FACULTAD DE CIENCIAS AGRARIAS

UNIVERSIDAD NACIONAL DE ROSARIO

"Anotación automática GO de productos génicos en SARS-CoV-2"

Bioq. Elizabeth Chiacchiera

TRABAJO FINAL PARA OPTAR AL TÍTULO DE ESPECIALISTA EN BIOINFORMÁTICA

Director: Dr. Flavio E. Spetale

2023

“Anotación automática GO de productos génicos en SARS-CoV-2”

Chiacchiera, Elizabeth

Bioquímica - UNR

Este trabajo final es presentado como parte de los requisitos para optar al grado académico de Especialista en Bioinformática, de la Universidad Nacional de Rosario y no ha sido previamente presentada para la obtención de otro título en ésta u otra Universidad. El mismo contiene los resultados obtenidos en investigaciones llevadas a cabo en CIFASIS-CONICET-UNR, durante el período comprendido entre septiembre de 2021 y abril de 2023 bajo la dirección del Dr. Flavio E. Spetale.

Bioq. Chiacchiera Elizabeth

Dr. Flavio E. Spetale

Defendida: .....de 2023.

Agradecimientos

Este trabajo está dedicado a todas las víctimas de la enfermedad producida por el virus del Covid 19, en especial a mi colega y papá Norberto Chiacchiera que lo extraño mucho.
A la Universidad pública, en especial a la UNR, a mis compañeros de especialidad y su cuerpo docente. Estoy infinitamente agradecida con el lujo de director que estuvo a cargo que se involucró para que esto se concrete superando mis expectativas.
A mi familia, que es el motivo de mi existencia.
Confío plenamente en Dios y espero que este aporte al conocimiento generado cumpla un propósito mayor.

Publicaciones y Presentaciones a Congresos

- Chiacchiera E., Tapia E. Spetale F.E. “Automatic GO prediction of proteins on SARS-CoV-2”. XI CAB2C. Argentina, 2021.
- Spetale F.E., Chiacchiera E., Iglesias N., Tapia E., Ponce S., Bulacio P. “Automatic GO annotation of gene products in SARS-CoV-2". XXIV Argentino de Bioingeniería y XIII Jornadas de Ingeniería Clínica - SABI 2023 (en revisión).

Abreviaturas y Símbolos

GO: Gene Ontology
FGGA: Factor Graph GO Annotation
SVM: Máquinas de Vectores Soporte
BP: Procesos Biológicos
MF: Funciones Moleculares
CC: Componente o localización Celular

Resumen

La anotación de funcionalidades biológicas de productos génicos, RNA y proteínas, es una tarea crítica en el desarrollo de proyectos de secuenciación genómica. En el caso de proyectos de genomas virales, estas anotaciones infieren el rol molecular de estos productos virales de interés durante la infección a sus células diana, indicando aquellos procesos biológicos en los que están involucrados y constituyen una herramienta útil para el desarrollo y mejoramiento de tratamientos antivirales. La velocidad actual a la que se generan nuevas secuencias de RNA y proteínas a partir de proyectos de secuenciación genómica genera un cuello de botella para los métodos de anotación tradicionales, basados en estudios experimentales exhaustivos. Este cuello de botella puede resolverse parcialmente mediante métodos computacionales de anotación. Es de interés global el estudio de virus y en particular, SARS-CoV-2, que causa la enfermedad COVID-19 y representa aún una amenaza para la salud mundial. Los esfuerzos para desarrollar medicamentos y vacunas eficaces frente a nuevas variantes se ven obstaculizados por el conocimiento limitado de los detalles moleculares de cómo el SARS-CoV-2 infecta y se propaga. En particular, en este trabajo se aborda el problema de anotación funcional automática de productos génicos para SARS-CoV-2 a través de ontologías y aprendizaje computacional. La ontología funcional de genes utilizada es Gene Ontology (GO) y el método de aprendizaje computacional utilizado se llama Factor Graph GO Annotation (FGGA). Este método de clasificación jerárquico toma como entrada un conjunto de atributos, características, extraídos desde las secuencias y devuelve un grafo consistente en los tres subdominios de GO. El proceso de extracción de atributos desde las secuencias se lo denomina caracterización. En este trabajo, se considera una caracterización básica que consiste en propiedades fisicoquímicas y una caracterización enriquecida, desarrollada en este proyecto, que agrega atributos virales. La incorporación de estos contribuye a mejorar la especificidad de predicción de las funcionalidades GO. Finalmente, se evalúa el rendimiento de las predicciones GO obtenidas y se compara los resultados obtenidos sobre 31 productos génicos anotados en forma experimental en Jungreis et al. (2021). Estos resultados validaron de forma exitosa las anotaciones existentes curadas manualmente y también generaron nuevas anotaciones in-silico que fueron avaladas por diversas fuentes bibliográficas disponibles en la actualidad.

Palabras clave: Funciones Biológicas, Aprendizaje Computacional, SARS-CoV-2

Abstract

The annotation of biological functionalities of gene products, RNA and proteins, is a critical task in the development of genomic sequencing projects. In the case of viral genome projects, these annotations infer the molecular role of these viral products of interest during the infection of their target cells, indicating the biological processes in which they are involved and constitute a useful tool for the development and improvement of antivirals treatments. The current rate at which new RNA and protein sequences are generated from genomic sequencing projects creates a bottleneck for traditional annotation methods, based on extensive experimental studies. This bottleneck can be partially resolved by computational annotation methods. The study of viruses is of global interest and, in particular, SARS-CoV-2, which causes the disease COVID-19 and still represents a threat to global health. Efforts to develop effective drugs and vaccines are hampered by limited knowledge of the molecular details of how SARS-CoV-2 infects cells. In particular, this paper addresses the problem of automatic functional annotation of gene products for SARS-CoV-2 through ontologies and computational learning. The functional gene ontology used is Gene Ontology (GO) and the computational learning method used is called Factor Graph GO Annotation (FGGA). This hierarchical classification method takes as input a set of attributes, characteristics, extracted from the sequences and returns a graph consisting of the three GO subdomains. The process of extracting attributes from the sequences is called characterization. In this work, a basic characterization is considered, which consists of physicochemical properties, and an enriched characterization, developed in this project, which adds viral attributes. The inclusion of these attributes contributes to improving the specificity of GO functionality predictions. Finally, the performance of the obtained GO predictions is evaluated and compared with the results obtained for 31 gene products experimentally annotated in Jungreis et al. (2021). These results successfully validated existing hand-curated annotations and also generated new in-silico annotations that were supported by various currently available literature sources.

Índice

1. Introducción

En esta sección se describirán los conceptos fundamentales para la realización del presente trabajo. En primer lugar, se realizará un breve estado del arte de la anotación de funciones biológicas en virus, particularizando en "SARS-CoV-2". En segundo lugar, se definirá la ontología Gene Ontology que serán las funciones biológicas a predecir. Por último, se detallará la metodología de análisis computacional utilizada para la predicción automática.

1.1. Anotación de funciones biológicas en virus

Los avances de las tecnologías de secuenciación de alto rendimiento (NGS) ocurridos durante la última década han impulsado el desarrollo de proyectos de secuenciación en una gran diversidad de organismos (Reuter et al., 2015; Kircher and Kelso, 2010). Durante este período, los costos y tiempos de estos proyectos han experimentado una reducción exponencial. Pero este éxito es parcial, sólo una fracción reducida de los datos generados por estos proyectos puede ser interpretada desde alguna dimensión biológica. Esto implica que para la mayoría de las secuencias obtenidas sólo es posible acceder de forma rápida a su potencial codificante y que sus anotaciones específicas (función biológica, ubicación celular o participación en algún proceso biológico) deben esperar estudios experimentales complementarios y/o curado experto por literatura. Esta situación se acentúa en proyectos de secuenciación de organismos no modelo. En soja (*Glycine max*), el porcentaje de anotación de genes que codifican para proteínas sólo alcanza el 65 % (Brown et al., 2020). La situación se repite en proyectos de secuenciación de genomas virales, críticos para la comprensión de mecanismos de infección y progresión de enfermedades que tienen como agentes etiológicos virus (Zhang et al., 2019), este entendimiento es clave para el posterior desarrollo de vacunas o tratamientos antivirales eficaces. La forma más precisa para anotar genes/proteínas virales es la realización de ensayos experimentales. No obstante, estos requieren mucho tiempo y dinero. Además, por las limitaciones de las complejas técnicas de laboratorio, se pueden omitir algunos genes que se expresan sólo en condiciones específicas del hospedador difíciles de replicar in vitro. En este contexto, los métodos computacionales para la anotación automática viral desempeñan un rol fundamental como instrumentos de asistencia y referencia en los procesos de anotación experimental. De forma estándar, los métodos computacionales para la anotación automatizada de genes que codifican proteínas comúnmente se basan en búsquedas de similitud de secuencia (Altschul et al., 1990; Dong and Strous, 2019) o de motivos (Lee et al., 2007; Blum et al., 2020). Sin embargo, la aplicación de este tipo de métodos para secuencias virales altamente divergentes es problemática, especialmente por las tasas de mutación inherentemente altas de algunos tipos virales como los virus de genoma de ARN (Skewes-Cox et al., 2014). Las tasas de mutación de estos están en el rango de 10-3 a 10-6 errores de copia por nucleótido incorporado en el producto de ARN incipiente (Domingo et al., 2021). Por ende, se deben considerar estrategias alternativas. Una opción es la utilización de técnicas ómicas de alto rendimiento para identificar de redes humano-virus de interacción de proteínas (PPI) para luego inferir las anotaciones funcionales faltantes (Szklarczyk et al., 2020; Lian et al., 2020). Sin embargo, estos métodos requieren interacciones altamente fiables junto a datos de anotación de base experimental (Lian et al., 2021). Estas dificultades motivan el desarrollo de métodos de anotación basados en técnicas del aprendizaje computacional que explotan ontologías biológicas como Gene

Ontology (GO) (Ashburner and et al, 2000) para la predicción de funcionalidades partiendo de sus secuencias proteicas (Jain and Kihara, 2018; Huerta-Cepas et al., 2018; Kieft et al., 2020). Estas ontologías, de carácter esencialmente jerárquico, se presentan bajo la forma de grafos acíclicos dirigidos (DAG) donde los nodos reflejan las funcionalidades admisibles para cada producto génico y las conexiones reflejan las relaciones entre ellos. Los métodos actuales de anotación automática GO en virus se basan en métodos de filogenia. Debe notarse, sin embargo, que a pesar de la enorme cantidad de datos genómicos que se recopilan en todo el árbol de la vida, la inferencia filogenética está restringida sólo a una pequeña parte de estos datos. Brevemente, las limitaciones de estos métodos obedecen principalmente a fallas en el modelado de carácter analítico o biológico. Las fallas en el modelado analítico afectan a la reconstrucción filogenética y se reflejan en la elección de criterios de optimalidad inadecuados, en el muestreo insuficiente de taxones y en violaciones de los supuestos en los modelos de evolución de las secuencias (Som, 2014). Por otro lado, las fallas en el modelado biológico causan incongruencia entre filogenias y se reflejan en violaciones de la ortología entre secuencias causada por la clasificación incorrecta de linajes, paralogía oculta y transferencia horizontal de genes. También pueden ocurrir errores estocásticos o sesgo en el muestreo de caracteres relacionado con la longitud de los genes y errores sistemáticos debido a la presencia de una señal no filogenética en los datos (Som, 2014; Smith and Hahn, 2021). En conjunto, estas limitaciones motivan al desarrollo de métodos computacionales específicos para la anotación automática de secuencias en proyectos genómicos virales. En esta propuesta se plantea abordar la predicción automática de funciones biológicas de productos génicos virales. Más específicamente, del Coronavirus 2 del Síndrome Respiratorio Agudo Grave (SARS-CoV-2).

Debido a su rápida transmisión de persona a persona por la falta de inmunidad poblacional y de una terapia antiviral efectiva, el COVID-19 ha causado una pandemia con más de 661 millones de casos y más de 6,7 millones de muertes en todo el mundo [1]. El contexto mundial de pandemia era el principal factor condicionante para la selección del mismo, pero no el único. En particular: i) las herramientas actuales para la anotación funcional de secuencias genómicas no se han diseñado específicamente para el análisis de genomas virales como SARS-CoV-2 (Chiara et al., 2020), ii) su contenido genético no se ha resuelto por completo con varios marcos de lectura abierta (ORF) cuya función o incluso estado de codificación de las proteínas se desconoce (Jungreis et al., 2021), y iii) no existe un recurso sistemático para interpretar el impacto funcional de las mutaciones del SARS-CoV-2 y priorizar los impulsores candidatos que pueden subyacer a las diferencias fenotípicas entre las variantes (Jungreis et al., 2021). Los estudios realizados sobre variantes indican que pueden constituir diferencias importantes en la función o el potencial de plegamiento del producto proteico (Dimonaco et al., 2021). Por último, la anotación actual de SARS-CoV-2 de las seis proteínas accesorias (3a, 6, 7a, 7b, 8 y 10, NC_045512.2) sus funcionalidades no se han confirmado experimentalmente (Finkel et al., 2021). Estas consideraciones, motivaron al siguiente objetivo general de investigación: *Predicción automática de funcionalidades biológicas sobre SARS-CoV-2 a partir de secuencias de productos génicos.*

[1] https://covid19.who.int

1.2. SARS-CoV-2

Es un virus, un parásito intracelular obligado, que se comporta como una organela extracelular y por ende se lo considera un organismo no vivo. Es un agente infeccioso de células permisivas donde dirige la replicación de su genoma y la síntesis de los componentes del virión (forma extracelular, funcionalmente inactiva) utilizando el aparato biosintético celular (Gardiol, 2011). SARS-CoV-2 pertenece a la familia Coronaviridae, género Betacoronavirus, subgénero Sarbecovirus. Es un virus envuelto, que posee un genoma que consiste en una molécula de ARN de sentido positivo (5'-3') monocatenario de aproximadamente 29.900 nucleótidos (secuencia de referencia NCBI: NC 045512.2) dispuesta en 14 marcos de lectura abierta (ORF) que codifican 31 proteínas. Entre ellas, cuatro proteínas son estructurales: membrana (M), espiga (S), envoltura (E), que forman parte de la envoltura viral y nucleocápside (N) que contiene el genoma ARN, y además proteínas no estructurales y accesorias, para traducirlas utiliza la maquinaria biosintética de la célula huésped (Hogue and Machamer, 2007). ORF1a y ORF1ab codifican poliproteínas que se dividen proteolíticamente dando lugar a 16 proteínas no estructurales (nsp1 - nsp16). En la Figura 1 se muestra a qué nivel actúan las proteínas virales nsp del coronavirus inhibiendo la respuesta inmune del húesped.

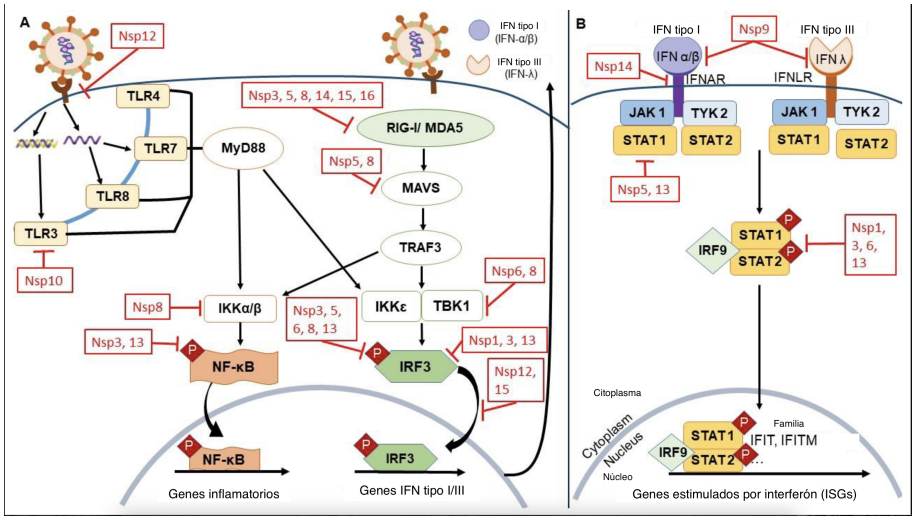

Figura 1: Descripción general de las proteínas no estructurales del SARS-CoV-2 que contribuyen al escape inmunitario del huésped (en recuadros rojos). (A) Los patrones moleculares asociados a patógenos (PAMP) del virus son reconocidos por varias células inmunitarias (macrófagos, monocitos, neutrófilos, células dendríticas y epiteliales). Tras la infección, estas células inmunitarias reconocen antígenos y moléculas virales extraños (PAMP) a través de receptores de reconocimiento de patrones (PRR), como los receptores tipo Toll (TLR) y los receptores tipo RIG-I (RLR), estimulando así las citoquinas y los IFN, y posteriormente induciendo respuestas inmunitarias del huésped. (B) Los IFN de tipo I y tipo III se producen y se unen a sus receptores de IFN de superficie celular específicos, lo que activa la señalización JAK/STAT para promover genes estimulados por IFN (ISG) para lograr respuestas antivirales (Low et al., 2022).

La proteína espiga o spike (S) está codificada por ORF2, sobre (E) por ORF4, membrana (M) por ORF5 y nucleocápside (N) por ORF9. 9 ORF adicionales codifican proteínas accesorias: ORF3a, ORF3b, ORF6, ORF7a, ORF7b, ORF8, ORF9b, ORF9c y ORF10 (ver Figura 2). El genoma del coronavirus es único entre los nidovirus porque codifica un número variable de proteínas accesorias cuya función no parece esencial para la replicación del virus, pero parece desempeñar un pa-

pel relevante en la patogénesis. Además, se han observado mutaciones en proteínas accesorias como ORF3a, ORF6, ORF7a, ORF8 o ORF10 en las variantes de preocupación que circulan actualmente, contribuyendo potencialmente al aumento de la patogénesis y la transmisibilidad en estas cepas del SARS-CoV-2 (Redondo et al., 2021). Muchas de las proteínas detalladas anteriormente no poseen funcionalidad demostrada experimentalmente.

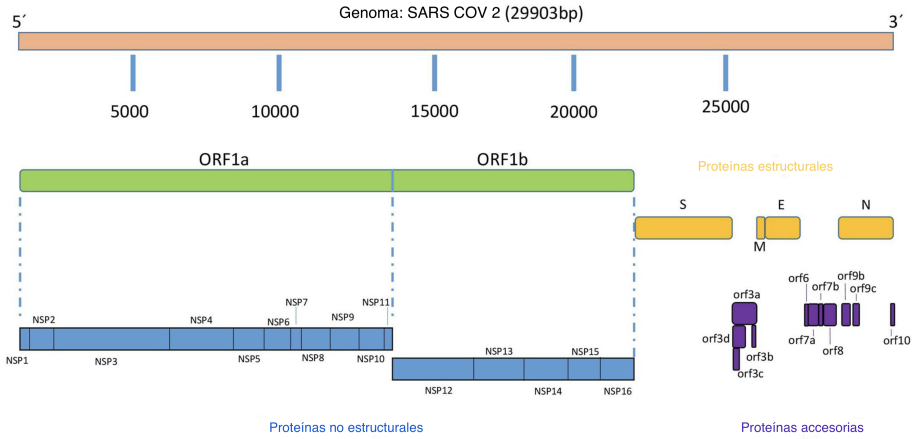

Figura 2: Organización genómica SARS-CoV-2. Las proteínas del SARS-CoV-2 comprenden dos grandes poliproteínas: ORF1a y ORF1ab que se divide por una proteasa codificada por virus en proteínas no estructurales complejas de replicación individual para formar 16 proteínas no estructurales (nsp 1-16) involucradas en la replicación del genoma y la regulación de la respuesta inmune.La replicación del ciclo de vida del SARS-CoV-2 comienza en las vesículas de doble membrana inducidas por virus derivadas del retículo endoplásmico (ER), que finalmente se integran para formar elaboradas redes de membranas enrevesadas. Aquí, el genoma de la cadena positiva entrante sirve como plantilla para el ARN de cadena negativa de longitud completa y el ARN subgenómico (sg). La traducción del ARN sg da como resultado tanto proteínas estructurales como proteínas accesorias (Redondo et al., 2021)

1.3. Gene Ontology

La ontología Gene Ontology (GO) se desarrolló para describir sistemáticamente las propiedades funcionales de los productos génicos entre especies y para facilitar la predicción computacional de la función génica (Ashburner and et al, 2000). Debido a que GO se actualiza de forma rutinaria, sirve como estándar y es hoy en día la principal fuente de conocimiento en genómica funcional. GO está formada por tres subdominios, sub-ontologías, que se corresponden con tres aspectos diferentes de la biología celular: *i*) Procesos Biológicos (BP), *ii*) Funciones Moleculares (MF) y *iii*) Componente o localización Celular (CC).

- Procesos Biológicos: son una serie conocida de eventos o funciones moleculares con un principio definido y final.
- Funciones Moleculares: son las tareas que hacen o las "habilidades" que tiene un producto génico. Los individuos (productos génicos) tienen diferentes habilidades o tareas (funciones) y trabajan en forma conjunta para alcanzar diferentes objetivos (procesos).
- Componentes Celulares: describe lugares en los niveles de las estructuras subcelulares y complejos macromoléculares. Además, incluye enzimas de múltiples

subunidades y otros complejos de proteínas pero no proteínas individuales o ácidos nucleicos.

