UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA QUÍMICA, BIOTECNOLOGÍA Y MATERIALES

EVALUACIÓN DE GRAFOS PARA LA REPRESENTACIÓN DE PROTEÍNAS COMO SOPORTE PARA LA APLICACIÓN DE ALGORITMOS DE INTELIGENCIA ARTIFICIAL EN LA PREDICCIÓN DE EFECTOS DE MUTACIONES Y SITIOS EPISTÁTICOS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL QUÍMICO E INGENIERO CIVIL EN BIOTECNOLOGÍA

JORGE LUIS MUÑOZ SOTO

PROFESOR GUÍA:
Álvaro Olivera Nappa

MIEMBROS DE LA COMISIÓN:
José Salgado Herrera
Barbara Andrews Farrow

SANTIAGO DE CHILE
2023

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL QUÍMICO
E INGENIERO CIVIL EN BIOTECNOLOGÍA
POR: JORGE LUIS MUÑOZ SOTO
FECHA: 2023
PROF. GUÍA: ÁLVARO OLIVERA NAPPA

EVALUACIÓN DE GRAFOS PARA LA REPRESENTACIÓN DE PROTEÍNAS COMO SOPORTE PARA LA APLICACIÓN DE ALGORITMOS DE INTELIGENCIA ARTIFICIAL EN LA PREDICCIÓN DE EFECTOS DE MUTACIONES Y SITIOS EPISTÁTICOS

El estudio de proteínas por medio de enfoques computacionales es de gran interés para diversas industrias que las utilizan debido a la diversidad funcional que estas poseen. Recientemente se han desarrollado modelos que se centran en el uso de grafos para representar estas moléculas, utilizándolos para estudiarlas a nivel estructural. En este contexto se desarrolla la presente memoria que tiene como fin explorar, diseñar e implementar estrategias computacionales para la aplicación de estructuras de grafos como método de representación de proteínas con estructura tridimensional conocida.

El trabajo tiene un fin exploratorio y es dividido en cuatro tareas principales: construir un programa que permita obtener un grafo desde un modelo proteico en formato pdb, realizar detección de comunidades sobre estos, utilizar las comunidades detectadas como unidad de comparación entre proteínas nativas y sus variantes, y generar una metodología para utilizar estas estructuras como inputs en modelos de aprendizaje profundo. La implementación fue llevada a cabo utilizando Python 3.9.7 y diversas librerías como son Networkx 2.6.3, Igraph 0.9.11 y Pytorch geometric 2.0.2.

Los primeros resultados obtenidos fueron cuatro tipos de grafos para representar proteínas: Carbono alfa-distancia, Centroide-distancia, Átomo-distancia e Interacciones intermoleculares. En estudio de comunidades se destaca Spinglass como el algoritmo con mejor modularidad para los grafos de distancia y el que presenta menos resultados anómalos. Por su parte, en los grafos de interacción intermolecular no se puede utilizar este algoritmo debido a que corresponden a estructuras no completas, así es Multilevel el que logra el mejor desempeño con este indicador. Con relación a la comparación de grafos, a partir del análisis de casos puntuales, se nota que los grafos de Centroide-distancia son los más sensibles a redistribución de elementos en comunidades, y que los grados de interacciones intermoleculares podrían ser útiles para identificar la redistribución de redes de puentes de hidrógeno. Respecto al aprendizaje profundo, se programa una red neuronal para clasificar enzimas según su función, entregando una exactitud de 16.3 %, una precisión de 2.7 % y un recall de 16.7 %, dicho desempeño se atribuye a no ajustar los hiperparámetros. A pesar de lo anterior, se destaca la metodología propuesta sobre el comportamiento de la red.

Se concluye que el objetivo exploratorio de la memoria se cumple, obteniendo una representación flexible para la proteínas, que puede ser utilizada en distintas tareas y con el potencial de generar representaciones más completas al incluir más información a las aristas.

A mi familia consanguínea y a mi familia elegida,
gracias por creer.

Agradecimientos

Esta memoria representa un hito culmine en esta etapa de mi vida, la cual no podría haber completado exitosamente sin la ayuda y compañía de varías personas a quienes me gustaría agradecer.

En primer lugar, a José y Lucy, mi padre y mi madre, quienes siempre me han acompañado, cuidado, proveído y amado. Gracias por su cariño incondicional, preocupación constante y la infinita confianza que han tenido en mí. Los amo mucho.

A mi hermano Cristian, quien desde pequeño me inculcó el amor por la ciencia, guió por el proceso de entender el universo y alimentó mi curiosidad. Gracias por tener la paciencia de responder incontables preguntas.

A mi hermano Germán, que me apoyó y enseñó cuando lo necesite y que desde su experiencia de vida me guió. Gracias por los consejos que me diste y la ayuda que me brindaste.

A mi padrino Aníbal, mi madrina Juana, mi tío Ramón y mi tía Nancy, gracias por abrirme las puertas de sus casas y dejarme formar parte de su hogar, donde siempre me sentí bienvenido. Gracias también, por el amor que me dieron y me siguen dando.

A mis primos, Angello y Camilo, quienes fueron mis otros maestros en la ciencia y que me ayudaron en distintas etapas de la vida. Gracias por darse el tiempo de explicarme y guiarme.

A mis primas, Katy y Coni, a quienes considero hermanas del alma. Gracias por todas las conversaciones, los buenos momentos, las risas y la compañía. A Coni especialmente, gracias por estar en las horas de frustración y convertirlas en momentos de distensión, alegría y amor.

A mi mami Minda y mi tía Cristel, quienes siempre se preocuparon por mi bienestar y me enviaron sus buenos deseos y cariños.

A mi segunda familia, aquellas personas que conocí lejos de casa, pero que me acompañaron y ayudaron a volverme una persona más completa. Ampi, Lore, Javi Chasco, Claudio, Javier, Gian, Cata, Felipe, Ulises, Roy, Sandra y Vicho, mis amigos de plan común, a quienes quiero y admiro mucho. Gracias por todas los momentos compartidos, carretes, tardes de juego, viajes y conversaciones, que hicieron que la universidad no fuese solo estudios. Espero haber aportado en su camino tanto como ustedes aportaron en el mío.

Una mención especial a Ampi, Lore y Javi Chasco, excelentes amigas, que me enseñaron, soportaron y dieron un segundo hogar, la huella que dejaron en mi alma me acompañará por

siempre.

A mis amigas de especialidad, Sicely, Macka y Vane, quienes me acompañaron en incontables horas de estudio y en los momentos de celebración. Gracias por las palabras de apoyo, las horas de chancheo y ayudarme a fortalecer habilidades para ser un mejor profesional.

A mis amigas del liceo, Pasi y Tefi, con quienes el contacto ya no es el mismo, pero que me enseñaron como enfrentar la vida desde un prisma más positivo y alegre.

Como no agradecerle también a mi profesor guía Álvaro Olivera, quien ha sido una figura de admiración desde los primeros cursos que rendí con él. Gracias por ser un maestro no solo en lo académico si no también en la vida.

Y a quienes pude haber olvidado, y es muy probable que así sea, porque este camino está formado de experiencias compartidas con muchas personas maravillosas. Gracias por haber formado parte de él ¡que la fiesta no termine!

Tabla de Contenidos

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Motivación y Generalidades

Las proteínas son las biomoléculas más abundante en las células humanas, donde cumplen diversas funciones que van desde lo estructural hasta lo catalítico [1]. Esta característica captó la atención de industrias tan diversas como la textil, la alimenticia y la minera [2]. Así nace la ingeniería de proteínas, una ciencia en constante movimiento que busca mejorar las propiedades de las proteínas para un fin específico. Esta se sirve de distintas estrategias computacionales para el estudio de estas macromoléculas, sus interacciones y la relación de las mutaciones con sus propiedades particulares.

Enfoques recientes de dicho estudio computacional se centran en el uso de modelamiento matemático discreto, usando estructuras de grafos para representar proteínas y comparar cambios a nivel estructural, siendo ejemplos clásicos la utilización de distancias como métricas para definir las conexiones de los átomos representados como nodos [3]. Otros enfoques se basan en entrenar modelos predictivos con técnicas de *graph neural network* para diseñar predictores empleando la información de estas estructuras [4]. Ventajas importantes de emplear estas representaciones se asocian a comparación de pockets estructurales relacionados a sitios de de unión o sitios prostáticos. No obstante, a la fecha y pese a su utilidad, no tenemos una metodología que realmente permita un buen estudio de mutaciones.

1.2. Descripción del proyecto

El proyecto consiste en utilizar estructuras de grafos con distintas propiedades en las aristas (fuerzas intermoleculares, distancias, etc) como representación de proteínas a nivel computacional. Esto facilitaría la identificación de zonas relevantes en la proteína y su comparación con variantes, reconociendo patrones estructurales desde un punto de vista simple y eficiente, que permitan observar los efectos de la mutaciones sobre la proteína.

Lo anterior ya que, la utilización de grafos como representación de las proteínas permite manejar conceptos abstractos como las interacciones de una manera más simple y permite obtener un modelo con información representativa, lo que permitiría generar modelos predictivos más completos. A su vez el desarrollo de estos modelos para el caso específico de las mutaciones puede permitir diseñar mejoras para proteínas de interés industrial, utilizando un enfoque semiracional que seleccionar las mejores mutaciones a comprobar en laboratorio.

Capítulo 2

Objetivos

2.1. General

Explorar, diseñar e implementar estrategias computacionales para la aplicación de estructuras de grafos como método de representación de proteínas con estructura tridimensional conocida.

2.2. Específicos

- Explorar y validar estrategias computacionales para la representación de proteínas con estructuras de grafos, contemplando el diseño de los nodos y las aristas.
- Diseñar, implementar y validar estrategias computacionales para la identificación de patrones en estructuras de proteínas mediante las representaciones de grafos aplicando métodos basados en identificación de comunidades.
- Diseñar, implementar y validar estrategias computacionales para la comparación de grafos con el fin de facilitar la identificación de cambios estructurales de la proteína a nivel de mutaciones o variantes y proponer una posible identificación de sitios epistáticos.
- Explorar, validar y proponer un pipeline de aplicación de arquitecturas *Graph Convolutional Networks* para la implementación de modelos predictivos de clasificación de grafos con uso en tareas de ingeniería de proteínas.

Capítulo 3

Antecedentes

3.1. Proteínas

Las proteínas son macromoléculas biológicas de alto peso molecular que cumplen diversas funciones que van desde un rol estructural hasta uno catalítico o de señalización. Su estructura está compuesta por una sucesión de moléculas llamadas aminoácidos, por lo se le denominan polímeros de aminoácidos [5].

3.1.1. Aminoácidos

Un aminoácido, es una molécula compuesta por un grupo carboxilo ($-CO_2$) y un grupo amino ($-NH_2$) unidos a un carbono central (carbono α), el que también se encuentra enlazado con un hidrógeno y un radical variable [5], como se muestra en la Fig. 3.1.

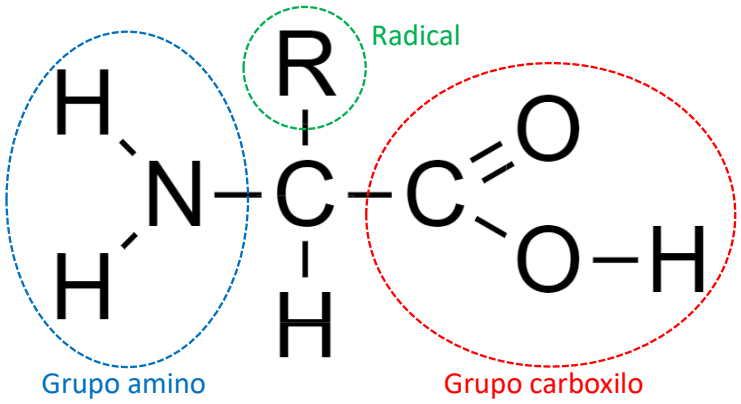


Figura 3.1: Estructura general de un aminoácido, donde R hace referencia al radical.

Dependiendo de radical que posea, el aminoácido tendrá un nombre y propiedades distintas, siendo los presentados en la Fig. 3.2 los que habitualmente se encuentran en las proteínas [5]. En esta se enseña, además, una clasificación según la naturaleza química del radical, siendo esta la que determinará las posibles interacciones entre estas zonas de la molécula.

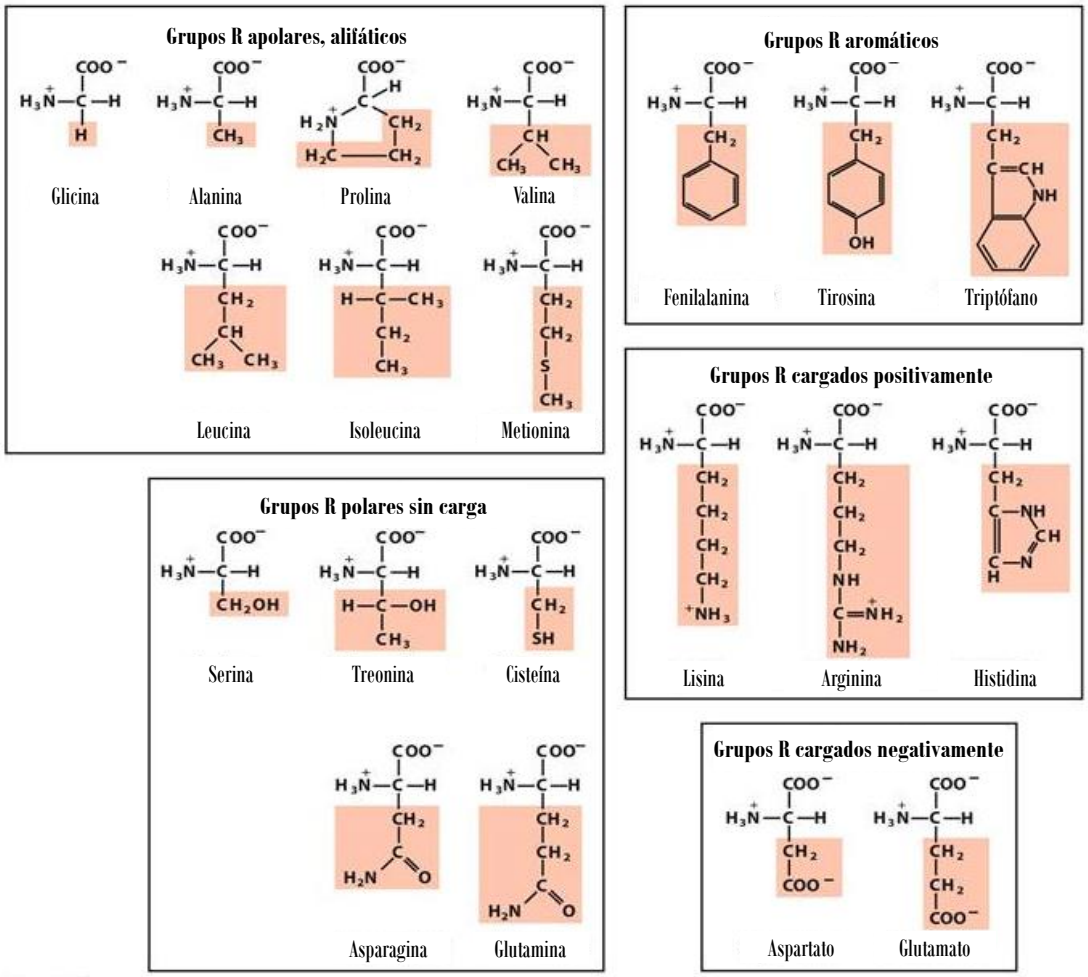


Figura 3.2: Clasificación de aminoácidos más comunes a partir de la naturaleza de su radical. Adaptado de [5].

Cuando se unen dos aminoácidos se forma una molécula denominada dipéptido, continuando con la denominación de péptido hasta que se supero el peso de 10000 Da, luego de lo cual la macromolécula se denomina proteína [5].

3.1.2. Estructura de las proteínas

Como se menciona anteriormente, las proteínas están compuestas por cadenas de aminoácidos, sin embargo, su estructura dista de ser una cadena lineal. En realidad, las interacciones entre sus átomos genera una disposición espacial que puede describirse como una serie de niveles interdependientes que son jerarquizados dependiendo de su complejidad [5].

3.1.2.1. Estructura primaria

La estructura primaria de una proteína corresponde a la secuencia lineal de aminoácidos que la componen, unidos por medio de enlaces peptídicos [5].

Un enlace peptídico es una enlace covalente formado a partir de una reacción de condensación en la que participa en grupo amino de un aminoácido y el grupo carboxilo de otro, luego

de lo cual los aminoácidos se denominan residuos aminoacídicos, como se puede observar en la Fig. 3.3. Cabe destacar que existe resonancia entre el oxígeno carbonílico y el nitrógeno amida, lo que provoca que el enlace sea rígido y plano, por lo que los residuos enlazados son coplanares [5].

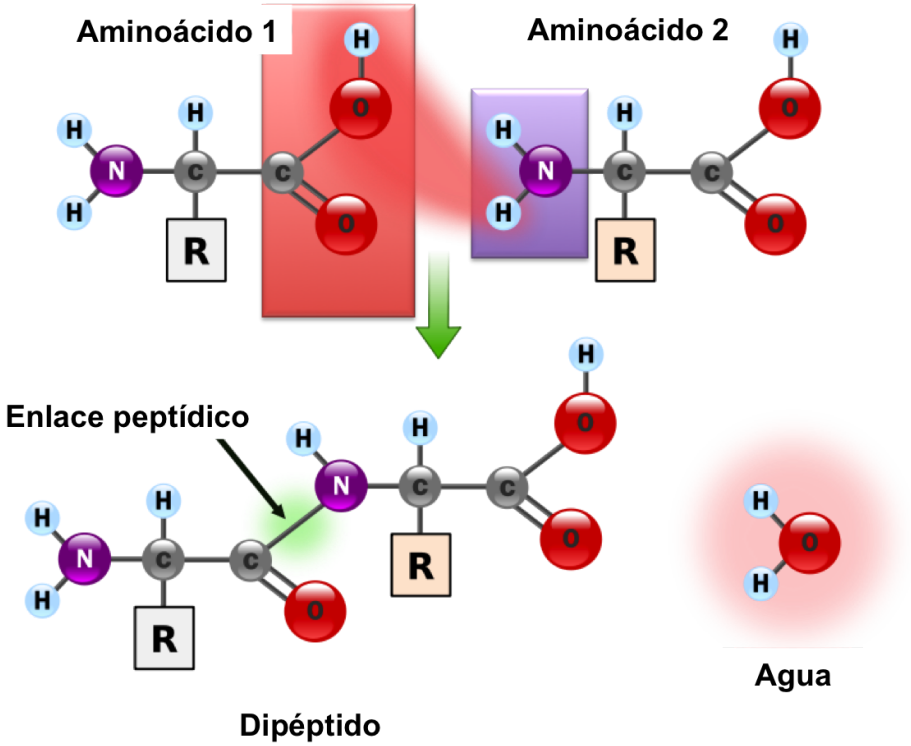


Figura 3.3: Reacción de condensación de dos aminoácidos que forma un enlace peptídico. Obtenido de [6].

3.1.2.2. Estructura secundaria

La estructura secundaria de una proteína corresponde a la distribución espacial local del esqueleto proteico (cadena de principal de carbonos sin considerar los residuos laterales), formada debido a la interacción entre los grupos del enlace peptídico por medio de puentes de hidrógeno [5].

Un puente de hidrógeno es una interacción atractiva (caso especial de una interacción dipolo - dipolo) que se establece entre un hidrógeno unido covalentemente a un átomo altamente electronegativo (aceptor) y un átomo electronegativo que regularmente posee un par de electrones no enlazados (donor). La energía contenida en esta interacción depende del tipo de donor y aceptor involucrados, así como de la geometría de ésta.

Las estructuras secundarias más comunes son:

- Hélice α (ver Fig. 3.4.a): Esta corresponde a una estructura helicoidal dextrógira en que los residuos se orientan hacia el exterior de esta. Cada vuelta de la hélice está conformada por 3.6 aminoácidos, que corresponde a aproximadamente 5.4 Å a lo largo del eje axial de la estructura. La naturaleza de los residuos aminoacídicos definirá en gran medida la

estabilidad de esta estructura, volviéndose inestable al tener contiguos aminoácidos de gran tamaño debido al efecto estérico, o aminoácidos con el mismo tipo de carga debido a la repulsión generada. Además, los aminoácidos prolina y glicina son poco proclives a formar parte de está estructura, el primero debido a que su geometría aporta una torsión indeseada a la hélice y el segundo ya que su flexibilidad característica genera un efecto desestabilizador. Finalmente, otro factor que influye en la estabilidad es el tipo de aminoácidos que forman parte del extremo amino y carboxilo terminal de la proteína, esto porque los enlaces de peptídico aportan un dipolo que crece con el largo de la hélice debido a los puentes de hidrógeno, lo que provoca que estructura este polarizada, existiendo un carga parcial positiva en el extremo amino terminal y una negativa en el carboxilo terminal, la cuales pueden ser estabilizadas si cerca de estos extremos se ubican aminoácidos con carga contraria o desestabilizadas si poseen la misma carga [5].

- Lamina u hoja β (ver Figs. 3.4.b y 3.4.c): Esta corresponde a la interacción entre varias hebras o conformaciones β adyacentes, siendo una hebra β un esqueleto polipetídico extendido en zigzag. Así, los puentes de hidrógeno se forman entre los enlaces peptídicos de distintas hebras, que pueden formar parte de una misma o de distintas cadenas polipeptídicas. En este caso, los residuos sobresalen de la cadena alternadamente hacia arriba y abajo de la lamina. Dependiendo de como se ubiquen las hebras β adyacentes, la lamina β puede ser paralela o antiparalela, siendo el primer caso cuando se tiene la misma orientación carbono-amino terminal y el segundo cuando esta es opuesta, lo que genera distinta orientación para los puentes de hidrógeno siendo lineales para la última y no lineales para la primera [5].

- Giro β (ver Fig. 3.4.d): Esta corresponde a una segmento de cadena de cuatro residuos que genera un giro cerrado de 180. En esta, el grupo carbonilo de primer residuo del giro forma un puente de hidrógeno el grupo amino con el cuarto. Los aminoácidos más comunes que forman parte de esta estructura son la prolina debido a la torsión que aportan y la glicina por su flexibilidad y se presentan comúnmente en la conexión de hebras β que forman una lamina β antiparalela [7].

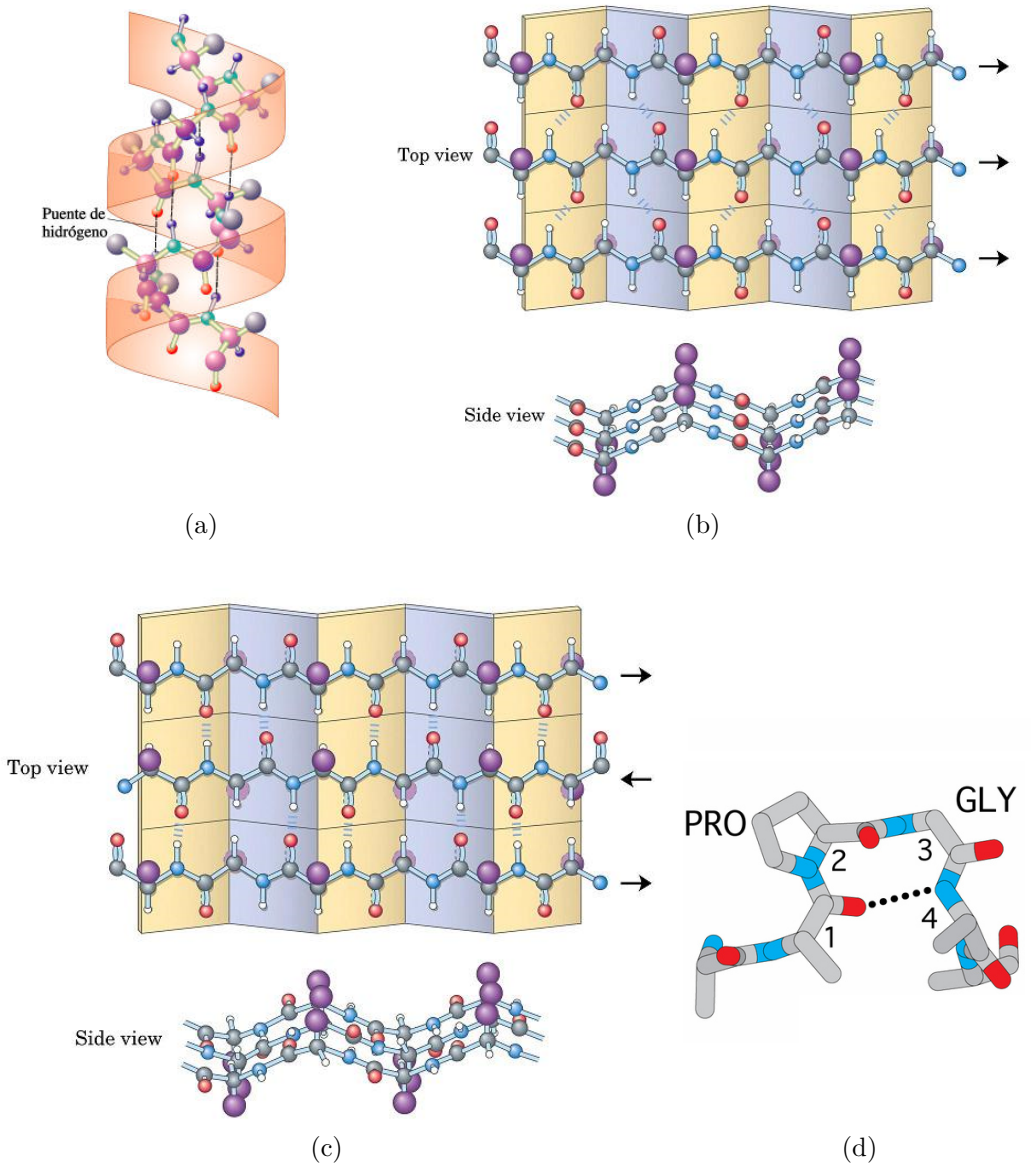
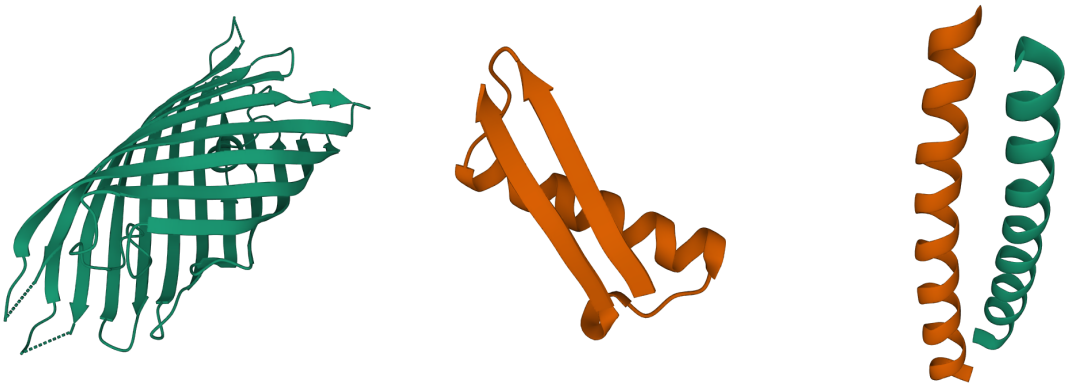


Figura 3.4: Esquemas de las estructuras secundarias más comunes. (a) Hélice α. (b) Lamina β paralela (b) Lamina β antiparalela. (c) Giro β. Obtenidos de [8] y [9]

A la combinación o repetición de estructuras secundarias que se encuentran de manera común en distintos tipos de proteínas se le llama *estructura supersecundaria* o *motivo estructural*, siendo ejemplos de estos los presentados en la Figura 3.5.



(a) (b) (c)

Figura 3.5: Ejemplos de estructuras supersecundarias. (a) Barril β, compuesto por la repetición de hojas β (proteína con id pdb: 2QOM [10]) (b) β-α-β, compuesto por la combinación de hojas β y hélices α (parte de la proteína con id pdb: 1GIF [11]) (c) Hélice superenrollada, compuesto por la interacción de hélices α (proteína con id pdb: 1ZIK [12]). Visualizaciones generadas con RCSB [13]

3.1.2.3. Estructura terciaria

La estructura terciaria de una proteína corresponde al plegamiento tridimensional global que adopta una cadena polipeptidíca, el que ocurre y se mantiene debido a distintos tipos de interacciones débiles que se establecen entre los radicales de los residuos, siendo las principales:

- Puente salino: Es la atracción que se establece entre dos moléculas con carga neta. Dándose entonces entre los residuos con carga positiva y negativa.
- Puente de hidrógeno: Como ya fue descrito en la Sección 3.1.2.2, se establece entre un hidrógeno unido covalentemente a un átomo altamente electronegativo y un átomo electronegativo que regularmente posee un par de electrones no enlazados. Por lo que se puede formar entre los aminoácidos polares sin carga (excluyendo la cisteína), los cargados positivamente, los cargados negativamente y dos de los aromáticos: tirosina y triptófano.
- Interacciones de Van de Waals: Son un conjunto de interacciones que incluyen las interacciones dipolo - dipolo, dipolo - dipolo inducido y dipolo inducido - dipolo inducido, estableciéndose dependiendo de la polaridad de la molécula, siendo la primera entre dos moléculas polares, la segunda entre una molécula polar y una apolar y la última fundamentalmente entre moléculas apolares. Estas si bien se dan entre todas las moléculas son más relevantes en los aminoácidos apolares, ya que no establecen otras interacciones más fuertes.
- Interacciones de nube π: Corresponde a la interacción de los orbitales de un sistema π rico en electrones con otro sistema. Son establecidas entre los animoácidos aromáticos ya que su anillo presenta electrones deslocalizados en los orbitales p lo que genera una nubes ricas en electrones bajo y sobre el anillo y una zona con baja densidad electrónica en el mismo anillo, pudiendo así interactuar el contorno de un anillo con la nube de otro.

Además existe la posibilidad de que la estructura presente puentes disulfuro, los que corresponden a enlaces covalentes entre azufre, pudiendo por esto solo ser formado entre cisteínas.

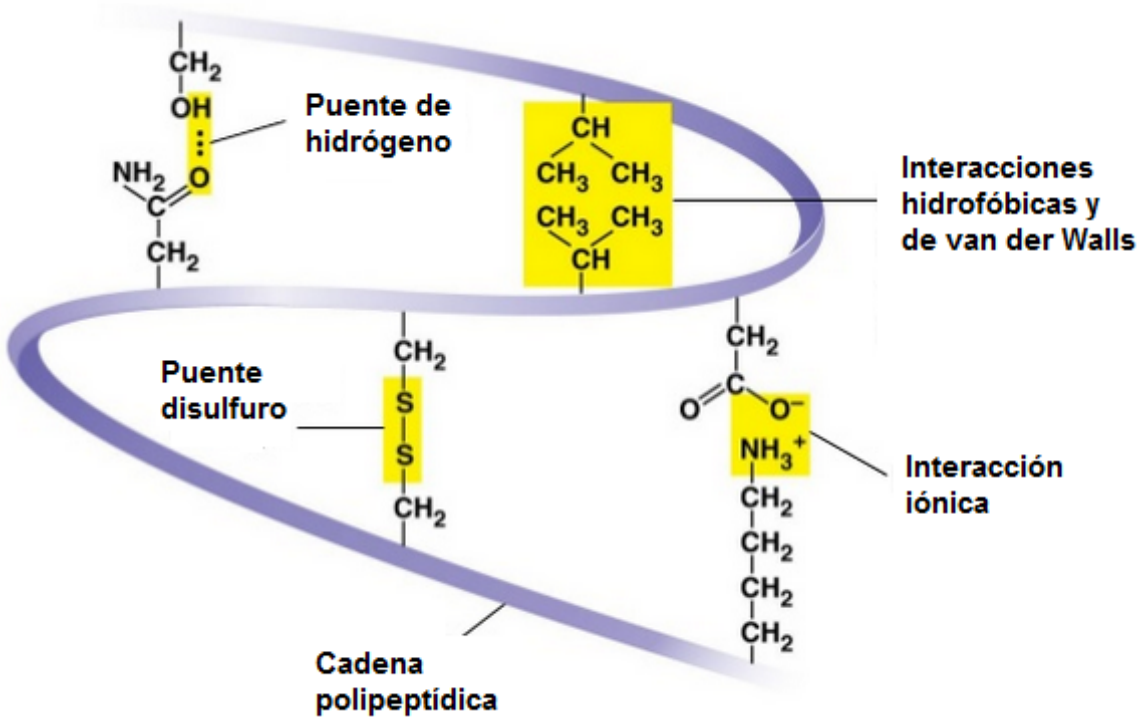


Figura 3.6: Diagrama que presenta las principales interacciones que mantienen las estructura terciaria. Obtenida de [6].

La estructura terciaria es de especial relevancia ya que es esta la que determina la función de la proteína, produciéndose una perdida de esta al perder dicha estructura (fenómeno denominado desnaturalización). Así, proteínas como las enzimas dependen de sitios determinados por su estructura terciaria para poder llevar a cabo su función catalítica (sitios activos) o de otros sitios para su regulación (sitio alostérico), las proteínas transportadoras cuentas con un sitio de unión para un sustrato indicado y otras proteínas como las fibrosas que cumplen funciones estructurales (e.g. colágeno) deben a esta estructura su rigidez o flexbilidad [5].

3.1.2.4. Estructura cuaternaria

La estructura cuaternaria de una proteína existe cuando esta está compuesta por dos o más cadenas separadas (subunidades), siendo este nivel organizacional el asociado al ordenamiento e interacción de ellas. La interacción entre subunidades puede provocar cambios en la actividad de la proteína, existiendo un fenómeno denominado cooperatividad, en el la unión de un substrato a un sitio presente en alguna subunidad genera un cambio de conformación que afecta la afinidad por el sustrato de los sitios restantes en las otras subunidades [5].

3.1.2.5. Técnicas para la determinación de la estructura terciaria

La estructura terciaria de una proteína se puede determinar a partir de métodos experimentales o computaciones Las principales técnicas experimentales utilizadas para esto:

- Cristalografía de rayos X: Esta técnica consiste en bombardear un sólido cristalino con rayos X, donde se produce difracción de los rayos, ya que la longitud de onda de esta radiación es del mismo orden que las distancias interatómicas, existiendo interferencia constructiva y destructiva debido al ordenamiento del cristal que permiten obtener una señal amplificada. Así, cada cristal genera un patrón de difracción que contiene información de la densidad electrónica del sólido, el que es captado por un detector y luego por medio de análisis computacional se genera un modelo de la estructura de la proteínas [14]. Una de las principales limitantes asociadas a esta técnica es la inhabilidad de cristalizar ciertas proteínas, lo que provoca que no se pueda estudiar su estructura con este método.

- Resonancia nuclear magnética: Esta técnica consiste en someter una solución de un compuesto a un campo magnético estable, lo que genera que los momentos magnéticos de los núcleos atómicos se alineen con este. Luego, se bombardea con pulsos de ondas de radio, lo que produce que los núcleos con un spin distinto a cero (y por tanto con momento magnético) resuenen a una frecuencia característica, que estará influenciada tanto por el núcleo que resuena como por su entorno químico inmediato. Se obtiene entonces, un espectro que puede ser traducido a un modelo estructural por medio de un análisis computacional [14]. Si bien está técnica no requiere de la cristalización de las proteínas, si requiere que la muestra a analizar posee un elevado nivel de pureza (mayor al 97 % [15]). Además, se requiere que los átomos que la proteínas de estudio posean un spin nuclear, el cual es 0 para C^{12}, N^{14} y O^{16}, por lo que en el caso de las proteínas que provienen de fuentes naturales el estudio se realiza solo en torno a los H. Esta última dificultad se puede subsanar en proteínas recombinantes las que pueden ser marcadas con formas isotópicas de carbono y nitrógeno que poseen un spin nuclear, lo que permite obtener una mejor resolución de la proteína [16]. Otra desventaja de esta técnica es que mientras mayor sea el peso de la proteínas mayor serán la intensidad del campo magnético requerido para separar los picos del espectro. Por lo cual se limita la separación efectiva de picos espectrales para proteínas de hasta 28-30 kDa.

- Criomicroscopía electrónica: Esta técnica consiste en una etapa de pretratamiento en que se somete a la solución acuosa a un proceso de vitrificación que genera una capa de hielo amorfo que contiene a los compuestos a estudiar (sólido amorfo). A continuación, la muestra vitrificada es bombardeada por un haz de electrones que interactúan con ella, los electrones emergentes son convertidos por medio de un sistema de lentes ópticos de electrones en imágenes en un detector, las que luego permiten obtener un modelo de la estructura molecular. Una de las desventajas de esta técnica es que se debe mantener el sólido amorfo, por lo que se debe trabajar a bajas temperaturas (-135°C) y este se debe formar de manera correcta ya que si la estructura presenta cristales de hielo generarían grandes obstáculos en las imágenes obtenidas [17].

Las principales técnicas computacionales utilizadas para esto:

- Modelado por homología: Consiste en predecir el plegamiento tridimensional de una secuencia proteica a partir de la estructura de proteínas homólogas. Dos proteínas son homólogas cuando poseen secuencias similares debido a que presentan un origen evolutivo en común. Así este método se basa en la noción de que la estructura tridimensional es fuertemente conservada entre proteínas homólogas [18].

- Modelado por *threading*: Consiste en predecir el plegamiento de la proteína a partir de su comparación con una librería de plegamientos, alineando cada aminoácido de la secuencia con una posición del plegamiento, permitiendo *gaps*, y luego evaluando el ajuste por medio de una función de puntuación [19].
- Modelado *ab initio*: Consiste en predecir el plegamiento de la proteína a partir de su estructura primaria usando principios fisicoquímicos y funciones de energía adecuadas.

3.1.3. Funciones de las proteínas

La variedad de aminoácidos y posibles combinaciones entre ellos aportan a las proteínas una gran versatilidad, llevando a cabo un amplio rango de funciones esenciales para la mantención y el bienestar de los seres vivos. Algunas de las más relevantes se listan a continuación:

- Estructural: A nivel celular las proteínas forman parte de distintas estructuras celulares, como el glicocálix en forma de glicoproteínas, los ribosomas que son conformados por la unión de proteínas y ARNr, el citoesqueleto compuesto por proteínas que le dan soporte a las células, entre otras. A nivel de tejido, proteínas como el colágeno aportan soporte y elasticidad formando parte de la matriz extracelular.
- Enzimática: Las enzimas corresponden a un tipo de proteínas que catalizan reacciones químicas, esto ya que poseen un sector especifico en su estructura denominado sitio activo que les permite unirse a un ligando y llevar a cabo la catálisis. Existen una gran variedad de enzimas, en la que se incluyen las transferasas, oxidoreductasas, hidrolasas, entre otras.
- Señalización: Algunas células secretan proteínas que luego se unirán a los receptores de otras células, permitiendo así una coordinación entre estas. Un caso específico de esta función son las hormonas de origen proteico que permiten transmitir señales y de esta manera regular las funciones de distintos órganos o sistemas.
- Transporte: Algunas proteínas, como la ferritina y la hemoglobina, se unen a otras moléculas y las transportan dentro del cuerpo o hacia células.
- Inmunológica: Los anticuerpos son un tipo de proteína cuya estructura les permite reconocer y unirse a moléculas extrañas, lo que activará mecanismos del sistema inmune, cumpliendo una función de defensa.

3.1.4. Mutaciones

Las mutaciones son cambios en la secuencia del ADN, los que se pueden traducir en un cambio en la proteína que se codifica. A nivel génico estas se pueden clasificar en:

- Deleciones: Corresponden a la eliminación de uno o más nucleótidos de la secuencia.
- Inserciones: Corresponden a la adición de uno o más nucleótidos a la secuencia.
- Sustituciones: Corresponde al cambio de uno o más nucleótidos por otros (el mismo número).

Las deleciones e inserciones pueden producir un cambio en el marco de lectura si es que no son los nucleótidos en cuestión no son un múltiplo de tres, lo cual se traduciría en una cambio significativos en los aminoácidos que componen la proteína, pudiendo perder su actividad. En cuanto a las sustituciones, como el código genético es degenerado, es posible que no afecten la secuencia de la proteína si es que se reemplaza un codón por otro que codifique la misma proteína (mutación silenciosa) e incluso cuando se sustituya una aminoácido por otro, si el nuevo radical establece interacciones similares al original es posible mantener minimizar el efecto en la estructura tridimensional (reemplazo conservativo).

3.2. Ingeniería de proteínas

Considerando la gran cantidad de funciones que poseen las proteínas no es de extrañar que sean moléculas de interés para distintas industrias como la alimenticia, farmacéutica, textil, entre otras. En este contexto nace la ingeniería de proteínas que corresponde a una rama de la ingeniería encargada de desarrollar proteínas útiles o valiosas para un fin específico [2]. Para esto se ocupan principalmente los siguiente enfoques.

3.2.1. Diseño racional

El diseño racional es un enfoque en que se modifican proteínas en base al conocimiento que se tiene de su relación estructura-función [20].

Generalmente, se estudia la estructura tridimensional de la proteína, los cambios conformacionales que sufre y los aminoácidos más relevantes para su función. A partir de este estudio, se identifica el o los aminoácidos que se quieren sustituir y por cuales, considerando como el cambio de los residuos podría afectar la función de la proteína [20]. Luego, se materializan las sustituciones utilizando, generalmente, mutagénesis sitio dirigida, esta técnica consiste en realizar una PCR utilizando *primers* mutados que contengan el cambio en el codón que codifica para el aminoácido a sustituir. Finalmente, la proteína es sintetizada utilizando microorganismos recombinantes y purificada, para evaluar si se obtuvo una mejora en la propiedad deseada. En caso de que la mejora no se haya logrado, se determina experimentalmente la estructura de la proteína recombinante y con esta se intenta explicar el motivo por el que la sustitución no entregó el resultado esperado, utilizando esta información en una nueva etapa de diseño. Así, este enfoque nos permite profundizar el conocimiento sobre las proteínas y sus mecanismos [2].

La principal desventaja asociada con este enfoque es la necesidad de conocer la estructura terciaria de la proteína y el rol que cumplen los aminoácidos en su función. Por lo que no es aplicable para proteínas que no poseen una estructura terciaria estable (proteínas intrínsecamente desestructuradas) o sobre las que no se posee un gran entendimiento de su relación estructura-función, ya que no se podría realizar una predicción efectiva [2].

Actualmente, este enfoque ha sido aplicado para el diseño de anticuerpos y de péptidos terapéuticos como Proleukin® y Betaseron® [21], donde sus estructuras nativas fueron modificadas sustituyendo las cisteínas libres en posiciones enterradas (protegido en el núcleo hidrofóbico) por serinas, logrando una disminución en su agregación y mejorando su vida útil [22].

3.2.2. Evolución dirigida

La evolución dirigida es un enfoque en que se mejoran las características de una proteína realizando un proceso que sigue los principios de la evolución darwiniana (variabilidad genética, ventaja competitiva y heredabilidad), induciendo mutaciones al azar y seleccionando aquellas que presenten las mejoras deseadas [2].

