



Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Química Biológica

**Enfoques integrados en la identificación de blancos terapéuticos:
aplicaciones bioinformáticas y quimioinformáticas en
antimicrobianos y otros contextos biomédicos**

Tesis presentada para optar al título de Doctor de la
Universidad de Buenos Aires en el área de Química Biológica

Gustavo Ezequiel Schottlender

Director de tesis: Dr. Darío Fernández Do Porto
Consejero de estudios: Dr. Lautaro Álvarez
Lugar de trabajo: Instituto de Cálculo

Buenos Aires, 21 de Agosto de 2025

Resumen

La resistencia antimicrobiana representa uno de los desafíos más críticos para la salud pública a nivel global, lo que subraya la necesidad urgente de desarrollar nuevas terapias efectivas. Esta tesis se centra en el diseño e implementación de metodologías bioinformáticas y quimioinformáticas para la identificación de blancos terapéuticos en relación con el descubrimiento de nuevos compuestos con actividad antimicrobiana.

Se desarrollaron y validaron dos metodologías complementarias, que pueden aplicarse tanto de forma independiente como en combinación, para abordar diferentes aspectos del descubrimiento de fármacos.

La primera integra aprendizaje automático no supervisado basado en la similaridad química de compuestos, análisis de esencialidad de blancos y docking molecular, y está diseñada para identificar blancos terapéuticos candidatos para compuestos con mecanismos de acción desconocidos. Esta estrategia se probó en patógenos como *Mycobacterium tuberculosis* y *Klebsiella pneumoniae*, demostrando su capacidad para aplicarse de manera sistemática a todo el proteoma de un organismo.

La segunda metodología, aplicada inicialmente al caso de estudio de los Citocromos P450 bacterianos (BacCYPs), combina análisis filogenético, contexto genético y *docking* molecular para predecir sustratos, productos e inhibidores. Aunque se desarrolló y validó en BacCYPs, esta metodología es adaptable a otras familias de proteínas, lo que amplía su aplicabilidad en diversos contextos biomédicos y biotecnológicos.

Los resultados obtenidos demuestran la eficacia de las metodologías propuestas, destacando su capacidad para identificar blancos terapéuticos y generar información clave para el diseño racional de fármacos. Además, las pruebas de concepto realizadas validan su potencial para integrarse en plataformas bioinformáticas existentes, optimizando el desarrollo de nuevos agentes antimicrobianos. Finalmente, se discuten las perspectivas futuras de este trabajo, incluyendo posibles mejoras en los algoritmos predictivos y su integración en estudios de mayor alcance, subrayando su relevancia para la investigación biomédica y biotecnológica.

Abstract

Integrated approaches for identifying therapeutic targets of bioactive compounds: bioinformatics and chemoinformatics applications in antimicrobials and other biomedical contexts.

Antimicrobial resistance represents one of the most critical challenges for global public health, underscoring the urgent need to develop new effective therapies. This thesis focuses on the design and implementation of bioinformatics and cheminformatics methodologies for the identification of therapeutic targets related to the discovery of new compounds with antimicrobial activity.

Two complementary methodologies were developed and validated, which can be applied independently or in combination to address different aspects of drug discovery. The first integrates unsupervised machine learning based on chemical similarity of compounds, target essentiality analysis, and molecular docking, and is designed to identify therapeutic targets for compounds with unknown mechanisms of action. This strategy was tested on pathogens such as *Mycobacterium tuberculosis* and *Klebsiella pneumoniae*, demonstrating its ability to systematically analyze the entire proteome of an organism.

The second methodology, initially applied to the case study of bacterial cytochrome P450s (BacCYPs), combines phylogenetic analysis, genetic context, and molecular docking to predict substrates, products, and inhibitors. Although developed and validated in BacCYPs, this methodology is adaptable to other protein families, broadening its applicability in diverse biomedical and biotechnological contexts.

The results demonstrate the efficacy of the proposed methodologies, highlighting their ability to identify therapeutic targets and generate key information for rational drug design. Additionally, the proof-of-concept studies validate their potential for integration into existing bioinformatics platforms, optimizing the development of new antimicrobial agents. Finally, the future perspectives of this work are discussed, including possible improvements in predictive algorithms and their integration into larger-scale studies, underscoring their relevance for biomedical and biotechnological research.

Resumen.....	2
Abstract.....	3
Abreviaturas.....	7
Glosario.....	8
1. Introducción general.....	10
1.1 Quimioinformática.....	12
1.2 Bioinformática.....	14
1.2.1 Bioinformática general.....	14
1.2.2 Bioinformática estructural.....	16
1.3 Descubrimiento de antimicrobianos.....	18
1.4 Estructura-función de proteínas.....	19
1.5 Citocromos P450.....	20
1.6 Bases de datos.....	23
1.6.1 UniProt.....	23
1.6.2 PFAM.....	24
1.6.3 Database of Essential Genes.....	25
1.6.4 PDB.....	26
1.6.5 MOAD.....	27
1.6.6 KEGG.....	28
1.6.7 ChEMBL.....	29
1.6.8 ChEBI.....	30
1.6.9 RHEA.....	30
1.7 Aprendizaje automático.....	32
1.8 Objetivos de la tesis.....	34
2. Métodos.....	35
2.1 Métodos de quimioinformática.....	35
2.1.1 Representación de compuestos químicos en formato de texto (SMILES).....	35
2.1.2 Fingerprints moleculares.....	37
2.1.3 Comparación de compuestos químicos.....	40
2.2 Métodos de bioinformática.....	42
2.2.1 BLAST.....	42
2.2.2 Alineamiento de secuencias múltiples.....	44
2.2.3 Árboles filogenéticos.....	46
2.2.4 Modelos ocultos de Markov y HMMER.....	48
2.2.5 Propiedades de las proteínas y sus reacciones (esencialidad, drogabilidad, puntos de estrangulamiento y centralidad).....	50
2.2.6 Búsqueda de sustratos y productos de proteínas por contexto genético.....	53
2.3 Modelado de proteínas.....	54
2.3.1 Modelado por homología.....	54
2.3.2 Modelado por AlphaFold.....	56
2.3.3 Validación de los modelos.....	57
2.4 Docking molecular.....	61
2.4.1 Docking molecular: descripción del algoritmo.....	61

2.4.2 Docking molecular sesgado (biased docking).....	64
2.4.3 Z-score.....	65
2.5 Métodos de Aprendizaje automático.....	68
2.5.1 Búsqueda de vecinos cercanos.....	68
2.5.2 Clustering de compuestos similares.....	70
2.5.3 Aglomeración jerárquica.....	73
2.5.4 Validación cruzada y leave-one-out.....	73
2.6 Interfaz gráfica - Front-End.....	75
3. Desarrollo de un algoritmo para predicción de blancos de compuestos antimicrobianos (ReverseLigQ).....	76
3.1 Introducción e hipótesis.....	76
3.2 Resultados.....	78
3.2.1 Caracterización de los genomas de cuatro organismos diferentes.....	78
3.2.2 Generación de un algoritmo predictivo basado en aprendizaje no supervisado para la búsqueda de predicción de targets para drogas con mecanismo de acción desconocido.....	81
3.3.3 Evaluación del algoritmo predictivo.....	83
3.3.4 Docking molecular.....	86
3.3.5 Caracterización de las proteínas de cada organismo según sus propiedades	89
3.3.6 Evaluación adicional de la metodología con drogas cuyos blancos moleculares no están presentes en la versión de las bases de datos utilizadas.....	94
Análisis del set de compuestos con actividad probada y mecanismo no elucidado publicados por GlaxoSmithKline.....	94
Evaluación de la metodología desarrollada en comparación con otros estudios publicados respecto al set de compuestos inhibitorios publicados por GlaxoSmithKline.....	95
Aplicación adicional de la metodología desarrollada para determinar el blanco con el que interactúa otro compuesto con mecanismo de acción recientemente descubierto: Fabimicina.....	97
3.3 Conclusiones.....	99
3.4 Métodos específicos del capítulo.....	100
3.4.1 Obtención de proteomas de los organismos y anotación.....	100
3.4.2 Asignación de dominios a las proteínas.....	100
3.4.3 Armado de la base de datos interna de proteínas, dominios y ligandos.....	101
3.4.4 Similitud química.....	103
3.4.5 Caracterización de los blancos candidatos.....	103
3.4.6 Docking molecular.....	104
3.4.7 Docking sesgado.....	104
3.4.8 Desarrollo de la interfaz gráfica.....	105
4. Descubrimiento de sustratos, productos e inhibidores de Citocromos P450 bacterianos: un enfoque integrado utilizando filogenia, bioinformática estructural y contexto genético.....	107
4.1 Introducción e hipótesis.....	107
4.2 Resultados.....	109
4.2.1 Análisis de la diversidad de secuencias de Citocromos P450 bacterianos....	109
4.2.2 Análisis de la diversidad de estructuras y evaluación del modelado estructural	

para Citocromos P450 bacterianos.....	115
Análisis de la diversidad de estructuras de Citocromos P450 bacterianos.....	115
Evaluación del modelado estructural.....	115
4.2.3 Análisis de la diversidad de sustratos, productos e inhibidores de Citocromos P450 bacterianos.....	117
4.2.4 Enfoque Predictivo 1: Correlación Filogenética de los Compuestos que unen BacCYPs.....	119
4.2.5 Enfoque predictivo 2: Búsqueda mediante Contexto Genético.....	122
4.2.6 Enfoque predictivo 3: Docking molecular.....	125
4.2.7 Diseño de un Flujo de Trabajo Integrado para los Tres Enfoques Desarrollados	126
4.2.8 Evaluación de 2 casos de estudio para probar la eficacia de la metodología	129
4.3 Conclusiones.....	134
4.4 Métodos específicos del capítulo.....	136
4.4.1 Obtención de secuencias de Citocromos P450 bacterianos.....	136
4.4.2 Alineamiento múltiple de secuencias (MSA) y reconstrucción de árboles filogenéticos.....	136
4.4.3 Búsqueda de sustratos, productos e inhibidores de Citocromos P450 bacterianos.....	137
4.4.5 Creación y evaluación de modelos estructurales construidos con Modeller y AlphaFold.....	140
4.4.6 Docking molecular.....	142
4.4.7 Docking sesgado.....	142
4.4.8 Búsqueda de sustratos por contexto genético.....	142
5. Discusión, perspectivas futuras y aplicaciones.....	144
5.1 Discusión general.....	144
5.2 Perspectivas futuras y limitaciones.....	146
5.3 Aplicaciones y alcance de los métodos desarrollados.....	148
Anexo: Tablas suplementarias.....	149
Bibliografía.....	156

Abreviaturas

- Å: Angstrom.
- API: Interfaz de Programación de Aplicaciones (*Application Programming Interface*).
- BacCYPs: Citocromos P450 bacterianos.
- CYPs (o CYP450s): Citocromos P450.
- DOPE: Energía Proteica Discreta Optimizada (*Discrete Optimized Protein Energy*).
- HMM: Modelos Ocultos de Markov (*Hidden Markov Models*).
- Ki: Constante de Inhibición.
- IDDT: Diferencia de Distancia Local Predicha (*Local Distance Difference Test*).
- ML: Aprendizaje Automático (*Machine Learning*).
- MSA: Alineamiento Múltiple de Secuencias (*Multiple Sequence Alignment*).
- NGS: Secuenciación de Próxima Generación (*Next Generation Sequencing*).
- NNS: Búsqueda de Vecinos Cercanos (*Nearest Neighbor Search*).
- pb: pares de bases nucleotídicas.
- pLDDT: Prueba de Diferencia de Distancia Local Predicha (*Predicted Local Distance Difference Test*).
- PRs: Proteínas representativas del universo de Citocromos P450 bacterianos, cada una interactúa con un grupo representativo de ligandos. Detallado en la sección 4.2.6.
- RMSD: Desviación Cuadrática Media (*Root-Mean-Square Deviation*).
- RMSD-C α : Desviación Cuadrática Media de Carbonos Alfa.
- SMILES: Especificación de Introducción Lineal Molecular Simplificada (*Simplified Molecular Input Line Entry System*).
- SPIs: Sustratos, Productos e Inhibidores.
- SPs: Sustratos y Productos.
- TI: Índice de similitud de Tanimoto (*Tanimoto Index*).

Glosario

- Blanco molecular: Macromolécula, como una proteína o ácido nucleico, cuya modulación por un ligando, inhibidor o fármaco genera una respuesta biológica específica. En el presente trabajo el término se aplica casi exclusivamente a proteínas; cuando su inhibición determina la actividad antimicrobiana de un compuesto se las denomina blancos terapéuticos.
- *Clustering*: Técnica de agrupamiento utilizada para organizar objetos o datos en grupos según su similitud, con el objetivo de identificar patrones o relaciones en conjuntos complejos. En este trabajo, los términos *clustering* y agrupamiento se utilizan como sinónimos.
- *Docking* molecular: Técnica computacional que predice cómo una molécula, como un ligando, interactúa con una proteína, evaluando su orientación, posición y afinidad de unión en un sitio activo.
- Dominio funcional: Región específica de una proteína que está estructural y funcionalmente organizada para llevar a cabo una tarea particular, como la unión a un ligando o la catálisis de una reacción. En este trabajo se utiliza el término en representación de una familia de proteínas, ya que según el agrupamiento basado en dominios Pfam, se considera que las proteínas que comparten un mismo dominio funcional pertenecen a una misma familia funcional.
- *E-value*: Valor esperado que indica la probabilidad de encontrar coincidencias por azar en herramientas de búsqueda de secuencias.
- Familia de proteínas: Proteínas que comparten entre sí un dominio funcional incorporado en la base de datos Pfam. A lo largo de este trabajo, el término *familia de proteínas* se emplea para referirse a agrupamientos basados en la presencia de un dominio Pfam común, asumiendo que esta conservación refleja una relación funcional y evolutiva.
- Filogenia: Estudio de relaciones evolutivas entre organismos y/o proteínas.
- Grupo de SPIs: Grupos de moléculas similares entre sí correspondientes a sustratos, productos o inhibidores de Citocromos P450 bacterianos.
- Grupo filogenético: Grupos de proteínas cercanas evolutivamente entre sí, derivados de un árbol filogenético.
- *In-silico*: Análisis o simulaciones realizadas mediante herramientas computacionales.
- Inhibidor: Molécula que reduce o bloquea la actividad de una proteína al unirse a su sitio activo o a otro sitio regulador.
- Ligando: Molécula que se une de manera específica a un blanco biológico, como una proteína o un receptor, para formar un complejo molecular y potencialmente modular su

función. Según el contexto, los ligandos pueden corresponder a drogas, fármacos, compuestos bioactivos, inhibidores, sustratos o productos.

- Péptido: Cadena de aminoácidos unidos por enlaces peptídicos, que puede desempeñar funciones biológicas específicas.
- Producto: Molécula resultante de la transformación química de un sustrato catalizada por una enzima.
- Proteína molde: Proteína con estructura determinada experimentalmente, que se emplea para realizar modelado por homología. Debe compartir suficiente identidad de secuencia (y por consiguiente conservación estructural) con la proteína que se quiere modelar para permitir la predicción confiable de su conformación tridimensional.
- Proteoma: Conjunto completo de proteínas que son o pueden ser expresadas por una célula, tejido u organismo.
- Receptor: Denominación que recibe la proteína sobre la cual se evalúa su unión con un ligando en el contexto de los algoritmos de *docking* molecular.
- *Screening*: Proceso de evaluación sistemática de grandes conjuntos de compuestos o datos para identificar candidatos potenciales, como fármacos o moléculas bioactivas.
- *Screening* virtual: Técnica computacional que utiliza simulaciones para evaluar grandes bibliotecas de compuestos en busca de posibles candidatos, como ligandos con afinidad hacia un blanco molecular.
- Sustrato: Molécula que interactúa con una enzima y es transformada durante una reacción química.

1. Introducción general

La resistencia antimicrobiana se ha convertido en uno de los retos más críticos para la salud pública mundial. La Organización Mundial de la Salud (OMS) ha advertido sobre la falta de avances significativos en el desarrollo de nuevos antibióticos, requiriéndose la urgente necesidad de investigar nuevos agentes antimicrobianos (World Health Organization 2022). En este contexto, la identificación de blancos moleculares efectivos cobra especial relevancia, pues constituye el primer paso hacia la creación de fármacos más eficientes contra patógenos resistentes. En respuesta a esta necesidad, nuestro Grupo de investigación ha desarrollado previamente la herramienta Target-Pathogen (Sosa et al. 2018), un recurso computacional que clasifica genes y proteínas de patógenos según su potencial como blancos de compuestos antimicrobianos.

Estudios anteriores han utilizado Target-Pathogen para seleccionar blancos moleculares en una variedad de patógenos, como *Mycobacterium tuberculosis* (Defelipe et al. 2018), *Klebsiella pneumoniae* (Ramos et al. 2018), *Bartonella bacilliformis* (Farfán-López et al. 2020), *Trypanosoma cruzi* (Rivara-Espasandín et al. 2023) y *Schistosoma mansoni* (Lobo-Silva et al. 2020). Sin embargo, cuando se identifican estos posibles blancos moleculares, el desafío clave que sigue a esta etapa es identificar moléculas pequeñas capaces de inhibir de forma efectiva la función de dichos blancos, lo que resulta fundamental para avanzar hacia el desarrollo de nuevos tratamientos antimicrobianos.

El *screening* fenotípico de compuestos antimicrobianos ha sido una herramienta esencial durante el último siglo, desempeñando un papel clave en el descubrimiento de medicamentos como la Daptomicina (Heidary et al. 2017). Este enfoque implica la evaluación *in vitro* de miles de compuestos para determinar su capacidad de inhibir el crecimiento microbiano. En particular, los productos naturales resultan ser una de las fuentes principales de compuestos bioactivos. El proceso comienza con la obtención de extractos crudos de materiales naturales, como hongos o bacterias, que luego son tamizados en busca de actividad antimicrobiana. Una vez identificada dicha actividad, se procede a purificar y caracterizar los compuestos responsables. Este método ha sido la base de descubrimientos emblemáticos, como el de la penicilina, que marcó un hito en la historia de la medicina (W. H. Hughes 1974).

Una vez identificado un compuesto con actividad antimicrobiana, el próximo gran desafío consiste en **determinar su blanco molecular**. Para abordar este problema, se suelen emplear métodos experimentales como la inducción del desarrollo de resistencia al fármaco, ya sea de forma natural o mediante algún mutágeno, y la posterior comparación genómica entre la cepa parental y la resistente (D. Hughes and Karlén 2014). En contraste

con la alta efectividad de estos métodos, son extremadamente costosos y demandan mucho tiempo, por lo cual en muchos casos resultan poco viables, especialmente en países en vías de desarrollo. Por ello, se requiere promover la investigación de diversos enfoques computacionales para agilizar el proceso.

En la primera parte de este trabajo, se propone una metodología novedosa diseñada para abordar uno de los principales retos en la investigación de nuevos agentes antimicrobianos: *la identificación de blancos moleculares para compuestos con mecanismos de acción desconocidos*.

La metodología combina herramientas computacionales avanzadas, como el análisis de similitud química, el *docking* molecular y la priorización de blancos candidatos basada en criterios genómicos y funcionales, integrando estos enfoques en un marco sistemático y escalable. Este enfoque permite explorar de manera exhaustiva los proteomas de organismos patógenos para identificar qué proteínas podrían actuar como blancos terapéuticos de dichos compuestos.

En la segunda parte de este trabajo, se presenta un flujo de trabajo que integra enfoques estructurales y de análisis de secuencias para refinar los resultados obtenidos de la predicción de blancos moleculares.

Utilizando la familia de los Citocromos P450 bacterianos como caso de estudio, se busca mejorar la precisión en la identificación de interacciones proteína-ligando relevantes mediante un enfoque integrado. Esta metodología combina el análisis de la diversidad de secuencias y estructuras, relaciones evolutivas, contexto genético y herramientas estructurales, permitiendo establecer un marco predictivo para facilitar la priorización de interacciones relevantes.

En esta primera sección, se ofrece una descripción introductoria de los conceptos teóricos clave abordados a lo largo de este trabajo.

1.1 Quimioinformática

La quimioinformática, un campo de estudio surgido de la necesidad de gestionar una creciente cantidad de información química disponible en bases de datos informáticas de manera sistematizada, ha evolucionado como un componente esencial en la investigación y desarrollo en diferentes campos de la química y la biología.

El término "quimioinformática" se ha consolidado en los últimos años, definiéndose como el empleo de herramientas computacionales para obtener información valiosa a partir de datos químicos con el propósito de acelerar el proceso de descubrimiento de fármacos y de moléculas útiles en diversas aplicaciones. (Engel* 2006)

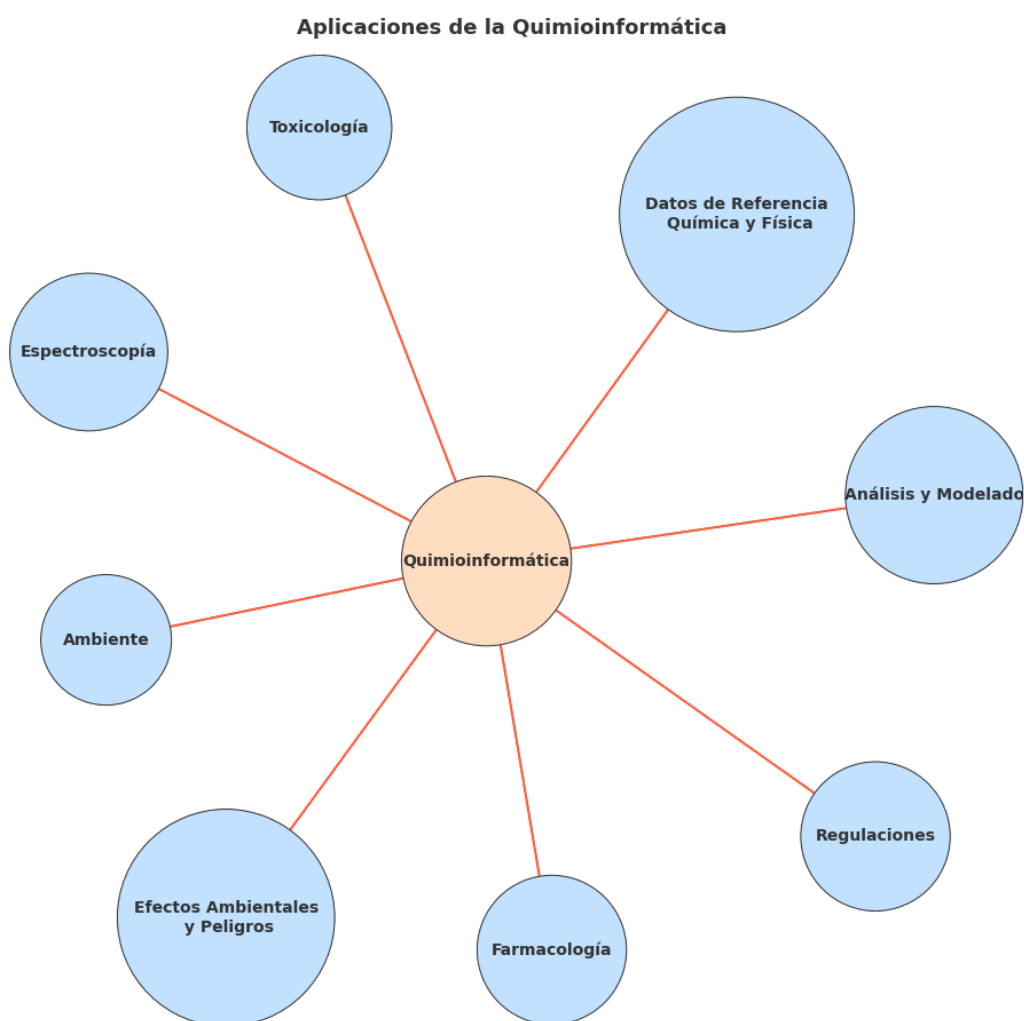


Figura 1.1: Aplicaciones de la quimioinformática.

La quimioinformática aborda varios aspectos cruciales relacionados con la representación, manipulación y procesamiento de información química. Un ejemplo relevante lo constituyen las distintas representaciones de compuestos químicos, entre ellas el formato SMILES, ampliamente utilizado. Estas representaciones son fundamentales para obtener descripciones informativas de moléculas en dos o tres dimensiones para estudios posteriores. (Wigh, Goodman, and Lapkin 2022). El formato SMILES se explicará con mayor detalle más adelante.

La relación estrecha entre quimioinformática y bioinformática es evidente, especialmente en el diseño de drogas. Los métodos de genómica se utilizan para identificar blancos moleculares proteicos, mientras que la quimioinformática hace su contribución al otorgar un conjunto de moléculas que pueden ser útiles para lograr una actividad farmacológica deseada.

Aunque tradicionalmente la quimioinformática ha centrado su atención en moléculas pequeñas con actividad farmacológica, su alcance se ha expandido hacia diferentes áreas de la química, como toxicología e impacto ambiental de compuestos químicos, entre otras, ampliando así su importancia y aplicabilidad (Figura 1.1). ("A Study on Cheminformatics and Its Applications on Modern Drug Discovery" 2012)

En el estudio de reacciones químicas, la quimioinformática permite tanto la búsqueda en bases de datos de reacciones ya estudiadas como la predicción de productos (o sustratos) y condiciones experimentales óptimas mediante herramientas computacionales. Se emplean enfoques como la Teoría de Orbitales Moleculares y Relaciones Lineales de Energía Libre para comprender las transformaciones químicas. Aunque la Quimioinformática de Reacciones está menos desarrollada que la representación de compuestos, ha experimentado avances significativos, especialmente con el establecimiento de bases de datos especializadas que almacenan información sobre reacciones químicas. (Engkvist et al. 2018)

La diversidad de datos en química, que abarca desde estructuras moleculares hasta espectros de absorción, presenta desafíos significativos en términos de estandarización. (Chen, Kogej, and Engkvist 2018) Formatos unificados como Molfile (utilizado para almacenar información sobre los átomos, enlaces, conectividad y coordenadas de una molécula) han desempeñado un papel crucial al facilitar el intercambio de información química entre diferentes plataformas y aplicaciones. (J. Z. Zhou 2011)

Dado el contexto actual, la quimioinformática representa un pilar indispensable para abordar la complejidad de la química contemporánea. Su capacidad para acelerar el descubrimiento de drogas y facilitar la colaboración interdisciplinaria entre la química y la biología la posiciona como un campo esencial en la investigación científica actual. Por otro lado, su papel crítico en la representación, procesamiento y análisis de datos químicos

promete seguir impulsando avances significativos en cuanto a la investigación y el descubrimiento de nuevas moléculas con aplicaciones útiles.

1.2 Bioinformática

1.2.1 Bioinformática general

En términos generales, la bioinformática se define como la disciplina que aborda problemas biológicos a través del empleo de herramientas computacionales. Este enfoque comprende tanto la creación de algoritmos y programas diseñados para estos fines, como el uso de plataformas ampliamente distribuidas. Desde su origen, la bioinformática ha evolucionado como un campo multidisciplinario, integrando no sólo la biología y la informática, sino también conceptos relacionados con la química, la matemática y la estadística.

Sus raíces se remontan a las décadas de 1960 y 1970, impulsadas por descubrimientos clave como el código genético y las primeras estructuras proteicas. La explosión de la bioinformática tuvo lugar a finales de la década de 1990 con proyectos de secuenciación de genomas completos, especialmente el Proyecto Genoma Humano ("Initial Sequencing and Analysis of the Human Genome" 2001), lo que involucró avances revolucionarios en las técnicas y métodos de análisis de datos biológicos. La introducción de secuenciadores de próxima generación (NGS) en 1999 marcó una escalada en esta revolución tecnológica. (Gauthier et al. 2018; Hood and Rowen 2013)



Figura 1.2: Aplicaciones de la bioinformática.

Hoy en día, la bioinformática no solo se enfoca en el análisis eficiente de grandes volúmenes de datos, sino también en la integración de la información biológica subyacente para generar conocimientos que impulsen nuevos descubrimientos científicos. Algunas de sus técnicas más comúnmente utilizadas involucran el análisis y alineamiento de secuencias, tanto de ácidos nucleicos como de proteínas, así como también la reconstrucción filogenética, ensamblados de genomas y comparaciones entre organismos según diferentes criterios establecidos. Su amplitud se refleja en desarrollos significativos en diversas áreas, como predicción de genes, anotación e interpretación de redes metabólicas, análisis filogenéticos, modelado de macromoléculas, análisis de expresión

génica, y descubrimiento de drogas o agroquímicos, entre otras (Figura 1.2). (Edwards, Stajich, and Hansen 2009)

La importancia de la bioinformática en la genómica va en incremento de la mano del continuo crecimiento de las bases de datos biológicas. Los proyectos de secuenciación genómica, los estudios poblacionales y la metagenómica generan volúmenes masivos de datos cuyo análisis requiere de métodos informáticos avanzados. La predicción de genes y unidades de transcripción, con el complemento de los algoritmos de alineamiento de secuencias y búsqueda por cercanía evolutiva, son fundamentales para asignar funciones y caracterizar genomas de manera eficaz. (Pevsner 2015)

En conclusión, la bioinformática se ha convertido en una herramienta esencial en genómica y áreas relacionadas, siendo esencial para abordar los desafíos actuales y futuros en la investigación biológica. La recopilación y estructuración eficiente de datos biológicos en bases de datos es determinante para respaldar investigaciones en curso y futuras.

1.2.2 Bioinformática estructural

La bioinformática estructural es una subdisciplina destacada que se centra en el análisis computacional de las estructuras tridimensionales de biomoléculas, con un enfoque particular en las proteínas.

Está orientada a investigar, analizar, predecir e interpretar las estructuras de distintos tipos de biomoléculas, utilizando recursos, algoritmos y herramientas informáticas específicas. (Wei et al. 2014)

Las proteínas son largas cadenas de aminoácidos que desempeñan funciones esenciales en los organismos. Son fundamentales para el funcionamiento y la regulación de los procesos celulares, así como también en el mantenimiento de la estructura de células y tejidos. Uno de los ejes centrales de la bioinformática estructural consiste en el análisis de las proteínas en el espacio tridimensional, aprovechando la creciente disponibilidad de estructuras en bases de datos especializadas (Gu and Bourne 2009).

Desde el surgimiento del Protein Data Bank (PDB) (Berman et al. 2000), establecido en 1971, que almacena las estructuras de macromoléculas (desde proteínas hasta ácidos nucleicos) determinadas experimentalmente, se ha avanzado significativamente tanto en el conocimiento de las conformaciones espaciales de este tipo de biomoléculas, como también posteriormente, en la predicción *in-silico* de la estructura. Se destaca en los últimos años el éxito de AlphaFold (Jumper et al. 2021) en las competiciones CASP (Evaluación Crítica de Predicción Estructural, un evento bienal destinado a evaluar y fomentar el desarrollo de métodos de predicción de estructuras proteicas). Esta herramienta, capaz de

predecir estructuras de proteínas con una precisión que en muchos casos es prácticamente idéntica a los métodos experimentales, requiere como punto de partida únicamente una secuencia de aminoácidos. Este avance, que fue motivo del Premio Nobel de Química en 2024 (Callaway 2024), se traduce en la generación masiva de modelos estructurales precisos de proteínas utilizando técnicas de aprendizaje automático, marcando un hito en la predicción estructural a nivel proteómico (Varadi et al. 2021).

Por otra parte, la biología estructural se encuentra muy presente en el campo de la predicción de interacciones entre drogas y blancos moleculares. Áreas como el *docking* y la dinámica molecular o métodos basados en inteligencia artificial permiten simular y analizar computacionalmente cómo las moléculas de un fármaco interactúan con proteínas específicas. Estos diferentes enfoques se suelen utilizar de manera combinada en flujos de trabajo compuestos de diferentes pasos con el fin de obtener moléculas con actividades relevantes a fines farmacológicos y biotecnológicos, entre otros (Figura 1.3). Estas herramientas han contribuido significativamente a obtener representaciones más precisas y consistentes de estos procesos (Santos, Ferreira, and Caffarena 2019). Estos métodos computacionales son fundamentales para el diseño de fármacos ya que contribuyen significativamente al entendimiento de las complejas interacciones a nivel molecular.

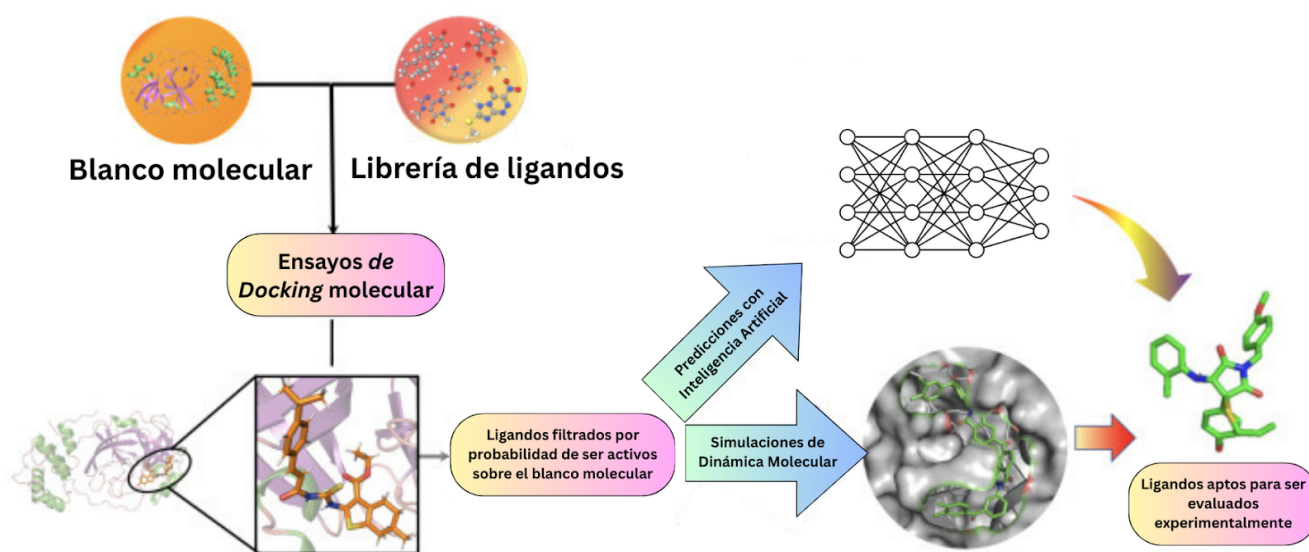


Figura 1.3: Esquema del uso del *docking* molecular en combinación con filtros posteriores como el uso de algoritmos de Inteligencia Artificial y las simulaciones de Dinámica Molecular en la predicción de moléculas con potencial inhibitorio de un blanco molecular. Adaptado de Joshi and Kumar 2021.

1.3 Descubrimiento de antimicrobianos

La búsqueda y el descubrimiento de moléculas con actividad antimicrobiana son pilares fundamentales de la investigación biomédica, ya que estos compuestos desempeñan un papel crucial en el tratamiento de infecciones bacterianas.

Los antimicrobianos, que en su mayoría son moléculas orgánicas con diferentes grados de complejidad, destacan por su extraordinaria diversidad estructural y funcional. Comprender mejor su naturaleza es esencial para combatir de manera efectiva la amplia variedad de patógenos existentes (Shinu et al. 2022). La investigación de esta clase de moléculas es fundamental en el ámbito clínico, dado que la rápida aparición y propagación de la resistencia bacteriana a los antimicrobianos genera una necesidad urgente de desarrollar nuevas terapias efectivas. El estudio de las relaciones entre la estructura y las propiedades físico-químicas de estas moléculas con sus mecanismos de acción es esencial para diseñar estrategias innovadoras para el descubrimiento de fármacos con actividad antimicrobiana.

La notable diversidad funcional y estructural de estos compuestos, junto con su capacidad para interactuar con blancos específicos en los microorganismos, plantea desafíos complejos que demandan enfoques interdisciplinarios. En este contexto, la bioinformática, la quimioinformática y las simulaciones a nivel molecular, combinadas con ensayos experimentales en laboratorio, desempeñan un papel crucial en el descubrimiento y reposicionamiento de compuestos con potencial antimicrobiano (Serral et al. 2021).

La bioinformática juega un papel esencial en el descubrimiento de antimicrobianos, proporcionando herramientas innovadoras y métodos analíticos que mejoran significativamente la eficiencia de la investigación biomédica. Gracias a su aplicación, es posible analizar extensas bases de datos genómicas y proteómicas para identificar posibles blancos terapéuticos en microorganismos, además de predecir el potencial antibiótico de moléculas presentes en bases de datos existentes (Xia 2018). Las simulaciones computacionales, por su parte, son fundamentales para modelar las interacciones entre antimicrobianos y sus blancos biológicos, ofreciendo un abordaje detallado para el entendimiento de los mecanismos de acción. Asimismo, las técnicas de aprendizaje automático y el análisis de datos experimentales permiten identificar patrones relevantes relacionados con la bioactividad, acelerando el descubrimiento de moléculas prometedoras con potencial terapéutico (Dara et al. 2021).

La integración de enfoques multidisciplinarios en el descubrimiento de antibióticos mejora la eficiencia de la investigación, ampliando significativamente las posibilidades de identificar nuevos compuestos con actividad antimicrobiana para enfrentar de manera más efectiva los desafíos persistentes asociados con la resistencia bacteriana.

1.4 Estructura-función de proteínas

Las proteínas son fundamentales para las funciones esenciales de la célula y el desarrollo de la vida, destacando por su extraordinaria versatilidad. A lo largo de miles de millones de años de evolución, han desarrollado una impresionante capacidad para cumplir roles esenciales en diversos procesos biológicos. Por estos motivos, profundizar en el conocimiento sobre las proteínas no solo es esencial para el desarrollo de nuevas terapias efectivas, sino que también representa un desafío considerable.

La estructura proteica es clave para comprender y determinar su funcionalidad. Uno de los principios fundamentales en el estudio de las proteínas consiste en que su función está directamente determinada por su estructura ("From Protein Structure to Function" 1999).

Entre la diversidad de estructuras proteicas, destaca especialmente la presencia de sitios conservados en proteínas diferentes que cumplen funciones relacionadas, en las cuales aminoácidos específicos se disponen de la misma manera en el espacio tridimensional. Este tipo de regiones, recurrentes en distintas familias proteicas, son fundamentales para su actividad biológica. La presencia de sitios conservados en combinación con otras regiones altamente variables permite a las proteínas actuar como catalizadores en una amplia variedad de reacciones químicas. Asimismo, los grupos funcionales desempeñan un papel clave en las funciones estructurales, de transporte y de regulación que las proteínas realizan en los organismos. La relación entre estructura y función, asociada con la diversidad de roles biológicos, se refleja en la notable riqueza y complejidad de las estructuras proteicas.

Por otro lado, el campo de estudio se expande considerablemente al considerar las interacciones tanto con moléculas pequeñas, conocidas como ligandos, como con otras proteínas. Este volumen de información representa un desafío significativo para la investigación, en donde la bioinformática y la simulación molecular se presentan como disciplinas complementarias para abordarlo de manera integral (Gallina, Bork, and Bordo 2014).

1.5 Citocromos P450

Los Citocromos P450 (CYPs o CYP450s) han evolucionado como catalizadores excepcionales en la naturaleza, llevando a cabo una enorme diversidad de reacciones químicas, involucradas en la biosíntesis y degradación de un gran número de compuestos. Tanto es así que su investigación es clave para el descubrimiento de nuevos medicamentos, biocombustibles y otros productos muy útiles para diversos fines. El potencial de los Citocromos es inmenso dada la enorme cantidad de nuevas proteínas y funciones asociadas que quedan por descubrir (Behrendorff 2021; F. P. Guengerich 1995).

El estudio de los Citocromos P450 se originó en la década de 1950 con la investigación de pigmentos en células hepáticas. El término 'Citocromo P450' surgió al identificar un pigmento unido al monóxido de carbono, con un pico de absorbancia a 450 nm en el espectro UV, P por pigmento y 450 por su absorbancia ("Cytochromes P450: History, Classes, Catalytic Mechanism, and Industrial Application" 2016; F. Peter Guengerich 2020).

Los CYPs están presentes en todos los reinos biológicos, desde bacterias hasta eucariotas, y su distribución abarca diversos tejidos dentro de un organismo, reflejando su notable versatilidad funcional. Las bacterias producen unas 20 variantes en promedio, los humanos sintetizan alrededor de 60, y las plantas, debido a su necesidad de alta diversidad funcional, generan cientos de CYPs para producir pigmentos y toxinas esenciales para su protección. En eucariotas, los CYPs se localizan principalmente en la mitocondria y el retículo endoplásmico, mientras que en bacterias son solubles y no asociados a membranas ("Functional Expression and Regulation of Eukaryotic Cytochrome P450 Enzymes in Surrogate Microbial Cell Factories" 2022).

La nomenclatura de los CYPs sigue una sistemática basada en criterios filogenéticos, asignándoles nombres compuestos por el prefijo CYP seguido de números y letras que indican familia, subfamilia y gen. Esta nomenclatura proporciona una clasificación precisa y ordenada de estas enzimas. (Nelson 2006)

El sitio activo de los CYPs, de tipo hemotiolato, les permite activar oxígeno molecular e insertarlo en una amplia variedad de sustratos, participando en numerosas reacciones oxidativas. Estas enzimas desempeñan un papel crucial en el metabolismo de compuestos xenobióticos como drogas, pesticidas y solventes orgánicos, así como en la biotransformación de sustancias endógenas, incluyendo colesterol, ácidos biliares, esteroides, vitaminas liposolubles y ácidos grasos (Peter Guengerich 2018). La estructura general de un citocromo P450 se observa en la Figura 1.4.

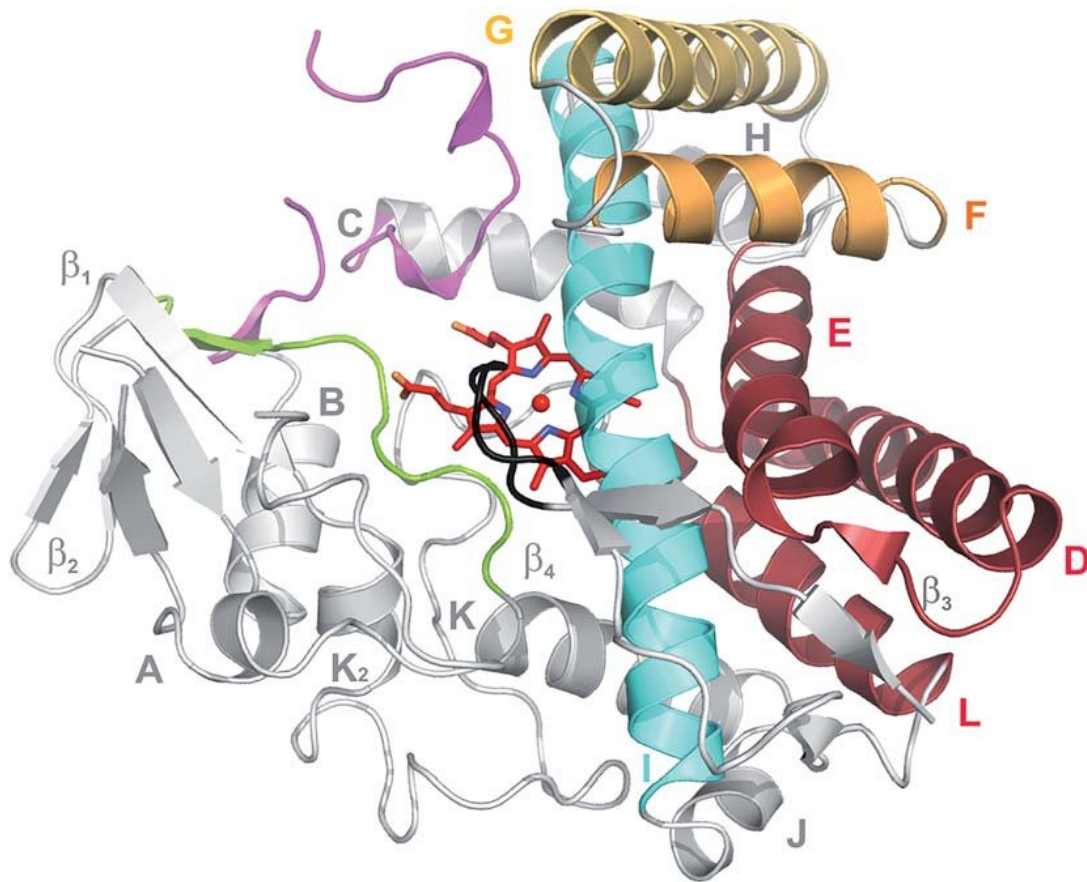
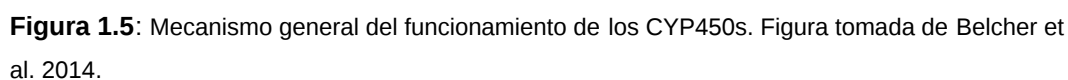


Figura 1.4: Estructura general de un Citocromo P450. Figura tomada de Greule et al. 2018.

La reacción catalizada por un CYP450 comienza con la unión del sustrato a su sitio activo, lo que provoca cambios en la enzima y permite la activación del oxígeno molecular. Este oxígeno es posteriormente incorporado al sustrato, llevando a cabo su modificación química, generalmente mediante hidroxilación. Una vez completada la reacción, el producto es liberado y la enzima regresa a su estado inicial, lista para repetir el ciclo catalítico. (Peter Guengerich* 2001). Este proceso se ilustra en la Figura 1.5.



Entender mejor las propiedades estructurales y funcionales de los CYPs es fundamental para explorar su potencial y aprovecharlo en diferentes campos de investigación y desarrollo.

1.6 Bases de datos

Las bases de datos son herramientas fundamentales para el almacenamiento, gestión y análisis de datos biológicos a gran escala. Estas bases permiten organizar información genética, estructural, funcional y evolutiva, facilitando el acceso a datos esenciales para la investigación en biomedicina, biotecnología y biología computacional. Existen distintos tipos de bases de datos según su contenido, como las de secuencias de ADN y proteínas (GenBank, RefSeq, UniProt), estructuras macromoleculares (PDB, AlphaFold DB), interacciones moleculares (STRING, BioGRID), y anotaciones funcionales (GO, KEGG). Gracias a estas herramientas, los investigadores pueden realizar comparaciones evolutivas, predecir funciones de genes y proteínas, identificar variantes genéticas asociadas a enfermedades y diseñar nuevos fármacos. El desarrollo y la actualización constante de estas bases de datos son esenciales para mantener la precisión y confiabilidad de los análisis bioinformáticos. A continuación se describen las bases de datos más relevantes para este trabajo.

1.6.1 UniProt

UniProt, la Base de Datos Universal de Proteínas (The UniProt Consortium et al. 2022), es un recurso esencial que ofrece acceso gratuito y centralizado a información detallada sobre secuencias y anotaciones funcionales de proteínas. Este proyecto, fruto de la colaboración entre el Instituto Europeo de Bioinformática (EBI), el Recurso de Información de Proteínas (PIR) y el Instituto Suizo de Bioinformática (SIB), abarca actividades clave como la curación manual de secuencias, el almacenamiento y gestión de información relevante, la operación de su plataforma web y la integración de referencias cruzadas con otras bases de datos.

Dado el constante aumento en las secuencias de proteínas, impulsado por algoritmos de predicción de genes, el avance de la proteómica a gran escala y la genómica estructural, UniProt se posiciona como el catálogo más completo y confiable de secuencias proteicas, proporcionando una plataforma robusta para interconectar y explorar datos provenientes de diversas fuentes. Esta base de datos se constituye de diferentes componentes (Figura 1.6):

- UniProt Knowledgebase (UniProtKB): Se compone de UniProtKB/Swiss-Prot y UniProtKB/TrEMBL. El primero, caracterizado por sus registros de alta calidad que incluye refinamiento mediante curación y anotación manual, contiene cientos de miles de descripciones de proteínas, que incluyen función, estructura de dominio, ubicación subcelular, modificaciones postraduccionales y variantes caracterizadas funcionalmente. UniProtKB/TrEMBL abarca secuencias de proteínas cuya anotación

fue generada automáticamente a partir de las secuencias genómicas despositadas en Ensembl, utilizando métodos predictivos avanzados que consideran también su caracterización funcional (Bairoch and Apweiler 2000).

- UniProt Reference Clusters (UniRef): Ofrece clusters de secuencias de UniProtKB y registros seleccionados de UniProt Archive, agrupándolos según la identidad de secuencia al 100%, 90% y 50%. Estos agrupamientos sirven para disminuir las redundancias y son actualizados con cada nueva versión de UniProtKB (Suzek et al. 2007).
- UniProt Archive (UniParc): Almacena todas las secuencias de proteínas de diversas fuentes, evitando la redundancia al fusionar secuencias idénticas. Se actualiza diariamente, siendo una fuente integral y sincronizada con los últimos cambios.
- UniProt Metagenomic and Environmental Sequences (UniMES): Creada para datos metagenómicos.

UniProt, construido sobre la infraestructura bioinformática y la experiencia científica de EBI, PIR y SIB, es de acceso libre y está diseñado para diversos usos, desde la anotación manual hasta la búsqueda de secuencias similares. Es esencial para la investigación en genómica, proteómica, biología estructural y áreas afines.

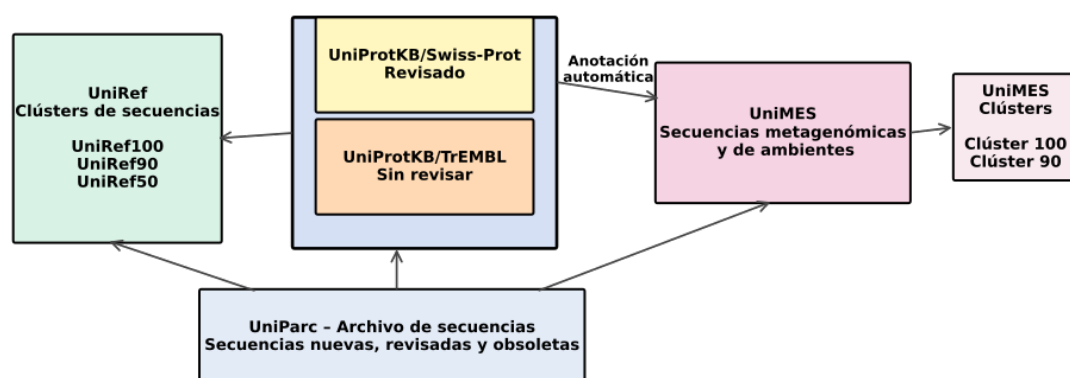


Figura 1.6: Esquema general de la base de datos UniProt, en el cual se detallan sus componentes principales.

1.6.2 PFAM

Pfam (Mistry et al., 2020) es una base de datos actualmente incluida en la plataforma de InterPro (Blum et al. 2025) que clasifica **dominios funcionales** de proteínas, es decir, regiones conservadas responsables de funciones específicas dentro de una proteína. Dado que estos dominios suelen conservarse en proteínas evolutivamente

relacionadas, su identificación permite inferir funciones y establecer relaciones entre proteínas. Los dominios funcionales suelen ser utilizados como base para **agrupar proteínas en familias**, ya que distintas proteínas que comparten un mismo dominio pueden desempeñar funciones similares. A lo largo de varios años de crecimiento, con la incorporación periódica de nuevos dominios, PFAM llegó a alcanzar una cobertura de las secuencias de la UniProt Knowledgebase (UniProtKB) cercana al 80%.

Cada dominio se representa mediante alineamientos múltiples de secuencias y Modelos Ocultos de Markov (HMM, por sus siglas en inglés). De manera simplificada, los HMM son herramientas basadas en análisis estadístico muy útiles que se emplean para modelar procesos y situaciones, basándose en la probabilidad de ocurrencia de etapas claramente definidas dentro de estos procesos y situaciones. Este concepto es particularmente aplicable al reconocimiento de patrones dentro de las cadenas aminoacídicas.

El proceso comienza con la creación de un alineamiento múltiple de proteínas caracterizadas por contener un mismo dominio funcional. Este primer paso constituye la base para la construcción de HMMs. Regularmente se van incorporando nuevas secuencias a los alineamientos desde Swiss-Prot, dando lugar a conjuntos que representan de manera cada vez más precisa a los dominios correspondientes. De esta manera, cada HMM puede representar un dominio funcional evolutivamente conservado que se encuentra presente en diversas proteínas.

Pfam también incluye agrupamientos de nivel superior para los dominios funcionales, dando lugar a “superfamilias” denominadas “clanes”. Estos establecen relaciones entre las diferentes clases de dominios que se encuentran anotadas en la base de datos, permitiendo agrupamientos más abarcativos.

