aminoácidos; los virus interactúan con el huésped en su mayoría mediante interacciones Dominio – motivo para realizar procesos de señalización celular que hacen posible la supervivencia y replicación del virus(Brito & Pinney, 2017; Garamszegi et al., 2013; Itzhaki, 2011).

Nuestro estudio se enfoca en la identificación de motivos compartidos entre los tres beta coronavirus: MERS-CoV, SARS-CoV y SARS-CoV-2, así como su relación ontológica en procesos celulares, y analizando el potencial de estos motivos como epitopes conservados que sean relevantes en la respuesta de células T mediante estudios de acoplamiento molecular con receptores HLA relevantes en la infección por SARS-CoV-2, contribuyendo así al análisis de nuevas estrategias terapéuticas contra estos virus en el futuro.

2. ANTECEDENTES

Los coronavirus pueden causar diferentes enfermedades respiratorias, como el Síndrome Respiratorio Agudo Grave (SARS), el Síndrome Respiratorio de Oriente Medio (MERS) y la reciente pandemia de COVID-19, causada por el nuevo coronavirus SARS-CoV-2. Este virus pertenece a la familia Coronaviridae, incluida en el género beta-Coronavirus (Lu et al., 2020). El genoma viral del SARS-CoV-2 incluye dos principales marcos de lectura abiertos (ORFs, por sus siglas en inglés). Dos tercios del genoma viral, localizados en el primer ORF (ORF1a/b), codifican dos poliproteínas, pp1a y pp1ab, y dieciséis proteínas no estructurales (nsp); mientras que los ORFs restantes codifican proteínas accesorias y estructurales conocidas como: glucoproteína de espícula (S), envoltura (E), membrana (M) y nucleocápside (N)(Mariano et al., 2020; Song et al., 2019). Entre las proteínas estructurales codificadas en los coronavirus, la proteína N es la más producida en las células infectadas y su función principal es participar en la formación de los nuevos viriones producidos durante la infección(McBride et al., 2014). En 2019, un nuevo virus se propagó por China y por todo el mundo; este virus fue identificado inicialmente como 2019-nCoV y luego renombrado como SARS-CoV-2 por el Comité Internacional de Taxonomía de Virus debido a su similitud genética con el coronavirus previamente conocido, SARS-CoV. El SARS-CoV-2 es el agente etiológico de un síndrome respiratorio clínico conocido como enfermedad por coronavirus 2019 (COVID-19). En marzo de 2020, la Organización Mundial de la Salud declaró oficialmente la

epidemia de COVID-19 como una emergencia de salud pública de preocupación internacional.

Actualmente, no existe un tratamiento específico para la COVID-19 y las vacunas disponibles todavía están bajo observación. Además, estas vacunas están principalmente diseñadas para activar la respuesta inmune del huésped a través de la proteína de espícula como objetivo antigénico. Sin embargo, el SARS-CoV-2 continúa evolucionando y mutando en la proteína de espícula, lo que ha reducido la efectividad de las vacunas actuales debido a la especificidad fina de los anticuerpos.

3. JUSTIFICACIÓN

El SARS-CoV-2 representa una problemática mundial con implicaciones significativas tanto a nivel de salud como económico. Actualmente, no existe un tratamiento específico para esta enfermedad, y las vacunas disponibles se encuentran en una fase de observación constante. Adicionalmente, el surgimiento de nuevas variantes del virus plantea un desafío, ya que podría reducir la eficacia de las vacunas existentes.

En este contexto, la identificación y el estudio de epítopes altamente inmunogénicos que puedan participar en las respuestas de inmunidad celular generadas por el SARS-CoV-2 se convierten en una prioridad científica. Este enfoque podría proporcionar una base teórica sólida para continuar la experimentación *in vitro* y en modelos animales, facilitando así el posible desarrollo de nuevas vacunas más efectivas y adaptables a la evolución del virus.

4. HIPÓTESIS

Existen epitopes conservados con potencial inmunogénico entre los Coronavirus MERS-CoV, SARS-CoV y SARS-CoV-2 y éstos participan en interacciones Dominio-motivo.

5. OBJETIVO

Identificar vía *in silico* epítopes inmunogénicos contra el SARS-CoV-2, relevantes en la respuesta de células T CD8+, para evaluar su capacidad inmunogénica en modelos de acoplamiento molecular.

6. ARTICULO EXPERIMENTAL

The analysis on the human protein domain targets and host-like interacting motifs for the MERS-CoV and SARS-CoV/CoV-2 infers the molecular mimicry of coronavirus.

Yamelie A. Martínez[1,2], Xianwu Guo[3], Diana P. Portales-Pérez[2], Gildardo Rivera[4], Julio E. Castañeda-Delgado[1,5], Carlos A. García-Pérez[6], José A. Enciso-Moreno[1], Edgar E. Lara-Ramírez[1]*

[1] Unidad de Investigación Biomédica de Zacatecas, Instituto Mexicano Del Seguro Social, Zacatecas, México.

[2] Laboratorio de Inmunología y Biología Celular y Molecular, Facultad de Ciencias Químicas, Universidad Autónoma de San Luis Potosí, San Luis Potosí, México.

[3] Laboratorio de Biotecnología Genómica, Centro de Biotecnología Genómica, Instituto Politécnico Nacional, Reynosa, México.

[4] Laboratorio de Biotecnología Farmacéutica, Centro de Biotecnología Genómica, Instituto Politécnico Nacional, Reynosa, México.

[5] Cátedras-CONACYT, Unidad de Investigación Biomédica de Zacatecas, Instituto Mexicano Del Seguro Social, Zacatecas, México.

[6] Information and Communication Technology department (ICT), Complex Systems, Helmholtz Zentrum München, Neuherberg, Germany.

* Corresponding author: doc_lara_ram@hotmail.com; (E.E.L.R)

ABSTRACT

The MERS-CoV, SARS-CoV, and SARS-CoV-2 are highly pathogenic viruses that can cause severe pneumonic diseases in humans. Unfortunately, there is a non-available effective treatment to combat these viruses. Domain-motif interactions (DMIs) are an essential means by which viruses mimic and hijack the biological processes of host cells. To disentangle how viruses achieve this process can help to develop new rational therapies. Data mining was performed to obtain DMIs stored as regular expressions (regexp) in 3DID and ELM databases. The mined regexp information was mapped on the coronaviruses' proteomes. Most motifs on viral protein that could interact with human proteins are shared across the coronavirus species, indicating that molecular mimicry is a common strategy for coronavirus infection. Enrichment ontology analysis for protein domains showed a shared biological process and molecular function terms related to carbon source utilization and potassium channel regulation. Some of the mapped motifs were nested on B, and T cell epitopes, suggesting that it could be as an alternative way for reverse vaccinology. The information obtained in this study could be used for further theoretic and experimental explorations on coronavirus infection mechanism and development of medicines for treatment.

Key words: Data mining, coronaviruses, domain-motif interactions, reverse vaccinology, molecular mimicry.

INTRODUCTION

Coronaviruses (CoV) are enveloped single-stranded, positive-sense RNA viruses, responsible very often for mild upper respiratory infections in humans. Nevertheless, remarkably pathogenic CoVs to humans have been reported. The first one appeared in 2003 in Guangdong, China, leading to an epidemic of severe acute respiratory syndrome (SARS) and this virus was named SARS-CoV [1]. In 2012, another CoV arose in Middle Eastern countries, causing pneumonic syndrome, called MERS-CoV [2]. At the end of 2019, a new CoV emerged in Wuhan, China, causing severe pneumonia [3] and was named SARS-CoV-2 due to its genomic similarity with the

past SARS-CoV [4]. This is the first CoV that caused a pandemic disease termed COVID-19. These three CoVs are zoonotic, and its primary origin was traced to bats and other animals [4,5]. We are still suffering from SARS-CoV-2. This is a serious public health concern, especially for the aged people with increased risk for complications such diabetes mellitus (DM), hypertension, and severe obesity, which cause the high morbidity-mortality rates of COVID-19 [6]. Humans infected by SARS-CoV-2 could be also asymptomatic, but they may transmit the virus [6]. Although numerous efforts are currently underway to develop drugs and vaccines to combat those viruses, there is no effective treatment available yet.

The study on molecular interactions of host-pathogen helps to find new targets for drug discovery or antigens for vaccine development. Host-pathogen relation is mainly explored through protein-protein interaction (PPI) studies. These studies can be experimentally and computationally aided [7]. The computational studies could be preliminary but quick to guide the rational selection of data for experimental confirmations. Experimental approaches have been carried out for SARS-CoV, MERS-CoV [8,9], and recently for SARS-CoV-2 [10]. A detailed literature mining that surveys experimental and predicted PPIs for several coronaviruses, including the viruses studied herein, was recently published [11]. Also, several computation-aided researches focused on predicting PPI of host and SARS-CoV-2 [7,12,13]. Such predictions provided valuable information to help the rational design of treatments against these viral infections.

However, the analysis of domain-motif interaction (DMI) has paid less attention to those CoVs. Domains in proteins are the functional units involved in the signaling networks within a cell [14]. Its length is up to 200 amino acids, and its folding patterns are independent of the rest of the whole protein [15]. In contrast, motifs are short plastic linear sequences with a length of 3 to 15 amino acids. DMIs are the preferential molecular mechanism by which viruses interact with host cells [16]. Motifs are employed by the viruses to mimic and hijack the host cell's essential process for its survival [17]. Currently, two studies have approached the role of motifs present on essential host proteins for SARS-CoV-2 infection. The research of

Mészáros et al . [18] consisted in the prediction of motifs retrieved from Eukaryotic Linear Motif (ELM) resource that were mapped onto the angiotensin-converting enzyme 2 (ACE2) and integrins of the human host. They found conserved motifs on the cytoplasmatic tails of ACE2 and integrin β3 that interacts with several critical regulatory protein domains. This motif information was tested later on experimental binding affinity measurements [19] and found that NHERF3 PDZ1, SHANK1 and SNX27 PDZ domains bind to synthetic peptides of the ACE2, and to the synthetic ATG8 domains, MAP1LC3s and GABARAPs, of integrin β3. Those studies exemplify the utility of motif predictions to guide experimental proposals.

Here contrariwise to the previous researches, we focused on the motifs mapped on the MERS-CoV, SARS-CoV, and SARS-CoV-2 proteomes linked to human protein domains. The frequently matched motifs were compared among the coronaviruses. The motif functionality was inferred through enrichment ontology analysis of its partner domains. The based-motif information obtained could be used as the starting point to develop new therapies to combat these viruses in the future.

MATERIALS AND METHODS

Protein sequence retrieval

The SARS-CoV (taxid:694009) and SARS-CoV-2 (taxid:2697049) sequences were retrieved from the NCBI virus repository (accessed on 01 September 2020) [20] using available predefined filters, such as human for host, length of proteins, and the completeness option for sequences. These sequences were firstly filtered based on its report date; then, sequences before 2019 were put on the SARS-CoV dataset. The redundant amino acid sequences were removed with the perl program “fasta_uniqueseq.pl” obtained from FASTA Tool list web page (http://www.ncbi.nlm.nih.gov/CBBresearch/Spouge/html.ncbi/fasta/list.html). The sequences for MERS-CoV were retrieved from the virus variation database [21], using the options as human host, sequence completeness, and collapse for removing redundant sequences. The final number of each viral protein in the datasets ordered by its arrangement on the genome are shown in Table 1. The

SARS-CoV protein sequences were grouped together with the SARS-CoV-2 dataset for the analysis due to its small number after eliminating the redundant sequences.

