Universidad del Bío-Bío, Chile

Facultad de Ciencias Empresariales

Departamento de Sistemas de Información

Análisis y comparación de algoritmos de Machine Learning para el estudio de la predicción de estructura, enfocado en la interacción entre proteínas.

Proyecto de título presentado por Edgar Araya Contreras
de la Carrera Ingeniería Civil Informática
Dirigida por Tatiana Gutiérrez Bunster

2023

Resumen

Las proteínas son cadenas de moléculas más pequeñas, llamadas aminoácidos, que se conectan de manera lineal en un orden específico. Los aminoácidos tienen distintas propiedades bioquímicas que contribuyen a que la proteína forme una estructura tridimensional, que finalmente especifica que función cumple la proteína.

Existen diversas técnicas experimentales para obtener la estructura tridimensional de una proteína, pero estas técnicas toman bastante tiempo. Debido a esto, nace el campo de predicción de estructuras a través de métodos computacionales.

El objetivo de este proyecto ha sido comparar distintas técnicas del estado del arte, que hacen uso de inteligencia artificial, deep learning, procesamiento del lenguaje natural y mecanismos de atención, para modelar la estructura tridimensional de las proteínas. El foco principal es el modelamiento de complejos proteicos de dos cadenas, estas técnicas son evaluadas en base a la estructura obtenida, comparada con la estructura experimental del Protein Data Bank.

En base a estos resultados, se evalúa si el mejor método es capaz de predecir con precisión la interfaz de interacción entre ambas cadenas del complejo. Posteriormente, se hace uso de una función que permite predecir la calidad de complejos de los que no se tiene una estructura experimental, y se utiliza para evaluar los modelos obtenidos. Se hace uso de esta función para evaluar el dataset elegido para modelar complejos y un dataset de pares proteínas no interactuantes, a modo de analizar la tasa de éxito que tiene la función para diferenciar entre pares de proteínas interactuantes y no interactuantes.

Finalmente, utilizando los cálculos y análisis obtenidos sobre las estructuras, se diseñó una aplicación web almacenada que realiza el cálculo de la función sigmoide sobre un archivo PDB obtenido de algunos de los métodos, en base a las distancias entre los aminoácidos de las cadenas de los complejos, y la puntuación de confianza de posición de los aminoácidos que se encuentran en la interfaz calculada.

Palabras Clave — Proteínas, Inteligencia Artificial, Deep Learning, Atención.

Índice general

Índice de figuras

Índice de Tablas

Capítulo 1

Introducción

Las proteínas son moléculas complejas que juegan roles fundamentales en el funcionamiento de los seres vivos, son uno de los bloques básicos que componen la vida. Están compuestas de unidades más pequeñas llamadas aminoácidos, que se unen entre ellas para formar una cadena lineal. Existen 20 aminoácidos diferentes que se combinan para formar las proteínas, y el orden lineal en el que se combinan determinan la forma tridimensional en la que se conforma. Esta forma tridimensional tiene muchas otras formas a nivel local en distintas regiones, puede tener hélices, pliegues y giros. La importancia de conocer la estructura de las proteínas radica en que la estructura de la proteína dicta la función de ésta, sin embargo, esto se ha hecho históricamente de manera experimental o por homología, lo que tiene sus limitaciones. Para ayudar en esta tarea, el rápido avance de la computación permite que la inteligencia artificial se convierta en una herramienta más a nuestra disposición para resolver problemas y en particular, resolver la predicción de estructuras.

La inteligencia artificial permite a los computadores replicar capacidades de la mente humana, y se puede observar como esta rama de las ciencias computacionales ya nos rodea por completo, ya sea en los dispositivos móviles, el uso de la navegación en Internet, o la planificación de rutas para llegar a un lugar determinado. En la actualidad existen diversas técnicas de inteligencia artificial que son utilizadas para ayudar a resolver el problema de predicción de estructuras de proteínas, y el avance que se ha hecho en esta área es enorme. Es un problema que fue propuesto en 1962, debido a la hipótesis de que la estructura de una proteína en su estado natural es la más termodinámicamente estable, y es dictada por su secuencia de aminoácidos.

Con los avances que han habido y que se van a discutir en este proyecto de título, se puede decir que el problema de predicción de la estructura de una proteína está casi resuelto. Encontrar una manera confiable de predecir las estructuras de las proteínas de manera consistente no sólo puede ayudar a combatir enfermedades, desarrollar medicamentos, y producir nuevas terapias, sino que

también puede ayudar a descifrar los mecanismos sobre los cuales la vida en si funciona.

1.1. Objetivo General

Analizar y comparar técnicas de deep learning para la predicción de estructuras de complejos proteicos, permitiendo identificar que técnicas son más adecuadas para su estudio y posterior utilización en análisis de interacción de proteínas.

1.2. Objetivos específicos

(a) Realizar marco teórico de técnicas de inteligencia artificial, deep learning y machine learning.
(b) Realizar marco teórico de estructuras e interacción de proteínas.
(c) Investigar el estado del arte de deep learning y machine learning en estudio de estructura de proteínas.
(d) Identificar técnicas de machine learning más utilizadas para el análisis y comparación propuesto.
(e) Seleccionar técnicas de machine learning a comparar.
(f) Implementar técnicas seleccionadas, sobre dataset seleccionado y especificar pruebas a realizar.
(g) Analizar los resultados obtenidos y presentar comparación de estas técnicas.

1.3. Fundamento del proyecto

En la actualidad existe una gran cantidad de algoritmos usando biología computacional para predecir las interacciones proteína-proteína y el campo ha avanzado rápidamente en la última década, sin embargo, no todas las interacciones se pueden predecir con alta precisión usando un solo método.

Las interacciones proteína-proteína son intrínsecas en la mayoría de los procesos biológicos, y la identificación de estas interacciones ayudan a dilucidar la función de las proteínas y el rol que desempeñan dentro de la comunicación celular (Westermarck et al., 2013). Dado el papel crítico que desempeñan las interacciones proteína-proteína, muchas enfermedades se deben a irregularidades que se presentan en estas interacciones (Zinzalla y Thurston, 2009), es por esto que el estudio y predicciones de las interacciones proteína-proteína y sus estructuras es muy

importante para la medicina, ya que expande las posibilidades de potenciales tratamientos en el futuro.

1.4. Actividades

- Revisión de la literatura respecto a las interacciones proteína-proteína y el uso de técnicas de inteligencia artificial
- Revisión de la literatura respecto a las técnicas utilizadas para la predicción de estructuras proteicas
- Análisis de las técnicas utilizas en el estado del arte de la predicción de estructuras de proteínas.
- Comparación de las técnicas utilizas en el estado del arte de la predicción de estructuras de proteínas.
- Aplicación de técnicas del estado del arte para la predicción de estructuras con un enfoque en la predicción de interacciones proteína-proteína.
- Comparación de las técnicas aplicadas en el estado del arte para la predicción de estructuras proteicas, en la predicción de interacciones proteína-proteína.

1.5. Detalle del informe

Este proyecto de título se encuentra divido en los siguientes capítulos:

- **Introducción:** Presentación del objetivo del proyecto, justificación y actividades a realizar.
- **Proteínas e Interacciones:** Se explican los conceptos relacionados a proteínas, las reglas biológicas que las rigen, sus clasificaciones y sus interacciones.
- **Inteligencia Artificial y Métodos Computacionales:** Se explican los conceptos relacionados con la inteligencia artificial, los tipos de inteligencias principales que se usan en la biología computacional y los métodos que se utilizan para aumentar la eficacia de las inteligencias.
- **La predicción de estructuras proteicas a través del tiempo:** Se describen las primeras predicciones de estructuras proteicas, las principales inteligencias artificiales basadas en machine learning y las técnicas usadas para predecir estructuras de proteínas.
- **Estudio y Desarrollo de Metodología:** Se presenta el contexto del problema y la metodología a llevar a cabo para responder a los objetivos del proyecto.
- **Resultados:** Se presentan los resultados obtenidos durante el desarrollo del proyecto.
- **Conclusión y Trabajo Futuro:** Conclusión general del proyecto y posibles proyectos futuros en base a los resultados obtenidos en el proyecto.

Capítulo 2

Proteínas e Interacciones

En el universo, la vida es posible gracias a las interacciones de moléculas orgánicas, específicamente el ADN (ácido desoxirribonucleico), el ARN (ácido ribonucleico) y las proteínas. Estas moléculas se comportan de distintas maneras debido a sus estructuras, es decir, como sus átomos se posicionan para formar los bloques tridimensionales que se combinan para producir estructuras más grandes.

En su forma más fundamental, una proteína es una cadena de moléculas más pequeñas, llamadas aminoácidos. Los aminoácidos son los bloques básicos de las proteínas, estas moléculas tienen un átomo de carbono central que se conecta a otros átomos en cuatro lados (Alberts B, 2002). En una dirección siempre hay un átomo de hidrógeno, en otro lado se encuentra el "grupo amino" (un átomo de nitrógeno conectado a dos átomos de hidrógeno), en el tercer lado se encuentra el "grupo carboxilo" (un átomo de carbono, dos de oxígeno y uno de hidrógeno) y en el cuarto lado se encuentra la "cadena lateral"; lo que se encuentra en este lado es lo que hace a cada aminoácido único, en la naturaleza existen 20 aminoácidos que conforman a las proteínas.

Una proteína consiste en una larga cadena de aminoácidos conectados entre sí en un orden específico, como por ejemplo, las letras en una palabra. Los aminoácidos se conectan cuando el grupo carboxilo de uno se une al grupo amino del aminoácido siguiente. Este proceso de unión libera una molécula de agua que lleva a la formación de un enlace peptídico. Al estar ambos aminoácidos unidos mediante un enlace covalente, este nuevo par de aminoácidos se llama dipéptido (Figura 2.1). La unión de tres aminoácidos se llama tripéptido, cuatro aminoácidos tetrapéptidos y así sucesivamente, las proteínas generalmente se consideran polipéptidos.

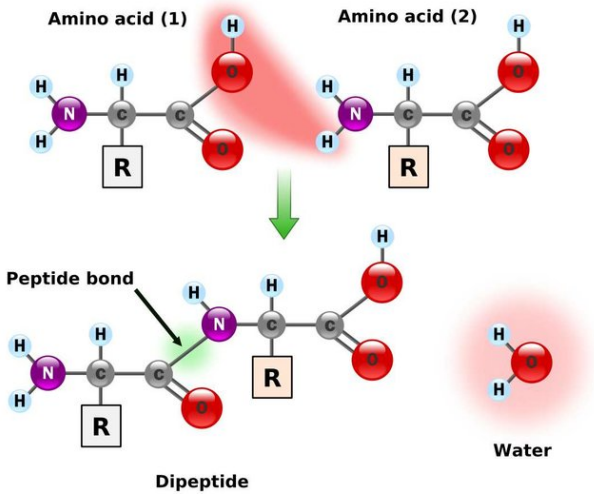


Figura 2.1: Estructura típica de aminoácidos y formación de enlace peptídico.(Bhattacharya, 2022)

La secuencia a lo largo del núcleo de la cadena del polipéptido, donde se encuentran los átomos de carbono y nitrógenos repetidos unidos por enlaces peptídicos, se conoce como la columna vertebral polipeptídica (Alberts B, 2002). Unidos a esta cadena se encuentran las porciones de los aminoácidos que no participan en el enlace peptídico (las cadenas laterales) y son las que dan a cada aminoácido sus propiedades únicas.

En el Anexo A.1 se puede observar la estructura general que tiene una proteína y en el Anexo A.2 se puede ver las propiedades que tienen las cadenas laterales.

2.1. Estructura de las proteínas

Los aminoácidos unidos por los enlaces peptídicos forman una cadena polipeptídica. Una o más de estas cadenas se retuercen en una forma tridimensional que forma una proteína. Estas formas pueden ser muy complejas y contener varios pliegues, hélices, bucles o curvas. Para ser capaces de realizar sus funciones, las proteínas se pliegan en estas distintas formas espaciales gracias a la gran cantidad de interacciones no covalentes, como los enlaces de hidrógeno, enlaces iónicos y fuerzas de Van der Waals (ver Anexo A.4).

2.1.1. Ángulos de Torsión

Los ángulos de torsión son formados por tres enlaces consecutivos en una molécula, y definidos por el ángulo creado entre los dos enlaces externos. La cadena principal de una proteína tiene tres ángulos de torsión diferentes:

- Ángulo *fi* (ϕ): alrededor del enlace N-C alfa
- Ángulo *psi* (ψ): alrededor del enlace C alfa-C
- Ángulo *omega* (ω): alrededor del enlace peptídico entre C y N

El enlace omega tiene un carácter de doble enlace y por lo tanto es casi siempre 180°. La estructura de una proteína es formada principalmente por los ángulos *fi* y *psi*. Los ángulos de torsión se encuentra en un rango específico forzado por la cadena principal debido a los elementos de estructura secundaria, esto se puede visualizar en un diagrama de Ramachandran.

Un diagrama de Ramachandran es un gráfico de dos dimensiones de los ángulos de torsión (ϕ) y (ψ) de una secuencia proteica (Coumar, 2021).

2.1.2. Estructura Primaria

La estructura primaria corresponde a la cadena de aminoácidos y la secuencia en que se encuentran conectados a través de enlaces peptídicos para formar una cadena polipeptídica. Los extremos de la cadena polipeptídica se conocen como N-terminal o C-terminal y los 20 aminoácidos diferentes pueden ser utilizados múltiples veces en la misma cadena peptídica para estructuras primarias especificas de una proteína. Para describir la secuencia de las proteínas se usan códigos de letras, en orden comenzando desde el grupo amino (N-terminal) hasta el grupo carboxilo (C-terminal), este puede ser un código de tres o una letra, como se puede observar en la Figura A.2.

Como ejemplo, en la Figura 2.2 se pueden observar en código las primeras 5 posiciones de la cadena A de la insulina humana, es decir, comenzando desde el extremo izquierdo, en N-terminal se tiene G (Glicina) seguido por I (Isoleucina), V (Valina), E (Ácido Glutámico) y Q (Glutamina).

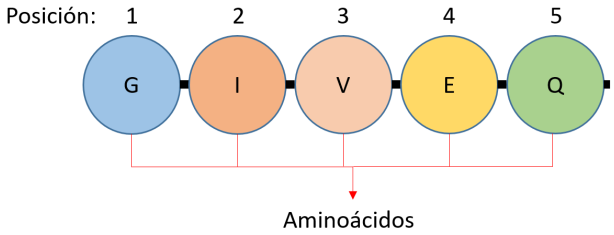


Figura 2.2: Estructura primaria, primeras 5 posiciones de la secuencia proteica de la cadena A de la insulina humana (código PDB:3I40)

2.1.3. Estructura Secundaria

La estructura secundaria corresponde a las estructuras de pliegues que se forman localmente en la cadena polipeptídica debido a interacciones entre los átomos de la cadena principal. Los enlaces de hidrógeno entre los grupos amino y carboxilo en regiones adyacentes de la cadena de proteína hacen que ocurran patrones de pliegue. Esto usualmente produce dos formas principales, una espiral llamada hélice alfa, y láminas beta, y dos formas secundarias, los giros beta y los bucles omega. Las láminas beta generalmente se encuentran cercanas unas a otras y forman láminas de mayor tamaño. En el Anexo A.3 se describen las formas y propiedades de las laminas beta, hélice alfa, giros beta y bucles omega.

2.1.4. Estructura Terciaria

La estructura terciaria corresponde a la forma tridimensional de la proteína, esta estructura surge cuando las estructuras secundarias se atraen entre sí, causando pliegues y formando módulos más grandes llamados “dominios”. Esto resulta en aminoácidos que quedan al interior de la estructura y otros que quedan en el exterior, donde pueden interactuar con otros dominios o moléculas. Los dominios son regiones en la cadena polipeptídica de las proteínas que es capaz de estabilizarse por si sola y se puede plegar independiente del resto. Cada dominio se compacta en un plegamiento tridimensional y las proteínas pueden contener varios dominios.

En este nivel de jerarquía estructural es donde se consideran las diferentes interacciones en la proteína para mantener estable la estructura, para formar estructuras estables deben haber más interacciones atractivas que interacciones repulsivas, y la estabilidad depende de la frecuencia y fuerza de estas interacciones. En el Anexo A.4 se describen las interacciones que dan forma a la estructura tridimensional de las proteínas.

2.1.5. Estructura Cuaternaria

La estructura cuaternaria es la interacción de dos o más cadenas polipeptídicas plegadas y a cada cadena se le llama subunidad. Si la proteína resultante esta compuesta de dos subunidades, se llama dímero (ver ejemplo en Figura 2.3), si contiene tres subunidades trímero; 4 subunidades forman un tetrámero (ver ejemplo en Figura 2.4) y así sucesivamente. Si las subunidades son idénticas, se usa el prefijo "homo" y se denominan homodímeros. Si las subunidades son diferentes, se usa el prefijo "hetero" y se denominan heterodímero (Godbey, 2021).

Figura 2.3: Proteína dimérica Bence-Jones (código PDB=1REI), cadena izquierda corresponde a la cadena A y la cadena derecha corresponde a la cadena B.

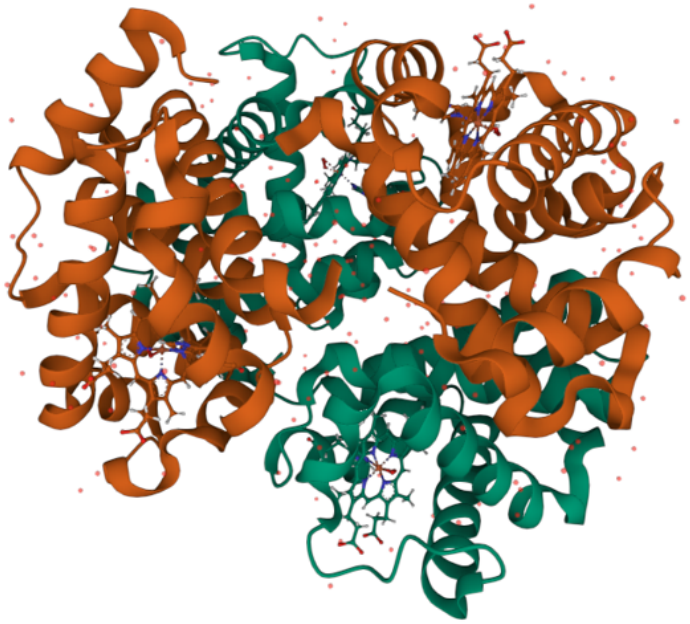


Figura 2.4: La hemoglobina humana (código PDB=6BB5) es una proteína tetramérica, donde cada subunidad es capaz de almacenar y enlazar una molécula de oxigeno.

La asociación de dos cadenas polipeptídicas se basan en el principio de la complementariedad de la forma geométrica y en las interacciones enlazantes. Una gran cantidad de interacciones no covalentes mantienen a las subunidades juntas.

La complementariedad de las interacciones enlazantes requieren que los donantes de enlaces de hidrógeno, grupos no polares, y cargas positivas sean opuestas a receptores de enlaces de hidrógeno, grupos no polares y cargas negativas. Una unión estable entre las cadenas se obtiene si el número de interacciones débiles es maximizada por la complementariedad geométrica de las superficies interactuantes (Pal, 2020).

2.2. Plegamiento

En orden para que las proteínas sean capaces de desempeñar sus funciones, estas deben adoptar una forma tridimensional específica, llamada estado nativo (Pal, 2020). El proceso de plegamiento es el proceso que lleva a una cadena polipeptídica desde su secuencia lineal de aminoácidos a una estructura espacial definida característica del estado nativo de la proteína.

Christian Anfinsen llevó a cabo experimentos en 1950 (Anfinsen, 1973), que permitieron estable-

cer que la secuencia de una proteína determina su estructura nativa, y de aquí surge la necesidad de resolver el problema del plegamiento de las proteínas. Existen tres problemas diferentes pero interrelacionados con el plegamiento de las proteínas:

- El "código de plegamiento" que indicaría que una particular combinación de fuerzas interatómicas dicta la estructura tridimensional de las proteínas.
- La **"predicción de estructura"** computacional de las proteínas a partir de su secuencia de aminoácidos.
- La cinética y termodinámica asociada al rápido "proceso de plegamiento".

Las interacciones responsables del plegamiento de las proteínas corresponden a las interacciones iónicas, los enlaces de hidrógeno, las fuerzas de van der Waals y las interacciones hidrofóbicas. Estas se describen en detalle en el Anexo A.5.

Entre todas estas interacciones, aún no hay una conclusión absoluta de que interacción es la dominante. En análisis experimentales y computacionales realizados (Ferenczy y Kellermayer, 2022) se ha encontrado que los enlaces de hidrógeno tanto intraproteicos como entre proteínas y moléculas de agua contribuyen de una manera al menos tanto como el efecto hidrofóbico. Además, cuando se consideran los grupos hidrofílicos e hidrofóbicos como sitios de interacciones, las fuerzas en los grupos hidrofílicos fueron considerablemente más fuertes que los grupos hidrofóbicos. Si bien no se sabe con claridad cuál de todas estas interacciones es la predominante en el proceso de plegamiento, se puede inferir que tanto los grupos hidrofílicos e hidrofóbicos desempeñan roles complementarios en acelerar el proceso de plegamiento.

2.3. Interacciones proteína-proteína

Las interacciones proteína-proteína (IPP) son mediadores centrales en todos los procesos biológicos, la mayoría de las interacciones son controladas por el arreglo tridimensional y la dinámica de las proteínas interactuantes. Estas interacciones pueden ser permanentes o temporales, algunas interacciones proteína-proteína son específicas a un par de proteínas, mientras que algunas proteínas otras pueden interactuar con muchas otras. Esta complejidad existente en las interacciones es un desafío tanto para métodos experimentales y computacionales. Estas IPP pueden ser clasificadas según (Acuner Ozbabacan et al., 2011):

- **Composición:** Si los complejos son homo-oligoméricos o hetero-oligoméricos.
- **Afinidad:** Si los complejos son obligados o no-obligados.
- **Vida útil o estabilidad:** Si los complejos son permanentes o transitorios.
- **Pliegue:** Si los complejos son dominio-dominio o dominio-péptido

A continuación se verán los tipos de IPP, en la Figura 2.5 se puede observar como se pueden clasificar las IPP según su afinidad y vida útil, la mayoría de las interacciones no-obligadas son transitorias pero existen algunas que son permanentes, como las interacciones entre enzimas e inhibidores.

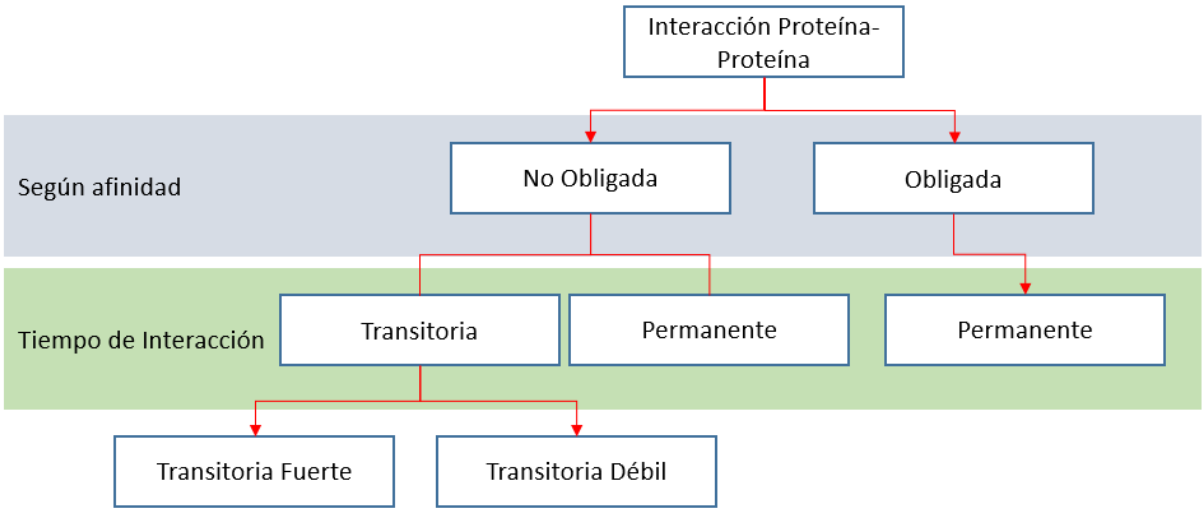


Figura 2.5: Relación de los tipos de interacciones proteína-proteína según su afinidad y su vida útil (Acuner Ozbabacan et al., 2011)

.

2.3.1. Complejos homo-oligoméricos y hetero-oligoméricos

Los complejos homo-oligoméricos y hetero-oligoméricos son diferenciados por la composición de las proteínas interactuantes, específicamente entre las cadenas, si las cadenas de las proteínas son idénticas, se forma un complejo homo-oligoméricos, y si las cadenas no son idénticas se forma un complejo hetero-oligomérico.

