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

COLEGIO DE CIENCIA Y TECNOLOGÍA

LICENCIATURA EN MODELACIÓN MATEMÁTICA

Algoritmos de programación en Python para el análisis de la estructura de proteínas del virus SARS-CoV-2 mediante una representación en redes de carbonos C_α y C_β de la estructura globular

TESIS

QUE PARA OPTAR POR EL TÍTULO DE

LICENCIADO EN MODELACIÓN MATEMÁTICA

PRESENTA:

OHTLI GERARDO QUIROZ SÁNCHEZ

DIRECTOR

DR. LUIS AGUSTÍN OLIVARES QUIROZ

Ciudad de México, octubre de 2023.

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

CA Coordinación Académica

SISTEMA BIBLIOTECARIO DE INFORMACIÓN Y DOCUMENTACIÓN

UNIVERSIDAD AUTÓNOMA DE LA CIUDAD DE MÉXICO
COORDINACIÓN ACADÉMICA

RESTRICCIONES DE USO PARA LAS TESIS DIGITALES

DERECHOS RESERVADOS©

Agradecimiento

A mis padres que siempre han manifestado su amor hacia mí y mis hermanos.

Índice general

Capítulo 1

Introducción

1.1. El virus SARS-CoV-2: estructura y mecanismo de infección.

La aparición en diciembre de 2019 de un nuevo virus en la ciudad de Wuhan, China, que se extendió en el mundo con más de 700 millones de casos confirmados y más de 6 millones de víctimas mortales [1] ha resultado en uno de los eventos con mayores repercusiones sociales, de salud pública y financieras de las últimas dos décadas. El SARS CoV-2 es un virus con un genoma de ARN monocatenario de sentido positivo perteneciente a la familia coronaviridae [2]. A la misma familia pertenecen dos virus potenciales conocidos en los últimos 17 años: el coronavirus causante del síndrome respiratorio de Oriente Medio (MERS-CoV) que tiene 50 % de similaridad genética con el nuevo virus [2], y causó la muerte de 858 personas [3]; y el coronavirus que ocasionó el síndrome respiratorio agudo severo (SARS-CoV) que cobró la vida de 774 personas [4] y tiene una alta similaridad con el nuevo virus en 79 % [2].

La estructura viral de SARS-CoV-2 posee 4 tipos de proteínas conocidos como proteínas estructurales: nucleocapside (N), membrana (M), envoltura (E) y spike (S) [5]. La cadena de ARN tiene aproximadamente 30 mil pares de nuclebases que codifican otras 16 proteínas conocidas como proteínas no estructurales (nsps) que ocupan las dos terceras partes del genoma y son responsables de la replicación viral, Figura 1.1. La tercera parte del genoma restante está dedicada a las proteínas estructurales y, además, 9 proteínas accesorios que participan indirectamente en la replicación viral [5].

El SARS-CoV-2 logra su entrada celular principalmente a través de la proteína estructural S [7]. Una vez que el genoma del SARS-CoV-2 es liberado en el citoplasma se traducirá en la poliproteína 1a o la poliproteína 1ab por el ribosoma huésped [8]. Por consiguiente, las poliproteínas son escindidas en proteínas más pequeñas que resultan ser las 16 nsps por la actividad proteolítica de la proteasa nsp5, llamada también proteasa

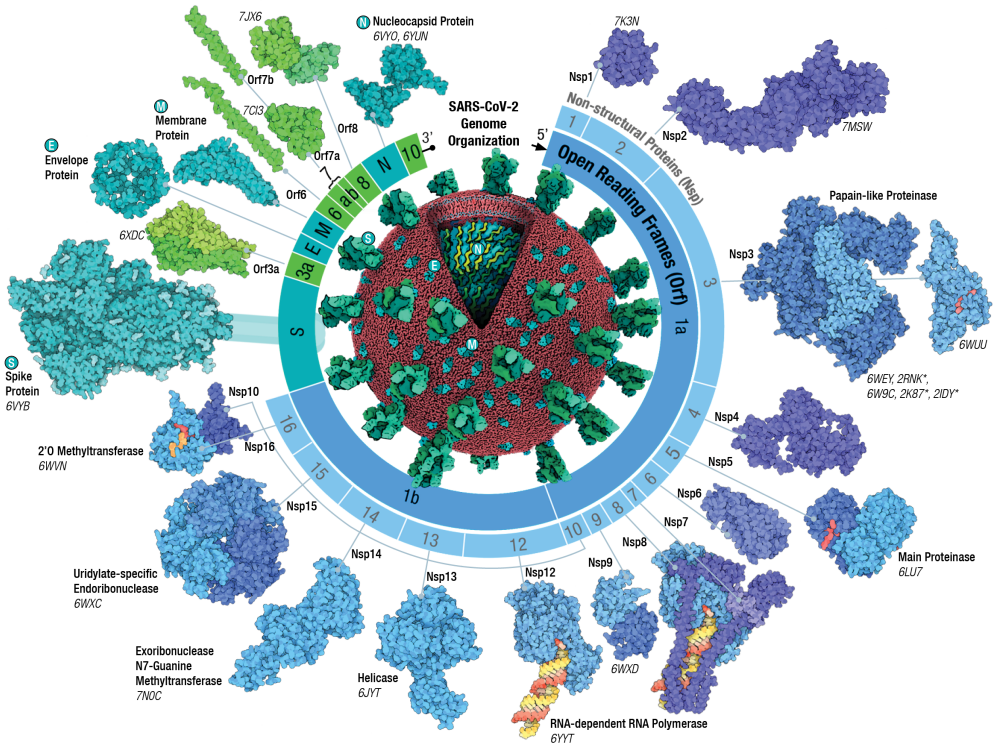


Figura 1.1: Esquema de la organización del genoma de SARS-CoV-2. Imagen tomada de [6].

principal (en inglés main protease, abreviada Mpro), y la proteasa nsp3 (abreviada PLpro en inglés por papain-like protease). Inhibir tan solo la función catalítica de la proteína nsp5 resultaría en interrumpir el ciclo de replicación viral debido a que las nsps componen la maquinaria de transcripción y replicación de ARN viral, y es por eso que es un objetivo potencial en el diseño farmacológico [9].

En este trabajo estudiaremos las proteínas nsp3, nsp5 y la proteína estructural S que tienen un papel clave en la replicación viral por lo que son objetivos potenciales en el diseño farmacológico. Nos enfocaremos en el análisis de la estructura globular y los sitios activos por las cuales las proteínas realizan su función. En el próximo cuadro se muestra la descripción, el código PDB (que veremos en la tercera sección 1.3) y el porcentaje de similaridad con las proteínas homólogas al SARS-CoV-1 con el que tiene mayor similitud genética. Cabe mencionar, que este antecedente con coronavirus anteriores dio la ventaja de poder conocer y enfrentar mejor el SARS-CoV-2 resultando en la elaboración de vacunas en tiempo record. En conjunto, se ha elaborado terapias antivirales como el PF-07321332 de administración oral que inhibe los sitios activos en la proteasa principal nsp5 reduciendo el riego de hospitalización o muerte en un 88 % en pacientes que se les administró durante los 5 días de presentar los síntomas [10].

Nombre	NSP3	NSP5	Spike (S)
PDB	7LG7	7C6S	6VYB y 6VXX
Porcentaje de similaridad con SARS-CoV	75.82 %	96.08 %	75.96 %
Número de aminoácidos	169	306	2875 y 2916
Cadenas	A	A	A, B y C
Función	Libera nsp1 y nsp2, interrumpe el sistema inmunológico e interactúa con nsp4 y nsp6 para modificar la membrana del retículo endoplasmático y proteger el complejo RTC.	Libera las proteínas replicasa nsp4-nsp16 de la poliproteína.	Se une al receptor ACE2 para después fusionarse con la membrana de la célula y desencadenar el proceso de replicación viral.

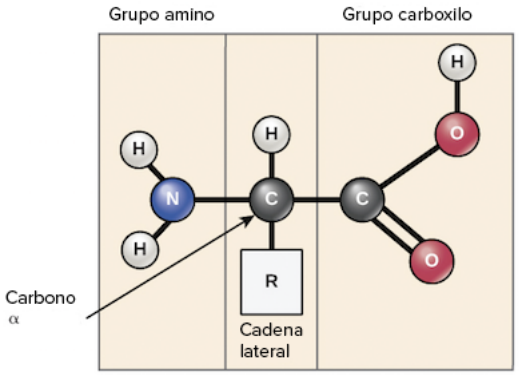


Figura 1.2: Esquema de la estructura general de un aminoácido. El carbono α, mostrada en la imagen, conecta a un grupo carboxilo, un grupo amino, un átomo de hidrógeno y una cadena lateral distinta para cada tipo de aminoácido. Imagen tomada de [11].

1.2. Plegamiento de proteínas y conformación nativa.

Las proteínas se caracterizan por su secuencia lineal de aminoácidos [12]. Hay principalmente 20 tipos de aminoácidos y cuentan con un grupo carboxilo $COOH$, un grupo

amino H_2N y una cadena lateral R (distinta para cada tipo de aminoácido) unidos por un átomo de carbono C_α como se muestra en la Figura (1.2). En la cadena R, el primer átomo es el carbono C_β a excepción de glicina que es el aminoácido más sencillo que posee un átomo de hidrógeno como cadena lateral. Los aminoácidos están dispuestos en una cadena lineal por la unión de un grupo carboxilo de un aminoácido y el grupo amino de otro en una reacción de condensación, el cual produce moléculas de agua [13]. Estos enlaces son conocidos como enlaces peptídicos y la cadena principal de carbonos C_α y los grupos que forman los enlaces peptídicos es conocido como esqueleto (en inglés, backbone). Para que una cadena de aminoácidos sea considerada una proteína debería tener mínimo 50 aminoácidos [12].

En la década de 1960, Christian Anfinsen llevó a cabo una serie de experimentos para descifrar la secuencia de aminoácidos de una proteína llamada ribonucleasa A, y averiguar qué determinaba la conformación final y estructuralmente funcional de una proteína, también conocido como conformación nativa [13]. La ribonucleasa A (también abreviada como RNasa-A) es una proteína de 124 aminoácidos, y resulta sencilla de obtener en estado purificado, cuya función es escindir el ácido ribonucleico (ARN) [13]. Posee 4 enlaces covalentes importantes para la estabilización de la estructura conocidos como enlaces disulfuro. Para desplegar la RNasa-A en su conformación primaria, Anfinsen mezcló dos desnaturalizantes conocidos como urea y mercaptoetanol. El resultado fue que al disminuir la concentración de los desnaturalizantes, sin más que el agua como solvente, la proteína volvía a su conformación final y los cuatro puentes de disulfuro volvían a establecerse. Anfinsen demostró que la estructura tridimensional de la proteína estaba determinada exclusivamente por la secuencia lineal de aminoácidos sin necesidad de la maquinaria celular [13, 14]. Anfinsen también demostraba que el plegamiento de la proteína es un proceso espontáneo y, por lo tanto, un estado de menor energía libre, es decir, un estado termodinámicamente estable [13].

Alrededor de 1970, Cyrus Levinthal sugería que el plegamiento de las proteínas tendría que propagarse a través de un grupo de aminoácidos cuya interacción sirve como núcleo [15]. Levinthal basaba su razonamiento a lo que hoy en día lleva su nombre, *la paradoja de Levinthal*. La paradoja indica que en el número astronómicamente grande de conformaciones que puede tener la cadena de aminoácidos y el tiempo bastante corto en conformarse el estado nativo [13]. Si cada aminoácido en la cadena tuviera dos posibles conformaciones, uno de acuerdo al estado nativo y otro para la conformación incorrecta. El tiempo en plegarse una proteína media de unos cientos de aminoácidos, considerando que un aminoácido pasa de una conformación a otra en términos de nanosegundos, conllevaría un tiempo mayor a la edad del universo [13].

Si observamos, los aminoácidos interactúan de cierta forma entre si y las moléculas

	Hidrofóbico	Neutral	Polar
7LG7	41.3 %	25.7 %	32.9 %
7C6S	40.2 %	30.2 %	29.6 %
6VYB	38.7 %	31.7 %	29.6 %
6VXX	38.8 %	31.5 %	29.7 %

Cuadro 1.2: Proporción de aminoácidos hidrofóbicos, neutrales y polares para cada una de las proteínas objetivo.

de agua, y estas pueden ser interacciones hidrofóbicas o interacciones electrostáticas. Los primeros son los aminoácidos no polares, también llamados hidrofóbicos por su incapacidad de interactuar con el agua, y los segundos son tanto los aminoácidos polares (también hidrófilos) como los aminoácidos neutrales. Los aminoácidos polares tienen una carga parcial positiva y parcial negativa que les permite tener mejor interacción electrostática que los neutrales que no poseen una carga neta. Sin embargo, los neutrales igualmente pueden interactuar con los moléculas de agua son polares [16, 17]. La Figura 1.3 muestra a qué tipo pertenece cada uno de los 20 aminoácidos: *polar*, *hidrofóbico* o *neutral*. En el Cuadro 1.2 se muestra la proporción de aminoácidos tipo hidrofóbico, polar y neutral para cada una de las proteínas objetivo: 7LG7, 7C6S, 6VYB y 6VXX. Podemos ver que los aminoácidos hidrofóbicos son mayoría en cualquiera de las proteínas a pesar de la dimensión. Cabe mencionar que una de las interacciones electrostáticas importantes son los llamados *puentes de hidrógeno*, que tienen un rango de energía mucho más débil al de los enlaces covalentes, pero que, sin embargo, proporcionan estabilidad a estructuras superiores en el proceso de la conformación nativa. Únicamente los aminoácidos hidrofóbicos no pueden crear puentes de hidrógeno con otras moléculas y es por ello que son repelidos por las moléculas polares como el agua. Un ejemplo adicional de la importancia de los puentes de hidrógeno es que son parte esencial para la formación de la doble hélice del ácido desoxirribonucleico (ADN) [17].

El proceso de transformación de la cadena de aminoácidos a la conformación nativa, llamado plegamiento de la proteína (protein folding, en inglés), se propaga a través de un núcleo, o glóbulo compacto, debido a los aminoácidos hidrofóbicos por la incapacidad de crear puentes de hidrógeno con las moléculas de agua, la cual se encuentra mayormente en el ambiente de la proteína [13]. Este efecto provoca que los aminoácidos hidrofóbicos tiendan a agruparse entre sí lejos del ambiente polar, también conocido como efecto hidrofóbico [13]. Es por esto que en este trabajo nos enfocaremos en parte en analizar los aminoácidos que están más aglutinados en el interior del glóbulo de la proteína, ya que podrían jugar un papel clave tanto en el plegamiento como en la estabilidad estructural que está íntimamente relacionada con la funcionalidad biológica de las proteínas [14, 18].

En 1951, Linus Pauling y Robert Corey propusieron, con fundamentos físicos y conociendo las propiedades estructurales de la cadena de aminoácidos, una de las conforma-

Aminoácido		Tipo
Nombre	Símbolo	
Isoleucina	I	Hidrofóbico
Valina	V	Hidrofóbico
Leucina	L	Hidrofóbico
Fenilalanina	F	Hidrofóbico
Cisteína	C	Hidrofóbico
Metionina	M	Hidrofóbico
Alanina	A	Hidrofóbico
Glicina	G	Neutral
Treonina	T	Neutral
Serina	S	Neutral
Triptófano	W	Neutral
Tirosina	Y	Neutral
Prolina	P	Neutral
Histidina	H	Polar
Glutamina	Q	Polar
Asparagina	N	Polar
Ácido Glutámico	E	Polar
Ácido aspártico	D	Polar
Lisina	K	Polar
Arginina	R	Polar

Figura 1.3: Un aminoácido es del tipo polar, hidrofóbico o neutral [19].

ciones superiores de la cadena de aminoácidos a la que llamaron hélice α, tan solo seis años antes de que la estructura fuera vista por primera vez en la proteína mioglobina por la difracción de rayos X [13, 17]. La hélice α toma la forma de un cilindro, y las cadenas laterales de todos los aminoácidos se extienden hacia afuera, véase la Figura 1.4. La estructura se estabiliza por medio de los puentes de hidrógeno entre los grupos NH y CO del backbone de cada aminoácido con excepción de los que están próximos al final de la estructura. Un grupo CO se enlaza con el grupo NH del cuarto aminoácido próximo, situado a 1.5 Å de distancia [17]. Pauling y Corey propusieron también la estructura hoja plegada β conformada por dos o más cadenas de aminoácidos, llamadas hebras, las cuales se presentan casi extendidas, véase la Figura 1.4. Las cadenas laterales de aminoácidos adyacentes de una hebra están extendidas en sentido contrario, véase la Figura 1.4. La hebra se estabiliza, al igual que los hélices α, por puentes de hidrógeno entre el grupo NH de un aminoácido y el grupo CO del aminoácido de la hebra adyacente, donde la distancia de un aminoácido a otro adyacente es aproximadamente de 3.5Å [17]. Existen otras estructuras pero las hélices α y hojas β son las dos estructuras más frecuentes que forman la cadena de aminoácidos encaminadas a la conformación biológicamente funcional de las proteínas [13].

En la Figura 1.7 podemos ver las estructuras hélice y hoja en las proteínas del SARS-CoV-2. En la proteína Spike, (c) y (d), se puede ver varias hélices entre las cuales hay tres llamadas HR1, responsables de fusionarse con la membrana celular huésped y la entrada

viral, por lo cual son objetivos potenciales en el diseño de fármacos [20, 21]. Por lo tanto, los sitios activos que trabajaremos, y de los que hablaremos más adelante, se encuentran en la estructura HR1 de la cadena A (en color azul). Por consiguiente, la parte superior de la proteína Spike está dominada por la estructura en hoja. Estas inician un proceso importante de infección viral interactuando con los receptores de la célula. Especialmente, la parte llamada RBD, de la que también hablaremos más adelante, se ancla al receptor ACE2 de la célula, y es crítica en el diseño de anticuerpos neutralizantes [20].

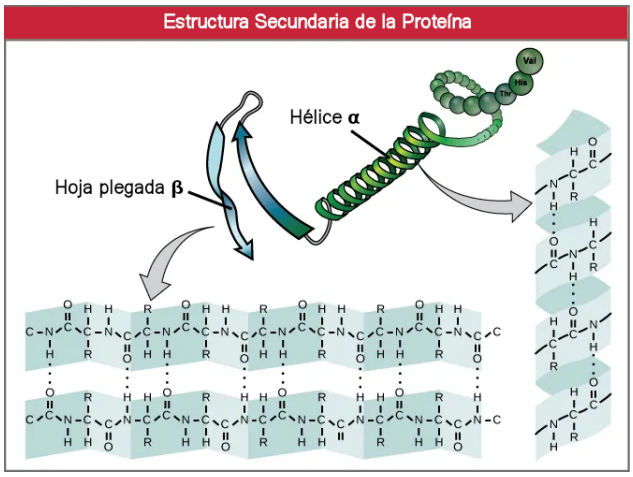


Figura 1.4: Dos conformaciones más frecuentes de la estructura secundaria de las proteínas. Los puntos sucesivos representan los puentes de hidrógeno, y las líneas continuas son los enlaces covalentes. Las cadenas laterales se muestran con la letra R. Imagen tomada de [11].

La unión de las conformaciones en la estructura secundaria, hélice y hoja, en una sola unidad corresponde al siguiente nivel organizacional conocido como estructura terciaria. Este nivel corresponde a la estructura biológicamente funcional, o conformación nativa, de la proteínas NSP3, NSP5 y Spike. Hasta este punto, la cadena de aminoácidos está completamente plegada donde los aminoácidos hidrofóbicos dominan en el interior de la proteína fuera del ambiente polar debido a que es el estado más estable termodinámicamente. Sin embargo, la cadena principal compuesta por el grupo CO y el grupo NH, unidos por el carbono principal C_α, es de carácter polar. Por lo tanto, son propensos a estar en un ambiente polar como el agua. Es en este momento, en que en la cadena principal se crean los puentes de hidrógeno dando lugar a las primeras estructuras, la hélice α y la hoja plegada β. Podemos ver que hay muchas hélices y hojas que tienen una cara hidrofóbica apuntando hacia el interior, y una cara polar apuntando hacia el solvente. Estas estructuras son llamadas anfipáticas [17]. De esta forma, podemos presenciar que a partir de un núcleo representado por los aminoácidos hidrofóbicos, dirigen la cadena principal a una conformación estable y funcional de la proteína [13, 17]. Sin embargo, el plegamiento incorrecto de las proteínas puede conducir a alguna alteración en el funcionamiento celular grave ya sea por el ambiente externo o por mutación genética. Las proteínas mal

plegadas pueden formar agregados que conllevarían a enfermedades patológicas como la enfermedad de Alzheimer, la enfermedad de Parkinson, la enfermedad de Huntington y la diabetes tipo 2 [22], entre otros.

En una proteína estructuralmente funcional hay aminoácidos que por su posición en la superficie globular se unen a un tipo de molécula, como llamado sustrato, y hay aminoácidos que catalizan esta misma molécula, es decir, la transforma [23]. Estas dos regiones en la proteína se les conoce como *sitios activos* (Figura 1.5). En el caso de nsp5 de SARS-CoV-2 tenemos una diada catalítica Cys145-His41 que rompen los enlaces peptídicos de la poliproteína 1ab para liberar las proteínas nsp4-nsp16 que permitirá al ARN viral replicarse [10]. Desactivar los sitios activos en la enzima conllevaría a desactivar o disminuir la actividad catalítica. Es por esto que la proteína nsp5 es un objetivo atractivo en el desarrollo de fármacos.

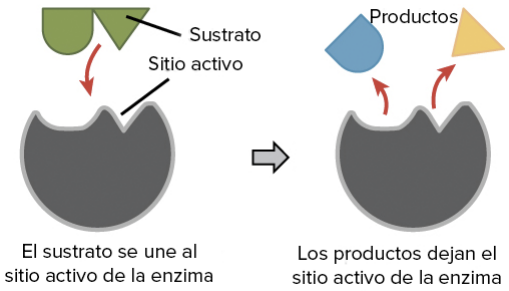


Figura 1.5: En el sitio activo hay regiones en la superficie de la enzima que permiten unirse al sustrato, y hay regiones catalíticas donde transforman el sustrato dando lugar a productos. Imagen tomada de [11]

Hay regiones, o cavidades, constituidos por sitios activos que se les conoce como *bolsillo de anclaje* (en inglés binding pocket), los cuales forman las condiciones adecuadas para unirse al sustrato [24]. Por ejemplo, la proteína 7LG7 posee un bolsillo de anclaje crítico conocido como distal de ribosa y adenosina, un objetivo en el diseño de fármacos [25]. La proteína Spike posee 3 bolsillos de anclaje críticos que tienen una interacción alta en la fusión con la membrana celular huésped [21].

1.3. Estructura macromolecular.

Hoy en día para saber la estructura de una proteína se usan una variedad de técnicas de laboratorio como la cristalografía de rayos X, la microscopía crioelectrónica y la resonancia magnética nuclear, pero se usa mayormente la primera. Este proceso consiste en emitir rayos X sobre estructuras cristalinas donde están alineados cada uno de los componentes moleculares de la proteína, y al ser difractados por la misma crean un patrón por el que se obtiene la posición y el tipo de átomo en la proteína. Por otro lado, se han

desarrollado técnicas computacionales basados en la inteligencia artificial para predecir la estructura de una proteína con precisión atómica a partir de la secuencia de aminoácidos como el proyecto AlphaFold2 de DeepMind que obtuvo una puntuación por arriba de los 90 en la prueba de distancia global (en inglés. *global distance test* o GDT) en dos tercios de las proteínas en las que no se sabía la estructura [26]

En 1971 se creó el Banco de Datos de Proteínas (en inglés Protein Data Bank bajo las siglas PDB) por los Laboratorios Nacionales de Brookhaven para almacenar información acerca de las estructuras cristalinas de macromoléculas biológicas [27]. En 1998, el Colaboratorio de Investigación en Bioinformática Estructural (en inglés, Research Collaboratory for Structural Bioinformatics o RCSB) pasó a gestionar los archivos pdb a la cual es posible acceder a su página de web (https://www.rcsb.org/) y permitir su descarga libre. En este trabajo usaremos las coordenadas de carbonos C_α y C_β contenidas en el archivo pdb. Las proteínas no estructurales 3 y 5 que trabajaremos corresponden al PDB 7LG7 y 7C6S, respectivamente. Para la proteína Spike trabajaremos con dos archivos PDB que son 6VYB y 6VXX debido a que hay dos conformaciones estructurales conocidos, respectivamente, como estado abierto y estado cerrado. Estos estados conformacionales corresponden a una sección llamada dominio de unión al receptor (en inglés, receptor-binding domain o por sus siglas, RBD). La proteína Spike está conformada por tres cadenas polipetídicas idénticas identificadas como cadena A, B y C, formando un trímero como estructura general. El RBD corresponde a los residuos 319-541 de la cadena B, véase la Figura 1.7 (c) y (d).

Los archivos PDB de las proteínas objetivo del repositorio de RCSB PDB se obtuvieron por el método cristalográfico de rayos X. Sin embargo, existen regiones flexibles o bucles que aún se mueven, lo que imposibilita la obtención de las coordenadas de los átomos involucrados [28]. Estas regiones, mayormente en proteínas grandes como la proteína Spike, pueden presentarse a través de la secuencia PDB y, frecuente, en los extremos [28]. Por lo tanto, las coordenadas de los átomos de las regiones dinámicas de la proteína no son registrados en el archivo PDB. En la proteína 7LG7 las coordenadas de dos aminoácidos, ubicados al final de la secuencia, no son registrados; para la proteína 7C6S son 5 aminoácidos no registrados, también ubicados al final de la secuencia; para las proteínas 6VYB y 6VXX, el 25.2 % y 24.1 % de los aminoácidos, distribuidos tanto en los extremos como a través de la secuencia, no fueron registrados.

En los cristales, los arreglos de las proteínas también pueden presentarse algunas diferencias debido a la dinámica de algunas cadenas laterales o sustratos [28]. Las coordenadas de estas regiones pueden registrarse y, además, mostrar la fracción de proteínas que tienen dicha conformación, conocido como ocupación (en inglés, *occupancy*). La décima columna del registro ATOM, Figura 1.6 (b), muestra la ocupación de cada átomo. Si la ocupación

```
HEADER    VIRAL PROTEIN                           25-FEB-20   6VYB
TITLE     SARS-COV-2 SPIKE ECTODOMAIN STRUCTURE (OPEN STATE)
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: SPIKE GLYCOPROTEIN;
COMPND   3 CHAIN: A, B, C;
COMPND   4 FRAGMENT: ECTODOMAIN;
COMPND   5 SYNONYM: S GLYCOPROTEIN,E2,PEPLOMER PROTEIN;
COMPND   6 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: SEVERE ACUTE RESPIRATORY SYNDROME CORONAVIRUS
SOURCE   3 2;
SOURCE   4 ORGANISM_COMMON: 2019-NCOV;
SOURCE   5 ORGANISM_TAXID: 2697049;
SOURCE   6 GENE: S, 2;
SOURCE   7 EXPRESSION_SYSTEM: HOMO SAPIENS;
SOURCE   8 EXPRESSION_SYSTEM_COMMON: HUMAN;
SOURCE   9 EXPRESSION_SYSTEM_TAXID: 9606
KEYWDS    CORONAVIRUS, SARS-COV-2, SARS-COV, SPIKE GLYCOPROTEIN, FUSION
KEYWDS   2 PROTEIN, STRUCTURAL GENOMICS, SEATTLE STRUCTURAL GENOMICS CENTER FOR
KEYWDS   3 INFECTIOUS DISEASE, SSGCID, VIRAL PROTEIN
EXPDTA    ELECTRON MICROSCOPY
AUTHOR    A.C.WALLS,Y.J.PARK,M.A.TORTORICI,A.WALL,SEATTLE STRUCTURAL GENOMICS
AUTHOR   2 CENTER FOR INFECTIOUS DISEASE (SSGCID),A.T.MCGUIRE,D.VEESLER
REVDAT   6   29-JUL-20 6VYB    1       COMPND REMARK HETNAM LINK
REVDAT   6 2                    1       SITE   ATOM
REVDAT   5   06-MAY-20 6VYB    1       COMPND SOURCE DBREF  SEQADV
REVDAT   4   29-APR-20 6VYB    1       JRNL
REVDAT   3   01-APR-20 6VYB    1       COMPND
REVDAT   2   25-MAR-20 6VYB    1       JRNL
REVDAT   1   11-MAR-20 6VYB    0
JRNL        AUTH   A.C.WALLS,Y.J.PARK,M.A.TORTORICI,A.WALL,A.T.MCGUIRE,
JRNL        AUTH 2 D.VEESLER
JRNL        TITL   STRUCTURE, FUNCTION, AND ANTIGENICITY OF THE SARS-COV-2
JRNL        TITL 2 SPIKE GLYCOPROTEIN.
JRNL        REF    CELL                          V. 181   281 2020
JRNL        REFN                   ISSN 1097-4172
JRNL        PMID   32155444
JRNL        DOI    10.1016/J.CELL.2020.02.058
REMARK   2
REMARK   2 RESOLUTION.    3.20 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   SOFTWARE PACKAGES      : LEGINON, RELION, RELION
```

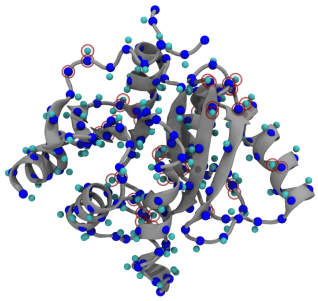
(a) El archivo muestra especificaciones generales.

```
ATOM    511  CA  ILE A 101     181.248 259.801 231.770  1.00 60.67           C
ATOM    512  C   ILE A 101     182.574 259.816 232.497  1.00 61.11           C
ATOM    513  O   ILE A 101     183.606 260.171 231.925  1.00 59.46           O
ATOM    514  CB  ILE A 101     181.243 258.690 230.694  1.00 60.30           C
ATOM    515  CG1 ILE A 101     179.865 258.623 230.027  1.00 60.24           C
ATOM    516  CG2 ILE A 101     181.607 257.335 231.338  1.00 58.49           C
ATOM    517  CD1 ILE A 101     179.818 257.837 228.754  1.00 57.04           C
ATOM    518  N   ARG A 102     182.547 259.457 233.771  1.00 59.93           N
ATOM    519  CA  ARG A 102     183.754 259.530 234.568  1.00 60.34           C
ATOM    520  C   ARG A 102     184.209 258.192 235.108  1.00 61.49           C
ATOM    521  O   ARG A 102     185.377 258.049 235.468  1.00 63.29           O
ATOM    522  CB  ARG A 102     183.561 260.487 235.722  1.00 62.53           C
ATOM    523  CG  ARG A 102     183.282 261.907 235.310  1.00 61.42           C
ATOM    524  CD  ARG A 102     183.041 262.786 236.459  1.00 62.21           C
ATOM    525  NE  ARG A 102     182.660 264.112 236.026  1.00 63.43           N
ATOM    526  CZ  ARG A 102     182.378 265.146 236.843  1.00 65.71           C
ATOM    527  NH1 ARG A 102     182.447 265.002 238.149  1.00 66.66           N
ATOM    528  NH2 ARG A 102     182.031 266.311 236.329  1.00 66.94           N
ATOM    529  N   GLY A 103     183.324 257.208 235.199  1.00 59.90           N
ATOM    530  CA  GLY A 103     183.784 255.995 235.854  1.00 57.79           C
ATOM    531  C   GLY A 103     182.949 254.745 235.672  1.00 59.01           C
ATOM    532  O   GLY A 103     181.989 254.712 234.903  1.00 56.60           O
ATOM    533  N   TRP A 104     183.384 253.693 236.364  1.00 57.26           N
ATOM    534  CA  TRP A 104     182.761 252.375 236.317  1.00 59.15           C
ATOM    535  C   TRP A 104     182.706 251.693 237.677  1.00 60.46           C
ATOM    536  O   TRP A 104     183.611 251.837 238.501  1.00 61.21           O
ATOM    537  CB  TRP A 104     183.530 251.439 235.384  1.00 57.94           C
ATOM    538  CG  TRP A 104     183.605 251.878 233.985  1.00 56.84           C
ATOM    539  CD1 TRP A 104     182.784 251.510 232.975  1.00 56.79           C
ATOM    540  CD2 TRP A 104     184.569 252.780 233.408  1.00 57.74           C
ATOM    541  NE1 TRP A 104     183.161 252.125 231.819  1.00 55.89           N
ATOM    542  CE2 TRP A 104     184.252 252.907 232.067  1.00 56.32           C
ATOM    543  CE3 TRP A 104     185.656 253.481 233.919  1.00 58.13           C
ATOM    544  CZ2 TRP A 104     184.981 253.713 231.217  1.00 54.96           C
ATOM    545  CZ3 TRP A 104     186.385 254.290 233.069  1.00 57.75           C
ATOM    546  CH2 TRP A 104     186.053 254.403 231.755  1.00 54.43           C
ATOM    547  N   ILE A 105     181.684 250.873 237.868  1.00 59.80           N
ATOM    548  CA  ILE A 105     181.587 250.004 239.032  1.00 59.24           C
ATOM    549  C   ILE A 105     181.562 248.562 238.570  1.00 60.60           C
ATOM    550  O   ILE A 105     180.779 248.214 237.690  1.00 53.08           O
ATOM    551  CB  ILE A 105     180.306 250.243 239.850  1.00 60.71           C
ATOM    552  CG1 ILE A 105     180.225 251.670 240.329  1.00 61.04           C
ATOM    553  CG2 ILE A 105     180.310 249.296 241.052  1.00 61.98           C
ATOM    554  CD1 ILE A 105     178.871 252.042 240.910  1.00 63.11           C
```

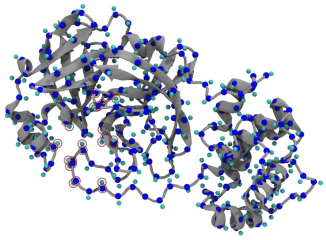
(b) Las coordenadas de cada átomo se muestran entre las 7-9 columnas.

