UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

FACULTAD DE CIENCIAS

Análisis de las medidas de centralidad de segundo y tercer orden en proteasas del SARS-CoV-2

T E S I S

QUE PARA OBTENER EL TÍTULO DE

FÍSICO

P R E S E N T A

Margarita Reyes García

A S E S O R

Dr. Luis Agustín Olivares Quiroz

Ensenada, B.C. Agosto 2024

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
FACULTAD DE CIENCIAS,
CAMPUS ENSENADA.

“Análisis de las medidas de centralidad de segundo y tercer orden en proteasas del SARS-CoV-2”

TESIS

PARA CUBRIR LOS REQUISITOS NECESARIOS PARA OBTENER EL TÍTULO DE

Físico

PRESENTA

Margarita Reyes García
357554

A quien el Comité de Tesis autoriza el trabajo terminal y de acuerdo con el

Art. 19 del R.G.E.P.E.P, emite los siguientes votos aprobatorios mediante rubrica:

Dr. Luis Agustín Olivares Quiroz
DIRECTOR

Dr. Juan Crisóstomo Tapia Mercado
SINODAL

Dr. Jesús Ramón Lerma Aragón
SINODAL

Dra. Eloísa del Carmen García Canseco
SINODAL

“Por la Realización Plena del Ser”

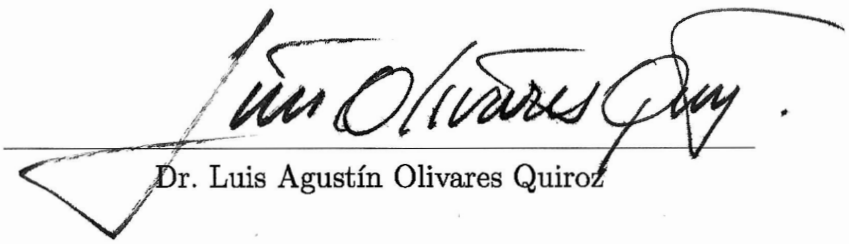
Resumen de la tesis presentada por **Margarita Reyes Garcia** como requisito parcial para obtener el **Título profesional de Físico**. Ensenada, Baja California, México. Agosto, 2024.

ANÁLISIS DE LAS MEDIDAS DE CENTRALIDAD DE SEGUNDO Y TERCER ORDEN EN PROTEASAS DEL SARS-COV-2

El estudio de las proteasas M^{pro} y PL^{pro} del nuevo virus SARS-CoV-2 se ha convertido en un tema central en las ciencias físicas y biomédicas desde la aparición de la pandemia ocasionada por este virus a finales de 2019, debido a que se ha demostrado que son las principales responsables de la replicación del virus. El propósito principal de este trabajo es analizar el sitio de replicación, conocido como sitio activo, dentro de la estructura de PL^{pro} mediante un enfoque en teoría de grafos y medidas de centralidad de redes. Para esto, se modeló la estructura tridimensional de PL^{pro} como una red con sus átomos de C_α como nodos, y las interacciones covalentes y no covalentes como enlaces. Como segunda parte del análisis, se obtuvieron las propiedades $\langle L \rangle$ y $\langle C \rangle$, y las distribuciones de grado para determinar la dinámica y robustez de la red. Se extendieron los cálculos a las estructuras con los inhibidores GRL-0617 de molécula pequeña y VIR250 basada en péptidos para determinar sus capacidades inhibitorias *in silico*. Al mismo tiempo, se analizaron mutaciones del residuo catalítico principal Cys111 a temperaturas de 100 K y 293 K. Con el método estadístico AUC-ROC se encontró que las medidas de orden superior C_E y LR-C_S son las más adecuadas para la predicción de los sitios activos en PL^{pro}, por otro lado, la red tiene propiedades de redes de mundo pequeño. Los métodos y herramientas mencionados, en conjunto, representan un muy buen método para predecir sitios activos en PL^{pro} del SARS-CoV-2 y para comprender la dinámica de este virus.

Palabras clave: **sitios activos, medidas de centralidad, subgrafos, PL^{pro} del SARS-CoV-2, red de mundo pequeño.**

Resumen aprobado por:



Dr. Luis Agustín Olivares Quiroz

Agradecimientos

Agradezco a mi asesor de tesis, el Dr. Luis Agustín Olivares Quiroz, por compartirme su conocimiento, por guiarme en todo el trayecto de mi trabajo, por su gran paciencia y por darme ánimos.

También agradezco a mis sinodales, el Dr. Juan Tapia, el Dr. Jesús Lerma y la Dra. Eloísa García por tomarse el tiempo de revisar mi tesis y por sus comentarios. Agradezco también al profe Claudio por asesorarme en algunas de mis dudas y por ser un gran profesor.

Quiero agradecer a mi familia por estar conmigo. A mis hermanos Chuyito, Salvador, Itaí y Esther por quererme y cuidarme. A mi abue Magdalena por sus consejos, y a todos mis tíos. A mi mamá Juana le agradezco mi vida entera.

Por último, agradezco a todos los profesores de la facultad por su gran labor de enseñanza.

Índice general

Resumen I

Agradecimientos II

Índice de figuras V

Índice de tablas VII

Notación IX

1. Contexto del problema 1

2. SARS-CoV-2 y sus proteasas virales 8
2.1. Estructura nativa de las proteínas . . . 8
2.2. Definición de proteasa . . . 11
2.3. Síntesis de proteínas del SARS-CoV-2 . . . 13
2.4. Enzima PL^{pro} . . . 15
2.4.1. Inhibidores GRL-0617 y VIR250 . . . 17
2.4.2. Mutaciones . . . 18

3. Conceptos básicos de redes 21
3.1. Teoría de grafos (Graph-theory) . . . 23
3.1.1. Matriz de adyacencia . . . 24
3.1.2. Medidas básicas de redes . . . 26
3.2. Propiedad de mundo pequeño (*small world*) en redes complejas . . . 28
3.2.1. Modelo de Erdős-Rényi . . . 30
3.2.2. Modelo de Watts-Strogatz . . . 33

4. Medidas de centralidad de orden superior 35
4.1. Medidas de centralidad de primer orden . . . 39
4.2. Medidas de centralidad de segundo orden . . . 40
4.2.1. Centralidad de eigenvector . . . 40
4.2.2. Centralidad de subgrafo . . . 43
4.2.3. Comunicabilidad de la red . . . 46
4.3. Medidas de tercer orden . . . 46
4.3.1. Centralidad de subgrafo de largo alcance . . . 46
4.3.2. Comunicabilidad de largo alcance . . . 48

5. Metodología computacional en bioinformática **50**
5.1. Gestión de archivos PDB 50
5.1.1. Estructura del archivo PDB 50
5.1.2. Verificación de la secuencia y corrección del archivo PDB 51
5.1.3. Lectura del archivo PDB con Biopython 54
5.2. Construcción del grafo y cálculo de las medidas de centralidad 55
5.3. Método AUC-ROC 58
5.4. Grafo aleatorio y regular 60
5.4.1. Longitud de ruta promedio 61
5.4.2. Coeficiente de agrupamiento promedio 63
5.5. Herramientas adicionales 64

6. Aplicación a las enzimas del SARS-CoV-2 **65**
6.1. Construcción de la red de residuos de proteína 66
6.2. Análisis de las medidas de centralidad en PL^{pro} (6W9C) 68
6.3. Inhibidores y mutaciones 73
6.4. Propiedad de mundo pequeño 80

7. Conclusiones **84**

Bibliografía **87**

Índice de figuras

2.1. Estructura base de los aminoácidos estándar . . . 9
2.2. Unión de un sustrato en el sitio activo de una enzima . . . 11
2.3. Proteasa viral M^{pro} del SARS-CoV-2 . . . 13
2.4. Proteasas codificadas por el genoma del SARS-CoV-2 . . . 14
2.5. Representación cristalográfica de PL^{pro} de SARS-CoV-2 . . . 16
2.6. Inhibidor de péptidos VIR250 en complejo con PL^{pro} . . . 18

3.1. Tipos de grafos: simples y no dirigidos, múltiples y dirigidos . . . 24
3.2. Ilustración de la esfera con radio R_c centrada en un nodo de la red . . . 25
3.3. Grafo de árbol no dirigido y su matriz de adyacencia asociada . . . 25
3.4. Variación de la longitud de ruta promedio $\langle L \rangle$ conforme varía el número de enlaces . . . 26
3.5. Coeficiente de agrupamiento promedio $\langle C \rangle$ conforme varía la distribución y el número de enlaces . . . 27
3.6. Redes aleatorias con número fijo de enlaces . . . 31
3.7. Distribuciones de Poisson para dos redes aleatorias . . . 33
3.8. Modelo de Watts y Strogatz para distintos valores de p . . . 34

4.1. *Network motifs* . . . 36
4.2. Diferencia entre *caminos cerrados* y subgrafos. . . . 37
4.3. Ciclos sin cuerda . . . 44

5.1. Estructura del archivo pdb. Sección ATOM *records*. . . . 51
5.2. Seleción de cadenas con PDBFixer . . . 53
5.3. Selección de residuos faltantes con PDBFixer . . . 54
5.4. Diagrama de flujo del método de la potencia para calcular C_E . . . 57
5.5. Ejemplo de una Curva ROC generada a partir de 100 umbrales distintos . 59
5.6. Algoritmo de Floyd-Warshall para el cálculo de $\langle L \rangle$ . . . 62
5.7. Matriz t del coeficiente de agrupamiento $\langle C \rangle$ . . . 63

6.1. Representación cristalográfica de PL^{pro} tipo *wild* y en su forma de red de átomos de C_α . . . 66
6.2. Comparación de las medidas de centralidad de PL^{pro} del SARS-CoV-2 para los radios de corte R_c de 7 Å, 9 Å y 11 Å. . . . 67
6.3. Medidas de centralidad de orden superior de PL^{pro} tipo *wild* del SARS-CoV-2 69
6.4. Cálculo del área bajo la curva (AUC) de las curvas ROC asociadas a las centralidades de orden superior. . . . 72

6.5. Medidas de centralidad de orden superior para inhibidores y mutaciones de PL^{pro} 74
6.6. Diferencia relativa porcentual de las medidas de centralidad de orden superior entre PL^{pro} tipo *wild* (PDB ID: 6W9C) y sus complejos con inhibidores y mutaciones 76
6.7. Distribución de grado de PL^{pro} y la red aleatoria de Erdös–Rényi equivalente 80

Índice de tablas

I. Convención de tres y una letra de los aminoácidos estándar 10
II. Sitios activos y de unión en PL^{pro} tipo *wild*, y en las estructuras con inhibidores y mutaciones 20

I. Centralidad de intermediación C_B, eigenvector C_E, subgrafo C_S y subgrafo de largo alcance Z_{pp} promedio para dos grafos con ciclos sin cuerda 49

I. Residuos máximos y mínimos en orden descendente en las medidas de centralidad de orden superior para PL^{pro} (PDB ID: 6W9C). Se obtuvieron los mismos resultados para C_E y LR-C_S. Las celdas rojas representan los residuos del sitio activo y las celdas azules los residuos del sitio de unión. . 71
II. Residuos aminoácidos de proteínas del SARS-CoV-2 con valores máximos en C_E en orden descendente. Se obtienen los mismos resultados para los máximos en LR-C_S. 75
III. Residuos aminoácidos de proteínas del SARS-CoV-2 con valores máximos en C_S en orden descendente. Las celdas rojas representan los sitios activos. 77
IV. Comparación de las medidas de centralidad promedio de la PL^{pro} tipo *wild* del SARS-CoV-2 (6W9C), PL^{pro}-GRL0617 (7CMD), PL^{pro}-VIR250 (6WUU), las mutaciones de Cys111 a 100 K (6WRH) y a 293 K (6XG3). . 79
V. Diferencia relativa porcentual entre PL^{pro} tipo *wild* y sus estructuras variantes: 7CMD[1], 6WUU[2], 6WRH[3] y 6XG3[4]. 79
VI. Longitud de ruta promedio $\langle L \rangle$ y coeficiente de agrupamiento $\langle C \rangle$ de PL^{pro} del SARS-CoV-2 (6W9C), y sus equivalentes en redes regular y aleatoria de Erdös–Rényi (ER) 82

Notación

Siglas

ACE2 Enzima Convertidora de Angiotensina 2 (*Angiontensin-Converting Enzyme 2*)

ADN Ácido desoxirribonucleico

ARN Ácido ribonucleico

AUC-ROC Área bajo la curva - Característica operativa relativa (*Area under the curve - Relative operating characteristic*)

COVID-19 Coronavirus 2019

M^{pro} Proteasa principal (*main protease*)

NSP Proteínas no estructurales (*non-structural proteins*)

ODLIS Diccionario en Línea de Bibliotecología y Ciencias de la Información (*Online Dictionary for Library and Information Sciencie*)

OMS Organización Mundial de la Salud

ORF Marco de lectura abierto (*open reading frame*)

PCR Reacción en cadena de la polimerasa (*polymerase chain reaction*)

PDB Banco de Datos de Proteínas (*Protein Data Bank*)

PL^{pro} Proteasa similar a la papaína (*papain-like protease*)

PPI Interacción proteína-proteína (*protein-protein interaction*)

PRN Red de residuos de proteína (*protein residue network*)

SARS-CoV-2 Coronavirus 2 del Síndrome Respiratorio Agudo Severo (*Severe Acute Respiratory Syndrome Coronavirus 2*)

SIDA Síndrome de Inmunodeficiencia Adquirida

SW Mundo pequeño (*small world*)

VIH Virus de Inmunodeficiencia Humana

Variables y funciones

A Matriz de adyacencia

C Coeficiente de agrupamiento

$\langle \boldsymbol{C} \rangle$ Coeficiente de agrupamiento promedio

$\boldsymbol{C}_{\mathrm{B}}$ Centralidad de intermediación

$\boldsymbol{C}_{\mathbf{D}}$ Centralidad de grado

$\boldsymbol{C}_{\mathbf{E}}$ Centralidad de eigenvector

$\boldsymbol{C}_{\mathbf{S}}$, $\boldsymbol{G_{pp}}$ Centralidad de subgrafo

$\boldsymbol{d_{ij}}$ Matriz de distancias entre los nodos n_i y n_j

$\boldsymbol{E}$ Número de enlaces

$\mathcal{E}$ Conjunto de enlaces de un grafo

G Matriz de la exponencial de A

$\mathcal{G}$ Grafo

$\langle \boldsymbol{G_{pq}} \rangle$ Comunicabilidad promedio de la red

H Función de Heaviside

$\boldsymbol{k}$ Grado de nodo

LR-$\boldsymbol{C}_{\mathbf{S}}$, $\boldsymbol{Z_{pp}}$ Centralidad de subgrafo de largo alcance

$\langle \boldsymbol{L} \rangle$ Longitud de ruta promedio

$\boldsymbol{N}$ Número de nodos

$\mathcal{N}$ Conjunto de nodos de un grafo

$\boldsymbol{n_i}$ i-ésimo nodo de un grafo

$\boldsymbol{p}$ Probabilidad

$\boldsymbol{R}_{\mathbf{c}}$ Radio de corte

$\boldsymbol{R_{ij}}$ Distancia euclidiana entre los nodos n_i y n_j

Z Matriz para medidas de tercer orden

$\langle \boldsymbol{Z_{pq}} \rangle$ Comunicabilidad promedio de largo alcance

$\boldsymbol{\theta_{pq}}$ Ángulo de comunicabilidad promedio

Capítulo 1

Contexto del problema

El estudio de enfermedades virales en los últimos años a tomado relevancia entorno a la enfermedad COVID-19 y a su agente viral causante SARS-CoV-2, debido a la reciente pandemia de principios de 2020 con origen en Wuhan, China (Mojica-Crespo y Morales-Crespo, 2020). De hecho, en 2003 ya se había desencadenado una pandemia por un virus SARS-CoV de la misma familia de coronavirus con un total de 8,096 infectados y 774 muertes distribuidas en 26 países alrededor del mundo, que finalizó en julio del mismo año (De Wit et al., 2016). Aunque se encontró que SARS-CoV y SARS-CoV-2 son similares en su secuencia en un 90 % (Osipiuk et al., 2021), SARS-CoV-2 originó una pandemia de gran escala que dejó de ser considerada una emergencia sanitaria de nivel internacional hasta mayo de 2023 (OMS, 2023). Esta pandemia dejó un saldo de 774 millones de casos de infección y 7 millones de muertes hasta enero de 2024. Tan sólo del 11 de diciembre de 2023 al 7 de enero de 2024 se detectaron 1.1 millones de infecciones nuevas (OMS, 2024).

La ejecución rápida de las medidas de contención del SARS-CoV fue el motivo principal por el que la pandemia ocasionada por este virus no escaló a una pandemia de gran magnitud, explicaron De Wit *et. al.* (De Wit et al., 2016). Sin embargo, la alta tasa de replicación viral de SARS-CoV-2, mucho mayor que la de SARS-CoV, también es un factor determinante en la persistencia prolongada de la pandemia por COVID-19 y los grandes estragos a nivel económico, social y de la salud. En la replicación del virus las proteasas

virales desempeñan la función principal. Estas son un tipo de proteínas que aceleran los procesos de replicación. Por ende, es suficiente con disminuir la velocidad de replicación de SARS-CoV-2 mediante el diseño y la aplicación de inhibidores potentes en sus proteasas virales para frenar los contagios. En el caso del SARS-CoV-2 las proteasas que juegan este papel importante son M^{pro} y PL^{pro} (Osipiuk et al., 2021).

Los virus, incluido el SARS-CoV-2, se caracterizan por su incapacidad de reproducirse por sí mismos. Para reproducirse requieren de maquinarias presentes únicamente en células vivas. La variante de coronavirus SARS-CoV-2 determinó que las células humanas son una opción muy viable y favorable que garantizan su supervivencia. En particular, aquellas células con el receptor ACE2 presentes en mayor cantidad en células del pulmón y del intestino delgado. De ahí que la invasión de este virus en la célula huésped ocasione enfermedades respiratorias o, en casos más graves, neumonía (Hamming et al., 2004).

El virus SARS-CoV-2 se acopla en el receptor ACE2 a través de la proteína *Spike*. Luego, libera su material genético y se sintetizan dos grupos de poliproteínas virales. En ambos grupos se encuentran las proteínas no estructurales (NSP) que necesitan ser liberadas para que el virus pueda replicarse (Yadav et al., 2021). En el caso específico de este virus se encontró que la proteasa principal M^{pro}, perteneciente a NSP5, escinde los enlaces peptídicos de NSP4-11 liberándolas como proteínas individuales. La PL^{pro}, perteneciente al multidominio NSP3, es una proteasa que cumple una función similar a M^{pro}, libera a NSP1-3. Además de esto, cumple la función de descomponer la ubiquitina, una proteína que también se encuentra en la célula humana y que se encarga de seleccionar proteínas dañadas o proteínas virales para su aniquilación (Osipiuk et al., 2021).

Por esta razón PL^{pro} es un objetivo farmacológico importante al igual que M^{pro}, sin embargo, las investigaciones para PL^{pro} son más escasas. A fin de inhibir la acción de esta enzima se han desarrollado inhibidores *in vitro* que muestran tener efectos positivos en

el frenamiento de la replicación viral. Entre estos se encuentran el inhibidor de péptidos VIR250 y el inhibidor de molécula pequeña GRL-0617, siendo este último el más potente. Sin embargo, aún hacen falta los ensayos *in silico* para determinar la eficacia de estos inhibidores (Calleja et al., 2022).

La catálisis de la reacción para llevar a cabo la escisión de NSP1-3 se da en una región de la PL^{pro} conocida como sitio activo. Esta región es característica de las proteínas catalíticas y puede encontrarse en forma de bolsillo en la superficie o en el interior de estas. En el caso de PL^{pro} estos bolsillos se encuentran en la superficie (Osipiuk et al., 2021). Los inhibidores son efectivos porque actúan sobre esta región para detener parcial o completamente la actividad catalítica, por esta razón es importante determinar la ubicación de los sitios activos. Aunque la cristalización de las proteínas por difracción de rayos-X nos permite identificar la ubicación de estos sitios activos, este proceso es muy complejo y costoso. Por esta razón, se está trabajando con herramientas computacionales que nos permitan predecir estos sitios activos.

La predicción de sitios activos en proteínas es un área activa de la investigación básica tanto en física, matemáticas como en biología molecular, y se puede estudiar desde tres enfoques: los métodos específicos de ligandos (Babor et al., 2008), métodos basados en la secuencia (Capra et al., 2009) y métodos basados en estructuras (Zhang et al., 2022). De estos tres métodos, los basados en estructura son los más precisos para la predicción de estos sitios importantes, ya que consideran las propiedades fisicoquímicas, como el efecto hidrofóbico, al considerar la geometría de la proteína. En particular, este tipo de método detecta bolsillos en la estructura que podrían estar relacionados con sitios de unión a ligandos y sitios activos (Aguilar-Pineda y Olivares-Quiroz, 2021). El análisis de estructuras proteicas y la predicción de sus sitios activos mediante teoría de grafos y medidas de centralidad de redes es un enfoque basado en estructura.

La teoría de grafos es una herramienta matemática efectiva para modelar casi cualquier problema de la vida real y extraer información importante de las interacciones entre sus elementos. Su origen se remonta al año 1741 cuando Euler intentó resolver el problema de los siete puentes de Königsberg (Euler, 1741). Este problema consistió en determinar si era posible cruzar exactamente una vez los siete puentes del río Pregolia y llegar al mismo punto de partida. Para esto, Euler representó a cada sección aislada de la ciudad por puntos (nodos), y conectó a dos de ellos con un enlace (arista) como representación de los puentes que unen a ambos. Actualmente a este tipo de modelo, en donde se representan a los elementos de un problema como nodos y a las relaciones entre estos como enlaces, se le conoce como grafo. Su modelo de grafo le permitió concluir que no era posible y que de serlo todos los nodos deben de tener un número par de enlaces. Este problema es actualmente conocido como el problema de ciclo euleriano (Sampath et al., 2021). Un siglo después, Kirchhoff aplicó la teoría con éxito para solucionar las ecuaciones involucradas con el problema de circuitos eléctricos (Kirchhoff, 1847). Desde entonces, han surgido otros problemas matemáticos similares y desafiantes en teoría de grafos. Por ejemplo, el problema del viajante que consiste en encontrar un ciclo hamiltoniano de menor coste, es decir, un ciclo que atraviese cada estación de una ciudad exactamente una vez de forma que el viajante recorra la menor cantidad de distancia posible. La teoría de grafos es una rama de las matemáticas versátil y altamente multidisciplinaria (Cheikhrouhou y Khoufi, 2021).

El enfoque de teoría de grafos para proteínas consiste en una representación de grano grueso en donde se eligen a los residuos aminoácidos como nodos dentro de una red con N aminoácidos y E enlaces. A esto se le conoce como red de residuos de proteína (PRN, por sus siglas en inglés). En segunda instancia, las medidas de centralidad son una parte importante de la teoría de grafos que se emplean cuando el objetivo es localizar a nodos o regiones clave dentro de una red. Mediante estas medidas se cuantifica la centralidad de manera local y promedio, en donde cada medida refleja un aspecto diferente de una red. Algunas de las más importantes son la centralidad de grado C_{D}, de cercanía C_{C}, de

intermediación C_{B} y de eigenvector C_{E} (Estrada, 2012).

Este enfoque ha sido estudiado exitosamente por Aguilar Pineda y Olivares Quiroz (Aguilar-Pineda y Olivares-Quiroz, 2021). Aplicaron C_{C}, C_{B} y C_{E} para la predicción de sitios activos en proteínas globulares. Por otro lado, Estrada (Estrada, 2020) diseñó nuevas medidas de centralidad basadas en subgrafos que definió como medidas de segundo y tercer orden. Mismas que aplicó para el análisis topológico de la proteasa viral $\mathrm{M}^{\mathrm{pro}}$ del SARS-CoV-2. En su estudio encontró que estas nuevas medidas predicen mejor los sitios activos en $\mathrm{M}^{\mathrm{pro}}$ que medidas de centralidad estándar.

Además del análisis mediante medidas de centralidad, es necesario un análisis de redes más exhaustivo para determinar propiedades globales. Este análisis involucra las distribuciones de grado y la comparación con modelos de redes teóricas. Esta idea parte del hecho de que diversas redes del mundo real, principalmente las tecnológicas y biológicas, a menudo exhiben comportamientos de redes de mundo pequeño (SW, por sus siglas en inglés). Este tipo de redes son conocidas por oscilar entre redes aleatorias y regulares. Mientras que la regularidad les proporciona un alto nivel de agrupamiento, la aleatoriedad proporciona una alta accesibilidad entre un par de nodos en una red. Es decir, en una red SW, los nodos tienen un alto grado de conectividad y existe una alta resistencia a la pérdida de información (Barabási, 2013). A esto se le conoce como robustez de la red. Determinar si la PRN de una proteína viral es una red SW nos daría información sobre la robustez de la red y, con ello, podríamos elegir el plan de acción adecuado para combatir un virus.

El objetivo principal de esta investigación es analizar la estructura tridimensional de la proteína $\mathrm{PL}^{\mathrm{pro}}$ de SARS-CoV-2 mediante un enfoque de teoría de redes para determinar si existe una correlación entre los sitios activos y las medidas de centralidad de segundo y tercer orden. A esto se agrega el estudio de sus inhibidores VIR250 y GRL-0617 para

determinar su eficacia, y también el de las mutaciones del residuo del sitio activo Cys111, el residuo con mayor actividad catalítica. Para esto, construiremos una red tomando como nodos a los átomos de C_α de los aminoácidos que conforman a PL^{pro}, y como enlaces a las interacciones covalentes y no covalentes entre todos los pares de nodos de la red. Para los enlaces construiremos esferas de radio R_c centradas en cada átomo de C_α, es decir, en cada nodo n_i de la red, de manera que si un átomo distinto n_j es contenido en la esfera, se enlazarán n_i y n_j. Después, se calculará la representación matricial del grafo conocida como matriz de adyacencia y, posteriormente, el par de funciones matriciales **G** y **Z** de las que dependen las medidas de centralidad de segundo y tercer orden, respectivamente. Se calcularán las medidas de centralidad locales para detectar a los nodos clave de la PL^{pro} y determinar las variaciones locales con respecto a los compuestos PL^{pro}-inhibidores y mutaciones. Por otro lado, se calcularán las medidas de centralidad promedio para determinar las variaciones a nivel global. Finalmente, se comparará a la red de la PL^{pro} con redes aleatorias y regulares para determinar el grado de fragilidad o robustez de la proteína.

Estructura de la tesis:

- En el capítulo 2 se describen conceptos básicos de fisicoquímica, una extensión más detallada del mecanismo de infección viral del SARS-CoV-2 y el rol de la proteasa viral PL^{pro} en este proceso. También se mencionan inhibidores y mutaciones significativos.

- En el capítulo 3 se discuten los conceptos básicos de la teoría de grafos para modelar a PL^{pro} como una red de residuos de proteína. También se introducen los modelos de redes teóricas necesarios para determinar si una red es de mundo pequeño.

- En el capítulo 4 se describen las medidas de centralidad de primer hasta tercer orden, y antecedentes de la aplicación de estas medidas en el análisis de sitios activos en proteínas.

- En el capítulo 5 se mencionan los programas y algoritmos empleados para la construcción del grafo y cálculos de las medidas de centralidad, así como para los modelos de redes.

- Finalmente, en el capítulo 6 se exponen e interpretan los resultados.

- Las conclusiones de los resultados se presentan en el capítulo 7.

