UACM
Universidad Autónoma
de la Ciudad de México

Nada humano me es ajeno

COLEGIO DE CIENCIAS Y HUMANIDADES

MAESTRÍA EN CIENCIAS DE LA COMPLEJIDAD

Análisis de la conectividad y estructura de proteínas del virus SARS-CoV-2 mediante teoría de grafos para la identificación de sitios activos

TESIS QUE PARA OBTENER EL GRADO DE
MAESTRA EN CIENCIAS DE LA COMPLEJIDAD

PRESENTA

Jessica Pereda Méndez

Director de la Tesis

Dr. Luis Agustín Olivares Quiroz

Ciudad de México, septiembre del 2023.

UACM
Universidad Autónoma de la Ciudad de México
Nada humano me es ajeno

CA Coordinación Académica

SISTEMA BIBLIOTECARIO DE INFORMACIÓN Y DOCUMENTACIÓN

UNIVERSIDAD AUTÓNOMA DE LA CIUDAD DE MÉXICO
COORDINACIÓN ACADÉMICA

RESTRICCIONES DE USO PARA LAS TESIS DIGITALES

DERECHOS RESERVADOS©

A mí,
por tanta paciencia y fortaleza para poder alcanzar mis sueños

Agradecimientos

Al finalizar un trabajo tan arduo y lleno de dificultades como es el desarrollo de una tesis de maestría, es inevitable que te asalte un sentimiento que te lleva a concentrar la mayor parte del mérito en el aporte que has hecho. Sin embargo, el análisis objetivo te muestra inmediatamente que la magnitud de ese aporte hubiese sido imposible sin la participación de personas y de la institución que te han facilitado las cosas para que este trabajo llegue a su fin. Por ello, es para mí un verdadero placer usar este espacio para ser justa y expresarles mis agradecimientos.

Debo agradecer de manera especial y sincera al Dr. Luis A. Olivares Quiroz por aceptarme para realizar esta tesis de maestría bajo su dirección, por brindarme las incontables horas de discusión, de análisis y profundo compromiso sobre este trabajo de tesis. Gracias por su confianza y su capacidad para guiar mis ideas, pues han sido un aporte invaluable y no solamente en el desarrollo de esta tesis, sino también en mi formación como investigadora.

Agradezco a la Universidad Autónoma de la Ciudad de México que ha hecho posible la realización del trabajo presentado en esta tesis, por la beca otorgada en los cuatro semestres de maestría. A todos los miembros del Posgrado en Ciencias de la Complejidad: a los profesores por compartir sus conocimientos y al personal administrativo por el apoyo brindado.

Quiero expresar también mi más sincero agradecimiento a la Dra. Rosa Margarita Álvarez González por motivarme a continuar con mis estudios de posgrado y a mis lectores por aceptar revisar este trabajo de investigación.

Finalmente, agradezco a mi familia por su comprensión y apoyo. De manera muy especial a Misael por la motivación para que me supere día con día, por el apoyo incondicional y la ayuda de siempre. A Dios por darme las fuerzas necesarias para llegar hasta el final.

Resumen

Se presenta un trabajo sobre cómo identificar residuos del sitio activo de proteínas relacionadas con el virus del SARS-CoV-2, mediante teoría de grafos y redes complejas. La identificación de residuos funcionales en proteínas es un tema complejo, incluso cuando se dispone de estructuras atómicas detalladas, de ahí que, las interacciones de los residuos de proteínas dentro y entre los sitios funcionales son cruciales para la actividad de las proteínas.

En bioquímica, se conoce una región en particular de una enzima donde las moléculas de sustrato se unen y experimentan una reacción química, se trata del sitio activo. El sitio activo consta de residuos que forman enlaces temporales con el sustrato (sitio de unión) y residuos que catalizan una reacción de ese sustrato (sitio catalítico). El sitio activo es la parte más importante ya que cataliza directamente la reacción química.

Las estructuras de proteínas se pueden representar y modelar como redes (grafos) donde los residuos son los nodos (vértices) y sus interacciones son los enlaces (aristas). La representación de las estructuras proteicas como redes complejas facilita la búsqueda de información, que pueden relacionarse con residuos funcionalmente importantes. Aquí, nuestro objetivo fue investigar el rendimiento de la centralidad de los residuos, en la identificación de residuos funcionalmente importantes en las familias de proteínas, más en específico los residuos del sitio activo, así mismo, unir esta información con los resultados que nos proporciona el análisis del interior de la proteína, mediante el análisis de la profundidad del átomo y el área de superficie accesible al solvente.

Abstract

A paper is presented on how to identify residues of the active site of proteins related to the SARS-CoV-2 virus, using graph theory and complex networks. The identification of functional residues in proteins is a complex issue, even when detailed atomic structures are available, hence the interactions of protein residues within and between functional sites are crucial for protein activity.

In biochemistry, a particular region of an enzyme where substrate molecules bind together and undergo a chemical reaction is known, it is the active site. The active site consists of residues that form temporary bonds with the substrate (binding site) and residues that catalyze a reaction of that substrate (catalytic site). The active site is the most important part as it directly catalyzes the chemical reaction.

Protein structures can be represented and modeled as networks (graphs) where the residuals are the nodes (vertices) and their interactions are the bonds (edges). The representation of protein structures as complex networks facilitates the search for information, which can be related to functionally important residues. Here, our objective was to investigate the performance of the centrality of the residues, in the identification of functionally important residues in the families of proteins, more specifically the residues of the active site, likewise, to unite this information with the results provided by the analysis of the interior of the protein, by analyzing the depth of the atom and the surface area accessible to the solvent.

Índice general

Introducción

Los seres vivos están conformados por una gran variedad de proteínas que se pueden agrupar de acuerdo con sus características físicas, químicas, estructurales o funcionales. Las proteínas son polímeros biológicos, es decir, están formadas por unidades similares unidas de extremo a extremo para formar moléculas muy grandes [**Engel**, **2020**]. Las proteínas son componentes importantes de nuestro organismo porque son los responsables de construir, mantener y regenerar las células de nuestro cuerpo. Teniendo en cuenta su función, las proteínas se pueden clasificar según el fenómeno biológico en el cual están involucradas. Hay proteínas que se encuentran catalizando la mayoría de las reacciones metabólicas que ocurren en las células, es importante tener en cuenta que casi la totalidad de las funciones biológicas que tienen lugar en una célula son llevadas a cabo por proteínas, ya que interactúan físicamente entre ellas y la forma en que lo hacen está relacionada a las funciones y procesos celulares específicos [**Bruce et al.**, **2015**].

En particular, nos interesan las proteínas que aceleran la velocidad de una reacción química específica en la célula, se trata de una enzima que en sí, es un catalizador biológico. El papel de las enzimas es catalizar reacciones químicas particulares, es decir, acelerarlas [**Pain**, **1994**]. La catálisis es el proceso por el cual se aumenta la velocidad de una reacción química, debido a la participación del catalizador, aquellas que desactivan la catálisis son denominados inhibidores [**Engel**, **2020**]. La enzima no se destruye durante la reacción y se utiliza una y otra vez. Una célula contiene miles de diferentes tipos de moléculas de enzimas específicas para diferentes reacciones en particular.

En biología molecular, se conoce una región en particular de una enzima donde las moléculas de sustrato se unen y experimentan una reacción química, se trata del sitio activo. El sitio activo consta de residuos o aminoácidos que forman enlaces temporales con el sustrato (sitio de unión) y residuos que catalizan una reacción de ese sustrato (sitio catalítico) [**Merchant & Larios**, **2003**]. El sitio activo es la parte más importante ya que cataliza directamente la reacción química. Por lo general, consta de tres a cuatro aminoácidos, mientras que se requieren otros aminoácidos dentro de la proteína para mantener la estructura terciaria

de las enzimas [**Bruce et al.**, **2015**]. Cada sitio activo se desarrolla para unirse a un sustrato particular y catalizar una reacción particular, lo que da como resultado una alta especificidad. A veces, las enzimas necesitan unirse con algunos cofactores para cumplir su función. Un sitio activo puede catalizar una reacción repetidamente ya que los residuos no se alteran al final de la reacción (pueden cambiar durante la reacción, pero se regeneran al final) [**Engel**, **2020**].

La identificación de residuos del sitio activo es crucial en el proceso de descubrimiento de fármacos. La estructura tridimensional de la enzima se analiza para identificar los residuos del sitio activo y diseñar fármacos que puedan caber en ellos. Las enzimas proteolíticas son el objetivo principal de algunos medicamentos, como los inhibidores de proteasa, que incluyen medicamentos contra el SIDA y la hipertensión. Estos inhibidores de proteasa se unen al sitio activo de una enzima y bloquean la interacción con sustratos naturales. Un factor importante en el diseño de fármacos es la fuerza de unión entre el sitio activo y un inhibidor enzimático [**Lorenzo**, **2020**]. Si la enzima que se encuentra en las bacterias es significativamente diferente de la enzima humana, entonces se puede diseñar un inhibidor contra esa bacteria en particular sin dañar la enzima humana [**Pazos**, **2022**]. Si un tipo de enzima solo está presente en un tipo de organismo, su inhibidor puede usarse para eliminarlos específicamente. Se puede mapear el sitio activo para ayudar al diseño de nuevos fármacos, como los inhibidores de enzimas [**Stepniewska et al.**, **2020**].

La identificación de residuos del sitio activo de proteínas sigue siendo una tarea difícil. Se han propuesto diferentes enfoques basados en características estructurales, identificando residuos del sitio activo en diversas proteínas [**Aguilar & Olivares**, **2021**]. La representación de las estructuras proteicas como redes que interactúan facilita el análisis de las características topológicas, que pueden contener cierta información sobre aminoácidos funcionalmente importantes. Además, este modelo permite investigar el papel de cada aminoácido individual dentro de la compleja red de interacción [**Vendruscolo et al.**, **2002**]. La representación de las estructuras proteicas como redes complejas facilita la búsqueda de determinantes topológicos, que pueden relacionarse con residuos funcionalmente importantes. Este enfoque, ha tenido un fuerte auge en el campo de la física, la biología molecular y la complejidad desde las publicaciones de Watts y Strogatz sobre redes de mundo pequeño y, la de Barabasi y Albert sobre el estudio de redes libres de escala [**Strogatz**, **2001**].

Como consecuencia directa, los sistemas biológicos complejos pueden representarse y analizarse como redes de interacción. Por ejemplo, los ecosistemas pueden modelarse como redes de especies o una proteína puede ser modelada como una *protein residue network* (red de residuos de proteína) donde los aminoácidos pueden representarse como una red de átomos

conectados entre sí, existen numerosos tipos de redes que pueden construirse y utilizarse para analizar e interpretar procesos biológicos.

Es interesante conocer las redes complejas ya que abundan en la naturaleza, son parte de nuestra vida diaria y se presentan a diferentes niveles de organización. Es un hecho sobresaliente el que todas estas redes, tan diferentes en naturaleza y en tamaño, tengan muchas propiedades estructurales similares. Este hecho tan simple como sorprendente, hace posible que se pueda formular modelos para entender y explicar las propiedades estructurales (y en algunos casos también las propiedades dinámicas) de las redes complejas. Es por eso que se presenta un estudio sobre cómo determinar residuos del sitio activo en la estructura de proteínas, utilizando la teoría de redes complejas mediante el análisis de la conectividad y de la estructura globular de una proteína. El sitio activo de una proteína está constituido por una serie de aminoácidos que interaccionan con el sustrato, es la zona de la enzima donde se producen reacciones químicas [**Aguilar & Olivares**, **2021**]. Dado este escenario se propone que los residuos del sitio activo de las proteínas se relacionan con algunas medidas de centralidad de la teoría de redes complejas, específicamente se analiza la centralidad de cercanía (*closeness*), centralidad de vector propio (*eigenvector centrality*), centralidad de intermediación (*betweenness centrality*) y, por último, el coeficiente de agrupamiento (*clustering coefficient*) [**Newman**, **2010**]. Este hallazgo apoya la idea de que la estructura de la proteína contiene información valiosa sobre residuos funcionalmente importantes. Aquí, nuestro objetivo fue estudiar la generalidad de este resultado mediante el análisis de cuatro proteínas biológicamente diferentes relacionadas con el virus del SARS-CoV-2 que involucran diferentes sitios funcionalmente importantes.

Por otra parte, la medida en que un aminoácido interactúa con el solvente y el núcleo de la proteína, es naturalmente proporcional al área de superficie expuesta a estos entornos [**Pintar et al.**, **2003**]. También, se analiza la profundidad del átomo, la profundidad es a menudo un medidor más útil para ubicar residuos en el interior o exterior de la proteína. Esto probablemente esté relacionado con el hecho de que el interior de la proteína y el solvente circundante difieren significativamente en polaridad y densidad de empaquetamiento [**Pintar et al.**, **2003**].

Hoy en día, las técnicas utilizadas para el estudio de conformación tridimensional de macromoléculas biológicas, tales como Difracción de Rayos X (XRD), Difracción de Neutrones (ND), y Resonancia Magnética Nuclear (NMR), han evolucionado notablemente, y esto ha permitido contar con información accesible a toda la comunidad científica mundial a través de las bases de datos de proteínas (PDB) [**Berman et al.**, **2000**], con acceso libre a través de

internet. A su vez, las técnicas de simulación computacional permiten el modelado y análisis detallado de la estructura microscópica y las propiedades dinámicas de estas macromoléculas, como también de su interacción con otras sustancias [**Dayhoff**, **1965**].

Con este trasfondo es que se inicia un estudio de la conectividad y de la estructura globular de una proteína, esto para la determinación de residuos del sitio activo. La técnica empleada para el estudio de este sistema es la simulación computacional por VMD (Visual Molecular Dynamics), herramienta de acceso libre y gratuito, de probada eficiencia en este campo, y ampliamente utilizada en todo el mundo, es un programa de modelamiento molecular y visualización de estructuras. VMD fue principalmente desarrollado como una herramienta para ver y analizar los resultados de las simulaciones de dinámica molecular, pero también incluye herramientas para trabajar con datos de volumen, secuencias y objetos gráficos arbitrarios como conos, cilindros o esferas [**Fackler D. et al.**, **2021**], asimismo, simulaciones realizadas mediante el *Software Wolfram Mathematica*.

Debido a esto, las proteínas se pueden estudiar en todos sus niveles estructurales por medio de diferentes técnicas experimentales. De acuerdo con lo anterior y teniendo en cuenta la importancia de estudiar las proteínas, su estructura y algunos métodos básicos experimentales que se pueden emplear para el estudio de proteínas relacionadas con el virus del SARS-CoV-2.

Los virus pueden describirse como nanopartículas autoensambladas, formadas principalmente por material genético (ARN o ADN), proteínas (codificadas por el propio virus) y lípidos (tomados por lo regular de la membrana lipídica de la célula infectada). Por lo regular, la capa superficial (envoltura o cápsula) de un virus está constituida por una doble capa lipídica asociada a glicoproteínas que pueden proyectarse en forma de espigas desde la superficie de la partícula viral hacia el exterior. La bicapa lipídica es el componente más débil de un virus, misma que se ensambla a través de interacciones no-covalentes muy débiles. En el caso de los coronavirus, el material genético viral es ARN, las proteínas, por otro lado, le sirven al virus para reconocer a la célula blanco, asistir la replicación viral y, en esencia, como bloque estructural. Los lípidos y las proteínas forman un recubrimiento alrededor del virus para protegerlo y ayudarle a su propagación, y en la invasión celular. Cuando un virus ataca una célula, usan esas células para multiplicarse (hacer copias de sí mismos). Este proceso también se llama replicación. El proceso puede matar, dañar o cambiar las células infectadas. Estas nuevas moléculas de ARN y proteínas se autoensamblan al interior de endosomas y luego, al emerger, remueven parte de los lípidos de la membrana de la célula infectada a través de interacciones débiles para formar las nuevas partículas virales. Una vez que el virus inicia el proceso de infección, nuestro sistema inmune busca contrarrestarlo, a través de la

acción de células especializadas que atacan al invasor. El virus es capaz de confundir (por su tamaño y por los azúcares con que superficialmente modifica a sus proteínas en membrana) al sistema de defensa, evitando así ser reconocido y continuando su proceso de infección [**McGorgan & Enrique**, **2020**].

El SARS-CoV-2 es un virus relativamente grande, muestra una estructura esférica con un halo coronado, de ahí su nombre: coronavirus. El material genético (ARN) codifica para al menos 29 proteínas distintas. Una de las proteínas características del coronavirus es la proteína de espiga o proteína S (una de las cuatro proteínas estructurales, además de la proteína E de ensamblaje, la M de membrana y la N de protección al ARN), la cual se encuentra en la cápsula superficial que protege al ARN y que le permite llevar a cabo procesos de reconocimiento hacia receptores específicos en distintos tejidos (en particular a receptores del tipo de las proteínas ACE2, que se encuentran en numerosos tejidos respiratorios humanos) [**Walls et al.**, **2020**]. La proteína S forma un trímero, que forma las puntas características y que dan nombre a los coronavirus y que es además la estructura responsable del proceso de reconocimiento molecular que ayuda a la infección viral. Otras proteínas accesorias (orf3a, orf3b, orf6, orf7a, orf8) ayudan a cambiar el ambiente dentro de la célula infectada para favorecer los procesos de replicación, abren agujeros en las membranas de las células infectadas para permitir el escape de las nuevas partículas virales y desencadenan el proceso inflamatorio, uno de los síntomas más peligroso de la enfermedad [**Villa**, **2020**]. La proteína N, en particular, es importante para la estabilización del ARN viral, formando fibras espirales que envuelven y protegen al material genético. También dieciséis proteínas no estructurales (NSP 1-16) modifican las membranas celulares internas para anclar el complejo a las membranas dentro de la célula huésped, otras interfieren con los mecanismos de defensa celular o la respuesta inmunitaria del huésped [**McGorgan & Enrique**, **2020**].

El siguiente trabajo de investigación se divide en seis capítulos importantes:

En el primer capítulo se va a realizar una breve descripción de conceptos básicos de biología molecular, como son las estructuras de las proteínas, descripción e importancia de las propiedades químicas de los aminoácidos. Se va a discutir como las proteínas son creadas en el interior de las células y son necesarias para la estructura, función, regulación de los tejidos y órganos del cuerpo, también, los diferentes niveles en los que suele describirse su estructura y por último, el plegamiento de las proteínas (*protein folding*), ya que éste es el proceso por el que una proteína alcanza su estructura tridimensional. La función biológica de una proteína depende de su correcto plegamiento. Si una proteína no se pliega correctamente será no funcional y, por lo tanto, no será capaz de cumplir su función biológica. También,

se va a analizar la importancia del plegamiento adecuado de las proteínas, la existencia y descripción del sitio activo, pues, el sitio activo es la zona de la enzima donde se producen las reacciones químicas. También, el papel que desempeña el sitio activo dentro del diseño de nuevos fármacos. De acuerdo a lo anterior, es de interés trabajar en el análisis de las proteínas que están relacionadas con el virus SARS-CoV-2.

Para continuar con el segundo capítulo, se va a introducir una descripción general del marco matemático de la teoría de grafos y redes complejas, desde definiciones, propiedades y características que son de interés para aplicarlo al estudio de proteínas y a partir de esto, analizar algunas métricas globales y medidas de centralidad de redes, ya que uno de los temas fundamentales en el estudio de redes, es determinar la accesibilidad de los nodos y el papel que desempeña cada uno de ellos dentro de la red. También, se analizan algunas de las propiedades de la superficie de la proteína y de como calcularlos, la profundidad del átomo, definida como la distancia (Å) de un átomo (C_α) de la proteína, al átomo del solvente más cercano que encuentre en la superficie, proporciona una descripción simple pero precisa del interior de la proteína, así como también, el área de superficie accesible al solvente (SASA): mide el nivel de exposición de un residuo al solvente (agua) en una proteína, estas propiedades se asocian con el valor de hidrofobicidad de cada residuo de la proteína. Estos parámetros son calculados mediante softwares on line.

Se van a aplicar los conceptos básicos anteriores al tercer capítulo, ya que se lleva a cabo el análisis de las estructuras de proteínas y su modelo mediante redes. Con frecuencia, estas redes incluyen un pequeño número de nodos importantes que son concentradores de información y a través de los cuales muchos nodos pueden conectarse indirectamente. Se buscó examinar si los nodos de las redes de interacción de residuos de proteínas corresponden a residuos funcionales. Por lo tanto, se expone, describe y analizan los resultados del coeficiente de agrupamiento y medidas de centralidad de la red de aminoácidos de cada una de las proteínas relacionadas con el virus del SARS-CoV-2. Se muestran los resultados gráficamente y se realiza un análisis e interpretación de los mismos.

En el cuarto capítulo se va a realizar un análisis del interior de la proteína mediante las propiedades estudiadas en el capítulo 2, usando software online (programas independientes o servicios web). Estos generalmente investigan un solo aspecto de la estructura molecular. Se van a generar resultados que se van a interpretar y mostrar de manera grafica.

Para cerrar este trabajo, en el capítulo 5 se analiza y discuten algunos resultados globales del coeficiente de agrupamiento y medidas de centralidad de las proteínas estudiadas, por

último, en el capítulo 6 se exponen algunas conclusiones y el trabajo a futuro.

El análisis de las *protein residue network* es necesario para un estudio más extenso de los procesos biológicos. Estos análisis pueden facilitarse mediante el conocimiento de los descriptores de red que caracterizan las propiedades de conectividad, organización, robustez y estabilidad, por ejemplo, la centralidad de intermediación representa el grado de centralización de una red y los puntos de influencia, el coeficiente de agrupamiento modela las tendencias estructurales. Hasta ahora, la teoría de redes establecida o teoría de grafos, desde el campo de las matemáticas y la informática, ha facilitado revelar patrones de enriquecimiento, entendimientos sistemáticos, relaciones de alto nivel y pistas basadas en redes en redes biológicas.

Capítulo 1

Conceptos básicos de biología molecular

En este capítulo, se va a realizar una breve descripción de las proteínas, las moléculas conocidas más estructuralmente complejas y funcionalmente más sofisticadas. Posteriormente, se va discutir la importancia y descripción del sitio activo, ya que éste ocupa una región de la estructura tridimensional de una proteína con características químicas especiales, pues es la zona de la enzima en la que se une el sustrato para ser catalizado. Al mismo tiempo se va a estudiar la estructura de las proteínas relacionadas con el virus del SARS-CoV-2 para después, llevar a cabo un estudio con el método de redes complejas para la determinación de residuos del sitio activo.

1.1. Proteínas y aminoácidos

Las proteínas son moléculas formadas por aminoácidos que están unidos por un tipo de enlaces conocidos como enlaces peptídicos. El orden y la disposición de los aminoácidos dependen del código genético de cada organismo. Las proteínas están compuestas por: carbono, hidrógeno, oxígeno, nitrógeno, y la gran mayoría contiene además azufre y fósforo [**Merchant & Larios**, **2003**].

Las proteínas desempeñan un papel fundamental en el organismo. Son esenciales para el crecimiento, para la síntesis y mantenimiento de diversos tejidos del cuerpo como la hemoglobina, las vitaminas, las hormonas y las enzimas (estas últimas actúan como catalizadores biológicos haciendo que aumenten la velocidad a la que se producen las reacciones químicas del metabolismo) [**Nelson**, **2004**].

Por otra parte, las proteínas están formadas por una cadena lineal de aminoácidos de manera que la diferencia entre una proteína y otra es la cantidad y el orden de la secuencia en que aparecen estos aminoácidos. A esta secuencia se le conoce como estructura primaria, y es fundamental en la determinación de las propiedades de la proteína [**Branden & Tooze**, **1999**]. Esta secuencia es diferente para cada tipo de proteína y está indicada en el código genético del ADN. En la figura (1.1) se muestra el nombre, la abreviatura (de 3 letras y de una) y la estructura atómica de los 20 aminoácidos naturales presentes en las proteínas.

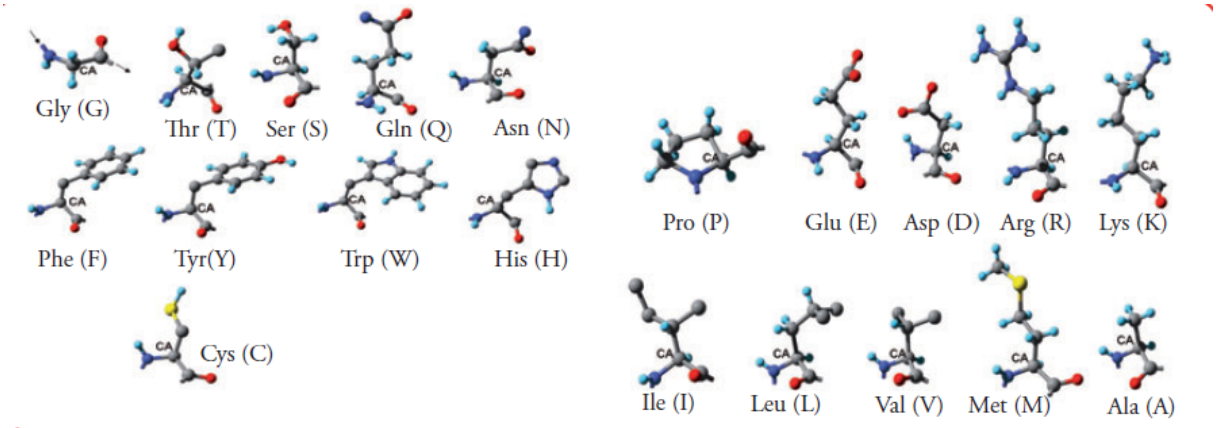


Figura 1.1: Estructura química de los 20 aminoácidos naturales presentes en las proteínas. Se representan con modelos de esferas y cilindros. Los átomos de carbono en gris, los de nitrógeno en azul, los de oxígeno en rojo y los de azufre se muestran en amarillo. Algunos de los átomos de hidrógeno se muestran en celeste. No se diferencian simples de dobles enlaces. Figura extraída del libro "Las proteínas" de Javier Santos. Colección: Las ciencias naturales y la matemática.

En el interior de las células existen unos complejos macromoleculares llamadas ribosomas, que se encargan de traducir este código y unir los aminoácidos para sintetizar las proteínas. Además, de todos los aminoácidos que existen en la naturaleza, sólo 20 distintos se utilizan para la conformación de proteínas en los seres vivos, y cada uno de estos aminoácidos está caracterizado por su cadena lateral o grupo residual R [**Fersht & Freeman**, **1999**].

La función en las proteínas depende en gran medida de la adquisición de estructuras tridimensionales específicas mediante el plegamiento en escalas de tiempo fisiológicas [**Olivares & Garcia**, **2004**]. Las células han desarrollado mecanismos controlados para mantener el plegamiento de proteínas nativas, que incluyen patrones de estructura tridimensional detallados y regiones desordenadas específicas. En este caso, la estabilidad estructural de las proteínas depende en gran medida de que los residuos hidrófobos se orienten hacia el núcleo de la proteína.

Esto ha permitido a la evolución desarrollar un sistema de alerta conservado, en el que la exposición de regiones hidrofóbicas proteicas se reconoce como un patrón molecular asociado a la presencia de citotoxicidad [**Bruce et al.**, **2015**].

En pocas palabras, las proteínas se pliegan espontáneamente y, el estado plegado corresponde al estado de mínima energía, es decir el más estable [**Olivares**, **2017**]. Así se podría, en principio, producir una proteína y ésta debería llegar a su forma correctamente plegada (nativa) y lo que resulta realmente importante, es que lo hace en ausencia de otros factores biológicos. Las condiciones adecuadas del solvente deberían ser suficientes para estabilizar dicha conformación [**Pain**, **1994**].

Aminoácidos polares e hidrofóbicos

Existen 20 aminoácidos diferentes en proteínas que están codificados directamente en el ADN de un organismo, cada uno con diferentes propiedades químicas [**Branden & Tooze**, **1999**]. La secuencia repetitiva de átomos a lo largo del núcleo de la cadena polipeptídica se denomina columna vertebral polipeptídica (Backbone). Unido a esta cadena repetitiva están aquellas porciones de los aminoácidos que no participan en la formación de un enlace peptídico y que le dan a cada aminoácido sus propiedades únicas: las 20 cadenas laterales de aminoácidos diferentes. Algunas de estas cadenas laterales son no polares e hidrófobas, otras tienen carga negativa o positiva, algunas forman fácilmente enlaces covalentes, etc.

En general, todos los aminoácidos están constituidos por un átomo central de carbono C_α al cual se unen un grupo funcional amino (NH_2), uno carboxilo ($COOH$), un hidrógeno (H) y un grupo R o lateral. Las diferencias entre los aminoácidos se debe a la estructura de sus grupos laterales o R (residuo o resto de la molécula). Con base en las propiedades físicas de los grupos residuales respecto del solvente en el que se encuentran inmersos ((H_2O)), es posible definir una clasificación estándar de los 20 aminoácidos. Esta clasificación los agrupa en dos categorías principales: la primera consiste en aquellos aminoácidos con propiedades hidrofóbicas, es decir, aquellos que presentan interacciones electrostáticas repulsivas hacia las moléculas de H_2O y la segunda agrupa a los aminoácidos polares, que son aquellos cuya interacción electrostática con las moléculas del solvente es atractiva [**Olivares & Garcia**, **2004**]. En la tabla (1.1) se muestra la lista de los 20 aminoácidos conocidos, en relación a las categorías mencionadas.

Es útil clasificar los aminoácidos en función de sus grupos R, ya que son estas cadenas laterales las que le dan a cada aminoácido sus propiedades características. Así, se puede espe-

Aminoácidos polares	Aminoácidos hidrofóbicos
Ácido aspártico (Asp, D)	Alanina (Ala, A)
Ácido glutámico (Glu, E)	Glicina (Gly, G)
Arginina (Arg, R)	Valina (Val, V)
Lisina (Lys, K)	Leucina (Leu, L)
Histidina (His, H)	Isoleucina (Ile, I)
Asparagina (Asn, N)	Prolina (Pro, P)
Glutamina (Gln, Q)	Fenilalanina (Phe, F)
Serina (Ser, S)	Metionina (Met, M)
Treonina (Thr, T)	Triptófano (Trp, W)
Tirosina (Tyr, Y)	Cisteína (Cys, C)

Cuadro 1.1: Clasificación de aminoácidos de acuerdo a sus cadena laterales: polares y no polares (hidrofóbicos) [**Bruce et al.**, **2015**].

rar que los aminoácidos con grupos laterales (químicamente) similares funcionen de manera similar, por ejemplo, durante el plegamiento de proteínas [**Olivares & Garcia**, **2004**].

1.2. Importancia del sitio activo

Las enzimas son catalizadores. Generalmente son proteínas, las enzimas disminuyen la energía de activación de una reacción, es decir, la cantidad de energía necesaria para que ocurra una reacción. Logran esto al unirse a un sustrato y sostenerlo de tal manera que permite que la reacción ocurra más eficientemente [**Engel**, **2020**]. La parte de la enzima donde se une el sustrato se llama el sitio activo (ya que ahí es donde sucede la "función" catalítica).