Figura 1.6: En (a), el archivo pdb muestra especificaciones básicas como la especie, género, cadenas; autores del experimento, el método de descripción estructural [27]. En (b) muestra las coordenadas de cada átomo, así como también el número de secuencia, aminoácido a la que pertenece, el tipo de átomo: si es C_α o C_β y se muestra como CA o CB, respectivamente (tercera columna).

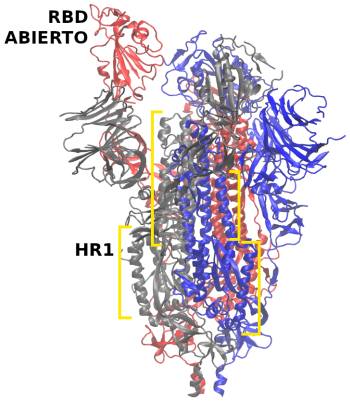
de un átomo es 1, quiere decir que en todas las proteínas se mostró que tiene la misma posición. Por lo general, el valor es 1 para todos los átomos. En cambio, si hay dos o más registros para un mismo átomo, entonces, la ocupación varía de acuerdo a la fracción de proteínas con dicha conformación, lo que sumado es 1. Por ejemplo, si tenemos dos registros para un átomo en común A y B, es decir, con dos coordenadas distintas para un mismo átomo. Además, si la ocupación es de 0.5 para cada uno, quiere decir que la mitad de las proteínas en los cristales tiene la conformación A y la otra mitad tiene la conformación B. En las proteínas 7LG7 y 7C6S podemos encontrar algunas cadenas laterales con distintas conformaciones. Escogeremos el primero de cada uno ya que la posición varía ligeramente para ambas proteínas objetivo.



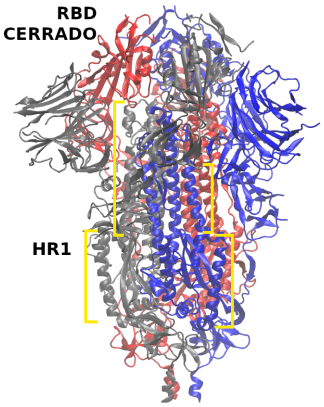
(a) Proteína NSP3 (PDB 7LG7).



(b) Proteína NSP5 (PDB 7C6S).



(c) Proteína S (PDB 6VYB, estado abierto).



(d) Proteína S (PDB 6VYB, estado cerrado).

Figura 1.7: Representación de las proteínas 7LG7, 7C6S, 6VYB y 6VXX. En la estructura proteica, las esferas de color azul fuerte se presentan los carbonos C_α, y en azul claro son los carbonos C_β. En (c) y (d) muestra la proteína S en estado abierto y estado cerrado, respectivamente. En color azul, rojo y gris corresponden a las cadenas A, B y C, respectivamente.

Capítulo 2

Teoría de grafos.

Este capítulo está dedicado a establecer el contexto de la dinámica de este trabajo. La primera sección, la dedicaremos al contexto histórico que llevó al interés por estudiar y crear modelos a partir de estas estructuras. La segunda sección sobre fundamentos de teoría de grafos, está dedicada a establecer la base de este trabajo. En esta parte, la última sección, llamada grafos aleatorios (en inglés random graphs), está dedicada a tres tipos importantes de grafos aleatorios que marcaron un antes y después en la teoría de grafos.

2.1. Contexto histórico de la teoría de grafos.

En la antigua ciudad de Königsberg (hoy en día Kaliningrado perteneciente a Rusia) existía un famoso problema referido a los siete puentes que conectaban la isla llamada Kneiphof formada por el cruce del río Pregolya, Figura (2.1). Ante la intersección del río, el problema consistía en atravesar los puentes solo una vez.

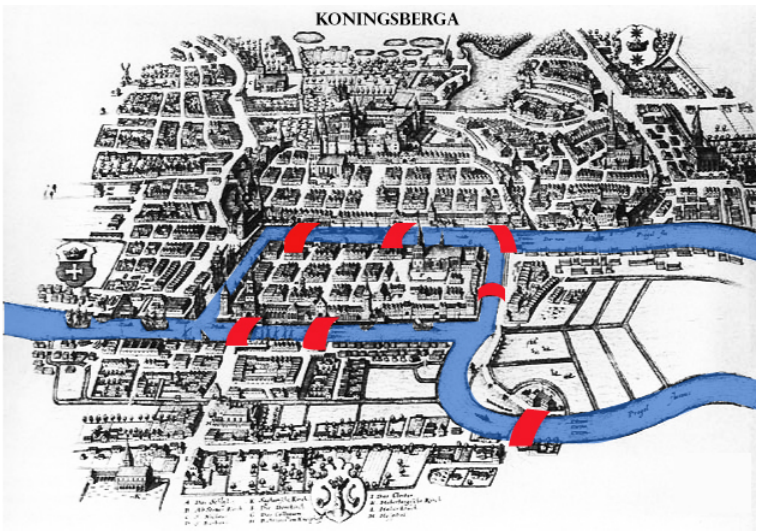
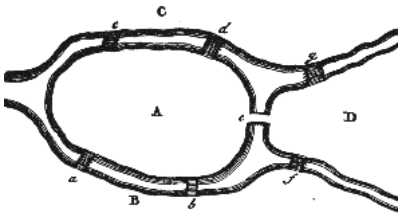


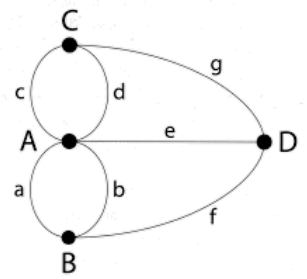
Figura 2.1: Mapa de la antigua ciudad de Königsberg. Marcados en rojo, la ubicación de los puentes que conectaban la ciudad. Imagen [29].

En 1736 salió a la luz un artículo escrito por el matemático Leonard Euler, en el que establece no solo una solución a este enigma, sino también en general a cualquier número de puentes, número de áreas de tierra y arreglos que se presenten.

Euler establece dos conjuntos: uno en letras en mayúsculas para representar las diferentes áreas de tierra, $N = \{A, B, C, D\}$; y otro conjunto con letras en minúsculas que representaría los puentes, $L = \{a, b, c, d, e, f, g\}$, Figura 2.2 (a). El par de conjuntos pasaría a conformar un grafo en términos modernos y pasaría a ser la base del método de Euler en la cual resolvería el enigma [30]. En consecuencia, el conjunto de áreas de tierra puede verse en una representación gráfica más simple por nodos y los puentes por líneas que inciden de ellos, Figura 2.2 (b).



(a) Diagrama de Euler.



(b) Diagrama de Euler.

Figura 2.2: En (a), el diagrama de Euler en su artículo [27]. En (b) es el grafo con los componentes elementales.

Una caminata del área A al C pasando por D, Euler lo representaría como la secuencia $AeDgC$, donde el número de puentes es el número de áreas de tierra menos uno. Entonces, la secuencia que representaría el cruce de los 7 puentes de Königsberg solo una vez tendría 8 letras mayúsculas.

Si el área de tierra A tiene un número impar de puentes, entonces A debería aparecer en la secuencia el número de puentes más uno dividido por 2. Por ejemplo, supongamos que sólo existiera las áreas A y B, conectados por 3 puentes. Entonces, el número de veces que aparecería A en la secuencia sería dos veces sin importar el punto de partida, es decir,

$$A \quad a \quad B \quad b \quad A \quad c \quad B$$

Si son 5 puentes, entonces

$$B \quad a \quad A \quad b \quad B \quad c \quad A \quad d \quad B \quad e \quad A$$

el área A aparecería 3 veces. En el problema de los 7 puentes de Königsberg, el área A tendría que parecer 3 veces y los demás 2 veces. Entonces, si sumamos, la secuencia ten-

dría que tener 9 letras mayúsculas en total, lo cual es una contradicción. Es aquí donde Euler prueba que no es posible cruzar los siete puentes una sola vez.

La resolución de los puentes de Königsberg por Euler dio inicio a una nueva rama matemática. Aunque primeramente conocida como geometría de posición, los sucesivos trabajos de investigadores como Kirchhoff en circuitos eléctricos, Cayley en isómeros químicos, Hamilton en circuitos que hoy en día lleva su nombre, Erdős y Rényi en grafos aleatorios, etc. pasarían a culminar lo que hoy en día llamamos Teoría de Grafos.

En 1864 el químico Alexander Brown propuso una representación molecular de sustancias químicas que hoy en día usamos, véase la Figura (2.3), y es muy similar a la representación de un grafo. Cada átomo de una molécula está representado por su nombre simbólico encerrado en un círculo, y cada enlace entre ellos está representado por una línea. Entre varias propuestas de la época como el de Couper (1958), Loschmidt (1861) y Kekulé (1861), la representación de Alexander destacó debido a que pudo explicar una propiedad candente de la época de sustancias como la isomería [30]: sustancias químicas que poseen la misma fórmula química pero con propiedades físicas distintas debido al orden en el que están dispuestos los átomos en la estructura molecular. Este es un claro ejemplo de que la sola representación gráfica del objeto de estudio puede tener un alto grado de importancia en la investigación y llevar a descubrir ciertas propiedades que de otra forma sería intrincado o imposible de ver. Cabe mencionar que Alexander Brown se refirió a su propuesta como "notación gráfica"que posteriormente inspiraría la terminología de grafo por Sylvester [30].

Diez años después, la representación molecular de Alexander y el fenómeno de la isomería se habían extendido hasta llegar al matemático Artur Cayley quien decidió aplicar sus recientes investigaciones en árboles en teoría de grafos y descubrir ciertas pautas que le permitió enumerar los isómeros.

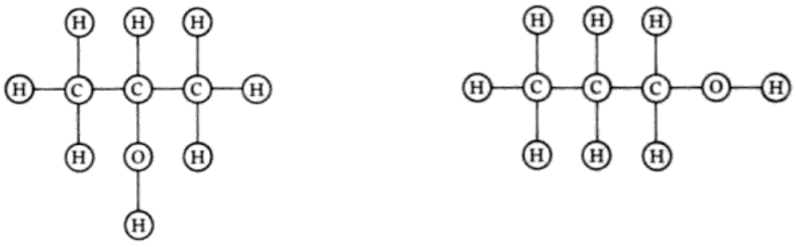


Figura 2.3: Representación bajo la notación de Alexander Brown. El Propan-1-ol y el Propan-2-ol son dos sustancias con propiedades físicas diferentes pero con la misma fórmula química C_3H_8O. Imagen [30].

En esta investigación, los nodos serán los carbonos C_α y C_β de los aminoácidos constituyentes de las proteínas globulares del SARS-CoV-2. Las líneas que uniría a un carbono

con un segundo será si éste está dentro de un radio R en el espacio tridimensional del primero, es decir, si son adyacentes el uno al otro en un radio R. A este radio le llamaremos *radio de interacción*.

2.2. Fundamentos de teoría de grafos

Un grafo es un par ordenado $G(\mathcal{N}, \mathcal{L})$, donde $\mathcal{N} = \{n_1, n_2, \cdots, n_m\}$ es un conjunto no vacío de m nodos, y $\mathcal{L}$ es un subconjunto de pares no ordenados de nodos distintos, es decir, $\mathcal{L} \subseteq \{\{n_i, n_j\} | n_i, n_j \in N \wedge i \neq j\}$. A lo largo de este trabajo también nos referiremos a esta estructura de grafo con sus componentes interconectados como una red. Una línea, o arista, $l \in \mathcal{L}$ es un par no ordenado de nodos $n_i, n_j \in N$ siempre que $i \neq j$. Si $\mathcal{L} = \emptyset$, entonces G es llamado *grafo vacío*. Si $l \in \mathcal{L}$ y $n_i, n_j \in l$ tal que $i \neq j$, se dice que n_i y n_j son *incidentes* a l, y viceversamente, l es incidente a los dos nodos. Además, dado que $n_i, n_j \in l = \{n_i, n_j\}$, el par de nodos n_i y n_j son llamados *adyacentes*. Si dos líneas tienen en común un nodo, ambas son líneas adyacentes.

Dado un grafo G, una secuencia de nodos y líneas $(n_0, l_0, n_1, l_1, \cdots, l_{m-1}, n_m)$ o simplemente $(n_0, n_1, \cdots, n_m)$ es llamada *caminata*, no importa si se repiten los nodos o líneas. Por el contrario, si una secuencia no tiene elementos repetidos, entonces es una *ruta*. Si en la secuencia donde $n_0 = n_m$ y los demás nodos son distintos, entonces es un ciclo. Además, si cada par de nodos del grafo está conectado por al menos una ruta, entonces, G es un *grafo conectado*. Si cada par de nodos es adyacente, entonces, G es un *grafo completo* [31].

Dada una caminada $(n_0, l_0, n_1, l_1, \cdots, l_{m-1}, n_m)$, la longitud de n_0 a n_m será igual al número de líneas en la secuencia, es decir, m. La *distancia geodésica* $d(n_i, n_j)$, o simplemente *distancia*, de dos nodos n_i y n_j es la longitud de la caminata más corta que conecta a ambos, por lo que también $d(n_i, n_j) = d(n_j, n_i)$. Para cualquier nodo $n_i \in N$, $d(n_i, n_i) = 0$. Si dos nodos no están conectados por alguna caminata, entonces, $d(n_0, n_m) = \infty$. Si G es completo, la función d es una métrica [31].

Hay varias formas de representar un grafo además de la notación gráfica: ya sea por listas, tuplas o forma matricial. El que usaremos será esta última llamada *matriz de adyacencia* $A = (A_{ij})$ donde

$$A_{ij} = \begin{cases} 1, & \text{si los nodos } n_i \text{ y } n_j \text{ son adyacentes} \\ 0, & \text{de otra forma} \end{cases} \tag{2.1}$$

Si sumamos la fila entera i, tenemos el número de líneas incidentes del nodo n_i, el cual hablaremos más en la siguiente sección. Debido a que cada línea es un conjunto no ordenado de nodos distintos, la diagonal principal de la matriz es cero.

2.2.1. Grado de un nodo

Dado un grafo G con n nodos y m líneas. El *grado* de un nodo n_i, denotado por $d(n_i)$, es el número de líneas incidentes de n_i. En términos de la matriz de adyacencia,

$$d(n_i) = \sum_{j=1}^{m} A_{ij} \qquad 0 \leq d(n_i) \leq n-1 \tag{2.2}$$

El primer teorema considerado de la teoría de grafos, relacionado al grado, es debido a una de las observaciones de Euler en su artículo de los puentes de Königsberg [30, 31]. Dado que toda línea es un conjunto de par de nodos, la suma de los grados es también un número par ya que la línea que incide de sus dos puntos se toma dos veces en la suma. Por lo tanto, la suma de los grados de todos los nodos en un grafo es par.

Teorema 2.2.1 ([31]). *La suma de los grados de los nodos de un grafo G es dos veces el número de líneas,*

$$\sum_{i=1}^{n} d(n_i) = 2m \tag{2.3}$$

En un grafo con n nodos, el grado promedio esta dado por

$$k = \frac{\sum_{i=1}^{m} d(n_i)}{n} = \frac{2m}{n} \tag{2.4}$$

Otra de las observaciones de Euler como consecuencia del teorema fue que para que la suma de grados sea par, aquellos nodos con grado impar deberían ser pares.

Un grafo G que tenga todos sus nodos con el mismo grado k es un grafo regular también llamado grafo k-regular. Un grafo 0-regular es también un grafo trivial con todos sus *nodos aislados*; si es 1-regular cada nodo incide una línea, también llamado *nodos terminales*; si es 2-regular tenemos un ciclo o una unión disjunta de ciclos; si es 3-regular es un grafo cúbico.

A partir de conocer los grados de cada nodo, es de interés saber qué tipo de distribución de grado tiene la red. Las redes del mundo real como en la World Wide Web y la cantidad de visitas en los sitios de web presentan una distribución llamada ley de potencia [32, 33] (escrito en inglés como Power Law), mientras hay otros tipos de redes que veremos más adelante que tienen una distribución de Poisson del grado.

2.2.2. Coeficiente de Agrupamiento

El coeficiente de agrupamiento es la probabilidad de que dado dos nodos que tienen en común a un nodo adyacente sean a la vez adyacentes. Hay varias formas de formular esta

idea, pero usaremos la idea de transitividad que resulta ser más óptimo y preciso [33]. Una relación $\circ$ es transitiva si $a \circ b$ y $b \circ c$, entonces, $a \circ c$ [33].

Sean $n_1, n_2, n_3 \in \mathcal{N}(G)$ tres nodos de una red donde n_1 y n_3 son adyacentes a un nodo en común n_2. Si n_1 y n_3 son también adyacentes, entonces, la ruta $n_1 n_2 n_3$ pasa a llamarse un triángulo o ruta cerrada; de otra forma es una triada de nodos conectados. Definimos el coeficiente de agrupamiento como

$$C = \frac{3\,(\text{el número de triángulos en la red})}{(\text{el número de triadas de nodos conectados})} \tag{2.5}$$

El 3 en el numerador es debido a que el número de triángulos en el denominador es contado tres veces.

El coeficiente de agrupamiento de una red dada G toma el valor máximo, $C = 1$, si G es un grafo completo donde cada par de nodos es adyacente. Por otro lado, si el coeficiente toma el valor mínimo $C = 0$, entonces, no existen triángulos en el conjunto de aristas $\mathcal{L}$.

El coeficiente de agrupamiento es otra de las propiedades de interés en el análisis de una red y marca una diferencia entre las redes basadas del mundo real con los grafos teóricos [33].

2.2.3. Grafos Aleatorios (random graphs)

A partir de 1959, Erdős y Rényi marcaron un antes y después en la teoría de grafos con la introducción de la teoría de la probabilidad al demostrar que ciertas propiedades ya conocidas son inherentes a los grafos y, más aún, permitiendo descubrir otras importantes [34].

El enfoque detrás del tratamiento de las propiedades de grafos es a partir de la evolución de espacios de probabilidad [34]. Erdős y Rényi definieron el primer modelo donde el espacio de probabilidad $\mathscr{G}(n, m)$ está constituido por el conjunto de todos los grafos $G(\mathcal{N}, \mathcal{L})$ con n nodos y m aristas, y una medida de probabilidad $\mathbb{P}$ que asigna el mismo valor a cada grafo. De esta forma, dado que hay $N = \binom{n}{2}$ aristas posibles en un grafo con n nodos, y que los grafos tienen exactamente m aristas, entonces hay en total $\binom{N}{m}$ grafos en el conjunto de todos los grafos del espacio de probabilidad $\mathscr{G}(n, m)$. Entonces, la probabilidad de que el grafo aleatorio $\mathbb{G}_{n,m}$ tome un grafo en particular G es

$$\mathbb{P}(\mathbb{G}_{n,m} = G) = \frac{1}{\binom{N}{m}} \tag{2.6}$$

En el mismo año que Erdős y Rényi, el matemático Edgar Gilbert propone su modelo de

grafo aleatorio $\mathscr{G}\{n, P(\text{línea}) = p\}$ [35]. Dado $0 \leq p \leq 1$, el modelo constituye el conjunto de todos los grafos con n nodos, donde cada par de nodos es enlazado independientemente con una probabilidad p, y una medida de probabilidad tal que para cada grafo G

$$\mathbb{P}(\mathbb{G}_{n,p} = G) = p^m(p-1)^{N-m} \tag{2.7}$$

y la probabilidad de elegir un grafo con exactamente m aristas está dado por

$$\mathbb{P}(|\mathcal{L}_{n,p}| = m) = \binom{N}{m} p^m(p-1)^{N-m} \tag{2.8}$$

Es decir, $(\mathcal{L}_{n,p})$ sigue una distribución binomial con parámetros (N, p), por lo cual el número de aristas promedio está dado por

$$\langle m \rangle = \sum_{m=0}^{N} m\mathbb{P}(|\mathcal{L}_{n,p}| = m) = \binom{n}{2} p \tag{2.9}$$

Por otro lado, el grado promedio de un grafo aleatorio está dado por

$$\begin{aligned} \langle k \rangle &= \sum_{m=0}^{N} k\mathbb{P}(|\mathcal{L}_{n,p}| = m) = \sum_{m=0}^{N} \frac{2m}{n}\mathbb{P}(|\mathcal{L}_{n,p}| = m) \\ &= \frac{2}{n}\binom{n}{2} p = (n-1)p \end{aligned} \tag{2.10}$$

donde $k = 2m/n$ por (2.4). La probabilidad p que tiene el nodo de unirse a otro, multiplicado por el resto de los $n-1$ nodos.

En el modelo de Gilbert, si elegimos un nodo la probabilidad de conectarse a cualquiera de los otros $n-1$ nodos es p. Entonces, la probabilidad de que tenga grado k es

$$P(k) = \binom{n-1}{k} p^k(1-p)^{n-1-k} \tag{2.11}$$

donde $\binom{n-1}{k}$ es el número de maneras de elegir los k nodos adyacentes con probabilidad $p^k(1-p)^{n-1-k}$. Por lo tanto, tenemos una distribución de grado binomial. Esta es una propiedad que nos interesaría ver qué sucede cuando el número de nodos se vuelve cada vez más grande, $n \to \infty$. Entonces, si aplicamos logaritmo tenemos

$$\begin{aligned} \ln[P(k)] &= \ln\left[\binom{n-1}{k} p^k(1-p)^{n-1-k}\right] \\ &= \ln\left[\binom{n-1}{k}\right] + \ln p^k + \ln\left[(1-p)^{n-1-k}\right] \end{aligned}$$

Si usamos 2.10 para sustituir p tenemos

$$\ln[P(k)] = \ln\left[\binom{n-1}{k}\right] + \ln\left[\left(\frac{c}{n-1}\right)^k\right] + (n-1-k)\ln\left[1-\frac{c}{n-1}\right]$$

El tercer logaritmo lo expresamos en series de Taylor. Entonces, resulta

$$\ln\left[1-\frac{c}{n-1}\right] = -\frac{c}{n-1} - \frac{1}{2}\left(-\frac{c}{n-1}\right)^2 + \cdots + \frac{(-1)^{2i-1}}{i}\left(-\frac{c}{n-1}\right)^i + \cdots$$

Cuando $n \to \infty$, tenemos que

$$(n-1-k)\ln\left[1-\frac{c}{n-1}\right] \simeq -(n-1-k)\frac{c}{n-1} \simeq -c,$$

lo cual es asintóticamente igual a -c ya que

$$\lim_{n\to\infty} \frac{(n-1-k)\ln\left[1-\frac{c}{n-1}\right]}{-(n-1-k)\frac{c}{n-1}} = 1$$

y

$$\lim_{n\to\infty} \frac{-(n-1-k)\frac{c}{n-1}}{-c} = 1$$

A continuación, aplicamos exponencial y, de esta forma, tenemos que

$$(1-p)^{n-1-k} = e^{-c} \tag{2.12}$$

Por otra parte,

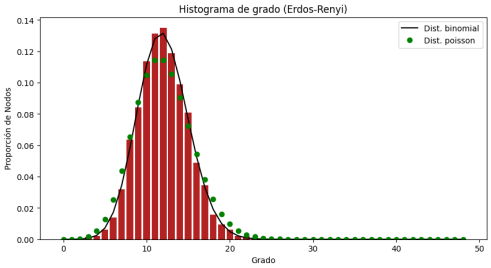
$$\binom{n-1}{k} = \frac{(n-1)!}{(n-1-k)!k!} \simeq \frac{(n-1)^k}{k!} \tag{2.13}$$

Por lo tanto, juntamos estos dos últimos resultados y, tenemos que para 2.11 con $n \to \infty$,

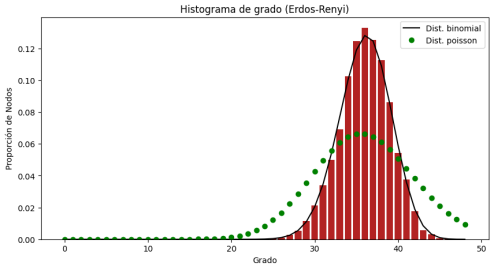
$$P(k) = \frac{(n-1)^k}{k!}\left(\frac{c}{n-1}\right)^k e^{-c} = e^{-c}\frac{c^k}{k!} \tag{2.14}$$

Tenemos que la distribución de grado de un grafo aleatorio es una distribución de Poisson con n grande. De esta forma, también son llamados grafos aleatorios de Poisson [33].

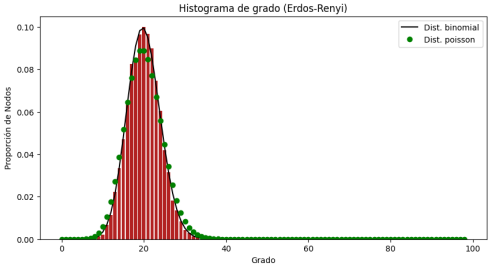
En la Figura 2.4 muestra la distribución de grado promedio de 100 grafos aleatorios de Erdő-Rényi y lo comparamos con la distribución binomial (línea negra) y la distribución de Poisson (círculos verdes). El parámetro de la distribución binomial, ecuación 2.11, es $p = m\binom{n}{2}$, el cual se obtiene a partir de 2.9.



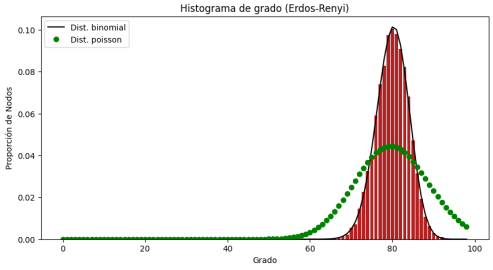
(a) 50 Nodos y 300 aristas



(b) 50 Nodos y 900 aristas



(c) 100 Nodos y 1000 aristas



(d) 100 Nodos y 4000 aristas

Figura 2.4: Distribución de grado de 100 grafos aleatorios de 50 y 100 nodos con 300 y 900 aristas. En la línea curvada representa la distribución binomial, y los puntos verdes la distribución de Poisson.

En (a) y (b), el número de nodos en ambos casos fue de 50 nodos y el número de líneas fue de 300 y 900, respectivamente. Podemos ver que la gráfica se recorre a la derecha si aumentamos el número de líneas en el grafo, y que la distribución de Poisson tiende a no coincidir con las demás distribuciones. En (c) y (d), el número de nodos fue de 100 para ambos casos con 1000 y 4000 lineas, respectivamente. Si aumentamos el número de nodos podemos observar que el número de barras también aumenta y se hallan mayormente dentro de la línea de la distribución binomial. Además, podemos observar que la distribución de grado de Erdő-Rényi sigue una distribución binomial al igual que el modelo de Gilbert.

2.2.4. Grafo de Mundo Pequeño (Small-world networks)

En 1998 Duncan Watts y Steven Strogatz escribieron un artículo llamado ”Collective dynamics of ‘small-world’ networks” en donde presentaban que diferentes sistemas en ciencias biológicas, sociales, informáticos, etc. poseían estructuras con características relativamente similares. En aquel entonces, se conocía que una de estas características es que dado dos nodos cualesquiera tienen en promedio caminos cortos, o distancias geodésicas, relativamente menores. El psicólogo Stanley Milgram los nombró redes de mundo pequeño (small-world networks) por su característica inherente, quien además hizo un experimento al mandar 96 paquetes por correo a su amigo por medio de distintas personas elegidas al azar en Kanzas y Nebraska que debían reenviar el paquete al destinatario o conocido que pudiera cumplir con cualquiera de los datos con el primer nombre, ocupa-

ción de corredor de bolsas y que vivía en la ciudad de Boston, Massachusetts. Al final, solo 18 paquetes llegaron al amigo de Milgram, el cual el número promedio de personas hasta llegar al destinatario objetivo fue de 5.9. Este estudio se popularizó en los Estados Unidos con la idea que para cualesquiera dos personas en el mundo hay solo seis pasos [33].

Duncan Watts y Steven Strogatz no solo definieron matemáticamente las redes de mundo pequeño si no que pudieron construirlos como un ente que está entre los grafos regulares y grafos aleatorios. Watts y Strogatz empezaron con un grafo regular formando un anillo de 60 nodos, cada nodo conectado a sus k vecinos más cercanos. El proceso consiste en desconectar una arista de un nodo y volverlo a conectar con probabilidad p con cualquier otro nodo, elegido de forma uniforme aleatoria. El proceso se repite hasta recorrer cada uno de los nodos en sentido de las manecillas del reloj. Por consiguiente, se repite la vuelta al anillo hasta que cada arista haya sido considerado. Si $p = 0$, el grafo regular no cambia, pero si $p = 1$ tenemos un grafo desordenado, vea la Figura (2.5). El resultado fue que con valores intermedios de p, el grafo adquiría un grado alto de agrupamiento y un promedio de distancias geodésicas cortas, es decir, lograron construir un grafo de mundo pequeño. El proceso tiene un comportamiento muy sensible en el promedio de las distancias geodésicas. En un grafo regular con 1000 nodos, el promedio de distancias geodésicas fue de 250. Al desconectar y conectar el 5 % de los nodos, el promedio cayó alrededor de 20 [32].

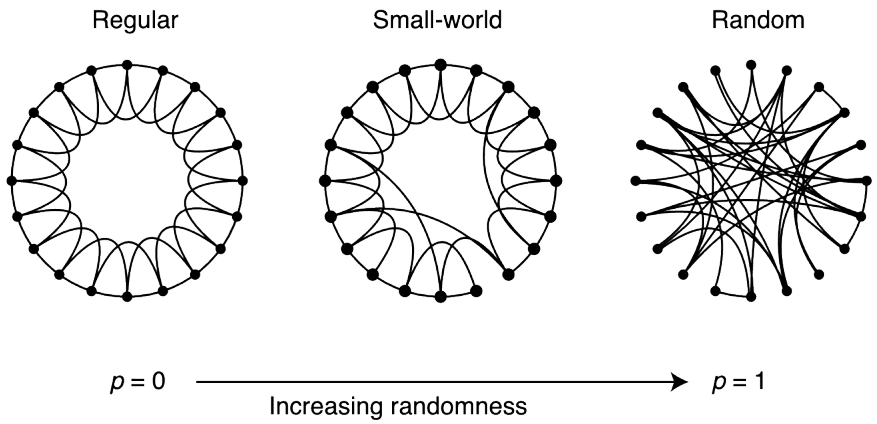


Figura 2.5: El primer grafo de la izquierda es regular. En un proceso de volver a conectar los enlaces con cualquier otro con probabilidad p podemos obtener un grafo aleatorio o un grafo de pequeño mundo. Imagen [36].

