

Tesis Doctoral

Decodificando la diversidad catalítica de los citocromos p450 bacterianos mediante métodos bioinformáticos

Dumas, Victoria Gisel

2016-08-12

Este documento forma parte de la colección de tesis doctorales y de maestría de la Biblioteca Central Dr. Luis Federico Leloir, disponible en digital.bl.fcen.uba.ar. Su utilización debe ser acompañada por la cita bibliográfica con reconocimiento de la fuente.

This document is part of the doctoral theses collection of the Central Library Dr. Luis Federico Leloir, available in digital.bl.fcen.uba.ar. It should be used accompanied by the corresponding citation acknowledging the source.

Cita tipo APA:

Dumas, Victoria Gisel. (2016-08-12). Decodificando la diversidad catalítica de los citocromos p450 bacterianos mediante métodos bioinformáticos. Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires.

Cita tipo Chicago:

Dumas, Victoria Gisel. "Decodificando la diversidad catalítica de los citocromos p450 bacterianos mediante métodos bioinformáticos". Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. 2016-08-12.

EXACTAS UBA

Facultad de Ciencias Exactas y Naturales



UBA

Universidad de Buenos Aires



UNIVERSIDAD DE BUENOS AIRES
Facultad de Ciencias Exactas y Naturales
Departamento de Química Biológica

Decodificando la diversidad catalítica de los citocromos p450 bacterianos mediante métodos bioinformáticos

Victoria Gisel Dumas

Trabajo de tesis presentado para optar por el título de doctora de la Universidad de Buenos Aires en el área Química Biológica

Director y consejero de estudios:

Dr. Marcelo Adrián Martí

Buenos Aires, 2016

Decodificando la diversidad catalítica de los citocromos p450 bacterianos mediante métodos bioinformáticos

Los citocromos p450 bacterianos (CYPs) son una familia de enzimas que poseen una enorme diversidad funcional y una amplia capacidad para llevar a cabo diferentes reacciones de oxidación sobre sus sustratos, que van desde pequeñas moléculas aromáticas como el alcanfor, pasando por algunas más grandes como el colesterol o la vitamina D, hasta moléculas enormes como los antibióticos macrólidos. Lo cierto es que estas enzimas contribuyen de manera significativa a la generación de una gran variedad de productos naturales, muchos de ellos con un enorme potencial biotecnológico. Es por ello que determinar el rol de los CYPs en la producción de estos metabolitos es esencial para comprender su biosíntesis y posible rol fisiológico.

Considerando que los CYPs desarrollan funciones sumamente importantes para la vida de los microorganismos, comprender su función constituye un objetivo de alta relevancia biológica. De la vasta cantidad de reacciones que son capaces de catalizar estas enzimas, algunas han sido ya extensamente estudiadas, pero muchas otras permanecen aún completamente desconocidas. Por esta razón, en la presente tesis, se procedió en primer término al estudio del mecanismo de reacción de un CYP esencial para la viabilidad de *Mycobacterium tuberculosis*, -denominado CYP121- cuya reacción es poco común en esta familia de enzimas. Mediante la aplicación de diversas técnicas de simulación computacional se consiguió arribar a una propuesta mecanística detallada de su funcionamiento con una mirada fisicoquímica. A continuación, se procedió a realizar un estudio general de los determinantes estructurales que dan lugar a la diversidad funcional en los CYPs, con el objeto de poder predecir *in-silico* el rango de sustratos y reacciones catalizadas por estas enzimas. Conseguir hacer tales predicciones es un objetivo clave, dado que el reciente crecimiento exponencial en la secuenciación de diferentes microorganismos ha resultado en la anotación de miles de secuencias a este grupo de proteínas, las cuales en su mayoría carecen de información funcional. Este estudio derivó en el desarrollo de un *pipeline* bioinformático que tomando como punto de partida la secuencia de un dado CYP y analizando una serie de propiedades, permite predecir la reacción que éste cataliza.

En su conjunto, la presente tesis presenta un estudio detallado del mecanismo de reacción del CYP121 de *Mycobacterium tuberculosis*, el cual dada la esencialidad de esta enzima, podría resultar importante para el desarrollo de drogas para el tratamiento de la tuberculosis; junto con una caracterización de la diversidad funcional de la familia de los citocromos p450s bacterianos y una herramienta que permite predecir la función de un determinado CYP partiendo solamente de su estructura primaria.