Este proceso está compuesto por tres etapas principales, inicialmente se realiza mutagénesis al azar, induciendo mutaciones en la secuencia de ADN que codifica la proteína. Para

lo anterior, generalmente, se utiliza PCR propensa a error, siendo esta una técnica de amplificación de ADN que utiliza una Taq polimerasa en un buffer en presencia de Mn^{2+} y un desbalance en la concentración de nucleótidos, lo que disminuirá la fidelidad con que se replica el ADN [23], generando así una librería de variantes. Luego, se expresan las proteínas utilizando microoganismos recombinates y se evalúa la actividad de estas, seleccionando aquellas que posean un mejora en la característica deseada. Finalmente, los genes que codifican para las proteínas con mejor desempeño se aíslan y se pueden someter a un nuevo ciclo de evolución [24].

Este método no requiere un conocimiento intensivo de la proteína y su estructura, lo que permite aplicarlo en proteínas cuya estructura terciaria no puede ser obtenida por métodos tradicionales. Sin embargo, está limitado por la necesidad de métodos de evaluación de desempeño (*screening*) y selección de alto rendimiento, debido a la amplitud de las librerías con las que se trabaja [25]. Estos regularmente son específicos para cada experimento y se aplican en una gran escala por lo que podrían traducirse en un alto costo.

Es importante notar que a pesar de que no se conoce directamente la razón por la que la sustitución mejora el desempeño de la proteína, el método si permite identificar posiciones de interés en que ocurre esto, las cuales posteriormente se pueden estudiar o sobre las que se pueden aplicar otras técnicas como mutagénesis de saturación [25].

Una investigación insignia en que se ocupo este enfoque es la realizada por Frances Arnold para mejorar el desempeño en solventes organoacuosos de una paranitrobencilo esterasa. El método de *screening* seleccionado utiliza paranitrofenil ester, un sustrato análogo al sustrato de interés (paranitrobencilo ester), debido a la incapacidad de realizar un *screening* rápido si se usaba el sustrato original. Aplicando solo evolución dirigida, se logró luego de cuatro ciclos, obtener una actividad 16 veces mayor que la original con 30 % de dimetilformamida en el medio [26].

Cabe destacar que si se selecciona un método de *screening* que utilice sustratos análogos al original, ya que no se conoce el mecanismo por el que mejora la enzima, se puede lograr una mejora de desempeño específica solo para el primero, no cumpliendo con el objetivo del proceso.

3.2.3. Diseño semiracional

Tambien llamado evolución dirigida focalizada, el método de diseño semiracional es una combinación del diseño racional y la evolución dirigida, aplicando el conocimiento e información que se posee de la relación estructura-función de la proteína para focalizar la evolución dirigida en zonas en que se cree se obtendrá una mayor cantidad de mutaciones benéficas [27].

Este enfoque realiza un estudio del estructura tridimensional de la proteína, identificando zonas de interés. Luego, generalmente, se inducen mutaciones en esta zona utilizando la técnica de mutagénesis de saturación del sitio, la que utiliza *primers* mutados para sustituir el codón que codifica para el aminoácido en la posición de interés por un codón que codifique para uno de los otros 19 aminoácidos, variando los *primers* de tal manera que existan variantes de la proteína con los 20 aminoácidos en el sitio deseado [28]. Con esto se obtiene un

set o librería focalizada que es más pequeño que el obtenido con evolución dirigida. Luego se expresan las proteínas utilizando microoganismos recombinantes y se evalúa la actividad de estas, seleccionando aquellas que posean una mejora en la característica deseada. Finalmente, se aíslan los genes que codifican para las proteínas seleccionadas [29]. De esta manera al combinar ambos enfoques se subsana en parte las limitaciones de ambas. Por una parte, no se requiere un conocimiento tan profundo de la enzima sino que el suficiente para saber qué sitio podría ser de interés, sin tener la necesidad de predecir el efecto de la mutación de antemano. Además, al reducirse la librería de mutantes, permite realizar el proceso de *screening* a una escala más baja [27].

Un ejemplo de la aplicación de este de enfoque es la mejora de la enantioselectividad y actividad de una esterasa de *Pseudomonas fluorescens* donde se seleccionaron cuatro aminoácidos cercanos al sitio activo y se determinaron las mutaciones a aplicar por medio de un software computacional que entrega la distribución de aminoácidos para la superfamilia que contiene a la proteína, identificando así los aminoácidos más comunes en posiciones análogas a las deseadas, y que por ende probablemente no producirán un mal plegamiento, así se realiza se realiza mutagénesis saturando las posiciones con los aminoácidos que se creen favorables, minimizando así la librería de variantes [30].

3.3. Métodos computacionales para proteínas

Como se menciona en secciones anteriores, las técnicas computacionales son utilizadas para diversas tareas relacionadas con proteínas que vas desde la creación de un modelo estructural a partir de datos experimentales, predecir una la estructura tridimensional de una proteína [31] u optimizar las mutaciones que se llevarán a cabo en un diseño con enfoque semiracional. Es en este contexto que es vital comprender el concepto de bioinformática, que es el diseño y aplicación de herramientas computacionales para la adquisición, almacenamiento, análisis y diseminación de información biológica [32]. Una herramienta utilizada en la actualidad es el Machine Learning, que es una rama de las ciencias en computación que busca desarrollar algoritmos que permitan el aprendizaje por parte de las maquinas a partir de los datos, en lugar de la programación explicita [33].

3.3.1. Estrategia de representación de proteínas

Si se quiere utilizar para la estructura proteica en algoritmos de ML es necesario contar con una manera adecuada de representarla computacionalmente. Una forma de clasificar las representaciones es a partir de si se conoce o no su estructura terciaria. En caso de que no se conozca y se posea únicamente la secuencia de aminoácidos que la conforma, se pueden ocupar varias formas para representarlas. Entre estas se encuentran: One-hot encoding [4], Natural Lenguage Processing [4] y Digital Signal Processing [34].

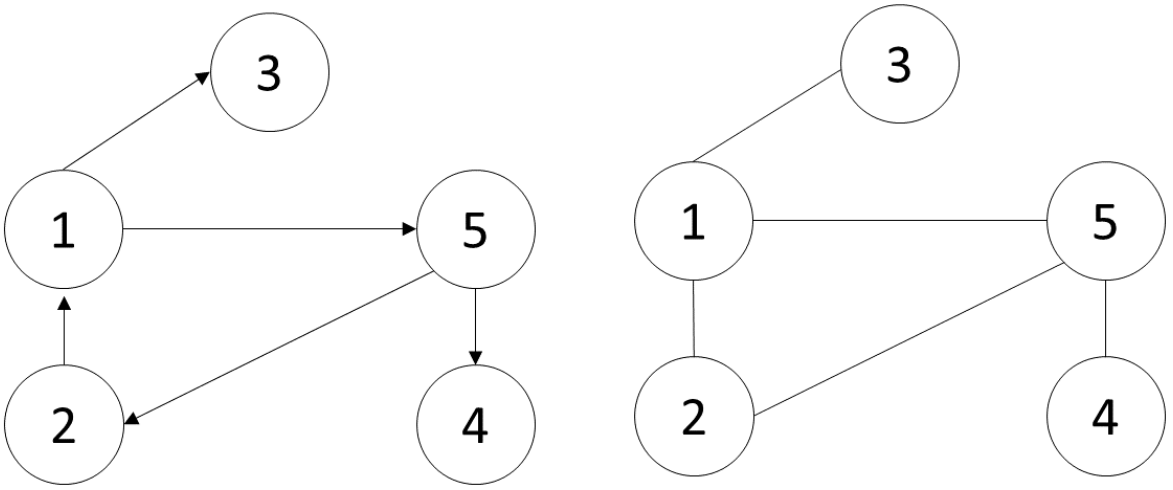
One-hot encoding corresponde a una codificación en que un elemento se representa como un vector con un bit alto (1) en una posición y todos los demás bajos (0), variando la posición en que se encuentra el bit alto según el elemento. Así, para los aminoácidos se tendrían 20 vectores distintos uno para cada uno, y las proteínas se representarían como un array de estos vectores. Una desventaja de esta representación es su *sparsity* lo cual que la hace ineficiente [4].

Natural Language Processing (NLP) son un grupo de algoritmos que buscan que los computadores puedan entender textos de la manera en que las personas lo hacen. BioVec es un modelo basado en los algoritmos de NLP Word2Vec que permite interpretar secuencias de tres aminoácidos no superpuestas como palabras, considerando entonces la secuencia de aminoácidos com una secuencia de palabras con un contexto del que se puede aprender [4].

Signal Processing considera los aminoácidos y sus propiedades asociadas. De está manera una proteína queda descrita por una secuencia de propiedades que es codificada y luego transformada en una señal por medio de la aplicación de la Transformada de Fourier [34].

Si se posee un modelo de la estructura terciaria, los grafos se presentan como una buena estrategia de representación [4], permitiendo representar interacciones de una manera sencilla e intuitiva, lo que se traduce en una ventaja frente a las otras formas presentadas. Los grafos son conjuntos de pares $G = (V, E)$ donde V son los vértices o nodos que representan objetos y E las aristas que relacionan a dichos objetos. Los grafos se puede ser dirigidos o no dirigidos, en los primeros sus aristas poseen una dirección, siendo entonces unilaterales. Así, si un nodo está relacionado con otro la relación inversa no necesariamente existe. Por otra parte, en los grafos no dirigidos las aristas no poseen dirección, por lo que todas las relaciones son

bilaterales. Una representación gráfica se puede observar en la Fig. 3.7, donde los números corresponden a los vértices y las rectas a las aristas [35].



(a) (b)

Figura 3.7: Diagrama de los distintos tipos de grafos (a) dirigido y (b) no dirigido.

En el caso de proteínas, estas se pueden representar con distinto nivel de granularidad. En un grado más fino se pueden utilizar los átomos como nodos y sus interacciones (distancias, interacciones débiles u otras propiedades) por las aristas [3]. Sin embargo, esto significaría un alto costo computacional para proteínas con secuencias muy largas. Considerando un modelo *coarse grained* (grano grueso), los residuos pueden ser representados por los nodos y sus interacciones por las aristas, obteniendo de esta manera un modelo más simplificado, pero que permite disminuir el costo computacional. La elección del nivel de granularidad es función de los objetivos del estudio y la capacidad computacional disponible [4].

Como se mencionó anteriormente, las aristas de los grafos pueden representar distintas interacciones entre los residuos o átomos, como distancia o interacciones débiles. Sin embargo, relacionar todos los nodos puede significar un alto nivel de ruido, por lo que un criterio de creación de aristas basado en un umbral mínimo puede ser útil para disminuirlo [3].

3.3.2. Implementación de grafos

Comúnmente, un grafo se puede representar computacionalmente por medio de una lista de adyacencia o una matriz de adyacencia. Las primeras corresponden a un *array* en que cada elemento corresponde a una lista que indica los elementos que se relacionan con el nodo asociado al índice del array [35]. Por ejemplo, si el nodo asociado con el indice 0 del array se relaciona con los nodos 1 y 4, el elemento que se encontraría en la posición 0 del array sería una lista con 1 y 4. Además, si las aristas poseen una peso w_{ij} cada uno de los elementos de la lista estaría compuesto por el nodo con el que se interacciona y dicho valor. Las segundas

son matrices A de tamaño NxN , donde N es el número de nodos del grafo, considerando un elemento A_{ij} su valor serán 1 si existe relación entre el nodo asociado al índice i y el asociado a j, y 0 si no. Al igual que en las listas de adyacencia, las aristas pueden poseer la intensidad de la interacción w_{ij}, lo que se traduce en elementos que corresponden a pares ordenados del tipo (1, w_{ij}) o (0, w_{ij}) [35].

Dependiendo del tipo de relaciones que existan se puede preferir utilizar una estructura o la otra, siendo preferible la matriz de adyacencia cuando existen relaciones entre todos o casi todos los nodos, teniendo un sparsity bajo; en el caso contrario son preferidas las listas.

3.3.3. Detección de comunidades en grafos

Sobre los grafos se pueden aplicar algoritmos de detección de comunidades que permiten comprender cómo se agrupan los nodos dentro de estos. En este contexto es importante comprender el concepto de modularidad (Ec. 3.1), siendo esta la fracción de las aristas que pertenecen a un grupo luego de dividirlos en una comunidad, menos la fracción esperada si se dividieran de manera aleatoria. Así, esta es una métrica de qué tanto se dividió el grafo en comunidades [36]. Dentro de los algoritmos previamente evaluados para una interacción proteína-proteína se encuentran los algoritmos de Louvain, Spinglass y Fast Greedy [37].A demás pueden ser usados para reconocer distintos motivos y sitios importantes en la estructura terciaria de una proteínas.

$$Q = \frac{1}{2w} \sum_{ij} \left(w_{ij} - \frac{w_i w_j}{2w} \right) \delta(c_i, c_j) \tag{3.1}$$

En la Ec. 3.1 w representa la suma de los pesos de todos las aristas, w_i el *strenght* del nodo i (i.e suma de los pesos de las aristas adjuntas al nodo), w_{ij} es el *strenght* de la arista que une al nodo i y j y $\delta(c_i, c_j$ es una función que toma el valor de 1 si el nodo i y j pertenecen a la misma comunidad y 0 si no 1 [38].

Una descripción de siete algoritmos comunes implementados en la librería igraph es presentada en la Tabla 3.1.

Tabla 3.1: Descripción de distintos algoritmos de detección de comunidades. Adaptación al español de lo presentado en el [39].

Algoritmo	**Descripción**	**Referencias**
Fast greedy	Cada nodo inicia en una comunidad individual. Luego se busca el par de comunidades que al unirse generen la máxima mejor de modularidad. Esto se repite hasta que ninguna unión de comunidades aumente la modularidad.	[40]

Tabla 3.1: Descripción de distintos algoritmos de detección de comunidades (cont.). Adaptación al español de lo presentado en el [39].

Algoritmo	**Descripción**	**Referencias**
Infomap	Emplea paseos aleatorios en la red para analizar el flujo de información en esta. Comienza codificando la red en módulos de manera de maximizar la cantidad de información de la red original. Entonces envía la señal a un decodificador a través de un canal de capacidad limitada. El decodificador intenta decodificar el mensaje y construir un set con los posibles candidatos para el grafo original. Mientras menos candidatos, más información ha sido transmitida.	[41]
Label propagation	Asume que cada nodo forma parte de la misma comunidad que la mayoría de sus vecinos. Comienza inicializando una etiqueta (comunidad) distinta para cada nodo. Luego, los nodos son listados en un orden secuencial al azar. A continuación, según la secuencia, cada nodo toma la etiqueta de la mayoría de sus vecinos. Esto se repite hasta que cada nodo tenga la misma etiqueta que la mayoría de su vecindad.	[42]
Leading eigenvector	Primero se calcula el vector propio principal de la matriz de modularidad, y entonces el grafo es dividido en dos partes de manera que la mejora de modularidad es maximizada basada en el vector propio principal. Después de eso, la contribución de modularidad es calculada en cada paso de la subdivisión de la red. Se detiene cuando esta no es positiva.	[43]
Multilevel	Cada nodo inicia en una comunidad individual. Luego un nodo es integrado a la comunidad del vecino con el que alcance la mayor contribución positiva de modularidad. Esto se repite para todos los nodos hasta que no se deje de lograr una mejora. Entonces, utilizando las comunidades como unidades, se unen al igual que se hizo previamente hasta que se obtiene una sola comunidad o la modularidad no se puede aumentar en una iteración.	[44]
Spinglass	El principio básico de este método es que las aristas deberían conectar nodos con el mismo estado de spin (representado por la comunidad en este caso); al contrario, nodos de diferentes estados deberían estar desconectados. Así, el objetivo de este algoritmo es encontrar el estado basal de un modelo de vidrio de spin (spin glass) con un hamiltoniano de Potts.	[45]
Walktrap	La idea básica de este método es los paseos aleatorios de corta distancia tienden a mantenerse dentro de la misma comunidad. Comenzando con una partición sin comunidades la distancia entre todos los nodos adyacentes es calculada. Luego, dos comunidades son elegidas, se unen en una nueva y las distancias son actualizadas.	[46]

3.3.4. Graph Neural Network

Una *Neural Network* es una serie de algoritmos cuyo fin es encontrar patrones y relaciones en un set de datos, que basa en la estructura de una red neuronal biológica con distintas neuronas artificiales (funciones) que someten al input a una transformación y cuyo resultado se transmite a otra hasta que se llega a la salida [47]. Específicamente, *Graph Neural Network* hace referencia a aquellas que operan sobre grafos [48]. Esta redes pueden llevar a cabo distintas tareas, dentro de las que se encuentran la clasificación de nodos, permitiendo asignar una característica a los nodos de un grafo; la predicción de relaciones, que busca determinar si dos entidades poseen una conexión y la clasificación de grafos, donde se asigna una categoría al grafo completo [49].

En las tareas de clasificación, a partir de las predicciones del modelo se puede construir una matriz, denominada matriz de confusión, que posee un registro de las veces que se predice cada una de las clases respecto a su valor real. En la Fig. 3.2 se observa un ejemplo de esta, aquí cinco elementos de la clase a fueron correctamente etiquetados, tres lo fueron erróneamente con la b y 4 con la c. En la diagonal se encuentran los casos en que la categoría real coincide con la predicha, caso que se considera un positivo verdadero (TP, True positive en ingles) [50].

Tabla 3.2: Ejemplo de matriz de confusión. Elaboración propia.

	Clase predicha			
Clase real		a	b	c
	a	5	3	4
	b	1	6	8
	c	6	10	0

Además, a partir de los valores de la matriz de confusión se pueden calcular métricas que permiten evaluar el desempeño de la red neuronal. Estas son:

- Accuracy (Exactitud): fracción de la suma de las predicciones correctas para todas las clases sobre el total de predicciones (Ec. 3.2) [50].

$$Accuracy = \frac{\sum_{i=1}^{N} TP_i}{Total\ de\ predicciones} \tag{3.2}$$

- Precision (Precisión): fracción de los elementos que han sido correctamente predichos para una clase sobre el total de predicciones que entregan como resultado esa clase (3.3).

$$Precision_i = \frac{TP_i}{Total\ de\ predicciones\ etiquetadas\ como\ i} \tag{3.3}$$

  En caso de ser binaria se calcula sobre la clase positiva. Por otra parte, para clasificaciones multicategoría es necesario obtener un valor promedio, el que puede ser macro (Ec. 3.4) o micro (3.5) [50].

$$Precision\ Promedio\ Macro = \frac{\sum_{i=1}^{N} Precision_i}{N} \tag{3.4}$$

$$Precision\ Promedio\ Micro = \frac{\sum_{i=1}^{N} TP_i}{Total\ de\ predicciones} \tag{3.5}$$

El primer caso corresponde al promedio aritmético de la precisión para cada clase y el segundo no considera las diferencias de las clases realizando un promedio general de todos las predicciones correctas.

- Recall (Exahustividad): fracción de las predicciones correctas de una clase sobre el total de elementos de esa clase (Ec. 3.6).

$$Recall_i = \frac{TP_i}{Total\ de\ elementos\ de\ clase\ i} \tag{3.6}$$

Al igual que con el indicador anterior, en caso de ser binaria se calcula sobre la clase positiva, y para clasificaciones multicategoría es necesario obtener un valor promedio, el que puede ser macro (Ec. 3.7) o micro (3.8) [50].

$$Recall\ Promedio\ Macro = \frac{\sum_{i=1}^{N} Recall_i}{N} \tag{3.7}$$

$$Recall\ Promedio\ Micro = \frac{\sum_{i=1}^{N} TP_i}{Total\ de\ predicciones} \tag{3.8}$$

- F1-score: corresponde al promedio armónico de precision y recall (Ec. 3.9)

$$F1 - score = 2 * \frac{recall * precision}{recall + precision} \tag{3.9}$$

Al igual que con el indicador anterior, en caso de ser binaria se calcula sobre la clase positiva, y para clasificaciones multicategoría es necesario obtener un valor promedio, el que puede ser macro (Ec. 3.10) o micro (3.11) [50].

$$F1 - score\ Promedio\ Macro = 2 * \frac{Recall\ Macro * Precision\ Macro}{Recall\ Macro + Precision\ Macro} \tag{3.10}$$

$$F1 - score\ Promedio\ Micro = \frac{\sum_{i=1}^{N} TP_i}{Total\ de\ predicciones} \tag{3.11}$$

Capítulo 4

Metodología

La ejecución del presente trabajo está dividida en una etapa de recolección de datos y cuatro partes principales, cada una asociada con un objetivo. A continuación se detalla la metodología realizada en cada una.

4.1. Adquisición de datos

Para el desarrollo de la memoria fue necesario contar con diversos conjuntos de datos de modelos proteicos. Estos fueron diseñados con el fin de llevar a cabo diversas tareas que permitiesen estudiar el potencial de los grafos como representación de proteínas.

Los modelos proteicos en su mayoría fueron adquiridos desde el banco de datos de RCSB (*Research Collaboratory for Structural Bioinformatics*) [13]. A través de la *Search API* [51], variando los parámetros de búsqueda, se obtuvieron listas de identificadores asociados a las proteínas que cumplían con las condiciones particulares del conjunto a generar. Luego, estos fueron utilizados para realizar una descarga remota de los archivos pdb que contenían las estructuras proteicas.

A continuación, se presentan los parámetros utilizados para cada conjunto de datos y el tratamiento específico que se le aplicó a cada uno.

4.1.1. Estructura secundaria

Este conjunto de datos se creó para llevar a cabo tareas de clasificación de modelos proteícos, sin considerar ligandos ni solventes, según las categorías de SCOP (*Structural Classification of Proteins*) [52]. Así, las clases presentes en este son:

- *All Alpha proteins*: Dominios que presentan predominantemente hélices alfa [52].
- *All beta proteins*: Dominos que presentan predominantemente hebras beta [52].
- *Alpha and beta proteins (a+b)*: Dominios con hélices alfa y hebras beta segregadas [52].
- *Alpha and beta proteins (a/b)*: Dominios con hélices alfa y hebras beta alternadas [52].
- *Small proteins*: Dominios con pequeñas o inexistentes estructuras secundarias [52].

Para adquirir los identificadores de este conjunto se hizo uso de los parámetros de búsqueda presentados en la Tabla 4.1. El valor del atributo rcsb_polymer_instance_annotation.annotation_lineage.name se varió según la clase que se deseaba buscar.

Tabla 4.1: Parámetros de selección para la adquisición de datos para la clasificación según SCOP.

Característica	**Atributo en API**	**Valor**
Resolución	rcsb_entry_info.resolution_combined	≤ 3 [Å]
Tamaño	entity_poly.rcsb_sample_sequence_length	≤ 1000 [kDa]
		≥ 150 [kDa]
Tipo de modelo	exptl.method	Excluir teóricos
Categoría de la anotación	rcsb_polymer_instance_annotation.type	SCOP
Valor de la anotación	rcsb_polymer_instance_annotation.annotation_lineage.name	Nombre de la clase

Posterior al estudio de los datos, se notó la existencia de elementos multicategoría, por lo que, para simplificar el conjunto, se procedió a filtrar dichos elementos, conservando sólo aquellos que pertenecen a una única clase. Finalmente, se seleccionó de manera aleatoria un subconjunto de 1500 elementos de cada categoría, utilizando la función *sample* de la librería *random* de Python con semilla 100. En el caso del conjunto de *Small proteins*, cuyo tamaño es menor a 1500 elementos se consideró la totalidad de estos.

4.1.2. Función enzimática

Este conjunto de datos se creó para llevar a cabo tareas de clasificación de modelos proteicos, sin considerar ligandos ni solventes, según las recomendaciones del Comité de Nomenclatura de la Unión Internacional de Bioquímica y Biología Molecular (IUBMB, *Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*) [53]. Así, las clases presentes en este son:

- *Oxidoreductase*: catalizan reacciones de oxidorreducción [53].
- *Transferase*: catalizan la transferencia de un grupo funcional de un sustrato a otro [53].
- *Hydrolase*: catalizan la hidrólisis de un enlace químico [53].
- *Lyase*: catalizan la ruptura de enlaces químicos con un mecanismo de eliminación, dejando doble enlaces o anillos, o, por el contrario, agregan grupos a doble enlaces [53].
- *Isomerase*: catalizan cambios geométricos o estructurales dentro de una molécula [53].
- *Ligase*: catalizan la formación de enlaces químicos para la unión de moléculas, acompañada de la hidrolisis de un ATP o un trifosfato similar [53].
- *Translocase*: asisten y/o regulan el movimiento de una molécula [53].

Para adquirir los identificadores de este conjunto se hizo uso de los parámetros de búsqueda presentados en la Tabla 4.2. El valor del atributo struct_keywords.pdbx_keywords se varió según la clase que se deseaba buscar, y se excluyó la palabra *Inhibitor* de las búsquedas,

debido a que se notó que, en el caso contrario, se obtenían tanto enzimas como inhibidores de la actividad enzimática.

Tabla 4.2: Parámetros de selección para la adquisición de datos para la clasificación de enzimas según su actividad.

Característica	**Atributo en API**	**Valor**
Resolución	rcsb_entry_info.resolution_combined	≤ 3 [Å]
Tamaño	entity_poly. rcsb_sample_sequence_length	≤ 1000 [kDa]
		≥ 150 [kDa]
Tipo de modelo	exptl.method	Excluir teóricos
Palabras claves	struct_keywords.pdbx_keywords	Excluir "inhibitor"
		Nombre de una clase

Posterior al estudio de los datos, se notó la existencia de elementos multicategoría, por lo que, para simplificar el conjunto, se procedió a filtrar dichos elementos, conservando sólo aquellos que pertenecen a una única clase. Finalmente, se seleccionó de manera aleatoria un subconjunto de 1500 elementos de cada categoría, utilizando la función *sample* de la librería *random* de Python con semilla 100. En el caso del conjunto de *Translocase*, cuyo tamaño es menor a 1500 elementos se consideró la totalidad de estos.

4.1.3. Unión con ARN o ADN

Este conjunto de datos fue concebido con el fin de realizar tareas de clasificación de proteínas según el tipo de ácido nucleico al que se unen. En este caso, si bien el ligando es de interés ya que define la función estudiada, no existe necesidad de representarlo, ya que se quiere estudiar cómo la estructura de la proteína define la interacción, no la interacción en sí. Así, las clases presentes en este son:

- *ADN binding*: Proteínas que poseen sitios de unión a ADN.
- *ARN binding*: Proteínas que poseen sitios de unión a ARN.

Para adquirir los identificadores de este conjunto se hizo uso de los parámetros de búsqueda presentados en la Tabla 4.3. El valor del atributo struct_keywords.pdbx_keywords se varió según la clase que se deseaba buscar.

Tabla 4.3: Parámetros de selección para la adquisición de datos para la clasificación según el tipo de ácido nucleico al que se unen.

Característica	**Atributo en API**	**Valor**
Resolución	rcsb_entry_info.resolution_combined	≤ 3 [Å]
Tamaño	entity_poly. rcsb_sample_sequence_length	≤ 1000 [kDa]
		≥ 150 [kDa]
Tipo de modelo	exptl.method	Excluir teóricos
Palabras claves	struct_keywords.pdbx_keywords	Nombre de la clase

Posterior al estudio de los datos, se notó la existencia de un único elemento multicategoría (4WZW), el que fue removido de ambos subconjuntos.

4.1.4. Variantes de p53 humana

Este conjunto de datos está compuesto por estructuras de la proteína p53 humana en su estado nativo y mutado. Fue concebido para ser utilizado en detección de comunidades, permitiendo identificar el efecto de las mutaciones en estas. La notación utilizada para las sustituciones es explicada en el Anexo A.

Para su creación se revisaron las mutaciones de p53 humana registradas en la plataforma Uniprot [54], dividiéndolas en variantes naturales e inducidas por mutagénesis dirigida (artificiales). Luego, se obtuvo la estructura nativa predicha por AlphaFold [55], sobre la que se aplicaron estas mutaciones puntuales por medio de la plataforma SDM [56], la que entrega como resultado modelos en formato pdb con las estructuras mutadas y estimaciones de su efecto, las que se presentan en el Anexo B.

4.1.5. Parámetros de unión

Este conjunto de datos se ideó con el fin de realizar tareas de regresión que permitan predecir el valor de distintos parámetros que cuantifican la unión de una proteína y un ligando. Dentro de estos parámetros se encuentran:

- IC_{50}: concentración de ligando que reduce la actividad enzimática en un 50 %.
- EC_{50}: concentración de compuesto que genera la mitad de la respuesta máxima.
- K_d: constante de disociación.
- K_a: constante de asociación.
- K_i: constante de inhibición enzimática.
- ΔG: Energía libre de Gibbs de la unión.

Para su obtención se utilizaron los parametros de búsqueda presentados en la Tabla 4.4. En el atributo rcsb_binding_affinity.type se variaron los valores según el parámetro de unión deseado.

Tabla 4.4: Parámetros de selección para la adquisición de datos para el conjunto parámetro de unión proteína-ligando.

Característica	**Atributo en API**	**Valor**
Resolución	rcsb_entry_info.resolution_combined	≤ 3 [Å]
Tamaño	entity_poly. rcsb_sample_sequence_length	≤ 1000 [kDa]
		≥ 150 [kDa]
Tipo de modelo	exptl.method	Excluir teóricos
Valor de parámetro de unión	rcsb_binding_affinity.value	Existe
Tipo de parámetro de unión	rcsb_binding_affinity.type	Nombre de uno de los parámetros

Luego, se ocupó la *Data API* de RCSB [57] para obtener el valor numérico de los distintos parámetros para cada proteína. Siendo necesario para lo anterior, indicar el identificador del

modelo deseado en *queries* a un servidor GraphQL.

Cabe destacar que en un modelo, para un mismo ligando, se puede presentar más de un valor para un parámetro, debido a la existencia de distintas fuentes. Se recomienda al usuario del *dataset* seleccionar como manejar esta característica (calcular el promedio, elegir el valor máximo, aquel que se repite más, etc).

4.1.6. Afinidad con distintos sustratos

Este conjunto posee datos de la actividad especifica de 145 esterasas con 96 sustratos, obteniendo así 13920 datos, los que se buscaron pensado en procesos de clasificación sobre la afinidad de enzima-sustrato. La información se obtuvo del paper Determinants and Prediction of Esterase Substrate Promiscuity Patterns [58], específicamente de la Tabla Suplementaria S3. A partir de representaciones SMILES de los sustratos presentes en dicha tabla, se generaron modelos 3D de estos. Para las proteínas se descargaron aquellas que se encuentran en RCSB, siendo necesario generar con herramientas como AlphaFold aquellas que no se encuentran disponibles en el banco de datos. Finalmente se debe realizar un docking proteínas-ligando para su uso en los modelos predictivos.

4.2. Generación de grafos

Actualmente los métodos y modelos computacionales son ampliamente utilizados para el estudio y diseño de proteínas, siendo un paso esencial para esto la selección de una representación adecuada [4]. Una manera de dividir las diversas representaciones que existen es considerando si hacen o no uso de la estructura tridimensional. En el primer grupo tenemos métodos como son: One-hot encoding [4], Natural Language Processing [4] y Digital Signal Processing [34] y en el segundo a los grafos. Estos últimos parecen ser una buena opción debido a la capacidad que tienen de incorporar las interacciones de una manera sencilla e intuitiva en sus aristas, lo que los posiciona como un objeto de estudio en la presente memoria.

El estudio se inició diseñando e implementando un programa que permitiese generar grafos a partir de los diversos modelos proteicos disponibles en línea. Así, en primera instancia, se determinó el tipo de grafos que se deseaba generar y los criterios de representación a utilizar. Los grafos no dirigidos fueron seleccionados como la estructura base, variando los elementos que se componen sus nodos y aristas según lo indicado en la Figura 4.1.

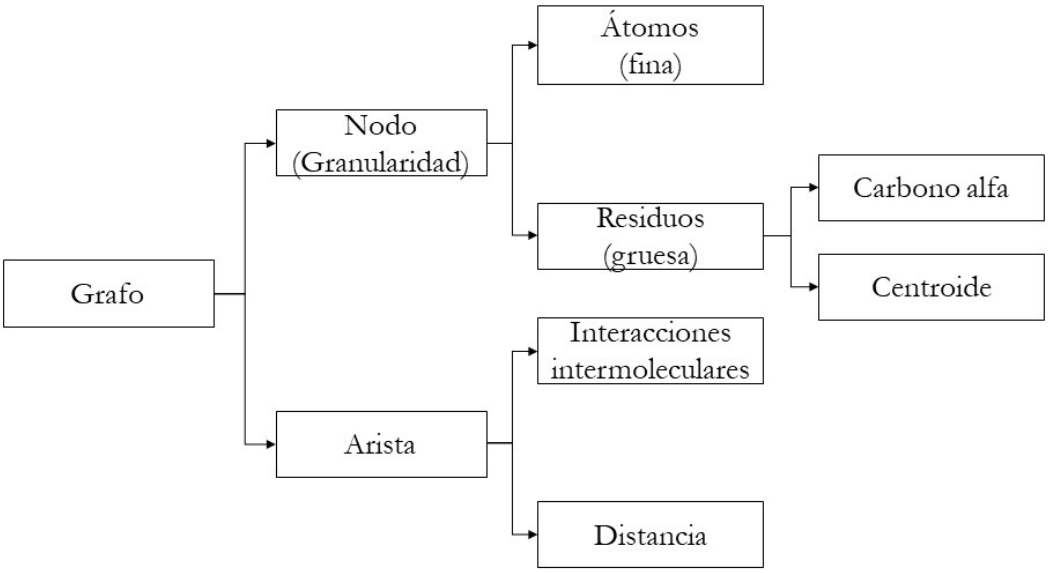


Figura 4.1: Criterios usados para la representación de proteínas por medio de grafos.

Según la granularidad se construyeron modelos finos y gruesos. En el primer grupo los nodos corresponden a los átomos de la molécula y en el segundo, a los residuos, los que a su vez se representan con su carbono alfa o centroide. Respecto a la información de las aristas se consideraron dos tipos: distancia euclidiana e interacciones intermoleculares, específicamente puentes de Hidrógeno. Estos últimos fueron utilizados solo a nivel de residuo con carbonos alfa. Así, a partir de una proteína se generaron cuatro grafos distintos: átomo-distancia, carbono alfa-distancia, centroide-distancia e interacción intermolecular.

A continuación, se diseñó un programa que genera el grafo correspondiente a partir de un modelo proteico en formato pdb, la granularidad y tipo de interacción. El flujo del programa se presenta en la Figura 4.2, donde se puede observar que el tipo de interacción divide el procedimiento que sigue el programa.

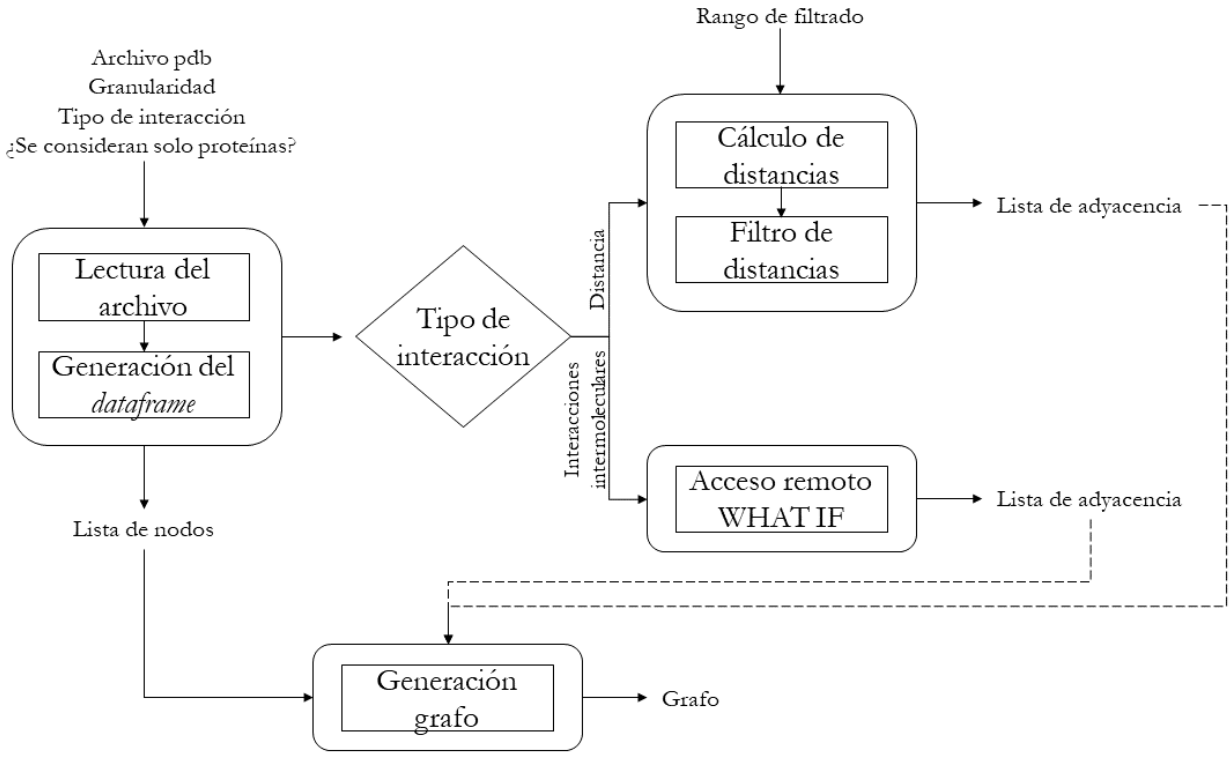


Figura 4.2: Flujo del programa para generación de grafos. La línea punteada indica que solo una de las dos listas de adyacencia entra a la etapa de construcción de grafos.

El programa en primera instancia lee el archivo pdb con un parser implementado en la librería *Biopython*, rescatando el primer modelo presentado en este. Dicha lectura es condicionada a partir de un parámetro entregado por el usario que indica si se quieren considerar todas las moléculas del modelo, incluyendo solvente y ligandos no proteicos, o solo las proteínas. Luego, se genera un *dataframe* que posee información de los nodos y su posición espacial. Para lo anterior, según la granularidad, se realiza una de las siguientes acciones:

- Identificar individualmente cada átomo de la proteína.
- Identificar los carbonos alfa de cada residuo.
- Calcular la posición del centroide de cada residuo. Para esto se modela al redisuo como un sistema de partículas sin volumen, calculando el promedio de la posición de los átomos en cada eje sin ponderarlas por el aporte de volumen da cada una.

Es importante destacar que el *parser* no considera los átomos de hidrógeno que forman parte de la estructura proteica.

El siguiente paso consiste en generar dos listas en formato csv: la lista de nodos y la lista de adyacencia. La primera se obtiene de manera directa de los datos y para la última se realiza un procedimiento diferente dependiendo de la información deseada en los vértices:

- Grafos de distancia: A partir de las posiciones indicadas en el *dataframe* se calcula la distancia de cada nodo versus los nodos restantes. Debido a la gran cantidad de relaciones

que establece este método y con el fin de disminuir el ruido se implementó un filtro que permite eliminar aquellas que se encuentren fuera de un rango.

- Grafos de interacciones intermoleculares: Se accede de manera remota al servicio *what if* [59] que permite estimar la red optimizada de puentes de Hidrógeno presentes en la molécula.

Finalmente, con las listas de adyacencia y de nodos se genera un grafo en formato gpickle para lo que se utilizan funciones implementadas en la librería *networkx*.

El *script* de generación de grafos fue ejecutado sobre los diversos conjuntos de datos presentados en la Sección 4.1, considerando solo las proteínas presentes en los modelos y filtrando las distancias menores a 2 Å y mayores a 15 Å en los grafos de distancia.

4.2.0.1. Implementación

Para la implementación de los *scripts* mencionados se utilizó Visual Studio Code 1.66.0 y Python 3.9.7. La lectura de los archivos pdb se realizó con la librería BioPython 1.79, el manejo de datos con Pandas 1.3.4, el acceso remoto con Selenium 4.1.3 y la construcción de los grafos con Networkx 2.6.3.

4.3. Detección de comunidades

Una labor importante en el estudio de proteínas es identificar patrones en su estructura que permitan entenderla desde un punto funcional, filogenético o estructural. Al ser representadas por medio de grafos nace la oportunidad de aplicar algoritmos de detección de comunidades sobre los modelos proteicos, que podrían permitir identificar patrones de interés.

La detección de comunidades se realizó por medio de un *script* que permite ejecutar distintos algoritmos de *clustering* a grafos en formato gpickle. El flujo del programa diseñado se puede observar en la Figura 4.3. Se inicia con la lectura del archivo, el que se transforma en un objeto compatible con la librería *igraph*. Este paso es necesario debido a que este último módulo posee una mayor variedad de algoritmos implementados. Luego, se procede a la ejecución del *clustering*, donde se recibe como input una lista con los nombres de los algoritmos a utilizar. Es importante usar algoritmos que soporten grafos con pesos, y entregarles los pesos respectivos como parámetro a la función. Se obtiene como resultado dos archivos en formato csv: el primero posee la distribución de los nodos en las distintas comunidades detectadas y el segundo las estadísticas de la partición (número de comunidades y modularidad). Un punto destacable en el flujo presentado es que algunos algoritmos como walktrap generan un objeto del tipo *VertexDendrogram* que debe ser transformado en un *VertexClustering* antes de ser interpretado.

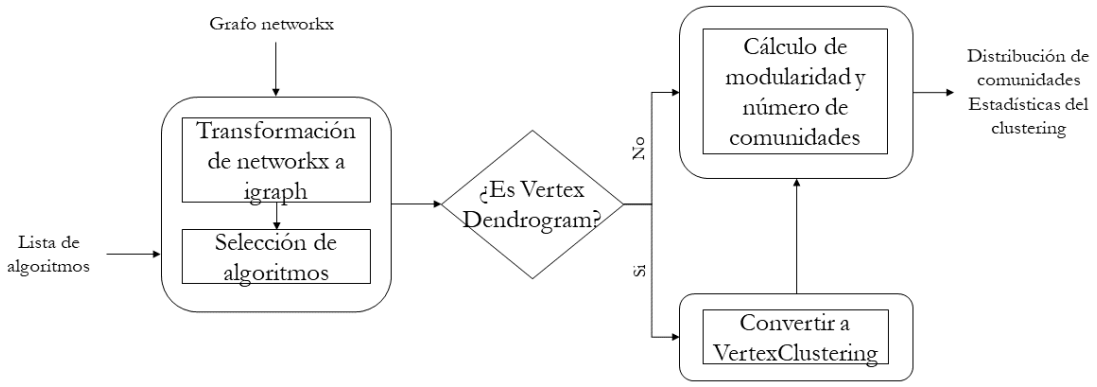


Figura 4.3: Flujo del programa de detección de comunidades. Elaboración propia.

Un punto de interés es que los algoritmos no funcionan con todos los tipos de grafos. Específicamente, el algoritmo Spinglass requiere que el grafo sobre el que se aplica sea conectado, inutilizándose en las representaciones que usan interacciones intermoleculares. Por otra parte, el algoritmo Walktrap puede entregar errores con ciertos grafos no conectados al transformar el *VertexDendrogram* en un *VertexClustering*, ocurriendo un problema similar al descrito en el caso anterior.