Determinar los aminoácidos que constituyen el sitio activo de una proteína es un problema de bastante desarrollo teórico y experimental. En la Web se pueden encontrar diversas páginas que proporcionan información general sobre proteínas, algunas incluyen información acerca de la ubicación en la cadena lineal de los residuos que forman el sitio activo de la proteína. En este caso se recurrió al sitio web más conocido, el *Protein Data Bank* [**Berman et al.**, **2000**] (http://www.rcsb.org/) es una base de datos de la estructura tridimensional de las proteínas y ácidos nucleicos. Estos datos son obtenidos mediante cristalografía de rayos X o resonancia magnética nuclear, son enviados por biólogos y bioquímicos de todo el mundo, están bajo el dominio público y pueden ser usados libremente. El formato *Protein Data Bank* (PDB) es un estándar para archivos que contienen coordenadas atómicas, es leído y escrito por muchos programas [**Berman et al.**, **2000**]. El archivo PDB completo proporciona una gran cantidad de información, incluidos autores, referencias bibliográficas y el método de determinación de la estructura. El formato PDB consta de líneas de información en un

archivo de texto, en donde cada línea de información del archivo se denomina registro. Un archivo PDB generalmente contiene varios tipos diferentes de registros, organizados en un orden específico para describir una estructura.

Existen otras bases de datos de proteínas, como lo es el *Universal Protein Resource* (UniProt) [**Lombardot et al.**, **2021**] (www.uniprot.org) es un repositorio central de datos sobre proteínas. La mayoría de entradas proviene de proyectos de secuenciación del genoma, y se encuentran publicadas en revistas científicas. Así como también PROSITE (https://prosite.expasy.org/) [**Bougueleret et al.**, **2022**] es una base de datos de familias y dominios de proteínas que consiste en entradas que describen dominios, familias y sitios funcionales así como patrones de aminoácidos. La base de datos ProRule se basa en las descripciones de dominio de PROSITE [**Sigrist et al.**, **2005**]. Esta proporciona información adicional acerca de funcionalidades o de aminoácidos estructuralmente críticos.

El sitio activo es una región de la estructura tridimensional de una proteína con características químicas especiales, pues es la zona de la enzima en la que se une el sustrato para ser catalizado. Las moléculas del sustrato se unen al sitio activo, donde tiene lugar la catálisis. Dentro del sitio activo hay ciertos aminoácidos que intervienen en la unión del sustrato a la enzima y se denominan residuos de unión, mientras que los que participan de forma activa en la transformación química del sustrato se conocen como residuos catalíticos.

El sustrato se adapta al sitio activo de una enzima como una llave a una cerradura [**Nelson**, **2004**]. Para que una enzima catalice al sustrato esta debe acomodarse de la forma en la cual el sitio de la enzima está conformado [**Engel**, **2020**].

1.3. Diseño de fármacos moleculares

La capacidad de predecir con precisión las estructuras de proteínas a partir de su secuencia de aminoácidos sería fundamental para las ciencias de la vida y la medicina, ya que aceleraría enormemente los esfuerzos para comprender los componentes básicos de las células y permitiría un descubrimiento de fármacos más rápido y avanzado. La farmacología molecular estudia a las características bioquímicas y biofísicas de las interacciones entre los fármacos y los blancos de las células. De algún modo, es la biología molecular aplicada a las preguntas farmacológicas y toxicológicas [**Delgado et al.**, **2002**].

La estructura de la molécula de un fármaco que puede interactuar específicamente con

las biomoléculas se puede modelar usando herramientas computacionales. Estas herramientas pueden permitir que una molécula de fármacos (drugs) sea construida dentro de la biomolécula usando lo que se conoce de su estructura y de la naturaleza de su sitio activo. La construcción de una molécula de fármaco se puede hacer desde dentro hacia fuera o desde afuera hacia dentro, dependiendo de si eligen primero el núcleo o a los grupos R. Sin embargo, muchos de estos enfoques están plagados de los problemas prácticos de la síntesis química [**Lorenzo**, **2020**].

Las enzimas son los catalizadores de los seres vivos. Sin ellos las reacciones químicas en las células serían demasiado lentas para ser útiles y la vida no se hubiera dado o sería totalmente distinta a la actual. A escala molecular, son diversos los mecanismos por los que un catalizador es capaz de disminuir la energía de activación de una reacción química. Entre ellos, cabe destacar su capacidad para proporcionar un entorno adecuado (el sitio activo) para que la reacción química tenga lugar [**Engel**, **2020**] . Para ello, participan en el acercamiento y orientación de los reactivos de modo que se pueda alcanzar más fácilmente (con menor gasto energético o energía de activación) el estado de transición del proceso. En ocasiones contribuyen a debilitar ciertos enlaces de los reactivos o bien participan en el mecanismo del proceso por formación de nuevos enlaces transitorios entre el sustrato y la enzima.

Los procesos enzimáticos son reversibles, es decir, las enzimas pueden catalizar tanto la reacción directa como la inversa, alcanzándose un equilibrio cuya posición será independiente del camino seguido [**Dayhoff**, **1965**]. La naturaleza de los mecanismos enzimáticos es inherente a las características estructurales de las propias enzimas. Así, por tratarse de proteínas, su estructura terciaria determina la naturaleza tridimensional del sitio activo y explica la especificidad respecto al sustrato que se observa en la mayoría de ellos. Por otra parte, en el sitio activo se encontrarán diversos aminoácidos cuyos grupos residuales serán determinantes para el establecimiento de interacciones con los reactivos, bien de tipo enlazante (consistentes en el anclaje y la orientación adecuada del sustrato) o bien de tipo catalítico, por las que ciertos restos de aminoácidos pueden participar en el mecanismo de la reacción. Los fármacos interactúan con las enzimas principalmente inhibiendo el proceso enzimático compitiendo con el ligando endógeno por el sitio activo. El diseño de inhibidores enzimáticos representa una de las estrategias de diseño de fármacos más actuales [**Lorenzo**, **2020**]. Una prueba de ello es que prácticamente la tercera parte de los cincuenta fármacos más vendidos en el mundo en la actualidad son inhibidores enzimáticos.

Desde un punto de vista farmacológico, el conocimiento de los receptores es importante para el diseño de nuevos fármacos más eficaces y selectivos que puedan modular las respues-

tas derivadas de la acción del ligando o mensajero químico natural. La aparición de una respuesta farmacológica asociada a la interacción de un fármaco sobre un receptor depende de la capacidad de aquel para inducir una serie de procesos bioquímicos resultantes de dicha interacción. Una consecuencia muy importante de la teoría de la adaptación inducida es que la conformación del fármaco que acaba unido al receptor puede ser distinta de la conformación más abundante de dicho fármaco en disolución.

El objetivo del diseño racional de fármacos es descubrir nuevos fármacos más rápido y a menor precio. Gran parte del esfuerzo se dedica a mejorar las metodologías de acoplamiento y puntuación. Sin embargo, la mayoría de las técnicas asumen que se conoce la ubicación exacta de los residuos del sitio de unión, también conocidos como bolsas o cavidades de unión. Tales bolsas pueden ubicarse tanto en una superficie de una sola proteína o en interfaces de interacción proteína-proteína (PPI) (y usarse para interrumpir la interacción). Esta tarea es muy desafiante y se carece de un método que prediga los residuos del sitio de unión con alta precisión: la mayoría de los métodos pueden detectar solo el 30 % y el 40 % de los ligandos de las proteínas [**Stepniewska et al.**, **2020**].

Las proteasas son enzimas cuya actividad es responsable de la hidrólisis de péptidos mediante diferentes mecanismos moleculares, criterio por el cual se clasifican en distintas sub-familias respectivamente. En la actualidad, un gran número de estas moléculas han sido identificadas, cuyas funciones biológicas son muy diversas, interviniendo en diferentes procesos muy dispares. En el caso concreto de agentes infecciosos como los virus, las proteasas permiten la maduración de proteínas víricas a partir de los productos de la traducción de su ARN mensajero. Su actividad proteolítica puede ser regulada y/o bloqueada en casos de actividad exacerbada mediante inhibidores de proteasas (IPs), los cuales constituyen importantes herramientas de la naturaleza, pues, se presentan en múltiples formas (por ejemplo tejidos de animales, plantas y microorganismos) y en diversos procesos fisiológicos normales al igual que en aquellos de carácter patológico.

El diseño de medicamentos a partir de estas estructuras considera las siguientes clasificaciones: inhibidores que imitan sustratos peptídicos naturales, e inhibidores basados en moléculas pequeñas obtenidos de modificaciones químicas de IPs existentes que posteriormente son evaluados por técnicas de cribado virtual considerando su estabilidad en términos de su energía mínima libre posible. Desafortunadamente esto lleva a muchos investigadores a dedicar innumerables horas a la identificación de compuestos biológicamente activos que después resultan inactivos [**Macchiagodena et al.**, **2020**] .

La proteasa 3CLpro, también conocida como Mpro, es la enzima principal que se requiere para la maduración proteolítica del coronavirus al participar en la escisión de las poliproteínas PP1a y PP1ab. Otras proteínas funcionales como la ARN polimerasa, la endoribonucleasa y la exoribonucleasa se generan mediante la escisión de dichos péptidos. Por lo tanto, las investigaciones referenciadas demuestran que utilizar a la enzima 3CLpro como blanco terapéutico inhibirá la maduración viral y mejorará la respuesta inmune innata del huésped contra COVID-19. Estas hipótesis no habrían sido posibles sin la obtención de la primera estructura cristalina de la proteasa 3CLpro de la cual inicialmente se reveló que cuenta con tres dominios denominados I, II y III. Los dominios I (residuos 8-101) y II (residuos 102-184) presentan una estructura en forma de láminas β, mientras que el dominio III (residuos 201-306) adopta la forma de hélices α. Este último está involucrado en la dimerización de esta misma, mecanismo necesario para su actividad catalítica [**Macchiagodena et al.**, **2020**]. Adicionalmente, se evidenció que el sitio activo de la proteasa en cuestión son clasificados como S1, S1, S2 y S4, respectivamente, así como la similitud en cuanto a su alta conservación con otros coronavirus como SARS-CoV y MERS-CoV.

1.4. Estructura general y mecanismos de acoplamiento del SARS-CoV-2

En cuanto al análisis genómico y proteómico, se ha evidenciado que SARS-CoV-2 codifica aproximadamente 25 proteínas necesarias para llevar a cabo la infección en humanos y consecuentemente, su replicación [**Gorvalenya**, **2020**]. Asimismo, el genoma del nuevo coronavirus es similar a otros de la misma familia pues está conformado por al menos diez marcos de lectura abiertos (ORFs, por sus siglas en inglés) de los cuales algunos son necesarios para la expresión de poliproteínas involucradas en la formación del complejo replicasa-transcriptasa (RTC, por sus siglas en inglés) cuyo objetivo es favorecer el ensamblaje de proteínas virales en el organismo infectado. No obstante, también codifican proteínas que a pesar de ser consideradas como no esenciales para la replicación, se ha demostrado que confieren cierta ventaja en cuanto a la sobrevivencia del virus en sistemas in vivo [**Drosten et al.**, **2003**].

Por otra parte, investigaciones pasadas han demostrado que la mayoría de los CoV requieren de al menos cuatro proteínas estructurales para producir una partícula viral estructuralmente completa, lo que sugiere que algunos CoV pueden codificar proteínas adicionales con funciones compensatorias superpuestas en el ciclo de replicación del virus. De esta manera, las proteínas estructurales comúnmente reportadas por su importancia estructural son las siguientes: S (espiga), M (matriz), E (envoltura) y N (nucleocápside). Cada una de ellas

Proteína estructural	**Características y propiedades**
Espiga (S)	Estructura con peso molecular aproximado de 150 kDa. Su ensamblaje en homotrímeros en la superficie del virus asemeja a un pico y pertenece a la clasificación de proteínas de fusión. Cuenta con dos dominios funcionales, S1 (facilita la unión al receptor) y S2 (funge como soporte estructural), respectivamente.
Matriz (M)	Estructura con peso molecular aproximado de 25-30 kDa. Cuenta con tres dominios transmembranales que mantienen la curvatura morfológica característica de los coronavirus y se encuentra unida a la nucleocápside.
Envoltura(E)	Estructura con peso molecular aproximado de 8-12 kDa. Se ha identificado por ser una proteína transmembrana que asemeja la actividad de un canal iónico posiblemente necesario para la patogénesis de SARS-CoV-2. Además, es fundamental para el ensamblaje del virión y liberación del material genético.
Nucleocápside (N)	Participa en el empaquetamiento del genoma viral al interactuar directamente con la proteína M y NSP3, el cual es uno de los componentes que favorecen su unión al complejo replicasa-transcriptasa (RTC).

Cuadro 1.2: Características y propiedades distintivas de las proteínas estructurales de SARS-CoV-2: espiga (S), M (matriz), E (envoltura) y N (nucleocápside) [**Walls et al.**, **2020**].

cuenta con características y propiedades distintivas como se observa en la tabla (1.2) que son reconocidas durante el ciclo de infección de SARS-CoV-2 [**Zaki et al.**, **2012**]. Inicialmente, es crucial considerar que la infección COVID-19 está condicionada por la expresión de la enzima convertidora de angiotensina 2 (ACE2, por sus siglas en inglés) misma que es altamente expresada en la superficie extracelular de distintos órganos como los pulmones, cerebro, corazón, riñones e intestino, lo que a su vez explicaría la fisiopatología sistémica de otros coronavirus [**Wu et al.**, **2020**]. Por consiguiente, la subunidad S1 de la proteína viral (S) es el segmento responsable de la unión al receptor ACE2 garantizando la entrada celular de SARS-CoV-2 en el huésped [**Walls et al.**, **2020**]. De manera simultánea, la subunidad S2 es activada por la TMPRSS2 (proteasa transmembrana de serina 2 asociada a la superficie del huésped). Juntas, estas acciones dan como resultado la fusión completa de la membrana viral y la liberación del genoma de ARN en el citoplasma de la célula huésped. Posteriormente, la maquinaria celular de traducción del huésped es secuestrada para la síntesis de las poliproteínas y proteasas virales esenciales. De las cuales, las poliproteínas (PP1a y PP1ab) se dividen en 16 proteínas efectoras no estructurales mediante las proteasas 3CLpro y PLpro, lo que les permite formar el complejo de replicación junto con la ARN polimerasa dependiente de ARN, el cual es necesario para cumplir dos funciones: replicar el genoma completo de ARN y generar las plantillas individuales de ARN subgenómico requeridas para la traducción de las proteínas estructurales y accesorias virales [**Towler et al.**, **2004**]. Las nuevas estructuras proteicas recién sintetizadas son transferidas desde el retículo endoplasmático al aparato de Golgi, donde se ensamblan los nuevos viriones [**Towler et al.**, **2004**]. Finalmente, los viriones maduros de SARS-CoV-2 se exocitan y se liberan de la célula huésped al ambiente para repetir el ciclo de infección.

Proteínas relacionadas con el SARS-CoV-2

Los coronavirus poseen los genomas más grandes entre todos los virus de ARN. El NCBI (Centro Nacional de Información Biotecnológica) proporciona información donde se resumen algunas características clave del virus SARS-CoV-2, y de la secuencia del genoma del virus. En particular, enumera algunos genes importantes que se han identificado en el genoma, junto con los dominios conservados que se han identificado en las proteínas transcritas [**McGorgan & Enrique**, **2020**].

El genoma contiene varios genes esenciales que codifican las proteínas virales necesarias para la replicación, transcripción y ensamblaje del virus infeccioso. Los genes esenciales comprenden los marcos de lectura abiertos 1a y 1b (ORF1ab) que se traducen para producir 16 proteínas no estructurales maduras (NSP1-NSP16, numeradas según su orden desde el

extremo N hasta el extremo C de las poliproteínas ORF 1) que participan en la replicación y transcripción del ARN viral.

El SARS-CoV-2 contiene genes que codifican cuatro proteínas estructurales que están involucradas en el ensamblaje del virus infeccioso: proteínas *(S)pike*, *(E)nvelope*, *(M)embrane* y *(N)ucleocapsid)* [**Villa**, **2020**]. La proteína N contiene el genoma de ARN y las proteínas S, E y M juntas crean la envoltura viral. La proteína espiga es responsable de permitir que el virus se adhiera y se fusione con la membrana de una célula huésped. Intercalados entre estos genes en el genoma del coronavirus hay varios otros genes llamados genes accesorios o específicos de grupo y sus productos genéticos se denominan proteínas accesorias que son prescindibles para el crecimiento del virus in vitro, pero, pueden desempeñar un papel importante en la modulación del huésped, dar respuesta a la infección por el virus y, por lo tanto, contribuir a la patogénesis.

La infección comienza cuando el virus tiene acceso al organismo a través de las mucosas del tracto respiratorio, donde el ciclo de vida del virus tiene lugar al generarse la unión de la proteína viral S a ACE2. Este reconocimiento provoca un cambio conformacional de S que fomenta la fusión de la membrana de SARS-CoV-2 siguiendo una ruta endosómica.Posteriormente, se libera el ARN+ del virus, el cual se traduce para generar las replicasas pp1a y 1 ab, que son proteolíticamente cortadas por proteasas virales en pequeños productos. Una polimerasa produce ARNm subgenómicos por transcripción discontinua (varios ORFs), traducidos a proteínas virales (S, E, M, N y accesorias). Las proteínas y el ARN (positivo y asociado a nucleoproteínas) son ensamblados en el retículo endoplásmico y el aparato de Golgi de la célula hospedera para generar viriones, para luego ser transportado a través de vesículas al exterior de la célula por exocitosis, y estos viriones infectarán nuevas células [**Andersen**, **2020**].

Es posible que las definiciones funcionales de las variantes del SARS-CoV-2 que se presentan aquí se modifiquen periódicamente para adaptarse a la evolución continua de este virus y los nuevos conocimientos al respecto. Siempre que sea necesario, cualquier variante que no cumpla todos los criterios que se mencionan en estas definiciones se podrá designar como variante preocupante, de interés o bajo vigilancia, mientras que aquellas que entrañen riesgos menores que otras variantes circulantes se podrán reclasificar, tras solicitar asesoramiento al Grupo Consultivo Técnico de la OMS sobre Evolución de los Virus, (este grupo se denominaba anteriormente Grupo de Trabajo sobre la Evolución de los Virus) [**Walls et al.**, **2020**]. En las actualizaciones epidemiológicas semanales de la Organización Mundial de la Salud (OMS), se proporciona regularmente información actualizada sobre las clasificaciones

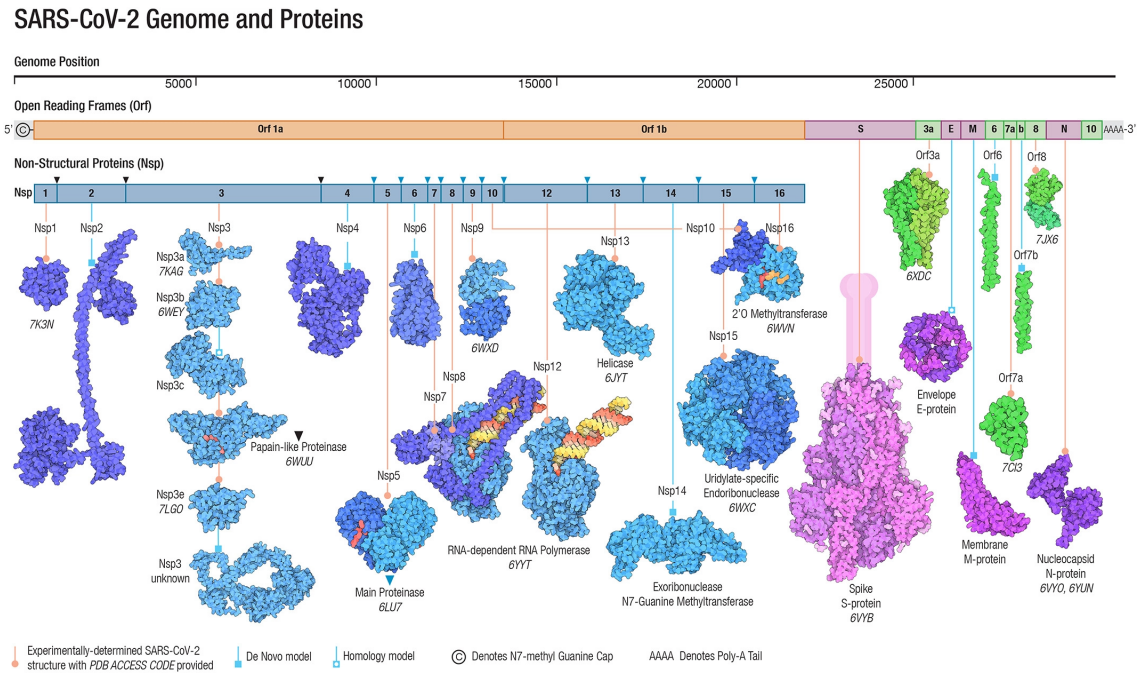


Figura 1.2: Arquitectura del genoma y proteoma del SARS-CoV-2. Asimismo se incluye los NSPs derivados de proteínas o pp1a y pp1ab (tonos de azul), proteínas estructurales de virión (rosa/púrpura) y proteínas de marco de lectura abierto (Orfs, tonos de verde). Los sitios de escisión de poliproteínas están indicados por triángulos invertidos para la proteinasa similar a la papaína (PLPro, negro) y la proteasa principal (NSP5). El complejo sustrato-producto de ARN de doble cadena de la ARN polimerasa dependiente de ARN (mostrado como NSP7-NSP8-NSP12 heterotetramer y por separado con sólo NSP12) está codificado por colores (amarillo: hebra de producto, rojo: hebra de plantilla). Las porciones transmembrana de la proteína Spike S se muestran en forma de dibujos animados (rosa). Se indica la fuente de los modelos estructurales utilizados para los análisis de todas las proteínas de estudio (determinado experimentalmente, modelo de homología computacional o modelo predicho de novo). Figura extraída de *Protein Data Bank* [**Berman et al.**, **2000**].

del SARS-CoV-2, la distribución geográfica de las variantes preocupantes y los resúmenes de sus características fenotípicas (transmisibilidad, gravedad de la enfermedad, riesgo de reinfección e impactos en el diagnóstico y la eficacia de la vacuna) basada en los estudios publicados.

Los virus como el SARS-CoV-2 evolucionan constantemente a medida que se producen cambios en el código genético (provocados por las mutaciones genéticas o la recombinación viral) durante la replicación del genoma. Un linaje es un grupo de variantes de virus estrechamente relacionados desde el punto de vista genético derivados de un ancestro en común. Una variante tiene una o más mutaciones en su material genético que la diferencian de las otras variantes del virus del SARS-CoV-2. Un recombinante es una variante creada por la

combinación de material genético de dos variantes diferentes. Tal como se preveía, se han documentado múltiples variantes del SARS-CoV-2 en los Estados Unidos y a nivel mundial durante esta pandemia [**Andersen**, **2020**]. Para fundamentar las investigaciones de brotes locales y comprender las tendencias nacionales, los científicos comparan las diferencias genéticas entre los virus para identificar las variantes (incluidos los recombinantes) y cuán estrecha es su relación entre sí.

Los linajes genéticos del SARS-CoV-2 se monitorean de manera rutinaria a través de investigaciones epidemiológicas, vigilancia de la secuencia genética de los virus y estudios de laboratorio. El 30 de noviembre del 2021, el grupo interagencias del SARS-CoV-2 (SIG) del gobierno de los EE. UU. clasificó a la variante ómicron como variante de preocupación (VOC). Esta clasificación se basó en lo siguiente: La detección de casos atribuidos a la variante ómicron en varios países, incluso entre quienes no habían viajado. Transmisión y reemplazo de la variante delta en Sudáfrica. La cantidad y ubicaciones de las sustituciones en la proteína S. Datos disponibles de otras variantes con menos sustituciones en la proteína S que indican una reducción en la neutralización por sueros de personas vacunadas o convalecientes [**Wu et al.**, **2020**]. Datos disponibles de otras variantes con menos sustituciones en la proteína S que indican una menor susceptibilidad a ciertos tratamientos de anticuerpos monoclonales.

El 14 de abril del 2022, el SIG del gobierno de los EE. UU. degradó a la variante delta de variante de preocupación a variante bajo monitoreo. Esta nueva clasificación se basó en lo siguiente: Reducción significativa y sostenida en sus proporciones nacionales y regionales en el tiempo. Evidencia que sugiere que la variante delta actualmente no representa un riesgo significativo para la salud pública en los Estados Unidos.

El esquema de clasificación de variantes del SIG define cuatro clases de variantes del SARS-CoV-2: Variante bajo monitoreo (VBM) que incluye Alpha (linajes B.1.1.7 y Q), Beta (linajes B.1.35 y descendientes), Gamma (linajes P.1 y descendientes), Delta (linajes B.1.617.2 y AY), Epsilon (B.1.43 y B.1.43), Eta (B.1.52), Iota (B.1.53), Kappa (B.1.617.1), Mu (B.1.621, B.1.621.1) y Zeta (P.2). Variante de interés (VOI, por sus siglas en inglés), Variante de preocupación (VOC, por sus siglas en inglés) que inluye a Ómicron (linajes B.1.1.529, BA.1, BA.1.1, BA.2, BA.3, BA.4 y BA.5) y Variante con grandes consecuencias (VOHC, por sus siglas en inglés) que hasta la fecha [**Andersen**, **2020**], no se ha identificado ninguna variante de gran consecuencia, es por eso que se siguen monitoreando todas las variantes que circulan.

En este caso se va a trabajar con cuatro de las dieciséis proteínas no estructurales

(NSP) que son codificadas por el genoma del SARS-CoV-2, los datos para este estudio como la ubicación de sus residuos del sitio activo se realiza de diversos bancos de datos como PDB [**Berman et al.**, **2000**], UNIPROT [**Lombardot et al.**, **2021**] Y SWISS MODEL [**Bienert et al.**, **2017**].

Los coronavirus son una extensa familia de virus que pueden causar enfermedades tanto en animales como en humanos. En los humanos, se sabe que varios coronavirus causan infecciones respiratorias que pueden ir desde el resfriado común hasta enfermedades más graves como el síndrome respiratorio de Oriente Medio (MERS) y el síndrome respiratorio agudo severo (SRAS). La enfermedad por coronavirus (COVID-19) es una enfermedad infecciosa causada por el virus SARS-CoV-2. Este nuevo virus como la enfermedad que provoca eran desconocidos antes de que estallara el brote en Wuhan (China) en diciembre de 2019, actualmente COVID-19 es una pandemia que afecta a muchos países de todo el mundo.

1.5. Métodos de determinación de residuos del sitio activo de proteínas relacionadas con el SARS-CoV-2

Existen métodos para la determinación de residuos de sitio activo de una proteína, en los laboratorios utilizan técnicas modernas de resonancia magnética nuclear y difracción de rayos X para conocer la estructura tridimensional de proteínas, así como de complejos proteína-proteína y proteína ligando. Por otra parte, están los algoritmos genéticos para ubicar residuos del sitio activo. Los algoritmos genéticos están basados en la teoría de la evolución, modifica al azar un conjunto de soluciones y se rescata a las mejores. Para el caso de proteínas, un individuo es un conjunto de residuos que incluye al sitio activo y sus vecinos cercanos. Estos individuos se comparan geométricamente con sitios activos conocidos, y al final el algoritmo señala los mejores candidatos.

Los sistemas complejos pueden ser analizados como redes de interacciones entre los componentes del sistema. El análisis de la red puede caracterizar todo el sistema y sus componentes individuales. Las estructuras de proteínas se perciben típicamente como elementos de estructura local que forman diversas topologías y plegamientos. Sin embargo, las estructuras de proteínas también se pueden representar como redes (grafos) donde los residuos son los nodos y sus interacciones son los enlaces. Este enfoque se utilizó para estudiar varios aspectos de las proteínas, incluida la profundidad del átomo y el área de superficie accesible al solvente para ubicar residuos funcionales.

La identificación de residuos funcionales (por ejemplo, residuos del sitio activo) en proteínas es un tema complejo, incluso cuando se dispone de estructuras atómicas detalladas. Sólo una fracción muy pequeña de todas las proteínas conocidas ha sido bioquímicamente bien estudiada [**Aguilar & Olivares**, **2021**]. La conservación evolutiva junto con la información de la estructura tiene éxito en la predicción de algunos residuos del sitio catalítico y de unión de ligandos en varias proteínas. La combinación de diferentes propiedades de residuos estructurales y evolutivos mejora la identificación de residuos del sitio activo [**Kahraman et al.**, **2008**], cada método tiene sus propias ventajas y limitaciones, pero, incluso la integración de diferentes enfoques no puede identificar todos los sitios buscados. Todavía se necesitan nuevas formas de caracterizar y predecir los residuos de proteínas clave. En nuestro caso, es de interés estudiar las proteínas relacionadas con el virus del SARS-CoV-2. Estudiar las interacciones de los residuos de estas proteínas dentro y entre los sitios funcionales, ya que son cruciales para la actividad de las proteínas.

Capítulo 2

Teoría de grafos, redes complejas y representación de la estructura proteica

En este capítulo se introducen y discuten las bases teóricas necesarias para el entendimiento de las redes complejas que serán utilizadas a lo largo de toda la tesis. En primer lugar, se estudia la teoría de grafos para caracterizar estructuralmente las redes bajo estudio. Más adelante, se va a hablar de las principales métricas asociadas al análisis de la teoría de grafos y las medidas de centralidad aplicadas a redes.

Las redes complejas son redes modeladas como grafos, es decir, un conjunto de vértices o nodos unidos por unos enlaces o aristas, que permiten representar relaciones entre los elementos de un conjunto, ya sean fenómenos físicos, biológicos o sociales, con el fin de llegar a modelos predictivos de estos fenómenos. Las redes complejas, a diferencia de los grafos, describen las interacciones presentes en un sistema complejo concreto, mientras que un grafo en general no tiene por qué estar asociado a un sistema complejo. Estas redes poseen ciertas propiedades estadísticas y topológicas no triviales que no se encuentran en redes simples, como por ejemplo su estructura jerárquica o comunitaria. Este panorama nos permitirá estudiar las *protein residue network* y así poder realizar los cálculos correspondientes de métricas y medidas de centralidad e interpretar los resultados obtenidos, lo cual es el tema central de este trabajo de investigación. Por otra parte, se aborda el análisis de la estructura tridimensional y el modelo con el que se representa la estructura de una proteína como un grafo.

2.1. Conceptos básicos

Es de importancia hablar de la teoría de grafos, debido a que, se puede representar y analizar estas redes de manera formal. La teoría de grafos es válida para cualquier número de nodos y enlaces en múltiples circunstancias [**Costa et al.**, **2007**], al definir una red se hace de igual manera que un grafo.