1.6.3 Database of Essential Genes

La base de datos DEG (Database of Essential Genes) (Luo et al. 2021) es un recurso especializado que recopila información sobre genes esenciales en diversos organismos. Los genes esenciales son aquellos cuya eliminación o mutación conduce a la inviabilidad o a la pérdida significativa de función en un organismo, lo que los convierte en elementos clave para el desarrollo y supervivencia celular.

DEG proporciona datos sobre genes esenciales identificados a través de estudios experimentales en escala genómica como mutagénesis, inactivación génica y experimentos de RNAi, entre otros. Esta base de datos incluye información detallada sobre los genes esenciales de organismos procariotas y eucariotas, junto con sus respectivas secuencias,

funciones biológicas y relevancia en distintos contextos, como el desarrollo de nuevos fármacos o el estudio de mecanismos básicos de la biología celular.

En el contexto de la investigación farmacológica, DEG es utilizada para identificar potenciales blancos terapéuticos, ya que los genes esenciales suelen estar involucrados en procesos biológicos fundamentales y, por lo tanto, pueden ser buenos candidatos para la intervención terapéutica.

1.6.4 PDB

El Banco de Datos de Proteínas (PDB) representa el primer recurso digital de acceso abierto en biología dedicado al almacenamiento de estructuras tridimensionales (3D) de proteínas. Establecido por primera vez en 1971 inicialmente con 7 estructuras, ha experimentado un crecimiento exponencial, como se puede observar en la Figura 1.7. Esta plataforma ofrece acceso en 2024 a más de 227,000 entradas de proteínas (y otras macromoléculas como ácidos nucleicos y carbohidratos) tanto *wild type* (variante natural o más común) como modificadas genéticamente, de las cuales una gran cantidad se encuentran formando complejos con moléculas pequeñas (lo cual es de suma importancia para el estudio de interacciones proteína-ligando).

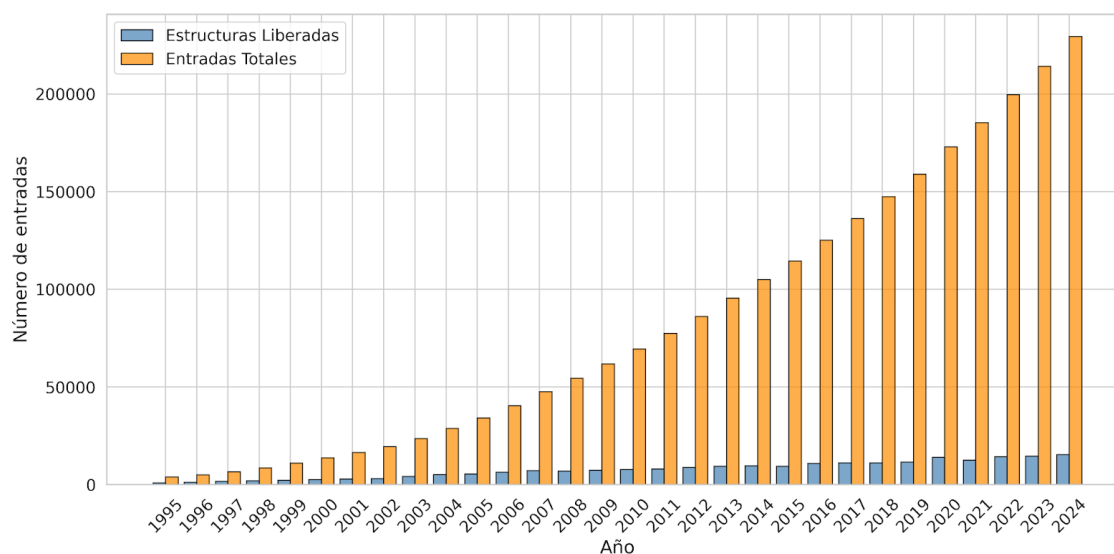


Figura 1.7: Evolución de la cantidad de estructuras depositadas en la base de datos PDB a lo largo de los años. Adaptado de "RCSB PDB",

<https://www.rcsb.org/stats/growth/growth-released-structures>.

Originalmente concebido para la comunidad de biología estructural, el PDB se ha vuelto una herramienta vital para biólogos, científicos computacionales, bioinformáticos, estudiantes, educadores y el público en general. Su archivo, gestionado por el Consorcio Mundial de Banco de Datos de Proteínas (wwPDB) (Behzadi and Gajdács 2021), se encuentra distribuido en centros regionales en Estados Unidos, Europa y Japón. Además, el Centro PDB RCSB (Kouranov et al. 2006), se centra en ofrecer recursos para investigación y educación.

El formato característico de esta base de datos (.pdb) es utilizado para registrar las estructuras 3D de macromoléculas determinadas mediante experimentos de cristalografía de rayos X, resonancia magnética nuclear (RMN) y/o microscopía electrónica 3D (3DEM). Estos documentos incluyen información tal como coordenadas atómicas, datos experimentales y detalles adicionales generados mediante herramientas especializadas.

Por otra parte, los curadores de PDB revisan, anotan y validan cada entrada. Todas las entradas de esta base de datos se encuentran vinculadas extensamente con información experimental y referencias cruzadas (como información sobre drogas, secuencias de proteínas a través de UniProt, dominios de Pfam, entre muchas otras).

La anotación uniforme del PDB facilita las búsquedas y análisis posteriores al organizar la información de manera estandarizada. Para lograrlo, se utilizan conjuntos de reglas y definiciones conocidos como diccionarios de datos, que actúan como guías para estructurar y representar la información. Por ejemplo, el PDB Exchange (PDBx)/Diccionario de Información Cristalográfica de Macromoléculas (mmCIF) define cómo se deben describir las macromoléculas en los archivos del PDB, mientras que el Diccionario de Componentes Químicos (CCD) establece cómo representar los componentes químicos asociados, como ligandos o cofactores. Validaciones basadas en estándares de la comunidad e informes de calidad estructural se proporcionan como metadatos de cada una de las estructuras. La base de datos alberga estructuras de diferentes calidades, y los informes de validación ayudan a los usuarios a seleccionar estructuras de mayor confianza en relación con sus estudios.

El sitio web RCSB PDB integra datos del PDB, información externa y herramientas para consulta, análisis y visualización. Las herramientas de búsqueda y navegación ayudan a refinar consultas, y se ofrecen representaciones de estructuras con visores 3D en línea e interactivos, incluyendo incluso interacciones moleculares tales como puentes de hidrógeno o del tipo pi-stacking. El sitio también contiene mucha información sobre ligandos, siendo posible su búsqueda y análisis en relación con las proteínas que unen.

1.6.5 MOAD

La base de datos MOAD (Mother of All Databases) (L. Hu et al. 2005) es un repositorio especializado que recopila información detallada sobre complejos proteína-ligando obtenidos de estructuras cristalográficas depositadas en el Protein Data Bank (PDB). Su objetivo es proporcionar una visión integral de las interacciones ligando-proteína, lo que es crucial en estudios de diseño de fármacos y *screening* virtual. MOAD se destaca por contener información sobre la afinidad de unión de ligandos, medidas como constantes de disociación (Kd), IC50 y otras métricas que describen la eficacia de la unión entre el ligando y su blanco molecular. Esto permite realizar análisis comparativos entre diferentes ligandos y proteínas, contribuyendo a una mejor comprensión de los factores que influyen en la interacción proteína-ligando.

Una de las principales aplicaciones de MOAD consiste en identificar compuestos biológicamente relevantes, clasificando como verdaderos ligandos a pequeñas moléculas orgánicas, cofactores, péptidos cortos y oligonucleótidos, dejando de lado moléculas irrelevantes como sales o iones.

1.6.6 KEGG

KEGG (Enciclopedia de Genes y Genomas de Kyoto) (Kanehisa and Goto 2000) es una base de datos diseñada para analizar sistemáticamente las funciones de los genes, integrando información genómica con funcionalidad biológica. Con el objetivo de abarcar estas áreas, KEGG consta de tres bases de datos principales: GENES, PATHWAY y LIGAND:

- GENES Database: Esta base de datos almacena catálogos de genomas parcial o completamente secuenciados.
- PATHWAY Database: Contiene representaciones de diversos procesos celulares, entre ellos el metabolismo, el transporte de membranas, la transducción de señales y el ciclo celular.
- LIGAND Database: Almacena información sobre compuestos químicos y sus relaciones con las reacciones enzimáticas.

Las bases de datos se actualizan regularmente y están disponibles de forma gratuita. Además, KEGG proporciona un enfoque único para la anotación funcional de genes, vinculando conjuntos de genes en el genoma con redes metabólicas y actividades enzimáticas.

La estructura del sistema KEGG permite a los usuarios explorar información comenzando desde rutas metabólicas o desde datos genómicos, según sus necesidades. Esto se logra a través de herramientas útiles como navegadores de mapas genómicos,

comparaciones de genomas y herramientas computacionales para la comparación de secuencias y predicción o anotación de vías metabólicas.

1.6.7 ChEMBL

ChEMBL (Zdrazil et al. 2023) es una base de datos de acceso abierto que reúne información sobre moléculas con actividad biológica relevante, desempeñando un papel clave en la investigación y desarrollo en bioquímica y farmacología. Su objetivo principal es compilar datos sobre moléculas pequeñas y su respectiva actividad biológica, complementando con información sobre su grado de avance (o aprobación) en el proceso de investigación clínica. Además, incluye detalles relevantes como mecanismos de acción, indicaciones terapéuticas y otros datos fundamentales para el diseño y evaluación de compuestos farmacéuticos.

La información de esta base de datos se obtiene tanto de la literatura científica como de repositorios de bioactividad, entre ellos PubChem (S. Kim et al. 2015), BioAssay (Y. Wang et al. 2013) y BindingDB (Gilson et al. 2015), consolidando una vasta cantidad de datos (Figura 1.8). La curación de estos datos se realiza de manera manual, lo que incluye procesos rigurosos como la identificación de estructuras químicas y la verificación de referencias bibliográficas.

ChEMBL estructura sus datos en secciones claramente definidas e interrelacionadas, que incluyen información sobre blancos, ligandos, documentos, ensayos y mecanismos de acción, entre otros. La base de datos ofrece una organización visualmente clara y sistemática, permitiendo una exploración eficiente a través de identificadores únicos.

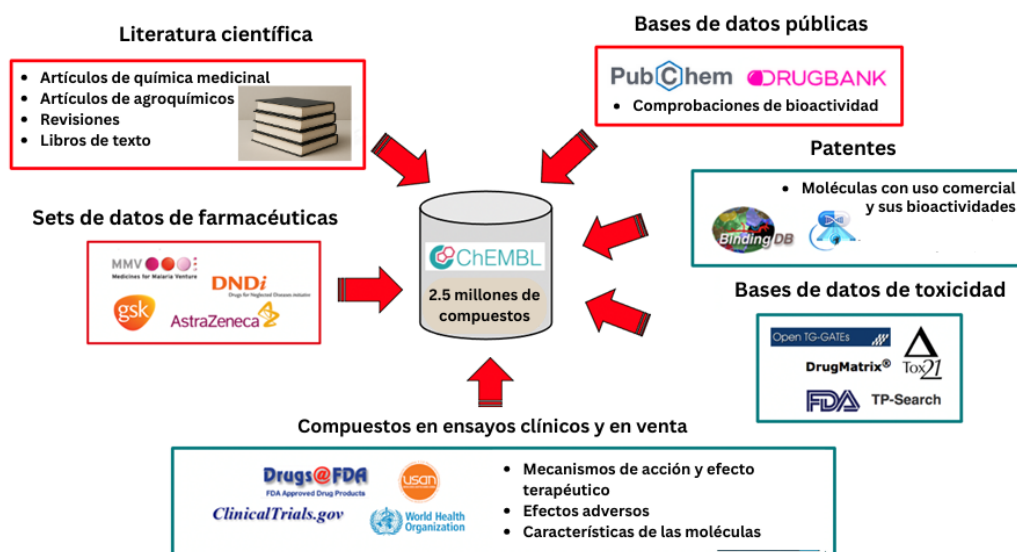


Figura 1.8: Esquema general de la composición de la base de datos ChEMBL. Se observa que esta consiste de la recopilación de datos de diversas fuentes, como literatura científica, sets de datos publicados por grandes empresas farmacéuticas, información sobre compuestos en ensayos clínicos, así como también registros de patentes y anotaciones relacionadas con la toxicidad. Adaptado de "What is ChEMBL?", por EMBL-EBI, <https://www.ebi.ac.uk/training/online/courses/chembl-quick-tour/what-is-chembl/>.

Esta base de datos, al igual que muchas otras, ofrece diferentes formas de acceso: descarga de información, una interfaz web para usuarios tradicionales y servicios web que requieren de habilidades de programación. Estos servicios web, accesibles a través de una API (Interfaz de Programación de Aplicaciones), permiten consultas y operaciones personalizadas y eficientes, lo que posibilita el procesamiento rápido de grandes volúmenes de datos sin la necesidad de descargar la totalidad de la base de datos localmente.

También es posible el uso local de la base de datos en formato SQL, para realizar las consultas y post procesamiento necesario independientemente del estado de la red.

1.6.8 CHEBI

Chemical Entities of Biological Interest (ChEBI) es una base de datos gratuita y accesible, orientada a compuestos químicos de bajo peso molecular relevantes en biología (Hastings et al. 2016). Esta base incluye entidades moleculares, ya sean productos naturales o compuestos sintéticos, que intervienen en procesos biológicos. No se incluyen moléculas directamente codificadas por el genoma, como ácidos nucleicos o proteínas.

ChEBI se organiza mediante una clasificación ontológica que especifica las relaciones entre entidades moleculares, facilitando el análisis y la categorización. Emplea

nomenclatura y terminología avaladas por organizaciones científicas internacionales, como la IUPAC y NC-IUBMB, asegurando un estándar de calidad y trazabilidad.

La base de datos se construye a partir de fuentes confiables, como IntEnz, KEGG COMPOUND, PDBeChem y ChEMBL, y se somete a procedimientos de depuración para eliminar redundancias. Todos los datos están disponibles bajo una licencia Creative Commons (CC BY 4.0) y son completamente referenciables.

ChEBI proporciona información detallada de cada entidad, incluyendo identificador único, nombres recomendados, estructura química, fórmula molecular, masa, carga, nombres IUPAC, sinónimos, marcas comerciales, vínculos a bases de datos externas y números de registro. Además, presenta datos ontológicos y visualizaciones gráficas que facilitan el análisis contextual y estructural.

1.6.9 RHEA

Rhea (Bansal et al. 2022) es una base de datos curada que proporciona información detallada y estandarizada sobre reacciones bioquímicas, incluidas tanto reacciones enzimáticas como de transporte. Los compuestos involucrados en estas reacciones están descritos mediante la ontología química ChEBI (Chemical Entities of Biological Interest), lo que garantiza consistencia y precisión en su representación.

Rhea es ampliamente utilizada en la anotación funcional de enzimas en bases de datos como UniProtKB, facilitando el análisis de su papel en rutas metabólicas específicas. Además, su aplicación se extiende al modelado de sistemas metabólicos, la biología sintética y la metabolómica, constituyéndose como un recurso esencial para el estudio y diseño de procesos metabólicos biológicos.

1.7 Aprendizaje automático

El aprendizaje automático es un campo dentro de la inteligencia artificial dedicado al desarrollo de algoritmos que tienen el objetivo de reconocer patrones y relaciones en grandes conjuntos de datos para poder realizar predicciones o identificar información relevante (Ray 2019).

Una técnica muy destacada dentro del aprendizaje automático es la búsqueda de vecinos cercanos (NNS). Este método se utiliza para la obtención de elementos similares entre sí utilizando alguna métrica específica. En el descubrimiento de fármacos, este enfoque permite predecir la posible bioactividad de una molécula en estudio: si es lo suficientemente parecida a otras drogas cuyos blancos moleculares ya se conocen, se infiere que podría actuar sobre los mismos (Willett 2011).

Uno de los requerimientos más importantes para las herramientas de aprendizaje automático es la conversión de los datos utilizados para las predicciones a un formato capaz de ser procesado por estos algoritmos. En el caso de las moléculas, existen diferentes descriptores moleculares para poder representarlas, dentro de los cuales se encuentran los *Fingerprints*, que serán explicados detalladamente en la sección de métodos (Lange et al. 2024).

La Figura 1.9 ilustra de manera general el proceso requerido para realizar la predicción del blanco molecular de un compuesto en estudio, el cual involucra la conversión de moléculas a un formato de descriptores moleculares procesables por los algoritmos de búsqueda de vecinos, el armado de una base de datos de compuestos con actividades conocidas y posteriormente la predicción sobre la droga con blanco molecular a determinar.

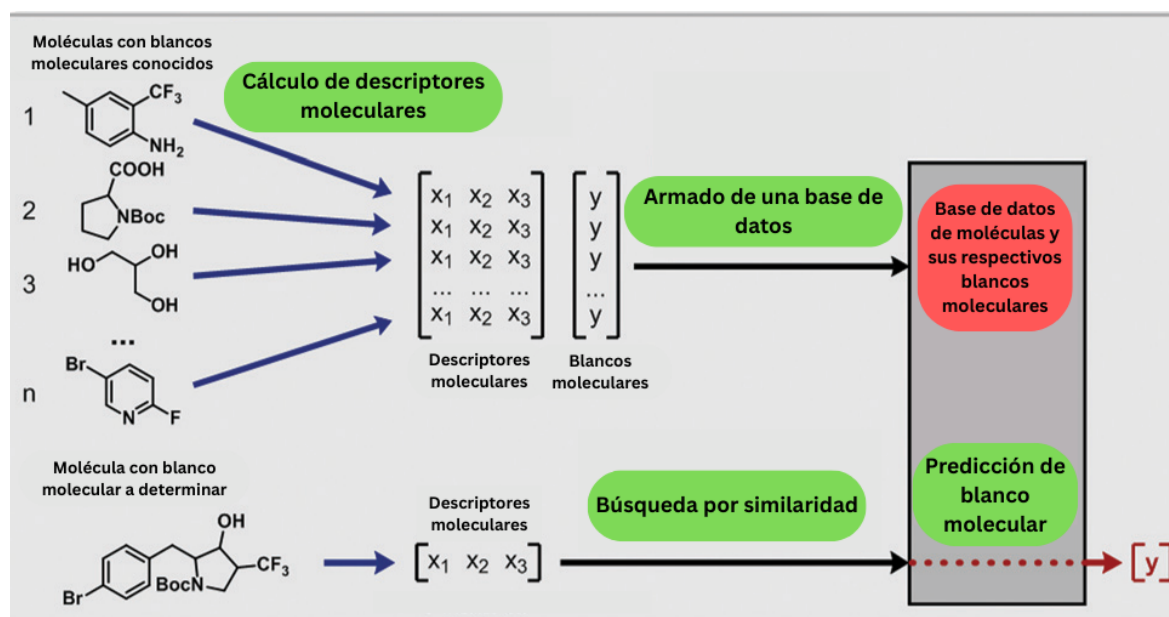


Figura 1.9: Flujo general para realizar predicciones mediante búsquedas de vecinos cercanos aplicado al procesamiento de compuestos químicos: 1) Conversión a descriptores: las moléculas con blanco conocido se representan como vectores numéricos de descriptores moleculares. 2) Construcción de la base de datos: los vectores y sus blancos se almacenan en una base para posteriormente establecer predicciones. 3) Búsqueda por similitud: la molécula con blanco molecular a determinar se transforma al mismo formato de descriptores y se compara con la base para localizar los vecinos más similares; los blancos asociados a estos vecinos se emplean para predecir el blanco molecular de la nueva molécula. Adaptado de Strieth-Kalthoff et al. 2020.

1.8 Objetivos de la tesis

Esta investigación tiene como objetivo principal desarrollar flujos de trabajo computacionales para asistir a la comunidad científica en la identificación de blancos moleculares de compuestos antimicrobianos y de sustratos o moléculas de interés biotecnológico en microorganismos, ya sean patogénicos o relevantes para procesos industriales. El proyecto integra herramientas de bioinformática, quimioinformática, ciencia de datos y aprendizaje automático, enfocándose en dos aspectos clave para abordar estos desafíos:

El primer enfoque se centra en el desarrollo de una plataforma diseñada para predecir blancos moleculares específicos de una droga experimental con actividad inhibitoria probada sobre un patógeno. Este trabajo adopta un nivel de análisis general, abordando el proteoma completo de un organismo y combinando técnicas de bioinformática, quimioinformática, ciencia de datos y aprendizaje automático.

El segundo enfoque, más específico, complementa la metodología desarrollada en la primera etapa, incorporando técnicas adicionales para la caracterización de sustratos, productos e inhibidores (SPIs) en conjuntos de proteínas pertenecientes a una familia determinada. Como caso de estudio, se utilizó la familia de los Citocromos P450 bacterianos (BacCYPs), pero se propone como una estrategia aplicable a diferentes familias de proteínas.

2. Métodos

2.1 Métodos de quimioinformática

En el ámbito de la química computacional, la representación adecuada de las estructuras moleculares es crucial para su análisis y comparación. Los formatos textuales como SMILES (Simplified Molecular Input Line Entry System) y las representaciones binarias conocidas como *Fingerprints* moleculares han permitido un avance significativo en la manipulación y procesamiento informático de compuestos químicos. Estas herramientas, junto con métricas como el Índice de Tanimoto, facilitan la comparación estructural y la identificación de relaciones moleculares, contribuyendo de manera esencial al descubrimiento de nuevos fármacos y al análisis de grandes bases de datos químicas. Librerías con una amplia diversidad de funcionalidades como RDKit (<http://www.rdkit.org>) facilitan la aplicación computacional de todos estos procesos.

A lo largo de esta sección, se detallan estas representaciones y métricas, su aplicabilidad y su importancia en el desarrollo de estrategias computacionales para el diseño de fármacos.

2.1.1 Representación de compuestos químicos en formato de texto (SMILES)

Con el fin de comparar estructuras de moléculas de manera computacional, se requiere un formato adecuado, ya que la representación usual en estructura 2D no es de fácil procesamiento en los contextos de aplicaciones informáticas y de algoritmos de búsqueda y almacenamiento. En las bases de datos que almacenan información sobre estructuras químicas, se desarrollaron representaciones alternativas para moléculas pequeñas. Una de las más extendidas es el formato SMILES (*Simplified Molecular Input Line Entry System*), una notación lineal que facilita la representación de moléculas para su procesamiento informático (Anstein and Kremer 2005).

A diferencia de las representaciones bidimensionales tradicionales, los SMILES contienen la misma información que se halla en una tabla de conectividad (representación tabular que contiene enlaces y tipos de átomos), pero ofrecen la ventaja de ser considerablemente más compactos y manejables. Esto se debe a que los SMILES son construcciones lingüísticas (en formato de texto) en lugar de estructuras de datos computacionales.

El proceso de generación de una molécula en formato SMILES consta de los siguientes pasos:

- **Comienzo con la estructura 2D completa:** se parte de la estructura en dos dimensiones de la molécula con todos los átomos y los hidrógenos explícitos.
- **Supresión de hidrógenos:** se omiten los átomos de hidrógeno (H), que quedan implícitos en el formato final. Solo se conservan los que aportan información especial (por ejemplo [nH]).
- **Detección de ciclos:** se identifican los anillos y se asignan dígitos a los átomos que marcan la apertura y el cierre de cada uno.
- **Inicio del recorrido:** se selecciona un átomo inicial y un orden de recorrido.
- **Redacción de la cadena:** se redacta la cadena del SMILES siguiendo el recorrido elegido cumpliendo con las siguientes pautas: los símbolos atómicos correspondientes a cada átomo se escriben en mayúscula, exceptuando los pertenecientes a anillos aromáticos que se escriben en minúscula; los enlaces dobles se denotan con "=", y los triples con "#"; las ramificaciones se encierran entre paréntesis; los átomos que indican apertura o cierre de anillos aromáticos se denotan con dígitos.
- **Anotaciones adicionales:** se añaden en caso de ser necesario quiralidades (@), configuraciones E/Z (/, \), cargas ([N+], [O-]) o isótopos ([13C]).

La Figura 2.1 ilustra el proceso de construcción de un SMILES para una molécula relativamente sencilla, como es la Aspirina (ácido acetilsalicílico), fármaco muy conocido y utilizado como anti inflamatorio.

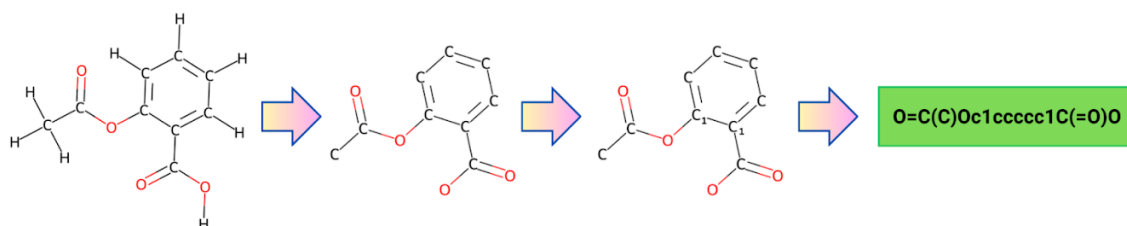


Figura 2.1: Conversión paso a paso de la estructura de la Aspirina al formato SMILES: se parte de la molécula en formato de dos dimensiones y, como primer paso, se ocultan los átomos de hidrógeno, que se excluyen en la notación. A continuación se identifican los seis carbonos del anillo aromático; los dos que enlazan con otros átomos de la molécula se marcan con el dígito 1, indicando en dónde se abre y se cierra el ciclo. Después de los ajustes se redacta la cadena SMILES: las ramificaciones se encierran entre paréntesis, los dobles enlaces se expresan con "=", y los átomos

aromáticos se escriben en minúscula, señalando como "c1" a los carbonos donde inicia y donde se cierra el ciclo. El resultado para la Aspirina es "O=C(C)Oc1ccccc1C(=O)O".

Entre las propiedades destacadas del formato SMILES se encuentra su capacidad para contener información detallada sobre los átomos y enlaces presentes en una molécula con una alta simplicidad, facilitando su manipulación. Esto ha impulsado su amplia adopción en el manejo computacional de información química. La versatilidad de los SMILES se evidencia en su aplicabilidad en algoritmos de búsqueda, almacenamiento informático y además funcionan como punto de partida para muchas herramientas utilizadas para la comparación estructural, lo que contribuye significativamente a avances en el diseño y análisis de moléculas en diversas disciplinas científicas (Weininger 2002).

Comúnmente, las herramientas ampliamente utilizadas para el procesamiento de moléculas, como RDkit, utilizan como entrada representaciones en formato SMILES para la generación de representaciones vectoriales adecuadas para establecer comparaciones entre compuestos químicos.

2.1.2 *Fingerprints* moleculares

Con el objetivo de facilitar la comparación eficiente de estructuras moleculares en química computacional, se recurre a los *Fingerprints* moleculares (huellas moleculares). Este tipo de representación vectorial desempeña un papel crucial al complementar las representaciones tradicionales, como el formato SMILES que corresponde a una cadena de texto. Los *Fingerprints* son particularmente útiles en aplicaciones informáticas relacionadas con algoritmos de búsqueda, comparación estructural y almacenamiento eficiente.

Los *Fingerprints* moleculares son representaciones numéricas binarias que codifican una molécula. A diferencia de las representaciones tridimensionales, que dependen de la disposición exacta de átomos en el espacio, los *Fingerprints* se centran en subestructuras de los compuestos químicos, es decir fragmentos o patrones específicos dentro de las moléculas que involucran a los tipos de átomos y sus enlaces. Por lo tanto, una representación molecular mediante *Fingerprints* corresponde a un vector binario donde cada posición indica la presencia (valor de 1) o ausencia (valor 0) de una subestructura química específica (Figura 2.2). Una de las principales ventajas de este formato es que cada posición del vector representa una misma subestructura química (o un mismo grupo de subestructuras), independientemente de la molécula representada. Esto implica que compuestos químicos diferentes son comparables con este método (Muegge and Mukherjee 2016).

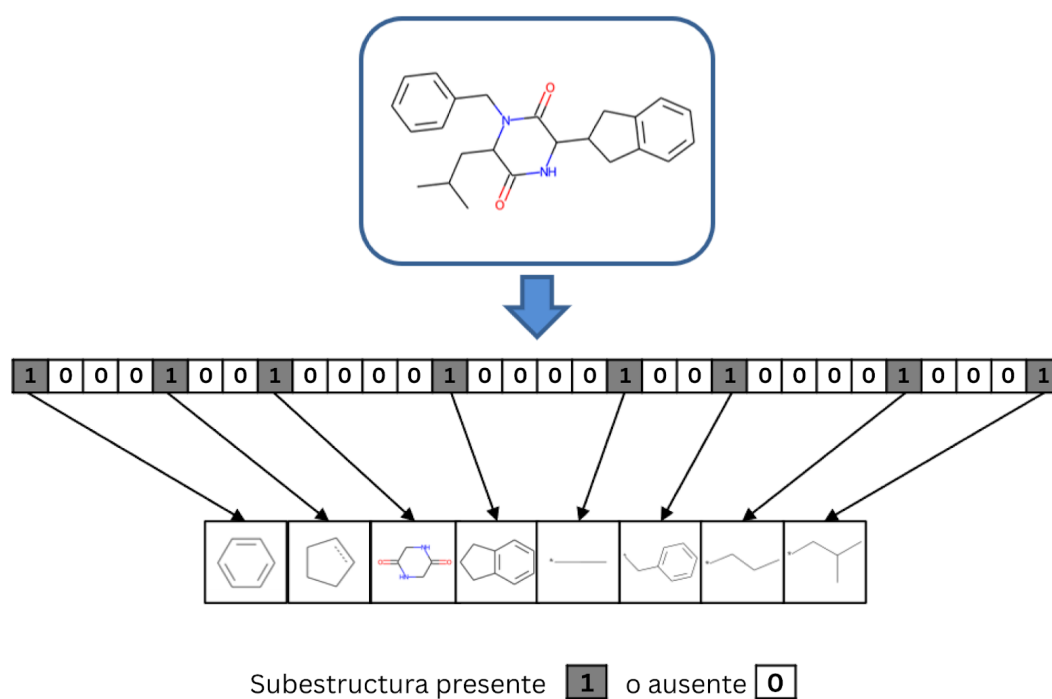


Figura 2.2: Representación de cómo una molécula química es convertida en un vector a través de la generación de *Fingerprints* Moleculares. Adaptado de Janela, Takeuchi y Bajorath 2022.

Existen diversos enfoques para calcular la representación de *Fingerprints* para una molécula, los cuales divergen en la manera por la cual se definen las subestructuras que contiene el compuesto químico. Uno de los más destacados y ampliamente utilizados corresponde a los *Fingerprints* de Morgan (Morgan 2002). Esta metodología considera como subestructura a cada átomo de la molécula en conjunto con su contexto químico inmediato.

El contexto máximo incluido en la subestructura correspondiente a un átomo determinado de una molécula está definido por un parámetro denominado radio (Figura 2.3), cuyos valores se interpretan de la siguiente manera:

- Radio 0: se considera sólo el átomo central.
- Radio 1: incluye el átomo central y los átomos directamente enlazados a él (vecinos directos).
- Radio 2: contiene el átomo central, sus vecinos directos y los átomos enlazados a esos vecinos, y así sucesivamente para radios mayores.

El proceso es jerárquico y acumulativo: al calcular un *Fingerprint* con un radio dado, el algoritmo incorpora automáticamente todas las subestructuras correspondientes a los

radios menores. Así, un *Fingerprint* generado con radio 2 incluye, además de las subestructuras de radio 2, las de radios 0 y 1.

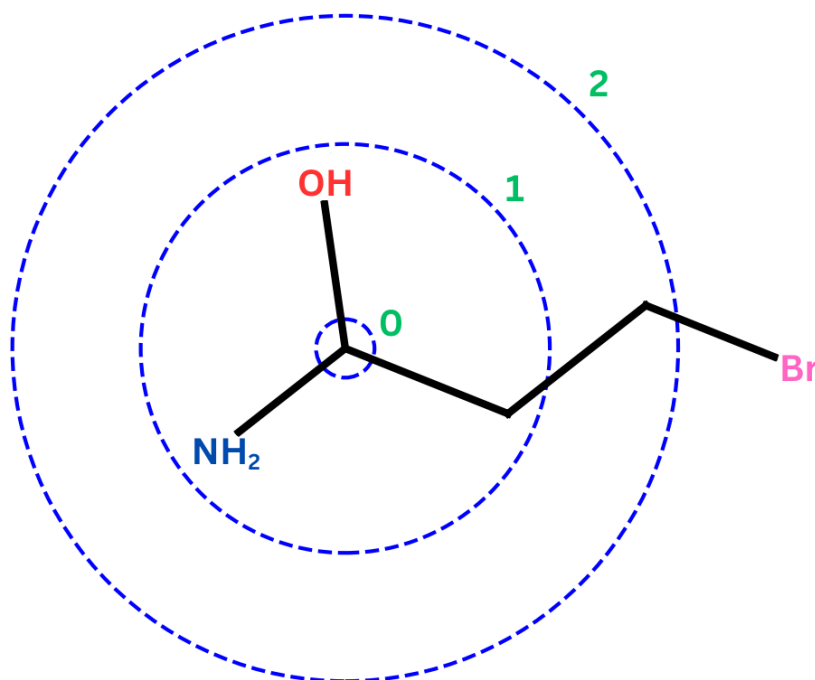


Figura 2.3: Definición del contexto químico máximo considerado utilizando *Fingerprints* de Morgan. El carbono central (radio 0) se toma como referencia y, a su alrededor, se definen círculos concéntricos (trazado discontinuo en azul) que delimitan los radios 1 y 2. Cada valor de radio (numerado en verde) toma en consideración como parte de la subestructura más grande posible el contexto químico englobado por su círculo correspondiente.

Dado que la expansión se realiza en círculos concéntricos alrededor de cada átomo, estos descriptores se conocen también como *Fingerprints* circulares. De esta manera, es posible obtener una representación precisa de la molécula considerando subestructuras con un grado de complejidad suficiente como para establecer comparaciones.

En contraposición, los *Fingerprints* de tipo Daylight (Wild and Blankley* 1999) describen a la subestructura como una secuencia lineal de átomos en lugar de entornos circulares.

Una vez definidas las subestructuras que representan la molécula, éstas se codifican en un vector binario de longitud fija, la cual representa un segundo parámetro relevante, que se denomina número de bits. Un vector de menor longitud (entre 256 y 512 bits) permite un bajo uso de memoria y mayor eficiencia computacional, pero puede traer

como consecuencia que diferentes subestructuras queden agrupadas en una misma posición del vector. Por otro lado, la elección de un número de bits mayor (por lo general 1024 bits o más) permite una diferenciación más eficiente de las subestructuras, pero implica un costo computacional mayor y un uso de memoria elevado (Gütlein and Kramer 2016).

En cuanto al uso práctico de esta metodología, un componente esencial en la implementación y manejo de *Fingerprints* moleculares es RDKit, una poderosa librería de código abierto ampliamente utilizada en la química computacional y el diseño de fármacos. RDKit proporciona una infraestructura robusta para manejar y procesar datos químicos. Ofrece herramientas avanzadas para la manipulación de estructuras moleculares, la conversión de moléculas proporcionadas en diferentes formatos (como SMILES) a *Fingerprints* moleculares.

Gracias a la flexibilidad y eficiencia de RDKit, es posible automatizar y escalar procesos complejos de análisis y comparación de grandes bases de datos químicas. En el contexto de este trabajo, RDKit resulta una herramienta clave para convertir representaciones moleculares almacenadas en formato SMILES a *Fingerprints* de Morgan de manera eficiente con los recursos computacionales disponibles.

2.1.3 Comparación de compuestos químicos

En el ámbito de la química computacional, el Índice de Tanimoto resulta una métrica esencial para evaluar la similitud química entre dos compuestos químicos. Esta medida de similitud se utiliza ampliamente en la comparación de *Fingerprints* moleculares, proporcionando un coeficiente que refleja la proporción de subestructuras compartidas entre dos moléculas (Bajusz, Rácz, and Héberger 2015). El Índice de Tanimoto (TI) se calcula mediante la fórmula 2.1:

$$TI = \frac{c}{(a+b-c)} \quad (2.1)$$

donde c es el valor numérico que corresponde a las subestructuras compartidas por ambos compuestos, mientras que a y b corresponden a los números respectivos de subestructuras que contiene cada una de las dos moléculas. Esto da como resultado una puntuación representada por un rango comprendido entre 0 y 1, en el cual cuanto más alto es el valor, mayor es la similaridad (Willett 2014).

En el contexto de los *Fingerprints* moleculares, esta fórmula se aplica a las representaciones binarias de las moléculas, donde cada posición del vector indica la presencia o ausencia (según si su valor es 1 o 0, respectivamente) de determinadas subestructuras.

Este enfoque es particularmente útil en el *screening* virtual de moléculas, donde la identificación de compuestos similares, que tienden a compartir actividades biológicas de interés, puede ser determinante para acelerar el proceso de descubrimiento de fármacos. Además, El Índice de Tanimoto encuentra aplicaciones en la clasificación y agrupación de compuestos químicos, facilitando la organización eficiente de grandes conjuntos de datos químicos.

En la Figura 2.4 se muestra un ejemplo ilustrativo del cálculo del Índice de Tanimoto aplicado a la comparación mediante *Fingerprints* moleculares de dos compuestos diferentes, resaltando las características compartidas y la contribución de esta métrica a la evaluación de similitud.

Debido a su practicidad y eficiencia, el Índice de Tanimoto se posiciona como una herramienta clave en química computacional, ofreciendo una medida cuantitativa precisa para la comparación de compuestos químicos utilizando sus representaciones de *Fingerprints* moleculares.

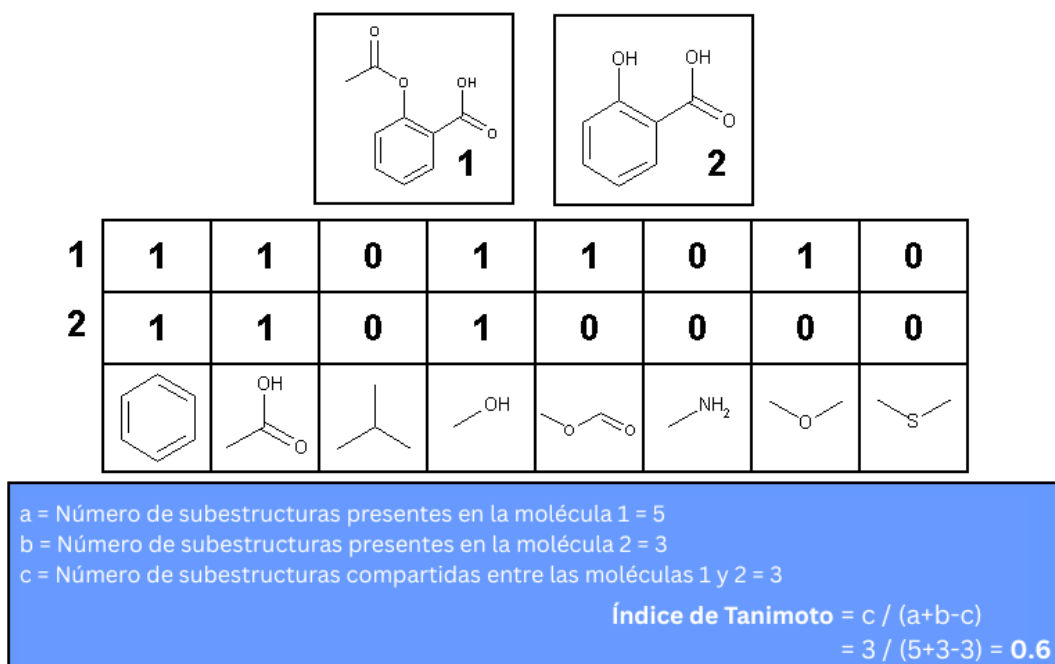


Figura 2.4: Ilustración representativa de la similitud dada por el Índice de Tanimoto entre dos compuestos químicos diferentes, destacando sus subestructuras compartidas y las diferencias, y su traducción correspondiente a vectores binarios. Adaptado de "Molecular Similarity Search: A Simple but Powerful Drug Discovery Tool," *Medium*, <https://medium.com/gsi-technology/molecular-similarity-search-a-simple-but-powerful-drug-discovery-tool-2b991d78c191>.

2.2 Métodos de bioinformática

Los métodos bioinformáticos juegan un papel fundamental en la investigación moderna, facilitando el análisis y la interpretación de secuencias biológicas de diverso tipo (como ADN y aminoácidos). Estas herramientas permiten desde el alineamiento de secuencias múltiples hasta la construcción de árboles filogenéticos, proporcionando información crucial sobre la evolución y funcionalidad de proteínas y otras estructuras de importancia en la biología. A lo largo de esta sección, se describen diversas metodologías representativas de la bioinformática, como el uso de modelos ocultos de Markov (HMMs), el análisis de redes metabólicas y la búsqueda de sustratos por contexto genético. Estos enfoques permiten explorar de manera eficiente desde la relación estructural y funcional de proteínas, hasta la identificación de reacciones químicas relevantes dentro de las redes metabólicas, aspectos clave en el estudio de patógenos y en el descubrimiento de nuevos compuestos terapéuticos.

2.2.1 BLAST

El algoritmo BLAST (*Basic Local Alignment Search Tool*) es una herramienta fundamental en bioinformática, diseñada para comparar secuencias biológicas y detectar regiones de similitud entre ellas. Su objetivo principal es identificar secuencias homólogas a una secuencia de consulta, tanto a nivel de nucleótidos como de proteínas, permitiendo inferir relaciones funcionales y evolutivas entre organismos (Boratyn et al. 2013).

Existen distintas variantes de BLAST, cada una diseñada para trabajar con distintos tipos de secuencias:

- BLASTn: Se utiliza para comparar secuencias de nucleótidos (ADN o ARN). Es comúnmente empleado en la búsqueda de homología entre secuencias genómicas o transcriptómicas.
- BLASTp: Está diseñado para comparar secuencias de proteínas. BLASTp es útil para identificar proteínas homólogas y analizar su similitud funcional.
- BLASTx: Permite comparar una secuencia de nucleótidos traducida en sus posibles marcos de lectura con una base de datos de secuencias de proteínas. Es ideal cuando se tienen secuencias de ADN o ARN y se quiere encontrar su homología con proteínas.
- TBLASTn: Utiliza una secuencia de proteína para buscar similitudes en una base de datos de secuencias de nucleótidos, traduciendo estos en los seis posibles marcos de lectura.

- TBLASTx: Traduce ambas secuencias, tanto la consulta como la base de datos de nucleótidos, en sus seis posibles marcos de lectura para identificar similitudes a nivel de proteínas.

BLAST funciona mediante la búsqueda de segmentos similares entre la secuencia de consulta y las secuencias almacenadas en bases de datos. Utiliza un enfoque heurístico para identificar alineamientos locales que usualmente consta de los siguientes dos pasos principales (Altschul et al. 1997):

- **Búsqueda de subsecuencias exactas**

Divide la secuencia de consulta en k -meros, subsecuencias cortas de largo k . Busca coincidencias exactas de estas subsecuencias en la base de datos. Esto filtra rápidamente más del 99 % de las secuencias.

- **Extensión guiada por puntaje**

El alineamiento de cada subsecuencia coincidente se extiende hacia ambos lados pero solo mientras el puntaje acumulado supere un umbral H . Si el puntaje cae por debajo de H , la extensión se detiene. Las extensiones que alcanzan un segundo umbral S se reportan como alineamientos candidatos.

Como ejemplo ilustrativo, para la secuencia de consulta **MEFGH** y $k=3$, BLAST encuentra las subsecuencias **MEF**, **EFG**, **FGH**. Supongamos que la base contiene la secuencia **AEFGHIK**. Coinciden las subsecuencias **EFG** y **FGH**; BLAST extiende alrededor de ellas y llega al alineamiento local **EFGH** con puntaje 34, por encima de S , mientras descarta coincidencias parciales de menor puntaje en otras secuencias.

Este enfoque, al ser heurístico no contempla todos los alineamientos posibles (se pierden los que no presentan coincidencias en las subsecuencias), pero reduce el tiempo de búsqueda de horas a segundos en bases de datos de gran tamaño.

Los resultados incluyen una lista de secuencias similares ordenadas por puntuación de similitud y significancia estadística, lo que permite identificar secuencias relevantes para estudios de homología, función o estructura.

El *blast score* o puntuación BLAST refleja la calidad del alineamiento entre la secuencia de consulta y las secuencias con las que ésta se compara, siendo una medida de similitud cuantitativa que evalúa tanto la longitud como la calidad del alineamiento. A mayor *blast score*, mayor es la probabilidad de homología funcional. Por otro lado, la identidad de secuencia es el porcentaje de coincidencia exacta de los residuos en las posiciones alineadas entre dos secuencias, lo que permite determinar qué tan estrechamente relacionadas están las secuencias en términos evolutivos. Otras métricas muy utilizadas para el análisis de resultados de esta herramienta son la cobertura, que

corresponde a la porción de cada secuencia que comprende el alineamiento, y el *e-value*, que indica la cantidad esperada de alineamientos de calidad similar que podrían ocurrir por azar.

2.2.2 Alineamiento de secuencias múltiples

Un alineamiento de secuencias es una manera de comparar secuencias de ADN, ARN o aminoácidos para identificar regiones similares, identificando similitudes y patrones conservados que son fundamentales para analizar relaciones evolutivas y características funcionales.

Los alineamientos múltiples permiten comparar simultáneamente tres o más secuencias biológicas. Se utilizan con diversos propósitos, entre ellos la identificación de patrones característicos en familias proteicas, la construcción de árboles filogenéticos, la asignación por homología de secuencias nuevas de proteínas a familias existentes, así como también para predecir estructura secundaria y terciaria de proteínas sin conocer la disposición tridimensional de sus átomos (Nguyen, Guo, and Pan 2016).

En comparación con los alineamientos de a pares, los alineamientos múltiples resultan más confiables para evaluar la similitud a gran escala, ya que la inclusión de un mayor número de secuencias tiende a resolver las ambigüedades presentes en la comparación entre pares de secuencias. Además, ofrecen información más detallada sobre las posiciones de aminoácidos individuales, incluyendo aspectos como la conservación general y las relaciones evolutivas, lo cual es especialmente relevante al emplear métodos de comparación para la construcción de árboles filogenéticos (Rychlewski et al. 2000). La Figura 2.5 muestra un ejemplo de alineamiento múltiple de secuencias de aminoácidos.

MAFFT (Katoh et al. 2002) es uno de los programas de referencia para alinear múltiples secuencias de ADN o proteínas. Su éxito radica en que combina velocidad y eficiencia: construye primero un alineamiento preliminar rápido y luego lo refina en varias rondas iterativas hasta que las posiciones conservadas quedan correctamente alineadas. Además, el programa puede incorporar datos estructurales cuando se dispone de estructuras experimentales.

En este contexto, MAFFT-DASH (Rozewicki et al. 2019) incorpora información estructural tomada de la base de datos DASH (*Database of Aligned Structural Homologs*). Para cada secuencia de entrada, el programa localiza estructuras homólogas en el PDB y las incorpora temporalmente como secuencias guía en el alineamiento; los residuos que coinciden en la superposición tridimensional se utilizan como 'anclas' durante las rondas de refinamiento. Estas anclas estructurales indican qué posiciones ocupan lugares equivalentes en el espacio y ayudan a mantenerlas alineadas incluso cuando las

secuencias comparten muy poca identidad. De esta manera, MAFFT-DASH genera alineamientos más confiables especialmente cuando se encuentran secuencias distantes entre sí, lo que resulta clave al reconstruir la evolución de grandes familias de proteínas.

Por último, otra de las representaciones más ampliamente utilizadas derivadas de los alineamientos múltiples consiste en construir un modelo oculto de Markov (HMM, por sus siglas en inglés), esto es, un modelo probabilístico que resume los patrones de conservación y variabilidad observados en el conjunto de secuencias (Schuster-Böckler, Schultz, and Rahmann 2004). Esta representación resulta muy útil al reconocer dominios funcionales en secuencias de proteínas, con el objetivo de predecir la funcionalidad de proteínas sobre las que no se dispone de información adicional asociada más allá de su composición aminoacídica.



Figura 2.5: Ejemplo de un alineamiento múltiple de secuencias de aminoácidos. Figura tomada de "Multiple Sequence Alignment," *Bioinformatics UPV*, <https://bioinf.comav.upv.es/courses/biotech3/theory/multiple.html>.

Otra aplicación a considerar en el marco de los alineamientos múltiples consiste en el agrupamiento de secuencias similares en conjuntos de datos extremadamente grandes. Esto resulta muy útil para poder realizar alineamientos múltiples evitando la sobrepoblación de ciertos grupos de secuencias muy similares entre sí, reduciendo significativamente el costo computacional. Este enfoque es crucial en diversos casos, como el ejemplo del universo de todos los Citocromos P450 bacterianos conocidos. Una herramienta representativa y ampliamente utilizada es *Cd-hit* (Li and Godzik 2006), que permite reducir la redundancia en bases de datos de secuencias y mejorar el rendimiento de los análisis subsecuentes.

Cd-hit emplea un algoritmo incremental y basado en estrategias de agrupamiento rápido, comenzando con la secuencia más larga como representante inicial y evaluando

cada secuencia restante en función de su similitud con las ya agrupadas. Las similitudes se estiman utilizando conteos comunes de palabras (*word indexing*) para filtrar alineamientos innecesarios, optimizando así la clasificación de secuencias en grupos. Este método es particularmente ventajoso cuando se trabaja con volúmenes masivos de datos, y ha sido mejorado para su implementación en entornos multi-núcleo, lo que permite procesar de forma paralela y eficiente grandes conjuntos de secuencias.

2.2.3 Árboles filogenéticos

Los árboles filogenéticos son herramientas esenciales en biología y genómica para visualizar y comprender las relaciones evolutivas. Estos árboles representan gráficamente la divergencia evolutiva y la ancestralidad común tanto a nivel de organismos como en secuencias nucleotídicas y aminoacídicas, ofreciendo una perspectiva clara de las relaciones filogenéticas alcanzadas a lo largo de la historia evolutiva (Kapli, Yang, and Telford 2020).

La construcción de árboles filogenéticos comienza con un alineamiento múltiple de secuencias (Figura 2.6), un paso esencial para identificar similitudes y diferencias en el conjunto de secuencias analizadas.

Para mejorar la calidad del análisis, se eliminan regiones poco informativas del alineamiento, reduciendo el impacto de inserciones, deleciones y regiones altamente variables que podrían introducir ruido en los resultados finales (Bogusz and Whelan 2016). En este proceso, es muy utilizada la herramienta BMGE (*Block Mapping and Gathering with Entropy*) (Criscuolo and Gribaldo 2010), un software diseñado para filtrar alineamientos eliminando posiciones ambiguas o poco informativas. BMGE utiliza un enfoque basado en la entropía de cada posición del alineamiento para determinar qué columnas deben conservarse, asegurando que sólo las regiones conservadas y evolutivamente relevantes sean retenidas en el análisis filogenético. Este enfoque mejora la precisión del árbol filogenético al reducir el ruido introducido por regiones altamente variables o mal alineadas, lo que es crucial para obtener una representación más fiable de las relaciones evolutivas.

Los parámetros más comúnmente utilizados en BMGE, que son importantes en este trabajo son los siguientes:

- -b (tamaño mínimo de bloque): Define el tamaño mínimo de las regiones conservadas que deben ser consideradas como bloques informativos.
- -w (tamaño de ventana): Especifica el número de posiciones consecutivas que se analizan conjuntamente para calcular la entropía local.
- -h (umbral de entropía): Determina el nivel máximo de variabilidad aceptable en cada posición.

- -g (proporción máxima de gaps): Indica el porcentaje máximo de gaps permitido en una columna para que esta sea conservada.

Existen numerosos métodos para la construcción de árboles filogenéticos a partir de los alineamientos múltiples de secuencias. En el presente trabajo se utilizó PhyML (Guindon et al. 2009). PhyML es un método de máxima verosimilitud ampliamente utilizado en estudios evolutivos. Este software optimiza la topología del árbol, las longitudes de las ramas y los parámetros del modelo evolutivo para obtener la mejor representación posible de las relaciones evolutivas. Este método es fundamental para garantizar la robustez y precisión del árbol generado, especialmente en conjuntos de datos complejos como los de los Citocromos P450 bacterianos.