Palabras clave: citocromos p450 bacterianos, mecanismo de reacción, predicción de función, bioinformática, QM-MM, docking, modelado por homología, contexto génico

ABSTRACT

Decoding bacterial p450 cytochromes catalytic diversity by bioinformatic methods

Bacterial p450 cytochromes (CYPs) are a family of enzymes that have an enormous functional diversity and great capacity to carry out different oxidation reactions on their substrates, which vary from small aromatic molecules such as camphor, to some larger as cholesterol or vitamin D up to huge molecules such as macrolide antibiotics. The fact is that these enzymes contribute significantly to the generation of a variety of natural products, many of them with huge potential biotechnological applications. Therefore, determining the CYPs' role in the production of these metabolites is essential to understand their biosynthesis and potential physiological role.

Whereas the CYPs develop extremely important functions for microorganisms life, understand their function is of uttermost biological relevance. Among the vast amount of reactions which are capable of catalyzing these enzymes, some have already been extensively studied, but many others remain still unknown. For this reason, in this thesis, we proceeded first to the study of the reaction mechanism of a CYP essential for the viability of *Mycobacterium tuberculosis*, -called CYP121- whose reaction is unusual in this family of enzymes. By applying different computer simulation techniques we were able to arrive at a detailed mechanistic proposal for the role of this enzyme. Following we proceeded to conduct a comprehensive study of the structural determinants that lead to functional diversity in CYPs, in order to be able to predict in-silico the range of substrates and reactions catalyzed by these enzymes. Becoming able to make such predictions is a key objective, since the recent exponential growth in the sequencing of different microorganisms has resulted in the entry of thousands of sequences to this group of proteins, mostly lacking of functional information. This study led to the development of a bioinformatic pipeline that taking as a starting point a given CYP's sequence and by analyzing a number of properties, is able to predict what reaction it catalyzes.

As a whole, this thesis presents a detailed study of the *Mycobacterium tuberculosis* CYP121 reaction mechanism, which given the essential nature of this enzyme, it may be important for the development of drugs for tuberculosis treatment; along with a characterization of the functional diversity of bacterial cytochrome P450s' family and a tool for predicting a particular CYP's function based only on its primary structure.

Keywords: bacterial cytochromes p450, reaction mechanism, function prediction, bioinformatics, QM-MM, docking, homology modelling, gene context

A Juan y a León, mi pequeña gran familia.

A mamá, papá, Flo, Joaco y Pau.

Índice general

Resumen	2
Abstract	3
1 Introducción general	8
1.1 La relación estructura-función en las proteínas	8
1.2 Bioinformática y simulación molecular	8
1.3 La importancia de predecir función a partir de la estructura	9
1.4 Los citocromos p450	11
1.4.1 Nomenclatura	12
1.4.2 Estructura	12
1.4.3 Ciclo catalítico típico	13
1.4.4 Aplicaciones	14
1.5 Objetivos de esta tesis	17
2 Métodos	19
2.1 Simulación computacional	19
2.1.1 Métodos basados en la mecánica cuántica	19
2.1.2 Métodos basados en la mecánica clásica	21
2.1.3 Métodos híbridos cuántico-clásicos (QM/MM)	24
2.1.4 Exploración de la superficie de energía potencial (SEP)	25
2.1.5 Estimación de energía libre: Integración termodinámica	28
2.2 Métodos bioinformáticos	29
2.2.1 Alineamiento de secuencias múltiples (ASM)	29
2.2.2 Alineamiento de estructuras proteicas	30
2.2.3 Comparación de estructuras de ligandos	30
2.3 Modelado por homología	32
2.3.1 Selección de la estructura molde más adecuada	33
2.3.2 Alineamiento de las secuencias de la proteína a modelar y el molde	33
2.3.3 Generación de un modelo estructural basado en el alineamiento de secuencias	33
2.3.4 Validación del modelo obtenido	35
2.4 Análisis del contexto génico	35
2.4.1 Obtención de información del contexto génico	35
2.5 Docking	37
2.5.1 Algoritmo de búsqueda	37
2.5.2 Función de puntuación	38
2.5.3 El algoritmo de docking utilizado en esta tesis	39
3 El mecanismo de reacción inusual de MtCYP121	40
3.1 Introducción	40