Este resultado logró producir un grafo de mundo pequeño cuyas propiedades son que el promedio de las distancias geodésicas son relativamente cortas y que el coeficiente de agrupamiento es grande. Sin embargo, no tiene mucho que decir sobre la distribución de grado [32].

Existen otras redes en la naturaleza que tienen la propiedad de mundo pequeño llamados redes de *libre escala* (free-scale networks en inglés) que tienen una distribución de grado singular conocido como *ley de potencia* (power-law en inglés). Además, poseen un coeficiente de agrupamiento alto [32]. Un ejemplo de ello es la red de hipervínculos entre páginas Web donde los nodos son las páginas web. El número de enlaces dirigidos a una página particular es su grado de entrada (en inglés es in-degree), mientras que los enlaces de esta página a otras es su grado de salida (en inglés es out-degree). En la Figura 2.6 muestra la distribución de grado de entrada y la distribución de grado de salida. Esta forma característica de la red muestra que muy pocos nodos poseen un grado bastante alto, mientras que la mayoría de los nodos tienen un grado muy bajo.

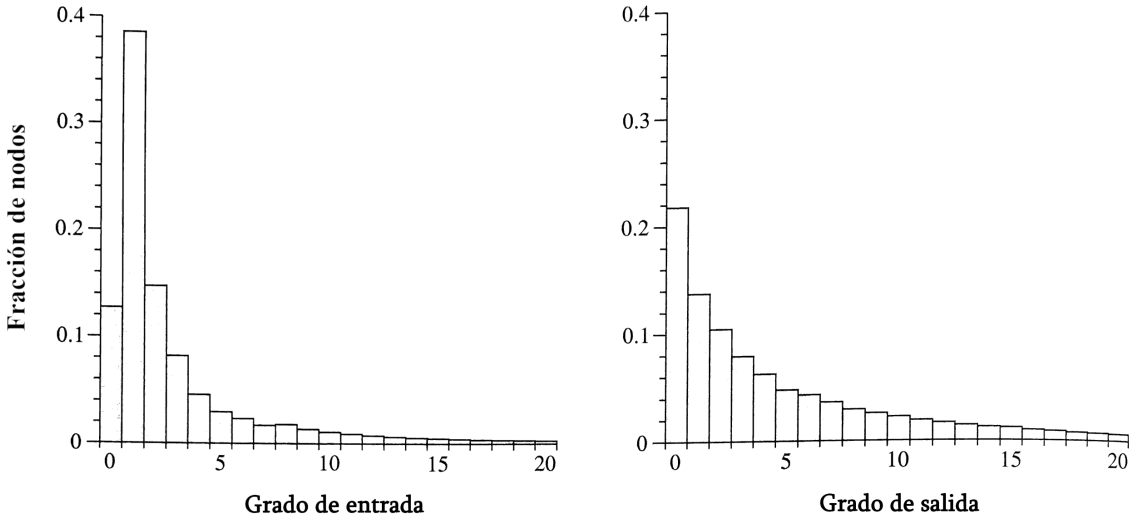


Figura 2.6: Distribución de grado de la Wold Wide Web. Imagen [33].

En 1999, los físicos Albert-László Barabási y Réka Albert publicaron un modelo para generar grafos aleatorios de mundo pequeño con una distribución de grado de Ley de Potencia. El modelo consiste en una serie de pasos comenzando con una red de n_0 nodos iniciales. En cada paso del tiempo, un nuevo nodo u con $l(l \leq n_0)$ aristas es agregado a la red. El nodo u es unido a un nodo existente v con una probabilidad proporcional al grado de éste, es decir,

$$P(v) = \frac{d(v)}{\sum_w d(w)} \tag{2.15}$$

donde $d(v)$ es el grado del nodo v, y el denominador es la suma de los grados de los nodos en la red actual. El proceso de enlazar un nodo a otro existente es conocido como *apego preferencial* (en inglés, preferential attachment).

2.3. Medidas de centralidad

2.3.1. Centralidad de vector propio

La centralidad de vector propio mide la importancia o influencia de un nodo en función de aquellos nodos adyacentes que son a la vez importantes en la red [33]. La medida de centralidad no solo puntúa el grado que tenga un nodo sino también la calidad de estas conexiones. Aunque dos nodos tengan igual número de conexiones en la red, no necesariamente tienen la misma centralidad. Por ejemplo, un nodo que tenga menor grado pero con conexiones a nodos con alta importancia en la red, puede tener mayor puntuación que un nodo con mayor grado pero con conexiones a nodos con baja importancia [33].

De esta forma, la centralidad del vector propio otorga a cada nodo una puntuación proporcional a la suma de las puntuaciones de los nodos adyacentes. Para reflejar esta idea, sea x_i el grado de importancia de un nodo i, entonces

$$x_i = k^{-1} \sum_{j \neq i} x_j \tag{2.16}$$

donde k^{-1} es la constante de proporcionalidad puesto de esta forma para propósito que veremos más adelante. Podemos usar la matriz de adyacencia de tal forma que el grado de importancia x_i para el nodo i es proporcional a la suma de las centralidades de los nodos adyacentes,

$$x_i = k^{-1} \sum_{j \neq i} a_{i,j} x_j \tag{2.17}$$

en notación matricial tenemos que $\mathbf{x} = k^{-1} A\mathbf{x}$, lo cual es lo mismo a

$$k\mathbf{x} = A\mathbf{x} \tag{2.18}$$

Podemos obtener una solución a partir de multiplicar reiteradamente $A\mathbf{x}$. Considérese la siguiente secuencia

$$x_n = k^{-1} A x_{n-1} \quad n \geq 1 \tag{2.19}$$

para x_0 cualquier vector inicial con elementos no negativos. Entonces,

$$x_1 = k^{-1} A x_0, \quad x_2 = k^{-1} A x_1 = k^{-2} A^2 x_0, \quad x_3 = k^{-1} A x_2 = k^{-3} A^3 x_0, \quad \cdots \tag{2.20}$$

En consecuencia, $x_n = k^{-n} A^n x_0$. Como tenemos una matriz cuadrada de tamaño mxm, tenemos m valores propios con sus respectivos vectores propios. Si escribimos x_0 como una combinación lineal bajo el espacio vectorial descrita por los vectores propios de A, entonces

$$x_n = k^{-n} A^n x_0 = k^{-n} A^n (c_1 v_1 + c_2 v_2 + \cdots + c_m v_m) = k^{-n} (c_1 A^n v_1 + c_2 A^n v_2 + \cdots + c_m A^n v_m)$$

Debido a que $Av_i = k_i$ y $A^2v_i = AAv_i = A(k_iv_i) = k_iAv_i = k_i(k_iv_i) = k_i^2v_i$ para todo valor propio k_i y vector propio v_i de la matriz A. Entonces,

$$x_n = k^{-n}(c_1k_1^nv_1 + c_2k_2^nv_2 + \cdots + c_nk_m^nv_m)$$

Supongamos que el valor propio más grande es k_1. De esta forma, podemos establecer que $k = k_1$. Entonces

$$x_n = k^{-n}k_1^n\left[c_1v_1 + \cdots + c_n\frac{k_m^n}{k_1^n}v_m\right] = c_1v_1 + \cdots + c_n\frac{k_m^n}{k_1^n}v_m \tag{2.21}$$

tenemos que $\frac{k_i}{k_1} < 0$ para toda $i \neq 1$. Entonces, para $n \to \infty$ tenemos que $x_n \to c_1v_1$. Por lo tanto, la secuencia vectorial $\{x_n\}$ tiende a un límite que es proporcional al vector propio principal propuesta por Bonacich en 1987 para ser la medida de centralidad.

El algoritmo para calcular la centralidad de vector propio de forma eficiente es conocido como iteración de potencia (power iteration en inglés) el cual usa el mismo procedimiento de multiplicar reiteradamente la matriz de adyacencia por un vector inicial con elementos no negativos. Debido a que los elementos del vector tienden a crecer a valores muy altos, tan grandes que la computadora no puede manejar, se normaliza en cada iteración.

Para comprobar la convergencia de la iteración, se ha planteado trabajar a partir de dos vectores iniciales distintos y ver si la diferencia absoluta está dentro de una tolerancia establecida. Por otro lado, para este propósito en lugar de iterar sobre la matriz A se propone trabajar sobre $A+I$, la matriz adyacencia más la matriz identidad, para avanzar en una iteración posterior y comparar que ambos vectores estén dentro del rango de tolerancia.

2.3.2. Centralidad de cercanía

La centralidad de cercanía se basa en lo cerca que está un nodo en relación a toda la red [37]. Entre más alejado está un nodo de los otros, tendrá una centralidad menor. Sabidussi en 1966 cuantifica de esta forma que la centralidad de un nodo es el inverso de su distancia promedio, es decir,

$$C_C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(v,u)} \tag{2.22}$$

donde n es el número de nodos en la red. En grafos no conectados esta ecuación tiene una desventaja si por lo menos un nodo w es aislado, al ser la centralidad de cercanía cero para todos los nodos debido a que $d(u,w) = \infty$. Por otro lado, si en la ecuación 2.22, n es el número de nodos en la fracción de red que contiene al nodo u, habría igualmente una desventaja. Si dos nodos w y z son adyacentes pero aislados al resto de la red,

$C_C(w) = C_C(z) = 1$. Es por eso que para un grafo no conectado, Wasserman y Faust propusieron una versión mejorada. Sea N el número total de nodos en la red y n el número de nodos en la fracción de red que contiene el nodo u.

$$C_{WF}(u) = \frac{n-1}{N-1}\frac{n-1}{\sum_{v=1}^{n-1} d(v,u)} \tag{2.23}$$

Si el grafo está completamente conectado, tenemos que $N = n$ y la formula se reduce a C_C.

Para ver más claramente la diferencia entre C_C y C_{WF}, usaremos la librería NetworkX para elegir un grafo de forma aleatoria uniforme de $N = 10$ nodos y 8 aristas [38], véase la Figura 2.7.

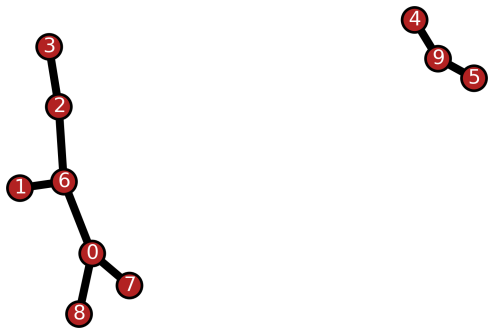


Figura 2.7: Centralidad de cercanía para un grafo aleatorio de 10 nodos.

El nodo 9 se encuentra conectado solo a $n - 1 = 2$ nodos, y tiene centralidad

$$C_C(9) = \frac{2}{\sum_v d(v,9)} = \frac{2}{d(4,9) + d(5,9)} = \frac{2}{1+1} = 1 \tag{2.24}$$

Mientras que la centralidad propuesta por Wasserman y Faust es

$$C_{WF}(9) = \frac{2}{9}\frac{2}{\sum_v d(v,9)} = \frac{2}{9}\left(\frac{2}{d(4,9) + d(5,9)}\right) = \frac{2}{9}\frac{2}{(1+1)} = \frac{2}{9} \tag{2.25}$$

El hecho de que $C_C(9) = 1$, quiere decir que el nodo 9 es el más cercano del grafo, véase el cuadro 2.1. No solo no podríamos caracterizar esta red con otras sino que perdería el concepto detrás de centralidad en la propia red; especialmente, en estructuras donde usamos las coordenadas de los átomos para crear la red, si los tres nodos fueran los carbonos periféricos de una estructura, la centralidad C_C puntuaría a uno de ellos como el más cercano a todos en la red.

Nodo	C_C	C_{WF}
0	0.6	0.3999

1	0.4286	0.2857
2	0.5	0.3333
3	0.3529	0.2857
4	0.6667	0.1481
5	0.6667	0.1481
6	0.6667	0.4444
7	0.4	0.2667
8	0.4	0.2667
9	1	0.2222

Cuadro 2.1: Comparación de la medida C_C y la medida mejorada C_{WF}.

Para hallar la distancia geodésica o ruta más corta de un nodo a los demás se usa un algoritmo eficiente conocido como *búsqueda en anchura* (breadth-first search en inglés, y abreviado como BFS). Consiste en recorrer la red por niveles a partir de un nodo inicial s conocido como raíz. El primer nivel está en visitar los nodos adyacentes a s, y después avanza a un nivel posterior hasta recorrer todos los nodos en la red. En cada visita a un nuevo nodo no explorado v, se le asigna la distancia geodésica igual al nivel en el que se encuentra $d_s(v) = d(s, v) =$ nivel i.

Sea $\{u_i\}, i = 1, .., k_s$ el conjunto de nodos adyacentes a s donde k_s es el grado de s. El primer nivel está en visitar los nodos adyacentes a s los cuales se les asigna la distancia geodésica igual al nivel en el que se encuentran, es decir $d_s(u_i) = 1$. Después avanza al nivel 2 en los nodos u_i, y de la misma manera, visita los nodos adyacentes de estos y se les asigna la distancia geodésica igual a 2. Así sucesivamente hasta recorrer toda la red. Note que cada nivel corresponde a la distancia geodésica desde el nodo raíz s al nodo actual. Por ejemplo, sea el siguiente grafo representado en la Figura 2.8. Si el nodo raíz es 0, en el primer nivel se visita a los nodos adyacentes 1 y 2, y se les asigna la distancia geodésica igual al primer nivel $d_0(1) = d_0(2) = 1$. Después avanzamos al siguiente nivel recorriendo a los nodos adyacentes 1 y 2, y visitando sus respectivos nodos adyacentes a los que se les asignará una distancia geodésica igual a 2, $d_0(3) = d_0(4) = d_0(5) = d_0(6) = 2$.

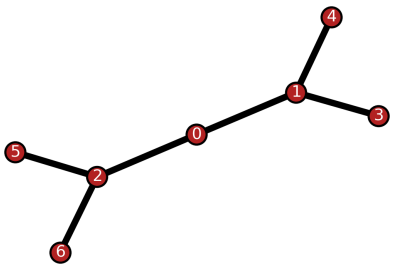


Figura 2.8: El algoritmo de búsqueda ampliada recorre cada nodo a partir de un nodo raíz.

2.3.3. Centralidad de intermediación

La centralidad de intermediación se basa en la frecuencia en que un nodo aparece en las distancias geodésicas de los demás. En otras palabras, un nodo será importante en la red si se encuentra entre los caminos más cortos de muchos nodos que no son adyacentes. La idea fue propuesta por Alex Bavelas en 1948 y Marvin Shaw en 1953 pero formalmente desarrollada por Anthonisse en 1971 (aunque no fue técnicamente publicada) y Freedman en 1977 [37].

En un análisis por Shimbel en 1953, y reconocida por Shaw en 1954, basada en la relación de comunicación entre los nodos en una red, Freeman establece formalmente la idea de centralidad de intermediación a partir de la definición clásica de probabilidad [37]. Sea σ_{st} el número de distancias geodésicas entre los nodos no adyacentes s y t. Si las rutas geodésicas tienen el mismo peso de ser elegidas, la probabilidad de tomar cualquiera de las σ_{st} rutas sería la misma, es decir, $1/\sigma_{st}$. Si $\sigma_{st}(v)$ es el número de rutas geodésicas que pasan por el nodo v entre la comunicación de s y t, entonces la probabilidad es

$$\frac{\sigma_{st}(v)}{\sigma_{st}} \quad s \neq t, v \tag{2.26}$$

Por lo tanto, la medida de centralidad de intermediación $C'_B(v)$ del nodo v está dada por

$$C'_B(v) = \sum_{s<t} \frac{\sigma_{st}(v)}{\sigma_{st}} \qquad s \neq t, \quad s \neq v \tag{2.27}$$

Si el nodo v está en todas las rutas geodésicas de los demás $n-1$ nodos, es decir, $\sigma_{st}(v) = \sigma_{st}$. Entonces,

$$C'_B(v) = \sum_{s<t} 1 = \binom{n-1}{2} = (n-1)(n-2)/2 \tag{2.28}$$

Por lo contrario, la puntuación sería cero si el nodo no se encuentra en ninguna de las distancias geodésicas.

Si estandarizamos esta medida dividiendo $C'_B(v)$ por el valor máximo. Entonces,

$$C_B(v) = \frac{C'_B(v)}{\binom{n-1}{2}} = \frac{2}{(n-1)(n-2)} \sum_{s<t} \frac{\sigma_{st}(v)}{\sigma_{st}} \qquad s \neq t, \quad s \neq v \tag{2.29}$$

De esta forma, el valor mínimo y máximo son 0 y 1, respectivamente. Esto da la ventaja de poder comparar la centralidad con otras redes sin importar el tamaño de las mismas [37].

En el año 2000, Ulrik Brandes publica un algoritmo para el cálculo de la centralidad de intermediación usando BFS como una primera parte para hallar las distancias y número de rutas geodésicas [39]. El algoritmo se basa en el criterio de Bellman por el cual divide el problema en subproblemas más simples (primera parte) para después obtener una solución general (segunda parte). En el apéndice A, puede ver el teorema y los lemas que Ulrik usa para establecer su algoritmo, y así evaluar la centralidad de intermediación de una red de forma eficiente.

Capítulo 3

Programación Python y los softwares DPX y VMD.

En el presente capítulo se hará una descripción de las herramientas de programación y los métodos de realización de los cálculos obtenidos para el análisis de la estructura de las proteínas globulares del SARS-CoV-2. En la primera sección hablaremos de las herramientas de programación indispensables de Python como NumPy, Pandas y NetworkX. En la segunda sección hablaremos de una herramienta para la visualización de las proteínas llamado *Visual Molecular Dynamics* (VMD). En la tercera sección hablaremos de dos softwares NACCESS y DPX para el análisis de la estructura globular proteica. En la cuarta sección hablaremos de la construcción de la red de carbonos C_α y C_β para distintos valores del parámetro radio.

3.1. Lenguaje de programación Python

En este trabajo utilizaremos el lenguaje de programación Python que es un lenguaje intérprete de código abierto. El proyecto Python empezó por Guido van Rossum con el propósito de crear un lenguaje de programación legible, óptimo al escribir y que facilitara el entendimiento entre programadores [40]. Guido van Rossum publicó su proyecto a finales de los años 80, y tal fue la aceptación de muchos que en la actualidad el proyecto Python está conformada por una gran comunidad en diversas áreas de la ciencia convirtiéndose en el lenguaje de programación más popular por varios años consecutivos [41,42].

Python es un lenguaje de programación multiparadigma entre los que se encuentra con mayor énfasis la programación orientada a objetos (en inglés, *object-oriented programming*). Este paradigma de programación se basa en la creación de **objetos** que tienen como elementos fundamentales una serie de datos conocidos como **atributos**, y una serie de funciones llamados **métodos**. Los objetos se crean a partir de una estructura (o modelo) llamado clase (en inglés, *class*). Por ejemplo, si creamos una clase que se llame

Proteína, y que nos sirva para contar el número de carbonos en la red:

```
class Proteina:
    def __init__(self, id_pdb, numero_CA, numero_CB):
        self.nombre_pdb = id_pdb
        self.numero_CA = numero_CA
        self.numero_CB = numero_CB

    def numero_total_de_carbonos(self):
        total = self.numero_CA + self.numero_CB
        print('El numero de carbonos en la red es',total)
```

Podemos ver que la clase se llama Proteína donde sus atributos son nombre_pdb, numero_CA y numero_CB que representan el nombre pdb, el número de carbonos C_α y el número de carbonos C_β, respectivamente. Creamos un método que realizará la suma, llamado numero_total_de_carbonos. A partir del modelo, podemos crear un objeto para la proteína Spike el cual tiene 2875 carbonos C_α y 2702 carbonos C_β

```
spike = Proteina("6vyb", 2875, 2702)
```

Si queremos saber el número de carbonos que habrá en la red C_α-C_β, invocamos el método *numero_total_de_carbonos*.

```
spike.numero_total_de_carbonos()
```

```
El número de carbonos en la red es 5577
```

En este trabajo creamos el modelo llamado de igual forma **Proteína**, que tiene como atributos el nombre PDB de la proteína, el número total de C_α en la proteína, el número total de C_β, el número de secuencia PDB de los aminoácidos que pertenecen a los sitios activos, y, por último, el atributo para el archivo PDB. Por lo tanto, un atributo puede ser una cadena de caracteres (el nombre PDB de la proteína), un número (la cantidad de C_α y C_β) u otro objeto del cual importamos el archivo PDB con la biblioteca Biopandas. Los métodos o funciones son aquellas que nos permitieron obtener: los cálculos, las gráficas de frecuencia de aparición de los tipos de aminoácidos en la proteína, los histogramas de grado, las gráficas de las medidas de centralidad, la visualización tridimensional tanto de la proteína como del grafo, etc.

La programación orientada a objetos nos ofrece la ventaja de reutilizar un mismo código. Cuando creamos un objeto, lo hacemos a partir del modelo Proteína. En este caso y en general, creamos tres objetos que corresponden a las tres proteínas objetivo de este trabajo que son 7LG7, 7C6S y 6VYB. Es decir, no es necesario crear un programa para

cada proteína sino que a partir del modelo Proteína podemos establecer los atributos y métodos del objetivo de nuestro proyecto para cada proteína. Este proceso se le conoce como **instanciar**.

La creciente comunidad en diversas áreas de la ciencia ha enriquecido la funcionalidad de Python con un amplio repertorio de bibliotecas con una estructura de datos importante. Una de estas bibliotecas es NumPy para el cálculo y análisis numérico de grandes bloques de datos, tal como su nombre lo indica de forma abreviada, *Numerical Python* [41]. Los datos se trabajan en un objeto llamado ndarray con el que es fácil trabajar en forma de vector o matriz. NumPy es muy eficiente cuando hacemos operaciones con grandes cantidades de datos debido a que ofrece mayor velocidad, simplicidad y ahorro de espacio. Por ejemplo, si queremos multiplicar dos matrices A y B con un millón de valores cada uno,

```
import numpy as np
from time import time

A = np.arange(1000000).reshape((1000,1000)) # Datos A
B = np.arange(1000000).reshape((1000,1000)) # Datos B

C = np.zeros(1000000).reshape((1000,1000)) # El resultado lo almacenaremos

def metodo_1(A, B):
    t0 = time()
    for i in range(1000):
        for j in range(1000):
            C[i,j] = A[i,j]*B[j,i]
    t1 = time()
    return t1-t0
```

donde primeramente, importamos la biblioteca como np. A y B son dos matrices cuadradas de tamaño 1000 y valores de 1 a un millón, y C es la matriz cuadrada con valores cero para almacenar el resultado del producto. El método 1 es una función que realiza la multiplicación de dos arreglos usando bucles, y retorna el tiempo estimado en hacer dicha operación. En cambio si usamos NumPy,

```
def metodo_2(A,B):
    t0 = time()
    A*B
    t1 = time()
    return t1-t0
```

Ahora, si repetimos 100 veces la misma operación. La siguiente fracción de código muestra

este proceso de multiplicar dos matrices de un millón de elementos:

```
tiempos_1 = np.zeros(100, dtype=float)
tiempos_2 = np.zeros(100, dtype=float)

for i in range(100):
    tiempos_1[i]=metodo_1(A,B)
    tiempos_2[i]=metodo_2(A,B)
```

El resultado fue que al repetir el mismo proceso fue de 3 minutos y 4 segundos para el primer método, y 0.27 segundos para el segundo método.

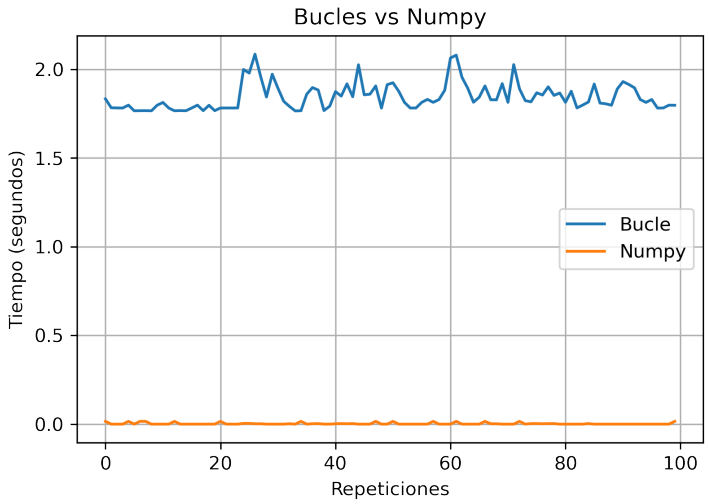


Figura 3.1: Comparando operación con bucles versus NumPy.

Usar NumPy también ofrece la ventaja de poder ahorrar memoria a diferencia de la estructura listas. NumPy no solo puede manejar números sino que también cadenas de caracteres, booleanos, listas o una combinación de éstos. Esta característica de NumPy permite hacer otras tareas a parte de operaciones numéricas como filtrar datos por medio de cadenas de caracteres y booleanos de forma eficiente como identificar, por ejemplo, los sitios activos según el número de PDB, la cadena a la que pertenece, etc.

Python posee otra estructura de datos llamado Pandas que ha sido esencial para este trabajo. Pandas es una biblioteca creada para el análisis y manejo de datos relacionales, similar a las hojas de cálculo. Es una herramienta que combina las características de cálculo de alto rendimiento de NumPy y la estructura de base de datos como SQL [41]. La biblioteca almacena los datos a partir de tres estructuras:

- la estructura *Series*: es una estructura unidimensional creada para manejar una columna de datos etiquetados, parecido a la estructura ndarray de NumPy con la

diferencia de que podemos relacionar los datos numéricos con un índice que puede ser una cadena de caracteres.

- la estructura *Dataframe*: es una estructura bidimensional similar a las tablas que comúnmente conocemos,
- la estructura *Panel*: es una estructura tridimensional que almacena los datos en forma de cubo. De hecho, el nombre de la biblioteca proviene del término datos panel en econometría y estadística. [41].

En nuestro caso, usaremos las dos primeras estructuras. La estructura Series nos sirvió para relacionar los carbonos C_α y C_β con sus respectivas centralidades. Esta estructura nos ofrece una variedad de métodos (o funciones) dentro de los cuales nos permitió organizar los datos de forma ascendente o descendente y poder identificar los primeros n carbonos C_α y C_β con valores máximos y mínimos. Por otro lado, la estructura Dataframe fue utilizado para almacenar los datos del registro ATOM del archivo PDB (figura 1.6 (b)). La estructura nos permite acceder a la información de cualquier átomo de la proteína objetivo como las coordenadas espaciales, el número de secuencia pdb, el aminoácido al que pertenece, etc. Por ejemplo, en la figura 3.2 tenemos una fracción de Dataframe de la proteína 7C6S.

	record_name	atom_number	...	atom_name	...	x_coord	y_coord	z_coord
0	ATOM	1		N		-2.52	-30.826	37.787
1	ATOM	2		CA		-2.977	-32.194	37.583
2	ATOM	3		C		-4.178	-32.204	36.65
3	ATOM	4		O		-4.914	-31.219	36.554
4	ATOM	5		CB		-3.328	-32.857	38.92

Figura 3.2: Los datos son almacenados en una estructura Dataframe. En la figura se muestra una fracción del Dataframe de la proteína 7C6S. En la columna atom_name están identificados los carbonos C_α y C_β como CA y CB, respectivamente.

La columna **atom_name** identifica el tipo de átomo por el cual podemos encontrar los carbonos C_α y C_β registrados como CA y CB, respectivamente. Las columnas **x_coord**, **y_coord** y **z_coord** corresponden a las coordenadas espaciales de cada átomo. La estructura Dataframe nos permite pasar los datos a una estructura Series o un ndarray. En particular, el segundo nos servirá para hacer cálculos a partir de las coordenadas espaciales.

Para importar los archivos PDB se ha creado la biblioteca Biopandas basándose, como su nombre lo dice, en la biblioteca Pandas. Por lo tanto, nos permite manipular los datos en la estructura Dataframe con los registros ATOM del archivo PDB. Uno de los métodos de la biblioteca es `fetch_pdb` que nos permite importar un archivo PDB desde la página

del banco de proteínas (RCSB PDB), por lo que es necesario tener acceso a internet. Podemos también importar directamente de nuestro equipo con `read_pdb`.

```
from biopandas.pdb import Pdb

pdb_7lg7 = Pdb().fetch_pdb("7lg7")
pdb_7c6s = Pdb().read_pdb("ruta_del_pdb/7c6s.pdb")

print(pdb_7lg7["ATOM"])
```

	record_name	atom_number	blank_1	...	element_symbol	charge	line_idx
0	ATOM	1		...	N	NaN	459
1	ATOM	2		...	N	NaN	461
2	ATOM	3		...	C	NaN	463
3	ATOM	4		...	C	NaN	465
4	ATOM	5		...	C	NaN	467
...	...	...	...	...	...	...	...
1319	ATOM	1320		...	C	NaN	3097
1320	ATOM	1321		...	C	NaN	3099
1321	ATOM	1322		...	O	NaN	3101
1322	ATOM	1323		...	C	NaN	3103
1323	ATOM	1324		...	O	NaN	3105

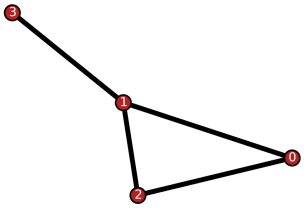
[1324 rows x 21 columns]

Figura 3.3: Dataframe de la proteína 7LG7.

En teoría de grafos, se ha creado la biblioteca NetworkX de código abierto para el análisis de redes con una capacidad de manejar incluso más de 10 millones de nodos y 100 millones de aristas [43]. La biblioteca tiene una estructura de datos eficiente y flexible por lo que puedes crear un objeto grafo a partir de una matriz de adyacencia con NumPy, e incluso con nodos que pueden ser cualquier tipo de objeto hashable (números, cadenas de caracteres, imágenes, otro grafo, etc).

Supongamos que tenemos la siguiente matriz de adyacencia,

$$\mathcal{A} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$



Podemos generar un objeto grafo G a partir de la matriz de adyacencia invocando el método `from_NumPy_array` de NetworkX. Para ello, necesitamos importar las bibliotecas NumPy y NetworkX.

```
import NumPy as np
import networkx as nx
```

```
A = np.array([[0,1,1,0],[1,0,1,1], [1,1,0,0],[0,1,0,0]],dtype=int)
G = nx.from_NumPy_array(A)

numero_de_nodos = G.number_of_nodes()
numero_de_aristas = G.number_of_edges()

print("El numero de nodos en el grafo son ", numero_de_nodos)
print("El numero de aristas en el grafo son ", numero_de_aristas)
```

```
El numero de nodos en el grafo son 4
El numero de aristas en el grafo son 4
```

Una vez que instanciamos el objeto grafo en la variable G, adquirimos una variedad de rutinas que nos proporciona información acerca del grafo como el número total de aristas y número de nodos. También nos proporciona información valiosa acerca de un nodo o un grupo de nodos como el grado, el número de aristas y los nodos vecinos. Esto atributos nos será de utilidad para construir la distribución de grado, conocer el grado promedio y generar grafos aleatorios a partir de las características de nuestras redes de carbonos C_α y C_β.

```
for n, d in G.degree((0,1)):
    print("El grado del nodo",n,"es:",d)
```

```
El grado del nodo 0 es: 2
El grado del nodo 1 es: 3
```

La biblioteca ofrece una amplia funcionalidad y algoritmos que nos permiten analizar la topología de la red como la distancia geodésica promedio, el coeficiente de agrupamiento, las medidas de centralidad, entre otros herramientas de análisis. Para el ejemplo anterior podemos obtener la distancia promedio y la transitividad (o coeficiente de agrupamiento) del grafo con average_shortest_path_length y transitivity, respectivamente.

```
l = nx.average_shortest_path_length(G)
c = nx.transitivity(G)

print("La distancia geodesica promedio del grafo es:",l)
print("El coeficiente de agrupamiento del grafo es:",c)
```

```
La distancia geodésica promedio del grafo es: 1.33
El coeficiente de agrupamiento del grafo es: 0.6
```

NetworkX incorpora los algoritmos para evaluar la centralidad de vector propio, cercanía e intermediación de una red con eigenvector_centrality, closeness_centrality y betweenness_centrality, respectivamente. Cada algoritmo requiere un grafo como parámetro. El algoritmo eigenvector_centrality también requiere el número máximo de iteraciones **max_iter** en el algoritmo de iteración de potencia, que por defecto son 100 iteraciones, y la tolerancia de convergencia, que por defecto es **tol** = 1.0e-6. Por ejemplo, en nuestro grafo G, la centralidad de vector propio de cada nodo es el siguiente:

```
EC = nx.eigenvector_centrality(G)

for n in EC:
    print("La centralidad del nodo",n,"es:",EC[n])
```

```
La centralidad del nodo 0 es: 0.5227
La centralidad del nodo 1 es: 0.6116
La centralidad del nodo 2 es: 0.5227
La centralidad del nodo 3 es:  0.2818
```

La biblioteca también posee algoritmos para la generación de grafos aleatorios a partir de los modelos Erdős-Rényi, Watts-Strogatz, Barabási-Albert y grafos tipo regular. En la tabla 3.1 se muestra un resumen de los respectivos comandos y parámetros. Por ejemplo, si queremos crear un grafo aleatorio tipo Erdős-Rényi con el número de nodos n y número de aristas m igual a los del grafo G invocamos el comando gnm_random_graph(n, m). La Figura 3.4 muestra dos grafos diferentes del modelo Erdős-Rényi al ejecutar una vez más el código.

```
n = G.number_of_nodes()
m = G.number_of_edges()

# Genera un grafo aleatorio con el modelo Erdos-Renyi
G_ER = nx.gnm_random_graph(n, m)
```

El amplio repertorio de bibliotecas de Python y la eficiencia de sus estructuras de datos, gracias a la amplia comunidad en distintos frentes de la ciencia, nos permite hacer un análisis de la topología estructural de las proteínas objetivo del SARS-CoV-2 y la relevancia de sus sitios activos en la estructura. A pesar de analizar proteínas grandes como la proteína Spike con 5577 carbonos C_α y C_β podremos hacer cálculos repetidas veces en muy poco tiempo.