- Los homo-oligomeros son simétricos y son un buen bloque para formar macromoléculas estables.
- Los hetero-oligomeros pueden variar en su estructura y sirven como base para reunir otras proteínas que cooperan en una sola macromolécula.

2.3.2. Complejos obligados y no-obligados

Los constituyentes de un complejo pueden ser protómeros o monómeros, los protómeros son la unidad estructural más pequeña de las proteínas oligoméricas, mientras que los monómeros son una sola molécula, por ejemplo, un único aminoácido. Si los constituyentes del complejo son inestables por si solos in vivo (que ocurren en un organismo) entonces esta es una interacción obligada, mientras que los componentes de las interacciones no-obligadas pueden existir de manera independiente.

2.3.3. Complejos permanentes y transitorios

Este tipo de interacción se basa en la vida útil del complejo, las interacciones permanentes son generalmente muy estables e irreversibles mientras que las interacciones transitorias se asocian y disocian in vivo. Las interacciones no-obligadas son predominantemente transitorias, sin embargo, igual existen algunas interacciones permanentes. En la literatura los términos permanente y obligado se usan como intercambiables. Las interacciones transitorias se pueden separar entre débiles y fuertes según la estabilidad de equilibrio oligomérico. Las transitorias fuertes desplazan el equilibrio de asociación y disociación sólo bajo ciertas perturbaciones, como por ejemplo, la unión con un ligando (sustancias que se unen a biomoléculas) (Perkins et al., 2010). Las transitorias débiles se forman y quiebran continuamente.

2.3.4. Complejos dominio-dominio y dominio-péptido

Según los pliegues de las proteínas interactuantes, estas se pueden clasificar como dominio-dominio o dominio-péptido. Los complejos formados por dominio-péptido tiene una naturaleza transitoria a medida que se forman por el reconocimiento de una estructura globular, un motivo proteico linear corto y una pequeña interfaz en el que ocurre la interacción. Las interacciones dominio-péptido también son llamadas interacciones transitorias mediada por péptidos. Las interacciones dominio-dominio reconocen y unen motivos proteicos específicos de péptidos, ya sea en los terminales o en regiones desordenadas de las proteínas (Acuner Ozbabacan et al., 2011). Estos dominios ya se encuentran listos para unirse ya que no sufren grandes cambios conformacionales en el proceso de unión. Estos dominios ensamblan las proteínas constituyentes en grandes complejos, juntando las diferentes combinaciones de dominios catalíticos con dominios regulatorios.

2.4. Docking

El docking, o docking molecular, es el estudio de como dos más estructuras moleculares (como una proteína) encajan juntas. Es una técnica de modelamiento usada para predecir como una proteína interactúa con ligandos (pequeñas moléculas) (Roy et al., 2015), o como interactúa con otras proteínas (docking proteína-proteína).

El docking proteína-proteína es la predicción de un complejo proteico (compuesto de dos o más proteínas), dadas las estructuras individuales de las proteínas (Vakser, 2014). La habilidad de una proteína para interactuar con moléculas pequeñas para formar un complejo más grande juega un rol importante en las dinámicas de una proteína. El docking se usa para identificar las posiciones correctas de ligandos en la unión con una proteína y para predecir la afinidad de un ligando y la proteína. Según el tipo de ligando, el docking se puede clasificar como:

- Proteína-ligando
- Proteína-ácido nucleico
- Proteína-proteína

El docking proteína-ligando es el más simple en cuanto a la complejidad, el docking proteína-proteína es típicamente mucho más complejo, debido a que las proteínas son flexibles y su espacio conformacional es mucho más grande.

Capítulo 3

Inteligencia Artificial y Métodos Computacionales

En este capítulo se explicarán los conceptos principales relacionados las diferentes técnicas de inteligencia artificial y métodos computacionales utilizadas en la predicción de estructuras.

3.1. Inteligencia Artificial

La inteligencia artificial es la habilidad que poseen los computadores modernos o robots controlados por computadores para realizar tareas comúnmente asociadas a seres inteligentes. Es un campo de estudio dentro de las ciencias de la computación que, generalmente, se asocia al desarrollo de sistemas con capacidades características de los seres humanos como la habilidad de razonar, buscar significado, generalizar, o aprender de la experiencia; esto no significa que la inteligencia artificial esté solamente limitada a emular los procesos de la mente humana, sino que abarca los procesos racionales entre un agente racional (Russell y Norvig, 2009) y su entorno.

Desde los inicios de la computación digital, los computadores han sido programados para realizar tareas cada vez más complejas, ya sea desde las primeras calculadoras digitales, o programas capaces de demostrar teoremas matemáticos. En la actualidad los usos de la inteligencia artificial se encuentran en casi todas las áreas de la sociedad moderna, como en los buscadores que se usan para navegar en la web, el reconocimiento del lenguaje hablado y escrito, el reconocimiento facial, o la capacidad de competir en juegos de estrategia como Ajedrez y Go al más alto nivel, superando incluso a las personas más adeptas en estos. Es por esto que no es una sorpresa que la inteligencia artificial también sea una herramienta importante para el área de la medicina y biología computacional, acelerando procesos de adquisición y evaluación de información de

manera considerable.

La necesidad de lograr que los computadores puedan aprender y adquirir conocimiento dio inicio a la búsqueda de distintos enfoques para lograr este objetivo, por mucho tiempo se creía que este nivel de inteligencia se podría lograr programando una gran cantidad de reglas explicitas para manipular grandes volúmenes de conocimientos en bases de datos explicitas. Este enfoque se conoce como inteligencia artificial simbólica (Chollet, 2017). Si bien este enfoque es eficiente para resolver problemas lógicos bien definidos, como Go (Silver et al., 2016) o el Ajedrez, resulta muy difícil definir reglas explicitas para resolver problemas más complejos y difusos, como la clasificación de imágenes, traducción de lenguajes, reconocimiento de voz, y la predicción de estructuras de proteínas. De esta dificultad surge un nuevo enfoque, llamado machine learning (aprendizaje automático).

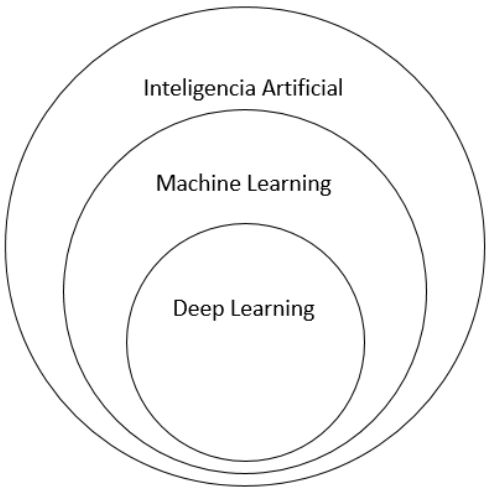


Figura 3.1: Deep learning es un subconjunto de machine learning y éste es un subconjunto del campo de Inteligencia Artificial.

3.1.1. Machine Learning

Machine learning es una aplicación de la inteligencia artificial que permite a los sistemas computacionales aprender y mejorar a través de la experiencia, sin tener la necesidad de ser programado en base a reglas explicitas (ver Figura 3.2). El campo de machine learning está dedicado al entendimiento y desarrollo de métodos capaces de aprendizaje sobre grandes volúmenes de datos y usar este aprendizaje para mejorar el desempeño de esta inteligencia artificial en tareas específicas. Un sistema basado en machine learning es **entrenado**, esto significa que al sistema se le presentan muchos ejemplos, llamados datos de entrenamiento, relevantes para la tarea a cumplir, y encuentra una estructura estadística con estos ejemplos que eventualmente permiten al sistema generar las reglas para la automatización de la tarea.

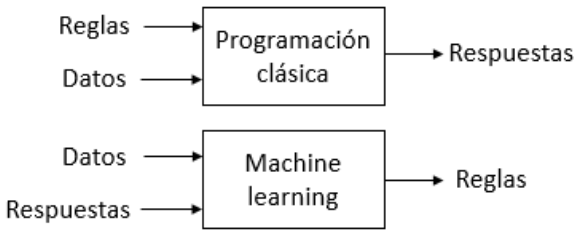


Figura 3.2: Diferencias entre programación clásica y machine learning. (Chollet, 2017)

Los datos de entrenamiento están diseñados para enseñarle al sistema, debe tener la información necesaria asociada para que el sistema pueda formar las correlaciones, por ejemplo, si se quiere entrenar un sistema para reconocer imágenes de gatos y perros automáticamente, las imágenes deben estar etiquetadas con el animal correspondiente y de manera correcta, de esta forma el sistema es capaz de formar reglas estadísticas y crear un modelo para asociar imágenes especificas con las etiquetas correspondientes. Para poder generar un modelo de machine learning, los algoritmos necesitan tres parámetros (Chollet, 2017):

- Datos de entrada.
- Ejemplos de los resultados esperados.
- Una manera para medir si el algoritmo está haciendo un buen trabajo.

Un modelo de machine learning transforma los datos de entrada en salidas significativas, este proceso de transformación es aprendido por la exposición a los ejemplos de datos de entrada y datos de salida. Esencialmente, lo que se debe resolver con machine learning es obtener información útil de los datos de entrada para transformar de manera significativa los datos y obtener una representación que se acerque al resultado esperado. Los enfoques de aprendizaje en machine learning (y en inteligencia artificial) se pueden dividir en tres categorías según la retroalimentación que acompaña a los datos de entrada (Russell y Norvig, 2009):

- **Aprendizaje supervisado:** Al sistema se le presentan datos de entrada junto a las salidas correspondientes y aprende una función para mapear el dato de entrada al dato de salida.
- **Aprendizaje no supervisado:** El sistema aprende patrones a partir de los datos de entradas sin retroalimentación directa, no existe etiquetado de los datos de salida.
- **Aprendizaje reforzado:** El sistema aprende en base a refuerzos: recompensas y castigos, en este caso el sistema debe decidir qué acciones tomar en base al reforzamiento obtenido, y maximizar las recompensas.

Con estos conceptos de machine learning se puede explicar en que consiste deep learning (aprendizaje profundo), que es una subcategoría de machine learning.

3.1.2. Deep Learning

Deep learning (aprendizaje profundo) es un subcampo de machine learning: es una nueva forma de aprender representaciones a partir de datos en el que el énfasis se encuentra en capas sucesivas con representaciones sucesivamente más significativas (Chollet, 2017). El deep learning moderno involucra comúnmente cientos de capas sucesivas de representaciones, y estas son aprendidas automáticamente a través de exposición a datos de entrenamiento. Mientras que en machine learning, otras técnicas tienden a enfocarse sólo en una o dos capas de representaciones de los datos; es por esto que también se les llama shallow learning (aprendizaje superficial).

En deep learning, las capas son casi siempre aprendidas en base a modelos llamados redes neuronales (o redes neuronales artificiales), compuestas de capas unas sobre la otra. El término de red neuronal es una referencia a la neurociencia, en parte inspirado por el modelo que se tiene de las redes de conexiones de neuronas en el cerebro, sin embargo, estas no son un modelo real del cerebro. Las redes neuronales están compuestas de capas de neuronas, contienen una capa de datos de entrada (input layer), una o más capas ocultas (hidden layer), y una capa de salida (output layer).

Las neuronas, también llamadas perceptrones o nodos, son los elementos básicos de una red neuronal, inspiradas por las neuronas biológicas que se encuentran en el cerebro. Estas pueden tener una o más entradas que son sumadas para producir la salida de la neurona. Cada entrada tiene un valor llamado peso (un parámetro que indica cuanta fuerza tiene este valor sobre el valor recibido), el núcleo de la neurona ejecuta cálculos en base a la suma de los valores multiplicados por sus pesos (función lineal), y este resultado pasa por la función (llamada función de activación) que produce la salida neuronal (ver Figura 3.3). La salida debe ser mayor a un umbral especifico de la neurona, si supera este umbral, la neurona se dice que esta activa.

La función de activación puede ser de muchos tipos, generalmente el propósito de esta es para introducir no linealidad en la red neuronal, es decir, sin una función de activación, la red neuronal sólo podría modelar relaciones lineales y no podría modelar relaciones no lineales presentes en los datos, que corresponde a la mayoría de los datos del mundo real.

La especificación de lo que hace cada capa a sus datos de entrada se almacena en los pesos de la capa, las transformaciones implementadas por una capa son parametrizadas por sus pesos (también llamados parámetros de capa). En este contexto, el aprendizaje significa encontrar un set de valores para los pesos de todas las capas de la red neuronal, de manera que tal que la red relacionará los datos de entrada a sus objetivos asociados (Chollet, 2017). Para lograr esto, existe una función llamada función costo (loss function o objective function), cuyo objetivo es realizar predicciones de la red y la verdadera salida de esta, y computar una puntuación de costo, para evaluar que tan bien la red lo ha hecho en este ejemplo en específico.

El resultado de esta función se usa como retroalimentación para efectuar leves ajustes en los

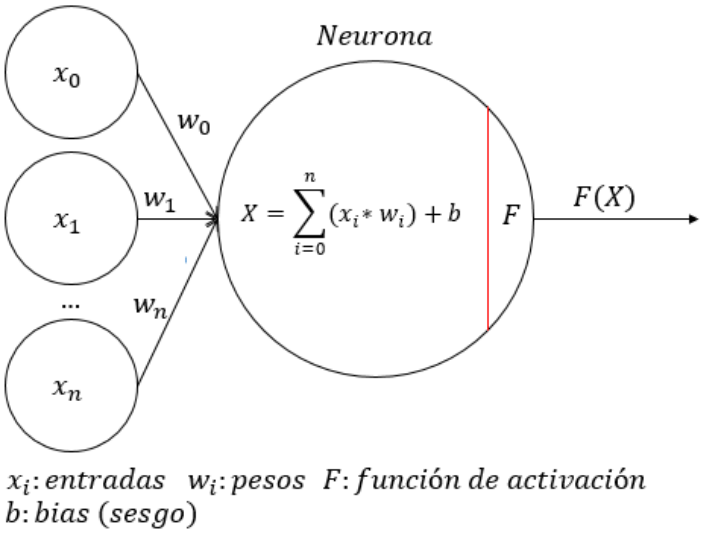


Figura 3.3: Neurona con múltiples entradas, en el núcleo se efectúan las operaciones lineales y se aplica la función de activación, que corresponde a la salida F(X).

pesos de las capas, en la dirección que reduzca el valor de la puntuación de pérdida para el ejemplo actual. El ajuste de los parámetros es realizado por el optimizador (ver Figura 3.4), y este implementa uno de los algoritmos fundamentales de deep learning: el algoritmo backpropagation.

El algoritmo de backpropagation funciona usando técnicas de cálculo multivariable, aplicando la regla de la cadena a cada computación de los valores del gradiente de una red neuronal, esencialmente permite el ajuste de los pesos de las capas de la red neuronal basados en las tasas de errores obtenidos en una iteración previa. El correcto ajuste de los pesos permite reducir la tasa de errores y hacer que el modelo sea más confiable al aumentar su generalización.

Inicialmente, los pesos de una red neuronal tienen asignados valores aleatorios, por lo que la red simplemente implementa una serie de transformaciones al azar. La salida se encuentra lejos de lo que debería ser y la puntuación de pérdida es alta, pero con cada ejemplo que la red neuronal procesa, los pesos son ajustados levemente en la dirección correcta, y la puntuación de costo disminuye. Si esto se repite el número suficiente de veces, los ajustes a los pesos minimizan la función de costo, una red con pérdidas mínimas es tal que sus salidas se encuentran lo más cercanas posibles a los objetivos, es decir, es una red neuronal entrenada. Las redes neuronales pueden clasificarse en distintas categorías:

- **Red neuronal feedforward (prealimentada):** Son redes compuestas por una capa de entrada, una capa de capas ocultas y una capa de salida. Los datos se ingresan en estos modelos para entrenarlos, y la información es de una sola dirección, es decir, viaja desde la capa de entrada hasta la capa de salida. Estas redes son la fundación de la visión

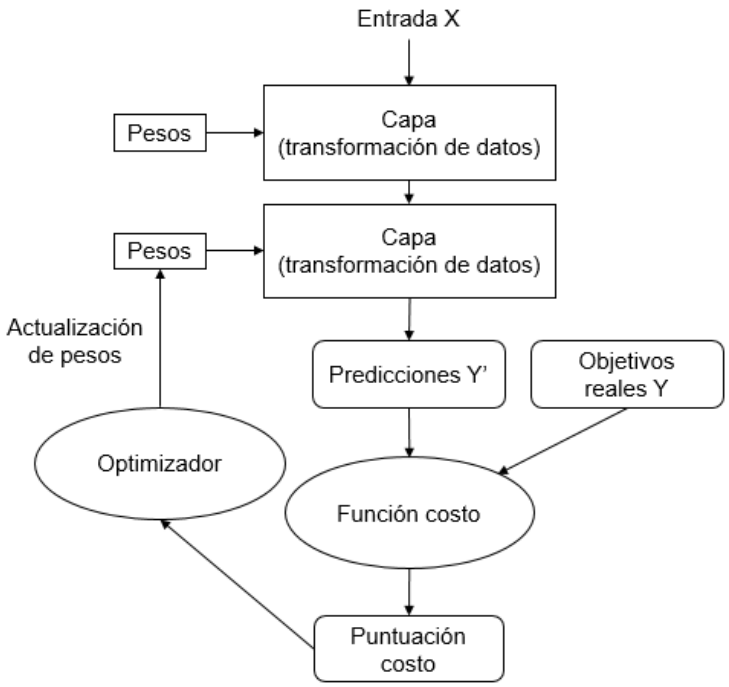


Figura 3.4: La puntuación obtenida por la función costo se usa como señal de retroalimentación para ajustar los pesos (Chollet, 2017)

.

computacional, el procesamiento del lenguaje y otros tipos de redes neuronales.

- **Redes neuronales convolucionales (convolutional neural network, CNN):** Tienen una arquitectura similar a las redes feedforward, con una capa de entrada, oculta y de salida. Estas redes emplean una operación matemática llamada convolución, las convoluciones son una operación linear especializada. Las redes convolucionales son simplemente redes neuronales que usan convolución en vez de multiplicación de matrices en al menos una de sus capas (Goodfellow et al., 2016). Generalmente se utilizan para reconocimiento de imágenes, visión de computadores y reconocimiento de patrones.
- **Redes neuronales recurrentes (recurrent neural network, RNN):** Estas redes se distinguen de las redes feedforward en que se permiten ciclos dentro de la red neuronal, es decir, los nodos pueden estar conectados no necesariamente en una sola dirección. Generalmente, cada ciclo tiene una demora para que los nodos puedan tomar como entrada un valor computado de su misma salida en un paso previo de la computación (sin una demora en el ciclo, el estado del circuito cíclico puede llegar a un estado inconsistente). Esto les permite a las redes neuronales recurrentes tener un estado interno, o memoria: las entradas recibidas en pasos temporales previos puede afectar la respuesta de la red a la entrada actual (Russell y Norvig, 2009). Este tipo de red puede ser usado cuando se tienen datos secuenciales dependientes del tiempo.

3.1.2.1. Atención

El concepto de atención es relativamente nuevo en el campo de deep learning, está inspirado en como los seres humanos tienden a enfocarse en partes distintivas al procesar grandes cantidades de información, y rápidamente se ha vuelto una de las técnicas más importantes en las nuevas redes neuronales del estado del arte en distintos ámbitos.

La atención es un proceso cognitivo complejo, que permite a los seres humanos concentrarse selectivamente en ciertas partes de la información, pero ignorar otra información que está siendo percibida en el mismo tiempo. Esta es una forma de seleccionar rápidamente información de alta importancia desde una cantidad masiva de información, usando recursos de procesamiento limitado. El mecanismo de atención mejora en gran medida la eficiencia y precisión del procesamiento de información perceptual (Niu et al., 2021).

Existen dos tipos de mecanismos de atención relevantes para deep learning, la atención inconsciente y la atención focalizada.

- La atención inconsciente es controlada por estímulos externos y no tiene un propósito definido.
- La atención focalizada se refiere a la atención que tiene un propósito determinado y depende de tareas específicas. Permite a los seres humanos enfocar la atención en un cierto objeto activa y conscientemente, la mayoría de los mecanismos de atención en deep learning están diseñados de acuerdo a tareas específicas para poder utilizar atención focalizada.

El mecanismo de atención sirve a la vez como un proceso para asignar recursos y resolver el problema de la sobrecarga de información. Además de proveer mejoras de desempeño, la atención también puede ser usada como herramienta para explicar el comportamiento anormal en una arquitectura neuronal. La atención es uno de los componentes innovadores que utilizan los algoritmos del estado del arte de la predicción de estructuras de proteínas.

3.2. Alineamiento de secuencias

El alineamiento de secuencias es un procedimiento que consiste en comparar dos o más secuencias de ADN, ARN o proteínas en búsqueda de patrones para identificar similitudes que pueden surgir en base a una relación estructural, funcional o evolutiva entre ellas (Mount, 2004). Las secuencias que son muy similares probablemente tienen la misma función, en el caso de las proteínas esto indica una estructura tridimensional y función bioquímica similar. Cabe mencionar que si bien las secuencias de aminoácidos tienden a cambiar con el tiempo debido a mutaciones, la estructura se conserva a través del tiempo (Illergård et al., 2009). Adicionalmente, si dos secuencias de

organismos diferentes son similares, esto puede significar que debe haber existido una secuencia ancestral común, y en este caso a las secuencias se les llama homólogas.

Hay dos tipos de alineamientos de secuencias según su alcance, global y local (ver Figura 3.5):

- **Global:** Se intenta alinear la secuencia en su totalidad, usando la mayor cantidad de residuos posibles, hasta ambos extremos de cada secuencia (ver Figura 3.5.a). Las secuencias que son similares y aproximadamente del mismo tamaño son candidatas para alineamiento global.
- **Local:** Se intenta alinear intervalos de secuencias con la mayor densidad de residuos que concuerdan, de manera de generar uno o más tramos de coincidencias (ver Figura 3.5.b). Estos alineamientos son más adecuados para alinear secuencias que son similares en tramos en alguna parte de su largo, pero disimilares en otros, en secuencias que tienen distinto tamaño, o en secuencias que comparten una región o dominio.

En la Figura 3.5 se puede ver la distinción entre el alineamiento local y global de dos secuencias, generalmente se usa como notación barras verticales para indicar similitudes (aminoácidos idénticos) y guiones o barras horizontales para indicar diferencias (Mount, 2004).

Figura 3.5: Distinción entre alineamiento local y global de dos secuencias.

Según el número de secuencias a alinear, se pueden identificar dos categorías, el alineamiento de secuencias múltiples (tres o más secuencias) y alineamiento de pareja (dos secuencias).

3.2.1. Alineamiento de pareja (PSA)

El alineamiento de pareja o par (Pairwise Alignment) consiste en el alineamiento de sólo dos secuencias. Existen tres métodos principales para identificar alineamientos de pareja:

- **Matriz de puntos:** Un análisis de matriz de puntos es un método principalmente para buscar posibles alineamientos entre residuos de secuencias. Para dos secuencias de largos M y N, se crea una matriz de dimensiones M x N, con una secuencia horizontal en la parte superior, de izquierda a derecha, y una secuencia vertical en el lado izquierdo, de arriba hacia abajo. Para cada posición de la matriz, se compara el residuo de la parte superior con el residuo de la parte izquierda. Sí y sólo si los residuos son idénticos, se ubica un punto en esa posición, y esto se repite hasta que se encuentran todos los residuos similares. Los gráficos de puntos generados por dos secuencias muy similares formarán una línea en la diagonal principal de la matriz. La ventaja principal de este método es exponer todas las coincidencias de residuos entre dos secuencias, y queda al criterio del investigador identificar las más significativas.
- **Programación dinámica:** Este es un método computacional en el que se compara cada par de residuos en las dos secuencias y se genera un alineamiento. Este alineamiento incluye los residuos que coinciden y no coinciden, y los espacios o vacíos entre las dos secuencias que son posicionadas de manera que el número de coincidencias entre residuos iguales sea el máximo posible. Este método ha sido probado matemáticamente para producir los mejores o más óptimos alineamientos entre dos secuencias al especificarse las condiciones de similitud, y puede aplicarse para alineamientos locales y globales. Sin embargo, este método es lento debido a la gran cantidad de pasos computacionales, que aumentan a una razón del cuadrado o el cubo de los largos de las secuencias. El requerimiento de memoria computacional también aumenta al cuadrado de los largos de las secuencias.
- **Palabra o tupla-k:** En estos métodos primero se buscan intervalos cortos idénticos de las secuencias (llamados palabras o tuplas-k) y luego se unen estas palabras en un alineamiento hecho por el método de programación dinámica. Estos métodos son lo suficientemente rápidos para buscar en una base de datos las secuencias que se alinean mejor a una secuencia de entrada de prueba. Este método es usado en algoritmos como los de FASTA y BLAST (McGinnis y Madden, 2004) (ambos son programas que identifican secuencias homólogas usando alineamiento de pareja, usando método de palabra heurístico). Si bien este tipo de método no garantiza encontrar una solución óptima como la programación dinámica, se considera más eficiente en el uso de recursos comparada con esta.