3.2	Resultados	42
3.2.1	Generación del primer radical	42
3.2.2	Generación del segundo radical	45
3.2.3	cY*Y* a mcic: la formación del enlace C-C ocurre en solución	49
3.3	Discusión	54
3.4	Conclusión	57
3.5	Metodología específica del capítulo	58
3.5.1	Integración termodinámica	58
3.5.2	Métodos cuánticos e híbridos cuántico-clásicos (QM/MM)	59
4	Análisis de la diversidad estructural y funcional de los p450s	60
4.1	Introducción	60
4.2	Resultados	60
4.2.1	Análisis filogenético de las secuencias	61
4.2.2	Generación de estructuras por modelado por homología	66
4.2.3	Análisis de la diversidad de sustratos de los CYPs	70
4.2.4	Contexto génico	71
4.2.5	Reacción	75
4.3	Conclusiones	76
4.4	Métodos específicos del capítulo	77
4.4.1	Obtención de los conjuntos de secuencias de partida	77
4.4.2	Filtrado para seleccionar solo las secuencias representativas	77
4.4.3	Alineamiento de secuencias teniendo en cuenta la información estructural	77
4.4.4	Construcción de los árboles filogenéticos	78
4.4.5	Construcción de las secuencias logo del alineamiento	78
4.4.6	Obtención de la distancia evolutiva entre todos los pares de secuencias del árbol filogenético	78
4.4.7	Generación de estructuras por modelado por homología y evaluación de los modelos	78
5	Predicción de la estructura del complejo CYP-sustrato	81
5.1	Introducción	81
5.2	Resultados	81
5.2.1	Evaluación de la capacidad para reproducir la interacción receptor-ligando observada en los cristales	81
5.2.2	Docking sobre modelos por homología	84
5.2.3	Selectividad	87
5.3	Conclusiones	89
5.4	Métodos específicos del capítulo	92
5.4.1	Obtención de las estructuras de complejos CYP-sustrato	92
5.4.2	Preparación de la estructura de los receptores	92
5.4.3	Preparación de las estructuras de los ligandos	93
5.4.4	Cálculo del RMSD de la estructura del ligando	93
5.4.5	Protocolo de docking	93
6	Integración del conocimiento generado en un esquema predictivo	95
6.1	Introducción	95
6.2	Resultados	95
6.2.1	Caso 1: CYP170A1	97
6.2.2	Caso 2: CYP125	98
6.2.3	Caso 3: CYP158A2	100
6.2.4	Aplicación a gran escala	100

6.3 Conclusiones	101
7 Discusión y perspectivas futuras	103
7.1 La predicción de función de los CYPs bacterianos	104
7.2 Perspectivas futuras	106
7.3 Aplicaciones	107
Publicaciones	108
Agradecimientos	109

INTRODUCCIÓN GENERAL

1.1 La relación estructura-función en las proteínas

La mayor parte de los roles estructurales y funcionales esenciales para la célula son mediados por proteínas. Estas moléculas grandes y complejas exhiben una versatilidad remarcable que les permite desempeñar una inmensidad de actividades que son fundamentales para la vida. De hecho ninguna otra macromolécula biológica podría asumir todas las funciones que las proteínas han acumulado durante miles de millones de años de evolución. Esto hace que comprender las proteínas no solo sea importantísimo para la biología, sino también muy interesante.

Cualquier consideración sobre la función proteica debe estar fundamentada en el entendimiento de la estructura proteica. Un principio fundamental en la ciencia de las proteínas es que *la estructura proteica determina la función*. Las diferentes estructuras proteicas permiten la ubicación de grupos funcionales particulares (tomando la definición química de los mismos) en sitios específicos en el espacio tridimensional. Es esta precisión la que permite a las proteínas actuar como catalizadores (en el caso de las enzimas) en una variedad sorprendente de reacciones químicas. El posicionamiento de grupos funcionales también permite a las proteínas desempeñar importantes funciones estructurales, de transporte y de regulación en los organismos. Dado que la estructura proteica conduce a la función, y las funciones proteicas son diversas, no es sorprendente que las estructuras proteicas sean semejantemente diversas. Además, la diversidad funcional de las proteínas es expandida a través de su interacción con moléculas pequeñas -denominadas ligandos-, así como también con otras proteínas. Para quienes estudiamos la estructura proteica, esta diversidad representa un desafío, y en este sentido, la bioinformática y la simulación molecular son dos disciplinas altamente complementarias al momento de abordar en profundidad dicho desafío.