En la Figura 3, se muestra la estructura GO, por ejemplo, para la subontología Procesos Biológicos. Los nodos, también llamados términos GO, representan los procesos biológicos relacionados con la inmunidad donde se encontraría actuando un gen codificante determinado y se identifican con un único término GO:xxxxxxx asociado al nombre de dichos procesos biológicos. Por otro lado, la relación existente entre los términos GO indica que el nodo hijo es un subtipo de sus nodos padres.

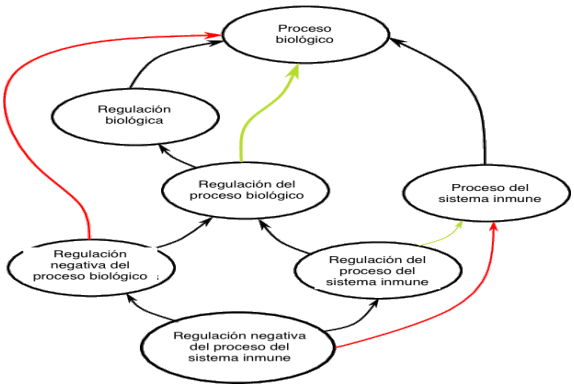

Figura 3: Subgrafo GO Biological Process donde GO:xxxxxxx representa los procesos biológicos y los enlaces representan las relaciones entre ellas.

Los términos se relacionan jerárquicamente: Un término más genérico ("término padre") puede tener asociados uno o más términos más específicos ("términos hijo"). Un término más específico ("término hijo") puede tener asociados uno o más términos menos específicos ("términos padre"). Siguiendo con el ejemplo de la subontología BP, un nodo padre podría ser: regulación biológica. Un nivel más abajo (específico) podría ser: regulación de procesos biológicos. Más específico, podría ser que dicho producto génico en cuestión involucre: regulación negativa del proceso del sistema inmunológico.

1.4. Método computacional FGGA utilizado

La predicción automática se realizó mediante un método computacional llamado FGGA -Factor Graph GO Annotation- diseñado para anotar genes que codifican para proteínas de organismos no modelo (Spetale et al., 2016, 2018). Brevemente, este método de clasificación jerárquica para la anotación automática de genes codificantes es del tipo ensamble, i.e. existen un conjunto de clasificadores basados en aprendizaje computacional que trabajan en forma conjunta para resolver un problema de clasificación donde las funciones biológicas a predecir están relacionadas. La Figura 4 muestra el diagrama de bloques del método FGGA donde la entrada es una secuencia de proteína y el grafo GO de referencia y la salida es un subgrafo GO consistente. El bloque de caracterización no forma parte del método y consiste en calcular un número fijo de atributos, características, a partir de la secuencia proteica. El bloque FG model genera un modelo gráfico de factores en base al grafo GO de referencia. El bloque de clasificadores individuales, en nuestro caso: Máquinas de Vectores Soporte (SVM), genera un modelo de predicción para cada función biológica. El bloque de Suma-Producto se encarga de obtener un subgrafo GO predicho consistente en base al modelo FG y las predicciones individuales de los SVMs

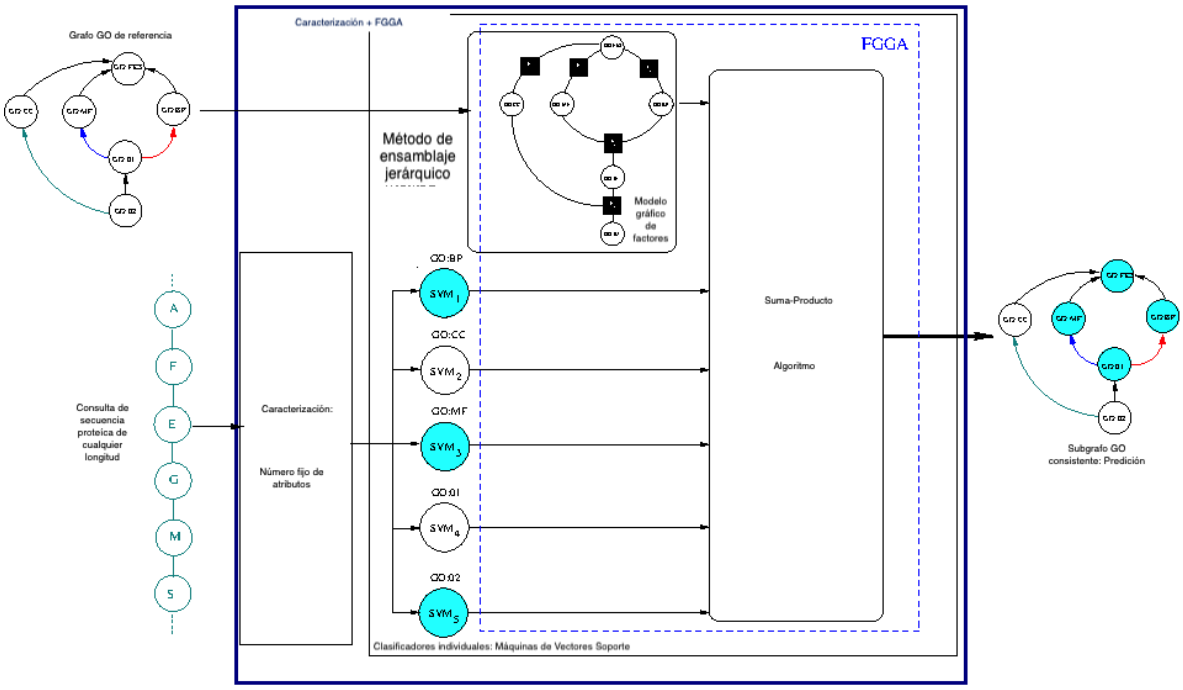

Figura 4: Método de ensamble jerárquico para la anotación automática de secuencias de genes codificantes. Factor Graph GO Annotation (FGGA)

mediante una versión modificada del algoritmo de pasaje de mensajes (Kschischang et al., 2001). Finalmente, esta metodología fue incorporada en un paquete de R llamado fgga y se encuentra disponible en el repositorio de Bioconductor (Spetale, 2022).

2. Objetivos

A continuación se detalla el objetivo general y los objetivos específicos para este trabajo final.
Objetivo general de investigación: *Predicción automática de funcionalidades biológicas sobre SARS-CoV-2 a partir de secuencias de productos génicos.*

Objetivos específicos:

1. **Base de datos y asociación de funciones GO**. Construcción de una base de datos de productos génicos de SARS-COV-2, secuencias y sus correspondientes anotaciones GO. Además, se propone relacionar las funcionalidades anotadas en Jungreis et al. (2021) para 31 productos génicos con las funciones de GO.

2. **Caracterización de datos**. Generación de un conjunto de atributos fijos relacionados con funciones virales a partir de las secuencias para enriquecer el modelo computacional.

3. **Predicción de productos génicos**. Aplicación del conjunto de datos caracterizados en un modelo de clasificación jerárquico basado en grafo de factores.

4. **Validación *in-silico***. Comparación del rendimiento sobre la predicción de los 31 productos génicos contra las anotaciones obtenidas en Jungreis et al. (2021).

3. Materiales y Métodos

Esta sección estará dividida en dos subsecciones. La primera presenta la construcción de la base de conocimiento. La segunda presenta el desarrollo de dos caracterizaciones de los datos, una general y otra enriquecida para virus.

3.1. Construcción de la base de conocimiento

Las secuencias proteicas de SARS-CoV-2, en formato fasta, fueron extraídas desde la base de datos Uniprot (Consortium, 2020) y sus anotaciones GO desde Gene Ontology (Ashburner and et al, 2000). Luego, se realizó una limpieza de las secuencias eliminando a todas aquellas que eran idénticas pero tenían diferentes nombres, Ids y/o poseían el carácter "X"; también fueron seleccionadas solo aquellas anotaciones GO que tenían tipo de evidencia[2]: electrónica, de similaridad o experimental. Los códigos de evidencia experimental considerados son: Inferencia Experimental (EXP), Inferencia de Ensayo Directo (IDA), Inferencia de Interacción Física (IPI), Inferencia de Fenotipo Mutante (IMP), Inferencia de Interacción Genética (IGI) e Inferencia de Patrón de Expresión (IEP). Mientras que los códigos de evidencia computacional considerados son: Inferencia de Secuencia o Similitud estructural (ISS), Inferencia de Ortología de la Secuencia (ISO), Inferencia de Alineación de la Secuencia (ISA) e Inferencia del Modelo de la Secuencia (ISM). Posteriormente, se seleccionaron aquellas funciones biológicas (GOs) cuya cantidad de anotaciones sea superior a 50 de forma garantizar una mínima cantidad de ejemplos positivos en cada clasificador individual (SVM). Finalmente, para ensamblar los conjuntos de datos de entrenamiento binarios convenientemente balanceados (Wei and Dunbrack, 2013), las secuencias de proteínas con anotaciones positivas se complementaron con contrapartes de proteínas con anotaciones negativas utilizando la política de separación inclusiva (Eisner et al., 2005). Este proceso da como resultado una base de datos curada de 3173 secuencias de SARS-CoV-2 con 211 funciones biológicas y se encuentra disponible en repositorio público[3].
Por otro lado, se construyó una segunda base de conocimiento, FiloDB, basada en los 32 productos génicos mencionados en Jungreis et al. (2021) para la comparación con el método computacional FGGA. Esta base de datos contiene por una lado, la secuencia proteica extraída desde la información suplementaria 2 del paper[4] y por otro, las funciones GO fueron curadas manualmente por la autora de este trabajo. El proceso de curación consistió en asociar las funcionalidades expresadas dentro de Jungreis et al. (2021) y vincularlas con funcionalidades GO en los tres subdominios que más se aproximaran desde el punto de vista biológico (ver Anexo I). Estas funcionalidades extraídas de Jungreis et al. (2021) fueron anotadas bajo el siguiente proceso. Primero, se detectaron las firmas evolutivas de codificación de proteínas y se distinguieron las regiones que evolucionan bajo la restricción de codificación de proteínas a través de la herramienta PhyloCSF (Lin et al., 2011). Brevemente, esta herramienta compara las sustituciones de codones y las frecuencias en alineaciones de genomas relacionados con modelos de evolución codificantes y no codificantes entrenados con datos del genoma completo, i.e., explota principalmente dos características principales de las firmas evolutivas de genes que codifican proteínas a lo largo del tiempo evolutivo: i), la preferencia por sustituciones sinónimas que preservan la traducción de aminoácidos y cambios conservativos de aminoácidos que

[2] http://geneontology.org/page/guide-go-evidence-codes
[3] https://www.cifasis-conicet.gov.ar/bioinformatica/dxCOVID.csv
[4] https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-021-22905-7/MediaObjects/41467_2021_22905_MOESM5_ESM.txt

preservan las propiedades biofísicas y *ii*) la evasión de codones de parada y las inserciones o eliminaciones que no sean múltiplos de tres, ya que interrumpirían el marco de lectura de la traducción. Específicamente, se cuantificó la restricción de codificación de proteínas calculando los valores del PhyloCSF para cada intervalo de tres nucleótidos en los tres marcos de lectura del genoma del SARS-CoV-2 utilizando 44 alineaciones del genoma completo de Sarbecovirus. Luego, se suavizaron estos valores utilizando un modelo oculto de Markov y se detectaron las regiones codificantes. Finalmente, para cada región detectada se consideraron las funciones biológicas existentes en SARS-CoV-2 dentro de la base de datos de Uniprot con cualquier tipo de evidencia. El proceso de curación manual arrojó como resultado final un total de 31 productos génicos debido a que se descartó el ORF2b ya que no hay una fuente confiable que proporcione la secuencia aminoacídica poniendo en evidencia la complejidad biológica y evolutiva que poseen los virus. Esta proteína accesoria posee una longitud de 39 codones y está codificada en un marco de lectura diferente (+1: el marco +1 indica que los codones se desplazan un nucleótido), en la dirección de 3′ de los codones en el ORF principal (más grande), que ocupa el marco +0. Por último, a este conjunto de validación se lo dividió en: proteínas no estructurales, estructurales y accesorias.

3.2. Caracterización de Datos

Esta etapa fue dividida en dos partes: *i*) caracterización general y *ii*) caracterización enriquecida. La caracterización general utiliza características, atributos básicos de las secuencias proteicas, mientras que, la caracterización enriquecida incorpora características relacionadas con los virus.

Caracterización general

A partir de las secuencias de proteínas se determina un número fijo de características utilizando dos métodos de caracterización llamada PhyChe^{+}: *i*) Signal^{+} y *ii*) fisicoquímica.

La caracterización Signal^{+} contiene 11 atributos extraídos desde la herramienta SignalP (Teufel et al., 2022) relacionados con la presencia de señales peptídicas y la ubicación de sus sitios de clivaje. Esta nueva versión del SignalP utiliza redes neuronales profundas para estimar sus predicciones.

La caracterización fisicoquímica estima 48 propiedades fisicoquímicas (Lee et al., 2009),(Chou and Fasman, 1974), 8 señales peptídicas obtenidas desde la herramienta SignalP (Teufel et al., 2022), 2 de estructura secundaria y la frecuencia de los aminoácidos la tomamos de forma individual y de a pares. En la práctica, esta caracterización fue desarrollada por el grupo de investigación en el lenguaje R y utiliza funciones del paquete Peptides (Osorio et al., 2015) para el cálculo de las propiedades fisicoquímicas. Este proceso dio como resultado una matriz de 3173 secuencias de SARS-CoV-2 con 479 atributos.

Caracterización enriquecida

Para poder enriquecer la caracterización con atributos relacionados con virus se realizó en primera instancia un análisis de las herramientas disponibles para determinar cuales son las más aptas para cumplir con la premisa que los atributos calculados puedan estar asociado a funciones biológicas. En la segunda instancia, se diseñó un método que es capaz de extraer información de esas herramientas seleccionadas y transformarlo en un vector de atributos fijo que será incorporado a la matriz de la

caracterización general.

A continuación se realiza una breve descripción conceptual de las herramientas analizadas.

1. **NetCorona 1.0** (Kiemer et al., 2004): Esta herramienta predice sitios de clivaje en proteínas de la proteasa 3CLpro del coronavirus (C30 endopeptidasa o Mpro) basada en redes neuronales artificiales a partir de secuencias de aminoácidos. La proteasa de tipo 3C es la principal proteasa de los coronavirus, corresponde a la proteína no estructural 5 (nsp5) y es capaz de escindir catalíticamente un enlace peptídico entre una glutamina en la posición P1 y un pequeño aminoácido (serina, alanina o glicina) en la posición P1'. Esta herramienta fue descartada porque solamente genera salidas en la poliproteìna P0DTC1 y P0DTD1, las proteínas de SARS-CoV-2 que utilizaremos incluyen aquellas que son el resultado de la escisión de las antes mencionadas (ver Figura 5).

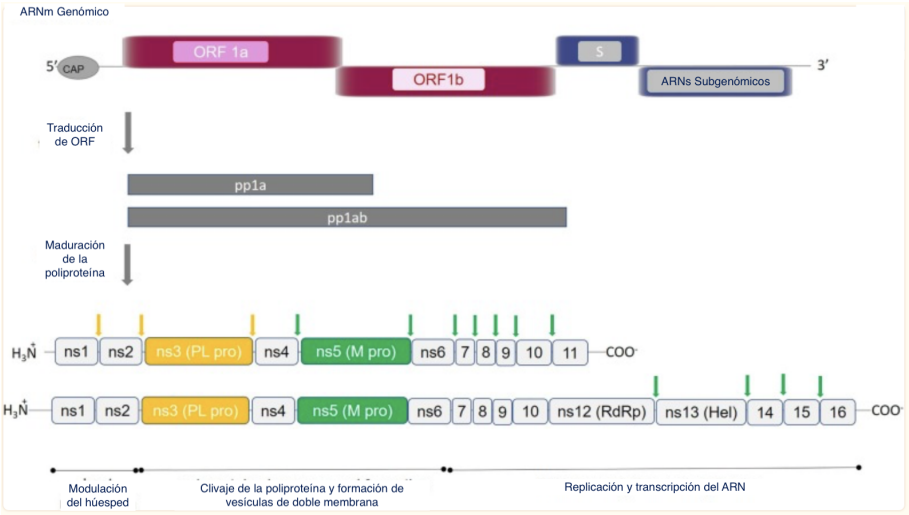

Figura 5: Poliproteínas SARS-CoV-2 codificadas por ORF1a y ORF1ab. Representación esquemática de los marcos de lectura abiertos 1a y 1ab, que codifican las poliproteínas pp1a y pp1ab. Se muestran las proteínas que componen cada poliproteína: (ns) indica proteínas no estructurales; La ARN polimerasa y la helicasa dependientes de ARN se indican mediante (RdRp) y (Hel), respectivamente. Los sitios proteolíticos escindidos por PLpro y Mpro se indican con flechas amarillas y verdes, respectivamente (Lam-Cabanillas et al., 2021).

2. **CPH models 3.2** (Nielsen et al., 2010): Esta herramienta es capaz de predecir la estructura 3D de proteínas con un modelo de estructura 3D desconocido donde el reconocimiento se basa en la alineación perfil-perfil guiado por la estructura secundaria y las predicciones de exposición. Esta herramienta fue descartada porque esta discontinuada.

3. **NetTurnP 1.0** (Petersen et al., 2010): Esta herramienta predice si un aminoácido se encuentra en una vuelta β o no, y discrimina los giros β en 9 tipos (I, I', II, II', IV, VIa1, VIa2, VIb y VIII) basados en ángulos ϕ, ψ, junto con una categoría IV miscelánea definidas en Hutchinson and Thornton (1994). NetTurnP es un método de aprendizaje computacional basado en 3 diferentes de redes neuronales artificiales (ANNs): giro β G (general), giro β S (específico) y giro β P (propensión). El método giro β G predice si un aminoácido se encuentra

en una región de giro β o no. El método giro β S predice si un aminoácido pertenece a alguno de los 9 tipos. El método giro β P predice si un aminoácido se encuentra en la posición 1, 2, 3 o 4 en un giro β. Cabe mencionar que algunos aminoácidos pueden ser asignados a múltiples posiciones dentro de una vuelta β. Esta herramienta fue seleccionada debido al rol importante que juega la formación de giros β en el plegamiento, la estabilidad de las proteínas y los procesos de reconocimiento molecular. Se ha descubierto una alta correlación entre la tendencia de una secuencia a formar un giro beta y la reactividad de las proteínas a los anticuerpos antipéptidos, y además se ha demostrado que hay una sobrerrepresentación de los giros beta en los epítopos de las células B (Petersen et al., 2010). Nota, la versión en línea permite correr hasta 500 secuencias por envío; mientras la versión de escritorio requiere descargar e instalar adicionalmente dos bases de datos de aproximadamente 4GB.

4. **DeepLoc 2.0** (Thumuluri et al., 2022): Esta herramienta predice la(s) localización(es) subcelular(es) de proteínas eucariotas. Las 10 localizaciones que pueden ser clasificada son diferentes: núcleo, citoplasma, extracelular, mitocondria, membrana celular, retículo endoplásmico, cloroplasto, aparato de Golgi, lisosoma/vacuola y peroxisoma. DeepLoc está basado en algoritmo de aprendizaje profundo entrenado con proteínas Uniprot que contienen evidencia experimental de sus localizaciones subcelulares, i.e., el modelo puede predecir si una proteína puede estar en una o varias localizaciones dentro de la célula eucariota partiendo de su secuencia. Las posiciones en la secuencia con un alto valor de probabilístico se consideran más relevantes para la predicción. Esto no significa que un aminoácido en particular sea muy importante para la predicción sino que una región en la vecindad de esas posiciones tiene más peso en la predicción final del modelo. Esta herramienta fue seleccionada debido a su capacidad predecir localizaciones subcelulares en distintos compartimentos de la célula infectada. Nota, la versión en línea permite correr hasta 500 secuencias por envío; mientras la versión de escritorio requiere un mínimo de 32GB RAM para funcionar. También la secuencia bajo estudio debe ser mayor que 10 y menor que 6000 aminoácidos.

5. **BepiPred 3.0** (Clifford et al., 2022): Esta herramienta predice epítopos de células B tanto lineales como discontinuos a partir de secuencias de proteínas. BepiPred está basado en un algoritmo de aprendizaje profundo que utiliza representaciones numéricas del modelo de lenguaje de proteínas ESM-2 para la generación de aminoácidos epítopos y no epítopos determinados a partir de estructuras cristalinas. El epítopo de células B de la proteína de superficie viral puede unirse específicamente al receptor de antígeno de células B del huésped e inducir al cuerpo a producir anticuerpos protectores y una respuesta inmunitaria humoral. Este modelo fue entrenado basándose en un conjunto de datos epítopedos validados experimentalmente derivados de 649 estructuras cristalinas antígeno-anticuerpo del Banco de Datos de Proteínas (PDB) Berman et al. (2000). Esta herramienta fue seleccionada porque el descubrimiento de epítopos es útil para el desarrollo de la vacuna contra el SARS-CoV-2 y la comprensión de la patogénesis del mismo Lon et al. (2020). Nota, la versión en línea permite correr hasta 50 secuencias por envío; mientras la versión de escritorio requiere un mínimo de 32GB RAM para funcionar y puede correr en CPU como en GPU. También la secuencia bajo estudio debe ser mayor que 10 y menor que 6000 aminoácidos, sin embargo, si es mayor que 1023 son truncadas

a esa longitud.