Table 1. The total number of non-redundant viral protein sequences for analysis

Protein	MERS-CoV	SARS-CoV	SARS-CoV-2
ORF1ab	162	4	4003
ORF1a	140		
S	98	5	1135
ORF3a	25	3	421
NS4a	22		
NS4b	36		
NS5	21		
E	6	1	45
M	18	6	125
ORF6		1	73
ORF7a		1	149
ORF8	18	1	146
N	44	1	539
ORF10		1	35
TOTAL	590	24	6671

Domain-motif data mining process

Our data mining process is based on our previous reported methodology [22], adapted to the data retrieved for the MERS-CoV and SARS-CoV/CoV-2 viruses. It includes three main steps. 1) Literature search. First, we obtained the human genes associated to the SARS-CoV/CoV-2 and MERS-CoV related diseases with pubtator [23]. This tool allows searching in a straightforward manner the reporting genes related to the infections by these viral pathogens in the PubMed literature. These gene names were compared and unified with the information from a recent research published by Perrin-Cocon et al., [11] to form a list of unique gene names. This list

was submitted into the UniProt database [24] to obtain the human UniProt IDs that match our query for the next process. 2) Pfam database [25] mining for human protein domains: From the Pfam we downloaded the latest version of the files “Pfam-A.regions.tsv” and “Pfam-A.clans.tsv”. The obtained UniProt IDs that match on the Pfam-A.regions.tsv file were extracted to mine the Pfam-A.clans file. Thereby, it was obtained the Pfam accession, clan ID, Pfam ID, and Pfam description columns that contain information associated with our UniProt ID list. 3) The domain-motif information was mined from the databases of three-Dimensional Interacting Domains (3DID) [26] and ELM [27]. The motif information for 3DID was retrieved from the 3DID-DMI flat 2019 version. From this file, the Pfam IDs, domain-motif name, and the regular expressions (regexp) were extracted and stored in local files which was used as the target file to draw out the information associated with the Pfam IDs previously obtained. In the ELM database, the information came from the files “elm_interaction_domains.tsv” and “elm_classes.tsv”. The first file was the target file to match the Pfam accessions IDs and was then used to take out the domain-motif name, Pfam accession, and the associated regexp from the elm_classes.tsv file. Each regexp was used to match motif amino acid sequences in the protein datasets with the patmatch software [28]. We used linux terminal for each query with the bash command “for ID in `cat file_of_IDs.txt`; do grep $ID target_file.txt; done > extracted_info_file.txt”. The obtained files were also checked manually for concordance with the query IDs.

Identification of potential functional host-like viral motifs

The potential functional motif identification was based on the percentage of regexp that matches a specific amino acid sequence. To this end, we followed 70% cut-off match as in the previous study [29]. For example, a total of 4003 ORF1ab non-redundant sequences were retrieved for SARS-CoV-2; consequently, a regexp present in more than 70% of ORF1ab proteins signifies that a specific motif matched more than 2802 sequences. Those frequent motifs were also queried on shuffled sequences versions of each protein dataset that was produced with the “shuffleseq”

function from the EMBOSS suite programs [30]. If those inferred motifs were found scarcely on the randomized sequences, it reinforces as functional motifs.

Protein domain enrichment analysis

The protein domain enrichment analysis was carried out with the dgOR package [31] for R statistical language. For this analysis, the Pfam accession numbers were used as input data and the first ten significant ($p < 0.05$) ontologies based on the hypergeometric test related to gene ontology biological process (GOBP) and Gene ontology molecular function (GOMF) were analyzed.

Identification of motifs as immune epitopes

The immune epitope database (IEDB) [32] was manually queried for motif sequences with ≥ 5 amino acids, setting the blast parameter of identity more than 70 %, and selecting the options “human host”, “all assay types”, and the disease option “COVID-19 and Severe acute respiratory syndrome” as filters. This query analysis was omitted for the MERS-CoV because there is not available information for this pathogen on the IEDB.

Statistics

The statistics rests on descriptive statistics of the frequent motifs. The obtained information was analyzed by its conjunction and disjunction relationships based on the matching patterns. This analysis was carried out with the help of the web tool for the calculation and drawing of custom Venn diagrams (http://bioinformatics.psb.ugent.be/webtools/Venn/).

RESULTS

Literature mining

After removing duplicate gene names among the reviewed publications (data in S1 File 1), 497 human genes for SARS-CoV/CoV-2 and 65 for MERS-CoV infection were found involved in pathogenesis (Table 2, data in S1 File2). The comparison of our mined information with Perrin-Cocon et al [11] showed overlapped gene information (n= 124), and the newly acquired (n= 438), especially for the MERS-CoV

viruses. After eliminating the duplicated the rest are theunique gene names (data in S1 File 2), which were used to search its corresponding UniProt IDs to mine the Pfam, 3DID, and ELM databases for the subsequent regexp match analysis.

Table 2. The total number of human gene names obtained from the PubMed literature and compared with Perrin-Cocon et al. [11]

	Present study	**Present study ∩ Perrin-Cocon et al.,**	**Perrin-Cocon et al.,**
MERS-CoV	55	10	7
SARS-CoV/CoV2	383	114	352

*∩ means the intersection in the conjunction-disjunction analysis

Identification of functional viral protein motifs

The functional regions of proteins are either structured or disordered. However, the proteins of coronaviruses were found mainly ordered according to IUPRED (S2 Fig 1) [33]. For example, most amino acids of the largest protein ORF1ab and the spike (S) protein were found below the 0.5 score. However, few regions of viral protein were disordered, such as the nucleocapsid (N) protein. In this study, the whole regexp lists obtained from the 3DID and ELM databases (data in S1 File 3) were mapped on the whole viral protein sequences. The frequent (>70%) regexps that matched amino acid motifs are shown in Table 3 and the data in S1 File 4.

Table 3. Total number of motifs frequently matched by regexp.

	3DID			ELM		
Protein	**M-CoV**	**M-CoV ∩ S-CoV/CoV-2**	**S-CoV/CoV-2**	**M-CoV**	**M-CoV ∩ S-CoV/CoV-2**	**S-CoV/CoV-2**
ORF1ab	65	148	31	8	78	5
S	47	50	38	11	44	12
ORF3a	4	6	24	1	13	28
NS4a	14			25		
NS4b	46			35		
NS5	20			28		
E	19	0	5	6	5	4
M	9	5	15	11	18	9
ORF6			4			18
ORF7a			20			27
ORF8	23	1	14	8	7	15
N	23	27	31	9	27	14
ORF10			2			12
TOTAL	270	237	184	142	192	144

*∩ means the intersection in the conjunction-disjunction analysis

The ORF1ab, S, and N sequences were matched by the regexp more than the other proteins from databases. A high number of motifs were shared among three CoVs in the ORF1ab (n=148 and 78), followed by the S (n=50 and 44) and the N (n= 27 and 27). The regexp motifs were redundant among the proteins or viral proteomes (data in S1 File 4); for example, the ORF1ab and S shared the same motifs (Fig 1A); and a high number of motifs shared between the MERS-CoV and SARS-CoV/CoV-2 after removing the redundant (Fig 1B, data in S1 File 5). Most of these motifs were

scarcely on the shuffled sequences; thus, all were considered in the subsequent analysis.

Figure 1. Venn diagrams show the redundant or non-redundant regexp motifs among the proteins or viral proteomes. (A) Venn diagram to show the redundant regexp numbers mapped on the ORF1ab and Spike proteins. (B) Venn diagram of total non-redundant regexp mapped in MERS-CoV and SARS-CoV-2 obtained from the two databases.

Protein domain enrichment analysis for non-redundant motifs

First, it was examined the conjunction-disjunction relationships for the total number of Pfam accessions associated with non-redundant motifs described above. A total of 78 non-redundant domains were shared for MERS-CoV and SARS-CoV/CoV-2 irrespective of the database source, and few were specific to MERS-CoV (n=8) and SARS-CoV/CoV-2 (n=9) (Fig 2A, data in S1 File 5). Protein domain enrichment analysis of the 78 shared domains for GOBP identifies general terms related to metabolic and cellular processes. Five GOBP significant terms were related to energy reserve and glycogen biosynthesis metabolism (Fig 2B, data in S1 File 6). GOMF analysis also identifies five important terms related to channel regulation in which potassium channel regulator activity was the most significant (Fig 2C, data in S1 File 6). The study of specific domains for MERS-CoV and SARS-CoV-2 also showed terms associated with the same biological processes and molecular functions of the 78 shared domains. Thus, those domains could be the primary targets for molecular mimicry generated by MERS-CoV and SARS-CoV/CoV-2 to manipulate the host cell machinery.

Figure 2. Protein domain enrichment analysis that produced the signicant gene ontology terms for non-redundant motifs. (A) Venn diagram for the non-redundant domains. (B) Gene ontology terms for biological processes and (C) molecular functions terms of the non-redundant domains. Nodes are colored according to adjusted p-values.

Analysis of significant domains present on distinct host proteins

The analysis described above allows us to identify specific proteins linked to the domains involved with significant ontology terms. Four domains (Pfam accession ID: PF00656, PF00026, PF00082, PF00089) related to the glycogen biosynthetic process were present in 26 proteins that matched our gene lists. Among them, the PF00089 related to trypsin domain function is the more promiscuous present on most of the proteins (Fig 3A). This domain was associated with the protease TMPRSS2, an endothelial cell surface protein involved in the entry and spread of CoVs and influenza virus [34], so that this protein has been proposed as a potential drug target to combat those viruses. It was also found the domains associated with the potassium channel regulator activity (Fig 3B).

Figure 3. Network representation of significant domains linked to proteins and their gene ontology terms. (A) Biological processes and (B) Molecular functions . The green light diamonds represent the domains, and the ellipses represent the protein names associated with the domains. The images were generated with the cytoscape software [35].

Identification of amino acid motif sequences as immune epitopes

The non-redundant motifs ≥ 5 amino acids were searched for a match with epitopes reported on the IEDB, which were experimentally confirmed. The amino acid sequences of several motifs matched on epitopes sequences for SARS-CoV/CoV-2 that recognize B and T cells specific to class I or II MHC (data in S1 File 7). These motifs had the following main characteristics. 1) The epitope linear motifs contain the nested motifs recognized by both B and T cells. For example, the motif matched with the regexp [DE]..[IMV].[ST] was found on the B cell and T cell epitope PK**EITVA**TSRTLSYYK (IEDB ID: 48052) in the M protein [36] of SARS-CoV and SARS-CoV-2 [37]. 2) Motifs matched by the same regexp are prone to occur in different protein structural locations. For example, the regexp motif P.{0,1}S.{1,2}K matches the amino acid sequences PLSETK and PVSMTK locating to varying coordinates on the S protein (Fig 4). 3) Motifs maintain its crucial amino acids, and

little variations occur at neighbor sites. For example, the **PVSMTK** motif nested on the B cell linear epitope IL**PVSMTKT**SVDCTMYICGD (IEDB ID:1309493) of SARS-CoV-2 (Fig 4A and D) [38] varied a little on the epitope sequence **PVSMAK**TSVDCNMYICGDS (IEDB ID: 49968) of the SARS-CoV, maintaining its main amino acid anchors P,S and K. PVSMAK was found only in one SARS-CoV-2 sequence (NCBI ID: QKV39263) isolated from Washington, Yakima County.