1.2 Bioinformática y simulación molecular

La bioinformática es la disciplina que utiliza y desarrolla métodos y herramientas computacionales para analizar, manipular y almacenar datos biológicos con el objetivo de poder transformarlos en conocimiento. El auge de la bioinformática viene de la mano del surgimiento de técnicas experimentales que generan grandes volúmenes de datos, tales como la secuenciación de ADN, la espectrometría de masas y los análisis de expresión génica, entre otros [1, 2, 3]. Estos datos son demasiado complejos para ser analizados manualmente, y por tanto hacen necesario recurrir a las herramientas computacionales. El rápido incremento en el número de estructuras tridimensionales disponibles en las bases de datos tales como el *Protein Data Bank (PDB)*, ha dado origen a una subdisciplina de la bioinformática, denominada bioinformática estructural. La misma se enfoca en la representación, almacenamiento, análisis y visualización de la información estructural a escala atómica. La bioinformática estructural, como muchas otras subdisciplinas dentro de la bioinformática, se caracteriza por dos objetivos: por un lado la creación de métodos de propósito general para manipular la información relativa a las macromoléculas.

las biológicas y por el otro la aplicación de esos métodos para resolver problemas de la biología y crear nuevo conocimiento. Estos dos objetivos están intrínsecamente relacionados, ya que gran parte de la validación de nuevos métodos involucra su uso exitoso en la resolución de problemas reales [4].

Ahora bien, las estructuras proteicas disponibles en las bases de datos como el PDB, que surgen de los resultados de estudios cristalográficos y de RMN, constituyen modelos estructurales estáticos de la estructura proteica. Sin embargo, la función proteica está intrínsecamente ligada a sus movimientos, o en otras palabras, a su dinámica. Por ello es necesario recurrir a técnicas de simulación en las que partiendo de una estructura promedio de alta calidad se analiza como la proteína va cambiando a lo largo del tiempo. De estas técnicas, la dinámica molecular es una de las metodologías más usadas en el campo de la bioinformática estructural. La dinámica molecular engloba métodos teóricos y técnicas computacionales para comprender, modelar y predecir el comportamiento de las moléculas basándose en las leyes físicas que gobiernan las interacciones entre los átomos. En este trabajo de tesis la combinación de herramientas de la bioinformática y de la simulación molecular ha permitido abordar el estudio de la relación estructura-función proteica comenzando con casos particulares y yendo hacia una visión general, extrayendo interpretaciones y conclusiones para toda una familia de proteínas.

1.3 La importancia de predecir función a partir de la estructura

Con el advenimiento de la genómica, la predicción de la función biológica a partir de la estructura se ha transformado en uno de los principales desafíos de la biología estructural y la bioinformática [5]. La asignación de función biológica constituye un valioso primer paso hacia la caracterización experimental de los roles celulares y fisiológicos de los productos génicos y podría contribuir a mejorar el análisis y anotación de los genomas.

Los proyectos de secuenciación de genomas a gran escala nos han provisto de detalles acerca de los genes que un organismo necesita para sobrevivir. A partir de esta información podemos traducir las secuencias de aminoácidos de todas las proteínas que el genoma puede producir. Los proyectos de genómica estructural apuntan a resolver las estructuras de todas esas proteínas aún cuando sus funciones sean desconocidas. Este proceso representa un cambio de paradigma respecto al modo usual de trabajar en la investigación experimental de proteínas, que consiste en tomar una proteína de interés, llevar a cabo experimentos bioquímicos para determinar la información funcional y luego utilizar la estructura para racionalizar dicha información funcional [6]. Al paso que van los proyectos de genómica estructural, resulta imposible estudiar todas las proteínas experimentalmente. Nos encontramos entonces con la necesidad vital de desarrollar métodos que permitan inferir su función partiendo de la secuencia, a fines de poder anotar los genomas con información funcional.