6. **NetSurfP 3.0** (Høie et al., 2022): Esta herramienta predice la accesibilidad del solvente, la estructura secundaria, el desorden estructural y los ángulos diedros ϕ/ψ de los aminoácidos en una secuencia proteica. NetSurfP está basado en una red neuronal convolucional profunda que utiliza representaciones numéricas del modelo de lenguaje de proteínas ESM-1b. El modelo fue entrenado con un conjunto de datos secuencias de 12185 estructuras cristalinas del PDB cuya resolución sea superior a 2,5 Å. Esta herramienta fue seleccionada porque es capaz de predecir la composición estructural del virión siendo la partícula viral madura consiste básicamente de un bloque de material genético rodeado de proteínas que lo protegen del medio ambiente y le sirven como vehículo para permitir su transmisión de una célula a otra. Nota, la versión en línea permite correr hasta 10000 secuencias o 10MB del archivo fasta por envío; mientras la versión de escritorio requiere un mínimo de 64GB RAM para funcionar y puede correr en CPU como en GPU. También la secuencia bajo estudio debe ser mayor que 10 y menor que 5000 aminoácidos.

viralFE: Método de caracterización para secuencias virales

En este trabajo se diseñó e implementó un método de caracterización, *viralFE*, que toma secuencias de proteínas virales y las transforma en un vector de atributos fijo utilizando las herramientas antes mencionadas. El método *viralFE* une las salidas de 6 funciones generadas donde cada una de ellas esta asociada a una herramienta. A continuación, se detallan las funciones, algoritmos, desarrolladas:

1. **NetTurnP 1.0**: Esta función toma como entrada el archivo generado por la herramienta y retorna una matriz de $nx20$ donde n son la cantidad de proteínas evaluadas y 20 son los atributos generados. Estas 20 variables se dividen en dos grupos de 10 atributos para ser calculados. Cada grupo considera el giro β y sus 9 tipos: I, I', II, II', IV, VIa1, VIa2, VIb y VIII. El grupo 1 cuenta la cantidad de giro β y de sus 9 tipos dentro de toda la secuencia proteica; mientras que el grupo 2 estima la cantidad de aminoácidos involucrados en cada una de estas 10 clases y luego la divide por la longitud de la secuencia. De esta forma obtenemos información sobre la existencia del giro β o de sus 9 tipos más su proporción dentro de la secuencia que se ve involucrada. La Tabla 1 muestra la cantidad de giros β y sus 9 tipos en 5 proteínas. Nota, esta herramienta en línea demora aproximadamente 8 horas para 50 secuencias.

Tabla 1: Cantidad de giros β y sus 9 tipos en 5 proteínas luego de aplicar la función generada a NetTurnP

Id	#β	#I	#I'	# II	#II'	#IV	#VIa1	#VIa2	#VIb	#VIII
A0A6M3A2H9	1	2	1	1	0	1	1	1	2	3
A0A6M3A2I0	5	7	6	36	289	0	0	285	7	234
A0A6M3A2I7	13	10	5	6	7	12	4	3	7	10
A0A6M3A2K2	2	1	0	0	0	1	1	0	2	3
A0A6M3A2J3	82	63	45	45	62	75	5	71	38	62

2. **Deeploc 2.0**: Esta función toma como entrada el archivo generado por la herramienta y retorna una matriz de $nx10$ donde n son la cantidad de proteínas evaluadas y 10 son las probabilidades asociadas a cada de las 10 localizaciones celulares definidas. La Tabla 2 muestra los resultados obtenidos en 5 proteínas a aplicar la función generada.

Tabla 2: Resultados obtenidos luego de aplicar la función generada a Deeploc en 5 proteínas. Localizaciones: núcleo (Nu), citoplasma (Ci), extracelular (Ex), mitocondria (Mi), membrana celular (Me-Ce), retículo endoplásmico (Re-En), cloroplasto (Cl), aparato de Golgi (Go), lisosoma/vacuola (Li/Va) y peroxisoma (Pe)

Id	Nu	Ci	Ex	Mi	Me-Ce	Re-En	Cl	Go	Li/Va	Pe
A0A6M3A2H9	0.41	0.48	0.10	0.54	0.22	0.16	0.12	0.17	0.06	0.08
A0A6M3A2I0	0.41	0.56	0.33	0.54	0.27	0.11	0.25	0.15	0.04	0.02
A0A6M3A2I7	0.37	0.54	0.24	0.55	0.36	0.08	0.34	0.16	0.03	0.03
A0A6M3A2K2	0.41	0.48	0.110	0.54	0.22	0.16	0.12	0.17	0.06	0.08
A0A6M3A2J3	0.45	0.56	0.31	0.47	0.24	0.11	0.21	0.14	0.03	0.03

3. **Bepipred 3.0**: Esta función toma como entrada el archivo generado por la herramienta y retorna una matriz de $nx2$ donde n son la cantidad de proteínas evaluadas y 2 son los atributos generados. El primer atributo indica el promedio de todas aquellas probabilidades superiores a 0.15 y el segundo atributo indica la proporción de la secuencia que se ve afecta a un epítope.

Tabla 3: Resultados obtenidos luego de aplicar la función generada a Bepipred en 5 proteínas.

Id	mean(Probabilidad)	%epítope
A0A6M3A2H9	0.18	0.29
A0A6M3A2I0	0.18	0.22
A0A6M3A2I7	0.18	0.23
A0A6M3A2K2	0.18	0.29
A0A6M3A2J3	0.18	0.18

La Tabla 3 muestra los resultados obtenidos en 5 proteínas a aplicar la función generada.

4. **NetSurfP 3.0**: Esta función toma como entrada el archivo generado por la herramienta y retorna una matriz de $nx25$ donde n son la cantidad de proteínas evaluadas y 25 son los atributos generados. El atributo RSA indica la accesibilidad relativa al solvente promedio; ASA indica la accesibilidad absoluta al solvente promedio; ϕ indica el promedio de los ángulos diédricos ϕ para cada aminoácido normalizado por la longitud de la secuencia; ψ indica el promedio de los ángulos diédricos ψ para cada aminoácido normalizado por la longitud de la secuencia; desorden (DO) indica el desorden promedio. Los atributos relacionados con las estructuras secundarias se dividen en dos grupos de acuerdo a Kabsch and Sander (1983). El grupo 1 considera los 3 estados generales de las estructuras secundarias: hélice (H), hebra (E) y espiral (C); mientras que el grupo 2 considera la extensión de los 8 estados posibles de las estructuras secundarias: 3_{10} hélice (G), hélice α(H), hélice π(I), hebra extendida en conformación paralela y/o anti-paralela (E), lámina β (B), giro de puente de hidrógeno (T), doblez (S) y espiral (C). En ambos grupos se calculan la cantidad de estados existentes y la proporción que ocupa cada estado dentro de la secuencia. La Tabla 4 muestran los resultados obtenidos en 5 proteínas a aplicar la función generada sin considerar los 8 estados posibles de estructuras secundarias.

4. Resultados y Discusión

Esta sección estará dividida en tres subsecciones. La primera presenta las métricas de rendimiento utilizadas en el trabajo para evaluar la calidad de las predicciones obtenidas. La segunda presenta los resultados obtenidos utilizando el método FGGA

Tabla 4: Resultados obtenidos luego de aplicar la función generada a NetSurfP en 5 proteínas.

Id	RSA	ASA	ϕ	ψ	DO	#H	avg(H)	#E	avg(E)	#C	avg(C)
A0A6M3A2H9	0.25	0.36	-0.51	0.03	0.03	2	0.05	5	0.64	7	0.31
A0A6M3A2I0	0.22	0.37	-0.58	-0.01	0.03	4	0.61	2	0.06	6	0.33
A0A6M3A2I7	0.23	0.36	-0.55	0.04	0.04	4	0.58	2	0.06	6	0.36
A0A6M3A2K2	0.25	0.36	-0.51	0.03	0.03	5	0.64	2	0.05	7	0.31
A0A6M3A2J3	0.22	0.40	-0.62	0.08	0.00	4	0.59	2	0.06	6	0.35

con las caracterizaciones general y enriquecida en datos de test. La tercera presenta los predicciones obtenidas utilizando el método FGGA con la caracterización general en datos de validación y su comparación con la base de datos FiloDB.

4.1. Métricas de rendimiento

Teniendo en cuenta las relaciones jerárquicas entre los términos GO objetivo, es necesario considerar métricas de rendimiento de clasificación específicas (Kiritchenko et al., 2005). Los errores de predicción cerca de la raíz de la jerarquía deberían castigarse con más severidad que aquellos en niveles más profundos (Kosmopoulos et al., 2015). Con este objetivo, las métricas de rendimiento de la clasificación jerárquica como las medidas de precisión jerárquica (HP), la exhaustividad jerárquico (HR) y el F-score jerárquica (HF) (Verspoor et al., 2006) reconocen adecuadamente las clasificaciones parcialmente correctas y penalizan en consecuencia los errores más distantes o más superficiales. Las fórmulas de las métricas jerárquicas se muestran a continuación:

$$HP(s) = \frac{1}{\mid l(P_G(s)) \mid} \sum_{q \in l(P_G(s))} \max_{c \in l(C_G(s))} \frac{\mid \uparrow c \cap \uparrow q \mid}{\uparrow q}$$

$$HR(s) = \frac{1}{\mid l(C_G(s)) \mid} \sum_{c \in (C_G(s))} \max_{q \in l(P_{rG}(s))} \frac{\mid \uparrow c \cap \uparrow q \mid}{\uparrow c}$$

$$HF(s) = \frac{2 \cdot HP \cdot HR}{HP + HR}$$

donde s es una secuencia de producto génico, G es el subgrafo GO, $P_G(s) \subset G$ es el subgrafo GO predicho de s, $C_G(s) \subset G$ es el subgrafo GO real de s, $l(P_G(s))$ es el conjunto de hojas de $P_G(s)$ y $l(C_G(s))$ es el conjunto de hojas de $C_G(s)$. Además, $\uparrow q$ es el conjunto de ancestros de un nodo q perteneciente a $P_G(s)$, y $\uparrow c$ es el conjunto de ancestros de un nodo c perteneciente a $C_{GO}(s)$. Se utilizaron métricas planas (no jerárquicas) de precisión promedio, exhaustividad promedio y F-score promedio de rendimiento. Específicamente, para cada secuencia de producto génico s, la precisión $P(s)$ se calculó como $\frac{tp(s)}{tp(s)+fp(s)}$, la exhaustividad $R(s)$ como $\frac{tp(s)}{tp(s)+fn(s)}$, la exactitud $Acc(s)$ como $\frac{tp(s)+tn(s)}{tp(s)+fp(s)+fn(s)+fn(s)}$ y el F-score $F(s)$ como $\frac{2 \cdot p(s) \cdot r(s)}{p(s)+r(s)}$, donde tp es el número de términos GO predichos correctamente como positivos (verdaderos positivos), fp es el número de términos GO predichos incorrectamente como positivos (falsos positivos), fp es el número de términos GO predichos correctamente como negativos (verdaderos negativos) y fn es el número de términos GO predichos incorrectamente como negativos (falsos negativos).

Adicionalmente, se utilizó el test no paramétrico de Friedman para determinar si los métodos de caracterización propuestos son significativos, mientras que se utilizó el test no paramétrico de prueba de suma de rangos de Wilcoxon con corrección de

Bonferroni para comparar los resultados obtenidos en las métricas individuales en tres funcionalidades GO virales.

4.2. Rendimiento de las predicciones GO en datos de test

Las predicciones de SARS-CoV-2 utilizando FGGA con la caracterización general, PhyChe$^+$, se evaluaron utilizando un enfoque de validación cruzada con k = 5 y métricas de rendimiento jerárquicas HP, HR y HF. Para el conjunto de 479 términos GO (BP-CC-MF) analizados, se obtuvo un 90 %, 93 % y 91 en HP, HR y HF, respectivamente. Dichos resultados fueron publicados en el congreso XI Chiacchiera et al. (2021).

La Tabla 5 muestra los resultados promedio de HP, HR, HF, P, R, F y Acc obtenidos aplicando el método FGGA con las caracterizaciones general y enriquecida. En primer lugar, se puede observar que la caracterización combinada, viralFE + PhyChe$^+$, presenta la leve mejora respecto de la caracterización individual PhyChe$^+$ aunque estos resultados no presentan una relevancia significativa ($p < 0.01$; prueba de Friedman). Sin embargo, debido a que la base de conocimiento no contiene funciones GO virales por la falta de secuencias en SARS-CoV-2 no pudimos ver si nuestra caracterización combinada realmente mejora en este tipo de funciones especificas como se preveía. Por tal motivo, se realizaron pruebas adicionales sobre funciones GO virales utilizando organismos cercanos al SARS-CoV-2 como: *Coronavirus de murciélago*, *Coronavirus de roedor*, *Coronavirus de paloma-dominante* y *Coronavirus de pato-dominante*. Tres funcionalidades GO, GO:0039694, GO:0046813 y GO:0075509, fueron seleccionadas a partir de las anotaciones más especificas obtenidas en la base de datos FiloDB.

Tabla 5: Resultados obtenidos por el método FGGA utilizando las diferentes caracterizaciones

Caracterización	HP	HR	HF	P	R	F	Acc
PhyChe$^+$	0.94	0.95	0.95	0.94	0.95	0.94	0.96
viralFE	0.90	0.90	0.90	0.91	0.94	0.92	0.92
viralFE + PhyChe$^+$	0.95	0.96	0.95	0.96	0.95	0.96	0.97

Por ejemplo, la Figura 6 muestra que la función biológica GO:0039694, replicación del genoma de ARN viral, seleccionada es una anotación especifica de los virus en el subdominio BP. Esto dio como resultado un nuevo conjunto de datos de 443 secuencias con 3 funciones GO virales para los ejemplos positivos y con funciones GO generales para los ejemplos negativos. Luego este conjunto fue caracterizado con ambos métodos y aplicado a un clasificador SVM binario para evaluar la influencia de la caracterización combinada en los niveles de exactitud, precisión, exhaustividad y F-score en el rendimiento del clasificador.

Los resultados observados en la Tabla 6 confirman la hipótesis anterior de la incorporación de características atribuibles a virus mejora la predicción en funcionalidades GO virales.

4.3. Rendimiento de la predicciones GO en datos de validación

La Tabla 7 muestra los resultados de las predicciones obtenidas con el método FGGA y las anotaciones GO curadas obtenidas desde Jungreis et al. (2021). A continuación, se hace un análisis comparativo de las predicciones obtenidas con

[5] https://www.ebi.ac.uk/QuickGO/term/GO:0039694

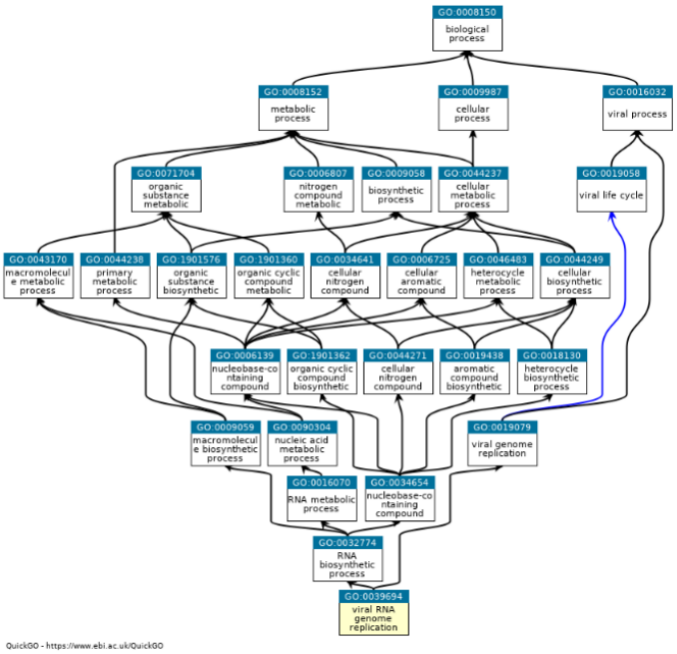

Figura 6: Grafo GO correspondiente a la funcionalidad replicación del genoma de ARN viral (GO:0039694) en el subdominio BP[5].

el FGGA y las funciones GOs curadas manualmente divididas según la categoría definida en 3.1.

4.3.1. Proteínas no estructurales

1. *ORF1a*: Poliproteína multifuncional que está involucrada en la transcripción, en la replicación del ARN viral y contiene a las proteínas responsables de su clivaje. De las predicciones obtenidas por el FGGA, nos enfocaremos en las funciones GO:0003968 (actividad polimerasa de ARN dependiente de ARN) y GO:0004175 (actividad endopeptidasa). Estas dos funciones predichas incluyen a las funciones GO experimentales GO:0032774 (proceso de biosíntesis de ARN) y GO:0008233 (actividad de peptidasa), i.e., son predicciones más específicas (ver Figuras 7 y 8).

   La funcionalidad GO:0003727 (unión de ARN monocatenario) predicha podría considerarse como correcta ya que su definición incluye el sustrato ARN propia de este proceso.

2. *ORF1ab*: Poliproteína multifuncional más grande, que está involucrada en la transcripción y en la replicación del ARN viral y contiene las proteinasas responsables de las escisiones de la poliproteína. De las predicciones obtenidas, nos enfocaremos en las funciones GO:0003968 (actividad polimerasa de ARN dependiente de ARN) y GO:0004175 (actividad endopeptidasa). Estas dos funciones predichas incluyen a las funciones GO experimentales GO:0032774 (proceso de biosíntesis de ARN) y GO:0008233 (actividad de peptidasa), i.e., son predicciones más específicas (ver Figuras 7 y 8). Las funciones GO:0001960 (regulación negativa de la vía de señalización mediada por citoquinas) y GO:0002832 (respuesta negativa al estímulo biótico) predichas involucradas con la desactivación de las defensas del huésped coinciden con la bibliografía O'Donoghue et al.

Tabla 6: El promedio de la Precisión (Prec), el exhaustividad (Rec), el F-score (Fscore) y la exactitud (Acc) de SVM método con kernel radial. Los términos GOs son: GO:0039694, GO:0046813 y GO:0075509. Las caracterizaciones son PhyChe^{+}, viralFE y PhyChe^{+} + ViralFE. Para cada término GO y caracterización, el método con mejor rendimiento según la prueba de suma de rangos de Wilcoxon con corrección de Bonferroni ($p_{value} = 0{,}01$) se muestra en negrita.

Término GO	Caracterización	Acc	Prec	Rec	Fscore
GO:0039694	PhyChe^{+}	0.89	0.96	0.91	0.94
	viralFE	0.92	0.93	0.98	0.95
	PhyChe^{+} + viralFE	**0.94**	0.96	0.97	**0.97**
GO:0046813	PhyChe^{+}	0.93	0.95	0.98	0.97
	viralFE	0.94	0.94	1.00	0.97
	PhyChe^{+} + viralFE	**0.96**	0.98	0.98	**0.98**
GO:0075509	PhyChe^{+}	0.93	0.96	0.96	0.96
	viralFE	0.94	0.94	1.00	0.97
	PhyChe^{+} + viralFE	**0.96**	0.98	0.99	**0.98**

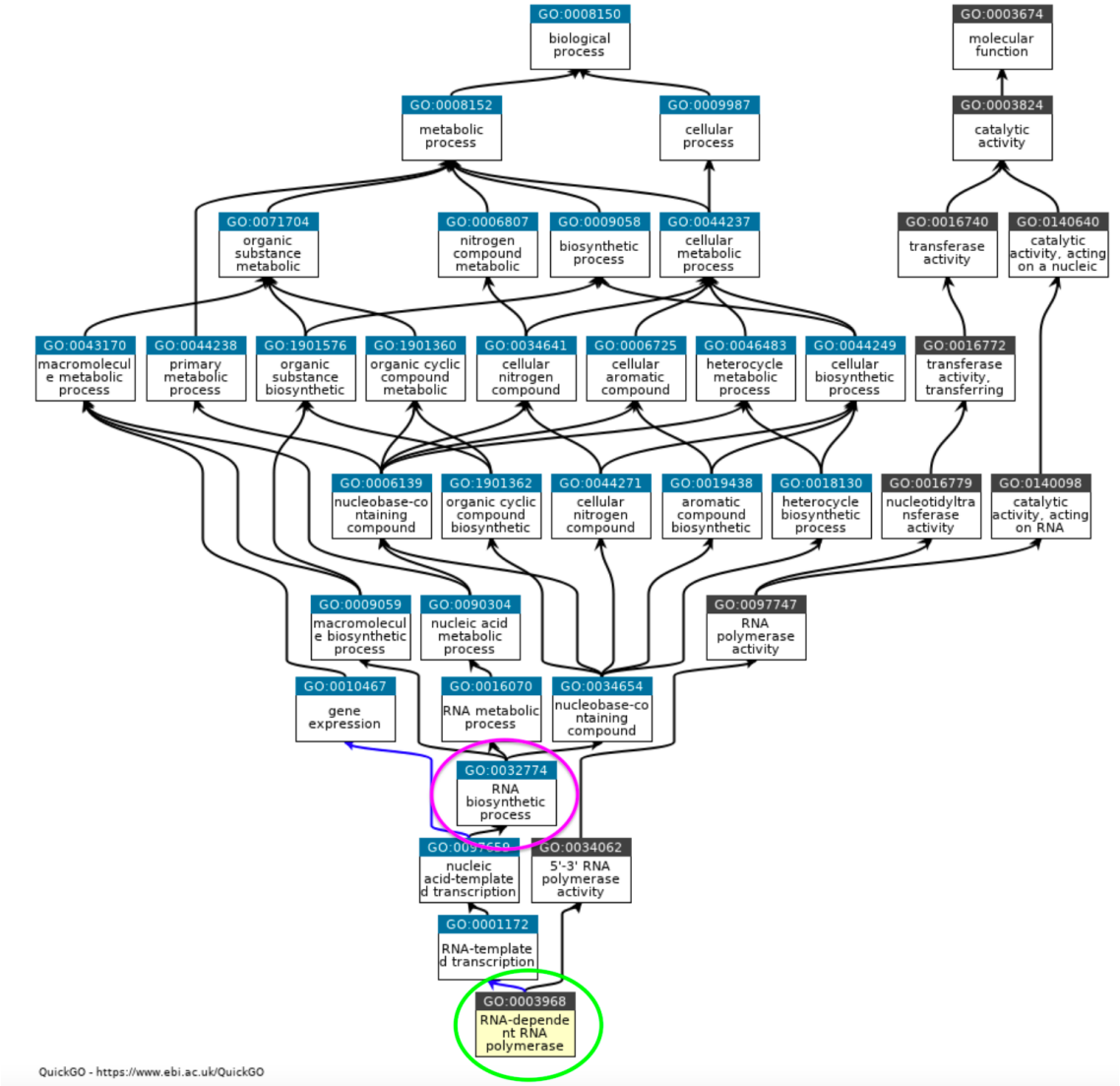

Figura 7: Grafo predicho para la proteína ORF1a. La funcionalidad más específica (GO:0003968) se encuentra identificada con un círculo verde. La funcionalidad GO experimental (GO:0032774) se encuentra identificada con círculo rosa.