Figure 4. Some motifs matched the epitopes on the Spike protein. (A) Spike protein of SARS-CoV-2 (PDBID:6XS6). (B) The regexp. (C) The motif PLSETK. Red balls indicate the PLSETK seqlogo motif mapped at amino acid positions 295 to 300 (D) The motif PVSMTK. Green balls and sticks showed the total length of the epitope ILPVSMTKTSVDCTMYICGD, including the PVSMTK seqlogo motif mapped (the balls) at amino acids positions 728 to 733.

DISCUSSION

In this work, we employed our previous data mining methodology [22] to identify potential functional motifs but applied to MERS-CoV and SARS-CoV/CoV-2 viruses. The main advantage of this method is the search restricted to human protein targets involved in the virus pathogenesis. The initial step allows us to reduce *a priori* the query on the 3DID and ELM databases. As a result, the unsheathed domain-motif information is potentially associated with human genes related to pathogenesis of the MERS-CoV and SARS-CoV/CoV2. Our approach is then similar to the methods used by Hagai, T., et al., Becerra, A. et al and Zhang, A et al [29,39,40] in predicting functional motifs. These methods include some distinctive features such as predicting disordered regions on the protein, the high frequency of amino acid motifs in the protein sequences datasets under study, and the scarcity of amino acid motifs on shuffled sequences. The filters were tailored according to the information obtained in each data mining process. All those filtered steps guided our analysis to a more specificity that linked the predicted functional motifs as part of immune epitopes as previously we did for influenza A viruses [22]. It is distinctive of our prediction approach, because it was used to reduce the high rate of false positives

associated with the computational prediction of motifs [41]. Furthermore, our method could be an alternative for computer-aided reverse vaccinology.

One interesting result is that the tendency of matched motifs occurred in the most variable proteins, the ORF1ab, and the S protein of the coronavirus proteomes. The ORF1ab contains the nonstructural proteins responsible for the translation machinery of viruses in the intracellular environment [42] and the S protein is essential for the virus's attachment to the host cell [43]. The tendency of motifs to appear on the proteins involved in virus replication was also observed in influenza viruses [44]. Thus, the high frequency of host-like motifs in those viral proteins suggests that such proteins could be the master kidnappers. Another finding is the high number of shared motifs across the proteome or distinct proteins of a proteome, reflecting the viral motifs to evolve independently in light of acquiring host-like mechanisms for the success in the invasion of host cells.

The domain enrichment analysis showed that the general biological processes, and molecular functions could be the consequence of the MERS-CoV and SARS-CoV/CoV-2 mimicry to hijack the host cell. The most significant ontology terms are the energy-saving and glycogen biosynthesis metabolism association. This result agrees with that viruses use the infected cells' carbon sources to achieve viral replication and virion production [45]. It is reasonable that glycogen, a storage form of glucose, is utilized in unexpected, exhausting cell activity [46] as infected. On the other hand, as this biosynthetic pathway is vital for the viruses' survival, targeting essential components such as the glycogen synthase kinase could help treat virus infections. It was reported that the use of two glycogen synthase inhibitors altered the hepatitis C virus assembly and release [47]. Hence, the proteins we found in the present study could be used to explore them as drug targets.

In another context, motifs have been suggested as potential immunogens [41]. It took our attention to search motif that matched with immune epitopes. Indeed we found that some motifs matched to the epitopes on the IEDB. Some of them were nested on the epitopes of earlier SARS-CoV and also present on those new SARS-CoV-2. It reaffirms the evidence of cross-reactive immune responses to coronavirus

infections by SARS-CoV and SARS-CoV-2 [48–51]. Additionally, our study identified the epitopes harboring motifs that could interact with human protein domains. It is quite relevant because such domain-motifs shared in the different coronavirus can trigger a common molecular mimicry process that could lead to autoimmune diseases. It was demonstrated that antibodies derived from Flu vaccinated patients react with homologous sequences of the nucleoprotein of influenza A virus and the hypocretin receptor 2 domain of humans, the latter of which was involved in narcolepsy, an autoimmune adverse effect attributed to the Flu-vaccine [52]. Influenza immunization is also attributed to Guillain-Barré syndrome [53], a disease in which its pathogenesis is associated with several bacterial and viral pathogens' molecular mimicry [54–56]. Thus, our results are vital to helping in the currently underway rational vaccine development efforts, mainly because several autoimmune diseases have been associated with COVID-19 [57].

CONCLUSIONS

In conclusion, this study showed that our method's adaptability and practicality could guide a rational inference of domain targets and their interacting host-like motifs on the MERS-CoV and SARS-CoV/CoV-2 proteomes. A high number of motifs were shared in the different CoVs, and it could interact with human proteins, indicating that molecular mimicry is a common strategy for CoVs. The finding of motifs as part of immune epitopes makes our method a suitable alternative for reverse vaccinology. The obtained information could be the starting point for future theoretic and experimental studies to develop new drugs and peptidic vaccines to combat those viruses.

References

1. Drosten C, Günther S, Preiser W, van der Werf S, Brodt H-R, Becker S, et al. Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome. New England Journal of Medicine. 2003;348: 1967–1976. doi:10.1056/NEJMoa030747

2. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi

Arabia. New England Journal of Medicine. 2012;367: 1814–1820. doi:10.1056/NEJMoa1211721

3. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. New England Journal of Medicine. 2020 [cited 29 Oct 2020]. doi:10.1056/NEJMoa2001017

4. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020;5: 536–544. doi:10.1038/s41564-020-0695-z

5. Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. Natl Sci Rev. 2020 [cited 15 Jun 2020]. doi:10.1093/nsr/nwaa036

6. Apicella M, Campopiano MC, Mantuano M, Mazoni L, Coppelli A, Prato SD. COVID-19 in people with diabetes: understanding the reasons for worse outcomes. The Lancet Diabetes & Endocrinology. 2020;8: 782–792. doi:10.1016/S2213-8587(20)30238-2

7. Khorsand B, Savadi A, Naghibzadeh M. SARS-CoV-2-human protein-protein interaction network. Informatics in Medicine Unlocked. 2020;20: 100413. doi:10.1016/j.imu.2020.100413

8. He R, Leeson A, Ballantine M, Andonov A, Baker L, Dobie F, et al. Characterization of protein–protein interactions between the nucleocapsid protein and membrane protein of the SARS coronavirus. Virus Res. 2004;105: 121–125. doi:10.1016/j.virusres.2004.05.002

9. Vidalain P-O, Jacob Y, Hagemeijer MC, Jones LM, Neveu G, Roussarie J-P, et al. A Field-Proven Yeast Two-Hybrid Protocol Used to Identify Coronavirus–Host Protein–Protein Interactions. Coronaviruses. 2014;1282: 213–229. doi:10.1007/978-1-4939-2438-7_18

10. Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature. 2020;583: 459–468. doi:10.1038/s41586-020-2286-9

11. Perrin-Cocon L, Diaz O, Jacquemin C, Barthel V, Ogire E, Ramière C, et al. The current landscape of coronavirus-host protein-protein interactions. J Transl Med. 2020;18: 319. doi:10.1186/s12967-020-02480-z

12. Sadegh S, Matschinske J, Blumenthal DB, Galindez G, Kacprowski T, List M, et al. Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. Nature Communications. 2020;11: 3518. doi:10.1038/s41467-020-17189-2

13. Zhou Y, Hou Y, Shen J, Huang Y, Martin W, Cheng F. Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discovery. 2020;6: 1–18. doi:10.1038/s41421-020-0153-3

14. Basu MK, Poliakov E, Rogozin IB. Domain mobility in proteins: functional and evolutionary implications. Brief Bioinform. 2009;10: 205–216. doi:10.1093/bib/bbn057

15. Lin MM, Zewail AH. Hydrophobic forces and the length limit of foldable protein domains. PNAS. 2012;109: 9851–9856. doi:10.1073/pnas.1207382109

16. Garamszegi S, Franzosa EA, Xia Y. Signatures of Pleiotropy, Economy and Convergent Evolution in a Domain-Resolved Map of Human–Virus Protein–Protein Interaction Networks. PLOS Pathog. 2013;9: e1003778. doi:10.1371/journal.ppat.1003778

17. Davey NE, Travé G, Gibson TJ. How viruses hijack cell regulation. Trends Biochem Sci. 2011;36: 159–169. doi:10.1016/j.tibs.2010.10.002

18. Mészáros B, Sámano-Sánchez H, Alvarado-Valverde J, Čalyševa J, Martínez-Pérez E, Alves R, et al. Short linear motif candidates in the cell entry system used by SARS-CoV-2 and their potential therapeutic implications. Sci Signal. 2021;14. doi:10.1126/scisignal.abd0334

19. Kliche J, Kuss H, Ali M, Ivarsson Y. Cytoplasmic short linear motifs in ACE2 and integrin β3 link SARS-CoV-2 host cell receptors to mediators of endocytosis and autophagy. Sci Signal. 2021;14. doi:10.1126/scisignal.abf1117

20. Brister JR, Ako-Adjei D, Bao Y, Blinkova O. NCBI viral genomes resource. Nucleic Acids Res. 2015;43: D571-577. doi:10.1093/nar/gku1207

21. Brister JR, Bao Y, Zhdanov SA, Ostapchuck Y, Chetvernin V, Kiryutin B, et al. Virus Variation Resource—recent updates and future directions. Nucl Acids Res. 2013; gkt1268. doi:10.1093/nar/gkt1268

22. García-Pérez CA, Guo X, Navarro JG, Aguilar DAG, Lara-Ramírez EE. Proteome-wide analysis of human motif-domain interactions mapped on influenza a virus. BMC Bioinformatics. 2018;19: 238. doi:10.1186/s12859-018-2237-8

23. Wei C-H, Kao H-Y, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res. 2013;41: W518–W522. doi:10.1093/nar/gkt441

24. Consortium TU. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47: D506–D515. doi:10.1093/nar/gky1049

25. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47: D427–D432. doi:10.1093/nar/gky995

26. Mosca R, Céol A, Stein A, Olivella R, Aloy P. 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. 2014;42: D374-379. doi:10.1093/nar/gkt887

27. Kumar M, Gouw M, Michael S, Sámano-Sánchez H, Pancsa R, Glavina J, et al. ELM—the eukaryotic linear motif resource in 2020. Nucleic Acids Res. 2020;48: D296–D306. doi:10.1093/nar/gkz1030

28. Yan T, Yoo D, Berardini TZ, Mueller LA, Weems DC, Weng S, et al. PatMatch: a program for finding patterns in peptide and nucleotide sequences. Nucleic Acids Res. 2005;33: W262–W266. doi:10.1093/nar/gki368

29. Hagai T, Azia A, Babu MM, Andino R. Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions. Cell Rep. 2014;7: 1729–1739. doi:10.1016/j.celrep.2014.04.052

30. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics. 2000;16: 276–277. doi:10.1016/S0168-9525(00)02024-2

31. Fang H. dcGOR: An R Package for Analysing Ontologies and Protein Domain Annotations. PLoS Comput Biol. 2014;10. doi:10.1371/journal.pcbi.1003929

32. Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, et al. The Immune Epitope Database 2.0. Nucleic Acids Res. 2010;38: D854–D862. doi:10.1093/nar/gkp1004

33. Mészáros B, Erdős G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Res. 2018;46: W329–W337. doi:10.1093/nar/gky384

34. Shen LW, Mao HJ, Wu YL, Tanaka Y, Zhang W. TMPRSS2: A potential target for treatment of influenza virus and coronavirus infections. Biochimie. 2017;142: 1–10. doi:10.1016/j.biochi.2017.07.016

35. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. 2003;13: 2498–2504. doi:10.1101/gr.1239303

36. He Y, Zhou Y, Siddiqui P, Niu J, Jiang S. Identification of immunodominant epitopes on the membrane protein of the severe acute

respiratory syndrome-associated coronavirus. J Clin Microbiol. 2005;43: 3718–3726. doi:10.1128/JCM.43.8.3718-3726.2005

37. Peng Y, Mentzer AJ, Liu G, Yao X, Yin Z, Dong D, et al. Broad and strong memory CD4+ and CD8+ T cells induced by SARS-CoV-2 in UK convalescent individuals following COVID-19. Nat Immunol. 2020;21: 1336–1345. doi:10.1038/s41590-020-0782-6

38. Yi Z, Ling Y, Zhang X, Chen J, Hu K, Wang Y, et al. Functional mapping of B-cell linear epitopes of SARS-CoV-2 in COVID-19 convalescent population. Emerg Microbes Infect. 2020;9: 1988–1996. doi:10.1080/22221751.2020.1815591

39. Becerra A, Bucheli VA, Moreno PA. Prediction of virus-host protein-protein interactions mediated by short linear motifs. BMC Bioinformatics. 2017;18: 163. doi:10.1186/s12859-017-1570-7

40. Zhang A, He L, Wang Y. Prediction of GCRV virus-host protein interactome based on structural motif-domain interactions. BMC Bioinformatics. 2017;18: 145. doi:10.1186/s12859-017-1500-8

41. Hraber P, O'Maille PE, Silberfarb A, Davis-Anderson K, Generous N, McMahon BH, et al. Resources to Discover and Use Short Linear Motifs in Viral Proteins. Trends in Biotechnology. 2020;38: 113–127. doi:10.1016/j.tibtech.2019.07.004

42. Yoshimoto FK. The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the Cause of COVID-19. Protein J. 2020;39: 198–216. doi:10.1007/s10930-020-09901-4

43. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell. 2020;181: 281-292.e6. doi:10.1016/j.cell.2020.02.058

44. Yang C-W. A Comparative Study of Short Linear Motif Compositions of the Influenza A Virus Ribonucleoproteins. PLoS One. 2012;7. doi:10.1371/journal.pone.0038637

45. Sanchez EL, Lagunoff M. Viral activation of cellular metabolism. Virology. 2015;479–480: 609–618. doi:10.1016/j.virol.2015.02.038

46. Berg JM, Tymoczko JL, Stryer L. Glycogen Metabolism. Biochemistry 5th edition. 2002 [cited 29 Oct 2020]. Available: https://www.ncbi.nlm.nih.gov/books/NBK21190/

47. Sarhan MA, Abdel-Hakeem MS, Mason AL, Tyrrell DL, Houghton M. Glycogen synthase kinase 3β inhibitors prevent hepatitis C virus

release/assembly through perturbation of lipid metabolism. Scientific Reports. 2017;7: 2495. doi:10.1038/s41598-017-02648-6

48. Mateus J, Grifoni A, Tarke A, Sidney J, Ramirez SI, Dan JM, et al. Selective and cross-reactive SARS-CoV-2 T cell epitopes in unexposed humans. Science. 2020;370: 89–94. doi:10.1126/science.abd3871

49. Tai W, Zhang X, He Y, Jiang S, Du L. Identification of SARS-CoV RBD-targeting monoclonal antibodies with cross-reactive or neutralizing activity against SARS-CoV-2. Antiviral Res. 2020;179: 104820. doi:10.1016/j.antiviral.2020.104820

50. Sette A, Crotty S. Pre-existing immunity to SARS-CoV-2: the knowns and unknowns. Nature Reviews Immunology. 2020;20: 457–458. doi:10.1038/s41577-020-0389-z

51. Grifoni A, Weiskopf D, Ramirez SI, Mateus J, Dan JM, Moderbacher CR, et al. Targets of T Cell Responses to SARS-CoV-2 Coronavirus in Humans with COVID-19 Disease and Unexposed Individuals. Cell. 2020;181: 1489-1501.e15. doi:10.1016/j.cell.2020.05.015

52. Ahmed SS, Volkmuth W, Duca J, Corti L, Pallaoro M, Pezzicoli A, et al. Antibodies to influenza nucleoprotein cross-react with human hypocretin receptor 2. Science Translational Medicine. 2015;7: 294ra105-294ra105. doi:10.1126/scitranslmed.aab2354

53. Schonberger LB, Bregman DJ, Sullivan-Bolyai JZ, Keenlyside RA, Ziegler DW, Retailliau HF, et al. Guillain-Barre syndrome following vaccination in the National Influenza Immunization Program, United States, 1976--1977. Am J Epidemiol. 1979;110: 105–123. doi:10.1093/oxfordjournals.aje.a112795

54. Rees JH, Soudain SE, Gregson NA, Hughes RAC. Campylobacter jejuni Infection and Guillain–Barré Syndrome. New England Journal of Medicine. 1995;333: 1374–1379. doi:10.1056/NEJM199511233332102

55. Steininger C, Popow-Kraupp T, Seiser A, Gueler N, Stanek G, Puchhammer E. Presence of Cytomegalovirus in Cerebrospinal Fluid of Patients with Guillain-Barré Syndrome. J Infect Dis. 2004;189: 984–989. doi:10.1086/382192

56. Rojas M, Restrepo-Jiménez P, Monsalve DM, Pacheco Y, Acosta-Ampudia Y, Ramírez-Santana C, et al. Molecular mimicry and autoimmunity. J Autoimmun. 2018;95: 100–123. doi:10.1016/j.jaut.2018.10.012

57. Ehrenfeld M, Tincani A, Andreoli L, Cattalini M, Greenbaum A, Kanduc D, et al. Covid-19 and autoimmunity. Autoimmun Rev. 2020;19: 102597. doi:10.1016/j.autrev.2020.102597

SUPPORTING INFORMATION

S1 FILE 1. THE LITERATURE MINED INFORMATION.

S1 FILE 2. THE MERGED GENE NAME LISTS FOR THE MERS-COV AND SARS-COV/COV-2.

S1 FILE 3. THE REGEXP LISTS OBTAINED FROM 3DID AND ELM.

S1 FILE 4. THE REDUNDANT REGEXP MATCHED ON THE MERS-COV AND SARS-COV/COV-2 PROTEOMES.

S1 FILE 5. THE NON-REDUNDANT MOTIFS WITH ITS DOMAIN ACCESSION PARTNER.

S1 FILE 6. GOBP AND GOMF FOR THE SIGNIFICANT DOMAINS.

S1 FILE 7. THE MOTIFS NESTED ON LINEAR SEQUENCES OF EPITOPES FROM IEDB.

S2 FIG. 1. ORDER AND DISORDER REGIONS FOR THE MERS-COV, SARS-COV, AND SARS-COV-2 PROTEINS ARRANGED BY ITS KNOWN GENOME ORDER.

7. CONCLUSIÓNES

7..1 Minería de datos sobre interacciones Dominio-Motivo

Para el análisis de minería de datos se obtuvieron las secuencias de las proteínas virales de los tres β-CoV de la base de datos de NCBI y la base Virus Variation Data Base, y se realizó la identificación de secuencias de aminoácidos redundantes; se realizó minería de datos para la identificación de expresiones regulares (regexp) según la metodología de García- Pérez (Garcia-Perez et al., 2018), la cual incluye los siguientes tres pasos: *1) Búsqueda en la literatura.* Se obtuvieron los nombres de genes asociados con la infección por MERS-CoV, SARS-CoV y SARS-CoV-2 con PubTator (Wei et al., 2019). La lista de genes obtenidos fue consultada en la base de datos UniProt para obtener los IDs correspondientes con las proteínas del huésped. *2) Minería de datos en la base de datos Pfam para la identificación de Dominios de proteínas humanas*. *3) Minería de datos para identificación de interacciones Dominio-motivo mediante regexp en las bases de datos 3DID y ELM (Kumar et al., 2024; Mosca et al., 2014)*. Se identificaron las expresiones regulares (regexp) y se evaluaron para la identificación de potenciales epítopes de reconocimiento, estos epítopes fueron posteriormente evaluados en términos de inmunogenicidad y toxicidad, además de identificar si éstos se encuentran en regiones desordenadas en la proteína para las consiguientes evaluaciones. Utilizamos la terminal de Linux para realizar cada una de las búsquedas utilizando el comando en bash: “For ID in ‘cat file_of_IDs.txt‘; do grep $ID target_file.txt; done > extracted_info_file.txt

7.2 Selección de secuencias de epítopes de relevancia

La base de datos de epítopes inmunes (IEDB) (Vita et al., 2019) se consultó manualmente para la identificación de secuencias de motivos con una longitud mayor a 8 aminoácidos, estableciendo el parámetro de porcentaje de identidad en más del 70% y seleccionando las opciones "huésped humano", "todos los tipos de ensayo"; la opción de enfermedad "COVID-19” y “enfermedades respiratorias

agudas graves” como filtros de búsqueda. Este análisis de consulta se omitió para el MERS-CoV ya que no había información disponible para este patógeno en la base de datos del IEDB, al momento de nuestro estudio. Posteriormente, las secuencias identificadas fueron evaluadas con la herramienta en línea de Toxinpred(Gupta et al., 2013) y AllerTOP (Dimitrov et al., 2013); seleccionando aquellas secuencias de aminoácidos con un potencial inmunogénico mayor del 60% y con un potencial de toxicidad negativo según el algoritmo de la herramienta AllerTOP.

7.3 Modelado molecular de los epitopes seleccionados

Posterior a la selección de los mejores candidatos se realizó el modelado molecular de dichas secuencias con ayuda del software Discovery Studio de BIOVIA(BIOVIA, 2020) como primer modelado para obtener la estructura de las moléculas. Para su visualización, se utilizó el software PyMOL (Schrödinger).

7.4 Evaluación de la interacción de los epitopes candidatos mediante acoplamiento molecular

Con los resultados de la minería de datos, se realizó un segundo filtrado para la identificación de los epítopes dirigidos contra la proteína N del SARS-CoV-2 con un potencial inmunogénico mayor al 60% con medida entre 11 y 30 aa y se evaluó su energía de unión a distintos receptores HLA correspondientes a células CD8+ con evidencia experimental de reconocer al SARS-CoV-2; por análisis de acoplamiento molecular tomando la afinidad de unión de estos cristales con su ligando original como punto de corte para seleccionar aquellos epítopes con un mayor potencial inmunogénico. El acoplamiento molecular se llevo a cabo con el programa AutoDock Vina, utilizando las coordenadas del centro x , y, z y el tamaño de caja X=, Y=, Z=, en lo que se denomina GridBox, una caja tridimensional en la que se encierra el sitio activo del ligando a evaluar y AutoDock Vina evalua las posibles conformaciones del ligando con su receptor, tomando como la conformación más estable aquella que llega al sitio activo y tiene una menor energía de unión(Goodsell & Olson, 1990).