Algoritmos	Comando	Parámetros
Centralidad de cercanía	closeness_centrality(G)	G:grafo
Centralidad de intermediación	betweenness_centrality(G)	G:grafo
Centralidad de vector propio	eigenvector_centrality(G, max_iter, tol)	G: grafo. max_iter: Numero de iteraciones máximo en el método de potencia. tol: tolerancia de convergencia.
Generador de grafos Erdős-Rényi	gnm_random_graph(n, m)	m = número de aristas
Generador de grafos Watts-Strogatz	watts_strogatz_graph(n, k, p)	n: número de nodos. p: probabilidad de volver a conectar una arista de cada nodo. k: grado inicial de cada nodo.
Generador de grafos Barabási-Albert	barabasi_albert_graph(n, m)	n: número de nodos. m = número de aristas de un nuevo nodo en el proceso PA.
Generador de grafos tipo regular	random_regular_graph(d, n)	n: número de nodos. d = grado de cada nodo.

Cuadro 3.1: Herramientas de NetworkX.

Figura 3.4: Grafos obtenidos con el modelo Erdős-Rényi con el número de nodos y número de aristas del grafo G.

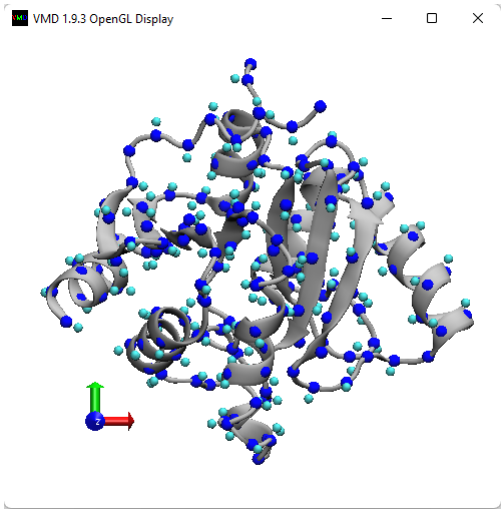
3.2. Software Visual Molecular Dynamics (VMD)

Para la visualización y el análisis estructural de las proteínas del SARS-CoV-2 se usó el software VMD (*Visual Molecular Dynamics*, por sus siglas en inglés) distribuido gratuitamente por el Grupo de Biofísica Teórica y Computacional de la Universidad de Illinois. El programa puede visualizar cualquier proteína desde sus componentes atómicos a través del archivo pdb, figura 3.5 (d), e identificar tanto las particularidades estructurales y complejos como las propiedades físicas y químicas. VMD posee tanto un interfaz gráfico como un interfaz de comandos basado en el lenguaje TCL o Python en donde se puede ingresar instrucciones de visualización, figura 3.5 (b) y (c). El interfaz de comandos a sido útil para asignar un mapa de colores de forma automática a los carbonos C_α y C_β con respecto a los valores de centralidad, descritos en el capítulo anterior. Por medio de un algoritmo de programación en el lenguaje TCL o Python en un archivo de texto es posible vincular VMD con otros programas. Si calculamos la medida de centralidad en NetworkX para cada átomo en la red, VMD puede asignar una escala de colores conforme a los valores. De esta forma, VMD expande su rango de funcionalidad por lo que se ha vuelto uno de los software más utilizados en la investigación científica.

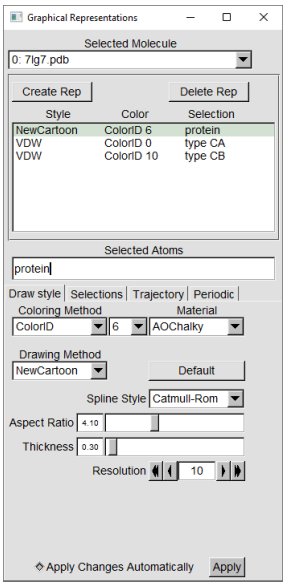
3.3. NACCESS y DPX

En 1971 Richards y Lee desarrollaron un método para describir el área de superficie accesible (ASA) de una molécula al hacer pasar una esfera (también conocido como sonda) sobre la superficie. La ruta que describe el centro de la esfera es el área de superficie accesible la cual es, en forma análoga, como un cascarón que envuelve a la molécula, Figura 3.6. Cuando el radio de la esfera es de 1.4 Å, que es el radio de una molécula de agua, la superficie se le conoce como superficie accesible al solvente (SASA) [44].

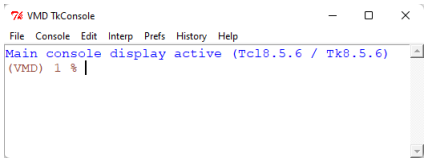
El software Naccess, de distribución gratuita, tiene como objetivo calcular el área de superficie accesible atómica para cualquier proteína o ácido nucleico. El programa hace



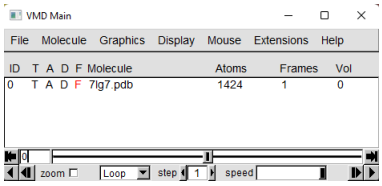
(a) Representación gráfica.



(b) Interfaz gráfico.



(c) Interfaz de comandos.



(d) VMD lee un archivo PDB

Figura 3.5: Software VMD.

sucesivos cortes al volumen de la molécula para calcular el asa de cada átomo [45]. Los parámetros del programa son el radio de la sonda que tiene por defecto 1.4Å para calcular la superficie accesible al solvente, y el radio atómico que por defecto es el radio de van der Waals [45]. Los resultados son presentados en tres archivos: uno con los cálculos ASA, un archivo con información relacionado al cálculo, y un tercer archivo que contiene los valores absolutos ASA por cada residuo de aminoácido y los valores relativos como la razón de la suma de los diversos valores de ASA y el valor máximo observado en un tripéptido ALA-X-ALA.

El área de superficie accesible atómica al solvente puede ser usado para obtener la medida de profundidad dpx_i en que un átomo i de un aminoácido está aglutinado en el interior de una proteína. Este es el objetivo del programa DPX, el cual usa NACCESS para el cálculo de ASA para identificar los átomos accesibles al solvente. Para cada átomo i, el programa calcula el mínimo de las distancias cartesianas al átomo accesible al solvente, es decir,

$$dpx_i = \text{mín}\{d_1, d_2, \cdots, d_n\} \tag{3.1}$$

donde d_j es la distancia del átomo i al átomo j accesible al solvente. De esta forma, la distancia $dpx_i = 0$ si el átomo i es accesible al solvente, de lo contrario $dpx_i > 0$.

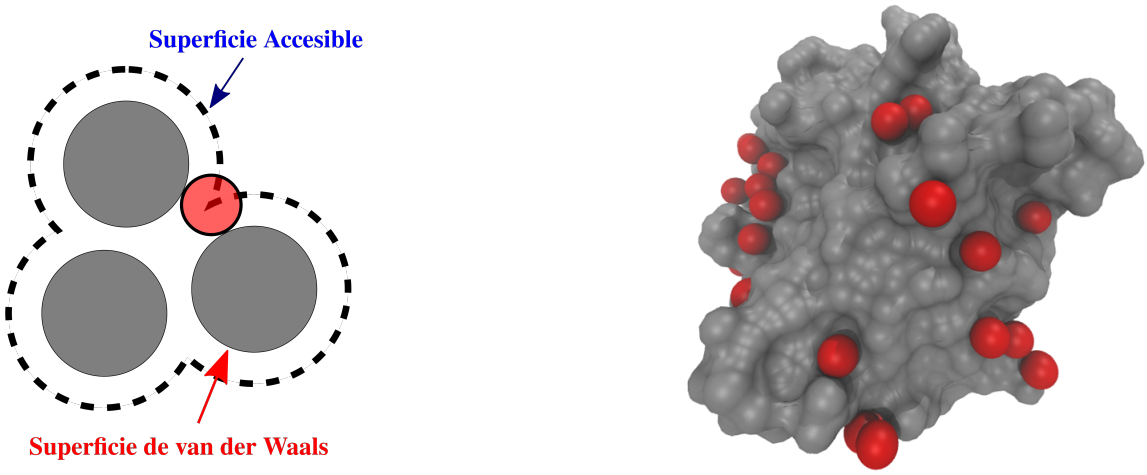


Figura 3.6: Representación de la superficie accesible atómica. En (a), las líneas consecutivas corresponden a la ruta que describe el centro de la esfera roja (o sonda) sobre la superficie de van de Waals de los átomos mostrados en gris, el área generada es conocida como superficie accesible [44]. En (b), tenemos la proteína 7LG7 en su representación de superficie accesible al solvente. Las esferas rojas corresponden a los átomos de oxígeno de las moléculas de agua.

El programa DPX al igual que NACCESS usa como parámetro 1.4Å para el radio de la sonda, Figura 3.7. El programa produce un archivo en formato PDB donde el campo del factor B ahora contiene el valor dpx [18]. Los residuos de aminoácidos que están más aglutinados en el interior de la proteína tienen una implicación importante tanto en el proceso de plegamiento como en la estabilidad termodinámica, y también da la posibilidad de clasificar los átomos de proteínas con independencia de sus propiedades físico químicas [18].

3.4. Red de carbonos C_α y C_β

Anteriormente vimos que podemos limpiar y filtrar los datos con Pandas y NumPy que nos permitirá obtener las coordenadas de los carbonos C_α y C_β de la secuencia PDB de aminoácidos. A partir de esto, crearemos tres matrices de distancia euclidiana $\mathcal{D}_\alpha$, $\mathcal{D}_\beta$ y $\mathcal{D}_{\alpha,\beta}$ correspondiente a los carbonos de tipo C_α, a los carbonos de tipo C_β y, por último, la combinación de estos dos.

Sean (x_i, y_i, z_i) y (x_j, y_j, z_j), respectivamente, las coordenadas cartesianas de los carbonos i y j de la secuencia PDB. Entonces, la distancia euclidiana es

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$

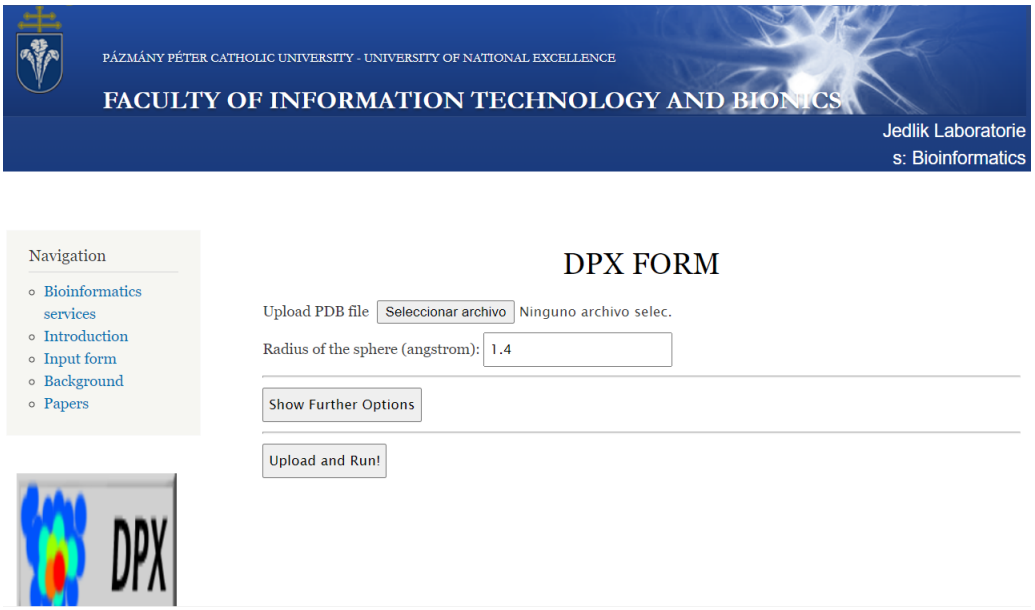


Figura 3.7: El servidor web de DPX puede ser usada para analizar los aminoácidos que están más aglutinados en el interior de una proteína globular [18].

La primera matriz D_α representará solo las distancias entre los carbonos C_α:

$$\mathcal{D}_\alpha = \begin{pmatrix} 0 & d_{1,2} & \cdots & d_{1,N} \\ d_{2,1} & 0 & \cdots & d_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N,1} & d_{N,2} & \cdots & 0 \end{pmatrix}$$

donde los elementos de la diagonal principal son cero ya que corresponde a la distancia de un carbono consigo mismo, entonces, $d_{i,j} = 0$ para $i = j$. La segunda matriz D_β representará las distancias entre los carbonos C_β

$$\mathcal{D}_\beta = \begin{pmatrix} 0 & d_{1,2}^{(\beta)} & \cdots & d_{1,N}^{(\beta)} \\ d_{2,1}^{(\beta)} & 0 & \cdots & d_{2,N}^{(\beta)} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N,1}^{(\beta)} & d_{N,2}^{(\beta)} & \cdots & 0 \end{pmatrix}$$

Esta matriz es más chica debido a que el aminoácido más sencillo llamado Glicina no posee carbono C_β sino un átomo de hidrógeno como cadena lateral. La tercera matriz $D_{\alpha,\beta}$ representará las distancias entre los carbonos C_α y C_β.

$$\mathcal{D}_{\alpha,\beta} = \begin{pmatrix} 0 & d_{1,2}^{(\alpha,\beta)} & \cdots & d_{1,N}^{(\alpha,\beta)} \\ d_{2,1}^{(\alpha,\beta)} & 0 & \cdots & d_{2,N}^{(\alpha,\beta)} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N,1}^{(\alpha,\beta)} & d_{N,2}^{(\alpha,\beta)} & \cdots & 0 \end{pmatrix}$$

Debido a que el orden de los carbonos en las filas y columnas corresponden a la secuencia PDB, el posicionamiento será el carbono C_α seguido por su correspondiente carbono C_β.

Observe que las tres matrices de distancia son simétricas y que la segunda diagonal contiene las distancias entre los carbonos sucesivos en la secuencia PDB, véase las gráficas 3.9. Además, en la tercera matriz podemos identificar tres distancias características con diferencias notables que corresponden a la distancia de un C_α con su respectivo C_β, la distancia de C_β con el C_α del residuo siguiente y la distancia de un carbono C_α de un residuo de glicina con el C_α del próximo residuo, véase las gráficas (c) y (f).

En la gráfica (a) de la figura 3.9 tenemos las distancias entre los carbonos C_α consecutivos de la proteína 7LG7 la cual tiene un promedio de 3.8070 Å; en (b), tenemos la distancia entre los carbonos C_β consecutivos de 7LG7 con un promedio de 5.652 Å ; y en (c) tenemos las distancias entre los carbonos C_α y C_β que tuvieron un promedio de 1.5318 Å para los carbonos de un mismo aminoácido (que son los que están alrededor de 1.5 en la gráfica), 4.589275 Å para los carbonos C_α y C_β de aminoácidos consecutivos; y, por último, 3.0881 Å en general entre todos los carbonos C_α y C_β, es decir, el promedio de la segunda diagonal de la matriz $\mathcal{D}_{\alpha,\beta}$. En el siguiente cuadro resumimos estos valores para cada proteína objetivo.

	C_α	C_β	$C_{\alpha,\beta}$
7LG7	3.8070Å	5.6520Å	1.5318Å * 4.5893Å** 3.0555Å***
7C6S	3.8050Å	5.7120Å	1.5314Å 4.6008Å 3.0953Å
6VYB	3.9891Å	5.8305Å	1.5266Å 4.7692Å 3.1476Å

Cuadro 3.2: En el cuadro representamos las distancias entre los carbonos C_α y C_β consecutivos de la secuencia PDB. En la columna $C_{\alpha,\beta}$ tenemos tres valores para cada proteína, el primero (*) se refiere a la distancia promedio entre los C_α y C_β de un mismo residuo de aminoácido; la segunda (**) se refiere a la distancia promedio entre los C_α y C_β de residuos diferentes consecutivos; y, por último, el tercero (***) se refiere a la distancia promedio entre los C_α y C_β en total.

A partir de las matrices de distancia crearemos otras tres matrices correspondientes, llamadas matrices de adyacencia. Entonces, sea $a_{i,j}^{(\alpha)}$ la componente ij de la matriz de adyacencia de carbonos alfa. Si el carbono j está dentro del radio R de distancia del carbono

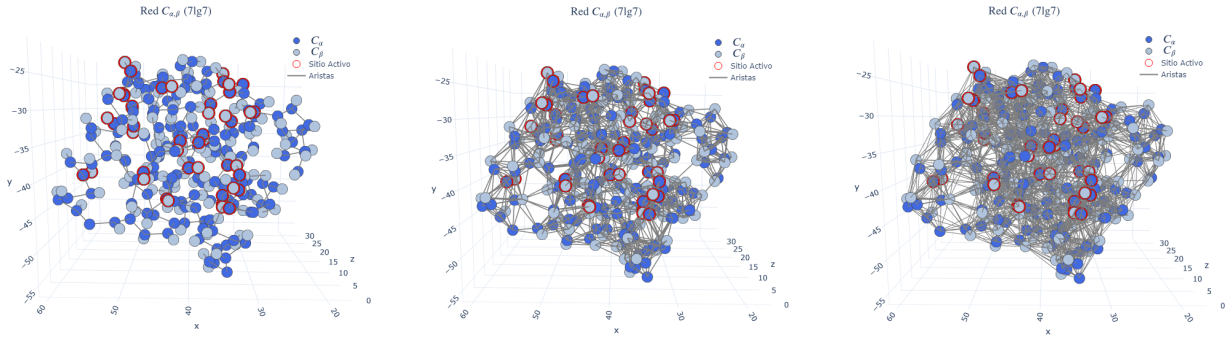
i entonces $a_{i,j}^{(\alpha)} = 1$, o de lo contrario es cero.

$$\mathcal{A}_{(\alpha)}(R) = \begin{pmatrix} 0 & a_{1,2}^{(\alpha)} & \cdots & a_{1,N}^{(\alpha)} \\ a_{2,1}^{(\alpha)} & 0 & \cdots & a_{2,N}^{(\alpha)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1}^{(\alpha)} & a_{N,2}^{(\alpha)} & \cdots & 0 \end{pmatrix}$$

Haremos lo mismo para la matriz del conjunto de carbonos beta y la matriz del conjunto de carbonos alfa y beta.

$$\mathcal{A}_R^{(\beta)} = \begin{pmatrix} 0 & a_{1,2}^{(\beta)} & \cdots & a_{1,N}^{(\beta)} \\ a_{2,1}^{(\beta)} & 0 & \cdots & a_{2,N}^{(\beta)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1}^{(\beta)} & a_{N,2}^{(\beta)} & \cdots & 0 \end{pmatrix} \quad \text{y} \quad \mathcal{A}_R^{(\alpha,\beta)} = \begin{pmatrix} 0 & a_{1,2}^{(\alpha,\beta)} & \cdots & a_{1,N}^{(\alpha,\beta)} \\ a_{2,1}^{(\alpha,\beta)} & 0 & \cdots & a_{2,N}^{(\alpha,\beta)} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1}^{(\alpha,\beta)} & a_{N,2}^{(\alpha,\beta)} & \cdots & 0 \end{pmatrix}$$

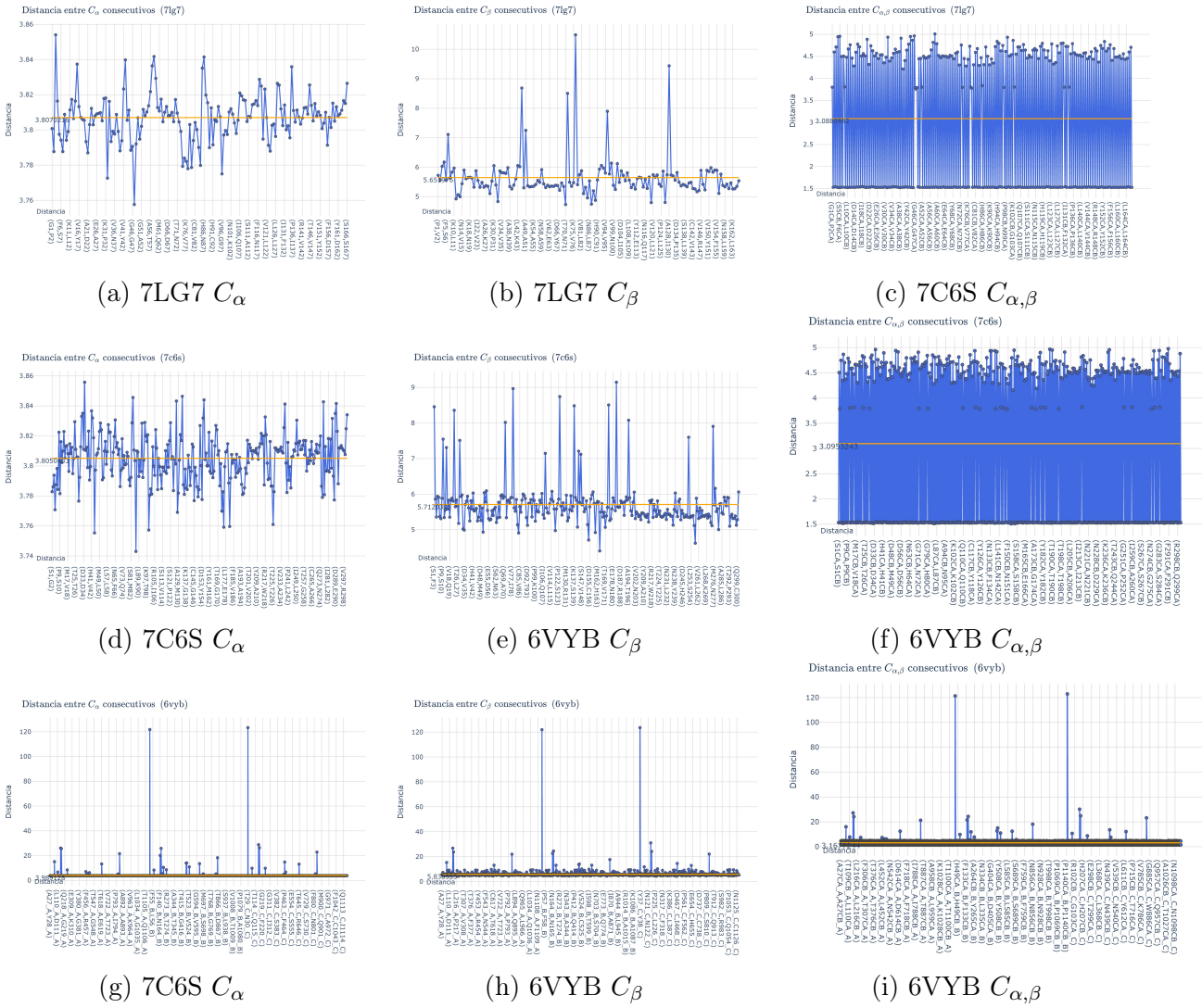
Esto nos proporciona la información necesaria para construir un grafo y una estructura en NetworkX donde los nodos representan los carbonos y las aristas las conexiones en función del radio de influencia R_i. La estructura en NetworkX nos permitirá calcular la centralidad de intermediación, centralidad de cercanía, centralidad de vector propio y centralidad de grado.



(a) R = 4 (b) R = 6 (c) R = 8

Figura 3.8: Red de carbonos $C_{\alpha,\beta}$ para la proteína 7LG7.

En la figura 3.9 (e) para 6VYB se muestran saltos de más de 6 Å hasta una distancia con más de 120 Å que corresponden al residuo 1147 de la cadena A con el residuo 27 de la cadena B, y el residuo 1147 de la cadena B con el 27 de la cadena C debido a que hay aminoácidos que no se observaron en los experimentos cristalográficos y las coordenadas no se incluyeron como registros ATOM.



(a) 7LG7 C_α (b) 7LG7 C_β (c) 7C6S $C_{\alpha,\beta}$

(d) 7C6S C_α (e) 6VYB C_β (f) 6VYB $C_{\alpha,\beta}$

(g) 7C6S C_α (h) 6VYB C_β (i) 6VYB $C_{\alpha,\beta}$

Figura 3.9: Gráficas de la distancias entre carbonos C_α y carbonos C_β para 7LG7, 7C6S y 6VYB.

Capítulo 4

Centralidades de las redes C_α, C_β y C_α-C_β.

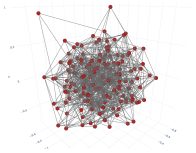
El enfoque de este capitulo es mostrar los datos que reflejan las características particulares de la estructura de red de carbonos C_α y C_β de las proteínas objetivo del SARS-CoV-2. La primera parte de este capítulo está enfocado en describir la estructura de red generado a partir del radio de influencia R visto en la última sección del capítulo anterior. En las restantes secciones aplicaremos las medidas de centralidad vistos en el capítulo 2 en la red de carbonos C_α y C_β de las proteínas objetivo tipo salvaje y tipo mutante. Las proteínas tipo mutante son aquellas que tienen uno o más carbonos bloqueados.

4.1. Estructura de red de carbonos C_α y C_β

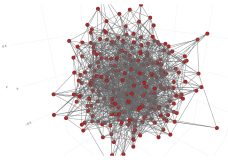
En el capítulo 2 vimos que las redes de origen teórico como el modelo de Erdős-Rényi contrastan a las redes del mundo real pero, sin embargo, se han creado modelos como Watts-Strogatz y el modelo Barabási-Albert que han podido suscitar ciertas propiedades encontradas en el mundo real. En este trabajo hemos generado 100 grafos aleatorio a partir de cada modelo con parámetros obtenidos de las redes de C_α, C_β y C_α-C_β de las proteínas objetivo como el grado promedio y el número de aristas.

Recordemos que en el modelo de Erdős-Rényi 2.7 un grafo con n nodos y m aristas es elegido aleatoriamente de entre los $\binom{\binom{n}{2}}{m}$ grafos. Entonces, elegimos el número nodos como el número de carbonos y el número de aristas como el número de conexiones encontradas en la red de radio de interacción $R = 8$Å para cada proteína. La figura 4.1 (a), (b) y (c) muestra la representación del modelo Erdős-Rényi con las características de la red C_α de las proteínas 7LG7, 7C6S y 6VYB, respectivamente.

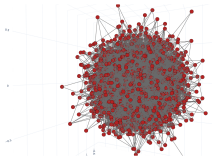
En el modelo Watts-Strogatz en el cual empieza con un grafo regular en una topología de anillo de n nodos unidos a m nodos vecinos. Los nodos son vuelto a conectarse



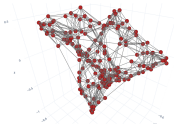
(a) Modelo Erdős-Rényi con 167 nodos.



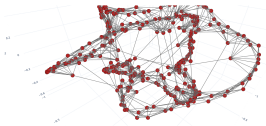
(b) Modelo Erdős-Rényi con 301 nodos.



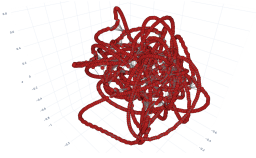
(c) Modelo Erdős-Rényi con 2875 nodos.



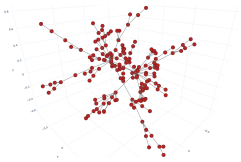
(d) Modelo Watts-Strogatz con 167 nodos.



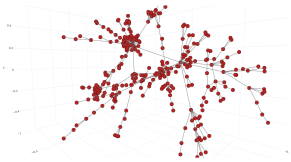
(e) Modelo Watts-Strogatz con 301 nodos.



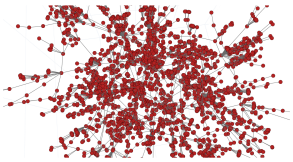
(f) Modelo Watts-Strogatz con 2875 nodos.



(g) Modelo Barabási-Albert con 167 nodos.



(h) Modelo Barabási-Albert con 301 nodos.



(i) Modelo Barabási-Albert con 2875 nodos.

Figura 4.1: Centralidad de vector propio para los modelo de grafo aleatorio tipo Erdős-Rényi, Watts-Strogatz y Barabási-Albert.

a nodos distintos de la red con probabilidad p. Los primeros dos parámetros n y m son elegidos con el número de carbonos y el número de aristas encontradas en la red de radio de interacción $R = 8$Å, respectivamente. El tercer parámetro p se obtiene a partir de la ecuación 2.10, es decir, $p = (n-1)p/\langle k \rangle$. En el Cuadro 4.1 la 4 columna se obtiene el grado promedio encontrada para cada red de las proteínas objetivo. En la Figura 4.1 (d), (f) y (e) muestra el grafo para el modelo Watts-Strogatz con parámetros de acuerdo a las propiedades encontradas en las proteínas 7LG7, 7C6S Y 6VYB.

En el modelo de Barabási-Albert los parámetros son el número de nodos n, el número de enlaces de un nuevo nodo agregado en la red el cual es el número promedio $\langle k \rangle$ en cada una de las redes C_α y C_β mostrado en el cuadro 4.1. Recordemos que el modelo empieza con un número inicial m_0 de nodos, la biblioteca NetworkX por defecto empieza con un grafo estrella con $m_0 = \langle k \rangle$ nodos, la cual es un grafo con un nodo central unido a los $m_0 - 1$ nodos por una sola arista. En la Figura 4.1 (g), (h) y (i) muestra el grafo producido con el modelo Barabási-Albert. Las gráficas no presentan el grafo real, la cual el número de enlaces al agregar un nodo nuevo es igual 1, para enfatizar la idea de red de escala libre donde se puede apreciar los nodos que son centrales, es decir, que tienen un grado alto que la mayoría de los nodos.