Además, para seleccionar el modelo evolutivo más adecuado, se desarrolló el módulo SMS (*Smart Model Selection*) integrado en PhyML (Lefort, Longueville, and Gascuel 2017). SMS evalúa automáticamente diferentes modelos evolutivos y selecciona el óptimo basándose en criterios de información como el AIC (Criterio de Información de Akaike) y el BIC (Criterio de Información Bayesiano), lo que permite una inferencia más eficiente y precisa. La selección automática del modelo adecuado es esencial para maximizar la calidad de las inferencias evolutivas, reduciendo el riesgo de seleccionar un modelo incorrecto que podría afectar la interpretación de las relaciones filogenéticas.

La construcción del árbol filogenético además de identificar las relaciones evolutivas entre proteínas de una familia, permite la identificación de grupos de proteínas con características similares. Estos grupos, definidos por patrones compartidos en sus secuencias, simplifican la comprensión de la diversidad funcional dentro de esta familia de proteínas y proporcionan un marco sólido para la caracterización de nuevas proteínas. Así, proteínas recién descubiertas pueden ser asignadas a categorías específicas basadas en su similitud de secuencia (Sharma, Jaloree, and Thakur 2018).

Por otro lado, dado que proteínas con alta similitud filogenética tienden a desempeñar funciones similares, como actuar sobre sustratos parecidos, es posible inferir información valiosa sobre los compuestos que podrían unir. La proximidad filogenética, revelada a través del árbol, se convierte así en una herramienta esencial para inferir funciones potenciales y comprender mejor la diversidad de interacciones de proteínas con sus sustratos, productos o inhibidores, contribuyendo significativamente al conocimiento de su funcionalidad.

En el contexto de nuestro Estudio, hemos aplicado esta metodología avanzada al estudio de la diversidad de Citocromos P450 bacterianos, como parte de un enfoque predictivo de sustratos, productos e inhibidores.

cutoff, un umbral específico otorgado por Pfam para cada perfil para separar coincidencias reales de falsos positivos—con algún perfil de la base de datos. Este procedimiento permite predecir funciones moleculares: por ejemplo, si una proteína de organismo patogénico muestra un dominio conservado de citocromo P450 o kinasa tras su comparación con los perfiles de Pfam, se infiere que actúa como tal.

En cuanto a las herramientas destacadas en el ámbito del análisis de secuencias, la más relevante es el programa HMMER (Potter et al. 2018), una aplicación diseñada específicamente para encontrar y anotar dominios funcionales en secuencias de proteínas y ácidos nucleicos utilizando Modelos Ocultos de Markov (HMMs). HMMER compara una secuencia de consulta con una base de datos de HMM, lo que permite identificar patrones conservados y asignar funciones a proteínas desconocidas.

HMMER también emplea perfiles generados a partir de alineamientos múltiples de secuencias relacionadas, capturando la variabilidad y la conservación de una familia de proteínas. Esto lo hace ideal para identificar dominios funcionales dentro de secuencias complejas y altamente divergentes. El programa es ampliamente utilizado en conjunto con la base de datos Pfam, que contiene modelos HMM para miles de familias de proteínas. Existen varios modos de uso de HMMER, entre ellos:

- *hmmsearch*: Busca secuencias de proteínas que coincidan con perfiles HMM almacenados en bases de datos como Pfam.
- *hmmsearch*: Permite buscar perfiles HMM específicos en bases de datos de secuencias, para encontrar todas las secuencias que correspondan a un modelo oculto en particular.
- *phmmer*: Realiza búsquedas rápidas de similitud de secuencias, similar a BLAST, pero utilizando perfiles HMM para una mayor precisión en la detección de homología.
- *jackhmmer*: Ejecuta una búsqueda iterativa similar a PSI-BLAST, lo que permite mejorar progresivamente los perfiles de secuencia a medida que se identifican nuevas homologías.

En este trabajo, HMMER se utilizó en combinación con los modelos HMM presentes en Pfam, lo que permitió la identificación de dominios conservados en diversas proteínas. Esta herramienta resulta fundamental para inferir la funcionalidad y posibles interacciones de diferentes proteínas.

2.2.5 Propiedades de las proteínas y sus reacciones (esencialidad, drogabilidad, puntos de estrangulamiento y centralidad)

La exploración de características como la esencialidad de proteínas, los puntos de estrangulamiento (*choke-points*) y la centralidad de las reacciones catalizadas en redes metabólicas resulta crucial para comprender la dinámica y función de los sistemas biológicos. Estos enfoques ayudan a identificar las proteínas clave dentro del proteoma de un organismo determinado y son fundamentales para determinar cuáles pueden ser blancos terapéuticos de antibióticos. Más relevante aún para este estudio, permiten identificar las proteínas que podrían verse afectadas por una droga con efecto antimicrobiano cuyo mecanismo de acción es desconocido.

Esencialidad:

La esencialidad de una proteína se refiere a su papel crítico para la supervivencia o el funcionamiento celular. Una proteína esencial desempeña funciones indispensables en procesos vitales, y su pérdida o inactivación puede tener graves consecuencias para la viabilidad de un organismo o célula (Jeong, Oltvai, and Barabási 2002). Sin embargo, la esencialidad no es una propiedad absoluta, sino que depende del contexto biológico y ambiental en el que se evalúe. Algunas proteínas pueden ser esenciales en ciertas condiciones, como en un medio de cultivo mínimo donde no hay rutas metabólicas alternativas, pero prescindibles en entornos ricos en nutrientes o en organismos con mecanismos compensatorios.

La esencialidad de las proteínas puede detectarse mediante ensayos experimentales de mutagénesis a gran escala, en los cuales se introducen múltiples inserciones o mutaciones aleatorias en los genomas de una población de organismos. Posteriormente, se secuencian los genomas de aquellos organismos que lograron sobrevivir. Los genes en los que no se detectan mutaciones en la población sobreviviente se consideran esenciales, ya que la interrupción de su función habría impedido la viabilidad celular (DeJesus et al. 2017).

En los casos donde no existen ensayos experimentales, es posible realizar predicciones sobre si una proteína es esencial al realizar comparaciones de identidad con proteínas reportadas como esenciales, las cuales se encuentran disponibles en bases de datos como Database of Essential Genes (DEG). Si una proteína presenta un homólogo esencial en otro microorganismo, existe una alta probabilidad de que también sea esencial en el organismo de interés, dado que los genes esenciales tienden a estar fuertemente conservados a lo largo de la evolución debido a su rol crítico en funciones celulares fundamentales.

Drogabilidad:

La drogabilidad es un concepto que describe la capacidad de una proteína para unirse a un compuesto con características de fármaco, modulando su función de manera deseada. Para que una proteína sea considerada drogable, debe poseer un bolsillo (cavidad en la estructura tridimensional de una proteína donde pueden unirse moléculas como ligandos o sustratos) bien definido con atributos fisicoquímicos adecuados que permitan predecir sitios de unión para compuestos. En este trabajo, la drogabilidad estructural se obtuvo de la plataforma Target-Pathogen (Sosa et al. 2018), que fue previamente calculada con el programa fpocket (Le Guilloux, Schmidtke, and Tuffery 2009) y el índice DrugScore (DS), una función de puntuación estadística que evalúa afinidad ligando-proteína mediante potenciales atómicos clasificando los bolsillos en categorías según su capacidad para unirse a compuestos (Velec, Gohlke, and Klebe 2005). Las proteínas con altos niveles de drogabilidad son las más susceptibles de ser moduladas por compuestos terapéuticos, lo que las convierte en posibles blancos para el desarrollo de fármacos.

Puntos de estrangulamiento:

Una red metabólica está compuesta por el conjunto de reacciones bioquímicas que ocurren en una célula, donde cada reacción convierte uno o más sustratos en productos, y es generalmente catalizada por una o más proteínas enzimáticas. Estas reacciones están interconectadas a través de los metabolitos que comparten, y juntas conforman el entramado funcional del metabolismo celular.

En este contexto, los puntos de estrangulamiento ("*choke-points*") son reacciones en una red metabólica que actúan como cuellos de botella, siendo las únicas capaces de consumir o producir un determinado metabolito. Se identifican a partir de una red metabólica curada, observando que un determinado sustrato o producto es consumido o producido por una única reacción química en el proceso de biosíntesis del metabolito. Su atractivo como blancos terapéuticos radica en que su inhibición puede provocar la acumulación de sustratos potencialmente tóxicos o la depleción de metabolitos esenciales, generando así un impacto significativo en la viabilidad celular del organismo objetivo. (Rahman and Schomburg 2006).

Centralidad de reacciones:

Una **red metabólica** puede representarse como un **grafo**, en el cual los **nodos corresponden a reacciones bioquímicas** y las **aristas** indican relaciones de dependencia entre ellas. Dos reacciones están conectadas si **comparten al menos un metabolito**, es

decir, si el producto de una reacción actúa como sustrato en otra. Esta representación permite modelar el **flujo metabólico** y analizar propiedades topológicas de la red, tales como la **conectividad**, la **centralidad** y la **modularidad**, proporcionando información clave sobre la organización y funcionamiento del metabolismo celular.

La centralidad de una reacción es un parámetro global relacionado con su importancia dentro de la red metabólica, en función de su rol como nodo de conexión entre múltiples trayectorias metabólicas. Una reacción representada por un nodo con alta centralidad participa como intermediaria de varios flujos metabólicos a la vez. Estas reacciones se pueden evaluar mediante medidas como la intermediación (betweenness), precalculada en Target Pathogen y utilizada en este trabajo (da Silva, Ma, and Zeng, 2008). Otras métricas habituales—grado, cercanía, eigenvector o PageRank—también se suelen utilizar en el contexto de redes metabólicas. Por otra parte, en redes medianas la centralidad se calcula con heurísticas dependientes del nodo sobre el cual inicia el cálculo para reducir su elevado coste computacional.

La **centralidad de intermediación** de un nodo v , denotada como $C_B(v)$, mide la cantidad de caminos más cortos entre pares de nodos que pasan por v . Se calcula con la fórmula detallada en la ecuación 2.2:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \quad (2.2)$$

en donde:

- σ_{st} es el número total de caminos más cortos entre los nodos s y t ,
- $\sigma_{st}(v)$ es el número de esos caminos que pasan por v ,
- la sumatoria $\sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$ se realiza sobre todos los pares de nodos s y t distintos de v .

Esta métrica identifica nodos que actúan como puentes funcionales o cuellos de botella, facilitando el flujo entre diferentes partes de la red.

La esencialidad, los puntos de estrangulamiento y la centralidad no son conceptos aislados sino propiedades complementarias que permiten identificar reacciones dentro de una red metabólica. Estas propiedades están interrelacionadas a través de los mecanismos de síntesis y degradación de metabolitos, y su análisis conjunto permite detectar vulnerabilidades funcionales en el metabolismo. En ese sentido, proteínas esenciales

pueden participar en reacciones de punto de estrangulamiento o con alta centralidad, lo que refuerza su relevancia biológica. Comprender estas interconexiones proporciona una visión más completa del papel que las proteínas desempeñan en el funcionamiento global del metabolismo. En el contexto de este trabajo, se analizaron estos parámetros para identificar proteínas que podrían ser blancos efectivos de fármacos, cuya inhibición afecte la supervivencia del patógeno.

2.2.6 Búsqueda de sustratos y productos de proteínas por contexto genético

En el contexto de genomas bacterianos, la proximidad física de genes adyacentes en varios casos muestra indicios importantes sobre relaciones entre sus funciones en la coordinación de procesos biológicos (Overbeek et al. 1999). Un fenómeno destacado es la presencia de operones, donde varios genes con funciones relacionadas están agrupados y son transcritos como una sola unidad funcional. Los operones bacterianos permiten una regulación eficiente y coordinada de la expresión génica, ya que comparten elementos regulatorios comunes. Estos grupos de genes, a menudo involucrados en vías metabólicas específicas, en muchos casos están involucrados en la adaptación rápida de las bacterias a su entorno, permitiendo respuestas coordinadas a cambios en las condiciones ambientales. Además, la conservación de la proximidad genética en genomas bacterianos relacionados sugiere la persistencia de funciones biológicas cruciales a lo largo de la evolución. Esta organización genómica ofrece una perspectiva valiosa en la anotación funcional y en la comprensión de las estrategias adaptativas de las bacterias en diversos entornos (Osborn and Field 2009).

Debido a estas relaciones de funcionalidad que existen entre genes próximos en el contexto genético, en muchos casos, por ejemplo en la vía metabólica de un compuesto químico determinado, hay alta similitud entre los sustratos y/o productos de las proteínas codificadas por genes cercanos. Este fenómeno recurrente en la naturaleza nos permite realizar predicciones de posibles sustratos y productos de una proteína catalizadora, basándose en su contexto genético.

2.3 Modelado de proteínas

El modelado estructural de proteínas ha evolucionado significativamente en los últimos años debido a la creciente disparidad entre el número de secuencias de proteínas disponibles y la cantidad de estructuras experimentales resueltas. En este contexto, el modelado por homología y herramientas avanzadas como AlphaFold se han consolidado como estrategias fundamentales para predecir estructuras tridimensionales a partir de secuencias de proteínas. El modelado por homología, que utiliza proteínas con estructuras conocidas como moldes, constituye una técnica históricamente muy empleada para obtener modelos estructurales confiables. Sin embargo, la llegada de AlphaFold significó una revolución en este campo, al emplear técnicas de aprendizaje profundo que permiten predecir estructuras con una confiabilidad sin precedentes, incluso en casos donde no existen proteínas conocidas con una homología evidente. En esta sección, se detallan ambas metodologías, sus principios fundamentales y las aplicaciones que permiten avanzar en la comprensión de la estructura y función de las proteínas.

2.3.1 Modelado por homología

La creciente disparidad entre la cantidad de secuencias de proteínas disponibles y las estructuras experimentales resueltas ha generado la necesidad imperante de aplicar métodos de modelado estructural, siendo el modelado por homología una estrategia destacada y ampliamente utilizada en los últimos años. Este proceso, también conocido como modelado comparativo, se fundamenta en el uso de proteínas con estructura conocida y secuencia similar como moldes para predecir la estructura tridimensional de proteínas relacionadas cuya estructura es desconocida. En la actualidad, el modelado por homología representa una herramienta eficiente para obtener modelos con una muy alta confianza de la estructura de una proteína a partir de su secuencia, pero es dependiente de que exista una proteína homóloga con estructura resuelta.

La base subyacente del modelado por homología consiste en la observación de que proteínas con estructura primaria estrechamente relacionada comparten una conformación espacial similar. Se postula que, si las secuencias aminoacídicas están evolutivamente emparentadas, la estructura tridimensional de una proteína puede inferirse a partir de las estructuras 3D de otras con secuencias evolutivamente cercanas. ("Methods for Protein Homology Modelling" 2006).

El proceso de modelado por homología consta de varias etapas clave. Como punto de partida, se requiere de la selección precisa de la estructura molde más adecuada. La calidad del modelado depende significativamente de esta elección, siendo una regla

general que una identidad de secuencia superior al 50% entre la proteína a modelar y el molde proporciona modelos más confiables.

La etapa crítica del alineamiento de las secuencias de la proteína a modelar y el molde establece las bases para la construcción del modelo. En esta fase, se busca minimizar la presencia de "gaps" fuera de las regiones de estructura secundaria del molde, optimizando así la correspondencia entre las dos secuencias.

La generación del modelo estructural se basa en el alineamiento de secuencias, y existen diversas aproximaciones, entre las que se destaca el ensamblaje de cuerpo rígido, el modelado por coincidencia de segmentos y el modelado por satisfacción de restricciones espaciales. Este último, implementado en programas como MODELLER ("Modeller: Generation and Refinement of Homology-Based Protein Structure Models" 2003), utiliza restricciones geométricas y de homología para generar modelos que minimizan las violaciones de estas restricciones, ofreciendo una perspectiva sólida en el modelado comparativo.

Además, el modelado por homología aborda desafíos específicos, como la predicción de conformaciones de loops, cruciales para la funcionalidad proteica. Se recurre a protocolos especializados, como el implementado por defecto en Modeller, para enfrentar este problema a menor escala, similar al desafío del plegado proteico.

En paralelo, la predicción de conformaciones de las cadenas laterales de aminoácidos se aborda considerando aspectos energéticos y estéricos, ampliando así la utilidad del modelado por homología en la caracterización detallada de estructuras proteicas.

El proceso general en el modelado por homología se ilustra en la Figura 2.7 A.

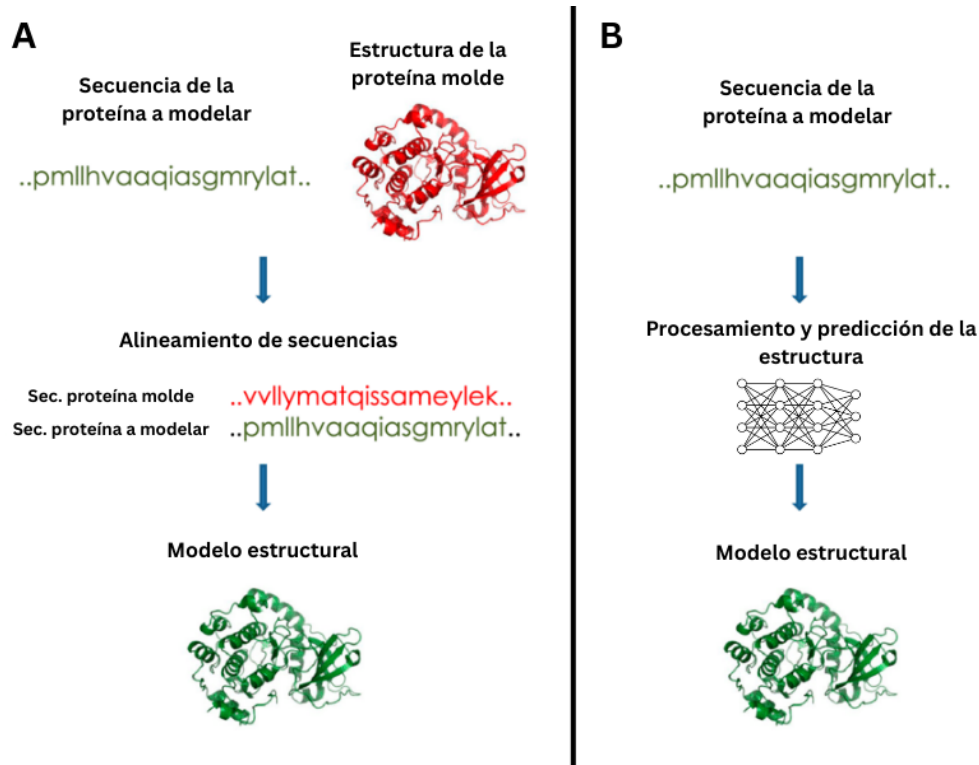


Figura 2.7: Comparativa representada de manera general de los procedimientos entre el modelado por homología (A), en el cual se requiere de una proteína molde, y el modelado por AlphaFold (B), enfoque que permite el modelado estructural de la proteína únicamente partiendo de la secuencia de aminoácidos. Adaptado de Sliwoski et al. 2014.

2.3.2 Modelado por Alphafold

AlphaFold (Jumper et al. 2021) constituye un avance disruptivo en la predicción de estructuras proteicas mediante el empleo de modelos de aprendizaje profundo entrenados sobre grandes volúmenes de datos estructurales disponibles públicamente. A diferencia de los enfoques clásicos de modelado por homología, no requiere una estructura molde, lo que permite modelar proteínas **sin necesidad de contar con secuencias homólogas de alta identidad de secuencia conocida**, ampliando así su aplicabilidad a familias menos caracterizadas (Figura 2.7 B).

Partiendo de la secuencia de aminoácidos de una proteína que se desea modelar, AlphaFold identifica secuencias homólogas en bases de datos extensas y construye un alineamiento múltiple. Esta información se combina con predicciones auxiliares, como la matriz de distancias D_{ij} y orientaciones relativas θ_{ij} entre residuos, que alimentan una red neuronal.

La red transforma estas entradas en una representación latente $\varphi(S) \in \mathbb{R}^{n \times d}$, donde cada vector captura información estructural y evolutiva de un residuo dado. A partir de esta codificación, el modelo predice un conjunto de coordenadas tridimensionales $\hat{X} = \{\hat{X}_1, \dots, \hat{X}_n\} \in \mathbb{R}^3$, que definen la estructura espacial de la proteína. El proceso se refina mediante una etapa de optimización geométrica aprendida, que mejora la conformación generada al minimizar una función de pérdida estructural basada en desviaciones de distancias y ángulos esperados.

Otro punto destacable en la practicidad de esta herramienta consiste en que para muchas de las proteínas con secuencia conocida y reportadas en bases de datos de gran escala como Uniprot, los modelos correspondientes se encuentran ya calculados y disponibles en la base de datos en línea de AlphaFold (Varadi et al. 2021). De esta manera es posible la obtención directa de modelos estructurales sencillamente sin la necesidad de capacidad de cómputo local.

2.3.3 Validación de los modelos

Métricas de evaluación de confiabilidad para modelos de proteínas.

En el modelado predictivo de estructura de proteínas, se desarrollaron diversos métodos para evaluar la confiabilidad de las estructuras generadas. Cada una de las herramientas de predicción estructural suele tener asociada una métrica determinada para poder definir la calidad de una predicción.

En el caso del software Modeller, utilizado para realizar modelado por homología, la métrica utilizada para este fin es DOPE (*Discrete Optimized Protein Energy*). DOPE es un potencial estadístico utilizado para evaluar la calidad de modelos tridimensionales de proteínas. Fue desarrollado para estimar qué tan probable es que una estructura predicha sea realista, en base a lo observado en proteínas con estructuras determinadas experimentalmente.

El método calcula una energía para el modelo teniendo en cuenta las distancias entre pares de átomos, comparando esas distancias con las que se encuentran en proteínas con estructura conocida. Para hacerlo, alrededor de cada átomo del modelo se trazan esferas concéntricas a diferentes distancias; se mide cuán frecuentemente aparecen ciertos tipos de átomos en cada rango de distancia, y se compara con lo que se espera ver en proteínas con estructuras obtenidas experimentalmente.

Si las posiciones relativas de los átomos en el modelo son coherentes con las distribuciones conocidas, el valor de energía será bajo, lo cual indica un modelo coherente.

DOPE es especialmente útil para comparar varios modelos de una misma proteína y seleccionar el más confiable entre ellos.

Para calcular DOPE, se agrupan los átomos en categorías llamadas tipos atómicos, que además del elemento químico (por ejemplo, carbono, oxígeno, nitrógeno), consideran su función estructural. El puntaje estadístico para cada par de átomos de tipos atómicos m y n separados por una distancia r se calcula según la ecuación 2.3:

$$-u_{mn}(r) = \ln\left(\frac{N_{mn}^{OBS}(r)}{N_{mn}^{REF}(r)}\right) \quad (2.3)$$

donde N_{mn}^{OBS} representa la cantidad de veces que un par de átomos de tipos atómicos m y n se encuentra a una distancia r en estructuras de proteínas reales, y N_{mn}^{REF} es el número esperado de ocurrencias para esa misma distancia si los átomos estuvieran distribuidos aleatoriamente en el volumen de la proteína. De esta manera, el puntaje final del modelo E_{DOPE} se obtiene según la ecuación 2.4, mediante la suma de los puntajes estadísticos individuales para todos los pares de átomos:

$$E_{DOPE} = \sum_{i < j} u_{m_i n_j}(r_{ij}) \quad (2.4)$$

Por otra parte, en el caso de AlphaFold, la métrica utilizada por defecto para evaluar la confiabilidad local de los modelos es pLDDT (*Predicted Local Distance Difference Test*), una medida de la confianza en la predicción de cada residuo del modelo. Si bien este parámetro no tiene una fórmula para su cálculo, ya que se trata de una predicción efectuada con una red neuronal, su diseño se basa en la métrica IDDT (*Local Distance Difference Test*) (Mariani et al. 2013).

El IDDT es una métrica que permite evaluar la similitud local entre un modelo estructural y la estructura experimental correspondiente, que funciona de la siguiente manera: para cada residuo, se consideran las distancias a otros residuos ubicados dentro de un radio predefinido (por defecto se utiliza el valor de 15 Å), y se comparan dichas distancias entre ambas estructuras. Los residuos suelen estar representados por un átomo de referencia, como el carbono alfa.

Luego, se evalúa cuántas de esas distancias se encuentran dentro de un margen de error aceptable (se suelen considerar los valores de 0.5 Å, 1 Å, 2 Å y 4 Å, y luego hacer un promedio de todas las comparaciones), lo que se considera como una predicción correcta. El puntaje de IDDT para el residuo evaluado se calcula como la proporción de distancias correctamente predichas respecto del total de comparaciones realizadas. De esta manera, el IDDT proporciona una estimación cuantitativa de la precisión local del modelo estructural,

sin requerir la superposición de las estructuras. De manera general, el valor de IDDT para toda la proteína se calcula como el promedio de IDDT por residuo, según la ecuación 2.5:

$$IDDT = \frac{1}{4N} \sum_{\tau \in \{0.5, 1.0, 2.0, 4.0\}} \sum_{i=1}^N \frac{1}{n_i} \sum_{j=1}^{n_i} \delta_{ij}^{(\tau)} \quad (2.5)$$

En donde N es el número total de residuos considerados, n_i es el número de átomos vecinos del residuo i dentro del radio predeterminado, y $\delta_{ij}^{(\tau)}$ es la función indicadora, que toma el valor de 1 si la diferencia entre la distancia predicha y la experimental es menor que el margen de error aceptable τ , y es igual a 0 en el caso contrario.

En el caso de AlphaFold, el pLDDT representa una predicción del IDDT generada por un modelo de inteligencia artificial entrenado para obtener un puntaje sin la necesidad de una estructura experimental de referencia.

El puntaje se presenta con un valor de 0 a 100, que refleja qué tan precisa se espera que sea la predicción en cada posición de la proteína, con valores más altos indicando mayor confianza y, generalmente, una predicción más confiable. En términos de interpretación, un pLDDT superior a 90 suele indicar una predicción de alta precisión tanto para la estructura del esqueleto como para las cadenas laterales. Por otro lado, valores entre 70 y 90 suelen corresponder a una predicción correcta del esqueleto, aunque con cierta inexactitud en las cadenas laterales (Wilson, Choy, and Karttunen 2022).

Validación de los modelos estructurales con respecto a sus estructuras experimentales correspondientes.

La manera más apropiada de validar la efectividad de un algoritmo para generación de modelos de proteínas resulta de comparar la estructura modelada con la estructura cristalizada. El método utilizado en este trabajo para efectuar dicha comparación es la desviación cuadrática media de las coordenadas (RMSD), representada por la ecuación 2.6, en la cual a_i y b_i corresponden a las coordenadas de cada par de átomos comparados (en este caso los pertenecientes a la estructura real y la modelada, respectivamente) y N es el número total de átomos considerados en la comparación:

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N \|a_i - b_i\|^2} \quad (2.6)$$

Esta métrica evalúa la diferencia absoluta media entre las posiciones de átomos específicos en dos estructuras tridimensionales. Esto implica que cuanto menor sea el RMSD, mayor será la similitud estructural entre ambas estructuras. Para simplificar este

cálculo y detenerse en posiciones más informativas, en muchos casos se utiliza el RMSD de carbonos alfa (RMSD-C α) entre las dos estructuras ("Significance of Root-Mean-Square Deviation in Comparing Three-Dimensional Structures of Globular Proteins" 1994) .

En el contexto del modelado por homología, la comparación entre una estructura predicha y una estructura de referencia se realiza de manera directa. Se puede seleccionar una proteína de referencia y generar un modelo utilizando otra proteína que comparte un considerable grado de identidad como molde. Luego, se calcula el RMSD entre la estructura modelada y la estructura real, sin introducir sesgos.

Sin embargo, en el caso de modelos generados por AlphaFold, la comparación se vuelve más compleja. Al modelar una proteína que ya ha sido cristalizada, existe la posibilidad de que ese cristal haya sido utilizado para generar el conjunto de datos utilizado para entrenar la red neuronal subyacente. Esto introduce un sesgo en la validación, ya que AlphaFold fue entrenado utilizando conocimiento previo sobre la proteína que se está intentando modelar. En el contexto de este trabajo, se aborda este problema utilizando como punto de evaluación para los modelos estructurales a proteínas que fueron cristalizadas con posterioridad a la fecha oficial de lanzamiento de las estructuras de PDB utilizadas para entrenar el modelo de AlphaFold.

2.4 *Docking* molecular

El *docking* molecular es una técnica utilizada para estudiar las interacciones entre proteínas y ligandos, permitiendo predecir configuraciones tridimensionales de estos complejos y estimar su estabilidad. Este enfoque se emplea ampliamente en bioinformática estructural y diseño de fármacos, ya que ayuda a identificar interacciones moleculares potencialmente relevantes. En esta sección se presentan tanto los algoritmos tradicionales, basados en enfoques geométricos y simulaciones como el *docking* molecular sesgado, que incorpora información biológica adicional para refinar las predicciones.

2.4.1 *Docking* molecular: descripción del algoritmo

Las estrategias de *docking* constituyen un componente esencial en la investigación estructural de proteínas, ya que buscan predecir la configuración tridimensional y estabilidad de los complejos proteína-ligando. Dos categorías principales de métodos de *docking* han surgido: aquellos que emplean coincidencia geométrica y aquellos que se basan en simulaciones detalladas de la formación del complejo.

En el primer enfoque, se crea un modelo del sitio activo de la proteína, incorporando áreas críticas para la interacción, como los sitios de puente de hidrógeno. Luego, se intenta colocar el ligando en una conformación específica dentro de este modelo. Este método, representado por programas como DOCK (Allen et al. 2015), se caracteriza por su eficiencia en el escaneo rápido de bibliotecas de compuestos.

En contraste, la segunda categoría de métodos de *docking* se sumerge en la descripción detallada de las estructuras del ligando y la proteína con la que interactúa, por convención llamada receptor en el contexto de los algoritmos de *docking*. Para evaluar el enorme espacio conformacional evitando un coste computacional prohibitivo, estos métodos recurren a estrategias heurísticas (no exhaustivas) que permiten explorar de forma eficiente las traslaciones, orientaciones y conformaciones del ligando en interacción con el receptor. Un ejemplo destacado de esta categoría es Autodock (Morris et al. 2009), herramienta de acceso libre bien documentada y utilizada en los protocolos de *docking* molecular por nuestro Grupo de Investigación.

Los algoritmos de *docking* mayormente utilizados consisten de dos elementos clave: un algoritmo de búsqueda heurística y una función de puntuación. Los algoritmos genéticos resultan uno de los algoritmos de búsqueda más relevantes. Inspirados en conceptos biológicos, generan una población inicial de ligandos con diversas conformaciones. Estos ligandos experimentan procesos de “entrecruzamiento” y “mutación”, y la selección se realiza en función de la energía de interacción. La búsqueda local implica la minimización de energía libre de interacción entre ligando y receptor.

La función de puntuación, crucial en el proceso de *docking*, es responsable de comparar y clasificar rápidamente las diversas conformaciones propuestas por el algoritmo. En Autodock, la función de puntuación, estima la energía libre de unión (ΔG) proteína-ligando, la cual se compone de diferentes términos, que se detallan de manera simplificada en la ecuación 2.7:

$$\Delta G_{total} = \Delta G_{vdW} + \Delta G_{hb} + \Delta G_{elec} + \Delta G_{desolv} + \Delta G_{tor} \quad (2.7)$$

en donde ΔG_{vdW} corresponde a la energía de van der Waals, ΔG_{hb} a la energía de enlaces de puente de hidrógeno, ΔG_{elec} es la energía electrostática, ΔG_{desolv} corresponde a la energía de desolvatación y ΔG_{tor} a la energía de torsión (Forli and Olson 2012).

Para facilitar la aplicación práctica, AutoDock utiliza grillas precalculadas con la herramienta AutoGrid para evaluar eficientemente la energía (átomo por átomo) de diferentes conformaciones, lo que contribuye significativamente a la eficiencia computacional del proceso de *docking*.

Por otra parte, el proceso de agrupamiento o población de clusters en AutoDock4 permite organizar las múltiples conformaciones generadas durante el *docking* en clústers de conformaciones similares, basándose en el desvío cuadrático medio (RMSD) entre ellas. Las conformaciones que muestran una estructura espacial similar se agrupan en un mismo clúster. Esto es especialmente importante porque la población de clusters proporciona una visión estadística de las conformaciones más frecuentes, que podrían ser las más relevantes biológicamente. Un cluster con una alta población indica qué poses pertenecientes a este conjunto de conformaciones similares fueron obtenidas con frecuencia, lo que sugiere que el ligando tiene una alta probabilidad de adoptar la orientación correspondiente en el sitio de unión con la proteína. Los clusters con mayor población y menor energía libre de unión son usualmente considerados las predicciones más confiables.

Por otro lado, en los últimos tiempos destaca el desarrollo de AutoDockGPU (Santos-Martins et al. 2021), una versión acelerada de AutoDock4 basada en la tecnología OpenCL, que mejora significativamente el rendimiento al aprovechar la arquitectura paralela de los procesadores gráficos (GPUs). Esta implementación permite reducir el tiempo de ejecución hasta 350 veces en comparación con procesos de un solo hilo, lo que resulta esencial para estudios a gran escala de *screening* virtual. La aceleración se logra mediante la distribución eficiente de las tareas de búsqueda global y local, y la integración de métodos avanzados de optimización como ADADELTA, optimizando así la precisión y eficacia de las simulaciones de *docking* molecular.

El proceso general de un ensayo de *docking* consiste en los siguientes pasos (Prieto et al. 2024), detallados en la Figura 2.8:

- Obtención de la estructura de la proteína a evaluar, llamada receptor en los protocolos de *docking*, la cual idealmente fue determinada experimentalmente. En el caso contrario es posible utilizar un modelo estructural.
- Obtención de la estructura tridimensional del ligando a evaluar, que puede ser descargada de diversas bases de datos (como PDB), o también puede ser generada desde el formato SMILES de la molécula con herramientas específicas.
- Procesamiento del receptor: se realizan diversas modificaciones para que la estructura sea apta para realizar el *docking*: 1) se eliminan moléculas pequeñas, como agua y compuestos químicos, 2) se agregan hidrógenos enlazados a átomos polares, importantes en las interacciones de puente de hidrógeno, 3) se asignan cargas parciales a cada uno de los átomos, 4) se definen los tipos de átomos, según si son donores o aceptores de puentes de hidrógeno, o carbonos alifáticos o aromáticos.
- Procesamiento del ligando: protocolo análogo al procesamiento del receptor, con la diferencia de que adicionalmente se seleccionan los enlaces rotables de la molécula. Es decir que en el proceso de *docking* se define que parte de la molécula puede ser rotada y que parte queda fija.
- Selección del sitio activo del receptor y armado de grilla: se define la zona en la cual va a ocurrir el ensayo de *docking*, determinada por una grilla cúbica. Dentro de esta grilla se precalcula la energía de interacción de cada tipo de átomo del ligando con la proteína.
- Definición de parámetros y corridas de *docking*: en este paso ocurre el proceso de docking propiamente dicho, en donde se exploran diversas conformaciones del ligando en interacción con el receptor hasta alcanzar las de mayor estabilidad según la función de puntuación.

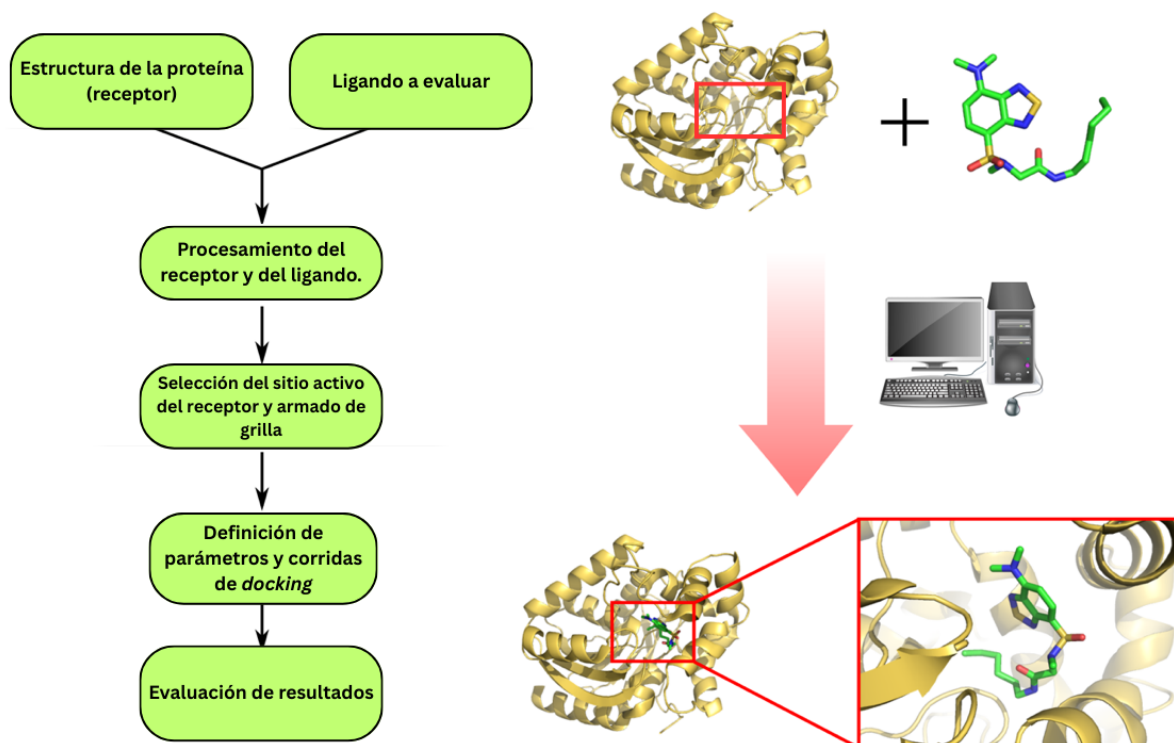


Figura 2.8: Esquema de la metodología de uso del *docking* molecular para la predicción de interacciones de proteína y ligando. Adaptado de Profacgen, <https://www.profacgen.com/protein-ligand-docking.html>.

2.4.2 Docking molecular sesgado (*biased docking*)

AutoDock Bias es un método basado en scripts adicionales al programa base, que posibilita la introducción de diversos tipos de sesgos para afinar las simulaciones de AutoDock, que fue desarrollado por nuestro Grupo de Investigación.

Los sesgos se incorporan como recompensas energéticas para cada átomo del ligando que participa en interacciones moleculares relevantes previamente identificadas entre la proteína y el ligando, generalmente, enlaces de tipo puente hidrógeno o interacciones hidrofóbicas (Figura 2.9).

El cálculo de la energía libre de unión modificada para un punto determinado de la grilla determinado por las coordenadas (x, y, z) , en el que se aplica el *docking* sesgado se detalla en la ecuación 2.8:

$$V_{bias} = V_{ori} + V_{set} \exp\{- [(x - x_i)^2 + (y - y_i)^2 + (z - z_i)^2]/r_i^2\} \quad (2.8)$$

donde V_{bias} corresponde a la energía libre de unión modificada. V_{ori} es la energía calculada por defecto por AutoDock. V_{set} representa la profundidad del pozo generado en la energía libre de unión. r_i es el radio respecto al centro del sitio de sesgo y (x_i, y_i, z_i) son las coordenadas del centro del sitio de sesgo (Arcon et al. 2019).

Este enfoque proporciona una herramienta flexible para mejorar la precisión y relevancia biológica de los resultados de *docking*. Este procesamiento adicional, que permite incorporar información conocida a los sistemas en evaluación, permite una mayor personalización de las simulaciones, ofreciendo también mejoras en su eficacia para ciertos casos.

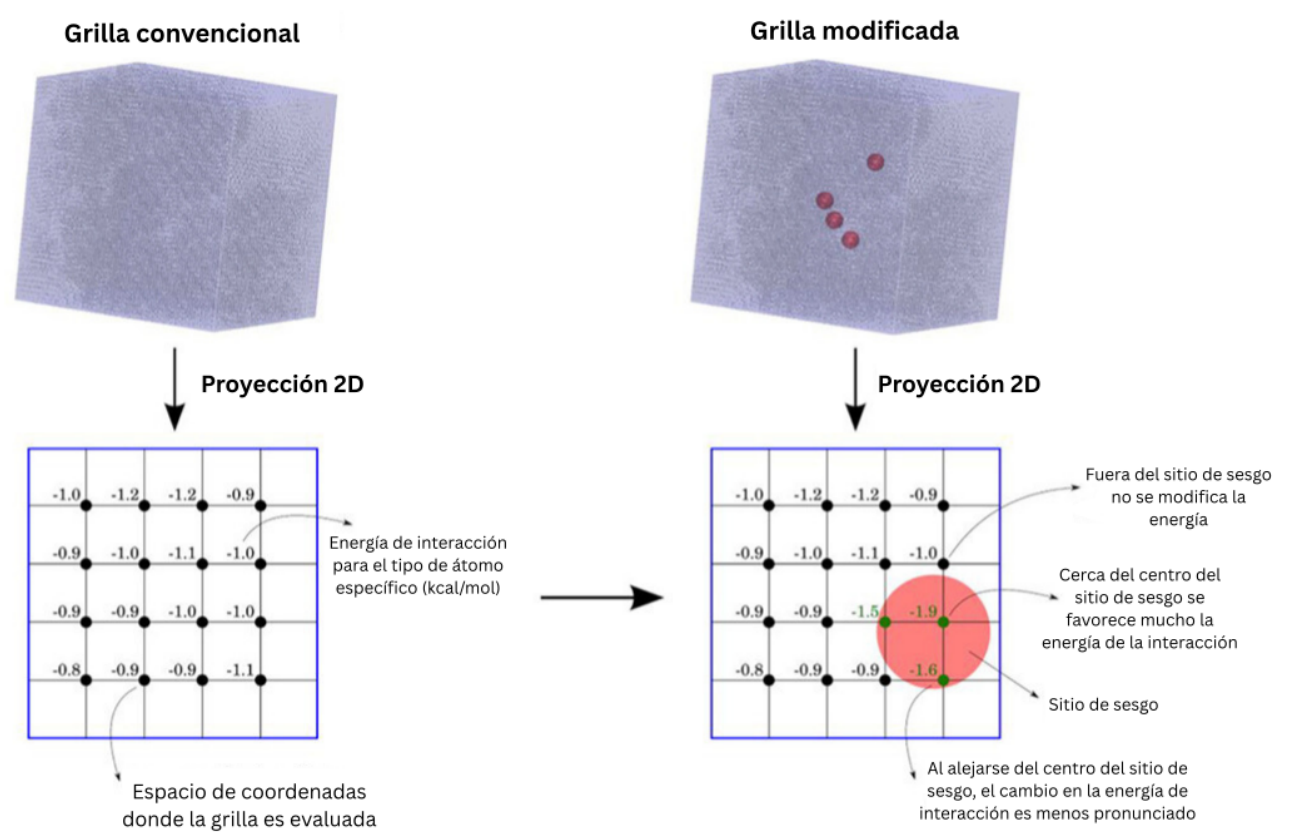


Figura 2.9: Esquema detallado de la implementación de sesgos en el proceso de *docking* molecular. Adaptado de Arcon et al. 2021.

2.4.3 Z-score

El z-score es una medida estadística que indica cuántas desviaciones estándar está un valor por encima o por debajo de la media de un conjunto de datos. Se utiliza

comúnmente para estandarizar diferentes distribuciones y facilitar la comparación entre ellas. Para el cálculo del Z-score de un valor determinado x , se le resta la media de la distribución (μ) y luego se divide el resultado obtenido por la desviación estándar (σ), tal como se detalla en la ecuación 2.9:

$$Z = \frac{x - \mu}{\sigma} \quad (2.9)$$

El Z-score permite identificar qué tan inusual o significativo es un valor dentro de su contexto, lo que lo convierte en una herramienta útil para la detección de valores atípicos o la evaluación de resultados en relación con una distribución de referencia (S. S. Kim, Aprahamian, and Lindert 2019).

En el caso del *docking* molecular realizado con Autodock4, el Z-score se aplica tanto a la energía libre predicha como también a la población de cluster, para cada una de las poses obtenidas. Finalmente se calcula un Z-score global sumando ambos términos. De esta manera es posible realizar un ranking de las poses más favorables, lo cual es muy útil en caso de evaluar la capacidad de unión de diferentes ligandos en una misma proteína.

Para las visualizaciones, se crean gráficos donde el eje X corresponde al Z-score de la energía libre predicha para una pose particular del ligando en interacción con la proteína, y el eje Y a la población de clusters para cada pose. De esta forma, los resultados más favorables se ubican en el cuadrante superior izquierdo del gráfico, lo cual indica una menor energía libre y una mayor población de clusters. Esta representación se ejemplifica en la Figura 2.10.

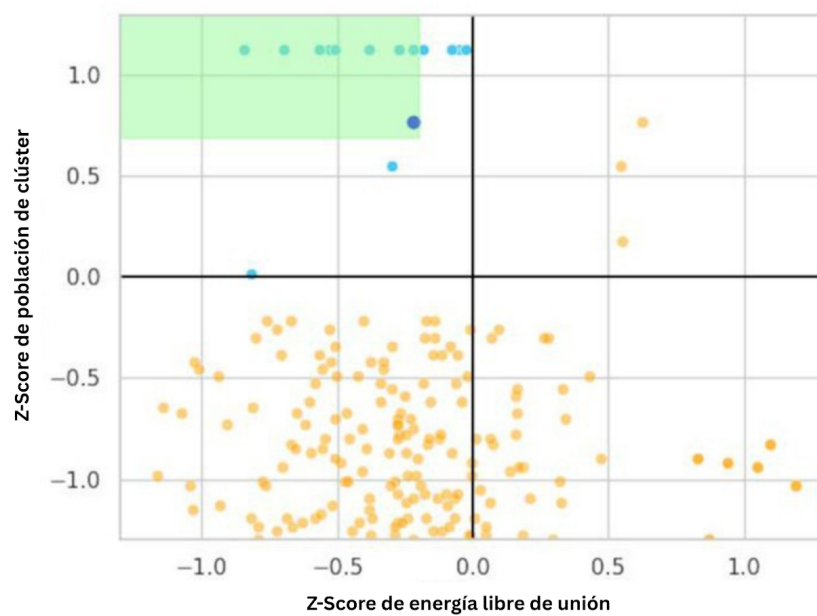


Figura 2.10: Ejemplo del gráfico de Z-score, en el cual se denotan los resultados de *docking* en diferentes cuadrantes. Como se explicó en la metodología, en el cuadrante superior izquierdo se encuentran los resultados de *docking* con mayor población de cluster y menor energía libre de unión predicha. Adaptado de Schottlender et al. 2024.

2.5 Métodos de Aprendizaje automático

El aprendizaje automático ha transformado el análisis de datos en una amplia variedad de campos, y su aplicación en la bioinformática y en la quimioinformática no son excepciones. En esta sección, se detallan técnicas clave como la búsqueda de vecinos más cercanos y el *clustering* de compuestos químicos. Estas metodologías permiten identificar patrones y relaciones dentro de grandes conjuntos de datos moleculares, facilitando el descubrimiento de compuestos con actividades biológicas de interés. El uso de algoritmos como la búsqueda de vecinos y el método de *clustering* Butina resultan herramientas poderosas para la organización y clasificación de ligandos, lo que a su vez contribuye al diseño eficiente de fármacos y la caracterización de nuevos compuestos en el ámbito farmacéutico.

2.5.1 Búsqueda de vecinos cercanos

La búsqueda de vecinos cercanos (NNS, por sus siglas en inglés) es un concepto fundamental en ciencia de datos y aprendizaje automático. Este enfoque consiste en identificar, dentro de un conjunto de datos, los puntos más próximos a un punto de referencia determinado (Figura 2.11). Esto permite analizar patrones locales y relaciones presentes en los datos. La diversidad de algoritmos de NNS se adapta a diversas aplicaciones, desde tareas de clasificación de imágenes, hasta la recomendación de contenido. Este proceso es fundamental para aprovechar la estructura inherente de los datos, lo cual contribuye a la toma de decisiones fundamentadas en diversos campos (Bhatia and Vandana 2010).

Una de las técnicas más conocidas derivadas de la búsqueda de vecinos más cercanos (NNS) es el algoritmo de clasificación por K-Vecinos más Cercanos (KNN). Este es un método de aprendizaje supervisado utilizado para clasificar un punto dado dentro de un espacio n-dimensional, definido por un conjunto de características (*features*). La clasificación se realiza en función de las clases predominantes entre los K vecinos más próximos al punto en cuestión dentro de dicho espacio multidimensional, utilizando una métrica de distancia, que transforma la diferencia entre los conjuntos de características en un número: cuanto menor es ese número, más cercanos son los puntos. En el caso de este trabajo, se utilizó una técnica operativamente muy parecida, llamada "*Fixed-Radius Near Neighbors*" (FNN), o Vecinos Cercanos con Radio Fijo ("*Fixed-Radius near Neighbors Search*" 1991).

El método de Vecinos dentro de un Radio Fijo (FNN, por sus siglas en inglés) es un enfoque fundamental dentro de la búsqueda de vecinos cercanos (NNS). Consiste en identificar, para cada punto de referencia en un conjunto de datos, aquellos puntos que se

encuentran dentro de una distancia predefinida o radio establecido. Esta técnica permite obtener información determinante desde la estructura local de los datos, resultando valiosa en diversas disciplinas.

En el algoritmo de FNN, se asigna un radio fijo a un punto (sobre el cual se desea realizar una predicción) dentro de un espacio n-dimensional de características. Luego, se buscan todos los puntos de la base de datos, que caen dentro de ese radio, considerándolos como vecinos cercanos. La información sobre estos vecinos se almacena para análisis posterior.

En este trabajo, un algoritmo de esta clase se utilizó para la búsqueda de compuestos químicos similares a un ligando “*query*” dentro de una base de datos.

Las características correspondientes en este caso son los *Fingerprints* de Morgan (explicados en la sección 2.1.1) que representan en forma de vector a cada una de las moléculas. Los vecinos cercanos están determinados por el Índice de Tanimoto, que actúa como métrica de distancia entre el compuesto en estudio y las moléculas que se encuentran en la base de datos.

El objetivo en nuestra aplicación de este enfoque no corresponde a un problema clásico de clasificación o regresión, sino a obtener el conjunto de las moléculas más parecidas al compuesto en estudio (representado por *Fingerprints* moleculares), dentro del mencionado radio fijo.

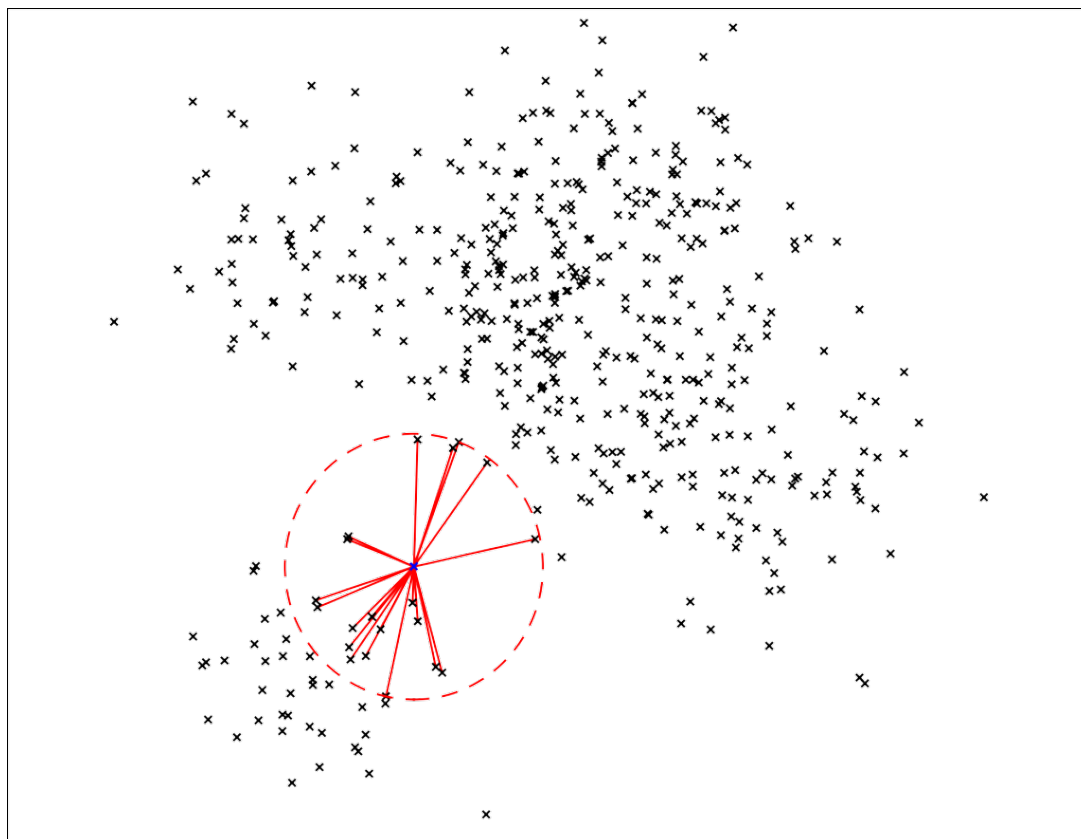


Figura 2.11: Ejemplo de la búsqueda de vecinos cercanos dado un punto del espacio en dos dimensiones. Tomado de "Nearest Neighbor Methods & Vector Models (Part 1)," por Erik Bernhardsson, publicado el 24 de septiembre de 2015, <https://erikbern.com/2015/09/24/nearest-neighbor-methods-vector-models-part-1.html>.