La minería de datos y las herramientas informáticas son una novedosa estrategia para identificación de epítopes reconocidos por el sistema inmune; en este proyecto se identificaron diversos epitopes dirigidos hacia la proteína N del SARS-CoV-2. El uso de herramientas inmunoinformáticas nos permite dirigir el enfoque de búsqueda de epítopes que podrían ser de relevancia para el diseño de vacunas según su potencial de unirse a receptores HLA para inducir una respuesta del sistema inmune. En este estudio se identificaron diversos epítopes dirigidos contra el SARS-CoV-2, partiendo de las interacciones Dominio-proteína entre el virus y el huésped, observando que los motivos de proteínas virales que interactúan con Dominios de proteínas del huésped se encuentran conservados entre los 3 beta Coronavirus. Proponiendo estos motivos conservados como potenciales epítopes para el diseño de vacunas contra estos virus.

7.5 La filogenética de la nucleocápside del SARS-CoV-2

Nuestro análisis de minería de datos demostró que la proteína N contiene motivos de interacción conservados entre los 3 Coronavirus capaces de infectar humanos, por lo que decidimos analizar la filogenética de esta proteína. Los primeros datos de secuencia disponibles de este nuevo virus ubicaron al SARS-CoV-2 en el subgénero Sarbecovirus de la familia Coronaviridae(Wu et al., 2020). Analizamos la proteína del SARS-CoV-2-N-CTD en relación con la proteína N de virus anteriores que causaron un brote global (SARS-CoV y MERS-CoV), otros humanos infectando coronavirus (HKU1, HCoV-OC43, HCoV-229E y HCoV-NL63), coronavirus de murciélago similares al SARS (bat-SL-CoV2XC21 y bat-SL-CoV2C45), el coronavirus de murciélago bat-CoV-HKU4 y bat-CoV- RaTG13 y el virus del pangolín (Philodota) PCoVGX-P2V. El árbol filogenético de la proteína N mostró que existen relaciones evolutivas entre las especies de coronavirus, por ejemplo, se observó que los coronavirus pangolín (PCoV-GX-P2V), murciélago (bat-CoV-RaTG13) y humano (SARS-CoV y SARS-CoV2) estan agrupados (Boni et al., 2020);también hubo una divergencia entre bat-CoV-RaTG13 y SARS-CoV-2. Estos resultados también se observaron en los informes de Li, et al. por tanto, estas

observaciones conducen hacia un origen zoonótico del SARS-CoV-2, ya que es su ancestro más fiable. Esta exploración filogenética también mostró que los mecanismos evolutivos pueden estar involucrados en las mutaciones de la proteína N (Li et al., 2020).

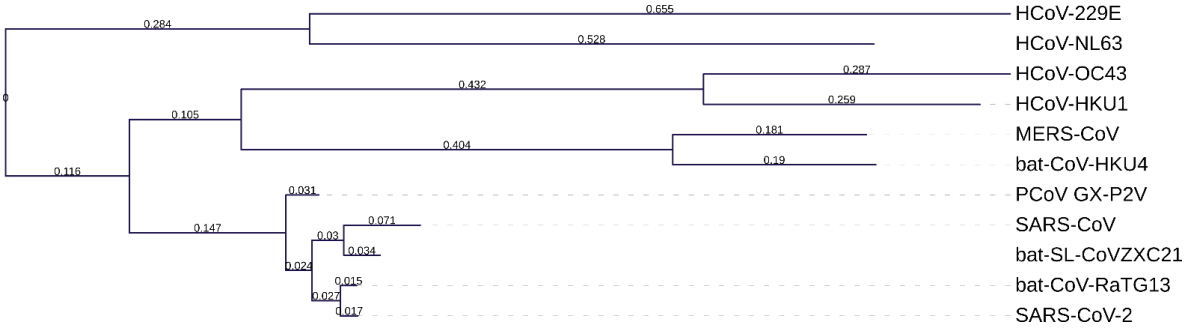

Figura 1. Los datos de la secuencia de la proteína N se descargaron de la base de datos de nucleótidos NCBI GenBank, números de acceso: SARS-CoV-2 (NC_045512), SARS-CoV (NC_004718), MERS-CoV (NC_019843), HKU1 (MH940245), HCoV-OC43 (KX344031) , HCoV-229E (KU291448), HCoV-NL63 (MK334047), bat-SL-CoVZC45 (MG772933), bat-SL-CoVZXC21 (MG772934), bat-CoV-HKU4 (NC_009019), bat-CoV-RaTG13 (MN996532) y PCoV_GX-P2V (MT072864). Las ramas se analizaron con una prueba de razón de verosimilitud aproximada (aLRT). Árbol filogenético estimado con SeaView versión 4.

7.6 Identificación de potenciales epítopes dirigidos hacía la proteína N del SARS-CoV-2 mediante minería de datos

Se realizó la minería de datos para la identificación de interacciones Dominio-motivo y de motivos compartidos entre los 3 β-CoV, encontrando que las proteínas estructurales M, E, S y N son las que comparten una mayor cantidad de motivos compartidos, por lo que se sugiere que estas proteínas se encuentran conservadas entre los 3 beta-coronavirus; además las proteínas S y N comparten una gran cantidad de motivos entre sí, sugiriendo que éstas se encuentran conservada en los CoVs (Fig.1). Debido a la gran cantidad de trabajo experimental sobre la proteína espiga y tomando en cuenta también que es la proteína con mayor cantidad de mutaciones en las variantes de interés (Ahn et al., 2024; Rojas Chavez et al., 2024)

, se seleccionó a la proteína N como nuestra proteína de interés en la identificación y evaluación de potenciales epítopes.

Se evaluó si los motivos identificados corresponden con secuencias de epítopes contra la Nucleocápside del SARS-CoV-2, y se encontraron 231 epítopes reportados en la base de IEDB; de los cuales 120 corresponden con epítopes reconocidos por células T y 111 corresponden con epítopes reconocidos por células B, además, se evaluó si estos epítopes se encuentran en regiones desordenadas dentro de la proteína nucleocápside, ya que las regiones desordenadas de las proteínas por su flexibilidad molecular, tienden a relacionarse con interacciones proteína-proteína y además el sistema inmune tiende a reconocer con una mayor afinidad a epítopes dentro de estas regiones, indicando el potencial que tienen estas secuencias de aminoácidos como el mecanismo por el cual el virus es capaz de mimetizar y secuestrar la maquinaria de replicación celular de la célula huésped para cumplir con su ciclo de replicación y supervivencia (Davey et al., 2011; Garamszegi et al., 2013; He et al., 2004; Khorsand et al., 2020).

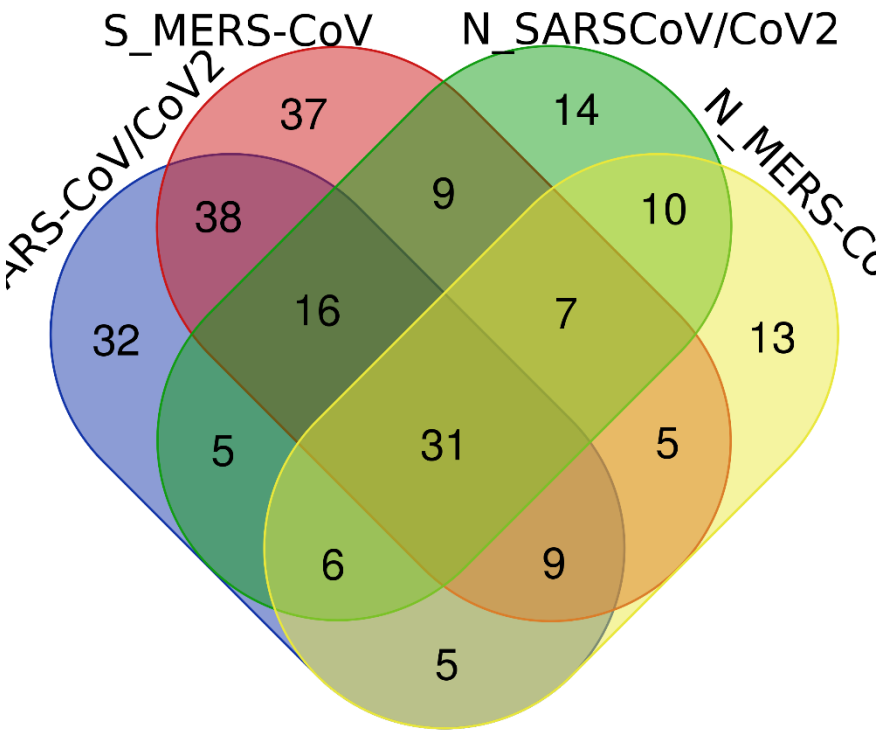

Figura 1. Diagrama de Venn que muestra las expresiones regulares redundantes mapeadas en la proteína nucleocapside (N) y proteína Spiga (S) de los CoVs SARS-CoV/CoV2 y MERS-CoV.

7.7 Evaluación de unión de epítopes de la proteína N con HLAs recurrentes en la infección por SARS-CoV-2

De acuerdo con el análisis de minería de datos, se seleccionaron los epítopes con un mayor potencial inmunogénico que estuvieran contenidos en la proteína N de los beta coronavirus, la selección de los HLA se realizó de acuerdo a aquellos que tuvieran evidencia bibliográfica de ser recurrentes en la infección por SARS-CoV-2.

Estos epítopes fueron modelados y evaluados mediante acoplamiento molecular tomando como control las energías de unión de los cristales de los HLAs con su

ligando original, de acuerdo con las estructuras obtenidas del PDB, los controles para el análisis de acoplamiento molecular se encuentran descritos en la tabla 1.

Tabla 1. ***Energía de unión de los HLA con su ligando control.*** *Se tomó en cuenta la energía de unión de cada ligando control con su HLA como punto de corte para los análisis de ligandos problema.*

HLA	PDB ID	Energía de unión (kcal/mol)
HLA-A*0301	3RL1	-7.7
HLA-B*0702	5EO1	-7.9
HLA-A*1101	1X7Q	-5.6
HLA-A*3001	6J1W	-4.7

En la evaluación por acoplamiento molecular se identificaron aquellos residuos de aminoácidos que se encuentran en el sitio activo del receptor para su posterior comparación con los epitopes problema. Tomando como un buen candidato a evaluar aquellos epítopes que interactuaron con una mayor cantidad de residuos de aminoácidos del sitio activo de cada receptor HLA. Las energías de unión obtenidas de las simulaciones de acoplamiento molecular se describen en la tabla 2.