Tabla 7: Anotación GO de SARS-CoV-2 con FGGA

Id	Términos GOs esperados	Términos GOs hoja predichos
ORF1ab	GO:0032774, GO:0008233	GO:0001960, GO:0002832, GO:0003724, GO:0003727, GO:0003968, GO:0004175, GO:0004482, GO:0004483, GO:0004519, GO:0004532, GO:0004843, GO:0005198, GO:0005216, GO:0005515, GO:0032559, GO:0035639
ORF3a	GO:0039656, GO:0019076, GO:0005216, GO:0039707, GO:0097194	GO:0001172, GO:0001510, GO:0004519, GO:0005737, GO:0006351, GO:0006370, GO:0006401, GO:0009896
ORF6	GO:0030683	GO:0001960, GO:0006955, GO:0008408, GO:0008757, GO:0019222, GO:0048518
ORF7a	GO:0019079	GO:0001172, GO:0001960, GO:0002682, GO:0005737, GO:0006914, GO:0008757, GO:0009057, GO:0009896
ORF7b	-	GO:0001172, GO:0001960, GO:0002682, GO:0009896
ORF8	GO:0005515	GO:0005524, GO:0009896, GO:0016788, GO:0043170, GO:0044238
ORF9b	GO:0005515	GO:0004519, GO:0004532, GO:0005524, GO:0009896
ORF9c	GO:0005515	GO:0001959, GO:0016070, GO:0044249, GO:0048518, GO:1901576
ORF10	-	GO:0001959, GO:0003824, GO:0016070
S1	GO:0046790, GO:0061025	GO:0001172, GO:0001960, GO:0002151, GO:0005737, GO:0006370, GO:0006402, GO:0006508, GO:0006811, GO:0006897, GO:0006914
S2	GO:0006897	GO:0001172, GO:0001960, GO:0002151, GO:0005737, GO:0006402, GO:0006508, GO:0006811, GO:0006897, GO:0009894, GO:0048518
S2'	GO:0075509, GO:0006508	GO:0001172, GO:0001960, GO:0002151, GO:0005737, GO:0006402, GO:0006508, GO:0006811, GO:0006897, GO:0009894
nsp1	GO:0039604, GO:0042783, GO:0019080	GO:0001959, GO:0003824, GO:0009059, GO:0010467,GO:0016070, GO:0018130, GO:0048518
nsp2	GO:0033554	GO:0001960, GO:0002832, GO:0003724, GO:0003727, GO:0003968, GO:0004482, GO:0004483, GO:0004532, GO:0005198, GO:0005216, GO:0005515, GO:0008234, GO:0032553, GO:0035639, GO:0101005
nsp3	GO:0016579, GO:0140526, GO:0019783, GO:0001960, GO:0008233, GO:0039503	GO:0001172, GO:0001960, GO:0002832, GO:0003727, GO:0004483, GO:0004519, GO:0004540, GO:0005216, GO:0005515, GO:0008234, GO:0019783, GO:0030554, GO:0035639, GO:0036265, GO:0097367
nsp4	GO:0140526	GO:0001172, GO:0004519, GO:0004527, GO:0006370, GO:0016772, GO:0042221, GO:0044260, GO:0090503
nsp5	GO:0004180, GO:0072570	GO:0001960, GO:0003968, GO:0004527, GO:0006370, GO:0036094, GO:0043168, GO:0048518, GO:0090503, GO:1901265
nsp6	GO:2000786, GO:1901096	GO:0001172, GO:0001959, GO:0004527
nsp7	GO:0019079, GO:0003968	GO:0001959, GO:0003824, GO:0009058, GO:0010467, GO:0016070
nsp8	GO:0019079, GO:0003968	GO:0001960, GO:0002832, GO:0003724, GO:0003727, GO:0003968, GO:0004197, GO:0004482, GO:0004483, GO:0004519, GO:0004532, GO:0004843, GO:0005198, GO:0005216, GO:0005515, GO:0005524
nsp9	GO:0003727, GO:0019079	GO:0009059, GO:0010467, GO:0016070, GO:0016787, GO:0018130, GO:0019221, GO:0060759
nsp10	GO:0019083, GO:0036451	GO:0001172, GO:0001959, GO:0003824, GO:0048518
nsp11	-	GO:0001959, GO:0003824, GO:0016070
nsp12	GO:0019083, GO:0019079, GO:0003968	GO:0003727, GO:0003968, GO:0006952, GO:0006955, GO:0008168, GO:0008270
nsp13	GO:0004386	GO:0001172, GO:0003723, GO:0005216, GO:0006370, GO:0006955, GO:0007049, GO:0007166, GO:0044260, GO:0048583
nsp14	GO:0004532, GO:0004482	GO:0006950, GO:0007049, GO:0007154, GO:0008234, GO:0048518, GO:0051716
nsp15	GO:0008663	GO:0009966, GO:0016070, GO:0016787, GO:0019221, GO:0060759
nsp16	GO:0008168, GO:0042783	GO:0001959, GO:0003824, GO:0016070
E	GO:0019069, GO:0039707, GO:0060139	GO:0000175, GO:0001172, GO:0001510, GO:0001960, GO:0002151, GO:0002683, GO:0009893, GO:0030554
M	GO:0019069	GO:0001172, GO:0001960, GO:0002151, GO:0002682, GO:0004519, GO:0004527, GO:0005737, GO:0009893
N	GO:0019068, GO:0003723, GO:0019083, GO:0019079	GO:0002376, GO:0003727, GO:0005793, GO:0005886

(2021). La funcionalidad GO:0003727 (unión de ARN monocatenario) predicha podría considerarse como correcta ya que su definición incluye el sustrato ARN propia de este proceso.

3. *nsp1*:

   Inhibe la traducción del huésped al interactuar con la subunidad ribosomal 40S, facilita la expresión génica viral eficiente en las células infectadas y la evasión

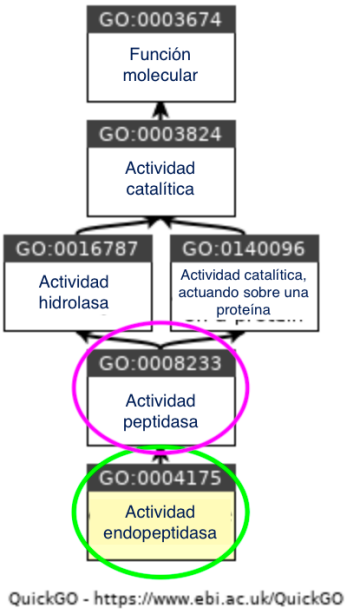

Figura 8: Grafo predicho para la proteína ORF1a. La funcionalidad más específica (GO:0004175) se encuentra identificada con un círculo verde. La funcionalidad GO experimental (GO:0008233) se encuentra identificada con círculo rosa.

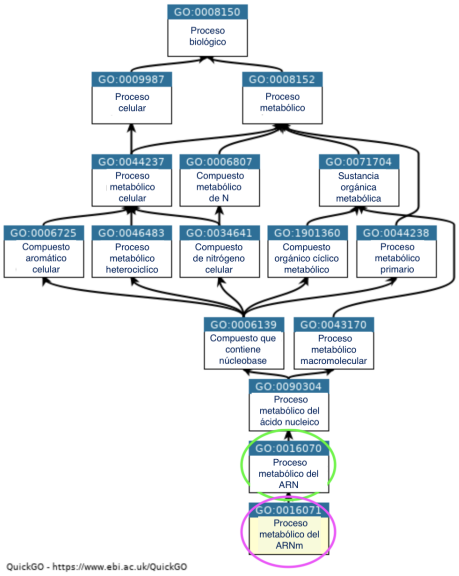

Figura 9: Grafo experimental de nsp1. La funcionalidad predicha GO:0016070 (proceso metabólico del ARN) está identificada con círculo verde. La funcionalidad GO experimental GO:0016071 (proceso metabólico del ARNm) está identificada con círculo rosa.

de la respuesta inmunitaria del huésped. De las predicciones obtenidas, nos enfocaremos en la función GO:0016070 (proceso metabólico del ARN) predicha que está incluida en la función GO experimental GO:0016071 (proceso metabólico del ARNm), i.e., es una predicción más general (ver Figura 9). Por otro lado, el GO:0039604 (supresión por virus de la traducción del huésped) experimental contiene los GO predichos GO:0048518 (positiva del proceso biológico),

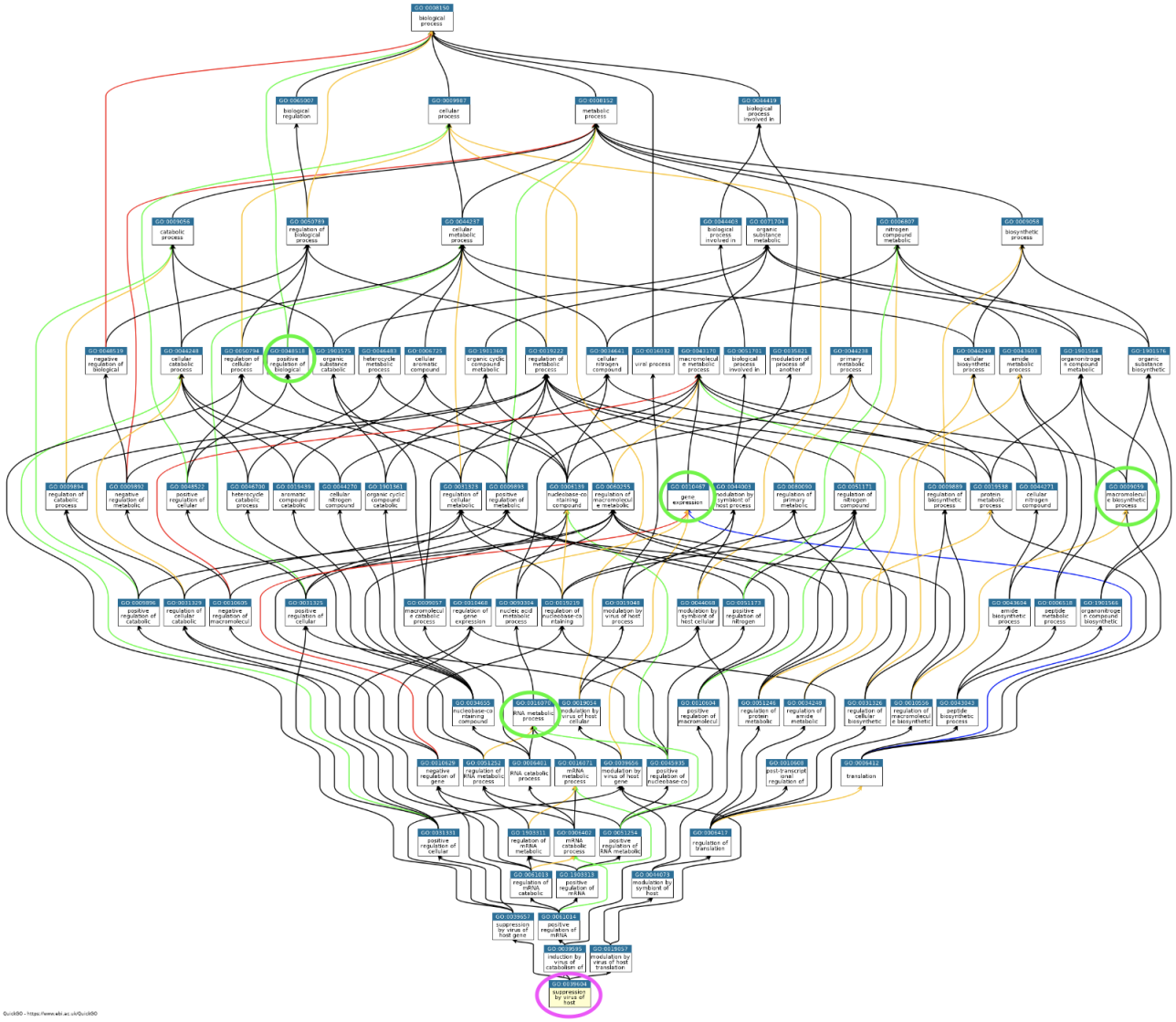

Figura 10: Grafo experimental de nsp1. Las funcionalidades predichas (GO:0048518, GO:0016070, GO:0010467, GO:0009059) están identificadas con círculos verdes. La funcionalidad GO experimental (GO:0039604) está identificada con círculo rosa.

GO:0016070 (proceso metabólico del ARN), GO:0010467 (expresión génica) y GO:0009059 (proceso biosintético de macromoléculas) como se observa en la Figura 10. En la Figura 11 puede verse la funcionalidad en común, GO:00500896, entre el grafo experimental GO:0042783 y el grafo predicho GO:0001959. El GO:0010467 (expresión génica) predicho es una anotación válida con evidencia experimental.

4. *nsp2*:

   Puede desempeñar un papel en la modulación de la vía de señalización de supervivencia de la célula huésped al interactuar con las proteínas PHB y PHB2 del huésped. Estas dos proteínas desempeñan un papel en el mantenimiento de la integridad funcional de las mitocondrias y en la protección de las células frente a diversos tipos de estrés. En este caso, podemos mencionar que la coincidencia entre el GO experimental, respuesta celular al estrés (GO:0033554) y el GO

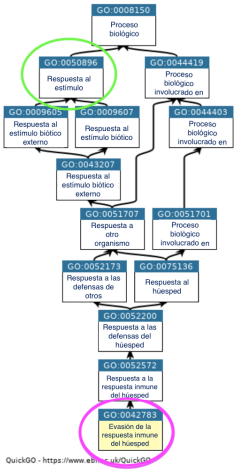

Figura 11: Grafo experimental de nsp1. La funcionalidad predicha regulación de la vía de señalización mediada por citoquinas (GO:0001959) comparte respuesta a estímulos (GO:00500896) identificada en la gráfica con círculo verde, con la funcionalidad experimental evasión de la respuesta inmune del huésped (GO:0042783) que está identificada con círculo rosa.

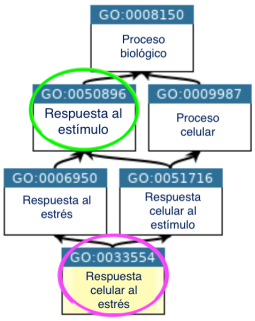

Figura 12: Grafo experimental para nsp2. La función GO:0050896 identificada en verde es común a la funcionalidad GO:0002832 predicha y al GO:0033554 experimental identificado con un círculo rosa.

predicho, regulación negativa de la respuesta al estímulo biótico (GO:0002832) es escasa, (ver Figura 12) solamente comparten el GO:0050896 (respuesta al estimulo).

5. *nsp3*:

   Es responsable de los clivajes localizados en el extremo N-terminal de la poliproteína replicasa y participa junto con el gen codificante nsp4 en el ensamblaje de vesículas citoplasmáticas de doble membrana inducidas por virus necesarias para la replicación viral. Las funcionalidades GO:0019783 (actividad de proteína peptidasa similar a la ubiquitina) y GO:0001960 (regulación negativa de la vía de señalización mediada por citoquinas) coinciden con los GO experimentales. Las funciones GO:0048585 (regulación negativa de la respuesta al estímulo) y GO:0008233 (actividad peptidasa) predichas están incluidas en los GO:0002832 (regulación negativa de la respuesta al estímulo biótico) y GO:0008234 (actividad de peptidasa tipo cisteína) experimentales respectiva-

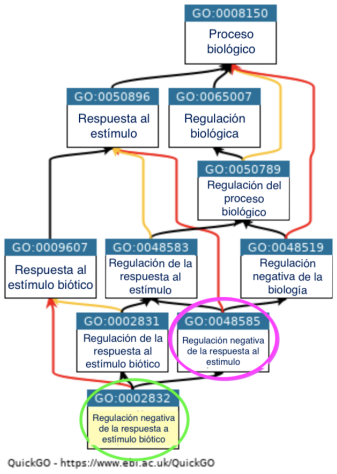

Figura 13: Grafo predicha de nsp3. La funcionalidad predicha GO:0048585 (regulación negativa de la respuesta al estímulo), identificada en verde está incluida en la funcionalidad experimental identificado con un círculo rosa GO:0002832 (regulación negativa de la respuesta al estímulo biótico).

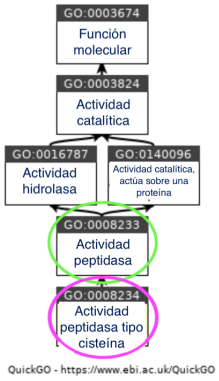

Figura 14: Grafo experimental para nsp3. La funcionalidad predicha GO:0008233 (actividad peptidasa) está identificada con círculo verde y la funcionalidad experimental GO:0008234 (actividad de peptidasa tipo cisteína) está identificada con círculo rosa.

mente, i.e., son predicciones más generales (ver Figuras 13 y 14).

6. *nsp4*:

   Participa en el ensamblaje de vesículas de doble membrana citoplasmáticas inducidas por virus necesarias para la replicación viral. En este caso, los GOs predichos no se encuentran incluidos en el GO:0140526 experimental, ya que, los términos relacionados con la organización de componentes o biogénesis no están presentes en la base de conocimiento del modelo entrenado.

7. *nsp5*:

   Reconoce sustratos que contienen una secuencia específica (Zhang et al., 2020) y también es capaz de unirse a un ADP-ribosa-1"-fosfato (ADRP). La función GO:0043168 (unión de aniones) predicha está incluida en el GO:0072570 (unión de ADP-D-ribosa) experimental, i.e., es una predicción más general (ver Figura 15). Por otro lado, se sabe que esta proteína antagoniza la producción de IFN-

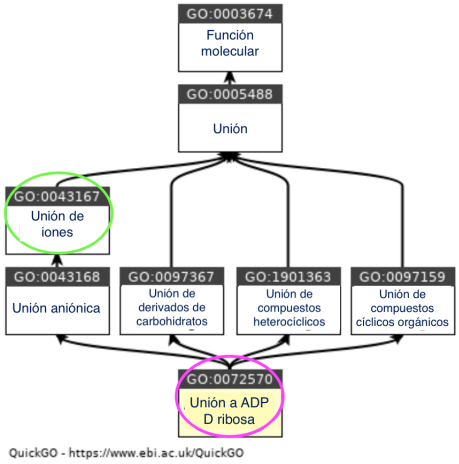

Figura 15: Grafo experimental para nsp5. La funcionalidad predicha unión de aniones (GO:0043168) está identificada con círculo verde y la funcionalidad GO experimental unión de ADP-D-ribosa (GO:0072570) está identificada con círculo rosa.

β, ya que, cliva a NEMO en múltiples sitios (E152, Q205 y Q231) debido a que es un modulador esencial de NF-kβ, factor de transcripción primario de acción rápida, i.e., primer respondedor ante un estímulo de daño celular que genera cambios muy rápidos en la expresión de distintos genes (Chen et al., 2022). Esta definición estaría contenida en el GO:0001960 (regulación negativa de la vía de señalización mediada por citoquinas) predicho. Además, experimentalmente la sobreexpresión de la proteína nsp5 promueve la replicación del virus (YZheng et al., 2022), suponiendo que puede estar incluida la predicción GO:0048518 (regulación positiva del proceso biológico). Finalmente, nsp5 también interactúa con el factor epigenético del huésped histona desacetilasa 2 (Naik et al., 2022) y esto hace suponer que la predicción GO:0036094 (unión de moléculas pequeñas) es correcta.