3.2.2. Alineamiento de secuencias múltiples - MSA

El alineamiento de secuencias múltiples (Multiple Sequence Alignment) consiste en el alineamiento de tres o más secuencias. De la alineación resultante, la homología de la secuencia puede ser inferida, y esto permite realizar un análisis filogenético (estudiar la historia evolutiva de las secuencias) para averiguar los orígenes evolutivos compartidos. Existen métodos heurísticos (que no necesariamente llegan a una solución óptima) y dinámicos para realizar MSA:

- **Programación dinámica:** Para este método se ocupan dos parámetros, una penalización por espacios vacíos (residuos que no tienen equivalente en las otras, este valor se debe

minimizar) y una matriz de sustitución para asignar puntuaciones basadas en las similitudes de las propiedades químicas y evolutivas de los aminoácidos a comparar, a modo de obtener una solución global óptima.

- **Construcción de alineamiento progresivo:** Funciona por una construcción sucesiva de alineamientos de parejas, este primer alineamiento queda fijo. Luego, se elige una tercera secuencia y se alinea con el primer alineamiento, este proceso se repite hasta todas las secuencias estén alineadas. El alineamiento progresivo es heurístico: no separa el proceso de puntuación de un alineamiento del algoritmo de optimización aplicado. Tampoco optimiza una función global de puntuación sobre el correcto alineamiento. La ventaja del alineamiento progresivo, es que es rápido y eficiente. Lo más importante en este método es la alineación de los pares de secuencias similares primeros, ya que estos son los alineamientos más confiables. Generalmente se usa una estructura de árbol binario para representar este método.
- **Alineamiento iterativo:** Estos métodos producen MSA mientras que reducen la cantidad de errores que tienen los métodos de alineamiento progresivos, funcionan de manera similar, pero repetidamente realinean las secuencias iniciales, como también añaden nuevas secuencias a la creciente MSA. Los métodos iterativos, a diferencia de los progresivos, pueden retornar a valores de parejas calculados anteriormente, o a sub secuencias múltiples alineadas, incorporando los subconjuntos de la secuencia de búsqueda a modo de optimizar una función general objetivo.
- **Alineamiento de consenso:** Estos métodos buscan encontrar la MSA óptima, teniendo múltiples diferentes alineamientos de las mismas secuencias. Uno de estos métodos es MergeAlign, que es capaz de generar alineamientos por consenso de cualquier número de alineamientos de entrada usando diferentes modelos de evolución de secuencias (Collingridge y Kelly, 2012).
- **Modelo oculto de Márkov ("hidden Markov model", HMM):** Es un modelo probabilístico de una MSA de proteínas. En el modelo, cada columna de símbolos en el alineamiento es representado por una frecuencia de distribución de símbolos (llamados estados), e inserciones y eliminaciones se representan con otros estados. El movimiento por el modelo es a través de un camino particular de un estado a otro en una cadena de Márkov (una transición de un estado a otro siguiendo reglas probabilísticas), intentado encontrar una coincidencia para la secuencia dada. El siguiente símbolo coincidente es elegido de cada estado, guardando su probabilidad y también la probabilidad de ir de ese estado a un estado previo. Las probabilidades de estado y transición son multiplicadas para obtener la probabilidad de la secuencia dada (Mount, 2009).

3.3. Conclusión del Capítulo

Las proteínas son largas cadenas de aminoácidos conectadas entre si, y su estructura se puede clasificar en primaria, secundaria, terciaria y cuaternaria.

- La estructura primaria corresponde a sólo a el orden lineal de la secuencia de aminoácidos.
- La estructura secundaria corresponde a las conformaciones locales que se forman a lo largo de la secuencia, principalmente hélices y láminas.
- La estructura terciaria corresponde a la estructura tridimensional de la proteína.
- La estructura cuaternaria es el complejo proteico formado por dos o más cadenas de aminoácidos.

El plegamiento tridimensional de la proteína es determinado por la secuencia de aminoácidos que la componen. Este principio es importante para la predicción de la estructura a través de sistemas computacionales. Además, las estructuras de las proteínas pueden interactuar y a esto se le llama una interacción proteína-proteína, es un componente central en todos los procesos biológicos. Las interacciones pueden ser de un carácter transitorio o permanente.

Una de las técnicas utilizadas para predecir estas estructuras formadas en base a interacciones proteína-proteína es el docking, esta es una forma de predecir un complejo proteico en base a las estructuras individuales de las proteínas interactuantes. Otro concepto importante para las predicciones de estructura de proteínas son los alineamientos. El alineamiento de secuencias es el análisis de secuencias proteicas para buscar similitudes, pueden ser de carácter global o local, y según la cantidad de secuencias puede ser alineamiento de pares o alineamiento de secuencias múltiples (MSA).

Los métodos modernos de predicción de estructuras proteicas hacen uso de los alineamientos, combinados con distintas técnicas de inteligencia artificial. La inteligencia artificial permite que los computadores puedan aprender de la experiencia, a modo de buscar significado y razonar frente a nueva información. Este aprendizaje a través de experiencia es parte importante de un subcampo de la inteligencia artificial, llamado machine learning.

El machine learning permite que los sistemas aprendan en base a datos de entrenamiento relevantes para llevar a cabo la función para la cual son entrenados. Los sistemas de predicción de estructuras hacen uso de un subcampo específico del machine learning, conocido como deep learning. El deep learning es un subconjunto del machine learning en el que el aprendizaje ocurre en capas de representaciones, que generalmente corresponden a redes neuronales. Las redes neuronales están compuestas por neuronas que reciben entradas de datos, y ejecutan una función para producir una salida que sirve de entrada a otras neuronas.

En las redes neuronales de predicción de estructuras proteicas se hace uso de mecanismos de

atención, que permiten optimizar el uso de recursos computacionales en la red y dar prioridad a los procesos que se consideren más importantes. La combinación de las redes neuronales con mecanismos de atención es lo que permite un gran avance en el campo de predicción de estructuras proteicas, esto será descrito en el capítulo siguiente.

Capítulo 4

La predicción de estructuras proteicas a través del tiempo

En este capítulo se describirán los avances que ha tenido el campo de la predicción de estructuras proteicas, desde sus inicios hasta el estado del arte actual, además de las primeras técnicas de predicción y experimentos realizados para obtener estructuras proteicas, finalizando con los métodos más avanzados de la actualidad, que hacen uso de inteligencia artificial.

Como se vio en el Capítulo 3.1, las proteínas consisten en una cadena lineal o secuencia de aminoácidos específica, y esta corresponde a la estructura primaria de una proteína, si bien existen métodos experimentales avanzados hoy en día para la determinación de estas, en los inicios del estudio de la estructura proteica esto no era una tarea sencilla.

El primer proceso de obtención de la secuencia de aminoácidos de una proteína se realizó con la insulina en 1949, por Sanger (Sanger, 1949) y su equipo. En este proceso se utilizaron múltiples métodos químicos, además de ensayo y error, para poder romper la cadena proteica en cadenas más pequeñas, para luego identificar los aminoácidos individuales de los extremos de cadena. Aquí se descubrió que la insulina tenía dos cadenas A y B. La determinación de la estructura completa culminó en 1958, se determinó la secuencia de los 51 aminoácidos totales de las dos cadenas y Sanger recibió el premio nobel de química (Sanger, 1958).

Las primeras estructuras secundarias predichas ocurrieron en este mismo periodo, en 1951, Linus Pauling, Robert Corey, y Herman Branson (Pauling et al., 1951) propusieron las estructuras secundarias de la alfa hélice y la lámina beta. Estas estructuras fundamentales y bloques básicos de las proteínas fueron deducidas a partir de las propiedades de las moléculas pequeñas, conocidas por estructuras cristalinas y por la teoría de resonancia los enlaces químicos de Pauling, que predecían los grupos planares peptídicos (Eisenberg, 2003). También propusieron estructuras y

funciones que aún no han sido observadas, como la hélice gamma.

La primera estructura proteica determinada en tres dimensiones fue la mioglobina (ver Figura 4.1) en 1957 por John Kendew, mediante el uso de difracción de rayos X. Esta técnica fue desarrollada y mejorada desde el 1913, cuando se descubrió que los cristales difractan los haces de rayos X. En el Anexo A.6 se describen los pasos para obtener estructuras proteicas usando rayos X.

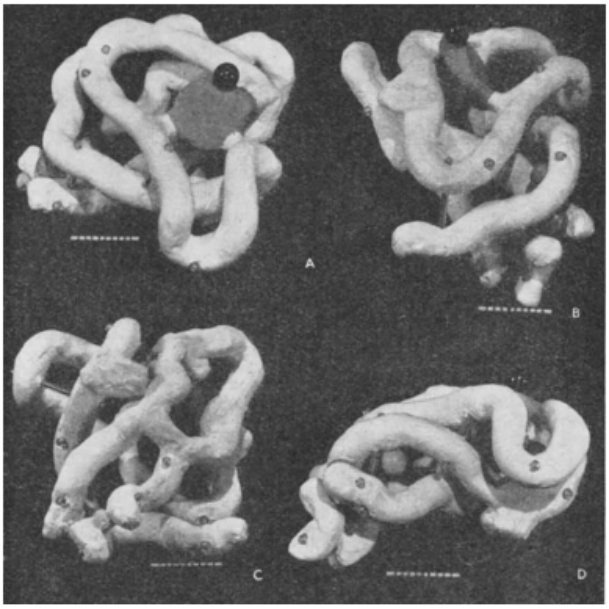


Figura 4.1: Fotografía del primer modelo obtenido de la molécula de mioglobina usando difracción de rayos X (Kendrew et al., 1958).

Debido a la cantidad de estructuras proteicas que se comenzaron a obtener experimentalmente, surge la necesidad de tener un repositorio con la información de estas, en 1971 se establece el Protein Data Bank (Banco de Datos de Proteínas, PDB), y almacena información de proteínas, ácidos nucleicos, y ensamblados complejos (wwPDB consortium, 2019).

Para almacenar estructuras tridimensionales en el Protein Data Bank, se creó el formato de archivo PDB, que provee la descripción y la notación de las proteínas, incluyendo sus coordenadas atómicas, asignaciones de estructuras secundarias, y también su conectividad atómica. Además de la cristalografía de rayos X, otro método experimental que existe para determinar la estructura de proteínas es la espectroscopia mediante resonancia magnética nuclear de proteínas (NMR). La NMR es frecuentemente uno de los métodos más usados para obtener información de alta resolución sobre la dinámica de las proteínas y su estructura en un solvente. Este método usa las propiedades magnéticas de los núcleos de los átomos, que poseen un spin. Para facilitar los experimentos, es necesario etiquetar isotópicamente la proteína con carbono 13 y nitrógeno 15,

para el hidrógeno no es necesario ya que tiene una abundancia del 99,9 %.

La proteína de interés se ubica en un campo magnético fuerte, y cada uno de los núcleos de los átomos se caracteriza por tener una frecuencia de resonancia única, dependiendo de la densidad electrónica del entorno químico local, pero también en combinación con el campo magnético local y el campo externo. El método más importante para la determinación de estructuras usa el efecto nuclear Overhauser, que se define como un cambio en la intensidad integrada de un espín cuando la población de equilibrio de otro espín es perturbada por saturación o inversión (Wüthrich, 2003). Se utiliza este efecto para determinar las distancias entre pares de átomos al interior de la molécula que no se encuentran conectados a través de enlaces químicos. Para determinar las distancias entre pares de átomos unidos por enlaces químicos se usan otros métodos experimentales (acoplamiento-J).

El objetivo final de la espectroscopia NMR es observar los cambios químicos desde los espectros multidimensionales a los núcleos de los átomos específicos en la proteína. Todos los valores luego son cuantificados y traspasados a restricciones de ángulo y distancia. La mayoría de estas restricciones se encuentran en un rango de posibles valores, no son restricciones precisas. Finalmente, estas restricciones son usadas para generar la estructura tridimensional de la proteína resolviendo problemas de distancia geométrica (Mishra et al., 2012).

Si bien la existencia de métodos experimentales como la cristalografía en rayos X y la espectroscopia NMR comenzaron a aumentar la base de datos de estructuras proteicas conocidas, estos métodos son lentos y no alcanzan a abarcar la gran cantidad de estructuras que se necesita estudiar. Para solucionar este problema existe el campo de la predicción de estructuras proteicas mediante métodos computacionales. El campo de predicción de estructuras se puede dividir en dos categorías, en predicción comparativa (basado en estructuras de proteínas ya conocidas) y predicción De Novo o Ab Initio (basado sólo en la secuencia de aminoácidos de la proteína).

4.1. Predicción Comparativa

La predicción comparativa predice la estructura tridimensional de una secuencia proteica basada principalmente en el alineamiento de una o más proteínas con estructura conocida (plantillas). El proceso de predicción consiste en asignación de pliegues, alineamiento de la plantilla objetivo, construcción del modelo y evaluación del modelo (Eswar et al., 2006). Los métodos comparativos se pueden dividir en dos grupos, en predicción por homología o predicción por enhebramiento.

4.1.1. Predicción por homología

El paso inicial para la predicción por homología es revisar si existe una proteína en la base de datos que tenga una secuencia similar a la proteína de interés. Si existe, la estructura de esta proteína será usada como plantilla. La búsqueda de la plantilla prosigue usando un algoritmo de comparación de la secuencia para identificar la similitud global de la secuencia. La secuencia de la proteína con estructura desconocida es alineada contra la secuencia de la proteína plantilla, de manera que ambas secuencias se superpongan en las regiones en que los aminoácidos son los mismos. Luego las coordenadas de los carbonos alfa de la plantilla se copian sobre la proteína objetivo de manera de formar la cadena principal. Las técnicas usadas de alineamiento son las vistas en el capítulo anterior, alineamiento de parejas y alineamiento de secuencias múltiples.

Para modelar la estructura de bucles se usan bases de datos de conformaciones de bucles o por modelamiento usando métodos ab initio. Para las cadenas laterales, se usan métodos de posicionamiento basados en librerías de rotámeros (los rotámeros son las diferentes posiciones que pueden tomar las cadenas laterales) con las conformaciones de las cadenas laterales. Las librerías de rotámeros contienen las conformaciones preferidas de todas las cadenas laterales de los aminoácidos, junto a los ángulos de torsión correspondientes.

El último paso es el refinamiento del modelo obtenido, generalmente al usar refinamiento enfocado solo en la minimización de energía conlleva a estructuras que son distintas a las obtenidas por cristalografía de rayos X. Para evitar estos problemas se usan combinaciones de técnicas, como dinámica molecular, potenciales estadísticos, o considerar los efectos de un solvente.

La calidad del modelo es dependiente de la calidad de la secuencia de alineamiento de la estructura plantilla usada. Los errores son significativamente altos en regiones de bucles, debido al aumento de flexibilidad de estas regiones. Errores en las cadenas laterales aumentan al disminuir la identidad de la secuencia de aminoácidos, y son causados por el hecho de que las cadenas laterales pueden existir en varias conformaciones.

4.1.2. Predicción por enhebramiento

La predicción por enhebramiento (threading) o reconocimiento de pliegues es usada para modelar la estructura de una proteína cuando no existen homólogos conocidos. El enhebramiento está basado en la idea de que existe un número limitado de plegamientos en la naturaleza, y por lo tanto una nueva estructura debe tener un plegamiento similar a los que ya se encuentran en el PDB.

Este método compara la secuencia objetivo contra una librería de posibles pliegues de plantillas usando potenciales de energía u otros métodos de puntuación similares. La plantilla con la

puntuación de energía más baja, o la similitud más alta, se utiliza para modelar el pliegue de la proteína objetivo.

La primera estructura proteica predicha usando predicción comparativa fue la alfa lactalbumina en 1969, basada en la estructura de rayos X de la lisozima. Comenzó con el alineamiento entre la plantilla y la proteína objetivo, seguida por la construcción de un modelo inicial creado por inserciones, eliminaciones y reemplazos de las cadenas laterales de la estructura plantilla, y terminó con el refinamiento del modelo usando minimización de energía.

4.2. Predicción a partir de la secuencia

Aunque los métodos de predicción comparativos tienen una gran tasa de éxito en sus predicciones, aún existen demasiadas proteínas que no tienen una proteína plantilla adecuada. Los métodos ab initio son más generales en este sentido, debido a que sólo dependen de conocer la estructura primaria de la proteína. Estos métodos están basados en la teoría de plegamiento de Anfinsen, en la que la estructura nativa de la proteína corresponde al mínimo de energía libre global.

El primer empuje a gran escala para el campo de la predicción de estructuras proteicas fue en 1994, donde se realizó y organizó por primera vez un evento dedicado al estudio de la predicción de estructuras proteicas, llamado Critical Assessment of Techniques for Protein Structure Prediction (CASP) (Evaluación crítica de las técnicas para la predicción estructural proteica). Este evento, encabezado por John Moult, tiene como objetivo ayudar a avanzar los métodos para identificar la estructura de las proteínas a partir de la secuencia de aminoácidos.

En el Anexo A.7 se describe como se desarrolló la primera iteración de CASP. En la actualidad, CASP es un centro de investigación, y está organizado para proveer medios de probar de manera objetiva estos métodos computacionales por un proceso de predicción ciega. Los experimentos realizados en CASP tienen como objetivo establecer el estado del arte actual en la predicción de estructuras proteicas, identificando que progreso se ha logrado, y destacando donde los futuros esfuerzos pueden ser más productivamente focalizados.

Esto nos lleva al estado del arte actual, donde en CASP14 realizado en 2020, se vio un avance enorme (ver Figura 4.2) en la precisión de modelos de una sola proteína y modelos de dominio, este avance se debe principalmente debido a la implementación exitosa de algoritmos de deep learning, particularmente por la inteligencia artificial propuesta por el equipo de DeepMind, Alphafold.

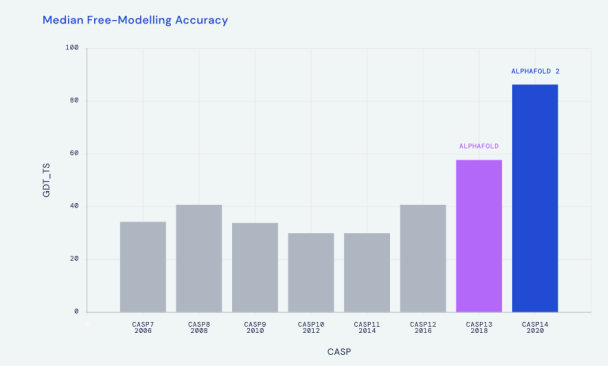


Figura 4.2: Resultados de CASP desde la séptima iteración (2006), en las versiones CASP13 y CASP14 la inteligencia artificial Alphafold de DeepMind obtuvo resultados significativos en la predicción de estructuras.

4.3. Alphafold2

Alphafold2 (llamado simplemente Alphafold) es un sistema de inteligencia artificial desarrollado por DeepMind, que es capaz de predecir la estructura tridimensional de una proteína a partir de su secuencia de aminoácidos (Jumper et al., 2021). Alphafold aumenta considerablemente la precisión de la predicción de estructuras de las proteínas al incorporar nuevas arquitecturas de redes neuronales y procedimientos de entrenamientos basados en las limitaciones evolutivas, físicas y geométricas de las estructuras de las proteínas.

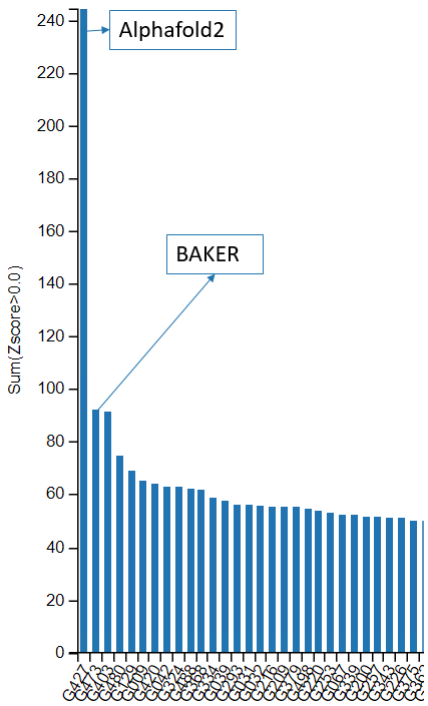


Figura 4.3: Alphafold obtuvo los mejores resultados promedio entre todas las categorías de CASP14, la diferencia entre el primer lugar y el segundo lugar es de un puntaje estándar de 153.1976.

Las estructuras predichas y resueltas por Alphafold fueron considerablemente (ver Figura 4.3) más precisas que los métodos de los competidores de CASP14. Las estructuras resueltas de Alphafold tenían una mediana con precisión de 0,96 $\AA$ r.m.s.d. $_{95}$ (un angstrom equivale a 10^{-10} metros).

- r.m.s.d. $_{95}$ (root-mean-square deviation): desviación de raíz cuadrada media del carbono alfa al 95 % de cobertura de residuos.
- Intervalo de confianza 95: 0,85–1,16 $\AA$.

Además de producir estructuras de dominio muy precisas, Alphafold es capaz de producir cadenas laterales con alta precisión cuando la cadena tiene alta precisión, y mejora considerablemente los métodos basados en estructuras de plantillas.

4.3.1. Red Alphafold

La red Alphafold predice directamente las coordenadas tridimensionales de todos los átomos pesados (átomos que no son hidrógeno) de una proteína dada usando su secuencia de aminoácidos y sus homólogos como datos de entrada.

La red neuronal tiene dos etapas principales. Primero, el tronco de la red procesa los datos de entrada a través de capas repetidas de un nuevo bloque neuronal llamado Evoformer, para producir un arreglo de N_sxN_r dimensiones (s es el número de secuencias y r el número de residuos) que representa una MSA procesada, y un arreglo de N_sxN_r que representa el número de pares de residuos.

El tronco de la red neuronal luego es seguido por un módulo estructural que introduce una estructura tridimensional explicita, en la forma de una rotación y traslación de cada residuo de la proteína (la rotación y traslación definen un marco global rígido, R y t). Estas representaciones se inicializan en un estado trivial, con todas las rotaciones en la posición identidad (ángulos en 0) y todas las posiciones se ubican en el origen, después de esto rápidamente se desarrolla y refina una estructura proteica con precisión de detalles atómico (Jumper et al., 2021).

Las innovaciones claves en esta sección de la red incluyen la habilidad de romper la estructura de cadena para permitir refinamiento local simultaneo de todas las partes de la estructura, una función de transformación novedosa para permitir que la red razone implícitamente sobre los átomos de las cadenas laterales que no son representados y un término de pérdida, que ofrece un peso sustancial sobre la correcta orientación de los residuos.

Al interior del módulo estructural y a través de toda la red, existe un refinamiento iterativo, aplicándose repetidamente la pérdida final de las salidas que luego son ingresadas recursivamente a estos mismos módulos. El refinamiento iterativo usado en toda la red, que en la red se llama reciclaje (recycling), contribuye considerablemente a la precisión, y sólo agrega un leve aumento de tiempo de entrenamiento.

4.3.2. Evoformer

El Evoformer (ver Figura 4.4), bloque básico de la red neuronal, tiene como principio clave la predicción de estructuras de proteínas como un problema de inferencia de grafos en el espacio tridimensional, en el que las aristas del grafo están definidos por residuos en la proximidad. Los elementos de una representación de pareja codifican información acerca de la relación de los residuos. Las columnas de las representaciones MSA codifican los residuos individuales de la secuencia de entrada, mientras que las filas representan las secuencias en que esos residuos aparecen.