Para realizar un estudio de cómo afectan las mutaciones a la detección de comunidades se ejecutó el programa sobre los grafos de distancia-carbono alfa, distancia-centroide e interacción intermoleculares del conjunto de datos: Variantes de p53 humana, con los que también observó el efecto que tienen los distintos algoritmos sobre el *clustering* generado.

4.3.0.1. Implementación

Para la implementación de los *scripts* mencionados se utilizó Visual Studio Code 1.66.0 y Python 3.9.7. El manejo de datos se realizó con Pandas 1.3.4, la construcción de los grafos con Networkx 2.6.3. y la detección de comunidades con Igraph 0.9.11.

4.4. Comparación de grafos

Para abordar la problemática de comparación de grafos se propone una metodología exploratoria basada en la comparación de subgrafos de un mismo esquema, pero utilizando las comunidades detectadas como el elemento comparativo. Se generó un *script* que recibe la distribución de los nodos en las distintas comunidades tanto en la proteína nativa como en

una mutante y por medio de manejo de texto se identifican las diferencias. Luego, se calcula la similitud entre las comunidades antes y después de la mutación a partir de los coeficientes presentados en las Ecs. 4.1 y 4.2, los que se aplican a todos los pares posibles de *clusters*. Similitud mut-nat indica el porcentaje del total elementos de la comunidad de la proteína nativa que se mantienen en la comunidad de la proteína mutada. Por su parte, similitud mut-nat señala el porcentaje de elementos de la comunidad de la proteína mutada que se encuentran en la proteína nativa.

$$\underset{nat-mut}{Similitud} = \frac{\#\ Nodos\ de\ la\ comunidad\ nativa\ presentes\ en\ la\ comunidad\ mutada}{\#\ Nodos\ en\ la\ comunidad\ nativa} * 100 \quad (4.1)$$

$$\underset{mut-nat}{Similitud} = \frac{\#\ Nodos\ de\ la\ comunidad\ mutada\ presentes\ en\ la\ comunidad\ nativa}{\#\ Nodos\ en\ la\ comunidad\ mutada} * 100 \quad (4.2)$$

Considerando una comunidad i de la proteína en estado nativo y una j en la variante, si se utiliza teoría de conjuntos, renomabrando conjunto I a la primera comunidad y J a la segunda, estos coeficientes corresponderían a la cardinalidad de la intersección de I con J sobre la cardinalidad del I para la similitud nat-mut o sobre la de J para la similitud mut-nat. Indicando así que tanto la intersección se asemeja a uno de los conjuntos En caso de que la similitud nat-mut sea 100 % se puede inferir que I será subconjunto de J, ya que el primero está formado sólo por elementos en común entre los dos (intersección). Un razonamiento análogo se puede plantear para una similitud mut-nat igual a 100 % concluyendo que J será subconjunto de I. A partir de los dos casos anteriores, un 100 % en ambos coeficientes establece una relación de equivalencia entre los conjuntos y, por tanto, entre las comunidades.

Finalmente, se seleccionaron mutantes de interés en el conjunto de variantes de p53 humana, eligiendo aquellas que provocan un mayor cambio en la estabilidad y aquellas que sufren variaciones en la modularidad una mayor cantidad de veces. Estas fueron estudiadas en detalle y graficadas para tener una aproximación visual de los cambios. Además, se estudió la vecindad de nodos de interés, como lo es el residuo mutado, con el fin de buscar modificaciones epistáticas.

4.4.0.1. Implementación

Para la implementación de los *scripts* mencionados se utilizó Visual Studio Code 1.66.0 y Python 3.9.7. El manejo de datos se realizó con Pandas 1.3.4.

4.5. Graph Convolutional Networks

Para finalizar el estudio de los grafos se buscó generar una metodología que permitiese utilizar estas representaciones en modelos de *deep learning* para predecir funciones y estructuras proteicas. Se utilizó un ambiente con pytorch geometric, para generar una modelo que clasificara las proteínas según función enzimática. El trabajo se dividió en tres pasos: elaboración del conjunto de datos, construcción de la red neuronal y ejecución y evaluación de desempeño.

La elaboración del conjunto de datos utilizó el dataset llamado función enzimática mencionado en la Sección 4.1, del cual se extrajeron 500 estructuras por tipo de dato. El desbalance que existe con la clase translocase fue manejado por medio de la eliminación de esta. Las estructuras fueron representadas utilizando grafos de Carbono-alfa distancia, que fueron guardados en una carpeta denominada *raw* según lo solicitado por la librería.

Luego, se programó una clase propia que hereda de de la clase *Dataset* de pytorch geometric. En está se procesaron los grafos creando un vector que contiene información sobre las propiedades de los aminoácidos, como son: peso molecular, radio atómico, hidrofobicidad y punto isoeléctrico, lo que permite representar los nodos con estos datos. Además, cada objeto de la clase posee información de las aristas y sus pesos asociados. Así, se generaron archivos de los grafos compatibles con la librería pytorch geometric, los que fueron ubicados en un carpeta denominada *processed*.

A continuación se construyó la red neuronal, la que consiste en un modelo convolucional de cuatro capas utilizando la función de activación ReLu para generar el embedding de los nodos. Asimismo se agregó una capa readout que permite unificar el embedding de los nodos en uno representativo para el grafo, específicamente se usó el promedio de estos. Finalmente, se incluyó una capa lineal con dropout de 0.5. Los hiperparámetros de la red son presentados en la Tabla 4.5. Es importante indicar que, debido a escasez de tiempo de ejecución y recursos computacionales, estos no fueron ajustados para obtener un rendimiento óptimo del modelo.

Tabla 4.5: Parámetros de selección para la adquisición de datos para la clasificación de enzimas según su actividad. Elaboración propia.

Hiperparámetro	**Valor**
Número de capas convolucionales	4
Número de épocas	100
Tasa de aprendizaje	0.1
Tamaño del batch	128
Canales de espacio oculto	128

Finalmente, la red se ejecutó utilizando el 70 % de los datos para el entrenamiento y el 30 % restante para testeo. Como función de costos se usó entropia cruzada y un optimizador Adam. Además se generaron gráficas y métricas de desempeño para evaluar el modelo generado.

4.5.1. Implementación

Para la implementación de los *scripts* mencionados se utilizó Visual Studio Code 1.66.0 y Python 3.9.7. La construcción de la red neuronal se realizó con Pytorch 1.10.0+cu113 y Pytorch geometric 2.0.2, el set de datos se armó con Networkx 2.6.3, Pandas 1.3.4 y Numpy 1.20.3 y el cálculo de las métricas de desempeño con Sklearn 0.0.post1 y Seaborn 0.11.2.

Capítulo 5

Resultados

En la presente sección se muestran los resultados obtenidos durante la investigación llevada a cabo en esta memoria.

5.1. Adquisición de datos

A continuación en las Figuras 5.1, 5.2, 5.3, 5.4, 5.5 y 5.6 se presentan gráficos que resumen la cantidad de datos obtenidos para cada uno de los conjuntos elaborados sin considerar los elementos multicategoría. Además, se describe el conjunto asociado con la afinidad de proteínas con distintos sustratos.

5.1.1. Estructura secundaria

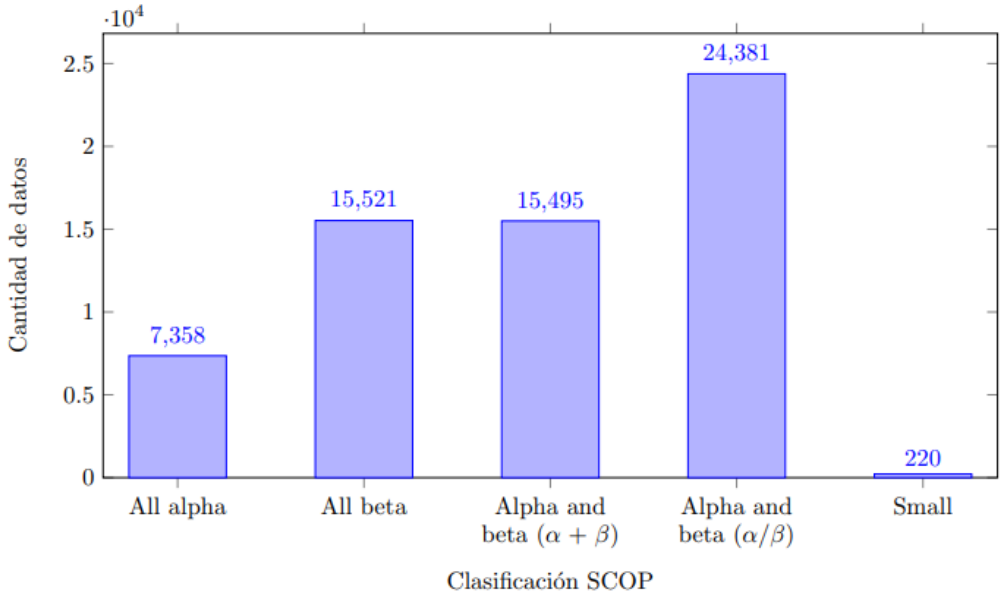


Figura 5.1: Distribución de datos por clase de SCOP sin considerar aquellos multicategoría.

5.1.2. Función enzimática

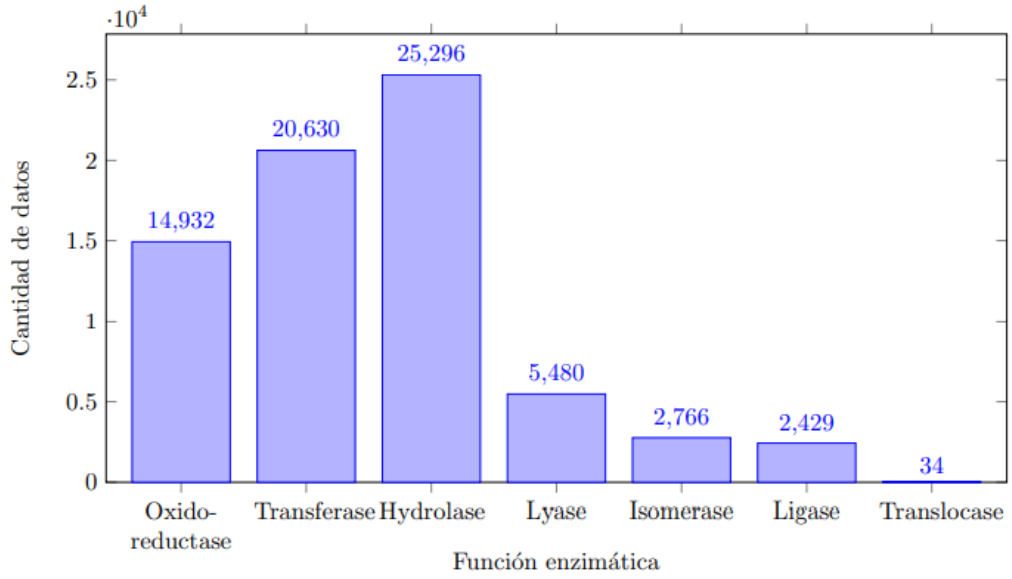


Figura 5.2: Distribución de datos por función enzimática sin considerar aquellos multicategoría.

5.1.3. Unión con ARN o ADN

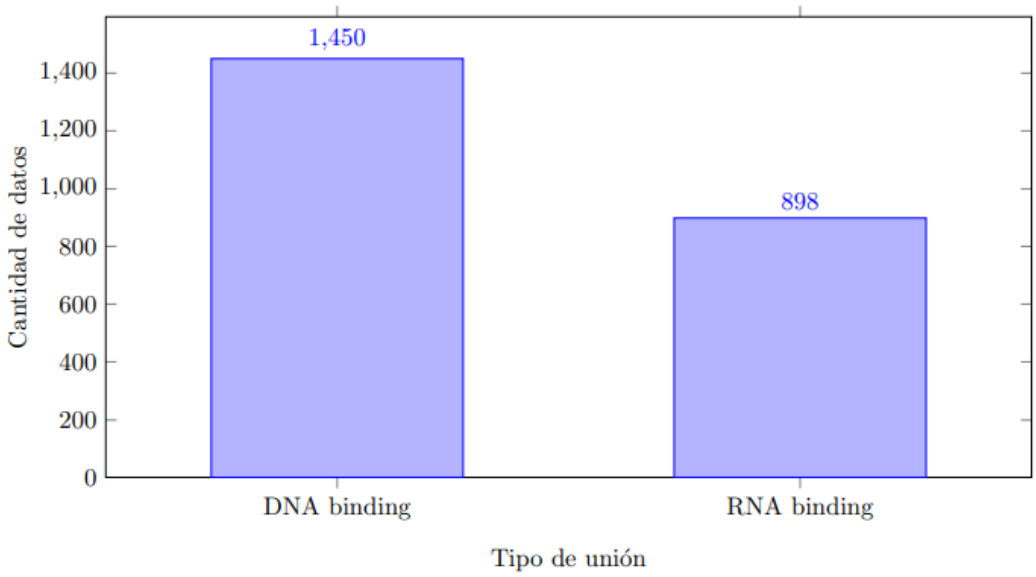


Figura 5.3: Distribución de datos por tipo de ácido nucleico al que se unen sin considerar aquellos multicategoría.

5.1.4. Variantes de p53 humana

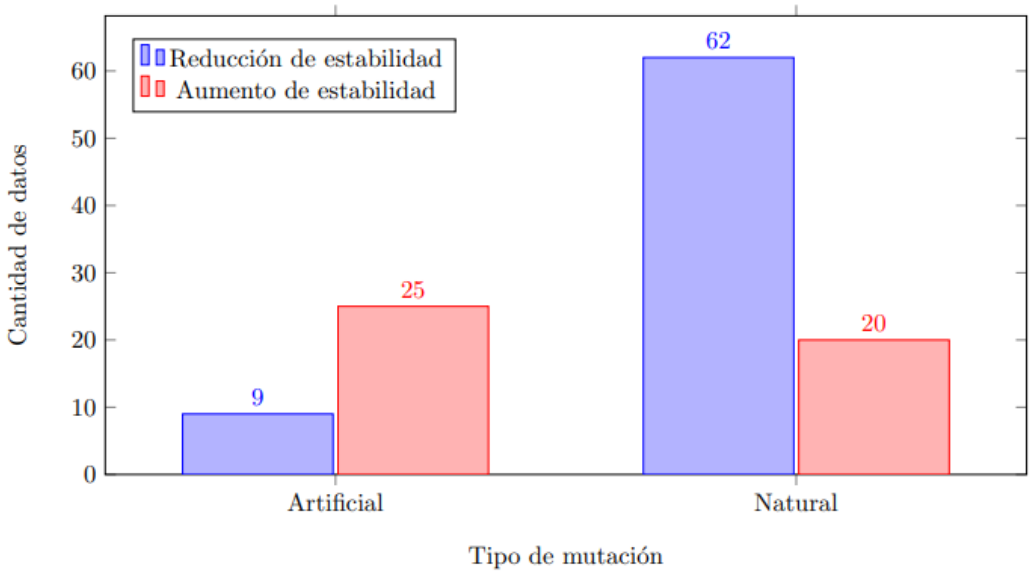


Figura 5.4: Distribución de datos según tipo de mutación y el efecto que poseen en la estabilidad proteica.

5.1.5. Parámetros de unión

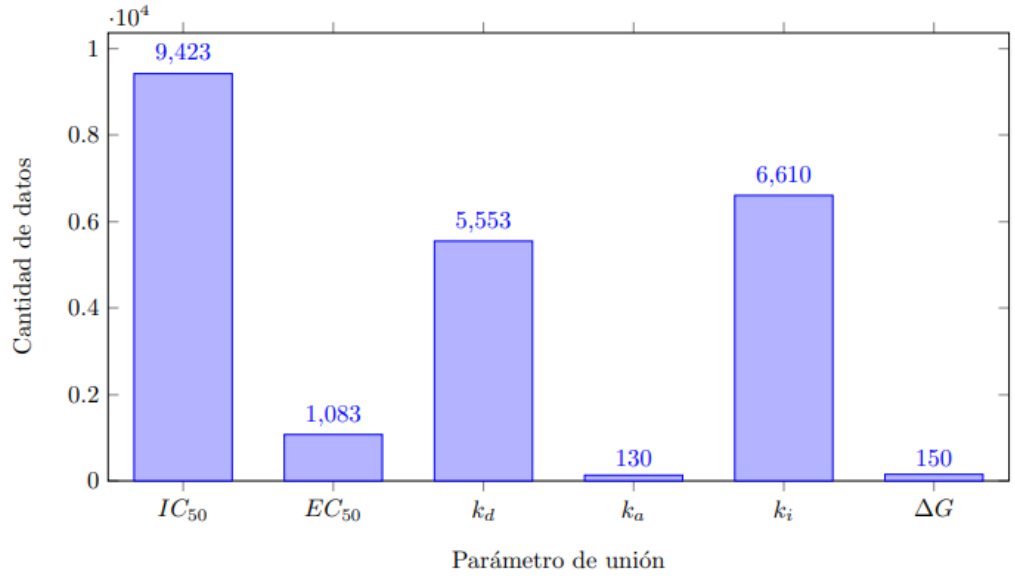


Figura 5.5: Distribución de datos según el tipo de parámetro de afinidad estudiado.

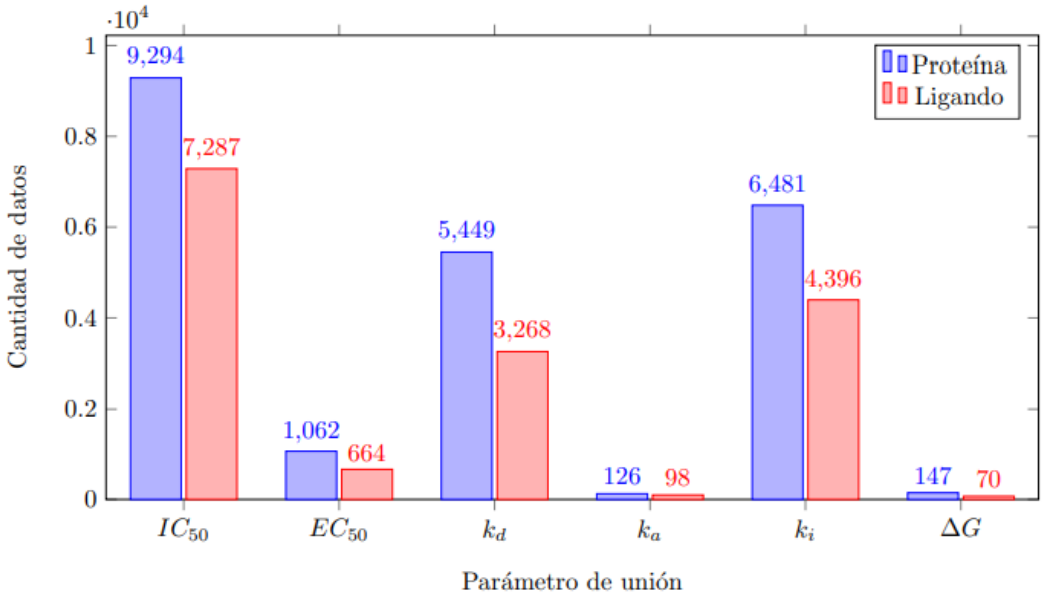


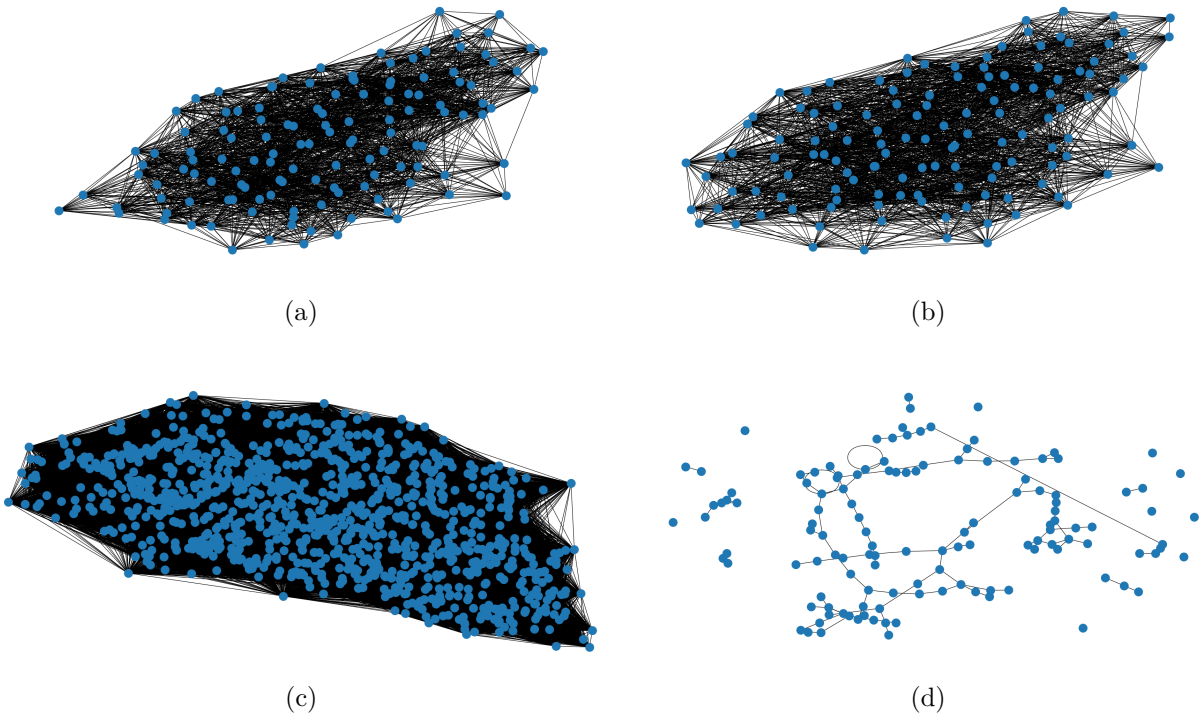
Figura 5.6: Distribución de datos según el tipo de molécula para cada parámetro de afinidad estudiado.

5.1.6. Afinidad con distintos sustratos

El conjunto resultante está compuesto por la información de la actividad enzimática específica de 145 esterasas al exponerse a 96 sustratos, poseyendo un total de 13920 datos.

5.2. Generación de grafos

Para estudiar los grafos construidos se utiliza un ejemplo debido a que la gran cantidad de resultados obtenidos impide presentarlos en su totalidad. Así, se presenta en la Figura 5.7 una comparación de las cuatro representaciones generadas para la lisozima humana (código pdb 1REX [60]) y el respectivo número de aristas y nodos en la Tabla 5.1.



(a) (b) (c) (d)

Figura 5.7: Comparación de distintas representaciones para la lisozima humana (pdb id: 1REX) (a) Grafo carbono alfa distancia (b) Grafo centroide distancia (c) Grafo átomo distancia (d) Grafo interacciones intermoleculares.

Tabla 5.1: Número de elementos de los distintos tipos de grafos creados para la lisozima humana.

Tipo de grafo	**Átomos Distancia**	**Carbono alfa Distancia**	**Centroide Distancia**	**Interacciones intermoleculares**
Nodos	1029	130	130	130
Aristas	180393	2935	2907	118

5.3. Detección de comunidades

La detección de comunidades se estudia en las Tablas 5.2, 5.3, 5.4, donde se pueden observar las estadísticas de la modularidad y número de comunidades de tres de las representaciones generadas para la proteína p53 humana: grafo de distancia-carbono alfa, grafo de distancia-centroide y grafo de interacciones intermoleculares.

Tabla 5.2: Estadísticas de detección de comunidades para grafos de distancia de p53 con carbonos alfa en los nodos.

Algoritmo	**Modularidad**		**Número de comunidades**	
	Promedio	**Desv. estándar**	**Promedio**	**Desv. estándar**
Fast greedy	0,4549	0,0002	4,0256	0,1581
Infomap	0,4982	0,0001	12,0000	0,0000
Label propagation	0,4782	0,0265	12,9658	0,2592
Leading eigenvector	0,4931	0,0001	9,0000	0,0000
Multilevel	0,5063	0,0003	7,0000	0,0000
Spinglass	0,5082	0,0005	8,0598	0,7656
Walktrap	0,4359	0,0102	29,0342	0,1817

Tabla 5.3: Estadísticas de detección de comunidades para grafos de distancia de p53 con centroides en los nodos.

Algoritmo	**Modularidad**		**Número de comunidades**	
	Promedio	**Desv. estándar**	**Promedio**	**Desv. estándar**
Fast greedy	0,4603	0,0017	5,3162	0,4650
Infomap	0,3763	0,1026	11,8462	0,3608
Label propagation	0,2908	0,0015	13,0000	0,2265
Leading eigenvector	0,4858	0,0451	6,9487	0,5523
Multilevel	0,5058	0,0006	7,0000	0,0000
Spinglass	0,5115	0,0005	7,8632	0,5048
Walktrap	0,4485	0,0027	32,8120	0,3907

Tabla 5.4: Estadísticas de detección de comunidades para grafos de interacciones de p53 con carbono alfa en los nodos.

Algoritmo	**Modularidad**		**Número de comunidades**	
	Promedio	**Desv. estándar**	**Promedio**	**Desv. estándar**
Fast greedy	0,9256	0,0020	214,2735	0,7116
Infomap	0,8975	0,0025	226,2821	0,5969
Label propagation	0,8252	0,0035	243,1453	0,4938
Leading eigenvector	0,9251	0,0019	213,4274	0,7432
Multilevel	0,9282	0,0019	213,3932	0,7036
Spinglass	-	-	-	-
Walktrap	-	-	-	-

Se estudió también cómo el cambio de modularidad se puede relacionar con las variaciones de estabilidad producidas por las mutaciones para grafos de distancia-carbono alfa, grafos de distancia-centroide y grafos de interacciones intermoleculares. Lo anterior se presenta respectivamente en las Figuras 5.8, 5.9 y 5.10. En las dos primeras se realizó un acercamiento sobre ciertas nubes de punto para observar la influencia de valores *outliers* en el análisis.

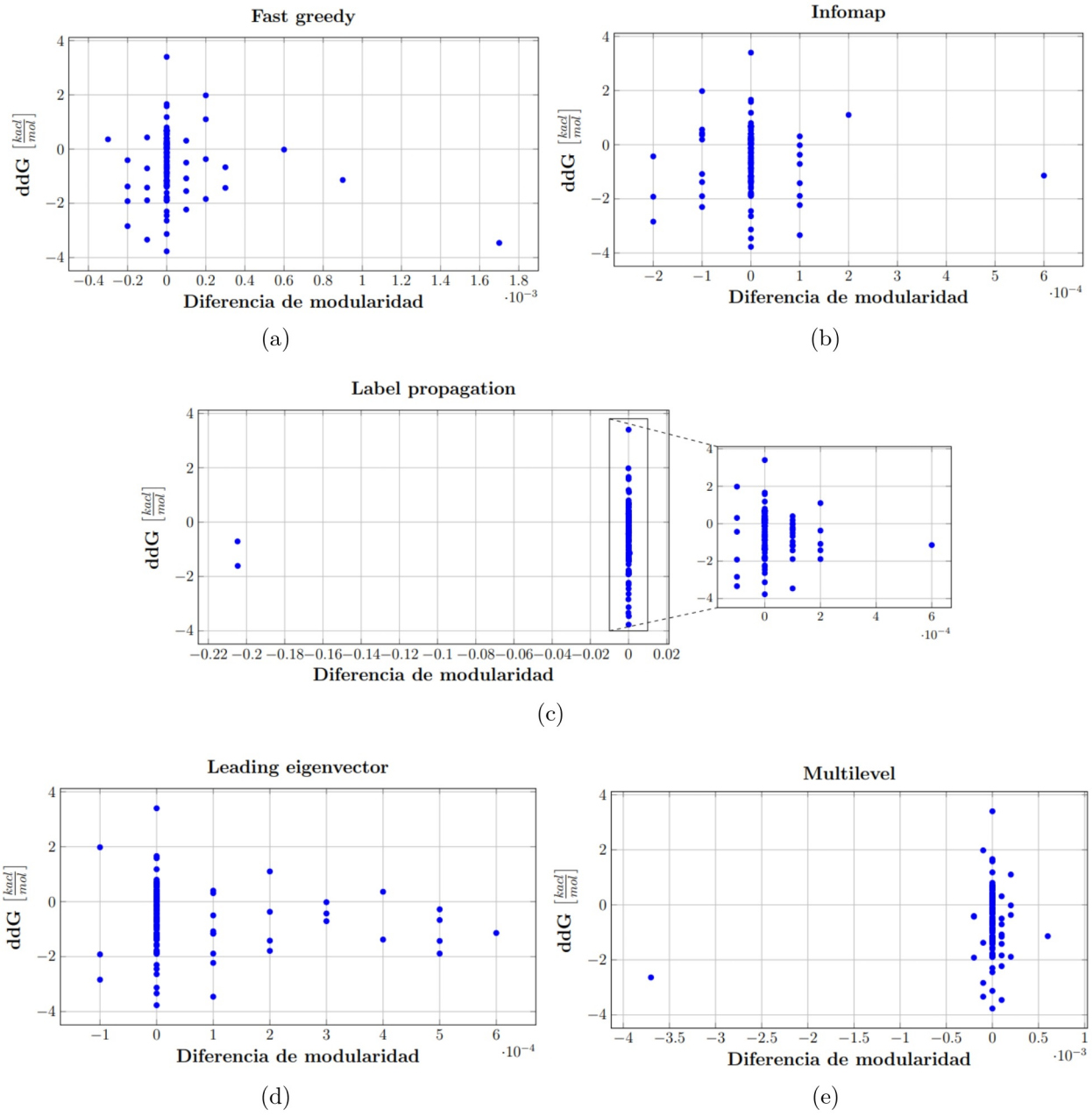
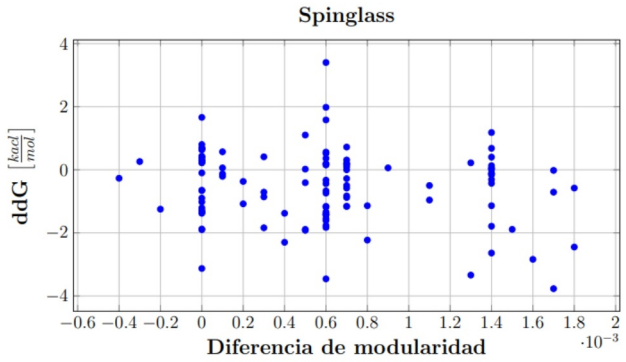
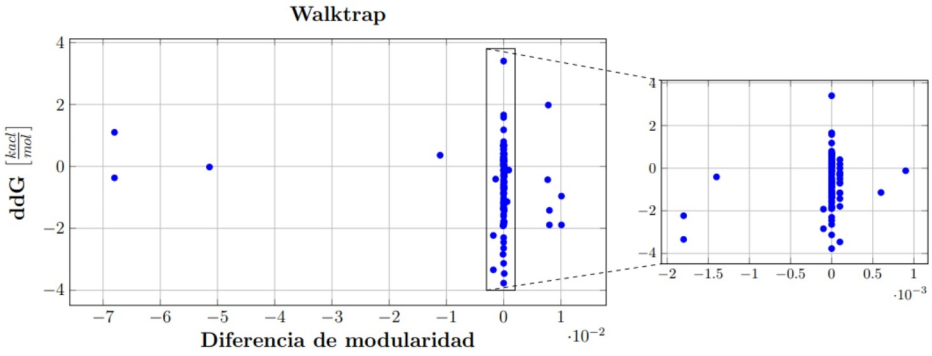


Figura 5.8: Variaciones de estabilidad según cambio de modularidad de los grafos de Carbono alfa-distancia para los distintos algoritmos (a) Fast greedy (b) Infomap (c) Label propagation (d) Leading eigenvector (e) Multilevel.



(f)



(g)

Figura 5.8: Variaciones de estabilidad según cambio de modularidad de los grafos de Carbono alfa-distancia para los distintos algoritmos (cont.) (f) Spinglass (g) Walktrap.

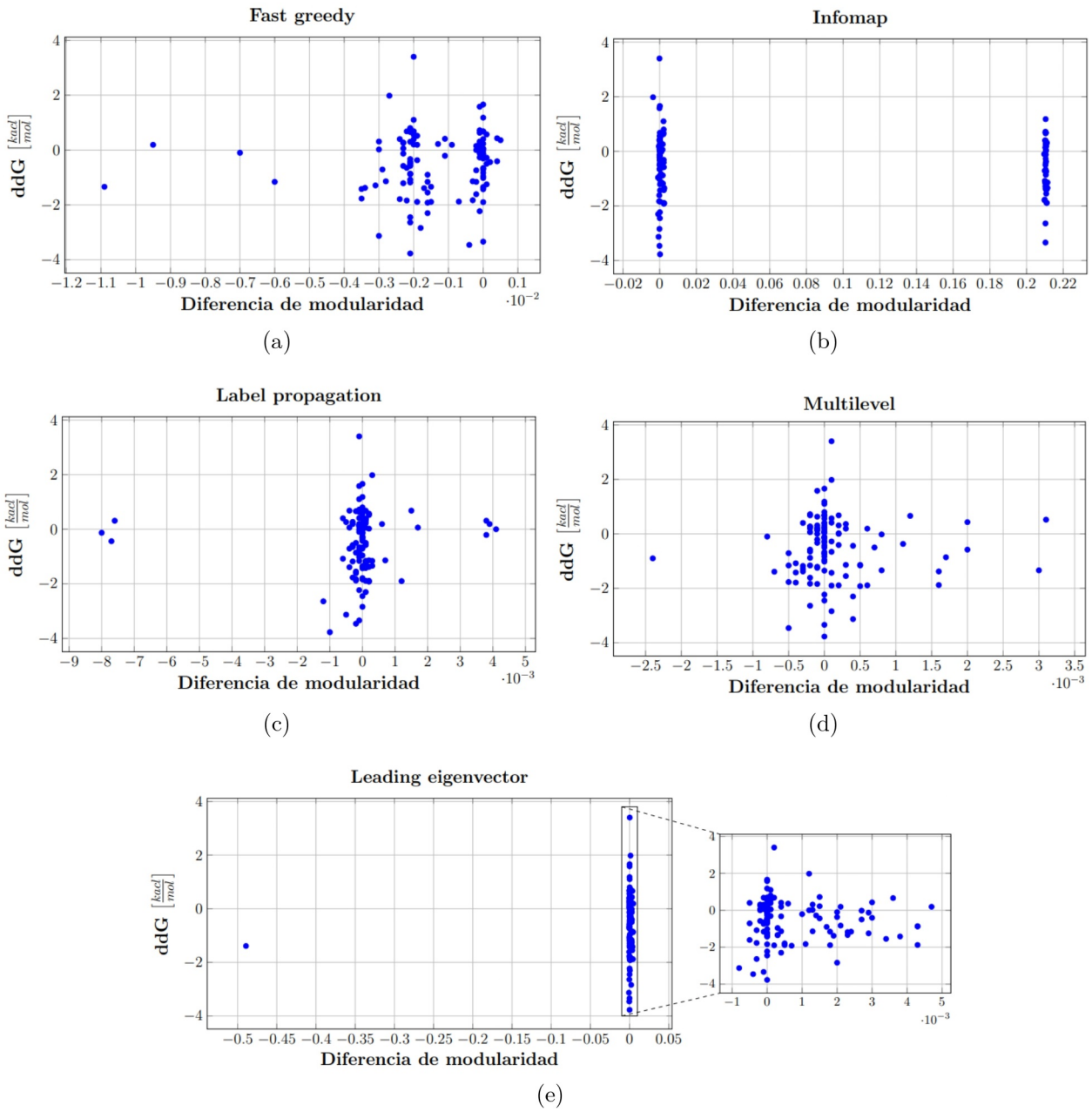
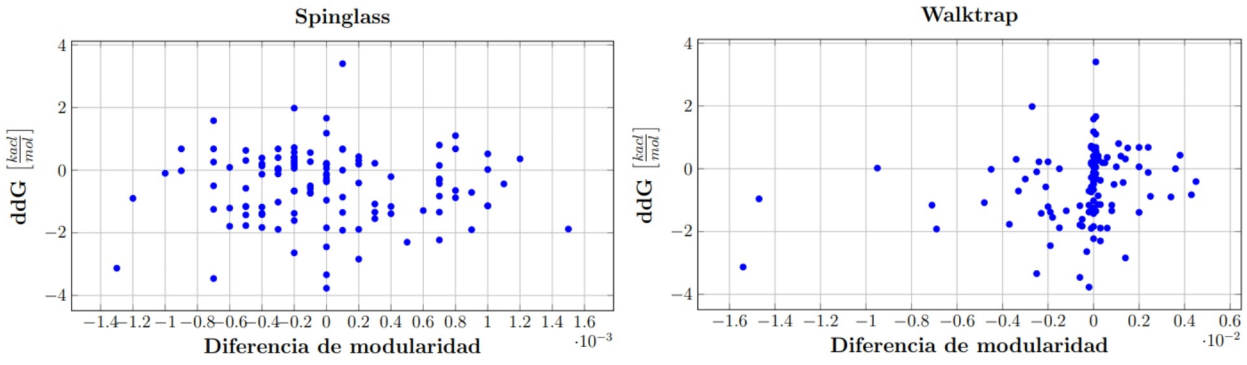
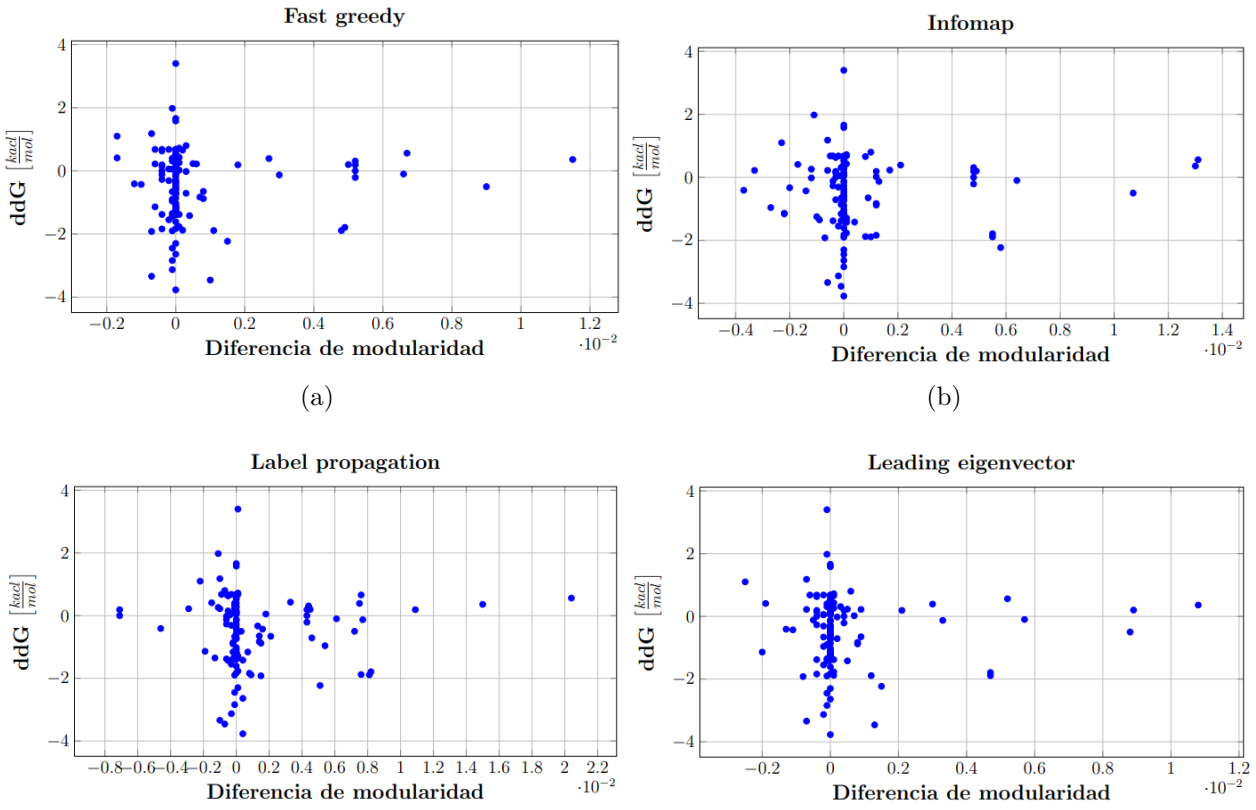


Figura 5.9: Variaciones de estabilidad según cambio de modularidad de los grafos de Centroide-distancia para los distintos algoritmos (a) Fast greedy (b) Infomap (c) Label propagation (d) Multilevel (e) Leading eigenvector.



(f) (g)

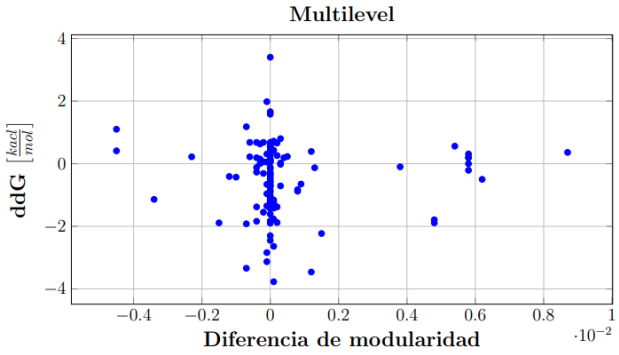
Figura 5.9: Variaciones de estabilidad según cambio de modularidad de los grafos de Centroide-distancia para los distintos algoritmos (cont.) (f) Spinglass (g) Walktrap.



(a) (b)

(c) (d)

Figura 5.10: Variaciones de estabilidad según cambio de modularidad de los grafos de Interacciones intermoleculares para los distintos algoritmos (a) Fast greedy (b) Infomap (c) Label propagation (d) Leading eigenvector.



(e)

Figura 5.10: Variaciones de estabilidad según cambio de modularidad de los grafos de Interacciones intermoleculares para los distintos algoritmos (cont.) (e) Multilevel.

5.4. Comparación de grafos

La comparación de grafos para observar el efecto de mutaciones sobre la p53 humana se realiza por medio del estudio de la similitud entre las comunidades de la proteína nativa y de una variante, para lo que se utilizan los indicadores similitud nat-mut y similitud mut-nat. Estos se presentan en tablas que consideran sólo los pares de comunidades que poseen alguna similitud y que sufrieron un cambio posterior a la mutación. Además, de destacar en color azul la fila que contiene el cambio del aminoácido mutado.

Es importante recordar que la similitud nat-mut indica que porcentaje del total elementos de la comunidad de la proteína nativa se mantienen en la comunidad de la proteína mutada. Por su parte, similitud mut-nat señala el porcentaje de elementos de la comunidad de la proteína mutada que se encuentran en la proteína nativa.

El análisis presentado corresponde a variantes de interés, siendo estas las mutantes con un mayor cambio en la estabilidad: G244D y P728L (como se puede observar en el Anexo B), y aquellas dos que sufrieron un cambio en la modularidad una mayor cantidad de veces con los distintos algoritmos: C141Y y T24E (como se muestra en el Anexo C).