8. *nsp6*:

   Desempeña un papel en la inducción inicial de autofagosomas desde el retículo endoplásmico del huésped. En este caso, los GOs predichos no se encuentran incluidos en los GO:1901096 y GO:2000786 experimentales. Sin embargo, nsp6 puede reducir la fosforilación de IRF3, lo que lleva a un efecto inhibitorio sobre la actividad del promotor de IFN-β (Low et al., 2022); esto sugiere que la funcionalidad GO:0001959 (regulación de la vía de señalización mediada por citoquinas) predicha es correcta.

9. *nsp7*:

   Forma un hexadecámero con nsp8 y puede participar en la replicación viral actuando como primasa. La función GO:0016070 (proceso metabólico del ARN) predicha está incluida en el GO:0003968 (actividad de polimerasa de ARN dependiente de ARN) experimental, i.e., es una predicción más general (ver Figura 16). Existen estudios que demuestran nsp7 junto con nsp8 funciona como un cofactor esencial que se une a nsp12 y forma un complejo que estabiliza el dominio de la polimerasa facilitando el reconocimiento de la plantilla y la polimerización de nucleótidos (Low et al., 2022). Además, se observa un antagonismo de la señalización IFN-α por SARS-CoV nsp7, esto coincide con la

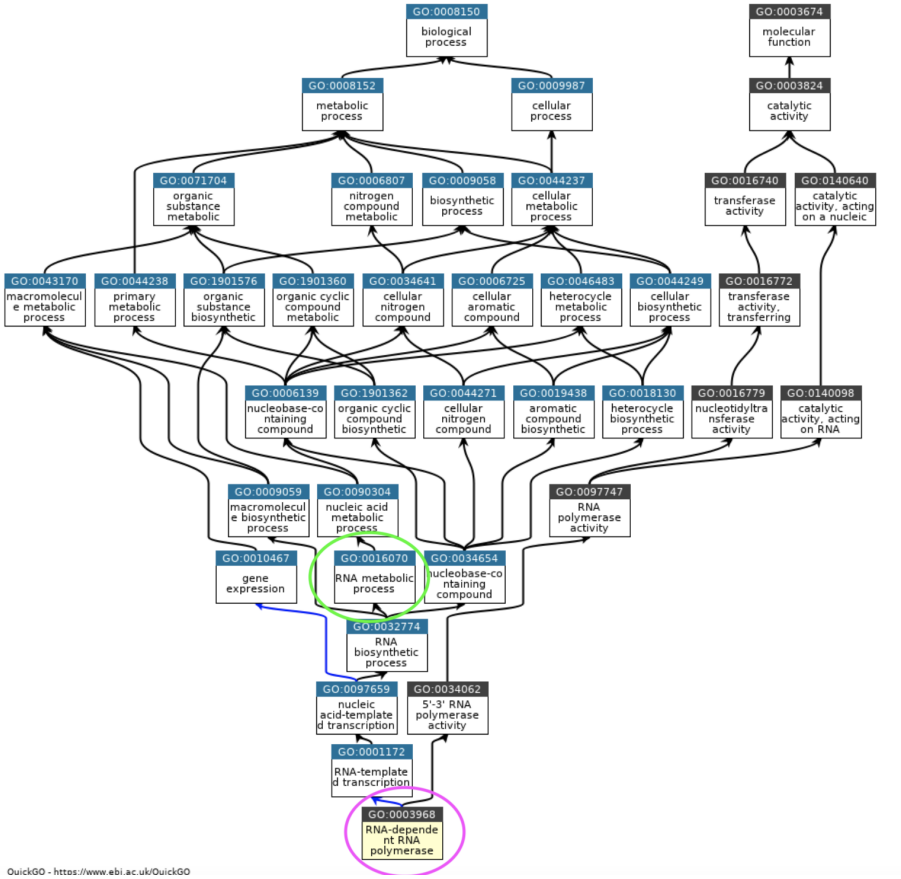

Figura 16: Grafo experimental para nsp7. La funcionalidad predicha proceso metabólico del ARN (GO:0016070) está identificada con círculo verde y la funcionalidad GO experimental actividad de polimerasa de ARN dependiente de ARN (GO:0003968) está identificada con círculo rosa.

funcionalidad GO:0001959 (regulación de la vía de señalización mediada por citoquinas) predicha.

10. *nsp8*:

    Forma un hexadecámero con nsp7 y puede participar en la replicación viral actuando como primasa. En este caso, los GOs predichos no se encuentran incluidos en el GO:0019079 experimental, ya que, los términos relacionados con el ciclo de vida viral no están presentes en la base de conocimiento del modelo entrenado por falta de anotaciones en SARS-COV-2. Sin embargo, la funcionalidad GO:0003968 (actividad de polimerasa de ARN dependiente de ARN) predicha coincide con el GO experimental. Por otro lado, nsp8 puede suprimir las respuestas antivirales dependientes de MAVS (Yang et al., 2020). Esto sugiere que la funcionalidad GO:0001959 (regulación de la vía de señalización mediada por citoquinas) predicha sería correcta. En te Velthuis et al. (2012) se demuestra que el hexadecámero nsp(7+8) es la conformación más probable de la segunda polimerasa SARS-CoV, dada la asociación casi completa de nsp7 y nsp8 cuando se mezcla 1:1 en solución; esto sugiere que la funcionalidad GO:0005515 (unión a una proteína) predicha es correcta. Además, se encontró que este complejo es capaz de unir moléculas de dsRNA y es capaz de extender parcialmente moldes de ARN doble cadena, por lo tanto, sugiere que la fun-

cionalidad GO:0003727 (unión de ARN monocatenario) predicha es correcta. Finalmente, en el desarrollo de la actividad de la polimerasa es necesario agregar el nucleótido ATP, lo que propone que la predicción GO:0005524 (unión a ATP) es correcta.

11. *nsp9*: Puede participar en la replicación viral al actuar como una proteína de

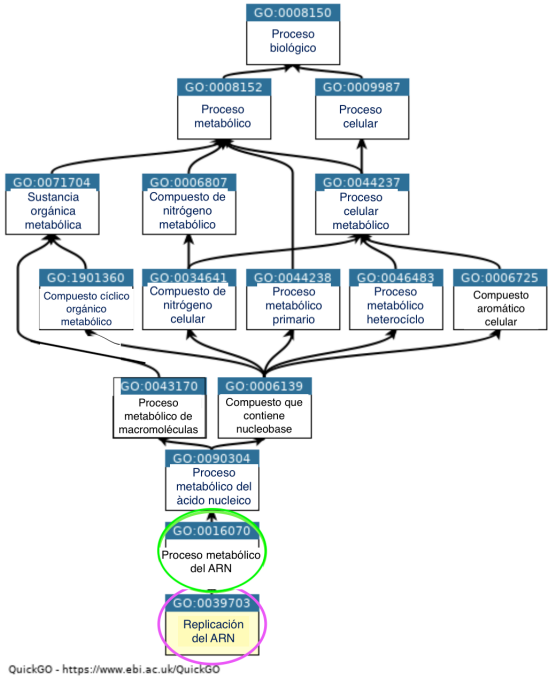

Figura 17: Grafo experimental para nsp9. La funcionalidad predicha proceso metabólico del ARN (GO:0016070), está identificada con círculo verde y está incluida en la funcionalidad experimental replicación de ARN (GO:0039703), identificada con círculo rosa.

unión a ssRNA. En este caso, los GOs predichos no se encuentran incluidos en el GO, replicación del genoma viral (GO:0019079), ya que, los términos relacionados con el ciclo de vida viral no están presentes en la base de conocimiento del modelo entrenado por falta de anotaciones en SARS-CoV-2. Sin embargo, replicación de ARN (GO:0039703) incluye la predicción proceso metabólico del ARN (GO:0016070). Los términos GOs predichos no pueden ser descartados, ya que, actualmente no hay evidencia que contradiga esas funcionalidades.

12. *nsp10*:

    Desempeña un papel fundamental en la transcripción viral al estimular las actividades de exoribonucleasa 3'-5' en nsp14 y 2'-O-metiltransferasa en nsp16; i.e., juega un papel esencial en la metilación de la caperuza de los ARNm virales. La funcionalidad GO:0001172 (transcripción con molde de ARN) coincide con el GO experimental. Los términos GOs restantes predichos no pueden ser descartados debido a que actualmente no hay evidencia que contradiga esas funcionalidades.

13. *nsp11*:

    No posee términos GO experimentales. Desde el punto de vista biológico, las funcionalidades GO:0001959 (regulación de la vía de señalización mediada por

citoquinas), GO:0003824 (actividad catalítica) y GO:0016070 (proceso metabólico del ARN) predichas pueden ser viables, sin embargo, no se encontró evidencia experimental reciente que pueda vincularlas.

14. *nsp12*:

    Responsable de la replicación y transcripción del genoma del ARN viral. La funcionalidad GO:0003968 (actividad de polimerasa de ARN dependiente de ARN) predicha coincide con el GO experimental. El resto de los GOs predichos no se encuentran incluidos en los GOs experimentales GO:0019079 y GO:0019083, ya que, los términos relacionados con el proceso viral no están presentes en la base de conocimiento del modelo entrenado por falta de anotaciones en SARS-COV-2. La funcionalidad GO:0003727 (unión de ARN monocatenario) predicha podría considerarse como correcta en vista de que su definición incluye el sustrato ARN en el proceso de transcripción del genoma. Por otro lado, se sabe que nsp12 suprime las respuestas antivirales del huésped y atenúa la activación del promotor del IFN tipo I inducida por el virus al inhibir la translocación nuclear de IRF3 (Wang et al., 2020). Esto sugiere que las funcionalidades GO:0006952 (respuesta de defensa) y GO:0006955 (respuesta inmune) pueden ser correctas. El resto de los términos GOs predichos no pueden ser descartados, ya que, actualmente no hay evidencia que contradiga esas funcionalidades.

15. *nsp13*:

    Presenta actividad de desenrollado de dúplex de ARN y ADN con polaridad de 5' a 3'. En este caso, los GOs predichos no se encuentran incluidos en el GO experimental GO:0004386. Sin embargo, en Jang et al. (2020) se demuestra que la acumulación de complejos dúplex de ARN/nsp13 aumentó a medida que aumentaba la concentración de nsp13, esto sugiere que la funcionalidad GO predicha GO:0003723 (unión de ARN) es correcta. El resto de los términos GOs predichos no pueden ser descartados, ya que, actualmente no hay evidencia que contradiga esas funcionalidades.

16. *nsp14*:

    Actúa como una exoribonucleasa correctora para la replicación del ARN, lo que reduce la sensibilidad del virus a los mutágenos del ARN y la N7-guanina metiltransferasa. En este caso, los GOs predichos no se encuentran incluidos en los GO:0004532 y GO:0004482 experimentales debido a que los términos relacionados con la actividad catalítica que actúa para modificar el ARN no están presentes en la base de conocimiento del modelo entrenado por falta de anotaciones en SARS-CoV-2. Los términos GOs predichos no pueden ser descartados, ya que, actualmente no hay evidencia que contradiga esas funcionalidades.

17. *nsp15*: Enzima específica de uridilato dependiente de Mn(2+), que deja fosfatos cíclicos 2'-3' 5' al enlace escindido. La función GO:0016787 (actividad hidrolasa) predicha está incluida en el GO:0008663 (actividad de 2',3'-nucleótido cíclico 2'-fosfodiesterasa) experimental, i.e., es una predicción más general (ver Figura 18). Por otro lado, se sabe que la endoribonucleasa específica de uridina nsp15 puede eliminar la región de 5′-poliuridina del ARN viral de cadena negativa (ARN de la PUN) para evitar ser detectado por sensores de dsRNA citosólico, incluidos MDA5, PKR y OAS/RNasaL. Por lo tanto, desempeña un papel fundamental en el bloqueo de los primeros eventos de la detección de antivirales del huésped Kasuga et al. (2021). Esto sugiere que las funcionalidades GO:0009966

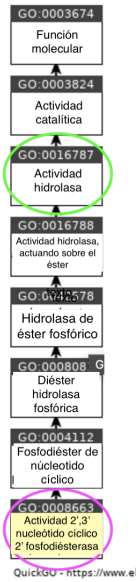

Figura 18: Grafo experimental para nsp15. La funcionalidad actividad hidrolasa (GO:0016787) predicha está identificada con círculo verde y la funcionalidad GO experimental actividad de 2',3'-nucleótido cíclico 2'-fosfodiesterasa (GO:0008663) está identificada con círculo rosa.

(regulación de la transducción de señales), GO:0016070 (proceso metabólico del ARN), GO:0019221 (vía de señalización mediada por citoquinas) y GO:0060759 (regulación de la respuesta al estímulo de citoquinas) pueden ser anotaciones válidas.

18. *nsp16*: Posee actividad de metiltransferasa y evasión de la respuesta inmune

Figura 19: Grafo experimental para nsp16. La funcionalidad actividad catalítica (GO:0003824) predicha está identificada con círculo verde y la funcionalidad GO experimental actividad de la metiltransferasa (GO:0008168) está identificada con círculo rosa.

del huésped. La función GO:0003824 (actividad catalítica) predicha está incluida en el GO:0008168 (actividad de la metiltransferasa) experimental, i.e., es una predicción más general (ver Figura 19). Por otro lado, se sabe que nsp16 al regular la 2′-O-metilación del ARN viral, esencial para el límite de 5′, evita la activación de sensores antivirales como MDA5 y las proteínas de la familia IFIT27 (Kasuga et al., 2021). Esto sugiere que las funcionalidades GO:0016070 (proceso metabólico del ARN) y GO:0001959 (regulación de la vía de señaliza-

ción mediada por citoquinas) pueden ser anotaciones válidas.

4.3.2. Proteínas estructurales

19. *E*:

    Juega un papel central en la morfogénesis y el ensamblaje del virus; actúa como una viroporina y se autoensambla en las membranas del huésped formando poros pentaméricos de proteínas y lípidos que permiten el transporte de iones. También juega un papel en la inducción de la apoptosis. En la Figura 20 puede verse la funcionalidad en común regulación del proceso celular (GO:0050794), entre el grafo experimental inducción por virus del proceso apoptótico del huésped (GO:0019051) y el grafo predicho regulación negativa de la vía de señalización mediada por citoquinas (GO:0001960). Aquí, se puede mencionar que el modelo no es capaz de alcanzar dicho GO debido a la falta de funcionalidades relacionadas con la muerte celular producto de una cantidad pequeñas de anotaciones en SARS-CoV-2. Un estudio reciente comparó la respuesta al estrés de las células infectadas con SARS-CoV con las células infectadas con SARS-CoVδE utilizando un enfoque basado en microarrays y se demostró que el virus que carecía de E induce una respuesta al estrés mucho más robusta que el virus de tipo salvaje. Por tanto, el virus que carece de E también causó un mayor grado de apoptosis en comparación con el virus de tipo salvaje (DeDiego et al., 2011). Esto sugiere que la funcionalidad GO:0002683 (regulación negativa del proceso del sistema inmunológico) puede ser correcta.

20. *M*:

    Juega un papel central en la morfogénesis y ensamblaje del virus a través de sus interacciones con otras proteínas virales. En este caso, los GOs predichos no se encuentran incluidos en el GO experimental GO:0019079, puesto que, los términos relacionados con el ciclo de vida viral no están presentes en la base de conocimiento del modelo entrenado por falta de anotaciones en SARS-CoV-2. Por otro lado, se demostró que la proteína de la membrana SARS-CoV (M) mitiga la formación del complejo activador NF-kB (TANK)-TBK1/IKKϵ asociado a los miembros de la familia TRAF3-TRAF, impidiendo así la activación de IRF3/IRF7 aguas abajo y la producción de IFN. Mecánicamente, la región TM1 (1-38 aminoácidos) de la proteína M es crítica para su localización en el aparato de Golgi, donde M interactúa con proteínas inmunes innatas como RIG-I, TBK1, IKKϵ y TRAF3, bloqueando las cascadas de señalización antiviral aguas abajo. Además, se mostró que el SARS-CoV-2 M suprime la producción de IFN de tipo I y III y que podría unirse directamente a moléculas esenciales de la vía de detección de ARN viral citosólico, como RIG-I, MDA5, MAVS y TBK1, y prevenir su interacción (Kasuga et al., 2021). Esto sugiere que las funcionalidades GO:0002682 (regulación del proceso del sistema inmunológico) y GO:0001960 (regulación negativa de la vía de señalización mediada por citoquinas) serían correctas.

21. *N*:

    Juega un papel fundamental durante el ensamblaje del virión a través de sus interacciones con el genoma viral y la proteína M de membrana. Además, un rol importante en la mejora de la eficiencia de la transcripción del ARN viral subgenómico como así como la replicación viral. La función GO:0003727

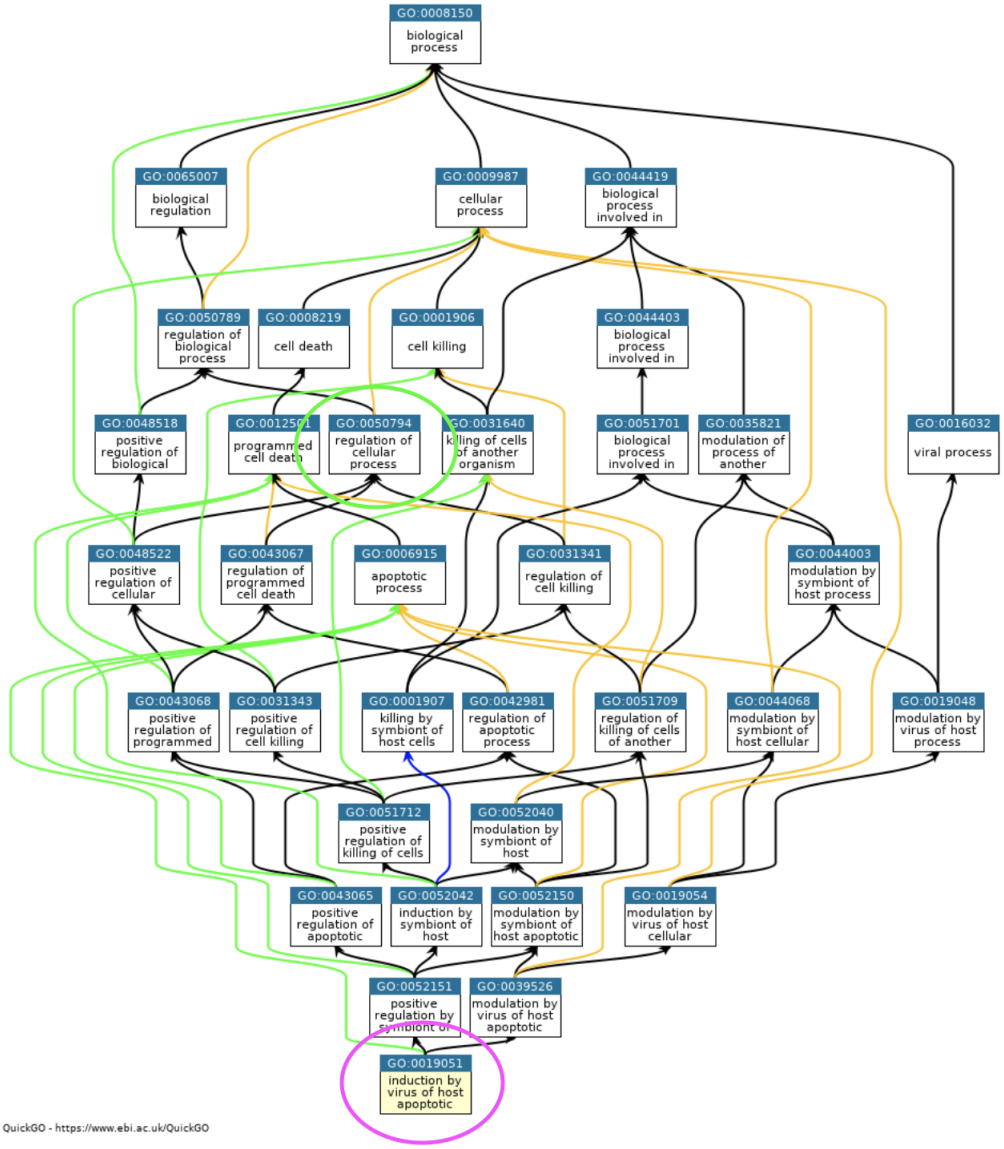

Figura 20: Grafo experimental para E. La funcionalidad más específica (GO:0019051) se encuentra identificada con un círculo rosa. La funcionalidad predicha común (GO:0050794) se encuentra identificada con un círculo verde. Esta funcionalidad proviene del grafo predicho GO:0001960.

(unión de ARN monocatenario) predicha incluyen a la función GO experimental GO:0003723 (unión de ARN), i.e., es predicción más específica (ver Figura 21). Por otro lado, la bibliografía define a la proteína N como un excelente epítopo inmunogénico para el sistema inmunitario adaptativo y, en particular, es objetivo de las células T CD8+ citotóxicas (Szeto et al., 2021). Adicionalmente, N desempeña un papel fundamental en la supresión de la inmunidad innata del huésped. Estudios con SARS-CoV mostraron que el N suprime la producción de IFN (Kasuga et al., 2021). Ambas puede justificar la predicción que la involucra con el proceso del sistema inmune (GO:0002376).