Tabla 2. ***Energía de afinidad de unión (kcal/mol).*** *Se muestra la afinidad de unión de cada epítope problema y su punto de corte según el cristal que se utilizó para la evaluación por acoplamiento molecular, además de la secuencia correspondiente de cada epítope, además de su score de desorden según la plataforma de IUPred.*

	Epitope	Energía de Unión (kcal/mol)	Secuencia	Posición	Región	Desorden (IUPRED score)
HLA-A*0301						
(-4.7 kcal/mol)						
	pep 10	-8.6	NTNSSPDDQIGYY	[76,82]	NTD	0.801
	pep 18	-8.1	LSPRWYFYYL	[105,111]	NTD	0.388
	pep 20	-7.8	SPRWYFYYL	[105,111]	NTD	0.388
	pep 22	-8	QRNAPRITF	[12,18]	N-arm	0.910
HLA-B*0702						
(-7.9) kcal/mol)						
	pep 8	-9.7	KTFPPTEPK	[359,366]	CTD	0.585
	pep 10	-8.2	NTNSSPDDQIGYY	[76,82]	NTD	0.801
	pep 18	-9.4	LSPRWYFYYL	[105,111]	NTD	0.388
	pep 20	-8.8	SPRWYFYYL	[105,111]	NTD	0.388
	pep 22	-8.6	QRNAPRITF	[12,18]	N-arm	0.910
HLA-A*1101						
(-5.6 kcal/mol)						

pep_2	-6.5	RIRGGDGKMK	[93,100]	NTD	0.611
pep_3	-5.9	AEGSRGGSQA	[173,180]	LKR	0.600
pep_7	-6.4	AYKTFPPTEPK	[359,366]	CTD	0.585
pep_8	-7.4	KTFPPTEPK	[359,366]	CTD	0.585
pep_15	-6.6	KAYNVTQAFGRRGPE	[268,274]	CTD	0.696
pep_17	-7	KMKDLSPRW	[76,81]	NTD	0.557
pep_18	-7.8	LSPRWYFYYL	[105,111]	NTD	0.388
pep_22	-6.6	LSPRWYFYYL	[12,18]	N-arm	0.388
HLA-A*3001 (-7.7 kcal/mol)					
pep 3	-6.8	AEGSRGGSQA	[173,180]	LKR	0.600
pep 8	-5.7	KTFPPTEPK	[359,366]	CTD	0.585
pep 11	-5.3	KMKDLSPRW	[102,108]	NTD	0.377
pep 17	-5.7	KMKDLSPRW	[76,81]	NTD	0.557

De acuerdo con el análisis de acoplamiento molecular, se seleccionaron aquellos epítopes cuya secuencia mostrara una energía de unión menor a la del ligando control de cada cristal de los receptores HLA (Tabla 2). Demostrando así que la proteína N es una proteína altamente inmunogénica y que ésta es reconocida por los HLA de las células T CD8+. Posterior a la evaluación por acoplamiento molecular, se utilizó el software PyMOL para la visualización de aquellas interacciones con un puntaje menor al control, como se muestra en la figura 2, mostrando el receptor en gris, el ligando en azul y los residuos de aminoácidos del sitio activo en color morado.

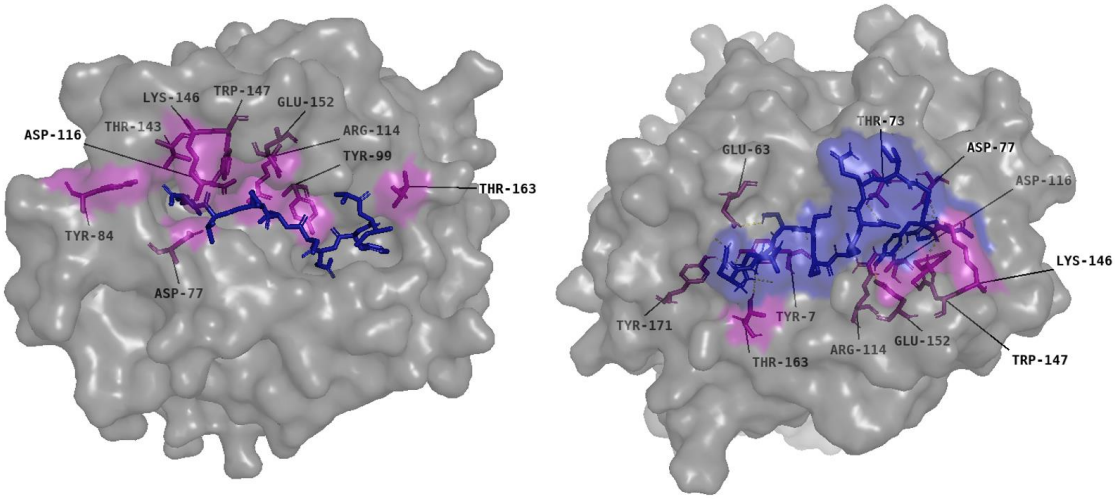

*Figura 2. Interacciones del HLA-A*0301 con su ligando control y péptido problema. Se muestra el ligando en color azul y los aa de interacción del receptor se resaltan en morado, se presenta el nombre de los aminoácidos que están en contacto con el sitio activo del receptor.*

7.8 Interacciones entre la proteína N y el huésped

La proteína N del SARS-CoV-2 es una proteína estructural de 46 kDa y puede dividirse en cinco Dominios: tres regiones intrínsecamente desordenadas y dos Dominios plegados; un Dominio intrínsecamente desordenado en el extremo N-terminal (NTD), un Dominio de unión al ARN (RBD) característico de la proteína N,

un enlace central predicho desordenado (LINK), un Dominio de dimerización y un Dominio intrínsecamente desordenado en el extremo C-terminal (CTD) (Cubuk et al., 2020) tal y como se muestra en la figura 3.

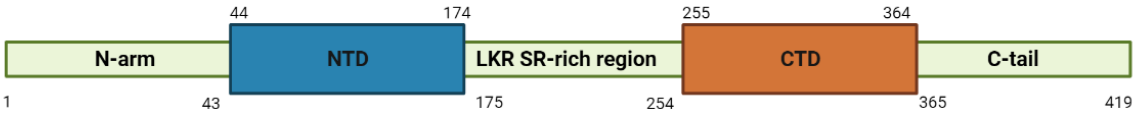

Figura 3. Representación esquemática de los Dominios de la nucleocápside del SARS-CoV-2. Se muestran el Dominio N-terminal (NTD), el Dominio C-terminal (CTD) y tres regiones intrínsecamente desordenadas (IDRs), es decir, el brazo N (N-arm), la región de enlace (LKR) y la cola C (C-tail).

Las proteínas tienen estados estructurales que incluyen Dominios globulares ordenados y regiones proteicas intrínsecamente desordenadas que existen como conjuntos conformacionales altamente flexibles cuando están aisladas (Dosztanyi, 2018; Dosztanyi et al., 2005; Meszaros et al., 2018).

Cubuk, et al., combinaron la espectroscopía de molécula única con simulaciones a escala atómica para obtener una descripción por residuos de las tres regiones desordenadas en la proteína N de SARS-CoV-2 con el fin de entender el conjunto conformacional de la proteína N y sus interacciones internas. Este estudio encontró que la proteína N contiene regiones altamente dinámicas y transitorias. Bajo simulaciones a escala atómica, se describe que el Dominio NTD contiene sitios de interacción atractivos y repulsivos, y que la lámina beta básica del Dominio RBD repele la región C-terminal rica en arginina del NTD. Mientras tanto, un residuo de fenilalanina en el NTD interactúa con el RBD a través de una cara hidrofóbica. Además, la región C-terminal del NTD contiene una región rica en arginina que forma una hélice alfa transitoria, hélice 2 (H2), proyectando tres residuos de arginina en la misma dirección.

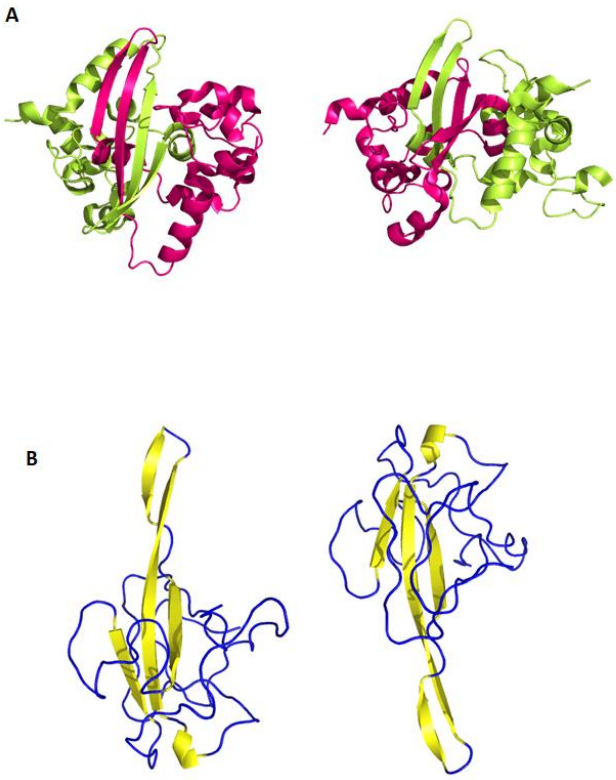

Figura 4. Estructura de la proteína N del SARS-CoV-2. A) Representación de cinta del dímero de N-CTD, donde un monómero está coloreado en rosa y el otro en verde. PDB: 6YUN. B) Representación de cinta del monómero de N-NTD. PDB: 7NU.

También encontraron dos regiones en el Dominio NTD que forman hélices transitorias, hélice 3 y 4 (H3 y H4), que contienen regiones ricas en serina-arginina, conocidas por mediar en interacciones proteína-proteína y proteína-ácido nucleico. Además, H2 y H3 flanquean el RBD y organizan un conjunto de residuos de arginina en la misma dirección, sugiriendo que estas hélices podrían facilitar la unión del ARN a la proteína N. La región rica en serina-arginina es necesaria para el reclutamiento de proteínas relacionadas con la replicación-transcripción, y se prefiere que esta región esté fosforilada, quizás permitiendo que la proteína N se una a los Dominios de las proteínas del huésped para llevar a cabo sus funciones en el secuestro de la maquinaria celular (Cubuk et al., 2020).

La N-NTD de los CoVs interactúa con el extremo 3' del ARN genómico viral, posiblemente a través de interacciones electrostáticas. Además, se han identificado varios residuos en el Dominio N-terminal como motivos de unión al ARN; lo que demuestra la relevancia de esta proteína en los procesos de replicación y supervivencia del virus al interior de la célula huésped(Grossoehme et al., 2009; Keane et al., 2012; Tan et al., 2006).

En nuestro análisis podemos observar que los epítopes identificados se encuentran principalmente en la región NTD de la proteína N y que éstos interactúan con receptores HLA (Tabla 2). Además, se observa que estos epítopes tienden a estar en regiones de tipo desordenadas dentro de la proteína N. Analizando a detalle estos epitopes, observamos que el epítope pep 15 de secuencia (KAYNVTQAFGRRGPE), que corresponde a la expresión regular (regexp) "...([ST])Q.." se encuentra asociado a interacciones entre Dominio y motivo relacionadas con procesos metabólicos, oxidación de compuestos orgánicos y generación de precursores de metabolitos y energía; aportando al panorama que tenemos actualmente sobre la interacción entre las proteínas virales y los Dominios de proteínas del huésped, además, el pep 15 identificado en este estudio, se encuentra desordenado en más de un 60%, lo que explicaría la flexibilidad y capacidad de éste para "imitar" las proteínas del huésped y activar procesos celulares que favorecen la replicación y supervivencia del virus (MacRaild et al., 2018).

7.9 Proteína nucleocápside como un potencial blanco terapéutico

Dada la función de la proteína N en la unión al ARN, el ensamblaje viral, la replicación viral y la respuesta inmune del huésped, y su expresión elevada durante las infecciones virales, la proteína N es un candidato prometedor para el desarrollo de nuevas estrategias terapéuticas (McBride et al., 2014; Tan et al., 2020; Verheije et al., 2010). Un estudio reciente realizado por Peng et al. identificó sitios conservados de unión al ARN entre MERS-CoV y SARS-CoV-2. Debido a que el

Dominio N-NTD contiene sitios de unión al ARN, el bloqueo de estas interacciones puede ser una estrategia antiviral interesante. Este estudio sugiere a PJ34 como una estrategia antiviral potencial, ya que PJ34 se une al sitio de unión al ARN en el Dominio N-NTD, imitando la unión de AMP al N-NTD. Además, Peng et al. también sugirieron que el 5-benzyloxi-gramine (P3) es una estrategia antiviral prometedora. Sus simulaciones mostraron que P3 se une al NTD, inhibiendo la oligomerización de la proteína N y la formación de la ribonucleoproteína (RNP)(Peng et al., 2008).