Una de las características definitorias de una red es la distribución de grado [33], por lo cual en la Figura 4.2 muestra este resultado para cada una de las redes de carbonos C_α

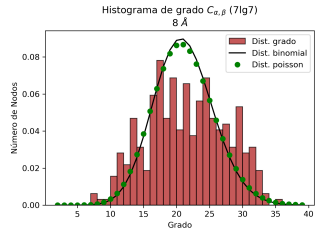
Red	PDB	m	$\langle k \rangle$	C	C_{ER}	C_{WS}	C_{RG}	l	l_{ER}	l_{WS}	l_{RG}
C_α-C_β	7LG7	3375	21.09	0.529	0.066	0.581	0.710	3.4	2.2	2.8	8.9
	7C6S	5797	20.09	0.518	0.035	0.639	0.710	4.9	2.4	3.4	14.9
	6VYB	58012	20.80	0.517	0.004	0.702	0.711	11.5	3.1	9.5	139.9
C_α	7LG7	850	10.18	0.494	0.061	0.552	0.666	3.7	3.5	3.4	8.8
	7C6S	1453	9.65	0.481	0.032	0.600	0.667	5.4	2.6	4.5	15.5
	6VYB	14367	9.99	0.483	0.004	0.659	0.667	12.9	3.7	18.7	144.2
C_β	7LG7	753	9.84	0.452	0.064	0.543	0.667	3.5	3.5	3.3	8.1
	7C6S	1301	9.43	0.439	0.034	0.579	0.643	5.1	2.8	5.1	17.7
	6VYB	13097	9.69	0.434	0.004	0.660	0.667	12.1	3.3	18.3	135.6

Cuadro 4.1: Propiedades para cada una de las redes C_α y C_β con un radio de interacción $R = 8$Å. En C, C_{ER}, C_{WS} y C_{RE} tenemos el coeficiente de agrupamiento para la red de carbonos, el modelo de Erdős-Rényi, el modelo de Watts-Strogatz, el grafo aleatorio regular, respectivamente. El promedio de las distancias geodésicas está representada como l, l_{ER}, l_{WS} y l_{RE} para la red de carbonos, el modelo de Erdős-Rényi, el modelo de Watts-Strogatz y el grafo aleatorio regular, respectivamente.

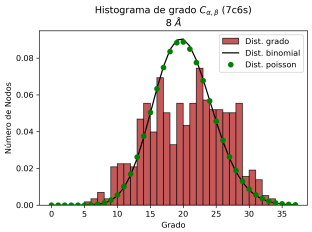
y C_β con radio $R = 8$Å de las proteínas objetivo tipo salvaje, y los comparamos con lo que esperaríamos en una distribución binomial (línea continua) y la distribución de Poisson (puntos en color verde), véase las gráficas 4.2. Sin embargo, podemos ver que no se ajusta a ninguna de estas dos distribuciones.

Por otro lado, una de las características topológicas de las redes del mundo real es que tienden a presentar la propiedad de mundo pequeño visto el capítulo 1, los cuales son: un promedio de distancias geodésicas menores y un coeficiente de agrupamiento alto. Duncan Watts y Steven Strogatz determinaron que las redes del mundo real tienden a estar entre la topología de redes aleatorias y las redes regulares. En el Cuadro 4.1 calculamos el coeficiente de agrupamiento y el promedio de las distancias geodésicas de 100 grafos aleatorios del modelo Erdős-Rényi, el modelo Watts-Strogatz y 100 grafos regulares, y los promediamos. Posteriormente, los comparamos con el coeficiente de agrupamiento C y el promedio de las distancias geodésicas l de cada una de las redes C_α, C_β y C_α-C_β. Sin embargo, los datos muestra que la topología de la red de carbonos no es la excepción a la propiedad de mundo pequeño y que ésta se encuentra en algún lugar entre las propiedades de los grafos aleatorios y grafos regulares

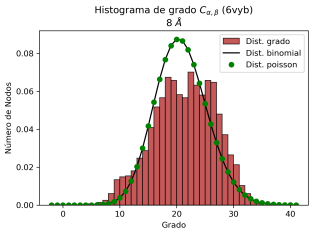
Podemos ver que en la red de carbonos C_α-C_β tienen mayor coeficiente de agrupamiento que la red de carbonos individuales mostradas en la segunda y tercer fila. Esto se debe a la corta distancia entre los C_α y C_β de un mismo aminoácido, si comparamos la distancia entre los C_α o C_β de residuos consecutivos o que son adyacentes en el espacio cartesiano. En la última sección del capítulo anterior registramos que la distancia promedio entre los C_α era de aproximadamente 4 Å, y 6 Å para los C_β. La diferencia de distancia de las redes individuales debe influir en el coeficiente de agrupamiento.



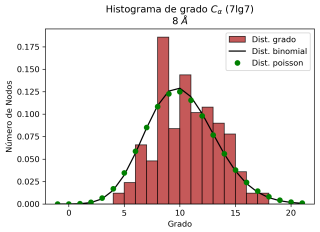
(a) Red de carbonos C_α y C_β de la proteína 7LG7.



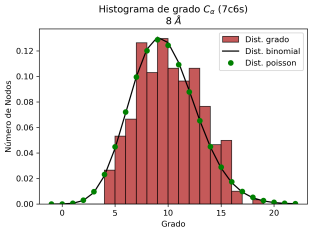
(b) Red de carbonos C_α y C_β de la proteína 7C6S.



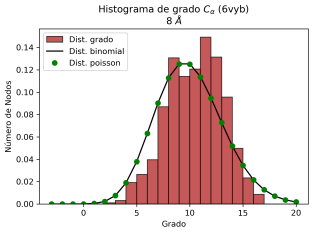
(c) Red de carbonos C_α y C_β de la proteína 6VYB.



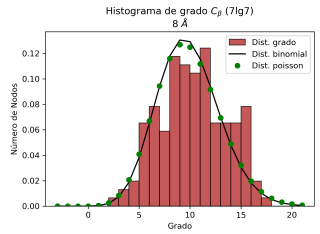
(d) Red de carbonos C_α de la proteína 7LG7.



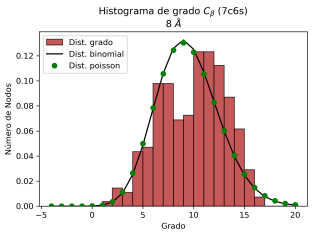
(e) Red de carbonos C_α de la proteína 7C6S.



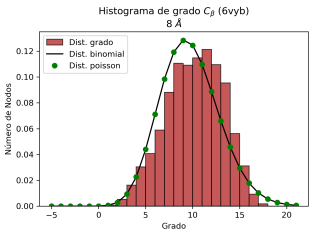
(f) Red de carbonos C_α de la proteína 6VYB.



(g) Red de carbonos C_β de la proteína 7LG7.



(h) Red de carbonos C_β de la proteína 7C6S.

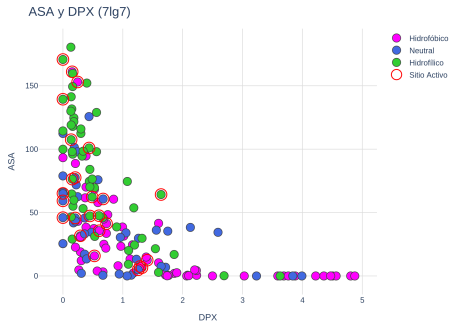


(i) Red de carbonos C_β de la proteína 6VYB.

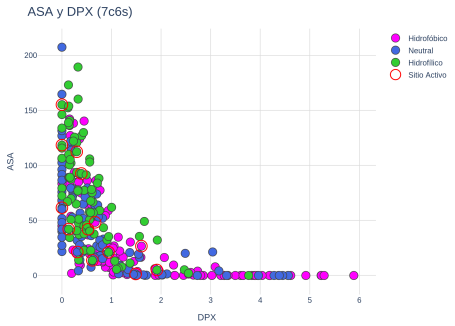
Figura 4.2: Histograma de grado de la estructura de red de carbonos C_α y C_β para las proteínas 7LG7, 7C6S y 6VYB.

4.2. Área Accesible Atómica y Profundidad Atómica

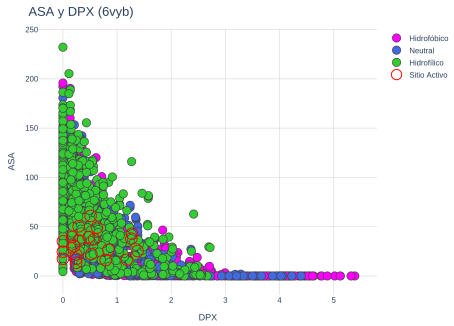
El área accesible atómica (ASA) calcula el área expuesta al solvente. En el capítulo anterior vimos que ASA se calcula al pasar una esfera (o sonda) a través de la proteína. Por defecto, el radio de esfera es de 1.4 Å, el cual es el radio de una molécula de agua ya que es el componente general del solvente [13]. Sin embargo, posee poca información sobre los átomos que están más aglutinados en el interior de la proteína, para ello se ha creado el algoritmo de profundidad atómica (DPX) para calcular la distancia entre un átomo a otro con área accesible al solvente más cercana. En la Figura 4.3 muestra las gráficas de la relación entre el ASA y DPX para cada una de las proteínas objetivo. Los círculos de color verde, azul y magenta corresponden a los aminoácidos polares, neutrales e hidrófobos, respectivamente. Los círculos con una corona roja pertenecen a los sitios activos de la proteína.



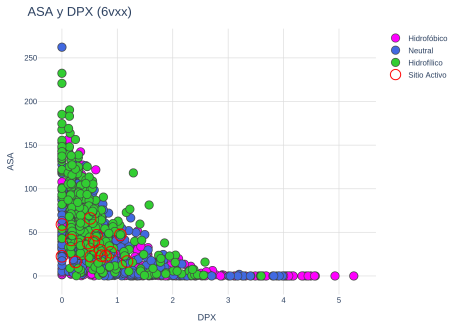
(a) Área accesible atómica para la proteína 7LG7.



(b) Área accesible atómica para la proteína 7C6S.



(c) Área accesible atómica para la proteína 6VYB.



(d) Área accesible atómica para la proteína 6VXX.

Figura 4.3: Medida del área accesible atómica para las proteínas 7LG7, 7C6S y 6VYB.

Podemos ver que en las gráficas (a) y (b) para las proteínas 7LG7 y 7C6S, respectivamente, los aminoácidos polares tienen una tendencia a un área accesible al solvente mayor mientras que los hidrofóbicos hacia una profundidad atómica más alta. En la proteína 7LG7, por ejemplo, el promedio de ASA de los de aminoácidos polares es 80.97 mientras que los hidrofóbicos es de 25.75, y los neutrales es 40.52. Los polares están más expuestos al solvente debido a que tienden a formar mejor puentes de hidrógeno e interacciones iónicas con el agua, mientras que los neutrales sólo pueden participar en formar puentes de hidrógeno con el agua [16,17]. En Cuadro 4.2 podemos ver el mismo efecto para las demás proteínas, los aminoácidos hidrofóbicos tienen un ASA promedio menor a 30 mientras que los polares dominan en el ambiente expuesto al solvente.

Por otra parte, en las cuatro proteínas la profundidad atómica promedio de los aminoácidos polares es menor de 0.22 Å, mientras que los hidrofóbicos dominan en el interior globular con un promedio mayor a 0.50 Å, véase el Cuadro 4.2. Los hidrofóbicos que tienen carácter no polar, no pueden formar puentes de hidrógeno mucho menos interacciones iónicas con el agua provocando una repulsión entre ellos y, por consiguiente, una tendencia de los aminoácidos hidrofóbicos a agruparse, lo que es conocido como efecto hidrofóbico. Esto es importante ya que juega un papel crucial tanto en el plegamiento de la proteína como en la estabilidad en el estado nativo [13].

PDB	Tipo	ASA (Å^2)	DPX (Å)
7LG7	Hidrofóbico	25.75	0.51
	polar	80.97	0.19
	Neutral	40.52	0.29
7C6S	Hidrofóbico	28.02	0.54
	polar	74.16	0.17
	Neutral	45.12	0.28
6VYB	Hidrofóbico	24.37	0.50
	polar	66.89	0.20
	Neutral	39.57	0.28
6VXX	Hidrofóbico	22.70	0.50
	polar	70.71	0.21
	Neutral	37.51	0.28

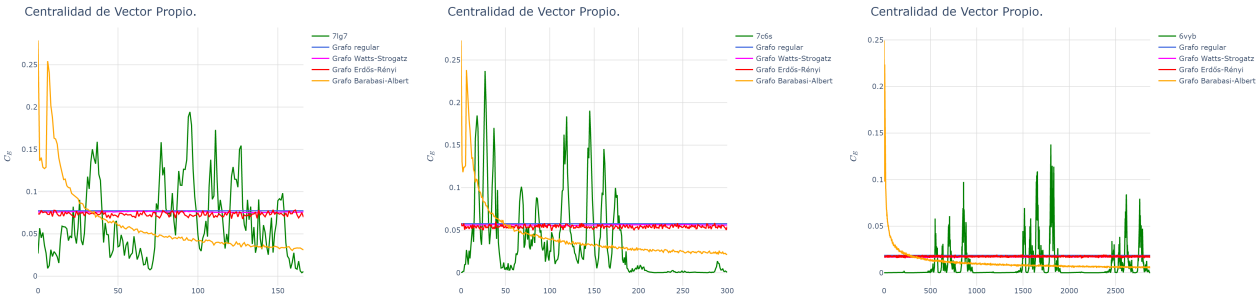
Cuadro 4.2: Promedio de ASA y DPX por tipo de aminoácido.

En las gráficas podemos observar sitios activos que no necesariamente tienen un área accesible al solvente grande. En la proteína 7C6S, gráfica (b), tenemos sólo dos sitios catalíticos que corresponde a H41 (polar) y C145 (hidrofóbico), y poseen un DPX de 0.94 Å y 1.60 Å, respectivamente, lo cual se hallan cerca del ambiente hidrofóbico que tiene un promedio de 1.69 Å. Podemos ver que hay un sitio de anclaje con mayor profundidad que es H164 con 1.92 Å. En la proteína 7LG7, gráfica (a), podemos ver sitios de anclaje más aglutinados en el interior de la proteína entre 1.2Å a 1.64Å, mientras que el promedio de profundidad atómica de los aminoácidos hidrofóbicos es 1.5Å. La proteína Spike, gráfica (c) y (d), posee tres bolsillos de anclaje que se muestra en la gráfica como sitios activos, color rojo. Recodemos que los bolsillos de anclaje son estos sitios de activos que forman una cavidad en la estructura. El DPX promedio de aminoácidos hidrofóbicos para la proteína Spike en estado abierto (6VYB) y estado cerrado (6VXX) es de 1.27Å y 1.30Å. Estos sitios activos forman parte de la sección conocida como S1 (residuos 14-685) cuya función es fusionarse con la membrana de la célula huésped [20, 21].

4.3. Centralidad de vector propio en redes de carbonos C_α y C_β

En la primera sección generamos 100 grafos aleatorios de tipo Erdős-Rényi, Watts-Strogatz, Barabási-Albert y grafos tipo regular con el número de nodos, el número de enlaces y grado promedio igual al de las proteínas objetivo del SARS-CoV-2. Posteriormente promediamos el coeficiente agrupamiento y la distancia geodésica promedio de los 100 grafos aleatorios y los comparamos con el de las proteínas. De la misma manera, evaluamos la centralidad de vector propio de cada uno de los 100 grafos aleatorios de cada tipo, y los promediamos para compararlos con las respectivas proteínas objetivo. Las comparaciones para las proteínas 7LG7, 7C6S y 6VYB se muestran en las gráficas la

Figura 4.4 (a), (b) y (c), respectivamente.



(a) Centralidad de vector propio con 167 nodos.

(b) Centralidad de vector propio con 301 nodos.

(c) Centralidad de vector propio con 2875 nodos.

Figura 4.4: Centralidad de vector propio para los modelo de grafo aleatorio tipo Erdős-Rényi, Watts-Strogatz, Barabási-Albert y grafos tipo regular.

Podemos notar que los grafos aleatorios de tipo Barabási-Albert (línea amarilla) en cada una de las gráficas (a), (b) y (c), sólo unos cuantos nodos poseen una centralidad bastante mayor, mientras que la mayoría tienen una centralidad muy baja. Esto es debido a que la centralidad de vector propio evalúa la importancia de un nodo con respecto al grado y a la calidad de sus conexiones a nodos adyacentes. Recordemos que en el modelo de Barabási-Albert cuando se agregaba un nuevo nodo al grafo, este se enlazaba a los nodos existentes de acuerdo al apego preferencial, es decir, por la probabilidad proporcional al grado de un nodo existente.

En cada gráfica de la Figura 4.4 podemos ver que la red C_α tiene una centralidad de vector propio más marcada con nodos que tienen una centralidad mucho mayor y nodos con centralidad mucho menor que están muy próximos a cero que los nodos de los grafos aleatorios Erdős-Rényi, Watts-Strogatz, Barabási-Albert y grafo tipo regular. La centralidad de vector propio en el grafo regular es el mismo para cada nodo debido a que todos tienen el mismo grado, mientras que la centralidad de los grafos tipo Erdős-Rényi y Watts-Strogatz puede verse que están más acotados alrededor de un intervalo muy estrecho. La gráfica (a) muestra que este intervalo está entre 0.067 a 0.079, en la gráfica (b) está entre 0.049 a 0.061 y, por último, para la gráfica (c) en 0.015 a 0.02.

En la Figura 4.5 muestra la centralidad de vector propio de la red de carbonos C_α para las proteínas objetivo. Los colores magenta, azul y verde representan los carbonos de aminoácidos tipo hidrofóbico, neutral e polar, respectivamente. En las gráficas podemos observar que los hidrofóbicos poseen una centralidad más alta, mientras que los polares tienden a tener menor centralidad. Para calcular la proporción de centralidad de vector propio P_E (%) para cada tipo de aminoácido, obtenemos primero la centralidad promedio para hacer una comparación entre los tres tipos de aminoácidos debido a que los aminoácidos hidrofóbicos son mayoría en las cuatro proteínas objetivo. Por ejemplo, la proteína

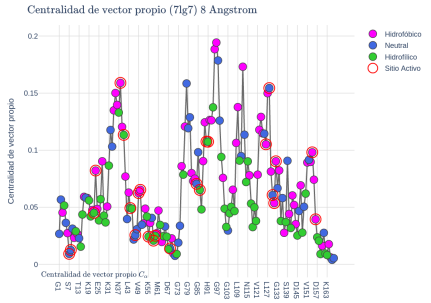
7LG7 posee el 41.3 % y 32.9 % de aminoácidos hidrofóbicos y polares, respectivamente. Mientras que los neutrales representan el 25.7 %, véase el Cuadro 1.2. De esta forma, dividimos la centralidad promedio de un tipo de aminoácidos entre la suma de los promedio de los tres tipos de aminoácidos. Es decir, la proporción de centralidad de aminoácidos hidrofóbicos está dado por

$$P_E\left(\frac{\sum_i C_E(h_i)}{N_h}\right) = 100\frac{\frac{\sum_i C_E(h_i)}{N_h}}{\frac{\sum_i C_E(h_i)}{N_h} + \frac{\sum_i C_E(p_i)}{N_p} + \frac{\sum_i C_E(n_i)}{N_n}} \tag{4.1}$$

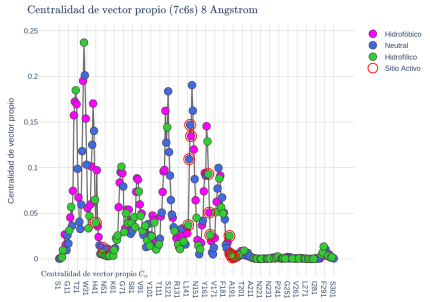
donde N_h, N_p y N_n es el número de aminoácidos tipo hidrofóbico, polar y neutral, respectivamente, y h_i es el i-ésimo aminoácido hidrofóbico. En la proteína 7LG7, la proporción de centralidad de vector propio de carbonos de aminoácidos tipo hidrofóbico representan el 41.5 %, mientras que los neutrales e polares representan el 32.7 % y 25.8 %, respectivamente; en la proteína 7C6S, la proporción es de 36.9 %, 36.7 % y 26.4 % para aminoácidos hidrófobos, neutrales y polares, respectivamente. Recordemos que la centralidad de vector propio califica de acuerdo al grado (o número de enlaces) y a la calidad de dichas conexiones. Entonces, podemos observar una mayor proporción de centralidad en los hidrofóbicos consecuencia del efecto hidrofóbico, una tendencia creciente de agruparse entre ellos evitando el agua, incluso más que otros solventes menos polares [17].

En la gráfica (b), el máximo es un Polar que corresponde a N28 que se encuentra completamente en el interior globular con un DPX igual a 3.77 Å, aproximadamente las dos terceras partes del carbono más enterrado que es C38 (hidrófobo) con 5.88 Å. Recordemos que el DPX promedio de los hidrofóbicos es 1.69, mientras que los polares es 0.56 Å, por lo tanto podemos decir que N28 se encuentra en el ambiente hidrofóbico, véase la gráfica (b) 4.3 de la sección anterior.

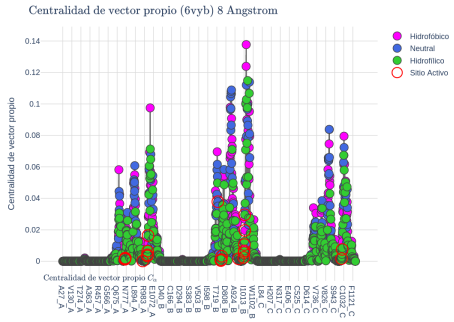
Las gráficas (c) y (d) de la Figura 4.5 representa la centralidad de vector propio para la proteína Spike en estado abierto y cerrado, respectivamente. A primera vista, podemos ver que cada cadena del estado cerrado poseen valores idénticos de centralidad debido a que cada cadena es una secuencia de aminoácidos idénticos que conforman la estructura en trímero. Por ejemplo, los tres máximos en la gráfica (c) del estado cerrado son los carbonos C_α del residuo L1034 en cada cadena. Podemos observar que los carbonos C_α con centralidad cercano a cero pertenecen a la subunidad S1 (residuos 27-685), y los C_α de la subunidad S2 (residuos 686-1142) tiene centralidad mucho más alta tanto en el estado cerrado como en el estado abierto. Recordemos que S1 se ancla al receptor de la célula huésped y S2 se encarga de fusionarse a la membrana celular. Podemos notar que los C_α de aminoácidos hidrofóbicos y neutrales poseen mayor centralidad debido al efecto hidrofóbico de los primeros que los hace agruparse entre si. En el Cuadro 4.4 muestra la proporción en porcentaje de centralidad de vector propio por tipo de aminoácido, donde los polares tienen una proporción menor que los neutrales y los hidrófobos.



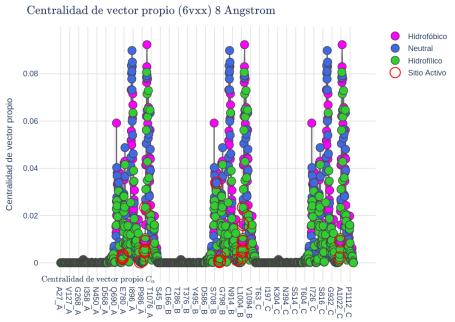
(a) Centralidad de vector propio de la red de C_α.



(b) Centralidad de vector propio de la red de C_α.



(c) Escala de centralidad de vector propio en la estructura de C_α.



(d) Escala de centralidad de vector propio en la estructura de C_α.

Figura 4.5: Centralidad de vector propio para la red de radio 8 de la proteína 7LG7. Los sitios activos están marcados con un círculo rojo en la gráfica cartesiana y en la estructura proteica.

PDB	Hidrofóbico	Polar	Neutral
7LG7	41.48	25.81	32.71
7C6S	36.91	26.36	36.73
6VYB	38.18	25.57	36.25
6VXX	38.22	25.31	36.47

Cuadro 4.3: Proporción en porcentaje de centralidad de vector propio por tipo de aminoácido.

Podemos notar fácilmente la diferencia con el estado abierto donde la cadena B posee los carbonos C_α con mayor centralidad en la misma región S1 que en el estado cerrado. Recordemos que el estado abierto se refiere a la conformación abierta del Dominio de Unión al Receptor (en inglés, Binding-Receptor Domain, o por sus siglas en inglés RBD) que destaca en la parte superior de la proteína (residuos 319-541, cadena B), véase la Figura 1.7. Aunque la región de RBD tanto en el estado abierto como cerrado posee una centralidad cercano a cero, podemos ver que influye en toda la red. La centralidad de vector propio de la conformación abierta de RBD, PDB 6VYB, en promedio representa el 11 % de la centralidad promedio del estado cerrado, PDB 6VXX. Esto es debido a que en esta región al estar más separada del resto, el grado de los C_α es menor.

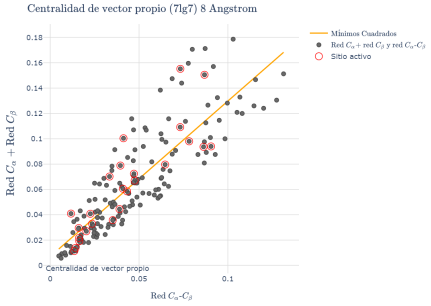
En esta sección vimos la centralidad de vector propio para la red C_α. Para las siguientes gráficas de la Figura 4.6 sumamos la centralidad de vector propio de la red C_α con la centralidad de vector propio de la red C_β y la suma lo dividimos entre dos, es decir,

$$\frac{C_E(\alpha_i) + C_E(\beta_i)}{2}$$

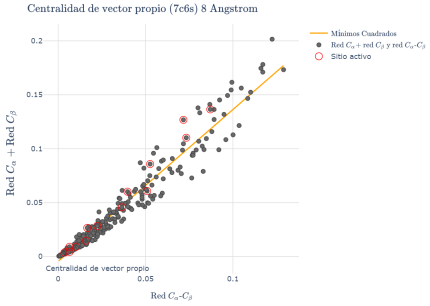
donde α_i y β_i son los carbonos del aminoácido i de la red C_α y la red C_β, respectivamente. El resultado lo comparamos con la centralidad de la red $C_\alpha - C_\beta$, el cual igualmente sumamos la centralidad de los C_α con C_β y lo dividimos entre dos, es decir

$$\frac{C_E(\alpha_i) + C_E(\beta_i)}{2}$$

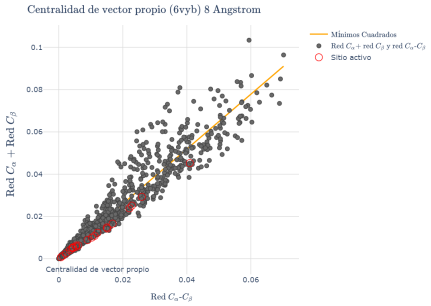
donde α_i y β_i son los carbonos del aminoácido i en la red $C_\alpha - C_\beta$.



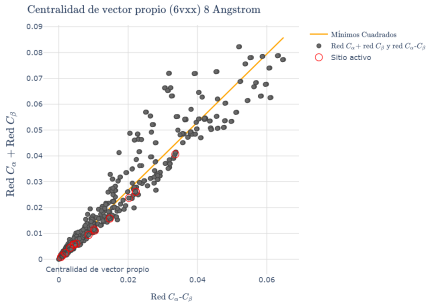
(a) Centralidad de vector propio para la proteína 7LG7.



(b) Centralidad de vector propio para la proteína 7C6S.



(c) Centralidad de vector propio para la proteína 6VYB.



(d) Centralidad de vector propio para la proteína 6VXX.

Figura 4.6: Correlación lineal entre la red $C_\alpha - C_\beta$ y la suma de la red C_α y la red C_β.

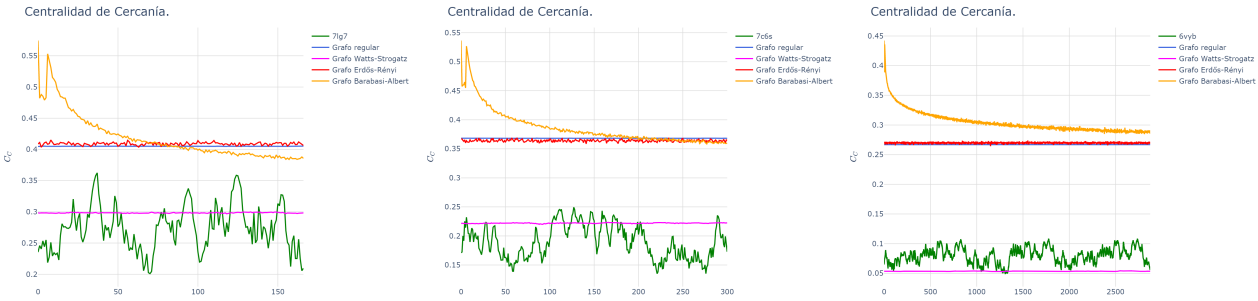
Proteína	m	c	CCP
7LG7	1.2278	0.007	0.8614
7C6S	1.4084	-0.0045	0.9782
6VYB	1.2963	0.0001	0.9742
6VXX	1.3187	0.0003	0.9739

Cuadro 4.4: Solución del método de mínimos cuadrados lineales donde m es la pendiente, c es la ordenada al origen y CCP es el coeficiente de correlación de Pearson. Centralidad de vector propio.

En cada una de las gráficas de la Figura 4.6 los datos muestran que hay cierta relación lineal entre los datos. Por consiguiente, evaluamos el coeficiente de correlación de Pearson (CCP) para medir el grado de correlación lineal entre los dos conjuntos de datos. La medida varía entre -1 a 1, donde cada extremo representa una relación lineal perfecta. Si el coeficiente es -1, los valores de una variable aumentan mientras que la otra disminuye; y si es 1, ambas variables aumentan. En la cuarta columna del Cuadro 4.4 muestra el CCP para cada proteína objetivo, el cual podemos ver que hay una relación fuerte de linealidad. La proteína más chica, el 7LG7, tiene un CCP menor lo cual muestra una mayor discrepancia de centralidad entre las redes. Recordemos que las proteínas están formadas principalmente por estructuras de hélice y hojas plegadas. Los carbonos C_β en las hélices están extendidas hacia fuera, mientras que en la estructura de hoja plegada están extendidas en direcciones opuestas entre aminoácidos contiguos. Esto puede tener mayor influencia en la red de carbonos C_α-C_β que en las redes por separado.

Usamos el método de mínimos cuadrados para encontrar una recta que mejor se ajustase a los datos, la cual se muestra en color naranja. En el Cuadro 4.4, se registra la pendiente m y la ordenada al origen c de las rectas. La pendiente para cada proteína es mayor a 1 por lo cual muestra una mayor puntuación en la suma de la red C_α y la red C_β.

4.4. Centralidad de cercanía en redes de carbonos C_α y C_β



(a) Modelo Erdős-Rényi con 167 nodos.

(b) Modelo Erdős-Rényi con 301 nodos.

(c) Modelo Erdős-Rényi con 2875 nodos.

Figura 4.7: Centralidad de cercanía para los modelo de grafo aleatorio tipo Erdős-Rényi, Watts-Strogatz y Barabási-Albert.

En la tabla 4.1 de la primera sección muestra que el promedio de las distancias geodésica l de los carbonos C_α es mayor al promedio de las distancias geodésicas de los 100 grafos aleatorios del tipo Erdős-Rényi, Watts-Strogatz y grafos tipo regular cada uno. Recordemos que la distancia geodésica es el número de aristas de la ruta más corta entre

un par de nodos.

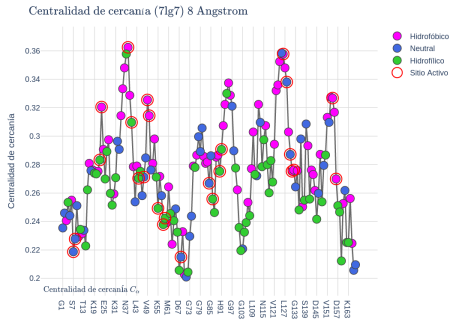
En la Figura 4.7 (a), (b) y (c) muestra la centralidad de cercanía de la red de carbonos C_α de las proteínas objetivo y la centralidad de cercanía promedio de los 100 grafos aleatorios del tipo Erdős-Rényi, Watts-Strogatz, Barabási-Albert y grafos tipo regular. Podemos ver que las proteínas del SARS-CoV-2 poseen una centralidad de cercanía mucho menor junto con los grafos tipo Watts-Strogatz que al resto de los grafos aleatorios. Entre la gráficas (a), (b) y (c) muestran una centralidad de cercanía con una separación de menor a mayor con los grafos aleatorios Erdős-Rényi, Barabási-Albert y grafos tipo regular. En (b) y (c) esta separación es más de 0.1.

Podemos observar que la diferencia entre las medidas de centralidad de las proteínas objetivo y el de los grafos tipo Watts-Strogatz son menores al resto de los grafos aleatorios corresponde al promedio de las distancias geodésicas visto en la tabla 4.1. En la gráfica (c) podemos observar que la centralidad de cercanía del grafo aleatorio tipo Watts-Strogatz es mucho menor que la mayoría de los nodos de la proteína Spike. Este hecho corresponde a que la distancia geodésica promedio del grafo aleatorio Watts-Strogatz es de 18.73 mientras que el de la proteína Spike fue menor en 12.86. La correspondencia es debido a que la centralidad de cercanía es el inverso de la distancia geodésica promedio.