Un grafo $G = (\nu, \epsilon)$ está constituido por dos conjuntos $\nu = \{v_1, ..., v_n\}$ cuyos elementos son los nodos o vértices del grafo y $\epsilon = \{e_1, ..., e_n\}$ cuyos elementos son los enlaces o aristas. Los enlaces vienen definidas por los órdenes de los nodos que unen, es decir, la unión entre los nodos v_i y v_j se denota por $e_k = (i, j) = (v_i, v_j) = e_{ij}$ [**Reka & Barabasi**, **2002**]. En cuanto a vocabulario, si hay un enlace o arista entre dos nodos, estos dos nodos se llaman vecinos o adyacentes. Cuando dos enlaces unen dos mismos nodos se refiere a enlaces paralelos. Y cuando existen enlaces que tienen en común un nodo o vértice, son enlaces adyacentes. También, se destaca que si a cada enlace se le asigna un peso o valor numérico que mide la intensidad de la unión, nos referimos a grafos pesados. En caso contrario es un grafo no pesado [**Strogatz**, **2001**] .

En particular, si se tiene un grafo completo en donde todos los vértices están unidos entre sí, y está formado por N vértices, como resultado, el número total de aristas será:

$$E = \binom{N}{2} = \frac{N(N-1)}{2} \tag{2.1}$$

Además, la densidad ρ de una red, se define como el número de aristas que forman la red, entre el número máximo posible de aristas que pueden formarse, esto es:

$$\rho = \frac{2E}{N(N-1)} \tag{2.2}$$

Se puede decir que la representación más habitual de una red compleja es mediante su matriz de adyacencia A. Esta matriz A es una matriz cuadrada de dimensión $N \times N$, donde las filas y las columnas hacen referencia a los nodos para almacenar en cada casilla la longitud entre cada par de nodos del grafo o la información si existe cierta conexión.

La relación entre un grafo y la matriz de adyacencia cuyos términos A_{ij} indican la correspondencia entre los vértices i y j, y se define como:

$$A_{ij} = \begin{cases} 1, \ si \text{ i , j } son\ vértices\ conectados \\ 0, \ si \text{ i , j } son\ vértices\ no\ conectados \end{cases} \tag{2.3}$$

Por el contrario, para grafos dirigidos, $A_{ij} = 1$ indica que hay una conexión que va de i a j, pero no necesariamente $A_{ij} = A_{ji}$ [**Aguilar**, **2019**]. Si el grafo es no dirigido, se cumple la igualdad y la matriz de adyacencia será siempre simétrica [**Trudeau**, **1994**], como se muestra en la figura (2.1).

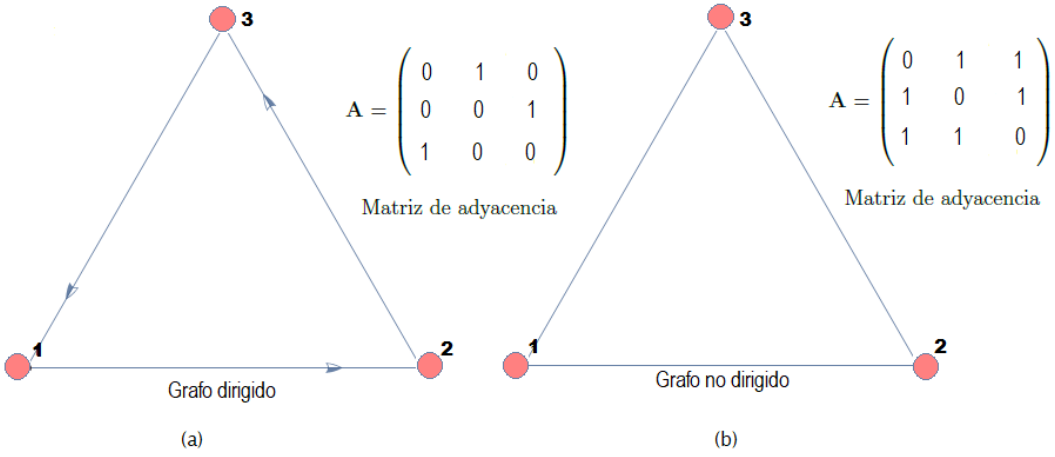


Figura 2.1: Matriz de adyacencia de grafos dirigidos y no dirigidos. En (a) se muestra el grafo y la matriz de adyacencia de un grafo dirigido, en (b) se muestra el grafo y la matriz de adyacencia de un grafo no dirigido. En el primer caso, la matriz es simétrica. Figura realizada con el programa *Wolfram Mathematica 11*.

Se define una red compleja como un conjunto de nodos y enlaces. Las redes complejas se pueden clasificar de acuerdo a diferentes aspectos como el tipo de conexión entre los nodos [**Watts & Strogatz**, **1998**]. Las redes complejas pueden ser caracterizadas y estudiadas usando algunos conceptos básicos, el más importante es probablemente el grado de un nodo, este es sólo el indicador del número de enlaces de cada nodo [**Costa et al.**, **2007**]. Es lo que hace que un nodo se destaque sobre los demás, cuanto mayor sea el grado, mayor será su capacidad de producir cambios en toda la red.

Para grafos *no dirigidos*, se define el *grado* k_i del vértice i, como el número de aristas

que inciden en él. Si A es la matriz de adyacencia asociada a una red no dirigida, entonces esta matriz es simétrica y el grado del vértice k_i se obtiene sumando todos los elementos del i-ésimo renglón de la matriz de adyacencia:

$$k_i = \sum_j A_{ij} \tag{2.4}$$

Además, la información del número total de enlaces que contiene una red E, se obtiene sumando el grado de todos los nodos de la misma y dividiendo entre 2, ya que cada enlace tiene dos vértices y es contado dos veces [**Strogatz**, **2001**]:

$$E = \frac{1}{2}\sum_i^N k_i = \frac{1}{2}\sum_{i,j} A_{ij} \tag{2.5}$$

Tambien, se puede calcular el *grado promedio de una red no dirigida* $< k >$, se define como el promedio de k sobre todos los vértices de la red, y se define como:

$$< k >= \frac{1}{N}\sum_i k_i = \frac{2E}{N} \tag{2.6}$$

Lo que significa que el grado promedio de la red se obtiene al dividir el doble del total de enlaces entre el número de vértices, pues cada enlace participa dos veces al definir el grado de los vértices.

En particular, es interesante estudiar el *coeficiente de agrupamiento* (mencionado en la literatura también como *clustering coefficient*) de un vértice en un grafo, pues, cuantifica qué tanto está de agrupado (o interconectado) con sus vecinos [**Strogatz**, **2001**]. Por lo que, el coeficiente de agrupamiento para el nodo i suele representarse formalmente como $\mathbb{C}_i$, y se define como la razón entre los enlaces que en verdad existen E_i entre los vecinos del nodo i, y la cantidad total que podrían formarse [**Atilgan et al.**, **2004**], esto es:

$$\mathbb{C}_i = \frac{2E_i}{k_i(k_i - 1)} \tag{2.7}$$

El coeficiente de agrupamiento local mide la relación que un nodo tiene con respecto a sus vecinos. Cuando el vecindario está completamente conectado el valor del coeficiente de agrupamiento es 1, cuando el coeficiente de agrupamiento vale 0 significa que el nodo está aislado y no posee ningún enlace con otro nodo.

Otro concepto clave es la *longitud de camino (path lenght)*, L entre dos nodos, se define como el número de enlaces por el número de veces que fueron cruzadas para ir de uno al otro [**Atilgan et al.**, **2004**]. Hay algunos conceptos relacionados con éste como la longitud de camino más corta, que es el número mínimo de enlaces que conectan dos nodos. Además, se puede promediar en toda la red para obtener la longitud de camino característica.

La *longitud de trayectoria característica o longitud media del camino* es el promedio sobre el número mínimo de conexiones que deben atravesarse para conectar el par de nodos i y j [**Atilgan et al.**, **2004**], es decir, es el promedio de las distancias más pequeñas entre dos nodos cualesquiera de la red [**Watts & Strogatz**, **1998**]:

$$L = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} L_{ij} \tag{2.8}$$

Por el contrario, la *distancia* d_{ij} entre los nodos i y j, se define como el número de enlaces que hay que cruzar por el camino más corto que conecte a los nodos i y j. En un grafo no dirigido $d_{ij} = d_{ji}$, mientras que en una red dirigida, lo anterior no siempre es cierto. Además, en un grafo, la distancia entre dos vértices es el menor número de aristas de un recorrido entre ellos. Por otra parte, el *diámetro del grafo* es la mayor distancia entre cualquier par de vértices en el grafo G.

$$diam(G) = max\left\{d(u,v)|u,v \in V(G)\right\} \tag{2.9}$$

Un diámetro infinito puede significar que el grafo tiene una infinidad de vértices o simplemente que no es conexo [**Strogatz**, **2001**]. También se puede considerar el diámetro promedio, como el promedio de las distancias entre dos vértices.

Cabe mencionar que el uso de la teoría de grafos tiene ciertas ventajas, ya que se pueden manejar grandes cantidades de datos relacionados de forma rápida y efectiva, soluciona múltiples problemas encontrando la solución más óptima, además, tiene una estructura flexible. Una desventaja podría ser el espacio que requiere escribir la matriz de adyacencia asociado al grafo.

2.2. Medidas de centralidad

La centralidad en teoría de grafos, mide según el criterio, la contribución o impacto de un nodo según su ubicación en la red [**Bavelas**, **1950**]. Existen cuatro medidas que son generalmente usadas en el análisis de redes, se va a estudiar tres de éstas medidas.

La *centralidad de cercanía (Closeness centrality)*, se define como el inverso del promedio de las distancias (o longitudes de los caminos más cortos) desde un nodo hacia todos los demás [**Strogatz**, **2001**]. Note que mientras mayor sea la distancia entre dos vértices, menor será la cercanía entre estos.

$$C_c(i) = \frac{N}{\sum_{i \neq j} d_{ij}} \tag{2.10}$$

En la red, los vértices más importantes son los que tienen el menor valor promedio de distancia a los demás, de esta forma $C_c(i)$ es más grande para los nodos que están a menor distancia de los demás. La cercanía mide de alguna forma la accesibilidad de un nodo en la red [**Vendruscolo et al.**, **2002**].

La *centralidad de intermediación (Betweenness centrality)* es una medida que cuantifica la frecuencia o el número de veces que un nodo actúa como un puente a lo largo del camino más corto entre otros dos nodos cualquiera [**Newman**, **2010**]. Si este puente resulta unir muchos nodos de la red por el camino más corto, este medida aumenta e indica una mayor importancia de este nodo en la red, ya que si este puente o nodo desapareciera, la unión del resto de nodos debería hacerse por otros caminos más largos[**Vendruscolo et al.**, **2002**]. Formalmente, la intermediación de un nodo i en la red se define como:

$$C_b(i) = \sum_{j \neq i \neq k} \frac{b_{jik}}{b_{jk}} \tag{2.11}$$

en donde b_{jk} es el número de caminos más cortos desde el nodo j hasta el nodo k, y b_{jik} es el número de caminos más cortos desde j hasta k que pasan a través del nodo i.

La *centralidad de vector propio (Eigenvector centrality)* $C_e(i)$ mide la influencia o conectividad de un nodo dentro de la red, es decir, un nodo que tenga pocas conexiones, puede poseer un valor alto de centralidad de vector propio, si estas pocas conexiones que tiene, a su vez están muy bien conectadas[**Vendruscolo et al.**, **2002**] . Por lo tanto, nodos que tengan un valor alto en esta medida de centralidad, serán buenos candidatos por ejemplo para difundir información en una red.

Los nodos más importantes en este sentido corresponden a centros de grandes grupos cohesivos. Mientras que en el caso de la centralidad de grado, cada nodo pesa lo mismo dentro de la red, en este caso la conexión de los nodos pesa de forma diferente [**Aguilar**, **2019**].

Sea x_i la centralidad de vector propio, se calcula mediante la suma de los elementos del renglón i de la matriz de adyacencia, multiplicados por el correspondiente x_j, definida formalmente como:

$$x_i = \frac{1}{\lambda} \sum_j A_{ij} x_j \tag{2.12}$$

donde λ es una constante. Se define el vector de centralidad $x = (x_1, x_2, ...)$ en forma matricial como:

$$\lambda x = Ax \tag{2.13}$$

donde x es el vector propio de A cuyo valor propio es λ, y para que x no tenga componentes negativas, se puede ver en [**Newman**, **2010**] que se debe elegir el valor propio más grande de la matriz A.

La *Centralidad de grado (Degree centrality)* mide la centralidad de un vértice por su grado, es decir, por su número de conexiones con otros vértices. Mientras más conexiones tenga un vértice, puede ser más eficiente al compartir información y puede ser parte de una mayor cantidad de trayectorias a lo largo de la red, esto se definió en la ecuación (2.4). Existen muchas medidas de centralidad que cuantifican diferentes características de la red, se va a trabajar con las tres primeras que son las que proporcionan mayor información y que además se han usado en diversos estudios.

En la siguiente sección se va aborda el análisis de la estructura tridimensional. La estructura de una proteína se determina en la actualidad por diferentes métodos en los cuales se obtienen una serie de datos que el científico utiliza para crear el modelo atómico final. Entre ellos, la cristalografía de rayos X, que genera un modelo de difracción de rayos X, como resultado se obtienen las coordenadas de los átomos de la proteína y estos son una buena aproximación de las posiciones medias de los átomos en solución. Sin embargo, en solución, los átomos pueden sufrir grandes fluctuaciones sobre estas posiciones y, por lo tanto, muchos átomos que aparentemente están justo debajo de la superficie en la estructura cristalina estática de rayos X pueden entrar en contacto transitorio con el solvente. Tales átomos y residuos podrían hacer una contribución diferente a la estabilidad de las proteínas que los residuos ubicados más profundamente. Siendo así, se va a analizar algunas propiedades de proteínas para detectar las ubicaciones y tipos de interacciones entre residuos, En particular, se muestra cómo modelar la estructura terciaria de una proteína con la metodología de redes y cómo vincular las características mencionadas de una red, con la determinación de los residuos del sitio activo de la proteína.

2.3. Modelos experimentales de estructuras de proteínas con Rayos X

En la actualidad existe una base de datos muy importante de modelos experimentales de estructuras de proteínas: el PDB o Banco de Datos de Proteínas [**Berman et al.**, **2000**]. En este banco de datos se almacenan, entre otras cosas, las coordenadas espaciales de los átomos de cada modelo. Estos datos se depositan en forma de archivos que reciben el nombre de archivos PDB. Por ejemplo, en 1EKG.pdb está archivada la estructura de la frataxina humana. Cada uno de estos archivos tiene un nombre particular que permite identificar y no confundir un modelo con otro. El nombre es único. En estos archivos, además de las coordenadas atómicas, se almacena muchísima información útil para los investigadores acerca de

los experimentos que dieron origen al modelo e información sobre el modelo en sí. Para ver los modelos y trabajar con ellos, se usan programas de computación especiales. Muchos de estos programas son gratuitos. Algunos de estos programas son Rasmol, SwissPDBviewer, MOL-MOL y VMD, entre otros.

Los archivos PDB tienen una estructura particular que es importante comprender, por ahora interesa estudiar una parte del archivo que incluye las coordenadas espaciales de cada átomo de carbono. Del *Protein Data Bank* [**Berman et al.**, **2000**] se va a trabajar con las poliproteínas (PP1a y PP1ab) que se dividen en 16 proteínas efectoras no estructurales, particularmente con cuatro de estas. Nuestro objetivo en este trabajo de investigación es identificar los residuos del sitio activo estrictamente de la estructura tridimensional de la proteína, es por eso que se va a transformar las estructuras de proteínas en gráficos de interacción de residuos, donde los residuos son nodos de grafos y sus interacciones entre sí son los enlaces de los grafos.

Ahora bien, con información del *Protein Data Bank*, se creó un script en *Wolfram Mathematica* que importa un archivo PDB desde el sitio web del banco de datos de proteínas RCSB, en este caso se toma como ejemplo la poliproteína no estructural NSP3 con PDB ID 6W9C (Disponible en https://www.rcsb.org/structure/6W9C) [**Berman et al.**, **2000**]. Antes de continuar, se sabe que una proteína esta formada por aminoacidos. Cada aminoácido tiene un átomo central de carbono que se llama C_α. Estos átomos C_α son los átomos que forman el backbone (esqueleto) de la proteína, el punto clave es extraer las coordenadas de estos átomos C_α para transformar la estructura de la proteína en una red.

Nuestro objetivo es estudiar la estructura de la *protein residue network* desde una perspectiva estadística a fin de revelar el papel de la disposición local en la estructura general y la dinámica de las proteínas. En el resto de este estudio, se presentan los resultados de las *protein residue network* que se construyen utilizando un radio de corte de 6 Å y 8 Å. Mediante el script elaborado en *Wolfram Mathematica* se lee el archivo PDB y devuelve una representación del backbone (columna vertebral) de la proteína en una forma estilizada, se extrae las coordenadas de los átomos C_α, a continuación, el problema que nos confiere aquí es análizar e ir seleccionando cada uno de estos residuos y ver con cuales de los restantes interactúa o se va a establecer un link(enlace), para esto se va a establecer una esfera de radio 8 Å con centro las coordenadas del átomo $C_{\alpha i}$ y con esto ver que átomo C_α pertenecen o que caen dentro de esta esfera y cuales no. Esta clasificación depende del umbral o radio de corte elegido, para evaluar la capacidad predictiva de un clasificador de este tipo es necesario realizar un análisis y comparación de resultados, es decir, si se toma un radio de corte de $R_c = 6$ Å el número de links para cada átomo es menor a comparación de cuando se toma un radio más grande.

El resultado de contar todos los átomos C_α que caen dentro de la esfera con centro el residuo i o que se encuentren a una distancia menor que la distancia de corte R_c, se está considerando la cercanía entre átomos C_α como la única razón que define si existe interacción entre aminoácidos y si se debe establecer un enlace entre ellos como se muestra en la figura (2.2).

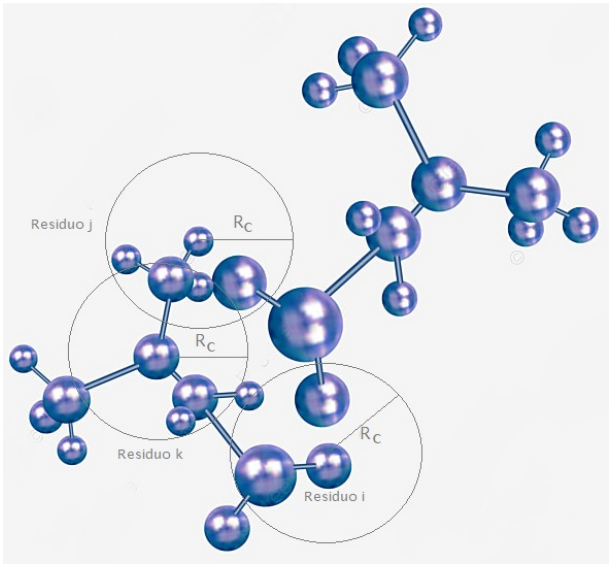
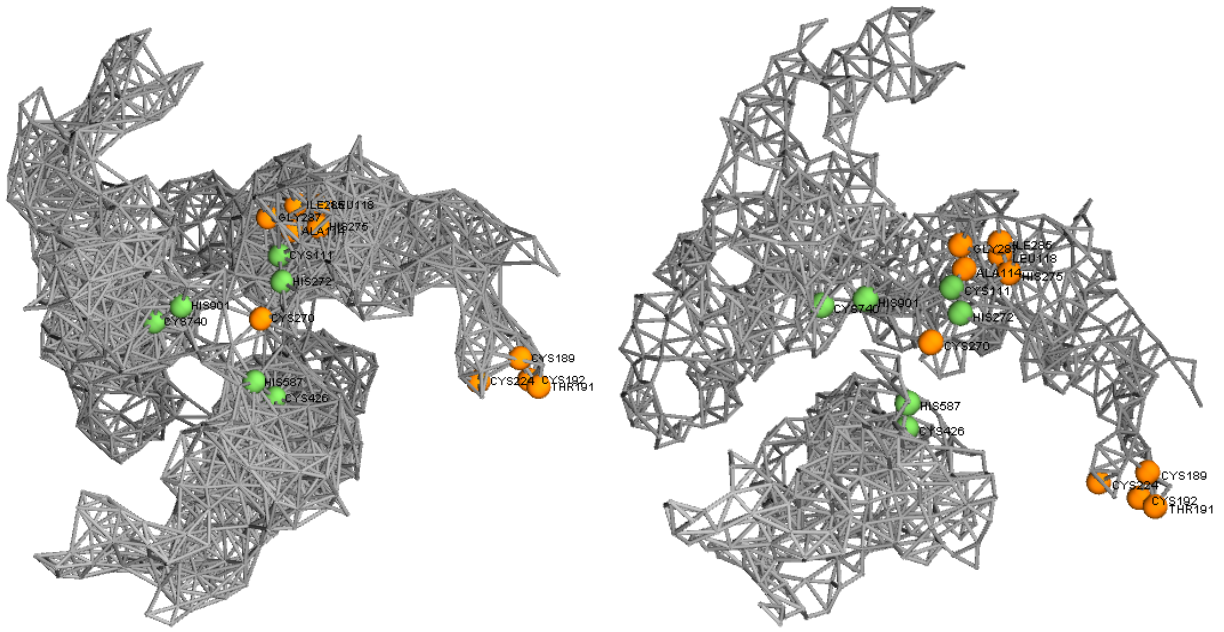


Figura 2.2: Esquema para mostrar cómo calcular los elementos de la matriz de adyacencia. Se toma un átomo $C_{\alpha i}$ y se cuenta cuántos átomos $C_{\alpha j}$ están a una distancia menor a R_c o que caen dentro de la esfera de radio R_c, los átomos que caen dentro de la esfera se les asigna un 1 dentro de la matriz de adyacencia indicando que existe un enlace de conexión, en el caso contrario se les asigna 0, esto se hace para todos los átomos $C_{\alpha i}$.

En este sentido, se tienen 926 átomos C_α y con información de sus coordenadas se construye una red con N nodos ubicados en dichas coordenadas. Como se mencionó anteriormente, en este caso se requiere que cada nodo n_i de la red, se le considere una esfera de radio 8 Å, y todos aquellos nodos n_j que caigan dentro de esta esfera tendrán una conexión (link) con el nodo inicial, también se consideró una esfera de radio 6 Å como se muestra en la figura (2.3).

Los sitios activos enzimáticos son regiones funcionales importantes y conservadas de proteínas cuya identificación puede ser un paso invaluable hacia la predicción de la función de las proteínas. La mayoría de los métodos existentes para esta tarea se basan en la similitud del sitio activo y presentan limitaciones, incluida la realización de coincidencias exactas en residuos de plantillas, restricciones de tamaño de plantilla, a pesar de no ser capaces de encontrar residuos del sitio activo entre dominios [**Kahraman et al.**, **2008**]. En el siguiente

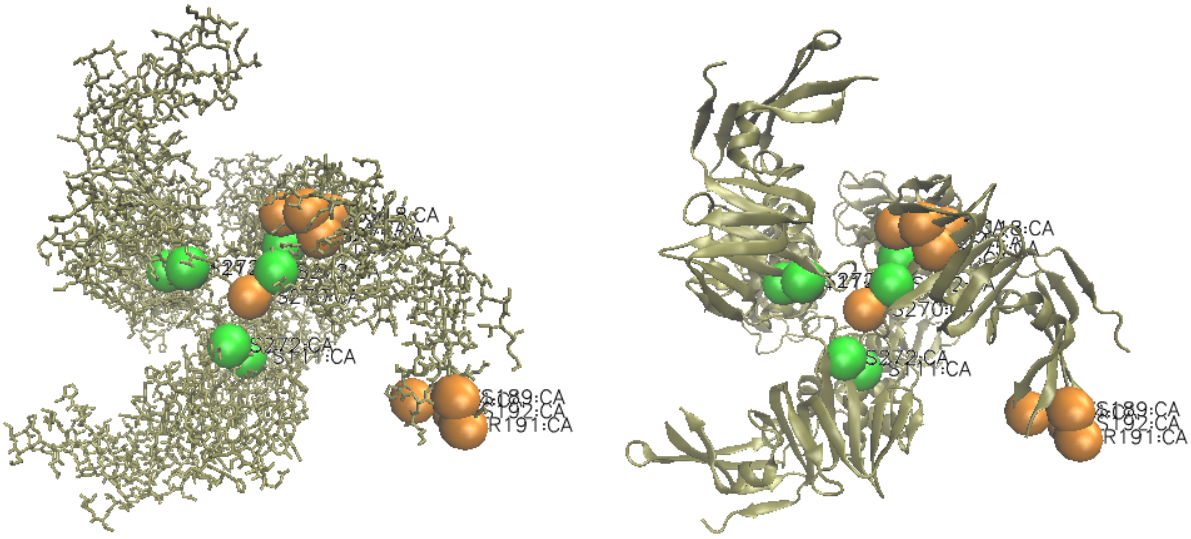


(a) (b)

Figura 2.3: Representación de la *protein residue network*. En (a) se muestra la representación de la *protein residue network* asociado a la matriz de adyacencia de la proteína 6W9C [**Berman et al.**, **2000**] y construida a partir de las coordenadas de los residuos C_α a un radio de 8 Å, en (b) se muestra la imagen de la *protein residue network* a un radio de 6 Å, en ambos casos se tiene una red con 926 nodos. Los residuos del sitio catalítico se marcan en color verde y los residuos del sitio de unión en color naranja. Figura elaborada con el *Software Wolfram Mathematica*.

capítulo se verán algunos conceptos propios de la Teoría de Redes, aplicado el modelo proteínas. De esta manera se podrán comparar ambos métodos entre sí (determinación del sitio activo mediante el modelado) y con resultados experimentales conocidos. En la figura (2.4) se muestra la estructura de la proteína de nuestro ejemplo anterior (PDB ID: 6W9C), de acuerdo al archivo PDB obtenido del *Protein Data Bank* [**Berman et al.**, **2000**], obtenidas con el programa Visual Molecular Dynamics (VMD).

La transformación de las estructuras proteicas en *protein residue network* nos va a permitir examinar la relación entre cada residuo o aminoácido y la estructura de la proteína en su conjunto. Las relaciones se evaluan por la centralidad de cada residuo dentro de la red de interacción. La matriz de adyacencia asociada a esta red de residuos de proteína será de gran importancia para poder aplicar las medidas de centralidad antes mencionadas. Asimismo, se estudió la superficie accesible al solvente y profundidad del átomo para complementar la información como se verá a continuación.



(a) (b)

Figura 2.4: Representación de la estructura tridimensional de la proteína. En (a) se muestra la representación Licorice, los residuos del sitio catalítico se marcan en color verde y los residuos del sitio de unión en color naranja, en (b) se muestra la representación con el método de representación de Cadenas New Cartoon, figuras elaboradas con el Software Visual Molecular Dynamics [**Fackler D. et al.**, **2021**].

2.4. Cálculo de área de superficie accesible al solvente

Al plegarse una proteína, una gran fracción de sus residuos se vuelven inaccesibles al solvente. La ubicación de los residuos se cuantifica convencionalmente mediante el *Solvent accessible surface area* (SASA) [**Chakravarty & Varadarajan**, **1999**]. SASA mide el nivel de exposición de un residuo, al solvente (agua) en una proteína [**Heffernan et al.**, **2015**]. Esta es una propiedad estructural importante ya que los residuos del sitio activo de las proteínas a menudo se encuentran en sus superficies.

El área de superficie accesible al solvente (SASA) de las proteínas siempre se ha considerado como un factor decisivo en los estudios de plegamiento y estabilidad de proteínas. Se define como el área de superficie atómica de una molécula (proteína) que es accesible a moléculas de agua, y generalmente se expresa en $Å^2$. SASA se calcula utilizando el algoritmo de "bola rodante"[**Mihel et al.**, **2008**], que utiliza una esfera (que representa el solvente) de un radio particular para "sondear"la superficie de la molécula. Un valor típico de un radio de

sonda es 1.4 Å, que se aproxima al radio de una molécula de agua. Según los valores de SASA, los residuos de una proteína se pueden clasificar como inmerso o expuestos [**Ali et al.**, **2014**]. El SASA de una proteína se puede estimar computacionalmente a partir de las coordenadas atómicas [**Heffernan et al.**, **2015**].

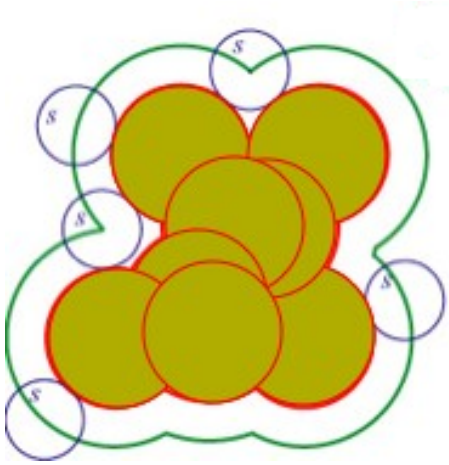


Figura 2.5: Superficie accesible al solvente. Los átomos de la proteína se representan como esferas y el área de superficie accesible al solvente se define por el centro de una bola rodante (S) que representa la molécula de solvente mientras pasa sobre la superficie de la proteína. Figura extraída del libro "Biología Molecular" de Robert F. Weaver. Capítulo 5: Herramientas moleculares para el estudio de genes y actividad génica.

Muy a menudo, un investigador quiere calcular rápidamente el área de superficie accesible al solvente, por ejemplo de una molécula de proteína, GETAREA [**Negi et al.**, **2015**] es un método eficiente de cálculo de la superficie accesible al solvente [**Fraczkiewicz & Braun**, **1998**] implementado en el programa FANTOM. Las coordenadas atómicas deben suministrarse en formato PDB. GETAREA es un servicio web proporcionado por el *Sealy Center for Structural Biology de la University of Texas Medical Branch*. Originalmente, GETAREA era una subrutina para el cálculo analítico eficiente del área de superficie accesible al solvente y su gradiente para proteínas [**Von et al.**, **1993**]. Este servicio permite al usuario enviar coordenadas cartesianas de átomos en una molécula, almacenadas en formato PDB en su disco local, y recuperar SASA o la energía de solvatación (dependiendo de la configuración del parámetro) en una variedad de formatos. De forma predeterminada, el formulario de envío está configurado para calcular SASA de átomos no hidrógeno en proteínas, pero un cambio apropiado de los parámetros de entrada permitirá calcular cualquier cantidad proporcional a SASA para cualquier tipo de molécula.

2.5. Cálculo de la profundidad del átomo

La profundidad del átomo es definida como la distancia (Å) de un átomo (C_α) de la proteína al átomo del solvente más cercano que encuentre, la profundidad del átomo proporciona una descripción simple pero precisa del interior de la proteína, se puede utilizar de una manera directa y efectiva para obtener una descripción sensible y precisa del interior de la proteína y, por lo tanto, complementar la información que se obtiene del cálculo de SASA. La profundidad átomo/residuo mide la ubicación de un residuo al átomo del solvente más cercano, esta medida ha encontrado una variedad de usos en la caracterización de las propiedades físicas y químicas de las estructuras proteicas [**Pintar et al.**, **2003**].