El problema de conocer la función biológica comienza por la definición misma del término, cuyo significado exacto varía dependiendo del contexto en el cual se utiliza [7]. Resulta obvio que la función biológica de una proteína engloba más de un aspecto. Tomemos por ejemplo una proteína quinasa: en el aspecto bioquímico la función de una quinasa sería la fosforilación de un sustrato específico. Sin embargo, las proteínas llevan a cabo su función dentro de un organismo y esto conlleva consecuencias que van desde la escala subcelular hasta el organismo entero. Entonces en el aspecto fisiológico, la misma quinasa puede formar parte de una vía de señalización donde esa proteína fosforila pero a la vez también es fosforilada por otra enzima con la cual interacciona. Al mismo tiempo, una mutación en esta quinasa puede causar una enfermedad, entonces surge otro aspecto que es el relacionado con el funcionamiento general del organismo, y en casos humanos con la salud. En consecuencia, cuando se habla de la función de una proteína, es necesario especificar el o los aspectos de la descripción funcional a los que uno se refiere [8]. En el contexto de este trabajo de tesis, consideraremos que la función de una enzima queda definida por el sustrato que une y la reacción que cataliza sobre el mismo, ciñéndonos de este modo a un enfoque molecular.

Una vez definidos los aspectos funcionales de interés, el siguiente asunto a resolver es cómo describir la función biológica de manera computacionalmente amigable. La información de secuencia y estructura proteica se puede describir fácilmente de modo legible para una computadora, no así la información funcional. Esto se debe a que la anotación acerca de la función de una proteína se escribe en lenguaje humano, lo cual conlleva no solo

las sutilezas y complejidades relativas a su función así como también la evidencia experimental que la soporta, su historia de investigación y otras características. Para hacer que la anotación funcional sea computacionalmente procesable, es necesario transmitirla de forma controlada y bien definida. Esta necesidad de un vocabulario controlado y relaciones bien definidas para describir la función fue reconocida en primer momento por la Unión Internacional de Bioquímica y Biología Molecular, que ha desarrollado una nomenclatura para identificar a las enzimas basada en los denominados Números EC. Los números EC (*Enzyme Commission numbers*) son un esquema de clasificación numérica para las enzimas basado en las reacciones químicas que catalizan. Como sistema de nomenclatura de enzimas, cada número EC está asociado a un nombre recomendado para dicha enzima. En realidad los números EC codifican reacciones catalizadas por enzimas. Enzimas diferentes (por ejemplo que procedan de organismos diferentes) que catalicen la misma reacción recibirán el mismo número EC. Cada código de enzimas consiste en las dos letras EC seguidas por 4 números separados por puntos. Estos números representan una clasificación progresivamente más específica. Por ejemplo, la enzima tripéptido aminopeptidasa tiene el código EC 3.4.11.4, construido así: 3 por hidrolasa (enzima que usa el agua para catalizar algunas reacciones); 3.4 por hidrolasa, que actúa sobre los enlaces peptídicos; 3.4.11 por aquellas que actúan sobre el terminal amino de aminoácido de un polipéptido; 3.4.11.4 por aquellas que actúan sobre el terminal amino de un tripéptido. Existen también otros esquemas de anotación funcional más allá del descripto. La característica compartida por todos ellos es el uso de un vocabulario controlado y en muchos casos una categorización que va de lo general a lo específico.

Ya definida la función y los medios para describirla, ahora analicemos las herramientas utilizadas para predecirla. Sin dudas, la más usada es la búsqueda por similitud en bases de datos de secuencias. Tradicionalmente se ha utilizado la identificación de secuencias de aminoácidos similares para inferir tanto la estructura como la función de una proteína, es decir, lo que se conoce como transferencia basada en homología (o falacia por asociación). La razón biológica detrás de la transferencia basada en homología está en que se creía que la estructura y la función pueden ser transferidas entre secuencias similares debido a que ambas han sido conservadas durante largos períodos de tiempo. Esta creencia se ha confirmado para la estructura [9], pero es más difícil de justificar para la función, particularmente cuando se desea definirla (o conocerla) con precisión. Además, a medida que las bases de datos crecen en número de secuencias, la transferencia basada en homología comienza a fallar, en primer lugar porque se sabe que hay casos en los que aún con alto grado de similitud la transferencia de función puede ser errónea y además porque cuantas más secuencias que ingresan a la base de datos son anotadas por transferencia por homología, más se incrementa la capacidad de que se amplifiquen las anotaciones erróneas en base a una única anotación equivocada inicial [10, 11]. Otra problemática asociada con este enfoque, es que si bien existen diversas bases de datos que contienen información acerca de la función proteica (Uniprot, KEGG, ENZYME, MetaCyc, BRENDA), los diferentes esquemas muestran diversos aspectos de la función. Además, es muy común encontrar que el número EC está incompleto (ej: 1.3.-.-), esto quiere decir que solo se cuenta con información muy general sobre la función enzimática. En consecuencia, la información funcional solo es capturada parcialmente por estas bases de datos, lo cual dificulta la transferencia de información funcional de una proteína homóloga a otra.