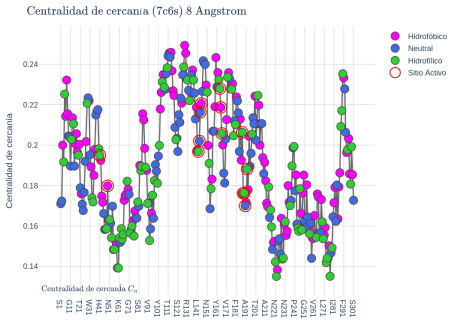
En la Figura 4.8 muestra las gráficas de centralidad de cercanía para cada proteína objetivo, y el tipo de aminoácido al que pertenece: magenta, azul y verde para los hidrofóbicos, neutrales e polares, respectivamente. Podemos observar en las gráficas que los hidrofóbicos ocupan mayormente los picos. Sin embargo, la proporción de centralidad entre los hidrofóbicos es ligeramente mayor que entre los polares y neutrales, véase la tabla 4.5.

Podemos observar que en el 10 % de los carbonos con centralidad mayor, los sitios activos ocupan una centralidad alta, principalmente si son hidrofóbicos. En la proteína 7LG7, gráfica (a), tenemos 4 sitios activos donde tres son hidrófobos y uno neutral. La proteína Spike en estado abierto y estado cerrado, gráfica (c) y (d), poseen 21 sitios activos, los cuales 10 son hidrófobos, 9 neutrales y 7 polares. Sin embargo, la proteína 7C6S, gráfica (b), que posee en total solo 3 sitios activos hidrofóbicos, 3 neutrales y 8 polares, tenemos que en el 20 % de los carbonos con centralidad mayor uno es polar y un hidrófobo.

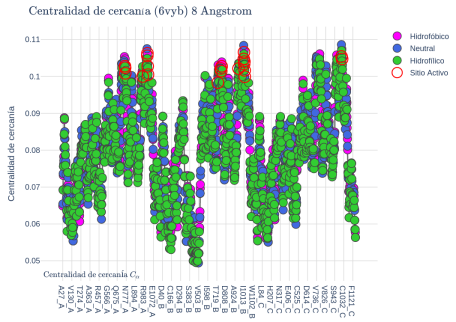
En (c) y (d) tenemos las gráficas de centralidad de cercanía para la proteína Spike en estado abierto y cerrado, respectivamente. La subunidad S2 (residuos 686-1142 de cada cadena) poseen los carbonos con mayor centralidad entre los que se encuentran los sitios de activos cuya función permitirá fusionarse con la membrana celular del huésped. Estos carbonos C_α poseen rutas geodésicas más cortas al resto de la red y, por lo tanto, mejor



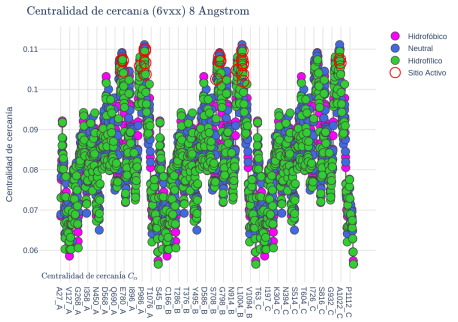
(a) Centralidad de cercanía de la red de C_α.



(b) Centralidad de cercanía de la red de C_α.



(c) Centralidad de cercanía de la red de C_α.



(d) Centralidad de cercanía de la red de C_α.

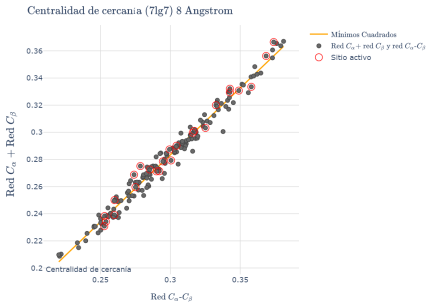
Figura 4.8: Centralidad de cercanía para la red de radio 8 de la proteína 7LG7. Los sitios activos están marcados con un círculo rojo en la gráfica cartesiana y en la estructura proteica.

conexión al resto de la red.

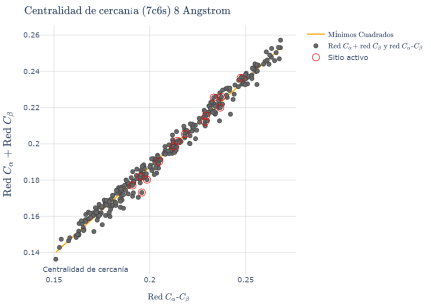
Las gráficas en ambas conformaciones de la proteína Spike se observa cierta similaridad, entre de 3 % y 7 % disminuyó la centralidad en la conformación abierta. Sin embargo, la centralidad de cercanía promedio de RBD en estado abierto a diferencia del estado cerrado disminuyó en 33 % debido a la separación del RBD como vimos en la sección anterior, por lo cual las rutas geodésicas fue afectado por la perdida de conexiones entre los C_α del RBD con carbonos circunstantes.

La proporción de centralidad de cercanía entre los C_α de aminoácidos hidrofóbicos, polares y neutrales es similar en ambas conformaciones de la proteína Spike. Sin embargo, puede observarse ligeramente una proporción mayor en hidrofóbicos.

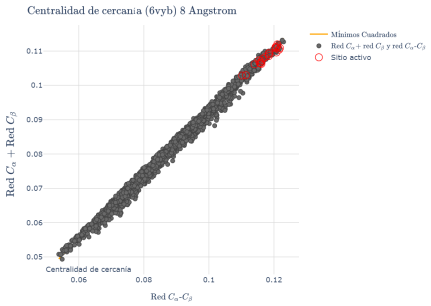
En la sección anterior sumamos la centralidad de vector propio de los nodos de la red C_α con los nodos de la red C_β, el resultado lo dividimos entre dos y lo comparamos con los nodos respectivos de la red $C_\alpha - C_\beta$. En la Figura 4.6 se muestra el ajuste lineal a los datos para la centralidad de cercanía. El cuadro 4.6 muestra la pendiente m y la ordenada al origen c de la recta, y el coeficiente de correlación de Pearson (CCP). Cada proteína



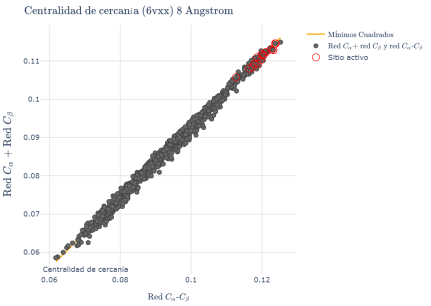
(a) Centralidad de cercanía para la proteína 7LG7.



(b) Centralidad de cercanía para la proteína 7C6S.



(c) Centralidad de cercanía para la proteína 6VYB.



(d) Centralidad de cercanía para la proteína 6VXX.

Figura 4.9: Correlación lineal entre la red $C_\alpha - C_\beta$ y la suma de la red C_α y la red C_β.

PDB	Hidrofóbico	Polar	Neutral
7LG7	35.43	31.77	32.79
7C6S	33.95	32.33	33.72
6VYB	33.78	33.07	33.15
6VXX	33.70	33.04	33.26

Cuadro 4.5: Proporción en porcentaje de centralidad de cercanía por tipo de aminoácido.

tiene una correlación casi perfecta cercano a 1. Si comparamos el coeficiente de correlación de la centralidad de vector propio, Cuadro 4.4, podemos observar que es mayor el CCP de la centralidad de cercanía. Recordemos que la centralidad de vector propio evalúa la importancia de un nodo según su grado y la calidad de los nodos adyacentes, por lo cual es una medida que depende más de la localidad de un nodo. Mientras que la centralidad de cercanía evalúa la importancia de un nodo según el promedio de las distancias geodésicas a los demás nodos, es decir, es una medida más general.

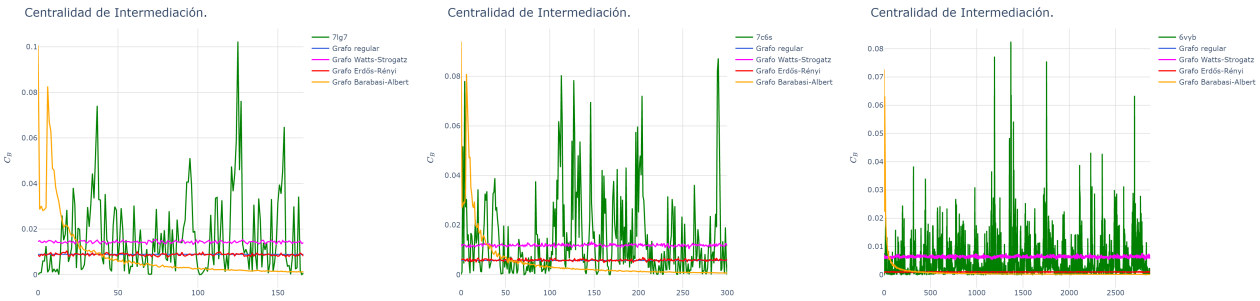
La pendiente de la recta que mejor ajusta a los datos en cada proteína muestra un valor casi cercano a 1, el cual muestra que el aumento de los valores de la centralidad es casi de la misma proporción entre las dos variables. Las gráficas (c) y (d) de la proteína Spike en sus dos estados muestra que los sitios activos (en circulo rojo) están entre los carbonos con mayor cercanía al resto de la red.

Proteína	m	c	CCP
7LG7	0.9871	-0.0128	0.9900
7C6S	0.9616	-0.005	0.9934
6VYB	0.9303	-0.0006	0.9963
6VXX	0.9244	0.0003	0.9961

Cuadro 4.6: Solución del método de mínimos cuadrados lineales donde m es la pendiente, c es la ordenada al origen y CCP es el coeficiente de correlación de Pearson. Centralidad de cercanía.

4.5. Centralidad de intermediación en redes de carbonos C_α y C_β

En la Figura 4.10 (a), (b) y (c) muestra la centralidad de intermediación de la red de carbonos C_α de las proteínas objetivo y la centralidad de cercanía promedio de los 100 grafos aleatorios del tipo Erdős-Rényi, Watts-Strogatz, Barabási-Albert y grafos tipo regular. En cada gráfica ningún grafo aleatorio tiene algún nodo con centralidad cero, es decir, que todos los nodos se encuentran por lo menos en una ruta geodésica de un par de nodos en la red. Por otro lado, en la red de carbonos C_α de las proteínas si existen nodos con centralidad igual a cero. Estos carbonos son los que tienen grado igual a 1, es decir, que solo hay un carbono en un radio de interacción $R = 8$Å. Por lo cual, estos carbonos pertenecen en las partes más exteriores de la superficie globular.



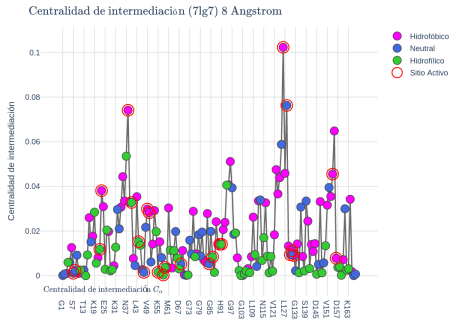
(a) Modelo Erdős-Rényi con 167 nodos.

(b) Modelo Erdős-Rényi con 301 nodos.

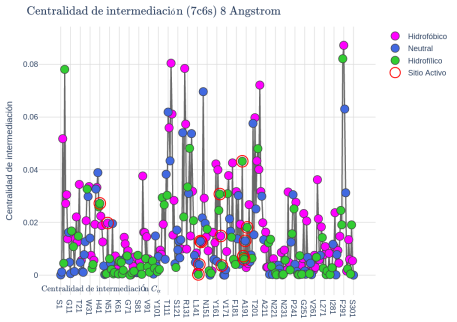
(c) Modelo Erdős-Rényi con 2875 nodos.

Figura 4.10: Centralidad de intermediación para los modelo de grafo aleatorio tipo Erdős-Rényi, Watts-Strogatz y Barabási-Albert.

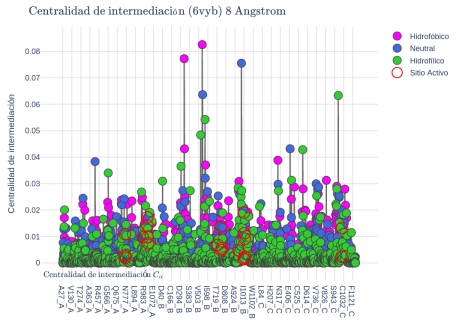
Podemos notar que los grafos aleatorios tipo Erdős-Rényi, Watts-Strogatz y regulares tienen una centralidad más acotadas al igual que las medidas anteriores a comparación de los grafos tipo Barabási-Albert y el de las proteínas del SARS-CoV-2. Las proteínas no solo tienen nodos con centralidad cero sino que también nodos con centralidad muy alta. En la gráfica (c) podemos notar que hay tres nodos con una centralidad que superan a los nodos del grafo Barabási-Albert.



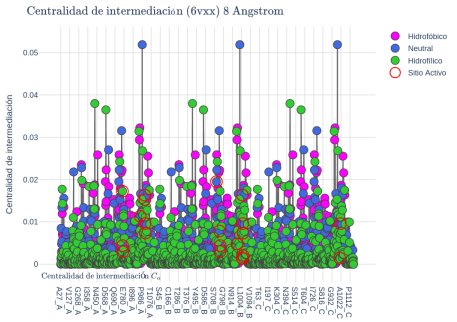
(a) Centralidad de intermediación de la red de C_α.



(b) Centralidad de intermediación de la red de C_α.



(c) Centralidad de intermediación de la red de C_α.



(d) Centralidad de intermediación de la red de C_α.

Figura 4.11: Centralidad de intermediación para la red de radio 8 de la proteína 7LG7. Los sitios activos están marcados con un círculo rojo en la gráfica cartesiana y en la estructura proteica.

La gráfica de la centralidad de intermediación de la proteína 7LG7, Figura 4.11 (a), muestra tres carbonos con valores máximos que son los sitios activos: L126 (hidrófobo), S128 (neutro) y A38 (hidrófobo), esto quiere decir que son los que se hallan más frecuentemente en las distancias geodésicas entre cada par de carbonos en la red. Sin embargo, también resulta que poseen en promedio las distancias geodésicas más chicas en la red, es decir, una centralidad de cercanía mayor, Figura 4.8 (a). En la centralidad de vector propio, A38 y S128 son los sitios activos con mayor puntuación, y L126 es el sexto, Figura 4.5 (a). Los tres sitios activos A38, L126 y S128 tienen un DPX cercano al promedio de los aminoácidos hidrófobos (1.5Å).

La centralidad de intermediación de la proteína 7C6S destaca al sitio de anclaje H164 (Polar) y a H41, uno de los dos sitios catalíticos. Por otra parte, H164, S128 y, por último, el sitio catalítico C145 poseen una alta puntuación en la centralidad de vector propio como la centralidad de cercanía. Es importante mencionar que estos tres sitios activos se hallan entre el DPX promedio de los aminoácidos hidrofóbicos (1.6Å).

Podemos notar que los aminoácidos hidrofóbicos (color magenta) tienden a encontrarse entre las rutas geodésicas de cada par de carbonos en la red para las cuatro proteínas.

PDB	Hidrofóbico	Polar	Neutral
7LG7	49.89	18.50	31.62
7C6S	42.49	26.88	30.62
6VYB	40.79	27.93	31.29
6VXX	39.80	28.52	31.68

Cuadro 4.7: Proporción en porcentaje de centralidad de intermediación por tipo de aminoácido.

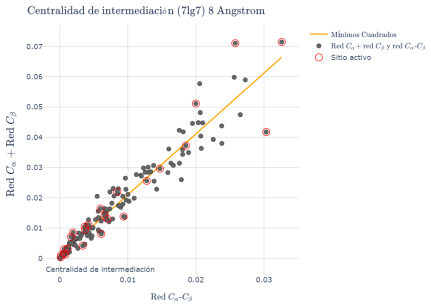
Recordemos que los hidrofóbicos tienden a agruparse formando una especie de núcleo durante el plegamiento de la proteína por el efecto hidrofóbico, mientras que los demás aminoácidos especialmente los polares dominan en la superficie globular ya que pueden formar fácilmente enlaces de hidrógeno y enlaces iónicos con las moléculas de agua. En la tabla 4.7 muestra la proporción de centralidad de intermediación por tipo de aminoácido para cada proteína objetivo.

La proteína Spike en estado cerrado, gráfica 4.11 (d), vuelve a mostrar la misma similaridad de centralidad entre las cadenas debido a la conformación en trímero de secuencias de aminoácidos idénticos. Sin embargo, con la conformación abierta del RBD de la proteína Spike, gráfica 4.11 (c), no solo influyó en la red total sino que favoreció la centralidad de carbonos C_α de la cadena B (en la misma cadena en la que se encuentra el RBD) igual que en la centralidad de vector propio. La centralidad de intermediación de la cadena B y C aumentó en un 20 % y 8.4 % de la conformación abierta, respectivamente. Mientras que en la cadena A disminuyó en 6.7 %. En la centralidad de vector propio, la cadena B aumentó en 33 %, específicamente de la subunidad S2. Mientras que las cadenas A y C disminuyeron en 33 %. Podemos ver que la conformación abierta del RBD, ubicada en la subunidad S1 de la cadena B, también implica que se mueva toda la cadena principalmente, afectando la centralidad de toda la red de C_α de la proteína.

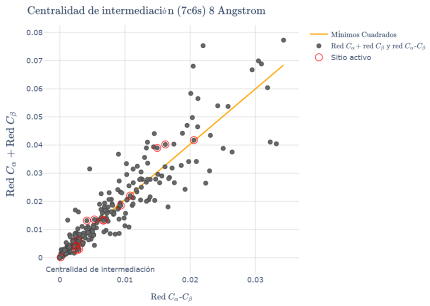
La siguiente Figura 4.12 muestra la relación entre la centralidad de intermediación de la red C_α-C_β con la suma de las centralidades de la red C_α con la red C_β. La línea naranjada de cada gráfica corresponde al ajuste lineal por el método de mínimos cuadrados. En el Cuadro 4.8, muestra la pendiente m y la ordenada al origen c de la línea, así como también el coeficiente de correlación lineal de Pearson para cada una de las proteínas objetivo.

La proteína 7LG7 muestra un mayor coeficiente de correlación con 0.9675 a comparación de las demás proteínas. La pendiente de la recta que mejor se ajustó a los datos es mayor a 2 por lo cual muestra una menor centralidad de intermediación en la red C_α-C_β. Esto muestra que al integrar ambos tipos de carbonos C_α y C_β que están a 1.5 Å de distancia afecta la centralidad entre ellos debido a que las rutas geodésicas que pudieron pasar por uno de ellos ahora pueden hacerlo también por el carbono contiguo. Este mismo efecto sucede con las demás proteínas con una pendiente aproximada a 2.

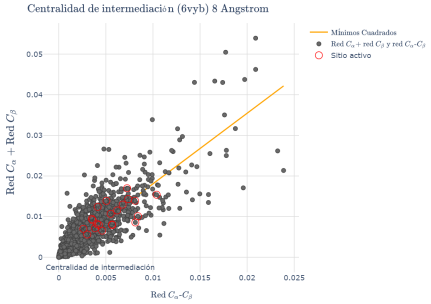
Podemos ver que la correlación de la proteína Spike en estado abierto y estado cerrado es menor que las otras dos proteínas, lo cual corresponde a la forma en que se dispersan los datos en las gráficas.



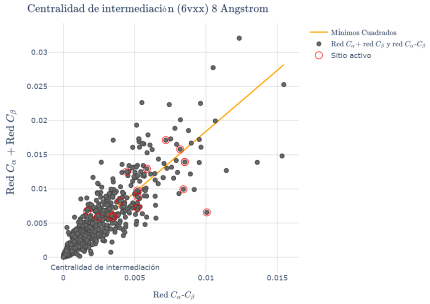
(a) Centralidad de intermediación para la proteína 7LG7.



(b) Centralidad de intermediación para la proteína 7C6S.



(c) Centralidad de intermediación para la proteína 6VYB.



(d) Centralidad de intermediación para la proteína 6VXX.

Figura 4.12: Correlación lineal entre la red $C_\alpha - C_\beta$ y la suma de la red C_α y la red C_β.

Proteína	m	c	CCP
7LG7	2.0107	0.0009	0.9675
7C6S	1.9607	0.0011	0.9283
6VYB	1.7369	0.0007	0.8743
6VXX	1.7878	0.0006	0.8862

Cuadro 4.8: Solución del método de mínimos cuadrados lineales donde m es la pendiente, c es la ordenada al origen y CCP es el coeficiente de correlación de Pearson. Centralidad de intermediación.

Capítulo 5

Bloqueo de sitios activos y el efecto en la centralidad.

En este capitulo observaremos el comportamiento en las medidas de centralidad para la red de carbonos C_α al efectuar el bloqueo en los sitios activos de las proteínas. Calculamos la diferencia de relativa (DR, %) entre la centralidad de la red C_α de la proteína en estado salvaje y la centralidad en estado mutante (bloqueando uno de los sitios activos), y lo comparamos con la centralidad del sitio activo bloqueado del estado salvaje. No efectuamos valor absoluto en la diferencia ya que nos muestra cuales C_α fueron afectados o cuales fueron beneficiados por dicho bloqueo.

Los sitios activos que bloqueamos para la proteína 7LG7 son los carbonos de los aminoácidos A38 (hidrofóbico), L126 (hidrofóbico) y S128 (neutral) los cuales son los que tienen mayor centralidad entre los sitios activos en cada medida de centralidad, incluso representaron ser los tres máximos en la centralidad de intermediación y la centralidad de cercanía (S128 fue el octavo máximo). Por esta razón, los tres C_α representan un objetivo central y de relevancia para el bloqueo. Para la proteína 7C6S bloquearemos los dos sitios catalíticos H41 (polar) y C145 (hidrofóbico), un sitio de anclaje H164, los cuales los últimos dos poseen una puntuación alta en cada una de las medidas de centralidad. La proteína Spike S en estado abierto y estado cerrado bloquearemos tres sitios de anclaje N764, A766 y T768 identificados como uno de los pocket de anclaje potenciales para el desarrollo de fármacos anticoronavirus [21].

5.1. Centralidad de vector propio en redes de proteínas mutantes.

En el capitulo anterior calculamos la centralidad de vector propio en una red de carbonos de una proteína tipo salvaje, es decir, tal como se encuentra en la naturaleza. Para evaluar el efecto de un bloqueo usaremos la *diferencia relativa* (DR) definida como la

diferencia entre la centralidad de vector propio del carbono C_α en la red de la proteína tipo salvaje y el carbono C_α en la red tipo mutante y lo compararemos con la centralidad del carbono bloqueado en el estado salvaje. Esto es,

$$DR(\alpha_i) = \frac{C_E(\alpha_i) - C'_E(\alpha_i)}{C_E(\alpha_k)} \tag{5.1}$$

donde $C'_E(\alpha_i)$ es la centralidad del carbono C_α del aminoácido i en la red mutante, y $C_E(\alpha_k)$ es la centralidad del sitio activo k en la red del estado salvaje. Si $C_E(\alpha_i) - C'_E(\alpha_i) > 0$, la importancia, o centralidad, en la red del carbono α_i fue mermado en la red al bloquear α_k. En cambio, si $C_E(\alpha_i) - C'_E(\alpha_i) < 0$ la importancia de α_i en la red es incrementada al bloquear el carbono α_k. Esto nos permite saber qué átomos de carbono y sitios activos fueron afectados o beneficiados en la red.

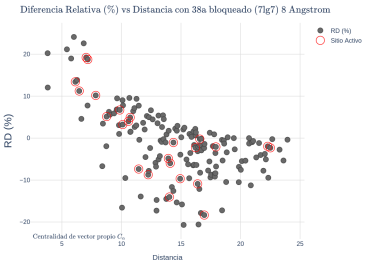
En la Figura 5.1 (a) muestra la relación entre la DR (%) y la distancia al carbono bloqueado A38 en la red C_α de la proteína mutante 7LG7, donde los círculos en rojos están marcados los sitios activos. Podemos observar que los carbonos más afectados se encuentran dentro del radio de interacción de A38, es decir, 8Å de distancia cartesiana. Esto es debido a que la centralidad de un nodo es proporcional al valor de sus nodos adyacentes, por lo cual no cabe de esperar que fueran mayormente afectados. Los sitios activos que fueron afectados dentro del radio de interacción en los tres bloqueos pertenecen junto con éstos al bolsillo de anclaje conocido como distal de ribosa y adenosina, un objetivo clave en el diseño de fármacos.

Las gráficas de la Figura 5.1 (a), (b) y (c) que corresponden a cada uno de los bloqueos, tenemos que a una distancia mayor al radio de interacción podemos observar carbonos con un porcentaje negativo, es decir, que fueron más beneficiados por el bloqueo. Estos carbonos que incrementaron su importancia en la red están en un radio entre 15Å a 25Å de distancia cartesiana donde se encuentran los sitios activos que no pertenecen al bolsillo de anclaje distal de ribosa y adenosina.

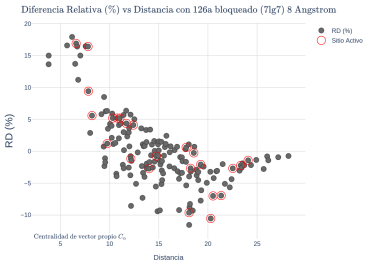
El bloqueo de S128, el cual tiene un DPX igual a 1.26 Å, tuvo un mayor efecto en toda la red tanto en los carbonos alrededor del radio de interacción como de los carbonos que fueron beneficiados debido a que se encuentra próximo al ambiente hidrofóbico, recuerde que los hidrofóbicos tienen un DPX promedio de 1.54 Å, y poseen una centralidad alta por su creciente tendencia a agruparse. En la proteína 7C6S, Figura 5.1 (i), el bloqueo que tuvo mayor efecto fue el carbono del aminoácido H164 ya que su DPX es 1.92 Å. Además, la Figura 5.1 (l) muestra que la mayoría de los sitios activos mermaron su importancia en la red, entre los cuales se encuentra la diada catalítica H41-C145. En cambio, en los bloqueos de H41 y C145 fueron solo sitios de anclaje los que empeoraron su centralidad.

En la Figura 5.2 muestra la DR para cada uno de los bloqueos en la proteína Spike

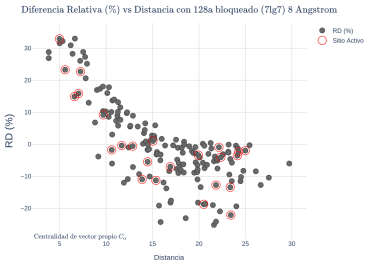
en sus dos estados estructurales, donde los C_α de la cadena A (color verde) están más afectados que las otras cadenas. Sin embargo, podemos ver carbonos de la cadena B y C (color azul y magenta, respectivamente) que tienen la misma distancia del carbono bloqueado que los de la cadena A pero con menor afectación debido a que los carbonos de una misma cadena están más agrupados, o juntos. Recuerde que la distancia promedio entre C_α en la secuencia PDB consecutivos es aproximadamente 4Å. Mientras que la distancia entre el carbono bloqueado al carbono más cercano de la cadena C y B que son Q314 y T1009 tienen una distancia de 9.8Å y 19.6Å, respectivamente. Entonces, podemos ver que la cadena C (color magenta) es la más afectada a pesar de que tenga carbonos a la misma distancia que los respectivos de la cadena B (color azul). En consecuencia, podemos ver que los sitios activos más afectados pertenecen a la cadena A, especialmente los que están más cercanos al sitio activo bloqueado. Aunque podemos notar que en cada gráfica hay carbonos con DR igual a cero, pero que en realidad tienen un cierto porcentaje de afectación. Si un carbono es bloqueado, el efecto se propaga a través de una la red conectada ya que la centralidad de vector propio de un nodo es proporcional al valor de sus nodos adyacentes. De tal forma que afectará a los nodos adyacentes y los nodos que sean adyacentes a éstos.



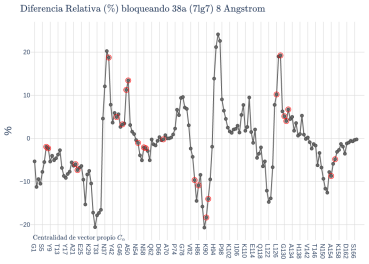
(a) Diferencia relativa de la red de C_α al bloquear el residuo A38.



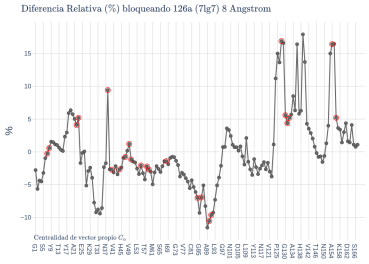
(b) Diferencia relativa de la red de C_α al bloquear el residuo L126.



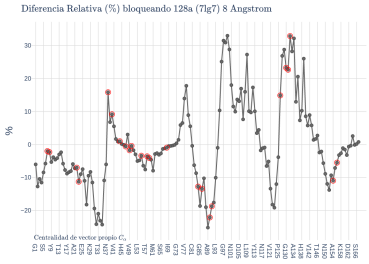
(c) Diferencia relativa de la red de C_α al bloquear el residuo S128.



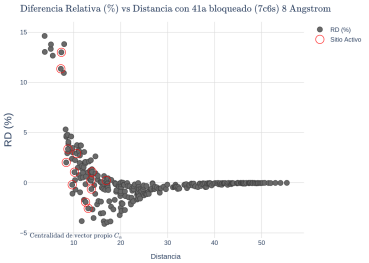
(d) Diferencia relativa de la red de C_α al bloquear el residuo A38.



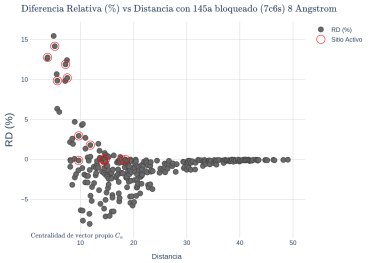
(e) Diferencia relativa de la red de C_α al bloquear el residuo L126.



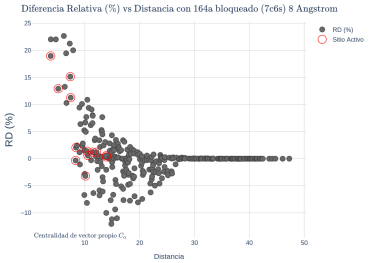
(f) Diferencia relativa de la red de C_α al bloquear el residuo S128.



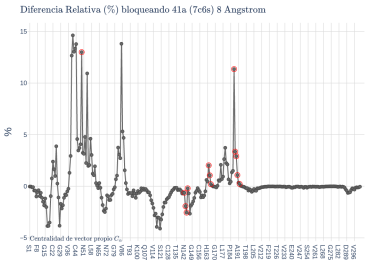
(g) Diferencia relativa de la red de C_α al bloquear el residuo H41.



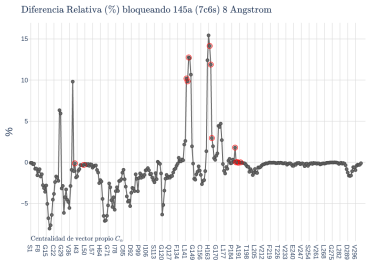
(h) Diferencia relativa de la red de C_α al bloquear el residuo C145.



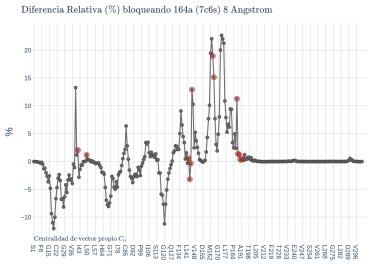
(i) Diferencia relativa de la red de C_α al bloquear el residuo H164.



(j) Diferencia relativa de la red de C_α al bloquear el residuo H41.

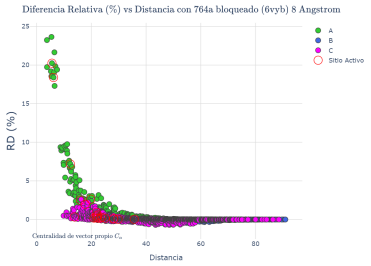


(k) Diferencia relativa de la red de C_α al bloquear el residuo C145.