Con motivo de simplificar la información entregada, esta sección solo contiene la comparación para los algoritmos de mayor interés según el tipo de grafo y excluye el fenómeno de renombramiento de comunidades (cuando una comunidad post mutación cambia su nombre pero no su composición) que no son de interés. Las tablas para todos algoritmos de detección estudiados considerando renombramiento se encuentran en el Anexo D.

5.4.1. Grafos Carbono alfa-Distancia

Los algoritmos seleccionados para esta representación corresponden a spinglass, debido a su alta modularidad promedio y label propagation, por la desviación estándar de su modularidad. Para el primero, las cuatro variantes de interés presentaron cambios, los que son presentados en las Tablas 5.5, 5.6, 5.7 y 5.8. Por el contrario, no existieron cambios con el

segundo algoritmo.

5.4.1.1. C141Y

Tabla 5.5: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Spinglass utilizando grafos carbono alfa-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
2	1	4,0	2,8
2	5	96,0	70,6
3	1	100,0	97,2
4	3	100,0	96,6
6	7	100,0	100,0
7	2	59,3	100,0
7	5	37,0	29,4
7	8	3,7	100,0
8	3	100,0	3,4

5.4.1.2. G244D

Tabla 5.6: Porcentaje de similitud entre comunidades de la proteína nativa y la variante G244D para el algoritmo Spinglass utilizando grafos carbono alfa-distancia. Elaboración propia

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
0	0	100,0	97,3
2	1	4,0	2,8
2	4	96,0	47,1
3	1	100,0	97,2
4	6	100,0	96,6
5	0	1,2	1,4
5	3	98,8	100,0
6	0	2,2	1,4
6	5	97,8	100,0
7	4	100,0	52,9
8	6	100,0	3,4

5.4.1.3. P278L

Tabla 5.7: Porcentaje de similitud entre comunidades de la proteína nativa y la variante P278L para el algoritmo Spinglass utilizando grafos carbono alfa-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
0	0	100,0	98,6
2	5	100,0	73,5
4	7	100,0	96,6
6	0	2,2	1,4
6	6	97,8	100,0
7	4	66,7	100,0
7	5	33,3	26,5
8	7	100,0	3,4

5.4.1.4. T284E

Tabla 5.8: Porcentaje de similitud entre comunidades de la proteína nativa y la variante T284E para el algoritmo Spinglass utilizando grafos carbono alfa-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
0	0	100,0	98,6
2	4	4,0	2,8
2	5	96,0	70,6
3	4	100,0	97,2
4	7	100,0	96,6
6	0	2,2	1,4
6	6	97,8	100,0
7	2	63,0	100,0
7	5	37,0	29,4
8	7	100,0	3,4

5.4.2. Grafos Centroide-Distancia

Los algoritmos seleccionados para esta representación corresponden a spinglass, debido a su alta modularidad promedio e infomap, por la desviación estándar de su modularidad. Para ambos, las cuatro variantes de interés presentaron cambios, los que son presentados en las Tablas 5.9, 5.11, 5.13 y 5.15 para el primero y en las Tablas 5.10, 5.12, 5.14 y 5.16 para el segundo.

5.4.2.1. C141Y

Tabla 5.9: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Spinglass utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
2	1	100,0	96,7
3	2	17,2	18,5
3	3	79,3	100,0
3	4	3,4	2,7
4	0	98,7	100,0
4	1	1,3	3,3
5	2	100,0	81,5
6	7	100,0	100,0
7	4	100,0	97,3

Tabla 5.10: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Infomap utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
1	1	100,0	92,0
2	1	20,0	8,0
2	2	80,0	33,3
3	2	100,0	66,7
6	5	42,4	100,0
6	6	35,4	100,0
6	7	22,2	97,8
7	7	2,6	2,2
7	8	97,4	68,5
9	8	100,0	31,5
10	10	100,0	84,2
11	10	17,6	15,8
11	11	82,4	100,0

5.4.2.2. G244D

Para esta variante en la Tabla 5.12 no se destaca ninguna fila, debido a que el aminoácido mutado no está involucrado en un cambio directo, manteniéndose en la comunidad 6.

Tabla 5.11: Porcentaje de similitud entre comunidades de la proteína nativa y la variante G244D para el algoritmo Spinglass utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
0	0	1,4	1,2
0	6	98,6	100,0
1	0	100,0	98,8
2	3	100,0	96,7
3	5	100,0	93,5
4	1	98,7	100,0
4	3	1,3	3,3
5	5	4,5	3,2
5	7	95,5	100,0
7	2	97,2	100,0
7	5	2,8	3,2

Tabla 5.12: Porcentaje de similitud entre comunidades de la proteína nativa y la variante G244D para el algoritmo infomap utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
10	10	100,0	84,2
11	10	17,6	15,8
11	11	82,4	100,0

5.4.2.3. P278L

Para esta variante en la Tabla 5.12 no se destaca ninguna fila, debido a que el aminoácido mutado no está involucrado en un cambio directo, manteniéndose en la comunidad 6.

Tabla 5.13: Porcentaje de similitud entre comunidades de la proteína nativa y la variante P278L para el algoritmo Spinglass utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
3	4	17,2	18,5
3	7	82,8	100,0
5	4	100,0	81,5
6	2	100,0	100,0

Tabla 5.14: Porcentaje de similitud entre comunidades de la proteína nativa y la variante P278L para el algoritmo Infomap utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
2	2	100,0	83,3
3	2	12,5	16,7
3	3	87,5	100,0
10	10	100,0	84,2
11	10	17,6	15,8
11	11	82,4	100,0

5.4.2.4. T284E

Tabla 5.15: Porcentaje de similitud entre comunidades de la proteína nativa y la variante T284E para el algoritmo Spinglass utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
0	6	98,6	100,0
0	7	1,4	1,2
1	7	100,0	98,8
2	0	100,0	96,7
3	3	17,2	18,5
3	5	82,8	96,0
4	0	1,3	3,3
4	4	98,7	100,0
5	3	100,0	81,5
6	2	100,0	100,0
7	1	97,2	100,0
7	5	2,8	4,0

Tabla 5.16: Porcentaje de similitud entre comunidades de la proteína nativa y la variante T284E para el algoritmo Infomap utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
6	6	100,0	99,5
7	6	2,6	0,5
7	7	97,4	100,0
8	8	100,0	100,0
9	9	100,0	100,0
10	10	100,0	84,2
11	10	17,6	15,8
11	11	82,4	100,0

5.4.3. Grafos Interacciones intermoleculares

Los algoritmos seleccionados para esta representación corresponden a multilevel, debido a su alta modularidad promedio y label propagation, por la desviación estándar de su modularidad. Para ambos, solo una de las cuatro variantes de interés presentó cambios, los que se muestran en las Tablas 5.17 y 5.18.

5.4.3.1. C141Y

Tabla 5.17: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo multilevel utilizando grafos de interacciones intermoleculares.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
82	82	73,3	84,6
82	101	26,7	30,8
90	82	10,5	15,4
90	90	42,1	80,0
90	101	47,4	69,2
103	90	100,0	20,0

Tabla 5.18: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Label propagation utilizando grafos de interacciones intermoleculares.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
93	93	20,0	33,3
93	95	80,0	100,0
109	93	100,0	66,6

En las Figuras 5.11, 5.12 y 5.13 se presenta una comparación visual de la detección de comunidades para las mutantes para el análisis realizado en las tablas anteriores. Es importante destacar que en los grafos de interacciones intermoleculares existe una gran cantidad de comunidades compuestas por un solo nodo, los cuales se colorearon gris oscuro para destacar las comunidades más grandes. Las figuras sin el cambio indicado anteriormente se encuentran en el Anexo E

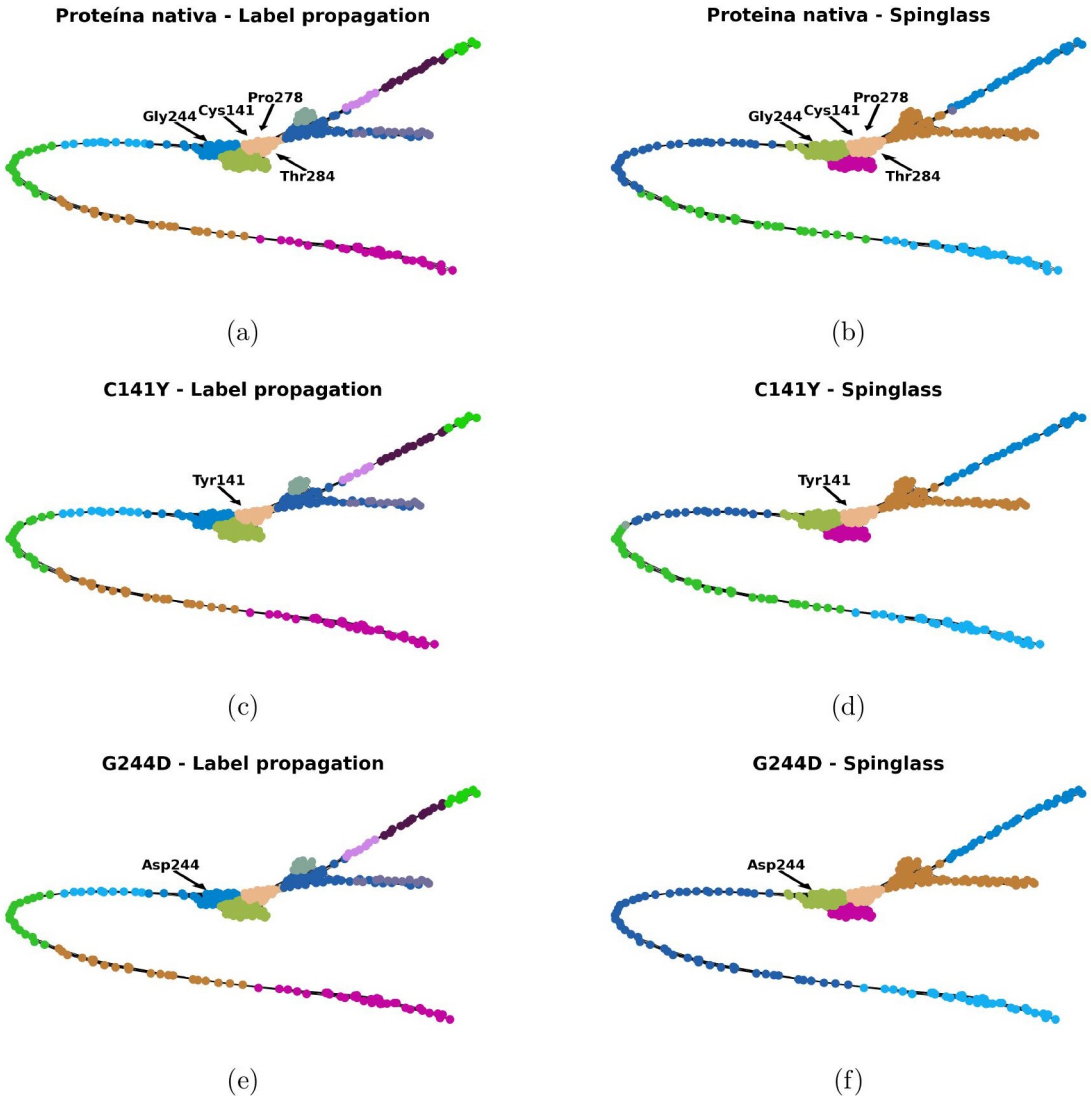
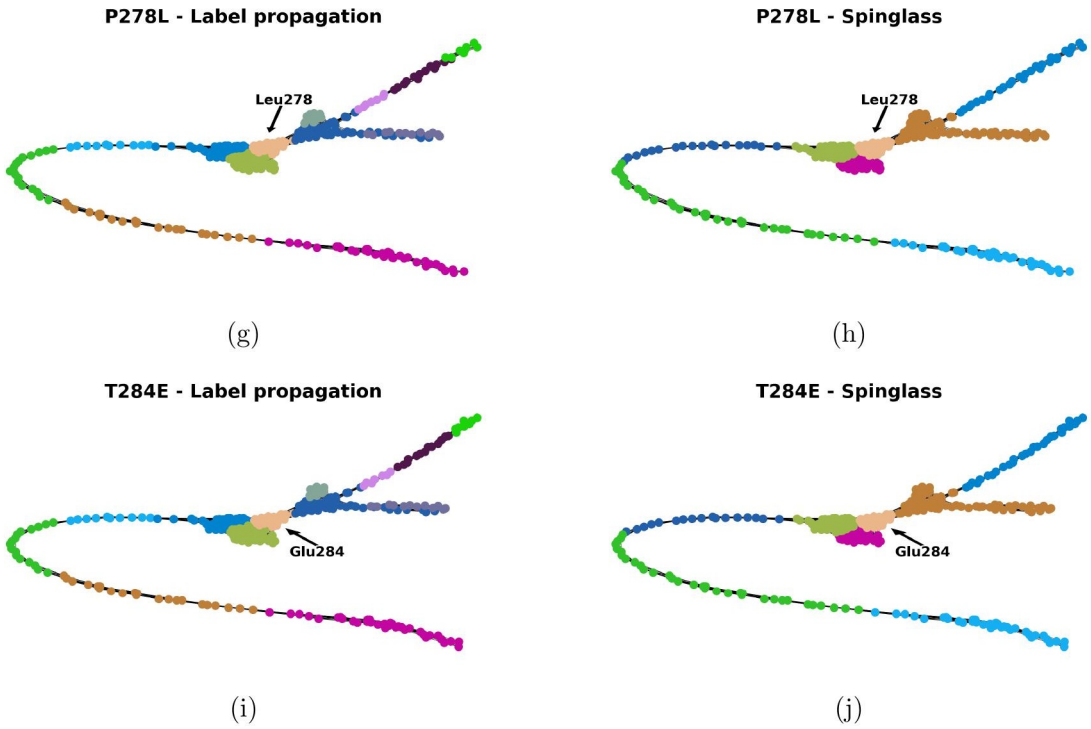


Figura 5.11: Comparación de detección de comunidades para la proteína p53 humana, representada con grafos de distancia-carbono alfa. Utilizando el algoritmo label propagation en (a) proteína nativa, (c) C141Y (e) G244D y spinglass en (b) proteína nativa, (d) C141Y (f) G244D.



(g) (h)

(i) (j)

Figura 5.11: Comparación de detección de comunidades para la proteína p53 humana, representada con grafos de distancia-carbono alfa (cont.). Utilizando el algoritmo label propagation en (g) P278L (i) T284E y spinglass en (h) P278L (j) T284E.

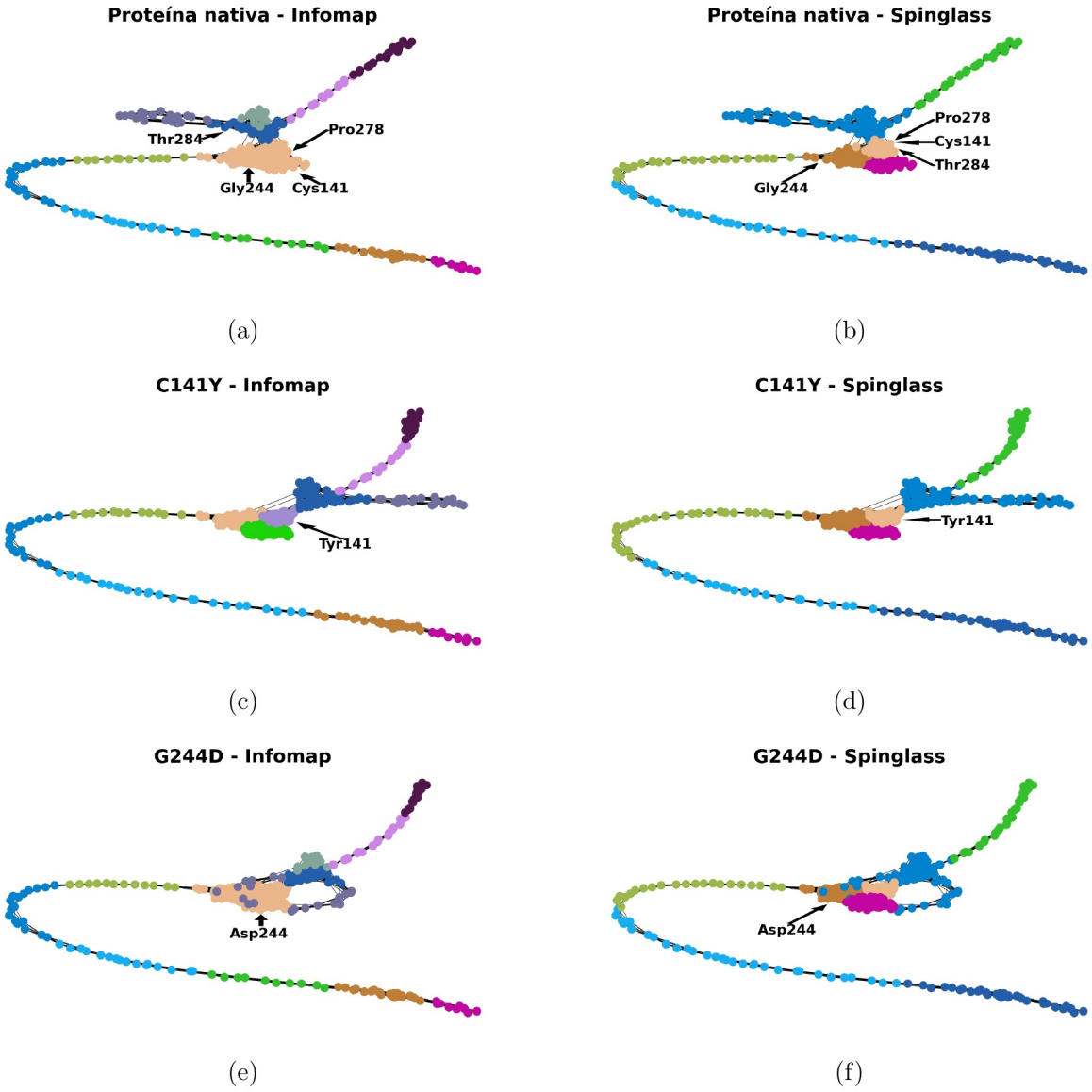
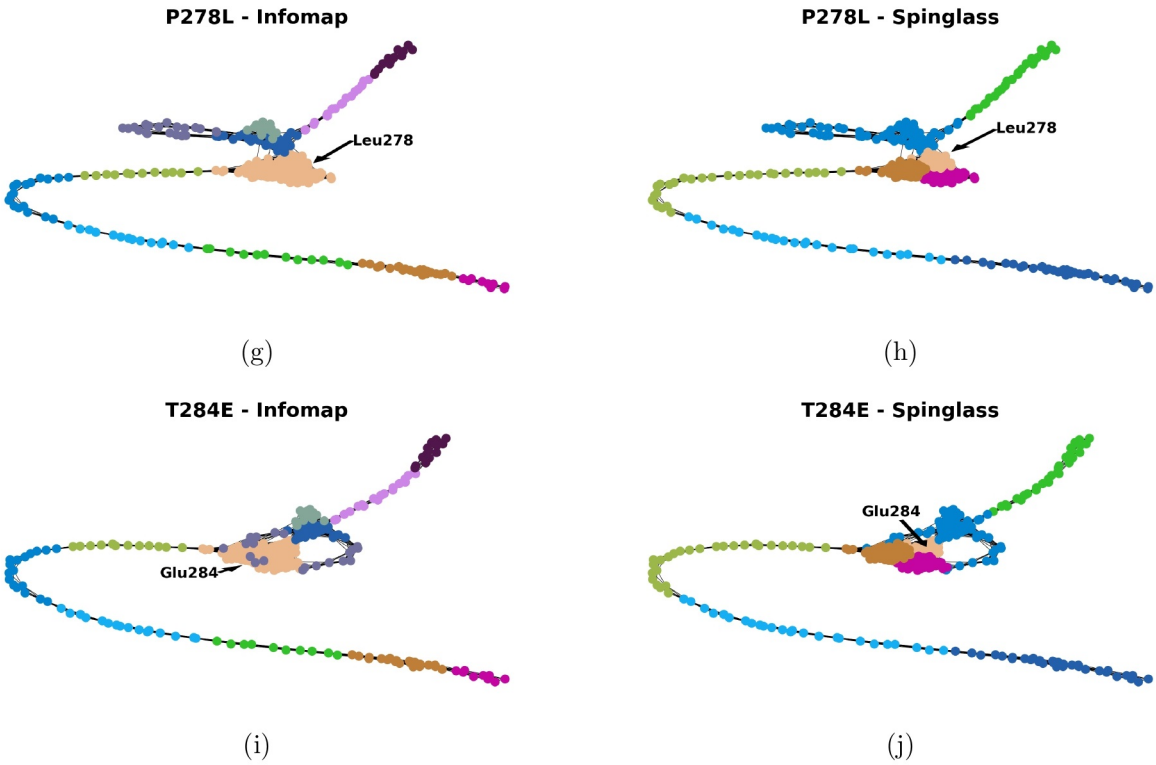


Figura 5.12: Comparación de detección de comunidades para la proteína p53 humana, representada con grafos de distancia-centroide. Utilizando el algoritmo infomap en (a) proteína nativa, (c) C141Y (e) G244D y spinglass en (b) proteína nativa, (d) C141Y (f) G244D.



(g) (h)

(i) (j)

Figura 5.12: Comparación de detección de comunidades para la proteína p53 humana, representada con grafos de distancia-centroide (cont.). Utilizando el algoritmo infomap en (g) P278L (i) T284E y spinglass en (h) P278L (j) T284E.

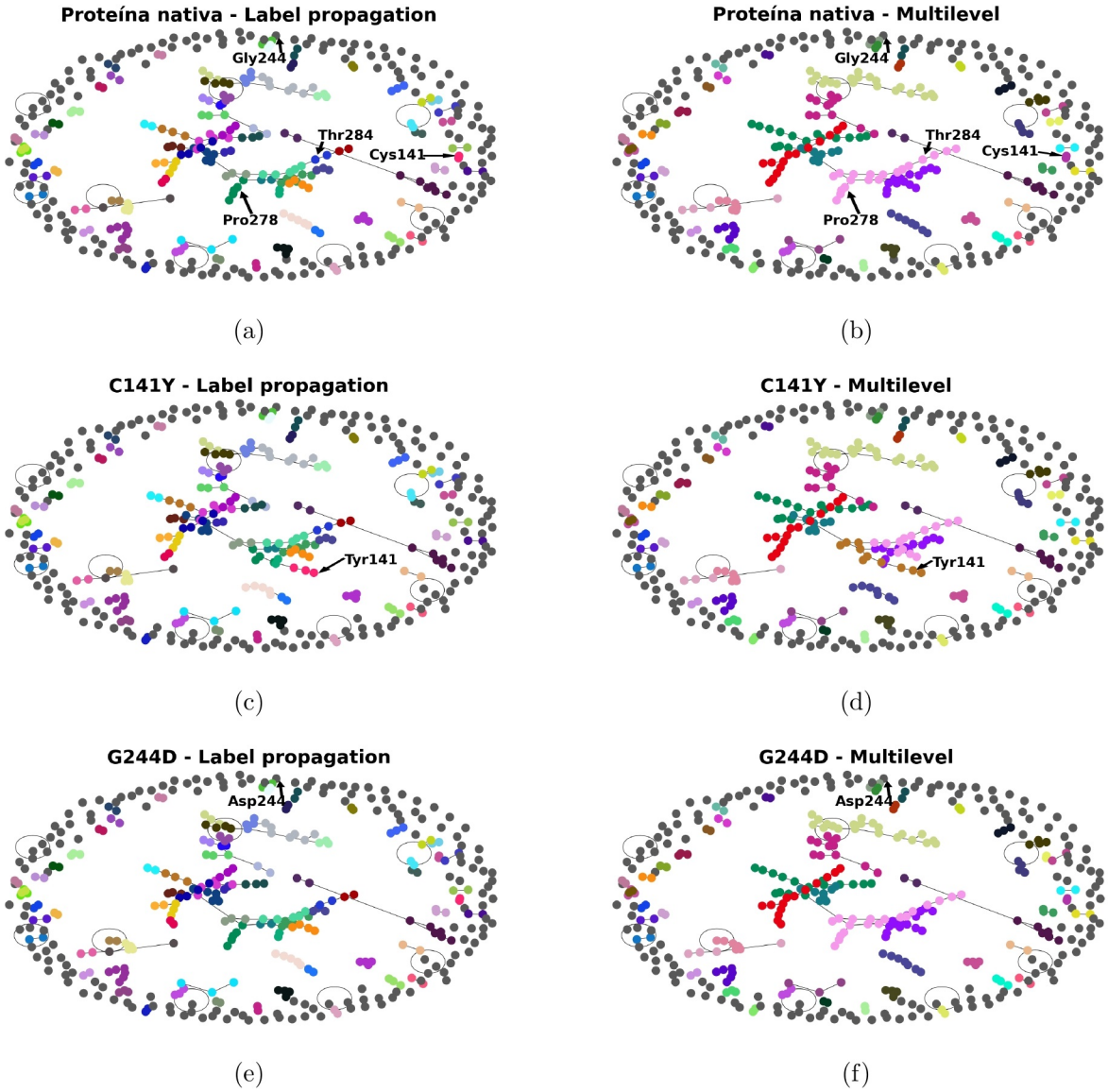
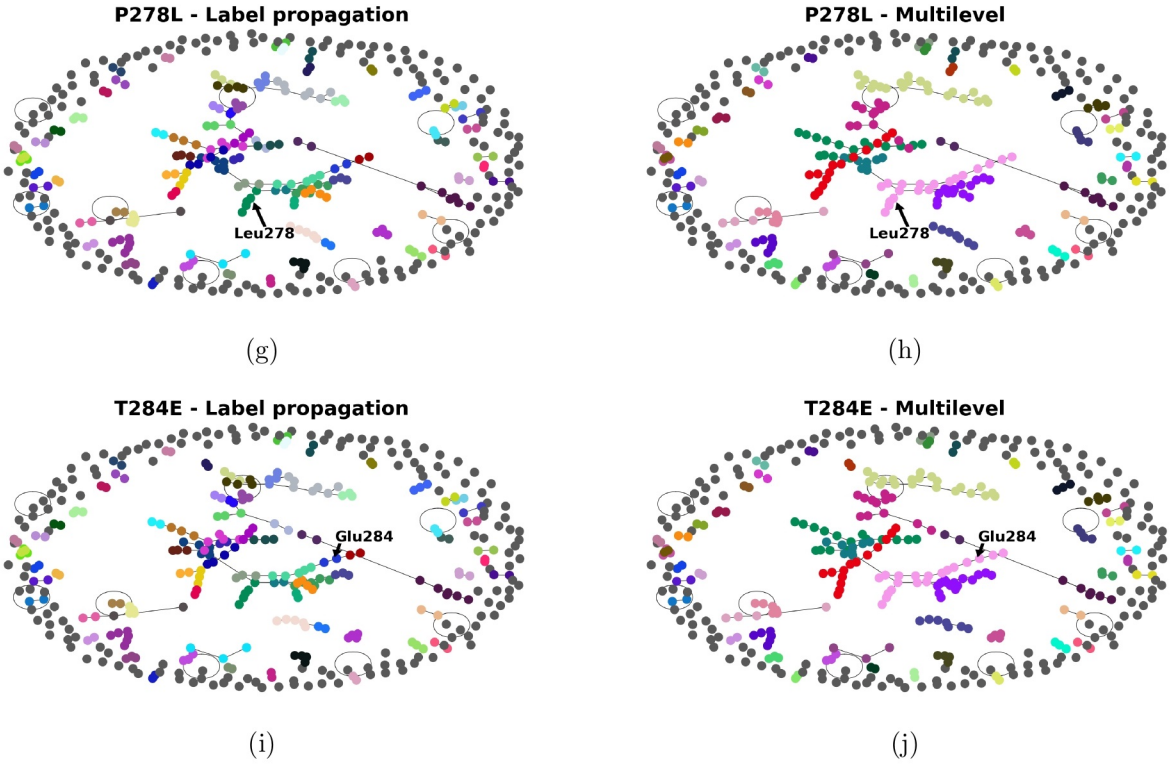


Figura 5.13: Comparación de detección de comunidades para la proteína p53 humana, representada con grafos de interacciones intermoleculares. Utilizando el algoritmo label propagation en (a) proteína nativa, (c) C141Y (e) G244D y multilevel en (b) proteína nativa, (d) C141Y (f) G244D.



(g) (h)

(i) (j)

Figura 5.13: Comparación de detección de comunidades para la proteína p53 humana, representada con grafos de interacciones intermoleculares (cont.). Utilizando el algoritmo label propagation en (g) P278L (i) T284E y multilevel en (h) P278L (j) T284E.

5.5. Graph Convolutional Networks

El desempeño de la red neuronal programada se puede observar a través del gráfico de accuracy que se encuentra en la Figura 5.14. Además, ya que la tarea ejecutada corresponde a clasificación, se presenta la matriz de confusión del modelo con los parámetros obtenidos en la época 100. Finalmente, las métricas clásicas de desempeño son indicadas en la Tabla 5.19.

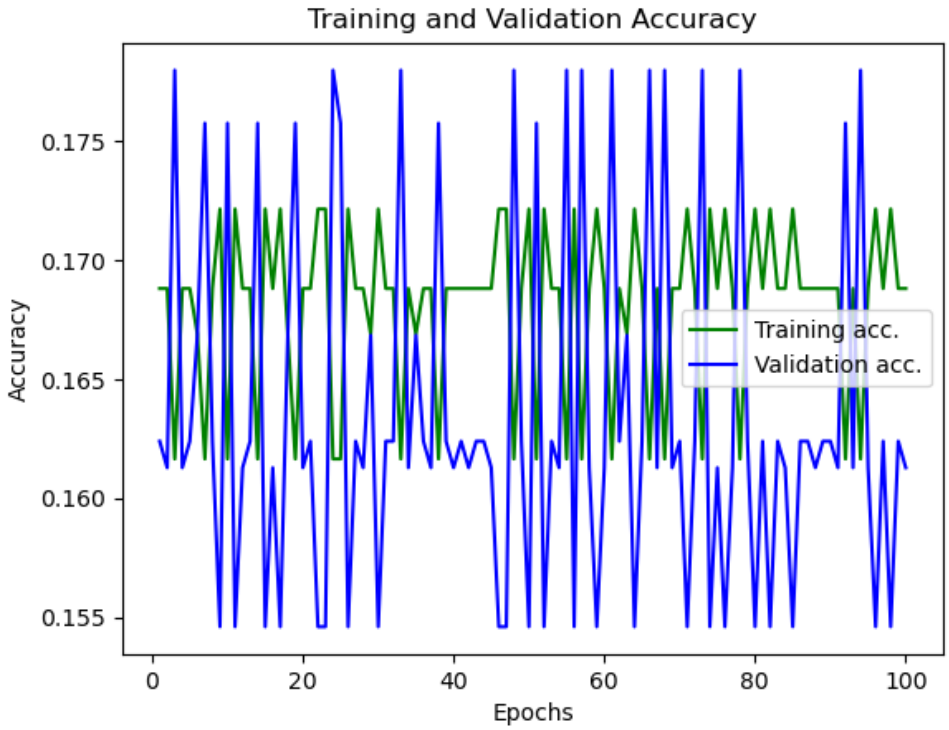


Figura 5.14: Accuracy del modelo a lo largo de las epochs.

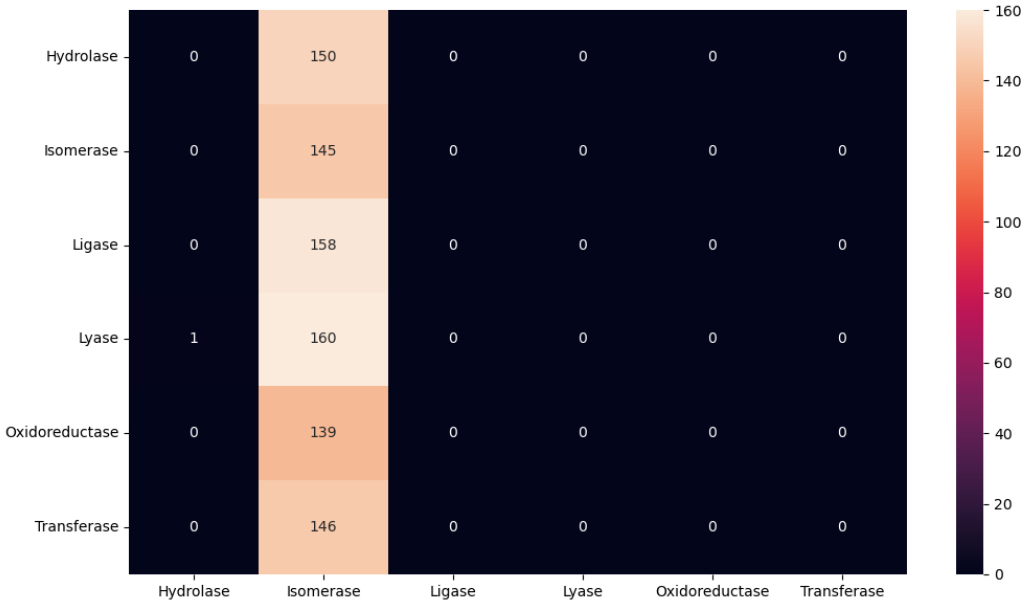


Figura 5.15: Matriz de confusión del modelo para la última época. Las clases reales se encuentran en las filas y las predichas en las columnas.

Tabla 5.19: Métricas de desempeño del modelo de clasificación de enzimas.

Métrica	**Accuracy**	**Recall**	**Precision**	**F1-score**
Valor	0.1613	0.16667	0.0269	0.0463

Capítulo 6

Discusiones

En la presente sección se analizan y discuten los resultados obtenidos durante la investigación llevada a cabo en esta memoria.

6.1. Adquisición de datos

Originalmente para la memoria se crearon seis conjuntos de datos, los cuales son presentados en la Sección 4.1. Estos se pensaron con el fin de llevar a cabo distintas tareas que permitiesen evaluar el desempeño de los grafos como representación de proteínas. Sin embargo, por limitaciones de tiempo sólo se utilizan dos de ellos, los conjuntos llamados Función enzimática y Variantes de p53 humana. Se decide mantener la totalidad de estos en la metodología como una proyección que facilite el trabajo de quienes deseen continuar con el estudio propuesto.

Un punto interesante respecto a los conjuntos de datos es la presencia de elementos multicategoria, sobre todo en el conjunto de datos: Estructura Secundaria. Respecto a este, la categorización SCOP se realiza considerando los dominios como la unidad de clasificación. Entonces, podrían existir dos dominios en una proteína que correspondan a distintas clasificaciones, sobre todo en el caso de proteínas grandes, entregando más de una categoría para una proteína. En base a esto, para la continuación del trabajo se podrían utilizar dominios como base de estudio en vez de proteínas, pudiendo analizar los grafos de cada una de manera individual, para luego realizar el análisis de la proteína completa, que puede ser representadas por la unión de distintos grafos donde los nodos estén etiquetados según su dominio y observar como se comportan las comunidades según esto.

6.2. Generación de grafos

En la generación de grafos, la selección de utilizar no dirigidos se sustenta debido a la bidireccionalidad de las relaciones seleccionadas. En el caso de la distancia, esta será la misma entre el punto A y B que entre B y A. En el caso de las interacciones electroestáticas, si bien existe un donor y un aceptor que podría establecer un sentido de la atracción, solo se desea estudiar la influencia de su intensidad y no su direccionalidad.

Respecto a la selección de información para nodos y aristas, no se consideran los grafos de interacciones intermoleculares a nivel de centroide, ya que la información aportada por el

servicio WHAT IF es a nivel de residuo, siendo independiente de cómo se represente este. De esta manera un grafo de carbono alfa-interacción intermolecular es equivalente a uno de centroide-interacción intermolecular, ya que su única diferencia corresponde a la ubicación del punto que representa al residuo, que es irrelevante para este tipo de grafo. WHAT IF también indica los átomos que establecen la interacción, por lo que se pudo considerar crear grafos átomo-interacción molecular, lo cual fue descartado para simplificar el manejo de datos en el presente trabajo. Otro tipo de representación no explorada, corresponde a multigrafos, que incorporen la información de distancia e interacciones intermoleculares a la vez. En este caso, el filtro de distancia debería considerar al menos el valor mínimo para generar un puente de hidrógeno.

Si bien en los resultados sólo se presenta una proteína a modo de ejemplo, a partir de su estudio se pueden determinar ciertas propiedades de las distintas representaciones propuestas. En un primer nivel, si se analiza la granularidad de los grafos se puede observar que la cantidad de nodos variará dependiendo del grosor seleccionado. Así, como se muestra en la Tabla 5.1 este será mayor mientras más fina es la granularidad, lo cual tiene sentido ya que un aminoácido luego de formar un enlace peptídico posee a lo menos cuatro átomos distintos de hidrógeno (Un carbono alfa, un carbono y un oxígeno en el grupo carboxilo y un nitrógeno en el grupo amino), valor que aumentará al considerar los átomos de los radicales. En la granularidad más gruesa se tiene que carbono alfa y centroide poseen una misma cantidad de nodos. Esto es debido a que ambos son una manera de representar lo mismo: un residuo aminoacídico, por lo que el número de nodos corresponderá a la cantidad de residuos de la proteína. Las interacciones intermoleculares usan carbonos alfas como granularidad, por lo que se suma al análisis anterior.

En el caso de los grafos de distancia, al comparar aquellos que usan átomos con uno de residuos se observa que el primero presenta más aristas. Esto es debido a que como se discutió anteriormente un residuo está compuesto por al menos cuatro átomos que interactúan entre sí y con otros nodos. Así al pasar de un grafo más grueso a uno más fino, si bien las interacciones entre los átomos de un mismo residuo pueden ser filtradas a causa de su cercanía, las que estos presentan con los nuevos nodos ubicados en torno a donde antes había un residuo vecino probablemente se mantendrán, ya que serán igual a la distancia original entre los dos residuos más o menos una variación. De esta manera, la disminución de granularidad producirá un aumento en el número de aristas.

Otro punto importante por observar es que, si bien los grafos de distancia generados con centroide y carbonos alfa son utilizados para representar residuos, ambos conceptos son distintos y, por ende, la distancia relativa entre dos nodos puede ser distinta si se usa uno o el otro, que en el ejemplo presentado se traduce en un distinto número de aristas luego del filtro.

Además, se pueden observar diferencias entre los grafos según la información que poseen en las aristas. Los grafos de distancia pueden variar entre conectados o no, dependiendo del rango de filtrado que se ocupe. Así, al usar un rango desde cero hasta el infinito (o no se filtra) cada nodo será vecino de todos los demás, obteniendo un grafo completo, y por tanto conectado. Al ir disminuyendo el rango, se van eliminando aristas entre los nodos pudiendo llegar a un punto en que un nodo quede sin interacciones, lo que se traduce en un grafo no conectado . En el caso específico del ejemplo presentado en la Figura 5.7, se intuye que, con

el rango de filtrado usado, existe la suficiente holgura para que cada nodo tenga al menos una conexión con otro. Por otra parte, en los grafos de interacciones intermoleculares solamente se consideran puentes de hidrógeno, los cuales son formados sólo por ciertos residuos, lo que produce que no todos los nodos tengan vecinos y por lo tanto sean grafos no conectados.

Finalmente, cabe destacar que mayor cantidad de aristas y nodos se traduce en un mayor peso de los archivos gpickle, debido a que se debe guardar una mayor cantidad de información, por lo que los grafos de átomos son generalmente más pesados que los demás.

6.3. Detección de comunidades

Una comunidad es un grupo de nodos altamente conectados entre ellos y con pocas aristas con los nodos pertenecientes a otros módulos. En los grafos de distancia, esto implica zonas cuyos elementos se encuentran a distancias dentro del rango de filtrado, pudiendo estar cerca o lejos dependiendo de como se configure este, y en los de interacciones intermoleculares representan sectores con redes de puentes de hidrógeno, pudiendo corresponder a sitios catalíticos o zonas de interacción con ligandos en modelos proteína-ligando. Por lo anterior, se cree que el estudio de las comunidades podría aportar a un mejor entendimiento de las propiedades estructurales y funcionales de las proteínas.

La tarea de detección de grafos se efectuó sobre tres de las cuatro representaciones de grafos propuestas, excluyendo aquellos que usan átomos en sus nodos debido al alto costo computacional que implicaba generarlos. No se pudo estudiar, entonces, el comportamiento de los algoritmos en estos.

En las Tablas 5.2, 5.3 y 5.4 se muestra el comportamiento de los algoritmos en las representaciones que sí fueron utilizadas. Para la modularidad se estudian dos medidas de interés; el promedio, que indica como el algoritmo aísla los módulos en la red proteica, y la desviación estándar, que entrega información sobre la dispersión de los datos, pudiendo servir como un indicador de la sensibilidad de este parámetro a las mutaciones. En cuanto al número de comunidades, este permite tener una intuición más simple de cómo cambian las comunidades detectadas.

En los grafos de Carbono alfa-distancia se identifican dos algoritmos de interés. Spinglass, presenta la mayor modularidad y una desviación estándar del orden más bajo respecto a los demás. Esto lo posiciona como el algoritmo que aísla de mejor manera los módulos presentes en la red proteica, y que posee una baja sensibilidad promedio a las mutaciones. Al observar el número de comunidades se nota que, a pesar de que estas poseen la desviación más alta de todos los algoritmos, la modularidad se mantiene estable, lo que indica una baja respuesta de este parámetro antes los cambios en la proteína. Por otra parte, Label propagation presenta la tercera modularidad promedio más baja y la desviación más alta, no teniendo el mejor desempeño al identificar los módulos, pero siendo a primera vista más sensible a las mutaciones. Esta última característica podría ser deseable para observar el efecto de mutaciones puntuales sobre proteínas, si se demuestra una conexión entre esta propiedad y cambios fisicoquímicos en la macromolécula. Combinando este análisis con lo mostrado en el eje de las ordenadas de los gráficos presentados en la Figura 5.8, se nota que en Label propagtion, existen dos puntos con un comportamiento anómalo respecto a los demás (R283C y C238S), con una

diferencia de modularidad de al menos tres órdenes de magnitud mayor que el resto (lo que puede ser comprobado en la Tabla C.3 del Anexo C), siendo estos los que condicionan el valor de la desviación estándar, Lo anterior termina descartando la sensibilidad del algoritmo a los cambios producidos por las mutaciones. En cambio, en el gráfico de Spinglass no se notan cambios de dimensiones tan bruscos.

De manera similar, en los grafos de Centroide-distancia Spinglass se mantiene como el algoritmo con modularidad promedio más alta y una desviación estándar del orden más bajo, con un comportamiento similar al caso anterior en el número de comunidades, lo que permite realizar un análisis análogo. En cambio, el algoritmo con mayor desviación es Infomap, siendo esta órdenes de magnitud mayor que los otros algoritmos. Esto se debe a la segmentación de puntos que se observa en la Figura 5.9, donde en el gráfico asociado al algoritmo se presentan dos nubes de puntos, una centrada en cero y la otra en 0.21 aproximadamente. Se nota que todas las mutantes con una comunidad menos se encuentran en el primer grupo (lo que puede ser comprobado en la Tabla C.9 del Anexo C). Sin embargo, este grupo también esta compuesto variantes en que el total de módulos se ha mantenido invariable.