Los residuos más profundos en el estado nativo invariablemente experimentan intercambio de hidrógeno por despliegue global de la proteína y a menudo están significativamente protegidos en los correspondientes estados de glóbulo fundido [**Hubbard et al.**, **1994**]. La profundidad es a menudo un medidor más útil de la ubicación de residuos [**Pintar et al.**, **2003**]. Esto probablemente esté relacionado con el hecho de que el interior de la proteína y el solvente circundante difieren significativamente en polaridad y densidad de empaquetamiento.

Recientemente se definio la "profundidad atómica"(dpx) como la distancia (Å) de un átomo de la proteína al átomo de la molécula de agua superficial más cercana [**Carugo & Pongor**, **2002**]. Usando esta definición, la profundidad de un átomo es, por lo tanto, cero para todos los átomos accesibles con solvente, $y > 0$ para los átomos inmersos en el interior de la proteína, átomos ubicados más profundamente tienen valores de dpx más altos [**Pearlman & Kollman**, **1995**].

Se han reportado enfoques algo diferentes dirigidos al cálculo de la profundidad del átomo. En la "molécula de agua más cercana", la proteína se coloca en una red 3D que contiene moléculas de agua y se mide la distancia entre los átomos y la molécula de agua más cercana. En nuestro enfoque, se utiliza *DPX Protusion Index Web Server* [**Nagy et al.**, **2016**] que es un servidor web de índice de protusión DPX. DPX es un algoritmo simple y rápido que calcula la profundidad de cada átomo en la estructura de una proteína. Como los archivos de salida se pueden mostrar directamente mediante programas de gráficos moleculares estándar, DPX se puede utilizar para la inspección visual del interior de la proteína de una manera sencilla. La profundidad del átomo/residuo se correlaciona con la estabilidad termodinámica de una proteína, con la formación de complejos proteína/proteína, con las tasas de intercambio de amida H/D medidas por NMR, con el grado de conservación de residuos en familias de proteínas estructuralmente relacionadas y con tipos de aminoácidos.

Métodos: La profundidad del átomo (dpx) se define como la distancia (Angstrom) de un átomo de la proteína al átomo del solvente accesible, donde la accesibilidad del átomo al solvente se calcula utilizando el algoritmo de esfera rodante. Por lo tanto, los átomos accesibles al solvente tienen dpx = 0 por definición, los átomos inmersos tendrán valores de dpx crecientes a medida que uno se mueve hacia el núcleo de la proteína. Los valores típicos de dpx están en el rango de $0 - 8$ Å. El valor máximo de dpx (dpx(máx)) encontrado para una proteína monomérica de cadena única depende en primera instancia del número de residuos en la cadena. Aunque, se debe de tener en cuenta, que diferentes proteínas, aunque tengan el mismo número de residuos, pueden mostrar valores bastante diferentes para dpx(máx).

DPX Protusion Index Web Server [**Nagy et al.**, **2016**] lee archivos de coordenadas PDB estándar como entrada. En la presente forma, DPX está dirigido al análisis del interior de la proteína monomérica de cadena única. Antes del envío, se debe de inspeccionar el archivo PDB y, si es el caso, editarlo para tener una sola molécula (cadena). El programa solo lee líneas ATOM. Por lo tanto, las líneas HETATM que describen residuos no estándar, cofactores, iones metálicos y moléculas de agua no se tienen en cuenta.

2.6. Proyecto Alpha Fold 2

La estructura de una proteína está ligada a su función y eso le permite determinar su rol. El resolver la estructura de una proteína puede tomar varios años y ser altamente costoso. Siendo esta información crucial para el desarrollo de fármacos terapéuticos y la investigación biológica, el plegamiento de proteínas es un tema principal en la investigación científica. Recientemente un hallazgo ha demostrado que el emplear la inteligencia artificial puede acelerar el desarrollo de nuevos fármacos y entender algunas enfermedades mediante el descifrar la estructura espacial de las proteínas, tan solo con la secuencia lineal de sus aminoácidos [**Das et al.**, **2016**]. El descifrar la estructura de las proteínas nos permite comprender como funcionan los procesos en la célula. Enfermedades como el Alzheimer, y el Parkinson, están relacionadas al mal plegamiento de las proteínas. Si se conoce la estructura de una proteína se puede aplicar este conocimiento a la implementación de terapias o desarrollo de enzimas para funciones de interés. Sin embargo, se necesita resolver el enigma del plegamiento, para ello actualmente existen técnicas costosas como la resonancia magnética nuclear, la cristalografía de rayos X, y la Crio-microscopia electrónica.

La competencia internacional CASP (*Critical Assesment of Protein Structure Prediction*), se realiza desde 1994 con la finalidad de investigar el plegamiento de las proteínas empleando

métodos computacionales. Cada dos años se celebra una competencia para elegir cuál es el software que predice mejor la estructura de proteínas a partir de sus secuencias de aminoácidos. El concurso se conoce por las siglas CASP (del inglés *Critical Assessment of Protein Structure Prediction*) [**Stephen**, **2020**]. En 2020 se celebró la decimocuarta edición de la competición (CASP14), y resultó vencedor el algoritmo AlphaFold 2, desarrollado por la empresa DeepMind, con una gran ventaja sobre sus competidores [**Jumper et al.**, **2021**].

Predecir la estructura 3D de una proteína a partir de su secuencia lineal de aminoácidos es un problema biológico complejo que debe involucrar entender las dinámicas de interacción espacial de los aminoácidos [**Das et al.**, **2016**]. La propuesta de DeepMind es resolver el problema del plegamiento basado en conversión de patrones de imágenes. Se propone encontrar las distancias entre todos los aminoácidos para formar una matriz de correlación [**Callaway**, **2020**], se obtiene una primera matriz con todas las distancias de los aminoácidos llamada PDM (*protein distance matrix*). Adicionalmente, se encuentran los aminoácidos cruciales en la estructura espacial empleando MSA (*"multiple sequence alignment "*). MSA es un alineamiento masivo en donde se comparan proteínas con una relación evolutiva. En el alineamiento, se comprueba si un aminoácido cambia ("evoluciona") debido a una mutación y que otro aminoácido de la misma proteína suele cambiar.

Estos cambios paralelos se dan debido a que estos interaccionan físicamente en la proteína y deben cambiar juntos para mantener la estructura de la proteína (coevolución estructural). Por lo tanto, DeepMind propuso combinar las distancias globales de todos los aminoácidos mediante PDM con distogramas (el distograma es una imagen que es una simple representación de la distancia tridimensional de todos los aminoácidos), y al mismo tiempo identificar los aminoácidos que interaccionan físicamente identificados por MSA para inferir la estructura espacial de las proteínas. Todos estos datos se analizan empleando algoritmos de redes neuronales convolucionales, usadas para reconocimiento de imágenes. El modelo "aprende" a generar una imagen, representa la estructura tridimensional de la proteína (distograma). Empleando un segundo algoritmo de inteligencia artificial llamado "descenso del gradiente" (*gradient descent*); se ajusta la estructura para plegarla optimizando los ángulos de torsión y los que interaccionan físicamente de los aminoácidos relacionados en coevolución.

A la fecha, la limitante al proceso de producción de proteínas de novo es la disponibilidad de estructuras cristalizadas y descritas en las bases de datos como *Protein Data Bank* [**Berman et al.**, **2000**]. Únicamente con esta colección los biólogos estructurales infieren moléculas 3D similares apoyados en algoritmos de comparación espacial y así predecir una función biológica. El juego ha cambiado y AlphaFold ofrece una alternativa poderosa para

predecir la estructura 3D de las proteínas y estudiar su función; por ello, se le ha catalogado como el avance tecnológico más relevante del año 2020 [**Jumper et al.**, **2021**].

Capítulo 3

Cálculo del coeficiente de agrupamiento $\mathbb{C}_i$ y medidas de centralidad para proteínas relacionadas con el virus del SARS-CoV 2

En este capítulo se representa la estructura tridimensional de una proteína como una red, para esto, cada aminoácido de la cadena proteica respresenta un nodo en la red(vértice) y, las conexiones entre ellos se representaban como aristas (links o enlaces). Con este método de análisis se relacionan las características mencionadas de una red para la determinación de los residuos del sitio activo de proteínas relacionadas con el virus del SARS-CoV-2, dado que es de relevancia determinar y cuantificar que tan importante es cada nodo o enlace dentro de la estructura de la red. En teoría de redes, las medidas de centralidad son justamente las variables que se usan para determinar esta importancia.

El genoma del virus SARS-CoV-2 es una sola cadena de ARN que codifica una pequeña cantidad de proteínas que son sintetizadas por la maquinaria ribosómica después de que el virus infecta una célula huésped, la figura (3.1) muestra ejemplos de proteínas del virus SARS-CoV-2 destacando el papel de éstas en la infección viral, estas proteínas están disponibles en PDB (se muestran con el ID de PDB, por ejemplo, 6W9C). Estas estructuras brindan información sobre los mecanismos de la infección por COVID-19 y se utilizan actualmente en el diseño de vacunas y terapias.

Las proteínas del virus SARS-CoV-2 se pueden agrupar en dos clases principales, pro-

teínas estructurales (SP) y no estructurales (NSP). Las proteínas estructurales forman la partícula viral, median la unión y fusión del virus a la célula huésped, así como su ensamblaje, mientras que las proteínas no estructurales participan en el ciclo viral del virus y favorecen su infección, además, tienen funciones en la replicación y maduración de las proteínas.

El virus tiene cuatro proteínas estructurales: la espícula (S, *Spike*), la de envoltura (E), la de membrana (M) y la nucleoproteína (N), la proteína S es la más antigénica y externa, además de que es la responsable, en gran parte, de la forma de corona al virión.

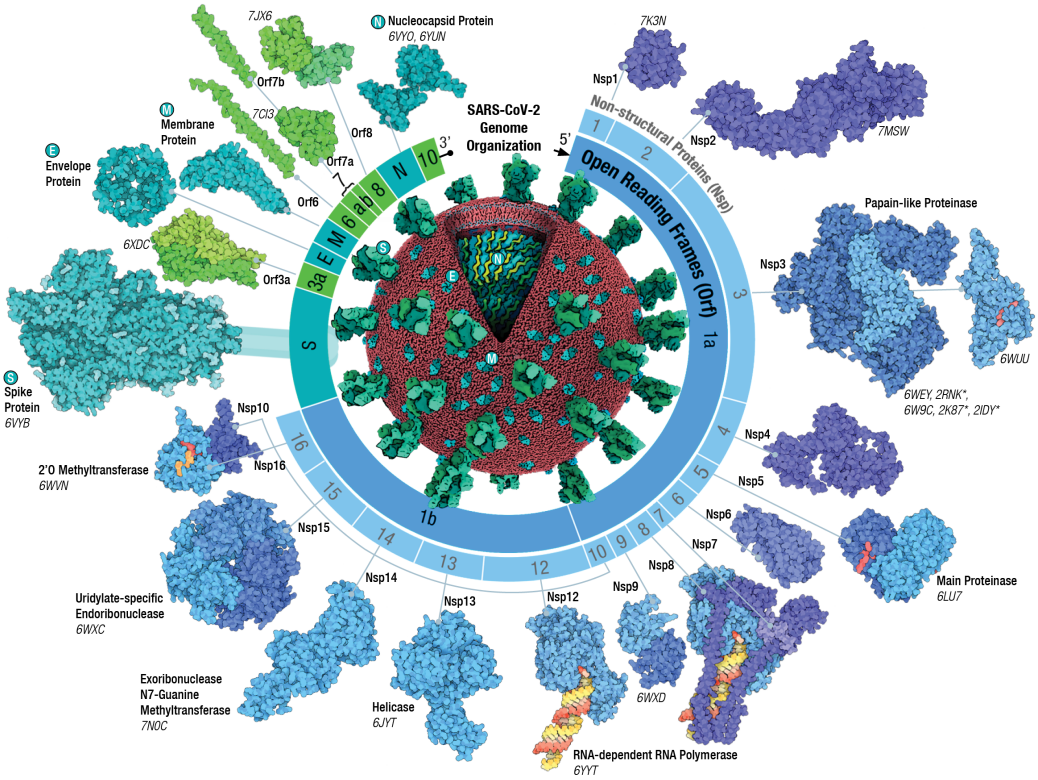


Figura 3.1: Proteínas del virus del SARS-CoV-2. La ilustración presenta ejemplos de proteínas del virus del SARS-CoV-2 disponibles en el PDB, y modelos desarrollados con base al genoma y otra información estructural disponible. La secuencia del genoma del SARS-CoV-2 se puso a disposición del público el 10 de enero de 2020 y el 5 de febrero de 2020 la primera estructura del SARS-CoV-2 determinada experimentalmente, la proteinasa principal se publicó en el archivo del *Protein Data Bank*. Figura extraída de PDB [**Berman et al.**, **2000**].

En el caso de SARS-CoV-2, la proteína S es la proteína que reconoce la proteína celular ECA2 (enzima convertidora de angiotensina 2) como receptor para fijarse a la célula hospedera. La proteína N proporciona protección al genoma e interviene en la síntesis del ARN viral,

Proteína	Cadenas	Total de residuos	Residuos del sitio activo reportados en la base de datos PDB	Residuos del sitio activo reportados en la base de datos UNIPROT	Residuos del sitio activo reportados en la base de datos SWISS MODEL
6W9C	3	926	CysA111, HisA272 CysB111, HisB272 CysC111, HisC272	CysA111, HisA272, AlaA114 LeuA118, HisA275, IleA285 GlyA287, CysA189, ThrA191 CysA192,CysA224, CysA270	CysA111 HisA272
6WXC	1	695	His235, His250 Lys290	His235, His250, Lys290 Asn1, Glu4, Pro24, Val25 Leu43, Glu45, Glu57, Asn75 Val78, Asp79, Ala95, Ile97 Cys103, Met105, Gly170 Glu171, Lys181, Gln245 Gly248, Val292, Ser294 Leu312, Val314, Tyr325 Gly337, Thr341, Tyr343 Lys345, Asn349 , Glu353 Pro373, Val376, Glu394 Glu406, Cys452, Gln594 Gly597, Val641, Ser643	His235 His250 Lys290
6WVN	2	422	Lys47, Asp131 Lys171, Glu204	Lys47, Asp131, Lys171, Glu204 Trp8, Pro15, Leu17, Cys28 Leu30, Tyr33, Asp35, Ser36 Leu39, Lys41, Gly42, Met44 Met45, Asn46, Tyr50, Arg69 Gly74, Gly76, Lys79, Pro83 Gly84, Asp102, Leu103, Asn104 Ile115, Asp117, Cys118, His122 Thr123, Asn125, Lys126, Tyr135 Lys140, Lys144, Glu150, Gly151 Lys163, Thr175, His177, Ser204 Gly216, Pro218, Tyr225, Pro239 Arg258, Lys280, Arg282, Arg286 Glu287, Asn289	Lys47 Asp131 Lys171 Glu204
6WXD	2	223	GluA68, AsnA95 AsnA96, LeuA97	GluA68, AsnA95, AsnA96 LeuA97, ArgB10, GlnB11 AlaB28, LeuB29, LeuB45 AsnB96	GluA68 AsnA95 AsnA96 LeuA97

Cuadro 3.1: Información de la ubicación de residuos del sitio activo de proteínas relacionadas con el SARS-CoV-2 [**Berman et al., 2000**], el número de cadenas y número de aminoácidos C_α de cada proteína se enumeran en las columnas 2 y 3 respectivamente. Las siguientes tres columnas muestran las comparaciones de información de ubicación de residuos del sitio activo de diferentes bases de datos, en este caso el *Protein Data Bank* [**Berman et al., 2000**], Uniprot [**Lombardot et al., 2021**] y Swiss Model [**Bienert et al., 2017**].

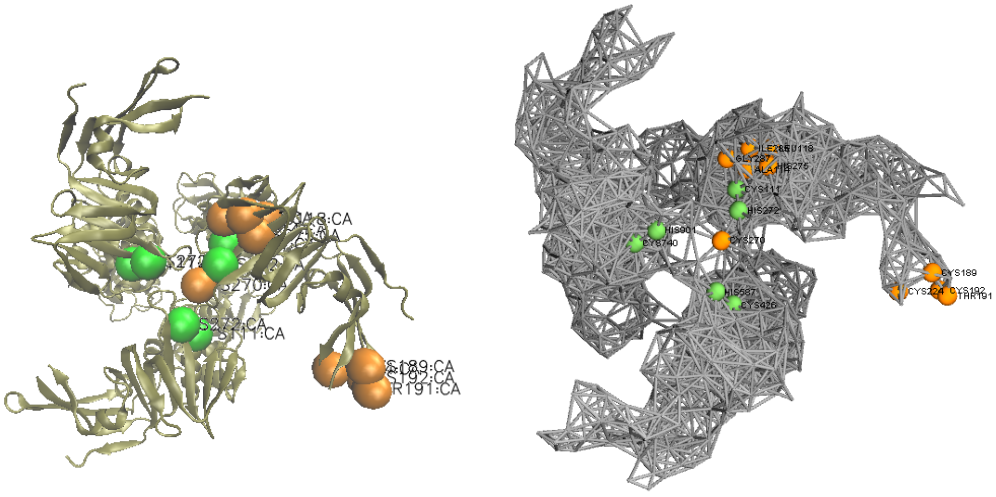
mientras que la M da estructura y estabilidad al virión; la E es un canal iónico con funciones durante el ensamble y la salida de los viriones de la célula hospedera. Entre las regiones codificantes de las proteínas estructurales se encuentran otros marcos de lectura (ORF, *open reading frame*) que codifican una serie de proteínas denominadas accesorias, aunque, estas proteínas se consideran no indispensables para el ciclo replicativo de los coronavirus, sin embargo, tienen diferentes funciones al interaccionar con proteínas del hospedero en diferentes vías de señalización relacionadas con la respuesta antiviral. Mientras que las proteínas no estructurales son al menos 16 y son conocidas como NSP1 a NSP16. Por otra parte, entre las 16 proteínas no estructurales se encuentran helicasa, trifosfatasa, metiltransferasa y nucleasa. Destaca especialmente la NSP5, la principal de dos proteasas que tiene el virus, conocida como Mpro (*main protease*) o 3CLpro (*3C like protease)*, proteasa del tipo de la quimiotripsina que interviene en la maduración de 12 proteínas, ya que actúa sobre 11 sitios de corte entre los productos de traducción del genoma viral.

En este caso, se va a trabajar con cuatro de las dieciséis proteínas no estructurales (NSP) que son codificadas por el genoma del SARS-CoV-2, los datos de la tabla (3.1) muestra la información de éstas cuatro proteínas, así cómo la ubicación de sus residuos del sitio activo en el formato PDB [**Berman et al.**, **2000**], UNIPROT [**Lombardot et al.**, **2021**] Y SWISS MODEL [**Bienert et al.**, **2017**].

3.1. Análisis de medidas de centralidad de la proteína 6W9C

La proteína no estructural 6W9C es una integrante del complejo NSP3, incluye un dominio de proteasa similar a la papaína (PLPro), que es responsable de la escisión (corta las proteínas en 3 proteínas NSP funcionales) de la cadena polipeptídica en tres sitios dentro de las porciones N-terminales de ambas proteínas. Es una de las dos proteasas virales responsables del procesamiento de los productos poliproteicos de traducción del genoma viral después de la infección [**Rimanshee et al.**, **2021**]. La proteasa similar a la papaína (PLpro) del coronavirus 2 del síndrome respiratorio agudo severo (SARS-CoV-2) desempeña un papel esencial en la replicación del virus y la evasión inmune, en particular, esta proteína está formada por tres cadenas semejantes (se tienen aminoácidos para la cadena A, para la cadena B y para C) como se muestra en la figura (3.2).

Esta proteína está formada por tres cadenas semejantes, nos interesa determinar los residuos cuya distancia promedio a los demás es mínima, es decir, los valores más altos en la



(a) (b)

Figura 3.2: Construcción de red a partir de una proteína. En (a) se muestra la estructura de la proteína 6W9C de acuerdo al archivo PDB, figura elaborada con el programa Visual Molecular Dynamics (VMD). En (b) se muestra la red construida a partir de los residuos C_α considerando un radio de corte de $R_c = 8$ Å. Los residuos del sitio catalítico se marcan en color verde y los residuos del sitio de unión en color naranja. Figura elaborada con el programa *Wolfram Mathematica 11*.

centralidad de cercanía [**Watts & Strogatz**, **1998**]. En la figura (3.3) se observa que para un radio de $R_c = 8$ Å los residuos que forman el sitio activo ocupan picos que se encuentran en la región que representa el 10 % de los valores más altos, en este caso se tiene que $CysA270$ es máximo local, mientras que $CysA111$, $CysA272$, $CysB111$, $CysB272$, $CysC111$ y $CysC272$ permanecen cercanos a él, lo que significa que se encuentran a una mínima distancia de los demás residuos. La centralidad de intermediación cuantifica la frecuencia o el número de oportunidades en las que un nodo actúa o sirve de conexión dentro de una ruta entre dos nodos determinados, para este caso los residuos $HisA272$, $HisB272$ e $HisC272$ son máximos locales que forman parte de los residuos por los que pasan la mayor cantidad de caminos que comunican al resto. En el cálculo de la centralidad de vector propio, los nodos que poseen un alto valor de esta medida están conectados de forma más eficiente dentro de un grafo determinado. En este caso el residuo $CysC111$ se encuentra entre los 5 picos más altos, el resto de residuos se localizan a un 30 % por arriba del promedio y se localizan en la región de los valores más altos, lo que significa que aunque no están altamente agrupados, sus vecinos sí están conectados a otros nodos relevantes o de alto grado.

Cabe mencionar que para este caso, la centralidad de cercanía e intermediación funcionan mejor para ubicar residuos del sitio activo en la estructura de la proteína, ya que hay más

residuos cercanos a los máximos locales. Por otro lado, se realizó el análisis de la *protein residue network* con $R_c = 6$ Å en donde se observa que los valores de la centralidad de cercanía e intermediación disminuyen al disminuir el radio de corte, mientras que los valores de centralidad de vector propio aumentan, sin embargo, la forma cualitativa de las gráficas no se modifica. Para este radio de corte se observa que las predicciones de los residuos funcionales mejoran al aumentar el radio de corte.

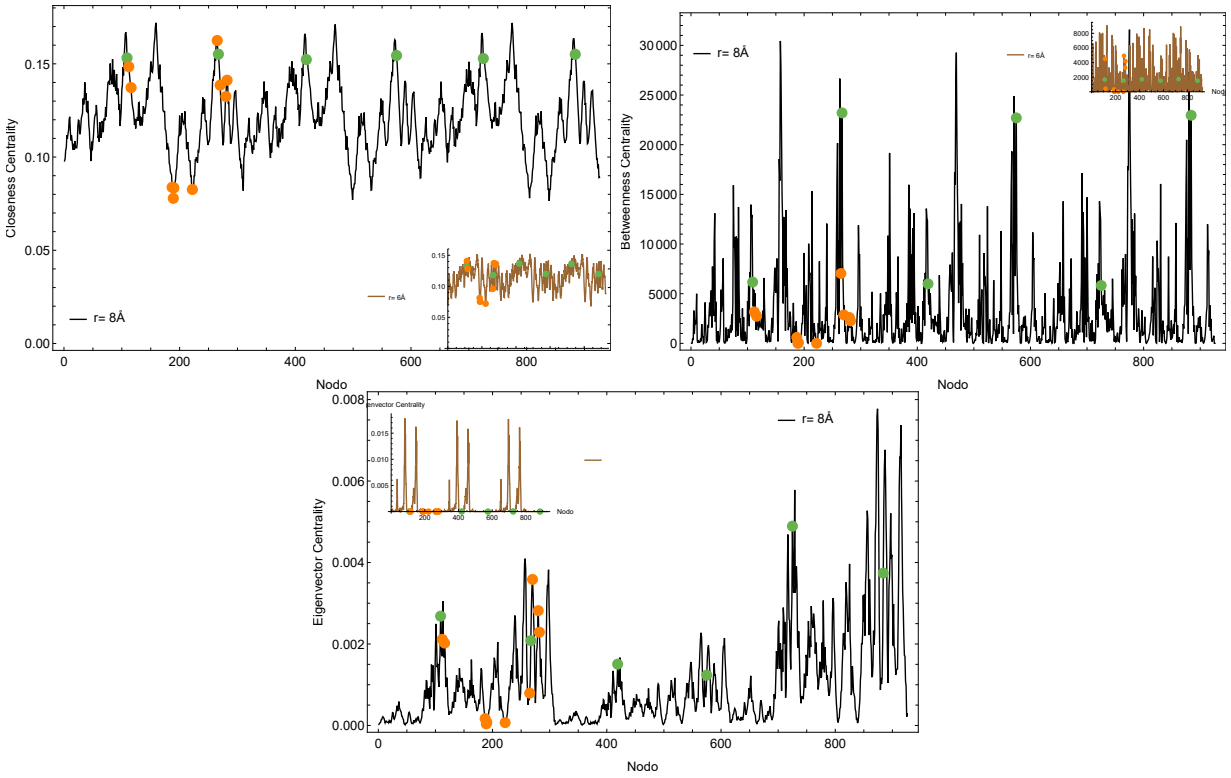
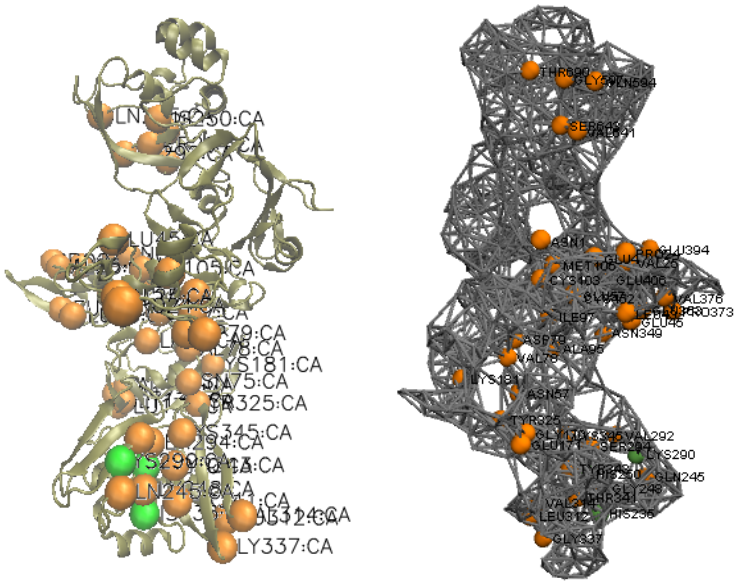


Figura 3.3: Gráficas de la centralidad de cercanía, intermediación y *eigenvector* de la proteína 6W9C. En todos los casos, la gráfica de líneas negras es para la *protein residue network* con $R_c = 8$ Å y la gráfica pequeña de líneas cafés ubicada en la parte inferior o superior es para $R_c = 6$ Å. Los residuos del sitio catalítico se marcan en color verde y los residuos del sitio de unión en color naranja. Figuras elaboradas con el programa *Wolfram Mathematica 11*.

Se observan picos máximos que corresponden a residuos que no están reportados como parte del sitio activo, pero, se podría explorar más a detalle con otras metodologías de análisis si existe la posibilidad de que estos residuos tengan alguna participación en la función de la proteína.

3.2. Análisis de medidas de centralidad de la proteína 6WXC

La proteína no estructural NSP15 (6WXC) endoribonucleasa del SARS-CoV-2, tiene la función de limpiador, es decir, de destruir el ARN viral sobrante para evadir a las defensas de la célula infectada [**Rimanshee et al.**, **2021**]. La figura (3.4) muestra la estructura de la proteína de acuerdo al archivo PDB y la estructura de la *protein residue network* con $R_c = 8$ Å.



(a) (b)

Figura 3.4: Construcción de red a partir de una proteína. En (a) se muestra la estructura de la proteína 6WXC de acuerdo al archivo PDB, figura elaborada con el programa Visual Molecular Dynamics (VMD). En (b) se observa la red construida a partir de los residuos C_α considerando un radio de corte de $R_c = 8$ Å. Los residuos del sitio catalítico se marcan en color verde y los residuos del sitio de unión en color naranja. Figura elaborada con el programa *Wolfram Mathematica 11*.

En la figura (3.5) se muestran los resultados para las medidas de centralidad para un radio de $R_c = 8$ Å, en particular, los residuos $Val78$ y $Asp79$ son máximos locales para la centralidad de cercanía e intermediación, los otros residuos reportados no están en la franja de valores máximos, pero, se localizan en un 20 % por arriba del promedio. Para la centralidad de vector propio, no es evidente la relación de estos residuos con valores extremos. Para este radio de corte, la centralidad de cercanía e intermediación funcionan mejor para ubicar

residuos del sitio activo en la estructura de la proteína, ya que hay más residuos cercanos a los máximos locales.

Por otro lado, se realizó el análisis de la *protein residue network* con $R_c = 6$ Å en donde se observa que los valores de la centralidad de cercanía e intermediación disminuyen al disminuir el radio de corte, mientras que los valores de centralidad de vector propio aumentan, sin embargo, la forma cualitativa de las gráficas no se modifica. Para este radio de corte se observa que las predicciones de los residuos funcionales mejoran al aumentar el radio de corte.