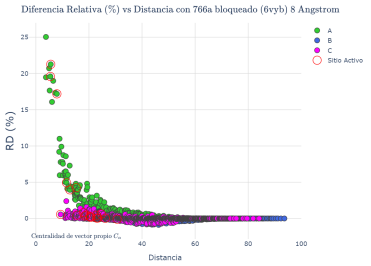


(l) Diferencia relativa de la red de C_α al bloquear el residuo H164.

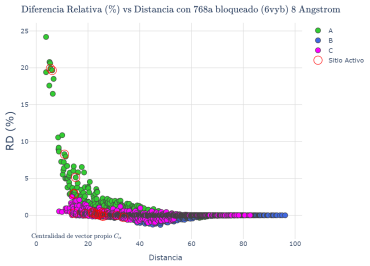
Figura 5.1: Diferencia relativa para la centralidad de vector propio para cada uno de los tres sitios activos bloqueados.



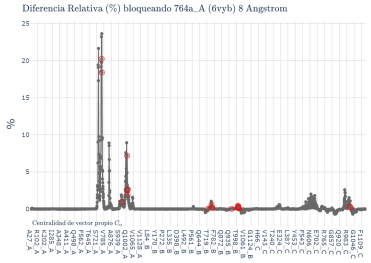
(a) Diferencia relativa de la red de C_α al bloquear el residuo N764.



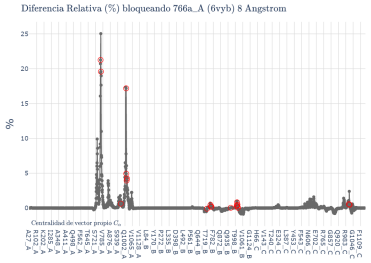
(b) Diferencia relativa de la red de C_α al bloquear el residuo A766.



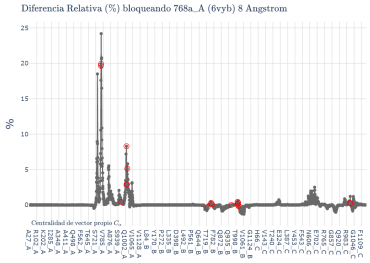
(c) Diferencia relativa de la red de C_α al bloquear el residuo T768.



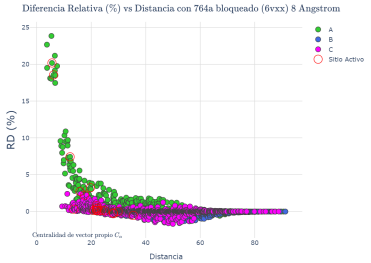
(d) Diferencia relativa de la red de C_α al bloquear el residuo N764.



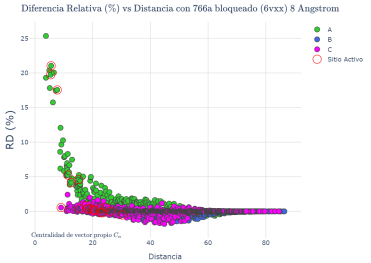
(e) Diferencia relativa de la red de C_α al bloquear el residuo A766.



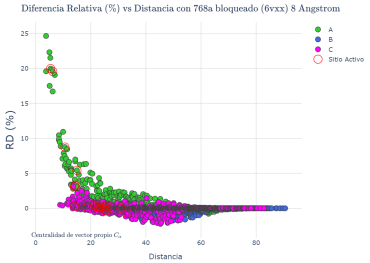
(f) Diferencia relativa de la red de C_α al bloquear el residuo T768.



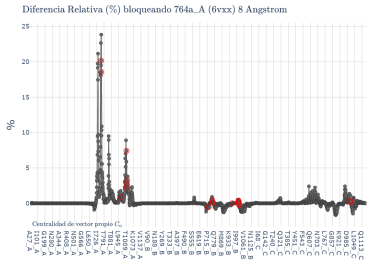
(g) Diferencia relativa de la red de C_α al bloquear el residuo N764.



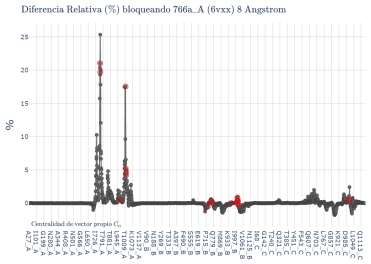
(h) Diferencia relativa de la red de C_α al bloquear el residuo A766.



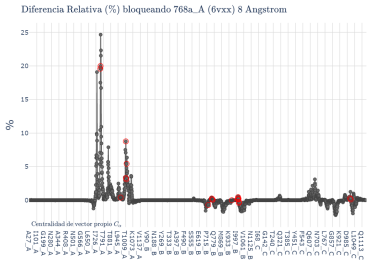
(i) Diferencia relativa de la red de C_α al bloquear el residuo T768.



(j) Diferencia relativa de la red de C_α al bloquear el residuo N764.



(k) Diferencia relativa de la red de C_α al bloquear el residuo A766.



(l) Diferencia relativa de la red de C_α al bloquear el residuo T768.

Figura 5.2: Diferencia relativa para la centralidad de vector propio para cada uno de los tres sitios activos bloqueados.

5.2. Centralidad de cercanía en redes de proteínas mutantes.

La centralidad de cercanía a diferencia de la centralidad de vector propio no hay carbonos que sean favorecidos a través de la red debido a que el bloqueo de un carbono indica que no pueda usarse entre las rutas geodésicas de los demás carbonos. Si un carbono A está entre la ruta geodésica de dos carbonos, digamos B y C, pero lo bloqueamos de la red. Entonces, la distancia geodésica de los carbonos B y C tendría que ocupar otra ruta más larga o puede ser que halla una ruta con la misma distancia. Sin embargo, para cada proteína objetivo la red de carbonos fue afectada en cada bloqueo por lo menos algún porcentaje distinto de 0 %, es decir, que todos los carbonos de toda la red tuvo que ocupar una ruta más larga.

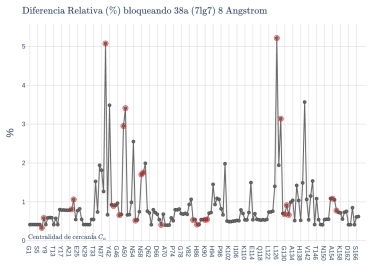
En las gráficas (a), (b) y (c) de la Figura 5.3 corresponden al DR de los bloqueos a los sitios activos A38, L126 y S128, respectivamente. Los más afectados al bloquear A38 que están por arriba de 2.5 % son sitios activos que pertenecen al bolsillo de anclaje distal de ribosa y adenosina. En (b) y (c) muestra que los sitios activos más afectados pertenecen al mismo bolsillo de anclaje por arriba de 2 % en diferentes sitios activos. Sin embargo, los carbonos que no son sitios activos pero que fueron mayormente afectados son de igual importancia ya que forman parte del soporte estructural que mantiene a los sitios activos, y un daño en la estructura podría afectar críticamente el funcionamiento biológico de la proteína, incluso su destrucción [14].

La diferencia relativa de los bloqueos de H41, C145 y H164 de la proteína 7C6S se muestran en la Figura 5.3 (g), (h) y (i), respectivamente. Podemos notar que los carbonos más afectados son sitios de anclaje que están a una distancia mayor a 7 Å, véase la Figura 5.3 (j), (k) y (l).

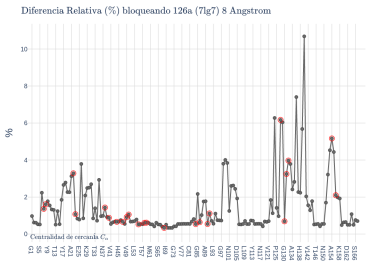
Los bloqueos en las proteínas Spike correspondiente a sus dos estados estructurales, cada carbono en la red posee un DR distinto de cero a pesar de tener carbonos con distancias de hasta 80 Å. En las gráficas de DR de la Figura 5.4 es notable que los carbonos más afectados pertenecen a la cadena A de la subunidad S2 seguido por los carbonos de la cadena C de la subunidad S2 debido a su proximidad. Sin embargo, al bloquear A766 de la cadena A, los carbonos de la subunidad S1 de la cadena B es más afectada que las respectivas subunidades de las cadenas A y C que están más cercanas al carbono bloqueado.

Los bloqueos en cada uno de las proteínas, podemos notar que a diferencia de la centralidad de vector propio, los carbonos adyacentes no son los más afectados sino los que están a una distancia alrededor de 7 Å o mayor al radio de interacción, después disminuye

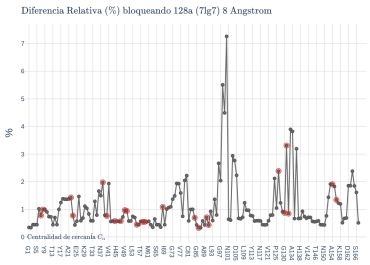
conforme aumenta la distancia. Esto es debido a que los carbonos más próximos al sitio activo bloqueado, en un radio entre 4 Å y 6 Å, poseen casi los mismos carbonos adyacentes. Entonces, estos nodos suelen aparecer entre las rutas geodésicas.



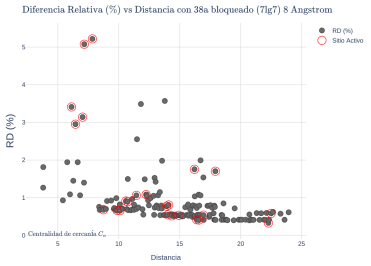
(a) Diferencia relativa de la red de C_α al bloquear el residuo A38.



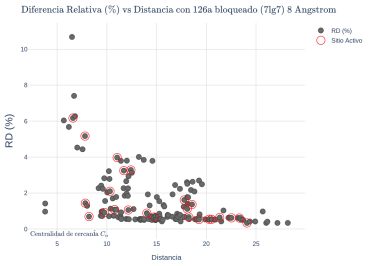
(b) Diferencia relativa de la red de C_α al bloquear el residuo N40.



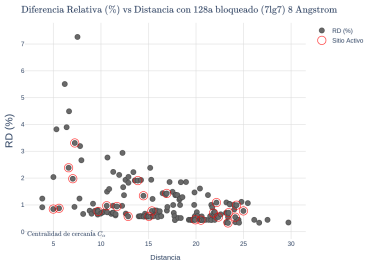
(c) Diferencia relativa de la red de C_α al bloquear el residuo S128.



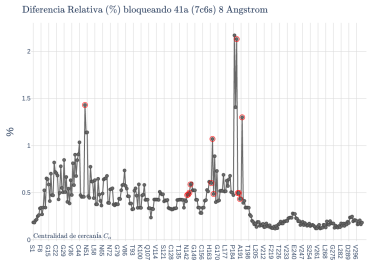
(d) Diferencia relativa de la red de C_α al bloquear el residuo A38.



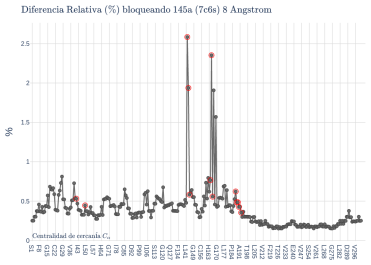
(e) Diferencia relativa de la red de C_α al bloquear el residuo N40.



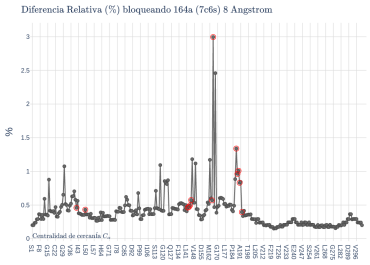
(f) Diferencia relativa de la red de C_α al bloquear el residuo S128.



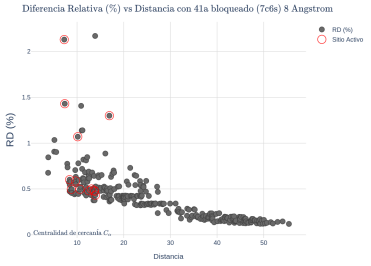
(g) Diferencia relativa de la red de C_α al bloquear el residuo H41.



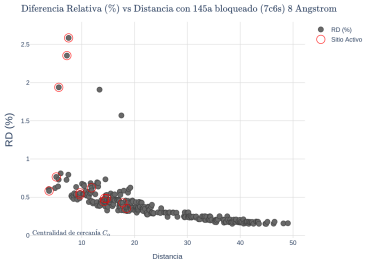
(h) Diferencia relativa de la red de C_α al bloquear el residuo C145.



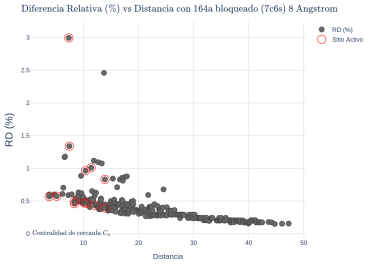
(i) Diferencia relativa de la red de C_α al bloquear el residuo C164.



(j) Diferencia relativa de la red de C_α al bloquear el residuo H41.

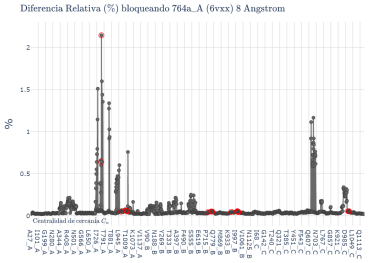


(k) Diferencia relativa de la red de C_α al bloquear el residuo C145.

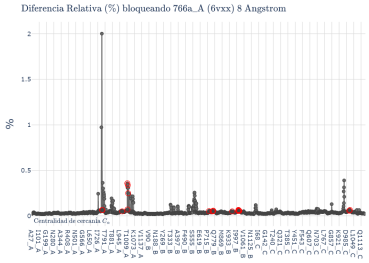


(l) Diferencia relativa de la red de C_α al bloquear el residuo C145.

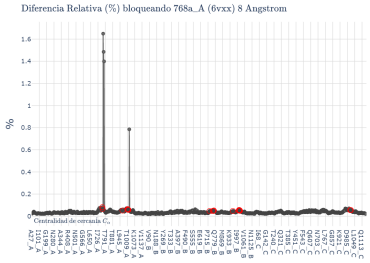
Figura 5.3: Diferencia relativa para la centralidad de cercanía para cada uno de los tres sitios activos bloqueados.



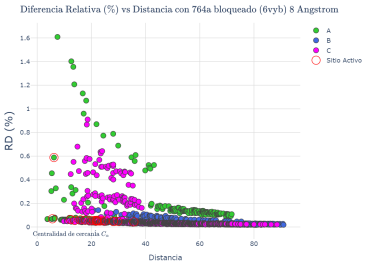
(a) Diferencia relativa de la red de C_α al bloquear el residuo N764.



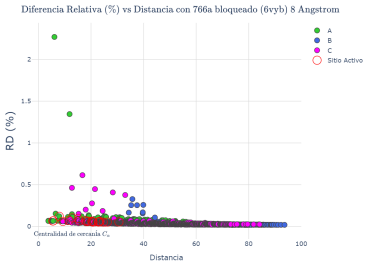
(b) Diferencia relativa de la red de C_α al bloquear el residuo A766.



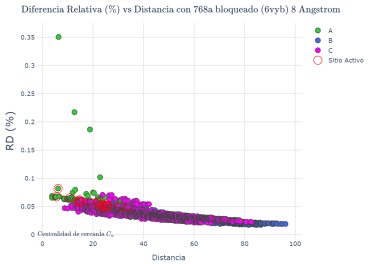
(c) Diferencia relativa de la red de C_α al bloquear el residuo T768.



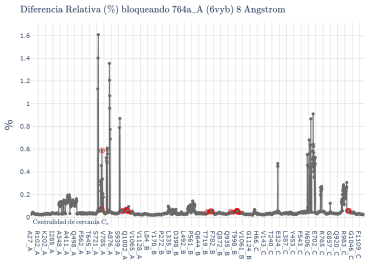
(d) Diferencia relativa de la red de C_α al bloquear el residuo N764.



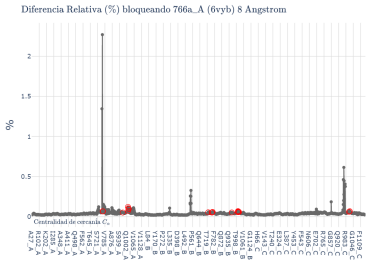
(e) Diferencia relativa de la red de C_α al bloquear el residuo A766.



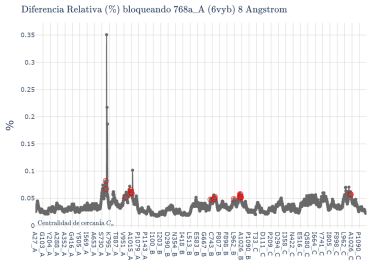
(f) Diferencia relativa de la red de C_α al bloquear el residuo T768.



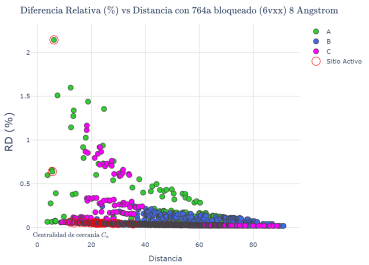
(g) Diferencia relativa de la red de C_α al bloquear el residuo N764.



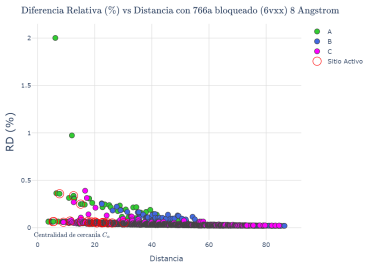
(h) Diferencia relativa de la red de C_α al bloquear el residuo A766.



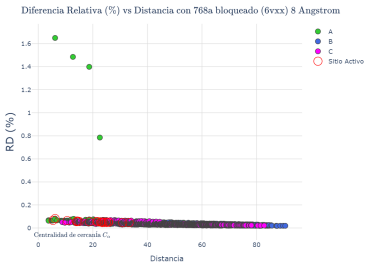
(i) Diferencia relativa de la red de C_α al bloquear el residuo T768.



(j) Diferencia relativa de la red de C_α al bloquear el residuo N764.



(k) Diferencia relativa de la red de C_α al bloquear el residuo A766.



(l) Diferencia relativa de la red de C_α al bloquear el residuo T768.

Figura 5.4: Diferencia relativa para la centralidad de cercanía para cada uno de los tres sitios activos bloqueados.

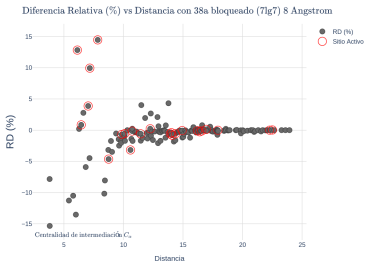
5.3. Centralidad de intermediación en redes de proteínas mutantes.

El efecto de los bloqueos en la centralidad de intermediación muestran un comportamiento distintivo a las anteriores medidas de centralidad. Podemos notar que el primer carbono C_α más favorecido por el bloqueo es un carbono adyacente, alrededor de 3.8 Å de distancia cartesiana, debido a que las rutas geodésicas que pasaban por el sitio activo encontraron como mejor medio de ruta después del bloqueo al carbono adyacente. Aunque, a simple vista, no necesariamente los más cercanos son los más favorecidos sino también lo son los más afectados debido a que el sitio activo bloqueado fue crucial para que las rutas geodésicas pasaran a través de los carbonos afectados. Recuerde que la centralidad de intermediación mide la importancia de un nodo de acuerdo a la frecuencia en que aparece entre las rutas geodésicas de cada par de nodos. Entonces, el bloqueo hace que las rutas que pasaban por un carbono afectado tengan que pasar por otros en la red. Podemos ver, incluso, que los carbonos más distantes del sitio activo bloqueado también presentan este mismo comportamiento pero en menor grado.

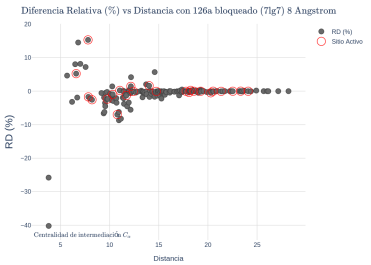
La gráfica de la Figura 5.5 (a) al bloquear A38 en la proteína 7LG7 muestra sitios activos cercanos que forman parte del pocket de anclaje distal de ribosa y adenosina, entre ellos N40 que es un sitio activo crítico y es el tercer máximo más afectado, véase la Figura 5.5 (d). Al bloquear L126 y S128 podemos observar que carbonos adyacentes son favorecidos en un 20 % más de centralidad, 40 % en el caso de L126. Los sitios activos que empeoraron también pertenecen al pocket de anclaje, véase la Figura 5.5 (e) y (f). En las gráficas podemos observar varias regiones en la secuencia PDB distantes de los bloqueos donde muestran las mismas fluctuaciones entre afectados y favorecidos. En los bloqueos de la proteína 7C6S, podemos observar que afecta mayormente a los carbonos cercanos dentro del radio de interacción (8 Å), los cuales son sitios de anclaje. Incluso los carbonos más distantes son en su mayoría afectados por cualquiera de los bloqueos, véase la Figura 5.5 (g), (h) y (i).

Los bloqueos de la proteína Spike en sus dos estados, los sitios activos que fueron más afectados y favorecidos son los que pertenecen al mismo pocket de anclaje con un porcentaje alrededor de 10 % y 20 %. Podemos ver en las gráficas de DR en relación a la distancia espacial, Figura 5.6, que tan solo un bloqueo puede afectar la importancia de un carbono a una distancia mayor al radio de interacción de hasta un 30 %. Incluso podemos ver carbonos pertenecientes a la subunidad S1 (secuencia PDB 27 a 685) de la cadena B que estando muy alejados, obtuvieron un cierto porcentaje de afectación no menor. Los carbonos más afectados en la proteína Spike, independiente de la distancia, están en la misma cadena A al que pertenecen los sitios activos bloqueados. Recordemos que la proteína Spike se ensambla por trímeros donde N764, A766 y T768 se hallan a la

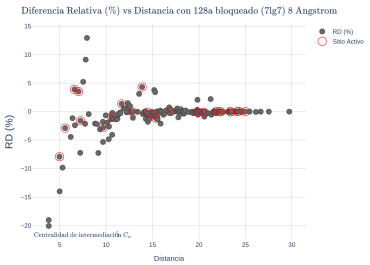
mitad de la estructura. Entonces, al bloquear alguno de estos, afectará con mayor grado las rutas geodésicas de los carbonos de la misma cadena por la corta proximidad entre carbonos adyacentes de la misma secuencia, tal como lo vimos con la centralidad de cercanía.



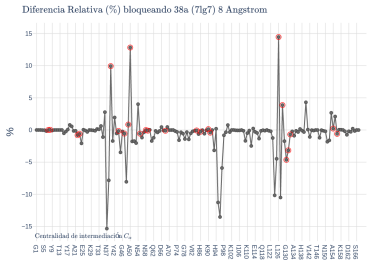
(a) Diferencia relativa de la red de C_α al bloquear el residuo A38.



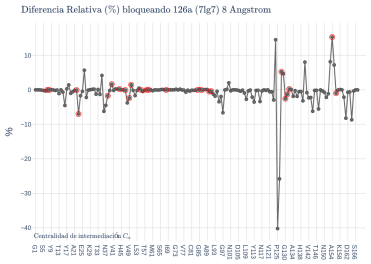
(b) Diferencia relativa de la red de C_α al bloquear el residuo L126.



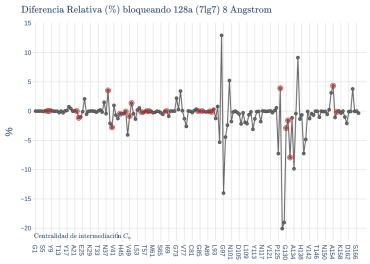
(c) Diferencia relativa de la red de C_α al bloquear el residuo S128.



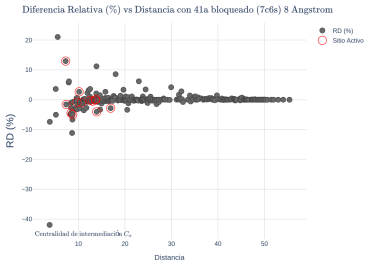
(d) Diferencia relativa de la red de C_α al bloquear el residuo A38.



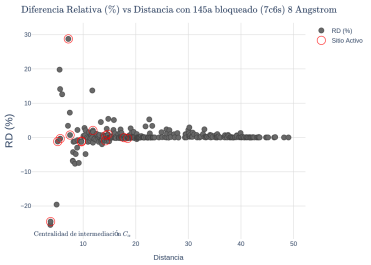
(e) Diferencia relativa de la red de C_α al bloquear el residuo L126.



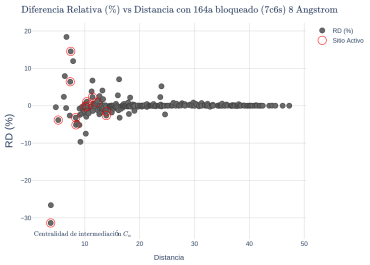
(f) Diferencia relativa de la red de C_α al bloquear el residuo S128.



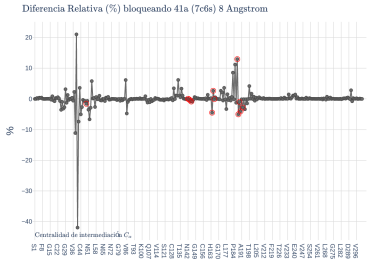
(g) Diferencia relativa de la red de C_α al bloquear el residuo H41.



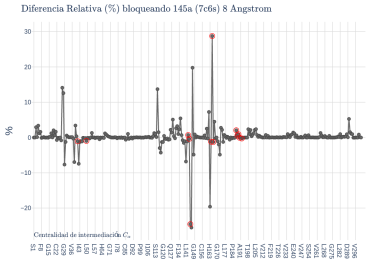
(h) Diferencia relativa de la red de C_α al bloquear el residuo C145.



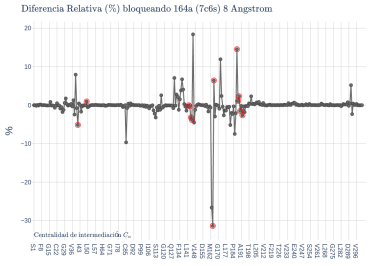
(i) Diferencia relativa de la red de C_α al bloquear el residuo Q192.



(j) Diferencia relativa de la red de C_α al bloquear el residuo H41.

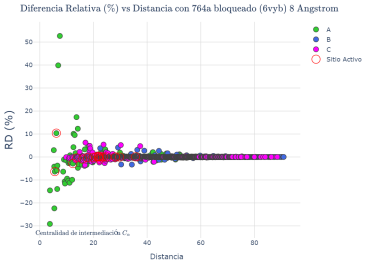


(k) Diferencia relativa de la red de C_α al bloquear el residuo C145.

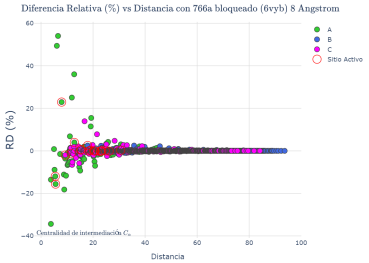


(l) Diferencia relativa de la red de C_α al bloquear el residuo Q192.

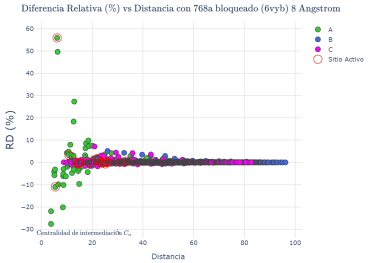
Figura 5.5: Diferencia relativa para la centralidad de intermediación para cada uno de los tres sitios activos bloqueados.



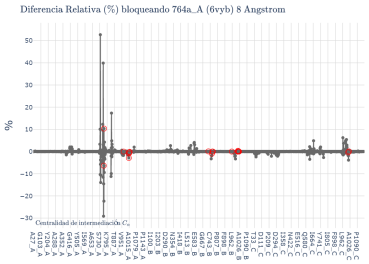
(a) Diferencia relativa de la red de C_α al bloquear el residuo N764.



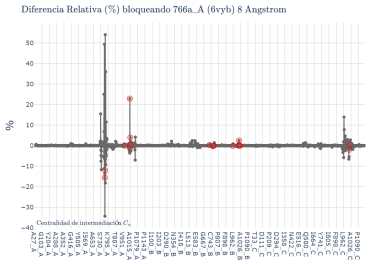
(b) Diferencia relativa de la red de C_α al bloquear el residuo A766.



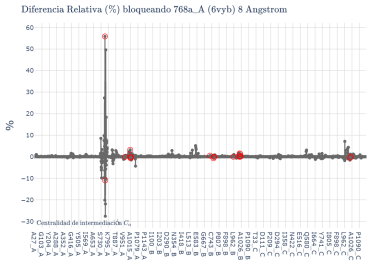
(c) Diferencia relativa de la red de C_α al bloquear el residuo T768.



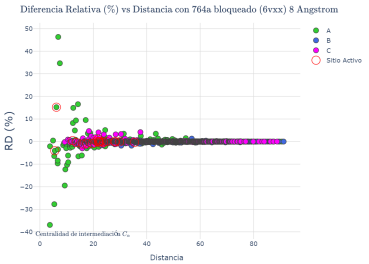
(d) Diferencia relativa de la red de C_α al bloquear el residuo N764.



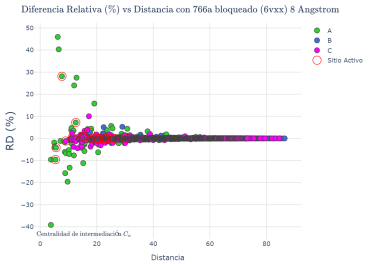
(e) Diferencia relativa de la red de C_α al bloquear el residuo A766.



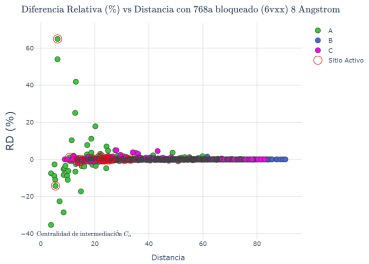
(f) Diferencia relativa de la red de C_α al bloquear el residuo T768.



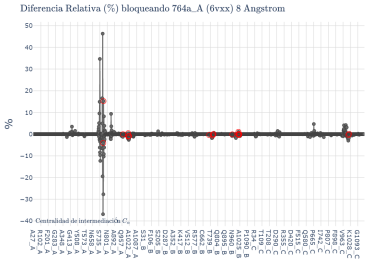
(g) Diferencia relativa de la red de C_α al bloquear el residuo N764.



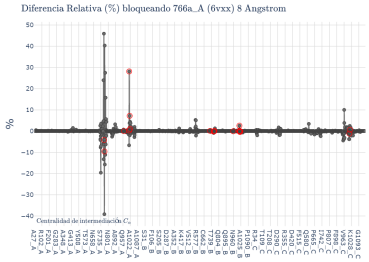
(h) Diferencia relativa de la red de C_α al bloquear el residuo A766.



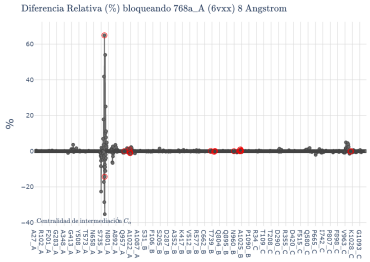
(i) Diferencia relativa de la red de C_α al bloquear el residuo T768.



(j) Diferencia relativa de la red de C_α al bloquear el residuo N764.



(k) Diferencia relativa de la red de C_α al bloquear el residuo A766.



(l) Diferencia relativa de la red de C_α al bloquear el residuo T768.

Figura 5.6: Diferencia relativa para la centralidad de intermediación para cada uno de los tres sitios activos bloqueados.