Además, un estudio realizado por Cai et al. sugiere que la proteína N puede regular el crecimiento de los gránulos de estrés en desarrollo. La proteína N alberga motivos RGG/RG, y se ha observado que estos motivos son metilados por proteínas arginin-metil transferasa tipo 1 (PMRT1). Cai et al. informaron que la metilación en R95 y R177 es un requisito para la asociación de la proteína N con ARN, lo que destaca su papel como blanco farmacológico. Su trabajo evaluó el efecto del inhibidor de la PRMT1, MS023, en la replicación de SARS-CoV-2, y encontraron que MS023 disminuyó el número de partículas virales en las células huésped infectadas con SARS-CoV-2(Cai et al., 2021).

Otra estrategia novedosa que se está explorando para investigar el potencial de compuestos activos para inhibir proteínas virales durante la infección es el reposicionamiento de fármacos. Ahamad et al. evaluaron medicamentos antivirales como posibles tratamientos para la enfermedad COVID-19 al inhibir la oligomerización de la proteína N, encontrando cinco compuestos con una unión prometedora al Dominio CTD(Ahamad et al., 2020). Aunque se requiere trabajo experimental, este tipo de datos computacionales es una indicación excelente para centrar la investigación futura en la búsqueda de nuevas estrategias terapéuticas para la enfermedad COVID-19.

Además, la proteína N es una proteína altamente inmunogénica, ya que puede inducir respuestas inmunológicas tanto humorales como celulares(McBride et al., 2014; Yasui et al., 2008). Debido a su potencial inmunogénico y su alta producción durante la infección, la proteína N es un excelente blanco para el desarrollo de vacunas. Varios estudios de vacunología inversa han dirigido la proteína N como un

posible inmunógeno para el desarrollo de vacunas(Kumar et al., 2021; Peng et al., 2008). Aunque los resultados experimentales de estos análisis inmunoinformáticos aún deben confirmarse, los datos computacionales proponen un gran potencial en la investigación y desarrollo de vacunas.

Las interacciones huésped-patógeno se estudian principalmente a través de interacciones proteína-proteína (PPI). Mientras que las proteínas humanas tienden a competir por sitios de unión de Dominio bajo similitud de secuencia, las proteínas virales tienden a competir por los sitios de unión de Dominio en ausencia de similitud de secuencia. Los virus utilizan motivos lineales cortos, para interactuar con Dominios en proteínas huésped(Garamszegi et al., 2013; Martinez et al., 2021). Estos motivos lineales actúan como sitios de reconocimiento para el anclaje proteolítico y proporcionan la especificidad necesaria para las modificaciones postraduccionales involucradas en los procesos biológicos en la célula huésped. Además, estos motivos lineales también funcionan como sitios de unión en procesos de señalización y regulación en el ciclo de vida viral(Davey et al., 2011). Los motivos lineales se encuentran principalmente en regiones desordenadas dentro de las proteínas virales (Diella et al., 2008); la proteína N contiene regiones altamente desordenadas que pueden contener los motivos lineales cortos necesarios para las interacciones con proteínas huésped necesarias para la unión al ARN. Además, la proteína N ha experimentado una mutación, en la región LKR, una región desordenada. Esta región es rica en residuos de serina/arginina con alta flexibilidad para interacciones proteína-proteína, se ha encontrado una mutación con el cambio de residuos de arginina (R203K/G204R); G204R introduce un aminoácido básico adicional en la región LKR, aumentando su carga positiva y, por lo tanto, mejorando la aptitud en la RNP con la proteína N y acelerando así la replicación viral (Wu et al., 2021). Mientras que la mayor atención se ha centrado en la proteína Spike para el diseño de vacunas y tratamientos, la proteína N ha sido parcialmente ignorada. La proteína N no solo es altamente inmunogénica y modula la respuesta inmune, sino que también es crucial para el secuestro y mimetismo del huésped a través de

las IPPs, y se debería prestar más atención a futuras investigaciones enfocadas en esta proteína. Por lo tanto, la proteína N puede tener algunas ventajas como candidata a vacuna para el SARS-CoV-2. La proteína N puede aumentar la inmunogenicidad y es una estrategia prometedora para el diseño de medicamentos debido a sus funciones en la respuesta inmune y el ciclo de vida del virus. Independientemente de las áreas restantes por explorar e investigar sobre la proteína N del SARS-CoV-2, se requiere más trabajo para dilucidar nuestra comprensión de los procesos biológicos e inmunológicos que implica, mejorando el diseño actual de vacunas y tratamientos contra COVID-19. Además, se debería prestar mucha más atención a las IPPs para aclarar los mecanismos exactos involucrados en la infección por SARS-CoV-2 para el secuestro y mimetismo de las células huésped.

8. REFERENCIAS BIBLIOGRAFICAS

Ahamad, S., Gupta, D., & Kumar, V. (2020). Targeting SARS-CoV-2 nucleocapsid oligomerization: Insights from molecular docking and molecular dynamics simulations. *J Biomol Struct Dyn*, 1-14. https://doi.org/10.1080/07391102.2020.1839563

Ahn, Y. M., Maddumage, J. C., Grant, E. J., Chatzileontiadou, D. S. M., Perera, W., Baker, B. M., Szeto, C., & Gras, S. (2024). The impact of SARS-CoV-2 spike mutation on peptide presentation is HLA allomorph-specific. *Curr Res Struct Biol, 7*, 100148. https://doi.org/10.1016/j.crstbi.2024.100148

Arslan, H. (2021, Nov). COVID-19 prediction based on genome similarity of human SARS-CoV-2 and bat SARS-CoV-like coronavirus. *Comput Ind Eng, 161*, 107666. https://doi.org/10.1016/j.cie.2021.107666

Basu, M. K., Poliakov, E., & Rogozin, I. B. (2009, May). Domain mobility in proteins: functional and evolutionary implications. *Brief Bioinform, 10*(3), 205-216. https://doi.org/10.1093/bib/bbn057

BIOVIA. (2020). *Discovery Studio Visualizer.* In (Version v21.1.0.20298) [Software de computadora]

Boni, M. F., Lemey, P., Jiang, X., Lam, T. T., Perry, B. W., Castoe, T. A., Rambaut, A., & Robertson, D. L. (2020, Nov). Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat Microbiol, 5*(11), 1408-1417. https://doi.org/10.1038/s41564-020-0771-4

Brito, A. F., & Pinney, J. W. (2017). Protein-Protein Interactions in Virus-Host Systems. *Front Microbiol, 8*, 1557. https://doi.org/10.3389/fmicb.2017.01557

Cai, T., Yu, Z., Wang, Z., Liang, C., & Richard, S. (2021). Arginine methylation of SARS-Cov-2 nucleocapsid protein regulates RNA binding, its ability to suppress stress granule formation, and viral replication. *J Biol Chem, 297*(1), 100821. https://doi.org/10.1016/j.jbc.2021.100821

Cubuk, J., Alston, J. J., Incicco, J. J., Singh, S., Stuchell-Brereton, M. D., Ward, M. D., Zimmerman, M. I., Vithani, N., Griffith, D., Wagoner, J. A., Bowman, G. R., Hall, K. B., Soranno, A., & Holehouse, A. S. (2020). The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA. *bioRxiv*. https://doi.org/10.1101/2020.06.17.158121

Davey, N. E., Trave, G., & Gibson, T. J. (2011). How viruses hijack cell regulation. *Trends Biochem Sci, 36*(3), 159-169. https://doi.org/10.1016/j.tibs.2010.10.002

Diella, F., Haslam, N., Chica, C., Budd, A., Michael, S., Brown, N. P., Trave, G., & Gibson, T. J. (2008). Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front Biosci, 13*, 6580-6603. https://doi.org/10.2741/3175

Dimitrov, I., Flower, D. R., & Doytchinova, I. (2013). AllerTOP--a server for in silico prediction of allergens. *BMC Bioinformatics, 14 Suppl 6*(Suppl 6), S4. https://doi.org/10.1186/1471-2105-14-S6-S4

Dosztanyi, Z. (2018). Prediction of protein disorder based on IUPred. *Protein Sci, 27*(1), 331-340. https://doi.org/10.1002/pro.3334

Dosztanyi, Z., Csizmok, V., Tompa, P., & Simon, I. (2005). The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol, 347*(4), 827-839. https://doi.org/10.1016/j.jmb.2005.01.071

Forni, D., Cagliani, R., Clerici, M., & Sironi, M. (2017). Molecular Evolution of Human Coronavirus Genomes. *Trends Microbiol, 25*(1), 35-48. https://doi.org/10.1016/j.tim.2016.09.001

Garamszegi, S., Franzosa, E. A., & Xia, Y. (2013). Signatures of pleiotropy, economy and convergent evolution in a domain-resolved map of human-virus protein-protein interaction networks. *PLoS Pathog, 9*(12), e1003778. https://doi.org/10.1371/journal.ppat.1003778

Garcia-Perez, C. A., Guo, X., Navarro, J. G., Aguilar, D. A. G., & Lara-Ramirez, E. E. (2018). Proteome-wide analysis of human motif-domain interactions mapped on influenza a virus. *BMC Bioinformatics, 19*(1), 238. https://doi.org/10.1186/s12859-018-2237-8

Goodsell, D. S., & Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins, 8*(3), 195-202. https://doi.org/10.1002/prot.340080302

Gordon, D. E., Jang, G. M., Bouhaddou, M., Xu, J., Obernier, K., White, K. M., O'Meara, M. J., Rezelj, V. V., Guo, J. Z., Swaney, D. L., Tummino, T. A., Huttenhain, R., Kaake, R. M., Richards, A. L., Tutuncuoglu, B., Foussard, H., Batra, J., Haas, K., Modak, M., Kim, M., Haas, P., Polacco, B. J., Braberg, H., Fabius, J. M., Eckhardt, M., Soucheray, M., Bennett, M. J., Cakir, M., McGregor, M. J., Li, Q., Meyer, B., Roesch, F., Vallet, T., Mac Kain, A., Miorin, L., Moreno, E., Naing, Z. Z. C., Zhou, Y., Peng, S., Shi, Y., Zhang, Z., Shen, W., Kirby, I. T., Melnyk, J. E., Chorba, J. S., Lou, K., Dai, S. A., Barrio-Hernandez, I., Memon, D., Hernandez-Armenta, C., Lyu, J., Mathy, C. J. P., Perica, T., Pilla, K. B., Ganesan, S. J., Saltzberg, D. J., Rakesh, R., Liu, X., Rosenthal, S. B., Calviello, L., Venkataramanan, S., Liboy-Lugo, J., Lin, Y., Huang, X. P., Liu, Y., Wankowicz, S. A., Bohn, M., Safari, M., Ugur, F. S., Koh, C., Savar, N. S., Tran, Q. D., Shengjuler, D., Fletcher, S. J., O'Neal, M. C., Cai, Y., Chang, J. C. J., Broadhurst, D. J., Klippsten, S., Sharp, P. P., Wenzell, N. A., Kuzuoglu-Ozturk, D., Wang, H. Y., Trenker, R., Young, J. M., Cavero, D. A., Hiatt, J., Roth, T. L., Rathore, U., Subramanian, A., Noack, J., Hubert, M., Stroud, R. M., Frankel, A. D., Rosenberg, O. S., Verba, K. A., Agard, D. A., Ott, M., Emerman, M., Jura, N., von Zastrow, M., Verdin, E., Ashworth, A., Schwartz, O., d'Enfert, C., Mukherjee, S., Jacobson, M., Malik, H. S., Fujimori, D. G., Ideker, T., Craik, C. S., Floor, S. N., Fraser, J. S., Gross, J. D., Sali, A., Roth, B. L., Ruggero, D., Taunton, J., Kortemme, T., Beltrao, P., Vignuzzi, M., Garcia-Sastre, A., Shokat, K. M., Shoichet, B. K., & Krogan, N. J. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature, 583*(7816), 459-468. https://doi.org/10.1038/s41586-020-2286-9