Un algoritmo que destaca en las dos representaciones previas es multilevel, que no presenta cambios en el número de comunidades y posee la segunda modularidad promedio más alta con una desviación estándar del orden más bajo. Esto podría indicar una baja sensibilidad del algoritmo a las variaciones de distancias en la proteína.

La diferencia de desempeño de los algoritmos en los grafos de carbono alfa y centroide se sustenta en que, a pesar de ser ambos una manera de representar los residuos, poseen distancias diferentes, lo que se traduce en una cantidad de aristas variables, cambiando la densidad de conexiones de las distintas zonas de la proteína.

Finalmente, en los grafos de Interacciones intermoleculares, debido a que corresponden a grafos no completos, no son aplicables dos algoritmos: Spinglass y Walktrap, siendo el primero de interés en las dos representaciones anteriormente discutidas, por lo mismo, no se puede analizar si este seguiría manteniendo su desempeño. Considerando lo anterior, el algoritmo con mayor modularidad promedio en este caso corresponde a multilevel. Respecto a la desviación estándar de este parámetro los cinco algoritmos entregan valores del mismo orden, destacando Label propagation como el que posee un mayor valor. En esta representación se obtiene un número de comunidades más elevado que en las anteriores debido a las escasas conexiones establecidas, existiendo módulos de sólo un nodo, siendo entonces de interés solamente aquellas comunidades con una mayor cantidad de residuos.

Respecto a la relación entre el cambio de estabilidad producido por las mutaciones y los cambios de modularidad, en las Figuras 5.8, 5.9 y 5.10 se muestra esta para los tres tipos de grafos analizados. En ningún caso en específico se puede notar una correlación directa entre ambas variables, apareciendo nubes altamente dispersas centradas principalmente en el eje de las abscisas. Lo anterior se puede deber a la escala de los gráficos, existiendo algunos puntos con cambios de modularidad órdenes de magnitud mayores a los *clusters* del centro, dificultando la visualización de los datos.

6.4. Comparación de grafos

Para la comparación de las comunidades se seleccionaron variantes de interés como G244D y P278L, debido a que estas son las que producen el mayor cambio de estabilidad. Este se justifica en el tipo de sustitución que ocurre. En el primer caso, una glicina que corresponde a un aminoácido pequeño y sin carga, por un ácido aspártico, que es más grande y con carga negativa, la que puede generar una desestabilización en la proteína al interaccionar con otros aminoácidos con carga. El segundo caso es una sustitución de una prolina por una leucina. La primera regularmente aporta rigidez a las cadenas debido a su anillo. Así, al ser reemplazada provoca que la molécula sea más laxa, pudiendo aumentar la estabilidad. Ambas mutaciones pueden cambiar considerablemente las distancias entre los residuos debido a las interacciones electroestáticas en el primer caso y a la variación de ángulo en el segundo .

Al observar el cambio de las comunidades que ocurre en estas mutantes para los grafos con Carbono alfa-distancia, se tiene que ambas sufrieron alteraciones considerables sólo para el algoritmo de Spinglass. Para G244D, el aminoácido mutado se encuentra originalmente en la comunidad 5, moviéndose a la 3 luego de la mutación. En la Tabla D.3 se presenta la similitud entre los módulos, para el par 5 original - 3 variante se observa que existe un 100 % de similitud mut-nat, pero sólo un 98,8 % en el sentido contrario. Esto indica que la comunidad 3 post mutación posee el 98,8 % de los aminoácidos de la comunidad 5 nativa, moviéndose el porcentaje restante a la comunidad 0. Además, se observa que todos los aminoácidos en la comunidad 3 son originalmente de la comunidad 5, entre estos el aminoácido mutado. También se nota que la mutación es acompañada por el movimiento de otros aminoácidos a nuevos módulos, lo que indica variaciones en las distancias entre los aminoácidos, vislumbrando un posible cambio epistático.

A su vez, P278L posee el aminoácido mutado inicialmente en el módulo 6, manteniéndose en este luego de la mutación. Semejante al caso anterior, se observa una similitud de 97,8 % entre la comunidad 6 nativa y la 6 mutada, y de un 100 % en el otro sentido, ocurriendo un cambio de algunos nodos a la comunidad 0 mutada, lo que se muestra en la Tabla D.4.

En la Figura 5.11 a), b), e), f), g) y h), se presentan las comunidades pre y post mutación, y sus cambios para Spinglass. Se incluye Label propagation para mostrar un estado sin cambios.

En relación a los grafos de Centroide-distancia, se nota que las variantes poseen cambios en más algoritmos, siendo los comunes Fast greedy, Infomap y Spinglass. Para este último, G244D posee inicialmente el nodo mutado en la comunidad 1 y luego pasa a estar en la comunidad 0. Según lo indicado en la Tabla D.16, ambas comunidades tienen una similitud del 100 % desde la proteína nativa a la mutada y 98,8 % a la inversa. Así, la comunidad 0 post mutación posee todos los nodos de 1 más algunos extra. La mutación genera un cambio en cadena en los otros módulos, existiendo renombramientos como es el caso del par 6-4 y resdistribuciones, lo que nuevamente indica cambios en la densidad de conexiones, vislumbrando cambios epistáticos.

A su vez, P278L ubica su nodo mutado inicialmente en el módulo 6 cambiando a 2 posterior a la sustitución. Si se observa la Tabla D.20, se nota que el coeficiente de similitud para el par 6-2 es de 100 % en ambos sentidos, por lo que no ocurrió una migración sino un renom-

bramiento de la comunidad, manteniéndose los mismos elementos. A pesar de esto, si ocurren redistribuciones de nodos en el resto de la proteína. Por ejemplo, el nuevo módulo 4 es una combinación de elementos del 3 y 5 original. Lo anterior indica cambios en las relaciones que implican variación de distancias mayor a 2 Å.

Ambas situaciones son observables en la Figura 5.12 a), b), e), f), g) y h), donde son las comunidades muestran cambios para Spinglass e Infomap.

Con respecto a las otras variantes de interés, C141Y y T284E, fueron seleccionadas por ser dos de las experimentan más variaciones en la modularidad. En la primera existe una sustitución de cisteína por tirosina, poseyendo la última un mayor volumen, lo cual puede generar una desestabilización en la molécula. Comúnmente, la cisteína genera puentes disulfuros, pero en esta proteína no están presentes, por lo que la sustitución tiene un efecto menor al que ocurriría si esta interacción existiese. En la segunda, la mutación corresponde al intercambio de una treonina por ácido glutámico. Este último es ionizable con carga negativa, lo que puede generar interacciones con otros aminoácidos que a su vez podría provocar una variación de las distancias y la estabilidad.

En los grafos Carbono alfa-distancia, ambas mutantes presentan cambios con el algoritmo Spinglass. En C141Y el aminoácido mutado originalmente se encuentra en el módulo 6 cambiando al 7 posterior a la mutación. Sin embargo, existe una similitud de 100 % en ambos sentidos, lo que se traduce en una equivalencia entre ambas comunidades, siendo este cambio esencialmente solo un renombramiento. A pesar de esto, existen variaciones en las otras comunidades, ocurriendo reparticiones como es el caso de la 7 original, cuyos nodos se dividen entre la 2, 5 y 8 mutadas, lo que es representado en la Tabla D.2. Por su parte, T284E conservó el aminoácido mutado en la comunidad 6, existiendo un movimiento de otros nodos similar al que ocurre con P278L, lo que se puede observar en la Tabla D.5.

En la Figura 5.11 a), b), c), d), i) y j), se presentan las comunidades pre y post mutación, y sus cambios para Spinglass. Se incluye Label propagation para mostrar un estado sin cambios.

En los grafos Centroide-distancia, ambas mutantes presentan cambios para los algoritmos de Fast greedy, Infomap, Leading eigenvector, Spinglass y Walktrap. En Spinglass, C141Y presenta el nodo mutado en la comunidad 6 previo a la sustitución y luego en la 7. Como se puede notar a partir de los coeficientes de similitud en la Tabla D.10, ocurrió un renombramiento de 6 por 7, no habiendo un cambio efectivo. Sin embargo, la variación de distancias sí produjo cambios en la distribución de los otros nodos. Para el mismo algoritmo, en T284E ocurre un cambio del aminoácido mutado del nodo 6 al 2, pero al observar la Tabla D.24 se puede concluir que esto ocurre debido a que la comunidad original se renombra. A pesar de esto, si suceden cambios en las distribuciones de los otros residuos producto de la mutación.

Ambas situaciones son observables en la Figura 5.12 a), b), c), d), i) y j), donde son las comunidades muestran cambios para Spinglass e Infomap.

Se considera relevante discutir que para las representaciones de distancia, un cambio en Carbono alfa posee un menor impacto que en Centroide. Esto se debe a que en los primeros la sustitución sólo considera cómo se mueve la cadena principal producto de las nuevas in-

teracciones establecidas y el cambio de ángulo generado. Los segundos incluyen las misma variaciones anteriores, pero agregan un nuevo factor: el tamaño y forma del nuevo aminoácido, no sólo representando la consecuencia del cambio sino que también la causa.

Finalmente, para los grafos de Interacciones intermoleculares de las cuatro variantes solamente C141Y presenta cambios, lo que se puede observar en la Figura 5.13. Esto es contrario a lo esperado, ya que en P278L y T284E el nodo mutado se encuentra estableciendo un puente de hidrógeno según la red calculada por WHAT IF, por lo que se predecía que se producirían cambios en la detección. Una razón para esto puede ser que si bien en los tres casos al ocurrir la mutación se mantiene la interacción, para C141Y ocurre el mayor cambio de intensidad del puente de hidrógeno correspondiendo a 0.255 (medido según la escala de WHAT IF) en comparación a -0.035 y 0.05 para P278L y T284E, respectivamente. Por su parte, en la posición 244 no se generan puentes de hidrógeno antes o después de la mutación, por lo que no existen cambios en este nodo para G244D y al parecer la variación de posición de la cadena principal que ocurre con el cambio de aminóacido no afecta las otras interacciones intermoleculares de la proteína, reflejándose en que no hay cambios en para esta mutante.

Para el algoritmo de multilevel, que es el que presenta mayor modularidad promedio para este tipo de representación, se nota que en C141Y el nodo mutado se encontraba inicialmente en la comunidad 103, siendo su único elemento, para transportarse a la 90 luego de la sustitución. Así, la comunidad 103 original desaparece y parte de los elementos de la comunidad 90 original se integran al nuevo módulo 101, reestructurando la red de interacciones.

Ya que se analizaron casos puntuales como ejemplo, es complejo deducir el comportamiento general de los algoritmos para la comparación de los grafos. Sin embargo, resulta notable que para los grafos de Carbono alfa-distancia, Spinglass que presenta una de las menores distribuciones estándar, demuestra cambios en los módulos, en contraposición a Label propagation, que tiene la mayor dispersión, pero no muestra cambios.

6.5. Graph Convolutional Network

El desempeño de la red neuronal programada puede ser evaluado a partir de la Fig. 5.14, donde se observa que la exactitud del modelo se mantiene bajo el 20 % a lo largo de las 100 épocas utilizadas para entrenar el modelo. Lo anterior está relacionado con lo mostrado por la matriz de confusión presentada en la Fig. 5.15, donde se indica que luego del entrenamiento el modelo predice casi únicamente la clase *isomerase*, la que corresponde a un sexto del conjunto de datos, ya que este se encuentra balanceado. El único otro caso es la predicción errónea de *hydrolase* cuando el valor real es *lyase* un 0,11 % de las veces. Además, en la Tabla 5.19, se muestra un Recall del 0.167, una Precisión de 0.026 y un F1-score de 0.0463, todos estos valores están asociados a que se predice una única clase y que la no predicción se maneja con valor cero para el calculo del Macro promedio de estos parámetros.

Motivos para el pobre desempeño de la red pueden ser el no ajustar los hiperparámetros, teniendo, por ejemplo, una tasa de aprendizaje no óptima, el uso de una baja cantidad de datos por clase o el definir un vector de caracteristicas no representativo. Esto fue resultado de la limitación de tiempo y recursos para el entrenamiento de la red, no pudiendo realizar tareas que implicaran una alta complejidad computacional. Otra causa del bajo rendimiento

de la red puede ser el problema denominado ReLu moribundo, donde una neurona "muere" cuando todas las entradas ponderadas a esta son negativas, lo que provoca que la salida siempre sea cero, afectando el descenso del gradiente de las siguientes iteraciones, ya que el gradiente de la función ReLU es 0 cuando su entrada es negativa.

Más allá de que los resultados obtenidos no son ideales, se logró probar una metodología para utilizar las representaciones generadas en tareas de *deep learning*. Es necesario incluir el paso de ajustar los hiperparámetros, para poder comprender a cabalidad la utilidad de las representaciones generadas. Sin embargo, el potencial de los grafos en aprendizaje profundo existe, como se demostró en modelos creados anteriormente por Claudio Guevara en su tesis "Aplicaciones de estructuras de grafos y aprendizaje profundo a sistemas de clasificación de interacción antígeno-anticuerpo" [61], donde se obtuvo un rendimiento del 51 %.

Capítulo 7

Conclusiones

La investigación llevada a cabo tuvo un fin principalmente exploratorio, como se indica en su objetivo general, el cual se considera llevado a cabo en su totalidad. Esto se debe a que se diseñaron e implementaron *scripts* que permitieron estudiar el potencial de las estructuras de grafos como representación de proteínas en distintas tareas computacionales.

Una de las principales limitantes para esta investigación fue la alta capacidad computacional requerida para la generación de los grafos y el entrenamiento de las redes, lo que produjo que ciertos puntos quedaran incompletos.

Si se analizan los objetivos específicos, se puede concluir que no todos fueron cumplidos en su totalidad. En el caso de de la generación de grafos, se diseñó, exploró y validó una estrategia que permite la representación de proteínas con estructura de grafos, variando la granularidad del modelo y la información de las aristas. Obteniendo específicamente, los cuatro tipos de grafos deseados: átomo-distancia, carbono alfa-distancia, centroide-distancia e interacción intermolecular. Sin embargo, los primeros no se utilizaron en la continuación del trabajo debido al costo computacional asociado a generarlos.

Los grafos generados poseen distintas propiedades dependiendo de su granularidad y tipo de interacción elegida. En cuanto al primer punto, los grafos de átomos son más pesados y poseen una mayor cantidad de nodos y aristas que los de carbono alfa y centroide, los que a su vez poseen el mismo número de nodos, pero no necesariamente el de aristas. Respecto al segundo punto los grafos de distancia son regularmente completos si se selecciona bien el rango de filtrado y los de interacciones generalmente no lo son.

A modo de proyección, solo se utilizaron dos posibles informaciones en las aristas, por lo que se recomienda considerar otras propiedades y/o interacciones, como ángulos de torsión, fuerzas de Van der Waals, puentes salinos, u otras cuantificaciones de las distancias distintas a la euclidiana, como son la semejanza de cosenos o *city block*. También, existen variables dentro de las representaciones generadas que se pueden variar como el rango de filtrado para los grafos de distancia. Además, se propone generar grafos que incluyan tanto distancia como interacciones intermoleculares, a modo de par ordenado.

Respecto a la detección de comunidades, se diseñó, implementó y validó una estrategia para identificar patrones en proteínas a partir de la detección de comunidades, probando el de-

sempeño de siete algoritmos distintos ya implementados en la librería *igraph* y estudiando la relación de los cambios de estabilidad con las variaciones de modularidad. Sin embargo, debido a la exclusión de los grados de átomo-distancia se considera inconcluso este objetivo.

Se consideraron dos parámetros de interés para esta tarea: la modularidad promedio asociada a cómo el algoritmo es capaz de separar la red en módulos bien definidos y la desviación estándar del mismo parámetro que se propone como un medidor de sensibilidad del algoritmo a las mutaciones. En caso de los grafos Carbono alfa-distancia Spinglass posee la mejor modularidad promedio y Label propagation la mayor desviación estándar para este medidor. A su vez, en los grafos de Centroide-distancia Spinglass vuele a ser tener el mejor desempeño en el primer medidor e Infomap en el segundo. Por último, en los grafos de Interacciones intermoleculares, son Multilevel y Label propagation los que destacan en los dos medidores, respectivamente.

Para esta los grafos de Interacciones intermoleculares, se propone extraer las comunidades más grandes y reenfocar el estudio en estas, eliminando así el ruido que generan las comunidades compuestas por un solo nodo.

En cuanto a la relación entre la variación de la estabilidad producida por las mutaciones y el consecuente cambio de modularidad en la detección de comunidades, no se pudo observar una correlación directa entre ambas variables.

En relación a la comparación de grafos para la identificación de cambios estructurales se diseñó e implementó una metodología basada en coeficientes de similitud que permite realizar dicha tarea por medio del uso de comunidades. Sin embargo, esta fue analizada para variantes puntuales, por lo que no pudo ser validada en su totalidad.

Para las mutantes que poseen mayores variaciones en las estabilidad de la proteína, se identifican alteraciones de nodos en las comunidades que se asocian con cambios epistáticos consistentes con la naturaleza de la mutación. A su vez, el mismo fenómeno ocurre en las mutantes que sufren cambio en la modularidad una mayor cantidad de veces. Principalmente, es Spinglass el algoritmo que presenta estas alteraciones para las proteínas analizadas. Existiendo una mayor cantidad de algoritmos que presentan este fenómeno para los grafos de centroide-distancia en comparación a los de carbono alfa-distancia.

Los grafos de Interacciones intermoleculares sólo presentan cambios para C141Y, a pesar de que los nodos involucrados en dos de las otras mutaciones establecen puentes de hidrógeno. Se vincula esto a la magnitud del cambio de la fuerza de la interacción, siendo C141Y donde ocurre el mayor.

En cuanto a la aplicación de grafos en redes neuronales convolucionales, se exploró y propuso un pipeline para aplicarlas en tareas de ingeniería de proteínas. Completando parcialmente el objetivo, siendo el principal problema el bajo desempeño obtenido en la red programada, prediciendo una sola clase el 100 % de las veces. Se identifica el no ajustar los hiperparámetros, el uso de ReLu como función de activación o la poca cantidad de datos usados como las posibles causas de esta situación. Sin embargo, el núcleo del objetivo se centra en la metodología más que en los resultados particulares.

Como proyección, utilizar para estas tareas otro tipo de representación además de los grafos carbono alfa-distancia, como lo son los grafos centroide-distancia y los de interacciones intermoleculares, permitiría comprender como se comportan y generar una comparación más detallada de cuando usar cada una.

Finalmente, se observa la versatilidad de los grafos como representación de proteínas, permitiendo incluir las interacciones entre elementos de manera sencilla y aportando diversas herramientas al estudio de estas macromoléculas como lo es la detección de comunidades o el uso de *graph convolutional networks*. Sin embargo, existe un alto costo computacional asociado a la generación y análisis de grafos de gran tamaño que se debe tener en cuenta al comenzar un estudio. Por otra parte, la gran utilidad de los grafos se puede extrapolar al estudio de otras moléculas y sus propiedades, por lo que el trabajo de la presente memoria corresponde a una base que se puede continuar y expandir.

Bibliografía

[1] Z. S. e. a. Lodish H, Berk A, "The molecules of life," in *Molecular Cell Biology* (W. H. Freeman, ed.), ch. 1.2, pp. 75 – 114, 4ta ed., 2000.

[2] A. J. Rosales, "Los retos actuales en la ingeniería de proteínas," *CIENCIA ergo sum*, vol. 26, pp. 1–11, Oct. 2019.

[3] A. R. Jamasb, P. Lió, and T. L. Blundell, "Graphein - a python library for geometric deep learning and network analysis on protein structures," July 2020.

[4] W. Gao, S. P. Mahajan, J. Sulam, and J. J. Gray, "Deep learning in protein structural modeling and design," *Patterns*, vol. 1, pp. 100–142, Dec. 2020.

[5] M. M. Cox and D. L. Nelson, "Aminoácidos, péptidos y proteínas," in *Lehninger Principios de Bioquímica* (Editorial Omega S.A., ed.), ch. 3, pp. 75 – 114, 6ta ed., 2014.

[6] E. Castaños. CienciadeLux. [En línea] https://cienciadelux.com/ [Consultado: 20 Julio de 2021].

[7] K.-C. Chou, "Prediction of tight turns and their types in proteins," *Analytical Biochemistry*, vol. 286, pp. 1–16, Nov. 2000.

[8] J. Sancho. Estructura de Macromoléculas. [En línea] http://jsancho.bifi.es/estructuramacromoleculas/ [Consultado: 20 Julio de 2021].

[9] L. Williams. MACROMOLECULAR STRUCTURE. Biopolymers in Three Dimensions. [En línea] https://williams.chemistry.gatech.edu/structure/index.html [Consultado: 20 Julio de 2021].

[10] T. J, Barnard, N. Dautin, P. Lukacik, H. D. Bernstein and Susan K Buchanan, "Autotransporter structure re- veals intra-barrel cleavage followed by conformational changes," Nature Structural & Molecular Biology, vol. 14, no. 12, pp. 1214–1220, 2007.

[11] Y. Kato, T. Muto, T. Tomura, H. Tsumura, H. Watarai, T. Mikayama, K. Ishizaka and R. Kuroki, "he crystal structure of human glycosylation-inhibiting factor is a trimeric barrel with three 6-stranded beta-sheets.," Proceedings of the National Academy of Sciences, vol. 93, no. 7, pp. 3007–3010, 1996.

[12] L. Gonzalez, D. N. Woolfson, and T. Alber, "Buried polar residues and structural specificity in the GCN4 leucine zipper," *Nature Structural Biology*, vol. 3, no. 12, pp. 1011–1018, 1996.

[13] H. M. Berman, "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[14] M. M. Cox and D. L. Nelson, "Estructura tridimensional de las proteínas," in *Lehninger*

Principios de Bioquímica (Editorial Omega S.A., ed.), ch. 4, pp. 115 – 156, 6ta ed., 2014.

[15] T. B. Acton, R. Xiao, S. Anderson, J. Aramini, W. A. Buchwald, C. Ciccosanti, K. Conover, J. Everett, K. Hamilton, Y. J. Huang, H. Janjua, G. Kornhaber, J. Lau, D. Y. Lee, G. Liu, M. Maglaqui, L. Ma, L. Mao, D. Patel, P. Rossi, S. Sahdev, R. Shastry, G. Swapna, Y. Tang, S. Tong, D. Wang, H. Wang, L. Zhao, and G. T. Montelione, "Preparation of protein samples for NMR structure, function, and small-molecule screening studies," in *Methods in Enzymology*, pp. 21–60, Elsevier, 2011.

[16] M. J. Howard, "Protein NMR spectroscopy," *Current Biology*, vol. 8, pp. R331–R333, May 1998.

[17] S. R. Carroni M, "Cryo electron microscopy to determine the structure of macromolecular complexes," *Methods*, vol. 95, pp. 78–85, 2016.

[18] S. Kaczanowski and P. Zielenkiewicz, "Why similar protein sequences encode similar three-dimensional structures?," *Theoretical Chemistry Accounts*, vol. 125, pp. 643–650, Oct. 2009.

[19] Y. Xu, Z. Liu, L. Cai, and D. Xu, "Protein structure prediction by protein threading," in *Computational Methods for Protein Structure Prediction and Modeling*, pp. 1–42, Springer New York, 2007.

[20] D. N. Rubingh and R. A. Grayling, "Protein engineering," in *BIOTECHNOLOGY - Volume III: Fundamentals in Biotechnology* (E. Publications, ed.), pp. 140 – 165, 2009.

[21] S. A. Marshall, G. A. Lazar, A. J. Chirino, and J. R. Desjarlais, "Rational design and engineering of therapeutic proteins," *Drug Discovery Today*, vol. 8, pp. 212–221, Mar. 2003.

[22] X. Xia, L. M. Longo, and M. Blaber, "Mutation choice to eliminate buried free cysteines in protein therapeutics," *Journal of Pharmaceutical Sciences*, vol. 104, pp. 566–576, Feb. 2015.

[23] D. S. Wilson and A. D. Keefe, "Random mutagenesis by pcr," *Current Protocols in Molecular Biology*, vol. 51, no. 1, pp. 8.3.1–8.3.9, 2000.

[24] R. E. Cobb, R. Chao, and H. Zhao, "Directed evolution: Past, present, and future," *AIChE Journal*, vol. 59, pp. 1432–1440, Jan. 2013.

[25] F. H. Arnold, "Directed evolution: Creating biocatalysts for the future," *Chemical Engineering Science*, vol. 51, pp. 5091–5102, Dec. 1996.

[26] J. C. Moore and F. H. Arnold, "Directed evolution of a para-nitrobenzyl esterase for aqueous-organic solvents," *Nature Biotechnology*, vol. 14, pp. 458–467, Apr. 1996.

[27] R. A. Chica, N. Doucet, and J. N. Pelletier, "Semi-rational approaches to engineering enzyme activity: combining the benefits of directed evolution and rational design," *Current Opinion in Biotechnology*, vol. 16, pp. 378–384, Aug. 2005.

[28] R. M. Siloto and R. J. Weselake, "Site saturation mutagenesis: Methods and applications in protein engineering," *Biocatalysis and Agricultural Biotechnology*, vol. 1, pp. 181–189, July 2012.

[29] E. Ordu and N. Gul, "Protein engineering applications on industrially important enzymes: Candida methylica FDH as a case study," in *Protein Engineering*, InTech, Feb.

2012.

[30] H. Jochens and U. T. Bornscheuer, "Natural diversity to guide focused directed evolution," *ChemBioChem*, vol. 11, pp. 1861–1866, Aug. 2010.

[31] J. Bienkowska, "Computational characterization of proteins," *Expert Review of Proteomics*, vol. 2, no. 1, pp. 129–138, 2005.

[32] A. M. Lesk, "Bioinformatics." Encyclopedia Britannica [En línea] https://www.britannica.com/science/bioinformatics [Consultado: 25 de Julio, 2021].

[33] IBM, "¿qué es machine learning?." [En línea] https://www.ibm.com/cl-es/analytics/machine-learning [Consultado: 25 de Julio, 2021].

[34] Medina-Ortiz, D., Contreras, S., Amado-Hinojosa, J., Torres-Almonacid, J., Asenjo, J. A., Navarrete, M., and Olivera-Nappa, Á. "Combination of digital signal processing and assembled predictive models facilitates the rational design of proteins," 2020. Under revision in Peer.

[35] P.-E. D. Koutrouli M, Karatzas E and P. GA, "A guide to conquer the biological network era using graph theory," *Front. Bioeng. Biotechnol.*, vol. 8, no. 34, 2020.

[36] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, pp. 8577–8582, May 2006.

[37] S. Rahiminejad, M. R. Maurya, and S. Subramaniam, "Topological and functional comparison of community detection algorithms in biological networks," *BMC Bioinformatics*, vol. 20, Apr. 2019.

[38] A. Christensen and H. Golino, "Estimating factors with psychometric networks: A monte carlo simulation comparing community detection algorithms," 2020.

[39] Z. Yang, R. Algesheimer, and C. J. Tessone, "A comparative analysis of community detection algorithms on artificial networks," *Scientific Reports*, vol. 6, no. 1, 2016.

[40] M. E. J. N. A. Clauset and C. Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 6, 2004.

[41] M. Rosvall and C. T. Bergstrom, "An information-theoretic framework for resolving community structure in complex networks," *Proceedings of the National Academy of Sciences*, vol. 104, no. 18, pp. 7327–7331, 2007.

[42] R. A. U. N. Raghavan and S. Kumara, "Near linear time algorithm to detect community structures in large-scale networks," *Physical Review E*, vol. 76, no. 3, 2007.

[43] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, no. 3, 2006.

[44] R. L. V. D. Blondel, J.-L. Guillaume and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[45] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," *Physical Review E*, vol. 74, no. 1, 2006.

[46] P. Pons and M. Latapy, "Computing communities in large networks using random walks," in *Computer and Information Sciences - ISCIS 2005*, pp. 284–293, Springer Berlin Hei-

delberg, 2005.

[47] IBM, "Neural networks." [En línea] https://www.ibm.com/cloud/learn/neural-networks [Consultado: 25 de Julio, 2021].

[48] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, "The graph neural network model," *IEEE Transactions on Neural Networks*, vol. 20, pp. 61–80, Jan. 2009.

[49] S. Hong, "An introduction to graph neural network(gnn) for analysing structured data." [En línea] https://towardsdatascience.com/an-introduction-to-graph-neural-network-gnn-for-analysing-structured-data-afce79f4cfdc [Consultado: 25 de Julio, 2021].

[50] E. B. M. Grandini and G. Visani, "Metrics for multi-class classification: an overview," 2020.

[51] RCSB PDB. Search API Documentation. [En línea] https://search.rcsb.org/#search-api [Consultado: 11 de Agosto de 2021].

[52] SCOP. Structural Classification of Proteins. [En línea] https://scop.mrc-lmb.cam.ac.uk/ [Consultado: 01 de Junio de 2022].

[53] Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. Enzyme Nomenclature. [En línea] https://iubmb.qmul.ac.uk/enzyme/ [Consultado: 01 de Junio de 2022].

[54] UniprotKB. Cellular tumor antigen p53 – Homo sapiens. [En línea] ttps://www.uniprot.org/uniprotkb/P04637/entry [Consultado: 15 de Marzo de 2022].

[55] AlphaFold Protein Structure Database. Cellular tumor antigen p53. [En línea] https://alphafold.ebi.ac.uk/entry/P04637 [Consultado: 15 de Marzo de 2022].

[56] C. L. Worth, R. Preissner, and T. L. Blundell, "SDM–a server for predicting effects of mutations on protein stability and malfunction," *Nucleic Acids Research*, vol. 39, no. suppl, pp. W215–W222, 2011.

[57] RCSB PDB. Data API Documentation. [En línea] https://data.rcsb.org/#data-api [Consultado: 15 Mayo de 2022].

[58] M. Martínez-Martínez, C. Coscolín, G. Santiago, J. Chow, P. J. Stogios, R. Bargiela, C. Gertler, J. Navarro-Fernández, A. Bollinger, S. Thies, C. Méndez-García, A. Popovic, G. Brown, T. N. Chernikova, A. García-Moyano, G. E. K. Bjerga, P. Pérez-García, T. Hai, M. V. Del Pozo, R. Stokke, I. H. Steen, H. Cui, X. Xu, B. P. Nocek, M. Alcaide, M. Distaso, V. Mesa, A. I. Peláez, J. Sánchez, P. C. F. Buchholz, J. Pleiss, A. Fernández-Guerra, F. O. Glöckner, O. V. Golyshina, M. M. Yakimov, A. Savchenko, K.-E. Jaeger, A. F. Yakunin, W. R. Streit, P. N. Golyshin, V. Guallar, M. Ferrer and The INMARE Consortium, "Determinants and prediction of esterase substrate promiscuity patterns," ACS Chemical Biology, vol. 13, no. 1, pp. 225–234, 2017.

[59] G. Vriend, "The what if web interface." [En línea] https://swift.cmbi.umcn.nl/servers/html/index.html [Consultado: 25 de Julio, 2021].

[60] M. Muraki, K. Harata, N. Sugita, and K. Sato, "Origin of carbohydrate recognition specificity of human lysozyme revealed by affinity labeling," Biochemistry, vol. 35, no. 42, pp. 13562–13567, 1996.

[61] C. Guevara, “Aplicaciones de estructuras de grafos y aprendizaje profundo a sistemas de clasificación de interacción antígeno-anticuerpo,” 2022.

Anexos

Anexo A

Nomenclatura de mutaciones

Durante el desarrollo de la memoria se utiliza la nomenclatura XPY para hacer referencia a las mutaciones de sustitución, donde X corresponde al aminoácido previo a la mutación, P a la posición de este e Y al aminoácido por el que es sustituido.

Anexo B

Predicción del efecto de mutaciones puntuales sobre la proteína p53 humana

En las Tablas B.1 y B.2 se presentan diversos parámetros de la proteína p53 antes y después de mutaciones puntuales. En estas WT indica la variante nativa o wildtype y MT a la mutadas, RSA corresponde a la accesibilidad del solvente al residuo mutado, Depth a la profundidad del residuo, OSP al empaquetamiento de la superficie ocluida del residuo y ddG presenta la variación de la energía libre de Gibbs de plegamiento entre la especie nativa y la mutada, lo que implica que si esta última tiene una mayor energía libre el valor de la variación será negativo, indicando que la estabilidad se redujo. Además, se indican parámetros asociados con la formación de puentes de Hidrogeno entre la cadena lateral de un residuo aminoacídico y otra cadena lateral (SS), una amina de la cadena principal (SN) o un grupo carbonil de la cadena principal (SO), siendo 1 si estos existen y 0 si no.

Tabla B.1: Efecto de las mutaciones artificiales en la proteína p53. Elaboración propia a partir de datos obtenidos via SDM [56]

Mutation	**WT DEPTH** (Å)	**WT OSP**	**MT DEPTH** (Å)	**MT OSP**	**Predicted ddG**	**Outcome**
G244D	3,55	0,233	3,34	0,165	-3,77	Reduced stability
R282G	5,12	0,555	6,32	0,547	-3,46	Reduced stability
G361E	3,33	0,117	3,24	0,049	-3,34	Reduced stability
R196P	5,98	0,436	6,59	0,498	-3,13	Reduced stability
L252P	4,65	0,449	5,04	0,454	-2,84	Reduced stability
G244V	3,55	0,233	3,29	0,212	-2,64	Reduced stability
L257Q	9,07	0,574	9,34	0,518	-2,45	Reduced stability
V197M	7,26	0,424	7,37	0,472	-2,30	Reduced stability
L344P	3,40	0,263	3,48	0,314	-2,23	Reduced stability
A138P	3,29	0,298	3,75	0,376	-1,92	Reduced stability
R306P	3,24	0,044	3,20	0,104	-1,90	Reduced stability
M133T	9,55	0,545	9,84	0,505	-1,89	Reduced stability

Tabla B.1: Efecto de las mutaciones artificiales en la proteína p53 (cont.). Elaboración propia a partir de datos obtenidos via SDM [56]

Mutation	WT DEPTH (Å)	WT OSP	MT DEPTH (Å)	MT OSP	Predicted ddG	Outcome
R213P	5,77	0,469	5,76	0,457	-1,89	Reduced stability
R175G	7,84	0,459	6,90	0,309	-1,88	Reduced stability
G105C	3,84	0,355	6,00	0,479	-1,84	Reduced stability
C238G	7,32	0,608	7,48	0,527	-1,83	Reduced stability
R213Q	5,77	0,469	5,25	0,432	-1,79	Reduced stability
I251M	10,60	0,464	11,07	0,480	-1,77	Reduced stability
C238S	7,32	0,608	7,96	0,571	-1,61	Reduced stability
R267Q	4,68	0,484	4,39	0,499	-1,55	Reduced stability
G245D	4,80	0,415	6,76	0,544	-1,43	Reduced stability
M133R	9,55	0,545	9,64	0,557	-1,42	Reduced stability
R248S	3,29	0,123	3,17	0,178	-1,39	Reduced stability
L265P	4,49	0,406	4,01	0,392	-1,38	Reduced stability
Y163C	7,14	0,538	8,07	0,411	-1,38	Reduced stability
G245S	4,80	0,415	6,29	0,500	-1,35	Reduced stability
Y234C	10,48	0,499	9,82	0,419	-1,34	Reduced stability
Y236C	11,48	0,518	10,33	0,420	-1,34	Reduced stability
Y220C	5,93	0,412	5,93	0,311	-1,29	Reduced stability
V272L	10,72	0,543	11,19	0,582	-1,25	Reduced stability
P151S	6,16	0,509	5,97	0,432	-1,21	Reduced stability
V172F	6,66	0,435	7,75	0,516	-1,18	Reduced stability
A138S	3,29	0,298	3,40	0,287	-1,16	Reduced stability
C275Y	5,31	0,473	4,18	0,280	-1,16	Reduced stability
V173M	9,59	0,460	9,80	0,519	-1,16	Reduced stability
C141Y	7,26	0,549	5,91	0,542	-1,14	Reduced stability
P359D	3,16	0,121	3,23	0,077	-1,08	Reduced stability
R337H	3,47	0,303	3,53	0,303	-1,02	Reduced stability
R175H	7,84	0,459	8,25	0,466	-0,96	Reduced stability
H233D	3,77	0,335	3,84	0,355	-0,90	Reduced stability
R181C	3,23	0,126	3,20	0,149	-0,88	Reduced stability
E180K	3,80	0,318	3,97	0,373	-0,86	Reduced stability
R181H	3,23	0,126	3,30	0,129	-0,83	Reduced stability
R248Q	3,29	0,123	3,38	0,133	-0,74	Reduced stability
R283C	3,51	0,284	3,30	0,310	-0,71	Reduced stability
R337C	3,47	0,303	3,51	0,391	-0,71	Reduced stability
G245C	4,80	0,415	6,30	0,532	-0,67	Reduced stability
N235S	4,79	0,461	4,52	0,446	-0,66	Reduced stability
R333KR335RK337K	3,46	0,298	3,62	0,328	-0,63	Reduced stability
P152L	3,55	0,269	3,25	0,167	-0,58	Reduced stability
P82L	3,14	0,085	3,30	0,057	-0,58	Reduced stability
E285Q	3,62	0,425	3,94	0,359	-0,50	Reduced stability
G245V	4,80	0,415	6,56	0,589	-0,50	Reduced stability
Q167K	3,47	0,185	3,12	0,097	-0,44	Reduced stability
H193R	6,65	0,516	6,40	0,538	-0,43	Reduced stability
E258K	4,25	0,484	4,72	0,468	-0,41	Reduced stability
R248W	3,29	0,123	3,40	0,176	-0,37	Reduced stability
L22QW23S	3,41	0,183	3,11	0,283	-0,36	Reduced stability

Tabla B.1: Efecto de las mutaciones artificiales en la proteína p53 (cont.). Elaboración propia a partir de datos obtenidos via SDM [56]

Mutation	WT DEPTH (Å)	WT OSP	MT DEPTH (Å)	MT OSP	Predicted ddG	Outcome
M237I	5,78	0,474	5,84	0,452	-0,33	Reduced stability
S37D	3,08	0,086	3,21	0,059	-0,31	Reduced stability
L22Q	3,45	0,233	3,53	0,313	-0,29	Reduced stability
M246V	9,30	0,523	8,09	0,523	-0,28	Reduced stability
T155N	6,16	0,521	5,97	0,515	-0,27	Reduced stability
R333K	3,23	0,088	3,17	0,098	-0,26	Reduced stability
R333KR335K	3,23	0,152	3,12	0,125	-0,25	Reduced stability
R273C	4,61	0,363	4,64	0,417	-0,21	Reduced stability
K381Q	3,22	0,023	3,18	0,03	-0,16	Reduced stability
R290H	3,44	0,265	3,54	0,237	-0,13	Reduced stability
T387A	3,21	0,083	2,94	0,071	-0,12	Reduced stability
T55A	3,32	0,117	3,09	0,042	-0,12	Reduced stability
E286A	3,93	0,413	3,55	0,344	-0,10	Reduced stability
R174G	4,61	0,312	5,42	0,245	-0,02	Reduced stability
R158H	4,18	0,421	4,28	0,432	0,00	Reduced stability
R273H	4,61	0,363	4,57	0,367	0,00	Reduced stability
S269E	4,64	0,418	4,25	0,399	0,02	Increased stability
K370R	3,16	0,072	3,28	0,084	0,05	Increased stability
K381R	3,22	0,023	3,25	0,027	0,05	Increased stability
K373R	3,13	0,037	3,26	0,035	0,06	Increased stability
T18A	3,73	0,225	3,33	0,27	0,06	Increased stability
L383A	3,27	0,089	3,02	0,069	0,09	Increased stability
F385A	3,27	0,071	3,13	0,08	0,13	Increased stability
K386A	3,18	0,078	2,97	0,071	0,15	Increased stability
S241T	3,33	0,258	3,50	0,253	0,15	Increased stability
K292I	3,40	0,165	3,55	0,222	0,19	Increased stability
R158G	4,18	0,421	6,52	0,396	0,19	Increased stability
R273G	4,61	0,363	5,74	0,369	0,19	Increased stability
D281N	4,11	0,475	4,28	0,485	0,20	Increased stability
R282W	5,12	0,555	4,81	0,587	0,22	Increased stability
S227T	3,50	0,280	3,53	0,324	0,22	Increased stability
H179Y	4,46	0,441	4,45	0,407	0,23	Increased stability
K24R	3,18	0,151	3,23	0,105	0,26	Increased stability
K291RK292R	3,39	0,166	3,42	0,158	0,26	Increased stability
K382R	3,16	0,038	3,25	0,035	0,27	Increased stability
K291R	4,14	0,286	4,09	0,189	0,30	Increased stability
S183E	3,00	0,09	3,16	0,062	0,30	Increased stability
G325V	3,81	0,204	3,99	0,278	0,31	Increased stability
R273L	4,61	0,363	4,58	0,429	0,31	Increased stability
K132E	5,39	0,575	5,30	0,544	0,36	Increased stability
R290L	3,44	0,265	3,59	0,262	0,39	Increased stability
S183A	3,00	0,09	2,88	0,097	0,40	Increased stability

Tabla B.1: Efecto de las mutaciones artificiales en la proteína p53 (cont.). Elaboración propia a partir de datos obtenidos via SDM [56]

Mutation	**WT DEPTH** (Å)	**WT OSP**	**MT DEPTH** (Å)	**MT OSP**	**Predicted ddG**	**Outcome**
S20D	3,25	0,246	3,32	0,257	0,40	Increased stability
P278S	6,80	0,654	7,34	0,568	0,41	Increased stability
R156H	3,94	0,351	4,09	0,459	0,43	Increased stability
S241F	3,33	0,258	3,36	0,227	0,52	Increased stability
K372R	3,17	0,034	3,26	0,077	0,57	Increased stability
E388A	3,20	0,081	2,96	0,07	0,63	Increased stability
R175L	7,84	0,459	8,46	0,461	0,66	Increased stability
K319A	3,17	0,053	3,1	0,092	0,68	Increased stability
K320A	3,11	0,056	2,9	0,093	0,68	Increased stability
K321A	3,08	0,072	3,05	0,138	0,68	Increased stability
S15A	3,38	0,26	3,17	0,251	0,68	Increased stability
S46A	3,17	0,159	3,11	0,157	0,68	Increased stability
Q144L	4,44	0,444	4,45	0,469	0,72	Increased stability
S269A	4,64	0,418	5,06	0,407	0,80	Increased stability
P278T	6,80	0,654	6,99	0,641	1,10	Increased stability
S362A	3,24	0,079	3,11	0,085	1,18	Increased stability
S20A	3,25	0,246	3,25	0,268	1,58	Increased stability
P309S	3,25	0,088	3,15	0,057	1,66	Increased stability
T284E	3,71	0,331	3,66	0,326	1,98	Increased stability
P278L	6,80	0,654	7,14	0,699	3,40	Increased stability

Tabla B.2: Efecto de las mutaciones artificiales en la proteína p53, segunda parte. Elaboración propia a partir de datos obtenidos via SDM [56]