2.5.2 *Clustering* de compuestos similares

La investigación y la búsqueda de nuevos compuestos químicos, especialmente en el ámbito de la química aplicada al descubrimiento de drogas y compuestos con propiedades de interés comercial, requieren herramientas avanzadas para organizar y comprender la complejidad de grandes conjuntos de ligandos. En este contexto, el *clustering* de ligandos resulta ser una estrategia ampliamente utilizada en química computacional. Este enfoque busca agrupar moléculas similares entre sí, proporcionando una perspectiva estructural y funcional que facilita la toma de decisiones en la búsqueda y caracterización de nuevos fármacos (Böcker†‡ et al. 2005).

El *clustering* de ligandos se basa en la premisa fundamental de que compuestos químicos estructuralmente similares tienden a compartir propiedades biológicas y métodos de interacción con los blancos moleculares. Esta relación estructura-actividad es esencial para el descubrimiento de fármacos, ya que compuestos con perfiles bioquímicos similares

pueden exhibir respuestas biológicas comparables (Geldenhuys et al. 2005; Friedman*† et al. 2006).

Existen diversas metodologías de *clustering*, cada una con sus propias ventajas y enfoques particulares. Desde métodos jerárquicos hasta enfoques basados en centroides, el *clustering* de ligandos ofrece diversas opciones que se pueden adaptar a la gran diversidad existente de datos moleculares (Lipkowitz 2002).

Uno de los casos más representativos, y el utilizado en este trabajo es el método de *clustering* Butina (Butina 1999).

En el ámbito de la química computacional, el *clustering* Butina resulta una técnica fundamental para la organización eficiente de conjuntos de compuestos químicos en clústers homogéneos. Este enfoque, basado en la similitud química, facilita la identificación de patrones y relaciones dentro de grandes conjuntos de datos moleculares.

El algoritmo de *clustering* Butina se centra en la construcción de clústers de moléculas que comparten similitudes estructurales significativas. Utiliza métricas de distancia molecular previamente diseñadas y probadas, como principalmente el Índice de Tanimoto, para evaluar la similitud entre pares de moléculas.

El método Butina consta de tres pasos principales (Figura 2.12):

1. Generación de *Fingerprints*: en primer lugar, con las metodologías previamente explicadas, se generan *Fingerprints* moleculares para cada molécula. Estos consisten en vectores binarios de longitud fija que codifican la presencia o ausencia de subestructuras moleculares en un mismo compuesto.
2. Identificación de potenciales centroides de clúster: se asume que la molécula con más vecinos dentro de un cierto umbral de similitud de Índice de Tanimoto es un buen candidato a centroide de clúster. Para cada molécula, se calcula el número de vecinos por encima del umbral de similitud de Tanimoto elegido. Las moléculas se ordenan de mayor a menor, según el número total de vecinos, quedando los potenciales centroides al principio del ranking.
3. Agrupamiento basado en esferas de exclusión: se parte con la molécula correspondiente al primer centroide según el orden. Las moléculas con similitud igual o superior al umbral elegido se convierten en miembros del clúster actual. Los compuestos identificados como miembros se marcan y excluyen de futuras comparaciones (no pueden ser centroides ni miembros de otros clústers). Esto se visualiza como una "esfera de exclusión" alrededor del clúster formado.

El proceso se repite con la siguiente molécula centroide (no excluida por pertenencia a algún otro clúster) de la lista ordenada. A medida que se van asignando moléculas a clústers, se requieren menos comparaciones, y el proceso se agiliza.

Las moléculas que no son asignadas a ningún clúster al final del proceso se convierten en *singletons*.

Al agrupar moléculas según su similitud química, este enfoque no solo simplifica la complejidad de grandes conjuntos de ligandos, sino que también se convierte en una herramienta muy útil en la toma de decisiones para el descubrimiento de nuevos fármacos. La premisa de que compuestos estructuralmente similares comparten propiedades biológicas es ampliamente aceptada y estudiada, y esta metodología ofrece una estructura eficiente para este propósito (Sheridan† 2002).

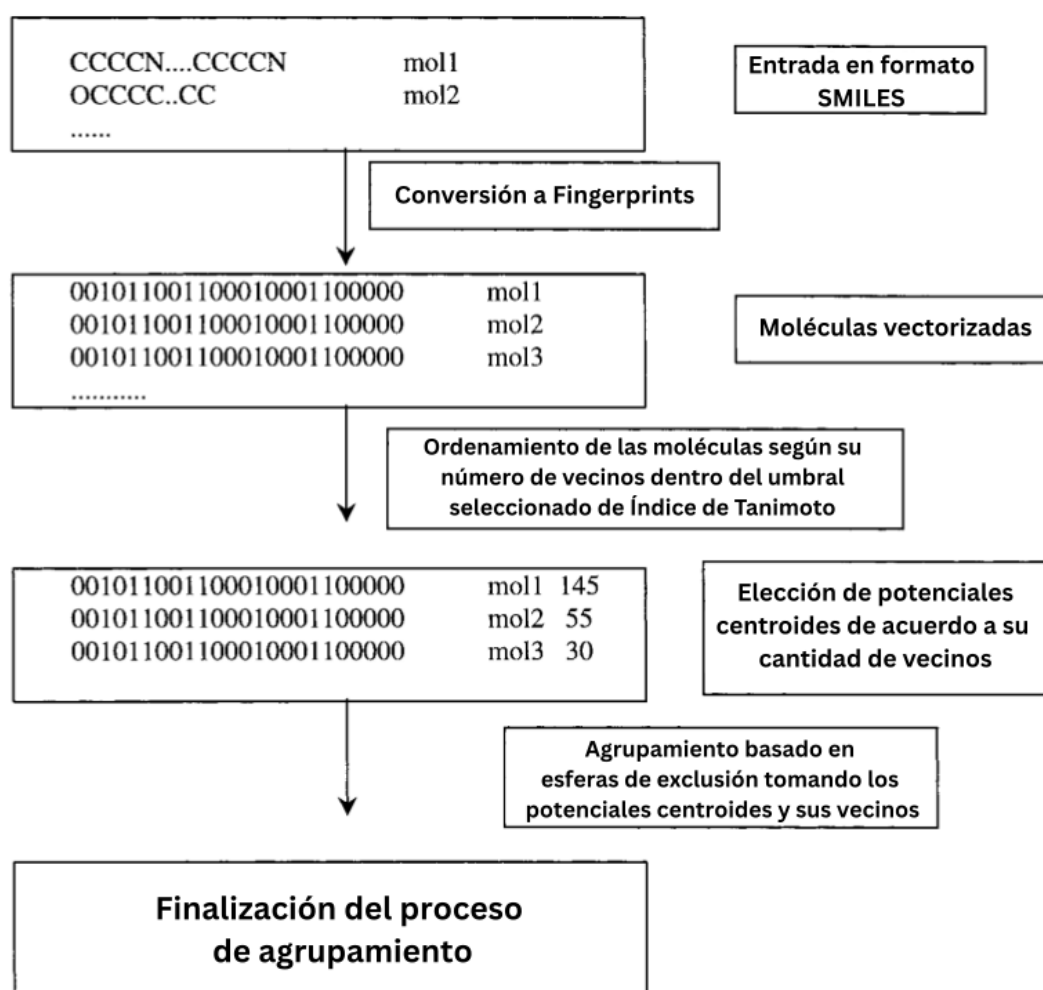


Figura 2.12: Esquema descriptivo del método de obtención de clusters de compuestos químicos con el enfoque de Butina. Adaptado de Butina 1999.

2.5.3 Aglomeración jerárquica

La aglomeración jerárquica es un método de aprendizaje automático utilizado para organizar datos en grupos de manera jerárquica, basándose en la similitud o distancia entre ellos (Murtagh and Contreras 2012). Este enfoque comienza considerando cada elemento como un grupo independiente y, a través de un proceso iterativo, fusiona los grupos más cercanos o similares en cada paso hasta que todos los elementos quedan reunidos en un solo conjunto. El resultado de este método se representa mediante un dendrograma, un diagrama en forma de árbol que visualiza las relaciones y similitudes entre los elementos agrupados.

La función *hclust* es una herramienta ampliamente utilizada para realizar este tipo de agrupamiento jerárquico, permitiendo elegir diferentes métodos de enlace, como el enlace simple, completo o promedio, que determinan cómo se mide la distancia entre los grupos a medida que se fusionan (Flynt and Dean 2016). Este tipo de análisis es fundamental para identificar estructuras o patrones en conjuntos de datos complejos, facilitando una mejor interpretación de las relaciones subyacentes.

En este contexto, se aplicó el método de aglomeración jerárquica principalmente para organizar secuencias de Citocromos P450, utilizando la distancia cofenética extraída mediante el paquete de R *ape* (Paradis and Schliep 2019). La distancia cofenética mide cuán similares deben ser dos objetos para ser agrupados en el mismo clúster, basada en la altura del dendrograma donde sus ramas se fusionan (Cardona et al. 2013). Este enfoque refleja las similitudes evolutivas capturadas en el dendrograma y proporciona una base sólida para identificar grupos filogenéticos de proteínas. De esta manera, se pueden inferir características comunes dentro de cada grupo, lo que contribuye a una comprensión más profunda de la diversidad y funcionalidad de los Citocromos P450 bacterianos.

Por otra parte, este método se utilizó para establecer relaciones significativas entre grupos de sustratos, productos e inhibidores de Citocromos P450 bacterianos, empleando el Índice de Tanimoto como métrica de distancia.

2.5.4 Validación cruzada y *leave-one-out*

La validación cruzada es una técnica ampliamente utilizada en el aprendizaje automático para evaluar la capacidad predictiva de un modelo y evitar problemas de sobreajuste. Consiste en dividir el conjunto de datos en múltiples subconjuntos (*folds*). El modelo se entrena en todos menos uno de los *folds*, y luego se evalúa su rendimiento en el *fold* que quedó fuera. Este proceso se repite iterativamente de manera que todos los *folds* actúen como conjunto de validación una vez, permitiendo obtener una evaluación general del desempeño del modelo utilizando todo el dataset. ("Cross-Validation Methods" 2000)

Un caso particular de validación cruzada es el enfoque conocido como *leave-one-out* (LOO). En este caso, cada entrada individual del conjunto de datos se trata como un *fold* separado, lo que significa que el modelo se entrena utilizando todas las entradas menos una, y luego se evalúa en la entrada excluida. Este procedimiento, ilustrado en la Figura 2.13, se repite tantas veces como entradas haya en el conjunto de datos. Si bien el *leave-one-out* proporciona una evaluación muy precisa del modelo, debido a que utiliza prácticamente todo el conjunto de datos en cada iteración, puede resultar computacionalmente costoso, especialmente en conjuntos de datos grandes. (“Performance Evaluation of Classification Algorithms by K-Fold and Leave-One-out Cross Validation” 2015)

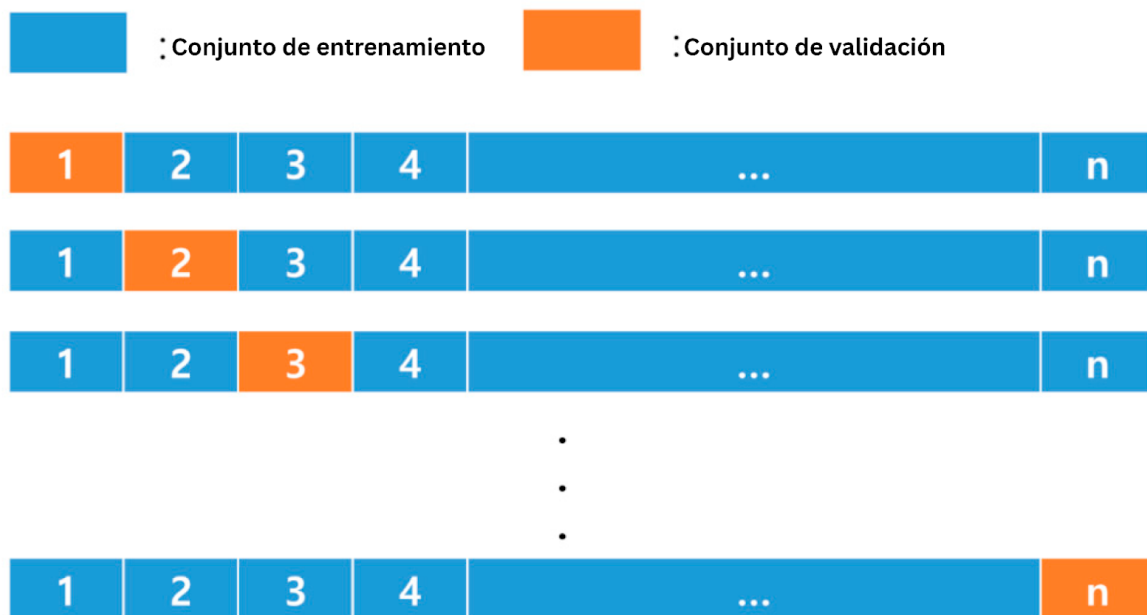


Figura 2.13: Esquema general del funcionamiento de la evaluación por *Leave-one-out Cross Validation*. Figura tomada de (Cha et al. 2020).

2.6 Interfaz gráfica - Front-End

Dado que la mayoría de la comunidad científica no cuenta con los conocimientos necesarios para utilizar aplicaciones mediante la línea de comandos, una solución efectiva consiste en el desarrollo de una interfaz gráfica que facilite su ejecución.

Por otro lado, el desarrollo de interfaces gráficas (*front-end*), representa una rama de la informática muy alejada de la ciencia de datos y la bioinformática, disciplinas en las que se enmarca este trabajo. De esta manera, integrar un *front-end* a las herramientas informáticas desarrolladas puede resultar altamente demandante y complejo. Como solución rápida y eficaz a este tipo de necesidades se creó la librería Streamlit (<https://streamlit.io/>). Streamlit es una librería de Python de código abierto que permite la creación de aplicaciones web personalizadas para herramientas relacionadas con las áreas de aprendizaje automático, ciencia de datos y también bioinformática, entre muchas otras más.

Con solo unas pocas líneas de código, se pueden construir interfaces interactivas que permiten a los usuarios manipular datos, ajustar parámetros y ver los resultados en tiempo real. Streamlit resulta muy sencillo de usar, requiere un código mínimo y permite centrarse en la parte más científica del proyecto, sin requerir mucha capacitación adicional para el desarrollo de *front-end* (Figura 2.14). Además, es muy accesible para principiantes, ya que no requiere conocimientos previos de HTML, CSS o JavaScript, facilitando la implementación de modelos de aprendizaje automático y proyectos de Python sin preocupaciones sobre la interfaz.

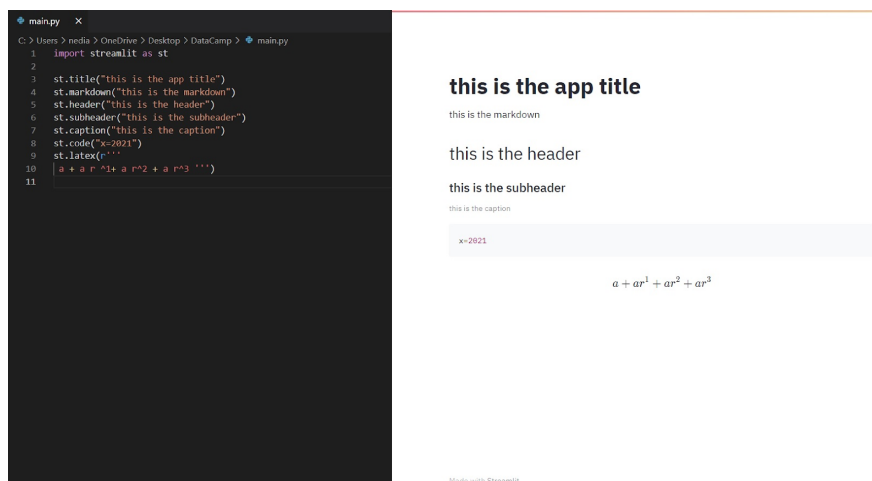


Figura 2.14: Ejemplificación del código requerido para obtener una página web sencilla utilizando Streamlit. Se observa que el desarrollo de una interfaz gráfica funcional se implementa enteramente en Python, con una cantidad mínima de líneas de código. Figura tomada de "Streamlit: A Beginner's Guide," *DataCamp*, <https://www.datacamp.com/tutorial/streamlit>.

3. Desarrollo de un algoritmo para predicción de blancos de compuestos antimicrobianos (ReverseLigQ)

3.1 Introducción e hipótesis

La resistencia antimicrobiana representa uno de los mayores desafíos de salud pública en la actualidad. Sorprendentemente, la Organización Mundial de la Salud (OMS) ha revelado recientemente que ha habido escasos avances en el desarrollo de nuevos antibióticos para combatir las infecciones resistentes a los medicamentos. En este contexto, la identificación de blancos moleculares adecuados es un paso clave en el desarrollo de nuevos fármacos. En la era postgenómica, los enfoques computacionales integrativos han facilitado la identificación y priorización de candidatos a blancos terapéuticos (Serral et al. 2021). Target-Pathogen (<http://target.sbg.qb.fcen.uba.ar>) es un recurso único desarrollado por nuestro Grupo de Investigación que integra datos de drogabilidad estructural, análisis de esencialidad, contexto metabólico, genómica y expresión para clasificar genes y proteínas según su potencial como blancos antimicrobianos (Sosa et al. 2018). Varios estudios han utilizado esta herramienta para seleccionar y priorizar blancos moleculares en patógenos relevantes desde el punto de vista clínico, como *Mycobacterium tuberculosis* (Defelipe et al. 2016), *Klebsiella pneumoniae* (Ramos et al. 2018), *Bartonella bacilliformis* (Farfán-López et al. 2020), *Trypanosoma cruzi* (Pinto et al. 2022) y *Schistosoma mansoni* (Lobo-Silva et al. 2020). Una vez seleccionado el blanco, el desafío es identificar una pequeña molécula capaz de inhibir su función, lo que permitiría su validación farmacológica y, eventualmente, el desarrollo de nuevos antibióticos.

El *screening* fenotípico de compuestos antimicrobianos ha sido una técnica poderosa que, a lo largo del último siglo, ha producido muchas moléculas que se convirtieron en medicamentos de primera línea para tratar enfermedades infecciosas (Ribeiro da Cunha, Fonseca, and Calado 2019). Por ejemplo, la daptomicina, descubierta a partir del cribado de compuestos derivados de *Streptomyces roseosporus* en la década de 1980 y aprobada por la FDA en 2003, es utilizada actualmente para tratar infecciones por Gram-positivos (Katz and Baltz 2016). En ensayos de *screening* fenotípico de alto rendimiento, se evalúan miles de compuestos por su capacidad para inhibir el crecimiento microbiano *in vitro*. Los compuestos más prometedores se someten a estudios más detallados en diversas condiciones de crecimiento y, en algunos casos, en modelos

animales. Además, en el ámbito de los productos naturales, los extractos crudos se tamizan rutinariamente en busca de actividades antimicrobianas, lo que ha llevado a la purificación y caracterización de moléculas activas, como fue el caso del descubrimiento de la penicilina, uno de los antibióticos más importantes del mundo.

Una vez encontrado un compuesto activo, el siguiente problema es identificar su blanco molecular, la biomolécula del huésped a la cual el compuesto se une para producir su efecto. Comprender el blanco molecular es crucial para el desarrollo posterior del compuesto, permitiendo su optimización como fármaco. Los blancos más comunes de los fármacos antimicrobianos son proteínas esenciales, que desempeñan funciones vitales en el patógeno. Al inhibir estas proteínas, el fármaco provoca la muerte del microorganismo.

Sin embargo, la identificación del blanco molecular de un compuesto es una tarea difícil. En la era pregenómica, se tardaba años en determinar el mecanismo de acción de un compuesto. Por ejemplo, el blanco de la isoniazida, un fármaco clave contra la tuberculosis, no se confirmó hasta 20 años después de su descubrimiento, cuando se demostró que inhibe la biosíntesis del ácido micólico (Barclay, Ebert, and Kochwaser 1953). Hoy en día, una técnica poderosa para identificar blancos es la inducción de resistencia en el microorganismo y la comparación de los genomas de las cepas parental y resistente. Este método ha sido utilizado con éxito para identificar el blanco de la bedaquilina, que inhibe la ATP sintasa en micobacterias (Andries et al. 2005). Sin embargo, este proceso es costoso y puede fallar por diversas razones, como la imposibilidad de obtener una cepa resistente o la dificultad para identificar correctamente la mutación relevante que causa la resistencia.

En respuesta a estos desafíos, los enfoques bioinformáticos han ganado relevancia para ayudar a los investigadores en la identificación de blancos moleculares. El *screening* virtual permite seleccionar compuestos que potencialmente se unen a un blanco determinado utilizando herramientas bioinformáticas y quimioinformáticas (Gimeno et al. 2019). El *screening* virtual inverso, por otro lado, se utiliza cuando ya se ha identificado un compuesto activo, y se busca predecir su blanco molecular.

En esta primera parte del trabajo se propone el desarrollo de un método informático novedoso de *screening* virtual inverso. **Se plantea como hipótesis que la combinación de métodos de similitud química, criterios de priorización de proteínas según su relevancia biológica —específicamente su esencialidad y su asociación a reacciones metabólicas críticas—, junto con el *docking* molecular permite otorgar una lista reducida de posibles blancos moleculares de un compuesto con probada actividad inhibitoria frente a un organismo patogénico aptos para avanzar hacia su validación en ensayos experimentales.**

3.2 Resultados

Los resultados de este capítulo se describen de la siguiente manera: en primer lugar se caracterizan los genomas de cuatro organismos que presentan diferencias significativas en el tamaño de su genoma y en la información disponible sobre ligandos y blancos moleculares.

Luego se describe la metodología propuesta para predecir cuáles pueden ser los posibles blancos de unión de una droga determinada, dentro de las proteínas codificadas en cada genoma, utilizando búsquedas de compuestos similares con mecanismos de acción conocidos, y se evalúa la capacidad predictiva.

Por último se evalúan otras metodologías adicionales para alcanzar una lista final de proteínas candidatas a ser blancos moleculares de la droga en estudio: por un lado el *docking* molecular, y por el otro la caracterización individual de proteínas respecto a su importancia para la supervivencia del organismo patógeno en cuanto a la esencialidad y características de las redes metabólicas en las que intervienen.

3.2.1 Caracterización de los genomas de cuatro organismos diferentes

Se seleccionaron cuatro organismos con marcada diferencia en cuanto a grado, tamaño genómico y pertenencia taxonómica: *Mycobacterium tuberculosis* (Mtb), *Bartonella bacilliformis* (Bb), *Klebsiella pneumoniae* (Kp) y *Trypanosoma cruzi* (Tc).

En una primera fase, se determinó el número de proteínas y la cantidad de dominios funcionales correspondientes presentes en cada organismo. Por una parte, muchos organismos patógenos están poco estudiados y apenas se conocen fármacos inhibidores con actividad reportada. Por otro lado, se considera importante el desarrollo de una herramienta que sea aplicable **a cualquier patógeno** con proteoma secuenciado y que permita **el hallazgo de mecanismos de acción novedosos**. Por estos motivos, en este primer análisis se procura maximizar la cobertura del proteoma de cada organismo en cuanto a la cantidad de dominios y, en consecuencia, de proteínas con compuestos inhibidores asociados.

Para abordar esta evaluación, las drogas asociadas con las proteínas de cada organismo se organizaron en cuatro grupos distintos. El grupo I incluye drogas que se unen a proteínas del organismo y que han sido cocrystalizadas en complejo con éstas, y tienen estructuras subidas al Protein Data Bank (PDB) (Berman et al. 2000). Por otro lado, el grupo II abarca drogas que se unen a proteínas de otros organismos, extraídas de esta base de datos (PDB), que presentan los mismos dominios de Pfam que las proteínas del patógeno de interés, es decir que son inhibidores relacionados a proteínas que se asemejan estructuralmente y posiblemente en funcionalidad a las del organismo en

cuestión. Los grupos III y IV replican la estructura de los grupos I y II, respectivamente, pero en cuanto a drogas extraídas de ensayos de ChEMBL con actividad inhibitoria probada (Zdrazil et al. 2023). Se aclara que en el caso de la información extraída de PDB, al disponerse la estructura del complejo ligando-proteína, se conoce con precisión el dominio con el que interactúa cada compuesto. Esto no ocurre en el caso de ChEMBL, donde únicamente se conoce el blanco molecular y los dominios de Pfam que éste contiene, pero se desconoce con precisión cuál de estos dominios interactúa con cada ligando. Para poder utilizar la vasta información contenida en ChEMBL se realizaron los siguientes supuestos. Si el blanco tiene un solo dominio, se asume que el ligando se une a dicho dominio dado que se ha demostrado previamente que aproximadamente solo el 2.5% de los sitios de unión de ligandos se encuentran por fuera de un dominio de Pfam (Kruger, Rostom, and Overington 2012). Por otro lado, si el blanco contiene más de un dominio, se utilizó la siguiente aproximación adicional: se identificaron los dominios presentes en PDB que interactúan con ligandos estructuralmente similares a cada fármaco de ChEMBL (Índice de Tanimoto ≥ 0.4). Si uno o más de estos dominios pertenecen al mismo blanco que figura en ChEMBL, todos ellos se asignan como sitios de unión del compuesto. Si bien se considera que esta aproximación puede representar una limitación en cuanto a la precisión en la asignación de dominios de unión para cada ligando, se decidió utilizarla con el fin de aprovechar la gran cantidad de datos provenientes de ensayos experimentales presentes en ChEMBL.

Organismo	<i>Mycobacterium tuberculosis</i>	<i>Bartonella bacilliformis</i>	<i>Klebsiella pneumoniae</i>	<i>Trypanosoma cruzi</i>
Número total de proteínas	4081	1143	5491	10288
Número total de dominios pfam (acumulativo de cada proteína por separado)	6293	1473	8737	8991
Número total de dominios pfam únicos	2141	1068	2966	2335
Número total de dominios pfam únicos que unen ligandos del grupo I (porcentaje sobre el total de dominios del organismo)	254 (11.9%)	0 (0.0%)	17 (0.57%)	16 (0.69%)

Número total de dominios pfam únicos que unen ligandos del grupo II (porcentaje sobre el total de dominios del organismo)	1117 (52.2%)	674 (63.1%)	1287 (43.4%)	1131 (48.4%)
Número total de dominios pfam únicos que unen ligandos del grupo III (porcentaje sobre el total de dominios del organismo)	23 (1.1%)	0 (0.0%)	0 (0.0%)	8 (0.3%)
Número total de dominios pfam únicos que unen ligandos del grupo IV (porcentaje sobre el total de dominios del organismo)	365 (17.0%)	100 (9.4%)	364 (12.3%)	403 (17.3%)
Número total de dominios pfam únicos que unen ligandos (todos los grupos)	1157 (54.0%)	679 (63.6%)	1316 (44.4%)	1154 (49.4%)
Número de ligandos totales considerando todos los grupos en conjunto	214728	99208	237345	264428

Tabla 3.1: Recuento total de pares ligando-proteína (blanco molecular) que se pueden asociar en *M. tuberculosis*, *B. bacilliformis*, *K. pneumoniae* y *T. cruzi*, utilizando los datos de PFAM y ChEMBL. Se muestran los datos correspondientes a cada uno de los grupos descritos.

Al realizar un análisis de manera global (Tabla 3.1), se observa que únicamente considerando las proteínas de los patógenos en estudio, es muy reducido el número de dominios proteicos que cuentan con ligandos en el Protein Data Bank (PDB) o en ChEMBL (grupos I y III). La representación es significativamente mayor para *Mycobacterium tuberculosis*, un organismo más estudiado, mientras que para los demás organismos, estos números resultan nulos o insignificantes, como en el caso de *Bartonella bacilliformis*, donde no se obtuvieron ligandos en absoluto.

No obstante, al considerar los grupos II y IV, con un enfoque más amplio, se logra cubrir aproximadamente la mitad del proteoma de cada organismo (en el caso de *B. bacilliformis*, esta representación alcanza el 63.6%).

Estos resultados permiten concluir que, para desarrollar una herramienta aplicable a cualquier organismo patógeno y con potencial para identificar mecanismos de acción novedosos, es necesario adoptar una aproximación basada en la *culpa por asociación*: las proteínas que son similares estructuralmente (en este caso aquellas que comparten un mismo dominio proteico), tienden a compartir función biológica, y en consecuencia a interactuar con moléculas similares. Esta premisa fue ampliamente documentada en investigaciones sobre la promiscuidad de compuestos bioactivos (Y. Hu and Bajorath 2015; Klabunde 2007) y en estudios sobre mecanismos de antibióticos de amplio espectro (Hooper and Jacoby 2016; Kocaoglu and Carlson 2015). Los análisis subsiguientes se llevarán a cabo bajo este supuesto.

3.2.2 Generación de un algoritmo predictivo basado en aprendizaje no supervisado para la búsqueda de predicción de targets para drogas con mecanismo de acción desconocido

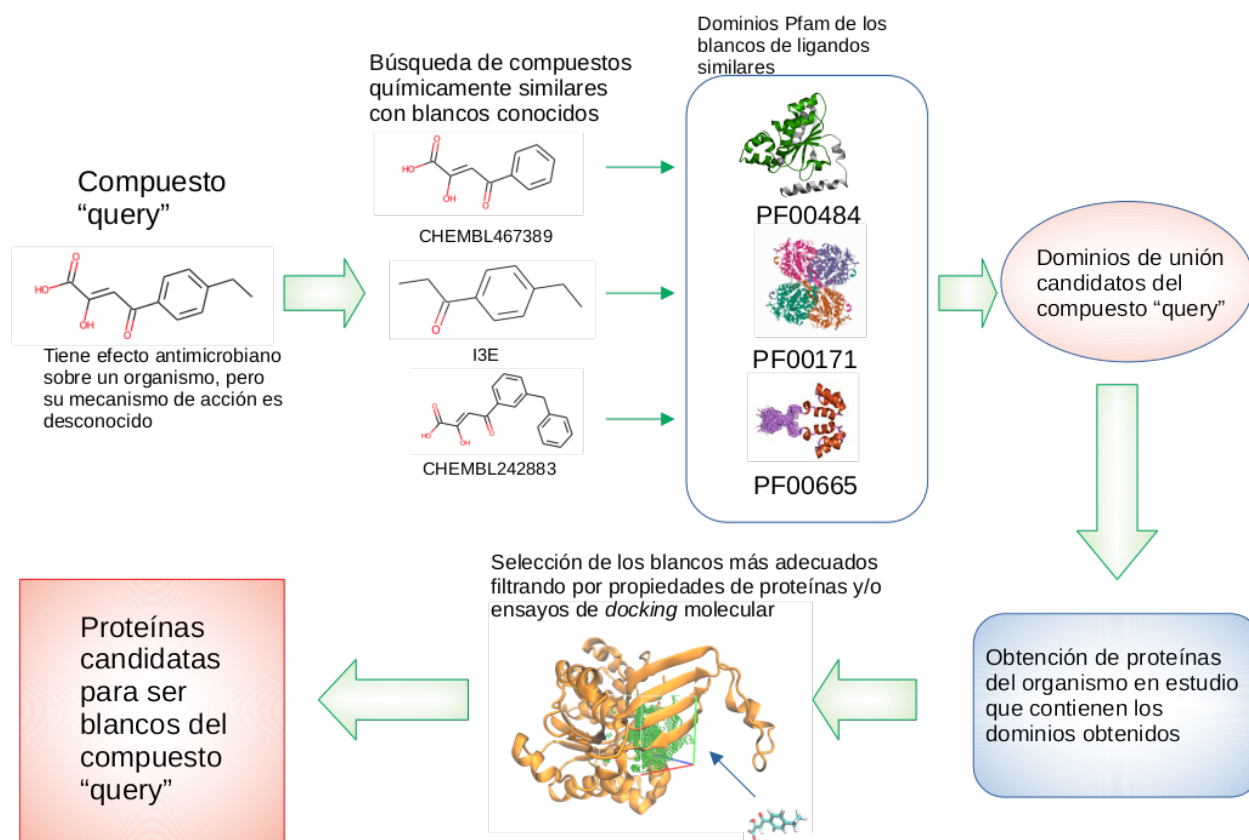


Figura 3.2: Esquema general de la metodología desarrollada.

El flujo de trabajo para el *screening* virtual inverso en un organismo patogénico parte de dos bases de datos: la primera (BD1) comprende fármacos y los dominios proteicos que estos afectan. Mientras que la segunda (BD2) contiene proteínas agrupadas por sus dominios de Pfam. Partiendo de esta información, el proceso se estructura en diferentes etapas, descritas a continuación:

1. **Ingreso de un Compuesto “query”:** Se parte de un compuesto al que llamamos “query”, que inhibe el crecimiento o elimina un organismo en estudio pero su mecanismo de acción no se ha determinado. Se realiza una búsqueda de compuestos con similitud química (por su Índice de Tanimoto) al compuesto “query” dentro de la base de datos BD1, utilizando un umbral de corte determinado por el usuario.
2. **Obtención de dominios candidatos:** A partir de los compuestos similares obtenidos en la búsqueda, se obtienen los dominios que estos unen de la misma base de datos BD1.
3. **Obtención de proteínas candidatas:** Partiendo de los dominios obtenidos en el paso anterior, se realiza una búsqueda dentro del proteoma del organismo de las proteínas que contienen cada uno de estos dominios (información contenida en la base BD2). Esto nos otorga una lista de proteínas candidatas a ser blanco del compuesto “query”.
4. **Filtro de proteínas candidatas:** Por ensayos de *docking molecular* y propiedades de la proteína (especialmente esencialidad), se eligen aquellas proteínas cuya interacción con el ligando “query” es más probable.
5. **Proteínas candidatas finales:** El resultado del proceso es la obtención de proteínas candidatas del organismo en estudio que podrían unirse al compuesto “query”. Estas proteínas son seleccionadas como blancos moleculares potenciales para evaluaciones experimentales y/o investigaciones más detalladas.

Se utilizó una idea similar a la descrita en un anterior trabajo del Grupo de Investigación, donde se demostró que utilizando la similitud química es posible encontrar un conjunto de ligandos capaces de inhibir una proteína dada partiendo de compuestos con actividad inhibitoria conocida (Radusky et al. 2017).

La capacidad del proceso para identificar blancos moleculares relevantes depende de varios factores clave: 1) la cantidad y diversidad de posibles blancos moleculares asociados a cada uno de los dominios candidatos, 2) el número y la variabilidad de los ligandos conocidos para los dominios presentes en el patógeno, 3) el enfoque de

similitud química empleado, junto con el umbral de corte seleccionado para definir similitudes significativas, y 4) la metodología aplicada para agrupar las proteínas del organismo.

En el caso de este trabajo, el criterio para definir la similitud química entre compuestos es el Índice de Tanimoto, y para clusterizar las proteínas se utilizó la asignación de dominios de Pfam, como se describe en la sección de métodos específicos de este capítulo. El esquema del proceso se encuentra ilustrado en la Figura 3.2.

3.3.3 Evaluación del algoritmo predictivo

La evaluación confiable del poder predictivo de esta metodología resulta ser un desafío importante, ya que está fuertemente condicionada por características individuales del ligando a evaluar, principalmente por la capacidad de encontrar en bases de datos drogas similares cuyo mecanismo de acción fue previamente determinado.

La validación utilizando pares ligando-dominio de unión (el dominio de Pfam sobre el que actúa cada droga y que representa a cada clúster de proteínas de un organismo en estudio) conocidos resulta ser una muy buena alternativa para corroborar la capacidad predictiva de este enfoque de aprendizaje no supervisado basado en similitud química de compuestos.

Para poder realizar estas evaluaciones, se seleccionaron compuestos de testeo aleatorios de nuestras propias bases de datos (PDB y ChEMBL por separado). En la evaluación se sostuvo el supuesto de *culpa por asociación* siguiendo la premisa especificada en la sección 3.1.1: si un ligando inhibe una proteína con un determinado dominio, se asume que puede inhibir a cualquier otra proteína que comparta ese dominio. De esta manera se permite la evaluación en organismos sin ligandos reportados, como *Bartonella bacilliformis*. El método elegido no depende de datos experimentales específicos del organismo, sino de proyectar información funcional desde compuestos estructuralmente similares y con mecanismo de acción conocido.

La estrategia de evaluación para cada compuesto de testeo consistió en primer lugar en excluir de la base de datos tanto el propio ligando como aquellos con un Índice de Tanimoto (TI) superior a 0.9. Posteriormente, se realizó una búsqueda de vecinos dentro del conjunto de datos restante, que contenía únicamente ligandos con una similitud menor a 0.9. Además, debido al gran tamaño del dataset, se utilizó un subset de 1000 ejemplos aleatorios para evaluar.

El objetivo de la evaluación fue determinar si por medio de algún otro compuesto similar de la base de datos es posible recuperar el dominio de unión de un ligando “*query*”.

Se probaron diferentes umbrales de similaridad máxima de TI, desde 0.8 hasta 0.3, para realizar las búsquedas de vecinos.

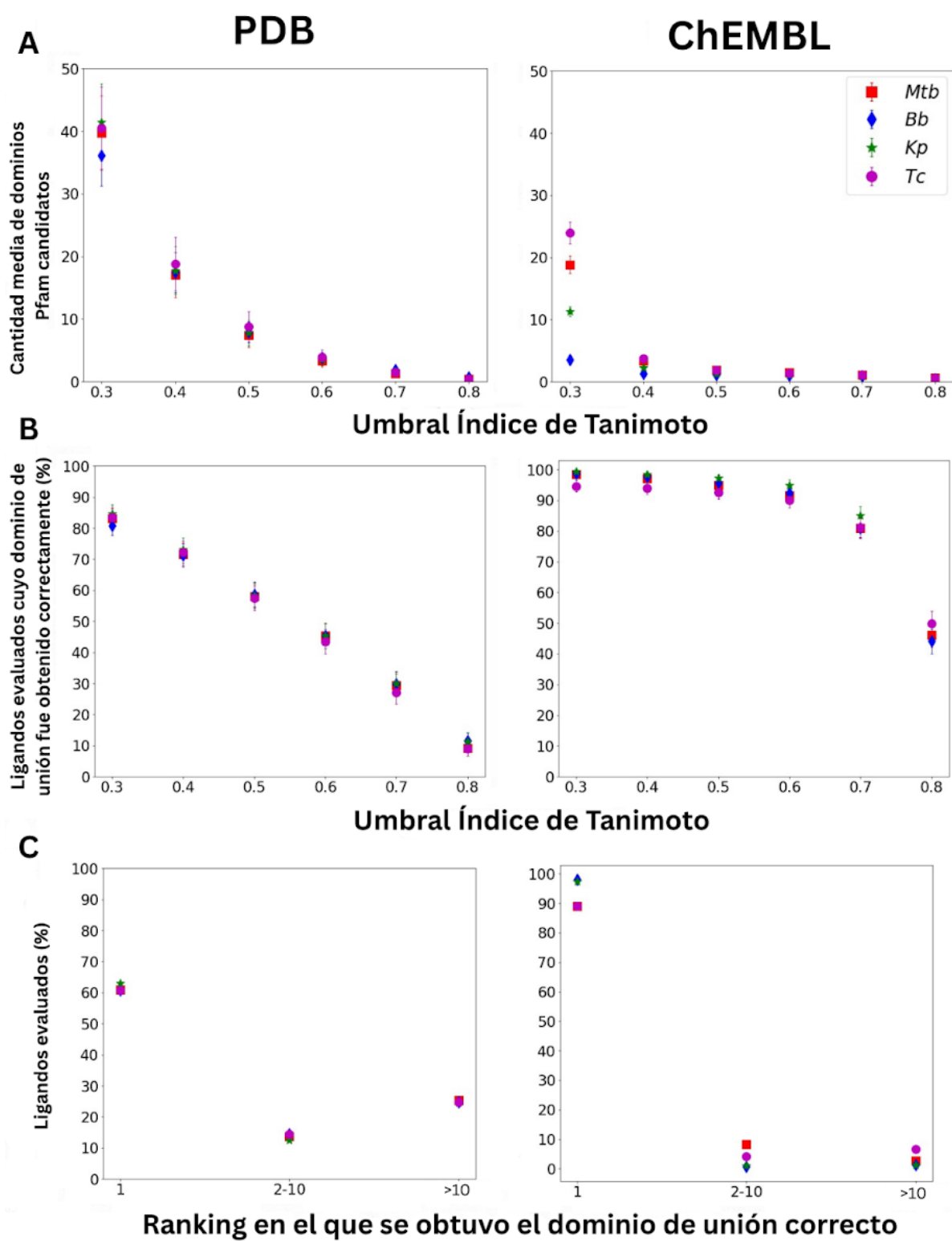


Figura 3.3: **A)** Media de la cantidad de dominios de Pfam encontrados para cada compuesto del set de testeo por umbral de Índice de Tanimoto (TI). **B)** Porcentaje de ligandos evaluados cuyo dominio

de unión se pudo recuperar exitosamente por medio de un compuesto similar para cada umbral de TI, tanto si se encontraron o no dominios de unión candidatos. **C)** Rankings más frecuentes, por orden de similitud, que indican la posición del compuesto similar mediante el cual se identifica el dominio de unión verdadero de cada ligando evaluado. Los gráficos se construyen en base a 1000 ejemplos de ligandos evaluados, extraídos de los conjuntos de datos del PDB (izquierda) y de ChEMBL (derecha).

En primer lugar, la evaluación de nuestra metodología reveló que hay diferencias importantes entre analizar los datos provenientes de PDB y de ChEMBL. ChEMBL contiene un gran número de ligandos (en promedio 191737 por organismo), pero estos en general interactúan con una cantidad reducida de dominios proteicos (que alcanzan una media del 14.0% de la cantidad total por patógeno), si bien estos dominios tienen un alto impacto clínico, ya que están asociados con procesos biológicos relevantes para el desarrollo de fármacos (Gaulton et al. 2012). Por otro lado, el set de datos de PDB se caracteriza por una mayor diversidad en términos de la cantidad de proteínas cubiertas (se cubre un 51.8% de los dominios presentes en cada organismo), abarcando un espectro más amplio de genomas de distintos organismos. Sin embargo, muchas de las interacciones documentadas en PDB no están necesariamente vinculadas a aplicaciones clínicas directas, sino que pueden representar estudios estructurales o funcionales de proteínas sin relevancia médica inmediata. Por otra parte, también la cantidad de ligandos en los datos extraídos del PDB es mucho menor a la que se encuentra en ChEMBL (12190 compuestos en promedio por organismo), es decir, que presenta un número bajo de ligandos asociados a cada proteína.

Los resultados obtenidos mostraron que, al reducir el umbral de similaridad de Tanimoto (TI), se incrementa tanto la variabilidad como el número de dominios Pfam identificados, algo que era de esperarse. Esta tendencia es lógica, ya que al incluir una mayor cantidad de compuestos en el análisis, es más probable que se encuentren interacciones con una mayor variedad de dominios (Figura 3.3 A). En el caso de ChEMBL, al aplicar un umbral de 0.5, se observó una cantidad media muy baja de dominios candidatos, de entre dos y tres por compuesto. Sin embargo, lo más destacable es que, para más del 90% de los ligandos evaluados, el dominio correcto estaba entre los obtenidos (Figura 3.3 B, derecha).

Por otro lado, en el caso de PDB, al reducir el umbral de TI a 0.5, se logró identificar el dominio de unión correcto en aproximadamente el 60% de los ligandos evaluados, entre un promedio de 10 dominios candidatos (Figura 3.3 B, izquierda).

Además, un resultado particularmente relevante consistió en que, cuando se obtienen resultados en los conjuntos de ligandos evaluados, el dominio de unión de cada

compuesto **suele coincidir (en más del 60% de los casos) con el dominio de unión del ligando más similar encontrado en la base de datos, es decir el primero del ranking, independientemente del umbral de TI utilizado.** Esta observación refuerza la premisa de que los compuestos con alta similaridad estructural tienden a compartir mecanismos de acción, interactuando de manera equivalente con las mismas proteínas (Figura 3.3 C).

Las tendencias identificadas en esta evaluación parecen mantenerse de manera similar para los cuatro organismos tomados como caso de estudio, lo cual nos sugiere que la metodología podría aplicarse de la misma manera a otros patógenos y/o organismos de interés.

El algoritmo genera, en esta primera etapa, una lista amplia de posibles blancos moleculares tras identificar los dominios de unión candidatos del compuesto bioactivo. Dado que dicha lista constituye el punto de partida para investigar el mecanismo de acción del fármaco, se optó por la inclusión del mayor número posible de proteínas potencialmente relevantes, asumiendo cierta permisividad respecto a la presencia de falsos positivos. Para acotar esta primera lista, se evaluaron posteriormente métodos destinados a priorizar los blancos candidatos más plausibles. Estos enfoques se detallan a continuación en las secciones 3.3.4 y 3.3.5.

3.3.4 *Docking* molecular

El primer método utilizado para reducir la lista de resultados posibles es el *docking* molecular. En esta etapa, se partió de la lista inicial de dominios de unión candidatos, obtenida de la búsqueda previa, con el objetivo de priorizar aquellos con mayor probabilidad de ser los verdaderos blancos. Para esta evaluación se utilizaron datos correspondientes a *Mycobacterium tuberculosis*. Se seleccionaron 20 ligandos del set de PDB y 10 del set de ChEMBL. Cada uno de ellos fue evaluado en 10 dominios diferentes, correspondientes al ranking generado por el modelo basado en similitud química. Dentro de estos 10 dominios, se encuentra el verdadero dominio de unión de cada ligando.

La evaluación de los resultados se realizó mediante el uso de técnicas de *docking* convencional y sesgado. Dado que no existían estructuras disponibles para todas las proteínas (y el uso de AlphaFold no era tan extendido en la época de realización de esta evaluación), se optó por tomar una estructura representativa de una proteína para cada dominio candidato. Dentro de todas las posibilidades existentes, se eligieron las proteínas que unen a los ligandos obtenidos a partir de nuestra búsqueda por similitud química (los que nos conducen a los dominios de unión candidatos de la búsqueda previa).

Para el caso del *docking* sesgado, los sesgos introducidos fueron determinados por interacciones conocidas entre átomos de la proteína representativa en el cristal correspondiente de PDB y los ligandos cocrystalizados con la misma.

Cada ligando fue evaluado en las proteínas representativas de los 10 dominios candidatos, incluyendo el dominio de unión correcto. Los resultados obtenidos se resumen en un gráfico de Z-score (Figura 3.4), que considera tanto la energía libre de unión como el tamaño del clúster de poses similares.

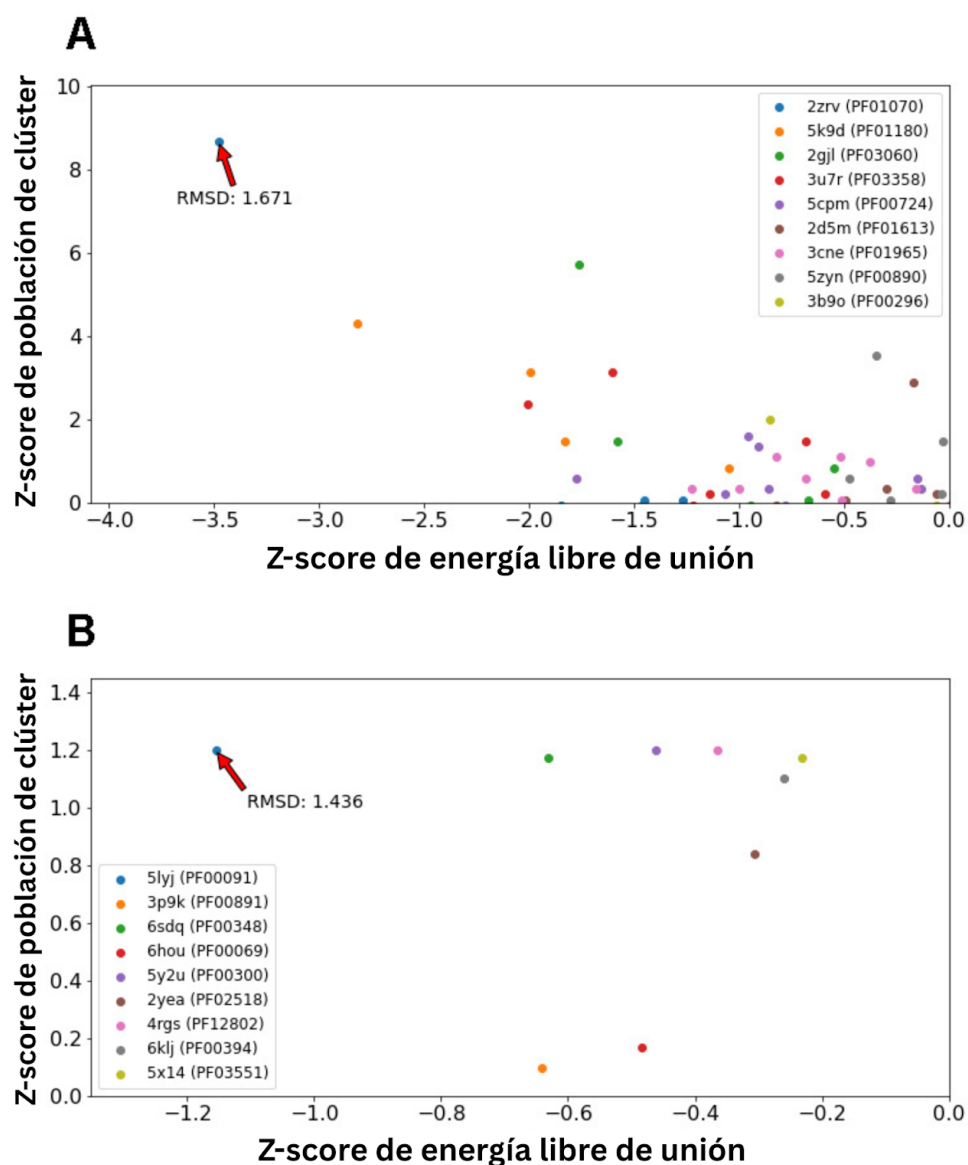


Figura 3.4: Gráfico normalizado de energía libre de unión predicha por el *docking* (eje x) y población de clúster de poses del ligando similares (eje y), en el cual se comparan todos los

dominios de unión candidatos (representados por una estructura para cada uno) con el dominio de unión correcto indicado para los compuestos 9OR (arriba) y 89U (abajo).

Tomando el Z-score se armó un ranking de las mejores poses. Dentro del ranking se incluyen todas las poses obtenidas en los ensayos de *docking* realizados en las 10 proteínas representativas de cada dominio candidato.