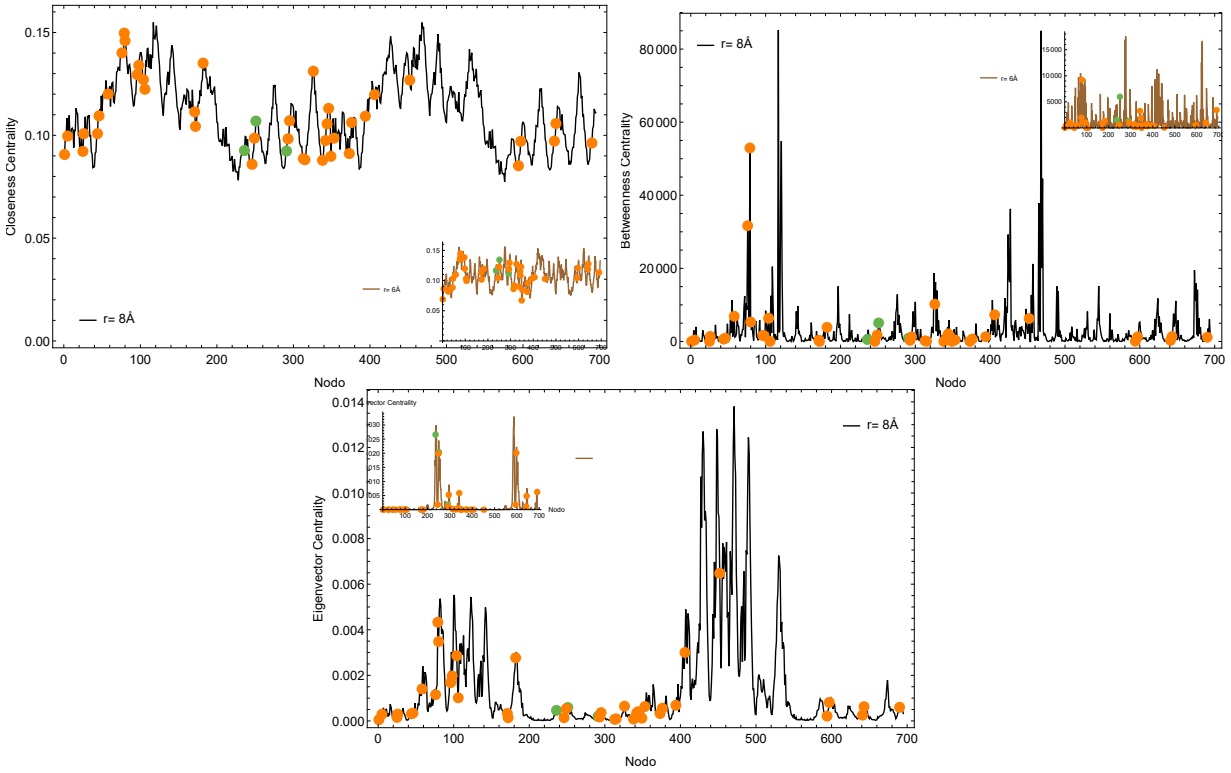
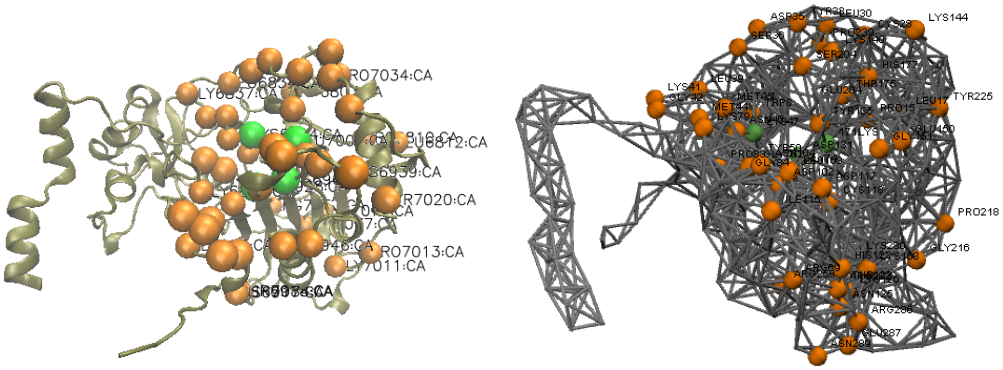


Figura 3.5: Gráficas de la centralidad de cercanía, intermediación y *eigenvector* de la proteína 6WXC. En todos los casos, la gráfica de líneas negras es para la *protein residue network* con $R_c = 8$ Å y la gráfica pequeña de líneas cafés ubicada en la parte inferior o superior es para $R_c = 6$ Å. Los residuos del sitio catalítico se marcan en color verde y los residuos del sitio de unión en color naranja. Figuras elaboradas con el programa *Wolfram Mathematica 11*.

3.3. Análisis de medidas de centralidad de la proteína 6WVN

La proteína no estructural NSP16 (6WVN) tiene la función de promover la proliferación viral, es decir, trabajar con NSP10 ocultando los genes virales de las defensas de la célula infectada [**Rimanshee et al., 2021**]. La figura (3.6) muestra la estructura de la proteína de acuerdo al archivo PDB y la estructura de la *protein residue network* con $R_c = 8$ Å.



(a) (b)

Figura 3.6: Construcción de red a partir de una proteína. En (a) se muestra la estructura de la proteína 6WVN de acuerdo al archivo PDB, figura elaborada con el programa Visual Molecular Dynamics (VMD). En (b) se observa la red construida a partir de los residuos C_α considerando un radio de corte de $R_c = 8$ Å. Los residuos del sitio catalítico se marcan en color verde y los residuos del sitio de unión en color naranja. Figura elaborada con el programa *Wolfram Mathematica 11.*

Esta proteína está formada por dos cadenas, la segunda es muy corta y no se tiene información de los residuos que forman el sitio activo. En la figura (3.7) se observa que para un radio de $R_c = 8$ Å los residuos $Asp133$ y $Thr173$ son máximos locales para la centralidad de cercanía, por lo que estos residuos se encuentran a una mínima distancia de los demás, mientras que el resto de residuos del sitio activo ocupan picos que se encuentan a un 14 % por arriba del promedio. En la centralidad de intermediación $Gly76$ no es máximo local, pero permanece cercano a ellos, mientras que el resto de los residuos que forman parte del sitio activo se ubican a un 20 % por arriba del promedio. Para la centralidad de *eigenvector* se observa que $Lys280$ y $Arg282$ no son máximos locales pero permanecen cercanos a ellos, el resto de residuos del sitio activo ocupan picos por arriba del 20 % del promedio por lo que están conectados a otros nodos que a su vez son muy relevantes. Para este radio de corte la centralidad de cercanía funciona mejor para ubicar residuos del sitio activo en la estructura

de la proteína, ya que hay más residuos cercanos al máximo global.

También, se realizó el análisis de la *protein residue network* con $R_c = 6$ Å y se observa que los resultados de la centralidad de cercanía son pequeños al disminuir el radio de corte, por el contrario, se tienen valores altos de centralidad de intermediación y de vector propio, sin embargo, la forma de las gráficas no se modifica. Para este radio de corte notamos que las predicciones de los residuos funcionales mejoran al aumentar el radio de corte.

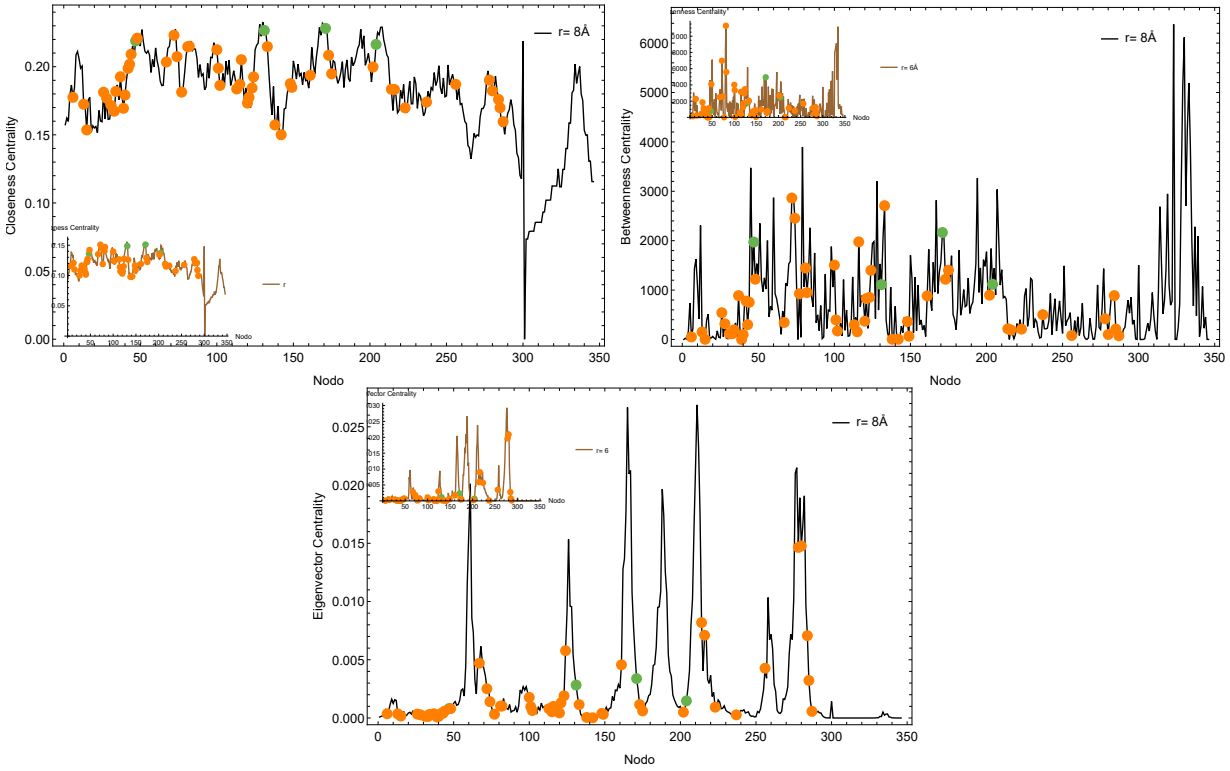
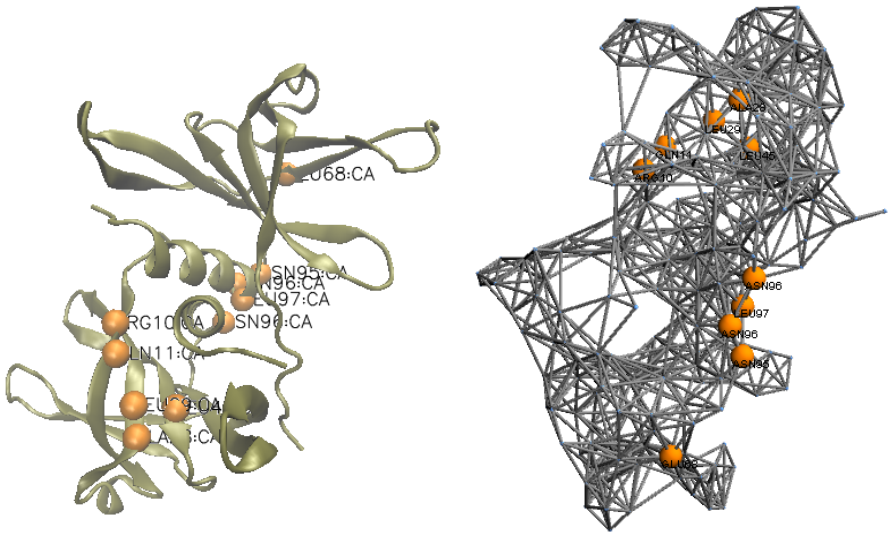


Figura 3.7: Gráficas de la centralidad de cercanía, intermediación y *eigenvector* de la proteína 6WVN. En todos los casos, la gráfica de líneas negras es para la *protein residue network* con $R_c = 8$ Å y la gráfica pequeña de líneas cafés ubicada en la parte inferior o superior es para $R_c = 6$ Å. Los residuos del sitio catalítico se marcan en color verde y los residuos del sitio de unión en color naranja. Figuras elaboradas con el programa *Wolfram Mathematica 11*.

3.4. Análisis de medidas de centralidad de la proteína 6WXD

La proteína no estructural NSP9 (6WXD) es responsable de producir pequeños canales en la membrana nuclear de la célula infectada [**Rimanshee et al., 2021**]. La figura (3.8)

muestra la estructura de la proteína de acuerdo al archivo PDB y la estructura de la *protein residue network* con $R_c = 8$ Å, en particular, esta proteína está formada por dos cadenas diferentes (se tienen aminoácidos´para la cadena A y para la cadena B).



(a) (b)

Figura 3.8: Construcción de red a partir de una proteína. En (a) se muestra la estructura de la proteína 6WXD de acuerdo al archivo PDB, figura elaborada con el programa Visual Molecular Dynamics (VMD). En (b) se observa la red construida a partir de los residuos C_α considerando un radio de corte de $R_c = 8$ Å. Los residuos del sitio de unión se marcan en color naranja. Figura elaborada con el programa *Wolfram Mathematica 11*.

La figura (3.9) ilustra los resultados para las medidas de centralidad en donde se observan valores extremos destacados para los residuos del sitio activo. En particular, cabe destacar que para la *protein residue network* con radio $R_c = 8$ Å la centralidad de cercanía si funciona para ubicar residuos funcionales en la estructura de la proteína, en este caso, $AsnA95$, $AsnA96$, $LeuA97$ y $AsnB96$ son máximos locales, además, se encuentran dentro de los 5 picos más altos. Para la centralidad de intermediación, $LeuA97$ es máximo global, $AsnA95$ y $AsnB96$ son máximos locales, estos valores están dentro de los 5 picos más altos y por los que pasan las trayectorias más cortas. Para los resultados de la centralidad de *eigenvector* $AsnA96$ es máximo global, mientras que $AsnA95$, $LeuA97$ $AsnB96$ son máximos locales y se encuentran dentro de los 5 picos más altos. Para este radio de corte las tres centralidades funcionan mejor para ubicar residuos del sitio activo en la estructura de la proteína, ya que hay más residuos destacados e incluso son máximo global.

En el análisis de la *protein residue network* con $R_c = 6$ Å se observa que los resultados de las tres medidas de centralidad son pequeños al disminuir el radio de corte. Para este radio de corte se encontró que las predicciones de los residuos funcionales mejoran al aumentar el radio de corte.

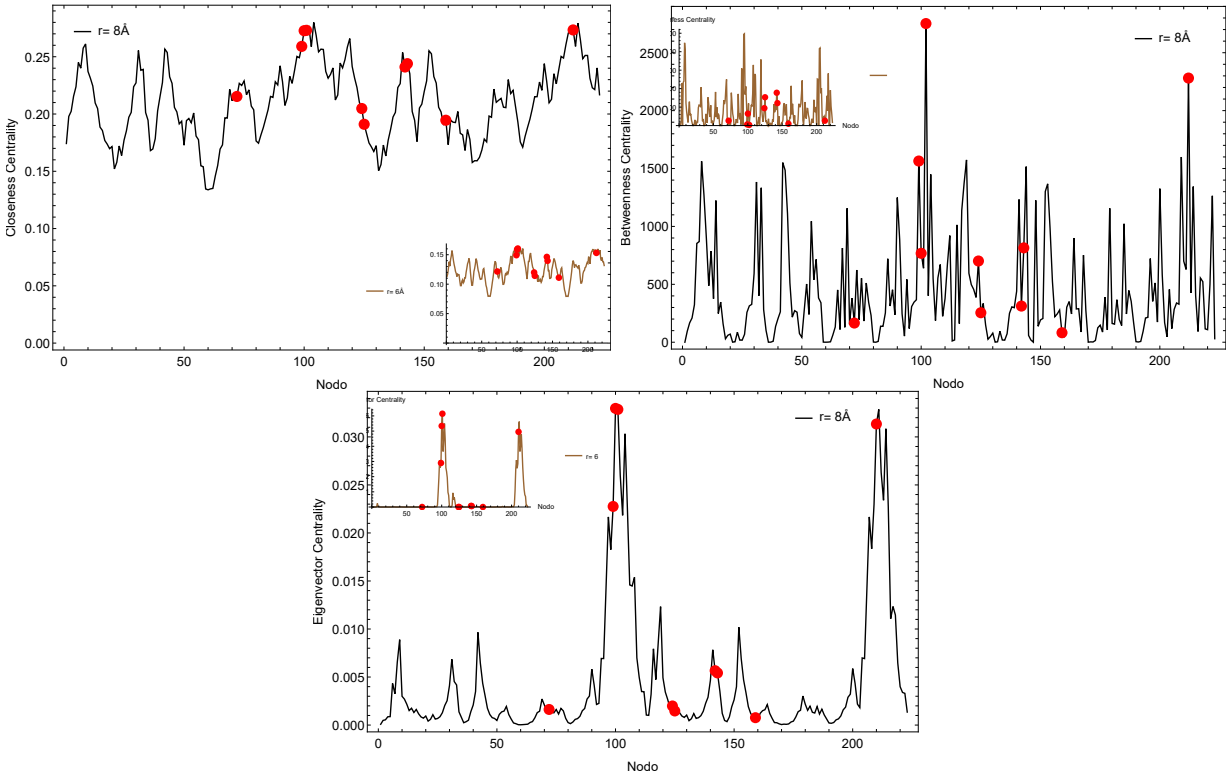


Figura 3.9: Gráficas de las medidas de centralidad de la proteína 6WXD. En todos los casos, la gráfica de líneas negras es para la *protein residue network* con $R_c = 8$ Å y la gráfica pequeña de líneas cafés ubicada en la parte inferior o superior es para $R_c = 6$ Å. Los residuos del sitio de unión se marcan en color naranja. Figuras elaboradas con el programa *Wolfram Mathematica 11*.

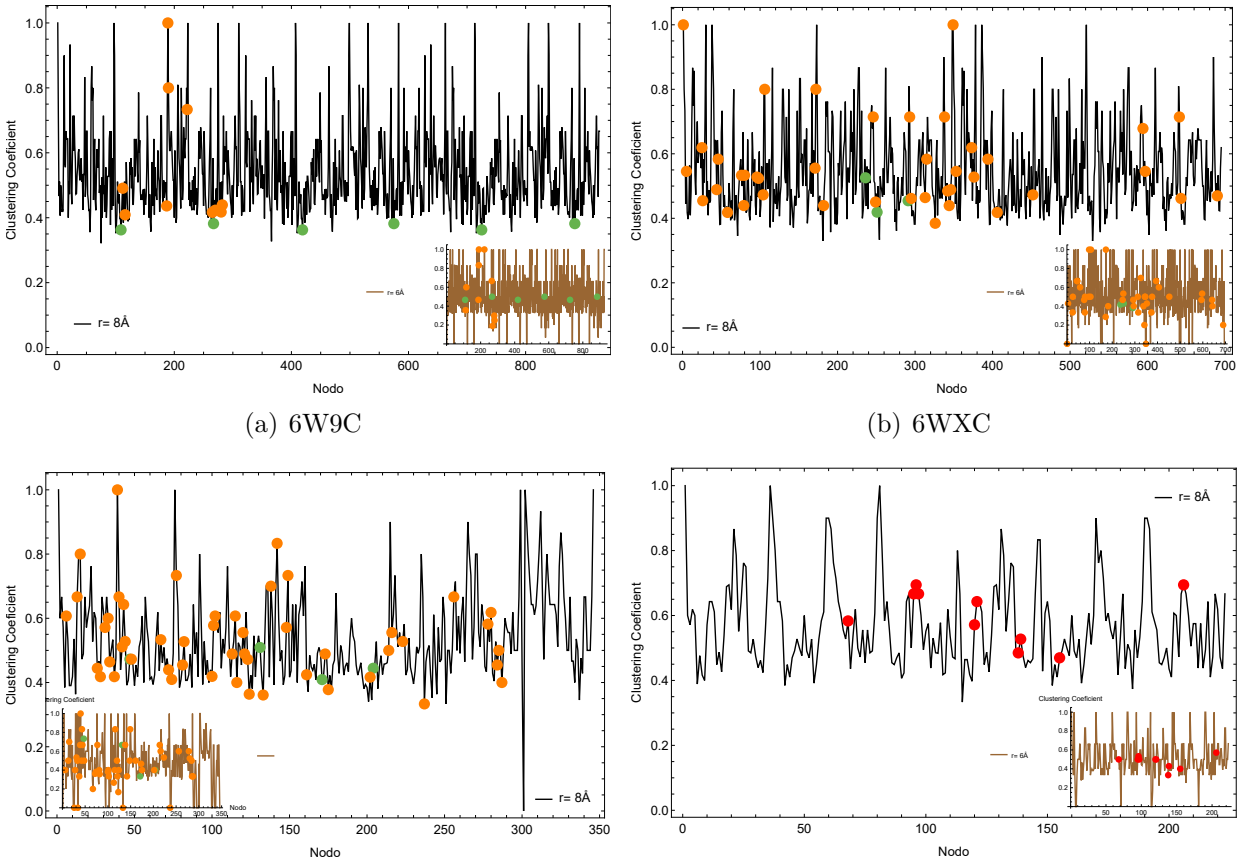
3.5. Análisis del coeficiente de agrupamiento $\mathbb{C}_i$ de las 4 proteínas

El coeficiente de agrupamiento $\mathbb{C}_i$ refleja la probabilidad de que los vecinos de un nodo también sean vecinos entre sí, y como tal, es una medida del orden local [**Watts & Strogatz, 1998**]. En la figura (3.10) se muestran los resultados del coeficiente de agrupamiento $\mathbb{C}_i$ para cada una de las proteínas de estudio. Para todos los casos, la gráfica de líneas negras es para

la *protein residue network* con $R_c = 8$ Å y la gráfica pequeña de líneas cafés ubicada en la parte inferior de cada figura es para $R_c = 6$ Å, asimismo, los residuos del sitio catalítico se marcan en color verde y los residuos del sitio de unión en color naranja.

Los resultados del coeficiente de agrupamiento para las *protein residue network* con $R_c = 8$ Å muestra que la mayoría de los residuos que forman el sitio activo son mínimos locales y ocupan picos que se encuentran por debajo del promedio, lo que significa que los vecinos de estos residuos no interactúan entre ellos, a excepción de algunos aminoácidos como se observa para la proteína 6W9C, $ThrA191$ tiene un valor alto de agrupamiento, así como $CysA192$ y $CysA224$ tienen vecinos altamente agrupados. En el caso de la proteína 6WXC se observa que $Trp1$ y $Lys368$ son valores destacados para esta medida, el residuo $Gly42$ de la proteína 6WVN tiene un valor alto para esta medida y por último, para la proteína 6WXD no se observan valores destacados para esta medida.

Algo similar ocurre con un radio de $R_c = 6$ Å, la mayoría de los residuos del sitio activo tienen valores mínimos de coeficiente de agrupamiento y se encuentran por debajo del promedio.



(a) 6W9C (b) 6WXC

(c) 6WVN (d) 6WXD

Figura 3.10: Gráficas del coeficiente de agrupamiento $\mathbb{C}_i$ de las proteínas 6W9C, 6WXC, 6WVN Y 6WXD. La gráfica de líneas negras es para la *protein residue network* con $R_c = 8$ Å y la gráfica pequeña de líneas cafés ubicada en la parte inferior es para $R_c = 6$ Å. Los residuos del sitio catalítico se marcan en color verde y los residuos del sitio de unión en color naranja. Figuras elaboradas con el programa *Wolfram Mathematica 11*.

Capítulo 4

Cálculo de profundidad del átomo y área de superficie accesible al solvente de proteínas relacionadas con el virus del SARS-CoV-2

En este capítulo se analiza las propiedades de la superficie de la proteína, como son: profundidad del átomo en donde se calcula la distancia entre el átomo (C_α) del i-ésimo residuo de la proteína, al átomo del solvente más cercano que encuentre en la superficie, además, se calcula SASA, que mide el área de superficie expuesta a moléculas de solvente. La interacción de los residuos es de importancia crítica para la función y estabilidad de una proteína, pero, también su posición y ubicación en la estructura.

4.1. 6W9C

Las interacciones entre los residuos y las conexiones de los residuos del sitio activo con aminoácidos polares e hidrofóbicos son de gran relevancia. De acuerdo a la clasificación de la tabla (1.1) en la figura (4.1) se muestra el grafo asociado a las conexiones entre los aminoácidos polares e hidrofóbicos (mapa HP) con los residuos del sitio activo para un radio de corte de $R_c = 8$ Å, se aprecia que la mayoría de aminoácidos hidrofóbicos tienden a agruparse en el interior de la estructura de la proteína para evitar el contacto con moléculas de agua, mientras que los aminoácidos polares se encuentran en el exterior de la estructura de la proteína donde tienen contacto con moléculas del solvente.

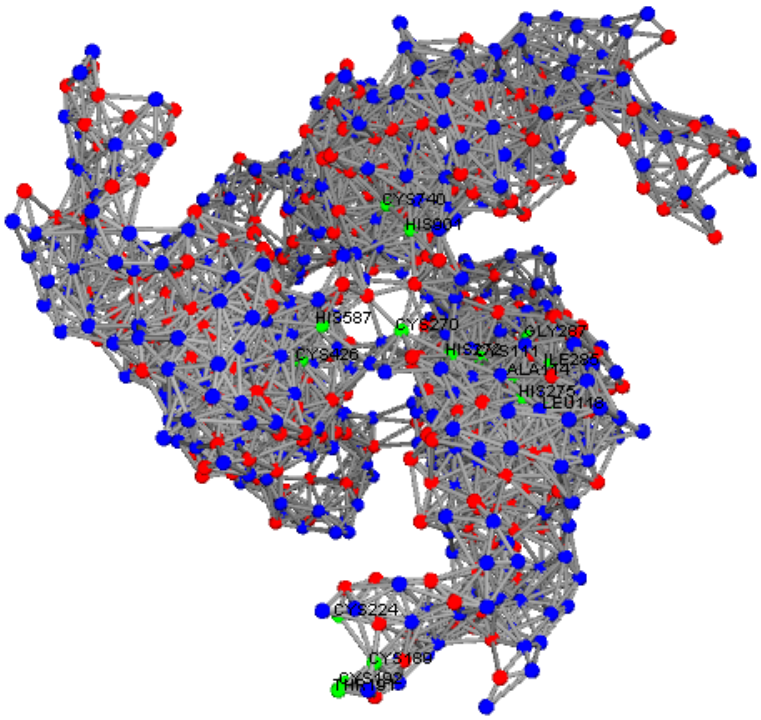
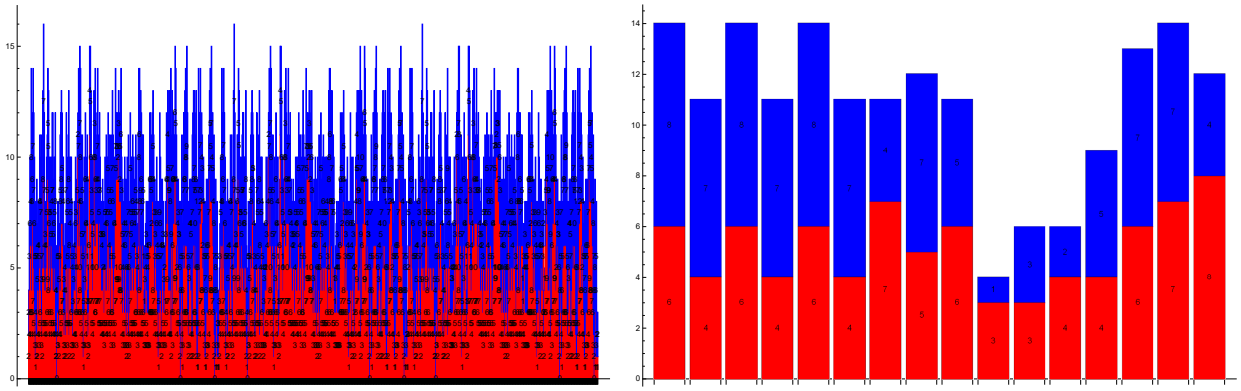


Figura 4.1: Grafo asociado al mapa de conexiones HP. En el grafo se identifican las conexiones entre los aminoácidos polares (nodos azules) y aminoácidos hidrofóbicos (nodos rojos), así como también los residuos del sitio activo (nodos verdes) de la proteína 6W9C. Figura realizada con el programa *Wolfram Mathematica 11*.

En la figura (4.2) se muestra el histograma del número de conexiones que tiene cada uno de los residuos C_α y residuos del sitio activo con aminoácidos hidrofóbicos y polares. Específicamente, los residuos del sitio catalítico (barra 1 - 6) están en promedio conectados con el 60.38 % de aminoácidos polares y el 39.61 % de aminoácidos hidrofóbicos, mientras que los residuos del sitio de unión (barra 7 - 16) están en promedio conectados con el 44.12 % de aminoácidos polares y el 55.87 % de aminoácidos hidrofóbicos, cabe mencionar que esto también depende de su ubicación dentro de la estructura de la proteína.

La ubicación de los residuos del sitio activo en la estructura de la proteína es primordial, es por eso que se hace uso de *DPX Protusion Index Web Server* para calcular la profundidad de cada residuo. En la figura (4.3) se ilustran los resultados de las distancias de cada residuo al átomo de la molécula de agua superficial más cercana, se observa que el 75 % de los residuos que forman el sitio activo tienen valores de profundidad cercana a cero, ya que se encuentran cerca de la superficie de la proteína. Por otra parte se muestran los resultados de SASA para cada residuo, con puntos verdes y naranjas se marca el área expuesta al solvente. Este resultado es interesante puesto que aunque podría esperarse que los residuos del sitio



(a) Histograma de conexión de cada residuo con aminoácidos polares e hidrofóbicos

(b) Histograma de conexión de cada residuo del sitio activo con aminoácidos polares e hidrofóbicos

Figura 4.2: Histogramas de conexión de residuos. En (a) se muestra el histograma del número de conexiones que tiene cada aminoácido C_α con los aminoácidos polares (azul) y aminoácidos hidrofóbicos (rojo), en (b) se muestra el número de conexiones que tiene cada residuo del sitio activo con los aminoácidos polares e hidrofóbicos. Figuras realizadas con el programa *Wolfram Mathematica 11*.

activo pudieran tener valores altos de SASA, este no es el caso. Por el contrario, la mayoría de los residuos del sitio activo están de alguna manera ocultos de la capa exterior de las moléculas del solvente, esto se debe a que no todos los residuos que forman parte del sitio activo están necesariamente en la superficie, algunos podrían tener actividad catalítica cerca del núcleo hidrofóbico y sólo el 25 % de residuos del sitio activo tienen valores altos de SASA, por lo que se encuentran en la parte exterior de la proteína.

Se analiza la correlación entre la profundidad de los residuos y el área de superficie accesible al solvente. Según los valores de SASA, los residuos o aminoácidos de una proteína se pueden clasificar como ocultos o expuestos a moléculas de agua y, de acuerdo a la profundidad de un átomo, esta es cero para todos los átomos accesibles al solvente y mayor que cero para los átomos que se ubican en el interior de la proteína, los átomos que se encuentran más profundamente tienen valores de profundidad más altos [**Carugo & Pongor**, **2002**].

En la figura (4.4) se muestra la gráfica de correlación entre la profundidad del átomo y SASA para la proteína 6W9C, en donde $ThrA191$, $CysA192$ y $CysA224$ son residuos que tienen valores altos de área de superficie accesible combinado con una profundidad cercana a cero, lo que significa que estos residuos se encuentran en el exterior de la proteína, a diferencia del resto que tienen valores pequeños de SASA combinado con valores altos de profundidad.

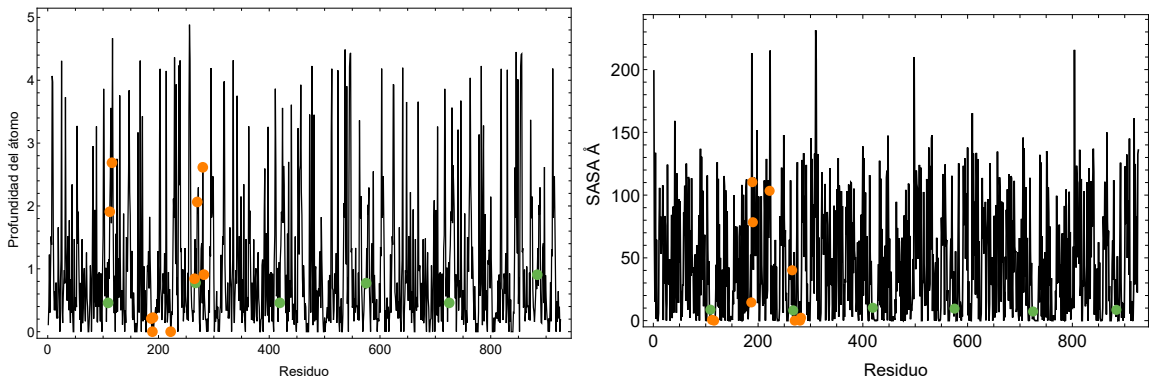


Figura 4.3: Gráficas de la profundidad del átomo y SASA. Valores de la profundidad del átomo para la proteína 6W9C (PDB) calculados por el *DPX Protusion Index Web Server*, así como los valores del área de superficie accesible al solvente calculados por el programa GetArea. Los residuos del sitio catalítico se marcan en color verde y los residuos del sitio de unión en color naranja. Figuras realizadas con el programa *Wolfram Mathematica 11*.