Capítulo 2

SARS-CoV-2 y sus proteasas virales

La llegada de la pandemia de COVID 19, provocada por el virus SARS-CoV-2 y originada en Wuhan China (Mojica-Crespo y Morales-Crespo, 2020), motivó a científicos en el mundo a estudiar sus proteasas virales. Este virus tiene una mayor tasa de transmisión que el SARS-CoV, de ahí que ha desembocado en la pandemia más catastrófica de este siglo con un saldo de 7 millones de descesos y 774 millones de casos de infección reportados hasta enero de 2024 (OMS, 2024). Este virus es miembro de la familia *Coronavidae* de género beta y afecta únicamente a humanos. Como en la mayoría de los procesos de replicación de todos los virus, se determinó que sus proteasas virales son las principales responsables de la replicación viral, especialmente PL^{pro} y M^{pro}. Dos inhibidores han demostrado disminuir significativamente la velocidad de replicación del virus *in vitro*: VIR250 y GRL-0617, diseñados para PL^{pro} (Calleja et al., 2022). A continuación se introducirán conceptos de bioquímica necesarios para comprender el mecanismo de replicación del SARS-CoV-2 y se discutirá el porqué se ha elegido a PL^{pro} como proteína de estudio. Además se discuten mutaciones significativas de este virus.

2.1. Estructura nativa de las proteínas

Un conjunto de macromoléculas importantes que participan en diversas funciones biológicas en seres vivos y en virus son las proteínas. Estos polímeros de aminoácidos

se sintetizan en la célula y participan en diversos procesos biológicos como los de transporte (p. ej., la hemoglobina que transporta el oxígeno en la sangre), estructura (p. ej., el colágeno presente en los huesos, piel y cartílagos), defensa (anticuerpos), catálisis (las que aceleran reacciones químicas), entre otros (Nelson et al., 2008). Su función está determinada primordialmente por su configuración de aminoácidos (aa) y su estructura tridimensional. Generalmente, su estructura se compone a partir de veinte aminoácidos estándar y alcanza longitudes de cientos a miles de aa. La estructura base de los aminoácidos consiste de un grupo amino, uno carboxilo, un átomo de hidrógeno y un grupo R o residuo unidos a un átomo de carbono alfa C_α, como se muestra en la figura 2.1. Se diferencían entre sí en sus cadenas laterales o grupos R (Olivares-Quiroz y García-Colín Scherer, 2004).

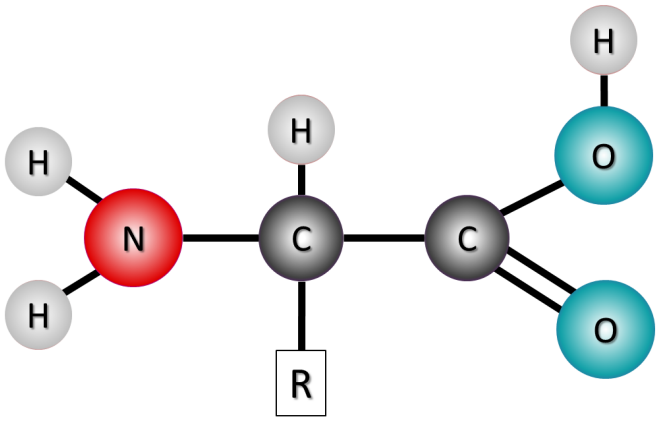


Figura 2.1: Estructura base de los aminoácidos estándar. A la derecha el grupo carboxilo COOH, a la izquiera el grupo amino NH_2, arriba el átomo de hidrógeno, abajo el grupo residuo R y en el centro el átomo de C_α (Nelson et al., 2008).

Los aminoácidos se unen covalentemente por enlaces peptídicos a través del grupo amino y carboxilo para formar una cadena lineal de aminoácidos conocida como la estructura primaria de una proteína. Al extremo de la secuencia que finaliza con el grupo amino se le conoce como extremo N-terminal y al extremo que finaliza con el grupo carboxilo como C-terminal. Dado que la unión de aminoácidos se da por la eliminación de una molécula de agua entre los grupos funcionales, la cadena se puede descomponer por hidrólisis, la adición de una molécula de agua. Por otro lado, los grupos R determinan la polaridad del aminoácido. Es decir, poseen propiedades físicas que definen interacciones

de tipo electrostáticas (repulsivas o atractivas) con el solvente en el interior de la célula, generando dos tendencias: residuos hidrofóbicos que repelen el agua o, el caso contrario, residuos hidrofílicos (clasificación en la tabla I) (Olivares-Quiroz y García-Colín Scherer, 2004).

Tabla I. Convención de tres y una letra de los aminoácidos estándar clasificados con base a su hidrofobicidad. La glicina no es un residuo hidrofóbico, puede considerarse neutro, sin embargo se mantiene en esa clasificación (Nelson et al., 2008).

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

Aunque las interacciones no covalentes (como el efecto hidrofóbico) son más débiles que las covalentes, son las que más favorecen la estabilidad de la estructura de la proteína. De hecho, el efecto hidrofóbico es el principal responsable del plegamiento de la secuencia (Finkelstein y Ptitsyn, 2016). Sin embargo, los puentes de hidrógeno son los principales formadores de las estructuras secundarias: α-hélices y β-láminas. Estas últimas son agrupaciones de estructuras más pequeñas denominadas hebras β. La estructura terciaria consiste en una estructura tridimensional única que se forma a partir de la unión de α-hélices y β-láminas mediante hebras de enlace para formar una sola cadena polipeptídica. Estas estructuras tienden a ubicarse en el centro de la proteína, mientras que las hebras de enlace se mantienen en la parte externa. La proteína debe alcanzar la estructura nativa para ser completamente funcional (Olivares-Quiroz y García-Colín Scherer, 2004).

2.2. Definición de proteasa

Las proteasas son un tipo de proteína catalítica que escinde los enlaces peptídicos de otra proteína por hidrólisis. Son especiales porque aceleran reacciones químicas en sistemas biológicos a velocidades del orden de 10^6 a 10^{12} más rápido que una reacción sin catálisis (Voet et al., 2016). La catálisis de una reacción en un compuesto químico, denominado el sustrato de la reacción, se da específicamente en una región dentro de la estructura tridimensional de la proteína denominada sitio activo. Sus sitios activos, usualmente, tienen forma de bolsillo y encajan únicamente con un tipo específico de compuesto químico (ver figura 2.2). Cuando la proteína se desnaturaliza, es decir, cuando pierde su estructura nativa, se pierde también su actividad catalítica (Nelson et al., 2008).

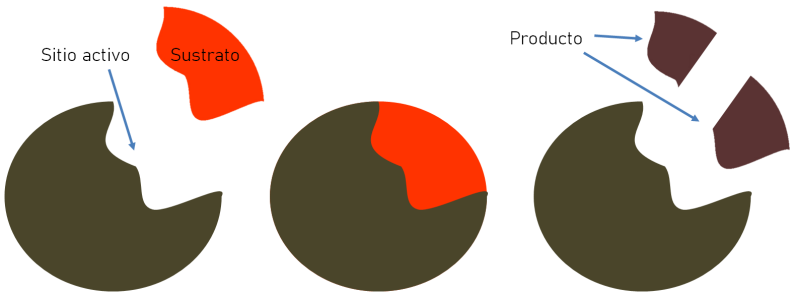


Figura 2.2: Representación de la acción catalítica de una enzima en un sustrato. La región destinada a la reacción solo actúa con un sustrato específico y se conoce como sitio activo. En el caso de las proteasas, el sustrato puede ser una poliproteína, de manera que la enzima divide a esta en proteínas de menor tamaño (Nelson et al., 2008).

Las proteasas se pueden clasificar en dos tipos: en proteasas celulares y en proteasas virales. Por un lado, las proteasas celulares están implicada en procesos biológicos en seres vivos, como la tripsina y la quimiotripsina. Ambas aceleran los procesos de digestión de alimentos y son sintetizadas en el páncreas (Nelson et al., 2008). Otro ejemplo son aquellas que pertenecen al complejo proteasomal con estructura en forma de barril y con múltiples sitios activos. El proteasoma se encarga de dar mantenimiento a las células a través de la eliminación de proteínas sobrantes o en mal estado para prevenir enfermedades a largo plazo (Jung y Grune, 2012). Por otro lado, las proteasas virales son aquellas que juegan un rol importantes en los procesos de maduración de los virus. Por ejemplo la proteasa del VIH causante de la enfermedad del SIDA. Este virus es uno de los más estudiados hasta

la actualidad. Pese a ello, aún no se ha logrado conseguir un tratamiento para erradicar por completo la enfermedad del SIDA debido a su alta tasa de mutación (Majerová y Konvalinka, 2022).

Las proteasas virales son consideradas como los objetivos terapéuticos más importantes para el diseño de medicamentos, más concretamente, sus sitios activos. La primera razón se debe a su participación activa en los procesos de replicación viral. La segunda, por la disminución de los efectos secundarios de los medicamentos al poder ser diseñados para ser aplicados directamente sobre la región catalítica del virus, y no en regiones aproximadas a este que puedan dañar a las células o tejidos de alrededor (Majerová y Konvalinka, 2022).

A menudo, los medicamentos o tratamientos de inhibición viral son diseñados considerando la posible resistencia por parte de estos virus. A una mayor dosis de estos medicamentos, la probabilidad de mutación se eleva. Esto es lo que ocurrió con la proteasa del VIH. Los primeros inhibidores contra este virus actuaron directamente en el sitio activo, sin embargo, debido a las altas dosis requeridas ocasionaron efectos secundarios adversos en humanos. Entre ellos, una mayor resistencia a los inhibidores que desencadenaron la rápida mutación del VIH. Por esta razón, se buscan inhibidores eficaces y muy específicos contra enzimas específicas, y la alternancia entre inhibidores y otros tipos de tratamientos antivirales para combatir los síntomas (Majerová y Konvalinka, 2022).

Tanto las proteasas celulares como las virales se pueden clasificar, por la triada catalítica que conforma su sitio activo, en seis tipos: cisteína, serina, treonina, glutámico, metalo y aspártico (Shen y Chou, 2009). Diferentes proteasas ejecutan un mecanismo de acción diferente para fragmentar secuencias de aminoácidos. La cisteína proteasa tiene en su sitio activo a la triada catalítica Cys, His y Asp, como la caspasa que participa en la regeneración de tejidos. Por otro lado, las serinas proteasas se caracterizan por la triada Ser, His y Asp. Este tipo de proteasas a menudo participan en los procesos de digestión,

como es el caso de la tripsina y la quimotripsina (Nelson et al., 2008). Algunas proteasas pueden formar una diada catalítica, como es el caso de la cisteína proteasa M^{pro} del SARS-CoV-2 con los residuos Cys145 e His41 en su sitio activo (ver figura 2.3). Al día de hoy no se conoce con certeza la relevancia del ácido aspártico en las cisteínas proteasas (Osipiuk et al., 2021).

Figura 2.3: Proteasa viral M^{pro} del SARS-CoV-2. El sitio activo Cys145-His41 se resalta en rojo.

2.3. Síntesis de proteínas del SARS-CoV-2

Durante años se ha cuestionado si los virus son entidades vivas o no vivas ya que no pueden reproducirse por sí mismos, por lo que se les ha etiquetado como entidades biológicas (Villarreal, 2004). Para reproducirse, acuden a células vivas y las programan a partir del material genético (ADN o ARN) para fabricar a sus nuevos viriones infecciosos. Este proceso es facilitado por las proteasas virales incorporadas inicialmente en el virus, proteasas que intervienen tanto en la unión al receptor de la célula huésped como en el ingreso del código genético. La invasión de la célula huésped por proteínas virales provoca daños en su funcionamiento que pueden llegar a ser muy perjudiciales para la salud.

El genoma de SARS-CoV-2 codifica tres grupos de proteínas importantes: cuatro proteínas estructurales *spike* (S), *envelope* (E), *nucleocapsid* (N) y *membrane* (M), dieciseis proteínas no estructurales (NSP1-NSP16) y seis accesorias (ORF, el total de proteínas accesorias aún es debatible) (figura 2.4). La proteína N protege el material genético del virus que es liberado luego del anclaje de la proteína S en el receptor ACE2 (*Angiontensin-Converting Enzyme 2*) de la célula humana (Yadav et al., 2021). Aunque este receptor se encuentra en células de varios órganos y tejidos del cuerpo humano como el estómago, el corazón, el cerebro, el riñón, también en la mucosa oral y nasal, y en arterias y venas, se ha encontrado en mayor abundancia en células del pulmón y del intestino delgado. Por esta razón se ha convertido en una enfermedad de vías respiratorias (Hamming et al., 2004).

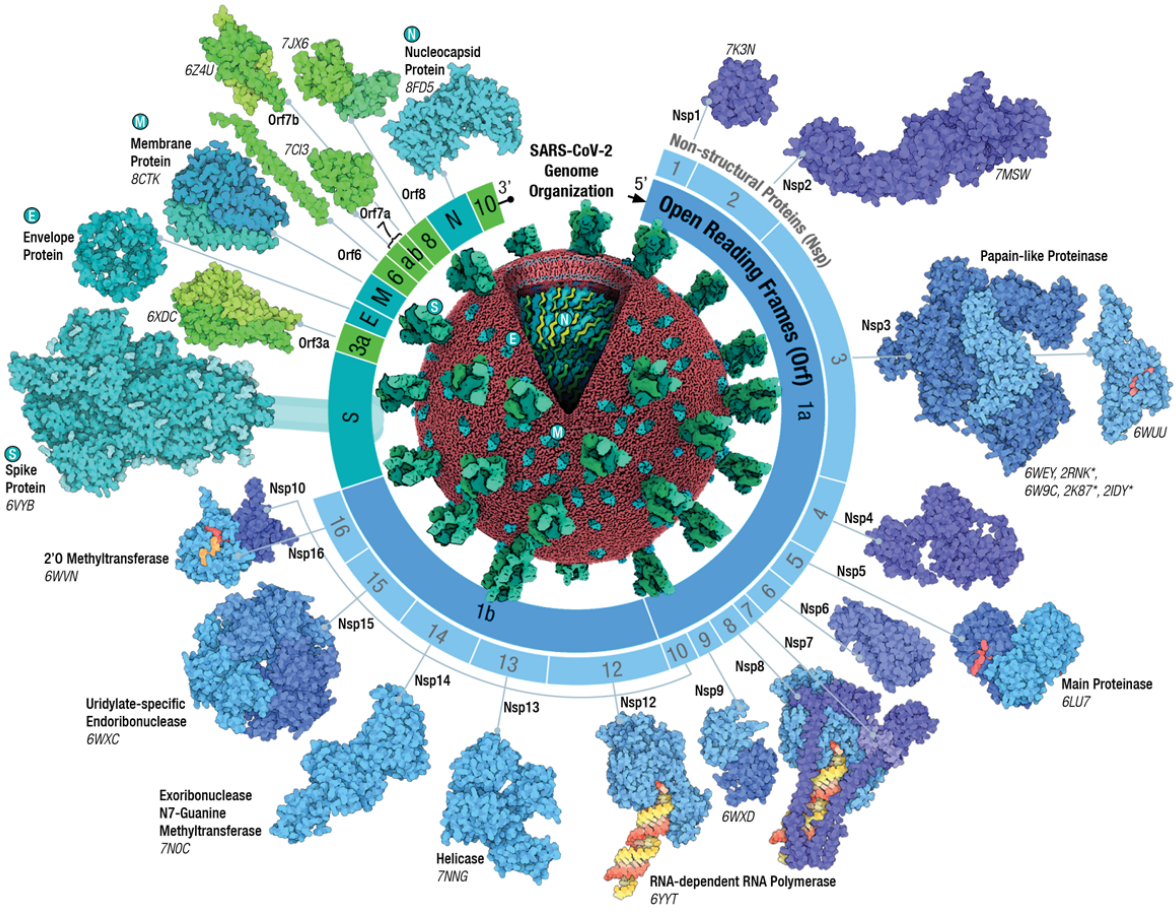


Figura 2.4: Proteasas codificadas por el genoma del SARS-CoV-2. Este virus pertenece a la familia de los coronavirus, los cuales presentan una forma coronada en la superficie de su estructura. Tomado de "*Evolution of the SARS-CoV-2 proteome in three dimensions (3D) during the first 6 months of the COVID-19 pandemic*" (Lubin et al., 2022).

Luego de la liberación del ARN, el par de proteínas accesorias ORF1a y ORF1b se convierten en dos poliproteínas que contienen a las dieciseis proteínas no estructurales, NSP1-11 y NSP12-16. La NSP3, una proteína multidominio que contiene a PL^{pro}, y la NSP5, conocida como M^{pro}, fragmentan a las poliproteínas en polímeros más cortos para liberar al resto de proteínas no estructurales, incluidas sus propias escinciones. Específicamente, NSP3 libera a NSP1-NSP3, y NSP5 al resto. Por esta razón a NSP3 y NSP5 también se les conoce como "tijeras moleculares". Después de la liberación de las NSP comienza la replicación viral (Yadav et al., 2021).

La PL^{pro} cumple una función adicional, participa en la hidrólisis de la ubiquitina, una proteína que trabaja en conjunto con el proteasoma. Mientras que la ubiquitina selecciona a las proteínas dañadas y proteínas virales, el proteasoma se encarga de eliminarlas (Osipiuk et al., 2021).

Por su capacidad de inhibir la ubiquitina y de acelerar el proceso de replicación de SARS-CoV-2 mediante la hidrólisis de proteasas virales, la PL^{pro} es un objetivo farmacológico importante al igual que M^{pro}. Sin embargo, el desarrollo de inhibidores y el estudio de los efectos de estos en PL^{pro} son más escasos que los llevados a cabo para M^{pro}, razón por la que se eligió a PL^{pro} como proteína diana en esta investigación.

2.4. Enzima PL^{pro}

La PL^{pro} es un tipo de cisteína proteasa con una estructura tridimensional que asemeja la forma de una mano derecha extendida (ver figura 2.5). Esta estructura se divide en los grupos *Palm* (palma), *Thumb* (pulgar) y *Fingers* (dedos). Su secuencia se compone de 315 aminoácidos con una alta concentración de residuos de cisteína (3.5 %). El subdominio *Palm* consiste en seis hebras β, el subdominio *Thumb* en seis α-hélices y un bucle β, y el subdominio *Fingers* en seis hebras β y dos α-hélices. En este último se halla el sitio de unión, conformado por Cys189, Cys192, Cys224 y Cys226, una región que participa

solamente en el ensablamiento de la proteína con otros compuestos. Además del dominio *Palm-Thumb-Fingers*, la PL^{pro} se caracteriza por el pequeño dominio N-terminal con los primeros 60 residuos de la cadena dispuestos en cinco hebras β y una α-hélice (Osipiuk et al., 2021).

El sitio activo de PL^{pro} se compone de la triada Cys111, His272 y Asp286 ubicada en la intersección *Palm-Thumb*. Cys111 participa como nucleófilo, es decir, atrae e hidroliza al sustrato, es el principal agente catalizador de la triada del sitio activo. Mientras que His272 (base) actúa como activador de la cisteína y Asp286 interviene principalmente como ayudante de His272. Es importante destacar que M^{pro} tiene la diada catalítica Cys145 e His41 en vez de una triada. Esto puede ser importante para el entendimiento más profundo de los efectos del residuo catalítico Asp y para dilucidar las diferencias en los mecanismos de inhibición de ambas proteasas virales (Osipiuk et al., 2021).

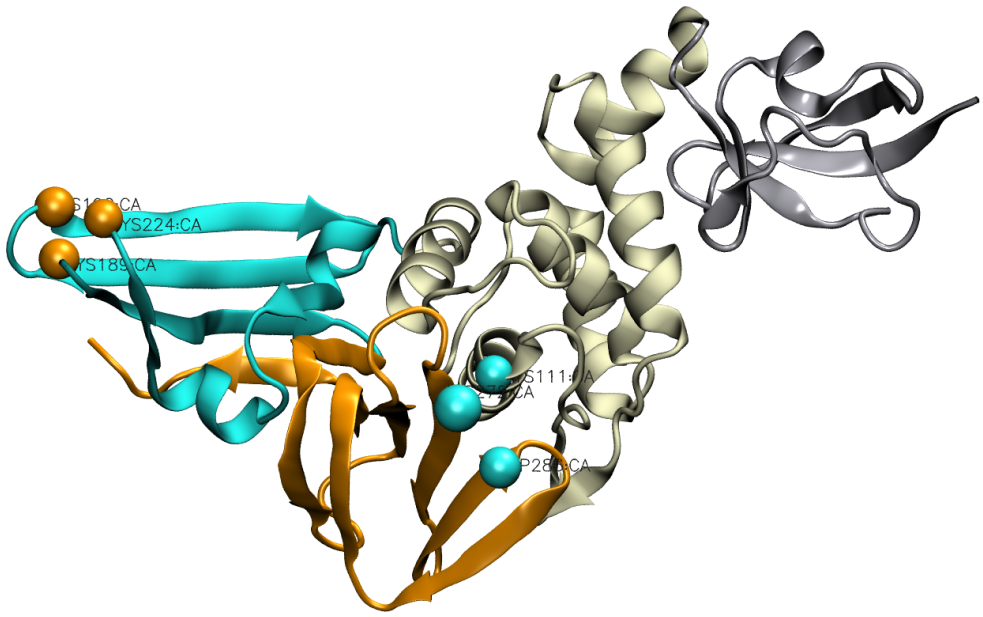


Figura 2.5: Representación cristalográfica de PL^{pro} (PDB ID: 6W9C). Es una proteína mediana de 315 aminoácidos distribuidos en cuatro subdominios: *Ubi-like* (1-60, gris), *Thumb* (61-178, blanco), *Fingers* (179-240, azul) y *Palm* (241-315, naranja). El sitio activo de esta cisteína proteasa consiste en la triada catalítica Cys111, His272 y Asp286 (esferas azules) en la intersección *Palm-Thumb*, y con residuos del sitio de unión Cys189, Cys192, Cys224 y Cys226 (esferas naranjas) en el subdominio *Fingers* (Ullrich y Nitsche, 2022).

2.4.1. Inhibidores GRL-0617 y VIR250

Las secuencias de PL^{pro} de SARS-CoV-2 y PL^{pro} de SARS-CoV son similares en un 90 % (Osipiuk et al., 2021) por lo que se ha intentado reutilizar inhibidores eficacez de PL^{pro} de SARS-CoV en PL^{pro} de SARS-CoV-2. Se encontró que algunos inhibidores fabricados para PL^{pro} de SARS-CoV también funcionan en PL^{pro} de SARS-CoV-2. El más potente de ellos es el inhibidor no covalente GRL-0617 de molécula pequeña. Así se le conoce por ser una molécula con un bajo peso molecular (unidades de Da). Este inhibidor actúa sobre un bolsillo en la región *Palm* de PL^{pro} destinada al sustrato junto al sitio activo, con un IC_{50} de aproximadamente 1 µM a 2 µM. Es decir, se requieren aproximadamente 1 µM a 2 µM de este inhibidor para inhibir la actividad catalítica de la proteasa viral en un 50 %. Es importante tener en cuenta que la reutilización de inhibidores no es más sencilla que el desarrollo desde cero de estos. Aún continúa siendo un tema complejo en términos éticos, financieros y principalmente de ensayos preclínicos, dado que estos fármacos son herramientas altamente específicas (Calleja et al., 2022).

Otros inhibidores consisten en estructuras basadas en péptidos. De hecho estas son las primeras versiones de inhibidores contra distintas especies de CoV, por ejemplo, VIR250 y VIR251. Los péptidos son cadenas de aminoácidos de longitud corta pero de mayor peso molecular que inhibidores de molécula pequeña, rondan unidades de kDa. Estos inhibidores actúan sobre Cys111 covalentemente y toman el lugar del sustrato, como se ve en la figura 2.6. Ambos inhibidores tienen un IC_{50} de aproximadamente 50 µM, alrededor de cincuenta veces más que el inhibidor GRL-0617. Es decir, es cincuenta veces menos potente que GRL-0617. Otro punto en contra es que su fabricación es considerada un gran desafío, mayor que el de inhibidores de molécula pequeña (Wang et al., 2023).

Los ensayos *in vitro* son los que han determinado que estos y otros inhibidores son eficacez y potentes, sin embargo, aún hacen falta los ensayos *in silico*. Por esta razón, es necesario un análisis exhaustivo de los efectos de los inhibidores de PL^{pro} de SARS-

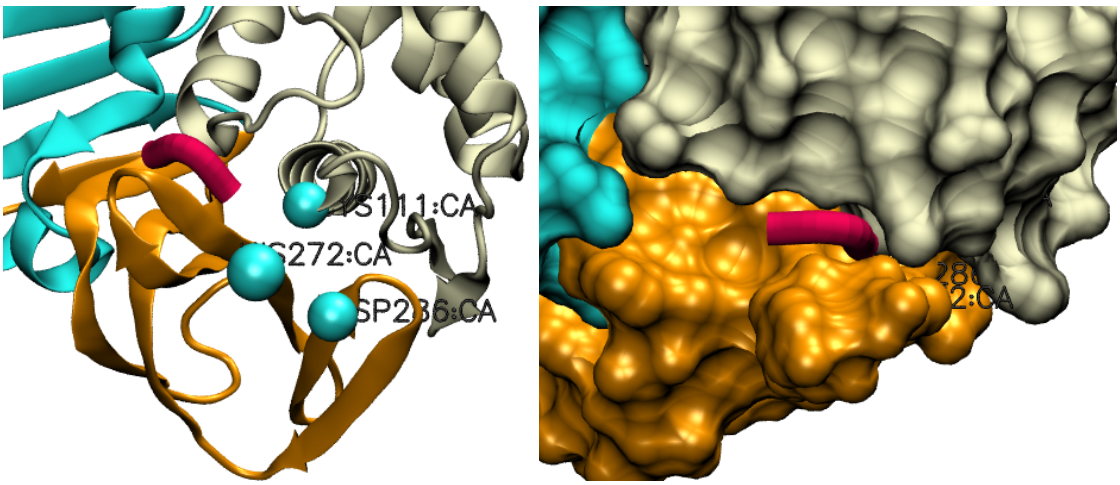


Figura 2.6: PL^{pro} en complejo con el inhibidor basado en péptidos VIR250 (PDB ID: 6WUU). El inhibidor consiste de cinco residuos aminoácidos (rojo) que actúa sobre la región destinada al sustrato en el sitio activo Cys111, His272 y Asp286.

CoV en PL^{pro} de SARS-CoV-2. Tanto los inhibidores basados en péptidos como los de molécula pequeña tienen ventajas y desventajas, por lo que Calleja *et. al.* discutieron sobre la posibilidad de fabricar un inhibidor como resultado de la combinación de ambos, sin embargo, aún se siguen realizando investigaciones sobre esto y se tienen altas espectativas.

Para esta investigación se seleccionó el inhibidor GRL-0617 por ser considerado un potente inhibidor, y el inhibidor VIR250 para comparar las diferencias en las potencias entre inhibidores de molécula pequeña y aquellos basados en péptidos.

2.4.2. Mutaciones

Las mutaciones son simultáneamente una causa y consecuencia de los mecanismos de inhibición fabricados para las proteasas virales. Por esta razón, el diseño de inhibidores no solo debe de considerar el funcionamiento de la proteasa tipo *wild* (proteasa sin mutaciones), también debe de considerar las variaciones evolutivas de las proteasas virales con el fin de minimizar la resistencia a estos fármacos (Yilmaz et al., 2016).

Las mutaciones que se dan en el sitio activo tienen efectos más cruciales en la estructura y funcionalidad de las proteínas. Una mutación importante en la PL^{pro} es la mutación

del residuo cisteína por el residuo serina en la posición 111, es decir, la mutación C111S. Hay dos versiones de esta mutación, la primera se da a una temperatura de 100 K (PDB ID: 6WRH) y la segunda a 293 K (PDB ID: 6XG3). Aunque ambas versiones inducen un mismo tipo de mutación, se encontraron diferencias en las estructuras del sitio de unión ubicado en el subdominio *Fingers*, y en el intervalo Gly266 a Gly271 en la secuencia de PL^{pro}. Estos cambios infieren que estas regiones son las más susceptibles a deformarse, es decir, las más flexibles (Osipiuk et al., 2021). En el presente trabajo se analizarán las estructuras 6WRH y 6XG3 para precisar cambios conformacionales en las regiones mencionadas y para determinar si existe una relación con la temperatura.