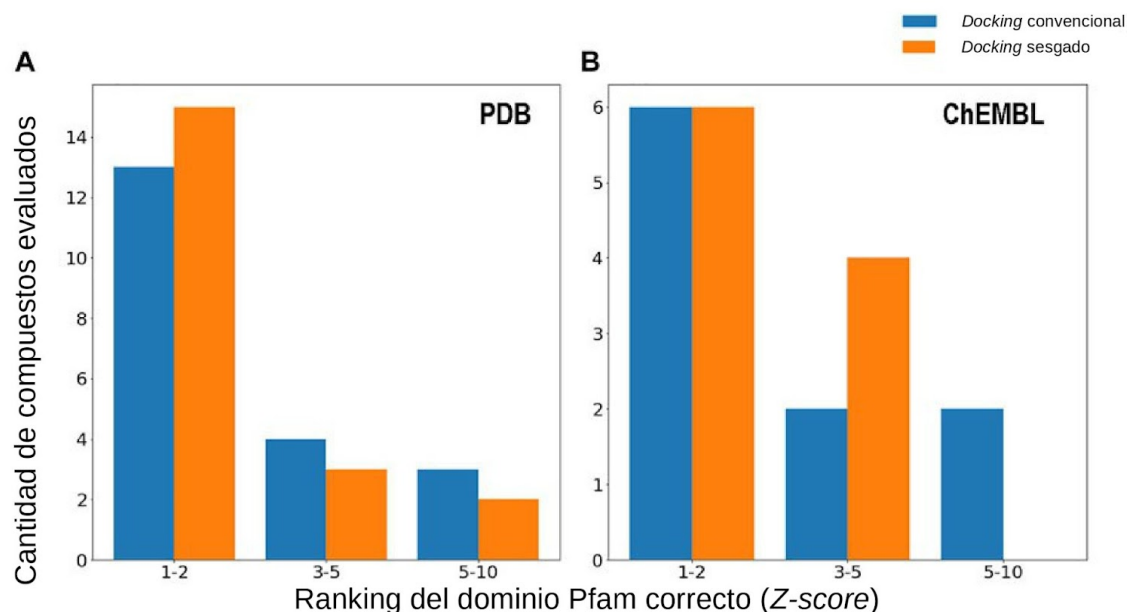


Figura 3.5: Posiciones donde se encontró el dominio de unión correcto en los rankings de Z-score. Se muestran los resultados obtenidos para *docking* sesgado y convencional. Las evaluaciones fueron realizadas para los 20 compuestos evaluados del conjunto de datos del PDB (A) y los 10 ligandos probados del conjunto de datos de ChEMBL (B).

Los resultados muestran que para la mayoría de los ligandos se determinó de manera consistente que el dominio correcto de unión se encuentra en una posición destacada del ranking de Z-score, generalmente entre los dos primeros puestos, tanto en el caso de los ligandos del PDB como en los de ChEMBL. En algunos casos aislados el dominio correcto se ubicó por debajo del puesto 5 (Figura 3.5).

Estos resultados indican que el *docking* molecular es una herramienta eficaz para maximizar las posibilidades de encontrar el dominio de unión de la droga dentro del genoma analizado.

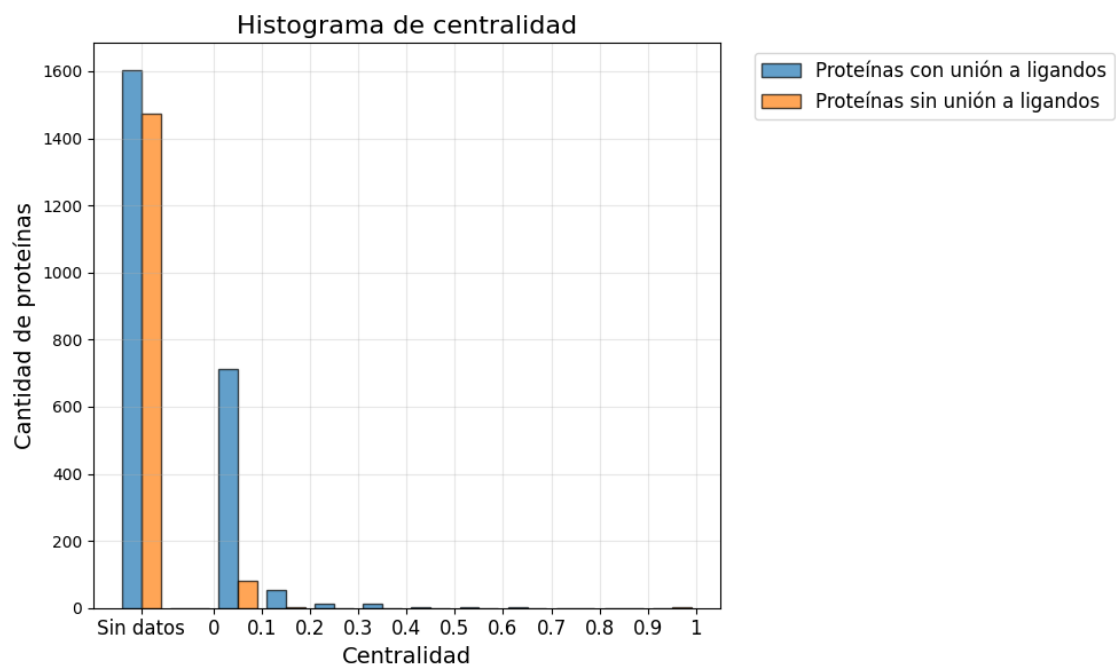
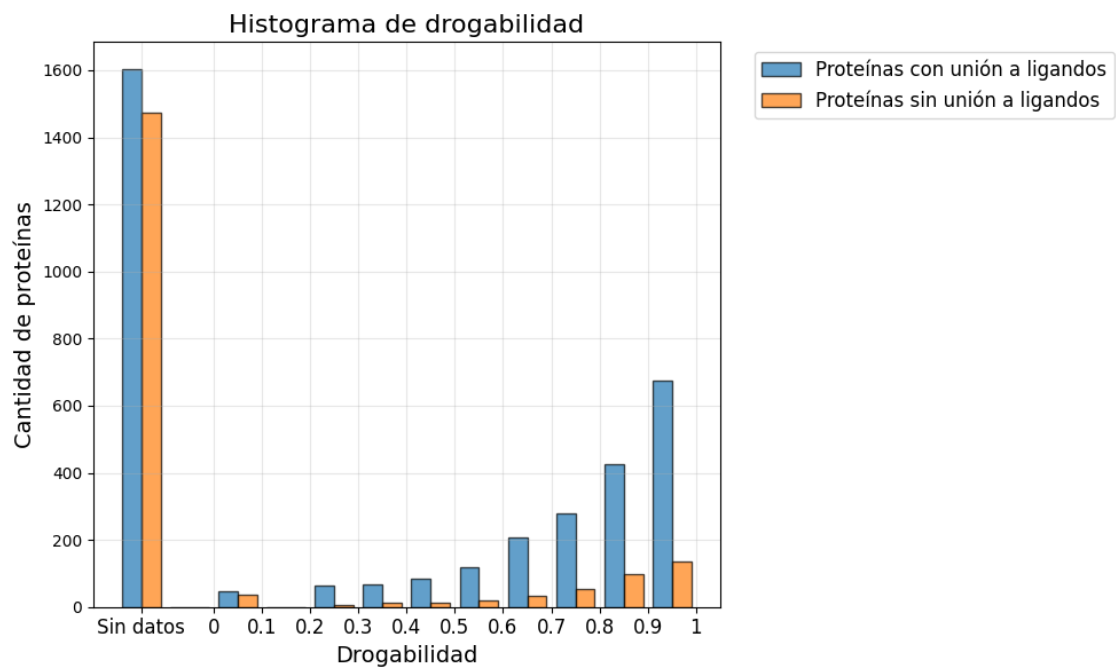
Es importante destacar que, en líneas generales, entre las pruebas con *docking* molecular convencional y con *docking* sesgado, no se obtuvieron diferencias significativas. Sin embargo, el enfoque sesgado demostró ser ligeramente más eficiente, especialmente al evaluar los ligandos de ChEMBL.

3.3.5 Caracterización de las proteínas de cada organismo según sus propiedades

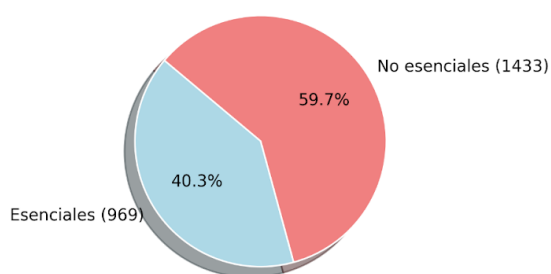
Hasta el momento, nuestro enfoque alcanza el nivel de inclusión de dominios, es decir que la unidad mínima representada en los resultados obtenidos corresponde a un grupo de proteínas con cierta similitud estructural y funcional. Debido a esto, se investigaron enfoques para, en un post-procesamiento, poder ampliar la “resolución” de los resultados hasta proteínas individuales de cada organismo.

Para refinar estos resultados se analizaron distintos parámetros de las proteínas de forma individual, para conocer cuáles de todas aquellas que fueron obtenidas en primera instancia son las más propensas a ser afectadas por el compuesto “*query*”.

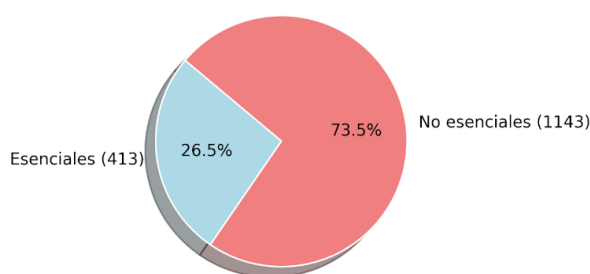
Se evaluaron cuatro características diferentes: esencialidad, drogabilidad, si la reacción catalizada es un punto de estrangulamiento y su valor de centralidad (de intermediación) en la red metabólica. Las dos primeras se relacionan directamente con las proteínas, mientras que las últimas dos están asociadas a las redes metabólicas en las que estas proteínas intervienen (Figura 3.6).



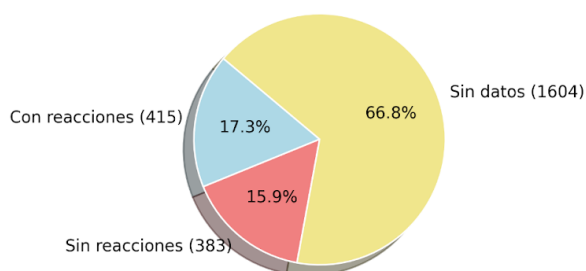
Proteínas esenciales o no (con unión a ligandos)



Proteínas esenciales o no (sin unión a ligandos)



Proteínas con o sin reacciones correspondientes a puntos de estrangulamiento (con unión a ligandos)



Proteínas con o sin reacciones correspondientes a puntos de estrangulamiento (sin unión a ligandos)

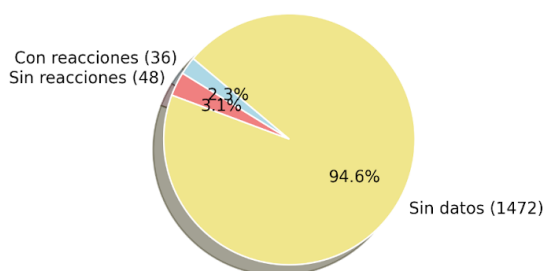


Figura 3.6: Histogramas y gráficos de torta que muestran los parámetros utilizados para la caracterización de las proteínas. Se comparan dos subconjuntos: por un lado, las proteínas que tienen uniones con ligandos documentadas en bases de datos, y por otro, aquellas para las que no se dispone de información al respecto. Este análisis se realizó para el proteoma de *Mycobacterium tuberculosis*.

También se realizó un análisis comparativo entre dos conjuntos de proteínas: aquellas que presentan ligandos asociados y aquellas que carecen de esta información. Estas evaluaciones revelaron diferencias significativas en cuanto a la potencialidad de cada grupo como posibles blancos terapéuticos.

En una primera observación, se constató que las proteínas con ligandos reportados cuentan con una mayor disponibilidad de información respecto a sus características, mientras que el grupo de proteínas sin esta asociación es considerablemente más "oscuro" y carece de datos detallados.

En términos de drogabilidad, se observó que la mayoría de las proteínas, independientemente de su categoría, exhiben valores generalmente elevados. No obstante, resultó paradójico que algunas proteínas que tienen ligandos asociados en bases de datos presentaron niveles de drogabilidad considerablemente bajos, contrariamente a lo esperado.

En cuanto a la centralidad en las reacciones metabólicas y su participación en puntos de estrangulamiento, se notó una escasez de información suficiente para realizar un análisis exhaustivo del proteoma de un organismo. Los valores de centralidad, en general, fueron bajos para las proteínas conocidas, y la información sobre su correspondencia con reacciones que son puntos de estrangulamiento es muy limitada. Estos dos parámetros, a primera vista, podrían no ser de gran utilidad a escala global; sin embargo, podrían ofrecer información valiosa en estudios específicos.

Finalmente, al considerar la esencialidad de las proteínas, se observó un alto porcentaje de proteínas esenciales en ambos grupos, con una tendencia aún más marcada en aquellas con ligandos reportados. Este hallazgo resalta la importancia de considerar la esencialidad como un criterio relevante al evaluar la factibilidad de las proteínas como posibles blancos terapéuticos.

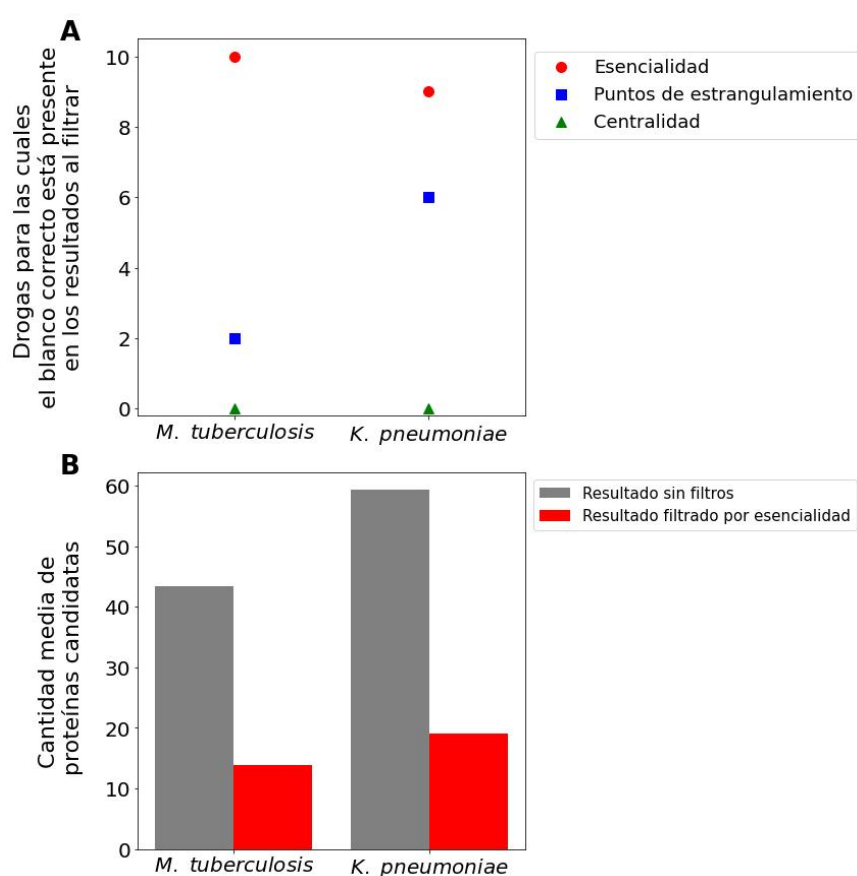


Figura 3.7: A) Pares de ligando-proteína (blanco molecular) en los que la proteína correcta se mantiene dentro de los resultados después de aplicar un filtrado basado en características proteicas, tales como esencialidad, implicación en puntos de estrangulamiento (*choke-points*) o participación en reacciones centrales dentro de vías metabólicas.

B) Promedio del número de dominios candidatos antes y después del filtrado por esencialidad. Los datos fueron obtenidos a partir de conjuntos de 10 fármacos conocidos, utilizados para el tratamiento de *M. tuberculosis* y *K. pneumoniae*.

Luego de detallar y caracterizar las características en términos de esencialidad, participación en puntos de estrangulamiento y alta centralidad, se procedió a evaluar su relevancia tanto para reducir la cantidad de proteínas candidatas como para corroborar que el blanco molecular correcto (la proteína que interactúa con el ligando de interés) pasara el filtro y permaneciera entre los resultados obtenidos (Figura 3.7).

Este análisis se realizó evaluando pares de ligando-blanco molecular correspondientes a fármacos conocidos por inhibir el crecimiento de dos organismos en estudio: *Mycobacterium tuberculosis* y *Klebsiella pneumoniae*, los cuales se detallan en la Tabla Suplementaria S.3.1, del Anexo.

Entre los parámetros evaluados, la esencialidad demostró ser la única característica efectiva para refinar los resultados. Al utilizar la esencialidad como filtro, se observó que, en general, el blanco molecular se mantenía dentro de los resultados, con la excepción de un caso particular en *K. pneumoniae*. En contraste, las otras características, cuando se emplearon como filtros, tendían a excluir el blanco molecular correcto de los candidatos, posiblemente debido a la falta de información detallada sobre su relación con la eficacia como blancos terapéuticos.

Es importante señalar que la drogabilidad fue excluida de esta evaluación, ya que se considera que los dominios conocidos por unirse a ligandos son intrínsecamente drogables.

3.3.6 Evaluación adicional de la metodología con drogas cuyos blancos moleculares no están presentes en la versión de las bases de datos utilizadas

Como parte fundamental de la validación de nuestra metodología, se aplicaron los enfoques desarrollados a casos de estudio documentados previamente, con el objetivo de evaluar su capacidad para recuperar blancos moleculares conocidos. En particular, se seleccionaron compuestos con actividad inhibitoria contra *Mycobacterium tuberculosis*, publicados en ChEMBL por la compañía farmacéutica GlaxoSmithKline (GSK) (Hersey 2013), junto con otros ejemplos relevantes extraídos de la literatura científica. En todos los casos, se trata de fármacos cuyos blancos moleculares fueron identificados posteriormente a la versión de ChEMBL utilizada para el desarrollo de nuestra metodología, lo que permite evaluar su capacidad predictiva. A continuación, se detallan los resultados obtenidos y se

comparan nuestras predicciones con investigaciones anteriores que utilizaron tanto enfoques *in-silico* como experimentales.

Análisis del set de compuestos con actividad probada y mecanismo no elucidado publicados por GlaxoSmithKline

La empresa GSK publicó los resultados de un exhaustivo *screening* fenotípico de compuestos, identificando 776 fármacos con actividad inhibitoria contra *Mycobacterium tuberculosis*, cuyos resultados fueron posteriormente integrados en la base de datos ChEMBL.

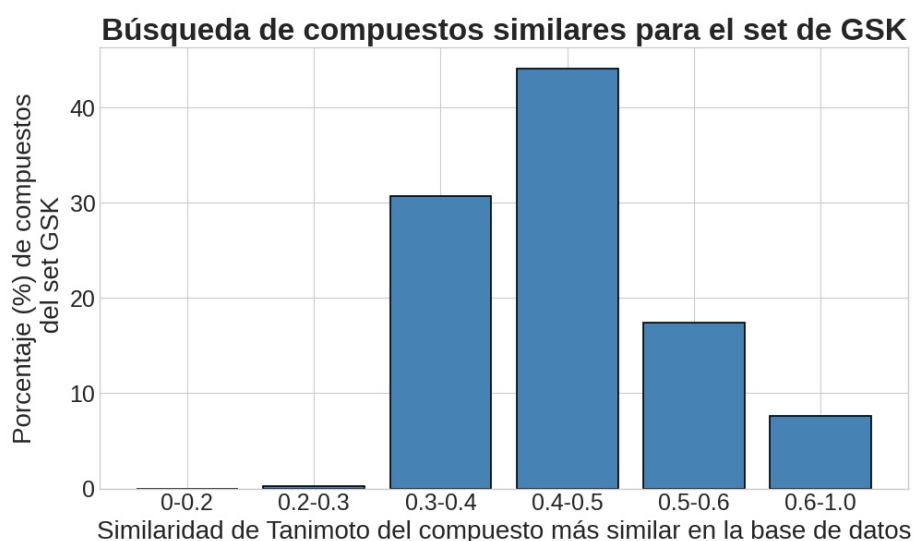


Figura 3.8: Histograma de distribución de similitud máxima entre compuestos del set de GSK con otros compuestos con bioactividad conocida extraídos de las bases de datos PDB y ChEMBL.

En una primera instancia, se realizó un análisis de comparación de similitud química entre los ligandos de nuestra base de datos de moléculas estudiadas que tienen blanco molecular conocido y los compuestos publicados por GSK. Se observó que la mayoría de los compuestos del set de GSK (69%) presentaban ligandos con una similitud al menos moderada dentro de nuestras bases de datos, con un Índice de Tanimoto (TI) superior a 0.4. Sin embargo, cabe destacar que la proporción de compuestos con blancos conocidos y una similitud mayor a 0.6 es considerablemente baja, lo que sugiere diferencias significativas entre este conjunto de drogas y otras previamente estudiadas (Figura 3.8).

Como un paso adicional en la validación de nuestro enfoque, se aplicó la metodología desarrollada para predecir los posibles blancos moleculares de estas drogas

con el objetivo de comparar nuestros resultados con otras investigaciones en donde se identificaron blancos potenciales para algunos de estos compuestos.

Evaluación de la metodología desarrollada en comparación con otros estudios publicados respecto al set de compuestos inhibitorios publicados por GlaxoSmithKline

Para evaluar la eficacia de nuestro método, se verificó si hubo coincidencias entre los blancos moleculares determinados utilizando nuestra metodología con los obtenidos experimentalmente o utilizando predicciones en otras investigaciones realizadas sobre el set de compuestos publicados por GSK con actividad inhibitoria probada sobre *Mycobacterium tuberculosis*. Dada la gran cantidad de moléculas publicadas en este estudio, se encontró poca información bibliográfica en la cual se muestran indicios de blancos moleculares relacionados. Dentro de la bibliografía disponible, se obtuvo un estudio *in-silico* (Martínez-Jiménez et al. 2013), en el cual se predijeron blancos moleculares para algunas de las drogas del set, y otros estudios experimentales (Mugumbate et al. 2015; Trofimov et al. 2018) en los que se determinó Dihidrofolato Reductasa (DHFR) como blanco de 30 compuestos del set, junto con otros blancos diversos para 11 compuestos adicionales.

Los resultados obtenidos mostraron ciertas coincidencias relevantes. Como en la metodología desarrollada las proteínas de cada organismo se agrupan por sus dominios, identificar el dominio de unión de un ligando implica, por consiguiente, recuperar la proteína correspondiente entre los resultados. En consecuencia, nuestra evaluación se simplificó a verificar coincidencias con los estudios mencionados en la identificación de los dominios presentes en los blancos moleculares propuestos para cada ligando.

En el caso del estudio *in-silico*, utilizando nuestra metodología se predijeron dentro de los primeros 10 resultados los mismos dominios de unión determinados por Martínez-Jiménez et al. para 19 de los 31 compuestos evaluados (resultados detallados en la Tabla Suplementaria S3.2, Anexo).

Por otra parte también hubo algunas coincidencias con los estudios experimentales expuestos en los trabajos de Mugumbate et al. y de Trofimov et al. Utilizando nuestra metodología, el dominio de DHFR (PF00186) fue identificado como el mejor resultado (obtenido a partir del compuesto con mayor similitud en la base de datos) para nueve de los 30 compuestos que se probó experimentalmente que actúan sobre esta proteína. Este dominio se ubicó entre los 10 principales resultados para otros tres representantes de este set de compuestos y dentro de los 20 mejores candidatos para otros seis ligandos activos sobre DHFR. Para las 11 drogas restantes que inhiben otros blancos moleculares, nuestro

enfoque produjo un resultado acertado solamente en dos casos: uno correspondiente a KasA (Pfam PF00109 y PF02801) y otro a MmpL3 (Pfam PF03176). En ambos casos, el compuesto con mayor similitud en la base de datos condujo al blanco real. Los resultados se ilustran en la Figura 3.9. Detalles adicionales se muestran en la Tabla Suplementaria S3.3, Anexo.

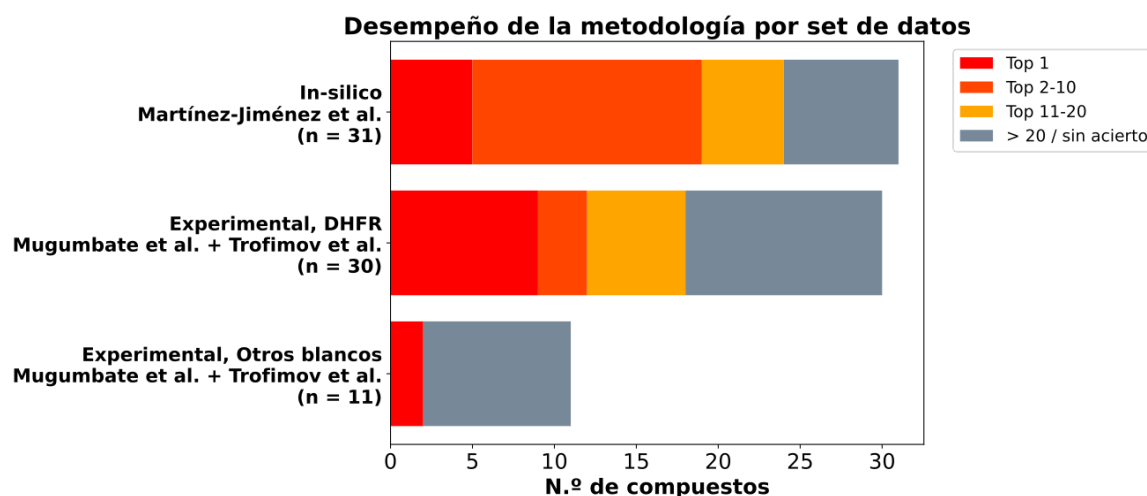


Figura 3.9: Posiciones del ranking en el que fueron encontrados con nuestra metodología los dominios de unión de los compuestos del set de GSK reportados en los diferentes estudios (experimentales e *in-silico*) utilizados como puntos de comparación.

No obstante, no se descarta que varios de los compuestos del set de GSK, que han sido difíciles de caracterizar a lo largo del tiempo, posean mecanismos de acción no relacionados con los blancos actualmente estudiados, lo que escapa al alcance de este método predictivo. Es esperable que exista una menor precisión en la predicción de blancos de drogas cuyos dominios de unión se encuentran poco estudiados. A medida que se realicen más ensayos experimentales y se descubran nuevos mecanismos de acción, se espera que la capacidad predictiva de nuestra metodología continúe mejorando.

La concordancia observada entre nuestra metodología y otros enfoques, especialmente los experimentales, refuerza la validez de nuestro método para predecir dominios de unión para un conjunto de drogas particularmente complejo, el cual, como se observó en nuestros análisis preliminares, se caracteriza por la escasez de drogas similares con mecanismos de acción conocidos.

Aplicación adicional de la metodología desarrollada para determinar el blanco con el que interactúa otro compuesto con mecanismo de acción recientemente descubierto: fabimicina

Por último, aplicamos nuestra estrategia de similitud química a fabimicina, un antibiótico recientemente descubierto cuyo mecanismo de acción también ha sido determinado recientemente (Parker et al. 2022). Este caso es especialmente interesante, ya que fabimicina no se encontraba presente en ChEMBL ni en PDB al momento de este estudio, por lo que resulta una prueba de concepto adecuada para nuestra metodología. fabimicina es activo contra bacterias Gram negativas (incluyendo *Klebsiella pneumoniae*, nuestro caso de estudio para esta última evaluación) y actúa inhibiendo la enzima FabI, perteneciente al dominio Reductasa enoil-ACP (Pfam PF13561). Según nuestros resultados, el ligando más similar a Fabimicina en nuestras bases de datos químicas es CHEMBL1652621 (TI = 0.78), compuesto presente en la base de datos ChEMBL, el cual actúa sobre la proteína FabI de *Staphylococcus aureus* y *Burkholderia pseudomallei*, que al tratarse a una proteína homóloga al blanco de *Klebsiella pneumoniae* también corresponde al dominio PF13561. Por otra parte, el dominio de unión correcto también fue obtenido por medio del compuesto OWE de la base de datos de PDB, pero con un TI de 0.54 respecto de fabimicina, inferior al de CHEMBL1652621. En este ejemplo, la búsqueda por medio de ambas bases de datos de compuestos con dominios de unión conocidos nos permitió recuperar el mecanismo de acción de fabimicina (Figura 3.10), probando que nuestra metodología resultó en la identificación correcta de FabI como uno de los posibles blancos de esta droga activa en *Klebsiella pneumoniae*.

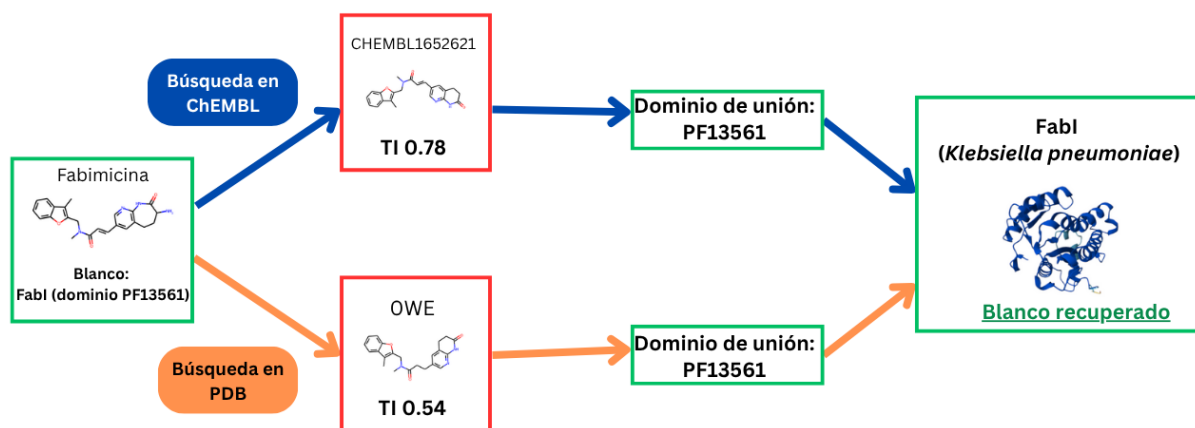


Figura 3.10: Aplicación de nuestra metodología para recuperar el blanco molecular del compuesto fabimicina. En este caso exitoso, el dominio correspondiente a FabI (PF13561) logró ser recuperado

por los compuestos más similares a fabimicina obtenidos tanto de la base de datos ChEMBL como de PDB. En consecuencia se obtiene el blanco FabI de *Klebsiella pneumoniae* entre otras proteínas candidatas que presentan el mismo dominio funcional.

3.3 Conclusiones

Se desarrolló una estrategia novedosa y considerablemente eficiente para la identificación de posibles blancos moleculares de compuestos con probada actividad inhibitoria en organismos patogénicos. Esta metodología, aplicable de manera integral a todo el proteoma de un organismo, ofrece resultados rápidos y con un bajo costo computacional. El proceso consta de tres etapas clave: la comparación por similitud química con compuestos de blanco conocido para identificar dominios candidatos, la selección de los más favorables entre estos mediante *docking* molecular y, finalmente, para aumentar la resolución del método, la aplicación de un filtro basado en la esencialidad de las proteínas resultantes.

En el análisis de los conjuntos de datos, se observó que en más del 60% de los compuestos evaluados fue posible identificar el dominio de unión a partir del compuesto más similar encontrado en la base de datos, independientemente del uso de un umbral de similitud química. Por otra parte, también se adiciona la posibilidad de elegir un corte de similaridad química mínimo para realizar las búsquedas, lo cual permite decidir por una mayor o menor diversidad de resultados.

Esta metodología demostró ser una herramienta eficaz para identificar posibles interacciones moleculares en el proteoma de diversos organismos. Su rapidez y bajo costo computacional, permitiendo generar e incorporar a la plataforma las bases de datos para un organismo a partir de su proteoma en un lapso de entre 6 y 10 horas, y realizar consultas para la obtención de una lista de blancos candidatos para un compuesto en estudio con una espera inferior a los cinco segundos, la hacen adecuada para la identificación preliminar de blancos terapéuticos en compuestos con actividad antimicrobiana. No obstante, sería muy valioso poder disponer de validaciones experimentales adicionales para corroborar los resultados y explorar más a fondo sus aplicaciones.

En la versión actual de la herramienta, además de otros patógenos, se expandió el análisis al proteoma del humano, con el objetivo de ampliar su aplicabilidad y obtener indicios sobre los mecanismos de acción de drogas en estudio. Asimismo, se desarrolló una interfaz gráfica y una versión web de la aplicación, lo que facilita su uso público y amplía su accesibilidad.

La plataforma es totalmente funcional y actualmente se encuentra disponible en GitHub: <https://github.com/gschottlender/ReverseLigQ>, incluyendo todas las bases de datos generadas.

3.4 Métodos específicos del capítulo

3.4.1 Obtención de proteomas de los organismos y anotación

Los proteomas de los organismos seleccionados fueron obtenidos de la base de datos Target Pathogen, reconocida y ampliamente utilizada en el ámbito de la investigación microbiológica. Se decidió el uso de esta base de datos para descargar los proteomas de referencia ya que uno de los objetivos planteados para una etapa futura consiste en integrar la metodología desarrollada en este trabajo con dicha base. De esta manera se evitarán contradicciones.

Para garantizar una identificación precisa de cada proteína, se realizó un análisis adicional utilizando el algoritmo de búsqueda de similitud de secuencias Blastp. Este análisis comparativo se llevó a cabo para cada una de las proteínas presentes en el proteoma descargado de Target Pathogen en relación con las secuencias reportadas en UniProt, base de datos que provee información detallada sobre proteínas de múltiples organismos.

El objetivo de este procedimiento fue asignar a cada proteína un identificador único proveniente de UniProt, lo que facilita su identificación y es útil para el agrupamiento de secuencias redundantes.

3.4.2 Asignación de dominios a las proteínas

La asignación de dominios se llevó a cabo mediante el uso del programa HMMER y la ejecución del comando hmmscan, que escanea las secuencias proteicas contra la base de datos de Pfam. Se utilizó esta herramienta porque es la empleada oficialmente por el equipo de desarrollo de Pfam para construir sus HMMs representativos de dominios, lo que garantiza plena compatibilidad con los umbrales de inclusión de esta base de datos (*gathering cutoffs*) (Finn et al. 2010).

Se seleccionaron aquellos resultados que superaron el *gathering cutoff* —umbral específico determinado por expertos para cada familia de Pfam que define la pertenencia a esa familia— y, adicionalmente, presentaron un *e-value* menor a 1×10^{-5} (indicador de la probabilidad de coincidencia por azar). Este doble filtro combina un criterio definido por expertos con otro estadístico, reduciendo las probabilidades de coincidencias aleatorias y de descartar secuencias pertenecientes a la familia.

3.4.3 Armado de la base de datos interna de proteínas, dominios y ligandos

Para cada organismo estudiado, se creó una base de datos interna compuesta principalmente por dos tablas interconectadas. La primera tabla contiene los dominios Pfam asociados a sus respectivas proteínas, mientras que la segunda incluye los ligandos y sus correspondientes dominios de unión. Esta estructura permite, para cada ligando, obtener información sobre sus dominios de unión probados experimentalmente, y luego identificar las proteínas del organismo en estudio que comparten dichos dominios.

Los datos de interacción entre dominios y ligandos se obtuvieron tanto de PDB como de ChEMBL. PDB proporciona estructuras experimentales en las cuales es posible conocer los contactos entre un ligando y los residuos específicos de su proteína blanco, además de ofrecer mapeos estandarizados a Pfam y UniProt, lo que facilita la identificación precisa del dominio de unión. Por otra parte, ChEMBL aporta información bioquímica curada sobre ensayos experimentales, describiendo la actividad inhibitoria sobre blancos moleculares de los ligandos allí registrados. Ambas bases de datos han sido empleadas en estudios previos de nuestro Grupo de Investigación (Radusky et al. 2017; Serral et al. 2021), lo cual respalda esta elección metodológica con experiencia previa y mayor confianza técnica.

En el caso de PDB, a través de un script específico (`domain_pdb_ligand.py`), se extrajeron los pares ligando-dominio Pfam (es decir que de cada estructura, también se extrajeron los dominios que se encuentran en contacto con cada compuesto), junto con su correspondiente ID de PDB.

Para los ligandos, se aplicó un filtro para retener aquellos biológicamente relevantes, clasificados como drogas, utilizando la base de datos MOAD. Esta base de datos considera compuestos tipo droga como péptidos de 10 o menos aminoácidos, oligonucleótidos de 4 o menos nucleótidos, pequeñas moléculas orgánicas y cofactores, eliminando sustancias irrelevantes como sales, iones y metales.

En el caso de ChEMBL, primero se asignaron los dominios Pfam a todos los blancos de la base de datos (aproximadamente 14,000) utilizando HMMER, siguiendo un procedimiento similar al empleado para agrupar las proteínas de cada organismo. A continuación, se seleccionaron los ligandos que presentan interacciones comprobadas con proteínas de las familias Pfam de cada organismo estudiado, que interactúan con proteínas únicas (excluyendo complejos proteicos) y tienen un valor de `pchembl` superior a 6. Se aclara que el `pchembl` (Bento et al. 2014) representa el logaritmo negativo de la constante de inhibición (K_i) y se utiliza como una métrica estándar para evaluar la capacidad del compuesto de inhibir la actividad de la proteína. El valor de corte de 6 equivale a una K_i de

1 μ M, lo que implica que todos los compuestos seleccionados como activos presentan una constante de inhibición inferior a este valor, asegurando que su actividad biológica es consistente (Brown et al. 2022).

En cuanto a la asignación de dominios de unión para estos ligandos, a diferencia del caso de PDB, los pares ligando-dominio Pfam en ChEMBL no son identificables directamente ya que no se dispone de estructuras experimentales de estos complejos. Adicionalmente, algunos de los blancos de ChEMBL tienen más de un dominio, lo cual dificulta más la tarea de determinar a cuál de ellos corresponde el sitio de unión de cada fármaco.

Sin embargo, es posible realizar aproximaciones para garantizar que, con una probabilidad considerablemente alta, los dominios de unión asignados a cada ligando sean correctos. En primer lugar, estudios previos determinaron que la probabilidad de que el sitio de unión de un ligando bioactivo se encuentre por fuera de un dominio es cercana al 2.5% (Kruger, Rostom, and Overington 2012). Estos resultados sugieren que es razonable la suposición de que si un ligando inhibe un blanco que tiene un único dominio, se puede considerar que el sitio de unión del fármaco corresponda a dicho dominio.

Los casos donde los blancos de un ligando tienen más de un dominio resultan más desafiantes, por lo cual se añade una aproximación adicional: para cada una de estas drogas, se realizó una búsqueda en PDB de moléculas similares con dominios de unión conocidos, utilizando un umbral de Índice de Tanimoto (TI) de 0.4. Todo dominio identificado en PDB que coincidiera con alguno presente en el blanco de ChEMBL se le asignó al ligando como dominio de unión. Esta aproximación se justifica en nuestra evaluación de la recuperación de dominios de unión por medio de ligandos similares en el conjunto de datos de PDB (en el cual se conocen con exactitud los dominios de unión de cada ligando). Nuestros resultados (Figura 3.3 A y B, izquierda) sugieren que para más del 70% de los ligandos evaluados, su dominio de unión es recuperado por medio de un compuesto similar usando un corte de TI de 0.4. Con este umbral se obtienen, en promedio, unos 20 dominios candidatos por compuesto. Sin embargo, dado que en la búsqueda se consideran únicamente los dominios anotados en el blanco de ChEMBL, la probabilidad de incorporar un dominio no relacionado por mera casualidad resulta muy baja, especialmente teniendo en cuenta que la cantidad de dominios de unión reportados en ChEMBL (308 por organismo, en promedio) es mucho menor que la registrada en la PDB (promedio de 1052 por organismo). Asimismo, los ligandos sin coincidencias en la búsqueda fueron descartados.

Por último, cabe señalar que, dado que el objetivo de esta etapa es la obtención de una **lista preliminar de blancos candidatos** que luego se afinará con otros criterios, se

optó por **mantener cierta permisividad ante algunos resultados inciertos para no descartar información potencialmente valiosa**.

3.4.4 Similaridad química

La similaridad química entre compuestos se evaluó mediante el Índice de Tanimoto (TI), una medida que cuantifica la similitud entre dos moléculas. Para calcular el TI, se utilizó el SMILES (Simplified Molecular Input Line Entry System) de cada compuesto, que fue convertido en *Fingerprints* Moleculares de Morgan de 1024 bits y radio 2 (valores estándar en quimioinformática) mediante el programa RDKit. Estos *Fingerprints* son secuencias de dígitos binarios que representan la presencia o ausencia de subestructuras químicas en los ligandos. Esta representación ha demostrado ser una de las más populares y eficaces en *screening* virtual y análisis de similitud química de fármacos (H. Zhou and Skolnick 2024), especialmente por su capacidad para capturar patrones estructurales relevantes asociados con bioactividad. Además, su accesibilidad a través de RDKit facilita considerablemente su implementación práctica e integración directa en las metodologías desarrolladas.

El cálculo del TI se realiza mediante la Fórmula 2.1, detallada en la sección 2.1.3 del capítulo de métodos generales.

3.4.5 Caracterización de los blancos candidatos

La información metabólica de las proteínas de cada organismo fue obtenida de Target Pathogen. Desde el punto de vista metabólico, la centralidad de una reacción mide la frecuencia con la que dicha reacción actúa como intermediaria en el flujo metabólico dentro de la red. Se supone que bloquear proteínas asociadas a reacciones con alta centralidad puede generar desequilibrios en diversas vías metabólicas. Por otro lado, los puntos de estrangulamiento (*choke-points*) son reacciones que producen o consumen un único producto o sustrato, y bloquear estas proteínas podría afectar críticamente el equilibrio de compuestos esenciales o generar acumulación de metabolitos tóxicos en la célula del patógeno. En este contexto, las proteínas asociadas con puntos de estrangulamiento se consideran relevantes para el descubrimiento de drogas inhibitorias.

La evaluación de la esencialidad de las proteínas se basó en datos experimentales recopilados de Target Pathogen, que identifican qué proteínas son fundamentales para la supervivencia del organismo. Dado que la información acerca de estos ensayos experimentales solo se encuentra disponible para el caso de *Mycobacterium tuberculosis*, para los demás organismos se empleó un enfoque predictivo. Este consistió en realizar un BLAST del proteoma de cada organismo contra la base de datos de genes esenciales

(DEG), utilizando criterios estrictos de selección para garantizar homologías funcionales claras, específicamente una cobertura superior al 80% y un *e-value* menor a 1×10^{-10} . Este proceso nos permitió inferir la esencialidad de las proteínas en organismos donde no se realizaron ensayos al respecto.

3.4.6 *Docking* molecular

Primero, se obtuvieron los 10 ligandos más similares a cada ligando evaluado, utilizando el índice de Tanimoto para cuantificar la similitud química, de acuerdo al algoritmo de aprendizaje automático no supervisado explicado previamente.

Una vez seleccionados los ligandos, se realizaron los ensayos de *docking* molecular utilizando las estructuras de PDB de las proteínas cocrystalizadas con dichos ligandos. Cada proteína representa un dominio Pfam diferente, tal como se mencionó en la sección de resultados. Las posiciones de estos ligandos en las estructuras de PDB se emplearon también como centros de grilla durante el proceso de *docking*. Este fue llevado a cabo con el programa AutoDockTools (versión 4.2.6), una herramienta de acceso libre reconocida por la comunidad científica y empleada de manera protocolar por nuestro Grupo de Investigación (Gauto et al. 2013) con una experiencia notable, la cual incluye el desarrollo de modificaciones personalizadas sobre el código original para ajustar los procesos a necesidades específicas (Clemente, Prieto, and Martí 2024; Arcon et al. 2019). AutoDockTools, por otra parte, ofrece prestaciones comparables a las de programas comerciales como GOLD y Glide, pero sin los elevados costos asociados a licencias pagas (Z. Wang et al. 2016).

Dado que se buscó un tratamiento equitativo para una gran diversidad de proteínas y ligandos, se utilizaron los parámetros *default* predefinidos por los desarrolladores de la herramienta: para la configuración de la cuadrícula de búsqueda, se estableció un tamaño de $15 \times 15 \times 15$ Å con un espaciado de 0.375 Å. Se llevaron a cabo 100 corridas del algoritmo genético (ga_run) con un umbral de tolerancia de 2 Å para el RMSD dentro de un mismo cluster de poses resultantes (rmstol).

3.4.7 *Docking* sesgado

La identificación de interacciones clave entre la proteína y drogas permitió incorporar restricciones en base a información conocida al *docking* molecular (cómo se explicó en la sección de métodos). Para ello se realizó una inspección visual detallada de los puentes de hidrógeno y/o interacciones aromáticas, presentes en cada estructura de PDB, utilizando el programa VMD ("VMD: Visual Molecular Dynamics" 1996), lo que

permitió una evaluación precisa de cómo los ligandos interactúan con cada proteína en cada caso.

El proceso de construcción de grillas para las corridas de *docking* molecular incluyó la aplicación de sesgo específico basado en las interacciones seleccionadas. Este sesgo fue implementado utilizando AutoDockBias con parámetros *default*, asegurando que las regiones donde se producen interacciones clave fueran especialmente consideradas durante la exploración de conformaciones ligando-proteína realizada durante el proceso de *docking* molecular.

En todas las corridas de *docking* molecular, se mantuvieron constantes los parámetros previamente mencionados en la sección 3.4.6.

3.4.8 Desarrollo de la interfaz gráfica

Partiendo del script utilizado para generar predicciones de blancos moleculares para un compuesto “*query*” por medio del método desarrollado de aprendizaje automático no supervisado, se desarrolló una interfaz gráfica.

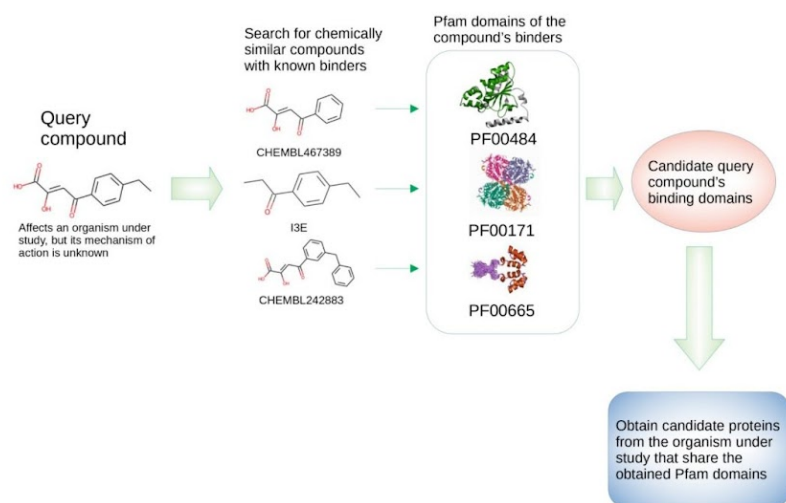
Esta interfaz gráfica permite seleccionar el organismo para la búsqueda de blancos entre una lista de 13 organismos, entre los cuales se incluyen bacterias (las especies estudiadas en detalle a lo largo del trabajo y otras adicionales), dos virus (Dengue y SARS-CoV-2) y como se ha mencionado anteriormente, también *Homo sapiens*, ya que la metodología también puede ser funcional para investigar el mecanismo de acción de ciertas drogas.

El desarrollo fue realizado mediante Streamlit en unas pocas líneas de código (como se mencionó en la sección de métodos, esta librería facilita en gran medida el desarrollo *front-end* para herramientas de ciencia de datos) que permiten a través de la interfaz gráfica correr los comandos del script original y mostrar su salida.

La Figura 3.11 muestra la pantalla correspondiente a la interfaz gráfica generada para el uso público de este algoritmo.

Reverse LigQ

Unsupervised learning tool for Compound Target Prediction based on chemical similarity



Select Organism

Bartonella bacilliformis

Enter compound in SMILES format

CCCCCOCOCOCO

Define the Tanimoto Index threshold

0.40

- +

Citation:

Schottlender, G., Prieto, J. M., Palumbo, M. C., Castello, F. A., Serral, F., Sosa, E. J., et al. (2022). From drugs to targets: reverse engineering the virtual screening process on a proteomic scale. Front. Drug Discov. 2:969983. doi: 10.3389/fddsv.2022.969983

Run Similarity Search

Figura 3.11: Captura de la interfaz gráfica elaborada para utilizar el algoritmo, desarrollada con Streamlit.

4. Descubrimiento de sustratos, productos e inhibidores de Citocromos P450 bacterianos: un enfoque integrado utilizando filogenia, bioinformática estructural y contexto genético

4.1 Introducción e hipótesis

En el capítulo 3 se desarrolló una herramienta diseñada para predecir los blancos de compuestos inhibidores con actividad antimicrobiana comprobada, cuyo mecanismo de acción se requiere investigar. Una de las limitaciones de esta estrategia es que, generalmente, el resultado final consiste en una lista de proteínas que comparten dominios funcionales, que suele ser amplia como para realizar evaluaciones experimentales directas. Para refinar estos resultados y priorizar con mayor precisión las interacciones proteína-ligando, y avanzar posteriormente con los ensayos experimentales, se requieren abordajes adicionales más específicos. En esta segunda parte de la investigación se propone un flujo de trabajo que combina diferentes enfoques, tanto a nivel estructural como derivados del análisis de secuencias, y que es aplicable a diversas familias de proteínas. Para su desarrollo, se seleccionó como caso de estudio la familia de los Citocromos P450 bacterianos, que incluye proteínas relevantes no sólo como posibles blancos de antibióticos, sino también como catalizadores de reacciones que generan productos con potencial biotecnológico.

Los Citocromos P450 bacterianos (BacCYPs) son proteínas hemoproteicas que, al igual que sus homólogos eucariotas, desempeñan un papel clave en la catalización de reacciones de oxidación sobre una amplia gama de sustratos químicos complejos. Estas reacciones dependen de la disponibilidad de oxígeno molecular y requieren la transferencia de dos electrones, lo que las convierte en procesos altamente especializados. Sin embargo, a diferencia de los Citocromos P450 eucariotas, que suelen estar anclados en membranas, los BacCYPs son solubles, lo que les confiere una mayor versatilidad en cuanto a su participación en diversas rutas metabólicas. Esta diferencia estructural y funcional les permite involucrarse en múltiples procesos bioquímicos, lo que contribuye significativamente a la síntesis de una amplia variedad de productos naturales de interés biotecnológico (Greule et al. 2018).

Una de las características más destacadas de los BacCYPs es su capacidad para realizar diversas reacciones de oxidación, lo que les permite actuar sobre una extensa

gama de sustratos. Este amplio espectro de acción no sólo es relevante desde un punto de vista biológico, sino que también tiene implicaciones directas para aplicaciones biotecnológicas. La comprensión detallada del papel de los BacCYPs en la biosíntesis de estos metabolitos es esencial para explotar su potencial, que abarca desde el desarrollo de nuevos productos farmacéuticos hasta la producción de biocombustibles y productos químicos industriales (Kelly and Kelly 2013).

En los últimos años, el aumento exponencial en la secuenciación de diversos microorganismos ha permitido la anotación de más de 100,000 secuencias de BacCYPs. A pesar de este avance, la mayoría de las anotaciones se limitan a clasificaciones generales dentro de familias de proteínas, lo que deja un vacío significativo en cuanto a la caracterización funcional de estos Citocromos. El desafío actual radica en predecir *in-silico* el rango de sustratos y las reacciones específicas que cada una de estas enzimas puede catalizar, una tarea que, aunque compleja, es crucial para avanzar en el entendimiento de su función biológica. A pesar del progreso en la secuenciación, la caracterización funcional de los BacCYPs sigue estando rezagada, lo que genera una brecha importante en nuestro conocimiento de sus capacidades catalíticas y especificidades de sustrato (Khmelevtsova et al. 2017).

Esta brecha es especialmente relevante cuando se busca entender cómo los BacCYPs interactúan con sustratos a nivel molecular y cómo estas interacciones conducen a la catálisis de reacciones específicas. El reciente desarrollo de técnicas de biología estructural de alto rendimiento, junto con avances en métodos computacionales, ha abierto nuevas oportunidades para superar estas limitaciones (Darabi, Seddigh, and Abarshahr 2017; Sirim et al. 2010; Bustamante et al. 2016). Estas tecnologías permiten el estudio de interacciones proteína-ligando a un nivel sin precedentes, proporcionando información detallada sobre las relaciones estructura-función.

En esta segunda parte de nuestra investigación, se propone un flujo de trabajo basado en la hipótesis de que la combinación de análisis evolutivo, la exploración del contexto genético, y la evaluación de la estabilidad de las interacciones proteína-ligando mediante *docking* molecular permite predecir los posibles sustratos, productos e inhibidores de Citocromos P450 bacterianos (BacCYPs) partiendo de su secuencia de aminoácidos. De esta manera se constituye una metodología aplicable potencialmente a otras familias de proteínas.

4.2 Resultados

Los resultados se presentan de la siguiente manera: inicialmente, se ofrece un análisis actualizado de la diversidad de secuencias y estructuras de BacCYPs, junto con una evaluación de la calidad y aplicabilidad de los modelos estructurales basados en homología y generados con AlphaFold. Luego, se desarrolla el esquema predictivo:

Primero, se analiza la diversidad de sustratos, productos e inhibidores (SPIs) que se unen a esta familia de proteínas en un contexto filogenético, esto con el objetivo de inferir nuevas interacciones a partir de relaciones evolutivas entre las proteínas de la familia en estudio.