Mutation	**WT SSE**	**WT RSA** (%)	**WT SS**	**WT SN**	**WT SO**	**MT SSE**	**MT RSA** (%)	**MT SS**	**MT SN**	**MT SO**
G244D	g	98,7	0	0	0	g	86,0	0	0	0
R282G	H	13,7	1	0	1	H	8,6	0	0	0
G361E	g	101,4	0	0	0	g	103,3	0	0	0
R196P	E	5,8	0	0	1	b	13,3	0	0	0
L252P	E	11,9	0	0	0	E	8,8	0	0	0
G244V	g	98,7	0	0	0	g	80,6	0	0	0
L257Q	E	0,1	0	0	0	E	0,6	0	0	0
V197M	E	2,4	0	0	0	E	2,6	0	0	0
L344P	H	69,6	0	0	0	H	58,9	0	0	0
A138P	g	42,5	0	0	0	g	29,9	0	0	0
R306P	b	92,9	0	0	0	b	71,0	0	0	0
M133T	E	1,0	1	0	0	E	6,3	0	0	1

Tabla B.2: Efecto de las mutaciones artificiales en la proteína p53, segunda parte (cont.). Elaboración propia a partir de datos obtenidos via SDM [56]

Mutation	**WT SSE**	**WT RSA (%)**	**WT SS**	**WT SN**	**WT SO**	**MT SSE**	**MT RSA (%)**	**MT SS**	**MT SN**	**MT SO**
R213P	p	2,8	0	0	1	p	9,5	0	0	0
R175G	p	2,2	0	0	1	p	18,9	0	0	0
G105C	e	6,8	0	0	0	e	0,9	0	1	1
C238G	b	0,0	1	1	0	b	5,6	0	0	0
R213Q	p	2,8	0	0	1	b	13,0	1	0	0
I251M	E	0,1	0	0	0	E	0,2	0	0	0
C238S	b	0,0	1	1	0	b	0,0	0	0	0
R267Q	E	16,5	0	0	1	E	9,8	0	0	1
G245D	b	8,6	0	0	0	b	6,4	1	1	1
M133R	E	1,0	1	0	0	E	0,8	1	0	0
R248S	g	85,1	0	0	0	l	79,3	0	0	0
L265P	E	23,1	0	0	0	a	18,6	0	0	0
Y163C	E	1,1	1	0	1	E	4,0	0	0	1
G245S	b	8,6	0	0	0	b	5,2	1	1	1
Y234C	E	0,9	0	0	0	E	6,9	0	0	0
Y236C	E	0,0	1	0	0	E	4,5	0	0	0
Y220C	p	5,6	0	0	0	p	14,8	0	0	0
V272L	E	0,0	0	0	0	E	0,0	0	0	0
P151S	p	0,0	0	0	0	p	0,0	1	0	0
V172F	p	1,4	0	0	0	b	1,5	0	0	0
A138S	g	42,5	0	0	0	g	48,9	1	0	0
C275Y	b	12,0	1	0	1	b	33,8	0	0	0
V173M	p	0,6	0	0	0	p	2,8	1	1	0
C141Y	E	0,0	1	0	1	E	1,8	0	1	1
P359D	p	88,5	0	0	0	p	111,9	0	0	0
R337H	H	58,7	0	0	1	H	65,2	0	0	0
R175H	p	2,2	0	0	1	p	4,0	0	0	1
H233D	E	48,2	0	0	0	E	36,5	0	0	0
R181C	a	86,3	0	0	1	H	93,5	0	0	1
E180K	H	40,5	0	0	0	H	25,0	0	0	0
R181H	a	86,3	0	0	1	H	93,1	0	0	0
R248Q	g	85,1	0	0	0	l	84,2	0	0	0
R283C	H	49,4	0	0	1	H	61,4	0	0	1
R337C	H	58,7	0	0	1	H	53,0	0	0	1
G245C	b	8,6	0	0	0	b	5,4	1	1	1
N235S	E	14,7	1	0	1	E	14,1	0	0	0
R333KR335KR337K	H	61,3	0	0	1	H	57,6	0	0	0
P152L	p	49,9	0	0	0	p	81,4	0	0	0
P82L	p	91,3	0	0	0	p	106,1	0	0	0
E285Q	H	31,0	1	0	0	H	45,0	0	0	0
G245V	b	8,6	0	0	0	b	2,9	0	0	0
Q167K	H	78,2	0	0	0	H	95,3	0	0	0
H193R	p	1,6	1	0	0	p	1,9	1	0	1
E258K	E	8,7	1	0	0	E	19,3	0	0	1
R248W	g	85,1	0	0	0	g	65,9	0	0	0
L22QW23S	H	75,0	0	0	0	H	61,0	0	0	1

Tabla B.2: Efecto de las mutaciones artificiales en la proteína p53, segunda parte (cont.). Elaboración propia a partir de datos obtenidos via SDM [56]

Mutation	**WT SSE**	**WT RSA (%)**	**WT SS**	**WT SN**	**WT SO**	**MT SSE**	**MT RSA (%)**	**MT SS**	**MT SN**	**MT SO**
M237I	a	11,1	0	0	0	a	8,4	0	0	0
S37D	p	109,2	0	0	0	b	110,0	0	0	0
L22Q	H	63,8	0	0	0	H	45,2	0	0	0
M246V	a	0,0	0	0	0	a	3,0	0	0	0
T155N	p	1,4	0	0	1	p	1,4	0	0	1
R333K	p	96,7	0	0	0	b	95,2	0	0	0
R333KR335K	H	89,1	0	0	0	H	104,2	0	0	0
R273C	E	39,8	1	0	0	E	21,8	0	0	0
K381Q	b	99,5	0	0	0	b	110,7	0	0	0
R290H	H	62,3	1	0	0	H	70,0	1	0	0
T387A	b	94,2	0	0	0	b	96,4	0	0	0
T55A	b	74,8	0	0	0	b	80,6	0	0	0
E286A	H	28,6	1	0	0	H	43,6	0	0	0
R174G	b	22,7	1	0	1	b	38,4	0	0	0
R158H	E	18,8	1	0	0	E	18,1	1	0	0
R273H	E	39,8	1	0	0	E	27,3	1	0	0
S269E	E	8,7	0	0	0	E	24,6	0	0	1
K370R	b	90,0	0	0	0	b	87,4	0	0	0
K381R	b	99,5	0	0	0	b	101,6	0	0	0
K373R	p	106,5	0	0	0	b	91,0	0	0	0
T18A	a	65,1	1	1	1	a	53,0	0	0	0
L383A	p	87,9	0	0	0	p	97,8	0	0	0
F385A	p	92,7	0	0	0	b	90,2	0	0	0
K386A	b	86,5	0	0	0	b	98,7	0	0	0
S241T	a	64,3	0	0	0	a	55,5	0	0	0
K292I	H	69,6	0	0	0	H	65,4	0	0	0
R158G	E	18,8	1	0	0	E	9,1	0	0	0
R273G	E	39,8	1	0	0	E	33,5	0	0	0
D281N	H	9,7	1	0	0	H	13,4	1	0	1
R282W	H	13,7	1	0	1	H	14,1	1	0	1
S227T	b	44,5	1	1	1	b	42,8	1	0	0
H179Y	H	20,1	1	0	0	H	26,4	0	0	0
K24R	H	82,2	0	0	0	H	91,2	0	0	0
K291RK292R	H	70,0	0	0	0	H	71,5	0	0	0
K382R	p	103,5	0	0	0	p	97,5	0	0	0
K291R	H	30,3	0	0	1	H	32,6	1	0	0
S183E	a	107,1	0	0	0	a	111,8	0	0	0
G325V	a	55,3	0	0	0	a	35,5	0	0	0
R273L	E	39,8	1	0	0	E	22,1	0	0	0
K132E	E	7,8	1	0	0	E	4,9	1	0	0
R290L	H	62,3	1	0	0	H	63,6	0	0	0
S183A	a	107,1	0	0	0	a	107,8	0	0	0

Tabla B.2: Efecto de las mutaciones artificiales en la proteína p53, segunda parte (cont.). Elaboración propia a partir de datos obtenidos via SDM [56]

Mutation	**WT SSE**	**WT RSA (%)**	**WT SS**	**WT SN**	**WT SO**	**MT SSE**	**MT RSA (%)**	**MT SS**	**MT SN**	**MT SO**
S20D	H	81,3	0	0	0	H	75,8	1	0	0
P278S	H	0,1	0	0	0	H	0,0	0	0	1
R156H	E	25,2	1	0	0	E	22,5	1	0	1
S241F	a	64,3	0	0	0	a	65,7	0	0	0
K372R	b	96,6	0	0	0	b	86,6	0	0	1
E388A	p	87,0	0	0	0	b	91,6	0	0	0
R175L	p	2,2	0	0	1	p	3,1	0	0	0
K319A	p	106,3	0	0	0	p	107,2	0	0	0
K320A	p	107,6	0	0	0	p	101,5	0	0	0
K321A	p	99,6	0	0	0	p	88,2	0	0	0
S15A	p	68,2	1	0	0	p	65,7	0	0	0
S46A	p	79,0	1	0	0	p	76,8	0	0	0
Q144L	E	20,3	1	0	0	E	13,3	0	0	0
S269A	E	8,7	0	0	0	E	6,9	0	0	0
P278T	H	0,1	0	0	0	H	0,0	0	0	1
S362A	t	81,9	0	0	0	t	74,1	0	0	0
S20A	H	81,3	0	0	0	H	77,9	0	0	0
P309S	b	90,7	0	0	0	b	101,3	0	0	0
T284E	H	47,4	0	0	1	H	40,8	0	0	0
P278L	H	0,1	0	0	0	H	0,1	0	0	0

Anexo C

Parámetros de detección de comunidades para cada variante

A continuación, en las Tablas C.1, C.2, C.3, C.4, C.5, C.6 y C.7 se presenta la detección de comunidades para los grafos de carbono alfa-distancia de las variantes de p53.

Tabla C.1: Parámetros de detección de comunidades con el algoritmo Fast greedy para grafos de Carbono alfa-distancia.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,4549	4	Wild
A138P	0,4547	4	Natural
A138S	0,4549	4	Natural
C141Y	0,4558	4	Natural
C238G	0,4549	4	Natural
C238S	0,4549	4	Natural
C275Y	0,4549	4	Natural
D281N	0,4549	4	Natural
E180K	0,4549	4	Natural
E258K	0,4547	4	Natural
E285Q	0,4550	4	Natural
E286A	0,4549	4	Natural
G105C	0,4551	4	Natural
G244D	0,4549	4	Natural
G244V	0,4549	4	Natural
G245C	0,4552	5	Natural
G245D	0,4552	5	Natural
G245S	0,4549	4	Natural
G245V	0,4549	4	Natural
G325V	0,4550	4	Natural
H179Y	0,4549	4	Natural
H193R	0,4549	4	Natural
H233D	0,4549	4	Natural

Tabla C.1: Parámetros de detección de comunidades con el algoritmo Fast greedy para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
I251M	0,4549	4	Natural
K132E	0,4546	4	Natural
K292I	0,4549	4	Natural
L252P	0,4547	4	Natural
L257Q	0,4549	4	Natural
L265P	0,4547	4	Natural
L344P	0,4550	4	Natural
M133R	0,4548	4	Natural
M133T	0,4549	4	Natural
M237I	0,4549	4	Natural
M246V	0,4549	4	Natural
N235S	0,4549	4	Natural
P151S	0,4549	4	Natural
P152L	0,4549	4	Natural
P278L	0,4549	4	Natural
P278S	0,4549	4	Natural
P278T	0,4551	4	Natural
P309S	0,4549	4	Natural
P82L	0,4549	4	Natural
Q144L	0,4549	4	Natural
Q167K	0,4549	4	Natural
R156H	0,4548	4	Natural
R158G	0,4549	4	Natural
R158H	0,4549	4	Natural
R174G	0,4555	5	Natural
R175G	0,4549	4	Natural
R175H	0,4549	4	Natural
R175L	0,4549	4	Natural
R181C	0,4549	4	Natural
R181H	0,4549	4	Natural
R196P	0,4549	4	Natural
R213P	0,4548	4	Natural
R213Q	0,4549	4	Natural
R248Q	0,4549	4	Natural
R248W	0,4551	4	Natural
R267Q	0,4550	4	Natural
R273C	0,4549	4	Natural
R273G	0,4549	4	Natural
R273H	0,4549	4	Natural

Tabla C.1: Parámetros de detección de comunidades con el algoritmo Fast greedy para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
R273L	0,4549	4	Natural
R282G	0,4566	4	Natural
R282W	0,4549	4	Natural
R283C	0,4548	4	Natural
R290H	0,4549	4	Natural
R290L	0,4549	4	Natural
R306P	0,4549	4	Natural
R337C	0,4549	4	Natural
R337H	0,4549	4	Natural
S227T	0,4549	4	Natural
S241F	0,4549	4	Natural
S241T	0,4549	4	Natural
T155N	0,4549	4	Natural
V172F	0,4549	4	Natural
V173M	0,4549	4	Natural
V197M	0,4549	4	Natural
V272L	0,4549	4	Natural
Y163C	0,4549	4	Natural
Y220C	0,4549	4	Natural
Y234C	0,4549	4	Natural
Y236C	0,4549	4	Natural
E388A	0,4549	4	Artificial
F385A	0,4549	4	Artificial
G361E	0,4548	4	Artificial
K24R	0,4549	4	Artificial
K291RK292R	0,4549	4	Artificial
K319A	0,4549	4	Artificial
K320A	0,4549	4	Artificial
K321A	0,4549	4	Artificial
K370R	0,4549	4	Artificial
K372R	0,4549	4	Artificial
K373R	0,4549	4	Artificial
K381Q	0,4549	4	Artificial
K381R	0,4549	4	Artificial
K382R	0,4549	4	Artificial
K386A	0,4549	4	Artificial
L22QW23S	0,4549	4	Artificial
L383A	0,4549	4	Artificial
P359D	0,4550	4	Artificial
R248S	0,4549	4	Artificial

Tabla C.1: Parámetros de detección de comunidades con el algoritmo Fast greedy para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
R333KR335KR337K	0,4549	4	Artificial
S15A	0,4549	4	Artificial
S183A	0,4549	4	Artificial
S183E	0,4549	4	Artificial
S20A	0,4549	4	Artificial
S20D	0,4549	4	Artificial
S269A	0,4549	4	Artificial
S269E	0,4549	4	Artificial
S362A	0,4549	4	Artificial
S37D	0,4549	4	Artificial
S46A	0,4549	4	Artificial
T18A	0,4549	4	Artificial
T284E	0,4551	4	Artificial
T387A	0,4549	4	Artificial
T55A	0,4549	4	Artificial

Tabla C.2: Parámetros de detección de comunidades con el algoritmo Infomap para grafos de Carbono alfa-distancia.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,4982	12	Wild
A138P	0,4980	12	Natural
A138S	0,4982	12	Natural
C141Y	0,4988	12	Natural
C238G	0,4982	12	Natural
C238S	0,4982	12	Natural
C275Y	0,4982	12	Natural
D281N	0,4982	12	Natural
E180K	0,4982	12	Natural
E258K	0,4982	12	Natural
E285Q	0,4982	12	Natural
E286A	0,4982	12	Natural
G105C	0,4982	12	Natural
G244D	0,4982	12	Natural
G244V	0,4982	12	Natural
G245C	0,4982	12	Natural
G245D	0,4982	12	Natural
G245S	0,4982	12	Natural

Tabla C.2: Parámetros de detección de comunidades con el algoritmo Infomap para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
G245V	0,4982	12	Natural
G325V	0,4983	12	Natural
H179Y	0,4982	12	Natural
H193R	0,4980	12	Natural
H233D	0,4982	12	Natural
I251M	0,4982	12	Natural
K132E	0,4981	12	Natural
K292I	0,4981	12	Natural
L252P	0,4980	12	Natural
L257Q	0,4982	12	Natural
L265P	0,4981	12	Natural
L344P	0,4983	12	Natural
M133R	0,4983	12	Natural
M133T	0,4983	12	Natural
M237I	0,4982	12	Natural
M246V	0,4982	12	Natural
N235S	0,4982	12	Natural
P151S	0,4982	12	Natural
P152L	0,4982	12	Natural
P278L	0,4982	12	Natural
P278S	0,4982	12	Natural
P278T	0,4984	12	Natural
P309S	0,4982	12	Natural
P82L	0,4982	12	Natural
Q144L	0,4982	12	Natural
Q167K	0,4982	12	Natural
R156H	0,4981	12	Natural
R158G	0,4982	12	Natural
R158H	0,4982	12	Natural
R174G	0,4983	12	Natural
R175G	0,4982	12	Natural
R175H	0,4982	12	Natural
R175L	0,4982	12	Natural
R181C	0,4982	12	Natural
R181H	0,4982	12	Natural
R196P	0,4982	12	Natural
R213P	0,4982	12	Natural
R213Q	0,4982	12	Natural
R248Q	0,4982	12	Natural
R248W	0,4983	12	Natural

Tabla C.2: Parámetros de detección de comunidades con el algoritmo Infomap para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
R267Q	0,4982	12	Natural
R273C	0,4982	12	Natural
R273G	0,4982	12	Natural
R273H	0,4982	12	Natural
R273L	0,4982	12	Natural
R282G	0,4982	12	Natural
R282W	0,4982	12	Natural
R283C	0,4983	12	Natural
R290H	0,4982	12	Natural
R290L	0,4982	12	Natural
R306P	0,4981	12	Natural
R337C	0,4982	12	Natural
R337H	0,4982	12	Natural
S227T	0,4982	12	Natural
S241F	0,4982	12	Natural
S241T	0,4982	12	Natural
T155N	0,4982	12	Natural
V172F	0,4982	12	Natural
V173M	0,4982	12	Natural
V197M	0,4981	12	Natural
V272L	0,4982	12	Natural
Y163C	0,4982	12	Natural
Y220C	0,4982	12	Natural
Y234C	0,4982	12	Natural
Y236C	0,4982	12	Natural
E388A	0,4982	12	Artificial
F385A	0,4982	12	Artificial
G361E	0,4983	12	Artificial
K24R	0,4982	12	Artificial
K291RK292R	0,4981	12	Artificial
K319A	0,4982	12	Artificial
K320A	0,4982	12	Artificial
K321A	0,4982	12	Artificial
K370R	0,4982	12	Artificial
K372R	0,4982	12	Artificial
K373R	0,4982	12	Artificial
K381Q	0,4982	12	Artificial
K381R	0,4982	12	Artificial
K382R	0,4982	12	Artificial
K386A	0,4982	12	Artificial

Tabla C.2: Parámetros de detección de comunidades con el algoritmo Infomap para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
L22QW23S	0,4982	12	Artificial
L383A	0,4982	12	Artificial
P359D	0,4981	12	Artificial
R248S	0,4982	12	Artificial
R333KR335KR337K	0,4982	12	Artificial
S15A	0,4982	12	Artificial
S183A	0,4982	12	Artificial
S183E	0,4982	12	Artificial
S20A	0,4982	12	Artificial
S20D	0,4982	12	Artificial
S269A	0,4982	12	Artificial
S269E	0,4982	12	Artificial
S362A	0,4982	12	Artificial
S37D	0,4982	12	Artificial
S46A	0,4982	12	Artificial
T18A	0,4982	12	Artificial
T284E	0,4981	12	Artificial
T387A	0,4982	12	Artificial
T55A	0,4982	12	Artificial

Tabla C.3: Parámetros de detección de comunidades con el algoritmo Label propagation para grafos de Carbono alfa-distancia.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,4817	13	Wild
A138P	0,4816	13	Natural
A138S	0,4817	13	Natural
C141Y	0,4823	13	Natural
C238G	0,4817	13	Natural
C238S	0,2771	11	Natural
C275Y	0,4818	13	Natural
D281N	0,4817	13	Natural
E180K	0,4817	13	Natural
E258K	0,4817	13	Natural
E285Q	0,4818	13	Natural
E286A	0,4817	13	Natural
G105C	0,4817	13	Natural
G244D	0,4817	13	Natural

Tabla C.3: Parámetros de detección de comunidades con el algoritmo Label propagation para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
G244V	0,4817	13	Natural
G245C	0,4818	13	Natural
G245D	0,4818	13	Natural
G245S	0,4817	13	Natural
G245V	0,4817	13	Natural
G325V	0,4816	13	Natural
H179Y	0,4817	13	Natural
H193R	0,4816	13	Natural
H233D	0,4817	13	Natural
I251M	0,4817	13	Natural
K132E	0,4817	13	Natural
K292I	0,4817	13	Natural
L252P	0,4816	13	Natural
L257Q	0,4817	13	Natural
L265P	0,4817	13	Natural
L344P	0,4817	13	Natural
M133R	0,4819	13	Natural
M133T	0,4819	13	Natural
M237I	0,4818	13	Natural
M246V	0,4818	13	Natural
N235S	0,4817	13	Natural
P151S	0,4817	13	Natural
P152L	0,4817	13	Natural
P278L	0,4817	13	Natural
P278S	0,4817	13	Natural
P278T	0,4819	13	Natural
P309S	0,4817	13	Natural
P82L	0,4817	13	Natural
Q144L	0,4817	13	Natural
Q167K	0,4817	13	Natural
R156H	0,4817	13	Natural
R158G	0,4817	13	Natural
R158H	0,4817	13	Natural
R174G	0,4818	13	Natural
R175G	0,4817	13	Natural
R175H	0,4818	13	Natural
R175L	0,4817	13	Natural
R181C	0,4817	13	Natural
R181H	0,4817	13	Natural

Tabla C.3: Parámetros de detección de comunidades con el algoritmo Label propagation para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
R196P	0,4817	13	Natural
R213P	0,4818	13	Natural
R213Q	0,4817	13	Natural
R248Q	0,4817	13	Natural
R248W	0,4819	13	Natural
R267Q	0,4817	13	Natural
R273C	0,4818	13	Natural
R273G	0,4818	13	Natural
R273H	0,4818	13	Natural
R273L	0,4817	13	Natural
R282G	0,4818	13	Natural
R282W	0,4817	13	Natural
R283C	0,2770	11	Natural
R290H	0,4817	13	Natural
R290L	0,4817	13	Natural
R306P	0,4817	13	Natural
R337C	0,4817	13	Natural
R337H	0,4817	13	Natural
S227T	0,4817	13	Natural
S241F	0,4817	13	Natural
S241T	0,4817	13	Natural
T155N	0,4817	13	Natural
V172F	0,4818	13	Natural
V173M	0,4818	13	Natural
V197M	0,4817	13	Natural
V272L	0,4817	13	Natural
Y163C	0,4817	13	Natural
Y220C	0,4817	13	Natural
Y234C	0,4817	13	Natural
Y236C	0,4817	13	Natural
E388A	0,4817	13	Artificial
F385A	0,4817	13	Artificial
G361E	0,4816	13	Artificial
K24R	0,4817	13	Artificial
K291RK292R	0,4817	13	Artificial
K319A	0,4817	13	Artificial
K320A	0,4817	13	Artificial
K321A	0,4817	13	Artificial
K370R	0,4817	13	Artificial

Tabla C.3: Parámetros de detección de comunidades con el algoritmo Label propagation para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
K372R	0,4817	13	Artificial
K373R	0,4817	13	Artificial
K381Q	0,4817	13	Artificial
K381R	0,4817	13	Artificial
K382R	0,4817	13	Artificial
K386A	0,4817	13	Artificial
L22QW23S	0,4817	13	Artificial
L383A	0,4817	13	Artificial
P359D	0,4819	13	Artificial
R248S	0,4817	13	Artificial
R333KR335KR337K	0,4817	13	Artificial
S15A	0,4817	13	Artificial
S183A	0,4818	13	Artificial
S183E	0,4817	13	Artificial
S20A	0,4817	13	Artificial
S20D	0,4817	13	Artificial
S269A	0,4817	13	Artificial
S269E	0,4817	13	Artificial
S362A	0,4817	13	Artificial
S37D	0,4817	13	Artificial
S46A	0,4817	13	Artificial
T18A	0,4817	13	Artificial
T284E	0,4816	13	Artificial
T387A	0,4817	13	Artificial
T55A	0,4817	13	Artificial

Tabla C.4: Parámetros de detección de comunidades con el algoritmo Leading eigenvector para grafos de Carbono alfa-distancia.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,4930	9	Wild
A138P	0,4929	9	Natural
A138S	0,4930	9	Natural
C141Y	0,4936	9	Natural
C238G	0,4930	9	Natural
C238S	0,4930	9	Natural
C275Y	0,4931	9	Natural
C238S	0,4930	9	Natural

Tabla C.4: Parámetros de detección de comunidades con el algoritmo Leading eigenvector para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
C275Y	0,4931	9	Natural
D281N	0,4930	9	Natural
E180K	0,4930	9	Natural
E258K	0,4930	9	Natural
E285Q	0,4931	9	Natural
E286A	0,4930	9	Natural
G105C	0,4930	9	Natural
G244D	0,4930	9	Natural
G244V	0,4930	9	Natural
G245C	0,4935	9	Natural
G245D	0,4935	9	Natural
G245S	0,4930	9	Natural
G245V	0,4930	9	Natural
G325V	0,4931	9	Natural
H179Y	0,4930	9	Natural
H193R	0,4933	9	Natural
H233D	0,4930	9	Natural
I251M	0,4930	9	Natural
K132E	0,4934	9	Natural
K292I	0,4930	9	Natural
L252P	0,4929	9	Natural
L257Q	0,4930	9	Natural
L265P	0,4934	9	Natural
L344P	0,4931	9	Natural
M133R	0,4932	9	Natural
M133T	0,4931	9	Natural
M237I	0,4930	9	Natural
M246V	0,4935	9	Natural
N235S	0,4930	9	Natural
P151S	0,4930	9	Natural
P152L	0,4930	9	Natural
P278L	0,4930	9	Natural
P278S	0,4930	9	Natural
P278T	0,4932	9	Natural
P309S	0,4930	9	Natural
P82L	0,4930	9	Natural
Q144L	0,4930	9	Natural
Q167K	0,4930	9	Natural
R156H	0,4930	9	Natural

Tabla C.4: Parámetros de detección de comunidades con el algoritmo Leading eigenvector para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
R158G	0,4930	9	Natural
R158H	0,4930	9	Natural
R174G	0,4933	9	Natural
R175G	0,4930	9	Natural
R175H	0,4930	9	Natural
R175L	0,4930	9	Natural
R181C	0,4930	9	Natural
R181H	0,4930	9	Natural
R196P	0,4930	9	Natural
R213P	0,4935	9	Natural
R213Q	0,4932	9	Natural
R248Q	0,4930	9	Natural
R248W	0,4932	9	Natural
R267Q	0,4930	9	Natural
R273C	0,4930	9	Natural
R273G	0,4930	9	Natural
R273H	0,4930	9	Natural
R273L	0,4930	9	Natural
R282G	0,4931	9	Natural
R282W	0,4930	9	Natural
R283C	0,4933	9	Natural
R290H	0,4930	9	Natural
R290L	0,4930	9	Natural
R306P	0,4930	9	Natural
R337C	0,4930	9	Natural
R337H	0,4930	9	Natural
S227T	0,4930	9	Natural
S241F	0,4930	9	Natural
S241T	0,4930	9	Natural
T155N	0,4930	9	Natural
V172F	0,4930	9	Natural
V173M	0,4931	9	Natural
V197M	0,4930	9	Natural
V272L	0,4930	9	Natural
Y163C	0,4930	9	Natural
Y220C	0,4930	9	Natural
Y234C	0,4930	9	Natural

Tabla C.4: Parámetros de detección de comunidades con el algoritmo Leading eigenvector para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
Y236C	0,4930	9	Natural
E388A	0,4930	9	Artificial
F385A	0,4930	9	Artificial
G361E	0,4930	9	Artificial
K24R	0,4930	9	Artificial
K291RK292R	0,4930	9	Artificial
K319A	0,4930	9	Artificial
K320A	0,4930	9	Artificial
K321A	0,4930	9	Artificial
K370R	0,4930	9	Artificial
K372R	0,4930	9	Artificial
K373R	0,4930	9	Artificial
K381Q	0,4930	9	Artificial
K381R	0,4930	9	Artificial
K382R	0,4930	9	Artificial
K386A	0,4930	9	Artificial
L22QW23S	0,4930	9	Artificial
L383A	0,4930	9	Artificial
P359D	0,4931	9	Artificial
R248S	0,4930	9	Artificial
R333KR335KR337K	0,4930	9	Artificial
S15A	0,4930	9	Artificial
S183A	0,4931	9	Artificial
S183E	0,4930	9	Artificial
S20A	0,4930	9	Artificial
S20D	0,4930	9	Artificial
S269A	0,4930	9	Artificial
S269E	0,4930	9	Artificial
S362A	0,4930	9	Artificial
S37D	0,4930	9	Artificial
S46A	0,4930	9	Artificial
T18A	0,4930	9	Artificial
T284E	0,4929	9	Artificial
T387A	0,4930	9	Artificial
T55A	0,4930	9	Artificial

Tabla C.5: Parámetros de detección de comunidades con el algoritmo Louvain para grafos de Carbono alfa-distancia.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,5063	7	Wild
A138P	0,5061	7	Natural
A138S	0,5063	7	Natural
C141Y	0,5069	7	Natural
C238G	0,5063	7	Natural
C238S	0,5063	7	Natural
C275Y	0,5064	7	Natural
D281N	0,5063	7	Natural
E180K	0,5063	7	Natural
E258K	0,5061	7	Natural
E285Q	0,5064	7	Natural
E286A	0,5063	7	Natural
G105C	0,5064	7	Natural
G244D	0,5063	7	Natural
G244V	0,5026	7	Natural
G245C	0,5063	7	Natural
G245D	0,5063	7	Natural
G245S	0,5063	7	Natural
G245V	0,5063	7	Natural
G325V	0,5064	7	Natural
H179Y	0,5063	7	Natural
H193R	0,5061	7	Natural
H233D	0,5063	7	Natural
I251M	0,5063	7	Natural
K132E	0,5063	7	Natural
K292I	0,5063	7	Natural
L252P	0,5062	7	Natural
L257Q	0,5063	7	Natural
L265P	0,5062	7	Natural
L344P	0,5064	7	Natural
M133R	0,5064	7	Natural
M133T	0,5065	7	Natural
M237I	0,5063	7	Natural
M246V	0,5063	7	Natural
N235S	0,5063	7	Natural
P151S	0,5063	7	Natural
P152L	0,5063	7	Natural
P278L	0,5063	7	Natural
P278S	0,5063	7	Natural
P278T	0,5065	7	Natural

Tabla C.5: Parámetros de detección de comunidades con el algoritmo Louvain para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
P309S	0,5063	7	Natural
P82L	0,5063	7	Natural
Q144L	0,5063	7	Natural
Q167K	0,5063	7	Natural
R156H	0,5063	7	Natural
R158G	0,5063	7	Natural
R158H	0,5063	7	Natural
R174G	0,5065	7	Natural
R175G	0,5063	7	Natural
R175H	0,5063	7	Natural
R175L	0,5063	7	Natural
R181C	0,5063	7	Natural
R181H	0,5063	7	Natural
R196P	0,5063	7	Natural
R213P	0,5063	7	Natural
R213Q	0,5063	7	Natural
R248Q	0,5063	7	Natural
R248W	0,5065	7	Natural
R267Q	0,5063	7	Natural
R273C	0,5063	7	Natural
R273G	0,5063	7	Natural
R273H	0,5063	7	Natural
R273L	0,5063	7	Natural
R282G	0,5064	7	Natural
R282W	0,5063	7	Natural
R283C	0,5064	7	Natural
R290H	0,5063	7	Natural
R290L	0,5063	7	Natural
R306P	0,5063	7	Natural
R337C	0,5063	7	Natural
R337H	0,5063	7	Natural
S227T	0,5063	7	Natural
S241F	0,5063	7	Natural
S241T	0,5063	7	Natural
T155N	0,5063	7	Natural
V172F	0,5063	7	Natural
V173M	0,5063	7	Natural
V197M	0,5063	7	Natural
V272L	0,5063	7	Natural
Y163C	0,5063	7	Natural

Tabla C.5: Parámetros de detección de comunidades con el algoritmo Louvain para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
Y220C	0,5063	7	Natural
Y234C	0,5063	7	Natural
Y236C	0,5063	7	Natural
E388A	0,5063	7	Artificial
F385A	0,5063	7	Artificial
G361E	0,5062	7	Artificial
K24R	0,5063	7	Artificial
K291RK292R	0,5063	7	Artificial
K319A	0,5063	7	Artificial
K320A	0,5063	7	Artificial
K321A	0,5063	7	Artificial
K370R	0,5063	7	Artificial
K372R	0,5063	7	Artificial
K373R	0,5063	7	Artificial
K381Q	0,5063	7	Artificial
K381R	0,5063	7	Artificial
K382R	0,5063	7	Artificial
K386A	0,5063	7	Artificial
L22QW23S	0,5063	7	Artificial
L383A	0,5063	7	Artificial
P359D	0,5064	7	Artificial
R248S	0,5063	7	Artificial
R333KR335KR337K	0,5063	7	Artificial
S15A	0,5063	7	Artificial
S183A	0,5063	7	Artificial
S183E	0,5063	7	Artificial
S20A	0,5063	7	Artificial
S20D	0,5063	7	Artificial
S269A	0,5063	7	Artificial
S269E	0,5063	7	Artificial
S362A	0,5063	7	Artificial
S37D	0,5063	7	Artificial
S46A	0,5063	7	Artificial
T18A	0,5063	7	Artificial
T284E	0,5062	7	Artificial
T387A	0,5063	7	Artificial
T55A	0,5063	7	Artificial

Tabla C.6: Parámetros de detección de comunidades con el algoritmo Spinglass para grafos de Carbono alfa-distancia.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,5076	9	Wild
A138P	0,5081	8	Natural
A138S	0,5082	8	Natural
C141Y	0,5084	9	Natural
C238G	0,5082	8	Natural
C238S	0,5082	8	Natural
C275Y	0,5083	8	Natural
D281N	0,5083	8	Natural
E180K	0,5079	8	Natural
E258K	0,5081	8	Natural
E285Q	0,5083	8	Natural
E286A	0,5076	9	Natural
G105C	0,5079	8	Natural
G244D	0,5093	7	Natural
G244V	0,5090	7	Natural
G245C	0,5082	8	Natural
G245D	0,5082	8	Natural
G245S	0,5076	9	Natural
G245V	0,5087	7	Natural
G325V	0,5083	7	Natural
H179Y	0,5076	9	Natural
H193R	0,5090	7	Natural
H233D	0,5076	9	Natural
I251M	0,5082	8	Natural
K132E	0,5082	8	Natural
K292I	0,5082	8	Natural
L252P	0,5092	7	Natural
L257Q	0,5094	7	Natural
L265P	0,5080	8	Natural
L344P	0,5084	8	Natural
M133R	0,5082	8	Natural
M133T	0,5091	7	Natural
M237I	0,5082	8	Natural
M246V	0,5083	8	Natural
N235S	0,5076	9	Natural
P151S	0,5076	9	Natural
P152L	0,5083	8	Natural
P278L	0,5082	8	Natural
P278S	0,5079	8	Natural
P278T	0,5081	8	Natural

Tabla C.6: Parámetros de detección de comunidades con el algoritmo Spinglass para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
P309S	0,5076	9	Natural
P82L	0,5094	7	Natural
Q144L	0,5083	8	Natural
Q167K	0,5082	8	Natural
R156H	0,5076	9	Natural
R158G	0,5083	8	Natural
R158H	0,5090	7	Natural
R174G	0,5093	7	Natural
R175G	0,5076	9	Natural
R175H	0,5087	7	Natural
R175L	0,5076	9	Natural
R181C	0,5083	8	Natural
R181H	0,5083	8	Natural
R196P	0,5076	9	Natural
R213P	0,5081	8	Natural
R213Q	0,5090	7	Natural
R248Q	0,5082	8	Natural
R248W	0,5078	9	Natural
R267Q	0,5082	8	Natural
R273C	0,5077	9	Natural
R273G	0,5082	7	Natural
R273H	0,5083	8	Natural
R273L	0,5076	9	Natural
R282G	0,5082	8	Natural
R282W	0,5076	9	Natural
R283C	0,5079	10	Natural
R290H	0,5077	8	Natural
R290L	0,5076	9	Natural
R306P	0,5076	9	Natural
R337C	0,5093	7	Natural
R337H	0,5076	9	Natural
S227T	0,5089	7	Natural
S241F	0,5082	8	Natural
S241T	0,5083	8	Natural
T155N	0,5072	8	Natural
V172F	0,5082	8	Natural
V173M	0,5083	8	Natural
V197M	0,5080	8	Natural
V272L	0,5074	9	Natural
Y163C	0,5076	9	Natural

Tabla C.6: Parámetros de detección de comunidades con el algoritmo Spin-glass para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
Y220C	0,5076	9	Natural
Y234C	0,5082	8	Natural
Y236C	0,5076	9	Natural
E388A	0,5076	9	Artificial
F385A	0,5090	7	Artificial
G361E	0,5089	7	Artificial
K24R	0,5073	9	Artificial
K291RK292R	0,5082	8	Artificial
K319A	0,5076	9	Artificial
K320A	0,5076	9	Artificial
K321A	0,5090	7	Artificial
K370R	0,5090	7	Artificial
K372R	0,5077	8	Artificial
K373R	0,5077	8	Artificial
K381Q	0,5090	7	Artificial
K381R	0,5090	7	Artificial
K382R	0,5076	9	Artificial
K386A	0,5082	8	Artificial
L22QW23S	0,5076	9	Artificial
L383A	0,5083	8	Artificial
P359D	0,5078	7	Artificial
R248S	0,5082	8	Artificial
R333KR335KR337K	0,5090	7	Artificial
S15A	0,5076	9	Artificial
S183A	0,5090	7	Artificial
S183E	0,5076	9	Artificial
S20A	0,5082	8	Artificial
S20D	0,5076	9	Artificial
S269A	0,5076	9	Artificial
S269E	0,5081	8	Artificial
S362A	0,5090	7	Artificial
S37D	0,5090	7	Artificial
S46A	0,5076	9	Artificial
T18A	0,5085	7	Artificial
T284E	0,5082	8	Artificial
T387A	0,5090	7	Artificial
T55A	0,5090	7	Artificial

Tabla C.7: Parámetros de detección de comunidades con el algoritmo Walktrap para grafos de Carbono alfa-distancia.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,4372	29	Wild
A138P	0,4371	29	Natural
A138S	0,4372	29	Natural
C141Y	0,4378	29	Natural
C238G	0,4372	29	Natural
C238S	0,4372	29	Natural
C275Y	0,4373	29	Natural
D281N	0,4372	29	Natural
E180K	0,4372	29	Natural
E258K	0,4358	29	Natural
E285Q	0,4373	29	Natural
E286A	0,4372	29	Natural
G105C	0,4372	29	Natural
G244D	0,4372	29	Natural
G244V	0,4372	29	Natural
G245C	0,4373	29	Natural
G245D	0,4373	29	Natural
G245S	0,4372	29	Natural
G245V	0,4372	29	Natural
G325V	0,4372	29	Natural
H179Y	0,4372	29	Natural
H193R	0,4449	29	Natural
H233D	0,4372	29	Natural
I251M	0,4372	29	Natural
K132E	0,4261	30	Natural
K292I	0,4372	29	Natural
L252P	0,4371	29	Natural
L257Q	0,4372	29	Natural
L265P	0,4372	29	Natural
L344P	0,4354	29	Natural
M133R	0,4452	29	Natural
M133T	0,4452	29	Natural
M237I	0,4373	29	Natural
M246V	0,4373	29	Natural
N235S	0,4372	29	Natural
P151S	0,4372	29	Natural
P152L	0,4372	29	Natural
P278L	0,4372	29	Natural
P278S	0,4372	29	Natural
P278T	0,3692	30	Natural

Tabla C.7: Parámetros de detección de comunidades con el algoritmo Walktrap para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
P309S	0,4372	29	Natural
P82L	0,4372	29	Natural
Q144L	0,4372	29	Natural
Q167K	0,4372	29	Natural
R156H	0,4372	29	Natural
R158G	0,4372	29	Natural
R158H	0,4372	29	Natural
R174G	0,3858	30	Natural
R175G	0,4372	29	Natural
R175H	0,4473	29	Natural
R175L	0,4372	29	Natural
R181C	0,4372	29	Natural
R181H	0,4372	29	Natural
R196P	0,4372	29	Natural
R213P	0,4473	29	Natural
R213Q	0,4373	29	Natural
R248Q	0,4372	29	Natural
R248W	0,3692	30	Natural
R267Q	0,4372	29	Natural
R273C	0,4373	29	Natural
R273G	0,4373	29	Natural
R273H	0,4373	29	Natural
R273L	0,4372	29	Natural
R282G	0,4373	29	Natural
R282W	0,4372	29	Natural
R283C	0,4373	29	Natural
R290H	0,4372	29	Natural
R290L	0,4372	29	Natural
R306P	0,4372	29	Natural
R337C	0,4372	29	Natural
R337H	0,4372	29	Natural
S227T	0,4372	29	Natural
S241F	0,4372	29	Natural
S241T	0,4372	29	Natural
T155N	0,4372	29	Natural
V172F	0,4372	29	Natural
V173M	0,4373	29	Natural
V197M	0,4372	29	Natural
V272L	0,4372	29	Natural
Y163C	0,4372	29	Natural

Tabla C.7: Parámetros de detección de comunidades con el algoritmo Walktrap para grafos de Carbono alfa-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
Y220C	0,4372	29	Natural
Y234C	0,4372	29	Natural
Y236C	0,4372	29	Natural
E388A	0,4372	29	Artificial
F385A	0,4372	29	Artificial
G361E	0,4354	29	Artificial
K24R	0,4372	29	Artificial
K291RK292R	0,4372	29	Artificial
K319A	0,4372	29	Artificial
K320A	0,4372	29	Artificial
K321A	0,4372	29	Artificial
K370R	0,4372	29	Artificial
K372R	0,4372	29	Artificial
K373R	0,4372	29	Artificial
K381Q	0,4372	29	Artificial
K381R	0,4372	29	Artificial
K382R	0,4372	29	Artificial
K386A	0,4372	29	Artificial
L22QW23S	0,4372	29	Artificial
L383A	0,4372	29	Artificial
P359D	0,4372	29	Artificial
R248S	0,4372	29	Artificial
R333KR335KR337K	0,4372	29	Artificial
S15A	0,4372	29	Artificial
S183A	0,4373	29	Artificial
S183E	0,4372	29	Artificial
S20A	0,4372	29	Artificial
S20D	0,4372	29	Artificial
S269A	0,4372	29	Artificial
S269E	0,4372	29	Artificial
S362A	0,4372	29	Artificial
S37D	0,4372	29	Artificial
S46A	0,4372	29	Artificial
T18A	0,4372	29	Artificial
T284E	0,4450	29	Artificial
T387A	0,4372	29	Artificial
T55A	0,4381	29	Artificial

En las Tablas C.8, C.9, C.10, C.11, C.12, C.13 y C.14 se presenta la detección de comunidades para los grafos de centroide-distancia de las variantes de p53.