22. *S1*: Une el virión a la membrana celular al interactuar con el receptor del huésped, iniciando la infección (por similitud). La unión al receptor ACE2 humano y la internalización del virus en los endosomas de la célula huésped induce cambios conformacionales en la glicoproteína Spike. También utiliza TMPRSS2 humano

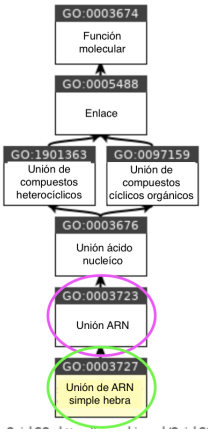

Figura 21: Grafo predicho para N. La funcionalidad más específica (GO:0003727) se encuentra identificada con un círculo verde. La funcionalidad experimental (GO:0003723) se encuentra identificada con círculo rosa.

para el cebado de células pulmonares humanas, que es un paso esencial para la entrada viral. Las funcionalidades GO:0006897 (endocitosis) y GO:0006508 (proteólisis) predichas coinciden con los GO experimentales.

23. *S2*: Media la fusión del virión y las membranas celulares al actuar como una proteína de fusión viral de clase I. Según el modelo actual, la proteína tiene al menos tres estados conformacionales: estado nativo previo a la fusión, estado intermedio anterior a la horquilla y estado posterior a la fusión. Durante la fusión de la membrana viral y de la célula diana, las regiones de la bobina enrollada (repeticiones de heptada) adoptan una estructura de trímero de horquillas, colocando el péptido de fusión muy cerca de la región C-terminal del ectodominio. La formación de esta estructura parece impulsar la aposición y la subsiguiente fusión de las membranas de las células virales y diana. La funcionalidad GO:0006897 (endocitosis) predicha coincide con el GO experimental.

24. *S2'*: Actúa como un péptido de fusión viral que se desenmascara después de la escisión S2 que ocurre con la endocitosis del virus. La funcionalidad GO:0006508 (proteólisis) predicha coincide con el GO experimental. La función GO:0006897 (endocitosis) predicha está incluida en el GO experimental GO:0075509 (endocitosis implicada en la entrada viral en la célula huésped), i.e., es una predicción más general (ver Figura 22).

4.3.3. Proteínas accesorias

25. *ORF3a*: Forma canales iónicos homotetraméricos en las membranas de las células huésped sensibles al potasio (viroporina), modula la liberación del virus, induce la apoptosis en cultivo celular, regula el aumento de la expresión de las subunidades de fibrinógeno FGA, FGB y FGG en las células epiteliales del pulmón huésped, y regula negativamente el receptor de interferón tipo 1 al inducir la fosforilación de serina dentro del motivo de degradación de la subunidad 1 del receptor de IFNα (IFNAR1) y aumenta la ubiquitinación de IFNAR1. En la Figura 23 puede verse la funcionalidad en común, proceso metabólico macromolécula (GO:0043170), entre el grafo experimental modulación por el virus de la expresión génica del huésped (GO:0039656) y el grafo predicho transcripción

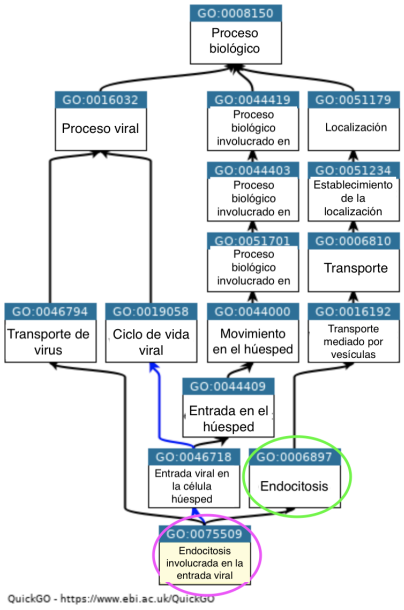

Figura 22: Grafo experimental para S2'. La funcionalidad experimental (GO:0075509) más específica se encuentra identificada con un círculo rosa. La funcionalidad GO predicha (GO:0006897) se encuentra identificada con un círculo verde.

con plantilla de ARN (GO:0001172). En Bianchi et al. (2021) se evidencia que la supresión genómica de ORF3a redujo la replicación del virus en organismos modelos de animales. Esto sugiere que la funcionalidad GO:0001172 (transcripción con plantilla de ARN) predicha sería correcta. El resto de los GOs predichos no se encuentran incluidos en los GOs experimentales GO:0019076, GO:0005216, GO:0039707 y GO:0097194 debido a que los términos relacionados con el proceso biológico implicado en la interacción con el huésped y el ciclo de vida viral no están presentes en la base de conocimiento del modelo entrenado por falta de anotaciones en SARS-CoV-2.

26. *ORF6*: Es determinante para la virulencia del virus, ya que, cuando se expresa en una cepa JHM atenuada de coronavirus murino y puede aumentar drásticamente su letalidad. La función GO:0006955 (respuesta inmune) predicha está incluida en el GO experimental GO:0075528 (modulación por virus de la respuesta inmune del huésped), i.e., es una predicción más general (ver Figura 24). Por otro lado, se ha demostrado experimentalmente que ORF6 puede bloquear la expresión de genes estimulados por interferón como por ejemplo ISG15 que tienen actividad antiviral (O'Donoghue et al., 2021). Esto sugiere que la funcionalidad GO:0001960 (Regulación negativa de la vía de señalización mediada por citoquinas) predicha sería correcta.

27. *ORF7a*: Es prescindible para la replicación del virus en cultivo celular. La funcionalidad GO:0001172 (transcripción con molde ARN) predicha coincide con el GO experimental. Por otro lado, se demostró que la unión de ORF7a a los monocitos CD14+ condujo a una disminución significativa en las moléculas HLA-DR/DP/DQ en los monocitos CD14+ sugieriendo que ORF7a puede suprimir la capacidad de presentación de antígenos de estos monocitos. Además, La coincubación del SARS-CoV-2 ORF7a con monocitos desencadena la regulación ascendente significativa de múltiples citocinas proinflamatorias, lo que

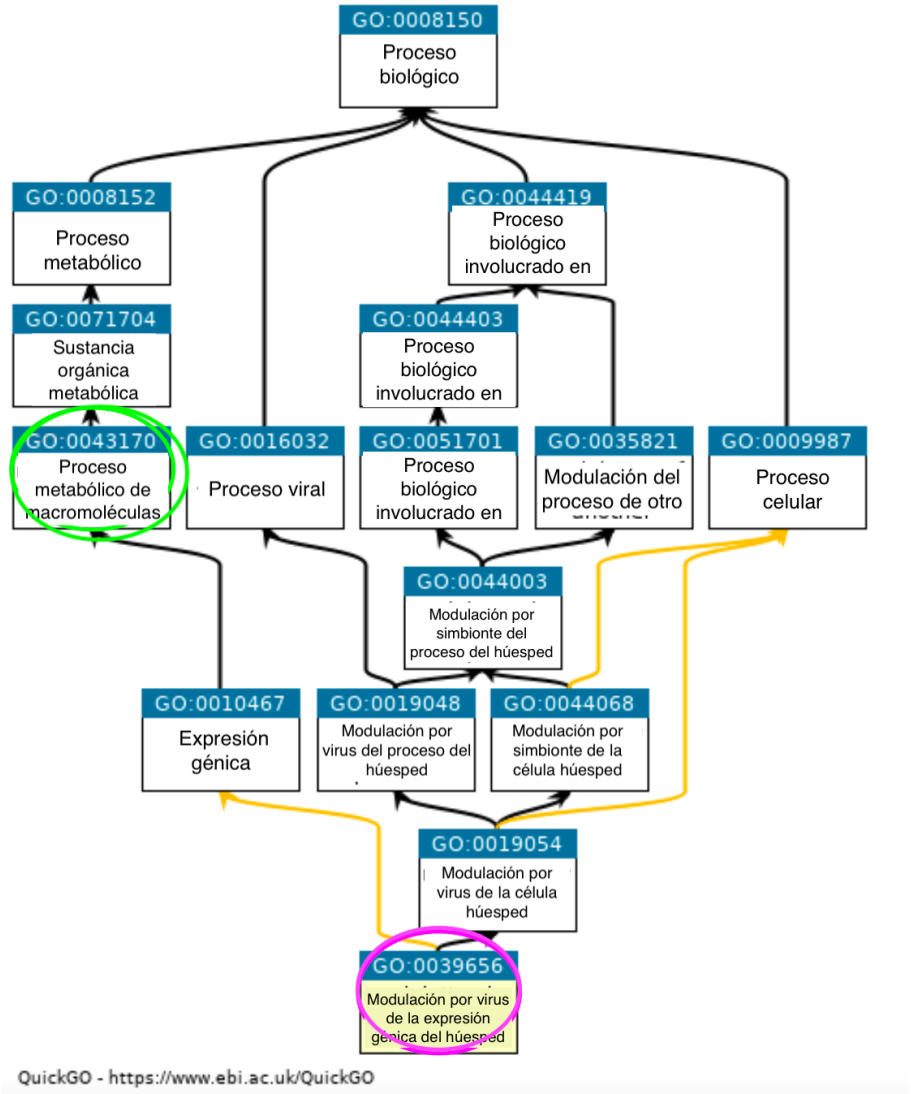

Figura 23: Grafo experimental para ORF3a. La funcionalidad más específica modulación por el virus de la expresión génica del huésped (GO:0039656) se encuentra identificada con un círculo rosa. La funcionalidad GO predicha común (GO:0043170) se encuentra identificada con un círculo verde. Esta funcionalidad proviene del grafo predicho (GO:0001172).

indica que ORF7a podría ser un factor clave en la progresión de la tormenta de citoquinas en COVID-19 (Qin et al., 2020). En conjunto, la capacidad inmunomoduladora potencial del SARS-CoV-2 ORF7a en los monocitos humanos sirve como una estrategia de escape inmune viral (Zhou et al., 2021). Esto sugiere que las funcionalidades GO:0001960 (regulación negativa de la vía de señalización mediada por citoquinas) y GO:0002682 (regulación del proceso del sistema inmunológico) predichas serían correctas.

28. *ORF7b*: Carece de GO experimentales. Desde el punto de vista biológico, las funcionalidades GO:0001172 (transcripción con molde de ARN), GO:0001960 (regulación negativa de la vía de señalización mediada por citoquinas), GO:0002682 (regulación del proceso del sistema inmunológico) y GO:0009896 (regulación positiva del proceso catabólico) predichas pueden ser viables, sin embargo, no se encontró evidencia experimental reciente que pueda vincularlas.

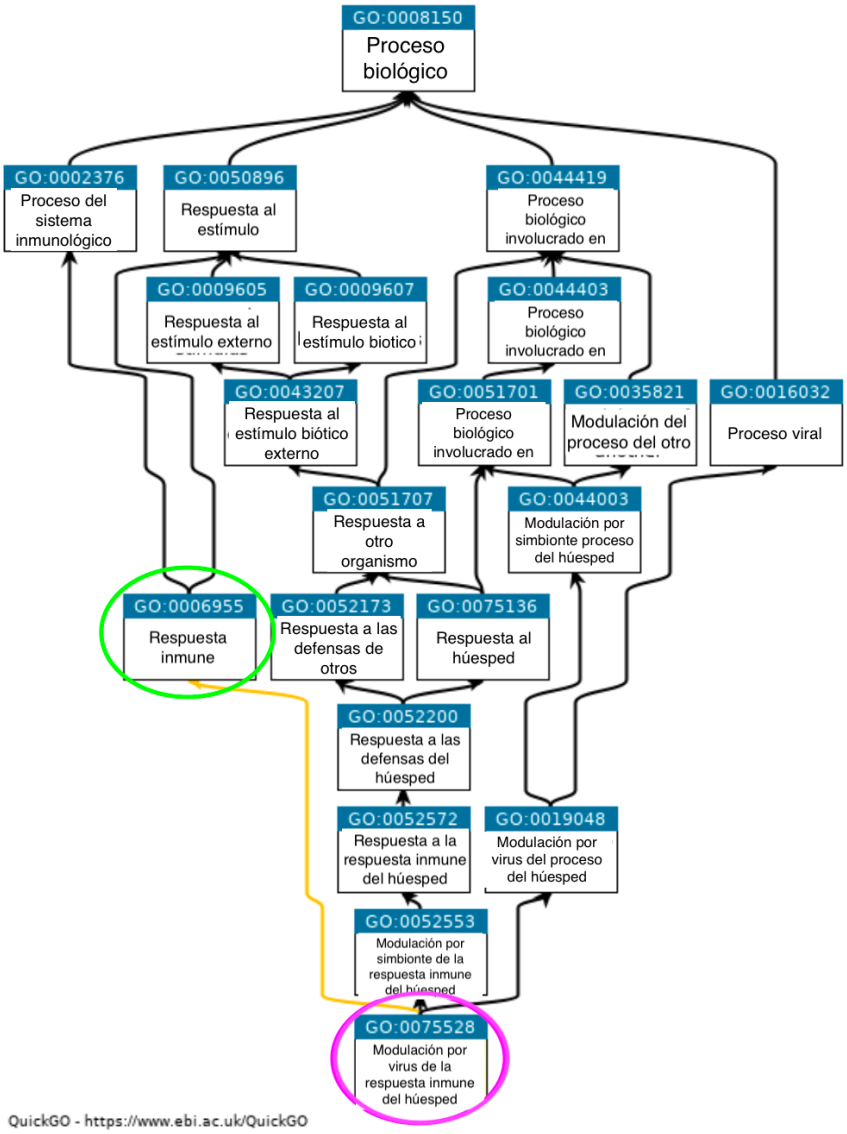

Figura 24: Grafo experimental para ORF6. La funcionalidad más específica modulación por virus de la respuesta inmune del huésped (GO:0075528) se encuentra identificada con un círculo rosa. La funcionalidad GO predicha respuesta inmune (GO:0006955) se encuentra identificada con un círculo verde.

29. *ORF8*: Puede desempeñar un papel en la interacción huésped-virus. En este caso, la funcionalidad GO:0005524 (unión a ATP) predicha comparte solamente el término GO:0005488 (unión) del GO experimental GO:0005515 (unión a proteínas) debido a que esta función es muy general (ver Figura 25). Por otro lado, en Zhang et al. (2021) se ha demostrado que la proteína ORF8 emplea sutilmente la vía de la autofagia que generalmente funciona como una estrategia antiviral para alcanzar su propósito. Esto sugiere que las funcionalidades GO:0044238 (proceso metabólico primario), GO:0043170 (proceso metabólico macromolécula), GO:0009896 (regulación positiva del proceso catabólico) y GO:0016788 (actividad hidrolasa actuando sobre enlaces éster) predichas serían correctas.

30. *ORF9b*: Desempeña un papel en la interacción huésped-virus. En este caso, la funcionalidad GO:0005524 (unión a ATP) predicha comparte solamente el

Figura 25: Grafo experimental para ORF8. La funcionalidad más específica unión a proteínas (GO:0005515) se encuentra identificada con un círculo rosa. La funcionalidad GO predicha común (GO:0005488) se encuentra identificada con un círculo verde. Esta funcionalidad proviene del grafo predicho GO:0005524 (unión a ATP).

término GO:0005488 (unión) del GO experimental GO:0005515 (unión a proteínas) debido a que esta función es muy general (ver Figura 25). Por otro lado, se sabe que la expresión transitoria de ORF9b condujo a una fuerte inducción de la autofagia en las células mediadas por ATG5, un regulador crítico de la autofagia (Chong-Shan et al., 2014). Esto sugiere que la funcionalidad GO:0009896 (regulación positiva del proceso catabólico) predicha sería correcta.

31. *ORF9c*: Desempeña un papel en la interacción huésped-virus. En este caso, todos los GOs predichos son del subdominio "Proceso Biológico"mientras que el GO:0005515 experimental es una función muy general y del subdominio "Función Molecular". Por otro lado, existe evidencia que la proteína ORF9c interactúa con los receptores Sigma que están implicados en la remodelación de los lípidos y la respuesta al estrés ER; y que también interactúa con las moléculas relacionadas con NF-kB (Gordon et al., 2020). Esto sugiere que las funcionalidades GO:0044249 (biosintético celular), GO:0048518 (regulación positiva de procesos biológicos) y GO:1901576 (biosintéticos de sustancias orgánicas) predichas serían correctas. Además, ORF9c suprime la respuesta antiviral. En particular, la expresión de ORF9c impide la señalización de interferón, el procesamiento y la presentación de antígenos, la señalización de complemento y de IL-6 inducida (Dominguez et al., 2020). Esto sugiere que la funcionalidad GO:0001959 (regulación de la vía de señalización mediada por citoquinas) predicha sería correcta.

32. *ORF10*: No posee hasta el momento GO experimentales. Desde el punto de vista biológico, las funcionalidades GO:0001959 (regulación de la vía de señalización mediada por citoquinas), GO:0003824 (actividad catalítica) y GO:0016070 (proceso metabólico del ARN) predichas pueden ser viables, sin embargo, no se encontró evidencia experimental reciente que pueda vincularlas.

5. Conclusiones

En la era de la secuenciación de próxima generación, las tecnologías de secuenciación de alto rendimiento se están explotando continuamente para producir una cantidad masiva de datos de secuenciación novedosos y no caracterizados. La anotación manual de dichos datos es una tarea compleja, laboriosa y que requiere mucho tiempo

y que sólo puede ser realizada por biocuradores experimentados. En tal sentido, en este trabajo se abordó la predicción automática de funcionalidades biológicas sobre SARS-CoV-2 a partir de secuencias de productos génicos. Los resultados obtenidos cumplieron satisfactoriamente los objetivos específicos planteados. En particular, se construyeron dos bases de datos públicas de productos génicos de SARS-COV-2 compuesta de las secuencias de aminoácidos y sus correspondientes anotaciones GO. Se desarrolló un nuevo método de caracterización basados en secuencias virales que permitió aumentar las métricas de rendimiento en funcionalidades GO específicas cuando se las combina con otros organismos virales. Asimismo, la predicción de los productos génicos de SARS-COV-2 obtenidas con FGGA superaron en todas las medidas jerárquicas el 90 % y esto nos permitió confirmar que el modelo de clasificación automática creado exclusivamente con secuencias de SARS-COV-2 es bueno. Por último, se validó el modelo generado con 31 productos génicos extraídos desde Jungreis et al. (2021) con anotación GO experimental, FiloDB, y se logró la verificación de muchas funcionalidades biológicas GO. Además, esta validación permitió generar nuevas anotaciones in-silico que cuenta con evidencia bibliográfica. Aquí debemos notar que la escasa evidencia experimental disponible hasta el momento sobre algunos productos génicos de SARS-CoV-2 no permitió validar muchas funciones biológicas predichas que a priori serían válidas.

Como línea futura de trabajo se plantea continuar con la inclusión de nuevas herramientas de estimación de parámetros virales partiendo de secuencias dentro del método desarrollado de caracterización pero que incluya varios organismos virales. También, se pretende colocar la herramienta en predicción automática en un servidor web para dar acceso a la comunidad.

6. Bibliografía

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990. ISSN 0022-2836.

M. Ashburner and et al. Gene ontology: Tool for the unification of biology. *The Gene Ontology Consortium. Nature genetics*, 25:25–29, 2000.

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

M. Bianchi, A. Borsetti, M. Ciccozzi, and S. Pascarella. Sars-cov-2 orf3a: Mutability and function. *International Journal of Biological Macromolecules*, 170:820–826, 2021. ISSN 0141-8130.

M. Blum, H.-Y. Chang, S. Chuguransky, T. Grego, Kandasaamy, and et al. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, 49(D1):D344–D354, 2020.

A. V. Brown, S. I. Conners, W. Huang, A. P. Wilkey, and et al. A new decade and new data at soybase, the usda-ars soybean genetics and genomics database. *Nucleic Acids Research*, 49(D1):D1496–D1501, 2020.

J. Chen, Z. Li, J. Guo, S. Xu, J. Zhou, Q. Chen, X. Tong, D. Wang, G. Peng, L. Fang, and S. Xiao. Sars-cov-2 nsp5 exhibits stronger catalytic activity and interferon antagonism than its sars-cov ortholog. *Journal of Virology*, 96(8):e00037–22, 2022.

E. Chiacchiera, E. Tapia, and F. E. Spetale. Automatic go prediction of proteins on sars-cov-2. In *Book Abstract: XI CAB2C*, page 26. A2B2C, 2021.

M. Chiara, F. Zambelli, M. A. Tangaro, P. Mandreoli, D. S. Horner, and G. Pesole. Corgat: a tool for the functional annotation of sars-cov-2 genomes. *Bioinformatics*, 36(22-23):5522–5523, 2020.

S. Chong-Shan, Q. Hai-Yan, B. Cedric, H. Ning-Na, M. Abu-Asab, S. James H., and K. John H. Sars-coronavirus open reading frame-9b suppresses innate immunity by targeting mitochondria and the mavs/traf3/traf6 signalosome. *The Journal of Immunology*, 193(6):3080–3089, 2014.

P. Y. Chou and G. D. Fasman. Prediction of protein conformation. *Biochemistry*, 13(2):222–245, 1974.

J. N. Clifford, M. H. Høie, S. Deleuran, B. Peters, M. Nielsen, and P. Marcatili. Bepipred-3.0: Improved b-cell epitope prediction using protein language models. *Protein Science*, 31(12):e4497, 2022.