Hemos visto que las proteínas son macromoléculas con una estructura nativa que se forma a partir del plegamiento de una secuencia de aminoácidos. Este plegamiento es favorecido principalmente por el efecto hidrofóbico, que aglomera a las subestructuras que repelen el agua en la región central de la proteína. Estas macromoléculas participan en procesos biológicos como los de transporte, estructura, defensa y catálisis. Las proteasas son un tipo especial de proteína catalítica debido a que desarma a otras proteínas a una velocidad elevada. Esta fragmentación la ejecuta en una región de su estructura conocida como sitio activo. Estos sitios usualmente consisten en una triada de aminoácidos ubicados en regiones accesibles en la superficie de una proteína o en cavidades en el interior.

Para esta investigación hemos elegido como proteasa de estudio a la PL^{pro} del SARS-CoV-2 debido a su destacada participación en los procesos de replicación viral, en la inhibición de la ubiquitina (una proteína que ayuda a dar mantenimiento a las células a través de la selección de proteínas en mal estado) y por las escasas investigaciones entorno a esta proteasa. Sus sitios activos y de unión son el foco de atención en este estudio ya que son las principales regiones que favorecen la replicación del virus. Por un lado, el sitio activo de PL^{pro} es la región en donde se dividen las poliproteínas virales que requieren ser liberadas para que el virus pueda replicarse a una velocidad elevada. Por el otro, el sitio de

unión es un sitio destinado únicamente al ensamblamiento de la proteasa viral con otras moléculas.

Además del estudio de la proteasa tipo *wild* (sin mutaciones ni inhibidores) se estudiarán los efectos de los inhibidores GRL-0617 de molécula pequeña y VIR250 basada en péptidos (alrededor de cincuenta veces menos potente que el anterior), a fin de determinar la capacidad inhibitoria *in silico* de estos en PL^{pro}. La eficacia de un inhibidor se mide por su potencia y también por su nivel de influencia en la mutación de un virus, dado que a mayor rango de afectación aumenta la probabilidad de que un virus mute. Por esta razón, también es necesario estudiar mutaciones de PL^{pro}, especialmente las que involucran su sitio activo. Para esto se han elegido dos mutaciones en el residuo cisteína del sitio activo, el residuo con una mayor actividad catalítica. En la tabla II se resume la información de las estructuras proteicas que se estudiarán en esta investigación.

Tabla II. Sitios activos y de unión en la estructura de PL^{pro} tipo *wild* (PDB ID: 6W9C) (sin mutaciones ni inhibidores) del SARS-CoV-2, en las estructuras de PL^{pro} con los inhibidores GRL-0617 (PDB ID: 7CMD) y VIR250 (PDB ID: 6WUU), y en las estructuras de las mutaciones de C111S a 100 K (PDB ID: 6WRH) y C111S a 273 K (PDB ID: 6XG3).

Proteína	**Residuos**	**Residuo del sitio activo**	**Residuo del sitio de unión**
6W9C	310	C111, H272, D286	C189, C192, C224
7CMD	315	C111, H272, D286	C189, C192, C224, C226
6WUU	318	C111, H272, D286	C189, C192, C224, C226
6WRH	316	S111, H272, D286	C189, C192, C224, C226
6XG3	313	S111, H272, D286	C189, C192, C224, C226

La PL^{pro} es una proteína con 315 aminoácidos, sin embargo, debido a complicaciones en la fase experimental y de recolección de datos, los datos reportados en los archivo PDB son incompletos. Todas las estructuras poseen la misma secuencia y sitios activos, excepto por Ser111 en las mutaciones, variando solo la estructura tridimensional.

En el siguiente capítulo se discutirán conceptos de teoría de grafos necesarios para modelar a PL^{pro} y sus estructuras variantes como una red de átomos de C_α. También se definirán los modelos de redes aleatoria y regular necesarios para determinar el grado de aleatoriedad y regularidad de la estructura de PL^{pro}.

Capítulo 3

Conceptos básicos de redes

La ciencia de redes es una ciencia que nos permite estudiar redes complejas, representaciones de sistemas complejos del mundo real. Su origen se remonta a 1741 con el surgimiento de la teoría de grafos, que a su vez aparece como consecuencia de la resolución del famoso "problema de los siete puentes de Königsberg" abordado por Euler (Euler, 1741). De hecho, la teoría de grafos no solo sentó las bases para la ciencia de redes, actualmente es una herramienta imprescindible para esta ciencia que nos permite modelar redes de cualquier tipo. No obstante, se consideró una ciencia independiente hasta apenas este siglo. Fue gracias a Erdös y Rényi con su artículo de redes aleatorias (Erdős y Rényi, 1959) que la ciencia de redes tuvo su *boom* en 2008.

Se dice que una red es compleja cuando sus propiedades globales no se pueden predecir a partir del comportamiento de sus elementos individuales. Es decir, no se puede predecir su estado más allá de un cierto intervalo de tiempo y/o espacio debido a que evolucionan dramáticamente su comportamiento. En ese sentido, la ciencia de redes busca expandir el conocimiento sobre estos tipos de redes a partir del análisis de la relación entre sus elementos (Boccaletti et al., 2006). Ejemplos de redes complejas son las redes de internet, las redes sociales, la *World Wide Web*, las redes celulares[1] y las redes ecológicas. El mo-

[1]Una red celular es aquella que surge de la interacción entre los componentes esenciales de células vivas, como los genes o las proteínas.

delado de estas redes es posible gracias a la teoría de grafos, un esquema que interpreta a los elementos de una red como nodos y a sus interacciones como enlaces (Barabási, 2013).

Se han encontrado similitudes entre distintos de estos tipos de redes aún al tratarse de distintos sistemas complejos. Es a lo que Barabási denomina *universalidad*, un concepto clave en la ciencia de redes que expone que las redes que exhiben comportamientos complejos convergen a un orden en común (Albert y Barabási, 2002). Algunos tipos de redes complejas muestran comportamientos oscilatorios entre redes aleatorias y redes regulares, más adelante trataremos este tema. Watts y Strogatz definieron como redes de mundo pequeño (SW) a estos tipos de redes (Watts y Strogatz, 1998). Barabási encontró otro patrón de distribuciones presentes en redes de interacción de proteínas (PPI), en la *World Wide Web* y en otras redes reales. A este segundo tipo los clasificó como redes libres de escala (Barabási, 2013). Con esto se pueden clasificar a las redes reales en redes SW o redes libres de escala. Existen más tipos pero estos son los más comunes.

Las redes de residuos de proteína (PRN[2]) suelen mostrar propiedades de redes SW, como es el caso de las proteínas *superóxido dismutasa* que cumplen el rol de antioxidante en las células, o la *enzima inactivadora de ribosoma* que detiene irreversiblemente la fabricación de otras proteínas esenciales en células de algunos tipos de hongos y plantas (Aguilar-Pineda y Olivares-Quiroz, 2021). Aunque diversos tipos de PRN suelen exhibir estas propiedades, es necesario un análisis riguroso para determinar si se pueden clasificar como tal dada su complejidad. Identificar el tipo de red nos permite determinar sus fortalezas y debilidades. Es decir, cómo se ve afectada dada la eliminación aleatoria o deliberada de alguno de sus elementos. En el caso de proteasas virales, conocer su estructura y dinámica nos permite desarrollar la estrategia adecuada para combatir un virus. En la sección 3.1 se definirán los conceptos necesarios para modelar a PL^{pro}, sus mutaciones y sus uniones con los inhibidores GRL-0617 y VIR250, como una PRN. En la sección 3.2 se definen los modelos de redes aleatorias y regulares para determinar si una red es SW.

[2] *PRN*: red que representa de la relación entre los aminoácidos de una proteína.

3.1. Teoría de grafos (Graph-theory)

La teoría de grafos es una rama de las matemáticas discretas que se ha aplicado a campos como la ingeniería, la sociología, la biología y la física (Barabási, 2013). Esta teoría busca representar en forma de grafo algún problema o fenómeno real para analizar las relaciones entre sus elementos. Su multidisciplinariedad se debe a que casi cualquier problema se puede representar en forma de grafo. Matemáticamente, un grafo $\mathcal{G}$ se define como un conjunto $\mathcal{N}$ de N nodos y un conjunto $\mathcal{E}$ de pares ordenados $\{(n_i, n_j)\}$ que representa los E enlaces entre los nodos n_i y n_j, es decir, $\mathcal{G} = \mathcal{G}(\mathcal{N}, \mathcal{E})$ (Newman, 2018).

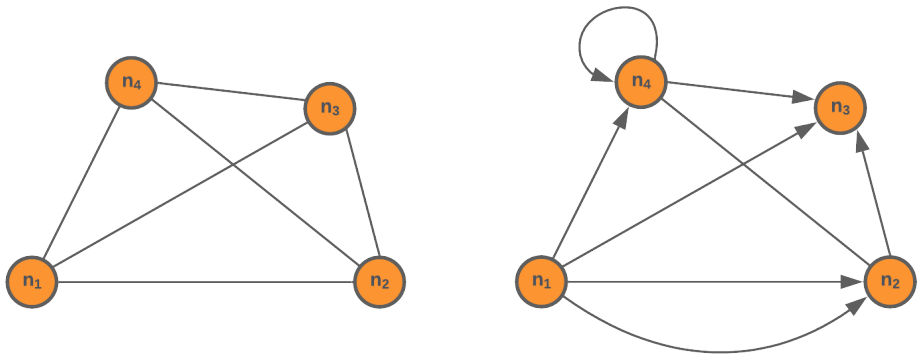
Los grafos se pueden clasificar, esencialmente, en dos categorías: por el sentido de sus enlaces y por su paralelismo (conexiones múltiples entre un par de nodos).

1.ª Categoría

- **Grafo dirigido:** También conocido como digrafo, es un tipo de grafo en el que los enlaces tienen una dirección, es decir, $\{n_i, n_j\} \neq \{n_j, n_i\}$, como es el caso del grafo de la figura 3.1b.
- **Grafo no dirigido:** En este tipo de grafo, como el de la figura 3.1a, el sentido de los enlaces no importa y, por lo tanto, $\{n_i, n_j\} = \{n_j, n_i\}$. Se consideran grafos simétricos.

2.ª Categoría

- **Grafo simple:** Es aquel en el que no existen las conexiones múltiples (paralelas) entre ninguno de sus nodos (figura 3.1a).
- **Multigrafo:** Los nodos pueden tener múltiples conexiones paralelas con otros nodos. Además puede contener bucles, es decir, conexiones de nodos consigo mismos como el grafo de la figura 3.1b.



(a) Grafo simple y no dirigido **(b)** Grafo dirigido múltiple

Figura 3.1: En (a) tenemos un grafo simple porque los nodos tienen, a lo mucho, una conexión con un mismo nodo. Es no dirigido ya que los enlaces no tienen dirección. En (b) el grafo es dirigido y multigrafo por el bucle en n_4 y la conexión doble entre n_1 y n_2.

3.1.1. Matriz de adyacencia

Un grafo $\mathcal{G}$ tienen asociada una representación matricial conocida como matriz de adyacencia $\mathbf{A}$, una matriz cuadrada de $N \times N$ que representa las conexiones entre los N nodos de un grafo. Se dice que dos nodos n_i y n_j son adyacentes cuando están conectados por un enlace, en cuyo caso se asigna un valor de uno en la celda a_{ij} de la matriz de adyacencia, y un valor de cero en el caso contrario (Estrada, 2012). Es decir,

$$a_{ij} = \begin{cases} 1 & \text{si} \quad (n_i, n_j) \in \mathcal{E} \\ 0 & \text{en otro caso.} \end{cases} \tag{3.1}$$

Una PRN se puede modelar como un grafo simple y no dirigido. Por lo tanto, se definen esferas centradas en cada nodo n_i y con radios R_c (radios de corte). Este radio representa la distancia de interacción promedio aproximada entre los aminoácidos de la proteína, de manera que aquellos nodos n_j adentro de la esfera son enlazados a n_i (ver figura 3.2) (Estrada, 2012). Además, se define una variable R_{ij} como la distancia euclidiana entre los nodos n_i y n_j, de manera que la matriz de adyacencia para una PRN equivale a:

$$A_{ij} = \begin{cases} H(R_c - R_{ij}) & i \neq j, \\ 0 & i = j. \end{cases} \tag{3.2}$$

La función de Heaviside **H** asigna un valor de 1 si se cumple la condición de que la distancia entre los nodos n_i y n_j distintos estén a una distancia menor a R_c, y un valor de 0 en caso contrario. La matriz de adyacencia de esta red es simétrica con ceros en su diagonal principal (Estrada, 2012).

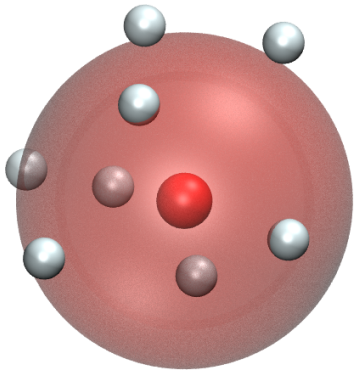


Figura 3.2: Ilustración de la esfera con radio R_c centrada en un nodo de la red. Dicho nodo será enlazado a los nodos que se encuentren en el interior de la esfera.

Como ejemplo, consideremos el grafo de árbol simple y no dirigido de 8 nodos y 7 enlaces de la figura 3.3. Su diagonal principal contiene ceros dado que no tiene bucles, además, es simétrica ya que sus enlaces no tienen dirección. Las definiciones de los capítulos siguientes se centran en grafos simples y no dirigidos. En resumen, en la PRN los nodos representan a los aminoácidos de la proteína y los enlaces representan las interacciones covalentes y no covalentes entre estos.

$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

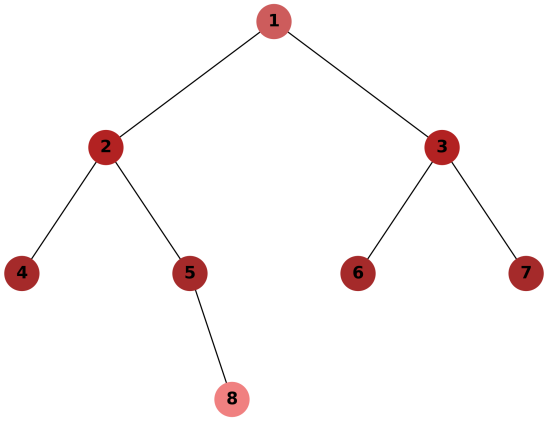


Figura 3.3: Grafo de árbol no dirigido (derecha) y su matriz de adyacencia asociada (izquierda). Una representación alternativa de un grafo es mediante su lista de enlaces, en este caso esta lista es: $\{(1,2),(1,3),(2,4),(2,5),(3,6),(3,7),(5,8)\}$

3.1.2. Medidas básicas de redes

La medida más básica de una red es el grado promedio $\langle k \rangle$. Consiste en el número de enlaces promedio y se obtiene como $2E/N$. Se considera que una red es dispersa cuando $\langle k \rangle \ll N$. A partir de esto, se calcula la densidad δ de una red que mide la proporción entre el total de nodos y el total de enlaces (Estrada, 2012). Se calcula como

$$\delta = \frac{\langle k \rangle}{(N-1)}. \tag{3.3}$$

Una red puede describirse a partir de dos métricas adicionales: la longitud de ruta promedio y el coeficiente de agrupamiento. Estas medidas se definirán a continuación y se profundizarán en la sección 3.2.

Longitud de ruta promedio

La longitud de ruta promedio se define como la separación promedio en aristas entre los nodos n_i y n_j. Para un grafo simple y no dirigido se calcula como (Barabási, 2013)

$$\langle L \rangle = \frac{1}{N(N-1)} \sum_{i<j} d_{ij}. \tag{3.4}$$

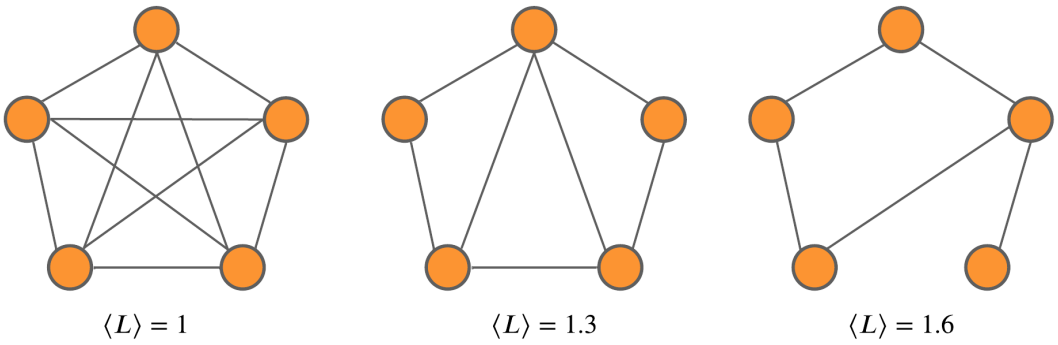


Figura 3.4: Variación de la longitud de ruta promedio $\langle L \rangle$ conforme varía el número de enlaces para un grafo de cinco nodos. Entre más aumenta el total de enlaces, más aumenta la accesibilidad entre un par de nodos de la red. El valor mínimo de $\langle L \rangle$ que puede tener cualquier tipo de red, siempre que no existan nodos aislados del resto, es 1. En este punto, todos los nodos están a una distancia de un arista de cualquier otro nodo de la red.

.

La variable N representa el total de nodos de la red y las distancias entre los nodos n_i y n_j están contenidos en la matriz de *distancias* d_{ij}. Esta matriz es simétrica, al igual que la matriz de adyacencia, por lo que la suma se calcula sobre los elementos arriba de la diagonal principal. Una longitud de ruta promedio alta denota una red en donde sus nodos están altamente separados entre sí (ver ejemplo de la figura 3.4) (Barabási, 2013).

Coeficiente de agrupamiento

Esta medida evalúa la frecuencia ν con la que los nodos adyacentes de n_i, con grado k_i, se vinculan a través de un enlace. Es decir, contabiliza las veces en las que el nodo n_i forma triángulos con otros nodos de la red. Para un grafo simple y no dirigido, el coeficiente de agrupamiento promedio tiene la siguiente expresión (Barabási, 2013)

$$\langle C \rangle = \frac{1}{N} \sum_{i=1}^{N} \frac{2\nu_i}{k_i(k_i - 1)}. \tag{3.5}$$

La frecuencia ν_i se calcula con un algoritmo computacional descrito en el capítulo 5. Los valores que puede tomar el coeficiente de agrupamiento rondan entre 0 y 1. Consideremos el ejemplo de la figura 3.5, mientras que un valor de 0 indica un nulo nivel de agrupamiento en donde los vecinos de n_i no están enlazados, un valor de 1 indica que todos los nodos de la red están conectados entre sí (Barabási, 2013).

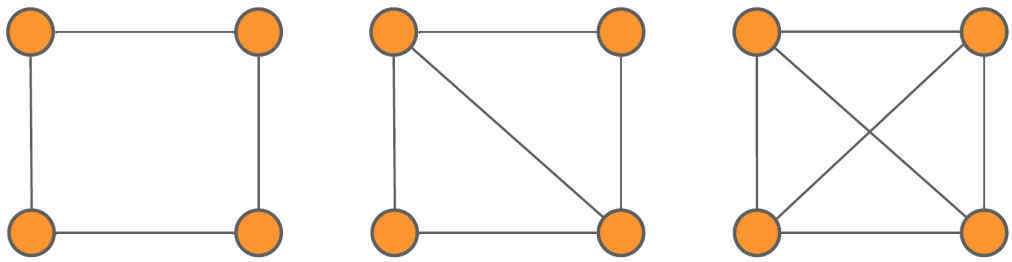


Figura 3.5: El coeficiente de agrupamiento promedio toma valores para $0 \leq \langle C \rangle \leq 1$. En el extremo inferior, los vecinos de cada nodo n_i de la red no están conectados, mientras que en el extremo superior todos los nodos están conectados entre sí, se forma la mayor cantidad de triángulos posibles en la red y la densidad de enlaces es más alta.

.

3.2. Propiedad de mundo pequeño (*small world*) en redes complejas

En la década de los 60 se popularizó el "problema de mundo pequeño", un problema que consistió en determinar la distancia de separación promedio (en personas) entre dos personas completamente desconocidas. Milgram abordó el problema desde un enfoque experimental. Para este experimento seleccionó aleatoriamente a 160 personas de una región de los EE.UU. y un segundo grupo de 160 personas de otra región lejana a la primer región. El objetivo fue trasmitir un mensaje del primer al segundo grupo a través de personas intermediarias, de manera que se formaran 160 cadenas. Es decir, los emisores debían de transmitir la carta a una segunda persona que consideraran más cercana a la persona objetivo, y así sucesivamente. Una de las restricciones era que conocieran a la segunda persona al menos por su nombre de pila (Milgram, 1967).

En este estudio, en donde participaron personas de los 200 millones de personas de los EE. UU., se encontró una distancia de separación media de seis personas. Por esta razón se le conoció como el experimento de "seis grados de separación". Este estudio se consideró no concluyente ya que de 160 cadenas solo se completaron 44. La distancia de separación promedio actualmente es conocida como longitud de ruta promedio. Por otro lado, el término "propiedad de mundo pequeño" hace referencia a una longitud de ruta promedio baja. De este problema surge la expresión "¡Qué pequeño es el mundo!" (Milgram, 1967).

Años después, Watts y Strogatz diseñaron su propio modelo de redes conocido como el "Modelo de Watts y Strogatz" con el que reprodujeron la propiedad de mundo pequeño descrita por Milgram en su experimento. Al indagar en redes teóricas encontraron que las redes aleatorias (redes cuyos enlaces se agregan progresivamente con una probabilidad que depende del estado actual de cada nodo) tienen la propiedad del mundo pequeño, sin

embargo, un coeficiente de agrupamiento bajo. Por otro lado, las redes regulares (todos sus nodos tienen el mismo grado de nodo) ofrecían el coeficiente de agrupamiento alto pero una longitud de ruta promedio alta. A partir de este fundamento, diseñaron el modelo de Watts y Strogatz como una combinación de redes regulares y aleatorias para obtener una red con una longitud de ruta promedio baja y un alto nivel de agrupamiento. A esta red la definieron red de mundo pequeño (SW) (Watts y Strogatz, 1998).

Retomando el ejemplo de una red social en donde los nodos representan a las personas y los enlaces las conexiones entre estas, la longitud de ruta promedio baja se relaciona con una gran cercanía entre un par de personas de esta red. Dos personas pueden comunicarse entre sí rápida y eficientemente. Por otro lado, el coeficiente de agrupamiento alto se relaciona con una alta cantidad de regiones densamente conectada, en donde un grupo de personas están altamente vinculados unos con otros, como los miembros de una familia o de una comunidad pequeña. A estas regiones densamente conectadas se les conoce como *clusters*. En general, estas dos propiedades se relacionan con una mayor eficiencia en la comunicación entre los nodos y una mayor robustez de la red (Barabási, 2013). Watts y Strogatz incentivaron a estudiar la propiedad del mundo pequeño como parte del análisis de redes ya que esta propiedad es muy común en redes reales, principalmente en redes tecnológicas y biológicas.

Se define formalmente como red SW a aquella que cumple las siguientes propiedades:

- Una longitud de ruta promedio $\langle L \rangle$ que se aproxima como $\ln(N)/\ln(\langle k \rangle)$, con $\ln(N) \ll N$. En otras palabras, este valor es proporcional logarítmicamente al tamaño de la red e inversamente proporcional al logaritmo del grado de nodo promedio de una red aleatoria. Entre más densa es una red, la distancia promedio entre un par de nodos disminuye. Se dice que $\langle L \rangle$ es "pequeña" cuando se cumple con la dependencia logarítmica del tamaño del sistema (Barabási, 2013).

- Un coeficiente de agrupamiento promedio $\langle C \rangle$ alto en comparación con el de una

red aleatoria de tamaño similar. En el caso de las redes reales, este coeficiente no depende del tamaño del sistema (Barabási, 2013).

A menudo, las PRN exhiben comportamientos de redes SW. Con el fin de determinar si una red es de este tipo, es necesario generar redes aleatorias y regulares equivalentes para compararlas con la red de interés. Para esto, se empleará el modelo de Erdős-Rényi para la red aleatoria y el modelo de Watts-Strogatz para la red regular. A continuación se definirán los modelos de redes para este propósito.

3.2.1. Modelo de Erdős-Rényi

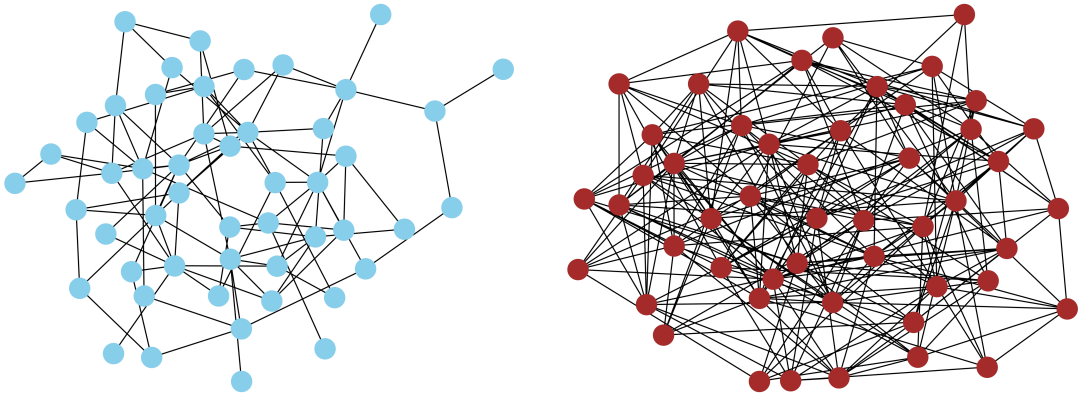
Los primeros en indagar en modelos de redes teóricas fueron los matemáticos Paul Erdős y Alfréd Rényi (Erdős y Rényi, 1959). Actualmente son conocidos como los fundadores de la teoría de redes aleatorias. Por esta razón, a las redes aleatorias también se les conoce como "Redes de Erdős-Rényi".

Aunque algunas redes reales son complejas e incluso pueden llegar a ser caóticas, esto no equivale a aleatoriedad. Sin embargo, pueden mostrar comportamientos de aleatoriedad en ciertos aspectos. El modelo de redes aleatorias nos permite cuantificar el grado de aleatoriedad de redes reales. Anteriormente se definió un grafo $\mathcal{G}$ como un conjunto de N nodos y E enlaces. En el modelo de red aleatoria existen dos definiciones distintas, una versión que depende de una probabilidad p y una segunda que depende del número de enlaces E.

- $\mathcal{G}_{N,p}$: A este modelo se le conoce como modelo de red aleatoria con probabilidad p. Consiste en una red de N nodos, en donde un par de nodos tienen una probabilidad p de conectarse. Cuando la probabilidad p es uno obtenemos un grafo completo, es decir, un grafo en donde todos los nodos están conectados entre sí. Este tipo de red tiene grado promedio $\langle k \rangle = p(N-1)$ (Frieze y Karoński, 2016).