Pero los métodos de transferencia por homología no son la única vía en la que puede utilizarse la información contenida en la secuencia para inferir función. Existen métodos que se basan en la información del contexto genómico de la secuencia para predecir función. Como se desarrollará más adelante en el capítulo 4, estos métodos son particularmente útiles en procariotas, dado que en tales genomas los genes funcionalmente relacionados tienden a estar colocalizados en el genoma, formando un mismo bloque de regulación transcripcional.

Ahora bien, mientras que los aspectos generales que hacen a la función proteica se pueden inferir de la secuencia, los detalles particulares requieren conocer y saber entender la relación de la estructura con la especificidad por cierto tipo de sustrato. Estudiar la estructura proteica es entonces clave para comprender la función, ya que en las enzimas frecuentemente ésta es conferida por unos pocos residuos conservados en el sitio activo, los cuales muchas veces no son detectados por los métodos basados en análisis de las secuencias. Estos residuos, sin embargo, pueden estar relacionados en tres dimensiones, entonces la comparación de las similitudes estructurales entre las proteínas tiene el potencial para identificar similitudes funcionales en proteínas no

homólogos.

1.4 Los citocromos p450

Los citocromos p450 (p450s o CYPs) están entre los catalizadores más extraordinarios de la naturaleza, frecuentemente llevan a cabo reacciones químicas que desafían aún las capacidades de los químicos más hábiles. El descubrimiento del citocromo P450 se remonta a la década de los años 50, en la que comenzaron a estudiarse unos pigmentos encontrados en células hepáticas, a los que se les denominó citocromos, del griego *ritos* (hueco, -en biología célula-) y *croma* (color). A principios de esta década Martin Klingenberg, investigador alemán, observó durante sus estudios con el citocromo b5, la presencia de un pigmento unido al monóxido de carbono con una banda de absorbancia máxima a 450nm de longitud de onda. ¿Podría ser una hemoproteína?. Sin embargo, esta característica no era propia de las hemoproteínas tipo Protoporfirina IX que se conocían hasta el momento [12]. No sería hasta 1964 cuando Omura y Sato [13], identifican la naturaleza hemoproteica de este pigmento, que se encontraba presente en los microsomas hepáticos de diferentes especies de mamíferos y que tras ser reducido por NADPH, era capaz de unirse al CO, mostrando un característico pico de absorbancia en el espectro UV a 450 nm (pico de Soret). Por este motivo a esta hemoproteína se la denominó citocromo P-450 (P por pigmento y 450 por su pico de absorbancia en el UV).

Citocromos por todas partes

Los citocromos p450 son ubicuos en todos los reinos biológicos, desde bacterias hasta eucariotas, así como también dentro de un mismo organismo, estando presentes en una gran variedad de tejidos. Cada organismo produce varios CYPs diferentes, cada uno de los cuales actúa sobre un grupo distinto de moléculas. Típicamente las bacterias producen alrededor de 20 formas diferentes, mientras que los humanos producimos unas 60. Pero las plantas producen aún más, en general tienen cientos de CYPs diferentes, y esto se debe a que las plantas fabrican pigmentos y toxinas exóticas para protegerse a sí mismas, y muchas de las reacciones necesarias para producir tales moléculas son llevadas a cabo por citocromos p450s especializados. En los humanos, en cambio, el rol principal de los p450s es el de detoxificar el organismo, solubilizando drogas y sustancias peligrosas para que puedan ser excretadas más fácilmente.

En organismos eucariotas se han detectado prácticamente en todas las membranas subcelulares [14], siendo la mitocondria, y principalmente el retículo endoplásmico las fuentes más importantes [15]. A diferencia de los eucariotas, en las bacterias, los citocromos no tienen una asociación a membrana y se solubilizan fácilmente [16]. Siendo estas proteínas tan ubicuas, la cantidad de secuencias de CYPs conocidas es enorme! Una simple inspección en la base de datos RefSeq arroja que se conocen actualmente nada menos que 28000 CYPs diferentes.