Tabla C.8: Parámetros de detección de comunidades con el algoritmo Fast greedy para grafos de Centroide-distancia.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,4617	5	Wild
A138P	0,4601	6	Natural
A138S	0,4601	6	Natural
C141Y	0,4614	6	Natural
C238G	0,4614	5	Natural
C238S	0,4615	5	Natural
C275Y	0,4557	6	Natural
D281N	0,4597	6	Natural
E180K	0,4596	5	Natural
E258K	0,4621	5	Natural
E285Q	0,4618	5	Natural
E286A	0,4547	6	Natural
G105C	0,4595	6	Natural
G244D	0,4596	6	Natural
G244V	0,4596	5	Natural
G245C	0,4617	5	Natural
G245D	0,4617	5	Natural
G245S	0,4617	5	Natural
G245V	0,4596	6	Natural
G325V	0,4616	5	Natural
H179Y	0,4617	5	Natural
H193R	0,4596	6	Natural
H233D	0,4601	5	Natural
I251M	0,4582	6	Natural
K132E	0,4622	5	Natural
K292I	0,4598	6	Natural
L252P	0,4599	5	Natural
L257Q	0,4596	5	Natural
L265P	0,4617	5	Natural
L344P	0,4616	5	Natural
M133R	0,4582	5	Natural
M133T	0,4602	5	Natural
M237I	0,4596	6	Natural
M246V	0,4618	5	Natural
N235S	0,4617	5	Natural
P151S	0,4594	6	Natural
P152L	0,4594	6	Natural

Tabla C.8: Parámetros de detección de comunidades con el algoritmo Fast greedy para grafos de Centroide-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
P278L	0,4597	6	Natural
P278S	0,4606	6	Natural
P278T	0,4597	6	Natural
P309S	0,4617	5	Natural
P82L	0,4596	6	Natural
Q144L	0,4616	5	Natural
Q167K	0,4619	5	Natural
R156H	0,4621	5	Natural
R158G	0,4522	5	Natural
R158H	0,4615	5	Natural
R174G	0,4617	5	Natural
R175G	0,4610	6	Natural
R175H	0,4617	5	Natural
R175L	0,4596	5	Natural
R181C	0,4596	5	Natural
R181H	0,4617	5	Natural
R196P	0,4587	5	Natural
R213P	0,4598	6	Natural
R213Q	0,4593	5	Natural
R248Q	0,4615	5	Natural
R248W	0,4598	6	Natural
R267Q	0,4601	5	Natural
R273C	0,4606	5	Natural
R273G	0,4608	5	Natural
R273H	0,4616	5	Natural
R273L	0,4587	6	Natural
R282G	0,4613	5	Natural
R282W	0,4604	6	Natural
R283C	0,4588	6	Natural
R290H	0,4594	5	Natural
R290L	0,4597	6	Natural
R306P	0,4617	5	Natural
R337C	0,4617	5	Natural
R337H	0,4617	5	Natural
S227T	0,4616	5	Natural
S241F	0,4598	6	Natural
S241T	0,4615	5	Natural
T155N	0,4616	5	Natural
V172F	0,4596	5	Natural
V173M	0,4615	5	Natural

Tabla C.8: Parámetros de detección de comunidades con el algoritmo Fast greedy para grafos de Centroide-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
V197M	0,4601	5	Natural
V272L	0,4618	5	Natural
Y163C	0,4583	6	Natural
Y220C	0,4586	5	Natural
Y234C	0,4508	5	Natural
Y236C	0,4602	5	Natural
E388A	0,4616	5	Artificial
F385A	0,4616	5	Artificial
G361E	0,4617	5	Artificial
K24R	0,4594	6	Artificial
K291RK292R	0,4597	6	Artificial
K319A	0,4597	5	Artificial
K320A	0,4597	6	Artificial
K321A	0,4595	5	Artificial
K370R	0,4617	5	Artificial
K372R	0,4618	5	Artificial
K373R	0,4616	5	Artificial
K381Q	0,4617	5	Artificial
K381R	0,4617	5	Artificial
K382R	0,4594	6	Artificial
K386A	0,4616	5	Artificial
L22QW23S	0,4595	6	Artificial
L383A	0,4616	5	Artificial
P359D	0,4596	5	Artificial
R248S	0,4600	5	Artificial
R333KR335KR337K	0,4589	6	Artificial
S15A	0,4617	5	Artificial
S183A	0,4617	5	Artificial
S183E	0,4596	6	Artificial
S20A	0,4616	5	Artificial
S20D	0,4593	6	Artificial
S269A	0,4596	5	Artificial
S269E	0,4587	5	Artificial
S362A	0,4617	5	Artificial
S37D	0,4617	5	Artificial
S46A	0,4617	5	Artificial
T18A	0,4594	6	Artificial
T284E	0,4590	5	Artificial
T387A	0,4616	5	Artificial
T55A	0,4617	5	Artificial

Tabla C.9: Parámetros de detección de comunidades con el algoritmo Infomap para grafos de Centroide-distancia.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,2932	12	Wild
A138P	0,2955	11	Natural
A138S	0,5041	12	Natural
C141Y	0,5046	12	Natural
C238G	0,2929	12	Natural
C238S	0,2930	12	Natural
C275Y	0,2937	12	Natural
D281N	0,2931	12	Natural
E180K	0,5037	12	Natural
E258K	0,2954	11	Natural
E285Q	0,5035	12	Natural
E286A	0,5028	12	Natural
G105C	0,2934	12	Natural
G244D	0,2934	12	Natural
G244V	0,5035	12	Natural
G245C	0,5036	12	Natural
G245D	0,2934	12	Natural
G245S	0,5036	12	Natural
G245V	0,2931	12	Natural
G325V	0,2934	12	Natural
H179Y	0,5036	12	Natural
H193R	0,5036	12	Natural
H233D	0,2932	12	Natural
I251M	0,5031	12	Natural
K132E	0,2935	12	Natural
K292I	0,5039	12	Natural
L252P	0,2931	12	Natural
L257Q	0,2933	12	Natural
L265P	0,5036	12	Natural
L344P	0,2933	12	Natural
M133R	0,2955	11	Natural
M133T	0,5042	12	Natural
M237I	0,2932	12	Natural
M246V	0,2932	12	Natural
N235S	0,2933	12	Natural
P151S	0,2934	12	Natural
P152L	0,2955	11	Natural
P278L	0,2931	12	Natural
P278S	0,2930	12	Natural
P278T	0,2953	11	Natural

Tabla C.9: Parámetros de detección de comunidades con el algoritmo Infomap para grafos de Centroide-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
P309S	0,2933	12	Natural
P82L	0,2950	11	Natural
Q144L	0,5035	12	Natural
Q167K	0,2933	12	Natural
R156H	0,2932	12	Natural
R158G	0,2928	12	Natural
R158H	0,2933	12	Natural
R174G	0,2941	12	Natural
R175G	0,5041	12	Natural
R175H	0,2923	12	Natural
R175L	0,5036	12	Natural
R181C	0,2949	11	Natural
R181H	0,2939	12	Natural
R196P	0,2926	12	Natural
R213P	0,2954	11	Natural
R213Q	0,5030	12	Natural
R248Q	0,5039	12	Natural
R248W	0,5038	12	Natural
R267Q	0,5039	12	Natural
R273C	0,2931	12	Natural
R273G	0,2932	12	Natural
R273H	0,2932	12	Natural
R273L	0,5043	12	Natural
R282G	0,2931	12	Natural
R282W	0,2935	12	Natural
R283C	0,5031	12	Natural
R290H	0,2931	12	Natural
R290L	0,5036	12	Natural
R306P	0,2957	11	Natural
R337C	0,5036	12	Natural
R337H	0,2931	12	Natural
S227T	0,2932	12	Natural
S241F	0,2936	12	Natural
S241T	0,5035	12	Natural
T155N	0,5036	12	Natural
V172F	0,2950	11	Natural
V173M	0,5034	12	Natural
V197M	0,2924	12	Natural
V272L	0,5036	12	Natural
Y163C	0,2956	11	Natural

Tabla C.9: Parámetros de detección de comunidades con el algoritmo Infomap para grafos de Centroide-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
Y220C	0,5035	12	Natural
Y234C	0,2957	11	Natural
Y236C	0,5046	12	Natural
E388A	0,2953	11	Artificial
F385A	0,5035	12	Artificial
G361E	0,5035	12	Artificial
K24R	0,2949	11	Artificial
K291RK292R	0,2932	12	Artificial
K319A	0,2933	12	Artificial
K320A	0,5037	12	Artificial
K321A	0,5036	12	Artificial
K370R	0,2931	12	Artificial
K372R	0,2936	12	Artificial
K373R	0,5037	12	Artificial
K381Q	0,2933	12	Artificial
K381R	0,2934	12	Artificial
K382R	0,5037	12	Artificial
K386A	0,2930	12	Artificial
L22QW23S	0,2931	12	Artificial
L383A	0,2931	12	Artificial
P359D	0,5032	12	Artificial
R248S	0,5042	12	Artificial
R333KR335KR337K	0,2942	12	Artificial
S15A	0,5036	12	Artificial
S183A	0,2933	12	Artificial
S183E	0,5036	12	Artificial
S20A	0,2931	12	Artificial
S20D	0,5033	12	Artificial
S269A	0,2954	11	Artificial
S269E	0,5038	12	Artificial
S362A	0,5036	12	Artificial
S37D	0,2954	11	Artificial
S46A	0,2953	11	Artificial
T18A	0,2948	11	Artificial
T284E	0,2896	12	Artificial
T387A	0,5036	12	Artificial
T55A	0,5036	12	Artificial

Tabla C.10: Parámetros de detección de comunidades con el algoritmo Label propagation para grafos de Centroide-distancia.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,2909	13	Wild
A138P	0,2911	13	Natural
A138S	0,2910	13	Natural
C141Y	0,2911	13	Natural
C238G	0,2907	13	Natural
C238S	0,2907	13	Natural
C275Y	0,2912	13	Natural
D281N	0,2908	13	Natural
E180K	0,2907	13	Natural
E258K	0,2909	13	Natural
E285Q	0,2907	13	Natural
E286A	0,2908	13	Natural
G105C	0,2909	13	Natural
G244D	0,2899	13	Natural
G244V	0,2897	13	Natural
G245C	0,2909	13	Natural
G245D	0,2909	13	Natural
G245S	0,2909	13	Natural
G245V	0,2910	13	Natural
G325V	0,2833	14	Natural
H179Y	0,2909	13	Natural
H193R	0,2909	13	Natural
H233D	0,2908	13	Natural
I251M	0,2906	13	Natural
K132E	0,2910	13	Natural
K292I	0,2915	13	Natural
L252P	0,2909	13	Natural
L257Q	0,2909	13	Natural
L265P	0,2909	13	Natural
L344P	0,2908	13	Natural
M133R	0,2910	13	Natural
M133T	0,2911	13	Natural
M237I	0,2909	13	Natural
M246V	0,2909	13	Natural
N235S	0,2908	13	Natural
P151S	0,2910	13	Natural
P152L	0,2910	13	Natural
P278L	0,2908	13	Natural
P278S	0,2908	13	Natural
P278T	0,2908	13	Natural

Tabla C.10: Parámetros de detección de comunidades con el algoritmo Label propagation para grafos de Centroide-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
P309S	0,2909	13	Natural
P82L	0,2906	13	Natural
Q144L	0,2908	13	Natural
Q167K	0,2832	14	Natural
R156H	0,2909	13	Natural
R158G	0,2906	13	Natural
R158H	0,2909	13	Natural
R174G	0,2908	13	Natural
R175G	0,2907	13	Natural
R175H	0,2909	13	Natural
R175L	0,2907	13	Natural
R181C	0,2908	13	Natural
R181H	0,2908	13	Natural
R196P	0,2904	13	Natural
R213P	0,2910	13	Natural
R213Q	0,2909	13	Natural
R248Q	0,2908	13	Natural
R248W	0,2909	13	Natural
R267Q	0,2907	13	Natural
R273C	0,2947	13	Natural
R273G	0,2948	13	Natural
R273H	0,2950	13	Natural
R273L	0,2947	13	Natural
R282G	0,2907	13	Natural
R282W	0,2910	13	Natural
R283C	0,2905	13	Natural
R290H	0,2829	14	Natural
R290L	0,2910	13	Natural
R306P	0,2921	12	Natural
R337C	0,2909	13	Natural
R337H	0,2908	13	Natural
S227T	0,2909	13	Natural
S241F	0,2911	13	Natural
S241T	0,2908	13	Natural
T155N	0,2909	13	Natural
V172F	0,2906	13	Natural
V173M	0,2908	13	Natural
V197M	0,2910	13	Natural
V272L	0,2910	13	Natural
Y163C	0,2911	13	Natural

Tabla C.10: Parámetros de detección de comunidades con el algoritmo Label propagation para grafos de Centroide-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
Y220C	0,2910	13	Natural
Y234C	0,2912	13	Natural
Y236C	0,2910	13	Natural
E388A	0,2908	13	Artificial
F385A	0,2908	13	Artificial
G361E	0,2908	13	Artificial
K24R	0,2904	13	Artificial
K291RK292R	0,2909	13	Artificial
K319A	0,2924	12	Artificial
K320A	0,2910	13	Artificial
K321A	0,2905	13	Artificial
K370R	0,2909	13	Artificial
K372R	0,2911	13	Artificial
K373R	0,2926	12	Artificial
K381Q	0,2909	13	Artificial
K381R	0,2910	13	Artificial
K382R	0,2906	13	Artificial
K386A	0,2908	13	Artificial
L22QW23S	0,2906	13	Artificial
L383A	0,2908	13	Artificial
P359D	0,2903	13	Artificial
R248S	0,2905	13	Artificial
R333KR335KR337K	0,2916	13	Artificial
S15A	0,2909	13	Artificial
S183A	0,2909	13	Artificial
S183E	0,2910	13	Artificial
S20A	0,2908	13	Artificial
S20D	0,2903	13	Artificial
S269A	0,2909	13	Artificial
S269E	0,2911	13	Artificial
S362A	0,2909	13	Artificial
S37D	0,2909	13	Artificial
S46A	0,2909	13	Artificial
T18A	0,2905	13	Artificial
T284E	0,2912	13	Artificial
T387A	0,2909	13	Artificial
T55A	0,2909	13	Artificial

Tabla C.11: Parámetros de detección de comunidades con el algoritmo Leading eigenvector para grafos de Centroide-distancia.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,4893	7	Wild
A138P	0,4900	7	Natural
A138S	0,4911	7	Natural
C141Y	0,4906	7	Natural
C238G	0,4904	7	Natural
C238S	0,4888	7	Natural
C275Y	0,4917	7	Natural
D281N	0,4892	7	Natural
E180K	0,4936	7	Natural
E258K	0,4923	7	Natural
E285Q	0,4920	7	Natural
E286A	0,4913	7	Natural
G105C	0,4893	7	Natural
G244D	0,4893	7	Natural
G244V	0,4890	7	Natural
G245C	0,4893	7	Natural
G245D	0,4893	7	Natural
G245S	0,4896	7	Natural
G245V	0,4893	7	Natural
G325V	0,4891	7	Natural
H179Y	0,4893	7	Natural
H193R	0,4893	7	Natural
H233D	0,4910	7	Natural
I251M	0,4890	7	Natural
K132E	0,4899	7	Natural
K292I	0,4897	7	Natural
L252P	0,4913	7	Natural
L257Q	0,4893	7	Natural
L265P	0,4893	7	Natural
L344P	0,4893	7	Natural
M133R	0,4931	7	Natural
M133T	0,4911	7	Natural
M237I	0,4897	7	Natural
M246V	0,4907	7	Natural
N235S	0,4892	7	Natural
P151S	0,4893	7	Natural
P152L	0,4893	7	Natural
P278L	0,4895	7	Natural
P278S	0,4897	7	Natural
P278T	0,4894	7	Natural

Tabla C.11: Parámetros de detección de comunidades con el algoritmo Leading eigenvector para grafos de Centroide-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
P309S	0,4893	7	Natural
P82L	0,4891	7	Natural
Q144L	0,4908	7	Natural
Q167K	0,4908	7	Natural
R156H	0,4923	7	Natural
R158G	0,4940	7	Natural
R158H	0,4905	7	Natural
R174G	0,4920	7	Natural
R175G	0,4936	7	Natural
R175H	0,4896	7	Natural
R175L	0,4929	7	Natural
R181C	0,4936	7	Natural
R181H	0,4914	7	Natural
R196P	0,4885	7	Natural
R213P	0,4898	7	Natural
R213Q	0,4898	7	Natural
R248Q	0,4892	7	Natural
R248W	0,4913	7	Natural
R267Q	0,4927	7	Natural
R273C	0,4903	7	Natural
R273G	0,4914	7	Natural
R273H	0,4891	7	Natural
R273L	0,4906	7	Natural
R282G	0,4889	7	Natural
R282W	0,4908	7	Natural
R283C	0,4888	7	Natural
R290H	0,4922	7	Natural
R290L	0,4894	7	Natural
R306P	0,4895	7	Natural
R337C	0,4893	7	Natural
R337H	0,4893	7	Natural
S227T	0,4893	7	Natural
S241F	0,4894	7	Natural
S241T	0,4892	7	Natural
T155N	0,4893	7	Natural
V172F	0,4916	7	Natural
V173M	0,4892	7	Natural
V197M	0,4897	7	Natural
V272L	0,4922	7	Natural

Tabla C.11: Parámetros de detección de comunidades con el algoritmo Leading eigenvector para grafos de Centroide-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
Y163C	0,4912	7	Natural
Y220C	0,4893	7	Natural
Y234C	0,4896	7	Natural
Y236C	0,4916	7	Natural
E388A	0,4893	7	Artificial
F385A	0,4893	7	Artificial
G361E	0,4892	7	Artificial
K24R	0,4891	7	Artificial
K291RK292R	0,4893	7	Artificial
K319A	0,4895	7	Artificial
K320A	0,4894	7	Artificial
K321A	0,4892	7	Artificial
K370R	0,4894	7	Artificial
K372R	0,4893	7	Artificial
K373R	0,4891	7	Artificial
K381Q	0,4893	7	Artificial
K381R	0,4894	7	Artificial
K382R	0,4892	7	Artificial
K386A	0,4893	7	Artificial
L22QW23S	0,4892	7	Artificial
L383A	0,4893	7	Artificial
P359D	0,4890	7	Artificial
R248S	0,0000	1	Artificial
R333KR335KR337K	0,4897	7	Artificial
S15A	0,4893	7	Artificial
S183A	0,4893	7	Artificial
S183E	0,4892	7	Artificial
S20A	0,4893	7	Artificial
S20D	0,4888	7	Artificial
S269A	0,4894	7	Artificial
S269E	0,4906	7	Artificial
S362A	0,4893	7	Artificial
S37D	0,4894	7	Artificial
S46A	0,4893	7	Artificial
T18A	0,4892	7	Artificial
T284E	0,4905	7	Artificial
T387A	0,4893	7	Artificial
T55A	0,4893	7	Artificial

Tabla C.12: Parámetros de detección de comunidades con el algoritmo Louvain para grafos de Centroide-distancia.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,5057	7	Wild
A138P	0,5062	7	Natural
A138S	0,5062	7	Natural
C141Y	0,5062	7	Natural
C238G	0,5055	7	Natural
C238S	0,5055	7	Natural
C275Y	0,5052	7	Natural
D281N	0,5055	7	Natural
E180K	0,5074	7	Natural
E258K	0,5059	7	Natural
E285Q	0,5064	7	Natural
E286A	0,5049	7	Natural
G105C	0,5056	7	Natural
G244D	0,5057	7	Natural
G244V	0,5055	7	Natural
G245C	0,5057	7	Natural
G245D	0,5058	7	Natural
G245S	0,5057	7	Natural
G245V	0,5057	7	Natural
G325V	0,5056	7	Natural
H179Y	0,5057	7	Natural
H193R	0,5057	7	Natural
H233D	0,5033	7	Natural
I251M	0,5052	7	Natural
K132E	0,5060	7	Natural
K292I	0,5060	7	Natural
L252P	0,5058	7	Natural
L257Q	0,5057	7	Natural
L265P	0,5073	7	Natural
L344P	0,5057	7	Natural
M133R	0,5053	7	Natural
M133T	0,5063	7	Natural
M237I	0,5056	7	Natural
M246V	0,5058	7	Natural
N235S	0,5058	7	Natural
P151S	0,5056	7	Natural
P152L	0,5077	7	Natural
P278L	0,5058	7	Natural
P278S	0,5058	7	Natural
P278T	0,5057	7	Natural

Tabla C.12: Parámetros de detección de comunidades con el algoritmo Louvain para grafos de Centroide-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
P309S	0,5057	7	Natural
P82L	0,5055	7	Natural
Q144L	0,5055	7	Natural
Q167K	0,5061	7	Natural
R156H	0,5077	7	Natural
R158G	0,5063	7	Natural
R158H	0,5059	7	Natural
R174G	0,5065	7	Natural
R175G	0,5073	7	Natural
R175H	0,5057	7	Natural
R175L	0,5069	7	Natural
R181C	0,5055	7	Natural
R181H	0,5057	7	Natural
R196P	0,5061	7	Natural
R213P	0,5059	7	Natural
R213Q	0,5053	7	Natural
R248Q	0,5057	7	Natural
R248W	0,5068	7	Natural
R267Q	0,5060	7	Natural
R273C	0,5056	7	Natural
R273G	0,5056	7	Natural
R273H	0,5056	7	Natural
R273L	0,5059	7	Natural
R282G	0,5052	7	Natural
R282W	0,5057	7	Natural
R283C	0,5052	7	Natural
R290H	0,5057	7	Natural
R290L	0,5057	7	Natural
R306P	0,5058	7	Natural
R337C	0,5057	7	Natural
R337H	0,5057	7	Natural
S227T	0,5056	7	Natural
S241F	0,5088	7	Natural
S241T	0,5057	7	Natural
T155N	0,5057	7	Natural
V172F	0,5054	7	Natural
V173M	0,5055	7	Natural
V197M	0,5061	7	Natural
V272L	0,5056	7	Natural
Y163C	0,5054	7	Natural

Tabla C.12: Parámetros de detección de comunidades con el algoritmo Louvain para grafos de Centroide-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
Y220C	0,5054	7	Natural
Y234C	0,5065	7	Natural
Y236C	0,5087	7	Natural
E388A	0,5056	7	Artificial
F385A	0,5056	7	Artificial
G361E	0,5057	7	Artificial
K24R	0,5055	7	Artificial
K291RK292R	0,5057	7	Artificial
K319A	0,5059	7	Artificial
K320A	0,5057	7	Artificial
K321A	0,5055	7	Artificial
K370R	0,5057	7	Artificial
K372R	0,5058	7	Artificial
K373R	0,5057	7	Artificial
K381Q	0,5057	7	Artificial
K381R	0,5057	7	Artificial
K382R	0,5055	7	Artificial
K386A	0,5056	7	Artificial
L22QW23S	0,5055	7	Artificial
L383A	0,5057	7	Artificial
P359D	0,5053	7	Artificial
R248S	0,5050	7	Artificial
R333KR335KR337K	0,5060	7	Artificial
S15A	0,5057	7	Artificial
S183A	0,5057	7	Artificial
S183E	0,5057	7	Artificial
S20A	0,5056	7	Artificial
S20D	0,5054	7	Artificial
S269A	0,5057	7	Artificial
S269E	0,5059	7	Artificial
S362A	0,5057	7	Artificial
S37D	0,5057	7	Artificial
S46A	0,5057	7	Artificial
T18A	0,5055	7	Artificial
T284E	0,5058	7	Artificial
T387A	0,5057	7	Artificial
T55A	0,5057	7	Artificial

Tabla C.13: Parámetros de detección de comunidades con el algoritmo Spinglass para grafos de Centroide-distancia.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,5115	8	Wild
A138P	0,5116	8	Natural
A138S	0,5119	8	Natural
C141Y	0,5125	8	Natural
C238G	0,5111	8	Natural
C238S	0,5113	8	Natural
C275Y	0,5110	8	Natural
D281N	0,5111	8	Natural
E180K	0,5116	8	Natural
E258K	0,5117	8	Natural
E285Q	0,5108	8	Natural
E286A	0,5105	8	Natural
G105C	0,5115	8	Natural
G244D	0,5115	8	Natural
G244V	0,5113	8	Natural
G245C	0,5113	8	Natural
G245D	0,5110	9	Natural
G245S	0,5116	8	Natural
G245V	0,5114	8	Natural
G325V	0,5110	8	Natural
H179Y	0,5113	8	Natural
H193R	0,5122	7	Natural
H233D	0,5103	8	Natural
I251M	0,5110	8	Natural
K132E	0,5127	7	Natural
K292I	0,5115	8	Natural
L252P	0,5117	8	Natural
L257Q	0,5115	8	Natural
L265P	0,5113	8	Natural
L344P	0,5122	7	Natural
M133R	0,5111	8	Natural
M133T	0,5117	8	Natural
M237I	0,5115	8	Natural
M246V	0,5122	7	Natural
N235S	0,5113	8	Natural
P151S	0,5109	8	Natural
P152L	0,5114	8	Natural
P278L	0,5116	8	Natural
P278S	0,5113	8	Natural
P278T	0,5123	7	Natural

Tabla C.13: Parámetros de detección de comunidades con el algoritmo Spinglass para grafos de Centroide-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
P309S	0,5115	8	Natural
P82L	0,5110	8	Natural
Q144L	0,5113	8	Natural
Q167K	0,5126	7	Natural
R156H	0,5117	8	Natural
R158G	0,5117	8	Natural
R158H	0,5116	8	Natural
R174G	0,5106	8	Natural
R175G	0,5130	7	Natural
R175H	0,5115	8	Natural
R175L	0,5116	8	Natural
R181C	0,5123	7	Natural
R181H	0,5122	7	Natural
R196P	0,5102	8	Natural
R213P	0,5112	8	Natural
R213Q	0,5109	8	Natural
R248Q	0,5114	8	Natural
R248W	0,5115	8	Natural
R267Q	0,5118	8	Natural
R273C	0,5119	7	Natural
R273G	0,5115	8	Natural
R273H	0,5112	8	Natural
R273L	0,5117	8	Natural
R282G	0,5108	8	Natural
R282W	0,5115	8	Natural
R283C	0,5114	7	Natural
R290H	0,5111	8	Natural
R290L	0,5111	8	Natural
R306P	0,5124	7	Natural
R337C	0,5124	7	Natural
R337H	0,5112	8	Natural
S227T	0,5118	7	Natural
S241F	0,5125	7	Natural
S241T	0,5113	8	Natural
T155N	0,5115	8	Natural
V172F	0,5111	8	Natural
V173M	0,5110	8	Natural
V197M	0,5120	8	Natural
V272L	0,5108	9	Natural
Y163C	0,5111	8	Natural

Tabla C.13: Parámetros de detección de comunidades con el algoritmo Spin-glass para grafos de Centroide-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
Y220C	0,5121	7	Natural
Y234C	0,5118	8	Natural
Y236C	0,5122	8	Natural
E388A	0,5110	9	Artificial
F385A	0,5111	8	Artificial
G361E	0,5115	8	Artificial
K24R	0,5108	9	Artificial
K291RK292R	0,5114	8	Artificial
K319A	0,5116	8	Artificial
K320A	0,5108	9	Artificial
K321A	0,5106	9	Artificial
K370R	0,5112	8	Artificial
K372R	0,5113	8	Artificial
K373R	0,5115	8	Artificial
K381Q	0,5115	8	Artificial
K381R	0,5112	8	Artificial
K382R	0,5114	8	Artificial
K386A	0,5122	7	Artificial
L22QW23S	0,5123	7	Artificial
L383A	0,5109	9	Artificial
P359D	0,5118	7	Artificial
R248S	0,5119	8	Artificial
R333KR335KR337K	0,5125	7	Artificial
S15A	0,5123	7	Artificial
S183A	0,5113	8	Artificial
S183E	0,5113	8	Artificial
S20A	0,5108	9	Artificial
S20D	0,5112	8	Artificial
S269A	0,5122	7	Artificial
S269E	0,5125	7	Artificial
S362A	0,5115	8	Artificial
S37D	0,5122	7	Artificial
S46A	0,5112	8	Artificial
T18A	0,5113	8	Artificial
T284E	0,5113	8	Artificial
T387A	0,5112	8	Artificial
T55A	0,5115	8	Artificial

Tabla C.14: Parámetros de detección de comunidades con el algoritmo Walktrap para grafos de Centroide-distancia.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,4490	33	Wild
A138P	0,4421	33	Natural
A138S	0,4419	33	Natural
C141Y	0,4492	32	Natural
C238G	0,4485	33	Natural
C238S	0,4485	33	Natural
C275Y	0,4498	32	Natural
D281N	0,4489	33	Natural
E180K	0,4492	33	Natural
E258K	0,4535	32	Natural
E285Q	0,4499	32	Natural
E286A	0,4465	33	Natural
G105C	0,4490	33	Natural
G244D	0,4488	33	Natural
G244V	0,4487	33	Natural
G245C	0,4490	33	Natural
G245D	0,4490	33	Natural
G245S	0,4491	33	Natural
G245V	0,4490	33	Natural
G325V	0,4492	33	Natural
H179Y	0,4490	33	Natural
H193R	0,4490	33	Natural
H233D	0,4524	32	Natural
I251M	0,4453	33	Natural
K132E	0,4496	33	Natural
K292I	0,4494	33	Natural
L252P	0,4504	32	Natural
L257Q	0,4471	33	Natural
L265P	0,4471	33	Natural
L344P	0,4490	33	Natural
M133R	0,4467	33	Natural
M133T	0,4496	33	Natural
M237I	0,4460	33	Natural
M246V	0,4489	32	Natural
N235S	0,4489	33	Natural
P151S	0,4470	33	Natural
P152L	0,4469	33	Natural
P278L	0,4491	33	Natural
P278S	0,4492	33	Natural
P278T	0,4491	33	Natural

Tabla C.14: Parámetros de detección de comunidades con el algoritmo Walktrap para grafos de Centroide-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
P309S	0,4491	33	Natural
P82L	0,4489	33	Natural
Q144L	0,4489	33	Natural
Q167K	0,4503	32	Natural
R156H	0,4528	32	Natural
R158G	0,4495	33	Natural
R158H	0,4526	32	Natural
R174G	0,4445	33	Natural
R175G	0,4475	33	Natural
R175H	0,4343	33	Natural
R175L	0,4505	32	Natural
R181C	0,4515	33	Natural
R181H	0,4533	32	Natural
R196P	0,4336	33	Natural
R213P	0,4493	33	Natural
R213Q	0,4484	32	Natural
R248Q	0,4489	33	Natural
R248W	0,4493	33	Natural
R267Q	0,4472	33	Natural
R273C	0,4490	33	Natural
R273G	0,4490	33	Natural
R273H	0,4475	33	Natural
R273L	0,4504	32	Natural
R282G	0,4484	33	Natural
R282W	0,4466	33	Natural
R283C	0,4457	33	Natural
R290H	0,4490	33	Natural
R290L	0,4490	33	Natural
R306P	0,4489	33	Natural
R337C	0,4488	33	Natural
R337H	0,4490	33	Natural
S227T	0,4470	33	Natural
S241F	0,4491	33	Natural
S241T	0,4489	33	Natural
T155N	0,4490	33	Natural
V172F	0,4484	33	Natural
V173M	0,4488	33	Natural
V197M	0,4493	33	Natural
V272L	0,4490	33	Natural
Y163C	0,4488	33	Natural

Tabla C.14: Parámetros de detección de comunidades con el algoritmo Walktrap para grafos de Centroide-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
Y220C	0,4490	33	Natural
Y234C	0,4498	32	Natural
Y236C	0,4478	32	Natural
E388A	0,4490	33	Artificial
F385A	0,4490	33	Artificial
G361E	0,4465	33	Artificial
K24R	0,4492	33	Artificial
K291RK292R	0,4491	33	Artificial
K319A	0,4489	33	Artificial
K320A	0,4491	33	Artificial
K321A	0,4510	32	Artificial
K370R	0,4490	33	Artificial
K372R	0,4491	33	Artificial
K373R	0,4510	32	Artificial
K381Q	0,4491	33	Artificial
K381R	0,4491	33	Artificial
K382R	0,4492	33	Artificial
K386A	0,4490	33	Artificial
L22QW23S	0,4489	33	Artificial
L383A	0,4490	33	Artificial
P359D	0,4442	33	Artificial
R248S	0,4510	32	Artificial
R333KR335KR337K	0,4493	33	Artificial
S15A	0,4490	33	Artificial
S183A	0,4490	33	Artificial
S183E	0,4456	33	Artificial
S20A	0,4490	33	Artificial
S20D	0,4502	33	Artificial
S269A	0,4501	32	Artificial
S269E	0,4395	33	Artificial
S362A	0,4490	33	Artificial
S37D	0,4491	33	Artificial
S46A	0,4514	32	Artificial
T18A	0,4500	33	Artificial
T284E	0,4463	33	Artificial
T387A	0,4490	33	Artificial
T55A	0,4514	32	Artificial

En las Tablas C.15, C.16, C.17, C.18 y C.19 se presenta la detección de comunidades para los grafos de interacciones intermoleculares de las variantes de p53. No se incluyen spinglass y walktrap debido a que estos algoritmos no entregaron resultados con este tipo de grafo.

Tabla C.15: Parámetros de detección de comunidades con el algoritmo Fast greedy para grafos de Interacciones intermoleculares.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,9250	214	Wild
A138P	0,9243	216	Natural
A138S	0,9250	214	Natural
C141Y	0,9244	213	Natural
C238G	0,9250	214	Natural
C238S	0,9250	214	Natural
C275Y	0,9250	214	Natural
D281N	0,9300	214	Natural
E180K	0,9250	214	Natural
E258K	0,9238	214	Natural
E285Q	0,9340	214	Natural
E286A	0,9316	214	Natural
G105C	0,9246	215	Natural
G244D	0,9250	214	Natural
G244V	0,9250	214	Natural
G245C	0,9250	214	Natural
G245D	0,9249	214	Natural
G245S	0,9249	214	Natural
G245V	0,9250	214	Natural
G325V	0,9249	214	Natural
H179Y	0,9255	215	Natural
H193R	0,9240	214	Natural
H233D	0,9249	214	Natural
I251M	0,9251	214	Natural
K132E	0,9365	214	Natural
K292I	0,9268	214	Natural
L252P	0,9249	214	Natural
L257Q	0,9249	214	Natural
L265P	0,9246	215	Natural
L344P	0,9265	215	Natural
M133R	0,9254	214	Natural
M133T	0,9261	214	Natural
M237I	0,9250	214	Natural
M246V	0,9250	214	Natural
N235S	0,9249	216	Natural
P151S	0,9250	214	Natural

Tabla C.15: Parámetros de detección de comunidades con el algoritmo Fast greedy para grafos de Interacciones intermoleculares (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
P152L	0,9250	214	Natural
P278L	0,9250	214	Natural
P278S	0,9233	213	Natural
P278T	0,9233	213	Natural
P309S	0,9250	214	Natural
P82L	0,9250	214	Natural
Q144L	0,9251	214	Natural
Q167K	0,9250	214	Natural
R156H	0,9251	215	Natural
R158G	0,9246	214	Natural
R158H	0,9246	214	Natural
R174G	0,9253	215	Natural
R175G	0,9252	217	Natural
R175H	0,9249	216	Natural
R175L	0,9252	217	Natural
R181C	0,9258	215	Natural
R181H	0,9257	215	Natural
R196P	0,9249	215	Natural
R213P	0,9298	216	Natural
R213Q	0,9299	216	Natural
R248Q	0,9250	214	Natural
R248W	0,9250	214	Natural
R267Q	0,9248	214	Natural
R273C	0,9302	214	Natural
R273G	0,9302	214	Natural
R273H	0,9302	214	Natural
R273L	0,9302	214	Natural
R282G	0,9260	214	Natural
R282W	0,9256	214	Natural
R283C	0,9253	215	Natural
R290H	0,9280	213	Natural
R290L	0,9277	213	Natural
R306P	0,9249	215	Natural
R337C	0,9250	214	Natural
R337H	0,9250	214	Natural
S227T	0,9244	214	Natural
S241F	0,9250	214	Natural
S241T	0,9250	214	Natural
T155N	0,9246	214	Natural

Tabla C.15: Parámetros de detección de comunidades con el algoritmo Fast greedy para grafos de Centroide-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
V172F	0,9250	214	Natural
V173M	0,9250	214	Natural
V197M	0,9250	214	Natural
V272L	0,9250	214	Natural
Y163C	0,9251	215	Natural
Y220C	0,9250	214	Natural
Y234C	0,9250	214	Natural
Y236C	0,9250	214	Natural
E388A	0,9246	215	Artificial
F385A	0,9250	214	Artificial
G361E	0,9243	215	Artificial
K24R	0,9251	214	Artificial
K291RK292R	0,9317	216	Artificial
K319A	0,9246	214	Artificial
K320A	0,9248	215	Artificial
K321A	0,9250	214	Artificial
K370R	0,9249	214	Artificial
K372R	0,9250	214	Artificial
K373R	0,9250	214	Artificial
K381Q	0,9250	214	Artificial
K381R	0,9250	214	Artificial
K382R	0,9250	214	Artificial
K386A	0,9246	215	Artificial
L22QW23S	0,9258	213	Artificial
L383A	0,9250	214	Artificial
P359D	0,9250	214	Artificial
R248S	0,9250	214	Artificial
R333KR335KR337K	0,9250	214	Artificial
S15A	0,9244	215	Artificial
S183A	0,9250	214	Artificial
S183E	0,9250	214	Artificial
S20A	0,9250	214	Artificial
S20D	0,9249	214	Artificial
S269A	0,9253	214	Artificial
S269E	0,9251	215	Artificial
S362A	0,9243	215	Artificial
S37D	0,9248	214	Artificial
S46A	0,9250	214	Artificial
T18A	0,9248	214	Artificial

Tabla C.15: Parámetros de detección de comunidades con el algoritmo Fast greedy para grafos de Centroide-distancia (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
T284E	0,9249	214	Artificial
T387A	0,9250	214	Artificial
T55A	0,9246	215	Artificial

Tabla C.16: Parámetros de detección de comunidades con el algoritmo Infomap para grafos de Interacciones intermoleculares.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,8969	226	Wild
A138P	0,8962	228	Natural
A138S	0,8947	227	Natural
C141Y	0,8947	226	Natural
C238G	0,8969	226	Natural
C238S	0,8969	226	Natural
C275Y	0,8968	226	Natural
D281N	0,9018	226	Natural
E180K	0,8968	226	Natural
E258K	0,8932	227	Natural
E285Q	0,9076	225	Natural
E286A	0,9033	226	Natural
G105C	0,8981	226	Natural
G244D	0,8969	226	Natural
G244V	0,8969	226	Natural
G245C	0,8969	226	Natural
G245D	0,8970	226	Natural
G245S	0,8960	226	Natural
G245V	0,8969	226	Natural
G325V	0,8968	226	Natural
H179Y	0,8986	225	Natural
H193R	0,8955	227	Natural
H233D	0,8969	226	Natural
I251M	0,8970	226	Natural
K132E	0,9099	225	Natural
K292I	0,8981	227	Natural
L252P	0,8969	226	Natural
L257Q	0,8969	226	Natural
L265P	0,8965	227	Natural
L344P	0,9027	226	Natural

Tabla C.16: Parámetros de detección de comunidades con el algoritmo Infomap para grafos de Interacciones intermoleculares (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
M133R	0,8973	226	Natural
M133T	0,8979	226	Natural
M237I	0,8949	227	Natural
M246V	0,8969	226	Natural
N235S	0,8968	228	Natural
P151S	0,8969	226	Natural
P152L	0,8969	226	Natural
P278L	0,8969	226	Natural
P278S	0,8952	226	Natural
P278T	0,8946	226	Natural
P309S	0,8969	226	Natural
P82L	0,8969	226	Natural
Q144L	0,8970	226	Natural
Q167K	0,8969	226	Natural
R156H	0,8970	227	Natural
R158G	0,8966	226	Natural
R158H	0,8966	226	Natural
R174G	0,8957	227	Natural
R175G	0,8977	228	Natural
R175H	0,8942	228	Natural
R175L	0,8977	228	Natural
R181C	0,8981	226	Natural
R181H	0,8981	226	Natural
R196P	0,8967	227	Natural
R213P	0,9024	227	Natural
R213Q	0,9024	227	Natural
R248Q	0,8969	226	Natural
R248W	0,8969	226	Natural
R267Q	0,8967	226	Natural
R273C	0,9017	226	Natural
R273G	0,9017	226	Natural
R273H	0,9017	226	Natural
R273L	0,9017	226	Natural
R282G	0,8968	227	Natural
R282W	0,8936	227	Natural
R283C	0,8966	227	Natural
R290H	0,8982	226	Natural
R290L	0,8990	226	Natural
R306P	0,8969	227	Natural

Tabla C.16: Parámetros de detección de comunidades con el algoritmo Infomap para grafos de Interacciones intermoleculares (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
R337C	0,8969	226	Natural
R337H	0,8969	226	Natural
S227T	0,8963	226	Natural
S241F	0,8969	226	Natural
S241T	0,8969	226	Natural
T155N	0,8965	226	Natural
V172F	0,8969	226	Natural
V173M	0,8969	226	Natural
V197M	0,8969	226	Natural
V272L	0,8959	226	Natural
Y163C	0,8968	227	Natural
Y220C	0,8970	226	Natural
Y234C	0,8970	226	Natural
Y236C	0,8970	226	Natural
E388A	0,8966	227	Artificial
F385A	0,8969	226	Artificial
G361E	0,8963	227	Artificial
K24R	0,8957	227	Artificial
K291RK292R	0,9100	227	Artificial
K319A	0,8965	226	Artificial
K320A	0,8967	227	Artificial
K321A	0,8969	226	Artificial
K370R	0,8968	226	Artificial
K372R	0,8969	226	Artificial
K373R	0,8969	226	Artificial
K381Q	0,8969	226	Artificial
K381R	0,8969	226	Artificial
K382R	0,8969	226	Artificial
K386A	0,8966	227	Artificial
L22QW23S	0,8978	225	Artificial
L383A	0,8969	226	Artificial
P359D	0,8969	226	Artificial
R248S	0,8969	226	Artificial
R333KR335KR337K	0,8969	226	Artificial
S15A	0,8964	227	Artificial
S183A	0,8969	226	Artificial
S183E	0,8969	226	Artificial
S20A	0,8969	226	Artificial
S20D	0,8969	226	Artificial

Tabla C.16: Parámetros de detección de comunidades con el algoritmo Infomap para grafos de Interacciones intermoleculares (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
S269A	0,8979	227	Artificial
S269E	0,8981	227	Artificial
S362A	0,8963	227	Artificial
S37D	0,8967	226	Artificial
S46A	0,8970	226	Artificial
T18A	0,8967	226	Artificial
T284E	0,8958	226	Artificial
T387A	0,8969	226	Artificial
T55A	0,8965	227	Artificial

Tabla C.17: Parámetros de detección de comunidades con el algoritmo Label propagation para grafos de Interacciones intermoleculares.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,8242	243	Wild
A138P	0,8257	244	Natural
A138S	0,8249	243	Natural
C141Y	0,8223	243	Natural
C238G	0,8242	243	Natural
C238S	0,8242	243	Natural
C275Y	0,8240	243	Natural
D281N	0,8287	243	Natural
E180K	0,8240	243	Natural
E258K	0,8196	244	Natural
E285Q	0,8314	243	Natural
E286A	0,8303	243	Natural
G105C	0,8250	243	Natural
G244D	0,8246	243	Natural
G244V	0,8246	243	Natural
G245C	0,8241	243	Natural
G245D	0,8237	243	Natural
G245S	0,8229	243	Natural
G245V	0,8245	243	Natural
G325V	0,8241	243	Natural
H179Y	0,8241	243	Natural
H193R	0,8258	243	Natural
H233D	0,8240	243	Natural