- $\mathcal{G}_{N,E}$: Este es el modelo de red aleatoria con número fijo de enlaces. Se construye a

partir de un conjunto de N nodos inicial y se añaden conexiones E veces entre pares de nodos de la red que se eligen de forma aleatoria, con $0 \leq E \leq \binom{N}{2}$. El coeficiente binomial $\binom{N}{2}$ representa el número máximo de enlaces. Como ilustración de este tipo de red aleatoria considere los grafos de la figura 3.6. El grado promedio para este tipo de red aleatoria es $\langle k \rangle = \frac{2E}{N}$ (Frieze y Karoński, 2016).



(a) $E = 100$ **(b)** $E = 250$

Figura 3.6: Redes aleatorias generadas con el modelo de red aleatoria con número fijo de enlaces. Ambas tienen el mismo número de nodos $N = 50$, variando solo el número de enlaces E.

La diferencia principal entre ambos modelos es que en el segundo se elige el número de enlaces E como parámetro y en el primero el número de enlaces es una variable que depende de p. La relación entre E y p equivale a $E = pN(N-1)/2$. Esta diferencia en sus parámetros se refleja en sus distribuciones de grado (distribuciones que cuantifican la probabilidad de que un nodo de una red tenga grado k_j). Mientras que la red aleatoria con probabilidad p sigue una distribución binomial, la red aleatoria con número fijo de enlaces E sigue una distribución hipergeométrica para $N \ll \langle k \rangle$. Cuando $N \gg \langle k \rangle$, ambas distribuciones se aproximan a una distribución de Poisson. En esta investigación se empleará el modelo de red aleatoria con número fijo de enlaces. La inclinación por un modelo u otro no tiene efectos significativos en los resultados, ya que por la naturaleza de la investigación $N \gg \langle k \rangle$. Las definiciones a continuación se centrarán en este tipo de red aleatoria (Frieze y Karoński, 2016).

Distribución de grado

Sea k una variable aleatoria que representa el total de éxitos en E ensayos binarios (de éxito o fracaso) sin reemplazo realizados sobre una población finita de tamaño $\binom{N}{2}$ y con $N-1$ como el tamaño de la muestra de la categoría de interés. La probabilidad de obtener $K = k$ sigue la distribución hipergeométrica

$$p(k) = \frac{\binom{N-1}{k}\binom{\binom{N-1}{2}}{E-k}}{\binom{\binom{N}{2}}{E}}. \tag{3.6}$$

Es decir, la probabilidad de que un nodo de la red aleatoria con número fijo de enlaces E tenga grado k. El tamaño de la población es el número de enlaces máximos de la red y el tamaño de la muestra de interés es el grado máximo que puede tener un nodo. Para un valor de $N \gg \langle k \rangle$, de forma que el tamaño de la población es grande en contraste con el tamaño de la muestra E, la distribución hipergeométrica se aproxima a la distribución de Poisson (Frieze y Karoński, 2016).

La distribución de Poisson modela la probabilidad de que ocurran x eventos independientes en un cierto intervalo de espacio o tiempo definido a una tasa promedio constante λ. En una red aleatoria, la tasa de ocurrencias promedio es $\langle k \rangle$, entonces, la probabilidad de que un nodo tenga grado k es

$$p(k) = \frac{e^{-\langle k \rangle}\langle k \rangle^k}{k!}. \tag{3.7}$$

La distribución de Poisson tiene las siguientes propiedades importantes, para esto, consideremos el ejemplo de la figura 3.7: (a) la probabilidad de que un nodo tenga grado k es máxima cuando $k = \langle k \rangle$, esta probabilidad decae exponencialmente conforme k se aleja de $\langle k \rangle$; (b) entre más densa es una red, más aumentan los posibles valores de k que puede tomar un nodo. Por lo tanto, conforme aumenta la densidad de la red disminuye la curtosis, y $\langle k \rangle$ se desplaza a la derecha. En síntesis, en las redes aleatorias la mayoría de los nodos tienen grado cercano a $\langle k \rangle$ y solo una pequeña parte tienen un grado k bajo

y alto. En consecuencia, estos tipos de redes son capaces de resistir ataques aleatorios ya que la probabilidad de eliminar aleatoriamente a alguno de sus nodos más influyentes es baja (Barabási, 2013).

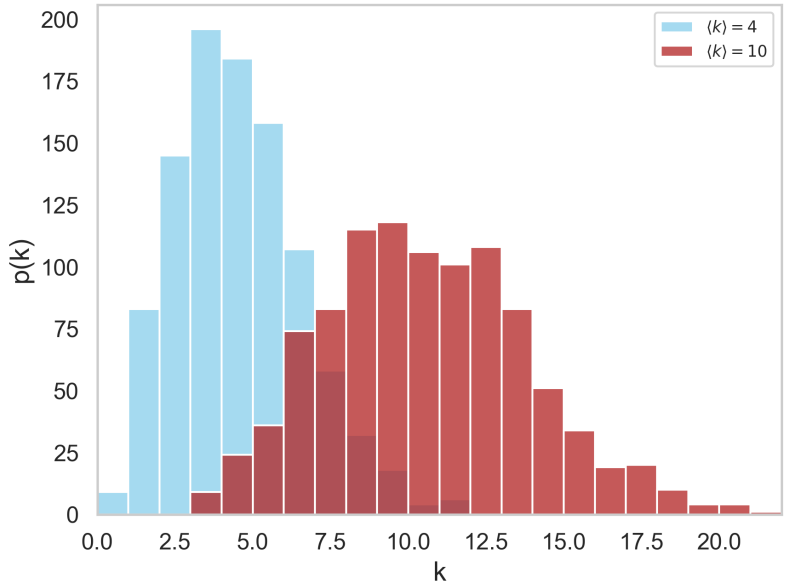


Figura 3.7: Distribuciones de Poisson para las redes aleatorias con grado promedio $\langle k \rangle = 4$ y $\langle k \rangle = 10$. Estas distribuciones tiene un pico en $\langle k \rangle$.

.

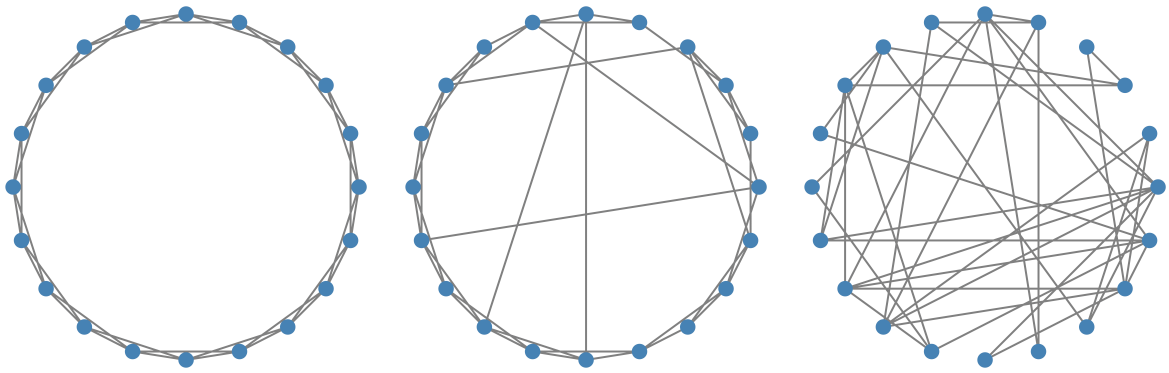
3.2.2. Modelo de Watts-Strogatz

El modelo de Watts y Strogatz $\mathcal{G}_{N,k,p}$ se considera una extensión de la red aleatoria al generar conexiones entre los nodos de la red con una probabilidad p. Es decir, una propabilidad de que un nodo se desconecte y reconecte a otro nodo de la red. Además de p, este modelo toma como parámetros a la cantidad de nodos N y el grado promedio k de la red. La idea principal de este modelo es generar una red SW, es decir, una red con $\langle L \rangle$ pequeño y $\langle C \rangle$ grande para valores fijos de N y k. Este modelo consiste en una red inicial regular de N nodos con grado promedio $\langle k \rangle$. Luego, se añaden conexiones con una probabilidad p de enlazarse con otros nodos de forma aleatoria (Barabási, 2013). Con este modelo se obtienen tres resultados destacables para distintos valores de p:

- Para $p = 0$. En este extremo se genera una red regular, un tipo de red en donde todos los nodos tiene el mismo grado de conexiones. En este caso en particular, la red regular se caracteriza por conectar a los nodos de la red con sus vecinos contiguos. Tanto $\langle C \rangle$ como $\langle L \rangle$ son altos (figura 3.8a).

- Para $p \approx 0$. Se mantiene el alto nivel de agrupamiento $\langle C \rangle$ mientras que la distancia promedio $\langle L \rangle$ entre los nodos disminuye. En este punto se considera que la red es SW (ver figura 3.8b).

- Para $p = 1$. Se genera una red aleatoria con valores bajos de $\langle C \rangle$ y $\langle L \rangle$ (figura 3.8c).



(a) Red regular **(b)** Red de mundo pequeño **(c)** Red aleatoria

Figura 3.8: Modelo de Watts y Strogatz para distintos valores de p.

En conclusión, este modelo toma la baja longitud de ruta promedio de la red regular y, por el otro, el alto nivel de agrupamiento característico de las redes aleatorias. Genera una red SW a partir de la interpolación entre una red regular y una red aleatoria. Dicho de otro modo, algunas redes del mundo real exhiben comportamientos de aleatoriedad y regularidad de forma simultánea.

En el siguiente capítulo se definirán las medidas de centralidad necesarias para determinar las variaciones a nivel local y global de una proteína, haciendo énfasis en medidas de segundo y tercer orden.

Capítulo 4

Medidas de centralidad de orden superior

A lo largo de los años se han analizado diversos tipos de redes reales y se ha encontrado que algunas de ellas siguen una misma tendencia estadística. Una de estas tendencias es la propiedad de mundo pequeño, como vimos en el capítulo anterior, la segunda se relaciona con los motivos de red o *network motifs*, estructuras geométricas que aparecen repetidas veces dentro de una red. Como hemos visto anteriormente, un grafo es un conjunto de N nodos y E enlaces, a partir de los cuales se forman subconjuntos más reducidos que generan distintas figuras geométricas (triángulos, cuadriláteros y demás figuras irregulares cuya forma es irrelevante). Estas figuras contenidas dentro de un grafo de mayor tamaño se conocen como subgrafos. Por otro lado, aquellos subgrafos de un cierto orden que aparecen con mayor frecuencia en una red que otros subgrafos de otros órdenes se les conoce como *network motifs* (ver figura 4.1) (Milo et al., 2002).

El término *network motifs* surgió a partir de investigaciones de Milo *et. al.* (Milo et al., 2002). En su artículo de 2002 estudiaron diversas redes tecnológicas y biológicas, entre las que se encuentran las redes genéticas de la bacteria *Escherichia coli* (presente en la microbiota de los intestinos de seres vivos), la *World Wide Web*, redes alimentarias (redes que representan la relación entre predadores y depredadores en comunidades del mundo

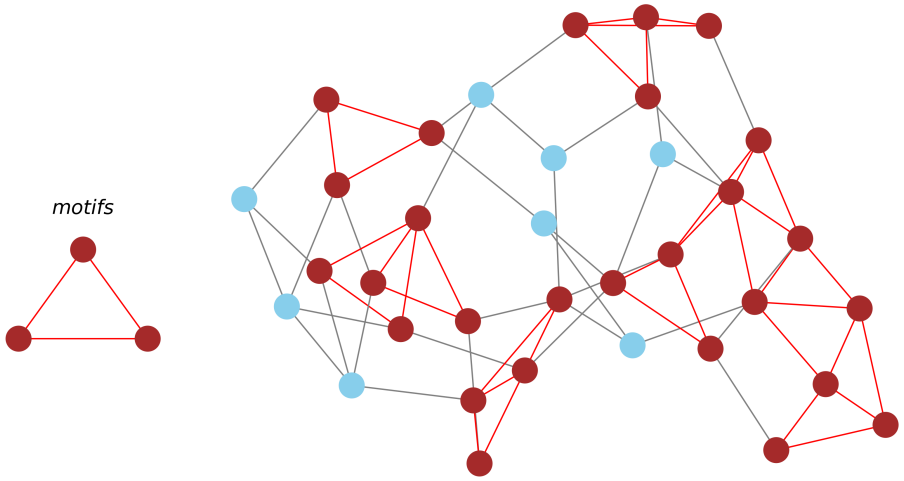


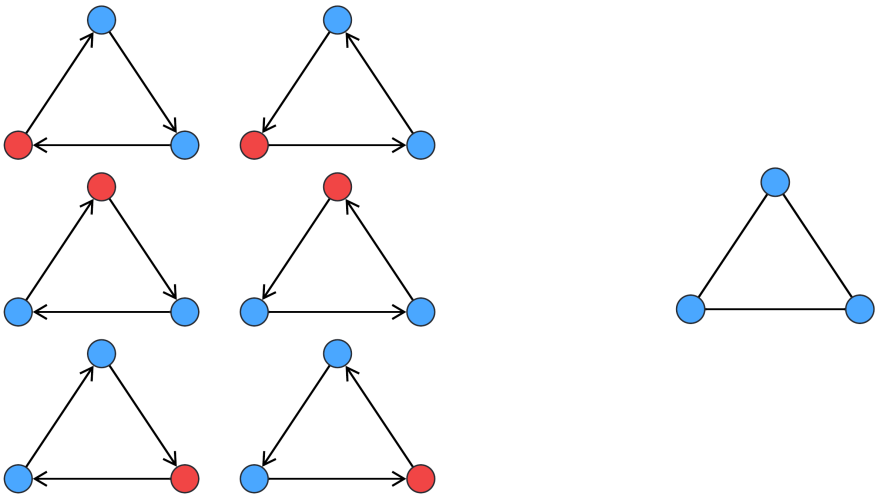
Figura 4.1: *Network motifs* en una red real. Los *network motifs* son figuras geométricas repetitivas dentro de un grafo, estas figuras pueden ser triángulos, cuadriláteros o cualquier otra figura con más aristas.

animal), entre otras. En este estudio encontraron estos patrones geométricos y repetitivos en todas las redes biológicas y de internet mencionadas. Lo remarcable de esto es que los elementos más importantes de estas redes intervienen en gran medida en los *network motifs*. Además, encontraron que los *network motifs* aparecen con menor frecuencia en redes aleatorias, por lo que estas propiedades pueden definirse como una caracterítica propia de redes reales (Milo et al., 2002).

Hasta la fecha la cuantificación de subgrafos en una red para identificar *network motifs* es un gran reto, tanto en términos de precisión como de eficiencia (es un problema computacionalmente costoso). Estrada y Rodríguez Velázquez se encuentran entre los científicos que han contribuido significativamente en investigaciones entorno a *network motifs* desde un enfoque de teoría de redes y medidas de centralidad. Las medidas de centralidad son herramientas matemáticas que forman parte de la ciencia de redes y nos permiten ubicar a nodos clave en una red. Por ejemplo, la centralidad de grado es una medida de centralidad estándar que categoriza a los nodos a partir del número de enlaces. Por ende, el nodo con el mayor número de enlaces será el nodo más central. Para identificar *network motifs* definieron una nueva medida de centralidad: la centralidad de subgrafo. Esta medida

consiste en cuantificar las veces en las que cada nodo de la red participa en subgrafos de todos los órdenes posibles, dando más importancia a los de menor tamaño (Estrada y Rodriguez-Velazquez, 2005).

Para cuantificar subgrafos emplean los conceptos de *camino* y *camino cerrado*. Un *camino* es un recorrido que atraviesa a todos los nodos de un conjunto $\{n_1, n_2, ..., n_u, n_{u+1}\}$ de nodos que inicia y finaliza en nodos distintos. En contraste, los *caminos cerrados* inician su recorrido y lo finalizan en el mismo nodo, es decir, $n_1 = n_{u+1}$. La variable u representa la longitud del *camino cerrado*. De este modo, los *caminos cerrados* de orden tres forman un subgrafo de grado 3 (un triángulo). Notemos que ambos son conceptos matemáticos distintos. Un subgrafo puede ser recorrido de diferentes formas, además, en la trayectoria se puede repetir nodos y enlaces. Esta es una de las limitaciones de la centralidad de subgrafo, cuantifica *caminos cerrados* (Estrada y Rodriguez-Velazquez, 2005).



(a) Caminos cerrados **(b)** Subgrafo

Figura 4.2: Un *camino cerrado* es un recorrido a través de un conjunto de nodos y enlaces (subgrafo) con la única restricción que inicie y finalice en el mismo nodo. Así, un subgrafo de grado tres (triángulo) puede ser recorrido de seis formas distintas al iniciar en tres nodos distintos (nodos en rojo) y con dos direcciones diferentes.

Además de la centralidad de grado y de subgrafo existen otras medidas de centralidad importantes. Entre estas medidas se encuentran: la centralidad de intermediación, que calcula la frecuencia con la que un nodo es intermediario en la comunicación entre otro par de nodos en la ruta más corta; la centralidad de eigenvector, que atribuye una mayor importancia a un nodo a partir de la densidad de enlaces de sus vecinos; entre otros.

En 2020, Estrada clasificó a las medidas de centralidad existentes por la manera en la que evalúan la participación de los nodos a través de distintas rutas. Catalogó a la centralidad de grado y de intermediación como medidas de primer orden debido a que consideran la influencia de un nodo en sus vecinos contiguos y/o aquellos que se comunican solo a través de las rutas mínimas. Por otro lado, a la centralidad de subgrafo y de comunicabilidad las clasificó como medidas de segundo orden, ya que consideran la comunicación entre cada par de nodos a cualquier distancia y por cualquier ruta posible. Finalmente, definió como medidas de tercer orden a la centralidad de subgrafo y de comunicabilidad de largo alcance. Ambas son variantes de la centralidad de subgrafo y de comunicabilidad de segundo orden que sancionan mucho menos las rutas más largas que las medidas de segundo orden (Estrada, 2020).

Volviéndo al objetivo principal, este consiste en determinar si existe alguna relación entre los sitios activos y de unión en PL^{pro} con las medidas de segundo y tercer orden. Estas medidas representan una forma de identificar *network motifs*, que a su vez reflejan características importantes en muchas redes biológicas y se relacionan con elementos destacables de una red. Entonces nos podrían dar información sobre lo sitios activos a través de sus máximos y mínimos.

A continuación se definirán las medidas de centralidad y su evolución de medidas de primer orden hasta tercer orden. Se discutirá el posible significado físico de los *network motifs* en las PRN.

4.1. Medidas de centralidad de primer orden

Hoy en día, las medidas de centralidad se aplican a una amplia variedad de campos, sin embargo, en un principio eran dirigidos mayormente al campo de las ciencias sociales. A finales de los años 40 nació el primer grupo de investigación que se dedicó al desarrollo de la teoría de centralidades. Este grupo centró sus investigaciones a la resolución de problemas de comunicación y liderazgo. Como consecuencia de ello, Bavelas determinó que una persona se encuentra en una posición central cuando es intermediaria en la comunicación de otro par de personas, en particular, en la ruta más corta. A partir de esto, surgió el término "centralidad de nodo" y eventualmente se formularon las que hoy conocemos como medidas de centralidad (Freeman, 1978).

En esta categoría, la primer medida de centralidad es la **centralidad de grado**. La centralidad de grado de un nodo n_i se define como el total de k conexiones que tiene con otros nodos. Desde esta perspectiva, el nodo más importante es aquél con el grado de nodo más alto. Sea a_{ij} el (i,j)-ésimo elemento de la matriz de adyacencia $\mathbf{A}$, la centralidad de grado del nodo n_i para un grafo no dirigido se calcula como (Estrada, 2012)

$$k_i = \sum_j a_{ij} \tag{4.1}$$

Freeman propuso un conjunto de nuevas medidas de centralidad basadas en la "centralidad de nodo" expuesta por Bavelas. Generalizó la noción de nodo intermediario como aquel que se interpone en el camino geodésico entre otros pares de nodos. Aquí, la distancia entre dos nodos es el número de enlaces contenidos en la ruta que une a ambos. Sean n_i y n_j dos nodos distintos alcanzables entre sí, exite una cantidad finita de geodésicas ρ_{ij} entre ambos. A esto se le conoce como grado de heterogeneidad, esta propiedad se calcula a partir de la centralidad de grado como (Freeman, 1977)

$$\rho = \sum_{(i,j \in E)} \left(k_i^{-1/2} - k_j^{-1/2} \right)^2 . \tag{4.2}$$

Este valor se puede normalizar como $\frac{\rho}{N-2\sqrt{N-1}}$ de manera que $0 \leq \rho \leq 1$. Cuando $\rho = 0$ nos encontramos ante una red regular, es decir, una red en donde todos los nodos tienen el mismo grado de nodo. Por otro lado, para $\rho = 1$ nos encontramos ante una red heterogénea con grados de nodo muy variables (Freeman, 1977).

A partir del grado de heterogeneidad, Freeman definió la **centralidad de intermediación** del nodo n_k como la razón entre el total de geodésicas ρ_{ij} que une a los nodos n_i y n_j que contienen a n_k, y el total de sus geodésicas que no necesariamente contiene a n_k. Por lo tanto, la centralidad de intermediación de un nodo n_i es baja cuando dicho nodo no participa significativamente en la comunicación entre otro par de nodos distintos. Para el valor promedio se divide la centralidad de intermediación entre el total de N nodos (Freeman, 1977).

$$\langle C_{\mathrm{B}} \rangle = \frac{1}{N} \sum_{i \neq j \neq k} \frac{\rho_{ikj}}{\rho_{ij}}. \tag{4.3}$$

A modo de conclusión, las medidas de primer orden reflejan la influencia de los nodos de la red en su entorno más próximo, tanto en los nodos adyacentes como aquellos que están a su alcance a través de las rutas geodésicas.

4.2. Medidas de centralidad de segundo orden

4.2.1. Centralidad de eigenvector

La centralidad de eigenvector fue introducida por Bonacich en 1972 como una extensión de la centralidad de grado. Introdujo esta nueva medida de centralidad como una forma de cuantificar el "poder" en los elementos de una red social (Bonacich, 1987).

Aunque nuestra intuición nos dice que una persona con una gran cantidad de conexiones sociales es la que tiene más "poder" y, por lo tanto, la más central, esto no ocurre exactamente así. Una persona tiene más poder por la calidad de sus relaciones y no por

la cantidad. Por esta razón, Bonacich definió esta nueva medida de centralidad que considera la relevancia de un miembro de una red con respecto a la relevancia de los miembros vecinos. En otras palabras, cuantifica la importancia de un nodo con respecto al grado de nodo k de sus nodos vecinos ponderada por su propia centralidad (Bonacich, 1987).

Para calcular esta centralidad, Bonacich asigna un vector centralidad $\bar{x}'$ al conjunto de nodos del grafo no dirigido $\mathcal{G}$ y lo calcula como el producto entre la matriz de adyacencia $\mathbf{A}$ y el vector centralidad $\bar{x}_0$ inicial de los nodos de la red

$$\bar{x}' = \mathbf{A}\bar{x}_0. \tag{4.4}$$

Después se realizan s iteraciones hasta que el vector centralidad converja a un valor estable, es decir, la centralidad inicial se reemplaza por la centralidad anterior s pasos

$$\bar{x}_s = \mathbf{A}^s\bar{x}_0. \tag{4.5}$$

Dado que $\mathbf{A}$ es la matriz de adyacencia asociada al grafo no dirigido $\mathcal{G}$, es simétrica y sus eigenvalores son reales. Entonces se puede expresar la ecuación de valor propio $\mathbf{A}\bar{x}_o = \lambda\bar{x}_0$, para algún valor propio λ y vector propio $\bar{x}_0$ de $\mathbf{A}$. Por conveniencia, se elige $\bar{x}_0$ como la combinación lineal de los eigenvectores v_i de $\mathbf{A}$ con eigenvalores λ_i: $\bar{x}_0 = \sum_i c_i v_i$, con constantes c_i adecuadas. Así (4.5)

$$\mathbf{A}^s\bar{x}_0 = \mathbf{A}^s\sum_i c_i v_i = \sum_i c_i(\lambda_i)^s v_i = (\lambda_1)^s\sum_i c_i\left(\frac{\lambda_i}{\lambda_1}\right)^s v_i, \tag{4.6}$$

con λ_1 como el eigenvalor más grande de $\mathbf{A}$ con eigenvector v_1. La razón entre los eigenvalores de $\mathbf{A}$ y el eigenvalor más grande decae exponencialmente para $s \gg 1$ y se puede aproximar $\mathbf{A}^s\bar{x}_0 \sim \lambda_1^s c_1 v_1 = \lambda_1^s\bar{x}_1$. Esto quiere decir que la centralidad definida como $\bar{x}'$ después de s pasos, digamos $\bar{x}$, es proporcional al eigenvalor más grande de la matriz de adyacencia $\mathbf{A}$, es decir, $\mathbf{A}\bar{x} = \lambda_1\bar{x}$. Finalmente, despejando el vector centralidad se obtiene

$$C_{\mathrm{E}}(i) = \lambda_1^{-1} \sum_i a_{ij} x_j, \tag{4.7}$$

donde a_{ij} es el (i, j)-ésimo elemento de $\mathbf{A}$ y x_j el j-ésimo elemento de $\bar{x}$. Esta es la **centralidad de eigenvector** definida por Bonacich, el eigenvector asociado al eigenvalor más grande de la matriz de adyacencia. Por el teorema de Perron-Frobenius se demuestra que la centralidad de eigenvector es positiva. Este teorema asegura dos cosas: (a) dado que λ_1 es el eigenvalor más grande de $\mathbf{A}$, cuadrada y no negativa, entonces $\lambda_1 > 0$, y (b) existe al menos un eigenvector positivo asociado a λ_1. En conclusión, el propósito de la centralidad de eigenvector es otorgar una mayor importancia a un nodo cuando se relaciona con otros nodos importantes (Bonacich, 1987).

Estrada aportó una nueva perspectiva de la centralidad de eigenvector a partir del concepto de *camino* definido anteriormente. La longitud de este recorrido es el total de u enlaces que atraviesa. Con base a esto, este medida evalúa la relación entre el total de *caminos* de longitud infinita que inician su recorrido en un nodo n_i $(M_u(i))$ y el resto de *caminos* que inician en otro nodo n_j $(M_u(j))$ (Estrada, 2020). Es decir,

$$C_{\mathrm{E}}(i) = lim_{u\to\infty} \frac{M_u(i)}{\sum_{j=1}^{N} M_u(j)} \tag{4.8}$$

Se ha encontrado que esta centralidad es útil para la identificación de residuos aminoácidos críticos en mecanismo alostéricos en proteínas (Negre et al., 2018) y para la predicción de sitios activos en enzimas. Particularmente, los sitios activos en proteínas se relacionan con los valores altos en la centralidad de eigenvector y de intermediación, sin embargo, se requieren de métricas adicionales para un análisis más preciso y profundo (Aguilar-Pineda y Olivares-Quiroz, 2021).

4.2.2. Centralidad de subgrafo

Como hemos visto, Estrada y Rodríguez Velázquez introdujeron la centralidad de subgrafo para identificar *network motifs* y como complemento a las medidas de centralidad de primer orden. En el mismo artículo aplicaron esta centralidad en el análisis de distintos tipos de redes reales, entre las que se encuentran las redes de levadura *Saccharomyces cerevisiae* (levadura de la cerveza), de la bacteria *Helicobacter pylori*, del diccionario en línea ODLIS, entre otros. El hallazgo más importante de esta investigación fue que aquellos elementos importantes de cada una de estas redes participan activamente en estos subgrafos (Estrada y Rodriguez-Velazquez, 2005).