Capítulo 6

Conclusiones y perspectivas

En este trabajo desarrollamos modelos matemáticos basados en la teoría de grafos para desentrañar la topología de la estructura de carbonos C_α y C_β de proteínas que son de suma importancia en el ciclo de replicación del virus SARS-CoV-2, los cuales se han convertido en blancos farmacológicos. Obtuvimos las coordenadas espaciales de los carbonos C_α y C_β a partir del archivo PDB. Construimos los grafos conectando un carbono con aquellos que están cercanos en un radio de interacción igual a 8 Å debido a que la distancia promedio entre carbonos C_α contiguos es de aproximadamente 4 Å. Bloqueamos sitios activos clave para analizar la influencia en la topología de la red a través de de las medidas centralidad. Esto fue posible usando la matriz de adyacencia al hacer cero en todas las entradas de la fila y columna que corresponde al carbono que bloqueamos.

Desde los trabajos de Steven Strogatz y Duncan Watts bajo un enfoque multidisciplinario al analizar redes de campos completamente diferentes como la red neuronal de los Caenorhabditis elegans, la red eléctrica del oeste de los Estados Unidos y la red de colaboración de los actores de cine [36], se sabe que la mayoría de las redes del mundo real tienden a estar entre la topología de un grafo regular y un grafo aleatorio. Las redes del mundo real presentan la propiedad de mundo pequeño donde las distancias geodésicas promedio son menores y el coeficiente de agrupamiento es alto. En nuestro trabajo, demostramos que la red de carbonos C_α y C_β no es la excepción a esta regla de mundo pequeño como se mostró en el Cuadro 4.1. Esto contrasta tanto a las redes aleatorias que poseen un coeficiente de agrupamiento menores, y las redes regulares que poseen distancias geodésicas promedio muy grandes.

Algunas redes del mundo real como la World Wide Web poseen una distribución de grado de Ley de Potencia con una predominancia en nodos con grado menor y muy pocos nodos con grado alto. Por otro lado, las redes aleatorias presentan una distribución de Poisson, los cuales los nodos tienden a tener un grado similar. Sin embargo, las redes de carbonos de las proteínas objetivo del SARS-CoV-2 no presentó ninguna de las anteriores.

Usamos las medidas topológicas de centralidad para describir la importancia de un carbono en la red. Calculamos la centralidad en tres redes correspondiente a la estructura únicamente de carbonos C_α, otra de carbonos C_β y, en conjunto, la estructura de carbonos C_α-C_β. Recordemos que el carbono C_α une al grupo amina y al grupo carboxilo, y posee una cadena lateral cuyo primer átomo, a excepción de glicina, es el carbono C_β. Ambas categorías de carbonos están enlazados covalentemente y tienen una distancia de separación de 1.5Å. Entonces, la pregunta propuesta es si ¿la medida de centralidad en la red C_α-C_β de ambos carbonos es igual a la suma de sus correspondientes centralidades independientes en la red C_α y la red C_β? Encontramos que la centralidad de cercanía tiene una relación lineal casi perfecta, con un coeficiente de correlación de Pearson de 0.99, y posee una pendiente y ordenada al origen aproximadamente de 1 y 0, respectivamente.

La centralidad de vector propio en las redes de carbonos poseen una correlación lineal alta de 0.97 a excepción de la proteína más pequeña 7LG7 que tuvo un coeficiente de 0.86 por la discrepancia de centralidad entre las redes. Recordemos que las proteínas están formadas principalmente por estructuras de hélice y hojas plegadas. Los carbonos C_β en las hélices están extendidas hacia fuera, mientras que en la estructura de hoja plegada están extendidas en direcciones opuestas entre aminoácidos contiguos. Esto puede tener mayor influencia en la red de carbonos C_α-C_β que en las redes por separado.

La centralidad de intermediación en las redes también muestran una alta relación lineal con un coeficiente de alrededor de 0.90, menor al de la centralidad de cercanía. La pendiente de la recta que mejor se ajustó a los datos fue alrededor de 2 mostrando una menor centralidad de intermediación en la red C_α-C_β. Esto muestra que ambos tipos de carbonos C_α y C_β en una misma red se ven afectada su centralidad debido a su aproximación, es decir, 1.5 Å de distancia en un mismo aminoácido, ya a que las rutas geodésicas que pudieron pasar por uno de ellos ahora pueden hacerlo también por el carbono contiguo.

La estructura biológicamente funcional de una proteína globular está determinada por la secuencia de aminoácidos, como vimos en la parte de introducción. Los aminoácidos hidrofóbicos al entrar en contacto con las moléculas de agua quedan mayormente ocultos en la estructura, a partir del cual se propaga la estructura nativa, un estado termodinámicamente estable. Por lo tanto, podemos ver que los aminoácidos hidrofóbicos juegan un papel definitivo en la estructura tridimensional y funcional. Sin embargo, esto no faltó en manifestarse en las medidas de centralidad donde podemos ver una mayor proporción de centralidad entre los carbonos de aminoácidos hidrofóbicos. Esto fue aún mayor tanto en la centralidad de vector propio como la centralidad de intermediación debido al efecto hidrofóbico, es decir, a la alta tendencia de agruparse entre las moléculas hidrofóbicas de tal forma que quedan más cerca uno de otros fuera del ambiente polar. Esto se ve reflejado en la centralidad de vector propio ya que mide la importancia de un nodo de acuerdo al

número de nodos adyacentes y al valor de los mismos. La centralidad de intermediación, por otra parte, los hidrofóbicos al estar mejor conectados entre ellos forman un núcleo por el cual la mayoría de las rutas geodésicas pueden pasar. Mientras que los polares que tuvieron menor centralidad se hallan mayormente en la superficie donde podemos encontrar carbonos con centralidad cero. Por otro lado, la proporción de centralidad de cercanía entre los carbonos de aminoácidos hidrofóbicos fue ligeramente mayor en las cuatro proteínas. Sin embargo, los sitios activos presentaron una centralidad de cercanía importante, encontrándose entre el 10 % de carbonos con mayor centralidad como se observó en las proteínas 7LG7, 6VYB y 6VXX.

Uno de los objetivos en este proyecto fue el efecto en la topología de la red al bloquear los sitios activos de la proteína a partir de las medidas de centralidad. Esto nos dio información interesante con respecto de carbonos que disminuyeron su importancia en la red y carbonos que fueron favorecidos al bloquear un sitio activo donde podemos ver un efecto global en la red, a distancias más grandes del radio de interacción. En una proteína con múltiples cadenas como la proteína Spike, los carbonos más afectados en un radio mayor al radio de interacción pertenecen a la misma cadena que los sitios bloqueados a pesar de que hay carbonos más cercanos pero que pertenecen diferentes cadenas. Esto es debido a que los carbonos de la cadena polipeptídica están más agrupados, donde los C_α contiguos están a una distancia de aproximadamente 4 Å por lo que están mejor conectados entre ellos.

La centralidad de intermediación a diferencia de las otras dos medidas, hay tanto carbonos afectados como beneficiados por los bloqueos, y disminuye el efecto conforme aumenta la distancia euclidiana. Recuerde que la centralidad de intermediación mide la importancia de un nodo de acuerdo a la frecuencia en que aparece entre las rutas geodésicas de cada par de nodos. Al bloquear un carbono hace que las rutas geodésicas que pasaban por éste tengan que pasar por otros más distantes. Estos otros carbonos son, por lo tanto, los más favorecidos por el bloqueo. Sin embargo, en estas nuevas rutas existen otros carbonos de la ruta anterior que ya no son tomados, por lo cual se ven disminuir su centralidad en la red.

Los sitios activos que fueron mayormente afectados en cada una de las medidas de centralidad, especialmente en la centralidad de vector propio y la centralidad de cercanía, fueron carbonos que pertenecen a los bolsillos de anclaje al que también pertenecen los carbonos bloqueados debido a su proximidad. Por otra parte, también se pudo observar carbonos que no pertenecen a los sitios activos que estuvieron entre el 10 % de los más afectados. Sin embargo, esto no es menor debido a que puede ser una parte crucial en la estructura que un daño podría afectar críticamente el funcionamiento biológico de la proteína, incluso su destrucción [14].

PDB	Tipo	N	m	BC	CC	EC
6VXX	C_α-C_β	5658	59006	00:04:53	00:01:10	00:00:04
	C_α	2916	14648	00:00:45	00:00:10	00:00:01
	C_β	2702	14986	00:00:48	00:00:07	00:00:01
6VYB	C_α-C_β	5577	58012	00:04:54	00:01:14	00:00:06
	C_α	2875	14367	00:00:45	00:00:08	00:00:01
	C_β	2702	13097	00:00:56	00:00:13	00:00:01

Cuadro 6.1: Tiempo de compilación para el calculo de la centralidad de intermediación (CB), la centralidad de cercanía (CC) y la centralidad de vector propio (CE). La tabla muestra el número de carbonos N y el número de aristas m.

La principal dificultad al trabajar con las proteínas fue el manejo del repositorio PDB por el cual es necesario conocer bien las distintas bibliotecas de Python para el manejo de datos, la optimización y la legibilidad del código. Usamos Biopandas principalmente para extraer información del archivo PDB y la biblioteca Pandas para filtrarlos ya que hay átomos que parecían repetidos debido al factor de ocupación como en las proteínas 7LG7 y 7C6S, o porque pertenecen a una cadena distinta como la proteína Spike, 6VYB y 6VXX. En la operación matemática, el manejo de matrices, el cálculo del coeficiente de correlación de Pearson, etc. se usó la biblioteca NumPy. La biblioteca NetworkX por la eficiencia de su estructura de datos dirigidos a grafos. En el próximo Cuadro 6.1, muestra el tiempo de compilación promedio de 100 repeticiones para cada una de las medidas de centralidad y cada tipo de red. El proceso se llevó acabo en una computadora con un procesador de gama media, AMD Ryzen 5 3450u, con una unidad de estado sólido (SSD) y una memoria RAM de 8 GB.

No sorprende que la centralidad de intermediación tardara más tiempo que la centralidad de cercanía ya que no solo usa el algoritmo de búsqueda en anchura (BFS) sino que también el algoritmo de Ulrik. Mientras que la centralidad de vector propio fue más rápido debido a que el algoritmo de iteración de potencias se basa en multiplicar reiteradamente la matriz de adyacencia por un vector inicial con elementos no negativos. Podemos observar que la red C_α-C_β en las proteínas Spike S, la red más grande con más de 5 mil nodos y más de 58 mil aristas, tardó aproximadamente 5 minutos para la centralidad de intermediación y 1 minuto para la centralidad de cercanía. Una vez calculado las centralidades podrían guardarse en una base de datos de forma que sea más rápido el acceso a los cálculos. Además, para ampliar aún más este trabajo en el caso de los bloqueos de carbonos pueden implementarse los *algoritmos incrementales* (en inglés, Incremental Algorithms) para reducir el número de BFS ante cambios topológicos [46], esto podría reducir aún más el tiempo tanto para la centralidad de cercanía como la centralidad de intermediación.

Los nodos y aristas menores a 3000 y 14000, respectivamente, tardan menos de un minuto. Por lo tanto, en proteínas típicas con centenas de aminoácidos [13] tardaría menos de un segundo. Es por este motivo que se tiene previsto nuevos algoritmos para manejar múltiples proteínas, incluso aquellos con más de mil aminoácidos como la proteína Spike, y analizar la topología de redes de carbonos C_α y C_β con más detalle, con distintos radios de interacción y, ¿por qué no?, con más átomos de los aminoácidos. Se tiene previsto también modelos matemáticos para la generación de grafos aleatorios a partir de la topología de las redes de carbonos C_α y C_β de las proteínas vistas en este trabajo, así como las medidas de centralidad. Por ejemplo, la centralidad de cercanía nos proporciona información de los sitios activos los cuales tienen una alta tendencia de estar más cercanos al resto de la red; o la centralidad de intermediación y la centralidad de vector propio que reflejan la alta agrupación de los aminoácidos hidrofóbicos. El estudio de la estructura biológicamente funcional de una proteína podría arrojar más resultados interesantes.

Apéndice A

Algoritmo de Ulrik

El algoritmo se basa en el criterio de Bellman por el cual divide el problema en subproblemas más simples (primer parte) para después obtener una solución general (segunda parte). Para ello se necesitarán los dos próximos lemas y un teorema.

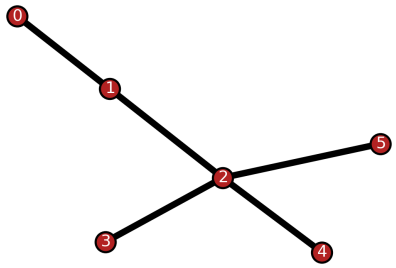


Figura A.1: El algoritmo de búsqueda ampliada recorre cada nodo a partir de un nodo raíz.

Lema 1 (Criterio de Bellman [39])**.** *Un vértice $v \in V$ se encuentra en una ruta geodésica entre vértices $s, t \in V$, si y sólo si $d(s,t) = d(s,v) + d(v,t)$*

Ulrik llama **dependencia del par** (en inglés, pair-dependency) a la fracción de rutas geodésicas donde cada nodo aparece entre los pares de nodos $\delta_{st}(v) = \frac{\sigma_{st}(v)}{\sigma_{st}}$. De esta forma, la ecuación A.1 puede escribirse como

$$C'_B(k) = \sum_{i<j} \delta_{st}(k) \qquad s \neq t \neq k \tag{A.1}$$

En cada recorrido de un nodo a otro nodo adyacente, Ulrik define el conjunto de nodos predecesores como

$$P_s(k) = \{v \in V : v, k \in V : d(s,k) = d(s,v) + 1\}. \tag{A.2}$$

Entonces, el conteo de rutas geodésicas puede escribirse como

Lema 2 ([39]). *Para* $s \neq v \in V$

$$\sigma_{sv} = \sum_{u \in P_s(k)} \sigma_{su}$$

La primera parte del algoritmo de Ulrik es usar BFS para que en cada nivel del recorrido se visite los nodos adyacentes w del nodo actual v, y se les asigne:

1. la distancia geodésica $d_s(w) = d_s(v) + 1$ del nodo raíz s al nodo adyacente w de v,
2. el predecesor $v \in P_s(w)$, el cual será almacenado en una lista de w, y
3. el número de rutas geodésicas $\sigma_{sw} = \sigma_{sv} + \sigma_{sw}$.

Cada nodo hallado se almacena en una estructura de datos llamado pila S de tal forma que el nodo raíz se agrega primero seguido por sus nodos adyacentes, así también en el siguiente nivel hasta almacenar el último nodo de la red. La estructura agrega los datos de tal forma que el último elemento es el primero que será extraído como una pila de objetos. Esta estructura de datos nos servirá para más adelante calcular la centralidad de intermediación. El siguiente ejemplo ilustra este procedimiento en un grafo con 6 nodos, véase la Figura A.1. Si el nodo raíz es 0, entonces

1. $d_0(1) = d_0(0) + 1 = 1$,
2. $0 \in P_0(1)$, y
3. $\sigma_{0,1} = \sigma_{0,0} + \sigma_{0,1} = 1$ y $\sigma_{0,2} = \sigma_{0,0} + \sigma_{0,2} = 1$.

Agregamos el nodo raíz a la estructura pila S, después agregamos el nodos 1 que fue encontrado. Posteriormente avanzamos al siguiente nivel en el nodo 1 repitiendo estos tres pasos al nodo adyacente 2, y agregamos el nodo 2 a la pila. En la tabla A.1 (a) muestra el siguiente nivel hasta recorrer toda la red. La pila termina de almacenar de forma ascendente empezando por el nodo raíz 0 hasta el último nodo 5.

Teorema A.0.1 ([39]). *La dependencia de* $s \in V$ *en cualquier* $v \in V$ *obedece*

$$\delta_{s\bullet}(v) = \sum_{w : v \in P_s(v)} \frac{\sigma_{sv}}{\sigma_{sw}} (1 + \delta_{s\bullet}(w)) \tag{A.3}$$

Demostración. Recordemos que $\delta_{st}(v) > 0$ es sólo para aquellos $t \in V \backslash \{s\}$ para la cual v se encuentra en al menos una ruta geodésica de s a t, y observe que en cualquiera de esos caminos hay exactamente una arista $\{v, w\}$ con $v \in P_s(w)$.

Extendemos la dependencia del par para incluir una arista $e \in E$ definiendo $\delta_{st}(v,e) = \frac{\sigma_{st}(v,e)}{\sigma_{st}}$ donde $\sigma_{st}(v,e)$ es el número de rutas geodésicas de s a t que contiene a ambos v y e. Entonces,

$$\delta_{s\bullet}(v) = \sum_{t \in V} \delta_{st}(v) = \sum_{t \in V} \sum_{w: v \in P_s(w)} \delta_{st}(v, \{v,w\}) = \sum_{w: v \in P_s(w)} \sum_{t \in V} \delta_{st}(v, \{v,w\}) \tag{A.4}$$

Sea w cualquier vértice con $v \in P_s(w)$. En las rutas geodésicas de s a w, σ_{sv} muchos primero van de s a v y entonces usan $\{v,w\}$. Esto sigue que la dependencia del par s a v y $\{v,w\}$ es

$$\delta_{st}(v, \{v,w\}) = \begin{cases} \frac{\sigma_{sv}}{\sigma_{sw}}, & \text{si } t = w \\ \frac{\sigma_{sv}}{\sigma_{sw}} \cdot \frac{\sigma_{st}(w)}{\sigma_{st}}, & \text{si } t \neq w \end{cases}$$

Si insertamos esto en la ecuación A.4 produce

$$\begin{aligned} \sum_{w: v \in P_s(w)} \sum_{t \in V} \delta_{st}(v, \{v,w\}) &= \sum_{w: v \in P_s(w)} \left(\frac{\sigma_{sv}}{\sigma_{sw}} + \sum_{t \in V} \frac{\sigma_{sv}}{\sigma_{sw}} \cdot \frac{\sigma_{st}(w)}{\sigma_{st}} \right) \\ &= \sum_{w: v \in P_s(w)} \frac{\sigma_{sv}}{\sigma_{sw}} \cdot (1 + \delta_{s\bullet}(w)) \end{aligned}$$

□

La segunda parte del algoritmo de Ulrik es un proceso inverso en la cual se extraen los nodos de la pila para calcular la dependencia del nodo raíz s de la ecuación A.3, y obtener la centralidad de intermediación. En la tabla A.2 (a) muestra el resultado para $s = 0$.

Estas dos partes del algoritmo se repite para los demás nodos. En la tabla A.2 (b) muestra el resultado de la primera parte para $s = 1$ en la cual la pila empieza con el nodo 1 seguido por 0, después 2 hasta terminar con el último nodo 5. En la tabla A.2 (b) muestra la segunda parte del algoritmo para el nodo raíz $s = 1$.

Nivel	Pasos	$s = 0$
1	1.	$d_0(1) = 1$
	2.	$\sigma_{0,1} = 1$
	3.	$P_0(1) = \{0\}$
2	1.	$d_0(2) = 2$
	2.	$\sigma_{0,2} = 1$
	3.	$P_0(2) = \{1\}$
3	1.	$d_0(3) = 3$
	2.	$\sigma_{0,3} = 1$
	3.	$P_0(3) = \{2\}$
	1.	$d_0(4) = 3$
	2.	$\sigma_{0,4} = 1$
	3.	$P_0(4) = \{2\}$
	1.	$d_0(5) = 3$
	2.	$\sigma_{0,5} = 1$
	3.	$P_0(5) = \{2\}$

(a) El 0 como el nodo raíz

Nivel	Pasos	$s = 1$
1	1.	$d_1(0) = 1$
	2.	$\sigma_{1,0} = 1$
	3.	$P_1(0) = \{1\}$
	1.	$d_1(2) = 1$
	2.	$\sigma_{1,2} = 1$
	3.	$P_1(2) = \{1\}$
2	1.	$d_1(3) = 2$
	2.	$\sigma_{1,3} = 1$
	3.	$P_1(3) = \{2\}$
	1.	$d_1(4) = 2$
	2.	$\sigma_{1,4} = 1$
	3.	$P_1(4) = \{2\}$
	1.	$d_1(5) = 2$
	2.	$\sigma_{1,5} = 1$
	3.	$P_1(5) = \{2\}$

(b) El 1 como el nodo raíz

Cuadro A.1: Resultados del Algoritmo de BFS.

$s = 0$
$\delta_{0\bullet}(2) = 1{,}0$
$C'_B(5) = 0{,}0$
$\delta_{0\bullet}(2) = 2{,}0$
$C'_B(4) = 0{,}0$
$\delta_{0\bullet}(2) = 3{,}0$
$C'_B(3) = 0{,}0$
$\delta_{0\bullet}(1) = 4{,}0$
$C'_B(2) = 3{,}0$
$\delta_{0\bullet}(0) = 5{,}0$
$C'_B(1) = 4{,}0$

(a) El 0 como el nodo raíz

$s = 1$
$\delta_{1\bullet}(2) = 2{,}0$
$C'_B(4) = 0{,}0$
$\delta_{1\bullet}(2) = 3{,}0$
$C'_B(3) = 0{,}0$
$\delta_{1\bullet}(2) = 3{,}0$
$C'_B(3) = 0{,}0$
$\delta_{1\bullet}(1) = 4{,}0$
$C'_B(2) = 6{,}0$
$\delta_{1\bullet}(1) = 5{,}0$
$C'_B(0) = 0{,}0$

(b) El 1 como el nodo raíz

Cuadro A.2: Resultados del Algoritmo de Ulrik.

Apéndice B

Mínimos Cuadrados Discretos.

El método de mínimos cuadrados consiste en hallar una recta $mx_i + c$ que mejor se ajuste a un conjunto de datos $\{(x_i, y_i)\}_i^N$, de tal forma que minimice el error total [47]:

$$E(m, c) = \sum_{i=1}^{m} |y_i - (mx_i + c)|^2$$

donde m es la pendiente y c es la ordenada al origen. Por consiguiente, consiste en resolver

$$\frac{\partial E}{\partial m} = 0 \qquad \text{y} \qquad \frac{\partial E}{\partial c} = 0 \tag{B.1}$$

de tal forma que el resultado

$$2\sum_{i=1}^{N}(y_i - mx_i - c)(-1) \qquad \text{y} \qquad 2\sum_{i=1}^{N}(y_i - mx_i - c)(-x_i) \tag{B.2}$$

Por lo tanto,

$$m = \frac{\sum_{i=1}^{N} x_i^2 \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i y_i \sum_{i=1}^{N} x_i}{N\left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)^2} \tag{B.3}$$

y

$$c = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)^2} \tag{B.4}$$

El coeficiente de correlación de Pearson (CCP) es una medida que evalúa la calidad general del ajuste lineal [48], el cual se define como

$$CCP = cc'$$

donde c' la ordenada al origen de la recta $m'y_i + c'$. El rango del coeficiente está entre -1 a 1. Si el coeficiente es -1, los valores de una variable aumentan mientras que la otra disminuye; y si es 1, ambas variables aumentan.

Bibliografía

[1] Geneva: World Health Organization. Who covid-19 dashboard. Recuperado el 1 de enero de 2023, de https://covid19.who.int/.

[2] Ali Rabaan, Shamsah Al-Ahmed, Shafiul Haque, Ranjit Sah, Ruchi Tiwari, Yashpal Singh Malik, Kuldeep Dhama, M Iqbal Yatoo, D Katterine Bonilla-Aldana, and Alfonso J Rodriguez-Morales. Sars-cov-2, sars-cov, and mers-cov: A comparative overview. *Le infezioni in medicina*, 28:174–184, 2020.

[3] World Health Organization (WHO). Middle East respiratory syndrome coronavirus (MERSCoV). Recuperado el 10 de agosto de 2023, de https://www.who.int/health-topics/middle-east-respiratory-syndrome-coronavirus-mers#tab=tab_1.

[4] World Health Organization (WHO). Respiratory syndrome coronavirus (SARS-CoV). Recuperado el 10 de enero de 2023, de https://www.who.int/emergencies/disease-outbreak-news/item/2004_01_05-en, 2004.

[5] Ritesh Gorkhali, Prashanna Koirala, Sadikshya Rijal, Ashmita Mainali, Adesh Baral, and Hitesh Kumar Bhattarai. Structure and function of major sars-cov-2 and sars-cov proteins. *Bioinformatics and biology insights*, 15, 2021.

[6] Lubin JH, Zardecki C, Dolan EM, Lu C, Shen Z, and Dutta S. 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, 2022.

[7] Taylor Heald-Sargent and Tom Gallagher. Ready, set, fuse! the coronavirus spike protein and acquisition of fusion competence. *Viruses*, 4(4):557–580, 2012.

[8] Ella Hartenian Divya Nandakumar, Azra Lari, Michael Ly, Jessica Tucker, and Britt Glaunsinger. The molecular virology of coronaviruses. *The Journal of biological chemistry*, 295(37):12910–12934, 2020.

[9] Matthew D Hall, James M Anderson, Annaliesa Anderson, David Baker, Jay Bradner, Kyle R Brimacombe, Elizabeth A Campbell, and Kizzmekia S Corbett. Report of the

national institutes of health sars-cov-2 antiviral therapeutics summit. *The Journal of infectious diseases*, 224(Supplement 1):S1–S21, 2021.

[10] Bethany Halford. How pfizer scientists transformed an old drug lead into a covid-19 antiviral. Recuperado el 10 de enero de 2023, de https://cen.acs.org/pharmaceuticals/drug-discovery/How-Pfizer-scientists-transformed-an-old-drug-lead-into-a-COVID-19-antiviral/100/i3, ene 2022.

[11] Biology OpenStax College. Recuperado el 26 de noviembre de 2021, de https://openstax.org/books/biology/pages/3-4-proteins. Imagen modificada.

[12] G.R. Choppin, L.R. Summerlin, and H.C. de Contin. *Química*. Publicaciones Cultural, México, (2003).

[13] Luis Olivares Quiroz. *Macromoléculas biológicas. El enigma entre la física y la biología molecular*. Universidad Autónoma de la Ciudad de México, (2020).

[14] A.V. Finkelstein and O.B. Ptitsyn. *Protein Physics: A Course of Lectures*. Series in soft condensed matter. Elsevier Science, 2nd edition, (2002).

[15] Marianne Rooman, Yves Dehouck, Jean Kwasigroch, Christophe Biot, and Dimitri Gilis. What is paradoxical about levinthal paradox? *Journal of biomolecular structure & dynamics*, 20:327–9, 01 2003.

[16] G.A. Petsko and D. Ringe. *Protein Structure and Function*. Primers in biology. New Science Press, (2004).

[17] J.M. Berg, L. Stryer, J.L. Tymoczko, and J.M. Macarulla. *Bioquímica*. Reverté, Barcelona, España., 6th edition, (2008).

[18] Alessandro Pintar, Oliviero Carugo, and Sándor Pongor. Dpx: For the analysis of the protein core. *Bioinformatics (Oxford, England)*, 19:313–4, 02 2003.

[19] Gita Sastria, C.-Y Liong, and Ishak Hashim. Recuperado el 3 de septiembre de 2022, de https://www.researchgate.net/figure/Kyle-and-Doolittle-K-D-hydrophobicity-scale_tbl1_234771184. Imagen modificada.

[20] Yuan Huang, Chan Yang, Xin-feng Xu, Wei Xu, and Shu-wen Liu. Structural and functional properties of sars-cov-2 spike protein: potential antivirus drug development for covid-19. *Acta Pharmacologica Sinica*, 41(9), 2020.

[21] A Fitri, H Basultan, and Iryani. Hydrophobic pocket of sars-cov-2 spike glycoprotein are potential as binding pocket. *Journal of Physics: Conference Series*, 1788(1):012021, feb 2021.

[22] Claudia Duran-Aniotz, Inés Moreno-Gonzalez, and Rodrigo Morales. Agregados amiloides: rol en desórdenes de conformación proteica. *Revista médica de Chile*, 141:495 – 505, 04 2013.

[23] Engel Paul. *Enzymes: A Very Short Introduction (Very Short Introductions)*. OUP Oxford, Great Britain, (2020).

[24] Antonia Stank, Daria B. Kokh, Jonathan C. Fuller, and Rebecca C. Wade. Protein binding pocket dynamics. *Accounts of Chemical Research*, 49(5):809–815, 2016. PMID: 27110726.

[25] Chris A. Brosey, Jerry H. Houl, Panagiotis Katsonis, Lakshitha P.F. Balapiti-Modarage, Shobanbabu Bommagani, Andy Arvai, Davide Moiani, Albino Bacolla, Todd Link, Leslie S. Warden, Olivier Lichtarge, Darin E. Jones, Zamal Ahmed, and John A. Tainer. Targeting sars-cov-2 nsp3 macrodomain structure with insights from human poly(adp-ribose) glycohydrolase (parg) structures with inhibitors. *Progress in Biophysics and Molecular Biology*, 163:171–186, 2021. DNA-Replication and Repair, Structures and Cancer.

[26] Will Douglas Heavenarchive page. Deepmind's protein-folding ai has solved a 50-year-old grand challenge of biology. Recuperado el 26 de noviembre de 2021, de https://www.technologyreview.com/2020/11/30/1012712/deepmind-protein-folding-ai-solved-biology-science-drugs-disease/.

[27] Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., and Bourne P. E. The protein data bank. *Nucleic acids research*, 28(1), Enero 2000.

[28] Christine Zardecki, Shuchismita Dutta, David S. Goodsell, Robert Lowe, Maria Voigt, and Stephen K. Burley. Pdb-101: Educational resources supporting molecular explorations through biology and medicine. *Protein Science*, 31(1):129–140, 2022.

[29] Merian-Erben. Königsberg. Recuperado el 10 de enero de 2023, de http://campusvirtual.cua.uam.mx/material/ipm/05_combinatoria_konigsberg_html/index.html#, 2015. Imagen por la unida Lite UAM Cuajimalpa.

[30] N. Biggs, E.K. Lloyd, and R.J. Wilson. *Graph Theory, 1736-1936*. Clarendon Press, United States, (1986).

[31] F. Harary. *Graph Theory*. Addison Wesley series in mathematics. Addison-Wesley, United States, (1971).

[32] M. Mitchell. *Complexity: A Guided Tour*. Oxford University Press, United States, (2009).

[33] M. E. J. Newman. *Networks. An Introduction*. Clásicos contemporáneos. Oxford University Press Inc., United States, (2010).

[34] B. Bollobás, B. Béla, W. Fulton, Cambridge University Press, A. Katok, F. Kirwan, P. Sarnak, and B. Simon. *Random Graphs*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2nd edition, (2001).

[35] E. N. Gilbert. Random Graphs. *The Annals of Mathematical Statistics*, 30(4):1141 – 1144, 1959.

[36] Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

[37] S. Wasserman, K. Faust, and J.L. Molina. *Análisis de redes sociales. Métodos y aplicaciones*. Clásicos contemporáneos. Centro de Investigaciones Sociológicas, (2013).

[38] Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. closeness_centrality. Recuperado el 28 de noviembre de 2021, de https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.centrality.closeness_centrality.html.

[39] Ulrik Brandes. A faster algorithm for betweenness centrality. *The Journal of Mathematical Sociology*, 25(2):163–177, 2001.

[40] Python Essays. Foreword for programming python. Recuperado el 10 de enero de 2023, de https://www.python.org/doc/essays/foreword/.

[41] William Wesley McKinney. *Python for Data Analysis*. O'Reilly Media, United States of America, (2012).

[42] TIOBE. Tiobe programming community index. Recuperado el 10 de enero de 2023, de https://www.tiobe.com/tiobe-index/, ago 2023.

[43] Salvatore Scellato. Networkx: Network analysis with python. Recuperado el 10 de agosto de 2023, de https://www.cl.cam.ac.uk/~cm542/teaching/2011/stna-pdfs/stna-lecture11.pdf, mar 2012.

[44] Keith Callenberg. Recuperado el 2 de mayo de 2022, de https://en.wikipedia.org/wiki/Accessible_surface_area#/media/File:Accessible_surface.svg, 05 2010. Imagen modificada.

[45] S. Hubbard and J. Thornton. Naccess. Recuperado el 26 de noviembre de 2021, de http://www.bioinf.manchester.ac.uk/naccess/.

[46] Ahmet Erdem Sariyüce, Kamer Kaya, Erik Saule, and Ümit V. Çatalyiirek. Incremental algorithms for closeness centrality. In *2013 IEEE International Conference on Big Data*, pages 487–492, 2013.

[47] R.L. Burden and J.D. Faires. *Análisis numérico*. Cengage Learning, México, 10th edition, (2002).

[48] Weisstein, Eric W. Least squares fitting. Recuperado el 8 de noviembre de 2022, de https://mathworld.wolfram.com/LeastSquaresFitting.html.