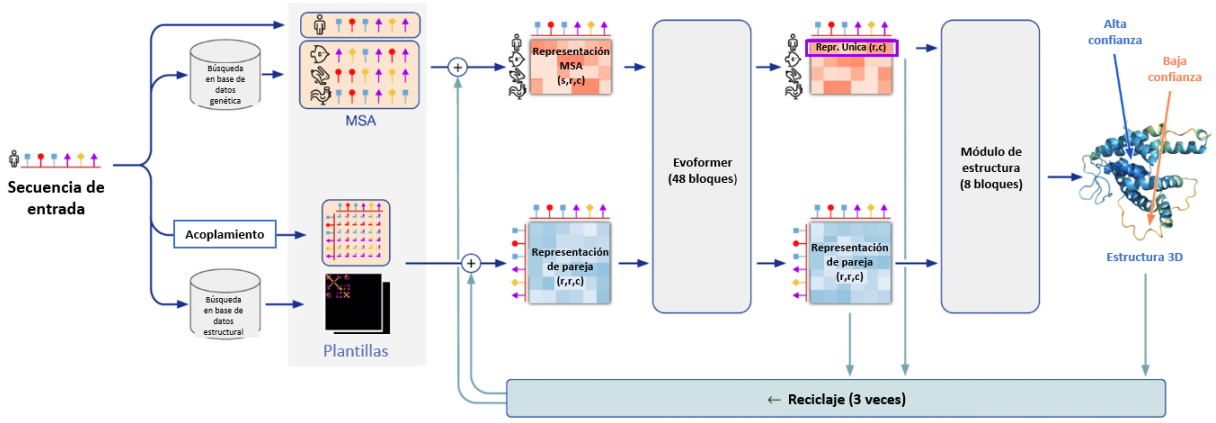


Figura 4.4: Arquitectura del modelo Alphafold.

En la Figura 4.4, Las flechas indican los flujos de información entre los varios componentes de la red. Las formas de los arreglos se denominan como: s, número de secuencias; r, número de residuos; c, número de canales (Jumper et al., 2021).

Dentro de esta arquitectura se definen un número de operaciones de actualización, que son aplicadas en cada bloque, en el cual las diferentes operaciones de actualización son realizadas en serie (Jumper et al., 2021). La representación MSA actualiza la representación de pareja a través de un producto exterior (cuyo resultado es una matriz) por elemento que es sumado sobre la dimensión de secuencia de la MSA. Esta operación se realiza sobre cada bloque de la red, lo que permite una comunicación continua de la representación MSA en evolución, con la representación de pareja.

4.3.3. Ángulos de cadenas laterales y marcos

Las predicciones de los ángulos de las cadenas laterales X, al igual que la precisión por residuo de la estructura (pLDDT) se calculan con pequeñas redes por residuo en las activaciones finales al final de la red. La puntuación de plantilla estimado (pTM) se obtiene de una predicción de error de pareja, que es computada en base a la representación de pareja final.

La pérdida final (llamada marco alineado por error de punto, “frame-aligned point error”, abreviado como FAPE) compara las posiciones predichas de los átomos con las verdaderas posiciones bajo varios diferentes alineamientos (ver Figura 4.5). Para cada alineamiento, definido por el marco predicho (R_k,T_k) con el marco verdadero, se calculan las distancias de todas las posiciones de los átomos x_i con las posiciones verdaderas.

Las distancias N_{marcos}x$N_{átomos}$ resultantes son penalizadas por una función de pérdida. Esto crea un sesgo fuerte para que los átomos se encuentren correctos posicionalmente relativo al marco local de cada residuo, y, por lo tanto, correcto con respecto con las interacciones de su cadena lateral; esto provee la principal fuente de quiralidad (la imagen reflejo es distinta a la original, es decir, la molécula no se puede superponer sobre la original mediante combinaciones de rotación, traslación o conformacionales) para Alphafold.

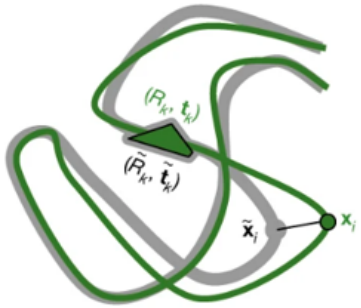


Figura 4.5: Marco alineado por error de punto (FAPE). En verde la estructura predicha, en gris la verdadera estructura; (R_k,t_k), marcos; x_i, posiciones de átomos (Jumper et al., 2021).

4.3.4. AlphafoldDB

En la actualidad, la cantidad de secuencias proteicas únicas archivadas son de aproximadamente 190 millones en la base de datos de UniProt (Consortium, 2020). Las bases de datos de UniProt existen para apoyar la investigación biológica y biomédica al proveer un compendio completo de todos los datos de secuencias de proteínas conocidas, junto con un sumario de la información funcional de la proteína obtenida por experimentos verificados o predicha computacionalmente.

Si bien la cantidad de secuencias de proteínas es alta, en el Protein Data Bank se almacenan aproximadamente 180.000 estructuras tridimensionales de proteínas (ver Figura 4.6), y esto limita severamente la cobertura de todo el espacio de secuencias de aminoácidos para investigación.

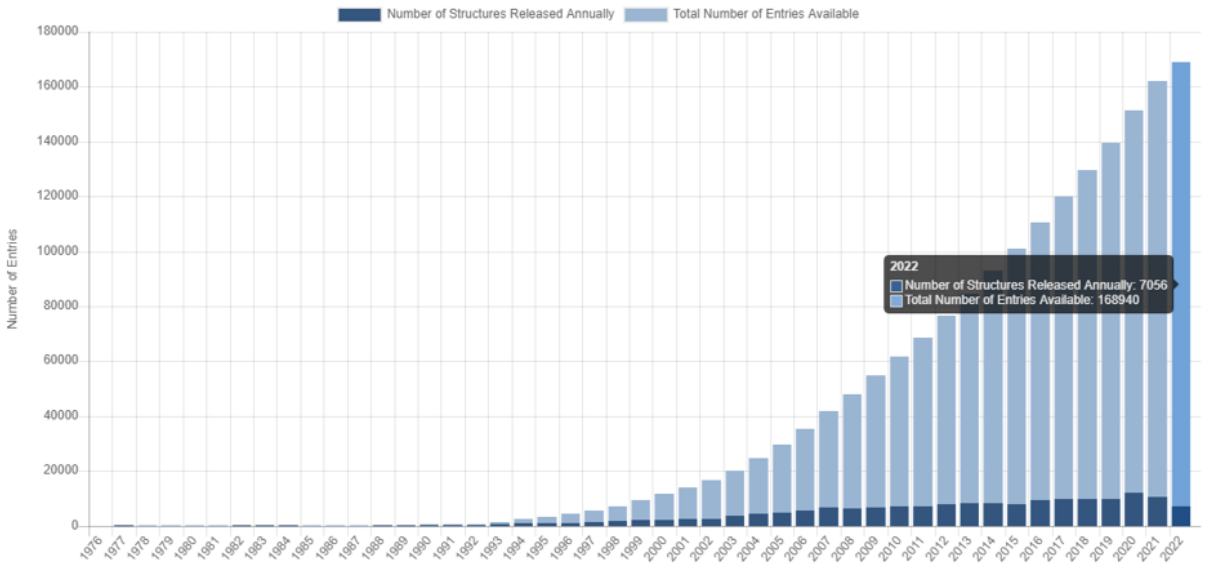


Figura 4.6: Estadísticas del PDB: Estructuras proteicas liberadas por año.

En la Figura 4.6, se pueden observar las cantidades de estructuras proteicas en el Protein Data Bank, en azul las que se agregan cada año y celeste el total hasta ese año. A la fecha actual existen 168.940 estructuras disponibles.

Alcanzar un aumento de la cobertura de todo el espacio de secuencias usando métodos experimentales para obtener estructuras de alta resolución es un trabajo muy intensivo. Frecuentemente requiere una gran cantidad de ensayo y error para encontrar condiciones en que una proteína sea susceptible a cristalización. Si bien avances recientes en los campos de la criomicroscopía electrónica (ver Figura 4.7), y métodos híbridos para la determinación de estructuras han acelerado la tasa a la que se determinan estructuras, la brecha entre las secuencias conocidas y las estructuras proteicas experimentales sigue expandiéndose.

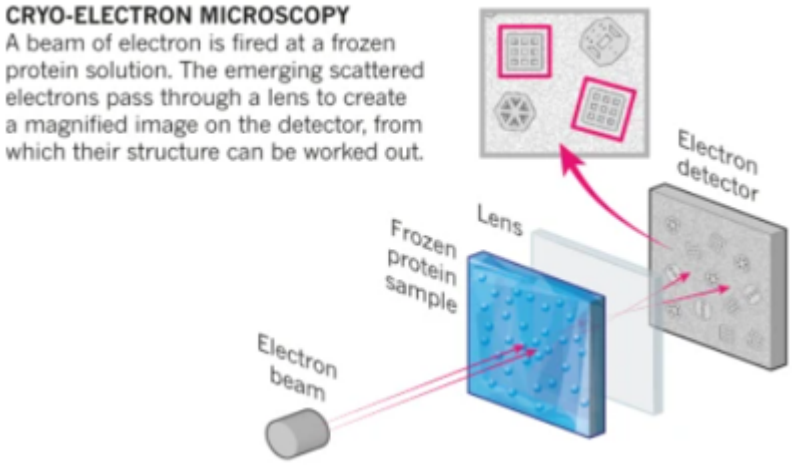


Figura 4.7: Un haz de electrones se hace incidir sobre una solución de proteínas congeladas. Los electrones que emergen dispersos a través de la solución pasar por un lente, para crear una imagen magnificada en un detector, desde el cual luego se puede trabajar para resolver la estructura (imagen de (Callaway, 2015).

Para disminuir esta brecha, se creó la base de datos de Alphafold, un nuevo recurso de datos de estructuras creado en conjunto con DeepMind y el Instituto Europeo de Bioinformática (EMBL-EBI) (Consortium, 2020). AlphafoldDB fue creada para hacer disponibles las predicciones de estructuras de manera libre a toda la comunidad científica. En su lanzamiento, la base de datos incluía el proteoma humano y otros 20 organismo, con un total de aproximadamente 360.000 predicciones de estructuras proteicas.

Al día de hoy, AlphafoldDB tiene un buscador que permite buscar estructuras (ver figura 4.8), y ha sido actualizado con casi la totalidad de las proteínas conocidas por la ciencia, con un total de estructuras sobre los 200 millones, un aumento considerable sobre la versión inicial, y vastamente superior a la cantidad de estructuras obtenidas experimentalmente, si bien aún se debe realizar un análisis sobre la confianza de la estructura proteica con la que se quiere trabajar, este potencial aumenta considerablemente el entendimiento que se tiene de biología.

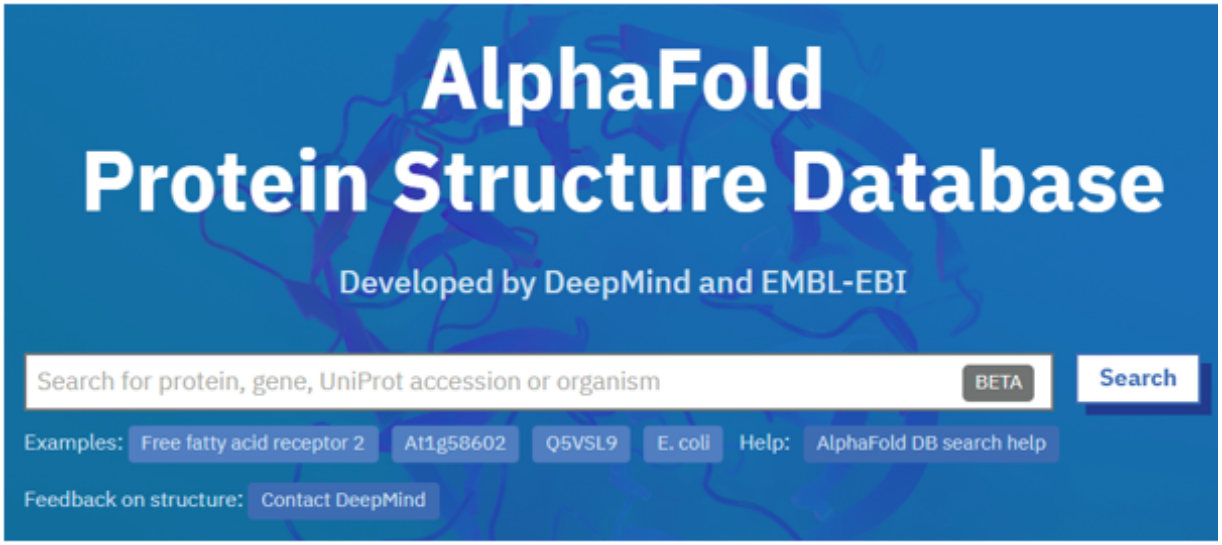


Figura 4.8: La base de datos AlphafoldDB es disponible para todos, y en la actualidad posee sobre 200 millones de predicciones de estructuras proteicas.

4.3.5. Predicción de múltiples cadenas

En la versión inicial de Alphafold, presentada en CASP14, esta solo podía predecir proteínas de una sola cadena, por lo que la predicción de complejos proteicos de múltiples cadenas aún representaba un desafío. Para solucionar esto, se entrenó un modelo de Alphafold específicamente para proteínas de entrada de múltiples cadenas con estequiometria conocida, llamado Alphafold-multimer (Evans et al., 2022).

Con Alphafold-multimer se aumenta la precisión de estructuras a partir de entradas de proteínas multiméricas, adaptadas de la versión original que permitía una única cadena, y aun manteniendo alta precisión para la cadena interior. Esto se logró modificando los parámetros de FAPE, cambios en la función de pérdida, una leve alteración a la construcción de MSA, la búsqueda de optimización del alineamiento de una permutación de múltiples cadenas, y usando genética de cadenas cruzadas, que consiste en proveer secuencias alineadas a la red, emparejando secuencias usando la notación de especies de UniProt. Cuando múltiples secuencias existen para una misma especie, se clasifican las filas candidatas de cada cadena según la similitud a su respectiva secuencia objetivo, y se concatenan los pares que tienen la misma clasificación.

En la actualidad, Alphafold-multimer ha sido incorporado a Alphafold, por lo que la inteligencia artificial tiene ambas capacidades, predicción de proteínas de una sola cadena y predicción de complejos proteicos de múltiples cadenas. Para la comunidad científica y el equipo de DeepMind, el éxito de Alphafold es uno de los hitos más importantes para el desafío de la predicción de

estructuras y para la inteligencia artificial. Al demostrar que la inteligencia artificial puede predecir precisamente la forma de una proteína con un nivel de precisión atómica, y en minutos, Alphafold no solo provee una solución al gran desafío de la predicción de estructuras a partir de la secuencia de aminoácidos de una proteína, sino que también se puede considerar como una de las grandes pruebas de que la inteligencia artificial puede acelerar dramáticamente los descubrimientos científicos, y al mismo tiempo avanzar la humanidad.

4.4. RoseTTAFold

Con el impacto de Alphafold en el campo de predicción de estructuras proteicas e inteligencia artificial, el interés en arquitecturas neuronales aumentó. Esto llevo a la exploración de distintas formas de construir una nueva red neuronal, inspirada en la red de Alphafold, llamada RoseTTAFold.

RoseTTAFold es una red neuronal de tres factores, que considera simultáneamente patrones en secuencias de proteínas, cómo los aminoácidos de una proteína interactúan entre sí y la posible estructura tridimensional de una proteína. En esta arquitectura, la información de una, dos y tres dimensiones fluye de un lado a otro, lo que permite a la red neuronal razonar colectivamente sobre la relación entre las partes químicas de una proteína y su estructura plegada (Baek et al., 2021).

Esta red neuronal de tres factores produce predicciones con precisión cercana o al nivel de la de Alphafold, y permite la solución rápida de problemas de modelamiento de estructuras de criomicroscopía electrónica y cristalografía de rayos X.

La red también permite la generación rápida de complejos proteína-proteína a partir de únicamente la secuencia, sin la necesidad de requerir modelado de subunidades individuales seguidas de docking, esta es una de las fortalezas de RoseTTAFold.

4.4.1. Generación directa de complejos proteína-proteína

La capa final de la red neuronal de RoseTTAFold genera estructuras tridimensionales combinando características de secuencias de cadenas discontinuas (dos segmentos de la proteína con un quiebre de cadena entre ellos). Debido a que la red puede manejar sin problemas quiebres de cadenas, es capaz de predecir la estructura de complejos proteína-proteína directamente de la secuencia de aminoácidos, con o sin la ayuda de estructuras plantillas, dos o más secuencias (y posibles plantillas para ellas) pueden ser los datos de entrada, y la salida serán las coordenadas de la vértebra de dos más cadenas proteicas. Esto es lo que permite a la red la construcción directa

de modelos para complejos proteína-proteína sólo a partir de información de la secuencia, sin necesidad de docking.

Una mayor cantidad de entrenamiento con datos de complejos proteína-proteína mejora el modelamiento de estructuras de complejos multiproteícos. Este método también puede ser utilizado en conjunto con metodología de diseño de unión de pequeñas moléculas con proteínas, para mejorar el descubrimiento computacional de nuevas proteínas y pequeños ligando moleculares para objetivos de interés.

4.5. Colabfold

Colabfold hace uso de las redes ya descritas, Alphafold y RoseTTAFold, en conjunto con búsqueda de homologos rápida usando MMseqs2. MMseqs2 ("Many-against-Many sequence searching", búsqueda de secuencias muchos contra muchos) es un software para buscar y agrupar grandes set de secuencias de nucleótidos y proteínas (Steinegger y Söding, 2017).

La búsqueda de ColabFold es rápida y la utilización optimizada del modelo de ColabFold permiten la predicción de cerca de 1000 estructuras por día en un servidor con una unidad de procesamiento de gráficos (Mirdita et al., 2022). Usando Google Colaboratory, ColabFold es una plataforma de uso libre para el plegamiento de proteínas. Colabfold consiste de tres partes:

- La primera es una búsqueda basada en homología usando MMseqs2 para construir diversas MSA y encontrar plantillas. El servidor alinea eficientemente las secuencias de entrada contra la base de datos UniRef100, PDB70.
- La segunda parte es una librería de Python que se comunica con el servidor de búsqueda MMseqs2, prepara las características de entrada para la inferencia de estructuras (que pueden ser de una sola cadena o complejos), y luego visualiza los resultados. Esta librería también implementa una línea de comandos.
- Y por último, el uso de cuadernos Jupyter para uso básico y avanzado usando la librería de Python.

ColabFold expone muchos de los parámetros que se usan internamente en Alphafold, como la cantidad de reciclajes (por defecto en Alphafold es 3), lo que controla el número de veces que la predicción es repetidamente ingresada al modelo. Para objetivos más difíciles como también para proteínas diseñadas sin homólogos, agregar iteraciones de reciclaje adicionales pueden resultar en una predicción de alta calidad.

ColabFold además mejora la búsqueda de secuencias ofrecida por Alphafold, provee herramientas para modelado de complejos homodímeros o heterodímeros, expone funcionalidades avanzadas (alterar parámetros de Alphafold) y habilita predicción de múltiples estructuras proteicas a gran

escala, a una rapidez mayor que Alphafold.

4.6. OmegaFold

OmegaFold es un método computacional basado en deep learning y procesamiento de lenguaje natural, cuya fundación es un modelo de lenguaje proteico que permite realizar predicciones a partir de sólo una secuencia, no siendo necesario ingresar MSA para realizar predicciones (Wu et al., 2022). Esta capacidad es importante, debido que no siempre se cuenta con información de alineamiento de secuencias de proteínas homólogas, lo que permite realizar predicciones con alta precisión sobre proteínas huérfanas (proteínas cuya función es desconocida o no se tiene información de su árbol de coevolucion filogenética).

La precisión demostrada en los modelos de una cadena predichos por OmegaFold, alcanza niveles similares o superiores a Alphafold, especialmente en proteínas en las que no se tiene una gran profundidad de MSA.

4.7. ESMFold

Al igual que OmegaFold, ESMFold (Evolutionary Scale Modeling) utiliza modelos de lenguaje proteico para realizar predicciones, sin necesidad de depender de MSA. ESMFold fue entrenado con modelos de lenguaje proteico de hasta 15 billones de parámetros. A medida que se escalan estos modelos de lenguaje, los modelos aprenden información que permite predecir la estructura tridimensional de una proteína con resolución a nivel atómico (Lin et al., 2022).

4.8. DGMFold

A diferencia de Alphafold y RosettaFold, DGMFold (también conocido como RocketX) realiza las predicciones a través de las distancias geométricas inter-residuales, y es clasificado como un plegamiento asistido por restricciones geometricas (Geometric Constraint Assisted Folding). El plegamiento asistido por restricciones geométricas incluye dos pasos, la predicción de la restricción geométrica y el plegamiento de la estructura. En la predicción de la restricción geométrica, una red neuronal se entrena para aprender relaciones de coevolución a partir de MSA de entrada de la secuencia a buscar, para predecir los contactos ínter-residuos, las distancias y las orientaciones.

Luego estas restricciones geométricas se convierten a potenciales de energía para guiar el plega-

miento, a través de minimización de energía (Liu et al., 2022).

4.9. Conclusión del Capítulo

Desde los inicios del estudio de la biología estructural en la década del 1950 hasta los tiempos actuales, el avance de este campo ha sido considerable gracias a la utilización de métodos experimentales, como la cristalografía de rayos X, y más recientemente, el uso de la inteligencia artificial.

La llegada de Alphafold en CASP14 causó una revolución que se ha extendido más allá del campo de la predicción de estructuras proteicas, sino que a distintos campos de la biología en general. Esto causó que varios equipos se inspiraran en la arquitectura de Alphafold y la usaran como base para desarrollar nuevos métodos de predicción.

Algunos métodos, como ESMFold y OmegaFold usaron las bases de la arquitectura de Alphafold y la combinaron con el procesamiento natural del lenguaje, creando modelos de lenguajes proteicos capaces de predecir estructuras a partir de sólo la secuencia de aminoácidos, analizándolas como lenguaje natural.

En el capítulo siguiente se pondrán a prueba los modelos computacionales del estado del arte sobre un dataset creado manualmente del PDB, a modo de comparar la precisión obtenida por cada método de inteligencia artificial con la estructura experimental.

Capítulo 5

Estudio y Desarrollo de Metodología

En este capítulo se describe la metodología y desarrollo de esta durante la comparación de los métodos de predicción.

5.1. Contexto del problema

Debido al impacto que tuvo Alphafold en el campo de la predicción de estructuras proteicas, ha surgido una gran cantidad de métodos nuevos inspirados en la arquitectura de Alphafold, combinando distintas técnicas como procesamiento natural del lenguaje o usando una forma distinta de buscar grandes cantidades de alineamientos para mejorar la predicciones.

Esto sumado a la creciente necesidad de modelar las secuencias proteicas que se obtienen experimentalmente, sin necesidad de realizar una obtención del modelo de manera experimental, conlleva a que resultaría de gran utilidad saber cual de todos estos métodos es el mejor para predecir estructuras de proteínas.

Además de la obtención de estructuras proteicas, también sería importante investigar que tan confiables son estos sistemas de predicción para predecir interacciones proteína-proteína, si bien no es algo para lo que fueron diseñados originalmente, la gran capacidad que tienen para modelar la estructura tridimensional de complejos quizás pueda dar indicio de posibles interacciones en base a la interfaz entre las proteínas modeladas.

5.2. Metodología propuesta

Como primera etapa se procederá a seleccionar complejos proteicos del Protein Data Bank para su posterior modelamiento usando las técnicas del estado del arte. Estos complejos elegidos deben tener una fecha de publicación posterior al 01 de abril del 2021, a modo de que los archivos PDB utilizados no sean parte del dataset de entrenamiento de cualquiera de estas técnicas.

Las técnicas utilizadas en estado del arte se pueden clasificar como predicciones end-to-end (Jumper et al., 2021), geometrías inter-residuo (Liu et al., 2022) y modelos de lenguaje proteico (Wu et al., 2022). Alphafold, ColabFold, RoseTTaFold y DGMFold hacen uso de MSA para producir una predicción de estructura, mientras que OmegaFold y ESMFold sólo necesitan la secuencia proteica para realizar la predicción.

Los métodos a comparar junto a su clasificación se pueden ver en la Tabla 5.1. De estos métodos, sólo Alphafold, ColabFold y RoseTTaFold pueden predecir complejos proteicos, sin embargo, para ver si existe la capacidad de modelar complejos en el sistema actual de los otros métodos, se empleará una unión de 21 aminoácidos de glicina-glicina-serina entre las cadenas proteínas del complejo, a esta unión se le llama linker, y se utiliza para que el input al sistema sea sólo una secuencia de aminoácidos unidos por un linker de 21 aminoácidos de glicina-glicina-serina (GGS).

Método de Predicción	Clasificación		
	End-to-end	Protein Language Model	Inter-residue geometries
Alphafold-Colab	X		
ColabFold unpaired+paired	X		
ColabFold paired	X		
RoseTTaFold	X		
ESMFold		X	
OmegaFold		X	
DGMFold			X

Tabla 5.1: Métodos de predicción a comparar.

El linker GGS tiene como objetivo unir las cadenas del complejo de manera que se modele la proteína como si fuera una sola cadena. Una vez que se tiene el modelo obtenido con la unión del linker, se procederá a separar las cadenas y obtener el complejo.