Definición de subgrafo

Un grafo $\mathcal{G}'(\mathcal{N}', \mathcal{E}')$ se considera subgrafo de otro grafo $\mathcal{G}(\mathcal{N}, \mathcal{E})$ cuando el conjunto de enlaces $\mathcal{E}'$ y nodos $\mathcal{N}'$ están contenidos en $\mathcal{E}$ y $\mathcal{N}$, respectivamente. Un subgrafo considera formas tanto abiertas como cerradas. Por ejemplo, un triángulo es una forma cerrada, sin embargo, si un par de vértices de esta figura no están enlazados entonces estamos ante una figura abierta. Mientras que los *caminos cerrados* cuantifican subgrafos cerrados, los *caminos* cuantifican subgrafos abiertos.

Consideremos el ejemplo de la levadura de la cerveza *Saccharomyces cerevisiae*. Esta levadura es un tipo de hongo que transforma el azúcar en alcohol (proceso conocido como fermentación) a partir de un conjunto de proteínas enzimáticas en su estructura. Entonces, las proteínas de esta levadura se pueden clasificar en proteínas esenciales (las enzimáticas que participan activamente en el proceso de fermentación) y menos esenciales (las no enzimáticas). La eliminación de alguna de las proteínas esenciales puede tener efectos letales en este organismo. Para estudiar este sistema, Estrada y Rodríguez Velázquez interpretaron a las proteínas que componen este organismo como nodos y a las relaciones entre estos como enlaces, a esto se le conoce como red de interacción de proteínas (PPI, por sus siglas en inglés). Con la centralidad de subgrafo cuantificaron la frecuencia con la

que cada proteína de la levadura participa (como nodo) en cada uno de los subgrafos de grado k. Encontraron que las proteínas enzimáticas participan en la gran mayoría de los subgrafos de menor tamaño en la red. Con la centraldiad de subgrafo detectaron 145 proteínas esenciales, mientras que con la centralidad de grado solo 135, una cantidad menor a la esperada (Estrada y Rodriguez-Velazquez, 2005).

Con la llegada de la pandemia por COVID-19, Estrada aplicó la centralidad de subgrafo al estudio de la proteína M^{pro} del SARS-CoV-2. Su decisión por emplear la centralidad de subgrafo se basa en la observación de que esta red, así como muchas de las PRN, tienen una alta presencia de ciclos sin cuerda, ciclos que pueden interpretarse como subgrafos. Este problema equivale a identificar *network motifs* (Estrada, 2020).

Ciclos sin cuerda

Un ciclo se define como una secuencia de nodos y enlaces de forma que solo se repite el primer y último nodo. Se dice que un ciclo es un ciclo sin cuerda cuando el total de nodos del ciclo es igual al total de enlaces, no existen vínculos adicionales. Estos ciclos son un tipo de subgrafo cerrado (ver figura 4.3). (Estrada, 2020).

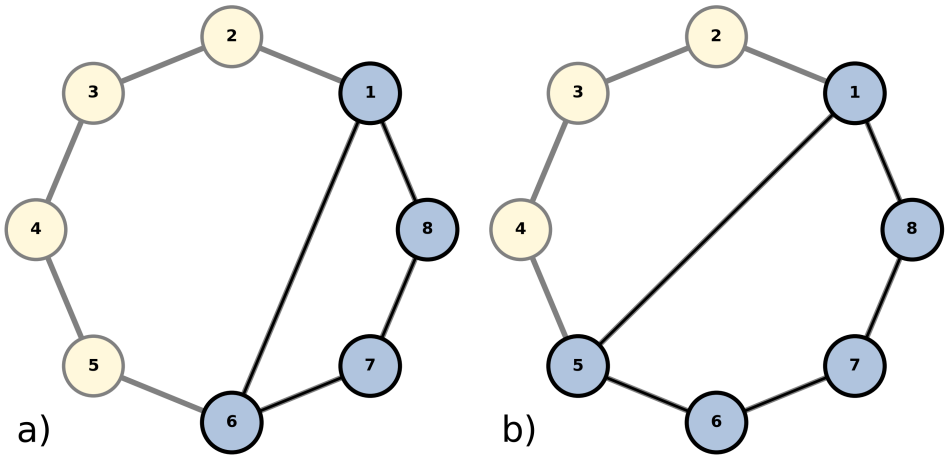


Figura 4.3: Grafos con la presencia de ciclos sin cuerda. El primer grafo (a) contiene dos ciclos sin cuerda de seis y cuatro nodos (1-2-3-4-5-6-1 y 1-6-7-8-1), y el segundo grafo (b) dos ciclos sin cuerda de cinco nodos (1-2-3-4-5-1 y 1-5-6-7-8-1).

A menudo, a los ciclos sin cuerda se les denomina “agujeros”. Recordemos que los sitios activos y sitios de unión a ligandos son regiones en el interior o en el exterior de una proteína en donde se lleva a cabo la catálisis o simplemente la unión del sustrato. Por lo tanto, estas regiones usualmente tienen forma de bolsillo para facilitar la unión con un sustrato específico. En ese sentido, los agujeros pudieran representar esos bolsillos de unión, de ahí la importancia de emplear la centralidad de subgrafo para analizar proteínas catalíticas (Estrada, 2020).

El total de *caminos cerrados* para el nodo n_i se define como los momentos espectrales de la matriz de adyacencia (Estrada, 2020)

$$\mu_u(i) = \sum_i (\mathbf{A}^u)_{ii}. \tag{4.9}$$

A partir de esto se define formalmente la **centralidad de subgrafo** de un nodo n_i como el total de *caminos cerrados* que empiezan y terminan en el nodo n_i, en donde se agrega el factor $1/u!$ para eliminar el problema de divergencia de la función

$$C_{\mathrm{S}}(i) = \sum_{u=0}^{\infty} \frac{\mu_u(i)}{u!}. \tag{4.10}$$

La expresión anterior equivale a los elementos de la diagonal de la exponencial de la matriz de adyacencia $[exp(\mathbf{A})]_{ii}$. Por conveniencia se define $\mathbf{G} = exp(\mathbf{A})$. Por lo tanto C_{S} toma la forma $C_{\mathrm{S}}(i) = G_{ii}$. El valor promedio de la centralidad de subgrafo se calcula como la suma de las centralidades locales entre el total de N nodos de la red

$$\langle C_{\mathrm{S}} \rangle = \frac{1}{N} \sum_i^N C_{\mathrm{S}}(i). \tag{4.11}$$

La centralidad de subgrafo considera las relaciones de un nodo con otros nodos a distancias más lejanas al considerar el $u-$ésimo momento espectral de la matriz de adyacencia cuando $u \to \infty$, en contraste con la centralidad de grado que considera el segundo momento espectral (Estrada, 2020).

4.2.3. Comunicabilidad de la red

Mientras que los subgrafos cerrados son cuantificados a partir de las entradas diagonales de **G**, los *caminos* son cuantificados por las entradas no diagonales de **G**. A esta medida de centralidad se le conoce como **comunicabilidad promedio** de la red, que mide la capacidad de un nodo de comunicarse con otros nodos a través de todas las trayectorias abiertas posibles. Al igual que C_{S}, da una mayor importancia a trayectorias de menor longitud (Estrada, 2020)

$$\langle G_{pq} \rangle = \frac{2}{N(N-1)} \sum_{p<q} G_{pq}. \tag{4.12}$$

Análogamente el **ángulo de comunicabilidad promedio** se define como las entradas no diagonales de la matriz θ que contiene a los ángulos entre los pares de nodos de la red. Esta centralidad mide la eficiencia con la que los nodos se comunican entre sí por cualquier ruta abierta (Estrada, 2020)

$$\langle \theta_{pq} \rangle = \frac{2}{N(N-1)} \sum_{p<q} \theta_{pq}. \tag{4.13}$$

En síntesis, las medidas de centralidad de segundo orden se diferencían de las medidas de primer orden por su capacidad de medir el grado de influencia de un nodo en otro por cualquier trayectoria posible entre ambos sin importar la longitud de dicha trayectoria. No solo consideran aquellas relaciones entre nodos próximos entre sí como es el caso de las medidas de primer orden.

4.3. Medidas de tercer orden

4.3.1. Centralidad de subgrafo de largo alcance

En 2017, Estrada y Silver propusieron una nueva función matricial similar a la función matricial **G** y de la que se obtienen nuevas centralidades de subgrafo y de comunicabilidad. Estas nuevas medidas de centralidad consideran las contribuciones de las centralidades a

través de todas las rutas posibles pero sancionan menos las de mayor longitud que las basadas en subgrafos de segundo orden (Estrada y Silver, 2017).

Con la llegada de la pandemia ocasionada por el virus SARS-CoV-2, Estrada analizó las estructura de las proteasas virales M^{pro} del SARS-CoV y del SARS-CoV-2 con el objetivo de compararlas y analizar el efecto de los inhibidores en M^{pro} del SARS-CoV-2. Aplicó las medidas de centralidad estándar y de orden superior para cuantificar las centralidades locales de cada nodo y las centralidades promedio. Como vimos en el capítulo 2, las secuencias de SARS-CoV y SARS-CoV-2 son similares en un 90 %, sin embargo, SARS-CoV-2 tuvo consecuencias más catastróficas. En este estudio, Estrada encontró que las medidas de centralidad estándar y de segundo orden no mostraron diferencias significativas en ambas variantes, mientras que las medidas de tercer orden promedio arrojaron diferencias del 1900 %. En otras palabras, se concluyó que las medidas de tercer orden son más apropiadas para obtener información de estas proteasas virales (Estrada, 2020).

Para distinguir con más claridad la interacción entre pares de nodos lejanos entre sí, esta nueva matriz utiliza el doble factorial del momento espectral en vez del factorial simple. Dado que $1/u! < 1/u!!$, el doble factorial penaliza menos los *caminos cerrados* de mayor tamaño. Esta matriz se denota por $\mathbf{Z}$ y se calcula como (Estrada, 2020)

$$\mathbf{Z} = \sum_{u=0}^{\infty} \frac{\mathbf{A}^u}{u!!}. \tag{4.14}$$

Estrada y Silver encontraron una aproximación de $\mathbf{Z}$ a una función que involucra la función error de una matriz. El cálculo de $\mathbf{Z}$ a partir de esta definición es más eficiente computacionalmente. La aproximación de $\mathbf{Z}$ es la siguiente (Estrada, 2020)

$$\mathbf{Z} = \frac{1}{2}\left[\sqrt{2\pi}\,\mathrm{erf}\left(\frac{\mathbf{A}}{\sqrt{2}}\right) + 2I\right]\exp\left(\frac{\mathbf{A}^2}{2}\right). \tag{4.15}$$

Con la definición de la función $\mathbf{Z}$ se define la **centralidad de subgrafo de largo**

alcance o LR-C_{S} como las entradas diagonales de $\mathbf{Z}$, Z_{pp}. Esta medida de centralidad difiere de la centralidad de subgrafo en su capacidad de percibir mejor los efectos un nodo en todas las rutas cerradas, pero principalmente las de mayor tamaño (Estrada, 2020)

$$\langle Z_{pp} \rangle = \frac{1}{N} \sum_{p=1}^{n} Z_{pp}. \tag{4.16}$$

4.3.2. Comunicabilidad de largo alcance

La **centralidad de comunicabilidad de largo alcance** promedio considera las contribuciones de un nodo en cualquier ruta abierta sin discriminar a las rutas de gran tamaño. Esta medida consiste en la suma de los elementos por encima o por debajo de la diagonal principal de la matriz $\mathbf{Z}$, Z_{pq} (Estrada, 2020),

$$\langle Z_{pq} \rangle = \frac{2}{N(N-1)} \sum_{p<q} Z_{pq}. \tag{4.17}$$

Para observar la diferencia entre la información obtenida por las medidas de centralidad de primer, segundo y tercer orden en redes que contienen ciclos sin cuerda, consideremos a los grafos de la figura 4.3. Ambos grafos consisten en estructuras de ocho nodos, nueve enlaces y un par de ciclos sin cuerda. Sin embargo, se diferencían en las cuerdas $(1, 6)$ y $(1, 5)$. Aunque existe esta ligera diferencia, la centralidad C_{B}, C_{E} y C_{S}, que corresponden a medidas de primer y segundo orden, no las detectan (estos resultados se muestran en la tabla I). Es decir, la presencia de ciclos sin cuerda no tiene ningún efecto directo en la red desde estas perspectivas. Sin embargo, la centralidad LR-C_{S} sí detecta estas variaciones mínimas, esto significa que la centralidad de tercer orden es más sensible a las perturbaciones. Las medidas de centralidad de tercer orden pudieran darnos información más esencial que medidas de centralidad de primer y segundo orden sobre la topología de las PRN.

Las medidas de centralidad de orden superior no reemplazan a las medidas de primer orden, más bien las complementan. La elección de unas sobre otras recae en el objetivo

Tabla I. Centralidad de intermediación C_{B}, eigenvector C_{E}, subgrafo C_{S} y subgrafo de largo alcance Z_{pp} promedio para los grafos de la figura 4.3.

Medida	Grafo 1	Grafo2
$\langle C_{\mathrm{B}} \rangle$	0.1786	0.1607
$\langle C_{\mathrm{E}} \rangle$	0.1786	0.1607
$\langle C_{\mathrm{S}} \rangle$	0.1786	0.1607
$\langle Z_{pp} \rangle$	5.4667	5.4371

de la investigación y en el tipo de red que se analiza. Por ejemplo, para redes con una alta presencia de ciclos sin cuerda es necesario emplear las medidas de centralidad de orden superior. Como es complicado determinar si una red contiene este tipo de ciclos, a menudo se emplea más de una de estas medidas para una comprensión más profunda de la estructura y comportamiento de una red. Como recordatorio, el objetivo de esta investigación es determinar si existe una correlación entre las medidas de centralidad de segundo y tercer orden, y los sitios activos en $\mathrm{PL}^{\mathrm{pro}}$ del SAR-CoV-2. Esto se debe a que las medidas de centralidad antedichas detectan "agujeros" en la red que podrían relacionarse con los sitios de unión al sustrato.

Capítulo 5

Metodología computacional en bioinformática

A continuación se describe la metodología necesaria para la edición y manipulación de archivos PDB (*Protein Data Bank*) para la construcción del grafo y el cálculo de las medidas de centralidad de las cinco estructuras de interés: PL^{pro}, los complejos PL^{pro}-GRL0617 y PL^{pro}-VIR250, y a las mutaciones PL^{pro} C111S a 100 K y PL^{pro} C111S a 273 K. Estas estructuras fueron extraídas del banco de datos RSCB (*Research Collaboratory for Structural Bioinformatics*): PDB con los códigos siguientes: 6W9C, 7CMD, 6WUU, 6WRH y 6XG3, en ese orden.

5.1. Gestión de archivos PDB

5.1.1. Estructura del archivo PDB

El archivo PDB es un archivo que se emplea principalmente para almacenar información sobre la estructura tridimensional de biomoléculas como las del ADN o de proteínas como resultado de su cristalización por rayos-X. Esta información incluye las coordenadas de los átomos de estas biomoléculas, los tipos de átomos, sus interacciones, entre otros. En la figura (5.1) se muestra la sección ATOM *records* destinada a las posiciones tridimen-

sionales de cada átomo de la proteína. Las primeras ocho columnas son las más relevantes para esta investigación. Primera columna: contiene el número de átomo. Segunda columna: el tipo de átomo, por ejemplo, CA para el C_α o CB para el C_β. Tercera columna: contiene el nombre del aminoácido al que pertenece cada átomo con la nomenclatura de tres letras. Cuarta columna: cadena a la que pertenece cada átomo. Una proteína se divide en cadenas de aminoácidos, estas se representan por letras (A, B, C, ...). Quinta columna: se muestra el número de aminoácido al que pertenece cada átomo. Finalmente, las siguientes tres columnas contienen las coordenadas x, y y z de cada átomo en unidades de Å (ángstroms).

```
ATOM      1  N   GLU A   1     -65.718  21.719 -10.231  1.00  0.00        N
ATOM      2  CA  GLU A   1     -64.244  21.484 -10.489  1.00  0.00        C
ATOM      3  CB  GLU A   1     -64.965  20.175 -10.958  1.00  0.00        C
ATOM      4  CG  GLU A   1     -63.953  19.107 -10.597  1.00  0.00        C
ATOM      5  CD  GLU A   1     -65.133  18.548  -9.856  1.00  0.00        C
ATOM      6  OE1 GLU A   1     -66.196  19.183  -9.717  1.00  0.00        O
ATOM      7  OE2 GLU A   1     -65.333  17.429 -10.263  1.00  0.00        O
ATOM      8  C   GLU A   1     -63.267  22.805 -10.967  1.00  0.00        C
ATOM      9  O   GLU A   1     -63.180  23.627 -10.003  1.00  0.00        O
ATOM     10  N   VAL A   2     -61.943  22.849 -11.595  1.00  0.00        N
ATOM     11  CA  VAL A   2     -60.550  22.423 -11.123  1.00  0.00        C
ATOM     12  CB  VAL A   2     -60.859  20.997 -11.674  1.00  0.00        C
ATOM     13  CG1 VAL A   2     -61.868  20.877 -12.828  1.00  0.00        C
```

Figura 5.1: Estructura del archivo pdb. Sección ATOM *records*.

5.1.2. Verificación de la secuencia y corrección del archivo PDB

La cristalizacion de una proteína es un proceso complejo y riguroso, dificultades en la fase experimental o en el modelado impiden la identificación de todos los residuos que componen una proteína. En consecuencia, los datos de las secuencias reportados en el archivo PDB pueden estar incompletos. Para la detección de este tipo de errores se utilizó la herramienta pdb-tools v2.5.0 (Rodrigues et al., 2018). Esta herramienta también ofrece la posibilidad de editar y manipular secuencias de proteínas, eliminar cadenas específicas,

eliminar átomos, obtener la intersección entre dos archivos PDB, entre otras funciones.

Las estructuras proteicas de esta investigación son variaciones de la proteína PL^{pro}, es decir, la PL^{pro} unida a inhibidores y mutaciones que modifican su estructura tridimensional pero no la secuencia. Por lo tanto, es necesario verificar que la composición de las cinco estructuras es la misma y que las secuencias estén completas. Como primer paso se instaló pdb-tools con el administrador de paquetes pip desde la consola (Anaconda Prompt) con el comando `pip install pdb-tools`. Luego, la dirección de la consola se redirigió a la ubicación de los *scrips* de pdb-tools. Esta dirección es la dirección de intalación de pdb-tools por defecto (es importante mencionar que los archivos PDB deben de colocarse en la misma carpeta que los *scripts*)

```
cd C:\Users\Usuario\anaconda3\Scripts
```

Enseguida, con la función `pdb_mkensemble` se combinaron las cinco estructuras en un ensamble y se almacenaron en un nuevo archivo `ensemble.pdb`

```
pdb_mkensemble 6w9c.pdb 7cmd.pdb 6wuu.pdb 6wrh.pdb 6xg3.pdb> ensemble
 .pdb
```

Creado el ensamble se verificó que las estructuras contenidas tuvieran la misma composición de átomos, residuos y cadenas con el código siguiente

```
pdb_chkensemble ensemble.pdb
```

Las diferencias entre las estructuras se muestran de par en par. Por ejemplo, cuando el modelo cuatro (6WRH) no contiene los átomos de los residuos MET y ARG contenidos en el modelo tres (6WUU), vemos algo similar a lo siguiente

```
2522      LYS A 315 Models 3 and 4 differ:
Atoms in model 3 only:
1  N   MET A  -1
```

```
 2   CA   MET A   -1
 3   C    MET A   -1
 4   O    MET A   -1
 5   CB   MET A   -1
 6   CG   MET A   -1
 7   SD   MET A   -1
 8   CE   MET A   -1
 9   N    ARG A    0
```

Una vez indentificadas las discrepancias en la composición de los archivos PDB se utilizó la herramienta PDBFixer para completar los residuos faltantes. Antes de la instalación de PDBFixer fue necesario instalar OpenMM, una colección de librerías destinadas a python. La v7.7.0 se instaló a través de la consola con el código

```
conda install -c conda-forge openmm
```

Una vez instalado OpenMM se instaló PDBFixer v1.7.1 en su formato de interfaz gráfica con el siguiente código

```
conda install -c conda-forge pdbfixer
```

Después se inició la aplicación desde la consola con el comando `pdbfixer`. En la primer ventana se seleccionó el archivo PDB de la proteína de interés. En la paso siguiente se mostró una tabla con la información de las cadenas (total de residuos por cadena, tipo de residuos), como se muestra en la figura 5.2.

Chain	# Residues	Content	Include?
A	310	Protein	☑
B	307	Protein	☑
C	309	Protein	☑
A	2	CL, ZN	☑
B	2	CL, ZN	☑
C	3	CL, ZN	☑

Figura 5.2: Seleción de cadenas con PDBFixer

En la misma figura se muestra la información para PL^{pro} tipo *wild* (PDB 6W9C). Aunque se muestran tres cadenas para esta proteína, PL^{pro} solo consiste en una cadena única de 315 nodos. Las cadenas son similares entre sí, por lo tanto, solo se tomó la cadena A. Finalmente, se mostraron los residuos faltantes en la secuencia y se seleccionaron todos ya que nos interesa completar la estructura (figura 5.3). El resto de las configuraciones se omitieron ya que son irrelevantes para esta investigación.

Chain	Residue Positions	Sequence	Add?
A	1 to 2	GLU, VAL	☑
A	225 to 227	THR, CYS, GLY	☑
A	316 to 317	ALA, ALA	☑

Figura 5.3: Selección de residuos faltantes con PDBFixer

5.1.3. Lectura del archivo PDB con Biopython

Para la lectura de las coordenadas espaciales de los átomos de C_α del archivo PDB se empleó el módulo PDB de la librería Biopython v1.82 (Cock et al., 2009). Biopython es una colección de librerías con diversas funciones destinadas a la bioinformática, estas funciones incluyen la lectura, escritura y manipulación de bases de datos en formatos FASTA, GenBank, PDB, entre otros (si utiliza la librería BioPandas es probable que presente errores de compatibilidad con el archivo modificado con PDBFixer). Para acceder a la librería de Biopython se ejecutó el siguiente código en la línea de comandos de Python:

```
1  from Bio import PDB
```

Para acceder al archivo PDB de la proteína, primero se creó el objeto PDBParser y luego se accedió al archivo PDB a través de su nombre.extensión (ej. 6W9C.pdb):

```
2  #Clase destinada a la lectura de los archivos PDB
3  pdb_parser = PDB.PDBParser(QUIET=True)
4  structure = pdb_parser.get_structure('6w9c', '6w9c.pdb')
```

Para acceder a las coordenadas espaciales de los C_α se debe de acceder a una serie

de información en el siguiente orden: estructura>modelo>cadena>residuo>átomo. A esto se le conoce como SMCRA. Algunos archivos PDB contienen más de una estructura de proteínas, por ejemplo, el ensamble creado para el análisis con PDB-tools es un archivo con cinco estructuras. Los modelos en un archivo PDB son las distintas conformaciones que puede tener una misma biomolécula, aquí se percibe un solo modelo, este tiene el identificador por defecto id 0. Los identificadores para las cadenas y los residuos son los mismos que proporciona el archivo PDB: nos interesan las átomos de C_α de la cadena A, por lo tanto, se emplea la abreviatura "CA" para referirnos a esta cadena. A través de un ciclo se almacenaron las coordenadas extraídas con el método `a.get_coord()`:

```
#Lista vacia para almacenar las coordenadas
coordinates = []

for model in structure.get_list():
  for chain in model.get_list():
    for residue in chain.get_list():
      if residue.has_id("CA"):
        ca = residue["CA"].get_coord()
        coordinates.append(ca)
```

5.2. Construcción del grafo y cálculo de las medidas de centralidad

Para la construcción del grafo de cada proteína se empleó el paquete de Networkx (Hagberg et al., 2008). Esta herramienta es popularmente utilizada para la generación, manipulación y cálculos de redes complejas. Se importó esta librería desde la línea de comando de Python

```
import networkx as nx
```

Como vimos en el capítulo 3, un grafo depende de un conjunto de nodos y de una

lista de enlaces. Sin embargo, también se puede generar un grafo a partir de la matriz de adyacencia. Se calculó la matriz de adyacencia con la definición dada por la ecuación 3.2 y con las coordenadas espaciales de los átomos de C_α que se obtuvieron del filtrado anterior. Después, se generó el grafo con la siguiente función

```
15  Grafo = nx.from_numpy_array(matriz_de_adyacencia)
```

Las medidas de centralidad y propiedades de un grafo se obtienen a partir de operaciones en la matriz de adyacencia **A**. Las únicas medidas de centralidad que nos dan información sobre la centralidad local de cada nodo y en promedio son la centralidad de intermediación C_{B}, de eigenvector, C_{E}, de subgrafo C_{S} y de subgrafo de largo alcance LR-C_{S}. El resto de medidas solo nos proporcionan información promedio. C_{B} local y promedio se calcularon directamente con la ecuación 4.3. Para el grado de heterogeneidad ρ se calculó primero la centralidad de grado con 4.1 y luego se obtuvo con 4.2. Para las medidas de segundo orden se calculó antes la matriz **G** que equivale a la exponencial de **A**. C_{S} local consiste en los elementos de la diagonal principal de **G** (para el valor promedio solo dividimos entre el total de nodos N) y la centralidad de comunicabilidad promedio G_{pq} se obtuvo de los elementos arriba de la diagonal de **G** (ecuación 4.13). C_E se obtuvo con el método de la potencia, algoritmo que veremos a continuación.

En el capítulo 4 se definió C_{E} como aquella que cuantifica la relevancia de un nodo en términos de la relevancia de sus nodos adyacentes con una ponderación determinada por su propia centralidad de grado. Para calcular esta centralidad se utilizó el método de la potencia, que consiste en iteraciones sobre la centralidad de grado C_{D}. El diagrama de flujo de este método se muestra en la figura 5.4. Para ello, primero se calculó C_{D} con la fórmula de la ecuación 4.1. Después se normalizó el vector y se almacenó en una variable c_1, y su constante de normalización en una variable s_1. Se tomó C_{D} normalizado como vector inicial y se multiplicó por la matriz de adyacencia **A**. Cada producto obtenido se multiplicó por la matriz **A** s pasos hasta que la constante de normalización siguiente

coincidió con la anterior. Aunque es un algoritmo iterativo, es muy eficiente. Para la PRN con un total de 315 nodos y una matriz de adyacencia de la misma dimensión se registraron 180 iteraciones con un tiempo de ejecución promedio de 21.15 ms.

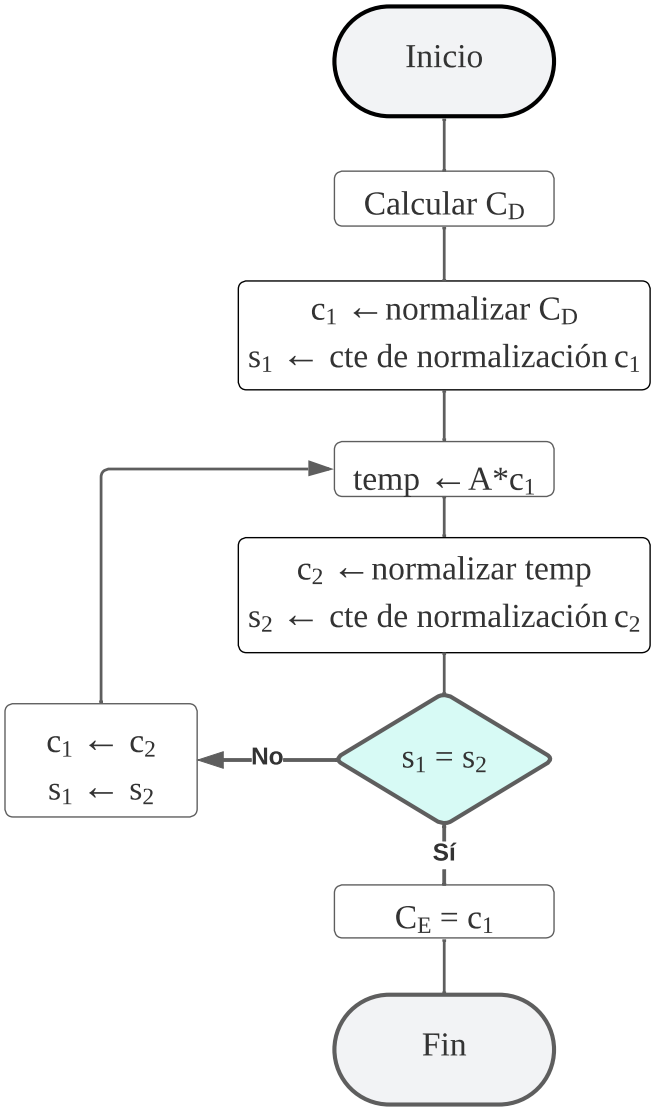


Figura 5.4: Diagrama de flujo del método de la potencia para calcular C_{E}

Finalmente, para las medidas de tercer orden se calculó primero la matriz **Z** con la definición de series de Taylor de la ecuación 4.14. Se utilizó esta definición ya que es más precisa que la definición 4.15 y la diferencia entre los tiempos de ejecución de ambas es mínima. LR-C_{S} se calculó como los elementos de la diagonal de **Z** y la comunicabilidad de largo alcance Z_{pq} como los elementos arriba de la diagonal.