Tabla C.17: Parámetros de detección de comunidades con el algoritmo Label propagation para grafos de Interacciones intermoleculares (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
I251M	0,8243	243	Natural
K132E	0,8392	242	Natural
K292I	0,8351	242	Natural
L252P	0,8241	243	Natural
L257Q	0,8241	243	Natural
L265P	0,8236	244	Natural
L344P	0,8293	243	Natural
M133R	0,8246	243	Natural
M133T	0,8251	243	Natural
M237I	0,8255	243	Natural
M246V	0,8242	243	Natural
N235S	0,8263	244	Natural
P151S	0,8242	243	Natural
P152L	0,8242	243	Natural
P278L	0,8243	243	Natural
P278S	0,8227	243	Natural
P278T	0,8220	243	Natural
P309S	0,8242	243	Natural
P82L	0,8242	243	Natural
Q144L	0,8243	243	Natural
Q167K	0,8242	243	Natural
R156H	0,8275	243	Natural
R158G	0,8171	244	Natural
R158H	0,8171	244	Natural
R174G	0,8236	244	Natural
R175G	0,8318	243	Natural
R175H	0,8296	243	Natural
R175L	0,8318	243	Natural
R181C	0,8257	243	Natural
R181H	0,8256	243	Natural
R196P	0,8239	244	Natural
R213P	0,8323	244	Natural
R213Q	0,8324	244	Natural
R248Q	0,8242	243	Natural
R248W	0,8242	243	Natural
R267Q	0,8239	243	Natural
R273C	0,8285	243	Natural
R273G	0,8285	243	Natural
R273H	0,8285	243	Natural

Tabla C.17: Parámetros de detección de comunidades con el algoritmo Label propagation para grafos de Interacciones intermoleculares (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
R273L	0,8286	243	Natural
R282G	0,8235	244	Natural
R282W	0,8213	244	Natural
R283C	0,8288	243	Natural
R290H	0,8319	242	Natural
R290L	0,8317	242	Natural
R306P	0,8241	244	Natural
R337C	0,8242	243	Natural
R337H	0,8242	243	Natural
S227T	0,8232	243	Natural
S241F	0,8242	243	Natural
S241T	0,8242	243	Natural
T155N	0,8236	243	Natural
V172F	0,8242	243	Natural
V173M	0,8242	243	Natural
V197M	0,8243	243	Natural
V272L	0,8241	243	Natural
Y163C	0,8240	244	Natural
Y220C	0,8243	243	Natural
Y234C	0,8243	243	Natural
Y236C	0,8243	243	Natural
E388A	0,8237	244	Artificial
F385A	0,8242	243	Artificial
G361E	0,8232	244	Artificial
K24R	0,8231	244	Artificial
K291RK292R	0,8446	242	Artificial
K319A	0,8236	243	Artificial
K320A	0,8239	244	Artificial
K321A	0,8242	243	Artificial
K370R	0,8260	242	Artificial
K372R	0,8242	243	Artificial
K373R	0,8242	243	Artificial
K381Q	0,8242	243	Artificial
K381R	0,8242	243	Artificial
K382R	0,8242	243	Artificial
K386A	0,8237	244	Artificial
L22QW23S	0,8256	242	Artificial
L383A	0,8242	243	Artificial
P359D	0,8242	243	Artificial
R248S	0,8242	243	Artificial

Tabla C.17: Parámetros de detección de comunidades con el algoritmo Label propagation para grafos de Interacciones intermoleculares (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
R333KR335KR337K	0,8241	243	Artificial
S15A	0,8233	244	Artificial
S183A	0,8242	243	Artificial
S183E	0,8242	243	Artificial
S20A	0,8242	243	Artificial
S20D	0,8241	243	Artificial
S269A	0,8235	244	Artificial
S269E	0,8238	244	Artificial
S362A	0,8232	244	Artificial
S37D	0,8239	243	Artificial
S46A	0,8243	243	Artificial
T18A	0,8239	243	Artificial
T284E	0,8231	243	Artificial
T387A	0,8242	243	Artificial
T55A	0,8236	244	Artificial

Tabla C.18: Parámetros de detección de comunidades con el algoritmo Leading eigenvector para grafos de Interacciones intermoleculares.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,9247	213	Wild
A138P	0,9239	215	Natural
A138S	0,9247	213	Natural
C141Y	0,9227	213	Natural
C238G	0,9247	213	Natural
C238S	0,9247	213	Natural
C275Y	0,9247	213	Natural
D281N	0,9336	213	Natural
E180K	0,9247	213	Natural
E258K	0,9234	213	Natural
E285Q	0,9335	214	Natural
E286A	0,9304	214	Natural
G105C	0,9243	214	Natural
G244D	0,9247	213	Natural
G244V	0,9247	213	Natural

Tabla C.18: Parámetros de detección de comunidades con el algoritmo Leading eigenvector para grafos de Interacciones intermoleculares (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
G245C	0,9247	213	Natural
G245D	0,9246	213	Natural
G245S	0,9246	213	Natural
G245V	0,9247	213	Natural
G325V	0,9246	213	Natural
H179Y	0,9252	214	Natural
H193R	0,9236	213	Natural
H233D	0,9246	213	Natural
I251M	0,9248	213	Natural
K132E	0,9355	214	Natural
K292I	0,9268	214	Natural
L252P	0,9246	213	Natural
L257Q	0,9246	213	Natural
L265P	0,9243	214	Natural
L344P	0,9262	214	Natural
M133R	0,9252	213	Natural
M133T	0,9259	213	Natural
M237I	0,9247	213	Natural
M246V	0,9247	213	Natural
N235S	0,9245	215	Natural
P151S	0,9247	213	Natural
P152L	0,9247	213	Natural
P278L	0,9246	213	Natural
P278S	0,9228	213	Natural
P278T	0,9222	213	Natural
P309S	0,9247	213	Natural
P82L	0,9247	213	Natural
Q144L	0,9248	213	Natural
Q167K	0,9247	213	Natural
R156H	0,9248	214	Natural
R158G	0,9243	213	Natural
R158H	0,9243	213	Natural
R174G	0,9249	214	Natural
R175G	0,9248	216	Natural
R175H	0,9245	215	Natural
R175L	0,9248	216	Natural
R181C	0,9255	214	Natural
R181H	0,9255	214	Natural
R196P	0,9245	214	Natural

Tabla C.18: Parámetros de detección de comunidades con el algoritmo Leading eigenvector para grafos de Interacciones intermoleculares (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
R213P	0,9294	215	Natural
R213Q	0,9294	215	Natural
R248Q	0,9247	213	Natural
R248W	0,9247	213	Natural
R267Q	0,9245	213	Natural
R273C	0,9251	214	Natural
R273G	0,9251	214	Natural
R273H	0,9251	214	Natural
R273L	0,9250	214	Natural
R282G	0,9260	214	Natural
R282W	0,9256	214	Natural
R283C	0,9249	214	Natural
R290H	0,9280	213	Natural
R290L	0,9277	213	Natural
R306P	0,9246	214	Natural
R337C	0,9247	213	Natural
R337H	0,9247	213	Natural
S227T	0,9240	213	Natural
S241F	0,9247	213	Natural
S241T	0,9247	213	Natural
T155N	0,9243	213	Natural
V172F	0,9247	213	Natural
V173M	0,9247	213	Natural
V197M	0,9247	213	Natural
V272L	0,9247	213	Natural
Y163C	0,9248	214	Natural
Y220C	0,9247	213	Natural
Y234C	0,9247	213	Natural
Y236C	0,9247	213	Natural
E388A	0,9243	214	Artificial
F385A	0,9247	213	Artificial
G361E	0,9240	214	Artificial
K24R	0,9248	213	Artificial
K291RK292R	0,9299	217	Artificial
K319A	0,9243	213	Artificial
K320A	0,9245	214	Artificial
K321A	0,9247	213	Artificial
K370R	0,9245	213	Artificial
K372R	0,9247	213	Artificial

Tabla C.18: Parámetros de detección de comunidades con el algoritmo Leading eigenvector para grafos de Interacciones intermoleculares (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
K373R	0,9247	213	Artificial
K381Q	0,9247	213	Artificial
K381R	0,9247	213	Artificial
K382R	0,9247	213	Artificial
K386A	0,9243	214	Artificial
L22QW23S	0,9256	212	Artificial
L383A	0,9247	213	Artificial
P359D	0,9247	213	Artificial
R248S	0,9247	213	Artificial
R333KR335KR337K	0,9247	213	Artificial
S15A	0,9241	214	Artificial
S183A	0,9247	213	Artificial
S183E	0,9247	213	Artificial
S20A	0,9247	213	Artificial
S20D	0,9246	213	Artificial
S269A	0,9253	214	Artificial
S269E	0,9254	214	Artificial
S362A	0,9240	214	Artificial
S37D	0,9245	213	Artificial
S46A	0,9247	213	Artificial
T18A	0,9245	213	Artificial
T284E	0,9246	213	Artificial
T387A	0,9247	213	Artificial
T55A	0,9242	214	Artificial

Tabla C.19: Parámetros de detección de comunidades con el algoritmo Louvain para grafos de Interacciones intermoleculares.

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
hp53	0,9278	213	Wild
A138P	0,9271	215	Natural
A138S	0,9278	213	Natural
C141Y	0,9244	213	Natural
C238G	0,9278	213	Natural
C238S	0,9278	213	Natural
C275Y	0,9279	213	Natural
D281N	0,9336	213	Natural
E180K	0,9278	213	Natural

Tabla C.19: Parámetros de detección de comunidades con el algoritmo Louvain para grafos de Interacciones intermoleculares (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
E258K	0,9266	213	Natural
E285Q	0,9340	214	Natural
E286A	0,9316	214	Natural
G105C	0,9274	214	Natural
G244D	0,9279	213	Natural
G244V	0,9279	213	Natural
G245C	0,9278	213	Natural
G245D	0,9278	213	Natural
G245S	0,9277	213	Natural
G245V	0,9278	213	Natural
G325V	0,9277	213	Natural
H179Y	0,9283	214	Natural
H193R	0,9268	213	Natural
H233D	0,9278	213	Natural
I251M	0,9279	213	Natural
K132E	0,9365	214	Natural
K292I	0,9282	214	Natural
L252P	0,9277	213	Natural
L257Q	0,9278	213	Natural
L265P	0,9274	214	Natural
L344P	0,9293	214	Natural
M133R	0,9279	213	Natural
M133T	0,9263	214	Natural
M237I	0,9278	213	Natural
M246V	0,9278	213	Natural
N235S	0,9277	215	Natural
P151S	0,9278	213	Natural
P152L	0,9278	213	Natural
P278L	0,9278	213	Natural
P278S	0,9233	213	Natural
P278T	0,9233	213	Natural
P309S	0,9278	213	Natural
P82L	0,9278	213	Natural
Q144L	0,9279	213	Natural
Q167K	0,9278	213	Natural
R156H	0,9279	214	Natural
R158G	0,9274	213	Natural
R158H	0,9275	213	Natural
R174G	0,9281	214	Natural

Tabla C.19: Parámetros de detección de comunidades con el algoritmo Louvain para grafos de Interacciones intermoleculares (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
R175G	0,9280	216	Natural
R175H	0,9277	215	Natural
R175L	0,9280	216	Natural
R181C	0,9286	214	Natural
R181H	0,9286	214	Natural
R196P	0,9277	214	Natural
R213P	0,9326	215	Natural
R213Q	0,9326	215	Natural
R248Q	0,9278	213	Natural
R248W	0,9278	213	Natural
R267Q	0,9276	213	Natural
R273C	0,9336	213	Natural
R273G	0,9336	213	Natural
R273H	0,9336	213	Natural
R273L	0,9336	213	Natural
R282G	0,9290	214	Natural
R282W	0,9255	214	Natural
R283C	0,9281	214	Natural
R290H	0,9291	213	Natural
R290L	0,9290	213	Natural
R306P	0,9278	214	Natural
R337C	0,9278	213	Natural
R337H	0,9278	213	Natural
S227T	0,9272	213	Natural
S241F	0,9278	213	Natural
S241T	0,9278	213	Natural
T155N	0,9274	213	Natural
V172F	0,9278	213	Natural
V173M	0,9278	213	Natural
V197M	0,9278	213	Natural
V272L	0,9279	213	Natural
Y163C	0,9280	214	Natural
Y220C	0,9279	213	Natural
Y234C	0,9279	213	Natural
Y236C	0,9278	213	Natural
E388A	0,9275	214	Artificial
F385A	0,9278	213	Artificial
G361E	0,9271	214	Artificial
K24R	0,9280	213	Artificial

Tabla C.19: Parámetros de detección de comunidades con el algoritmo Louvain para grafos de Interacciones intermoleculares (cont.).

Proteína	**Modularidad**	**Número de Comunidades**	**Tipo de Mutación**
K291RK292R	0,9332	216	Artificial
K319A	0,9274	213	Artificial
K320A	0,9276	214	Artificial
K321A	0,9278	213	Artificial
K370R	0,9277	213	Artificial
K372R	0,9278	213	Artificial
K373R	0,9278	213	Artificial
K381Q	0,9278	213	Artificial
K381R	0,9278	213	Artificial
K382R	0,9278	213	Artificial
K386A	0,9275	214	Artificial
L22QW23S	0,9287	212	Artificial
L383A	0,9278	213	Artificial
P359D	0,9278	213	Artificial
R248S	0,9278	213	Artificial
R333KR335KR337K	0,9278	213	Artificial
S15A	0,9272	214	Artificial
S183A	0,9278	213	Artificial
S183E	0,9278	213	Artificial
S20A	0,9278	213	Artificial
S20D	0,9278	213	Artificial
S269A	0,9281	214	Artificial
S269E	0,9281	214	Artificial
S362A	0,9271	214	Artificial
S37D	0,9276	213	Artificial
S46A	0,9279	213	Artificial
T18A	0,9276	213	Artificial
T284E	0,9277	213	Artificial
T387A	0,9278	213	Artificial
T55A	0,9274	214	Artificial

Anexo D

Similitud entre las comunidades de la proteína nativa y ciertas variantes de interés

Se muestra en las tablas presentadas a continuación la comparación de grafos para el estudio del efecto de las mutaciones en proteínas. Para lo que se calculó la similitud entre el estado pre y post mutación por medio de los indicadores similitud nat-mut y similitud mut-nat. En estas se destaca en azul la fila en que se encuentra el cambio del aminoácido mutado.

D.1. Grafos Carbono alfa-Distancia

Las cuatro variantes de interés presentaron cambios en la detección de comunidades.

D.1.1. C141Y

Esta variante solo presentó cambios en los algoritmos Fast greedy y Spinglass, los que se muestran en las Tablas D.1 y D.2 respectivamente.

Tabla D.1: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Fast greedy utilizando grafos carbono alfa-distancia.

Comunidad original	**Comunidad variante**	**Similitud nat-mut**	**Similitud mut-nat**
1	1	99,2	100,0
1	3	0,8	0,7
3	3	100,0	99,3

Tabla D.2: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Spinglass utilizando grafos carbono alfa-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
1	4	100,0	100,0
2	1	4,0	2,8
2	5	96,0	70,6
3	1	100,0	97,2
4	3	100,0	96,6
5	6	100,0	100,0
6	7	100,0	100,0
7	2	59,3	100,0
7	5	37,0	29,4
7	8	3,7	100,0
8	3	100,0	3,4

D.1.2. G244D

Esta variante solo presentó cambios en el algoritmo Spinglass, los que se muestran en la Tabla D.3.

Tabla D.3: Porcentaje de similitud entre comunidades de la proteína nativa y la variante G244D para el algoritmo Spinglass utilizando grafos carbono alfa-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
0	0	100,0	97,3
1	2	100,0	100,0
2	1	4,0	2,8
2	4	96,0	47,1
3	1	100,0	97,2
4	6	100,0	96,6
5	0	1,2	1,4
5	3	98,8	100,0
6	0	2,2	1,4
6	5	97,8	100,0
7	4	100,0	52,9
8	6	100,0	3,4

D.1.3. P278L

Esta variante solo presentó cambios en el algoritmo Spinglass, los que se muestran en la Tabla D.4.

Tabla D.4: Porcentaje de similitud entre comunidades de la proteína nativa y la variante P278L para el algoritmo Spinglass utilizando grafos carbono alfa-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
0	0	100,0	98,6
1	3	100,0	100,0
2	5	100,0	73,5
3	2	100,0	100,0
4	7	100,0	96,6
5	1	100,0	100,0
6	0	2,2	1,4
6	6	97,8	100,0
7	4	66,7	100,0
7	5	33,3	26,5
8	7	100,0	3,4

D.1.4. T284E

Esta variante solo presentó cambios en los algoritmos Spinglass y Walktrap, los que se muestran en las Tablas D.5 y D.6 respectivamente.

Tabla D.5: Porcentaje de similitud entre comunidades de la proteína nativa y la variante T284E para el algoritmo Spinglass utilizando grafos carbono alfa-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
0	0	100,0	98,6
2	4	4,0	2,8
2	5	96,0	70,6
3	4	100,0	97,2
4	7	100,0	96,6
5	3	100,0	100,0
6	0	2,2	1,4
6	6	97,8	100,0
7	2	63,0	100,0
7	5	37,0	29,4
8	7	100,0	3,4

Tabla D.6: Porcentaje de similitud entre comunidades de la proteína nativa y la variante T284E para el algoritmo Walktrap utilizando grafos carbono alfa-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
14	14	100,0	84,8
15	14	12,0	13,9
15	15	88,0	100,0
16	14	2,4	1,3
16	16	97,5	100,0

D.2. Grafos Centroide-Distancia

Las cuatro variantes de interés presentaron cambios en la detección de comunidades.

D.2.1. C141Y

Esta variante presentó cambios en los algoritmos Fast greedy, Infomap, Leading eigenvector, Spinglass y Walktrap, los que se muestran en las Tablas D.7, D.8, D.9, D.10 y D.11 respectivamente.

Tabla D.7: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Fast greedy utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
2	2	98,2	100,0
2	3	1,8	1,0
3	3	84,0	99,0
3	4	14,3	100,0
3	5	1,7	6,5
4	5	100,0	93,5

Tabla D.8: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Infomap utilizando grafos de centroide-distancia.

Comunidad original	**Comunidad variante**	**Similitud nat-mut**	**Similitud mut-nat**
1	1	100,0	92,0
2	1	20,0	8,0
2	2	80,0	33,3
3	2	100,0	66,7
4	3	100,0	100,0
5	4	100,0	100,0
6	5	42,4	100,0
6	6	35,4	100,0
6	7	22,2	97,8
7	7	2,6	2,2
7	8	97,4	68,5
8	9	100,0	100,0
9	8	100,0	31,5
10	10	100,0	84,2
11	10	17,6	15,8
11	11	82,4	100,0

Tabla D.9: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Leading eigenvector utilizando grafos de centroide-distancia.

Comunidad original	**Comunidad variante**	**Similitud nat-mut**	**Similitud mut-nat**
2	2	98,6	100,0
2	3	1,4	2,2
3	3	100,0	97,8

Tabla D.10: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Spinglass utilizando grafos de centroide-distancia.

Comunidad original	**Comunidad variante**	**Similitud nat-mut**	**Similitud mut-nat**
0	5	100,0	100,0
1	6	100,0	100,0
2	1	100,0	96,7
3	2	17,2	18,5
3	3	79,3	100,0
3	4	3,4	2,7
4	0	98,7	100,0
4	1	1,3	3,3
5	2	100,0	81,5
6	7	100,0	100,0
7	4	100,0	97,3

Tabla D.11: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Walktrap utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
12	12	100,0	33,3
13	12	100,0	66,7
14	13	100,0	100,0
15	14	100,0	100,0
16	15	100,0	100,0
17	16	100,0	77,9
18	16	26,3	22,1
18	17	73,8	100,0
19	18	100,0	100,0
20	19	100,0	100,0
21	20	100,0	100,0
22	21	100,0	100,0
23	22	100,0	100,0
24	23	100,0	100,0
25	24	100,0	100,0
26	25	100,0	100,0
27	26	100,0	100,0
28	27	100,0	100,0
29	28	100,0	100,0
30	29	100,0	100,0
31	30	100,0	100,0
32	31	100,0	100,0

D.2.2. G244D

Esta variante presentó cambios en los algoritmos Fast greedy, Infomap, Label propagation, Leading eigenvector, Spinglass y Walktrap, los que se muestran en las Tablas D.12, D.13, D.14, D.15, D.16 y D.17 respectivamente. En las Tablas D.12 y D.13 no se destaca ninguna fila, debido a que el aminoácido mutado no está involucrado en un cambio directo, manteniéndose en la comunidad 1 y 6, correspondientemente.

Tabla D.12: Porcentaje de similitud entre comunidades de la proteína nativa y la variante G244D para el algoritmo Fast greedy utilizando grafos de centroide-distancia.

Comunidad original	**Comunidad variante**	**Similitud nat-mut**	**Similitud mut-nat**
3	3	90,8	93,9
3	4	9,2	100,0
4	3	24,1	6,1
4	5	75,9	100,0

Tabla D.13: Porcentaje de similitud entre comunidades de la proteína nativa y la variante G244D para el algoritmo infomap utilizando grafos de centroide-distancia.

Comunidad original	**Comunidad variante**	**Similitud nat-mut**	**Similitud mut-nat**
10	10	100,0	84,2
11	10	17,6	15,8
11	11	82,4	100,0

Tabla D.14: Porcentaje de similitud entre comunidades de la proteína nativa y la variante G244D para el algoritmo Label propagation utilizando grafos de centroide-distancia.

Comunidad original	**Comunidad variante**	**Similitud nat-mut**	**Similitud mut-nat**
5	5	90,9	100,0
5	6	9,1	0,5
6	6	100,0	99,5

Tabla D.15: Porcentaje de similitud entre comunidades de la proteína nativa y la variante G244D para el algoritmo Leading eigenvector utilizando grafos de centroide-distancia.

Comunidad original	**Comunidad variante**	**Similitud nat-mut**	**Similitud mut-nat**
2	2	98,6	98,6
2	3	1,4	2,2
3	3	100,0	97,8
5	2	2,5	1,4
5	5	97,5	100,0

Tabla D.16: Porcentaje de similitud entre comunidades de la proteína nativa y la variante G244D para el algoritmo Spinglass utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
0	0	1,4	1,2
0	6	98,6	100,0
1	0	100,0	98,8
2	3	100,0	96,7
3	5	100,0	93,5
4	1	98,7	100,0
4	3	1,3	3,3
5	5	4,5	3,2
5	7	95,5	100,0
6	4	100,0	100,0
7	2	97,2	100,0
7	5	2,8	3,2

Tabla D.17: Porcentaje de similitud entre comunidades de la proteína nativa y la variante G244D para el algoritmo Walktrap utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
16	16	100,0	80,0
17	16	1,4	20,0
17	17	98,6	100,0
18	18	100,0	100,0
19	19	100,0	100,0
20	20	100,0	100,0
21	21	100,0	100,0
22	22	100,0	100,0
23	23	100,0	100,0
24	24	100,0	100,0
25	25	100,0	100,0
26	26	100,0	100,0
27	27	100,0	100,0
28	28	100,0	100,0
29	29	100,0	100,0
30	30	100,0	100,0
31	31	100,0	100,0
32	32	100,0	100,0

D.2.3. P278L

Esta variante presentó cambios en los algoritmos Fast greedy, Infomap y Spinglass, los que se muestran en las Tablas D.18, D.19 y D.20 respectivamente. En la Tabla D.19 no se destaca ninguna fila, debido a que el aminoácido mutado no está involucrado en un cambio directo, manteniéndose en la comunidad 6.

Tabla D.18: Porcentaje de similitud entre comunidades de la proteína nativa y la variante P278L para el algoritmo Fast greedy utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
3	3	90,8	93,9
3	4	9,2	100,0
4	3	24,1	6,1
4	5	75,9	100,0

Tabla D.19: Porcentaje de similitud entre comunidades de la proteína nativa y la variante P278L para el algoritmo Infomap utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
2	2	100,0	83,3
3	2	12,5	16,7
3	3	87,5	100,0
10	10	100,0	84,2
11	10	17,6	15,8
11	11	82,4	100,0

Tabla D.20: Porcentaje de similitud entre comunidades de la proteína nativa y la variante P278L para el algoritmo Spinglass utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
1	3	100,0	100,0
2	6	100,0	100,0
3	4	17,2	18,5
3	7	82,8	100,0
4	1	100,0	100,0
5	4	100,0	81,5
6	2	100,0	100,0
7	5	100,0	100,0

D.2.4. T284E

Esta variante presentó cambios en los algoritmos Fast greedy, Infomap, Leading eigenvector, Spinglass y Walktrap, los que se muestran en las Tablas D.21, D.22, D.23, D.24 y D.25 respectivamente. En la Tabla D.25, no se destaca ninguna fila, debido a que el aminoácido mutado no está involucrado en un cambio directo, manteniéndose en la comunidad 19.

Tabla D.21: Porcentaje de similitud entre comunidades de la proteína nativa y la variante T284E para el algoritmo Fast greedy utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
3	3	100,0	94,4
4	3	24,1	5,6
4	4	75,9	100,0

Tabla D.22: Porcentaje de similitud entre comunidades de la proteína nativa y la variante T284E para el algoritmo Infomap utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
6	6	100,0	99,5
7	6	2,6	0,5
7	7	97,4	100,0
8	8	100,0	100,0
9	9	100,0	100,0
10	10	100,0	84,2
11	10	17,6	15,8
11	11	82,4	100,0

Tabla D.23: Porcentaje de similitud entre comunidades de la proteína nativa y la variante T284E para el algoritmo Leading eigenvector utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
1	1	98,9	100,0
1	2	1,1	1,4
2	2	98,6	97,3
2	3	1,4	2,2
3	3	100,0	97,8
5	2	2,5	1,4
5	5	97,5	100,0

Tabla D.24: Porcentaje de similitud entre comunidades de la proteína nativa y la variante T284E para el algoritmo Spinglass utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
0	6	98,6	100,0
0	7	1,4	1,2
1	7	100,0	98,8
2	0	100,0	96,7
3	3	17,2	18,5
3	5	82,8	96,0
4	0	1,3	3,3
4	4	98,7	100,0
5	3	100,0	81,5
6	2	100,0	100,0
7	1	97,2	100,0
7	5	2,8	4,0

Tabla D.25: Porcentaje de similitud entre comunidades de la proteína nativa y la variante T284E para el algoritmo Walktrap utilizando grafos de centroide-distancia.

Comunidad original	Comunidad variante	Similitud nat-mut	Similitud mut-nat
20	20	89,5	100,0
20	25	10,5	20,0
25	25	100,0	80,0

D.3. Grafos Interacciones intermoleculares

Sólo una de las cuatro variantes de interés presentó cambios en la detección de comunidades.

D.3.1. C141Y

Está variante presentó cambio en los algoritmos Fast greedy, Infomap, Label propagation, Leading eigenvector y multilevel, lo que son presentados en las Tablas D.26, D.27, D.28, D.29 y D.30, respectivamente.

Tabla D.26: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Fast greedy utilizando grafos de interacciones intermoleculares.

Comunidad original	**Comunidad variante**	**Similitud nat-mut**	**Similitud mut-nat**
90	90	100	80
104	90	100	20
105	104	100	100
106	105	100	100
107	106	100	100
108	107	100	100
109	108	100	100
110	109	100	100
111	110	100	100
112	111	100	100
113	112	100	100
114	113	100	100
115	114	100	100
116	115	100	100
117	116	100	100
118	117	100	100
119	118	100	100
120	119	100	100
121	120	100	100
122	121	100	100
123	122	100	100
124	123	100	100
125	124	100	100
126	125	100	100
127	126	100	100
128	127	100	100
129	128	100	100
130	129	100	100
131	130	100	100
132	131	100	100
133	132	100	100
134	133	100	100
135	134	100	100
136	135	100	100
137	136	100	100
138	137	100	100
139	138	100	100
140	139	100	100

Tabla D.26: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Fast greedy utilizando grafos de interacciones intermoleculares (cont.).

Comunidad original	**Comunidad variante**	**Similitud nat-mut**	**Similitud mut-nat**
141	140	100	100
142	141	100	100
143	142	100	100
144	143	100	100
145	144	100	100
146	145	100	100
147	146	100	100
148	147	100	100
149	148	100	100
150	149	100	100
151	150	100	100
152	151	100	100
153	152	100	100
154	153	100	100
155	154	100	100
156	155	100	100
157	156	100	100
158	157	100	100
159	158	100	100
160	159	100	100
161	160	100	100
162	161	100	100
163	162	100	100
164	163	100	100
165	164	100	100
166	165	100	100
167	166	100	100
168	167	100	100
169	168	100	100
170	169	100	100
171	170	100	100
172	171	100	100
173	172	100	100
174	173	100	100
175	174	100	100
176	175	100	100
177	176	100	100
178	177	100	100
179	178	100	100

Tabla D.26: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Fast greedy utilizando grafos de interacciones intermoleculares (cont.).

Comunidad original	**Comunidad variante**	**Similitud nat-mut**	**Similitud mut-nat**
180	179	100	100
181	180	100	100
182	181	100	100
183	182	100	100
184	183	100	100
185	184	100	100
186	185	100	100
187	186	100	100
188	187	100	100
189	188	100	100
190	189	100	100
191	190	100	100
192	191	100	100
193	192	100	100
194	193	100	100
195	194	100	100
196	195	100	100
197	196	100	100
198	197	100	100
199	198	100	100
200	199	100	100
201	200	100	100
202	201	100	100
203	202	100	100
204	203	100	100
205	204	100	100
206	205	100	100
207	206	100	100
208	207	100	100
209	208	100	100
210	209	100	100
211	210	100	100
212	211	100	100
213	212	100	100

Tabla D.27: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Infomap utilizando grafos de interacciones intermoleculares.

Comunidad original	**Comunidad variante**	**Similitud nat-mut**	**Similitud mut-nat**
92	92	12,5	33,3
92	94	87,5	100,0
94	95	100,0	100,0
95	96	100,0	100,0
96	97	100,0	100,0
97	98	100,0	100,0
98	99	100,0	100,0
99	100	100,0	100,0
100	101	100,0	100,0
101	102	100,0	100,0
102	103	100,0	100,0
103	104	100,0	100,0
104	105	100,0	100,0
105	106	100,0	100,0
106	107	100,0	100,0
107	92	100,0	66,6

Tabla D.28: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Label propagation utilizando grafos de interacciones intermoleculares.

Comunidad original	**Comunidad variante**	**Similitud nat-mut**	**Similitud mut-nat**
93	93	20,0	33,3
93	95	80,0	100,0
95	96	100,0	100,0
96	97	100,0	100,0
97	98	100,0	100,0
98	99	100,0	100,0
99	100	100,0	100,0
100	101	100,0	100,0
101	102	100,0	100,0
102	103	100,0	100,0
103	104	100,0	100,0
104	105	100,0	100,0
105	106	100,0	100,0
106	107	100,0	100,0
107	108	100,0	100,0
108	109	100,0	100,0
109	93	100,0	66,6

Tabla D.29: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Leading eigenvector utilizando grafos de interacciones intermoleculares.

Comunidad original	**Comunidad variante**	**Similitud nat-mut**	**Similitud mut-nat**
81	81	95,2	100,0
81	209	4,8	33,3
100	209	100,0	66,6
101	100	100,0	100,0
102	101	100,0	100,0
103	102	100,0	100,0
104	103	100,0	100,0
105	104	100,0	100,0
106	105	100,0	100,0
107	106	100,0	100,0
108	107	100,0	100,0
109	108	100,0	100,0
110	109	100,0	100,0
111	110	100,0	100,0
112	111	100,0	100,0
113	112	100,0	100,0
114	113	100,0	100,0
115	114	100,0	100,0
116	115	100,0	100,0
117	116	100,0	100,0
118	117	100,0	100,0
119	118	100,0	100,0
120	119	100,0	100,0
121	120	100,0	100,0
122	121	100,0	100,0
123	122	100,0	100,0
124	123	100,0	100,0
125	124	100,0	100,0
126	125	100,0	100,0
127	126	100,0	100,0
128	127	100,0	100,0
129	128	100,0	100,0
130	129	100,0	100,0
131	130	100,0	100,0
132	131	100,0	100,0
133	132	100,0	100,0
134	133	100,0	100,0
135	134	100,0	100,0

Tabla D.29: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Leading eigenvector utilizando grafos de interacciones intermoleculares (cont.).

Comunidad original	**Comunidad variante**	**Similitud nat-mut**	**Similitud mut-nat**
136	135	100,0	100,0
137	136	100,0	100,0
138	137	100,0	100,0
139	138	100,0	100,0
140	139	100,0	100,0
141	140	100,0	100,0
142	141	100,0	100,0
143	142	100,0	100,0
144	143	100,0	100,0
145	144	100,0	100,0
146	145	100,0	100,0
147	146	100,0	100,0
148	147	100,0	100,0
149	148	100,0	100,0
150	149	100,0	100,0
151	150	100,0	100,0
152	151	100,0	100,0
153	152	100,0	100,0
154	153	100,0	100,0
155	154	100,0	100,0
156	155	100,0	100,0
157	156	100,0	100,0
158	157	100,0	100,0
159	158	100,0	100,0
160	159	100,0	100,0
161	160	100,0	100,0
162	161	100,0	100,0
163	162	100,0	100,0
164	163	100,0	100,0
165	164	100,0	100,0
166	165	100,0	100,0
167	166	100,0	100,0
168	167	100,0	100,0
169	168	100,0	100,0
170	169	100,0	100,0
171	170	100,0	100,0
172	171	100,0	100,0
173	172	100,0	100,0
174	173	100,0	100,0

Tabla D.29: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo Leading eigenvector utilizando grafos de interacciones intermoleculares (cont.).

Comunidad original	**Comunidad variante**	**Similitud nat-mut**	**Similitud mut-nat**
175	174	100,0	100,0
176	175	100,0	100,0
177	176	100,0	100,0
178	177	100,0	100,0
179	178	100,0	100,0
180	179	100,0	100,0
181	180	100,0	100,0
182	181	100,0	100,0
183	182	100,0	100,0
184	183	100,0	100,0
185	184	100,0	100,0
186	185	100,0	100,0
187	186	100,0	100,0
188	187	100,0	100,0
189	188	100,0	100,0
190	189	100,0	100,0
191	190	100,0	100,0
192	191	100,0	100,0
193	192	100,0	100,0
194	193	100,0	100,0
195	194	100,0	100,0
196	195	100,0	100,0
197	196	100,0	100,0
198	197	100,0	100,0
199	198	100,0	100,0
200	199	100,0	100,0
201	200	100,0	100,0
202	201	100,0	100,0
203	202	100,0	100,0
204	203	100,0	100,0
205	204	100,0	100,0
206	205	100,0	100,0
207	206	100,0	100,0
208	207	100,0	100,0
209	208	100,0	100,0
210	211	100,0	100,0
211	210	100,0	100,0

Tabla D.30: Porcentaje de similitud entre comunidades de la proteína nativa y la variante C141Y para el algoritmo multilevel utilizando grafos de interacciones intermoleculares.

Comunidad original	**Comunidad variante**	**Similitud nat-mut**	**Similitud mut-nat**
82	82	73,3	84,6
82	101	26,7	30,8
90	82	10,5	15,4
90	90	42,1	80,0
90	101	47,4	69,2
101	102	100,0	100,0
102	103	100,0	100,0
103	90	100,0	20,0

Anexo E

Representación visual de detección de comunidades para grafos de interacción intermolecular

En la Figura E.1 se muestra la detección de comunidades de grafos de interacción intermolecular de las variantes de interés sin colorear gris los módulos de tamaño uno.

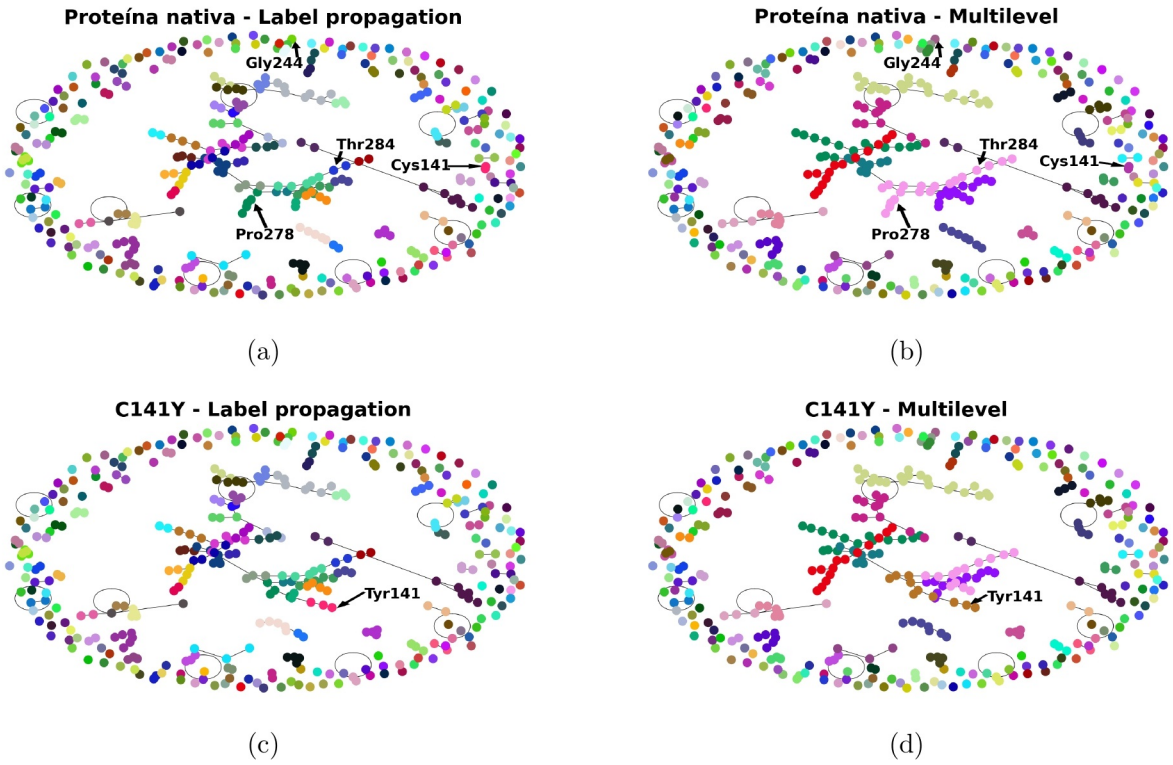
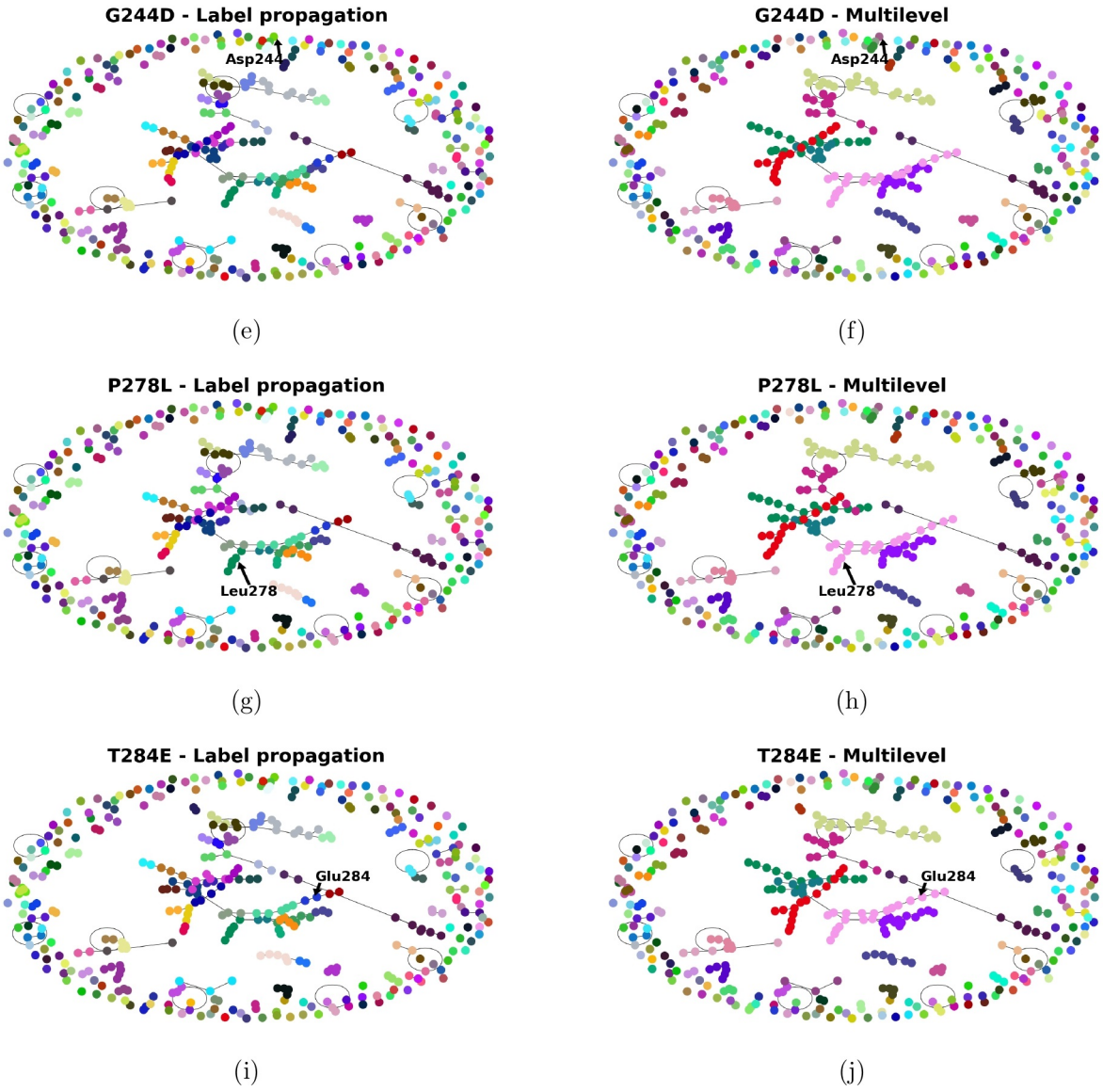


Figura E.1: Comparación de detección de comunidades para la proteína p53 humana, representada con grafos de interacciones intermoleculares. Utilizando el algoritmo label propagation en (a) proteína nativa (c) C141Y y multilevel en (b) proteína nativa (d) C141Y.



(e) (f) (g) (h) (i) (j)

Figura E.1: Comparación de detección de comunidades para la proteína p53 humana, representada con grafos de interacciones intermoleculares (cont.). Utilizando el algoritmo label propagation en (e) G244D (g) P278L (i) T284E y multilevel en (f) G244D (h) P278L (j) T284E.