T. U. Consortium. Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 2020.

M. L. DeDiego, J. L. Nieto-Torres, J. M. Jiménez-Guardeño, J. A. Regla-Nava, E. Alvarez, J. C. Oliveros, J. Zhao, C. Fett, S. Perlman, and L. Enjuanes. Severe acute respiratory syndrome coronavirus envelope protein regulates cell stress response and apoptosis. *PLoS Pathog*, 7:e1002315–e1002315, 2011.

N. J. Dimonaco, M. Salavati, and B. B. Shih. Computational analysis of sars-cov-2 and sars-like coronavirus diversity in human, bat and pangolin populations. *Viruses*, 13(1), 2021.

E. Domingo, C. Garcia-Crespo, R. Lobo-Vega, and C. Perales. Mutation rates, mutation frequencies, and proofreading-repair activities in rna virus genetics. *viruses*, pages 1–15, 2021.

A. A. Dominguez, Y. Feng, A. R. Campos, J. Yin, C.-C. Yang, B. James, R. Murad, H. Kim, A. J. Deshpande, D. E. Gordon, N. Krogan, R. Pippa, and Z. A. Ronai. Sars-cov-2 orf9c is a membrane-associated protein that suppresses antiviral responses in cells. *bioRxiv*, 2020.

X. Dong and M. Strous. An integrated pipeline for annotation and visualization of metagenomic contigs. *Frontiers in Genetics*, 10:999, 2019.

R. Eisner, B. Poulin, D. Szafron, P. Lu, and R. Greiner. Improving protein function prediction using the hierarchical structure of the gene ontology. In *Proc. IEEE CIBCB*, 2005.

Y. Finkel, O. Mizrahi, A. Nachshon, and et al. The coding capacity of sars-cov-2. *Nature*, 589:125–130, 2021.

D. N. Gardiol. *Diagnóstico en virología*. UNR Editora, Rosario, 2011.

D. Gordon, G. Jang, and M. e. a. Bouhaddou. A sars-cov-2 protein interaction map reveals targets for drug repurposing. *Nature*, 583(6):459–468, 2020.

B. G. Hogue and C. E. Machamer. *Coronavirus Structural Proteins and Virus Assembly*, chapter 12, pages 179–200. John Wiley & Sons, Ltd, 2007. ISBN 9781683671534.

J. Huerta-Cepas, D. Szklarczyk, D. Heller, A. Hernández-Plaza, and et al. eggnog 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1):D309–D314, 2018.

E. G. Hutchinson and J. M. Thornton. A revised set of potentials for β-turn formation in proteins. *Protein Science*, 3(12):2207–2216, 1994.

M. H. Høie, E. N. Kiehl, B. Petersen, M. Nielsen, O. Winther, H. Nielsen, J. Hallgren, and P. Marcatili. NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Research*, 50(W1):W510–W515, 2022.

A. Jain and D. Kihara. Phylo-pfp: improved automated protein function prediction using phylogenetic distance of distantly related sequences. *Bioinformatics*, 35(5): 753–759, 2018.

K. J. Jang, S. Jeong, D. Y. Kang, N. Sp, Y. M. Yang, and D. E. Kim. A high atp concentration enhances the cooperative translocation of the sars coronavirus helicase nsp13 in the unwinding of duplex rna. 10 (1), 2020.

I. Jungreis, R. Sealfon, and M. Kellis. Sars-cov-2 gene content and covid-19 mutation impact by comparing 44 sarbecovirus genomes. *Nature Communications*, 12:2642, 2021.

W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

Y. Kasuga, B. Zhu, K.-J. Jang, and J.-S. Yoo. Innate immune sensing of coronavirus and viral evasion strategies. *Experimental Molecular Medicine*, 53:723–736, 2021.

K. Kieft, Z. Zhou, and K. Anantharaman. Vibrant: Automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequence. *Microbiome*, 8:90, 2020.

L. Kiemer, O. Lund, S. Brunak, and N. Blom. Coronavirus 3clpro proteinase cleavage sites: possible relevance to sars virus pathology. *BMC bioinformatics*, 5(72), 2004.

M. Kircher and J. Kelso. High-throughput dna sequencing, concepts and limitations. *Bioessays*, 32(6):524–536, 2010. ISSN 1521-1878.

S. Kiritchenko, S. Matwin, and A. F. Famili. Functional annotation of genes using hierarchical text categorization. In *in Proc. of the BioLINK SIG: Linking Literature, Information and Knowledge for Biology (held at ISMB-05*, 2005.

A. Kosmopoulos, I. Partalas, E. Gaussier, G. Paliouras, and I. Androutsopoulos. Evaluation measures for hierarchical classification: A unified view and novel approaches. *Data Min. Knowl. Discov.*, 29(3):820–865, 2015.

F. R. Kschischang, B. J. Frey, and H. A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Inf. Theor.*, 47(2):498–519, 2001.

E. Lam-Cabanillas, A. León-Risco1, K. León-Risco1, G. Llamo-Hoyos1, R. López-Zavaleta, E. Luzuriaga-Tirado, and J. Mendoza-Blas, Alex andHuamán-Saavedra. Bases moleculares de la patogÉnesis de covid-19 y estudios in silico de posibles tratamientos farmacolÓgicos. *Fac. Med. Hum.*, 21:417–432, 2021.

B. Lee, M. Shin, Y. Oh, H. Oh, and K. Ryu. Identification of protein functions using a machine-learning approach based on sequence-derived properties. *Proteome Science*, 7(1):27, 2009. ISSN 1477-5956.

D. Lee, O. Redfern, and C. Orengo. Predicting protein function from sequence and structure. *Nat Rev Mol Cell Biol*, 8:995–1005, 2007.

X. Lian, X. Yang, J. Shao, F. Hou, S. Yang, D. Pan, and Z. Zhang. Prediction and analysis of human-herpes simplex virus type 1 protein-protein interactions by integrating multiple methods. *Quant. Biol.*, 8:312–324, 2020.

X. Lian, X. Yang, S. Yang, and Z. Zhang. Current status and future perspectives of computational studies on human–virus protein–protein interactions. *Briefings in Bioinformatics*, 2021.

M. F. Lin, I. Jungreis, and M. Kellis. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, 27(13):i275–i282, 2011.

J. Lon, Y. Bai, B. Zhong, F. Cai, and H. Du. Prediction and evolution of b cell epitopes of surface protein in sars-cov-2. *Virology Journal*, 17(165), 2020.

Z. Y. Low, N. Z. Zabidi, A. J. W. Yip, A. Puniyamurti, V. T. K. Chow, and S. K. Lal. Sars-cov-2 non-structural proteins and their roles in host immune evasion. *Viruses*, 14(9), 2022.

N. G. Naik, S.-C. Lee, B. H. S. Veronese, Z. Ma, and Z. Toth. Interaction of hdac2 with sars-cov-2 nsp5 and irf3 is not required for nsp5-mediated inhibition of type i interferon signaling pathway. *Microbiology Spectrum*, 10(5):e02322–22, 2022.

M. Nielsen, C. Lundegaard, O. Lund, and T. N. Petersen. Cphmodels-3.0: remote homology modeling using structure-guided sequence profiles. *Nucleic Acids Research*, 38(suppl2):W576–W581, 2010.

D. Osorio, P. Rondon-Villarreal, and R. Torres. Peptides: A package for data mining of antimicrobial peptides. *The R Journal*, 7(1):4–14, 2015.

S. I. O'Donoghue, A. Schafferhans, N. Sikta, C. Stolte, S. Kaur, B. K. Ho, S. Anderson, J. B. Procter, C. Dallago, N. Bordin, M. Adcock, and B. Rost. Sars-cov-2 structural coverage map reveals viral protein assembly, mimicry, and hijacking mechanisms. *Molecular Systems Biology*, 17(9):e10079, 2021.

B. Petersen, C. Lundegaard, and T. N. Petersen. Netturnp - predicción de redes neuronales de vueltas beta mediante el uso de información evolutiva y características de secuencia de proteínas predicha. *PLoS ONE*, 5:e15079–e15079, 2010.

C. Qin, L. Zhou, Z. Hu, S. Zhang, S. Yang, Y. Tao, and et al. Dysregulation of immune response in patients with coronavirus 2019 (covid-19) in wuhan, china. 71(15):762–768, 2020.

N. Redondo, S. Zaldívar-López, J. Garrido, and M. Montoya. Sars-cov-2 accessory proteins in viral pathogenesis: Knowns and unknowns. *Frontiers in Immunology*, 12, 2021.

J. A. Reuter, D. V. Spacek, and M. P. Snyder. High-throughput sequencing technologies. *Molecular Cell*, 58(4):586 – 597, 2015.

P. Skewes-Cox, T. J. Sharpton, K. S. Pollard, and J. L. DeRisi. Profile hidden markov models for the detection of viruses within metagenomic sequence data. *PLOS ONE*, 9(8):1–12, 2014.

M. L. Smith and M. W. Hahn. New approaches for inferring phylogenies in the presence of paralogs. *Trends in Genetics*, 37(2):174–187, 2021.

A. Som. Causes, consequences and solutions of phylogenetic incongruence. *Briefings in Bioinformatics*, 16(3):536–548, 2014.

F. Spetale, E. Tapia, F. Krsticevic, F. Roda, and P. Bulacio. A Factor Graph Approach to Automated GO Annotation. *PLoS ONE*, 11(1):1–16, 2016.

F. Spetale, D. Arce, F. Krsticevic, P. Bulacio, and E. Tapia. Consistent prediction of go protein localization. *Scientific Reports*, 8, 05 2018.

F. E. Spetale. fgga: Hierarchical ensemble method based on factor graph. *Bioconductor*, 2022. URL http://www.bioconductor.org/packages/fgga/.

C. Szeto, D. S. M. Chatzileontiadou, A. T. Nguyen, H. Sloane, C. A. Lobos, D. Jayasinghe, H. Halim, C. Smith, A. Riboldi-Tunnicliffe, E. J. Grant, and S. Gras. The presentation of sars-cov-2 peptides by the common hla-a∗02:01 molecule. *iScience*, 24(2):102096, 2021.

D. Szklarczyk, A. L. Gable, K. C. Nastou, D. Lyon, and et al. The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1):D605–D612, 2020.

A. te Velthuis, S. van den Worm, and E. Snijder. The sars-coronavirus nsp7+nsp8 complex is a unique multimeric rna polymerase capable of both de novo initiation and primer extension. *Nucleic Acids Res*, 40(1737-1747), 2012.

F. Teufel, J. J. Almagro Armenteros, and A. R. Johansen. Signalp 6.0 predicts all five types of signal peptides using protein language models. *Nature Biotechnology*, 40:1023–1025, 2022.

V. Thumuluri, J. J. Almagro Armenteros, A. R. Johansen, H. Nielsen, and O. Winther. Deeploc 2.0: multi-label subcellular localization prediction using protein language models. *Nucleic Acids Research*, 50(W1):W228–W234, 2022.

K. Verspoor, J. Cohn, S. Mnizewski, and J. C. A categorization approach to automated ontological function annotation. *Protein Science*, 15:1544–1549, 2006.

Q. Wang, J. Wu, H. Wang, Y. Gao, Q. Liu, A. Mu, W. Ji, and et al. Structural basis for rna replication by the sars-cov-2 polymerase. *Cell*, 182(2):417–428.e13, 2020.

Q. Wei and R. L. Dunbrack. The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PloS one*, 8(7), 2013. ISSN 1932-6203.

Z. Yang, X. Zhang, F. Wang, P. Wang, E. Kuang, and X. Li. Suppression of mda5-mediated antiviral immune responses by nsp8 of sars-cov-2. *bioRxiv*, 2020.

Y. YZheng, J. Deng, and L. e. a. Han. Sars-cov-2 nsp5 and n protein counteract the rig-i signaling pathway by suppressing the formation of stress granules. *Sig Transduct Target Ther*, 7(22):165–174, 2022.

K.-Y. Zhang, Y.-Z. Gao, M.-Z. Du, S. Liu, C. Dong, and F.-B. Guo. Vgas: A viral genome annotation system. *Frontiers in Microbiology*, 10:184, 2019.

L. Zhang, D. Lin, X. Sun, U. Curth, C. Drosten, L. Sauerhering, S. Becker, K. Rox, and R. Hilgenfeld. Crystal structure of sars-cov-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science*, 368:409–412, 2020.

Y. Zhang, Y. Chen, Y. Li, F. Huang, and et al. The orf8 protein of sars-cov-2 mediates immune evasion through down-regulating mhc-ı. *Proceedings of the National Academy of Sciences*, 118(23):e2024202118, 2021.

Z. Zhou, C. Huang, Z. Zhou, Z. Huang, L. Su, S. Kang, X. Chen, Q. Chen, S. He, X. Rong, F. Xiao, J. Chen, and S. Chen. Structural insight reveals sars-cov-2 orf7a as an immunomodulating factor for human cd14+ monocytes. 24 (3):331 – 337, 2021.

Anexo I

Tabla 8: Base de conocimiento, FiloDB, basada en Jungreis and et al (2021)

Proteína	**Secuencia**	**GOs**
ORF1ab	MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQHLKDGTCGLVEVEK GVLPQLEQPYVFIKRSDARTAPHGHVMVELVAELEGIQYGRSGETLGVLVPHVGEIPVA YRKVLLRKNGNKGAGGHSYGADLKSFDLGDELGTDPYEDFQENWNTKHSSGVTRELM RELNGGAYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQLDFIDTKRGVYCCR EHEHEIAWYTERSEKSYELQTPFEIKLAKKFDTFNGECPNFVFPLNSIIKTIQPRVEKKK LDGFMGRIRSVYPVASPNECNQMCLSTLMKCDHCGETSWQTGDFVKA	GO:0032774, GO:0008233
ORF3a	MDLFMRIFTIGTVTLKQGEIKDATPSDFVRATATIPIQASLPFGWLIVGVALLAVFQSAS KIITLKKRWQLALSKGVHFVCNLLLLFVTVYSHLLLVAAGLEAPFLYLYALVYFLQSINF VRIIMRLWLCWKCRSKNPLLYDANYFLCWHTNCYDYCIPYNSVTSSIVITSGDGTTSPIS EHDYQIGGYTEKWESGVKDCVVLHSYFTSDYYQLYSTQLSTDTGVEHVTFFIYNKIVDE PEEHVQIHTIDGSSGVVNPVMEPIYDEPTTTTSVPL	GO:0039656, GO:0019076, GO:0005216, GO:0039707, GO:0097194
ORF6	MFHLVDFQVTIAEILLIIMRTFKVSIWNLDYIINLIIKNLSKSLTENKYSQLDEEQPMEID	GO:0030683
ORF7a	MKIILFLALITLATCELYHYQECVRGTTVLLKEPCSSGTYEGnspFHPLADNKFALTCFST QFAFACPDGVKHVYQLRARSVSPKLFIRQEEVQELYSPIFLIVAAIVFITLCFTLKRKTE	GO:0019079
ORF7b	MIELSLIDFYLCFLAFLLFLVLIMLIIFWFSLELQDHNETCHA	
ORF8	MKFLVFLGIITTVAAFHQECSLQSCTQHQPYVVDDPCPIHFYSKWYIRVGARKSAPLIEL CVDEAGSKSPIQYIDIGNYTVSCLPFTINCQEPKLGSLVVRCSFYEDFLEYHDVRVVLDFI	GO:0005515
ORF9b	MDPKISEMHPALRLVDPQIQLAVTRMENAVGRDQNNVGPKVYPIILRLGSPLSLNMARK TLNSLEDKAFQLTPIAVQMTKLATTEELPDEFVVVTVK	GO:0005515
ORF9c	MLQSCYNFLKEQHCQKASTQKGAEAAVKPLLVPHHVVATVQEIQLQAAVGELLLLEWL AMAVMLLLLCCCLTD	GO:0005515
ORF10	MGYINVFAFPFTIYSLLLCRMNSRNYIAQVDVVNFNLT	
S1	SQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFSNVTWFHAIHVS GTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIVNNATNVVIKVC EFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLEGKQGNFKNL REFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQTLLALHRSYLTP GDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETKCTLKSFTVE KGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISNCVADYSVLY NSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIADYNYKLPDD FTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFN CYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVNFNFNG LTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITPGTNTS NQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECD IPIGAGICASYQTQTnspRRAR	GO:0046790, GO:0061025
S2	SVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSVDCTMYICGDST ECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGFNFSQILP DPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFNGLTVLPPLLTD EMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIAN QFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISSVLNDILSRLDK VEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLGQSKRVDFCGK GYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREGVFVSNGTHW FVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEELDKYFKNHTSP DVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKWPWYIWLGFIA GLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLHYT	GO:0006897
S2’	NCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSYECDIPIGAGICASY QTQTnspRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTISVTTEILPVSMTKTSV DCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIK DFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDCLGDIAARDLICAQKFN GLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQN VLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALNTLVKQLSSNFGAISS VLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRASANLAATKMSECVLG QSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPAICHDGKAHFPREG VFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDPLQPELDSFKEEL DKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDLQELGKYEQYIKW PWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDDSEPVLKGVKLH YT	GO:0075509, GO:0006508

Tabla 8: (Continuación) Base de conocimiento, FiloDB, basada en Jungreis and et al (2021)