5.3. Método AUC-ROC

Para evaluar la precisión del método basado en teoría de grafos y medidas de centralidad para la predicción de sitios activos se empleó el método del área bajo la curva de la curva ROC (AUC-ROC). La curva ROC es una representación de la relación entre los verdaderos positivos (TP), los verdaderos negativos (TN), los falsos positivos (FP) y los falsos negativos (FN) a lo largo de diferentes umbrales. El TP es la cantidad de elementos que un método predijo como positivos y que realmente son positivos, TN son los que el método predijo como negativos y sí lo son, FP son los que el método predijo como positivos pero son falsos y FN son los que el método predijo como negativos pero son positivos. De manera que la curva ROC se grafica como la tasa de TP (TPR), también conocida como sensitividad o *recall*, contra la tasa de FP (FPR), ambas definidas a continuación (Narkhede, 2018):

$$TPR = \frac{TP}{TP + FN}, \tag{5.1}$$

$$FPR = \frac{FP}{TN + FP}. \tag{5.2}$$

El área bajo la curva ROC máxima es de uno, de manera que se considera que un método es perfectamente preciso en este punto. En el caso opuesto cuando el área es cero, se considera que un método es completamente impreciso. En general, se consideran los siguientes rangos: 0.0-0.5 (malo), 0.5-0.6 (insatisfactorio), 0.6-0.7 (satisfactorio), 0.7-0.8 (bueno), 0.8-0.9 (muy bueno) y 0.9-1 (excelente) (ver figura 5.5). El método AUC-ROC se diseñó para determinar la precisión de métodos de aprendizaje profundo, sin embargo, dada su versatilidad se ha aplicado a otras áreas, por ejemplo, en el campo de la salud para evaluar la precisión de las pruebas de PCR en personas para la detección de la infección por COVID-19 (Narkhede, 2018).

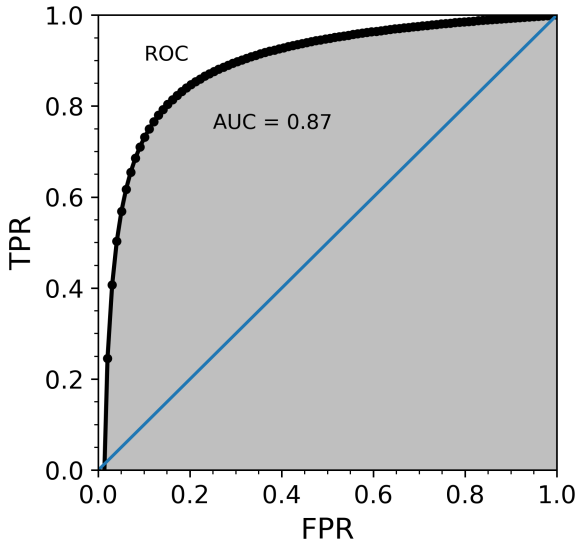


Figura 5.5: Ejemplo de una Curva ROC generada a partir de 100 umbrales distintos con un área bajo la curva AUC en el rango *muy bueno* con base al *test* de calidad.

La razón por la que se eligió este método es porque considera una gran cantidad de umbrales. Ya que la idea es determinar si existe una relación entre los máximos de las medidas de centralidad locales y los sitios activos de las proteínas, es difícil determinar un único umbral dentro de todos los posibles que sean cercanos al máximo. Este método considera un total de u umbrales que van del valor mínimo al máximo, de manera que se grafica un punto con un TPR y FPR dados para cada umbral y se evalúa la efectividad del método considerando los u umbrales.

Para aplicar este método en el contexto de sitios activos se consideraron como positivos a los residuos con una centralidad por arriba del umbral variable u y como negativos a aquellos con una centralidad por debajo de este umbral. De manera que los sitios activos con una centralidad por arriba de u son cuantificados por TP y aquellos residuos arriba de u que no participan en el sitio activo son cuantificados por FP. En el caso contrario, los residuos del sitio activo por debajo de u son cuantificados por FN y los residuos que no participan en el sitio activo por debajo de u son cuantificados por TN. En el contexto de sitios de unión se invirtieron los umbrales y se consideraron como postivos a aquellos residuos por debajo de u y como negativos a aquellos por arriba del umbral, ya que se

observó una relación entre los sitios de unión y los mínimos en las centralidades de orden superior, esto se discute en el siguiente capítulo.

5.4. Grafo aleatorio y regular

La última parte del análisis de la PRN consistió en determinar las distribuciones de grado, además de verificar la longitud de ruta promedio $\langle L \rangle$ y el coeficiente de agrupamiento $\langle C \rangle$ para determinar si la PRN es una red SW. Para este análisis se emplearon los modelos de redes teóricas cuyas propiedades son bien conocidas: la red aleatoria y la red regular, descritos en el capítulo 3. Para que la comparación entre las tres redes sea válida es necesario que cumplan con dos condiciones: que tengan el mismo número de nodos N y mismo número de enlaces E. Por esta razón se empleó el modelo de Erdös-Rényi con número fijo de enlaces. La paquetería Networkx proporciona un modelo para generar este tipo de grafo aleatorio

```
1  Grafo = nx.gnm_random_graph(N, E)
```

Las funciones de Networkx para grafos emplean la notación $\mathcal{G}(N, E)$ para referirse a un grafo con N nodos y E enlaces, no confundir con la definición matemática de grafo. Las propiedades de grafos aleatorios son variables aún para unos mismos parámetros. Por lo tanto, para obtener valores precisos se evaluaron estas propiedades sobre el promedio de un ensamble de $S = 150$ grafos aleatorios. Por ejemplo, para calcular la distribución de grado se calcularon S matrices de adyacencia $\mathbf{A}$ y S centralidades de grado C_{D}.

Recordemos que las redes regulares se obtienen a partir del modelo de Watts-Strogatz $\mathcal{G}_{N,k,p}$ en donde N es el total de C_α, k es el grado promedio y p es la probabilidad de que un par de nodos de la red se conecten. La red regular se obtiene cuando $p = 0$. Por otro lado, el grado promedio se obtiene a partir del total de enlaces como $2E/N$. La paquetería Networkx también ofrece una función para este modelo

```
Grafo = nx.watts_strogatz_graph(N, k, p)
```

Para obtener la matriz de adyacencia de estos grafos se utilizó la siguiente función de Networkx

```
A = np.array(nx.convert_matrix.to_numpy_matrix(Grafo))
```

5.4.1. Longitud de ruta promedio

Este medida calcula la distancia promedio entre un par de nodos de una red, como vimos en el capítulo 3. Para calcular $\langle L \rangle$ primero se calculó la matriz de distancias d con el algoritmo de Floyd-Warshall del diagrama 5.6. La matriz de distancias d inicial equivale a la matriz de adyacencia excepto porque se intercambian los ceros de la matriz por un valor infinito (los elementos de la diagonal se mantienen en ceros). Aquí se definió como infinito simplemente un valor muy grande: el número de enlaces máximos. Después de calcular esta matriz d inicial se utilizó un bucle for anidado. La primer condición es que las tres variables de iteración i, j, k sean diferentes entre sí, ya que no interesan los elementos de la diagonal al tratarse de un grafo simple. Como recordatorio, la distancia entre un par de nodos se mide como el total de enlaces necesarios para llegar del nodo n_i al nodo n_j. Entonces, si $d_{ik} = 1$ y $d_{ji} = 1$ indica que n_j está enlazado a n_k. Por lo tanto, se busca que esta intersección actualice su valor infinito por un entero como indicio de la conexión. El valor en la celda d_{jk} aumenta siempre exista una vía para llegar a él a través de otro par de nodos cualesquiera. Se finalizó el algoritmo hasta recorrer todos los elementos de la matriz. Finalmente, se calculó $\langle L \rangle$ con la ecuación 3.4.

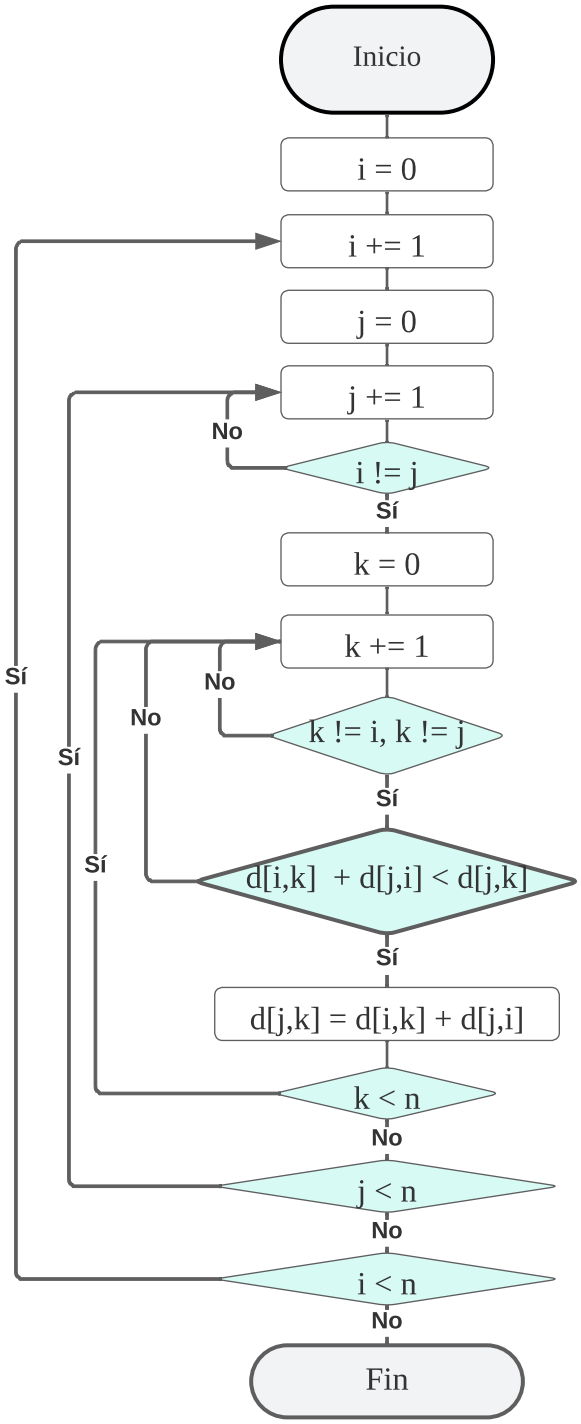


Figura 5.6: Algoritmo de Floyd-Warshall para el cálculo de $\langle L \rangle$

5.4.2. Coeficiente de agrupamiento promedio

Como vimos en el capítulo 3, el coeficiente de agrupamiento promedio $\langle C \rangle$ de un nodo se mide como la cantidad de subgrafos cerrados de grado 3 (triángulos) de los que forma parte. Para calcular la matriz de frecuencias t que contiene estas cantidades se utilizó el algoritmo del diagrama 5.7. Para esto se empleó la matriz de adyacencia $\mathbf{A}$. Si se cumple que $A_{ij} = A_{jk} = A_{kj} = 1$ significa que los nodos n_i, n_j y n_k forman un triángulo. Finalmente, se calculó $\langle C \rangle$ con la ecuación 3.5.

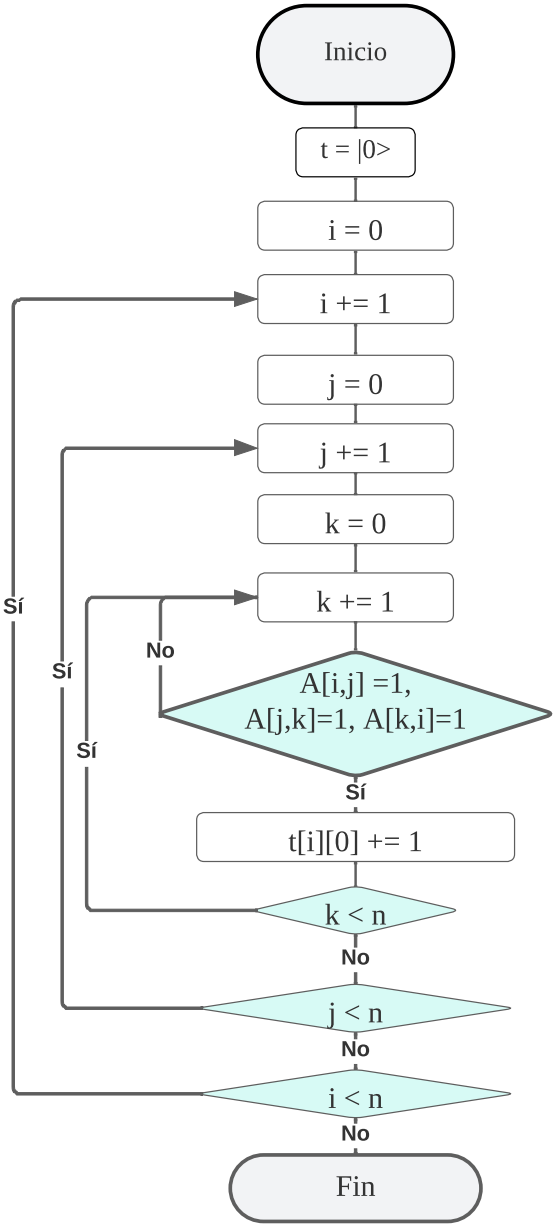


Figura 5.7: Matriz t del coeficiente de agrupamiento $\langle C \rangle$

5.5. Herramientas adicionales

Además de los programas y librerías mencionadas anteriormente, se emplearon los siguientes: para el almacenamiento de matrices se importó la biblioteca h5py v3.9.0, esta es una herramienta muy útil que permite almacenar más de una matriz de gran tamaño en un solo archivo en formato h5 y asignar un identificador a cada matriz para su manipulación individual.Para graficar la PRN de las proteínas se utilizó la biblioteca Mayavi v4.8.1 de Python, esta herramienta es útil cuando se desean escalar los gráficos a un vector (por ejemplo, para escalar los diámetros de los nodos con su centralidad de eigenvector). Para la visualización de la estructura cristalográfica de las proteínas se empleo el programa VMD.

Capítulo 6

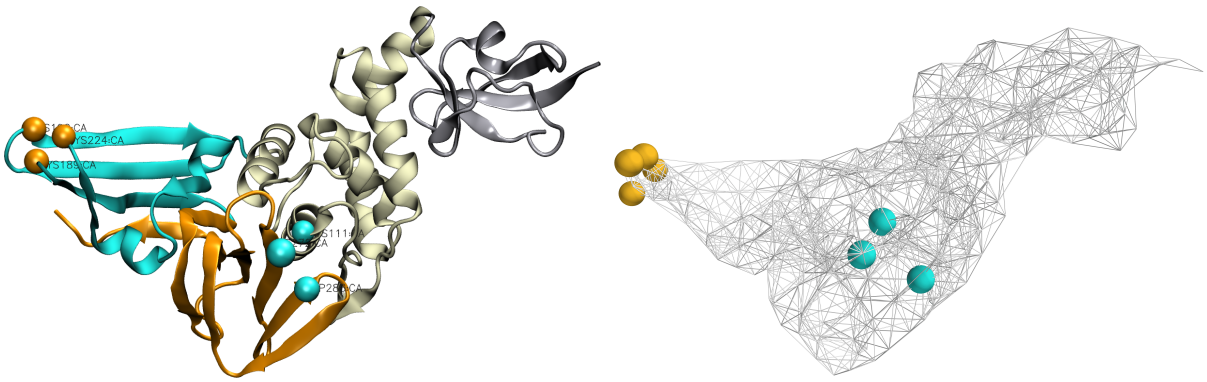
Aplicación a las enzimas del SARS-CoV-2

Aquí se exhiben los resultados y el análisis de las medidas de centralidad promedio y locales: de intermediación C_{B}, de eigenvector C_{E}, de subgrafo C_{S} y de subgrafo de largo alcance LR-C_{S} aplicadas a la red de residuos de proteína (PRN) de la proteasa viral $\mathrm{PL}^{\mathrm{pro}}$ del SARS-CoV-2. Se explica la relación entre los máximos y mínimos de estas centralidades con los sitios activos y de unión de la proteasa. También se añade el análisis de las propiedades de la estructura desde la perspectiva de la ciencia de redes, es decir, sus distribuciones de grado, propiedad de mundo pequeño y coeficiente de agrupamiento promedio $\langle C \rangle$ vistos en el capítulo 3. Se extienden los cálculos a los complejos con inhibidores $\mathrm{PL}^{\mathrm{pro}}$-GRL-0617 y $\mathrm{PL}^{\mathrm{pro}}$-VIR250 para determinar las variaciones en la actividad catalítica de la proteasa viral. También se analizan las variaciones estructurales de las mutaciones de $\mathrm{PL}^{\mathrm{pro}}$ en el residuo catalítico Cys111 a temperaturas de 100 K y 293 K.

6.1. Construcción de la red de residuos de proteína

Como ya hemos visto, la PL^{pro} es una cisteína proteasa con una secuencia conformada por 315 aminoácidos y con la triada catalítica Cys111, His272 y Asp286 en su sitio activo. El modelo para estos tipos de redes consiste en una representación de grano grueso, como hemos visto en el capítulo 3, por lo que los nodos pueden representar a los átomos de C_α o C_β de los aminoácidos. Incluso se puede optar por tomar sus centros de masas sus cadenas laterales Estrada (2012). Debido a la simplicidad se eligió la representación por C_α.

Por otro lado, para determinar las conexiones de la red es necesario considerar todas las interacciones implicadas en el plegamiento de la proteína, estas interacciones pueden ser de dos tipos: covalentes, como los enlaces peptídicos o enlaces disulfuro; o no covalentes, como las interacciones hidrofóbicas, fuerzas de Van der Waals, enlaces de hidrógeno, interacciones electrostáticas o interacciones pi-pi, como hemos visto en el capítulo 2. En la figura 6.1b se muestra la red de átomos de C_α para PL^{pro} en donde los enlaces representan todas las interacciones mencionadas, y en la figura 6.1a su representación cristalográfica.

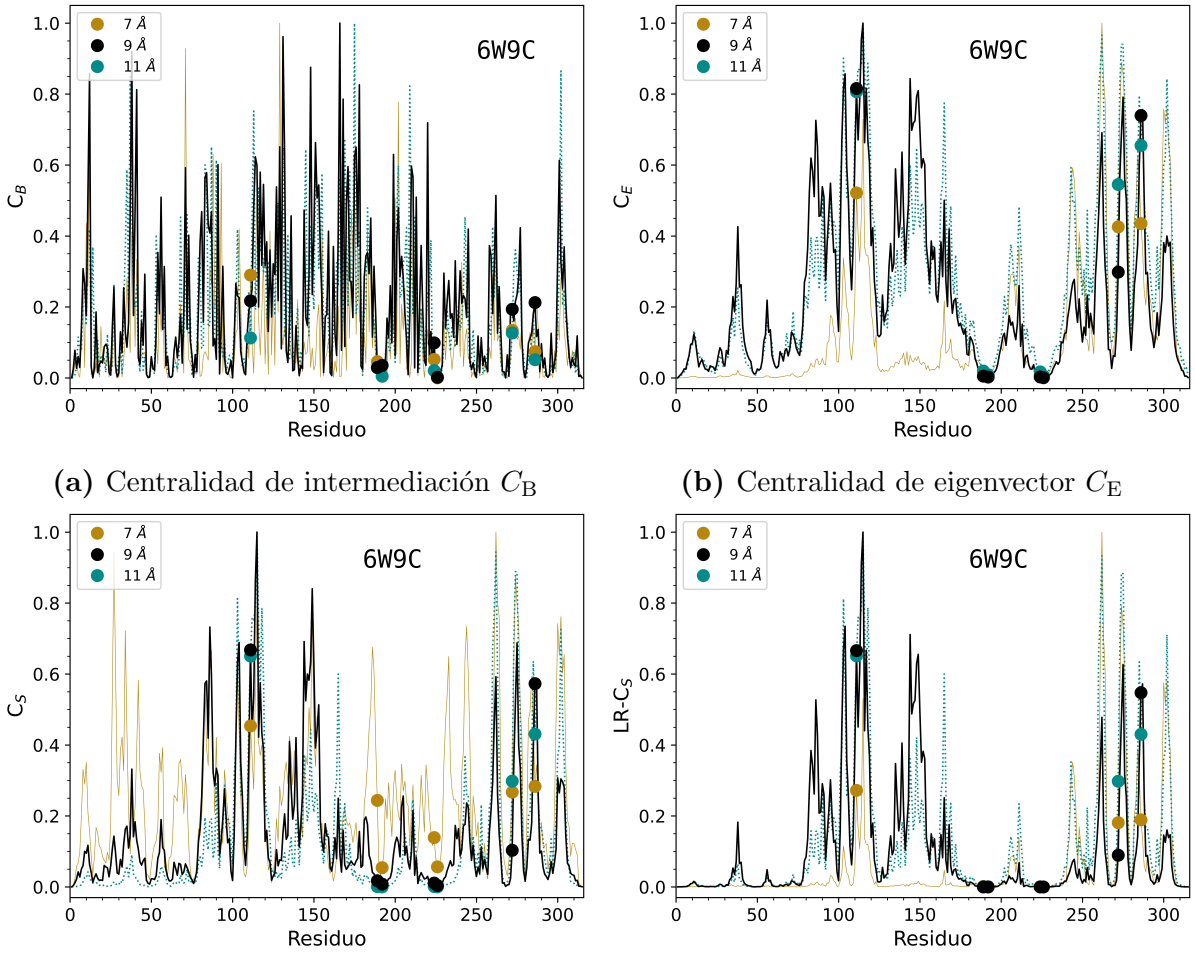


(a) **(b)**

Figura 6.1: Representación (a) cristalográfica de PL^{pro} tipo *wild* y (b) en su forma de red de átomos de C_α. Su sitio activo consiste en los residuos Cys111, His272 y Asp286 (esferas azules), y su sitio de unión en Cys189, Cys192, Cys224 y Cys226 (esferas amarillas)

En la literatura se menciona que el radio de interacción promedio que considera todos los tipos de interacciones sin ser demasiado grande ni demasiado pequeño ronda los 7

Å Estrada (2012, 2020). Para determinar el radio de interacción R_c más óptimo para la red de átomos de C_α de la PL^{pro}, se analizaron las medidas de centralidad locales para tres radios R_c distintos: 7 Å, 9 Å y 11 Å. En las gráficas de la figura 6.2 se muestran las medidas de centralidad locales para los tres radios mencionados. Las centralidades están normalizadas de cero a uno, de forma que el cero representa la centralidad correspondiente al nodo con la centralidad mínima y el uno corresponde a la centralidad máxima.



(a) Centralidad de intermediación C_B

(b) Centralidad de eigenvector C_E

(c) Centralidad de subgrafo C_S

(d) C. de subgrafo de largo alcance LR-C_S

Figura 6.2: Comparación de las medidas de centralidad de PL^{pro} del SARS-CoV-2 para los radios de corte R_c de 7 Å, 9 Å y 11 Å.

Como primera observación, se destaca que la centralidad de intermediación C_B no diferencía significativamente a aquellos residuos con una alta actividad catalítica de cualquier

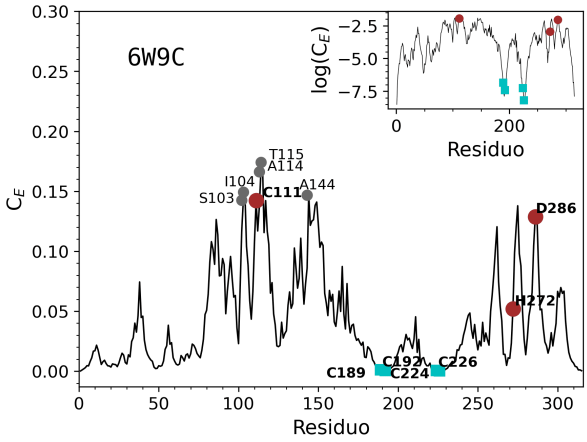
otro residuo de la red, en general, no se destacan residuos importantes para ninguno de los tres radios de corte. Anteriormente, se definió a C_{B} como aquella que determina la importancia de un nodo en términos de su presencia en el camino geodésico entre todos los pares de nodos de la red. En ese sentido, se concluye que los sitios activos y de unión no son centrales y, por lo tanto, C_{B} no es apropiada para la predicción de sitios activos en $\mathrm{PL}^{\mathrm{pro}}$ del SARS-CoV-2.

Como segunda observación, las centralidades C_{E}, C_{S} y LR-C_{S} marcan notoriamente una diferencia entre los residuos del sitio activo y el resto de residuos para los radios de 9 Å y 11 Å. Sin embargo, para $R_{\mathrm{c}} = 11$ Å se obtiene una red demasiado densa, además, se encuentra por arriba del umbral de la distancia de interacción máxima entre dos aminoácidos. Por esta razón, se determina que el radio de corte más adecuado es el de 9 Å.

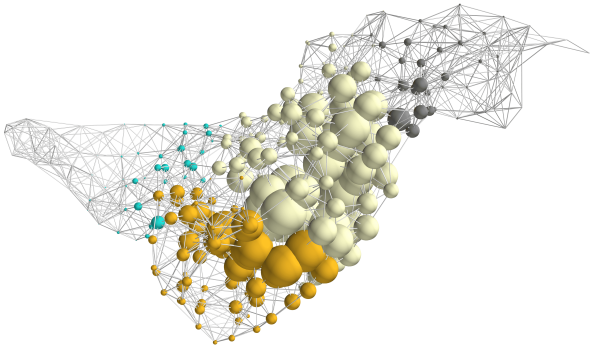
Las proteasas seleccionadas para esta investigación consisten en $\mathrm{PL}^{\mathrm{pro}}$ y sus complejos con inhibidores y mutaciones que modifican la estructura tridimensional de la proteína pero no su secuencia. Son estructuras diferentes tridimensionalmente pero muy similares entre sí y, por homogeneidad, se aplicó el mismo radio de corte en todas las estructuras proteicas. En síntesis, para el resto del análisis se excluye C_{B} por no mostrar relación alguna con los sitios activos, además, se emplea un R_{c} de 9 Å.