Se consideró instalar el sistema completo de predicción de algunos de estos métodos en el cluster de la facultad, pero debido a los altos requerimientos computacionales se decidió usar los servidores y cuadernos de Google oficiales para cada uno.

- Para modelar con Alphafold se usó el cuaderno de Google oficial de DeepMind, ingresando las secuencias que conforman la proteína. ColabFold consiste en varios cuadernos de Google con distintas capacidades, se usará el cuaderno que utiliza MM2seqs2 para la búsqueda de alineamientos, con 3 como número de reciclajes y se usaran los modos paired y unpaired+paired.

- Para OmegaFold y ESMFold también se utilizaron las versiones proveídas por ColabFold, ingresando las secuencias unidas por linker.
- Para RoseTTaFold se usó el servidor oficial Robetta, en este es necesario ingresar con una cuenta al sistema, especificar las secuencias y además proveer el MSA (de un solo organismo) a utilizar.
- Para DGMFold se utilizó el servidor oficial, ingresando las secuencias unidas por linker.

Una vez obtenidas las predicciones de las distintas técnicas sobre el dataset seleccionado, se procedió a comparar cada una de estas predicciones con el modelo experimental del Protein Data Bank, usando DockQ (Basu y Wallner, 2016). Con estas comparaciones entre las predicciones y las estructuras experimentales se obtuvieron puntajes de DockQ, según los resultados de estos puntajes, se seleccionó el mejor método y se utilizó este método para evaluar interacciones proteína-proteína en un dataset positivo y negativo.

El dataset positivo esta compuesto de los complejos obtenidos de la predicciones del mejor método según los resultados de puntajes. El dataset negativo consiste en cadenas de proteínas que no interactúan según la literatura, las interacciones en las que exista valores valores que discrepan de lo documentado en la literatura serán corroborados en el servidor PEPPI (Bell et al., 2021). Finalmente, se desarrolló una aplicación en la que se ingresan los complejos predichos y se puede recibir como respuesta si las proteínas interactúan basado en los cálculos de pDockQ (Bryant et al., 2022), y se interpretaron los resultados obtenidos.

La metodología a seguir se puede visualizar en la Figura 5.1.

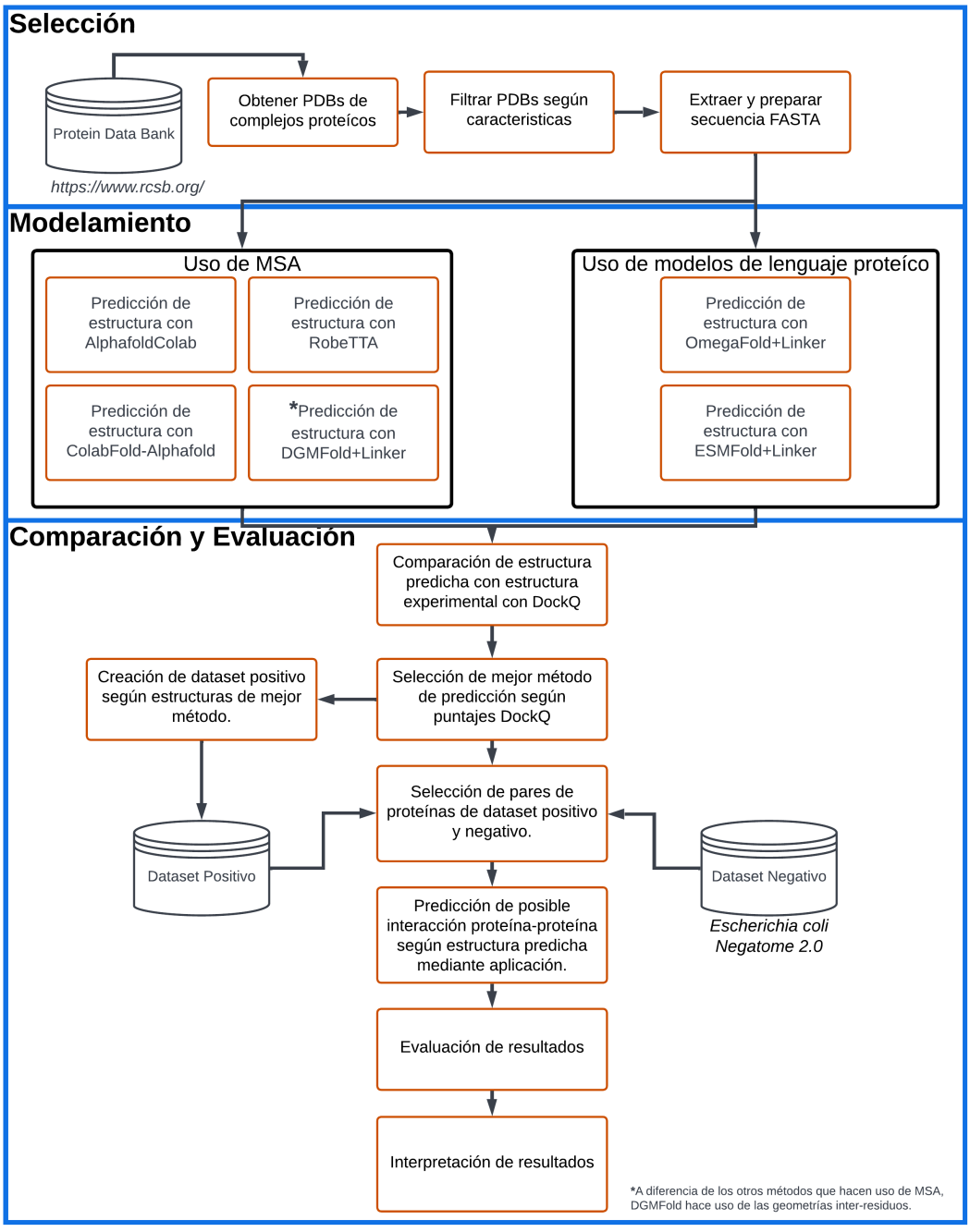


Figura 5.1: Metodología a seguir.

5.3. Selección de estructuras

Las estructuras a predecir se obtienen del Protein Data Bank, se seleccionaron 20 estructuras que hayan sido publicadas despues del 1 de abril del 2021, a modo de que las estructuras no sean parte del dataset de entrenamiento de algunos de los métodos, que tengan dos cadenas, no contengan fragmentos de ARN y que la cantidad de residuos se encuentre entre 230 y 500, para que los tiempos computacionales de los métodos no sean tan extensos y disminuir la posibilidad de perder conexión con los servidores de Google. Esta restricción en la cantidad de residuos y cadenas es debido a que inicialmente se tenia como limite 700 residuos, sin embargo, el tiempo computacional para Alphafold excedía el tiempo asignado, y provocaba que las sesiones de conexión terminaran abruptamente.

Una vez seleccionadas las estructuras objetivo, se extrae la secuencia FASTA y se utiliza las secuencias completas, no sólo los aminoácidos observados experimentalmente. Estas secuencias son ingresadas como datos de entrada a Alphafold, Colabold paired y unpaired+paired, Robetta, Dgmfold y OmegaFold, adecuandose a los requerimientos de entrada para complejos. En el caso de ESMFold y OmegaFold, primero las secuencias se unen mediante un linker de 21 residuos GGS, y se ingresa la secuencia resultante cómo si fuera una sola secuencia. Los IDs de los complejos, junto a la cantidad de residuos y cantidad de cadenas se pueden observar en la Tabla B.1.

5.4. Predicción y preparación de estructuras

Una vez obtenidas las predicciones, se procede a revisar las estructuras usando el software ChimeraX (Pettersen et al., 2021) y renombrar las cadenas en los casos que sea necesario para que coincidan con las estructuras experimentales, en el caso de las estructuras obtenidas por OmegaFold y ESMFold, se utiliza PyMol (The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC.) para visualizar, eliminar el linker agregado, y renombrar las cadenas para obtener un complejo con dos proteínas en un archivo PDB.

Esto se puede ejemplificar con la estructura 7KP1 obtenida por OmegaFold (ver Figura 5.2) mediante unión con linker, luego esta estructura es separada en las dos cadenas que forman el complejo (ver Figura 5.3) y se renombran las cadenas al nombre de las cadenas experimentales.

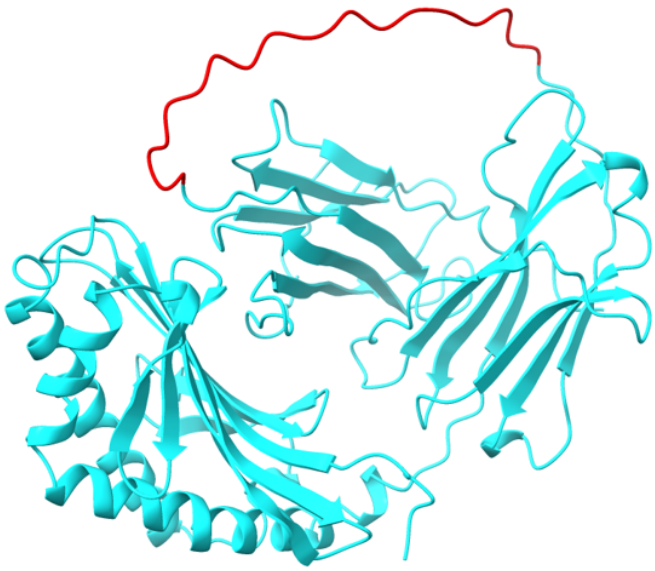


Figura 5.2: Estructura predicha para 7KP1 con OmegaFold mediante unión por linker, en color rojo.

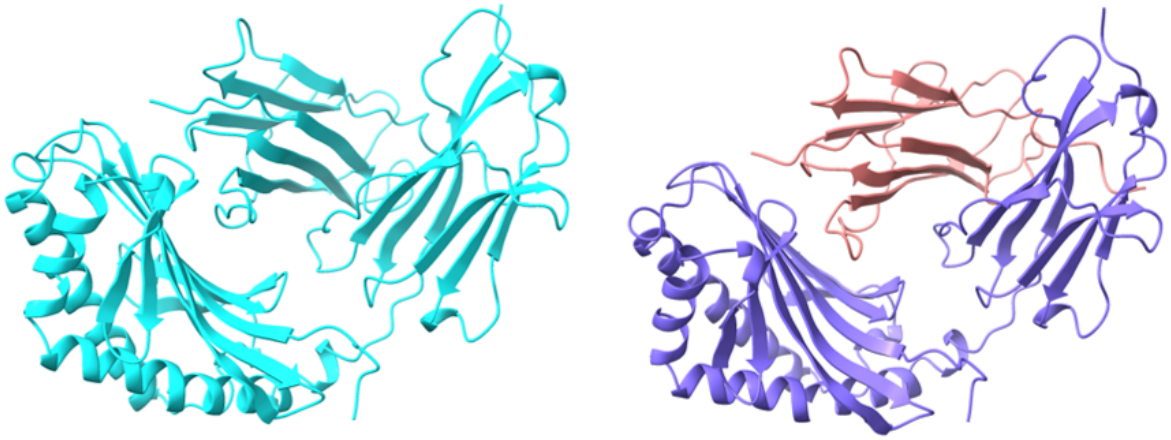


Figura 5.3: A la izquierda la estructura que se obtiene luego de quitar el linker, a la derecha las cadenas que forman el complejo según color, PDB ID: 7KP1.

En el caso de los complejos experimentales, se utiliza ChimeraX para quitar los ligandos pequeños que no se evaluan en la comparación y almacenar los PDBs.

Al realizar las predicciones con ColabFold no sólo se obtienen los archivos PDB con la estructura predicha, también se obtienen métricas importantes para interpretar la estructura obtenida, estas son la cobertura de secuencias (Sequence Coverage), el error de alineamiento predicho (Predicted Aligned Error. PAE) y el pLDDT (Predicted Local Distance Difference Test).

La cobertura de la secuencia indica cuantas secuencias en la busqueda de MSA se encontraron para cada posición de las secuencias, generalmente, para indicar buena cobertura esta debe ser mayor a 30 secuencias. También indica la identidad de cada secuencia, es decir, que tan similares son las secuencias que encontró en la búsqueda de MSA, esto se puede visualizar en la Figura 5.4, los colores azules indican gran similitud, y en el caso en que las líneas horizontales son moradas, indica que encontró una secuencia idéntica. La línea vertical indica que son secuencias diferentes, estas son concatenadas al mostrar el gráfico.

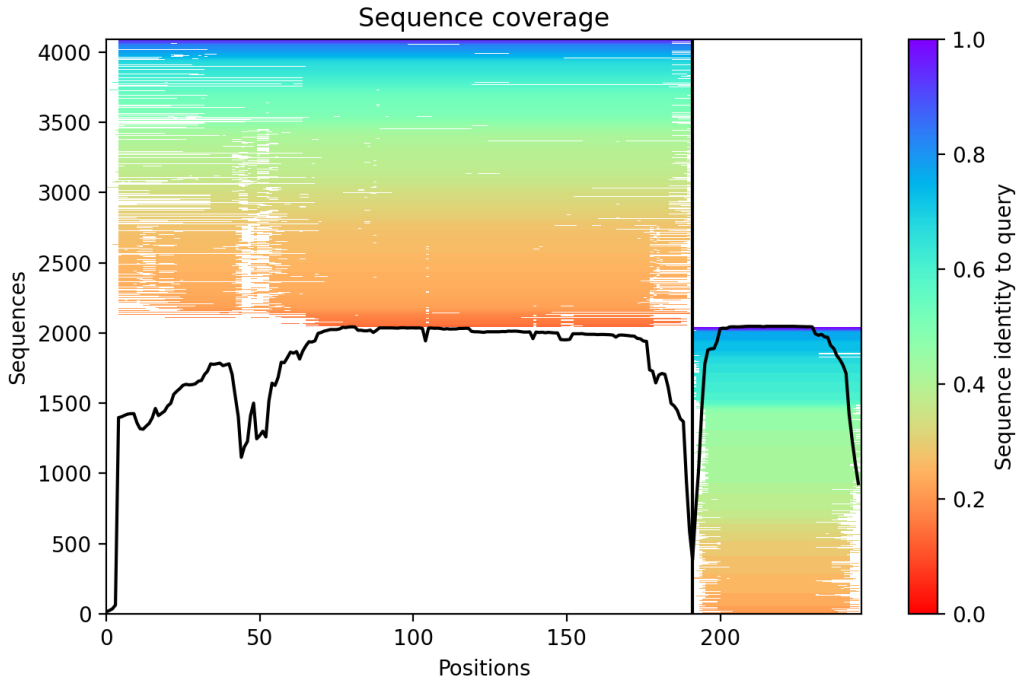
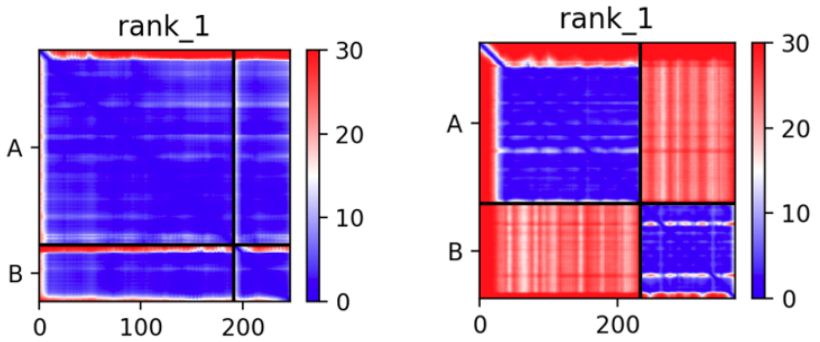


Figura 5.4: Cobertura de secuencia con alta identidad y gran cantidad de secuencias, PDB ID : 6WUD.

El PAE es una escala en Å que indica el error de posición esperado de un residuo x alineado con un residuo y, donde x e y corresponden a la posición horizontal y vertical en el gráfico. Mientras más bajo el valor, más confianza existe en que la posición relativa y orientación entre los residuos es la correcta.

En la Figura 5.5.a se puede visualizar el PAE que genera ColabFold para el mejor modelo, en este caso existe un PAE bajo en general para cada dominio, A y B, incluyendo la posición relativa entre ambos dominios. Esto se puede contrastar con la Figura 5.5.b, en la que si bien existe un PAE bajo en una gran región de cada dominio, el PAE entre las posiciones relativas entre los dominios es alta.



(a) PAE bajo (b) PAE alto

Figura 5.5: (a) El error de alineamiento predicho para el mejor modelo generado (rank 1) por ColabFold, PDB ID: 6WUD. (b) Error de alineamiento predicho para 7EOW, existe un alto PAE entre las posiciones relativas de los dominios.

El pLDDT es una puntuación que indica que tanta confianza existe en la posición tridimensional asignada para el residuo, en los archivos PDB generados, este valor se almacena en la columna bfactor de cada átomo de la estructura.

Generalmente las regiones de pLDDT se interpretan como:

- **$90 \leq$ pLDDT ≤ 100 :** Alta confianza y se espera alta precisión en el modelamiento.
- **$70 \leq$ pLDDT < 90 :** Se espera un modelamiento "bueno" y alta confianza en la vértebra principal de la cadena.
- **$50 \leq$ pLDDT < 70 :** Regiones de baja confianza y se deben interpretar con precaución.
- **$0 \leq$ pLDDT < 50 :** Regiones de muy baja confianza.

En la Figura 5.6 se puede visualizar el pLDDT para las secuencias de 6WUD, de los 5 modelos generados por ColabFold, del terminal N al terminal C, donde la mejor puntuación corresponde a rank 1, esto se puede visualizar en la estructura tridimensional (ver Figura 5.7) mediante el coloreado según el valor de pLDDT, donde el color azul indica alta confianza y el color amarillo baja confianza.

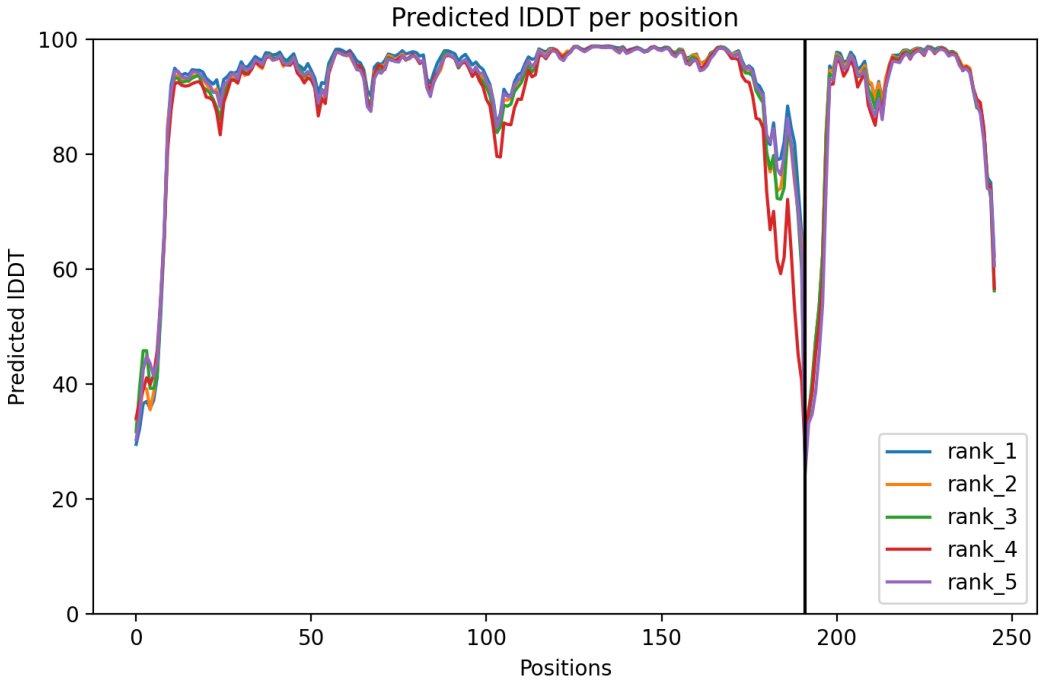


Figura 5.6: pLDDT para los 5 modelos generados por ColabFold, el mejor modelo es el que tiene mejor pLDDT, y se asigna al rango 1 (rank 1), PDB ID:6WUD.

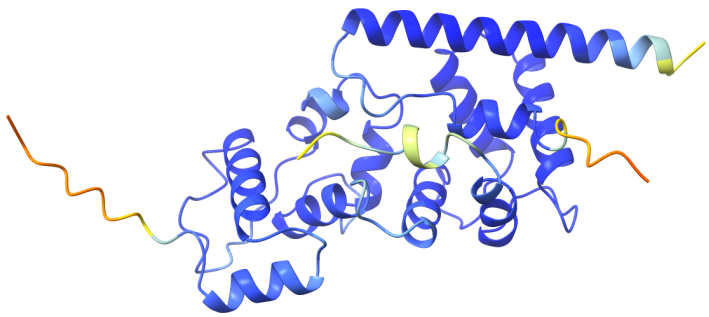


Figura 5.7: Complejo 6WUD coloreado según pLDDT, azul indica regiones con alta confianza.

5.5. DockQ

Para evaluar la calidad estructura de modelos obtenidos por docking se utilizan tres parámetros, Fnat, LRMS y iRMS, estos fueron estandarizados por CAPRI, que es un experimento a nivel global para evaluar las capacidades de métodos de docking de proteínas para predecir interacciones. Debido a la necesidad de poder evaluar la calidad de los modelos en base a los parámetros de CAPRI en una sola función numérica, surge DockQ, que consiste en una forma de medir la calidad de los modelos proteína-proteína obtenidos por docking, combinando los valores de F_{nat}, LRMS y iRMS en único valor en el rango 0 a 1, que puede ser utilizado para juzgar la calidad de los modelos obtenidos por docking (Basu y Wallner, 2016).

Para calcular los parámetros de medición, la interfaz entre dos proteínas interactuantes se define como cualquier par de átomos pesados entre las dos proteínas que se encuentran a 5 $\AA$ entre si.

- **Fnat:** Se define como la fracción de los contactos nativos de la interfaz que se preservan en la interfaz del complejo predicho.
- **LRMS:** Es la raíz de la desviación cuadrática del ligando, (Ligand Root Mean Square) calculada para la vértebra de la cadena más corta del complejo después de la superposición de la cadena más larga.
- **iRMS:** Corresponde raíz de la desviación cuadrática de la interfaz (interface Root Mean Square), la interfaz receptor-ligando se redefine en el objetivo (nativo) con un limite de contacto atómico de 10 Å, doble del valor utilizado para interfaz en el caso de Fnat. Los átomos de los residuos de la vértebra de esta interfaz se superpone en sus residuos equivalentes en el complejo predicho (modelo) para luego calcular la iRMS.

Luego el puntaje DockQ se define como:

$$DockQ(F_{nat}, LRMS, iRMS, d_1, d2_2) = \frac{(F_{nat} + RMS_{scaled}(LRMS, d_1) + RMS_{scaled}(iRMS, d_2))}{3} \tag{5.1}$$

Donde RMS_{scaled} (Ecuación 5.2) representa desviaciones escaladas correspondientes para LRMS o iRMS, y d_i es un factor de escala, d_1 es para LRMS y d_2 es para iRMS, optimizados para d_1=8.5 $\AA$ y d_2=1.5 $\AA$ respectivamente.

$$RMS_{scaled}(RMS, d_i) = \frac{1}{1 + (\frac{RMS}{d_i})^2} \tag{5.2}$$

Según la puntuación DockQ, la predicción se puede clasificar como se indica en la Tabla 5.2. Una

Valor	Clasificación
$0{,}00 \leq \text{DockQ} < 0{,}23$	Incorrecto
$0{,}23 \leq \text{DockQ} < 0{,}49$	Calidad aceptable
$0{,}49 \leq \text{DockQ} < 0{,}80$	Calidad mediana
$0{,}80 \leq \text{DockQ} \leq 1{,}00$	Calidad alta

Tabla 5.2: Clasificación de predicción según valor de DockQ.

puntuación sobre 0,90 se puede considerar a un nivel de calidad ya experimental.

Para efectuar este cálculo sobre la predicción y la estructura experimental se utiliza el código del repositorio de DockQ, esta además incluye un script para realizar un alineamiento de secuencias entre ambas estructuras y corrige los residuos que no se encuentran alineados apropiadamente en las estructuras, a modo de llevar a cabo la comparación entre la estructura experimental y la predicción.

Una vez ejecutado el software sobre la predicción y la estructura experimental, se obtienen los resultados de Fnat, Fnonnat, iRMS, LRMS y DockQ. Este proceso se repite para todas las estructuras seleccionadas del Protein Data Bank, y el total de estos resultados es comparado, para evaluar cual es el método que mejor realiza predicciones.