En segundo lugar, se implementa el análisis del contexto genético para predecir posibles SPIs que puedan unirse a BacCYPs.

En tercer lugar, se examina la capacidad del *docking* molecular para evaluar la viabilidad de los complejos BacCYP-SPIs, diferenciando los compuestos que se unen a BacCYPs de otros ligandos.

Finalmente, se desarrolla y evalúa un esquema integrador de predicción de SPIs de BacCYPs, considerando su potencial aplicabilidad en función de la información disponible.

4.2.1 Análisis de la diversidad de secuencias de Citocromos P450 bacterianos.

El análisis inicia examinando la diversidad total de secuencias de BacCYPs, realizando un Alineamiento Múltiple de Secuencias (MSA) que contiene 1809 secuencias definidas como *conjunto 1*, representativas de todo el universo de secuencias de BacCYPs conocidos. Este conjunto incluye 202 estructuras cristalizadas. Teniendo en cuenta únicamente las columnas informativas del alineamiento múltiple (proceso detallado en métodos), se obtuvo un árbol filogenético completo (Figura 4.1 A). Partiendo de este último, se clasificaron los BacCYPs en 14 grupos monofiléticos (grupos filogenéticos) diferentes (etiquetados de la A a la N), 11 de los cuales contienen al menos una estructura cristalizada representativa. Por otra parte también se obtuvo el Modelo Oculto de Markov (HMM) correspondiente a cada uno de los grupos filogenéticos.

El MSA y los HMMs evidencian aquellos motivos de secuencia que están conservados en todos los BacCYPs (ejemplos representativos se muestran en la Figura 4.1 B y resultados detallados en la Figura 4.2), así como los sitios específicos de cada grupo filogenético (Tabla 4.3).

La longitud del dominio BacCyp450 presenta una gran variación, con un promedio de 403 residuos. Dentro de este dominio, se destacan 33 posiciones conservadas que

aportan un alto contenido informativo. Entre ellas se encuentran: i) el motivo de unión al hemo (posiciones 364-373) cuya secuencia de aminoácidos tiene el patrón F-[G/S]-X-G-X-[H/R]-X-C-X-G, donde la glicina de la última posición se sustituye por alanina en el grupo filogenético N; ii) el motivo EXXR en las posiciones 292-295; iii) el "Proton Relay Motif" en las posiciones 236-241, con la secuencia de aminoácidos [A/G]-GX-[D/E]-T-[T/S] (Werck-Reichhart and Feyereisen 2000); y iv) las posiciones 332-335 que presentan el motivo [N/H]-R-D-P, el cual distingue a los grupos filogenéticos A, B, C, D, E, F, M y N, que presentan principalmente una histidina, de los demás, que muestran principalmente una asparagina.

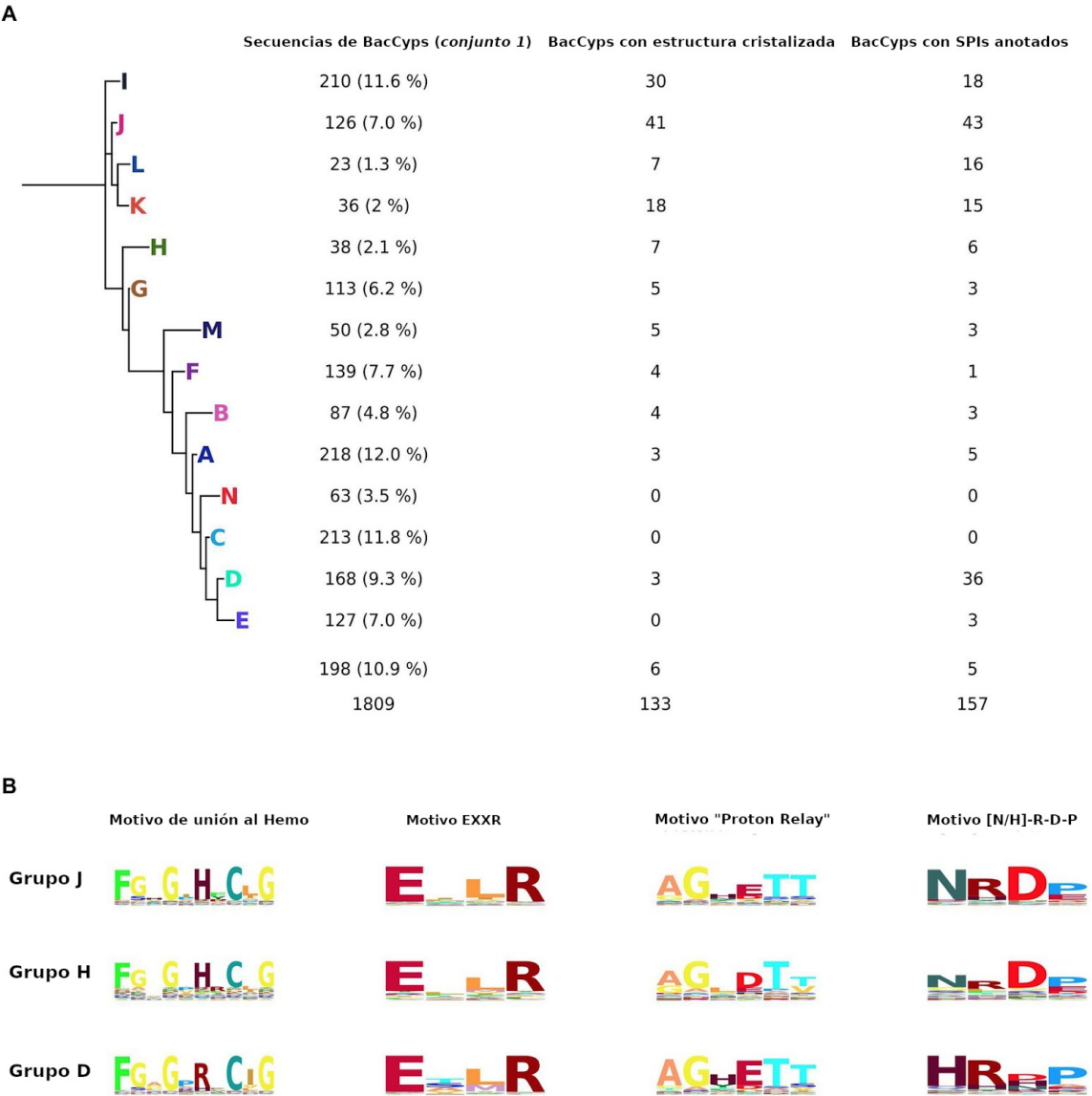


Figura 4.1: A) Árbol filogenético de los principales grupos de BacCYPs, mostrando el número de proteínas con estructuras cristalizadas y SPIs anotados. B) Logos HMM de los grupos filogenéticos

J, H y D, destacando variaciones clave de la secuencias de aminoácidos en los motivos de unión al hemo, EXXR, "Proton Relay Motif" y [N/H]-R-D-P.



Figura 4.2: Comparativa de los principales sitios conservados en cada grupo filogenético de BacCYPs.

El sitio activo de BacCYP (consideramos que todos los aminoácidos de la cadena que están involucrados en interacciones con ligandos en cristales de PDB corresponden al sitio activo) contiene aproximadamente 24 residuos, como se observa en la Figura 4.3 A. Entre estos residuos, 5 de ellos (representados como bastones magenta) muestran un notable nivel de conservación. Específicamente, la posición 57 está ocupada principalmente por fenilalanina (F) o tirosina (Y). La posición 320 está ocupada principalmente por alanina (A) o lisina (K). Además, dentro de estos residuos altamente conservados, se encuentran las posiciones 236 y 240, pertenecientes al “*Proton Relay Motif*”, que generalmente están ocupadas por alanina (A) y treonina (T) respectivamente. Por último, la posición 334, incluida en el motivo [N/H]RDP, comúnmente está ocupada por ácido aspártico (D).

Es notable que, en algunas estructuras de PDB, ciertos aminoácidos, en posiciones altamente conservadas que típicamente participan en interacciones con ligandos (representados como esferas naranjas en la Figura 4.3 A), son reemplazados por aminoácidos menos comunes cuando se observa dicha interacción. Por ejemplo, en la posición 179, que usualmente es ocupada por alanina (A), se puede encontrar al ácido aspártico (D); en la posición 239 (que es parte del “*Proton Relay Motif*”), la cual normalmente contiene ácido glutámico (E) o ácido aspártico (D), aparece histidina (H); y en la posición 335, típicamente ocupada por prolina (P) en el motivo [N/H]RDP, se observa también ácido aspártico (D). Los 16 residuos restantes (representados como esferas verdes en la Figura 4.3 A) en el sitio activo muestran una alta variabilidad y no están relacionados con sitios conservados. Estas observaciones concuerdan con la versatilidad de los BacCYPs, permitiéndoles interactuar con una amplia gama de compuestos químicos. Como ejemplo ilustrativo, en la posición 232, el BacCYP GcoA de *Amycolatopsis sp.* (cepa ATCC 39116 / 75iv2) presenta una valina (V), que en su estructura representativa de PDB se encuentra interactuando con el ligando guaiacol (PDB ID JZ3) (Figura 4.3 B). Mientras que en contraste, en el Cyp7863 de *Streptomyces peucetius*, esta posición corresponde a una treonina (T), que establece una unión de puente hidrógeno con un compuesto significativamente diferente, bisphenol A (PDB ID 2OH) (Figura 4.3 C).

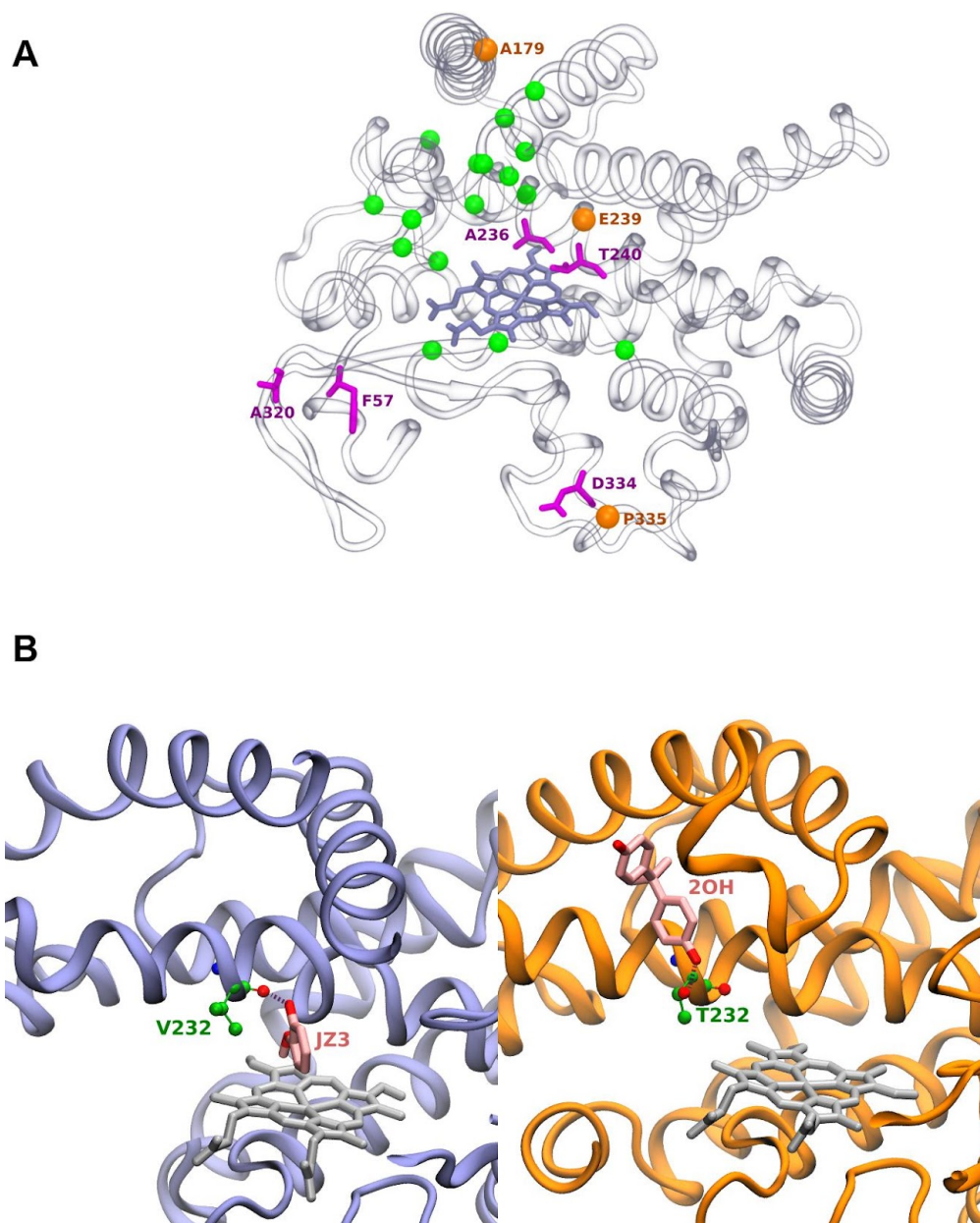


Figura 4.3: **A)** Estructura general de los BacCYPs resaltando los residuos del sitio activo. Los residuos representados en magenta representan posiciones altamente conservadas que no se encontraron modificadas en sus interacciones con ligandos en estructuras de PDB. Las esferas naranjas indican posiciones conservadas en las que se observaron aminoácidos menos comunes en las interacciones con ligandos analizadas. Las esferas verdes corresponden a los residuos restantes del sitio activo en posiciones no conservadas. **B)** Estructura de GcoA de *Amycolatopsis* sp. (cepa ATCC 39116 / 75iv2) unida a guaiacol (PDB ID JZ3), destacando su interacción con la cadena principal de V232. **C)** Estructura de Cyp7863 de *Streptomyces peucetius* unida a bisfenol A (PDB ID 2OH), resaltando su interacción con la cadena lateral de T232.

Posteriormente, se analizaron las posiciones conservadas distintivas de cada grupo filogenético (Tabla 4.4), las cuales no se encuentran en la totalidad de los BacCYPs. Se observa que la mayoría de estas posiciones no están relacionadas con las detalladas en la Figura 4.3 A, las cuales, mediante análisis de interacciones con ligandos, se determinó que corresponden al sitio activo de la proteína. Estos resultados refuerzan la hipótesis de que los sitios de interacción proteína-ligando específicos para cada BacCYP son muy específicos de cada proteína, y difíciles de predecir conociendo únicamente su secuencia.

Grupo filogenético de BacCYPs	Posiciones conservadas distintivas
A	-
B	L65, W171, E360
C	P121, T168
D	-
E	-
F	-
G	A128
H	W155
I	G116, A128, D149, W155
J	D121, A128, P144, W155
K	P131, L132, P133, I137, P144
L	P131, L132, P133, I137, P144, R149, W155, S156
M	P162
N	-

Tabla 4.4: Aminoácidos conservados que son grupo-específicos y sus posiciones respectivas en el dominio BacCYP.

4.2.2 Análisis de la diversidad de estructuras y evaluación del modelado estructural para Citocromos P450 bacterianos

Análisis de la diversidad de estructuras de Citocromos P450 bacterianos

En cuanto a la cobertura del genoma respecto a la disponibilidad de estructuras cristalizadas de referencia, el mapeo de las secuencias de BacCYPs con estructura disponible en el árbol filogenético, revela que la mayoría de los grupos filogenéticos principales cuentan con al menos una estructura cristalizada representativa (Figura 4.1 A).

La media del RMSD entre carbonos alfa (RMSD-C α) entre estructuras dentro de un mismo grupo filogenético, es de aproximadamente 1.51 Å (el rango varía entre 0.262 Å y 2.703 Å). En cambio, al comparar estructuras de diferentes grupos, el RMSD-C α promedio aumenta a 2.52 Å, con valores que oscilan entre 1.517 Å y 4.579 Å. Esto evidencia una diversidad estructural significativa, incluso dentro de un mismo grupo filogenético.

Evaluación del modelado estructural

Para evaluar la eficacia de la predicción estructural mediante modelado por homología, se analizó la calidad de los modelos en función de la distancia filogenética con las proteínas molde utilizadas. Se seleccionaron 12 estructuras de proteínas como casos de prueba y, para cada una de ellas, se generaron nueve modelos: tres con un molde de alta identidad (correspondientes a las tres proteínas con estructura más emparentadas dentro del árbol filogenético), tres con identidad moderada (las proteínas con estructura más distantes evolutivamente que corresponden al mismo grupo filogenético, o en caso de que no haya, proteínas con estructura del grupo filogenético más cercano) y tres con baja identidad (proteínas con estructura pertenecientes a grupos filogenéticos distantes). Este proceso se describe detalladamente en la sección de métodos (4.4.5). Los modelos resultantes se evaluaron realizando una comparación con la estructura real correspondiente mediante el RMSD-C α .

Por otra parte, complementando el análisis del modelado por homología, se generaron 20 modelos de estructuras de proteínas utilizando AlphaFold, una herramienta novedosa y ampliamente utilizada en la época del desarrollo de este trabajo. Para evaluar este método, se seleccionaron cuidadosamente proteínas cuya estructura cristalizada fue depositada en el PDB posteriormente a la fecha de corte oficial empleada para el entrenamiento de los modelos de esta plataforma (30 de abril de 2018) (https://github.com/google-deepmind/alphafold/blob/main/docs/technical_note_v2.3.0.md). Los resultados, presentados en la Figura 4.5 A, muestran que, como era de esperarse, en los modelos construidos con proteínas molde de alta identidad, más del 90% presenta un

RMSD-C α menor a 2 Å respecto a la estructura real, indicando una excelente predicción (Eramian et al. 2008). Para los moldes de identidad moderada, la precisión de los modelos disminuye, resultando en un RMSD-C α promedio de alrededor de 2.5 Å en comparación con las estructuras cristalizadas. Los modelos generados con moldes de baja identidad presentan valores de RMSD-C α más bajos, que probablemente no son lo suficientemente confiables para realizar análisis posteriores. La conclusión más destacable es que los modelos de AlphaFold son notablemente superiores a los generados mediante modelado por homología, incluso superando, aunque en menor medida, a aquellos basados en moldes de alta identidad.

Los resultados de los modelos de AlphaFold y los correspondientes a modelado por homología con moldes de alta identidad demuestran que estas predicciones son lo suficientemente confiables para realizar análisis adicionales, como por ejemplo cálculos de *docking* y dinámicas moleculares.

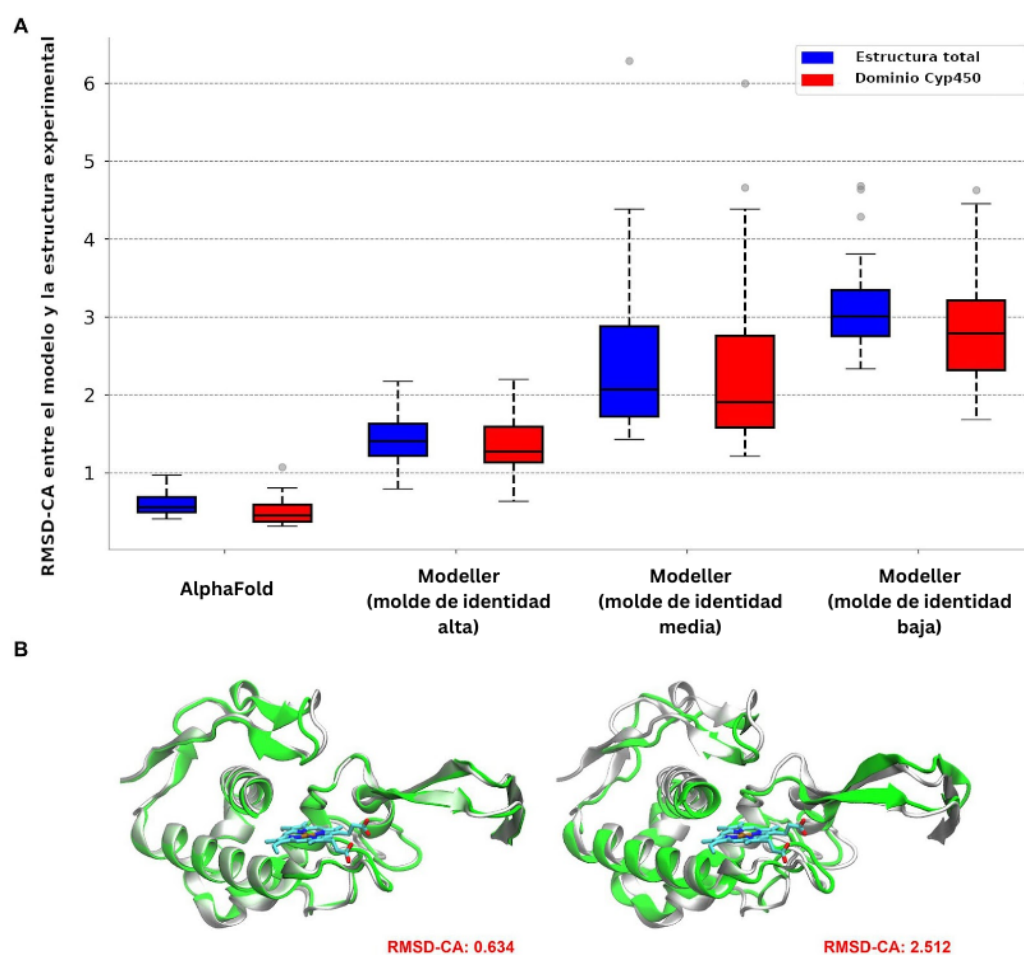


Figura 4.5: A) Comparación de calidad entre estructuras de referencia y modeladas utilizando AlphaFold y Modeller. Se proporcionan valores de RMSD-C α tanto para la proteína completa como

para la región específica correspondiente al dominio BacCYP. **B)** Representación visual de la comparación entre los modelos estructurales realizados por AlphaFold (izquierda) y por homología con proteínas molde de identidad media (derecha) en comparación con la estructura real.

4.2.3 Análisis de la diversidad de sustratos, productos e inhibidores de Citocromos P450 bacterianos

Con el fin de analizar la diversidad de sustratos, productos e inhibidores (SPIs) de BacCYPs, se agruparon estos compuestos utilizando el Índice de Tanimoto (TI) como métrica de similitud. A partir de este agrupamiento, se generó un dendrograma que refleja la organización obtenida -como se muestra en la Figura 4.6 A- quedando definidos diez grupos de SPIs principales, cuatro de ellos con subgrupos (metodología detallada en la sección 4.4.4). Los grupos se muestran en detalle en la tabla suplementaria S4.1 del Anexo.

Entre los diez grupos principales de SPIs, se identifican categorías bien definidas, como compuestos similares a esteroides (grupo 4), ácidos grasos (grupo 1), alcanfor y moléculas relacionadas (grupo 3), y macrólidos (grupo 2). En contraste con los grupos de SPIs 5, 6, 7 y 9, que presentan estructuras más diversas, con variedad de anillos aromáticos y diferentes tipos de átomos pesados.

Por otro lado, se investigó más detalladamente la diversidad de estos ligandos, analizando tanto las variaciones dentro de cada grupo de SPIs como entre compuestos de diferentes grupos. Para ello, se calculó el Índice de TI entre todos los pares de ligandos. En cuanto a la variabilidad dentro de un mismo grupo, el análisis reveló tres subconjuntos distintos (Figura 4.6 B, a la izquierda). El primer subconjunto agrupa pares de compuestos con alta similitud ($TI > 0.6$), indicando una estructura molecular casi idéntica. El segundo subconjunto contiene ligandos de similitud moderada, lo que sugiere una diversidad estructural intermedia. Finalmente, el tercer y más amplio subconjunto incluye pares de moléculas con diferencias estructurales significativas (TI entre 0.20 y 0.35), reflejando así la diversidad química inherente a estos compuestos que se unen a los BacCYPs.

En cambio, al analizar la similitud química entre SPIs de diferentes grupos (Figura 4.6 B, a la derecha), se observa en general una baja identidad entre los pares de ligandos, presentando valores de TI que varían mayormente entre 0.0 y 0.3. Predominan los pares de compuestos con un TI menor a 0.2, lo cual refleja diferencias significativas en la estructura química de estos ligandos.

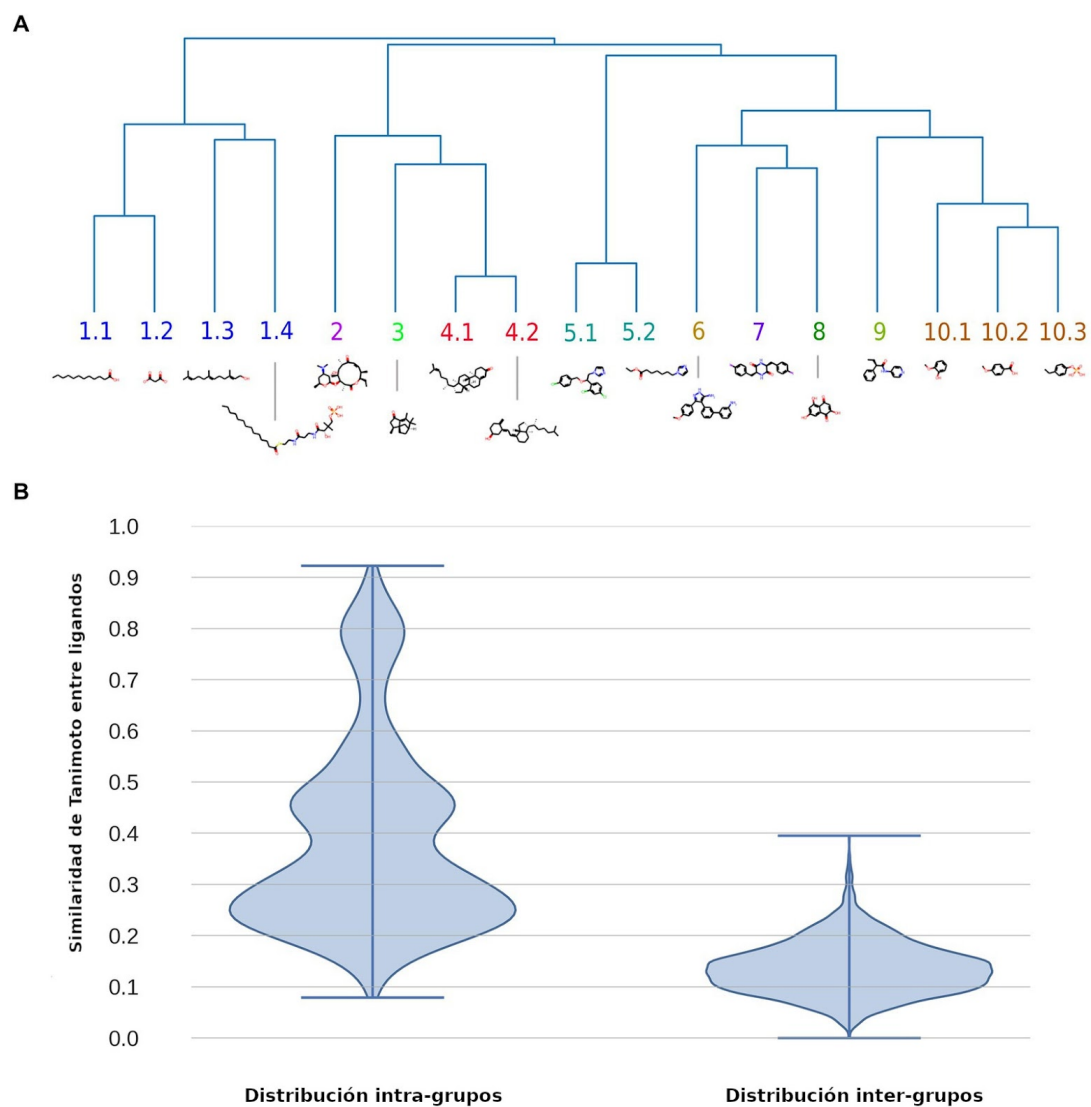


Figura 4.6: **A)** Dendrograma que muestra los 10 grupos principales de SPIs de BacCYPs y sus respectivos subgrupos, con un ejemplo ilustrativo de cada uno para observar las relaciones de similitud fisicoquímica. **B)** Distribuciones del Índice de Tanimoto entre todos los pares de ligandos dentro de un mismo grupo de SPIs (distribución intra-grupo) y entre compuestos de diferentes grupos (distribución inter-grupo).

4.2.4 Enfoque Predictivo 1: Correlación Filogenética de los Compuestos que unen BacCYPs

Como primera alternativa para predecir qué compuestos pueden unirse a un BacCYP específico, se propone obtener esta información a partir de las relaciones evolutivas con proteínas de la familia que disponen de sustratos, productos e inhibidores (SPIs) conocidos.

Inicialmente, como se mencionó previamente, y se esquematiza en la Figura 4.1 A, se realizó un mapeo de todos SPIs conocidos de proteínas de esta familia dentro el árbol filogenético más abarcativo, que es representativo de todas las secuencias conocidas de BacCYPs.

Como se observa en la Figura 4.7, la distribución de las similitudes entre SPIs correspondientes a un mismo grupo filogenético (representado en color rojo), es muy similar a la distribución general de todos los SPIs conocidos que unen BacCYPs (en azul). La sorprendente similitud entre ambas distribuciones sugiere que un análisis filogenético tan amplio, que contempla todo el universo de BacCYPs, por sí solo no proporciona información suficientemente consistente para predecir la especificidad de los SPIs hacia proteínas de un grupo filogenético particular.

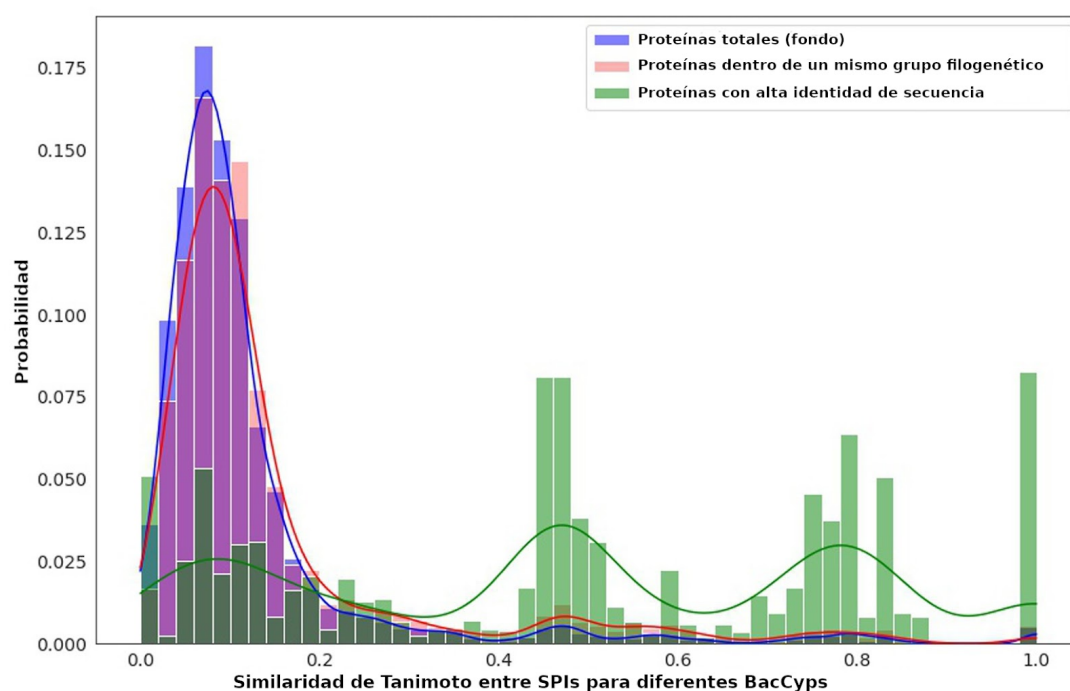


Figura 4.7: Diagramas de barras que representan la similitud mediante el Índice de Tanimoto entre SPIs que se unen a diferentes subconjuntos de proteínas: en azul, a todos los BacCYPs anotados,

en rojo a BacCYPs pertenecientes a un mismo grupo filogenético, y en verde, a BacCYPs altamente similares, que comparten más de un 70% de identidad de secuencia en BLAST.

Sin embargo, al observar los SPIs de los BacCYPs con una identidad de secuencia superior al 70% (Figura 4.7, en verde), se destacan picos distintivos en los valores de TI de 0.5, 0.8 y cercanos a 1. Estos resultados sugieren que, si bien no resulta sencillo inferir el sustrato de un BacCYP únicamente a partir de su grupo filogenético en el árbol más abarcativo, la obtención de BacCYPs con una alta identidad de secuencia que tienen SPIs conocidos, pueden proporcionar información valiosa sobre los SPIs de nuevas proteínas.

Para reforzar la idea de que un análisis menos abarcativo puede hacer más eficiente el enfoque basado en relaciones evolutivas, se construyó un árbol filogenético que incluye únicamente las proteínas con SPIs conocidos. Este árbol, que representa un subconjunto mucho más reducido en comparación con el evaluado anteriormente, es considerablemente más fácil de interpretar. Esto se debe a que, como se determinó en análisis previos y se muestra en la Figura 4.1 A, el conocimiento sobre los SPIs está mayormente concentrado en ciertos grupos filogenéticos, tales como el grupo I, el grupo J y el grupo D. Por lo tanto, este árbol refleja una representación mayoritaria de estos grupos, con agrupaciones mucho más específicas, que representan la porción más estudiada del universo de BacCYPs.

En la Figura 4.8 se muestra este árbol con los SPIs mapeados en cada proteína. Se puede observar que, en muchos casos, las proteínas cercanas evolutivamente están asociadas con los mismos grupos de SPIs, lo que respalda la utilidad de este enfoque para inferir interacciones potenciales basadas en la proximidad filogenética. Como ejemplos ilustrativos, esta tendencia se aprecia en el grupo filogenético que contiene a las proteínas con ID de Uniprot P9WPP9 y P0A513 entre otros representantes (en rojo), que se asocian mayoritariamente con el grupo 4.1 de SPIs, y en el grupo filogenético que incluye Q93H81 y A0A286SBY7 (en marrón), que tiende a unirse al grupo 2 de SPIs. Los grupos de SPIs fueron detallados en la Figura 4.6 A.

4.2.5 Enfoque predictivo 2: Búsqueda mediante Contexto Genético

En las bacterias, muchas vías metabólicas relacionadas con diversos compuestos suelen estar organizadas en operones, donde es común, entre otros ejemplos, que el producto de la proteína codificada por un gen determinado funcione como sustrato de la proteína codificada por un gen cercano posicionalmente en el genoma. Este principio sugiere que los genes vecinos podrían ser útiles para inferir posibles SPIs de un BacCYP en estudio.

Con este objetivo, se analizó la información disponible sobre los sustratos y/o productos (SPs) de todos los BacCYPs conocidos y sus genes vecinos en la base de datos KEGG. Como resultado, se obtuvo que para aproximadamente 4,815 BacCYPs se disponía de información sobre los SPs de al menos uno de sus genes vecinos; de estos BacCYPs, solamente para 120 se conocían los SPs verdaderos. Este conjunto es el que se utilizó para las evaluaciones de este método, ya que se dispone de SPs conocidos para validar la metodología.

Para analizar la relación entre los SPs de los 120 BacCYPs con SPs conocidos y los SPs de sus genes vecinos, se calculó el índice de similitud química entre ellos. Con este enfoque, se identificaron SPs candidatos a partir de los SPs de genes adyacentes en 74 de los casos evaluados, de los cuales en 59 de ellos, uno de los SPs recuperados corresponde exactamente al sustrato o producto del BacCYP. Considerando también los casos en los que no se obtuvo el compuesto exacto, se encontró algún SP con una similitud química significativa (Índice de Tanimoto superior a 0.4) respecto al SP real para un total de 61 BacCYPs.

Al ampliar la búsqueda a genes ubicados hasta 2,000 pares de bases (pb) de distancia del gen que codifica cada BacCYP, la capacidad de identificar SPs candidatos aumentó a 96 BacCYPs con SPs conocidos. Dentro de esta cobertura ampliada, en 75 casos dentro de los SPs recuperados se encontró el SP exacto del BacCYP, y en 77 casos se obtuvo algún ligando con una similitud química considerable ($TI > 0.4$) respecto al SP verdadero (Figura 4.9). No obstante, esta ampliación del rango de búsqueda también incrementó el número promedio de compuestos candidatos: de 2.72 compuestos al considerar sólo genes adyacentes a 3.33 compuestos por BacCYP en el rango extendido.

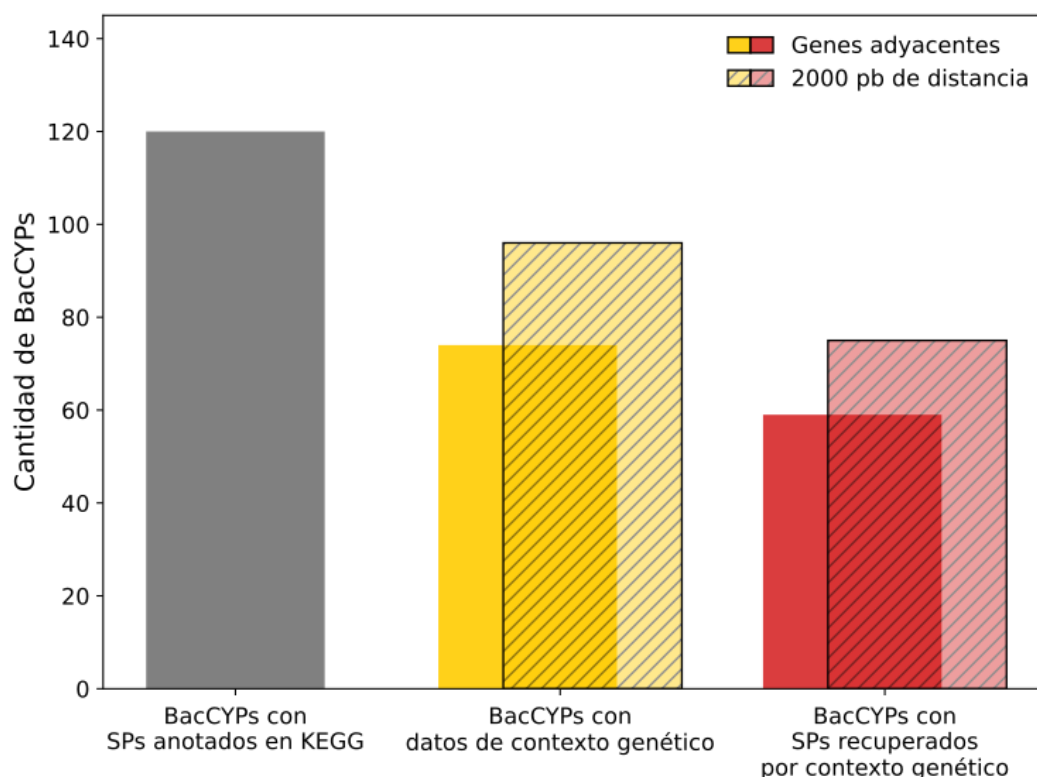


Figura 4.9: Evaluación de la recuperación de sustratos y/o productos de BacCYPs anotados por medio de contexto genético, considerando las reacciones químicas de proteínas codificadas por genes adyacentes (barras lisas) o encontrados dentro de un rango de distancia genómica de 2000 pares de bases (barras sombreadas con líneas en diagonal). Se observa que el criterio ampliado (distancia genómica ≤ 2000 pares de bases) aumenta tanto la cantidad de datos obtenidos como la capacidad de recuperación de sustratos y productos de BacCYPs en estudio.

Es importante destacar que, tanto en la búsqueda limitada a genes adyacentes como en el rango ampliado hasta 2000 pb, en aproximadamente el 80 % de los casos en los que se identifican SPs candidatos, se recuperó el mismo SP verdadero que interactúa con el BacCYP en estudio.

Como ejemplo ilustrativo, Cyp158A2 de *Streptomyces coelicolor* (ID Uniprot Q9FCA6) sintetiza 3,3'-biflaviolina utilizando flaviolina como sustrato. Entre las dos proteínas codificadas por genes adyacentes, una es la sintasa de 1,3,6,8-tetrahidroxinaftaleno (ID Uniprot Q9FCA7) y la otra es una proteína con un dominio tipo-1 Cupin (ID Uniprot Q9FCA5). La primera produce naftaleno-1,3,6,8-tetrol, un compuesto estructuralmente similar al sustrato de Cyp158A2 (Figura 4.10). Este es uno de los pocos ejemplos en los que la aplicación del enfoque de contexto genético resultó en un

compuesto similar al sustrato real del BacCYP en estudio, ya que el resultado más común fue la obtención de exactamente el SP verdadero.

Estos resultados destacan la efectividad de este enfoque para la identificación de SPIs en nuevos BacCYPs en estudio, con la posibilidad de extender estos análisis a más de 4,600 BacCYPs cuyos SPIs aún no se conocen, y para los cuales se dispone de información sobre los SPs de sus genes vecinos.

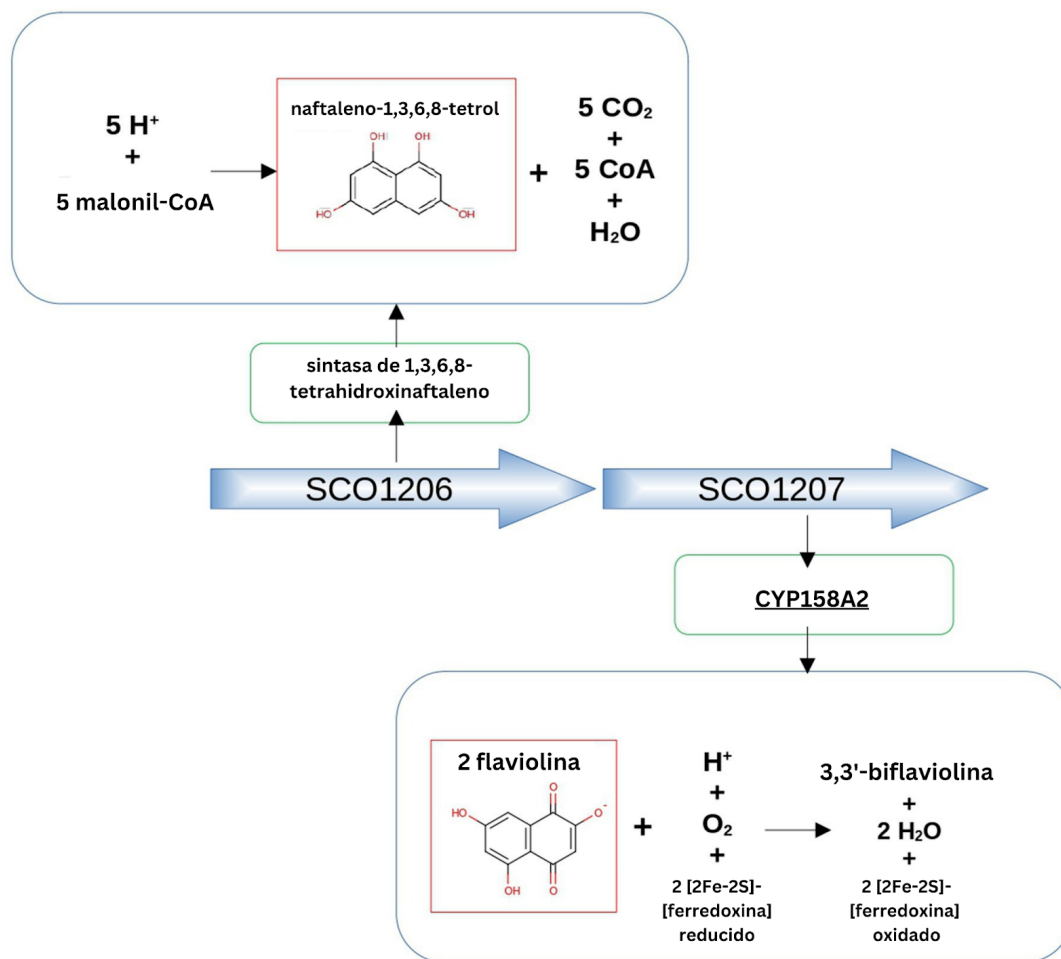


Figura 4.10: Enfoque del contexto genético ilustrado utilizando el ejemplo de los genes adyacentes SCO1206 y SCO1207 en *Streptomyces coelicolor*. SCO1206 codifica la proteína sintasa de 1,3,6,8-tetrahidroxinaftaleno, mientras que SCO1207 codifica el citocromo P450 CYP158A2, que sintetiza el compuesto 3,3'-biflavolina. La figura resalta la similitud química entre el naftaleno-1,3,6,8-tetrol, un producto de la reacción catalizada por la proteína sintasa de 1,3,6,8-tetrahidroxinaftaleno, y flaviolina, el sustrato de la proteína codificada por el gen adyacente, el citocromo P450 CYP158A2.

4.2.6 Enfoque predictivo 3: *Docking* molecular

El último enfoque evaluado para predecir SPIs para nuevos BacCYPs consiste en la predicción de la estabilidad de las interacciones de potenciales SPIs mediante *docking* molecular.

Dado que el universo de posibles SPIs de BacCYPs es enorme y abarca innumerables posibilidades, se optó por seleccionar, como compuestos de referencia para estas pruebas, dos representantes aleatorios de cada uno de los subgrupos de SPIs determinados en la sección 4.2.3 y ejemplificados en la Figura 4.6 A.

Para evaluar estos compuestos preseleccionados, se eligieron 15 estructuras de proteínas representativas (PRs), cada una interactúa con un subgrupo diferente de SPIs (a excepción del subgrupo 1.4 y del grupo 6, para los cuales no se pudieron obtener cristales representativos adecuados para la evaluación). Las proteínas seleccionadas como ejemplos y los subgrupos de SPIs con los que éstas interactúan se muestran en la Tabla 4.11.

Proteína (ID de Uniprot)	Estructura de PDB correspondiente	Grupo de SPIs que une la proteína
O31440	1IZO	1.1
Q8VQF6	3BDZ	1.2
P9WPP3	6T0J	1.3
Q9KIZ4	1Q5D	2
P00183*	1AKD	3
Q06069	5IKI	4.1
P18326	3CV9	4.2
P00183*	1PHA	5.1
P9WPN8	5LI8	5.2
P9WPP7	3G5H	7
Q9KZF5	2NZ5	8
P9WPP9	2CI0	9
A0A076MY51	5OMU	10.1
Q6N8N2	4DNJ	10.2

Q8NSW2	5GWE	10.3
--------	------	------

Tabla 4.11: Proteínas elegidas para las evaluaciones de *docking* molecular. Se muestra además cuál de los subgrupos determinados de SPIs unen y el cristal representativo de PDB.

*P00183 une tanto el grupo 3 de SPIs como el 5.1.

A continuación, el objetivo de la evaluación fue determinar si los ligandos del grupo de SPIs que se une a cada PR se posicionan alto en el ranking de resultados de *docking*. Este ranking se define a través del *Z-score*, y los ligandos con alta probabilidad de unirse a una proteína se consideran aquellos cuyo resultado de *docking* se ubica en el cuadrante superior izquierdo del gráfico de *Z-score*, indicando una baja energía libre y alta población de clusters. Este proceso se detalla en la sección 2.4.3, y un ejemplo de este tipo de gráfico se presenta en la Figura 2.10 correspondiente a dicha sección.

Como resultado, en 13 de las 15 PRs (las excepciones fueron las PRs representativas de los grupos de SPIs 1.3 y 8), al menos uno de los SPIs evaluados de los grupos que estas unen se situó en la zona del cuadrante superior izquierdo en el gráfico de *Z-score*.

En términos generales, los resultados sugieren que el *docking* molecular es una herramienta valiosa para explorar la afinidad de unión de sustratos a BacCYPs y para analizar la selectividad en ausencia de datos experimentales. A grandes rasgos, nos permite descartar tipos de moléculas con baja probabilidad de unión a una proteína específica en estudio.

4.2.7 Diseño de un Flujo de Trabajo Integrado para los Tres Enfoques Desarrollados

Finalmente, se propone integrar los 3 enfoques predictivos para el descubrimiento de nuevos SPIs de BacCYPs desarrollados a lo largo de nuestra investigación en un mismo flujo de trabajo. El protocolo sugerido se ilustra en la Figura 4.12.

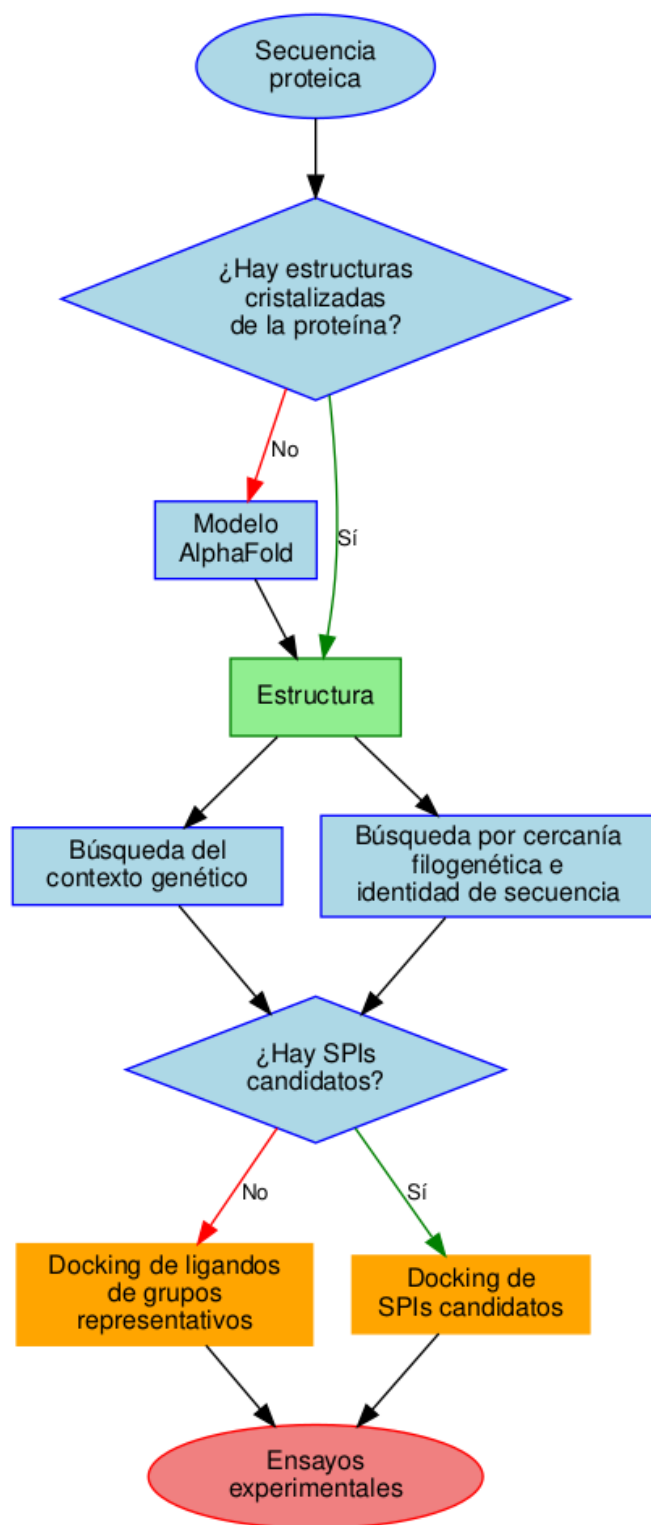


Figura 4.12: Esquema de nuestro flujo de trabajo propuesto para descubrir sustratos para BacYPs no anotados.

Este proceso requiere de la secuencia un BacCYP en estudio como punto de partida. Partiendo de la secuencia, se pueden realizar varios caminos diferentes para extraer de ésta la mayor cantidad de información útil.

En primer lugar, para la aplicación de enfoques estructurales, como el *docking* molecular, es fundamental determinar la estructura tridimensional de la proteína. Para ello, se verifica inicialmente la disponibilidad de una estructura cristalizada en el PDB. Si se encuentra una estructura adecuada, se utiliza para continuar con el análisis. En caso de no disponer de una estructura, se recurre a la generación de un modelo mediante AlphaFold, lo que permite contar con una estructura confiable de referencia para los análisis subsecuentes.