6.2. Análisis de las medidas de centralidad en $\mathrm{PL}^{\mathrm{pro}}$ (6W9C)

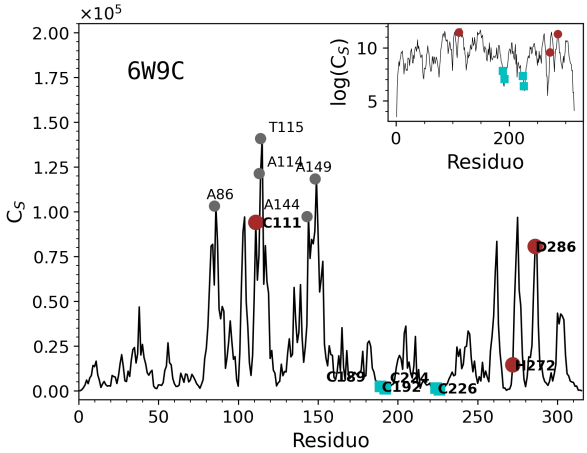
El objetivo principal de esta investigación es determinar si existe una correlación entre los máximos de las medidas de centralidad con los sitios activos de $\mathrm{PL}^{\mathrm{pro}}$. Recapitulando, esta idea surge del hecho de que los sitios activos de una enzima son regiones con una mayor actividad catalítica y pudieran relacionarse con nodos con una alta centralidad. En la figura 6.3 se muestran las medidas de centralidad de orden superior C_{E}, C_{S} y LR-C_{S}.



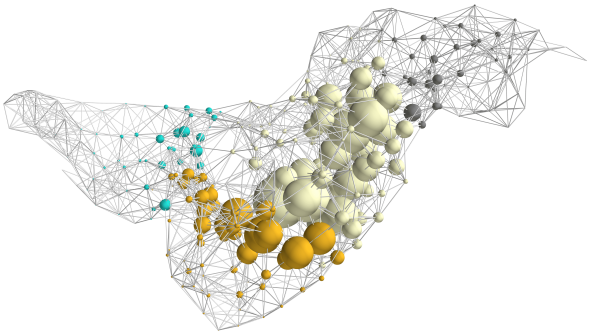
(a) Centralidad de eigenvector C_{E}



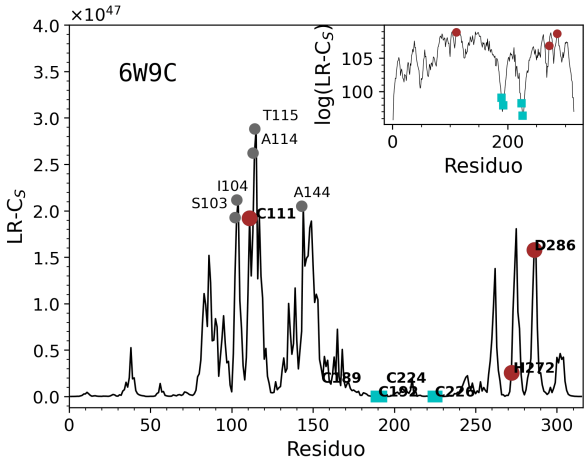
(b) PRN para C_{E}



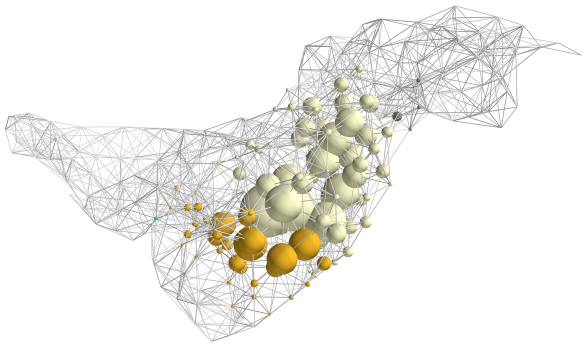
(c) Centralidad de subgrafo C_{S}



(d) PRN para C_{S}



(e) C. de subgrafo de largo alcance LR-C_{S}



(f) PRN para LR-C_{S}

Figura 6.3: Medidas de centralidad de orden superior de $\mathrm{PL}^{\mathrm{pro}}$ tipo *wild* con los residuos del sitio activo en rojo y los del sitio de unión en azul. En (b), (d) y (f) se encuentran las PRN adaptadas a cada una de estas medidas con los subdominios *Ubi-like* (1-60, gris), *Thumb* (61-178, blanco), *Fingers* (179-240, azul) y *Palm* (241-315, naranja). Los tamaños de las esferas son proporcionales a las centralidades.

De forma inmediata se observa que hay dos regiones de nodos con alta centralidad en las tres categorías, y otro par de regiones con baja centralidad. Recordemos que los subdominios de PL^{pro} se encuentran en el siguiente orden: *Ubi-like* (1-60), *Thumb* (61-178), *Fingers* (179-240) y *Palm* (241-315). Por consiguiente, las regiones de alta centralidad corresponden a los subdominios *Thumb* y *Palm*, regiones que contienen a los sitios activos. También se observa que muy cerca de los centros de dichos subdominios se encuentran los sitios activos y los picos más altos. En contraste, la segunda región con menor centralidad en las tres categorías es la región *Ubi-like* y la primera la región *Fingers*. Este último subdominio contiene a los cuatro sitios de unión de la proteína, que se encuentran ubicados en las dos cavidades de la región.

Dadas las gráficas logarítmicas en 6.3, los sitios de unión se encuentran entre los principales nodos con valores mínimos. Además, en la tabla I se localizan a estos sitios de unión dentro de los primeros quince residuos con valor mínimo en C_{E} y LR-C_{S}, y en C_{S} solo Cys226 se encuentra dentro de los primeros quince. En consecuencia, los sitios de unión se relacionan con mínimos en las medidas de centralidad anteriores.

Para los sitios activos, se observa que los residuos catalíticos Cys111 y Asp286 tienen valores altos en las tres medidas de centralidad. Principalmente, C_{E} y LR-C_{S} discriminan mejor a ambos residuos del resto de aminoácidos de la proteína. Esto se visualiza en los resultados de la tabla I, en donde los resultados de ambas medidas son las mismas. En ambas, Cys111 toma la séptima posición dentro de los residuos con valores máximos, y Asp286 la treceava. Sin embargo, los resultados no parecen ser satisfactorios para His272 en ninguno de los tres casos.

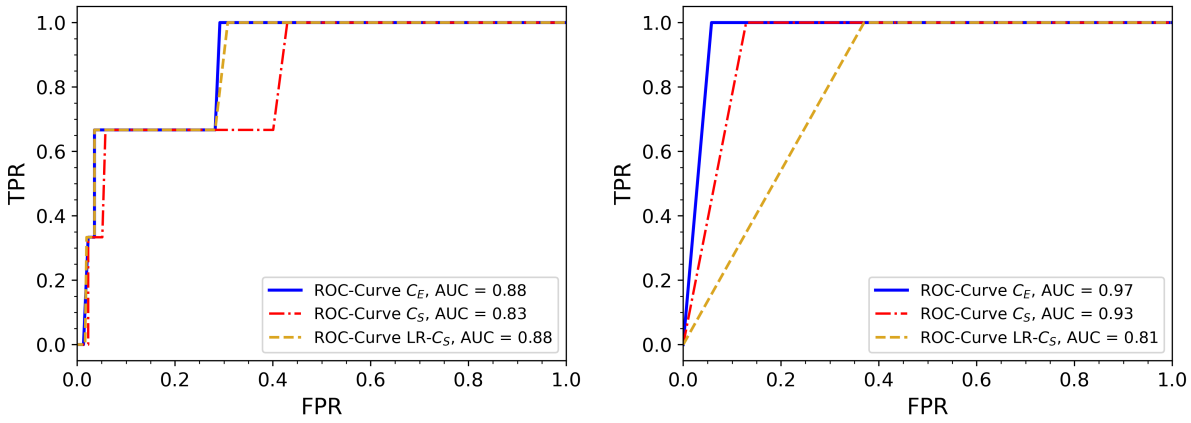
Tabla I. Residuos máximos y mínimos en orden descendente en las medidas de centralidad de orden superior para PL^{pro} (PDB ID: 6W9C). Se obtuvieron los mismos resultados para C_E y LR-C_S. Las celdas rojas representan los residuos del sitio activo y las celdas azules los residuos del sitio de unión.

Posición	Máximos		Mínimos	
	C_E	C_S	C_E	C_S
1	T115	T115	E1	E1
2	A114	A114	G227	K315
3	I104	A149	C226	Y268
4	A144	A86	T191	T313
5	S103	A144	K315	I314
6	L117	I104	K228	G227
7	C111	H275	T225	V2
8	A149	C111	K190	T191
9	C148	C148	C192	C226
10	H275	S103	P223	K190
11	L113	L150	G193	K228
12	G287	N146	C224	Q269
13	D286	L87	Q229	C270
14	A86	S262	T313	G193
15	F147	G287	C189	T225

Los residuos del sitio activo no corresponden exactamente a los máximos, especialmente His272. Existen dos posibles razones para esto: la primera, es que se consideró un mismo radio para analizar todos los tipos de interacciones, esto nos puede dar una relación menos precisa de la verdadera relación entre cada aminoácido de la proteína. Además, si un nodo es el nodo más central de la red, entonces sus vecinos más cercanos compartirán en gran medida esa alta centralidad, por lo tanto, Thr115 y Ala114 podrían ser falsos positivos. Segundo, podría ser un mecanismo de protección del virus el no posicionar a los residuos del sitio activo como los más centrales por la siguiente razón: al ser los nodos más centrales tendrán una mayor comunicación con otros nodos de la red y percibirán los efectos desde cualquier región, efectos que pueden ser tanto positivos como perjudiciales.

Para determinar la precisión del método basado en teoría de grafos y medidas de centralidad de orden superior locales para la predicción de sitios activos en la PL^{pro} del SARS-CoV-2, se calculó el área bajo la curva (AUC) de la curva ROC.

Como vimos en el capítulo 5, este método grafica la tasa de verdaderos positivos (TPR) contra la tasa de falsos positivos (FPR). Estos resultados se muestran en la figura 6.4.



(a) AUC-ROC para sitios activos **(b)** AUC-ROC para sitios de unión

Figura 6.4: Cálculo del área bajo la curva (AUC) de las curvas ROC asociadas a las centralidades de orden superior.

Con este método, se obtienen valores de AUC en el rango de parámetros de especificidad y sensitividad muy buenos (0.8-0.9), excepto por C_{E} y C_{S} correspondientes a los sitios de unión (figura 6.4b), que se encuentran en el rango excelente (0.9-1). Con esto, se advierte una estrecha correlación entre los máximos en C_{E} y LR-C_{S}, y los sitios activos, y en menor medida entre los máximos con C_{S}. Por otro lado, la relación es aún más estrecha entre los mínimos en C_{E} y C_{S} con los sitios de unión, y en menor medida entre los mínimos en LR-C_{S} y los mismos. En conclusión, el método basado en teoría de grafos y medidas de centralidad de orden superior es un muy buen método para la predicción de sitios activos y de unión en la $\mathrm{PL}^{\mathrm{pro}}$ del SARS-CoV-2.

Como se vió en el capítulo 4, C_{E} es interpretada por Estrada como aquella medida que cuantifica la importancia de un nodo en términos de los recorridos o *caminos* de grado infinito, mucho más allá de los vecinos más cercanos. Mientras que C_{S} cuantifica la participación de cada nodo en relación con los subgrafos de todos los tamaños pero sanciona menos los de longitud más corta. Por otro lado, LR-C_{S} cuantifica menos los de longitud

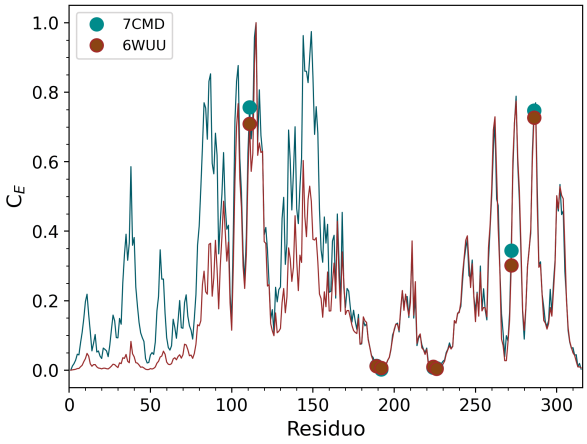
más grande. Dicho de otra manera, los sitios activos participan más activamente en subgrafos de un grado muy alto, principalmente los de grado infinito, y un poco menos en subgrafos de menor grado.

Los residuos catalíticos Cys111 y Asp286 son nodos con una alta conectividad a todos los nodos de la red y, por lo tanto, son capaces de percibir perturbaciones a distancias muy lejanas. Cambios estructurales que ocurren en regiones alejadas al sitio activo son percibidas por el par de residuos catalíticos. Por otro lado, los sitios de unión participan mínimamente en subgrafos de cualquier grado, se encuentran en una región aislada del resto de residuos y no perciben perturbaciones más allá de sus vecinos más cercanos. Con esto, se verifica que la centralidad de eigenvector C_{E} es la más apropiada para la identificación de sitios activos y de unión, seguida de LR-C_{S}.

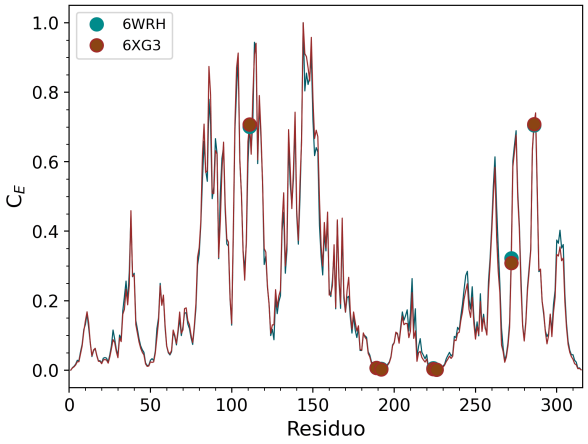
6.3. Inhibidores y mutaciones

La siguiente parte del análisis consiste en determinar la correlación entre las medidas de centralidad de orden superior y las variaciones en la actividad catalítica de los sitios activos como consecuencia de la aplicación de los inhibidores GRL-0617 de molécula pequeña y VIR250 basada en péptidos, y de las mutaciones. Con ello, determinar la capacidad inhibitoria *in silico* de los inhibidores diseñados para $\mathrm{PL^{pro}}$ y los efectos de las mutaciones del residuo Cys111 en la dinámica y estructura de la proteasa.

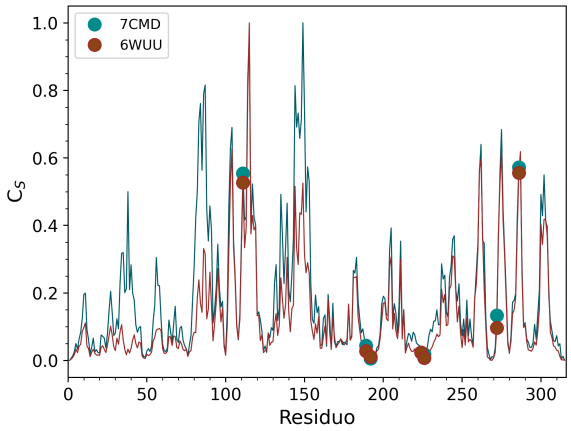
En las gráficas de la figura 6.5 se muestran las medidas de centralidad C_{E}, C_{S} y LR-C_{S} locales normalizadas para los complejos $\mathrm{PL^{pro}}$-GRL0617, $\mathrm{PL^{pro}}$-VIR250 y el par de mutaciones de $\mathrm{PL^{pro}}$ en Cys111 a 100 K y a 293 K. Recordando del capítulo 2, el inhibidor GRL-0617 para $\mathrm{PL^{pro}}$ tiene un $\mathrm{IC_{50}}$ de aproximadamente 1 µM a 2 µM y es cerca de cincuenta veces más potente que VIR250 *in vitro*.



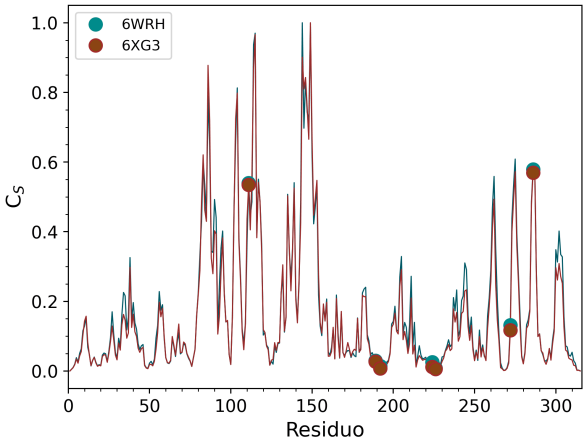
(a) Centralidad de eigenvector C_{E}: inhibidores



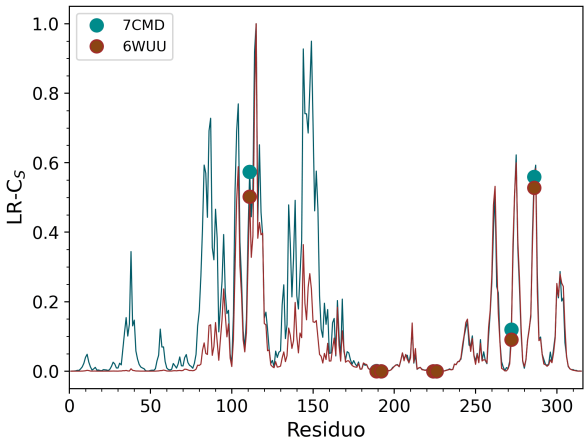
(b) Centralidad de eigenvector C_{E}: mutaciones



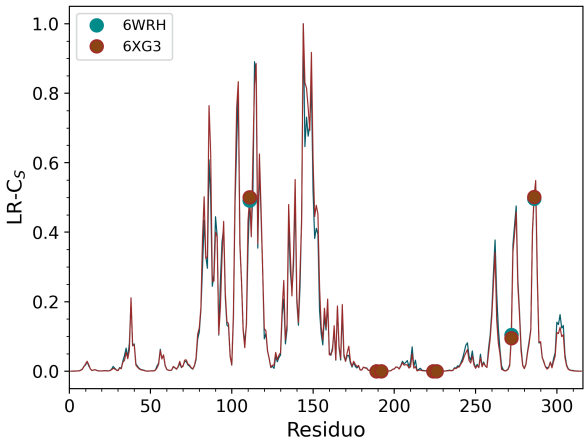
(c) Centralidad de subgrafo C_{S}: inhibidores



(d) Centralidad de subgrafo C_{S}: mutaciones



(e) Centralidad de subgrafo de largo alcance LR-C_{S}: inhibidores



(f) Centralidad de subgrafo de largo alcance LR-C_{S}: mutaciones

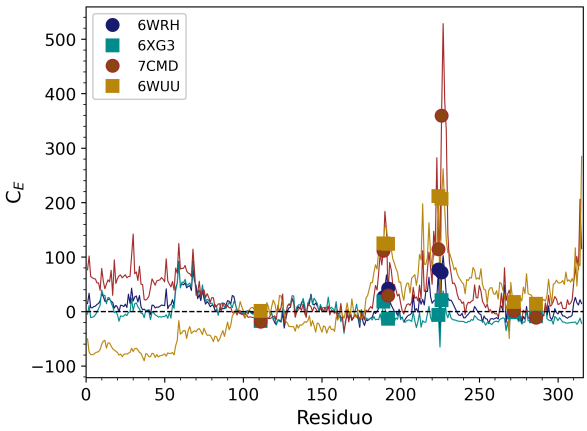
Figura 6.5: Medidas de centralidad de orden superior para los complejos con inhibidores $\mathrm{PL^{pro}}$-GRL0617 (PDB ID: 7CMD) y $\mathrm{PL^{pro}}$-VIR250 (PDB ID: 6WUU), y para las mutaciones de $\mathrm{PL^{pro}}$ en el residuo Cys111 del sitio activo a 100 K (PDB ID: 6WRH) y a 293 K (PDB ID: 6XG3).

Las diferencias entre las centralidades de los sitios activos predispuestas por ambos inhibidores muestran poca diferencia en las tres centralidades locales de orden superior. Sin embargo, se observa un aumento considerable en las centralidades del resto de nodos del domino *Thumb*. Además, en la tabla II se observa que el residuo Cys111 pasa de la posición nueve para la estructura 6WUU a la diecinueve para 7CMD en las centralidades C_{E} y LR-C_{S}. Por otro lado, las diferencias son casi insignificantes en el par de mutaciones de $\mathrm{PL}^{\mathrm{pro}}$ dadas en Cys111 con base a la figura 6.5. A nivel local aparentemente no hay diferencias significativas, sin embargo, más adelante veremos que no es así.

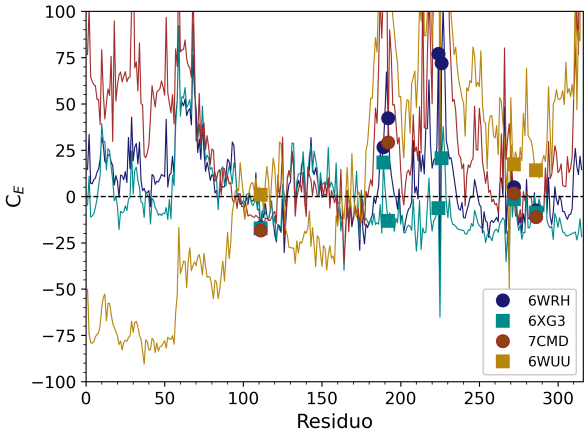
Tabla II. Residuos aminoácidos de proteínas del SARS-CoV-2 con valores máximos en C_{E} en orden descendente. Se obtienen los mismos resultados para los máximos en LR-C_{S}. Los sitios activos se muestran en rojo.

Posición	6W9C	7CMD	6WUU	6WRH	6XG3
1	T115	T115	T115	A144	A144
2	A114	A149	A114	A114	A149
3	I104	A144	H275	T115	T115
4	A144	A114	I104	A149	I104
5	S103	C148	G287	I104	A145
6	L117	I104	S262	N146	A114
7	C111	A145	D286	C148	N146
8	A149	N146	K274	F147	A86
9	C148	L87	C111	S103	F147
10	H275	L150	S103	A145	S103
11	L113	A86	A261	A86	C148
12	G287	S103	L117	L117	L87
13	D286	F147	T119	G287	L117
14	A86	L117	L118	L150	L113
15	F147	H275	L113	L113	L150
16	N146	Y83	A116	A139	A139
17	L150	G287	I276	L87	G287
18	A145	L152	I285	D286	T102
19	S262	C111	A144	S111	Y83
20	K274	M84	Y273	T102	D286
21	T102	D286	Y112	H275	S111

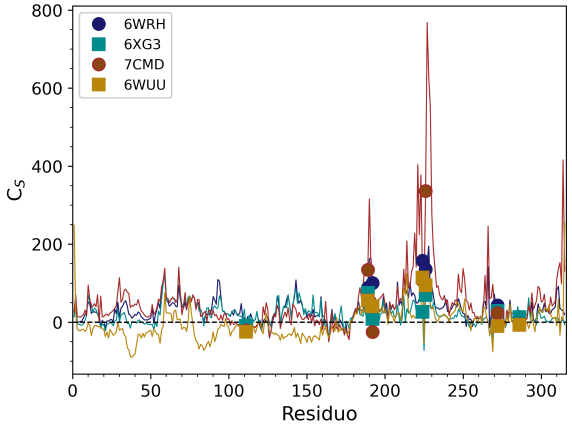
En las gráficas de la figura 6.6 se muestra la diferencia relativa porcentual entre las centralidades del par de estructuras con inhibidores y el par de mutaciones mencionados con respecto a $\mathrm{PL}^{\mathrm{pro}}$. La línea en negro representa a $\mathrm{PL}^{\mathrm{pro}}$ tipo *wild*.



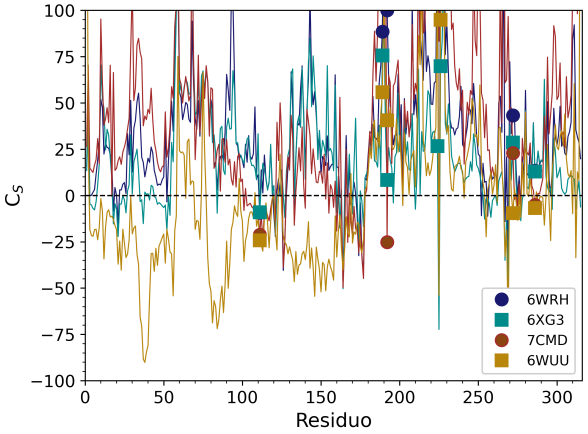
(a) Centralidad de eigenvector C_{E}



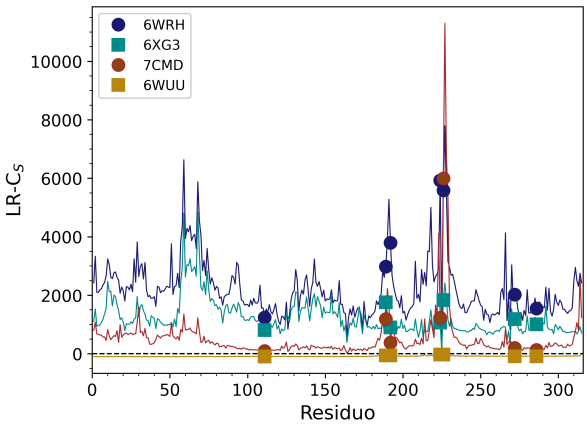
(b) Centralidad de eigenvector C_{E}



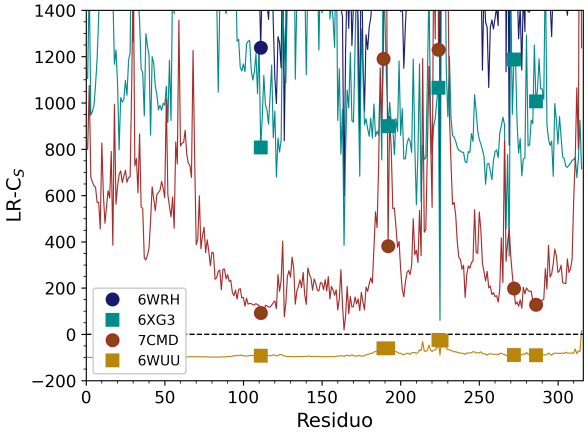
(c) Centralidad de subgrafo C_{S}



(d) Centralidad de subgrafo C_{S}



(e) C. de subgrafo de largo alcance LR-C_{S}



(f) C. de subgrafo de largo alcance LR-C_{S}

Figura 6.6: Diferencia relativa porcentual de las medidas de centralidad de orden superior entre $\mathrm{PL}^{\mathrm{pro}}$ tipo *wild* (PDB ID: 6W9C) y sus complejos con inhibidores y mutaciones. Las centralidades a), c) y e) de la $\mathrm{PL}^{\mathrm{pro}}$ con inhibidores 7CMD y 6WUU, y las centralidades b), d) y f) de la $\mathrm{PL}^{\mathrm{pro}}$ con mutaciones 6WRH y 6XG3.

Comenzando por la centralidad C_{E} de los complejos con inhibidores, se observa una disminución en las centralidades de los residuos a la izquierda de Cys111 y un aumento a la derecha de este para 6WUU. Mientras que para 7CMD hay un aumento a ambos lados de Cys111, además de una disminución del 20 % con respecto a Cys111 de $\mathrm{PL}^{\mathrm{pro}}$ tipo *wild*. En otras palabras, el inhibidor de molécula pequeña GRL-0617 entierra a Cys111 mientras que la acción del inhibidor de péptidos se puede interpretar como un semienterramiento. Este enterramiento también se nota en los resultados de las tablas II y III.