5.6. Dataset positivo y negativo de interacciones proteína-proteína

El dataset positivo corresponde a complejos con pares de proteínas que interactúan y el dataset negativo corresponde a pares de proteínas que se han catalogado que no interactúan.

- El dataset positivo (ver Tabla B.1) está compuesto por los complejos predichos por el mejor método de predicción, evaluando tanto la puntuación de DockQ promedio, como los tiempos de ejecución computacionales.
- El dataset negativo (ver Tabla B.2) está compuesto por 20 proteínas no interactuantes, 10 de estos pares son del organismo Escherichia coli, del dataset compuesto de 3987 pares proteicos no interactuantes de “Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences” (Green et al., 2021), y 10 pares son de ”The Negatome Database 2.0"(Blohm et al., 2013).

Estos pares de proteínas no interactuantes son modeladas con el método que haya obtenido la mejor evaluación final.

5.7. Predicción de interacciones

Para predecir si dos proteínas interactúan o no en los datasets, se hace uso de la función pDockQ (ver Ecuación 5.3) formulada en el paper "Improved prediction of protein-protein interactions using Alphafold2" (Bryant et al., 2022).

$$pDockq = \frac{L}{1 + e^{-k(x-x_0)}} + b \tag{5.3}$$

Donde x = promedio de la puntuación pLDDT de la interfaz·log(número de contactos en la interfaz) y los parámetros calculados corresponden a $L = 0{,}774$, $x_0 = 152{,}611$, $k = 0{,}052$ y $b = 0{,}018$. Utilizando el rango de DockQ en que los modelos se consideran aceptables, para verificar si dos proteínas interactúan o no, se tomó el valor 0,23 como umbral de posible interacción, un valor sobre este umbral indica que las proteínas si interactúan, y un valor bajo el umbral indica que las proteínas no interactúan.

5.8. Aplicación de pDockQ

Utilizando como framework Angular (Jain et al., 2014) y FastAPI (framework en lenguaje Python de backend), se desarrolló una aplicación web (ver Figura 5.8) en la que se puede subir un archivo PDB obtenido de predicción, que debe contener sólo dos cadenas, para luego realizar los cálculos sobre la estructura, con énfasis en la interfaz de interacción, para obtener los resultados de pDockQ. Además, se muestran los residuos que forman parte de la interfaz de interacción, definida como los residuos que contienen carbonos alfa a menos de 8 Å, y luego se calcula la menor la distancia entre los átomos que forman parte del residuo con el residuo de la otra proteína. Esto se logra usando la librería de python Biopython (Cock et al., 2009).

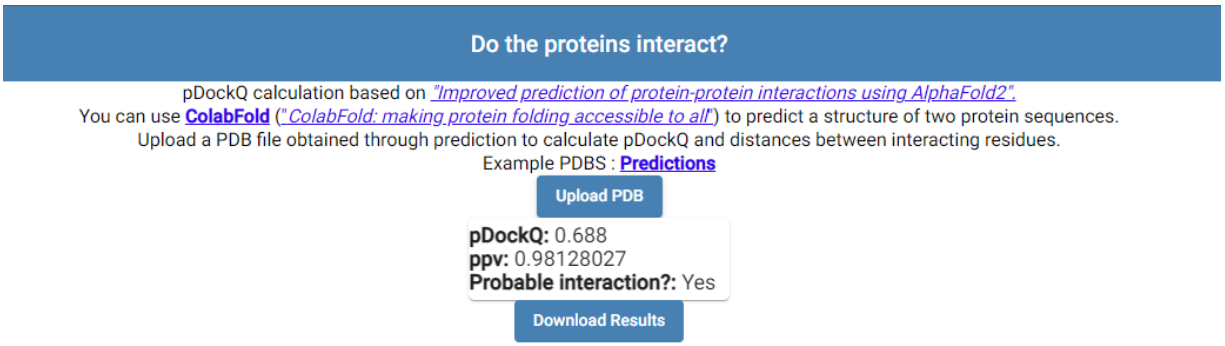
Luego los resultados de interfaz de interacción se pueden descargar como un archivo csv. Todos los códigos utilizados y predicciones obtenidas se encuentran en GitHub.

Repositorios:

https://github.com/orgs/BioCodeLabs/repositories

Aplicación:

http://146.83.194.142:1120/



> Residue 1	Position 1	Confidence 1	Residue 2	Position 2	Confidence 2	Distance
TYR	38	96.32%	VAL	104	85.24%	3.2366014
LEU	39	97.13%	VAL	104	85.24%	4.150605
LEU	39	97.13%	GLY	105	92.03%	6.6864057
THR	40	97.4%	VAL	104	85.24%	2.831015
THR	40	97.4%	GLY	105	92.03%	3.4984367

Figura 5.8: Predicción de posible interacciones a partir de archivo PDB y residuos en interfaz de interacción.

Capítulo 6

Resultados

En este capítulo se describen los resultados obtenidos en la comparación de puntuaciones DockQ de los métodos y la evaluación sobre dataset positivo y negativo de interacciones proteína-proteína, además de los ajustes realizados en cuanto a los métodos a comparar durante el transcurso de la metodología establecida.

6.1. Comparación de Predicciones

Durante la etapa de predicciones, se decidió suspender la utilización de los métodos RosettaFold y DGMFold, debido a los altos tiempos de espera del servidor Robetta, tomando aproximadamente tres semanas o más en entregar resultados, y los distintos resultados outlier que estaba entregando el servidor DGMFold.

Esto conllevó a que la comparación se realizara con Alphafold, ColabFold unpaired+paired y paired, Omegafold y ESMFold, con cada uno de estos métodos se logro predecir los 20 complejos elegidos en la etapa de selección.

El mejor promedio DockQ lo obtuvo Alphafold, con un promedio de 0,67, seguido por colabfold unpaired+paired con un promedio de 0,65. La tasa de éxito se define como el número de predicciones que superan el umbral en que se considera una estructura aceptable por DockQ, que corresponde a 0,23. La tasa de éxito para Alphafold es 92,86 % y la tasa de éxito de colabfold unpaired+paired es 80,00 %. Los resultados obtenidos según calidad se pueden visualizar en la Figura 6.1

Si bien Alphafold obtuvo los mejores resultados en cuanto a puntajes DockQ, se decidió utilizar

Método de Predicción	Tiempo Estimado (450 Residuos)	Clasificación
Alphafold	2 Horas	Alto
Colabfold	20 Minutos	Medio
OmegaFold	3 Minutos	Bajo
ESMFold	3 Minutos	Bajo

Tabla 6.1: Tiempos de ejecución aproximados en servidor de Google con 12 GB RAM y una GPU NVIDIA T4, para 450 residuos.

ColabFold unpaired+paired para realizar el análisis de interfaz de interacción en dataset positivo y negativo, esto debido a los bajos tiempos de ejecución computaciones (ver Tabla 6.1) en los servidores de Google Colab, y también a que los resultados son muy similares a Alphafold.

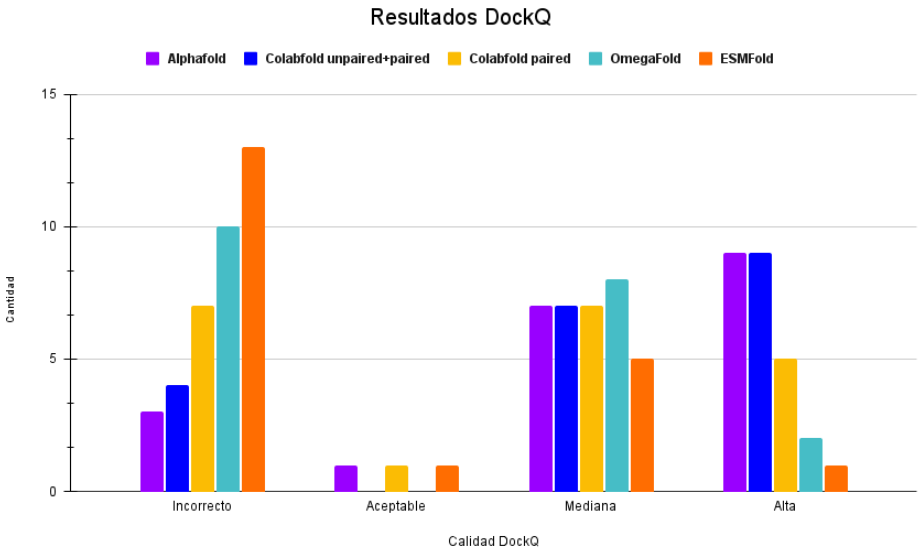


Figura 6.1: Cantidad de estructuras según la calidad obtenida en puntuación DockQ.

Las puntuaciones obtenidas por complejo para Colabfold unpaired+paired se pueden ver en la Tabla 6.2, los demás resultados se pueden encontrar en Anexo (B.3 a B.6)

Dentro de los resultados sólo 4 de los complejos obtuvieron una calidad calificada como incorrecta, se puede observar que en general los resultados producen modelos de calidad aceptable a mediana, con algunos modelos que se acercan a la calidad de un complejo obtenido experimentalmente (6WX1, 7FD0 y 6WUD).

Todos los métodos evaluados son capaces de obtener predicciones de alta calidad (ver Tabla 6.4), es destacable que los métodos OmegaFold y ESMFold pudieron producir complejos con DockQ de 0,834 y 0,854 respectivamente (ver Figura 6.3 (d) y (e)), aún sin tener oficialmente la

PDB ID	Fnat	Fnonnat	iRMS	LRMS	DockQ
7EOW	0,021	0,981	12,140	40,757	0,026
7JOD	0,881	0,342	1,129	4,317	0,772
7KP1	0,864	0,315	3,020	3,496	0,639
6WX1	0,921	0,094	0,578	0,711	0,928
6WRP	0,838	0,326	1,462	0,795	0,795
6ZH1	0,731	0,123	1,102	3,440	0,746
7SH4	0,885	0,353	3,625	4,690	0,599
7RYL	0,845	0,218	1,153	1,708	0,811
7KKH	0,832	0,210	1,179	1,430	0,807
7FDO	0,907	0,235	0,597	0,501	0,922
7EJO	0,152	0,787	12,299	15,052	0,136
7F4Q	0,894	0,218	1,000	1,213	0,855
6TKC	0,802	0,250	0,955	1,178	0,832
6WUD	0,922	0,053	0,442	0,561	0,946
6X28	0,000	1,000	17,506	51,036	0,011
6XOD	0,870	0,041	0,697	3,538	0,848
6ZBK	0,712	0,087	1,039	2,413	0,771
6ZWA	0,774	0,282	1,007	1,756	0,807
7ANQ	0,000	1,000	15,958	53,463	0,011
7B0W	0,830	0,182	1,315	1,547	0,788
				Promedio	**0,653**
				Mediana	**0,792**
				Tasa de exito %	**80,000**

Tabla 6.2: Resultados de Colabfold unpaired+paired

Método	Promedio	Mediana	Tasa de Éxito
Alphafold	0,667	0,797	92,857 %
Colabfold unpaired+paired	0,653	0,792	80,000 %
Colabfold paired	0,497	0,682	71,429 %
OmegaFold	0,359	0,370	42,857 %
ESMFold	0,250	0,074	35,714 %

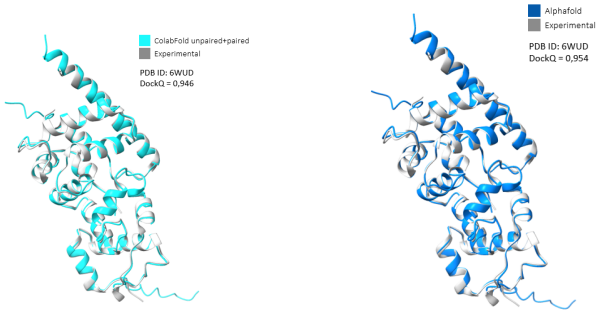
Tabla 6.3: Promedio, mediana y tasa de éxito de los métodos comparados.

cualidad de modelar complejos. Es muy probable que la capacidad para modelar multímeros para estos métodos sea mejorada y permita obtener modelos con alta precisión y con bajo tiempos de ejecución.

Alphafold y ColabFold obtuvieron modelos de calidad experimental para 6WUD (246 residuos modelados), estos se pueden visualizar en la Figura 6.2. Uno de los mejores modelos obtenidos para los 5 métodos es 6WX1 (ver Figura 6.3) , en este caso ambas cadenas del complejo pertenecen al organismo Mus Musculus, y corresponde a un complejo de 438 residuos modelados, mayor a la cantidad promedio de residuos modelados del dataset (333 residuos).

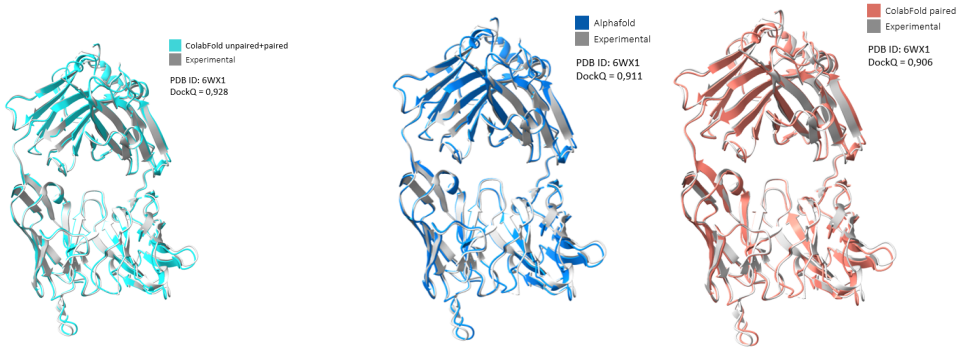
Método	Mejor Predicción	DockQ
Alphafold	6WUD	0,954
Colabfold paired+unpaired	6WUD	0,946
Colabfold paired	6WX1	0,906
OmegaFold	6WX1	0,834
ESMFold	6WX1	0,854

Tabla 6.4: Mejores predicciones obtenidas por cada uno de los métodos junto a la puntuación pDockQ.

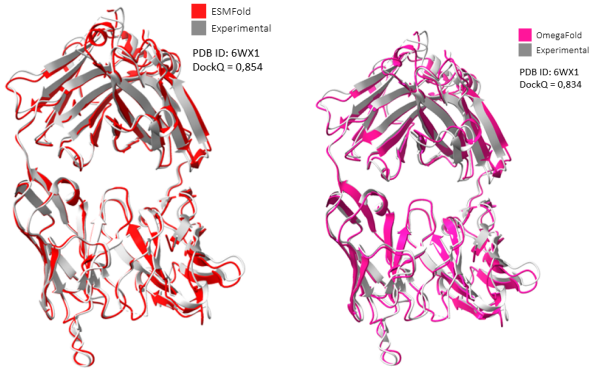


(a) (b)

Figura 6.2: (a) Predicción obtenida para 6WUD por ColabFold unpaired+paired, este corresponde al mejor modelo obtenido. (b) Predicción obtenida para 6WUD por Alphafold, este corresponde al mejor modelo obtenido.



(a) Colabfold vs Exp. (b) Alphafold vs Exp. (c) Colabfold paired vs Exp.



(d) ESMFold vs Exp. (e) OmegaFold vs Exp.

Figura 6.3: Comparación entre los 5 métodos para el complejo 6WX1.

En la Figura 6.3 se pueden observar los modelos obtenidos para 6WX1:

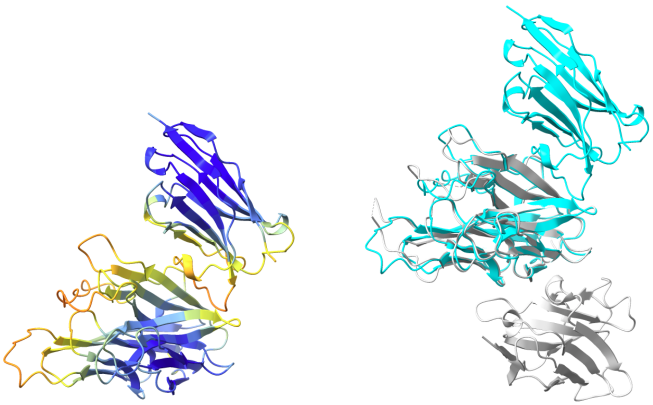
(a) Predicción obtenida para 6WX1 por ColabFold unpaired+paired, este modelo obtiene el mejor puntaje DockQ entre los 5 métodos.
(b) Predicción obtenida para 6WX1 por AlphaFold.
(c) Predicción obtenida para 6WX1 por ColabFold paired.
(d) Predicción obtenida para 6WX1 por ESMFold, este corresponde al mejor modelo obtenido por ESMFold, con una puntuación DockQ igual a 0,854.
(e) Predicción obtenida para 6WX1 por OmegaFold, este corresponde al mejor modelo obtenido por OmegaFold, con una puntuación DockQ igual a 0,834.

6.2. Predicción de interacciones Proteína-Proteína

La función pDockQ predice correctamente la no-interacción en un 90 % de los pares de proteínas del dataset negativo (ver Tabla B.7), reconociendo como interacciones positivas dos pares de proteínas, específicamente los ID NEGATIVEH01 y NEGATIVEH01. Estos dos pares de proteínas pertenecen a las interacciones negativas registradas de The Negatome 2.0, y ambas pertenecen al organismo Homo Sapiens. Se puede observar que para NEGATIVEH01 el promedio pLDDT de los residuos pertenecientes a la interfaz (IpLDDT) es de 84,41 % (ver Tabla B.10), lo que es alto comparado a los IpLDDT en general de los pares de proteínas negativos. Para NEGATIVEH08 si bien el promedio de IpLDDT es 67,14 % y es bajo comparado con el promedio de una interfaz positiva, en la interfaz de interacción se encuentran 70 residuos, lo que conlleva a que el pDockQ supere el umbral de interacción levemente.

Cabe destacar que sobre los pares de proteínas de Escherichia Coli el puntaje pDockQ se ajusta correctamente a los 10 pares no interactuantes, sin embargo, es probable que esto se deba a que la función fue calibrada con proteínas no interactuantes de Escherichia Coli. Para el dataset positivo el pDockQ igualmente predice correctamente la interacción en el 90 % de los casos, con alta confianza en todos estos pares.

Los complejos en que la predicción no se hizo correctamente corresponden a 7ANQ y 7EOW, para 7ANQ se puede ver que en promedio la interfaz de interacción tiene un valor bajo, lo que contribuye a que la interacción sea detectada como negativa según pDockq, sin embargo, el factor de fondo que contribuye a esta situación se puede ver en la Figura 6.4, ya que la interfaz de interacción modelada no es la correcta, las posiciones de ambas proteínas del complejo no se encuentran modeladas correctamente según la estructura cristalográfica experimental.



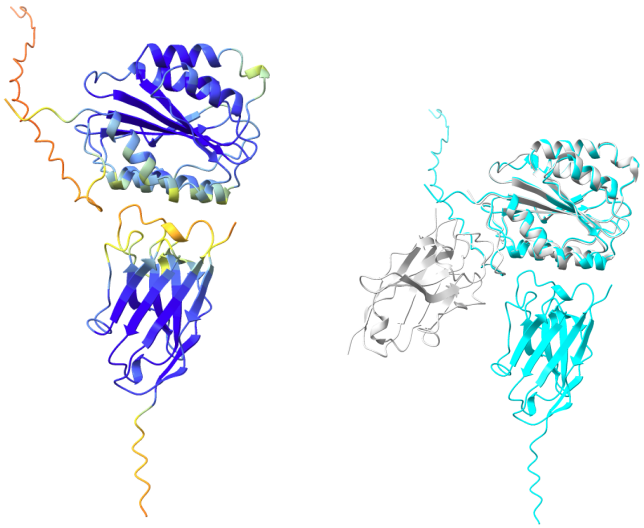
(a) 7ANQ pLDDT (b) 7ANQ + E. Exp.

Figura 6.4: (a) Predicción obtenida para 7ANQ, el pLDDT en la interfaz es bajo. (b) Estructura obtenida para 7ANQ (celeste) superpuesta con estructura experimental (gris) de 7ANQ, en el modelo obtenido la cadena B no se encuentra en la posición correcta.

Una situación similar ocurre con 7EOW, la interfaz tiene baja confianza, pero se encuentra en la posición incorrecta (ver Figura 6.5), este complejo tiene otra peculiaridad que puede contribuir al modelamiento incorrecto, la cadena B corresponde a una proteína diseñada, uno de los primeros nanocuerpos utilizados para medicina, llamado Caplacizumab.

En la Figura 6.5:

(a) Predicción obtenida para 7EOW, el pLDDT en la interfaz es bajo y se pueden observar secuencias en espiral con baja confianza.
(b) Estructura obtenida para 7EOW (celeste) superpuesta con estructura experimental (gris) de 7EOW, en el modelo obtenido la cadena B (Caplacizumab) no se encuentra en la posición correcta.



(a) 7EOW pLDDT (b) 7EOW + E. Exp.

Figura 6.5: Predicción obtenida para 7EOW, a la izquierda la estructura coloreada según su pLDDT y la derecha se encuentra junto a la estructura experimental (gris).

El resto de las interacciones del dataset positivo son predichas correctamente, además, los residuos participantes en la interfaz de interacción concuerdan en su mayor parte en la que se documenta en la literatura, lo que da buenos prospectos futuros a medida que mejore el sistema de Alphafold para el modelamiento de complejos y detección de interacciones con mayor precisión.

Estas interacciónes se pueden obtener a nivel de los átomos de los residuos interactuantes de ambas cadenas con la aplicación realizada, sin embargo, las distancias calculadas pueden ser distintas a las establecidas en los complejos experimentales, se espera que a medida que se siga actualizando ColabFold y Alphafold, los complejos multiméricos sean modelados con distancias más precisas en la interfaz.

6.3. Discusión

En el mes de diciembre de 2022 Alphafold tuvo su actualización a la versión 2.3, con un énfasis en la mejora de predicción de complejos, y fue reentrenado con estructuras del Protein Data Bank liberadas hasta el 30 de septiembre del 2021, además de aumentar la cantidad de residuos que poseen los PDBs en el entrenamiento. Es probable que esto mejore la predicción de complejos de mayor tamaño, pero en este proyecto se utilizó la versión previa. Hubiera sido interesante poder obtener el dataset de comparación con RosettaFold, pero lamentablemente esto no se

pudo completar, debido a la alta demanda que tiene el servidor Robetta y las largas colas de espera.

En cuanto a las interacciones, si bien los resultados obtenidos en dataset positivo y negativo tienen una alta tasa de éxito, cabe destacar que en estos datasets los pares de proteínas interactuantes se consideraron sólo como un complejo dimérico, es decir, si se quiere investigar si dos cadenas son capaces de interactuar en un complejo de más de dos cadenas, es probable que los resultados no sean confiables.

Además, el valor pDockQ no se debe utilizar estrictamente como una forma de predecir con exactitud si dos pares de proteínas interactúan o no, la interpretación de este valor se debe asociar principalmente a la confianza que existe en el complejo dimérico modelado, si se quiere examinar la confianza de interfaz de interacción esto se debe complementar con el PAE obtenido por ColabFold, junto a un análisis sobre la estructura obtenida.

Alphafold recomienda utilizar PAE para estudiar las interfaces de interacción, y una acotación importante que se debe mencionar es la interpretación de las regiones con baja confianza (bajo pLDDT), se ha estudiado que en una gran cantidad de caso estas regiones sirven para identificar regiones en las proteínas que son intrínsecamente desordenadas.

Finalmente, es posible que al modelar complejos de más de dos cadenas con ColabFold, se pueda obtener si dos cadenas interactúan, junto a la interfaz de interacción. Esto requiere modificar la función pDockQ utilizada en este proyecto, y esto permitiría evaluar distintos pares de cadenas en un mismo complejo. En investigaciones futuras se podría abordar este problema, enfocándose en un sólo método de predicción, lo que permitiría obtener un dataset más grande tanto negativo como positivo.

Capítulo 7

Conclusión y Trabajo Futuro

En este proyecto de título se compararon distintas técnicas para modelar estructuras proteicas, con énfasis en la predicción de complejos proteicos. Se utilizaron métodos capacitados para predecir complejos, como también métodos diseñados para proteínas de una sola cadena, en este caso, utilizando un linker para ver si existe la posibilidad de modelar complejos.

Usando ColabFold, se realizó un estudio enfocado a la interfaz de interacción entre los pares de cadenas modelados, y se creó una pequeña aplicación para evaluar la calidad de los modelos predichos.

Los resultados obtenidos muestran que las actuales técnicas para modelar complejos de proteínas son muy poderosas, incluyendo las técnicas que no fueron diseñadas para complejos, y el rápido avance que ha tenido este campo desde la llegada de Alphafold promete buenos resultados futuros en las predicciones de complejos más grandes.