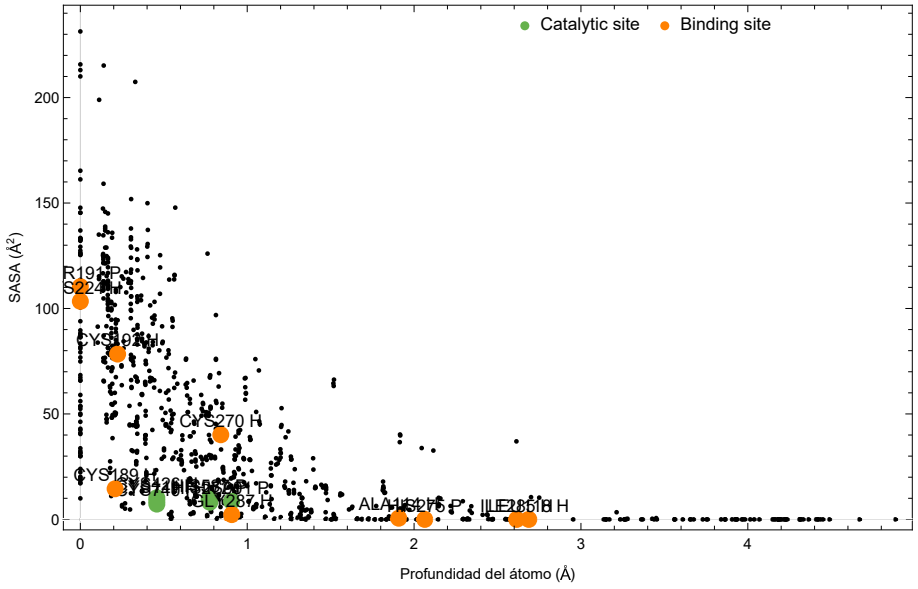


Figura 4.4: Correlación de la profundidad de átomo y el área de superficie accesible al solvente para la proteína 6W9C. Los residuos del sitio catalítico se marcan en color verde y los residuos del sitio de unión en color naranja. Figura realizada con el programa *Wolfram Mathematica 11*.

4.2. 6WXC

Para la *protein residue network* asociada a la proteína 6WXC y de acuerdo a la clasificación de la tabla (1.1), en la figura (4.5) se muestra el grafo asociado a las conexiones entre los aminoácidos polares e hidrofóbicos (mapa HP) con los residuos del sitio activo con un radio

de corte de $R_c = 8$ Å, las interacciones hidrofóbicas también son importantes, lo interesante es que éstos se agrupan juntos en el interior de la proteína, dejando a los aminoácidos polares en el exterior para interactuar con las moléculas de agua circundantes.

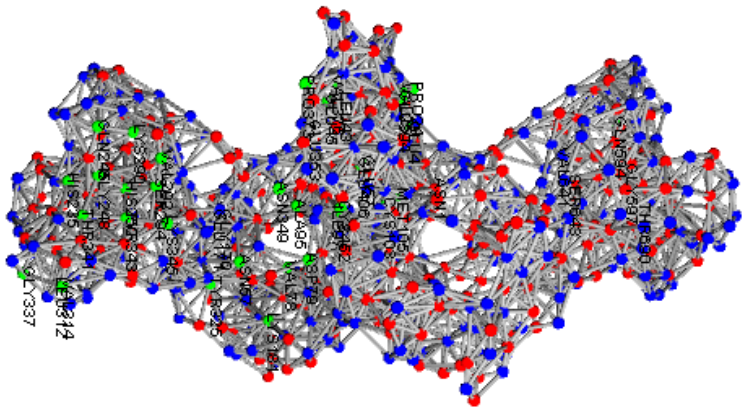
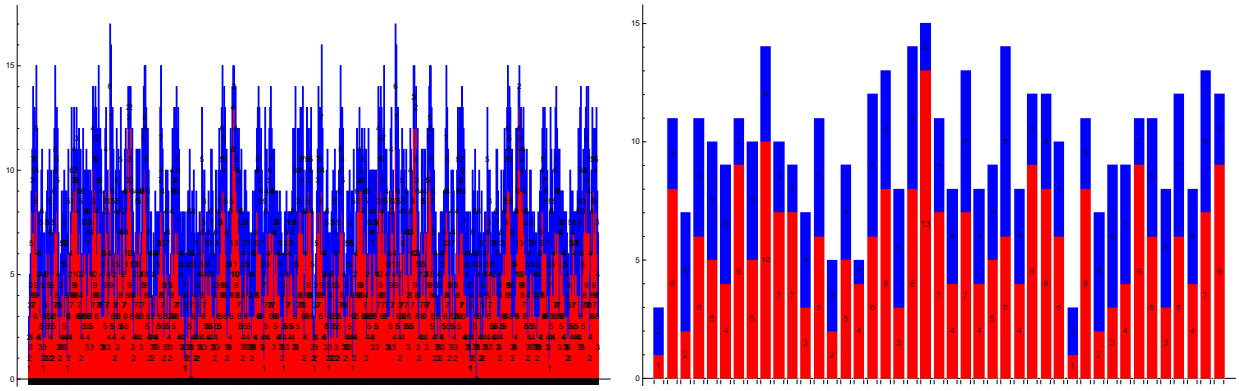


Figura 4.5: Grafo asociado al mapa de conexiones HP. En el grafo se muestran las conexiones entre los aminoácidos polares (nodos azules) y aminoácidos hidrofóbicos (nodos rojos), así como también los residuos del sitio activo (nodos verdes) de la proteína 6WXC. Figura realizada con el programa *Wolfram Mathematica 11*.

Por otra parte, nos interesa conocer el porcentaje de residuos hidrofóbicos y polares que se conectan con cada residuo C_α. En la figura (4.6) se muestra el histograma del número de conexiones que tiene cada uno de los aminoácidos y, en particular, el número de conexiones que tiene cada residuo que forma el sitio activo. Específicamente, los residuos del sitio catalítico (barra 1 - 3) están en promedio conectados con el 29.38 % de aminoácidos polares y el 70.61 % de aminoácidos hidrofóbicos, mientras que los residuos del sitio de unión (barra 4 - 40) están en promedio conectados con el 45.71 % de aminoácidos polares y el 54.28 % de aminoácidos hidrofóbicos.

Por otro lado, se calculó la profundidad de cada residuo como se describió anteriormente. En la figura (4.7) se ilustran los resultados de las distancias a las que se encuentra cada uno de los residuos a la molécula de agua más cercana que rodea a la proteína. Para este análisis, claramente se observa que el 80 % de los residuos que forman el sitio activo tienen valores de profundidad cero y cercanos a este valor, lo que implica que se sitúan cerca en la superficie de la proteína. Además, en la gráfica de la derecha se muestran los resultados obtenidos de SASA y se observa que los residuos del sitio activo se encuentren potencialmente en la región de mayor accesibilidad al solvente, puesto que tienen valores altos de SASA, lo que significa que se encuentran en la superficie y tienen contacto con moléculas de agua, el resto de resi-



(a) Histograma de conexión de cada residuo con aminoácidos polares e hidrofóbicos

(b) Histograma de conexión de cada residuo del sitio activo con aminoácidos polares e hidrofóbicos

Figura 4.6: Histograma de conexiones de residuos. En (a) se muestra el histograma del número de conexiones que tiene cada aminoácido C_α con los aminoácidos polares (azul) y aminoácidos hidrofóbicos (rojo), en particular, en (b) se muestra el número de conexiones que tiene cada residuo del sitio activo con los aminoácidos polares e hidrofóbicos. Figuras realizadas con el programa *Wolfram Mathematica 11*.

duos del sitio activo tienen valores pequeños de SASA y están ocultos de la capa exterior de las moléculas de agua.

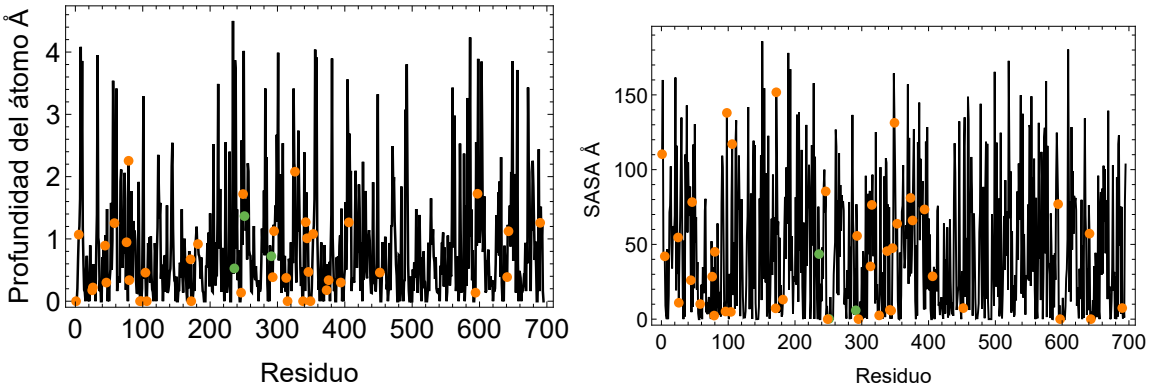


Figura 4.7: Gráficos de la profundidad del átomo y SASA. Valores de la profundidad de átomo para la proteína 6WXC (PDB) calculados por el *DPX Protusion Index Web Server*, así como los valores del área de superficie accesible al solvente calculados por el programa GetArea. Los residuos del sitio catalítico se marcan en color verde y los residuos del sitio de unión en color naranja. Figuras realizadas con el programa *Wolfram Mathematica 11*.

En la figura (4.8) se muestra la correlación de la profundidad del átomo vs SASA, para la proteína 6WXC, en donde se visualiza que el 85 % de residuos del sitio activo (puntos

verdes y naranjas) usualmente tienen valores altos de SASA combinado con una profundidad cercana a cero, lo que indica que los residuos del sitio activo se encuentran en la superficie de la proteína y tienen contacto con moléculas del solvente.

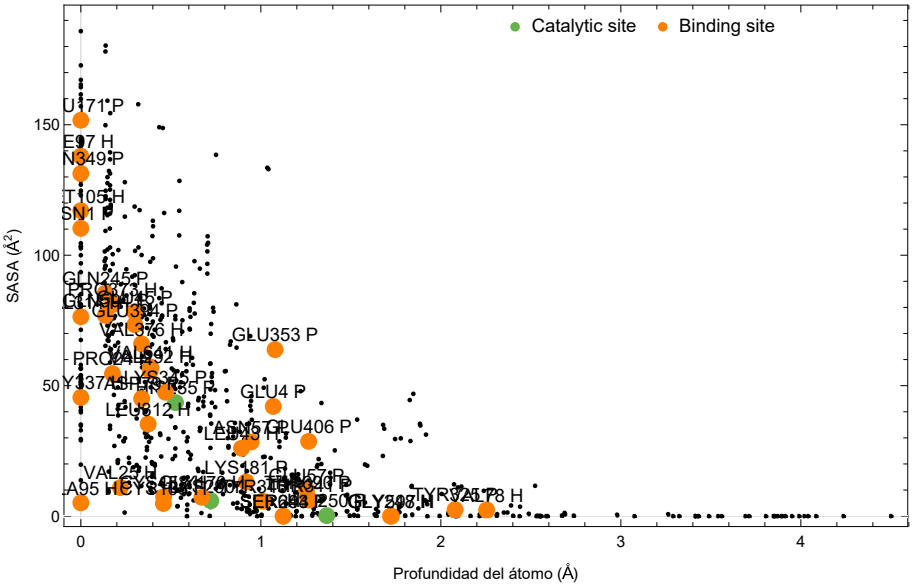


Figura 4.8: Correlación de la profundidad de átomo y el área de superficie accesible al solvente para la proteína 6WXC. Los residuos del sitio catalítico se marcan en color verde y los residuos del sitio de unión en color naranja. Figura realizada con el programa *Wolfram Mathematica 11*.

4.3. 6WVN

Calculamos el porcentaje de aminoácidos polares e hidrofóbicos que se conectan con los residuos del sitio activo dentro de la *protein residue network* con $R_c = 8$ Å de la proteína 6WVN. Basándonos en la tabla (1.1), marcamos en el grafo asociado a la *protein residue network* la conexión entre aminoácidos HP, en color rojo se marcan los aminoácidos hidrofóbicos, en color azul los polares y en verde los correspondientes al sitio activo como se observa en la figura (4.9), en donde los aminoácidos polares tienden a estar en el exterior de la proteína donde pueden interactuar con el agua, mientras que los aminoácidos hidrofóbicos se encuentran en el interior formando un núcleo hidrofóbico muy compacto de átomos que están ocultos del agua.

En la figura (4.10) se muestra el histograma del número de conexiones que tiene cada uno de los residuos C_α y en concreto, el número de conexiones que tiene cada residuo del

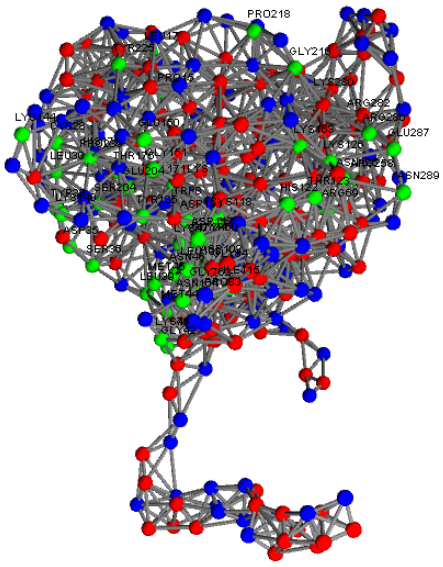
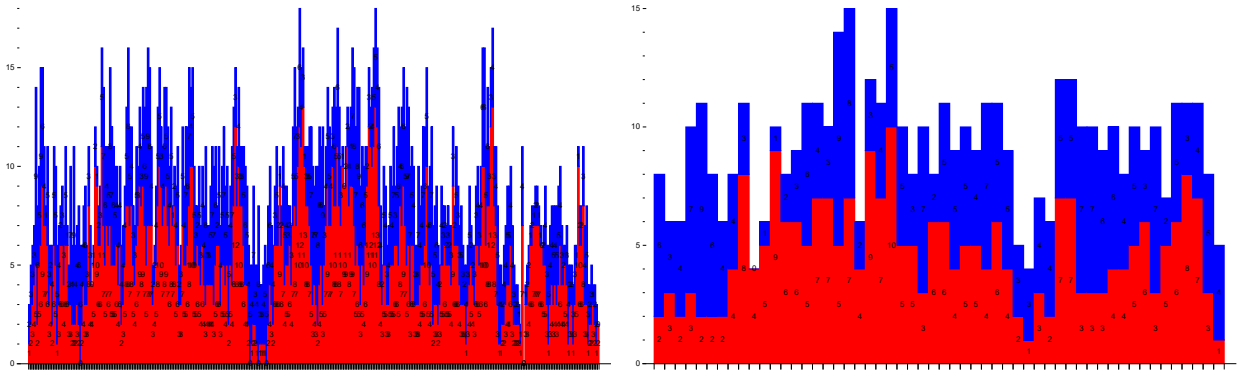


Figura 4.9: Grafo asociado al mapa de conexiones HP. En el grafo se identifican las conexiones entre los aminoácidos polares (nodos azules) y aminoácidos hidrofóbicos (nodos rojos), así como también los residuos del sitio activo (nodos verdes) de la proteína 6WVN. Figura realizada con el programa *Wolfram Mathematica 11*.

sitio activo con los aminoácidos polares e hidrofóbicos. Específicamente, los residuos del sitio catalítico (barra 1 - 4) están en promedio conectados con el 50.41 % de aminoácidos polares y el 49.58 % de aminoácidos hidrofóbicos, mientras que los residuos del sitio de unión (barra 5 - 54) están en promedio conectados con el 48.94 % de aminoácidos polares y el 51.05 % de aminoácidos hidrofóbicos.

Por otra parte, se tienen los valores de la profundidad de cada residuo, en la figura (4.11) se ilustran los resultados en donde se observa que el 87 % de los residuos que forman el sitio activo tienen valores de profundidad cero y cercanos a este, lo cual indica que estos residuos se localizan en el exterior de la proteína, el resto de residuos que forman el sitio activo se van acercando al núcleo hidrofóbico. Otro aspecto que es de nuestro interés se muestra en la gráfica de la derecha, donde se observa los valores de SASA y el 70 % de residuos del sitio activo se encuentran potencialmente en la región de mayor accesibilidad al solvente, lo que significa que se encuentran en la superficie de la proteína, mientras que el resto de residuos del sitio activo tienen valores pequeños de SASA dado que pueden tener actividad catalítica cerca del núcleo hidrofóbico de la proteína.

En la figura (4.12) se muestra la correlación de la profundidad del átomo vs SASA, para la proteína 6WVN, para este caso se observa que el 78 % de residuos del sitio activo (puntos



(a) Histograma de conexión de cada residuo con aminoácidos polares e hidrofóbicos

(b) Histograma de conexión de cada residuo del sitio activo con aminoácidos polares e hidrofóbicos

Figura 4.10: Histograma de conexiones de residuos. En (a) se muestra el histograma del número de conexiones que tiene cada aminoácido C_α con los aminoácidos polares (azul) y aminoácidos hidrofóbicos (rojo), en particular, en (b) se observa el número de conexiones que tiene cada residuo del sitio activo con los aminoácidos polares e hidrofóbicos. Figuras realizadas con el programa *Wolfram Mathematica 11*.

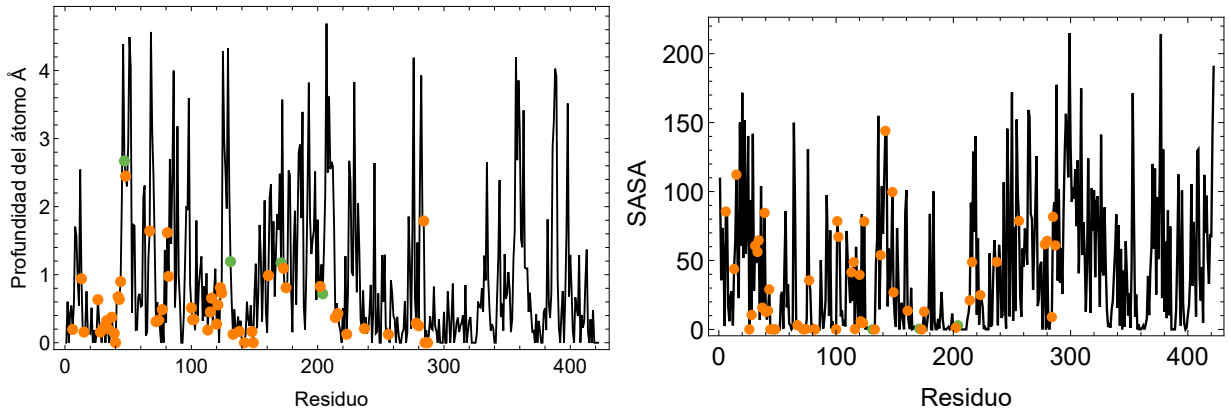


Figura 4.11: Gráficas de la profundidad del átomo y SASA. Valores de la profundidad del átomo para la proteína 6WVN (PDB) calculados por el *DPX Protusion Index Web Server*, así como los valores del área de superficie accesible al solvente calculados por el programa GetArea. Los residuos del sitio catalítico se marcan en color verde y los residuos del sitio de unión en color naranja. Figuras realizadas con el programa *Wolfram Mathematica 11*.

verdes y naranjas) usualmente tienen un valor grande de SASA combinado con una profundidad cercana a cero, lo que significa que estos se encuentran en el exterior de la proteína y tienen contacto con moléculas de agua. Aunque, por otro lado, se observa que el resto de los residuos del sitio activo tienen valores pequeños de SASA combinado con un valor alto de profundidad, lo que indica que los residuos del sitio activo están ocultos de la capa exterior

de la proteína.

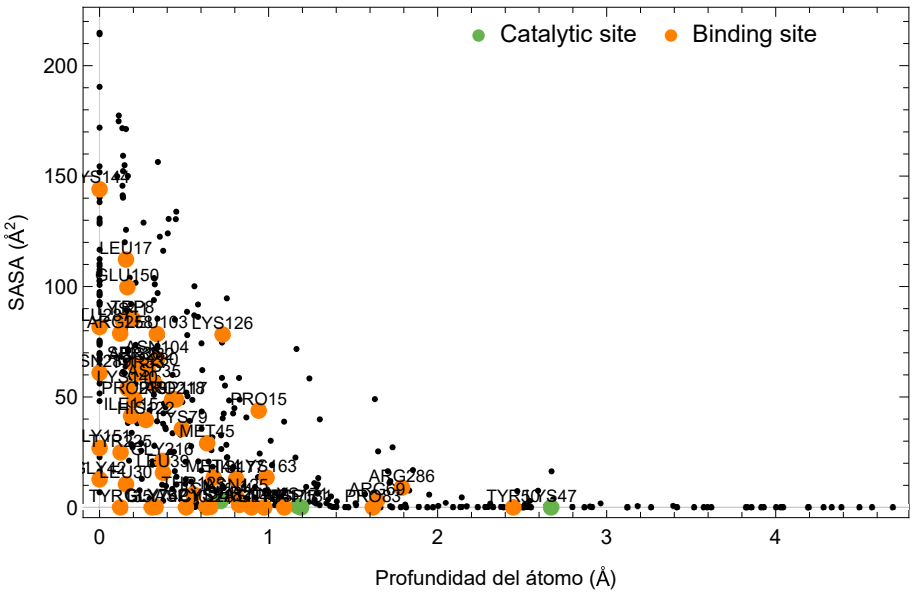


Figura 4.12: Correlación de la profundidad de átomo y el área de superficie accesible al solvente para la proteína 6WVN. Los residuos del sitio catalítico se marcan en color verde y los residuos del sitio de unión en color naranja. Figura realizada con el programa *Wolfram Mathematica 11*.

4.4. 6WXD

Nos interesa conocer las interacciones entre los residuos y, en particular, con que tipo de aminoácidos interactúan los residuos que forman el sitio activo, de acuerdo a la clasificación de la tabla (1.1), en la figura (4.13) se muestra el grafo asociado a las conexiones entre los aminoácidos polares e hidrofóbicos (mapa HP), en color rojo se marcan los aminoácidos hidrofóbicos, en color azul los polares y en verde los correspondientes al sitio activo, esto para un radio de corte de $R_c = 8$ Å. En este caso se observa que los aminoácidos hidrofóbicos se encuentran en el interior de la estructura de la proteína para evitar el contacto con moléculas de agua, mientras que los aminoácidos polares se encuentran en el exterior de la estructura de la proteína donde tienen contacto con moléculas del solvente.

Por otro lado, nos interesa conocer con qué tipo de aminoácidos se conectan los residuos que forman el sitio activo. En la figura (4.14) se observa que los residuos del sitio activo están rodeados por el 51 % de aminoácidos hidrofóbicos y el 49 % de aminoácidos polares, esto da pie para considerar del porqué los residuos del sitio activo tienen pequeños coeficientes de

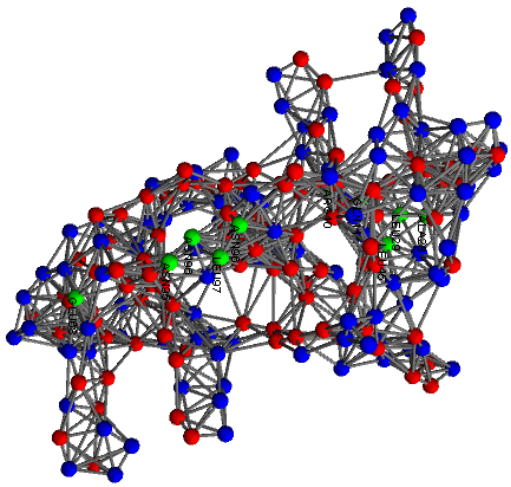
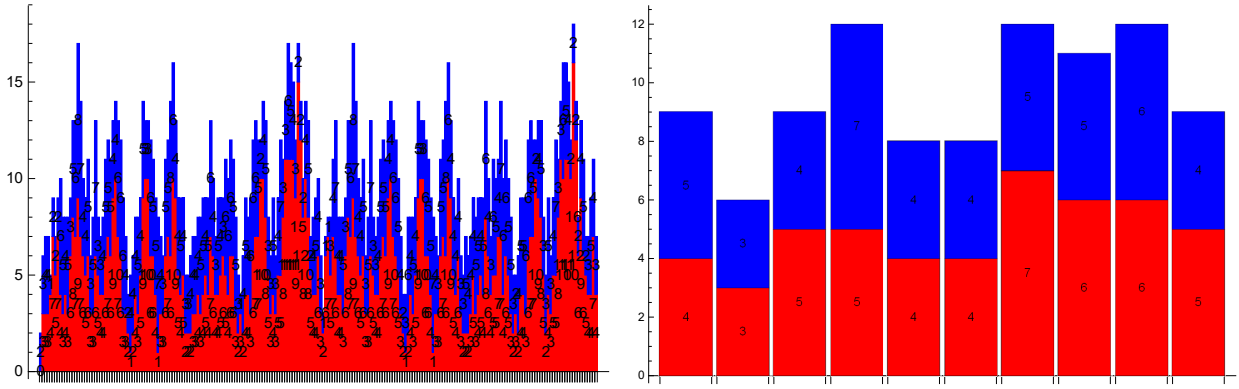


Figura 4.13: Grafo asociado al mapa de conexiones HP. En el grafo se identifican las conexiones entre los aminoácidos polares (nodos azules) y aminoácidos hidrofóbicos (nodos rojos), así como también los residuos del sitio activo (nodos verdes) de la proteína 6WXD. Figura realizada con el programa *Wolfram Mathematica 11*.

agrupamiento, por lo que se debe de estudiar la ubicación de los vecinos con respecto a su profundidad.



(a) Histograma de conexión de cada residuo con aminoácidos polares e hidrofóbicos

(b) Histograma de conexión de cada residuo del sitio activo con aminoácidos polares e hidrofóbicos

Figura 4.14: Histograma de conexión de residuos. En (a) se muestra el histograma del número de conexiones que tiene cada aminoácido C_α con los aminoácidos polares (azul) y aminoácidos hidrofóbicos (rojo), en (b) se muestra el número de conexiones que tiene cada residuo del sitio activo con los aminoácidos polares e hidrofóbicos. Figuras realizadas con el programa *Wolfram Mathematica 11*.

Se hace uso de *DPX Protusion Index Web Server* para calcular la profundidad de cada

residuo. En la figura (4.15) se ilustran los resultados de las distancias de cada residuo al átomo de la molécula de agua superficial más cercana. En este caso se observa que la mayoria de los residuos que forman el sitio activo tienen valores de profundidad cercana a cero, lo cual significa que se encuentran cerca de la superficie de la proteína. En la figura anterior también se muestra el valor de SASA para cada residuo, con puntos verdes y naranjas se marcan la ubicación de los residuos que forman parte del sitio activo. Este resultado es interesante, puesto que en este caso la mayoría de los residuos del sitio activo tienen valores altos de SASA, por lo que se encuentran en la parte exterior de la proteína.

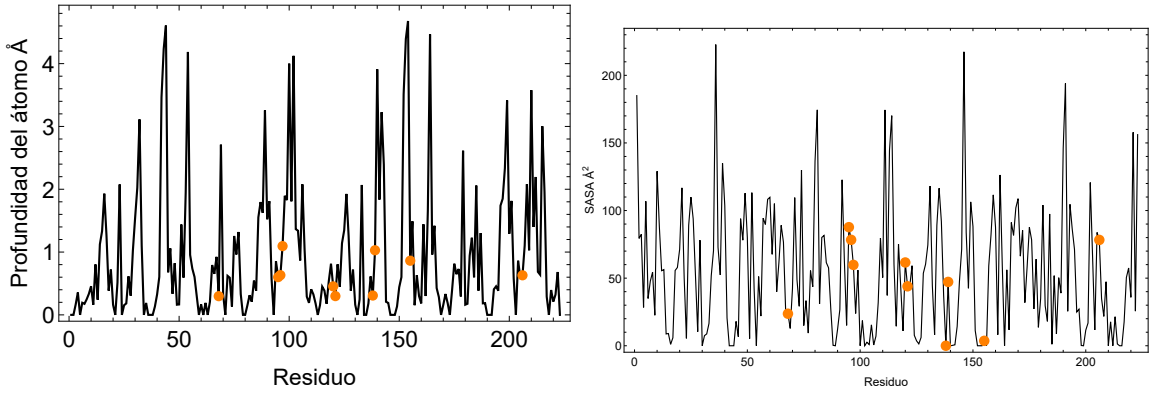


Figura 4.15: Gráficas de la profundidad del átomo y SASA. Valores de la profundidad del átomo para la proteína 6WXD (PDB) calculados por el *DPX Protusion Index Web Server*, asimismo, se tienen los valores del área de superficie accesible al solvente calculados por el programa GetArea. Los residuos del sitio de unión se marcan en color naranja. Figuras realizadas con el programa *Wolfram Mathematica 11*.

Los residuos catalíticos y de unión por lo regular están ubicados sobre la superficie de la proteína donde la catálisis de la reacción tiene lugar, por lo que preferentemente se podría esperar que los residuos del sitio activo pudieran tener valores altos de SASA y profundidad cero, en este caso, la figura (4.16) muestra que los residuos del sitio activo están ubicados en el exterior de la proteína, puesto que se localizan a una mínima distancia de la superficie y se encuentran en la zona de mayor accesibilidad al solvente. Al mismo tiempo, los datos de correlación encajan bien a una función potencial como veremos en el siguiente capítulo. La accesibilidad al solvente disminuye abruptamente a una profundidad de aproximadamente 1 Å. En este rango de profundidad, los residuos con profundidades similares pueden diferir en accesibilidad en aproximadamente un 15 %. Los residuos más profundos que 2 Å tienen cero accesibilidad, pero difieren en profundidad.