Con la secuencia clasificada filogenéticamente y la estructura de la proteína ya determinada, se procede a la identificación de SPIs candidatos utilizando dos enfoques complementarios: el análisis de contexto genético y la búsqueda por correlación filogenética e identidad de secuencias. El análisis de contexto genético, explicado en detalle en la sección 4.2.5, examina genes adyacentes y cercanos, fundamentándose en la típica organización de operones en bacterias. Por otro lado, la búsqueda por correlación filogenética e identidad de secuencias, descrita en la sección 4.2.4, permite predecir interacciones potenciales mediante comparaciones con proteínas homólogas que ya tienen SPIs conocidos.

Después de aplicar ambos enfoques, se verifica la presencia de posibles SPIs. Si no se identificaron compuestos candidatos, se procede con el *docking* molecular utilizando un conjunto representativo de los ligandos conocidos que unen BacCYPs, como se describió en la sección 4.2.6. Por el contrario, si se identificaron compuestos candidatos a ser SPIs, se realiza el *docking* molecular en el dominio funcional de la proteína con estas moléculas. Este enfoque permite priorizar, entre todos los candidatos, aquellos con mayor probabilidad de ser verdaderos SPIs del BacCYP. Finalmente, como resultado de este flujo de trabajo se obtiene una selección de compuestos aptos para poder realizar ensayos experimentales.

La automatización completa de este flujo de trabajo queda pendiente para una etapa futura; no obstante, cada módulo puede ejecutarse secuencialmente de forma independiente, con un tiempo de procesamiento de ~1 minuto/BacCYP en condiciones favorables. Este esquema considera:

1. Búsqueda rápida de BacCYPs cercanos evolutivamente mediante BLAST para recuperar SPIs anotados.
2. Análisis de contexto genético a través de llamadas a la API de KEGG.
3. Docking molecular acelerado con AutoDock GPU sobre una estructura disponible—ya sea cristalográfica o descargada de la base de datos de AlphaFold.

4.2.8 Evaluación de 2 casos de estudio para probar la eficacia de la metodología

Como paso final de este trabajo, se prueba la capacidad predictiva de la estrategia propuesta con dos ejemplos de BacCYPs con SPIs conocidos.

El primer caso de ejemplo corresponde al CYP121 de la cepa *Mycobacterium tuberculosis* H37Rv (MtCYP121, cuya ID de Uniprot es P9WPP7), que está involucrado en la síntesis del compuesto micociclosina, y utiliza como sustrato a ciclo(L-tirosil-L-tirosil), que pertenece al grupo 7 de SPIs.

El análisis de secuencia revela que MtCYP121 pertenece al grupo filogenético J, cuyas proteínas se asocian con SPIs de los grupos 1.1, 1.2, 1.3, 2, 4.1, 4.2, 5.1, 5.2, 6, 7 y 8. Como es de esperarse según los análisis previos, el análisis filogenético basado en el árbol que abarca el universo completo de BacCYPs no proporciona indicios claros sobre el grupo específico de SPIs que esta proteína puede unir. Por otra parte, las únicas proteínas muy cercanas evolutivamente con SPIs anotados corresponden a BacCYPs homólogos de diferentes cepas de *Mycobacterium tuberculosis*, que se descartan como ejemplos de estudio debido a que su similitud extrema no representa un ejemplo probable en la práctica como para evaluar adecuadamente la eficacia del enfoque.

Se generó un modelo estructural de MtCYP121 utilizando Modeller, a pesar de que las conclusiones de la metodología sugieren el uso de AlphaFold para obtener estructuras confiables. En este caso específico, se prefirió Modeller para evitar sesgos, dado que información estructural experimental de MtCYP121 se utilizó en el entrenamiento de AlphaFold v2.0. Para el modelado, se seleccionó como molde la estructura de MgMycG de *Micromonospora griseorubida* (PDB ID: 2YGX), que comparte un 50% de identidad de secuencia con MtCYP121 y también pertenece al grupo filogenético J. Esta elección se basó en que MgMycG, aunque no es la proteína con estructura conocida más similar, representa un caso realista en el que no es posible obtener un molde con una identidad de secuencia extremadamente alta. El modelo resultante mostró alta calidad, con un RMSD-C α de 1.192 Å respecto a la estructura cristalizada, como se ilustra en la Figura 4.13 A.

El análisis del contexto genético revela que los genes adyacentes al MtCYP121 corresponden a la sintasa de ciclo(L-tirosil-L-tirosil) (ID de Uniprot: P9WPF9) río arriba y a la proteína no caracterizada Rv2277c (ID de Uniprot: P9WLF1) río abajo. La proteína sintasa de ciclo(L-tirosil-L-tirosil) sintetiza el sustrato conocido de CYP121, ciclo(L-tirosil-L-tirosil), a partir de L-tirosil-ARNt(Tir). Durante esta reacción, la sección ARNt(Tir) del sustrato se separa, completando el proceso de síntesis. Por lo cual este compuesto resulta en el único SPI candidato para la proteína en estudio.

El *docking* del sustrato potencial ciclo(L-tirosil-L-tirosil) con la estructura de MtCYP121 da un resultado que indica una interacción favorable (perteneciente al cuadrante superior izquierdo en el gráfico de *Z-score*, como se detalló en la sección 4.2.6), y por lo tanto, favorece su identificación como resultado positivo, obteniéndose una pose muy similar a la interacción real entre el compuesto y la proteína (Figura 4.13 B).

Este caso de estudio representa un panorama favorable, ya que por medio de la búsqueda por contexto genético, se obtuvo un único compuesto candidato (Figura 4.13 C), que corresponde al verdadero sustrato de la proteína evaluada en el ejemplo. El resultado positivo del ensayo de *docking* molecular respalda la obtención de ciclo(L-tirosil-L-tirosil) como sustrato en este ejemplo.

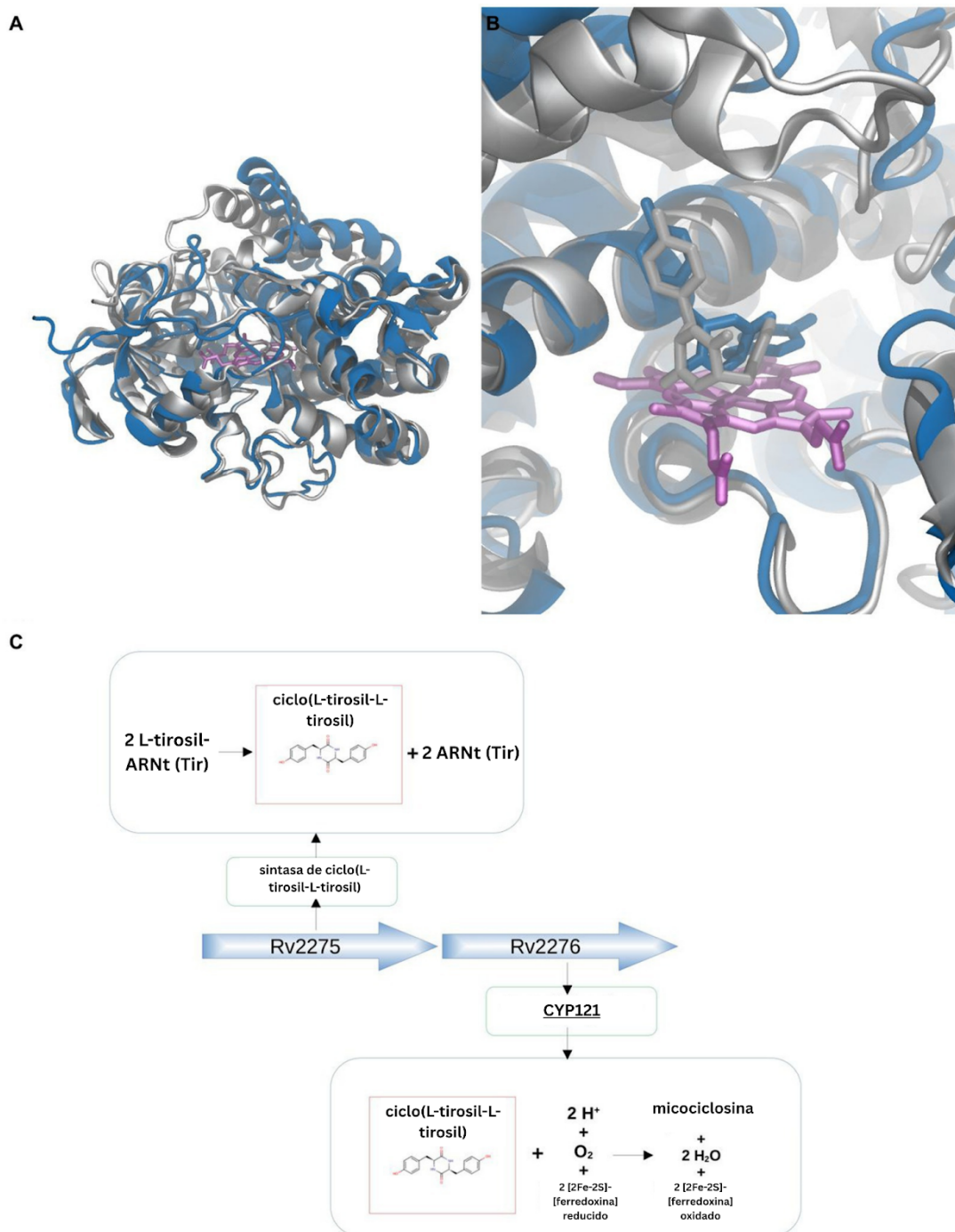


Figura 4.13: **A)** Flujo de trabajo para el descubrimiento de sustratos aplicado a MtCYP121. Análisis comparativo entre la estructura real (plateado) y el modelo generado con Modeller (azul). **B)** Comparación entre la pose real (plateado) y la pose predicha por *docking* (azul) del ligando ciclo(L-tyrosil-L-tyrosil) en complejo con la proteína MtCYP121. **C)** Enfoque de contexto genético que muestra que ciclo(L-tyrosil-L-tyrosil), sustrato de CYP121, es el producto de la proteína sintasa de ciclo(L-tyrosil-L-tyrosil), codificada por el gen adyacente Rv2275.

El segundo ejemplo corresponde a la subunidad citocromo P450 de la enzima O-demetilasa aromática (GcoA) de *Amycolatopsis* sp. ATCC 39116 (ID de Uniprot: P0DPQ7). Esta proteína es capaz de unir diferentes SPIs de los grupos 10.1 y 10.2, como guaiacol, 3-metoxicatecol y guaetol.

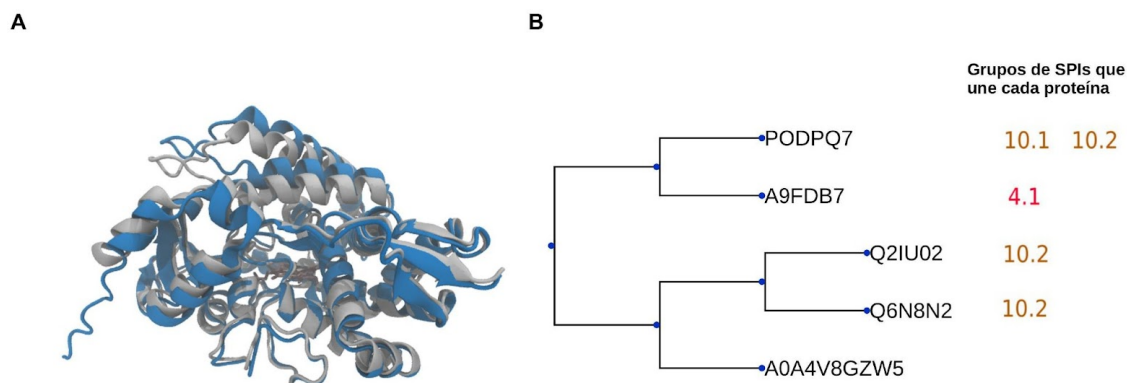


Figura 4.14: A) Comparación entre estructura cristalizada (plateado) y modelo generado por AlphaFold (azul).

B) Sección del árbol filogenético de BacCYPs con SPIs conocidos, en la que se demuestra que proteínas cercanas evolutivamente con GcoA de *Amycolatopsis* sp. ATCC 39116 (P0DPQ7) tienen como SPIs a compuestos del grupo 10.2.

En el árbol filogenético del conjunto total de BacCYPs (representado en la Figura 4.1 A), este polipéptido se clasifica dentro del grupo filogenético I, cuyas proteínas se asocian con SPIs de hasta nueve grupos distintos. Sin embargo, al considerar un árbol más específico que agrupa únicamente BacCYPs con SPIs conocidos (representado en la Figura 4.8), se observa que la proteína se encuentra relacionada con otras que unen los grupos de SPIs 10.2, como la proteína Q2IU02 de *Rhodopseudomonas palustris*, y 4.1, como la 1 α -hidroxilasa de esteroides C-19 de *Sorangium cellulosum* (ID de Uniprot: A9FDB7). Se excluyeron del análisis las proteínas con identificación de Uniprot A0A3B6UEE9 y A0A076MY51, que también unen al grupo 10.1 de SPIs, pero debido a su alta similitud de secuencia con la proteína en estudio podrían corresponder a otras identificaciones del mismo polipéptido.

La primera estructura cristalizada de esta subunidad fue publicada en el PDB el 4 de julio de 2018, lo que la hace adecuada para el modelado predictivo mediante AlphaFold, ya que su estructura no formó parte de los datos de entrenamiento de AlphaFold versión 2.0. Con esta herramienta se obtuvo un modelo de alta calidad, con un RMSD-C α de 0.726 en comparación con la estructura cristalizada (Figura 4.14 A).

El análisis de contexto genético no fue aplicable en este caso debido a la falta de información sobre los sustratos o productos de los genes vecinos. Sin embargo, el *docking* molecular con ligandos representativos mostró que el grupo de SPIs 10.1 ocupaba el segundo lugar en el ranking de *Z-score* (se detalla la metodología en la sección 2.4.3), por debajo del grupo 8, y el grupo 10.2 se posicionó en quinto lugar.

Este ejemplo ilustra que, aunque el análisis filogenético amplio no proporcionó pistas claras, la consideración de un árbol más restrictivo (limitado a proteínas con SPIs conocidos) permitió recuperar el grupo 10.2 como uno de los SPIs asociados (Figura 4.14 B). Por otro lado, con el *docking* molecular se logró identificar correctamente al grupo 10.1 de SPIs. Así, aunque el contexto genético no otorgó resultados en este caso, los otros dos enfoques permitieron inferir correctamente los grupos de SPIs que interactúan con esta proteína.

4.3 Conclusiones

Los BacCYPs son proteínas cruciales en las vías de síntesis y de degradación de diversos metabolitos, y por lo tanto, determinar su papel preciso en estos procesos es esencial tanto para el descubrimiento de inhibidores como para el aprovechamiento de su potencial biotecnológico.

En este trabajo se obtuvieron tres enfoques independientes para el descubrimiento de nuevos SPIs de BacCYPs. Mostramos que los tres métodos pueden ser utilizados tanto de manera individual como de manera integrada en un flujo de trabajo para ampliar y otorgar solidez a los resultados obtenidos.

El análisis filogenético, aunque resultó poco informativo al aplicarse al universo completo de BacCYPs, mostró una gran eficacia para la predicción de SPIs en grupos filogenéticos más pequeños, formados por proteínas evolutivamente más relacionadas. Esto quedó demostrado al construir el árbol filogenético de secuencias de BacCYPs con SPIs conocidos. El fortalecimiento de este método dependerá en gran medida de la investigación y caracterización de los grupos filogenéticos de BacCYPs menos estudiados, lo que permitirá extender su aplicabilidad a nuevas proteínas al aumentar la probabilidad de tener referencias más cercanas de BacCYPs con SPIs conocidos.

Por otro lado, la búsqueda de sustratos y productos (SPs) basada en el contexto genético ha mostrado una notable eficacia en la identificación precisa de SPs para BacCYPs, a partir de los SPs que interactúan con proteínas codificadas por genes vecinos. Cabe destacar que, a medida que se amplíe la anotación de genes, se espera que la cantidad de información disponible aumente, mejorando así la capacidad predictiva de este tipo de enfoques.

El *docking* molecular demostró ser una herramienta eficaz para descartar grupos de moléculas con baja probabilidad de ser SPIs verdaderos, especialmente en casos donde los otros dos enfoques no proporcionaron información adicional. Además, puede utilizarse para evaluar la estabilidad de la unión de un sustrato candidato previamente identificado y refinar los resultados obtenidos a través de los otros métodos presentados en este trabajo.

Un aspecto destacable de este trabajo es la implementación de AlphaFold para la construcción de modelos de estructuras de proteínas. Nuestros resultados demuestran que esta herramienta puede generar modelos de BacCYPs con una precisión sobresaliente, incluso superando la calidad de los modelos obtenidos mediante el modelado por homología partiendo de proteínas molde con una alta identidad de secuencia. La generación de modelos de alta calidad es fundamental para los enfoques de biología estructural, como el *docking* molecular y otros métodos adicionales, como dinámicas moleculares. La creciente capacidad de construir modelos confiables para BacCYPs sólo

conociendo su secuencia amplía de manera significativa el potencial predictivo a través de estrategias basadas en la estructura proteica.

En trabajos anteriores se han propuesto ideas similares a las que dieron lugar a los enfoques presentados en nuestra investigación. Por ejemplo, la definición de operones de genes que codifican proteínas responsables de la biosíntesis de varios metabolitos en una amplia gama de organismos, incluidos procariotas y eucariotas, es una estrategia bien conocida para la anotación de la función génica, incluida la especificidad de sustratos y/o productos (Chavali and Rhee 2018). Algunos estudios también analizaron operones de genes relacionados con el metabolismo de varios tipos de compuestos, y entre los genes involucrados se encuentran anotados varios BacCYPs (Greule et al. 2018).

En cuanto a las interacciones específicas entre BacCYPs y ligandos, las estrategias de *docking* molecular y *screening* virtual se han aplicado a diversos estudios y búsquedas de inhibidores de CYP450s, incluyendo BacCYPs encontrados en bacterias patogénicas (Snow Setzer, Sharifi-Rad, and Setzer 2016; Reddy et al. 2013). Además, otros trabajos han empleado herramientas de bioinformática estructural para la evaluación de interacciones moleculares entre BacCYPs con mutaciones relevantes y compuestos de interés (Park et al., 2010).

Finalmente, desde una perspectiva global, la estrategia presentada demuestra la efectividad de combinar distintos métodos bioinformáticos y quimioinformáticos, cada uno aportando información adicional para poder cumplir un objetivo general. Además, este protocolo, si bien fue desarrollado tomando el caso de estudio de los BacCYPs, puede ser aplicado para fortalecer las investigaciones en otras familias de proteínas de interés.

4.4 Métodos específicos del capítulo

4.4.1 Obtención de secuencias de Citocromos P450 bacterianos

Para identificar todas las secuencias de Citocromo P450, se realizó una búsqueda en la base de datos UniProt utilizando la consulta que incluye en el campo "protein_name" la frase "*Cytochrome P450*" y el filtro por taxonomía de secuencias correspondientes a "Bacteria (Eubacteria) [2]". El resultado de la búsqueda constituye el conjunto completo de secuencias de BacCYPs conocidas.

En segundo lugar, se aplicó el filtro adicional "*proteins with 3D structure*", con el fin de obtener las proteínas que tienen estructuras disponibles en el PDB. Este conjunto corresponde al conjunto de datos de BacCYPs con estructura experimental.

Por último, se utilizó el filtro "*proteins with catalytic activity*" para encontrar aquellos BacCYP450s que poseen reacciones químicas anotadas con sus respectivos sustratos y productos. Esta búsqueda se complementó recopilando información adicional de la base de datos KEGG (Kanehisa y Goto, 2000). El conjunto de datos resultante corresponde a los BacCYPs con Reacciones Conocidas.

La Tabla 4.15 muestra el número de secuencias obtenidas en cada conjunto de datos. Para identificar el dominio funcional CYP450 en cada secuencia, se siguió el mismo protocolo utilizado en la parte anterior del trabajo: se llevó a cabo una búsqueda utilizando el HMM correspondiente a esta familia, descargado de Pfam, mediante la herramienta HMMER.

BacCYPs totales	BacCYPs con estructura experimental	BacCYPs con reacciones conocidas
118225	133	120

Tabla 4.15: Número de secuencias de BacCYPs asignadas a cada subconjunto.

4.4.2 Alineamiento múltiple de secuencias (MSA) y reconstrucción de árboles filogenéticos

Los alineamientos múltiples de secuencias (MSAs) realizados para posteriormente establecer los análisis filogenéticos, se llevaron a cabo mediante la herramienta MAFFT, ampliamente utilizada y destacada por su comprobada relación rapidez-precisión (Katoh et al. 2002), por la opción --dash que integra datos estructurales del PDB y mejora la alineación de homólogos lejanos (Rozewicki et al. 2019) y por su bajo consumo de

memoria, lo que facilita su uso en distintos entornos computacionales (Nakamura et al. 2018).

La selección de las columnas más informativas del alineamiento múltiple -necesario para realizar un análisis filogenético preciso- se llevó a cabo utilizando la herramienta BMGE, con la matriz *default* BLOSUM62 y los siguientes parámetros: tamaño mínimo de bloque de 1, tamaño de ventana de 1, umbral de entropía de 1, y una proporción máxima de gaps de 0.4. Este conjunto de parámetros permite que el alineamiento resultante conserve las posiciones informativas —tanto las muy conservadas como las moderadamente variables— y descarte las columnas excesivamente variables que no aportan información evolutiva relevante.

Para realizar un análisis filogenético representativo pero manejable, se seleccionó una muestra de 2,000 secuencias de proteínas del conjunto completo de BacCYPs conocidos. Este subconjunto se generó utilizando Cd-hit para agrupar secuencias similares, de modo que cada grupo de secuencias similares entre sí estuviera reducido a un único representante. Esto garantizó una diversidad adecuada y la representatividad del universo total de BacCYPs.

Posteriormente, los árboles filogenéticos se construyeron con el programa PhyML, que presenta la capacidad de optimizar el proceso permitiendo la selección automática del modelo evolutivo más adecuado mediante el enfoque *Smart Model Selection* (SMS). Los grupos filogenéticos resultantes se identificaron mediante Aglomeración Jerárquica (hclust), utilizando una matriz de distancia basada en la distancia cofenética obtenida a través del paquete de R ape. El criterio para elegir la cantidad de grupos, consistió en maximizar el número de grupos que contengan cada uno, al menos el 1% de las secuencias totales.

Además, se realizó un árbol más simple que consideró exclusivamente los BacCYPs con sustratos, productos o inhibidores anotados. En este caso, no se aplicó la estrategia de agrupamiento con Cd-hit ni se realizó una subdivisión posterior en grupos.

Todas las técnicas empleadas se describen detalladamente en los métodos generales (Capítulo 2).

4.4.3 Búsqueda de sustratos, productos e inhibidores de Citocromos P450 bacterianos

Los sustratos, productos e inhibidores conocidos de BacCYPs fueron recopilados de diversas fuentes.

En primer lugar, se extrajeron sustratos y productos implicados en la actividad catalítica de los BacCYPs a partir de la base de datos Rhea, una fuente de reacciones bioquímicas curada, estandarizada y con enlaces cruzados a bases de datos clave.

Para identificar los ligandos que se unen a los BacCYPs en estructuras determinadas experimentalmente, se realizó una búsqueda en el Protein Data Bank (PDB), el principal recurso especializado en el área. Posteriormente, esta información fue filtrada con ayuda de la base de datos MOAD, que permite seleccionar exclusivamente ligandos biológicamente relevantes. MOAD clasifica estos compuestos como péptidos de hasta 10 aminoácidos, oligonucleótidos de hasta 4 nucleótidos, pequeñas moléculas orgánicas y cofactores, excluyendo sustancias no pertinentes como sales, iones y metales.

Para enriquecer aún más la recopilación de ligandos, se consultó la base de datos ChEMBL (versión 32), enfocándose en aquellos con evidencia de actividad inhibitoria en ensayos experimentales. Se seleccionaron ligandos con un valor pchembl superior a 6, el cual indica una afinidad inhibitoria significativa sobre la proteína.

Además, se obtuvieron los sustratos y productos asociados a BacCYPs anotados en la base de datos KEGG, especializada en reacciones enzimáticas y rutas metabólicas, para completar el conjunto de compuestos asociados con esta familia de proteínas.

Por último, todas las estructuras químicas de los compuestos recopilados se almacenaron en formato SMILES, para poder realizar los análisis posteriores, que requieren este tipo de representación como punto de partida.

4.4.4 Agrupación de sustratos, productos e inhibidores por similitud

La agrupación de los sustratos, productos e inhibidores (SPIs) de BacCYPs se realizó mediante la implementación del método de *clustering* Butina, incluido en la librería RDkit, que permite una agrupación personalizada con un umbral de corte de Índice de Tanimoto (TI).

Para implementar este método, fue necesario convertir todas las estructuras de SPIs en *Fingerprints* de Morgan de 1048 bits y radio 2 (parámetros usados de manera estándar en quimioinformática), generados a partir de las representaciones SMILES correspondientes.

Con el fin de optimizar el proceso de agrupamiento, se realizó una evaluación sistemática de umbrales de TI desde 0.20 hasta 0.50, en incrementos de 0.05. Este rango se seleccionó porque los *Fingerprints* de Morgan son altamente sensibles a pequeños cambios estructurales, lo que significa que incluso diferencias menores entre dos estructuras, como el intercambio de un grupo funcional, pueden resultar en un TI relativamente bajo. Así, umbrales de similitud aparentemente bajos (0.3 a 0.5) pueden ser lo suficientemente inclusivos para agrupar moléculas con estructuras y funciones similares (Rogers and Hahn 2010).

Se decidió la obtención de un umbral que equilibrara dos aspectos clave: primero, garantizar la cohesión interna de los clústeres, asegurando que las moléculas dentro de cada grupo de SPIs compartieran características estructurales y funcionales; segundo, mantener la cobertura y representatividad del conjunto de datos, minimizando el número de SPIs no agrupados. Una alta proporción de compuestos sin grupos asignados podría indicar una fragmentación excesiva, dificultando la extracción de conclusiones generales.

Para definir el umbral definitivo, se verificó que cada grupo de SPIs bien definido (como esteroides, ácidos grasos y compuestos similares al alcanfor) formara un clúster cohesivo, sin dejar compuestos pertenecientes sin agrupar o dispersos en diferentes clusters.

Tras una exhaustiva evaluación de los diferentes umbrales tomando estos criterios, se decidió seleccionar el valor de 0.35 como el más adecuado. A este nivel de similitud, los clústers formados mostraron una cohesión interna apropiada, evidenciando que las moléculas agrupadas compartían características químicas relevantes. Como así también, se logró reducir el número de sustratos no agrupados, incrementando así la cobertura y representatividad del conjunto de datos. Algunos ejemplos de los clústers obtenidos se muestran en la Figura 4.16.

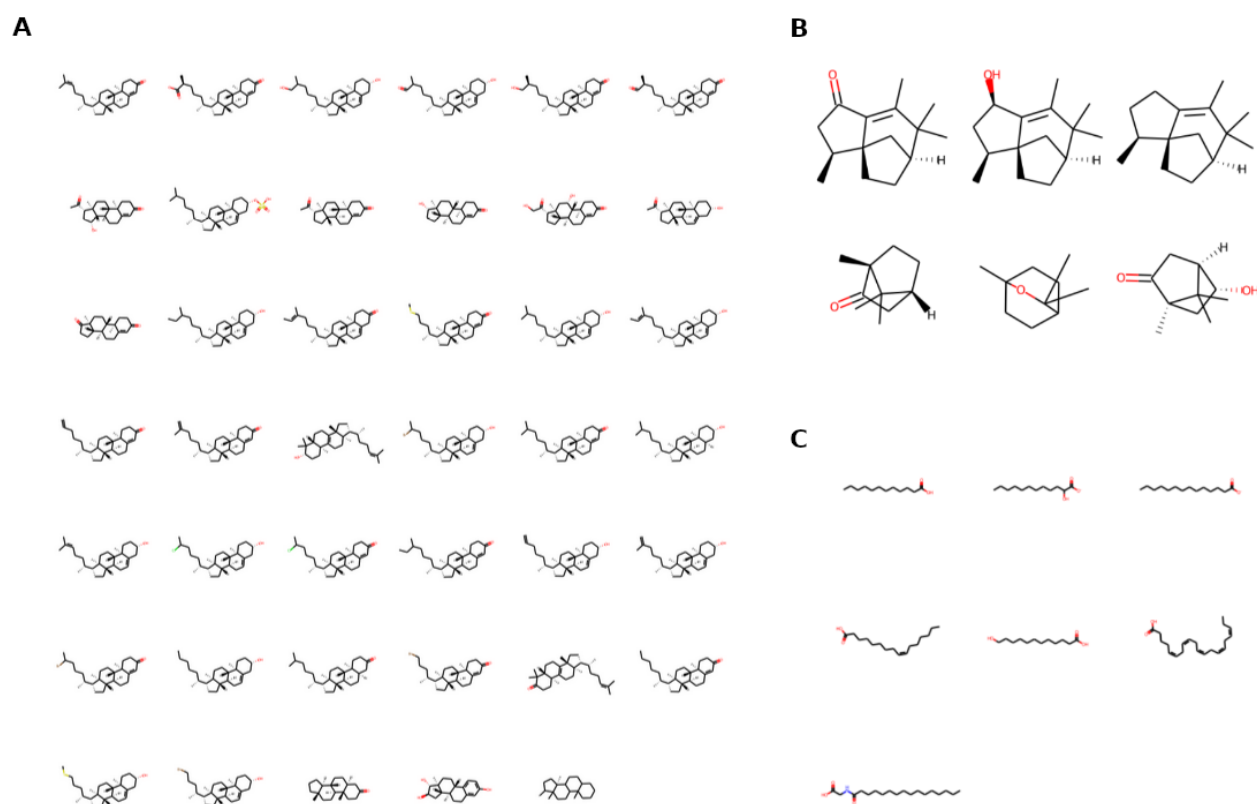


Figura 4.16 Se ilustran de manera general ejemplos de agrupamientos de compuestos obtenidos utilizando un umbral de índice de Tanimoto de 0.35:

A) Grupo correspondiente a los SPIs del grupo de los esteroides. **B)** SPIs agrupados dentro de la clase de ligandos similares al alcanfor. **C)** SPIs que conforman el grupo de ácidos grasos y otros ligandos afines.

4.4.5 Creación y evaluación de modelos estructurales construidos con Modeller y AlphaFold

Para la generación de modelos de proteínas mediante modelado por homología, se empleó el software Modeller, elegido tanto por la experiencia previa del Grupo de Investigación (Serral et al. 2021; Palumbo et al. 2022) como por ser una de las herramientas más reconocidas y robustas para la obtención automatizada de modelos tridimensionales a partir de alineamientos y proteínas molde.

El enfoque para la evaluación consistió en seleccionar una muestra de 12 proteínas con estructuras cristalizadas, con el fin de evaluar la precisión de los modelos empleando proteínas molde con diferentes distancias filogenéticas.

La selección de proteínas molde se basó en el árbol filogenético representativo del universo de todos los BacCYPs conocidos construido en la sección 4.2.1. A partir de este árbol, se definieron tres clases distintas de moldes:

- Moldes de corta distancia: Se seleccionaron las tres proteínas con estructura conocida más cercanas evolutivamente a la proteína que se desea modelar.
- Moldes de distancia media: Estas proteínas pertenecen al mismo grupo filogenético que la proteína a modelar, pero no son cercanas evolutivamente. En casos donde no se disponía de suficientes proteínas con estructura dentro del mismo grupo filogenético, se optó por las proteínas más similares evolutivamente del grupo filogenético más próximo.
- Moldes de larga distancia: Se seleccionaron proteínas de grupos filogenéticos distantes en términos evolutivos.

Cada proteína de la muestra se modeló nueve veces, distribuyéndose las simulaciones de la siguiente manera: tres modelos se construyeron con proteínas molde de corta distancia, tres con moldes de distancia media y tres con moldes de larga distancia. Para cada uno de estos modelos, se generaron diez réplicas y se seleccionó la mejor de estas con el criterio de puntuación DOPE (*Discrete Optimized Protein Energy*).

Para los modelos generados con AlphaFold, se utilizó la versión local de ColabFold v1.3.0 (Mirdita et al. 2022), que utiliza el flujo de trabajo de AlphaFold v2.0. Esta versión de la herramienta se entrenó con datos que incluyen estructuras depositadas en el PDB hasta el 30 de abril de 2018. Para evitar el sesgo derivado de que las estructuras experimentales de las proteínas a modelar hubieran sido utilizadas en el entrenamiento del modelo subyacente de esta herramienta, seleccionamos 18 BacCYPs cuyas estructuras cristalizadas fueron publicadas con posterioridad a la fecha indicada. Para cada proteína modelada, se generaron 5 réplicas sin utilizar moldes (para evitar incorporar datos sobre estructuras experimentales de las proteínas evaluadas) y se seleccionó la mejor de ellas de acuerdo a su puntaje pLDDT (la métrica por defecto utilizada por AlphaFold para determinar la confianza de los modelos).

La evaluación final de la calidad de las estructuras generadas con cada método se llevó a cabo realizando una comparación con la estructura experimental de referencia, tomando como métrica el RMSD. Las métricas utilizadas para definir la calidad de los modelos se explican detalladamente en la sección 2.3.3.

4.4.6 Docking molecular

Para realizar los ensayos de *docking*, se eligió un conjunto de ligandos que incluyó dos compuestos por cada uno de los grupos de SPIs definidos en la sección 4.2.3. Asimismo, por cada grupo de SPIs, se eligió una proteína cristalizada en el PDB que, con comprobación experimental, se probó que une algún ligando del grupo correspondiente. Con el objetivo de evaluar si es posible identificar correctamente el grupo de SPIs que cada proteína une, se realizaron ensayos de docking con el conjunto completo de ligandos para cada una de estas proteínas representativas.

Las grillas para las corridas del *docking* se construyeron utilizando AutoGrid, con un tamaño adaptativo, de $40 \times 40 \times 40$ Å para ligandos con menos de 35 átomos, y de $60 \times 60 \times 60$ Å para ligandos con más de 35 átomos. El espaciado de la grilla se estableció en 0.375 Å, y el centro se posicionó 10 Å por encima del grupo hemo de cada Citocromo. Los cálculos de *docking* se realizaron con AutoDockGPU (versión de AutoDockTools acelerada por GPU) utilizando los siguientes parámetros: $ga_run = 100$, $rmstol = 2.5$ Å. En este caso se modificaron los parámetros *default* (especialmente el tamaño de la grilla) debido a la presencia de algunos SPIs de mayor tamaño y flexibilidad estructural (grupos 1, 2, 4 y 5 por ejemplo, ilustrados en la Figura 4.6 A), que de otro modo hubiesen quedado desfavorecidos para la búsqueda conformacional frente a otros ligandos de menor tamaño.

4.4.7 Docking sesgado

El *docking* sesgado se realizó con AutoDock Bias, una herramienta previamente desarrollada por nuestro Grupo de Investigación. Este método introduce sesgos energéticos dirigidos a átomos específicos de las proteínas, favoreciendo aquellas interacciones proteína-ligando previamente identificadas como relevantes.

Para cada uno de los BacCYPs sobre los cuales se aplicó este protocolo, se obtuvieron las estructuras correspondientes de PDB en complejo con compuestos. Los sitios de aplicación de los sesgos se obtuvieron de las interacciones proteína-ligando no covalentes identificadas en dichos cristales.

Para los ensayos de *docking* luego de la introducción de los sesgos, se utilizaron los mismos parámetros de grilla y de corrida detallados en la sección 4.4.6.

4.4.8 Búsqueda de sustratos por contexto genético

Para el enfoque de la búsqueda de posibles SPIs por contexto genético, se desarrolló un script personalizado en lenguaje Python que, dados un gen codificante de un BacCYP y un organismo, realiza una consulta en la API de KEGG para identificar genes

cercanos, hasta un umbral especificado de distancia en pares de bases a lo largo del genoma.

El script recupera los genes (y las proteínas que codifican) ubicados río arriba y río abajo del gen correspondiente al BacCYP, y extrae de KEGG las anotaciones sobre reacciones que involucran productos y sustratos (de la sección MODULE), así como sus respectivos números de la Comisión de Enzimas (EC). También, como paso adicional este script luego de identificar los genes cercanos en el genoma, amplía la información de SPIs recopilando datos de ChEMBL mediante consultas locales en SQL y de PDB a través de extracción vía API.

Los cálculos de Índice de Tanimoto de similitud entre los SPIs conocidos de BacCYPs y los de los genes cercanos se realizaron utilizando RDKit, partiendo de moléculas en formato SMILES.

5. Discusión, perspectivas futuras y aplicaciones

5.1 Discusión general

En este trabajo se presentan dos grandes herramientas que pueden utilizarse tanto de manera independiente como complementaria en la investigación de fármacos novedosos (Schottlender et al. 2022, 2024). Aunque estas metodologías están enfocadas principalmente en la identificación de blancos moleculares de inhibidores para organismos patogénicos, se pueden aplicar a otros contextos diferentes, como el estudio de sustratos y productos de proteínas con relevancia biotecnológica.

La primera herramienta propuesta es una metodología diseñada para identificar blancos terapéuticos candidatos de fármacos con mecanismos de acción desconocidos, un desafío especialmente difícil de abordar si únicamente se utilizaran métodos experimentales. Esta estrategia -que integra aprendizaje automático no supervisado basado en similitud química, *docking* molecular y análisis de esencialidad de posibles blancos- resulta ser una solución relativamente rápida y económica desde el punto de vista computacional. Una de las principales fortalezas del enfoque consiste en su aplicabilidad de manera sistematizada a todo el proteoma de un organismo. Los resultados obtenidos destacan la capacidad para obtener los dominios correctos de unión de ligandos en un porcentaje considerable de los casos evaluados.

Estos resultados son particularmente relevantes en un contexto de recursos limitados para la investigación de fármacos, destacando el potencial de esta herramienta para acelerar el proceso de descubrimiento de nuevos agentes terapéuticos, permitiendo reducir la cantidad de ensayos experimentales y sus costos asociados. Sin embargo, una limitación importante de esta metodología es que en muchos casos, el resultado final es una lista muy extensa de proteínas candidatas, lo que dificulta su interpretación y la selección de blancos concretos para estudios experimentales posteriores. Por esta razón, resulta imprescindible su complemento con enfoques adicionales.

Para abordar esta limitación, se desarrolló una segunda metodología mucho más específica, aplicada a un caso de estudio particular: los Citocromos P450 bacterianos. Aunque probada en este caso particular, el objetivo es que esta herramienta sea aplicable de manera general a diferentes familias de proteínas.

Esta segunda herramienta integra tres enfoques principales para el descubrimiento de sustratos, productos e inhibidores (SPIs) de los Citocromos P450 bacterianos (BacCYPs): 1) el análisis filogenético, 2) el análisis de contexto genético y 3) el *docking*

molecular. Cada uno de estos métodos aporta una perspectiva diferente, y su combinación permite obtener resultados más sólidos y precisos.

El análisis filogenético demostró ser particularmente efectivo cuando se aplica a grupos pequeños de BacCYPs evolutivamente relacionados, permitiendo inferir posibles SPIs basados en similitudes estructurales y funcionales dentro de cada grupo. Este enfoque es especialmente prometedor cuando se cuenta con referencias cercanas de BacCYPs con SPIs previamente caracterizados. Sin embargo, su aplicabilidad al universo completo de BacCYPs es limitada, lo que refuerza la necesidad de investigar subgrupos de proteínas menos estudiadas para mejorar la eficacia de este método.

Por otro lado, el análisis de contexto genético se basó en la identificación de genes vecinos al que codifica el BacCYP de interés, aprovechando la organización común en operones bacterianos. Este enfoque permitió inferir SPIs basándose en las interacciones conocidas de los productos de los genes adyacentes, demostrando una notable capacidad para identificar con precisión SPIs en BacCYPs previamente no caracterizados. A medida que se amplíen las anotaciones genéticas disponibles, se espera que este enfoque gane aún más potencia predictiva.

Finalmente, el *docking* molecular se utilizó tanto para evaluar la probabilidad de unión de los SPIs candidatos como para refinar los resultados obtenidos con los otros dos enfoques. Además, demostró ser una herramienta eficaz para descartar moléculas con baja probabilidad de interacción y priorizar aquellas con mayor relevancia biológica. El uso de técnicas avanzadas como AlphaFold para la generación de modelos tridimensionales de BacCYPs contribuye significativamente en la precisión de este enfoque, al proporcionar estructuras de alta calidad incluso en ausencia de estructuras experimentales.

La integración de estos tres métodos en un flujo de trabajo permite abordar las limitaciones individuales de cada enfoque y maximizar la cantidad y calidad de la información obtenida. Aunque el caso de estudio se centró en los BacCYPs, la flexibilidad y robustez de este protocolo lo hacen aplicable a otras familias de proteínas de interés, contribuyendo significativamente al avance del descubrimiento racional de fármacos y al diseño de estrategias terapéuticas más efectivas.

Las herramientas desarrolladas en este trabajo ofrecen enfoques complementarios que abordan distintos aspectos del descubrimiento de fármacos. La primera metodología proporciona un marco general para identificar blancos específicos de drogas con mecanismos de acción desconocidos, que se puede aplicar eficientemente de manera sistemática a todo el proteoma de un organismo. La segunda, en cambio, está diseñada para profundizar en el análisis de los compuestos químicos que pueden interactuar con las diferentes proteínas de una familia, ofreciendo una caracterización más precisa y detallada de los resultados generados por la primera herramienta. En conjunto, estas metodologías

conforman un enfoque integrado que combina una primera perspectiva amplia con un análisis posterior más focalizado para la identificación y caracterización de interacciones proteína-ligando de interés.

5.2 Perspectivas futuras y limitaciones

A lo largo de este trabajo, surgieron múltiples interrogantes y alternativas metodológicas que, por razones de tiempo y alcance, no pudieron ser abordadas en su totalidad. En esta sección, se presentan algunas de esas ideas y perspectivas futuras que podrían ampliar y fortalecer los enfoques desarrollados en esta tesis.

En primer lugar, respecto de los métodos de agrupamiento de proteínas de función similar, sería adecuado evaluar nuevas estrategias alternativas al *clustering* por dominios de Pfam. Si bien esta estrategia resulta funcional al trabajar con organismos que contienen pocas proteínas (en el orden de los pocos miles) en su proteoma, como las bacterias, en casos de organismos más complejos, que tienen muchas proteínas por familia, los resultados nos traen en algunos casos números imposibles de manejar (por ejemplo más de 300 proteínas por familia). Como posibles abordajes, por un lado, se podría implementar un criterio de agrupamiento más específico que permita reducir el tamaño de los grupos generados, identificando subgrupos funcionales o evolutivos más reducidos dentro de cada familia. Por otro lado, sería valioso integrar métodos de agrupamiento estructural, utilizando métricas como el RMSD entre estructuras tridimensionales.

Por otra parte, respecto a la similaridad y el agrupamiento de compuestos químicos, si bien se utilizó el Índice de Tanimoto basado en *Fingerprints* moleculares como criterio estándar, existen diversos enfoques complementarios o alternativos que se podrían evaluar. Algunos de ellos podrían ser las máximas subestructuras compartidas (Cao, Jiang, and Girke 2008), los andamios de Bemis-Murcko (Malakhov and Pogodin 2024), propiedades físico-químicas de los compuestos, o métodos de análisis más exhaustivo derivados de información bibliográfica.

Por último, es necesario investigar nuevas maneras de filtrar las proteínas obtenidas como posibles blancos de una determinada droga a partir de los dominios candidatos. Si bien se comprobó que el *docking* molecular y la esencialidad contribuyen a una identificación más precisa, no son herramientas que funcionan de manera adecuada en todos los casos. No se conocen con exactitud todas las proteínas esenciales para la supervivencia de cada organismo, y el *docking* molecular presenta muchas simplificaciones (Pagadala, Syed, and Tuszynski 2017), como considerar la estructura de la proteína rígida y no dinámica, o también la limitación de enlaces rotables de cada molécula. Es necesario

incorporar nuevos enfoques, que con el tiempo se irán disponibilizando para la comunidad científica. Muchos de estos están relacionados con técnicas muy avanzadas de aprendizaje automático. Un claro ejemplo resulta ser la predicción de las interacciones entre proteínas y ligandos realizada a través de técnicas avanzadas de aprendizaje profundo, como AlphaFold3 (Abramson et al. 2024), que en caso de liberarse para su uso público, puede resultar en una herramienta mucho más poderosa que el *docking* molecular utilizado en este trabajo.

5.3 Aplicaciones y alcance de los métodos desarrollados

Los trabajos presentados en esta tesis establecen un marco metodológico adaptable para futuras investigaciones en diversos organismos y proteínas.

La primera herramienta, diseñada para predecir blancos de drogas con mecanismos de acción desconocidos, puede extenderse a diferentes contextos. La incorporación del proteoma humano en etapas finales de desarrollo metodológico (García 2024) permite su aplicación en estudios de toxicidad, mecanismos de acción y análisis de *off-targets* en compuestos preclínicos, ampliando su utilidad más allá del campo antimicrobiano.

La segunda herramienta, enfocada en el flujo de trabajo para la predicción de SPIs de Citocromos P450 bacterianos, como se mencionó a lo largo de este trabajo, puede adaptarse a otras familias de proteínas, principalmente en bacterias. Aunque el análisis de contexto genético tiene limitaciones en organismos eucariotas debido a su distinta organización genómica, los enfoques de agrupamiento filogenético para identificar proteínas con funcionalidad similar y las herramientas de biología estructural para predecir la estabilidad de complejos proteína-ligando siguen siendo altamente aplicables y relevantes para expandir su alcance.

En conclusión, esta tesis presenta dos herramientas originales que han demostrado su efectividad en casos de estudio específicos y que pueden ser potenciadas por los continuos avances en inteligencia artificial, bioinformática y quimioinformática.

Anexo: Tablas suplementarias

Tabla suplementaria S3.1: Pares blanco molecular - compuesto inhibitorio conocidos para evaluar los filtros de proteínas candidatas por esencialidad, puntos de estrangulamiento y centralidad.

Organismo	Blanco	Compuesto inhibitorio	Fuente
<i>M.tuberculosis</i>	rpoB	Rifampicina	(Taniguchi et al. 1996)
<i>M.tuberculosis</i>	inhA	Isoniazida, modificada por KatG	(Unissa et al. 2016)
<i>M.tuberculosis</i>	gyrA	Fluoroquinolonas	(Mdluli and Ma 2007)
<i>M.tuberculosis</i>	atpC	Bedaquilina	(Kundu et al. 2016)
<i>M.tuberculosis</i>	mHf	Clofazimina	(Shi et al. 2017)
<i>M.tuberculosis</i>	MenG	GSK1733953A	(Sukheja et al. 2017)
<i>M.tuberculosis</i>	KasA	GSK3011724A	(Trofimov et al. 2018)
<i>M.tuberculosis</i>	MmpL3	GSK2043267A	(Trofimov et al. 2018)
<i>M.tuberculosis</i>	LeuS	GSK3036656	(Shetye, Franzblau, and Cho 2020a)
<i>M.tuberculosis</i>	MurX	Capuramicina	(Siricilla et al. 2015)
<i>K.pneumoniae</i>	LpxH	Sulfonil piperazina	(Cho et al. 2020)
<i>K.pneumoniae</i>	mrcB	Ampicilina (y otros betalactámicos)	(Sutaria et al. 2018)
<i>K.pneumoniae</i>	mrdA	Ampicilina (y otros betalactámicos)	(Shetye, Franzblau, and Cho 2020b)
<i>K.pneumoniae</i>	mrcA	Ampicilina (y otros betalactámicos)	(Sutaria et al. 2018)
<i>K.pneumoniae</i>	pbpC	Ampicilina (y otros betalactámicos)	(Sutaria et al. 2018)
<i>K.pneumoniae</i>	gyrA	Levofloxacin (y otras Fluoroquinolonas)	(Geetha et al., 2020)
<i>K.pneumoniae</i>	gyrB	Levofloxacin (y otras Fluoroquinolonas)	(Geetha et al., 2020)
<i>K.pneumoniae</i>	parC	Levofloxacin (y otras Fluoroquinolonas)	(Geetha et al., 2020)
<i>K.pneumoniae</i>	parE	Levofloxacin (y otras Fluoroquinolonas)	(Geetha et al., 2020)
<i>K.pneumoniae</i>	FabI	Fabimicina	(Parker et al. 2022)

Tabla suplementaria S3.2: Resultados obtenidos con la metodología desarrollada en comparación con los publicados en un estudio de búsqueda de blancos in-silico (Martínez-Jiménez et al. 2013)

Compuesto	Blanco Predicho en el estudio de Martínez-Jiménez et al.	Dominio Pfam del blanco Predicho en el estudio de Martínez-Jiménez et al.	Ranking del blanco reportado obtenido con nuestra metodología	Índice de Tanimoto con el compuesto más similar encontrado en nuestra base de datos
GSK975784A	Rv2182c, Rv2483c	PF01553	2	0.37
GSK975810A	Rv2182c, Rv2483c, Rv2299c	PF01553, PF00775	6	0.33
GSK975840A	Rv2182c, Rv2483c	PF01553	6	0.33
GSK975842A	Rv2182c, Rv2483c, Rv2045c, Rv2139, Rv2299c, Rv2483c	PF01553, PF00775, PF00135, PF01180	6	0.42
GSK547481A	Rv0194	PF00664, PF00005	9	0.29
GSK547490A	Rv0194	PF00664, PF00005	25	0.27
GSK547491A	Rv0194	PF00664, PF00005	7	0.32
GSK547499A	Rv0194	PF00664, PF00005	17	0.26
GSK547500A	Rv0194	PF00664, PF00005	4	0.30
GSK547511A	Rv0194	PF00664, PF00005	8	0.29
GSK547512A	Rv0194	PF00664, PF00005	7	0.32
GSK547527A	Rv1640c, Rv3598c, Rv0194	PF09924, PF00152, PF16995, PF01336, PF00152, PF01336, PF00664, PF00005	nd	nd
GSK547528A	Rv1640c, Rv3598c, Rv0194	PF09924, PF00152, PF16995, PF01336, PF00152, PF01336, PF00664, PF00005	nd	nd
GSK547543A	Rv0194	PF00664, PF00005	7	0.31
GSK1829727A	Rv0053, Rv0379, Rv0650	PF01250, PF07311, PF00480	nd	nd
GSK1829729A	Rv3855, Rv0053, Rv0379, Rv0650	PF00440, PF01250, PF07311, PF00480	nd	nd
GSK1829816A	Rv0053, Rv0379, Rv0650	PF01250, PF07311, PF00480	nd	nd
GSK479031A	Rv0053, Rv0379, Rv0650	PF01250, PF07311, PF00480	nd	nd
GSK957094A	Rv3170, Rv0053, Rv0379, Rv0650	PF01593, PF01250, PF07311, PF00480	4	0.48
GSK1188379A	Rv0194	PF00664, PF00005	15	0.31
GSK1188380A	Rv0194	PF00664, PF00005	16	0.30
GSK1825940A	Rv0194	PF00664, PF00005	2	0.37
GSK1825944A	Rv0194	PF00664, PF00005	1	0.37

BRL-10143SA	Rv1649, Rv2763c, Rv2764c	PF02912, PF01409, PF00186, PF00303	1	0.37
BRL-51093AM	Rv2763c, Rv2764c	PF00186, PF00303	1	0.36
GSK1402290A	Rv1640c, Rv3598c, Rv3834c, Rv3105c, Rv3135	PF09924, PF00152, PF16995, PF01336, PF00152, PF01336, PF02403, PF00587, PF03462, PF00472, PF00823	nd	nd
GSK270671A	Rv1284, Rv3588c, Rv3273, Rv1707	PF00484, PF00484, PF00916, PF01740, PF00916	20	nd
GSK1365028A	Rv0014c	PF03793, PF00069	1	0.42
GSK1598164A	Rv0014c	PF03793, PF00069	1	0.56
GSK275628A	Rv0014c	PF03793, PF00069	3	0.40
GW664700A	Rv0014c	PF03793, PF00069	2	0.34

Tabla suplementaria S3.3: Resultados obtenidos con la metodología desarrollada en comparación con los publicados en ensayos experimentales. (Mugumbate et al. 2015; Trofimov et al. 2018)

Compuesto	Blanco identificado en ensayos experimentales	Dominio Pfam del blanco identificado en ensayos experimentales	Ranking del blanco comprobado experimentalmente obtenido con nuestra metodología	Índice de Tanimoto con el compuesto más similar encontrado
GSK2111534A	DHFR	PF00186	nd	nd
GSK353496A	DHFR	PF00186	nd	nd
BRL-51091AM	DHFR	PF00186	1	0.51
BRL-8903SA	DHFR	PF00186	1	0.38
BRL-8088SA	DHFR	PF00186	1	0.58
GSK957094A	DHFR	PF00186	nd	nd
BRL-7940SA	DHFR	PF00186	1	0.36
BRL-10143SA	DHFR	PF00186	1	0.38
GV187303X	DHFR	PF00186	10	0.40
GSK892651A	DHFR	PF00186	nd	nd
BRL-51093AM	DHFR	PF00186	1	0.36
GSK1829660A	DHFR	PF00186	nd	nd
SB-435634	DHFR	PF00186	18	0.32
GSK381407A	Seq. recG	PF00270, PF00271, PF17191	nd	nd
GSK1829820A	QcrB	PF13631	nd	nd
BRL-10988SA	DHFR	PF00186	1	0.36
GW369335X	DHFR	PF00186	1	1.0
GR135486X	DHFR	PF00186	10	0.36
GSK2043267A	MmpL3	PF03176	nd	nd
SB-354364	DHFR	PF00186	17	0.28
GSK798463A	DHFR	PF00186	nd	nd
GSK345724A	KasA	PF00109, PF02801	1	1.0
GSK1302651A	DHFR	PF00186	19	0.33
GSK921190A	DHFR	PF00186	nd	nd
GSK847920A	MurC	PF01225, PF02875 , PF08245	nd	nd
GSK831784A	MurC	PF01225, PF02875 , PF08245	nd	nd
GSK1826247A	DHFR	PF00186	nd	nd
GSK468214A	DHFR	PF00186	7	0.39
GSK2200150A	MmpL3	PF03176	1	1.0
GW859039X	FbpA	PF00756	nd	nd
GSK124576A	DHFR	PF00186	18	0.36
GSK1758774A	DHFR	PF00186	nd	nd
GW360240X	DHFR	PF00186	1	0.63

GR135487X	DHFR	PF00186	11	0.39
SB-829405	DHFR	PF00186	23	0.28
GSK2032710A	DprE1	PF04030, PF01565	nd	nd
GSK1742694A	Glpk	PF02782, PF00370	nd	nd
GW356807A	Seq. recG	PF00270, PF00271, PF17191	nd	nd
GSK463114A	DHFR	PF00186	15	0.39
GSK731389A	DHFR	PF00186	nd	nd
GSK1365028A	DHFR	PF00186	nd	nd

Tabla suplementaria S4.1: Grupos de sustratos, productos e inhibidores de Citocromos P450 bacterianos en detalle. Se utiliza la nomenclatura de los compuestos correspondiente a su base de datos de origen.