En esta investigación se utilizaron complejos que poseen de 232 a 457 aminoácidos, por lo que sería interesante a futuro poder realizar predicciones de complejos con mayor cantidad aminoácidos y más cadenas, utilizando ColabFold o bien implementando una versión local.

Alphafold ha revolucionado y generado impacto en casi todas las áreas de investigación de la Biología, y además ha revolucionado el campo de la inteligencia artificial utilizando conceptos novedosos como los mecanismos de atención en deep learning. Es probable que en la próxima década la incorporación de nuevos cálculos a los métodos de machine learning existentes, y los nuevos que surjan, permitan acercarnos a la solución al problema del plegamiento de las proteínas.

7.1. Trabajo Futuro

De acuerdo a los resultados obtenidos y temas investigados durante el transcurso proyecto, surgen varias lineas interesantes de investigación futura. Los resultados de esta investigación pueden ser complementados con los resultados en las distintas categorías de CASP15, cuyos resultados fueron publicados en el mes de diciembre del 2022.

ColabFold sigue siendo una herramienta muy poderosa y que da resultados muy precisos, sin necesidad de disponer equipo computacional avanzado, también existe la versión local, que es posible implementar en un computador o en un clúster. A continuación se describen posibles temas de investigación para trabajo futuro:

- **Función pDockQ para varias cadenas:** Implementar la función pDockQ para realizar cálculos de interacciones proteína-proteína con proteínas de más de dos cadenas, esto seria modificar la función actual y agregar la capacidad para examinar las cadenas interacción y estudiar cuales son las que participan en la interacción, una vez modificada la función obtener la tasa de éxito de la función sobre un dataset de proteínas interactuantes multiméricas, utilizando ColabFold.
- **Cálculo de PPI en base a pLDDT y PAE:** Implementar una función que contenga un módulo de cálculo de interacción similar a pDockQ, en base a el pLDDT de los dominios e interfaz de interacción, sumando a un módulo de cálculo en base al error de alineamiento predicho, con énfasis en el error de la interfaz de ambos dominios, utilizar dataset positivo y negativo para calibrar la nueva función.
- **Diseño de proteínas y anticuerpos con ColabDesign:** Estudiar sitios de interacción de proteínas de interés y usando ColabDesign u otros métodos del estado del arte, diseñar estructuras que se puedan unir a estos sitios de interacción.

Referencias

Saliha Ece Acuner Ozbabacan, Hatice Billur Engin, Attila Gursoy, y Ozlem Keskin. Transient protein-protein interactions. *Protein Eng Des Sel*, 24(9):635–648, 2011.

Lewis J et al Alberts B, Johnson A. *Molecular Biology of the Cell. 4th edition. New York: Garland Science.* 2002.

C B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, y David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021. doi:10.1126/science.abj8754. URL `https://www.science.org/doi/abs/10.1126/science.abj8754`.

Sankar Basu y Björn Wallner. Dockq: A quality measure for protein-protein docking models. *PLoS ONE*, 11:e0161879, 2016. doi:10.1371/journal.pone.0161879.

Eric Bell, Jacob Schwartz, Peter Freddolino, y Yang Zhang. Peppi: Whole-proteome protein-protein interaction prediction through structure and sequence similarity, functional association, and machine learning. 2021. doi:10.1101/2021.12.02.470917.

Sourangshu Bhattacharya. Computational protein structure analysis : Kernel and spectral methods. 2022.

Philipp Blohm, Goar Frishman, Pawel Smialowski, Florian Goebels, Benedikt Wachinger, Andreas Ruepp, y Dmitrij Frishman. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res*, 42(Database issue):D396–400, 2013.

Patrick Bryant, Gabriele Pozzati, y Arne Elofsson. Improved prediction of protein-protein interactions using alphafold2. *Nature Communications*, 13(1):1265, 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-28865-w. URL `https://doi.org/10.1038/s41467-022-28865-w`.

Ewen Callaway. The revolution will not be crystallized: a new method sweeps through structural biology. *Nature*, 525(7568):172–174, 2015. ISSN 1476-4687. doi:10.1038/525172a. URL `https://doi.org/10.1038/525172a`.

Samuel Chackalamannil, David Rotella, Simon E. Ward, Ana Martinez, Carmen Gil, Helmut Buschmann, Norbert Handler, Andrea Wolkerstorfer, Andrew M. Davis, Colin Edge, y et al. *6.03.06*. Elsevier, 2017.

François Chollet. *Deep Learning with Python*. Manning, 2017. ISBN 9781617294433.

Peter JA Cock, Tiago Antao, Jeffrey T Chang, Brad A Chapman, Cymon J Cox, Andrew Dalke, Iddo Friedberg, Thomas Hamelryck, Frank Kauff, Bartek Wilczynski, et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

Peter W Collingridge y Steven Kelly. MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. *BMC Bioinformatics*, 13:117, 2012.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489, 2020. ISSN 0305-1048. doi:10.1093/nar/gkaa1100. URL `https://doi.org/10.1093/nar/gkaa1100`.

M S Coumar. *Molecular docking for computer-aided drug design: Fundamentals, Techniques, resources and applications*. Academic Press, 2021.

David Eisenberg. The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. *Proc Natl Acad Sci U S A*, 100(20):11207–11210, 2003.

Narayanan Eswar, Ben Webb, Marc A Marti-Renom, M S Madhusudhan, David Eramian, Min-Yi Shen, Ursula Pieper, y Andrej Sali. Comparative protein structure modeling using modeller. *Curr Protoc Bioinformatics*, Chapter 5:Unit–5.6, 2006.

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John Jumper, y Demis Hassabis. Protein complex prediction with alphafold-multimer. 2022. doi:10.1101/2021.10.04.463034. URL `https://www.biorxiv.org/content/early/2022/03/10/2021.10.04.463034`.

György G Ferenczy y Miklós Kellermayer. Contribution of hydrophobic interactions to protein mechanical stability. *Comput Struct Biotechnol J*, 20:1946–1956, 2022.

J S Fetrow. Omega loops: nonregular secondary structures significant in protein function and stability. *FASEB J*, 9(9):708–717, 1995.

W.T Godbey. *3.2.4 Quaternary structure*. Academic Press, 2021.

Ian Goodfellow, Yoshua Bengio, y Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Anna G. Green, Hadeer Elhabashy, Kelly P. Brock, Rohan Maddamsetti, Oliver Kohlbacher, y Debora S. Marks. Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. *Nature Communications*, 12(1):1396, 2021. ISSN 2041-1723. doi:10.1038/s41467-021-21636-z. URL `https://doi.org/10.1038/s41467-021-21636-z`.

Kristoffer Illergård, David H. Ardell, y Arne Elofsson. Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, 77(3):499–508, 2009. doi:https://doi.org/10.1002/prot.22458. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.22458`.

Nilesh Jain, Ashok Bhansali, y Deepak Mehta. Angularjs: A modern mvc framework in javascript. *Journal of Global Research in Computer Science*, 5(12):17–23, 2014.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, y Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596:583–589, 2021. ISSN 1476-4687. doi:10.1038/s41586-021-03819-2.

J C Kendrew, G Bodo, H M Dintzis, R G Parrish, H Wyckoff, y D C Phillips. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*, 181(4610):662–666, 1958.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, y Alexander Rives. Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*, 2022. doi:10.1101/2022.07.20.500902. URL `https://www.biorxiv.org/content/early/2022/10/31/2022.07.20.500902`.

Jun Liu, Guang-Xing He, Kai-Long Zhao, y Gui-Jun Zhang. De novo protein structure prediction by incremental inter-residue geometries prediction and model quality assessment using deep learning. *bioRxiv*, 2022. doi:10.1101/2022.01.11.475831. URL `https://www.biorxiv.org/content/early/2022/01/12/2022.01.11.475831`.

Scott McGinnis y Thomas L Madden. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res*, 32(Web Server issue):W20–5, 2004.

Milot Mirdita, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, y Martin Steinegger. Colabfold: making protein folding accessible to all. *Nature Methods*, 19(6):679–682, 2022. ISSN 1548-7105. doi:10.1038/s41592-022-01488-1. URL `https://doi.org/10.1038/s41592-022-01488-1`.

Sushil Mishra, Gabriel Demo, Jaroslav Koc?a, y Michaela Wimmerová. *In Silico Engineering of Proteins That Recognize Small Molecules*. 2012. ISBN 978-953-51-0037-9. doi:10.5772/28001.

John Moult, Jan T. Pedersen, Richard Judson, y Krzysztof Fidelis. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Bioinformatics*, 23(3):ii–iv, 1995. doi:https://doi.org/10.1002/prot.340230303. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.340230303`.

David W. Mount. *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, 2004.

David W Mount. Using hidden markov models to align multiple sequences. *Cold Spring Harb Protoc*, 2009(7):db.top41, 2009.

Zhaoyang Niu, Guoqiang Zhong, y Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021. ISSN 0925-2312. doi:https://doi.org/10.1016/j.neucom.2021.03.091. URL `https://www.sciencedirect.com/science/article/pii/S092523122100477X`.

Subrata Pal. *Proteins 5.2.5*, pág. 110–111. Academic Press, 2020.

L Pauling, R B Corey, y H R Branson. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A*, 37(4):205–211, 1951.

James R. Perkins, Ilhem Diboun, Benoit H. Dessailly, Jon G. Lees, y Christine Orengo. Transient protein-protein interactions: Structural, functional, and network properties. *Structure*, 18(10):1233–1243, 2010. ISSN 0969-2126. doi:https://doi.org/10.1016/j.str.2010.08.007. URL `https://www.sciencedirect.com/science/article/pii/S0969212610003035`.

Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Elaine C Meng, Gregory S Couch, Tristan I Croll, John H Morris, y Thomas E Ferrin. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.*, 30(1):70–82, 2021.

K R Rajashankar y S Ramakumar. Pi-turns in proteins and peptides: Classification, conformation, occurrence, hydration and sequence. *Protein Sci*, 5(5):932–946, 1996.

Kunal Roy, Supratik Kar, y Rudra Narayan Das. *Understanding the basics of QSAR for applications in pharmaceutical sciences and risk assessment*. Elsevier/Academic Press, 2015.

Stuart J. Russell y Peter Norvig. *Artificial Intelligence: a modern approach*. Pearson, 3 ed$^{\underline{\text{ón}}}$., 2009.

F Sanger. The terminal peptides of insulin. *Biochem J*, 45(5):563–574, 1949.

Frederick Sanger. The nobel prize in chemistry 1958. 1958. URL `https://www.nobelprize.org/prizes/chemistry/1958/sanger/lecture/`.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, y Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. ISSN 1476-4687. doi:10.1038/nature16961. URL `https://doi.org/10.1038/nature16961`.

Martin Steinegger y Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017. ISSN 1546-1696. doi:10.1038/nbt.3988. URL `https://doi.org/10.1038/nbt.3988`.

Ilya A Vakser. Protein-protein docking: from interaction to interactome. *Biophys J*, 107(8):1785–1793, 2014.

Jukka Westermarck, Johanna Ivaska, y Garry L. Corthals. Identification of Protein Interactions Involved in Cellular Signaling. *Molecular & Cellular Proteomics : MCP*, 12(7):1752, 2013. doi:10.1074/mcp.R113.027771.

Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, y Jian Peng. High-resolution de novo structure prediction from primary sequence. *bioRxiv*, 2022. doi:10.1101/2022.07.21.500999. URL `https://www.biorxiv.org/content/early/2022/07/22/2022.07.21.500999`.

Kurt Wüthrich. NMR studies of structure and function of biological macromolecules (nobel lecture). *J Biomol NMR*, 27(1):13–39, 2003.

wwPDB consortium. Protein data bank: the single global archive for 3D macromolecular structure data. *Nucleic Acids Res*, 47(D1):D520–D528, 2019.

Giovanna Zinzalla y David E. Thurston. Targeting protein–protein interactions for therapeutic intervention: a challenge for the future. *Future Med. Chem.*, 2009. doi:10.4155/fmc.09.12.

Apéndice A

Proteínas

A.1. Estructura de una Proteína

En la Figura A.1 se puede observar como una proteína es una estructura que consiste de una vértebra polipeptídica con cadenas laterales. Cada tipo de proteína difiere en su secuencia y número de aminoácidos. Los dos extremos de una cadena polipeptídica son químicamente diferentes: el extremo con el grupo amino libre se denomina amino terminal o N-terminal (amino terminus o N-terminus), y el extremo con el grupo carboxilo libre se denomina carboxilo terminal o C-terminal (carboxyl terminus o C-terminus). La secuencia de aminoácidos de una proteína siempre se presenta en la dirección N a C (grupo amino a grupo carboxilo), y se lee de izquierda a derecha.

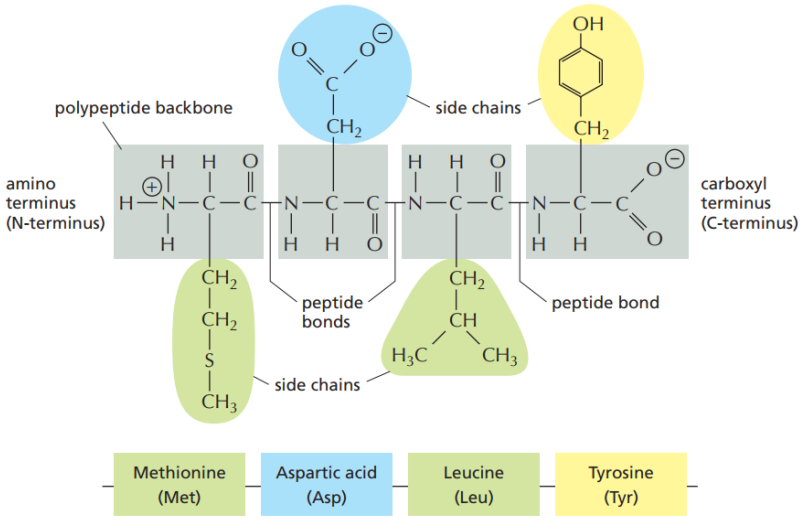


Figura A.1: Una proteína consiste en una vértebra polipeptídica (polypeptide backbone) con cadenas laterales (side chains) (Alberts B, 2002).

A.2. Cadenas Laterales

Algunas de estas cadenas laterales son no polares e hidrofóbicas, otras se encuentran cargadas positiva o negativamente, de esta manera pueden tener distintas características. En la Figura A.2 se puede ver la tabla con los nombres de aminoácidos y sus abreviaciones.

AMINO ACID			SIDE CHAIN
Aspartic acid	Asp	D	negative
Glutamic acid	Glu	E	negative
Arginine	Arg	R	positive
Lysine	Lys	K	positive
Histidine	His	H	positive
Asparagine	Asn	N	uncharged polar
Glutamine	Gln	Q	uncharged polar
Serine	Ser	S	uncharged polar
Threonine	Thr	T	uncharged polar
Tyrosine	Tyr	Y	uncharged polar
POLAR AMINO ACIDS			

AMINO ACID			SIDE CHAIN
Alanine	Ala	A	nonpolar
Glycine	Gly	G	nonpolar
Valine	Val	V	nonpolar
Leucine	Leu	L	nonpolar
Isoleucine	Ile	I	nonpolar
Proline	Pro	P	nonpolar
Phenylalanine	Phe	F	nonpolar
Methionine	Met	M	nonpolar
Tryptophan	Trp	W	nonpolar
Cysteine	Cys	C	nonpolar
NONPOLAR AMINO ACIDS			

Figura A.2: Los 20 aminoácidos que conforman las proteínas y sus abreviaciones (Alberts B, 2002)

.

Dado que cada uno de los 20 aminoácidos existentes es químicamente distinto y cada uno puede, teóricamente, ocurrir en cualquier posición de la cadena de la proteína, hay 20^n combinaciones posibles de cadenas de polipéptidos, donde$n$ es la cantidad de aminoácidos. Las proteínas se encargan de las funciones más esenciales en las células, como la estructura, el transporte, el comportamiento enzimático y regulatorio al interior de esta. Las funciones de las proteínas son generalmente determinadas a través de sus estructuras, que se pueden organizar en cuatro niveles de jerarquía, cada una con mayor complejidad.

El enlace omega tiene un carácter de doble enlace y por lo tanto es casi siempre 180°. La estructura de una proteína es formada principalmente por los ángulos *fi* y *psi*. Los ángulos de torsión se encuentra en un rango específico forzado por la cadena principal debido a los elementos de estructura secundaria, esto se puede visualizar en un diagrama de Ramachandran.

A.3. Estructuras Secundarias

A.3.1. Hélices Alfa

La estructura secundaria más común corresponde a las hélices alfa, son formadas por los enlaces de hidrógeno entre el grupo carbonilo y el grupo amino de diferentes aminoácidos, a una distancia de 4 posiciones de la cadena polipeptídica. Por ejemplo, el grupo carbonilo de un aminoácido en la posición 1 de una cadena puede formar un enlace de hidrógeno con el grupo amino del aminoácido en la posición 5 de la cadena. Este patrón de enlaces hace que la estructura de la cadena polipeptídica forme una estructura helicoidal, donde cada vuelta de la hélice en sentido horario contiene 3.6 aminoácidos (Alberts B, 2002) y cada giro de 100° contiene un aminoácido (360°/3.6=100°). Las hélices también se pueden formar en sentido anti-horario, sin embargo, esto es menos común ya que la estructura es menos estable. Los grupos R de los aminoácidos en esta estructura apuntan hacia afuera de la hélice alfa, donde pueden interactuar libremente.

En la Figura A.3 se puede observar la conformación de una hélice alfa en una cadena polipeptídica, el grupo amino (N-H) de cada enlace peptídico se une con un enlace de hidrógeno al grupo carbonilo (C=O) del enlace péptido ubicado a 4 enlaces péptidos de distancia en la misma cadena. Todos los grupos aminos apuntan hacia arriba en el diagrama y los grupos carbonilo apuntan hacia abajo, es decir, al C-terminal; esto le da polaridad a la hélice, donde la carga negativa se encuentra en el C-terminal y la carga positiva se encuentra en el N-terminal.

Debido a la gran cantidad de enlaces de hidrógeno que se forman siguiendo el patrón helicoidal, los enlaces de hidrógeno corresponden a una de las fuerzas principales de estabilización en esta estructura secundaria.

Cabe destacar que existen distintos tipos de hélices que se pueden formar en las proteínas,en particular, la hélice alfa también se llama hélice 3.6_{13} y existen las hélice pi (Rajashankar y Ramakumar, 1996) y la hélice 3_{10}.

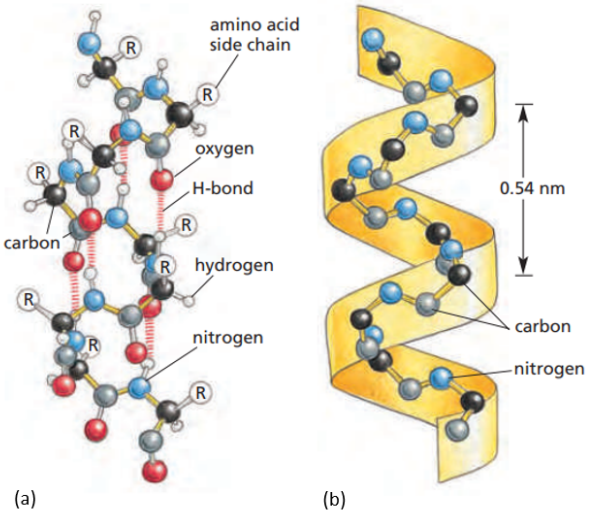


Figura A.3: Conformación de una hélice alfa en una cadena polipeptídica, en (a) se muestran todos los átomos de la cadena polipeptídica y en (b) se muestran solo los átomos de carbono y nitrógeno de la cadena principal (Alberts B, 2002).

A.3.2. Láminas Beta

En una lámina beta, dos o más filamentos de una cadena polipeptídica se encuentran una al lado de la otra, formando una estructura en forma de lámina, unida por enlaces de hidrógeno. Los enlaces de hidrógenos se forman entre los grupos carbonilo y amino de la cadena principal, mientras que los grupos R se extiendo sobre o bajo el plano de la lámina.

Los filamentos de la cadena polipeptídica pueden encontrarse de manera paralela, apuntando hacia la misma dirección, de modo que sus terminales N y C coinciden, o antiparalela, apuntando en direcciones opuestas de modo que el N-terminal de una cadena se encuentra posicionada justo al lado del C-terminal de la otra cadena. En la figura A.4 se observan cadenas peptídicas orientadas de manera opuesta (cadenas antiparalelas). Los enlaces de hidrógeno entre los diferentes filamentos mantienen a las cadenas del polipéptido individual juntos en una lámina beta, y las cadenas laterales de los aminoácidos en cada filamentos se proyectan sobre y bajo el plano de la lámina de manera alternativa.

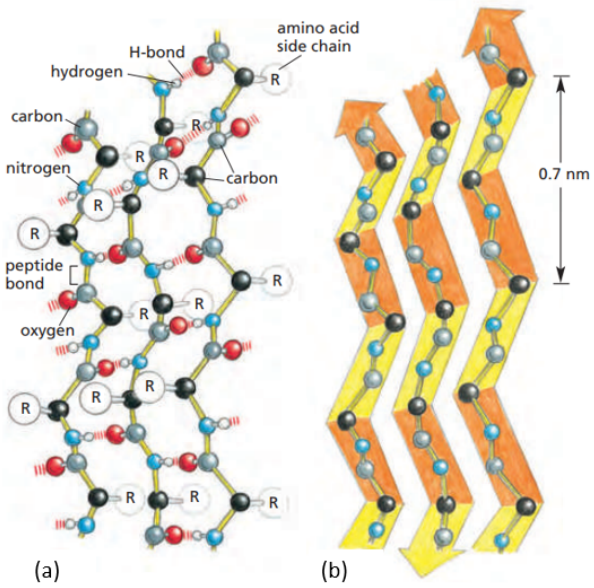


Figura A.4: Conformación de una lámina beta en una cadena polipeptídica, en (C) se muestran todos los átomos de la cadena polipeptídica y en (D) se muestran solo los átomos de carbono y nitrógeno de la cadena principal (Alberts B, 2002).

A.3.3. Giros Beta

Los giros beta son una de las formas estructurales más comunes en las proteínas y permite que la cadena principal cambie de dirección en casi 180°.

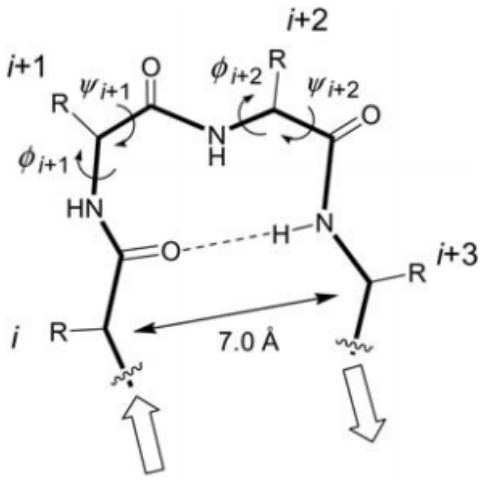
Cada giro consiste de 4 residuos aminoácidos que se pueden nombrar como (i), $(i+1)$, $(i+2)$ y $(i+3)$, y se pueden clasificar según el enlace de hidrógeno entre el grupo carbonilo del aminoácido i y el grupo amino del aminoácido $i+3$, o la distancia entre los átomos C alfa (Carbono alfa, primer átomo de carbono que tiene un grupo funcional) del residuo i y el residuo $i+3$, que debe ser menor a $7\mathring{A}$ (un $\mathring{A}$ equivale a $10^{-10}m$)(Chackalamannil et al., 2017).

Los giros beta también se pueden clasificar según los ángulos de torsión de la cadena principal (ϕ_{i+1} , ψ_{i+1}, ϕ_{i+2} , y ψ_{i+2}) en los residuos $i+1$ y $i+2$.

Usando la clasificación del enlace de hidrógeno para los giros beta surgen 4 categorías, llamadas tipo I, II, I'y I", en la figura A.5 se pueden observar estas clasificaciones y los respectivos ángulos de torsión .

Los tipo I y II son los más comunes y se estabilizan por el enlace de hidrógeno entre el grupo

carbonilo del aminoácido *i* y el grupo amino del aminoácido *i+3*, la diferencia entre estos dos tipos se debe a la orientación del enlace amida entre los aminoácidos *i+1* y *i+2* en el plano en que se encuentra el giro beta. Los tipos I'y I" corresponden a imágenes espejo de las conformaciones I y II respectivamente. Todos estos tipos ocurren regularmente, siendo el tipo I el más común entre estos, debido a que se parece a un hélice alfa, en general ocurre cada 310 hélices y en los extremos de los hélices alfa. Los giros beta tipo II ocurren asociados a las láminas beta, como parte de uniones entre láminas beta.