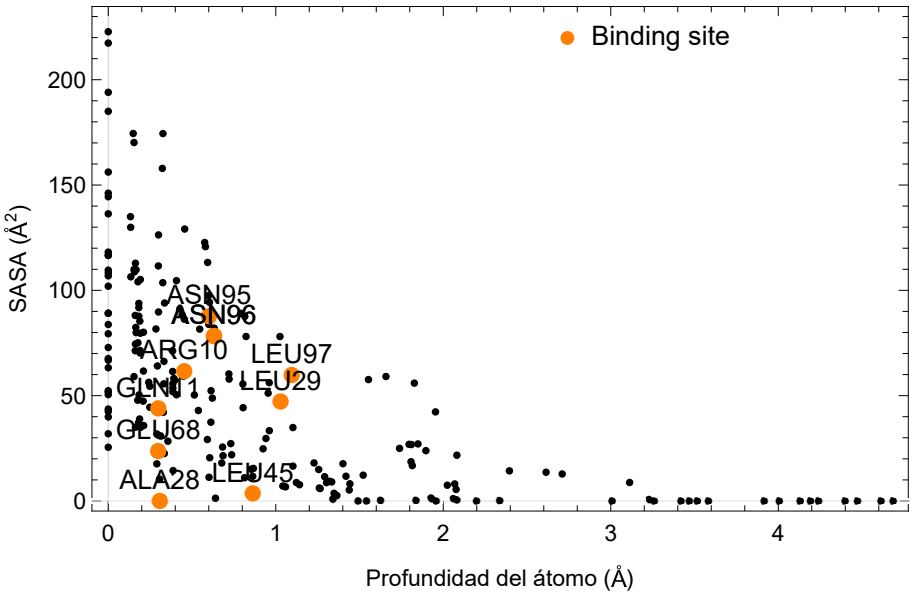


Figura 4.16: Correlación de la profundidad de átomo y el área de superficie accesible al solvente para la proteína 6WXD. Los residuos del sitio de unión están marcados en color naranja. Figura realizada con el programa *Wolfram Mathematica 11*.

Capítulo 5

Análisis y discusión de resultados

Como se expuso en el capítulo 3, a partir de la representación de la estructura tridimensional de una proteína como una red, se calculó el coeficiente de agrupamiento $\mathbb{C}_i$ y medidas de centralidad para el estudio de los residuos del sitio activo de las proteínas del virus SARS-CoV-2 y en la mayoría de nuestros resultados observamos la relación de estos residuos con valores extremos o mínimos locales.

A continuación se muestra un análisis de la comparación de los resultados promedios del coeficiente de agrupamiento y medidas de centralidad de los residuos del sitio activo y del resto de residuos de las proteínas de nuestro estudio. En la figura (5.1) se observa que los residuos del sitio activo de las proteínas de estudio con ID PDB: 6W9C, 6WXC, 6WVN y 6WXD, tienen en promedio valores pequeños de coeficiente de agrupamiento lo que significa que los vecinos de éstos nodos no son vecinos entre sí.

Asimismo, se observa que aunque los vecinos de los residuos del sitio activo no están bien agrupados, están conectados de forma más eficiente con nodos que a su vez son muy relevantes ya que poseen valores de centralidad de vector propio por arriba del promedio del resto de residuos, además, tienen en promedio valores altos y destacados de cercanía, lo cual indica que los residuos del sitio activo de estas proteínas se encuentran a una mínima distancia de los demás, es por eso que están dentro de las rutas más cortas y hacen el papel de puentes una mayor cantidad de veces. Los residuos del sitio activo de estas proteínas tienen en promedio valores más altos en todas las medidas de centralidad realizadas a comparación del resto de residuos.

Por otra parte, se calculó el promedio de los valores de SASA de los residuos del sitio activo y del resto de residuos de las proteínas consideradas en este trabajo de investigación. SASA es una cantidad física que mide el área de superficie expuesta a moléculas de solven-

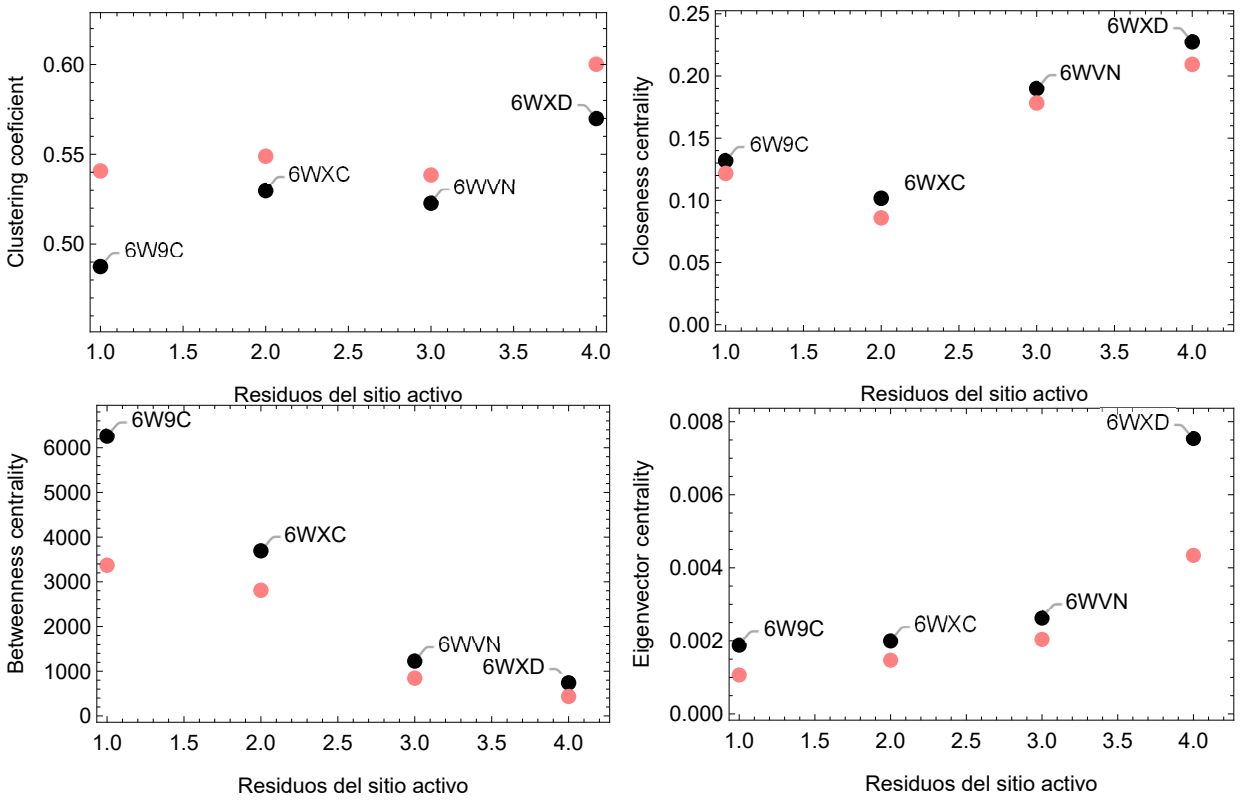
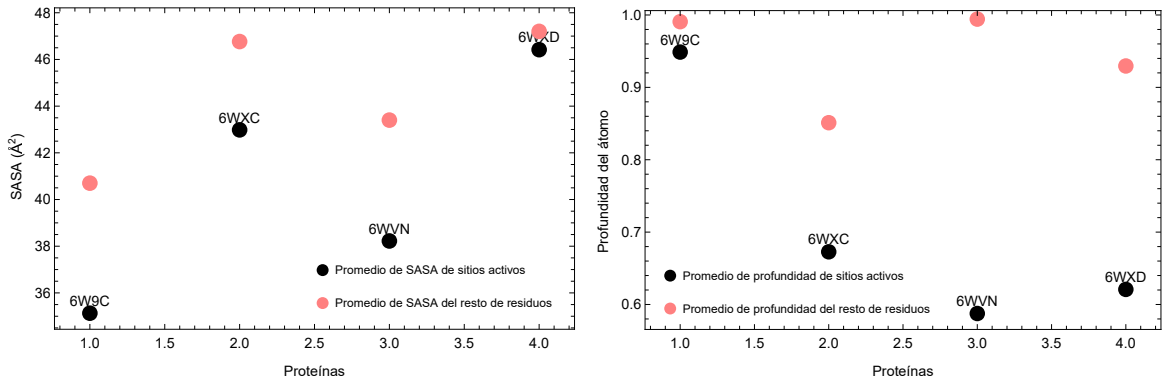


Figura 5.1: Gráficas de coeficiente de agrupamiento y medidas de centralidad. Resultados promedios del coeficiente de agrupación y medidas de centralidad de residuos del sitio activo (puntos negros) y del resto de residuos (puntos rosados) de proteínas relacionadas con el virus del SARS-CoV-2

te (moléculas de agua). En (a) de la figura (5.2) se muestran los resultados de los valores promedios de SASA de los residuos del sitio activo y del resto de residuos de cada proteína, es interesante observar que los residuos del sitio activo tienen valores altos de SASA pero que se encuentran por debajo de los resultados de los demás, esto se debe a que no todos los residuos que forman el sitio activo están necesariamente en la superficie, algunos podrían tener actividad catalítica cerca del núcleo hidrofóbico.

Por otro lado, estudiamos la profundidad del átomo, que es definida como la distancia dada en Å de un átomo (C_α) de la proteína al átomo del solvente más cercano que rodea la proteína, los residuos expuestos al solvente generalmente tienen valores de profundidad bajos, mientras que los residuos menos expuestos son más profundos en la proteína. En (b) de la figura (5.2) se muestran los resultados promedios de la profundidad de los residuos que forman parte del sitio activo y del resto de residuos, este resultado muestra que los residuos que forman parte del sitio activo evidentemente están más expuestos a las molécu-

las de agua circundantes por lo que se ubican en la superficie de la proteína, mientras que el resto de residuos que no forman parte del sitio activo se sitúan cerca del núcleo hidrofóbico.



(a) (b)

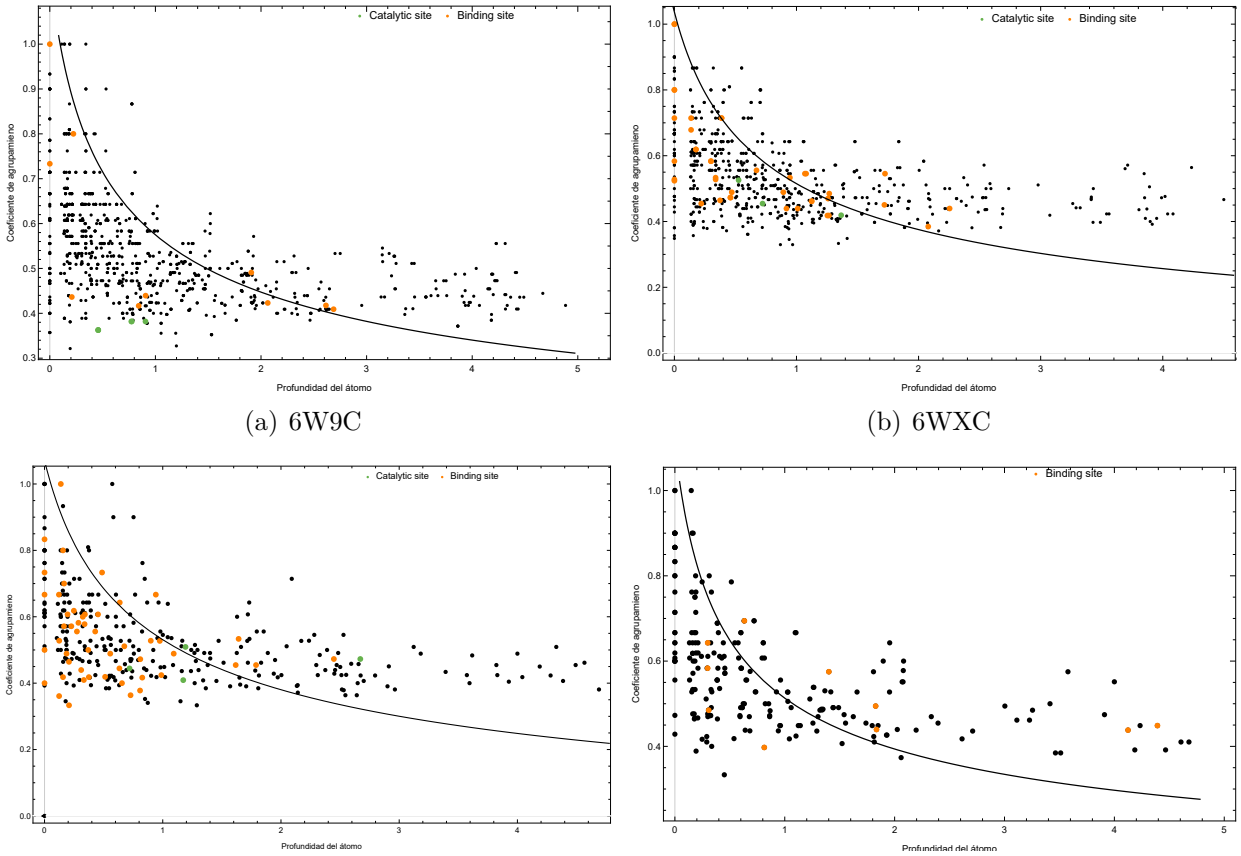
Figura 5.2: Gráficas de profundidad del átomo y SASA. Resultados promedios de la profundidad de átomo y área de superficie accesible al solvente de residuos del sitio activo y el resto de residuos de la proteína. Figuras realizadas con el programa *Wolfram Mathematica 11*.

También, se ha investigado la distribución de conectividad de los residuos del núcleo y la superficie. Anteriormente utilizamos el programa DEPTH que brinda información de la ubicación de los residuos en la proteína.

En la figura (5.3) se muestran los resultados de la correlación de profundidad de átomo y coeficiente de agrupamiento $\mathbb{C}_i$, un resultado interesante es que los residuos del sitio activo que se encuentran en la superficie de la proteína (profundidad cero) son aquellos que tienen valores altos de agrupamiento, mientras que aquellos que se encuentran cercanos al núcleo, tienen valores pequeños de agrupamiento.

Parece ser que para los residuos superficiales que están expuestos al solvente y en donde practicamente están rodeados por pocos residuos, el valor del coeficiente de agrupamiento depende de la ubicación y es algo superior a 0.509. Mientras que a mayores profundidades, donde los residuos están completamente rodeados por otros residuos y no están expuestos al solvente, la organización local de la proteína tiende siempre al mismo coeficiente de agrupamiento. Por otro lado, a profundidades mayores de 1 Å el coeficiente de agrupamiento $\mathbb{C}_i$ alcanza un valor promedio de 0.509 independientemente de la ubicación del residuo.

Asimismo, en la misma figura (5.3) se muestra para cada proteína una curva que posiblemente cumpla una función de ajuste. En este caso, la función de ajuste es una función



(a) 6W9C (b) 6WXC

(c) 6WVN (d) 6WXD

Figura 5.3: Gráficas de correlación de la profundidad y coeficiente de agrupamiento. Resultados de la correlación entre la profundidad del átomo y el coeficiente de agrupamiento $\mathbb{C}_i$ de proteínas relacionadas con el virus del SARS-CoV-2. Los residuos del sitio catalítico se marcan en color verde y los residuos del sitio de unión en color naranja. Figuras realizadas con el programa *Wolfram Mathematica 11*.

potencial del tipo $y = ax^b$ en donde los paramétros a y b para cada proteína no varían, los resultados se muestran en la tabla (5.1).

En la figura (5.4) se muestra un resultado que es parecido al que encontramos, se trata del artículo *Small-World Communication of Residues and Significance for Protein Dynamics* [**Atilgan et al.**, **2004**] en donde se estudió la correlación de la profundidad y el coeficiente de agrupamiento para proteínas de tamaños fijos (N = 150: cuadrados, 24 proteínas; N = 210: triángulos, 15 proteínas; N = 310: círculos, 15 proteínas), en donde a profundidades mayores a 4 Å, el coeficiente de agrupamiento alcanza un valor fijo de aproximadamente 0.35 independientemente del tamaño del sistema y la ubicación del residuo. Incluso para los residuos superficiales, el coeficiente de agrupamiento es independiente del tamaño del sistema,

	Parámetro a	Parámetro b
6W9C	0.60	0.40
6WXC	0.52	0.47
6WVN	0.50	0.42
6WXD	0.55	0.45

Cuadro 5.1: Resultados de los parámetros de la función potencial

aunque su valor depende de la ubicación y es algo superior a 0.35.

Quizás mucho menos esperado, es que el coeficiente de agrupamiento se acerque a un valor fijo más allá de una cierta profundidad independientemente del tamaño de las proteínas estudiadas. Como bien mencionamos anteriormente, a mayor profundidad, donde los residuos están completamente rodeados por otros residuos y están ocultos de las moléculas de agua, la organización local de la proteína es siempre la misma.

5.1. Predicción de residuos del sitio activo usando las metodologías previas

El análisis de redes complejas es un área emergente de interés en la disciplina de la ciencia y corresponde al análisis de redes complejas del mundo real desde el punto de vista de la teoría de grafos. Entre las diversas métricas utilizadas para el análisis de redes complejas, las medidas de centralidad de nodos es una métrica utilizada prominentemente de inmenso interés teórico y valor práctico. En nuestro caso, las aplicaciones para las métricas de centralidad, es ubicar residuos del sitio activo que es en donde se llevan a cabo las funciones catalíticas de la proteína. Los resultados muestran que existe una relación entre los residuos funcionales y las medidas de centralidad, por lo que se cree que puede ser de utilidad encontrar con esta metodología residuos con características de residuos del sitio activo y que además, se podrían complementar estos resultados con el estudio del interior y superficie de la proteína.

Se ha encontrado un uso generalizado de métodos de clasificación basados en redes (es decir, análisis de centralidad) para predecir sitios funcionales. En este estudio, trabajamos con cuatro proteínas relacionadas con el virus del SARS-CoV-2, posteriormente, se utilizaron 3 medidas de centralidad para la priorización de los nodos en todas las proteínas. Estas medidas de centralidad ayudan a extraer características relevantes y las relaciones correspondientes para cuantificar la conectividad en las redes biológicas. En otras palabras, el análisis de los componentes principales aclara qué medidas tienen los valores de contribución más altos, es

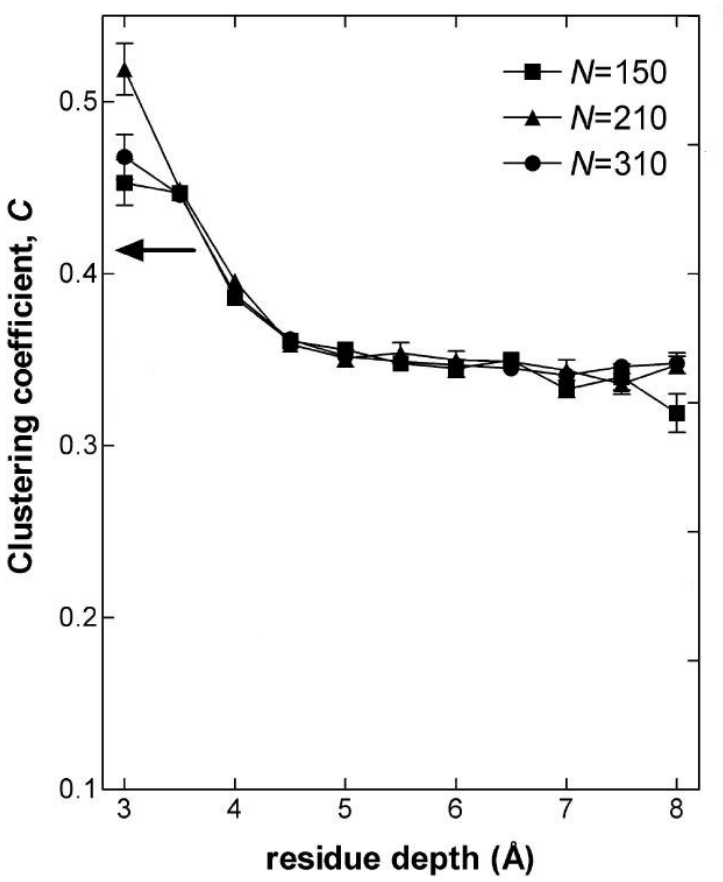


Figura 5.4: Coeficiente de agrupamiento $\mathbb{C}_i$ en función de la profundidad. Figura extraída del artículo *Small-World Communication of Residues and Significance for Protein Dynamics* [**Atilgan et al.**, **2004**].

decir, qué medidas comprenden mucha más información sobre la centralidad.

Un caso especial es el análisis de la proteína 6WXD que se estudió en el capítulo cuatro. Es notorio que de las medidas de centralidad, la de cercanía ubica una mayor cantidad de residuos funcionales y que además, para las otros medidas de centralidad se tienen valores extremos destacados, pero, no están reportados experimentalmente como parte de los residuos del sitio activo, sin embargo, este resultado puede representar un punto de inicio para analizar con otras técnicas, la opción de que esas regiones tengan alguna participación activa en la función de la proteína. En particular, cabe destacar que para la *protein residue network* con radio $R_c = 8$ Å se aprecia que la centralidad de cercanía funciona mejor para ubicar residuos funcionales en la estructura de la proteína, en este caso, $LeuA97$ es máximo global, mientras que $AsnA95$, $AsnA96$ y $AsnB96$ son máximos locales. Para la centralidad de intermediación, $AsnA95$ es máximo global. En la mayoría de los casos estudiados, se encontraron

más residuos del sitio activo cercanos a los máximos globales y locales de la centralidad de cercanía e intermediación. Para los resultados de la centralidad de *eigenvector*, en la mayoría de los casos obtuvimos una menor similitud entre los máximos o mínimos de esta cantidad con los residuos que forman el sitio activo. Cabe mencionar que de nuestro estudio esta proteína es la más pequeña en cuanto a la cantidad de residuos que la componen (223 residuos), además, su sitio activo está formado únicamente por residuos de unión que están ubicados sobre la superficie de la proteína, pues de acuerdo a nuestros resultados anteriores, éstos se localizan a una mínima distancia de la superficie. También, su sitio activo está conectado con el 62 % de aminoácidos polares. Con esto se podría suponer que estás característicos hicieron que esta metodología funcionara.

5.2. Frecuencia de aparición de residuos del sitio activo en proteínas del virus SARS-CoV-2

Presentamos un análisis de la distribución de frecuencia observada en los residuos catalíticos en comparación con todos los residuos en el conjunto de datos de proteínas relacionadas con el virus SARS-CoV-2. Este análisis se realizó por otros autores en el artículo *Analysis of Catalytic Residues in Enzyme Active Sites* [**Bartlett et al.**, **2002**], en donde realizaron un estudio de los residuos directamente implicados en la catálisis en 178 proteínas con un total de 615 residuos catalíticos, lo que da a cada proteína un promedio de 3.5 residuos catalíticos, en nuestro caso, consideramos 4 proteínas con un total de 134 residuos catalíticos. Se analizaron los resultados de la distribución de frecuencia y para ese caso los resultados indicaron el predominio de un pequeño conjunto de residuos en catálisis. Se espera que esta información proporcione una mejor comprensión de los mecanismos moleculares implicados en la catálisis y una base para predecir residuos catalíticos en proteínas de función desconocida.

La Figura (5.5) muestra la distribución de frecuencias observada de los diferentes tipos de residuos catalíticos, en comparación con la de todos los residuos en el conjunto de datos. Cabe resaltar que las primeras dos barras de cada residuo corresponden a la distribución de frecuencia observada en proteínas del SARS-CoV-2 y las siguientes dos barras corresponden a los resultados de Bartlett. Para obtener la barra oscura y verde, de todas las proteínas que se estudiaron se cuentan por ejemplo cuantas histidinas hay del total de aminoácidos (cada proteína tiene un total de aminoácidos) y se obtiene un porcentaje. Para la barra gris y rosa, se cuentan todos los sitios catalíticos y se revisa que composición tienen, si son histidina, se cuentan cuantas están participando del total.

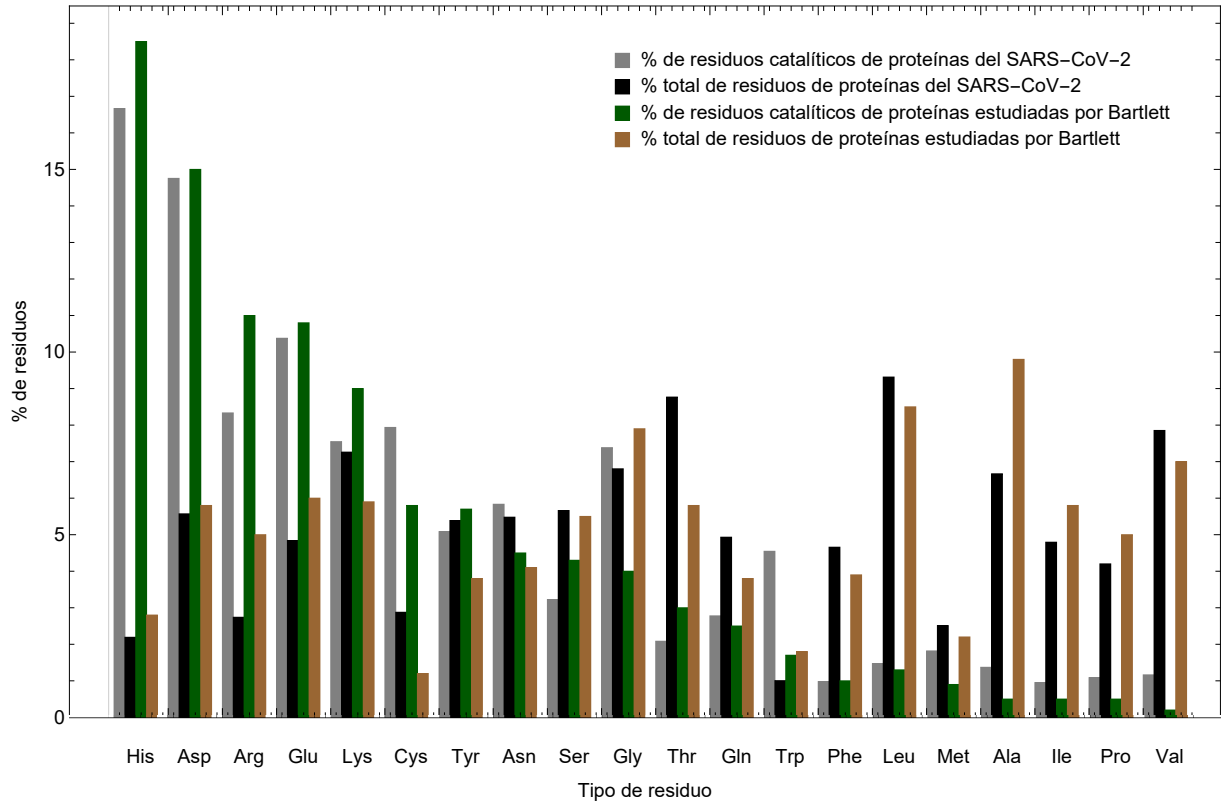


Figura 5.5: Distribución de frecuencias del tipo de residuo. Distribución de frecuencia de los diferentes tipos de residuos catalíticos del estudio de Bartlett en comparación con la de todos los residuos en el conjunto de datos de las 4 proteínas del virus del SARS-CoV-2. Figura realizada con el programa *Wolfram Mathematica 11*.

En nuestros resultados se observa que la histidina constituye el 16.7 % de todos los residuos catalíticos, aunque tiene un porcentaje global bajo (2.3 %). La histidina tiene versatilidad estructural y catalítica, puesto que está presente en el sitio activo de varias enzimas asociado a la transferencia de electrones. Los residuos de ácido aspártico y ácido glutámico constituyen el 15 % y 10.8 % de residuos catalíticos respectivamente. Su abundancia natural es casi idéntica (5.5 % y 5.1 %, respectivamente). Puede ser que el residuo aspartato se favorece ligeramente sobre el residuo de glutamato porque tiene una cadena lateral más corta por un grupo de metileno, haciendo que la cadena lateral sea menos flexible para que pueda mantenerse en su lugar, ayudando a la catálisis.

Este análisis se concentra en los residuos directamente involucrados en la catálisis. Las técnicas para ubicar residuos del sitio activo han mejorado, por lo tanto, es conveniente examinar los residuos implicados en la catálisis de proteínas del virus SARS-CoV-2 y el tipo de residuo que mayor predomina en la participación de la actividad catalítica. En nuestro

estudio obtuvimos la distribución de frecuencia del tipo de residuo y comparamos nuestros resultados con el de Bartlett. Obtuvimos resultados parecidos, además, sin importar el número de proteínas de estudio, tienen en promedio la misma frecuencia en los residuos y de participación en la actividad catalítica. Se espera que estos datos nos puedan ayudar a identificar posibles residuos catalíticos de la estructura y desarrollar herramientas para predecir el mecanismo a partir de la estructura. Estas herramientas son la base para predecir la función de las estructuras producidas por iniciativas de genómica estructural.

5.3. Correlación entre profundidad del átomo y SASA

Es conveniente volver a examinar la ubicación de los residuos catalíticos y de unión de las proteínas, es por ello que analizamos la correlación de profundidad de átomo y área expuesta al solvente de estos residuos. En la figura (5.6) se muestran todas los residuos catalíticos y de unión de las cuatro proteínas estudiadas, claramente se observa que el 50 % de residuos del sitio de únión tiene profundidad cero combinado con un área expuesta al solvente mayor, es decir, tienden a ubicarse en la superficie de la proteína. Por el contrario, la mayoría de los sitios catalitícos están en cierto modo ocultos a la capa exterior de las moléculas de agua, esto se debe a que tienen actividad catalítica cerca del núcleo hidrofóbico.

La Figura (5.6) muestra cómo los valores de SASA se correlacionan con los valores de profundidad de átomo. Como puede verse, los residuos del sitio de unión se encuentran en la zona de mayor accesibilidad, pero, como se mencionó anteriormente, esto no es una regla estricta, depende de la estructura tridimensional de la proteína, ya que hay valores SASA relativamente grandes combinados con valores más altos de profundidad.

Es de gran relevancia mencionar que la predicción exacta tanto de los residuos del sitio activo como de su ubicación específica en la estructura no es una tarea sencilla. Es por eso que consideramos que nuestro modelo combinado con algunas medidas físicas como SASA y profundidad de átomo puede ayudar a tener mejores resultados para esta tarea. En particular, observamos que una combinación de valores pequeños de profundidad combinado con valos altos de SASA pueden proporcionar una mejor estimación para la identificación y ubicación de residuos del sitio activo en proteínas globulares.