Recapitulando, los inhibidores actúan sobre el sitio activo provocando un desplazamiento de los residuos catalíticos. Este desplazamiento en la estructura tridimensional los desplaza a una región menos privilegiada. Es decir, Cys111 se desconeta de otros residuos y, por lo tanto, disminuye su centralidad. En ese sentido, con enterramiento nos referimos a la disminución de la accesibilidad de un residuo catalítico para unirse a un sustrato.

Tabla III. Residuos aminoácidos de proteínas del SARS-CoV-2 con valores máximos en C_{S} en orden descendente. Las celdas rojas representan los sitios activos.

Posición	6W9C	7CMD	6WUU	6WRH	6XG3
1	T115	A149	T115	A144	A149
2	A114	T115	A114	A149	T115
3	A149	A114	I104	T115	A144
4	A86	L87	H275	A114	A86
5	A144	A144	G287	N146	A114
6	I104	A86	S262	I104	N146
7	H275	M84	A261	A86	A145
8	C111	L150	D286	F147	I104
9	C148	N146	C111	C148	F147
10	S103	C148	A149	A145	S103
11	L150	Y83	K274	S103	L87
12	N146	A145	S103	L87	L150
13	L87	I104	A144	L150	C148
14	S262	H275	F147	H275	Y83
15	G287	F147	C148	Y83	G287
16	F147	S262	L117	G287	H275
17	M84	S103	D302	D286	D286
18	L117	G287	V303	S262	L113
19	Y83	L152	I276	L117	A153
20	D286	D286	I300	A139	L117
21	L113	S85	T119	S111	S111

Continuando con LR-C_S, el enterramiento de Cys111 (tanto en la figura 6.6f como en las tablas II y III) es aún más marcado para la estructura con el inhibidor GRL-0617, mientras que con VIR250 las variaciones entre Cys111 y el resto de residuos son despreciables en comparación con las inducidas por el inhibidor anterior.

Anteriormente, se determinó que C_E es la medida de centralidad más apropiada para la predicción de sitios activos, y C_S es la menos apropiada de las medidas de centralidad de orden superior. Aquí ocurre lo mismo, C_S es menos útil que C_E y LR-C_S para determinar la potencia de los inhibidores de PL^{pro}. Por otro lado, no se perciben variaciones significativas entre las mutaciones de Cys111 a 100 K y a 293 K, pero se detecta un ligero enterramiento de Cys111 por lo que la mutación es desfavorable para la proteína.

En síntesis, el inhibidor GRL-0617 modifica la accesibilidad del residuo Cys111 de PL^{pro}, una característica importante para los sitios activos ya que una alta accesibilidad permite la unión con un sutrato específico. La disminución de la accesibilidad se ve reflejada por la disminución en las centraldidades de orden superior pero, principalmente, por el aumento en las centralidades de los residuos vecinos a los residuos catalíticos como una forma de enterramiento de estos. Además, este inhibidor tiene una mayor capacidad inhibitoria que el inhibidor basado en péptidos, ya que modifica la dinámica y estructura de la proteína no solo en regiones muy próximas a Cys111 sino en regiones aún más lejanas. En el caso de las mutaciones, el ligero enterramiento de Cys111 nos habla de una mutación desventajosa para PL^{pro}.

Ahora se discutirán las variaciones a nivel promedio. En la tabla IV se reportan los resultados de las medidas de centralidad promedio de todas las categorías y en la tabla V las diferencias relativas porcentuales de las variantes con respecto a PL^{pro} tipo *wild*.

Tabla IV. Comparación de las medidas de centralidad promedio de la PL^{pro} tipo *wild* del SARS-CoV-2 (6W9C), PL^{pro}-GRL0617 (7CMD), PL^{pro}-VIR250 (6WUU), las mutaciones de Cys111 a 100 K (6WRH) y a 293 K (6XG3).

Medida	6W9C	7CMD	6WUU	6WRH	6XG3
δ	0.041	0.042	0.041	0.042	0.041
ρ	0.016	0.0168	0.0155	0.0168	0.0168
$\langle C_{\mathrm{B}} \rangle$	0.0124	0.0122	0.0128	0.0122	0.0124
$\langle C_{\mathrm{E}} \rangle$	0.040	0.0427	0.0396	0.0406	0.0398
$\langle C_{\mathrm{S}} \rangle$	21.262 $\times 10^3$	25.607 $\times 10^3$	17.580 $\times 10^3$	26.891 $\times 10^3$	25.701 $\times 10^3$
$\langle G_{pq} \rangle$	6.470 $\times 10^3$	7.890 $\times 10^3$	5.205 $\times 10^3$	8.344 $\times 10^3$	7.799 $\times 10^3$
$\langle \theta \rangle$	68.86	68.22	69.77	68.03	68.47
$\langle Z_{pp} \rangle$	3.015 $\times 10^{46}$	8.696 $\times 10^{46}$	2.354 $\times 10^{45}$	5.79 $\times 10^{47}$	3.999 $\times 10^{47}$
$\langle Z_{pq} \rangle$	1.543 $\times 10^{46}$	4.992 $\times 10^{46}$	1.161 $\times 10^{45}$	3.00 $\times 10^{47}$	1.989 $\times 10^{47}$

Las diferencias de las medidas de centralidad de primer orden no superan el 5 %, por lo tanto, estas medidas tampoco son adecuadas para determinar variaciones en estas estructuras. Las medidas de segundo orden muestran diferencias menores al 29 % mientras que las de tercer orden por arriba del 100 %. Por otro lado, el inhibidor GRL-0617 aumenta aproximadamente dos veces la cantidad disminuida por el inhibidor VIR250 en las centralidades de largo alcance para PL^{pro}. En otras palabras, el inhibidor de molécula pequeña tiene efectos a un nivel estructural más amplio que el inhibidor basado en péptidos. Aunque este aumento para 7CMD parezca contraproducente, anteriormente notamos un enterramiento de Cys111 a nivel local, mucho mayor que el de 6WUU.

Tabla V. Diferencia relativa porcentual entre PL^{pro} tipo *wild* y sus estructuras variantes: 7CMD[1], 6WUU[2], 6WRH[3] y 6XG3[4].

Medida	Δ_{rel1}(%)	Δ_{rel2}(%)	Δ_{rel3}(%)	Δ_{rel4}(%)
δ	-2.44	0.0	-2.44	0.0
ρ	-5.0	3.13	-5.0	-5
$\langle C_{\mathrm{B}} \rangle$	1.61	-3.23	1.61	0.0
$\langle C_{\mathrm{E}} \rangle$	-6.75	1.0	-1.50	5.0
$\langle C_{\mathrm{S}} \rangle$	-20.44	17.32	-26.47	-20.88
$\langle G_{pq} \rangle$	-21.95	19.55	-28.96	-20.54
$\langle \theta \rangle$	0.93	-1.61	1.21	0.57
$\langle Z_{pp} \rangle$	-188.42	92.19	-1,820.40	-1,226.37
$\langle Z_{pq} \rangle$	-223.53	92.48	-1,844.26	-1,189.05

En el caso de las mutaciones, para la proteína 6WRH se observa un aumento de las medidas de tercer orden por arriba del 1800 % y para la proteína 6XG3 un aumento que ronda en el rango de 1100 %-1300 %. Es decir, la mutación a temperatura ambiente es menos favorable para el virus. Por otro lado, el aumento de LR-C_S indica que la mutación mejora la comunicación entre todos los residuos de la red, haciéndola más sensible. Los aminoácidos de la proteína mutada son capaces de percibir mejor cambios estructurales a distancias más lejanas que los percibidos por la proteína sin mutaciones.

6.4. Propiedad de mundo pequeño

Como hemos visto en el capítulo 3, las redes aleatorias siguen una distribución de Poisson cuando el total de nodos N tiende a infinito. En la gráfica de la figura 6.7 se muestra la distribución de grado de la proteína y la distribución de la red aleatoria de Erdös-Rényi equivalente.

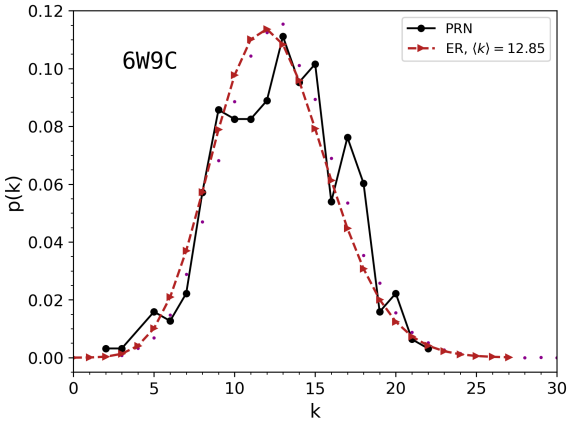


Figura 6.7: Distribución de grado de la proteína 6W9C y del ensamble de 150 grafos aleatorios de Erdös–Rényi equivalentes. La equivalencia consiste en un mismo número de nodos y enlaces, por lo tanto, ambas redes se conforman de $N = 315$ y $E = 1997$.

Ambas redes tienen un comportamiento similar en sus distribuciones de grado, por lo que los aminoácidos de PL^{pro} se distribuyen de la siguiente forma: en la red existen pocos aminoácidos con un bajo y alto número de conexiones, la gran mayoría tienden a un grado promedio $\langle k \rangle$. Por lo tanto, la probabilidad de que las modificaciones aleatorias

sean efectuadas sobre un nodo muy importante o poco importante es baja. Es decir, es más probable que el nodo eliminado tenga grado cercano a $\langle k \rangle$. En pocas palabras, PL^{pro} tiene una capacidad media para resistir cambios a nivel estructural. Cabe mencionar que esto no implica que la red de átomos de C_α de PL^{pro} sea una red aleatoria. Más bien es una red que exhibe comportamientos de aleatoriedad en la forma en la que se distribuyen las conexiones entre sus nodos. Las redes reales no pueden ser redes aleatorias, como hemos visto en capítulo 3.

Ahora se analizará la propiedad de mundo pequeño en las proteínas de interés. Recordemos que una red SW es una red que oscila entre una red regular y una red aleatoria, pero con una mayor proximidad a las redes regulares. Las propiedades de las redes SW consisten en un alto coeficiente de agrupamiento promedio $\langle C \rangle$, como el de las redes regulares, y una baja longitud de ruta promedio $\langle L \rangle$ (propiedad de mundo pequeño), como el de las redes aleatorias. En la tabla VI se muestran los resultados de estos cálculos para PL^{pro} tipo *wild* y sus redes aleatoria y regular equivalentes, los resultados se extienden a las estructuras con inhibidores y mutaciones.

Con base a estos resultados, se observa que para la proteína 6W9C, $\langle L \rangle$ escala aproximadamente como $\ln(N)/\ln(\langle k \rangle)$, entonces cumple la propiedad de mundo pequeño. Por otro lado, el coeficiente $\langle C \rangle$ es alto en comparación con el de la red aleatoria, ya que es aproximadamente siete veces más alto. Para el resto de estructuras los resultados son muy similares, por lo tanto, se concluye que la proteína PL^{pro} es una red SW al igual que las estructuras con inhibidores 7CMD y 6WUU. La conclusión es la misma para las mutaciones 6WRH y 6XG3.

Tabla VI. Longitud de ruta promedio $\langle L \rangle$ y coeficiente de agrupamiento $\langle C \rangle$ de PL^{pro} del SARS-CoV-2 (6W9C), y sus equivalentes en redes regular y aleatoria de Erdös–Rényi (ER). Para que las redes sean equivalentes deben de compartir el mismo número de nodos N y enlaces E al de la red de residuos de proteína (PRN). Para la red regular se utilizó el modelo de Watts y Strogatz $\mathcal{G}_{N,k,p}$ con un grado de nodo promedio k y una probabilidad $p = 0$.

Proteína	Red regular	ER	PRN
6W9C	$\langle L \rangle = 13.5860$	$\langle L \rangle = 1.4861$	$\langle L \rangle = 4.8856$
	$\langle C \rangle = 0.6818$	$\langle C \rangle = 0.0410$	$\langle C \rangle = 0.5603$
	$k = 13$		
7CMD	$\langle L \rangle = 13.5860$	$\langle L \rangle = 2.5137$	$\langle L \rangle = 4.8099$
	$\langle C \rangle = 0.6818$	$\langle C \rangle = 0.0418$	$\langle C \rangle = 0.5639$
	$k = 13$		
6WUU	$\langle L \rangle = 13.5860$	$\langle L \rangle = 2.5404$	$\langle L \rangle = 4.9938$
	$\langle C \rangle = 0.6818$	$\langle C \rangle = 0.0403$	$\langle C \rangle = 0.5607$
	$k = 13$		
6WRH	$\langle L \rangle = 13.5860$	$\langle L \rangle = 2.5204$	$\langle L \rangle = 4.8315$
	$\langle C \rangle = 0.6818$	$\langle C \rangle = 0.0418$	$\langle C \rangle = 0.5604$
	$k = 13$		
6XG3	$\langle L \rangle = 13.5860$	$\langle L \rangle = 2.5264$	$\langle L \rangle = 4.8673$
	$\langle C \rangle = 0.6818$	$\langle C \rangle = 0.0408$	$\langle C \rangle = 0.5598$
	$k = 13$		

Lo anterior implica que, dado que $\langle L \rangle$ es bajo, los nodos de la red se encuentran cercanos entre sí y la comunicación fluye rápidamente a través de toda la red. Por otro lado, dado que $\langle C \rangle$ es alto, existen regiones con una alta agrupación, regiones aún mucho más conectadas. En otras palabras, la red es resistente a ataques directos (como es el caso de la aplicación de los inhibidores en regiones específicas como sus sitios activos) y a variaciones estructurales (por ejemplo, las mutaciones) tanto por ser una red SW como por la forma en la que se distribuyen sus enlaces, como una distribución de Poisson.

Ahora se compararán estas propiedades entre las cinco estructuras. Se observa que $\langle L \rangle$ disminuyó para 7CMD mientras que aumentó para 6WUU, por otro lado, los cambios en $\langle C \rangle$ son mínimos. Desde esta perspectiva, el inhibidor de molécula pequeña GRL-0617 reduce la comunicación entre algunos nodos de la red al desconectar a algunos de estos, mientras que el inhibidor basado en péptidos produce el efecto contrario. Por otro lado, las mutaciones tienen efectos perjudiciales en términos de $\langle L \rangle$ pero casi nulos en $\langle C \rangle$.

En esta sección se concluye que PL^{pro} es una red SW con una distribución de Poisson, lo que le confiere una gran robustez. Aunque el inhibidor de molécula pequeña GRL-0617 demostró tener una mayor capacidad inhibitoria para PL^{pro} que el inhibidor VIR250, notamos que la robustez en la estructura impidió una mayor afectación a esta. Sin embargo, los ataques directos dados por inhibidores son más significativos que las variaciones producidas por el mismo virus, es decir, las mutaciones.

Capítulo 7

Conclusiones

En este trabajo se analizó el sitio activo y sitio de unión de la proteasa viral PL^{pro} del SARS-CoV-2 mediante un enfoque de teoría de redes y medidas de centralidad. Se modeló a la proteasa a partir de sus átomos de C_α como nodos y sus interacciones covalentes, y no covalentes como enlaces. La información cistalográfica de estas estructuras se obtuvo del banco de datos *Protein Data Bank* (PDB). Para esto se seleccionó un $R_c = 9$ Å como la distancia más óptima a la que interactúan los aminoácidos de la proteína. Se aplicaron las medidas de centralidad de redes de primer orden (C_B), segundo orden (C_E y C_S) y de tercer orden (LR-C_S), además de otras medidas promedio. Se extendió el análisis a las estructuras de PL^{pro} con los inhibidores GRL-0617 de molécula pequeña y VIR250 basada en péptidos, y a las mutaciones de PL^{pro} en el residuo catalítico Cys111 a 100 K y a 293 K. Además de esto, se analizaron las propiedades de longitud de ruta promedio $\langle L \rangle$ y coeficiente de agrupamiento $\langle C \rangle$ para determinar si la red es una red de mundo pequeño (SW) y, con ello, verificar la robustez de la red.

Con el método del área bajo la curva AUC-ROC se determinó que los residuos del sitio activo se relacionan con los máximos en C_E, LR-C_S y C_S, dado que se obtuvieron valores para AUC en el rango de parámetros de especificidad y sensitividad muy buenos. Al mismo tiempo, se estableció una relación similar entre los mínimos en las medidas de centralidad mencionadas y los residuos del sitio de unión, con mejores resultados de AUC,

en el rango muy bueno y excelente. En particular, los mejores resultados se obtuvieron con C_{E} y, seguidamente, con LR-C_{S}. El método AUC-ROC es ampliamente empleado para determinar la eficacia de métodos de aprendizaje profundo. Sin embargo, dada su versatilidad y por la naturaleza de la investigación se pudo emplear exitosamente. Este método se puede considerar una versión más precisa y confiable que otros métodos como el *recall* al considerar una gran cantidad de umbrales.

Con base a estos resultados se concluye que los residuos del sitio activo de $\mathrm{PL}^{\mathrm{pro}}$ participan activamente en subgrafos de grado muy alto a infinito. Por lo tanto, son capaces de percibir perturbaciones a distancias más allá de sus vecinos más cercanos, mientras que ocurre lo contrario con los sitios de unión. Los residuos catalíticos son dotados de una mayor sensibilidad para percibir alteraciones en la estructura de la proteína que el resto de residuos en ella. En consecuencia, el enfoque en teoría de grafos y, particularmente, de las medidas de centralidad de orden superior C_{E} y LR-C_{S} representan un buen método para la predicción del sitio activo y de unión en $\mathrm{PL}^{\mathrm{pro}}$ del SARS-CoV-2.

Como parte de la comparación entre las estructuras unidas a los inhibidores y de las mutaciones con respecto a $\mathrm{PL}^{\mathrm{pro}}$ tipo *wild*, se obtuvieron variaciones menores al 4 % en propiedades de primer orden. En propiedades de segundo orden menores al 29 % y en el rango de 92 %-1844 % en propiedades de tercer orden. En particular, el inhibidor GRL-0617 aumentó LR-C_{S} en un 188 %, sin embargo, a nivel local se advirtió un enterramiento del residuo catalítico principal del sitio activo, mucho mayor que el inducido por VIR250. Dicho de otro modo, el inhibidor de molécula pequeña demostró tener una mayor capacidad inhibitoria que el inhibidor basado en péptidos debido a su alta especificidad. Por otro lado, los hallazgos más significativos con respecto a las mutaciones se presentaron en las medidas de centralidad promedio de orden superior. En ambas mutaciones se obtuvieron variaciones por arriba de 1200 % en medidas de tercer orden, lo que implica un aumento en la sensibilidad de $\mathrm{PL}^{\mathrm{pro}}$ en afectaciones a nivel estructural de larga distancia.

Como última parte del análisis, se obtuvo que PL^{pro} es una red SW porque cumple con dos propiedades importantes: (a) la propiedad de mundo pequeño ya que $\langle L \rangle$ escala aproximadamente como $\ln(N)/\ln(k)$, con k como el grado promedio de la red aleatoria; (b) un valor alto de $\langle C \rangle$. El resto de estructuras también cumplieron con esta propiedad. Además de esto, PL^{pro} sigue una distribución de grado similar a la distribución de Poisson, por lo tanto, es resistente a modificaciones aleatorias a nivel estructural. Ambos resultados se traducen en una mayor capacidad de los aminoácidos de PL^{pro} en comunicarse eficientemente entre sí y en resistir perturbaciones, como las inducidas por los inhibidores, y en resistir variaciones genéticas provocadas por el mismo virus.

El análisis de la estructura tridimensional de PL^{pro} del SARS-CoV-2 mediante medidas de centralidad de orden superior es apropiado para localizar a residuos importantes en la proteína. Estos residuos pueden ser tanto los que conforman sus sitios activos y de unión. Sin embargo, también fue necesario un análisis de sus distribuciones de grado y propiedades $\langle L \rangle$ y $\langle C \rangle$ para determinar comportamientos a nivel global. Estas herramientas son valiosas para obtener información sobre la manera en la que la proteína actúa frente a ataques aleatorios o dirigidos, o a transformaciones internas. Determinan el nivel de fragilidad de una red. Esto nos da una idea general sobre cómo podemos abordar el diseño de inhibidores para frenar las afectaciones ocasionadas por la replicación del virus.

Bibliografía

Aguilar-Pineda, G. E., y Olivares-Quiroz, L. (2021). Catalytic and binding sites prediction in globular proteins through discrete markov chains and network centrality measures. *Physical Biology*, *18*(6), 066002.

Albert, R., y Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of modern physics*, *74*(1), 47.

Babor, M., Gerzon, S., Raveh, B., Sobolev, V., y Edelman, M. (2008). Prediction of transition metal-binding sites from apo protein structures. *Proteins: Structure, Function, and Bioinformatics*, *70*(1), 208–217.

Barabási, A.-L. (2013). Network science. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *371*(1987), 20120375.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., y Hwang, D.-U. (2006). Complex networks: Structure and dynamics. *Physics reports*, *424*(4-5), 175–308.

Bonacich, P. (1987). Power and centrality: A family of measures. *American journal of sociology*, *92*(5), 1170–1182.

Calleja, D. J., Lessene, G., y Komander, D. (2022). Inhibitors of sars-cov-2 plpro. *Frontiers in Chemistry*, *10*, 876212.

Capra, J. A., Laskowski, R. A., Thornton, J. M., Singh, M., y Funkhouser, T. A. (2009). Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3d structure. *PLoS computational biology*, *5*(12), e1000585.

Cheikhrouhou, O., y Khoufi, I. (2021). A comprehensive survey on the multiple traveling salesman problem: Applications, approaches and taxonomy. *Computer Science Review*, *40*, 100369.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., ... others (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422.

De Wit, E., Van Doremalen, N., Falzarano, D., y Munster, V. J. (2016). Sars and mers: recent insights into emerging coronaviruses. *Nature reviews microbiology*, *14*(8), 523–534.

Erdős, P., y Rényi, A. (1959). On random graphs i. *Publ. math. debrecen*, *6*(290-297), 18.

Estrada, E. (2012). *The structure of complex networks: theory and applications*. Oxford University Press, USA.

Estrada, E. (2020). Topological analysis of sars cov-2 main protease. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *30*(6).

Estrada, E., y Rodriguez-Velazquez, J. A. (2005). Subgraph centrality in complex networks. *Physical Review E*, *71*(5), 056103.

Estrada, E., y Silver, G. (2017). Accounting for the role of long walks on networks via a new matrix function. *Journal of Mathematical Analysis and Applications*, *449*(2), 1581-1600.

Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 128–140.

Finkelstein, A. V., y Ptitsyn, O. (2016). Protein physics: a course of lectures. , 64.

Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 35–41.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, *1*(3), 215-239.

Frieze, A., y Karoński, M. (2016). *Introduction to random graphs*. Cambridge University Press.

Hagberg, A. A., Schult, D. A., y Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx. En G. Varoquaux, T. Vaught, y J. Millman (Eds.), *Proceedings of the 7th python in science conference* (p. 11 - 15). Pasadena, CA USA.

Hamming, I., Timens, W., Bulthuis, M., Lely, A., Navis, G. v., y van Goor, H. (2004). Tissue distribution of ace2 protein, the functional receptor for sars coronavirus. a first step in understanding sars pathogenesis. *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland*, *203*(2), 631–637.

Jung, T., y Grune, T. (2012). Structure of the proteasome. *Progress in molecular biology and translational science*, *109*, 1–39.

Kirchhoff, G. (1847). Ueber die auflösung der gleichungen, auf welche man bei der untersuchung der linearen vertheilung galvanischer ströme geführt wird. *Annalen der Physik*, *148*(12), 497–508.

Lubin, J. H., Zardecki, C., Dolan, E. M., Lu, C., Shen, Z., Dutta, S., ... others (2022). Evolution of the sars-cov-2 proteome in three dimensions (3d) during the first 6 months of the covid-19 pandemic. *Proteins: Structure, Function, and Bioinformatics*, *90*(5), 1054–1080.

Majerová, T., y Konvalinka, J. (2022). Viral proteases as therapeutic targets. *Molecular Aspects of Medicine*, *88*, 101159.

Milgram, S. (1967). The small world problem. *Psychology today*, *2*(1), 60–67.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., y Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, *298*(5594), 824–827.

Mojica-Crespo, R., y Morales-Crespo, M. (2020). Pandemia covid-19, la nueva emergencia sanitaria de preocupación internacional: una revisión. *Medicina de Familia. SEMERGEN*, *46*, 65–77.

Narkhede, S. (2018). Understanding auc-roc curve. *Towards data science*, *26*(1), 220–227.

Negre, C. F., Morzan, U. N., Hendrickson, H. P., Pal, R., Lisi, G. P., Loria, J. P., ... Batista, V. S. (2018). Eigenvector centrality for characterization of protein allosteric pathways. *Proceedings of the National Academy of Sciences*, *115*(52), E12201–E12208.

Nelson, D. L., Lehninger, A. L., y Cox, M. M. (2008). *Lehninger principles of biochemistry*. Macmillan.

Newman, M. (2018). *Networks*. Oxford university press.

Olivares-Quiroz, L., y García-Colín Scherer, L. (2004). Plegamiento de las proteínas: Un problema interdisciplinario. *Revista de la Sociedad Química de México*, *48*(1), 95–105.

OMS. (2023). Declaración acerca de la decimoquinta reunión del comité de emergencias del reglamento sanitario internacional (2005) sobre la pandemia de enfermedad por coronavirus (covid-19).

OMS. (2024). Covid-19 epidemiological update – 19 january 2024.

Osipiuk, J., Azizi, S.-A., Dvorkin, S., Endres, M., Jedrzejczak, R., Jones, K. A., ... others (2021). Structure of papain-like protease from sars-cov-2 and its complexes with non-covalent inhibitors. *Nature communications*, *12*(1), 743.

Rodrigues, J., Teixeira, J., Trellet, M., y Bonvin, A. (2018). pdb-tools: a swiss army knife for molecular structures [version 1; peer review: 2 approved]. *F1000Research*, *7*(1961). doi: 10.12688/f1000research.17456.1

Sampath, R., Srinivasan, K., Tharaniya, P., y Elumalai, P. (2021). Alternative solution for konigsberg bridge problem through the concept of matching. *International Journal of Pharmaceutical Research (09752366)*, *13*(1).

Shen, H.-B., y Chou, K.-C. (2009). Identification of proteases and their types. *Analytical biochemistry*, *385*(1), 153–160.

Ullrich, S., y Nitsche, C. (2022). Sars-cov-2 papain-like protease: structure, function and inhibition. *ChemBioChem*, *23*(19), e202200327.

Villarreal, L. P. (2004). Are viruses alive? *Scientific American*, *291*(6), 100–105.

Voet, D., Voet, J. G., y Pratt, C. W. (2016). *Fundamentos de bioquímica: La vida a nivel molecular*. Editorial Médica Panamericana.

Wang, Q., Chen, G., He, J., Li, J., Xiong, M., Su, H., ... Xu, Y. (2023). Structure-based design of potent peptidomimetic inhibitors covalently targeting sars-cov-2 papain-like protease. *International Journal of Molecular Sciences*, *24*(10), 8633.

Watts, D. J., y Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. *nature*, *393*(6684), 440–442.

Yadav, R., Chaudhary, J. K., Jain, N., Chaudhary, P. K., Khanra, S., Dhamija, P., ... Handu, S. (2021). Role of structural and non-structural proteins and therapeutic targets of sars-cov-2 for covid-19. *Cells*, *10*(4), 821.

Yilmaz, N. K., Swanstrom, R., y Schiffer, C. A. (2016). Improving viral protease inhibitors to counter drug resistance. *Trends in microbiology*, *24*(7), 547–557.

Zhang, X., Wu, F., Yang, N., Zhan, X., Liao, J., Mai, S., y Huang, Z. (2022). In silico methods for identification of potential therapeutic targets. *Interdisciplinary Sciences: Computational Life Sciences*, 1–26.