Grupo de SPIs	Base de datos de origen			
	PDB	CHEMBL	CHEBI	KEGG
1.1	DAO, PAM, D12, 12H, EPA, 140, VGJ		CHEBI:141772, CHEBI:30807, CHEBI:70838, CHEBI:37257	pimelate
1.2	MLI, MLA		CHEBI:83955, CHEBI:76928	pyruvate, oxaloacetate, acetoacetate
1.3	RWZ, DXJ		CHEBI:16619, CHEBI:83951, CHEBI:175763, CHEBI:83958, CHEBI:17580	dimethylallyl-PP, farnesyl-PP
1.4	ZMP, ZMO		CHEBI:136878	pimeloyl-CoA, propanoyl-CoA, methylmalonyl-CoA, malonyl-CoA, isobutyryl-CoA
2	PXI, Z18, NRB, XJO, 1D2, QLE, MIV, MY8, EY5, ERB, EPB, EPD, F4O, 9LF, EFO, FLI, E53		CHEBI:76801, CHEBI:63307, CHEBI:63677, CHEBI:16089, CHEBI:64258, CHEBI:27977, CHEBI:80029, CHEBI:31549	erythromycin A
3	EL3, CAM, CAH		CHEBI:51460, CHEBI:51479	
4.1	C3S, STR, TES, C0R, PLO, ASD, NQ8, ESL, A9H	CHEMBL2048335, CHEMBL2048320, CHEMBL2048329, CHEMBL2048327, CHEMBL112570, CHEMBL2048319, CHEMBL2048331, CHEMBL2048328, CHEMBL225111, CHEMBL2048323, CHEMBL63243, CHEMBL2048325, CHEMBL455876, CHEMBL2048322, CHEMBL2048332, CHEMBL2048330, CHEMBL2048321, CHEMBL2048318, CHEMBL2048333, CHEMBL2048316, CHEMBL69851, CHEMBL2048334, CHEMBL2048336, CHEMBL2048326, CHEMBL2048317, CHEMBL2048324	CHEBI:71541, CHEBI:17703, CHEBI:84145, CHEBI:83860, CHEBI:83863, CHEBI:59418, CHEBI:16496, CHEBI:58734, CHEBI:28113, CHEBI:17791, CHEBI:138029	

4.2	VD3, VDY, VDX, 7ZU			
5.1	PFZ, FJQ	CHEMBL808, CHEMBL91		
5.2	M65, 6XD	CHEMBL330775		
6		CHEMBL3804971, CHEMBL3805159, CHEMBL3806115		
7	YTT	CHEMBL4437128, CHEMBL4565125, CHEMBL4471315, CHEMBL4467901	CHEBI:67269	
8	FLV, NQ, 226		CHEBI:58696, CHEBI:77877, CHEBI:77840	
9	1CM, CII, CM9, CM6			
10.1	JZ3, 261, V55, 3DM		CHEBI:141700, CHEBI:141702	catechol
10.2	ANN, 8ZU, 81J, FIV, 81M, TWO, 1F1, EGM, 81P, 4MI, 4IA, 8QM, EH1, 8ZX, EJV, IZP		CHEBI:16579, CHEBI:16639	tyrosine, benzoate, methylbenzoate
10.3	88L, GWM, FW6			
No agrupados	PIY, 17Q, PIW, BEN, DBE, 9AP, PIM, M8N, CMW, TCZ, NCT, CNL, RFA, SPK, TMH, PP9, 6XF, HIS, REA, 5MK, EMZ, ID3, HX2, SRT, 2OH, CRR, P3Z, AUI, JDJ, C8B, C8H, 2CC, II4, 2UO, KLN	CHEMBL104, CHEMBL106	CHEBI:58307	biotin, glutamine, pantothenate, serine, aspartate, quinolate, histidine, avermectin A1a, isopentenyl-PP, cyclooctatin, pentalenolactone, glucose, fructose-6P, tryptophan, staurosporine, calicheamicinone, calicheamicin, arginine, nocardicin A, cysteine, leucine, chorismate, benzoxazolinyl-CoA

Bibliografía

- Abramson, Josh, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, et al. 2024. "Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3." *Nature* 630 (8016): 493–500.
- Allen, William J., Trent E. Balius, Sudipto Mukherjee, Scott R. Brozell, Demetri T. Moustakas, P. Therese Lang, David A. Case, Irwin D. Kuntz, and Robert C. Rizzo. 2015. "DOCK 6: Impact of New Features and Current Docking Performance." *Journal of Computational Chemistry* 36 (15): 1132–56.
- Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25 (17): 3389–3402.
- Andries, Koen, Peter Verhasselt, Jerome Guillemont, Hinrich W. H. Göhlmann, Jean-Marc Neefs, Hans Winkler, Jef Van Gestel, et al. 2005. "A Diarylquinoline Drug Active on the ATP Synthase of Mycobacterium Tuberculosis." *Science (New York, N.Y.)* 307 (5707): 223–27.
- Anstein, Stefanie, and Gerhard Kremer. 2005. *Analysing Names of Organic Chemical Compounds: From Morpho-Semantics to SMILES Strings and Classes*.
- "Applications of Microbial Cytochrome P450 Enzymes in Biotechnology and Synthetic Biology." 2016. *Current Opinion in Chemical Biology* 31 (April):136–45.
- Arcon, Juan Pablo, Carlos P. Modenutti, Demian Avendaño, Elias D. Lopez, Lucas A. Defelipe, Francesca Alessandra Ambrosio, Adrian G. Turjanski, Stefano Forli, and Marcelo A. Marti. 2019. "AutoDock Bias: Improving Binding Mode Prediction and Virtual Screening Using Known Protein–ligand Interactions." *Bioinformatics* 35 (19): 3836–38.
- Arcon, Juan Pablo, Adrián G. Turjanski, Marcelo A. Martí, and Stefano Forli. 2021. "Biased Docking for Protein-Ligand Pose Prediction." *Methods in Molecular Biology* 2266:39–72.
- "A Study on Cheminformatics and Its Applications on Modern Drug Discovery." 2012. *Procedia Engineering* 38 (January):1264–75.
- Bairoch, Amos, and Rolf Apweiler. 2000. "The SWISS-PROT Protein Sequence Database and Its Supplement TrEMBL in 2000." *Nucleic Acids Research* 28 (1): 45–48.
- Bajusz, Dávid, Anita Rácz, and Károly Héberger. 2015. "Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations?" *Journal of Cheminformatics* 7 (1): 1–13.
- Bansal, Parit, Anne Morgat, Kristian B. Axelsen, Venkatesh Muthukrishnan, Elisabeth Coudert, Lucila Aimo, Nevila Hyka-Nouspikel, et al. 2022. "Rhea, the Reaction Knowledgebase in 2022." *Nucleic Acids Research* 50 (D1): D693–700.
- Barclay, W. R., R. H. Ebert, and D. Kochwaser. 1953. "Mode of Action of Isoniazid." *American Review of Tuberculosis* 67 (4): 490–96.
- Behrendorff, James B. Y. H. 2021. "Reductive Cytochrome P450 Reactions and Their Potential Role in Bioremediation." *Frontiers in Microbiology* 12 (April):649273.
- Behzadi, Payam, and Márió Gajdács. 2021. "Worldwide Protein Data Bank (wwPDB): A Virtual Treasure for Research in Biotechnology." *European Journal of Microbiology & Immunology* 11 (4): 77–86.
- Belcher, James, Kirsty J. McLean, Sarah Matthews, Laura S. Woodward, Karl Fisher, Stephen E. J. Rigby, David R. Nelson, et al. 2014. "Structure and Biochemical Properties of the Alkene Producing Cytochrome P450 OleTJE (CYP152L1) from the *Jeotgalicoccus* Sp. 8456 Bacterium." *The Journal of Biological Chemistry* 289 (10): 6535–50.
- Bento, A. Patrícia, Anna Gaulton, Anne Hersey, Louisa J. Bellis, Jon Chambers, Mark Davies, Felix A. Krüger, et al. 2014. "The ChEMBL Bioactivity Database: An Update." *Nucleic Acids Research* 42 (Database issue): D1083–90.
- Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. "The Protein Data Bank." *Nucleic Acids Research* 28 (1): 235–42.
- Bhatia, Nitin, and Vandana. 2010. "Survey of Nearest Neighbor Techniques." <http://arxiv.org/abs/1007.0085>.
- Blum, Matthias, Antonina Andreeva, Laise Cavalcanti Florentino, Sara Rocio Chuguransky, Tiago Grego, Emma Hobbs, Beatriz Lazaro Pinto, et al. 2025. "InterPro: The Protein Sequence Classification Resource in 2025." *Nucleic Acids Research* 53 (D1): D444–56.
- Böcker†‡, Alexander, †. Svetlana Derksen, †. Elena Schmidt, ‡. And Andreas Teckentrup, and †. Gisbert Schneider*. 2005. "A Hierarchical Clustering Approach for Large Compound Libraries," May. <https://doi.org/10.1021/ci0500029>.

- Bogusz, Marcin, and Simon Whelan. 2016. "Phylogenetic Tree Estimation With and Without Alignment: New Distance Methods and Benchmarking." *Systematic Biology* 66 (2): 218–31.
- Boratyn, Grzegorz M., Christiam Camacho, Peter S. Cooper, George Coulouris, Amelia Fong, Ning Ma, Thomas L. Madden, et al. 2013. "BLAST: A More Efficient Report with Usability Improvements." *Nucleic Acids Research* 41 (Web Server issue): W29–33.
- Brown, Rebecca M., Sameer Farouk Sait, Griffin Dunn, Alanna Sullivan, Benjamin Bruckert, and Daochun Sun. 2022. "Integrated Drug Mining Reveals Actionable Strategies Inhibiting Plexiform Neurofibromas." *Brain Sciences* 12 (6). <https://doi.org/10.3390/brainsci12060720>.
- Bustamante, Juan P., Leandro Radusky, Leonardo Boechi, Darío A. Estrin, Arjen ten Have, and Marcelo A. Martí. 2016. "Evolutionary and Functional Relationships in the Truncated Hemoglobin Family." *PLoS Computational Biology* 12 (1): e1004701.
- Butina, Darko. 1999. "Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets," June. <https://doi.org/10.1021/ci9803381>.
- Callaway, Ewen. 2024. "Chemistry Nobel Goes to Developers of AlphaFold AI That Predicts Protein Structures." *Nature* 634 (8034): 525–26.
- Cardona, Gabriel, Arnau Mir, Francesc Rosselló, Lucía Rotger, and David Sánchez. 2013. "Cophenetic Metrics for Phylogenetic Trees, after Sokal and Rohlf." *BMC Bioinformatics* 14 (January):3.
- Cha, Gi-Wook, Hyeun Jun Moon, Young-Min Kim, Won-Hwa Hong, Jung-Ha Hwang, Won-Jun Park, and Young-Chan Kim. 2020. "Development of a Prediction Model for Demolition Waste Generation Using a Random Forest Algorithm Based on Small DataSets." *International Journal of Environmental Research and Public Health* 17 (19). <https://doi.org/10.3390/ijerph17196997>.
- Chavali, Arvind K., and Seung Y. Rhee. 2018. "Bioinformatics Tools for the Identification of Gene Clusters That Biosynthesize Specialized Metabolites." *Briefings in Bioinformatics* 19 (5): 1022–34.
- Chen, Hongming, Thierry Kogej, and Ola Engkvist. 2018. "Cheminformatics in Drug Discovery, an Industrial Perspective." *Molecular Informatics* 37 (9-10): e1800041.
- Cho, Jae, Minhee Lee, C. Skyler Cochrane, Caroline G. Webster, Benjamin A. Fenton, Jinshi Zhao, Jiyong Hong, and Pei Zhou. 2020. "Structural Basis of the UDP-Diacylglycerolamine Pyrophosphohydrolase LpxH Inhibition by Sulfonyl Piperazine Antibiotics." *Proceedings of the National Academy of Sciences of the United States of America* 117 (8): 4109–16.
- Clemente, Camila M., Juan M. Prieto, and Marcelo Martí. 2024. "Unlocking Precision Docking for Metalloproteins." *Journal of Chemical Information and Modeling* 64 (5): 1581–92.
- Criscuolo, Alexis, and Simonetta Gribaldo. 2010. "BMGE (Block Mapping and Gathering with Entropy): A New Software for Selection of Phylogenetic Informative Regions from Multiple Sequence Alignments." *BMC Evolutionary Biology* 10 (July):210.
- "Cross-Validation Methods." 2000. *Journal of Mathematical Psychology* 44 (1): 108–32.
- "Cytochromes P450: History, Classes, Catalytic Mechanism, and Industrial Application." 2016. In *Advances in Protein Chemistry and Structural Biology*, 105:105–26. Academic Press.
- Darabi, Maryam, Samin Seddigh, and Mina Abarshahr. 2017. "Structural, Functional, and Phylogenetic Studies of Cytochrome P450 (CYP) Enzyme in Seed Plants by Bioinformatics Tools." *Caryologia*, January. <https://doi.org/10.1080/00087114.2016.1274866>.
- Dara, Suresh, Swetha Dhamecherla, Surender Singh Jadav, C. H. Madhu Babu, and Mohamed Jawed Ahsan. 2021. "Machine Learning in Drug Discovery: A Review." *Artificial Intelligence Review* 55 (3): 1947–99.
- Defelipe, Lucas A., Dario Fernández Do Porto, Pablo Ivan Pereira Ramos, Marisa Fabiana Nicolás, Ezequiel Sosa, Leandro Radusky, Esteban Lanzarotti, Adrián G. Turjanski, and Marcelo A. Martí. 2016. "A Whole Genome Bioinformatic Approach to Determine Potential Latent Phase Specific Targets in Mycobacterium Tuberculosis." *Tuberculosis* 97 (March):181–92.
- Defelipe, Lucas A., Dario Fernandez Do Porto, Pablo Ivan Pereira Ramos, Marisa Fabiana Nicolas, Ezequiel Sosa, Leandro Radusky, Esteban Lanzarotti, Adrian G. Turjanski, and Marcelo A. Martí. 2018. "Corrigendum to 'A Whole Genome Bioinformatic Approach to Determine Potential Latent Phase Specific Targets in Mycobacterium Tuberculosis' [Tuberculosis 97 (March 2016) 181-192]." *Tuberculosis* 110 (May):121.
- DeJesus, Michael A., Elias R. Gerrick, Weizhen Xu, Sae Woong Park, Jarukit E. Long, Cara C. Boutte, Eric J. Rubin, et al. 2017. "Comprehensive Essentiality Analysis of the Mycobacterium Tuberculosis Genome via Saturating Transposon Mutagenesis." *mBio* 8 (1). <https://doi.org/10.1128/mBio.02133-16>.
- Eddy, S. R. 1998. "Profile Hidden Markov Models." *Bioinformatics (Oxford, England)* 14 (9): 755–63.

- Edwards, David, Jason Stajich, and David Hansen. 2009. *Bioinformatics: Tools and Applications*. Springer Science & Business Media.
- Engel*, Thomas. 2006. "Basic Overview of Chemoinformatics†," November. <https://doi.org/10.1021/ci600234z>.
- Engkvist, Ola, Per-Ola Norrby, Nidhal Selmi, Yu-Hong Lam, Zhengwei Peng, Edward C. Sherer, Willi Amberg, Thomas Erhard, and Lynette A. Smyth. 2018. "Computational Prediction of Chemical Reactions: Current Status and Outlook." *Drug Discovery Today* 23 (6): 1203–18.
- Eramian, David, Narayanan Eswar, Min-Yi Shen, and Andrej Sali. 2008. "How Well Can the Accuracy of Comparative Protein Structure Models Be Predicted?" *Protein Science : A Publication of the Protein Society* 17 (11): 1881–93.
- Farfán-López, Mariella, Abraham Espinoza-Culupú, Ruth García-de-la-Guarda, Federico Serral, Ezequiel Sosa, María Mercedes Palomino, and Darío A. Fernández Do Porto. 2020. "Prioritisation of Potential Drug Targets against Bartonella Bacilliformis by an Integrative in-Silico Approach." *Memorias Do Instituto Oswaldo Cruz* 115 (August): e200184.
- Finn, Robert D., Jaina Mistry, John Tate, Penny Coghill, Andreas Heger, Joanne E. Pollington, O. Luke Gavin, et al. 2010. "The Pfam Protein Families Database." *Nucleic Acids Research* 38 (Database issue): D211–22.
- "Fixed-Radius near Neighbors Search." 1991. *Information Processing Letters* 39 (4): 201–3.
- Flynt, Abby, and Nema Dean. 2016. "A Survey of Popular R Packages for Cluster Analysis." *Journal of Educational and Behavioral Statistics: A Quarterly Publication Sponsored by the American Educational Research Association and the American Statistical Association* 41 (2): 205–25.
- Forli, Stefano, and Arthur J. Olson. 2012. "A Force Field with Discrete Displaceable Waters and Desolvation Entropy for Hydrated Ligand Docking." *Journal of Medicinal Chemistry* 55 (2): 623–38.
- Friedman*†, Mendel, †. Bruce E. Mackey, †. Hyun-Jeong Kim, †. In-Seon Lee, and †. Kap-Rang Lee. 2006. "Structure–Activity Relationships of Tea Compounds against Human Cancer Cells," December. <https://doi.org/10.1021/jf062276h>.
- "From Protein Structure to Function." 1999. *Current Opinion in Structural Biology* 9 (3): 374–82.
- "Functional Expression and Regulation of Eukaryotic Cytochrome P450 Enzymes in Surrogate Microbial Cell Factories." 2022. *Engineering Microbiology* 2 (1): 100011.
- Gallina, Anna M., Peer Bork, and Domenico Bordo. 2014. "Structural Analysis of Protein-Ligand Interactions: The Binding of Endogenous Compounds and of Synthetic Drugs." *Journal of Molecular Recognition: JMR* 27 (2): 65–72.
- García, L. Gabriel. 2024. "Estudio sobre la inhibición de la actividad ATPasa de HSP90." Tesis de licenciatura, Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales.
- Gaulton, Anna, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, Yvonne Light, et al. 2012. "ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery." *Nucleic Acids Research* 40 (Database issue): D1100–1107.
- Gauthier, Jeff, Antony T. Vincent, Steve J. Charette, and Nicolas Derome. 2018. "A Brief History of Bioinformatics." *Briefings in Bioinformatics* 20 (6): 1981–96.
- Gauto, Diego F., Ariel A. Petruk, Carlos P. Modenutti, Juan I. Blanco, Santiago Di Lella, and Marcelo A. Martí. 2013. "Solvent Structure Improves Docking Prediction in Lectin-Carbohydrate Complexes." *Glycobiology* 23 (2): 241–58.
- Geetha, Pacha Venkataramana, Kayanam Vijaya Lalitha Aishwarya, Shanthi Mariappan, and Uma Sekar. 2020 "Fluoroquinolone Resistance in Clinical Isolates of Klebsiella Pneumoniae." *Journal of Laboratory Physicians* 12 (2): 121–25. Accessed November 28, 2024.
- Geldenhuys, Werner J., Sarel F. Malan, Jeffrey R. Bloomquist, Alan P. Marchand, and Cornelis J. Van der Schyf. 2005. "Pharmacology and Structure-Activity Relationships of Bioactive Polycyclic Cage Compounds: A Focus on Pentacycloundecane Derivatives." *Medicinal Research Reviews* 25 (1): 21–48.
- Gilson, Michael K., Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. 2015. "BindingDB in 2015: A Public Database for Medicinal Chemistry, Computational Chemistry and Systems Pharmacology." *Nucleic Acids Research* 44 (D1): D1045–53.
- Gimeno, Aleix, María José Ojeda-Montes, Sarah Tomás-Hernández, Adrià Cereto-Massagué, Raúl Beltrán-Debón, Miquel Mulero, Gerard Pujadas, and Santiago Garcia-Vallvé. 2019. "The Light and Dark Sides of Virtual Screening: What Is There to Know?" *International Journal of Molecular Sciences* 20 (6). <https://doi.org/10.3390/ijms20061375>.
- Greule, Anja, Jeanette E. Stok, James J. De Voss, and Max J. Cryle. 2018. "Unrivalled Diversity: The Many Roles and Reactions of Bacterial Cytochromes P450 in Secondary Metabolism." *Natural Product Reports* 35 (8): 757–91.

- Guengerich, F. P. 1995. "Cytochrome P450 Proteins and Potential Utilization in Biodegradation." *Environmental Health Perspectives* 103 Suppl 5 (Suppl 5): 25–28.
- Guengerich, F. Peter. 2020. "A History of the Roles of Cytochrome P450 Enzymes in the Toxicity of Drugs." *Toxicological Research* 37 (1): 1–23.
- Guindon, Stéphane, Frédéric Delsuc, Jean-François Dufayard, and Olivier Gascuel. 2009. "Estimating Maximum Likelihood Phylogenies with PhyML." *Methods in Molecular Biology* 537:113–37.
- Gu, Jenny, and Philip E. Bourne. 2009. *Structural Bioinformatics*. John Wiley & Sons.
- Gütlein, Martin, and Stefan Kramer. 2016. "Filtered Circular Fingerprints Improve Either Prediction or Runtime Performance While Retaining Interpretability." *Journal of Cheminformatics* 8 (October):60.
- Habibie, Fathurrochman, Afiahayati, Guntur Budi Herwanto, Sri Hartati, and Aufaclav Zatu Kusuma Frisky. 2018. "A Parallel ClustalW Algorithm on Multi-Raspberry Pis for Multiple Sequence Alignment." In *2018 1st International Conference on Bioinformatics, Biotechnology, and Biomedical Engineering - Bioinformatics and Biomedical Engineering*. IEEE. <https://doi.org/10.1109/biomic.2018.8610612>.
- Hastings, Janna, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. 2016. "ChEBI in 2016: Improved Services and an Expanding Collection of Metabolites." *Nucleic Acids Research* 44 (D1): D1214–19.
- Heidary, Mohsen, Azar Dohkt Khosravi, Saeed Khoshnood, Mohammad Javad Nasiri, Saleh Soleimani, and Mehdi Goudarzi. 2017. "Daptomycin." *Journal of Antimicrobial Chemotherapy* 73 (1): 1–11.
- Hersey, Anne. 2013. "ChEMBL Deposited Data Set - GSK TB Set." EMBL-EBI. <https://doi.org/10.6019/chembl2095176>.
- Hood, Leroy, and Lee Rowen. 2013. "The Human Genome Project: Big Science Transforms Biology and Medicine." *Genome Medicine* 5 (9): 1–8.
- Hooper, David C., and George A. Jacoby. 2016. "Topoisomerase Inhibitors: Fluoroquinolone Mechanisms of Action and Resistance." *Cold Spring Harbor Perspectives in Medicine* 6 (9). <https://doi.org/10.1101/cshperspect.a025320>.
- Hughes, Diarmaid, and Anders Karlén. 2014. "Discovery and Preclinical Development of New Antibiotics." *Upsala Journal of Medical Sciences* 119 (2): 162–69.
- Hughes, William Howard. 1974. *Alexander Fleming and Penicillin*. Hodder Wayland.
- Hu, Liegi, Mark L. Benson, Richard D. Smith, Michael G. Lerner, and Heather A. Carlson. 2005. "Binding MOAD (Mother of All Databases)." *Proteins* 60 (3): 333–40.
- Hu, Ye, and Jürgen Bajorath. 2015. "Quantifying the Tendency of Therapeutic Target Proteins to Bind Promiscuous or Selective Compounds." *PLoS One* 10 (5): e0126838.
- "Initial Sequencing and Analysis of the Human Genome." 2001. *Nature* 409 (6822): 860–921.
- Janela, Tiago, Kosuke Takeuchi, and Jürgen Bajorath. 2022. "Introducing a Chemically Intuitive Core-Substituent Fingerprint Designed to Explore Structural Requirements for Effective Similarity Searching and Machine Learning." *Molecules* 27 (7): 2331.
- Jeong, Hawoong, Zoltan N. Oltvai, and Albert-László Barabási. 2002. "Prediction of Protein Essentiality Based on Genomic Data." *Complexity* 1 (1): 19–28.
- Joshi, Rajendra P., and Neeraj Kumar. 2021. "Artificial Intelligence for Autonomous Molecular Design: A Perspective." *Molecules (Basel, Switzerland)* 26 (22). <https://doi.org/10.3390/molecules26226761>.
- Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature* 596 (7873): 583–89.
- Kanehisa, Minoru, and Susumu Goto. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research* 28 (1): 27–30.
- Kapli, Paschalia, Ziheng Yang, and Maximilian J. Telford. 2020. "Phylogenetic Tree Building in the Genomic Age." *Nature Reviews. Genetics* 21 (7): 428–44.
- Katoh, Kazutaka, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. 2002. "MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform." *Nucleic Acids Research* 30 (14): 3059–66.
- Katz, Leonard, and Richard H. Baltz. 2016. "Natural Product Discovery: Past, Present, and Future." *Journal of Industrial Microbiology & Biotechnology* 43 (2-3): 155–76.
- Kelly, Steven L., and Diane E. Kelly. 2013. "Microbial Cytochromes P450: Biodiversity and Biotechnology. Where Do Cytochromes P450 Come From, What Do They Do and What Can They Do for Us?" *Philosophical Transactions of the Royal Society of London. Series B, Biological*

- Sciences* 368 (1612): 20120476.
- Khmelevtsova, L. E., I. S. Sazykin, M. A. Sazykina, and E. Yu Seliverstova. 2017. "Prokaryotic Cytochromes P450 (Review)." *Applied Biochemistry and Microbiology* 53 (4): 401–9.
- Kim, Stephanie S., Melanie L. Aprahamian, and Steffen Lindert. 2019. "Improving Inverse Docking Target Identification with Z-Score Selection." *Chemical Biology & Drug Design* 93 (6): 1105–16.
- Kim, Sunghwan, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, et al. 2015. "PubChem Substance and Compound Databases." *Nucleic Acids Research* 44 (D1): D1202–13.
- Klabunde, T. 2007. "Chemogenomic Approaches to Drug Discovery: Similar Receptors Bind Similar Ligands." *British Journal of Pharmacology* 152 (1): 5–7.
- Kocaoglu, Ozden, and Erin E. Carlson. 2015. "Profiling of β -Lactam Selectivity for Penicillin-Binding Proteins in Escherichia Coli Strain DC2." *Antimicrobial Agents and Chemotherapy* 59 (5): 2785–90.
- Kouranov, Andrei, Lei Xie, Joanna de la Cruz, Li Chen, John Westbrook, Philip E. Bourne, and Helen M. Berman. 2006. "The RCSB PDB Information Portal for Structural Genomics." *Nucleic Acids Research* 34 (suppl_1): D302–5.
- Kruger, Felix A., Raghd Rostom, and John P. Overington. 2012. "Mapping Small Molecule Binding Data to Structural Domains." *BMC Bioinformatics* 13 Suppl 17 (Suppl 17): S11.
- Kundu, Subhashri, Goran Biukovic, Gerhard Grüber, and Thomas Dick. 2016. "Bedaquiline Targets the ϵ Subunit of Mycobacterial F-ATP Synthase." *Antimicrobial Agents and Chemotherapy* 60 (11): 6977–79.
- Lange, Justus Johann, Andrea Anelli, Jochem Alsenz, Martin Kuentz, Patrick J. O'Dwyer, Wiebke Saal, Nicole Wyttenbach, and Brendan T. Griffin. 2024. "Comparative Analysis of Chemical Descriptors by Machine Learning Reveals Atomistic Insights into Solute-Lipid Interactions." *Molecular Pharmaceutics* 21 (7): 3343–55.
- Lefort, Vincent, Jean-Emmanuel Longueville, and Olivier Gascuel. 2017. "SMS: Smart Model Selection in PhyML." *Molecular Biology and Evolution* 34 (9): 2422–24.
- Le Guilloux, Vincent, Peter Schmidtke, and Pierre Tuffery. 2009. "Fpocket: An Open Source Platform for Ligand Pocket Detection." *BMC Bioinformatics* 10 (June):168.
- Lipkowitz, Kenny B. 2002. *Reviews in Computational Chemistry: Volume 18*. PDF. Edited by Kenny B. Lipkowitz and Donald B. Boyd. Reviews in Computational Chemistry. Nashville, TN: John Wiley & Sons.
- Li, Weizhong, and Adam Godzik. 2006. "Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences." *Bioinformatics* 22 (13): 1658–59.
- Lobo-Silva, Jessica, Fernanda J. Cabral, Murilo S. Amaral, Patrícia A. Miyasato, Rafaela Paula de Freitas, Adriana S. A. Pereira, Mariana I. Khouri, et al. 2020. "The Antischistosomal Potential of GSK-34, an H3K27 Demethylase Inhibitor: Insights from Molecular Modeling, Transcriptomics and in Vitro Assays." *Parasites & Vectors* 13 (1): 140.
- Luo, Hao, Yan Lin, Tao Liu, Fei-Liao Lai, Chun-Ting Zhang, Feng Gao, and Ren Zhang. 2021. "DEG 15, an Update of the Database of Essential Genes That Includes Built-in Analysis Tools." *Nucleic Acids Research* 49 (D1): D677–86.
- Mariani, Valerio, Marco Biasini, Alessandro Barbato, and Torsten Schwede. 2013. "IDDT: A Local Superposition-Free Score for Comparing Protein Structures and Models Using Distance Difference Tests." *Bioinformatics (Oxford, England)* 29 (21): 2722–28.
- Martínez-Jiménez, Francisco, George Papadatos, Lun Yang, Iain M. Wallace, Vinod Kumar, Ursula Pieper, Andrej Sali, James R. Brown, John P. Overington, and Marc A. Marti-Renom. 2013. "Target Prediction for an Open Access Set of Compounds Active against Mycobacterium Tuberculosis." *PLoS Computational Biology* 9 (10): e1003253.
- Mdluli, Khisimuzi, and Zhenkun Ma. 2007. "Mycobacterium Tuberculosis DNA Gyrase as a Target for Drug Discovery." *Infectious Disorders Drug Targets* 7 (2): 159–68.
- "Methods for Protein Homology Modelling." 2006. In *Applied Mycology and Biotechnology*, 6:37–59. Elsevier.
- Mirdita, Milot, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. 2022. "ColabFold: Making Protein Folding Accessible to All." *Nature Methods* 19 (6): 679–82.
- "Modeller: Generation and Refinement of Homology-Based Protein Structure Models." 2003. In *Methods in Enzymology*, 374:461–91. Academic Press.
- Morgan, H. L. 2002. "The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service," May. <https://doi.org/10.1021/c160017a018>.

- Morris, Garrett M., Ruth Huey, William Lindstrom, Michel F. Sanner, Richard K. Belew, David S. Goodsell, and Arthur J. Olson. 2009. "AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility." *Journal of Computational Chemistry* 30 (16): 2785–91.
- Muegge, Ingo, and Prasenjit Mukherjee. 2016. "An Overview of Molecular Fingerprint Similarity Search in Virtual Screening." *Expert Opinion on Drug Discovery*, February. <https://doi.org/10.1517/17460441.2016.1117070>.
- Mugumbate, Grace, Katherine A. Abrahams, Jonathan A. G. Cox, George Papadatos, Gerard van Westen, Joël Lelièvre, Szymon T. Calus, et al. 2015. "Mycobacterial Dihydrofolate Reductase Inhibitors Identified Using Chemogenomic Methods and in Vitro Validation." *PloS One* 10 (3): e0121492.
- Murtagh, Fionn, and Pedro Contreras. 2012. "Algorithms for Hierarchical Clustering: An Overview." *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery* 2 (1): 86–97.
- Nakamura, Tsukasa, Kazunori D. Yamada, Kentaro Tomii, and Kazutaka Katoh. 2018. "Parallelization of MAFFT for Large-Scale Multiple Sequence Alignments." *Bioinformatics (Oxford, England)* 34 (14): 2490–92.
- Nelson, David R. 2006. "Cytochrome P450 Nomenclature, 2004." *Cytochrome P450 Protocols*, 1–10.
- Nguyen, Ken, Xuan Guo, and Yi Pan. 2016. *Multiple Biological Sequence Alignment: Scoring Functions, Algorithms and Evaluation*. John Wiley & Sons.
- Osbourne, Anne E., and Ben Field. 2009. "Operons." *Cellular and Molecular Life Sciences: CMLS* 66 (23): 3755–75.
- Overbeek, R., M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev. 1999. "The Use of Gene Clusters to Infer Functional Coupling." *Proceedings of the National Academy of Sciences of the United States of America* 96 (6): 2896–2901.
- Pagadala, Nataraj S., Khajamohiddin Syed, and Jack Tuszynski. 2017. "Software for Molecular Docking: A Review." *Biophysical Reviews* 9 (2): 91–102.
- Palumbo, Miranda, Ezequiel Sosa, Florencia Castello, Gustavo Schottlender, Federico Serral, Adrián Turjanski, María Mercedes Palomino, and Darío Fernández Do Porto. 2022. "Integrating Diverse Layers of Omic Data to Identify Novel Drug Targets in *Listeria Monocytogenes*." *Frontiers in Drug Discovery* 2 (September):969415.
- Paradis, Emmanuel, and Klaus Schliep. 2019. "Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R." *Bioinformatics* 35 (3): 526–28.
- Parker, Erica N., Brett N. Cain, Behnush Hajian, Rebecca J. Ulrich, Emily J. Geddes, Sulyman Barkho, Hyang Yeon Lee, et al. 2022. "An Iterative Approach Guides Discovery of the FabI Inhibitor Fabimycin, a Late-Stage Antibiotic Candidate with Efficacy against Drug-Resistant Gram-Negative Infections." *ACS Central Science* 8 (8): 1145–58.
- "Performance Evaluation of Classification Algorithms by K-Fold and Leave-One-out Cross Validation." 2015. *Pattern Recognition* 48 (9): 2839–46.
- Peter Guengerich*, F. 2001. "Common and Uncommon Cytochrome P450 Reactions Related to Metabolism and Chemical Toxicity," May. <https://doi.org/10.1021/tx0002583>.
- Peter Guengerich, F. 2018. "Mechanisms of Cytochrome P450-Catalyzed Oxidations." *ACS Catalysis*, October. <https://doi.org/10.1021/acscatal.8b03401>.
- Pevsner, Jonathan. 2015. *Bioinformatics and Functional Genomics*. John Wiley & Sons.
- Pinto, Lilian, Alejandro G. Schijman, Julio Alonso-Padilla, Daniel Lozano, Mary Cruz Torrico, Pietro Gamba, Margarita Torrez, et al. 2022. "Molecular Detection and Parasite Load of *Trypanosoma Cruzi* in Digestive Tract Tissue of Chagas Disease Patients Affected by Megacolon." *Acta Tropica* 235 (November):106632.
- Potter, Simon C., Aurélien Luciani, Sean R. Eddy, Youngmi Park, Rodrigo Lopez, and Robert D. Finn. 2018. "HMMER Web Server: 2018 Update." *Nucleic Acids Research* 46 (W1): W200–204.
- Prieto, Juan Manuel, Gustavo Schottlender, Camila M. Clemente, Rafael Betanzos, Dario Fernández Do Porto, and Marcelo A. Martí. 2024. "Docking and Bias Docking." In *Computer-Aided Drug Discovery and Design*, 127–48. Cham: Springer International Publishing.
- Rabiner, L. R. 1989. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers* 77 (2): 257–86.
- Radusky, Leandro, Sergio Ruiz-Carmona, Carlos Modenutti, Xavier Barril, Adrian G. Turjanski, and Marcelo A. Martí. 2017. "LigQ: A Webserver to Select and Prepare Ligands for Virtual Screening." *Journal of Chemical Information and Modeling* 57 (8): 1741–46.
- Rahman, Syed Asad, and Dietmar Schomburg. 2006. "Observing Local and Global Properties of Metabolic Pathways: 'load Points' and 'choke Points' in the Metabolic Networks." *Bioinformatics* 22 (14): 1767–74.

- Ramos, Pablo Ivan Pereira, Darío Fernández Do Porto, Esteban Lanzarotti, Ezequiel J. Sosa, Germán Burguener, Agustín M. Pardo, Cecilia C. Klein, et al. 2018. "An Integrative, Multi-Omics Approach towards the Prioritization of Klebsiella Pneumoniae Drug Targets." *Scientific Reports* 8 (1): 10755.
- Ray, Susmita. 2019. "A Quick Review of Machine Learning Algorithms." In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*. IEEE. <https://doi.org/10.1109/comitcon.2019.8862451>.
- Reddy, Karnati Konda, Sanjeev Kumar Singh, Sunil Kumar Tripathi, Chandrabose Selvaraj, and Venkatesan Suryanarayanan. 2013. "Shape and Pharmacophore-Based Virtual Screening to Identify Potential Cytochrome P450 Sterol 14 α -Demethylase Inhibitors." *Journal of Receptor and Signal Transduction Research* 33 (4): 234–43.
- Ribeiro da Cunha, Bernardo, Luís P. Fonseca, and Cecilia R. C. Calado. 2019. "Antibiotic Discovery: Where Have We Come From, Where Do We Go?" *Antibiotics (Basel, Switzerland)* 8 (2). <https://doi.org/10.3390/antibiotics8020045>.
- Rivara-Espasandín, Martín, Miranda Clara Palumbo, Ezequiel J. Sosa, Santiago Radío, Adrián G. Turjanski, José Sotelo-Silveira, Darío Fernández Do Porto, and Pablo Smircich. 2023. "Omics Data Integration Facilitates Target Selection for New Antiparasitic Drugs against TriTryp Infections." *Frontiers in Pharmacology* 14 (April):1136321.
- Rogers, David, and Mathew Hahn. 2010. "Extended-Connectivity Fingerprints," April. <https://doi.org/10.1021/ci100050t>.
- Rozewicki, John, Songling Li, Karlou Mar Amada, Daron M. Standley, and Kazutaka Katoh. 2019. "MAFFT-DASH: Integrated Protein Sequence and Structural Alignment." *Nucleic Acids Research* 47 (W1): W5–10.
- Rychlewski, Leszek, Weizhong Li, Lukasz Jaroszewski, and Adam Godzik. 2000. "Comparison of Sequence Profiles. Strategies for Structural Predictions Using Sequence Information." *Protein Science: A Publication of the Protein Society* 9 (2): 232–41.
- Santos, Lucianna H. S., Rafaela S. Ferreira, and Ernesto R. Caffarena. 2019. "Integrating Molecular Docking and Molecular Dynamics Simulations." *Docking Screens for Drug Discovery*, 13–34.
- Santos-Martins, Diogo, Leonardo Solis-Vasquez, Andreas F. Tillack, Michel F. Sanner, Andreas Koch, and Stefano Forli. 2021. "Accelerating AutoDock4 with GPUs and Gradient-Based Local Search." *Journal of Chemical Theory and Computation* 17 (2): 1060–73.
- Schottlender, Gustavo, Juan Manuel Prieto, Camila Clemente, Claudio David Schuster, Victoria Dumas, Darío Fernández Do Porto, and Marcelo Adrian Martí. 2024. "Bacterial Cytochrome P450s: A Bioinformatics Odyssey of Substrate Discovery." *Frontiers in Microbiology* 15 (February):1343029.
- Schottlender, Gustavo, Juan Manuel Prieto, Miranda Clara Palumbo, Florencia A. Castello, Federico Serral, Ezequiel J. Sosa, Adrián G. Turjanski, Marcelo A. Martí, and Darío Fernández Do Porto. 2022. "From Drugs to Targets: Reverse Engineering the Virtual Screening Process on a Proteomic Scale." *Frontiers in Drug Discovery* 2 (October):969983.
- Schuster-Böckler, Benjamin, Jörg Schultz, and Sven Rahmann. 2004. "HMM Logos for Visualization of Protein Families." *BMC Bioinformatics* 5 (1): 1–8.
- Serral, Federico, Florencia A. Castello, Ezequiel J. Sosa, Agustín M. Pardo, Miranda Clara Palumbo, Carlos Modenutti, María Mercedes Palomino, et al. 2021. "From Genome to Drugs: New Approaches in Antimicrobial Discovery." *Frontiers in Pharmacology* 12 (June):647060.
- Sharma, Akansha, Shailesh Jaloree, and Ramjeevan Singh Thakur. 2018. "Review of Clustering Methods: Toward Phylogenetic Tree Constructions." *Proceedings of International Conference on Recent Advancement on Computer and Communication*, 475–80.
- Sheridan†, Robert P. 2002. "The Most Common Chemical Replacements in Drug-Like Compounds," January. <https://doi.org/10.1021/ci0100806>.
- Shetye, Gauri S., Scott G. Franzblau, and Sanghyun Cho. 2020a. "New Tuberculosis Drug Targets, Their Inhibitors, and Potential Therapeutic Impact." *Translational Research : The Journal of Laboratory and Clinical Medicine* 220 (June):68–97.
- . 2020b. "New Tuberculosis Drug Targets, Their Inhibitors, and Potential Therapeutic Impact." *Translational Research : The Journal of Laboratory and Clinical Medicine* 220 (June):68–97.
- Shinu, Pottathil, Abdulaziz K. Al Mouslem, Anroop B. Nair, Katharigatta N. Venugopala, Mahesh Attimarad, Varsha A. Singh, Sreeharsha Nagaraja, Ghallab Alotaibi, and Pran Kishore Deb. 2022. "Progress Report: Antimicrobial Drug Discovery in the Resistance Era." *Pharmaceuticals* 15 (4): 413.
- Shi, Wanliang, Shuo Zhang, Jie Feng, Peng Cui, Wenhong Zhang, and Ying Zhang. 2017. "Clofazimine Targets Essential Nucleoid Associated Protein, Mycobacterial Integration Host

- Factor (mLHF), in *Mycobacterium Tuberculosis*." *bioRxiv*. bioRxiv. <https://doi.org/10.1101/192161>.
- "Significance of Root-Mean-Square Deviation in Comparing Three-Dimensional Structures of Globular Proteins." 1994. *Journal of Molecular Biology* 235 (2): 625–34.
- Silva, Márcio Rosa da, Hongwu Ma, and An-Ping Zeng. 2008 "Centrality, Network Capacity, and Modularity as Parameters to Analyze the Core-Periphery Structure in Metabolic Networks." Accessed December 26, 2023. <https://doi.org/10.1109/JPROC.2008.925418>.
- Siricilla, Shajila, Katsuhiko Mitachi, Bajoie Wan, Scott G. Franzblau, and Michio Kurosu. 2015. "Discovery of a Capuramycin Analog That Kills Nonreplicating Mycobacterium Tuberculosis and Its Synergistic Effects with Translocase I Inhibitors." *The Journal of Antibiotics* 68 (4): 271–78.
- Sirim, Demet, Michael Widmann, Florian Wagner, and Jürgen Pleiss. 2010. "Prediction and Analysis of the Modular Structure of Cytochrome P450 Monooxygenases." *BMC Structural Biology* 10 (1): 1–12.
- Slivoski, Gregory, Sandeepkumar Kothiwale, Jens Meiler, and Edward W. Lowe Jr. 2014. "Computational Methods in Drug Discovery." *Pharmacological Reviews* 66 (1): 334–95.
- Snow Setzer, Mary, Javad Sharifi-Rad, and William N. Setzer. 2016. "The Search for Herbal Antibiotics: An In-Silico Investigation of Antibacterial Phytochemicals." *Antibiotics (Basel, Switzerland)* 5 (3). <https://doi.org/10.3390/antibiotics5030030>.
- Sosa, Ezequiel J., Germán Burguener, Esteban Lanzarotti, Lucas Defelipe, Leandro Radusky, Agustín M. Pardo, Marcelo Marti, Adrián G. Turjanski, and Darío Fernández Do Porto. 2018. "Target-Pathogen: A Structural Bioinformatic Approach to Prioritize Drug Targets in Pathogens." *Nucleic Acids Research* 46 (D1): D413–18.
- Strieth-Kalthoff, Felix, Frederik Sandfort, Marwin H. S. Segler, and Frank Glorius. 2020. "Machine Learning the Ropes: Principles, Applications and Directions in Synthetic Chemistry." *Chemical Society Reviews* 49 (17): 6154–68.
- Sukheja, Paridhi, Pradeep Kumar, Nisha Mittal, Shao-Gang Li, Eric Singleton, Riccardo Russo, Alexander L. Perryman, et al. 2017. "A Novel Small-Molecule Inhibitor of the Demethylmenaquinone Methyltransferase MenG Is Bactericidal to Both Growing and Nutritionally Deprived Persister Cells." *mBio* 8 (1). <https://doi.org/10.1128/mBio.02022-16>.
- Sutaria, Dhruvitkumar S., Bartolome Moya, Kari B. Green, Tae Hwan Kim, Xun Tao, Yuanyuan Jiao, Arnold Louie, George L. Drusano, and Jürgen B. Bulitta. 2018. "First Penicillin-Binding Protein Occupancy Patterns of β -Lactams and β -Lactamase Inhibitors in *Klebsiella Pneumoniae*." *Antimicrobial Agents and Chemotherapy* 62 (6). <https://doi.org/10.1128/AAC.00282-18>.
- Suzek, Baris E., Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H. Wu. 2007. "UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters." *Bioinformatics* 23 (10): 1282–88.
- Taniguchi, H., H. Aramaki, Y. Nikaido, Y. Mizuguchi, M. Nakamura, T. Koga, and S. Yoshida. 1996. "Rifampicin Resistance and Mutation of the rpoB Gene in Mycobacterium Tuberculosis." *FEMS Microbiology Letters* 144 (1): 103–8.
- The UniProt Consortium, Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Shadab Ahmad, Emanuele Alpi, et al. 2022. "UniProt: The Universal Protein Knowledgebase in 2023." *Nucleic Acids Research* 51 (D1): D523–31.
- Trofimov, Valentin, Sébastien Kicka, Sabrina Mucaria, Nabil Hanna, Fernando Ramon-Olayo, Laura Vela-Gonzalez Del Peral, Joël Lelièvre, et al. 2018. "Antimycobacterial Drug Discovery Using Mycobacteria-Infected Amoebae Identifies Anti-Infectives and New Molecular Targets." *Scientific Reports* 8 (1): 3939.
- Unissa, Ameeruddin Nusrath, Selvakumar Subbian, Luke Elizabeth Hanna, and Nagamiah Selvakumar. 2016. "Overview on Mechanisms of Isoniazid Action and Resistance in Mycobacterium Tuberculosis." *Infection, Genetics and Evolution : Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases* 45 (November):474–92.
- Varadi, Mihaly, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, et al. 2021. "AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models." *Nucleic Acids Research* 50 (D1): D439–44.
- Velec, Hans F. G., Holger Gohlke, and Gerhard Klebe. 2005. "DrugScore(CSD)-Knowledge-Based Scoring Function Derived from Small Molecule Crystal Data with Superior Recognition Rate of near-Native Ligand Poses and Better Affinity Prediction." *Journal of Medicinal Chemistry* 48 (20): 6296–6303.
- "VMD: Visual Molecular Dynamics." 1996. *Journal of Molecular Graphics* 14 (1): 33–38.
- Wang, Yanli, Tugba Suzek, Jian Zhang, Jiyao Wang, Siqian He, Tiejun Cheng, Benjamin A. Shoemaker, Asta Gindulyte, and Stephen H. Bryant. 2013. "PubChem BioAssay: 2014 Update."

- Nucleic Acids Research* 42 (D1): D1075–82.
- Wang, Zhe, Huiyong Sun, Xiaojun Yao, Dan Li, Lei Xu, Youyong Li, Sheng Tian, and Tingjun Hou. 2016. "Comprehensive Evaluation of Ten Docking Programs on a Diverse Set of Protein-Ligand Complexes: The Prediction Accuracy of Sampling Power and Scoring Power." *Physical Chemistry Chemical Physics : PCCP* 18 (18): 12964–75.
- Wei, Dongqing, Qin Xu, Tangzhen Zhao, and Hao Dai. 2014. *Advance in Structural Bioinformatics*. Springer.
- Weininger, David. 2002. "SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules," May. <https://doi.org/10.1021/ci00057a005>.
- Werck-Reichhart, D., and R. Feyereisen. 2000. "Cytochromes P450: A Success Story." *Genome Biology* 1 (6): REVIEWS3003.
- Wigh, Daniel S., Jonathan M. Goodman, and Alexei A. Lapkin. 2022. "A Review of Molecular Representation in the Age of Machine Learning." *Wiley Interdisciplinary Reviews. Computational Molecular Science* 12 (5): e1603.
- Wild and, D. J., and C. J. Blankley*. 1999. "Comparison of 2D Fingerprint Types and Hierarchy Level Selection Methods for Structural Grouping Using Ward's Clustering," November. <https://doi.org/10.1021/ci990086j>.
- Willett, Peter. 2011. "Similarity Searching Using 2D Structural Fingerprints." *Methods in Molecular Biology (Clifton, N.J.)* 672:133–58.
- . 2014. "The Calculation of Molecular Structural Similarity: Principles and Practice." *Molecular Informatics* 33 (6-7): 403–13.
- Wilson, Carter J., Wing-Yiu Choy, and Mikko Karttunen. 2022. "AlphaFold2: A Role for Disordered Protein/Region Prediction?" *International Journal of Molecular Sciences* 23 (9): 4591.
- World Health Organization. 2022. *The WHO AWaRe (Access, Watch, Reserve) Antibiotic Book*. World Health Organization.
- Xia, Xuhua. 2018. *Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Proteomics, and Transcriptomics*.
- Yoon, Byung-Jun. 2009. "Hidden Markov Models and Their Applications in Biological Sequence Analysis." *Current Genomics* 10 (6): 402–15.
- Zdrazil, Barbara, Eloy Felix, Fiona Hunter, Emma J. Manners, James Blackshaw, Sybilla Corbett, Marleen de Veij, et al. 2023. "The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods." *Nucleic Acids Research*, November. <https://doi.org/10.1093/nar/gkad1004>.
- Zhou, Hongyi, and Jeffrey Skolnick. 2024. "Utility of the Morgan Fingerprint in Structure-Based Virtual Ligand Screening." *The Journal of Physical Chemistry. B* 128 (22): 5363–70.
- Zhou, Joe Zhongxiang. 2011. "Chemoinformatics and Library Design." *Chemical Library Design*, 27–52.