Proteína	**Secuencia**	**GOs**
nsp1	MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSVEEVLSEARQHLKDGTCGLVEVEK GVLPQLEQPYVFIKRSDARTAPHGHVMVELVAELEGIQYGRSGETLGVLVPHVGEIPVA YRKVLLRKNGNKGAGGHSYGADLKSFDLGDELGTDPYEDFQENWNTKHSSGVTRELM RELNGG	GO:0039604, GO:0042783, GO:0019080
nsp2	AYTRYVDNNFCGPDGYPLECIKDLLARAGKASCTLSEQLDFIDTKRGVYCCREHEHEIA WYTERSEKSYELQTPFEIKLAKKFDTFNGECPNFVFPLNSIIKTIQPRVEKKKLDGFMG RIRSVYPVASPNECNQMCLSTLMKCDHCGETSWQTGDFVKATCEFCGTENLTKEGAT TCGYLPQNAVVKIYCPACHNSEVGPEHSLAEYHNESGLKTILRKGGRTIAFGGCVFSYV GCHNKCAYWVPRASANIGCNHTGVVGEGSEGLNDNLLEILQKEKVNINIVGDFKLNEEI AIILASFSASTSAFVETVKGLDYKAFKQIVESCGNFKVTKGKAKKGAWNIGEQKSILSPL YAFASEAARVVRSIFSRTLETAQNSVRVLQKAAITILDGISQYSLRLIDAMMFTSDLATNN LVVMAYITGGVVQLTSQWLTNIFGTVYEKLKPVLDWLEEKFKEGVEFLRDGWEIVKFI STCACEIVGGQIVTCAKEIKESVQTFFKLVNKFLALCADSIIIGGAKLKALNLGETFVTHS KGLYRKCVKSREETGLLMPLKAPKEIIFLEGETLPTEVLTEEVVLKTGDLQPLEQPTSE AVEAPLVGTPVCINGLMLLEIKDTEKYCALAPNMMVTNNTFTLKGG	GO:0033554
nsp3	APTKVTFGDDTVIEVQGYKSVNITFELDERIDKVLNEKCSAYTVELGTEVNEFACVVAD AVIKTLQPVSELLTPLGIDLDEWSMATYYLFDESGEFKLASHMYCSFYPPDEDEEEGDC EEEEFEPSTQYEYGTEDDYQGKPLEFGATSAALQPEEEQEEDWLDDDSQQTVGQQDG SEDNQTTTIQTIVEVQPQLEMELTPVVQTIEVNSFSGYLKLTDNVYIKNADIVEEAKKVK PTVVVNAANVYLKHGGGVAGALNKATNNAMQVESDDYIATNGPLKVGGSCVLSGHNL AKHCLHVVGPNVNKGEDIQLLKSAYENFNQHEVLLAPLLSAGIFGADPIHSLRVCVDTV RTNVYLAVFDKNLYDKLVSSFLEMKSEKQVEQKIAEIPKEEVKPFITESKPSVEQRKQD DKKIKACVEEVTTTLEETKFLTENLLLYIDINGNLHPDSATLVSDIDITFLKKDAPYIVGD VVQEGVLTAVVIPTKKAGGTTEMLAKALRKVPTDNYITTYPGQGLNGYTVEEAKTVL KKCKSAFYILPSIISNEKQEILGTVSWNLREMLAHAEETRKLMPVCVETKAIVSTIQRKY KGIKIQEGVVDYGARFYFYTSKTTVASLINTLNDLNETLVTMPLGYVTHGLNLEEAARY MRSLKVPATVSVSSPDAVTAYNGYLTSSSKTPEEHFIETISLAGSYKDWSYSGQSTQLGI EFLKRGDKSVYYTSNPTTFHLDGEVITFDNLKTLLSLREVRTIKVFTTVDNINLHTQVVD MSMTYGQQFGPTYLDGADVTKIKPHNSHEGKTFYVLPNDDTLRVEAFEYYHTTDPSFL GRYMSALNHTKKWKYPQVNGLTSIKWADNNCYLATALLTLQQIELKFNPPALQDAYY RARAGEAANFCALILAYCNKTVGELGDVRETMSYLFQHANLDSCKRVLNVVCKTCGQ QQTTLKGVEAVMYMGTLSYEQFKKGVQIPCTCGKQATKYLVQQESPFVMMSAPPAQY ELKHGTFTCASEYTGNYQCGHYKHITSKETLYCIDGALLTKSSEYKGPITDVFYKENSY TTTIKPVTYKLDGVVCTEIDPKLDNYYKKDNSYFTEQPIDLVPNQPYPNASFDNFKFVC DNIKFADDLNQLTGYKKPASRELKVTFFPDLNGDVVAIDYKHYTPSFKKGAKLLHKPIV WHVNNATNKATYKPNTWCIRCLWSTKPVETSNSFDVLKSEDAQGMDNLACEDLKPVS EEVVENPTIQKDVLECNVKTTEVVGDIILKPANNSLKITEEVGHTDLMAAYVDNSSLTIK KPNELSRVLGLKTLATHGLAAVNSVPWDTIANYAKPFLNKVVSTTTNIVTRCLNRVCT NYMPYFFTLLLQLCTFTRSTNSRIKASMPTTIAKNTVKSVGKFCLEASFNYLKSPNFSKL INIIIWFLLLSVCLGSLIYSTAALGVLMSNLGMPSYCTGYREGYLNSTNVTIATYCTGSIPC SVCLSGLDSLDTYPSLETIQITISSFKWDLTAFGLVAEWFLAYILFTRFFYVLGLAAIMQL FFSYFAVHFISNSWLMWLIINLVQMAPISAMVRMYIFFASFYYVWKSYVHVVDGCNSST CMMCYKRNRATRVECTTIVNGVRRSFYVYANGGKGFCKLHNWNCVNCDTFCAGSTFI SDEVARDLSLQFKRPINPTDQSSYIVDSVTVKNGSIHLYFDKAGQKTYERHSLSHFVNLD NLRANNTKGSLPINVIVFDGKSKCEESSAKSASVYYSQLMCQPILLLDQALVSDVGDSAE VAVKMFDAYVNTFSSTFNVPMEKLKTLVATAEAELAKNVSLDNVLSTFISAARQGFVD SDVETKDVVECLKLSHQSDIEVTGDSCNNYMLTYNKVENMTPRDLGACIDCSARHINAQ VAKSHNIALIWNVKDFMSLSEQLRKQIRSAAKKNNLPFKLTCATTRQVVNVVTTKIALK GG	GO:0016579, GO:0140526, GO:0019783, GO:0001960, GO:0008233, GO:0039503
nsp4	KIVNNWLKQLIKVTLVFLFVAAIFYLITPVHVMSKHTDFSSEIIGYKAIDGGVTRDIASTD TCFANKHADFDTWFSQRGGSYTNDKACPLIAAVITREVGFVVPGLPGTILRTTNGDFL HFLPRVFSAVGNICYTPSKLIEYTDFATSACVLAAECTIFKDASGKPVPYCYDTNVLEGS VAYESLRPDTRYVLMDGSIIQFPNTYLEGSVRVVTTFDSEYCRHGTCERSEAGVCVSTS GRWVLNNDYYRSLPGVFCGVDAVNLLTNMFTPLIQPIGALDISASIVAGGIVAIVVTCLA YYFMRFRRAFGEYSHVVAFNTLLFLMSFTVLCLTPVYSFLPGVYSVIYLYLTFYLTNDV SFLAHIQWMVMFTPLVPFWITIAYIICISTKHFYWFFSNYLKRRVVFNGVSFSTFEEAAL CTFLLNKEMYLKLRSDVLLPLTQYNRYLALYNKYKYFSGAMDTTSYREAACCHLAKAL NDFSNSGSDVLYQPPQTSITSAVLQ	GO:0140526
nsp5	SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDVVYCPRHVICTSEDMLNPNYEDL LIRKSNHNFLVQAGNVQLRVIGHSMQNCVLKLKVDTANPKTPKYKFVRIQPGQTFSVLA CYNGSPSGVYQCAMRPNFTIKGSFLNGSCGSVGFNIDYDCVSFCYMHHMELPTGVHAG TDLEGNFYGPFVDRQTAQAAGTDTTITVNVLAWLYAAVINGDRWFLNRFTTTLNDFN LVAMKYNYEPLTQDHVDILGPLSAQTGIAVLDMCASLKELLQNGMNGRTILGSALLEDE FTPFDVVRQCSGVTFQ	GO:0004180, GO:0072570
nsp6	SAVKRTIKGTHHWLLLTILTSLLVLVQSTQWSLFFFLYENAFLPFAMGIIAMSAFAMMFV KHKHAFLCLFLLPSLATVAYFNMVYMPASWVMRIMTWLDMVDTSLSGFKLKDCVMYA SAVVLLILMTARTVYDDGARRVWTLMNVLTLVYKVYYGNALDQAISMWALIISVTSNY SGVVTTVMFLARGIVFMCVEYCPIFFITGNTLQCIMLVYCFLGYFCTCYFGLFCLLNRYF RLTLGVYDYLVSTQEFRYMNSQGLLPPKNSIDAFKLNIKLLGVGGKPCIKVATVQ	GO:2000786, GO:1901096

Tabla 8: (Continuación) Base de conocimiento, FiloDB, basada en Jungreis and et al (2021)

Proteína	**Secuencia**	**GOs**
nsp7	SKMSDVKCTSVVLLSVLQQLRVESSSKLWAQCVQLHNDILLAKDTTEAFEKMVSLLSVL LSMQGAVDINKLCEEMLDNRATLQ	GO:0019079, GO:0003968
nsp8	PSTQYEYGTEDDYQGKPLEFGATSAALQPEEEQEEDWLDDDSQQTVGQQDGSEDNQT TTIQTIVEVQPQLEMELTPVVQTIEVNSFSGYLKLTDNVYIKNADIVEEAKKVKPTVVV NAANVYLKHGGGVAGALNKATNNAMQVESDDYIATNGPLKVGGSCVLSGHNLAKHCL HVVGPNVNKGEDIQLLKSAYENFNQHEVLLAPLLSAGIFGADPIHSLRVCVDTVRTNVY LAVFDKNLYDKLVSSFLEMKSEKQVEQKIAEIPKEEVKPFITESKPSVEQRKQDDKKIKA CVEEVTTTLEETKFLTENLLLYIDINGNLHPDSATLVSDIDITFLKKDAPYIVGDVVQEG VLTAVVIPTKKAGGTTEMLAKALRKVPTDNYITTYPGQGLNGYTVEEAKTVLKKCKS AFYILPSIISNEKQEILGTVSWNLREMLAHAEETRKLMPVCVETKAIVSTIQRKYKGIKIQ EGVVDYGARFYFYTSKTTVASLINTLNDLNETLVTMPLGYVTHGLNLEEAARYMRSLK VPATVSVSSPDAVTAYNGYLTSSSKTPEEHFIETISLAGSYKDWSYSGQSTQLGIEFLKRG DKSVYYTSNPTTFHLDGEVITFDNLKTLLSLREVRTIKVFTTVDNINLHTQVVDMSMTY GQQFGPTYLDGADVTKIKPHNSHEGKTFYVLPNDDTLRVEAFEYYHTTDPSFLGRYMS ALNHTKKWKYPQVNGLTSIKWADNNCYLATALLTLQQIELKFNPPALQDAYYRARAG EAANFCALILAYCNKTVGELGDVRETMSYLFQHANLDSCKRVLNVVCKTCGQQQTTLK GVEAVMYMGTLSYEQFKKGVQIPCTCGKQATKYLVQQESPFVMMSAPPAQYELKHGT FTCASEYTGNYQCGHYKHITSKETLYCIDGALLTKSSEYKGPITDVFYKENSYTTTIKPV TYKLDGVVCTEIDPKLDNYYKKDNSYFTEQPIDLVPNQPYPNASFDNFKFVCDNIKFAD DLNQLTGYKKPASRELKVTFFPDLNGDVVAIDYKHYTPSFKKGAKLLHKPIVWHVNNA TNKATYKPNTWCIRCLWSTKPVETSNSFDVLKSEDAQGMDNLACEDLKPVSEEVVENP TIQKDVLECNVKTTEVVGDIILKPANNSLKITEEVGHTDLMAAYVDNSSLTIKKPNELSR VLGLKTLATHGLAAVNSVPWDTIANYAKPFLNKVVSTTTNIVTRCLNRVCTNYMPYFF TLLLQLCTFTRSTNSRIKASMPTTIAKNTVKSVGKFCLEASFNYLKSPNFSKLINIIIWFLL LSVCLGSLIYSTAALGVLMSNLGMPSYCTGYREGYLNSTNVTIATYCTGSIPCSVCLSGL DSLDTYPSLETIQITISSFKWDLTAFGLVAEWFLAYILFTRFFYVLGLAAIMQLFFSYFAV HFISNSWLMWLIINLVQMAPISAMVRMYIFFASFYYVWKSYVHVVDGCNSSTCMMCYK RNRATRVECTTIVNGVRRSFYVYANGGKGFCKLHNWNCVNCDTFCAGSTFISDEVAR DLSLQFKRPINPTDQSSYIVDSVTVKNGSIHLYFDKAGQKTYERHSLSHFVNLDNLRANN TKGSLPINVIVFDGKSKCEESSAKSASVYYSQLMCQPILLLDQALVSDVGDSAEVAVKMF DAYVNTFSSTFNVPMEKLKTLVATAEAELAKNVSLDNVLSTFISAARQGFVDSDVETKD VVECLKLSHQSDIEVTGDSCNNYMLTYNKVENMTPRDLGACIDCSARHINAQVAKSHNI ALIWNVKDFMSLSEQLRKQIRSAAKKNNLPFKLTCATTRQVVNVVTTKIALKGGKIVNN WLKQLIKVTLVFLFVAAIFYLITPVHVMSKHTDFSSEIIGYKAIDGGVTRDIASTDTCFAN KHADFDTWFSQRGGSYTNDKACPLIAAVITREVGFVVPGLPGTILRTTNGDFLHFLPRV FSAVGNICYTPSKLIEYTDFATSACVLAAECTIFKDASGKPVPYCYDTNVLEGSVAYESL RPDTRYVLMDGSIIQFPNTYLEGSVRVVTTFDSEYCRHGTCERSEAGVCVSTSGRWVLN NDYYRSLPGVFCGVDAVNLLTNMFTPLIQPIGALDISASIVAGGIVAIVVTCLAYYFMRF RRAFGEYSHVVAFNTLLFLMSFTVLCLTPVYSFLPGVYSVIYLYLTFYLTNDVSFLAHIQ WMVMFTPLVPFWITIAYIICISTKHFYWFFSNYLKRRVVFNGVSFSTFEEAALCTFLLNK EMYLKLRSDVLLPLTQYNRYLALYNKYKYFSGAMDTTSYREAACCHLAKALNDFSNSG SDVLYQPPQTSITSAVLQSGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDVVYCPR HVICTSEDMLNPNYEDLLIRKSNHNFLVQAGNVQLRVIGHSMQNCVLKLKVDTANPKTP KYKFVRIQPGQTFSVLACYNGSPSGVYQCAMRPNFTIKGSFLNGSCGSVGFNIDYDCVS FCYMHHMELPTGVHAGTDLEGNFYGPFVDRQTAQAAGTDTTITVNVLAWLYAAVING DRWFLNRFTTTLNDFNLVAMKYNYEPLTQDHVDILGPLSAQTGIAVLDMCASLKELLQ NGMNGRTILGSALLEDEFTPFDVVRQCSGVTFQSAVKRTIKGTHHWLLLTILTSLLVLV QSTQWSLFFFLYENAFLPFAMGIIAMSAFAMMFVKHKHAFLCLFLLPSLATVAYFNMVY MPASWVMRIMTWLDMVDTSLSGFKLKDCVMYASAVVLLILMTARTVYDDGARRVWT LMNVLTLVYKVYYGNALDQAISMWALIISVTSNYSGVVTTVMFLARGIVFMCVEYCPIF FITGNTLQCIMLVYCFLGYFCTCYFGLFCLLNRYFRLTLGVYDYLVSTQEFRYMNSQGL LPPKNSIDAFKLNIKLLGVGGKPCIKVATVQSKMSDVKCTSVVLLSVLQQLRVESSSKLW AQCVQLHNDILLAKDTTEAFEKMVSLLSVLLSMQGAVDINKLCEEMLDNRATLQAIASE FSSLPSYAAFATAQEAYEQAVANGDSEVVLKKLKKSLNVAKSEFDRDAAMQRKLEKMA DQAMTQMYKQARSEDKRAKVTSAMQTMLFTMLRKLDNDALNNIINNARDGCVPLNIIP LTTAAKLMVVIPDYNTYKNTCDGTTFTYASALWEIQQVVDADSKIVQLSEISMDnspNLA WPLIVTALRANSAVKLQ	GO:0003727, GO:0019079
nsp9	NNELSPVALRQMSCAAGTTQTACTDDNALAYYNTTKGGRFVLALLSDLQDLKWARFP KSDGTGTIYTELEPPCRFVTDTPKGPKVKYLYFIKGLNNLNRGMVLGSLAATVRLQ	GO:0019083, GO:0036451
nsp10	AGNATEVPANSTVLSFCAFAVDAAKAYKDYLASGGQPITNCVKMLCTHTGTGQAITVT PEANMDQESFGGASCCLYCRCHIDHPNPKGFCDLKGKYVQIPTTCANDPVGFTLKNTV CTVCGMWKGYGCSCDQLREPMLQ	
nsp11	SADAQSFLNRVCG	

Tabla 8: (Continuación) Base de conocimiento, FiloDB, basada en Jungreis and et al (2021)

Proteína	**Secuencia**	**GOs**
nsp12	SADAQSFLNRVCGVSAARLTPCGTGTSTDVVYRAFDIYNDKVAGFAKFLKTNCCRFQE KDEDDNLIDSYFVVKRHTFSNYQHEETIYNLLKDCPAVAKHDFFKFRIDGDMVPHISRQ RLTKYTMADLVYALRHFDEGNCDTLKEILVTYNCCDDDYFNKKDWYDFVENPDILRV YANLGERVRQALLKTVQFCDAMRNAGIVGVLTLDNQDLNGNWYDFGDFIQTTPGSGV PVVDSYYSLLMPILTLTRALTAESHVDTDLTKPYIKWDLLKYDFTEERLKLFDRYFKYW DQTYHPNCVNCLDDRCILHCANFNVLFSTVFPPTSFGPLVRKIFVDGVPFVVSTGYHFR ELGVVHNQDVNLHSSRLSFKELLVYAADPAMHAASGNLLLDKRTTCFSVAALTNNVAF QTVKPGNFNKDFYDFAVSKGFFKEGSSVELKHFFFAQDGNAAISDYDYYRYNLPTMCD IRQLLFVVEVVDKYFDCYDGGCINANQVIVNNLDKSAGFPFNKWGKARLYYDSMSYED QDALFAYTKRNVIPTITQMNLKYAISAKNRARTVAGVSICSTMTNRQFHQKLLKSIAAT RGATVVIGTSKFYGGWHNMLKTVYSDVENPHLMGWDYPKCDRAMPNMLRIMASLVL ARKHTTCCSLSHRFYRLANECAQVLSEMVMCGGSLYVKPGGTSSGDATTAYANSVFNI CQAVTANVNALLSTDGNKIADKYVRNLQHRLYECLYRNRDVDTDFVNEFYAYLRKHFS MMILSDDAVVCFNSTYASQGLVASIKNFKSVLYYQNNVFMSEAKCWTETDLTKGPHEF CSQHTMLVKQGDDYVYLPYPDPSRILGAGCFVDDIVKTDGTLMIERFVSLAIDAYPLTK HPNQEYADVFHLYLQYIRKLHDELTGHMLDMYSVMLTNDNTSRYWEPEFYEAMYTPH TVLQ	GO:0019083, GO:0019079, GO:0003968
nsp13	AVGACVLCNSQTSLRCGACIRRPFLCCKCCYDHVISTSHKLVLSVNPYVCNAPGCDVTD VTQLYLGGMSYYCKSHKPPISFPLCANGQVFGLYKNTCVGSDNVTDFNAIATCDWTNA GDYILANTCTERLKLFAAETLKATEETFKLSYGIATVREVLSDRELHLSWEVGKPRPPL NRNYVFTGYRVTKNSKVQIGEYTFEKGDYGDAVVYRGTTTYKLNVGDYFVLTSHTVM PLSAPTLVPQEHYVRITGLYPTLNISDEFSSNVANYQKVGMQKYSTLQGPPGTGKSHFA IGLALYYPSARIVYTACSHAAVDALCEKALKYLPIDKCSRIIPARARVECFDKFKVNSTLE QYVFCTVNALPETTADIVVFDEISMATNYDLSVVNARLRAKHYVYIGDPAQLPAPRTLL TKGTLEPEYFNSVCRLMKTIGPDMFLGTCRRCPAEIVDTVSALVYDNKLKAHKDKSAQ CFKMFYKGVITHDVSSAINRPQIGVVREFLTRNPAWRKAVFISPYNSQNAVASKILGLPT QTVDSSQGSEYDYVIFTQTTETAHSCNVNRFNVAITRAKVGILCIMSDRDLYDKLQFTSL EIPRRNVATLQ	GO:0004386
nsp14	AENVTGLFKDCSKVITGLHPTQAPTHLSVDTKFKTEGLCVDIPGIPKDMTYRRLISMMG FKMNYQVNGYPNMFITREEAIRHVRAWIGFDVEGCHATREAVGTNLPLQLGFSTGVNL VAVPTGYVDTPNNTDFSRVSAKPPPGDQFKHLIPLMYKGLPWNVVRIKIVQMLSDTLK NLSDRVVFVLWAHGFELTSMKYFVKIGPERTCCLCDRRATCFSTASDTYACWHHSIGF DYVYNPFMIDVQQWGFTGNLQSNHDLYCQVHGNAHVASCDAIMTRCLAVHECFVKRV DWTIEYPIIGDELKINAACRKVQHMVVKAALLADKFPVLHDIGNPKAIKCVPQADVEW KFYDAQPCSDKAYKIEELFYSYATHSDKFTDGVCLFWNCNVDRYPANSIVCRFDTRVLS NLNLPGCDGGSLYVNKHAFHTPAFDKSAFVNLKQLPFFYYSDSPCESHGKQVVSDIDYV PLKSATCITRCNLGGAVCRHHANEYRLYLDAYNMMISAGFSLWVYKQFDTYNLWNTF TRLQ	GO:0004532, GO:0004482
nsp15	SLENVAFNVVNKGHFDGQQGEVPVSIINNTVYTKVDGVDVELFENKTTLPVNVAFELW AKRNIKPVPEVKILNNLGVDIAANTVIWDYKRDAPAHISTIGVCSMTDIAKKPTETICAP LTVFFDGRVDGQVDLFRNARNGVLITEGSVKGLQPSVGPKQASLNGVTLIGEAVKTQF NYYKKVDGVVQQLPETYFTQSRNLQEFKPRSQMEIDFLELAMDEFIERYKLEGYAFEHI VYGDFSHSQLGGLHLLIGLAKRFKESPFELEDFIPMDSTVKNYFITDAQTGSSKCVCSVID LLLDDFVEIIKSQDLSVVSKVVKVTIDYTEISFMLWCKDGHVETFYPKLQ	GO:0008663
nsp16	SSQAWQPGVAMPNLYKMQRMLLEKCDLQNYGDSATLPKGIMMNVAKYTQLCQYLNT LTLAVPYNMRVIHFGAGSDKGVAPGTAVLRQWLPTGTLLVDSDLNDFVSDADSTLIGD CATVHTANKWDLIISDMYDPKTKNVTKENDSKEGFFTYICGFIQQKLALGGSVAIKITE HSWNADLYKLMGHFAWWTAFVTNVNASSSEAFLIGCNYLGKPREQIDGYVMHANYIF WRNTNPIQLSSYSLFDMSKFPLKLRGTAVMSLKEGQINDMILSLLSKGRLIIRENNRVVIS SDVLVNN	GO:0008168, GO:0042783
E	MYSFVSEETGTLIVNSVLLFLAFVVFLLVTLAILTALRLCAYCCNIVNVSLVKPSFYVYSR VKNLNSSRVPDLLV	GO:0019069, GO:0039707, GO:0060139
M	MADSNGTITVEELKKLLEQWNLVIGFLFLTWICLLQFAYANRNRFLYIIKLIFLWLLWPV TLACFVLAAVYRINWITGGIAIAMACLVGLMWLSYFIASFRLFARTRSMWSFNPETNILL NVPLHGTILTRPLLESELVIGAVILRGHLRIAGHHLGRCDIKDLPKEITVATSRTLSYYKL GASQRVAGDSGFAAYSRYRIGNYKLNTDHSSSSDNIALLVQ	GO:0019069
N	MSDNGPQNQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNNTASWFTALTQ HGKEDLKFPRGQGVPINTNSSPDDQIGYYRRATRRIRGGDGKMKDLSPRWYFYYLGTG PEAGLPYGANKDGIIWVATEGALNTPKDHIGTRNPANNAAIVLQLPQGTTLPKGFYAE GSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGDAALALLLLDRLNQLESKM SGKGQQQQGQTVTKKSAAEASKKPRQKRTATKAYNVTQAFGRRGPEQTQGNFGDQE LIRQGTDYKHWPQIAQFAPSASAFFGMSRIGMEVTPSGTWLTYTGAIKLDDKDPNFKD QVILLNKHIDAYKTFPPTEPKKDKKKKADETQALPQRQKKQQTVTLLPAADLDDFSKQ LQQSMSSADSTQA	GO:0019068, GO:0003723, GO:0019083, GO:0019079