Grossoehme, N. E., Li, L., Keane, S. C., Liu, P., Dann, C. E., 3rd, Leibowitz, J. L., & Giedroc, D. P. (2009). Coronavirus N protein N-terminal domain (NTD) specifically binds the transcriptional regulatory sequence (TRS) and melts TRS-cTRS RNA duplexes. *J Mol Biol, 394*(3), 544-557. https://doi.org/10.1016/j.jmb.2009.09.040

Gupta, S., Kapoor, P., Chaudhary, K., Gautam, A., Kumar, R., Open Source Drug Discovery, C., & Raghava, G. P. (2013). In silico approach for predicting toxicity of peptides and proteins. *PLoS One, 8*(9), e73957. https://doi.org/10.1371/journal.pone.0073957

He, R., Leeson, A., Ballantine, M., Andonov, A., Baker, L., Dobie, F., Li, Y., Bastien, N., Feldmann, H., Strocher, U., Theriault, S., Cutts, T., Cao, J., Booth, T. F., Plummer, F. A., Tyler, S., & Li, X. (2004). Characterization of protein-protein interactions between the nucleocapsid protein and membrane protein of the SARS coronavirus. *Virus Res, 105*(2), 121-125. https://doi.org/10.1016/j.virusres.2004.05.002

Itzhaki, Z. (2011). Domain-domain interactions underlying herpesvirus-human protein-protein interaction networks. *PLoS One, 6*(7), e21724. https://doi.org/10.1371/journal.pone.0021724

Keane, S. C., Liu, P., Leibowitz, J. L., & Giedroc, D. P. (2012). Functional transcriptional regulatory sequence (TRS) RNA binding and helix destabilizing determinants of murine hepatitis virus (MHV) nucleocapsid (N) protein. *J Biol Chem, 287*(10), 7063-7073. https://doi.org/10.1074/jbc.M111.287763

Khorsand, B., Savadi, A., & Naghibzadeh, M. (2020). SARS-CoV-2-human protein-protein interaction network. *Inform Med Unlocked, 20*, 100413. https://doi.org/10.1016/j.imu.2020.100413

Kumar, M., Michael, S., Alvarado-Valverde, J., Zeke, A., Lazar, T., Glavina, J., Nagy-Kanta, E., Donagh, J. M., Kalman, Z. E., Pascarelli, S., Palopoli, N., Dobson, L., Suarez, C. F., Van Roey, K., Krystkowiak, I., Griffin, J. E., Nagpal, A., Bhardwaj, R., Diella, F., Meszaros, B., Dean, K., Davey, N. E., Pancsa, R., Chemes, L. B., & Gibson, T. J. (2024). ELM-the Eukaryotic Linear Motif resource-2024 update. *Nucleic Acids Res, 52*(D1), D442-D455. https://doi.org/10.1093/nar/gkad1058

Kumar, V., Kancharla, S., Kolli, P., & Jena, M. (2021). Reverse vaccinology approach towards the in-silico multiepitope vaccine development against SARS-CoV-2. *F1000Res, 10*, 44. https://doi.org/10.12688/f1000research.36371.1

Li, Y., Yang, X., Wang, N., Wang, H., Yin, B., Yang, X., & Jiang, W. (2020). The divergence between SARS-CoV-2 and RaTG13 might be overestimated due to the extensive RNA modification. *Future Virology, 15*(6), 341-347. https://doi.org/10.2217/fvl-2020-0066

Lin, M. M., & Zewail, A. H. (2012). Hydrophobic forces and the length limit of foldable protein domains. *Proc Natl Acad Sci U S A, 109*(25), 9851-9856. https://doi.org/10.1073/pnas.1207382109

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., Bi, Y., Ma, X., Zhan, F., Wang, L., Hu, T., Zhou, H., Hu, Z., Zhou, W., Zhao, L., Chen, J., Meng, Y., Wang, J., Lin, Y., Yuan, J., Xie, Z., Ma, J., Liu, W. J., Wang, D., Xu, W., Holmes, E. C., Gao, G. F., Wu, G., Chen, W., Shi, W., & Tan, W. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet, 395*(10224), 565-574. https://doi.org/10.1016/S0140-6736(20)30251-8

MacRaild, C. A., Seow, J., Das, S. C., & Norton, R. S. (2018). Disordered epitopes as peptide vaccines. *Pept Sci (Hoboken), 110*(3), e24067. https://doi.org/10.1002/pep2.24067

Mariano, G., Farthing, R. J., Lale-Farjat, S. L. M., & Bergeron, J. R. C. (2020). Structural Characterization of SARS-CoV-2: Where We Are, and Where We Need to Be. *Front Mol Biosci, 7*, 605236. https://doi.org/10.3389/fmolb.2020.605236

Martinez, Y. A., Guo, X., Portales-Perez, D. P., Rivera, G., Castaneda-Delgado, J. E., Garcia-Perez, C. A., Enciso-Moreno, J. A., & Lara-Ramirez, E. E. (2021). The analysis on the human protein

domain targets and host-like interacting motifs for the MERS-CoV and SARS-CoV/CoV-2 infers the molecular mimicry of coronavirus. *PLoS One, 16*(2), e0246901. https://doi.org/10.1371/journal.pone.0246901

McBride, R., van Zyl, M., & Fielding, B. C. (2014). The coronavirus nucleocapsid is a multifunctional protein. *Viruses, 6*(8), 2991-3018. https://doi.org/10.3390/v6082991

Meszaros, B., Erdos, G., & Dosztanyi, Z. (2018). IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res, 46*(W1), W329-W337. https://doi.org/10.1093/nar/gky384

Mosca, R., Ceol, A., Stein, A., Olivella, R., & Aloy, P. (2014). 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res, 42*(Database issue), D374-379. https://doi.org/10.1093/nar/gkt887

Mousavizadeh, L., & Ghasemi, S. (2021). Genotype and phenotype of COVID-19: Their roles in pathogenesis. *J Microbiol Immunol Infect, 54*(2), 159-163. https://doi.org/10.1016/j.jmii.2020.03.022

Peng, T. Y., Lee, K. R., & Tarn, W. Y. (2008). Phosphorylation of the arginine/serine dipeptide-rich motif of the severe acute respiratory syndrome coronavirus nucleocapsid protein modulates its multimerization, translation inhibitory activity and cellular localization. *FEBS J, 275*(16), 4152-4163. https://doi.org/10.1111/j.1742-4658.2008.06564.x

Perrin-Cocon, L., Diaz, O., Jacquemin, C., Barthel, V., Ogire, E., Ramiere, C., Andre, P., Lotteau, V., & Vidalain, P. O. (2020). The current landscape of coronavirus-host protein-protein interactions. *J Transl Med, 18*(1), 319. https://doi.org/10.1186/s12967-020-02480-z

Rojas Chavez, R. A., Fili, M., Han, C., Rahman, S. A., Bicar, I. G. L., Gregory, S., Helverson, A., Hu, G., Darbro, B. W., Das, J., Brown, G. D., & Haim, H. (2024). Mapping the Evolutionary Space of SARS-CoV-2 Variants to Anticipate Emergence of Subvariants Resistant to COVID-19 Therapeutics. *PLoS Comput Biol, 20*(6), e1012215. https://doi.org/10.1371/journal.pcbi.1012215

Schrödinger, L. *The PyMOL Molecular Graphics System.* In (Version Version 1.2r3pre) [Software de computadora]

Song, Z., Xu, Y., Bao, L., Zhang, L., Yu, P., Qu, Y., Zhu, H., Zhao, W., Han, Y., & Qin, C. (2019). From SARS to MERS, Thrusting Coronaviruses into the Spotlight. *Viruses, 11*(1). https://doi.org/10.3390/v11010059

Tan, W., Lu, Y., Zhang, J., Wang, J., Dan, Y., Tan, Z., He, X., Qian, C., Sun, Q., Hu, Q., Liu, H., Ye, S., Xiang, X., Zhou, Y., Zhang, W., Guo, Y., Wang, X.-H., He, W., Wan, X., Sun, F., Wei, Q., Chen, C., Pan, G., Xia, J., Mao, Q., Chen, Y., & Deng, G. (2020). Viral Kinetics and Antibody Responses in Patients with COVID-19. *medRxiv*, 2020.2003.2024.20042382. https://doi.org/10.1101/2020.03.24.20042382

Tan, Y. W., Fang, S., Fan, H., Lescar, J., & Liu, D. X. (2006). Amino acid residues critical for RNA-binding in the N-terminal domain of the nucleocapsid protein are essential determinants for the infectivity of coronavirus in cultured cells. *Nucleic Acids Res, 34*(17), 4816-4825. https://doi.org/10.1093/nar/gkl650

Verheije, M. H., Hagemeijer, M. C., Ulasli, M., Reggiori, F., Rottier, P. J., Masters, P. S., & de Haan, C. A. (2010). The coronavirus nucleocapsid protein is dynamically associated with the replication-transcription complexes. *J Virol, 84*(21), 11575-11579. https://doi.org/10.1128/JVI.00569-10

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., & Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Res, 47*(D1), D339-D343. https://doi.org/10.1093/nar/gky1006

Wei, C. H., Allot, A., Leaman, R., & Lu, Z. (2019). PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res, 47*(W1), W587-W593. https://doi.org/10.1093/nar/gkz389

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., Hu, Y., Tao, Z. W., Tian, J. H., Pei, Y. Y., Yuan, M. L., Zhang, Y. L., Dai, F. H., Liu, Y., Wang, Q. M., Zheng, J. J., Xu, L., Holmes, E. C., & Zhang, Y. Z. (2020). A new coronavirus associated with human respiratory disease in China. *Nature, 579*(7798), 265-269. https://doi.org/10.1038/s41586-020-2008-3

Wu, H., Xing, N., Meng, K., Fu, B., Xue, W., Dong, P., Tang, W., Xiao, Y., Liu, G., Luo, H., Zhu, W., Lin, X., Meng, G., & Zhu, Z. (2021). Nucleocapsid mutations R203K/G204R increase the infectivity, fitness, and virulence of SARS-CoV-2. *Cell Host Microbe, 29*(12), 1788-1801 e1786. https://doi.org/10.1016/j.chom.2021.11.005

Yasui, F., Kai, C., Kitabatake, M., Inoue, S., Yoneda, M., Yokochi, S., Kase, R., Sekiguchi, S., Morita, K., Hishima, T., Suzuki, H., Karamatsu, K., Yasutomi, Y., Shida, H., Kidokoro, M., Mizuno, K., Matsushima, K., & Kohara, M. (2008). Prior immunization with severe acute respiratory syndrome (SARS)-associated coronavirus (SARS-CoV) nucleocapsid protein causes severe pneumonia in mice infected with SARS-CoV. *J Immunol, 181*(9), 6337-6348. https://doi.org/10.4049/jimmunol.181.9.6337