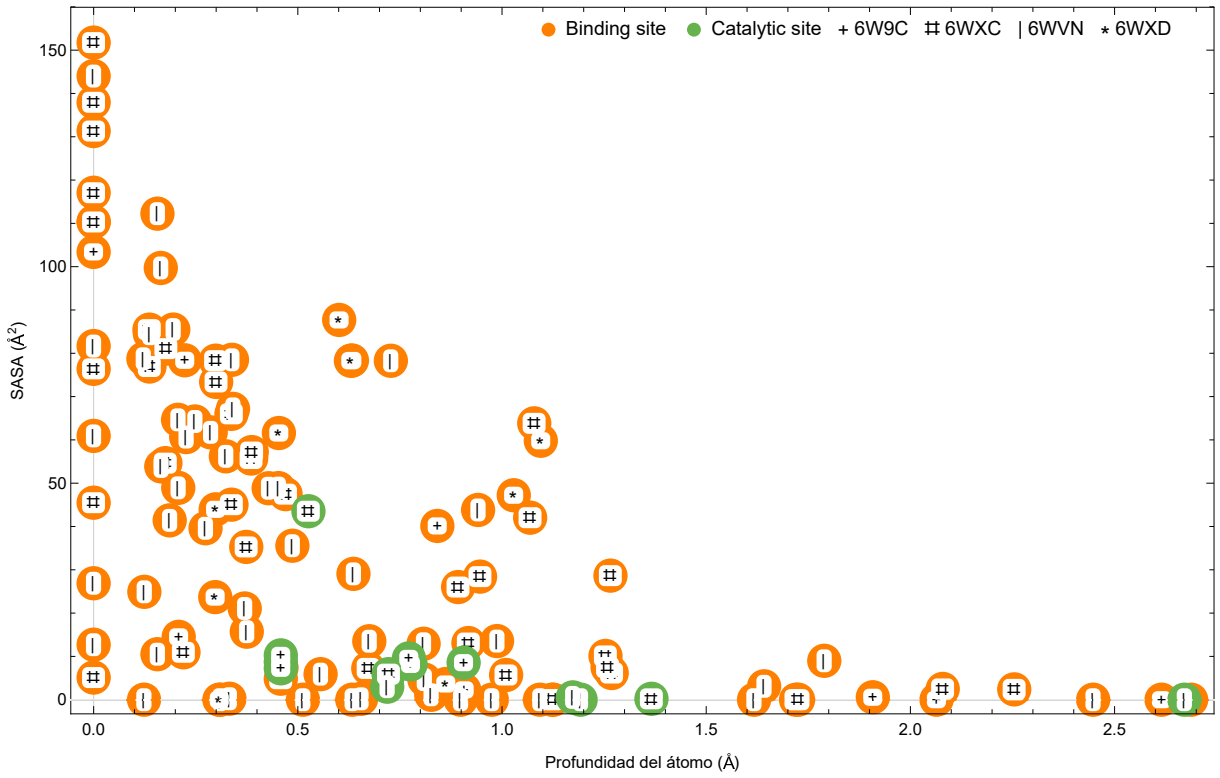


Figura 5.6: Correlación de la profundidad de átomo y el área de superficie accesible al solvente para residuos del sitio activo de proteínas del virus SARS-CoV-2. Los residuos del sitio catalítico están marcados en color verde y los residuos del sitio de unión en color naranja. Figura realizada con el programa *Wolfram Mathematica 11*.

Capítulo 6

Conclusiones y perspectivas

Los sistemas complejos pueden ser analizados como redes de interacciones entre los componentes del sistema. El análisis de la red puede caracterizar todo el sistema y sus componentes individuales. En particular, las estructuras de las proteínas se pueden representar como redes (grafos) donde los residuos o aminoácidos son los nodos (vértices) y sus interacciones son los enlaces (aristas), esto se conoce como un PRN (protein residue network en inglés). La representación de las estructuras proteicas como redes complejas facilita la búsqueda de información que pueden relacionarse con residuos funcionalmente importantes.

La identificación de residuos funcionales (residuos del sitio activo) en proteínas es un tema complejo, incluso cuando se dispone de estructuras atómicas detalladas. Las interacciones de los residuos de proteínas dentro y entre los sitios funcionales son cruciales para la actividad de las proteínas. Nuestro objetivo fue realizar un estudio de modelos de redes complejas para determinar la ubicación de residuos del sitio activo en proteínas relacionadas con el virus del SARS-CoV-2. Para esto, analizamos el rendimiento de la centralidad de los residuos en la identificación de residuos funcionalmente importantes, más en específico los residuos del sitio activo, así mismo, unir esta información con los resultados que nos proporciona el análisis del interior de la proteína mediante el análisis de la profundidad del átomo y el área de superficie accesible al solvente.

En el análisis de las estructuras de proteínas relacionadas con el virus del SARS-CoV-2 se encontró que estas redes incluyen un pequeño número de nodos importantes que son concentradores a través de los cuales muchos nodos pueden conectarse indirectamente. Se buscó examinar si los nodos de las redes de interacción de residuos de proteínas corresponden a residuos funcionales. Se encontró que la centralidad de cercanía de un residuo en la red caracteriza muchos sitios de proteínas funcionales, debido a que se encarga de determinar las rutas más eficientes que se deben recorrer para llegar desde un nodo a otro dentro de un

grafo. En particular, encontramos que para un radio de 8 Å los residuos que forman el sitio activo ocupan picos que se encuentran en la región que representa el 10 % de los valores más altos y forman parte de los máximos locales e incluso son máximo global.

La centralidad por cercanía o proximidad es una forma de detectar vértices que están dentro de la estructura de un grafo y que interactúan directamente o por unos pocos intermedios con todos los demás residuos de la proteína. Este calculo de proximidad de un nodo, también permite conocer su lejanía promedio (distancia inversa) del resto de los nodos. Por lo tanto, se pueden utilizar con otras medidas para analizar las estructuras de las proteínas. Asimismo, la centralidad de intermediación cuantifica la capacidad de un vértice para intervenir en la comunicación entre otros vértices. Cada vértice que forma parte de un camino más corto entre otros dos vértices puede intervenir en la comunicación o el flujo entre ellos. Para esta medida de centralidad, encontramos que los residuos funcionales forman parte el 5 % de los valores que destacan y, son parte de máximos locales, lo que significa que los residuos con valores altos de intermediación intercambian información con una buena cantidad de elementos dentro de la estructura de la proteína. La centralidad de vector propio cuantifica la cantidad de vértices vecinos que tiene cada uno de los residuos del sitio activo y de los vecinos que tienen esos vértices adyacentes, los residuos que forman el sitio activo se sitúan en un 10 % por debajo del máximo global, a excepción de la proteína 6WXD en donde encontramos que los residuos del sitio activo ocupan picos que corresponden a máximos locales e incluso uno de ellos es máximo global. Esta centralidad se determina a partir de sus conexiones directas con otros nodos y su potencia se deriva de las centralidades de estos vecinos directamente y de otros nodos de la red indirectamente.

En la tabla (6.1) se observa el rendimiento de cada uno de las centralidades y el rendimiento promedio en la identificación de residuos del sitio activo. Como se puede ver, la centralidad de cercanía funciona mejor para identificar residuos del sitio activo y con mayor presición en la proteína 6W9C, seguido de la centralidad de intermediación y por último, la centralidad de vector propio. Cabe destacar que las tres medidas de centralidad funcionan muy bien en la identifican residuos funcionales en la proteína 6WXD. Para estos casos, encontramos que los residuos del sitio activo de éstas proteínas están en la región que contiene los valores más altos y forman parte de los máximos locales e incluso son máximo global.

Nuestros resultados son parecidos a los encontrados en el artículo de Aguilar en donde se calcularon algunas medidas de centralidad buscando una relación entre éstas y la ubicación de sitios activos [**Aguilar & Olivares**, **2021**]. Se probó con la centralidad de cercanía, de intermediación y de vector propio, se encontró que la centralidad de cercanía funciona mejor

	6W9C	6WXC	6WVN	6WXD	Promedio
Closeness Centrality	0.83	0.52	0.71	0.70	0.681
Betweenness Centrality	0.20	0.047	0.018	0.3	0.141
Eigenvector Centrality	0.06	0	0.037	0.3	0.099
Promedio	0.36	0.189	0.255	0.433	

Cuadro 6.1: Resultados promedios para las medidas de centralidad en la identificación de residuos del sitio activo de proteínas relacionadas con el virus SARS-CoV-2.

para ubicar residuos funcionales en la estructura de la proteína. En los casos analizados, se encontraron más residuos cercanos a los máximos locales de la centralidad de cercanía, que a los de intermediación y para la centralidad de eigenvector se obtuvo una menor similitud entre los máximos o mínimos de esta cantidad con los residuos que forman los sitios activos.

El radio de corte utilizado en este estudio es de 6 Å y 8 Å, con esta distancia se determina que haya o no una conexión entre un par de átomos. Cabe mencionar que si cambiamos el radio de corte la red se modifica, pues, mientras más grande sea el radio más conectada será la red. En particular, aumenta la cantidad de enlaces (conexiones) y por lo tanto de caminos. También, pudimos ver que este cambio hace que las distancias sean más cortas, por lo tanto, existen trayectorias cortas y de mayor eficiencia. Entonces, es importante determinar el radio de corte de mayor eficiencia. En cuanto a la forma de las gráficas, no se observa ningún cambio cualitativo en los resultados, dicho eso, para este caso el radio de 8 Å funciona mejor, puesto que la red está más conectada.

A manera de investigación, realizamos el análisis de la PRN con un radio de corte $R_c = 10$ Å, en donde se encontró que la PRN está más conectada, pero, se ha comprobado que las conclusiones generales de este trabajo no se ven afectadas cuando se utiliza en su lugar este radio de corte, puesto que ubica los residuos del sitio activo de igual forma que con un radio de 8 Å, para las tres centralidades la mayoría de los residuos del sitio activo ocupan máximos destacados que son máximos locales e incluso máximos globales, el resto permanece cerca de ellos en un 5 %.

Con referencia al análisis del mapa HP, observamos que los residuos del sitio activo están conectados en promedio con el 53.84 % de aminoácidos polares y el 46.15 % corresponde a aminoácidos hidrofóbicos. La ubicación de los residuos del sitio activo es importante para complementar esta información, es por eso que se hizo un estudio del área de superficie accesible al solvente, este se utiliza ampliamente en los análisis de la estructura y estabilidad de las proteínas, en nuestro estudio encontramos que el 80 % de los residuos funcionales se localizan en las zonas de mayor accesibilidad. Sin embargo, para complementar la información del

área de superficie accesible calculamos la profundidad de cada uno de los residuos y residuos del sitio activo. Para nuestro estudio encontramos que el 75 % de los residuos que forman el sitio activo tiene profundidad cero, lo que signifca que se encuentran en la superficie de la proteína y por lo tanto, están en la zona de mayor accesibilidad.

Otro resultado interesante que encontramos es que los residuos que forman el sitio activo generalmente tienen valores pequeños de coeficiente de agrupamiento $\mathbb{C}_i$, lo que significa que los vecinos de estos residuos no interactúan directamente, pues, podrían conectarse por unos pocos nodos intermedios con todos los demás residuos de la proteína. Es por eso que estudiamos la correlación del coeficiente de agrupamiento y profundidad del átomo, resulta que los residuos superficiales que están expuestos al solvente son aquellos que tienen valores altos de coeficiente de agrupamiento, mientras que a mayores profundidades, donde los residuos están completamente rodeados por otros residuos y no están expuestos al solvente, la organización local de la proteína es siempre la misma en términos del coeficiente de agrupamiento.

También, realizamos un análisis de la frecuencia de aparición de los residuos directamente implicados en la catálisis, los resultados indicaron el dominio de un pequeño conjunto de residuos en catálisis, resulta que el porcentaje de residuos que participan en la catálisis de proteínas del SARS-CoV-2 son parecidos a los encontrados en el artículo de Bartlett, lo mismo ocurre con el porcentaje total de residuos de estás proteínas, se desconoce la función biológica de las 178 proteínas, pero, lo interesante es que la distribución de frecuencias de sus residuos son similares a los nuestros, siendo que las proteínas de nuestro estudio son proteínas no estructurales (NSPs) necesarias para la replicación del virus.

Cabe señalar que la implementación de este método es rápida en términos computacionales, pues un código que realice los cálculos que este estudio requiere, no debe tardar más de 1 minuto en una computadora común con la versión 11 de Mathematica, para una proteína con unos 200 residuos. Además, el modelo de redes complejas resulta interesante para determinar la ubicación de residuos del sitio activo mediante el análisis de medidas de centralidad en conjunto con el estudio del interior de la proteína. Por otra parte, encontramos residuos destacados en las medidas de centralidad que no están reportados experimentalmente como parte de sitios activos. Con esta metodología no se puede asegurar que lo sean, sin embargo, se podría explorar con otras técnicas la posibilidad de que esos residuos tengan alguna participación activa en la función de la proteína.

Los estudios funcionales de proteínas con frecuencia requieren el uso de múltiples enfo-

ques, particularmente, a nivel biológico y celular pueden mejorarse en gran medida mediante la exploración de las teorías, descriptores y modelos de redes desarrollados en otros campos que ofrece perspectivas y vías mucho más amplias para la comprensión de los sistemas biológicos, su organización interna y comportamiento dinámico. El progreso hacia estudios confiables basados en redes puede verse limitado por la información insuficiente sobre las redes, la capacidad limitada de los métodos de análisis y modelado de redes disponibles y los recursos computacionales inadecuados para facilitar el análisis y modelado de redes biológicas. El proporcionar facilidad de cálculo, complementa los otros recursos en la información, herramientas de modelado y los datos de PDB, Uniprot, Swiss Model sirven para facilitar colectivamente la investigación de las funciones de las proteínas y la dinámica de red.

En este trabajo estudiamos algunas medidas de centralidad para identificar nodos importantes en grandes redes, estas medidas nos permitieron clasificar los nodos influyentes y al mismo tiempo analizar qué medida puede adaptarse mejor al análisis de una red determinada. Además, el análisis de estas medidas es crucial para comprender las propiedades estructurales de las redes complejas. Tanto la centralidad de cercanía como la de intermediación se usan generalmente en muchas aplicaciones de análisis de redes. Las métricas de centralidad miden ciertas características de una red determinada, como las distancias entre nodos, las conexiones entre vecinos o las rutas cortas entre nodos, pero, es necesario desarrollar métricas de centralidad que adopten un concepto más amplio de influencia, aunque, se ha explorado un rico volumen de métricas de centralidad en la literatura, la mayoría de ellas se basan en el concepto de centralidad basada en la conectividad.

Un tema interesante que se puede estudiar más adelante es la conformación espacial del complejo enzima-sustrato y conocer las interacciones que intervienen en esta unión, ya que es un elemento de conocimiento importante tanto en el proceso de descubrimiento de nuevos fármacos como en el desarrollo de sustratos de mayor afinidad por la enzima. El establecimiento de un modelo conformacional se obtiene mediante la aplicación de técnicas de acoplamiento molecular, en las que el sustrato adopta diferentes poses dentro del centro catalítico de la enzima, resultando las más estables aquellas cuya energía de enlace sea menor. La técnica de acoplamiento molecular o técnica *docking* busca encontrar el mejor acoplamiento entre dos o más moléculas de tal forma que la afinidad entre ellas sea óptima. Esta técnica se podría correlacionar con los modelos de redes complejas para determinar la interacción enzima-sustrato.

El SARS-CoV-2, un nuevo coronavirus conocido principalmente que causa la enfermedad del COVID-19, se ha convertido en un asunto de preocupación crítica en todo el mundo. Ade-

más, han surgido métodos basados en redes para analizar y comprender el comportamiento complejo en sistemas biológicos con un enfoque computacional. En las últimas décadas, los métodos de clasificación basados en la red han proporcionado un análisis sistemático para predecir los residuos del sitio activo de influencia y proponer candidatos a fármacos en el tratamiento. Estudiar y comparar estas redes puede ser un paso efectivo para identificar nuevos compuestos farmacológicos para objetivos biológicos o de análisis.

Índice de figuras

Índice de cuadros

Bibliografía

[Bruce et al., 2015] Bruce A., Alexander J., Julian L., David M., Martin R., Keith R., & Peter W. (2015). *Molecular biology of the cell.* (6^a ed.). Garland Science.

[Engel, 2020] Engel P. (2020). *Enzymes: A Very Short Introduction*. Oxford University Press.

[Merchant & Larios, 2003] Larios L.F., & Merchant H. (2003). *Biología Celular y Molecular*. Pearson Educación.

[Aguilar & Olivares, 2021] Aguilar Pineda G.E., & Olivarez Quiroz L. (2021). *Catalytic and binding sites prediction in globular proteins through discrete Markov chains and network centrality measures.* Phys Biol, 18, 066002. doi: 10.1088/1478-3975/ac211b-.

[Vendruscolo et al., 2002] Vendruscolo, M., N. V. Dokholyan, E. Paci, & M. Karplus. (2002). *Smallworld view of the amino acids that play a key role in protein folding.* Phys. Rev. E 65:061910

[Pazos, 2022] Pazos F. (2022). *Predicción computacional de sitios funcionales de proteínas-Aplicaciones en biotecnología y biomedicina*. Adv Proteína Química Struct Biol. 130:39-5.

[Strogatz, 2001] Strogatz S. H. (2001). *Exploring complex networks*. Nature. 410:268–276.

[Watts & Strogatz, 1998] Watts, D. J., & S. H. Strogatz. (1998). *Collective dynamics of "smallworld" networks*. Nature. 393:440–442.

[Newman, 2010] Newman M. (2010). *Networks An Introduction*. Oxford University press.

[Trudeau, 1994] Trudeau R. J. (1994). *Introduction to graph theory*. Dover publications; 2nd edición.

[Newman, 2010] Newman M. (2010). *Networks An Introduction*. Oxford University press.

[Pintar et al., 2003] Alessandro P., Oliviero C., & Sándor P. (2003). *Atom Depth as a Descriptor of the Protein Interior*. Biophys J. 84(4): 2553–2561. doi: 10.1016/S0006-3495(03)75060-7

[Berman et al., 2000] Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, & P. E. Bourne. (2020). *The protein data bank.* Nucleic Acids Res. 28:235–242.

[Dayhoff, 1965] Dayhoff Margaret O. (1965). *National Biomedical Research Foundation*, Atlas of protein sequence and structure. Silver Spring, MD.

[Fackler D. et al., 2021] Fackler D., Tajkhorshid E., rizada Laxmikant C., Luthey Schulten Z. & Aksimentiev A. (2021). *Visual Molecular Dynamics LATEST ALPHA Version 1.9.4* [Software]. https://www.ks.uiuc.edu/Research/vmd/

[McGorgan & Enrique, 2020] McGorgan C., & Enrique R. (2020). *Reliable scientific information and COVID- 19.* Revista Cubana de Información en Ciencias de la Salud, 31(3), e1609.

[Walls et al., 2020] Walls, A. C., Park, Y.J., Tortorici, M. A., Wall, A., McGuire, A. T., & Veesler, D. (2020). *Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein.* Cell, 1–12. https://doi.org/10.1016/j.cell.2020.02.058

[Macchiagodena et al., 2020] Macchiagodena, M., Pagliai, M. & Procacci, P. (2020). *Identification of Potential Binders of the Main Protease 3CLpro of the COVID-19 via Structure-Based Ligand Design and Molecular Modeling.* Chemical Physics Letters, 137489.

[Villa, 2020] Villa Clara, J. (2020). *Impacto de la COVID-19 sobre la salud mental de las personas.* Medicentro Electrónica, 24(3), 578-594.

[Nelson, 2004] Nelson P. (2004). *Física Biológica.* Ed. Reverté.

[Branden & Tooze, 1999] Branden C & Tooze J. (1999). *Introduction to Protein Structure* (2^a ed.). Garland Publishing.

[Fersht & Freeman, 1999] Fersht A. & Freeman W. (1999). *Structure and Mechanism in Protein Science.* A more advanced coverage.

[Pain, 1994] Pain R. H. (Ed.) (1994). *Mechanisms of Protein Folding.* Oxford University Press.

[Olivares, 2017] Olivares Quiroz L. (2017). *Protein folding and unfolding pathways: The role of energy barriers, configurational entropy and internal energy: Comment on There and back again: Two views on the protein folding puzzle by Alexei V.* Physics of Life Reviews, 21, 75-76.

[Olivares & Garcia, 2004] Olivarez Quiroz L., & Garcia Colin L.S. (2004). *Plegamiento de las proteínas. Un problema interdisciplinario.* Journal of the Mexican Chemical Society. 48; 95-105.

[Torrance et al., 2008] Torrance J.W., Thornton J.M., & Bujnicki JM. (2008). *Predicción basada en la estructura de las enzimas y sus sitios activos.* John Wiley Sons.

[Lombardot et al., 2021] Lombardot T., Morgat A., Coudert E., Axelsen K., Neto T., Gehant S., Bansal P., Bolleman J., Gasteiger E., de Castro E., Baratin D., Pozzato M., Xenarios I., Poux S., Redaschi N., Bridge A., & UniProt Consortium. (2021). *UniProt Knowledgebase.* https://www.uniprot.org/help/uniprotkb

[Bienert et al., 2017] Bienert S., Waterhouse A., de Beer T.A.P., Tauriello G., Studer G., Bordoli L., Schwede T. & The SWISS-MODEL Repository (2017). *The SWISS-MODEL Repository - new features and functionality.* Nucleic Acids Res. 45, D313-D319

[Bougueleret et al., 2022] Bougueleret L., Sigrist C., de Castro E., Cerutti L., Cuche B., Hulo N., Puente A., & Xenarios I. (2022).*Desarrollos nuevos y continuos en PROSITE - ProRule.* https://prosite.expasy.org/prorule.html

[Sigrist et al., 2005] Sigrist CJ, De Castro E, Langendijk-Genevaux PS, Le Saux V., Bairoch A., & Hulo N. (2005). *ProRule: a new database containing functional and structural information on PROSITE profiles.* Bioinformatics. 21(21): 4060-4066.

[Delgado et al., 2002] A. Delgado, C. Minguillón & J. Joglar. (2002). *Introducción a la Síntesis de Fármacos.* Ed. Síntesis. ISBN: 84-9756-029-9.

[Lorenzo, 2020] Lorenzo P. (2020). *Farmacogenética y farmacogenómica.* Farmacología Básica y Clínica. (19^a ed.). Editorial Médica Panamericana.

[Stepniewska et al., 2020] Stepniewska Dziubinska, M.M., Zielenkiewicz, P., & Siedlecki, P. (2020) *Mejora de la detección de sitios de unión proteína-ligando con segmentación 3D.* Sci Rep 10, 5035 .https://doi.org/10.1038/s41598-020-61860-z

[Gorvalenya, 2020] Gorbalenya, A. E. (2020). *Severe acute respiratory syndrome-related coronavirus , The species and its viruses, a statement of the Coronavirus Study Group.* BioRxiv. https://doi.org/10.1101/2020.02.07.937862

[Drosten et al., 2003] Drosten, C., Günther, S., Preiser, W., Van der Werf, S., Brodt, H. R., Becker, S., Rabenau, H., Panning, M., Kolesnikova, L., Fouchier, R. A. M., Berger, A., Burguière, A. M., Cinatl, J., Eickmann, M., Escriou, N., Grywna, K., Kramme, S., Manuguerra, J. C., Müller, S., & Doerr, H. W. (2003). *Identification of a novel coronavirus*

in patients with severe acute respiratory syndrome. New England Journal of Medicine. https://doi.org/10.1056/NEJMoa030747

[Zaki et al., 2012] Zaki, A. M., Van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. M. E., & Fouchier, R. A. M. (2012). *Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia*. New England Journal of Medicine. https://doi.org/10.1056/NEJMoa1211721

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

[Towler et al., 2004] Towler, P., Staker, B., Prasad, S. G., Menon, S., Tang, J., Parsons, T., Ryan, D., Fisher, M., Williams, D., Dales, N. A., Patane, M. A., & Pantoliano, M. W. (2004). *ACE2 X-Ray Structures Reveal a Large Hinge-bending Motion Important for Inhibitor Binding and Catalysis*. Journal of Biological Chemistry. https://doi.org/10.1074/jbc.M311191200

[Andersen, 2020] Andersen, K. G. (2020). *The proximal origin of SARS-CoV-2*. Nature Medicine, 44–48. https://doi.org/10.1038/s41591-020-0820-9

[Rimanshee et al., 2021] Rimanshee A., Shweta K., Bharati P., Hiral M., Subhash C., Amit D., Vishal P., Gagan D., Lata P. & Mukesh K. (2021). *Structural insights into SARS-CoV-2 proteins*. Journal of Molecular Biology, Volume 433. https://doi.org/10.1016/j.jmb.2020.11.024.

[Kahraman et al., 2008] Kahraman A., Thornton J.M.. Schwede T., & Peitsch M. (2008). *Métodos para caracterizar las estructuras de los sitios de unión enzimática*. Biología Estructural Computacional: Métodos y Aplicaciones. Londres: World Scientific Publishing; 189–221.

[Costa et al., 2007] Costa F., Rodriguez F.A, Travieso G. & Villas Boas P.R. (2007). *I Characterization of Complex Networks: A Survey of measurements*. Advances in Physics.

[Reka & Barabasi, 2002] Reka A., & Barabasi A. (2002). *Statistical Mechanics of Complex Networks*. Reviews of Modern Physics.

[Aguilar, 2019] Aguilar G.E. (2019).*Sitios activos en macromoléculas biológicas mediante Cadenas de Markov y Teoría de Redes Complejas.*[Tesis de Maestría, Universidad Autónoma de la Ciudad de México].

[Atilgan et al., 2004] Atilgan R., Akan P., & Canan Baysal. (2004). *Small-World Communication of Residues and Significance for Protein Dynamics.* Biophysical Journal Volume 86. 85–91

[Newman, 2003] Mark E.J. Newman. (2003). *The Structure and Function of Complex Networks*. SIAM Review

[Bavelas, 1950] Bavelas A. (1950). *Communication Patterns in Task Oriented Groups*. The Journal of the Acoustical Society of America, 22(6), 725–730.

[Bonacich & Lu, 2012] Bonacich P., & Lu P. (2012). *Introduction to Mathematical Sociology*, Princeton University Press.

[Das et al., 2016] Das Gupta D, Kaushik R., & Jayaram B. (2016). *Protein folding is a convergent problem!* Biochem Biophys Res Commun. 480(4):741-744. doi: 10.1016/j.bbrc.2016.10.119. PMID: 27983988

[Jumper et al., 2021] Jumper J., Evans R. & Pritzel A. (2021). *Highly accurate protein structure prediction with AlphaFold* Nature 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

[Stephen, 2020] Stephen C. (2020). *DeepMind has not solved protein folding* Reciprocal Space.

[Callaway, 2020] Callaway E. (2020). *It will change everything: DeepMind's AI makes gigantic leap in solving protein structures*. Nature 588, 203-204 . doi: https://doi.org/10.1038/d41586-020-03348-4

[Heffernan et al., 2015] Heffernan R., Paliwal K., & Lyons J. (2015). *Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning*. Sci Rep 5, 11476 . https://doi.org/10.1038/srep11476

[Ali et al., 2014] Ali S.A., Hassan M.I., Islam A., & Ahmad F. (2014). *Una revisión de los métodos disponibles para estimar las áreas de superficie accesibles a los solventes de las proteínas solubles en los estados plegado y desplegado*. Curr Protein Pept Sci. 15(5):456-76. doi: 10.2174/1389203715666140327114232. PMID: 24678666.

[Mihel et al., 2008] Mihel J., Sikic M., & Tomić, S. (2008). *PSAIA – Protein Structure and Interaction Analyzer*. BMC Struct Biol . https://doi.org/10.1186/1472-6807-8-21

[Negi et al., 2015] Negi S., Zhu H., Fraczkiewicz R., & Braun W. (2015).*Calculation of Solvent Accessible Surface Areas, Atomic Solvation Energies and Their Gradients for Macromolecules: Sealy Center for Structural Biology, University of Texas Medical Branch, Galveston, TX 77555* [Software online]. En línea: http://curie.utmb.edu/getarea.html

[Fraczkiewicz & Braun, 1998] Fraczkiewicz, R., & Braun, W. (1998). *Exact and Efficient Analytical Calculation of the Accessible Surface Areas and Their Gradients for Macromolecules*. J. Comp. Chem., 19, 319-333.

[Von et al., 1993] von Freyberg, B., Richmond, T.J., & Braun, W. (1993). *Surface area included in energy refinement of proteins: a comparative study on atomic solvation parameters*. J. Mol. Biol. 233(2):275-292.

[Chakravarty & Varadarajan, 1999] Chakravarty S., & Varadarajan R. (1999). *Residue depth: a novel parameter for the analysis of protein structure and stability.* Struct. Fold. Des. 7:723–732.

[Pearlman & Kollman, 1995] Pearlman D.A., & Kollman P. (1995). *AMBER: Mecánica molecular, análisis de modo normal, dinámica molecular y cálculos de energía libre para simular las propiedades estructurales y energéticas de las moléculas.* Universidad de California.

[Hubbard et al., 1994] Hubbard S., Gross K.H., & Argos P. (1994). *Cavidades intramoleculares en proteínas globulares*. (7^a ed.). Prot. Eng.

[Carugo & Pongor, 2002] Carugo O., & Pongor S. (2002) *Protein fold similarity estimated by a probabilistic approach based on C(alpha)-C(alpha) distance comparison.*. J Mol Biol. 315(4):887-98. doi: 10.1006/jmbi. 2001.5250. PMID: 11812155.

[Nagy et al., 2016] Nagy T., Vera R., Hudaiberdiev S., Kumari S., Ligeti B. & Pongor S. (2016).*DPX Protusion Index Web Server* [Software online]. http://pongor.itk.ppke.hu/protein/dpx.html/DPX$_i$*ntro*

[Izidoro et al., 2015] Izidoro S.C., Lacerda A.M., & Pappa G.L. (2015) *Búsqueda de sitios activos genéticos multiobjetivos.* Conferencia de Computación Genética y Evolutiva (Madrid - España). NY, ACM (pp. 905–910).

[Watts, 1999] Watts, D. J. (1999). *Small Worlds*. Princeton University Press.

[Pedersen et al., 1991] Pedersen T., Sigurskjold g., Andersen k., & Redfield C. (1991). *A nuclear magnetic resonance study of the hydrogen-exchange behaviour of lysozyme in crystals and solution.* J. Mol. Biol. 218:413–426.

[Furnham et al., 2020] Furnham N., Holliday G.L., de Beer T.A., Jacobsen J.O., Pearson W.R., & Thornton J.M. (2020). *El Catalytic Site Atlas 2.0: catalogación de sitios catalíticos y residuos identificados en enzimas.* Ácidos nucleicos Res. 42:D 485–D489.

[Hage & Harary, 1995] Hage P. & Harary F. (1995). *Eccentricity and centrality in networks.* Social Netw., vol. 17, núm. 1, págs. 57 a 63.

[Bartlett et al., 2002] Bartlett G., Porter C., Borkakoti G. & Thornton J. (1995). *Analysis of Catalytic Residues in Enzyme Active Sites.*J. Mol. Biol. (2002) 324, 105–121

[Olivares & Garcia, 2007] Olivares Quiroz L., & Garcia Colin L.S. (2007). *Protein's native state stability in a chemically induced denaturation mechanism.* J. Theor. Biol. 246(2). 214-224. ISSN: 0022-5193.

[Zukang et al., 2000] Helen M. B, John W, Zukang F, Gary G, T. N. Bhat, Helge W, Ilya N. Shindyalov, & Philip E. Bourne. (2000). *The protein data bank*, Nucleic Acids Research, 28:235-242.