	ϕ_{i+1}	ψ_{i+1}	ϕ_{i+2}	ψ_{i+2}
type I	-60	-30	-90	0
type I'	-60	120	80	0
type II	60	30	90	0
type II'	60	-120	-80	0

Figura A.5: Estructura y clasificación de los giros beta, (Chackalamannil et al., 2017)

A.3.4. Bucles Omega

Los bucles omega (Omega loops) son formas estructurales no regulares en las proteínas globulares, caracterizadas por una cadena polipeptídica que genera una estructura en forma de bucle en tres dimensiones. No contienen ángulos de torsión que se repitan y no siguen un patrón en sus enlaces de hidrógeno (Fetrow, 1995). Se encuentran generalmente exclusivamente en la superficie de las proteínas. La característica principal de esta estructura es que los residuos del inicio y del fin del bucle se encuentran muy cercanos en el espacio, sin intervalos de distancia regulares que definan otra estructura secundaria.

A.4. Interacciones Estabilizantes

- **Efecto hidrofóbico:** Es la tendencia que existe en las moléculas no polares o hidrofóbicas de asociarse entre al encontrarse en una disolución acuosa (una disolución donde el principal componente es agua) y la tendencia a repeler el agua. Este efecto surge en las proteínas debido a que las secuencias de proteínas son combinaciones de regiones de aminoácidos hidrofóbicos e hidrofílicos.
  Esta fuerza es muy importante en el proceso de plegamiento de las proteínas que tienen aminoácidos hidrofóbicos, en este caso, se forman núcleos hidrofóbicos en el que las cadenas laterales se encuentran en el interior del estado plegado. Las cadenas laterales polares o hidrofílicas se encuentran expuestas en la superficie donde interactúan con las moléculas del agua que rodean a la proteína.
- **Enlaces de disulfuro:** Los enlaces de disulfuro dictan el plegamiento de las proteínas formando fuertes enlaces covalentes entre cadenas laterales de cisteína (C, Cys) que generalmente se encuentran a gran distancia en la estructura primaria. Un enlace de disulfuro no se puede formar entre residuos de cisteínas consecutivos y lo normal es que cada cisteína se encuentre separada por al menos 5 residuos. Los enlaces de disulfuro sólo se rompen a altas temperaturas, pH ácido o en la presencia de agentes reductores.
  En la FiguraA.6 se puede observar la cadena primaria de la proteína somastotina 14 y el enlace disulfuro que se forma entre las cisteínas 3 y 14 en la estructura ternaria (Figura: A.7), las interacciones **no** forman parte de la estructura primaria.
- **Enlaces de hidrógeno:** Los enlaces de hidrógeno contribuyen significativamente a la estabilidad de las hélices alfa y las interacciones de los filamentos beta para formar láminas beta paralelas o antiparalelas. Como resultado, estos enlaces son unos de los principales responsables de la estabilidad de la estructura terciaria o estado plegado. Los enlaces de hidrógeno se forman entre el grupo amino y el grupo carboxilo de la cadena principal, pero existe la posibilidad en el plegamiento para formar enlaces entre grupos de las cadenas laterales.
- **Enlaces de iónicos:** Se forman cuando los átomos de los aminoácidos que poseen cargas eléctricas se encuentran en proximidad. Los enlaces iónicos son importantes para la estructura de la proteína porque son atracciones electrostáticas potentes, y en los interiores hidrofóbicos de las proteínas, estos enlaces se acercan en fuerza a los enlaces covalentes. Estas interacciones ocurren entre las cadenas laterales de residuos cargados eléctricamente opuestamente, al igual que entre los grupos amino y carboxilo cargados a los extremos de la cadena polipeptídica. Las interacciones iónicas importantes corresponden a las cadenas laterales de lisina, arginina, e histidina junto a las cadenas laterales de ácido aspártico y ácido glutámico, y en menor medida tirosina y cisteína.
- **Fuerzas de Van der Waals:**
  Las fuerzas de Van der Waals son fuerzas de atracción y repulsión eléctricas débiles de un átomo a otro. Estas atracciones existen porque todos los átomos tienen nubes eléctricas que pueden fluctuar, lo que forma un dipolo temporal.

La formación de un dipolo temporal puede inducir a un dipolo complementario en otro átomo, si es que los átomos se encuentran cerca.
Si las nubes de electrones de átomos adyacentes se encuentran muy cerca, se formas fuerzas de repulsión debido la carga eléctrica negativa de los electrones.
A diferencia de los enlaces iónicos, estas fuerzas no siguen la ley de los cuadrados inversos y su contribución a la estabilidad varia en la totalidad de las atracciones intermoleculares.
Existen tres contribuciones a la estabilidad de la proteína, en orden decreciente de fuerza:

- El efecto de la orientación entre dipolos permanentes.
- El efecto de inducción entre dipolos permanentes y temporales.
- El efecto de dispersión (fuerza de dispersión de London), que es la interacción entre los dipolos temporales inducidos.

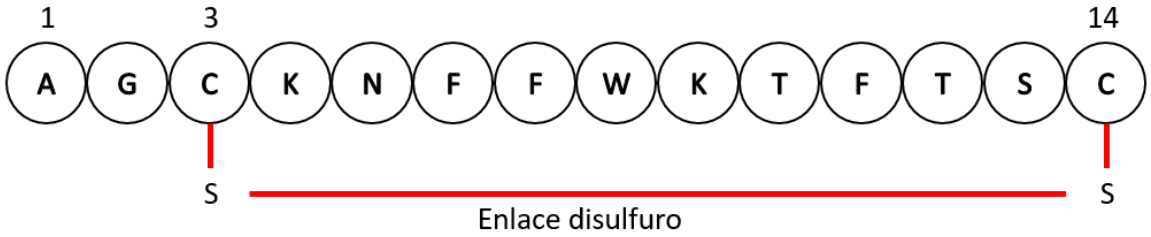


Figura A.6: Secuencia primaria de aminoácidos de proteína somastotina 14 (código PDB: 2MI1), el enlace disulfuro se forma entre los residuos de cisteína (C, Cys) de la posición 3 y la 14.

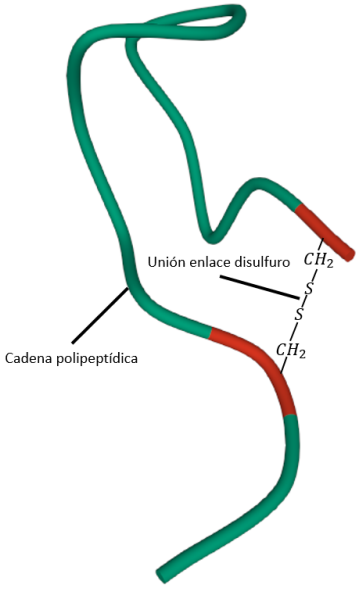


Figura A.7: Cadena polipeptídica de proteína somastotina 14 (código PDB: 2MI1), en color rojo los residuos de cisteína en las posiciones 3 y 14.

A.5. Interacciones en el Plegamiento Proteico

El plegamiento de las proteínas es acelerado y encaminado por una rápida formación de interacciones locales, que subsecuentemente determinan más plegamiento del polipéptido. Las fuerzas ejercidas por los átomos de las cadenas laterales de la cadena polipeptídica son las que guían el plegamiento de las proteínas, ya que estas son las que cambian de una proteína a otra. Las fuerzas en ciertos átomos y grupos de átomos surgen de interacciones con otros átomos de la proteína misma como también moléculas solventes.

Las interacciones pueden presentarse como enlaces de hidrógenos, enlaces iónicos, contactos de van der Waals y el efecto hidrofóbico (mediado por el agua). Existe debate en cuanto a cuál de todas las interacciones es la dominante en el proceso de plegamiento:

- Las interacciones iónicas, como se vio anteriormente en las fuerzas de la estructura ternaria, surgen de las cadenas laterales que se encuentran cargadas eléctricamente. El plegamiento es probable que no esté dominado por estas interacciones ya que la mayoría de las proteínas contienen una cantidad relativamente baja de residuos localizados cargados, y también la estabilidad de las proteínas se ha encontrado que es en su gran parte independiente del pH dentro de un rango y la concentración de sales.
- Los enlaces de hidrógeno fueron considerados como uno de los factores dominantes, ya que casi todas las interacciones de enlaces de hidrógeno son formadas en las estructuras nativas. Los enlaces de hidrógeno entre las amidas y carbonilos de la cadena principal son componentes clave de todas las estructuras secundarias. También se encontró que la fuerza de los enlaces de hidrógeno entre un donante y receptor en una proteína es debilitada por la formación de un enlace de hidrógeno con una molécula de agua disolvente. Por lo que se concluyó que no se pueden considerar como una fuerza significativa en lo que concierne al plegamiento de las proteínas.
- La energía de las interacciones de van der Waals en proteínas plegadas y fuertemente empaquetadas fue encontrada comparable a la de las interacciones hidrofóbicas. Por lo que se concluyó que la fuerzas de van de Waals en el plegamiento de las proteínas es significativo, pero no un factor dominante.
- En cuanto a las interacciones hidrofóbicas, se encontraron dos observaciones que permiten evidenciar a esta interacción como una de las dominantes: (a) la presencia de gran cantidad de energía libre negativa (es decir, la reacción es termodinámicamente favorable hacia los productos) con la transferencia de los solutos no polares del agua a un solvente orgánico y (b) en la mayoría de las proteína nativas con un núcleo hidrofóbico, las cadenas laterales no polares tienden a estar enterradas en el núcleo, alejadas del entorno polar de las moléculas de agua (Pal, 2020).

A.6. Obtención de Estructura Cristalográfica Usando Rayos X

El procedimiento para obtener la estructura en tres dimensiones usando cristalografía de rayos X consiste de esencialmente cuatro pasos:

- Primero al cristal proteico (sólido cristalino que contiene billones de proteínas idénticas) se le hacen incidir haces de rayos X, que, al colisionar con el cristal, se difractan en un patrón especifico basado en la red cristalina.
- Luego el patrón de difracción se usa para desarrollar un mapa de densidades de electrones, de manera en que se puede obtener la información de las posiciones promedio de los átomos en el cristal.
- Se computa y calcula el mapa de densidades de electrones resultante, que es es un promedio de las densidades de electrones de todas las moléculas del cristal.
- Finalmente, mediante la información obtenida por el mapa de densidades de electrones, se generan modelos estructurales que mejor se adecuen a este mapa. Durante este proceso, en la actualidad se usan herramientas automatizadas para construir cadenas laterales, ligandos, y detección de agua (Mishra et al., 2012).

A.7. Primera iteración de CASP

La primera iteración de CASP consistió de tres partes:

- La recolección de proteínas objetivo para predicción de la comunidad científica experimental.
- La recolección de predicciones de la comunidad de modelamiento de estructuras proteicas.
- La evaluación y discusión de los resultados.

La información de estructuras proteicas a predecir por los participantes, se obtuvo trabajando en conjunto con cristalógrafos en rayos X y espectroscopistas NMR. Los 33 objetivos proteicos propuestos fueron predichos en tres categorías: predicción comparativa, enhebramiento o reconocimiento de pliegues y plegamiento ab initio (Moult et al., 1995).

Apéndice B

Resultados

PDB ID	Residuos Totales	Cadenas
7EOW	369	2
7JOD	389	2
7KP1	391	2
6WX1	438	2
6WRP	451	2
6ZH1	266	2
7SH4	393	2
7RYL	457	2
7KKH	434	2
7FDO	257	2
7EJO	249	2
7F4Q	389	2
6TKC	445	2
6WUD	246	2
6X28	233	2
6XOD	339	2
6ZBK	232	2
6ZWA	402	2
7ANQ	358	2
7B0W	351	2

Tabla B.1: Los 20 complejos seleccionados para obtener predicciones.

Uniprot ID 1	Uniprot ID 2	ID Negativo
Q6ZNK6	Q9Y4K3	NEGATIVEH01
Q9NR31	Q15797	NEGATIVEH02
Q9NPY3	P02745	NEGATIVEH05
P03211	Q15796	NEGATIVEH06
Q99576	O43524	NEGATIVEH08
Q9BZM4	P16757	NEGATIVEH11
Q29983	P16757	NEGATIVEH12
Q8TD07	P16757	NEGATIVEH13
Q16611	P45880	NEGATIVEH14
P11799-3	P62149	NEGATIVEH19
PTHP_ECOLI	UBIE_ECOLI	NEGATIVE04
GLGA_ECOLI	IRAP_ECOLI	NEGATIVE05
WECH_ECOLI	YAHF_ECOLI	NEGATIVE09
NAPD_ECOLI	UDG_ECOLI	NEGATIVE10
UVRB_ECOLI	RSD_ECOLI	NEGATIVE11
ETTA_ECOLI	MOBA_ECOLI	NEGATIVE12
TSAD_ECOLI	YCGL_ECOLI	NEGATIVE13
RUVA_ECOLI	RSXG_ECOLI	NEGATIVE14
ALAC_ECOLI	WCAC_ECOLI	NEGATIVE15
GNSA_ECOLI	CODA_ECOLI	NEGATIVE18

Tabla B.2: Pares de IDs UniProt de proteínas no interactuantes de dataset negativo.

PDB ID	Residuos	Fnat	Fnonnat	iRMS	LRMS	DockQ
7EOW	369	0,250	0,755	11,972	25,690	0,121
7JOD	389	0,864	0,400	1,317	4,807	0,729
7KP1	391	0,898	0,307	3,092	3,639	0,644
6WX1	438	0,889	0,132	0,628	0,773	0,911
6WRP	451	0,829	0,348	1,397	1,349	0,780
6ZH1	266	0,769	0,143	1,281	2,538	0,755
7SH4	393	0,874	0,321	3,854	4,845	0,587
7RYL	457	0,863	0,178	1,136	1,565	0,822
7KKH	434	0,841	0,188	1,153	1,225	0,816
7FDO	257	0,930	0,091	0,576	0,635	0,932
7EJO	249	0,636	0,023	11,949	16,775	0,285
7F4Q	389	0,904	0,161	0,871	1,164	0,878
6TKC	445	0,828	0,220	1,032	1,262	0,828
6WUD	246	0,940	0,035	0,424	0,523	0,954
6X28	233	0,000	1,000	17,504	51,089	0,011
6XOD	339	0,889	0,040	0,697	3,524	0,855
6ZBK	232	0,780	0,115	0,864	1,944	0,827
6ZWA	402	0,778	0,262	1,062	1,964	0,798
7ANQ	358	0,000	1,000	18,005	53,745	0,010
7B0W	351	0,837	0,163	1,297	1,622	0,795

Tabla B.3: Resultados de Alphafold.

PDB ID	Residuos	Fnat	Fnonnat	iRMS	LRMS	DockQ
7EOW	369	0,104	0,894	13,289	23,208	0,078
7JOD	389	0,000	1,000	24,201	38,125	0,017
7KP1	391	0,886	0,271	3,029	3,454	0,647
6WX1	438	0,905	0,136	0,691	0,974	0,906
6WRP	451	0,838	0,321	1,472	1,519	0,772
6ZH1	266	0,000	0,000	22,187	56,633	0,009
7SH4	393	0,897	0,297	3,779	4,913	0,594
7RYL	457	0,857	0,198	0,970	1,480	0,844
7KKH	434	0,823	0,205	1,226	1,590	0,796
7FDO	257	0,884	0,208	1,140	1,349	0,831
7EJO	249	0,606	0,070	11,973	11,397	0,326
7F4Q	389	0,798	0,210	1,789	2,143	0,717
6TKC	445	0,871	0,217	1,019	1,275	0,844
6WUD	246	0,069	0,882	10,564	13,761	0,122
6X28	233	0,000	1,000	17,477	50,340	0,012
6XOD	339	0,852	0,042	0,693	3,363	0,847
6ZBK	232	0,000	1,000	15,262	40,676	0,017
6ZWA	402	0,783	0,272	1,099	1,941	0,795
7ANQ	358	0,000	1,000	19,524	66,128	0,007
7B0W	351	0,815	0,173	1,474	1,848	0,760

Tabla B.4: Resultados de Colabfold paired.

PDB ID	Residuos	Fnat	Fnonnat	iRMS	LRMS	DockQ
7EOW	369	0,104	0,783	5,869	25,715	0,088
7JOD	389	0,000	1,000	11,297	38,737	0,021
7KP1	391	0,750	0,353	3,082	4,082	0,585
6WX1	438	0,889	0,074	1,113	1,564	0,834
6WRP	451	0,847	0,338	1,995	1,775	0,722
6ZH1	266	0,000	1,000	21,136	66,480	0,007
7SH4	393	0,805	0,327	3,713	4,792	0,568
7RYL	457	0,335	0,325	8,924	12,245	0,227
7KKH	434	0,814	0,240	4,243	6,805	0,512
7FDO	257	0,000	1,000	16,428	48,953	0,013
7EJO	249	0,030	0,935	13,222	22,890	0,055
7F4Q	389	0,327	0,646	11,501	20,327	0,164
6TKC	445	0,879	0,197	1,240	1,543	0,814
6WUD	246	0,112	0,838	16,761	32,788	0,061
6X28	233	0,000	1,000	17,519	49,297	0,012
6XOD	339	0,778	0,023	1,102	4,298	0,741
6ZBK	232	0,424	0,194	1,966	4,999	0,512
6ZWA	402	0,566	0,224	1,789	3,788	0,604
7ANQ	358	0,014	0,786	11,603	22,165	0,053
7B0W	351	0,511	0,355	1,986	2,818	0,592

Tabla B.5: Resultados de Omegafold.

PDB ID	Residuos	Fnat	Fnonnat	iRMS	LRMS	DockQ
7EOW	369	0,000	1,000	11,366	26,704	0,036
7JOD	389	0,000	1,000	11,297	38,736	0,021
7KP1	391	0,000	1,000	15,317	26,256	0,035
6WX1	438	0,841	0,152	0,885	1,211	0,854
6WRP	451	0,694	0,238	2,316	3,132	0,623
6ZH1	266	0,000	1,000	23,134	67,623	0,007
7SH4	393	0,000	1,000	15,886	26,958	0,033
7RYL	457	0,665	0,164	2,841	2,939	0,592
7KKH	434	0,805	0,216	7,083	11,164	0,405
7FDO	257	0,000	1,000	20,330	53,175	0,010
7EJO	249	0,167	0,796	10,443	21,777	0,106
7F4Q	389	0,577	0,538	10,045	27,582	0,228
6TKC	445	0,810	0,190	1,622	2,274	0,735
6WUD	246	0,009	0,972	13,414	25,659	0,040
6X28	233	0,000	1,000	18,073	47,745	0,013
6XOD	339	0,000	1,000	14,850	70,345	0,008
6ZBK	232	0,017	0,875	5,927	14,864	0,108
6ZWA	402	0,561	0,279	2,329	4,682	0,540
7ANQ	358	0,014	0,769	15,167	25,704	0,041
7B0W	351	0,481	0,278	2,140	2,777	0,572

Tabla B.6: Resultados de ESMFold.

ID	Grupo	Método	pDockQ	PPV	Interacción?
NEGATIVE04	negativo	colabfold	0,1563	0,730443	no
NEGATIVE05	negativo	colabfold	0,018639	0,558902	no
NEGATIVE09	negativo	colabfold	0,025818	0,558902	no
NEGATIVE10	negativo	colabfold	0,064254	0,635554	no
NEGATIVE11	negativo	colabfold	0,056563	0,635554	no
NEGATIVE12	negativo	colabfold	0,066687	0,635554	no
NEGATIVE13	negativo	colabfold	0,061102	0,635554	no
NEGATIVE14	negativo	colabfold	0,04893	0,635554	no
NEGATIVE15	negativo	colabfold	0	0	no
NEGATIVE18	negativo	colabfold	0,027491	0,558902	no
NEGATIVE20	negativo	colabfold	0	0	no
NEGATIVE21	negativo	colabfold	0,05966	0,635554	no
NEGATIVEH01	negativo	colabfold	0,322917	0,822382	si
NEGATIVEH02	negativo	colabfold	0,045788	0,635554	no
NEGATIVEH05	negativo	colabfold	0,022994	0,558902	no
NEGATIVEH06	negativo	colabfold	0,029245	0,558902	no
NEGATIVEH08	negativo	colabfold	0,32393	0,822382	si
NEGATIVEH11	negativo	colabfold	0,028362	0,558902	no
NEGATIVEH12	negativo	colabfold	0,022371	0,558902	no
NEGATIVEH13	negativo	colabfold	0,025409	0,558902	no
NEGATIVEH14	negativo	colabfold	0,020543	0,558902	no
NEGATIVEH19	negativo	colabfold	0,110802	0,686355	no

Tabla B.7: Resultados de cálculo de pDockQ sobre dataset negativo.

ID	Grupo	Método	pDockQ	PPV	Interacción?
6TKC	positivo	colabfold	0,684484	0,98128	si
6WRP	positivo	colabfold	0,68544	0,98128	si
6WUD	positivo	colabfold	0,650521	0,963225	si
6WX1	positivo	colabfold	0,688406	0,98128	si
6X28	positivo	colabfold	0,337696	0,833678	si
6XOD	positivo	colabfold	0,494918	0,891955	si
6ZBK	positivo	colabfold	0,434223	0,878221	si
6ZH1	positivo	colabfold	0,477706	0,891955	si
6ZWA	positivo	colabfold	0,718399	0,98128	si
7ANQ	positivo	colabfold	0,108967	0,686355	no
7B0W	positivo	colabfold	0,627069	0,95333	si
7EJO	positivo	colabfold	0,317376	0,822382	si
7EOW	positivo	colabfold	0,087015	0,667285	no
7F4Q	positivo	colabfold	0,660861	0,98128	si
7FDO	positivo	colabfold	0,332642	0,833678	si
7JOD	positivo	colabfold	0,572897	0,924203	si
7KKH	positivo	colabfold	0,681369	0,98128	si
7KP1	positivo	colabfold	0,580487	0,924203	si
7RYL	positivo	colabfold	0,658317	0,98128	si
7SH4	positivo	colabfold	0,594405	0,93173	si

Tabla B.8: Resultados de pDockQ sobre dataset positivo.

PDB ID	Grupo	N° Residuos Interfaz	Prom. pLDDT	Prom. IpLDDT
6TKC	positivo	71	94.05999	93.57869
6WRP	positivo	89	94.59753	90.36159
6WUD	positivo	56	92.24886	94.0934
6WX1	positivo	76	96.14272	95.76297
6X28	positivo	6	90.63815	92.22093
6XOD	positivo	29	90.96466	96.61911
6ZBK	positivo	25	94.44972	94.26636
6ZH1	positivo	23	90.44018	96.44413
6ZWA	positivo	107	94.12005	94.83636
7ANQ	positivo	23	79.57104	60.91669
7B0W	positivo	62	90.54567	85.55758
7EJO	positivo	18	79.90539	96.31761
7EOW	positivo	23	82.62247	64.15708
7F4Q	positivo	57	92.03289	94.92919
7FDO	positivo	30	87.67408	94.82549
7JOD	positivo	39	94.24884	94.35912
7KKH	positivo	74	95.24662	93.49759
7KP1	positivo	49	90.97316	86.7035
7RYL	positivo	84	87.37087	88.11786
7SH4	positivo	59	91.12435	89.04067

Tabla B.9: Número de residuos en interfaz, promedio de pLDDT para el complejo completo y para la interfaz de interacción (IpLDDT) del dataset positivo.

ID	**Grupo**	**N° Residuos Interfaz**	**Prom. pLDDT**	**Prom. IpLDDT**
NEGATIVE04	negativo	10	91.04921	85.79506
NEGATIVE05	negativo	0	89.53206	0
NEGATIVE09	negativo	9	88.34129	62.98282
NEGATIVE10	negativo	32	87.37767	51.6069
NEGATIVE11	negativo	37	79.58652	54.13352
NEGATIVE12	negativo	27	79.4547	55.09332
NEGATIVE13	negativo	27	85.85415	55.65068
NEGATIVE14	negativo	10	83.39346	73.13765
NEGATIVE15	negativo	0	89.34133	0
NEGATIVE18	negativo	4	90.12006	57.13154
NEGATIVEH01	negativo	34	81.72884	84.4137
NEGATIVEH02	negativo	22	68.82425	53.66491
NEGATIVEH05	negativo	0	62.93973	0
NEGATIVEH06	negativo	30	49.57812	44.96987
NEGATIVEH08	negativo	70	38.82393	67.14207
NEGATIVEH11	negativo	11	48.72201	42.96222
NEGATIVEH12	negativo	9	50.56352	38.63871
NEGATIVEH13	negativo	25	48.92241	43.19076
NEGATIVEH14	negativo	2	75.91723	51.95
NEGATIVEH19	negativo	34	69.33913	66.11023

Tabla B.10: Número de residuos en interfaz, promedio de pLDDT para el complejo completo y para la interfaz de interacción (IpLDDT) del dataset negativo.