Agregando oxígeno

El sitio activo de los CYPs es de tipo hemotiolato (el hierro del grupo hemo está enlazado a la proteína por medio del tiolato de un residuo de cisteína) con propiedades redox. Se las conoce también como monooxigenasas debido a su usual capacidad para activar dioxígeno molecular produciendo especies de oxígeno monoatómicas altamente reactivas, y luego insertar el oxígeno en una amplia variedad de sustratos, aunque también son capaces de catalizar otras reacciones oxidativas. Un rasgo característico de este tipo de proteínas es precisamente su gran versatilidad funcional. Como ya se mencionó, se los ha hallado involucrados en el metabolismo de diversos compuestos xenobióticos como drogas, pesticidas, procarcinógenos, anestésicos, solventes orgánicos, así como también ciertas sustancias endógenas como colesterol, ácidos biliares, esteroides, vitaminas liposolubles y ácidos grasos, lo que hace que sean moléculas altamente interesantes para un amplio espectro de aplicaciones que se comentarán más adelante (ver sección 7.3).

1.4.1 Nomenclatura

Cuando solo se conocían unos pocos citocromos p450s, no había ningún problema para nombrarlos, ya que el nombre les era dado por el sustrato que unían o por la función que desempeñaban. A modo de ejemplo, la p450cam (hoy conocida también como CYP101) toma su nombre de su sustrato: el alcanfor (*camphor*, en inglés), y CYP51A1 también es llamada LDM (Lanosterol DeMetilasa) dado que une el sustrato Lanosterol y la función que realiza es la DesMetilación.

Sin embargo, en la medida en que se fueron resolviendo y caracterizando más y más estructuras de p450s, dado que el número de genes y proteínas p450s es demasiado grande, se vio la necesidad de implementar un sistema de nomenclatura sistemática. En 1987 se propuso entonces el sistema de nomenclatura y clasificación que se utiliza hoy en día, el cual obedece a criterios filogenéticos y se basa en la identidad de la secuencia de aminoácidos de las diferentes enzimas [17]. Según este criterio, los p450 se identifican con las siglas CYP seguido de un número que designa la familia, una letra que identifica la subfamilia y otro número que se corresponde con el gen. Con este sistema de nomenclatura quedan totalmente identificados todos los p450s, tanto procariotas como eucariotas. En una misma familia se agrupan aquellas enzimas cuya secuencia de aminoácidos tiene una similitud mayor al 40 %, independientemente de la especie de procedencia. A su vez las p450s de cada familia se agrupan en diferentes subfamilias que, siempre que haya más de una, se denominan correlativamente empezando siempre por la letra A (por ejemplo: CYP2A, CYP2B, CYP2C, etc). En este caso, el requisito para que dos p450 pertenezcan a la misma subfamilia es que tengan una homología en la secuencia de aminoácidos superior al 55 %. Por último, dentro de la misma subfamilia, las enzimas individuales se designan según números empezando siempre por el 1 (p. e. CYP1A1, CYP1A2), teniendo en cuenta que dos p450 se consideran como diferentes siempre y cuando sus respectivas secuencias difieran en más de un 3 % (ver Figura 1.1).

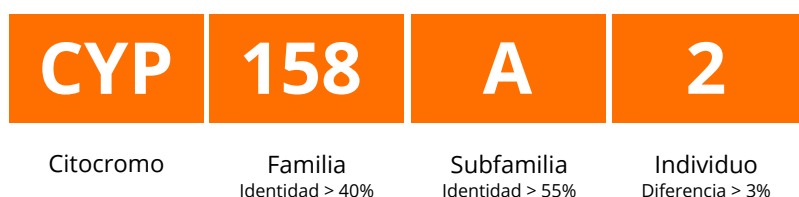


Figura 1.1: Significado de la nomenclatura de los citocromos p450s

1.4.2 Estructura

Las p450s comparten un plegado común que es único de esta clase de enzimas (Figura 1.2). Su estructura secundaria está compuesta mayormente por hélices α , y poseen un cofactor hemo encerrado entre un dominio más grande rico en hélices α (designadas de A a L) conocido como dominio alfa y uno más pequeño rico en hojas β (designadas de 1 a 4) que se denomina dominio beta. La estructura central que rodea al cofactor hemo provee el andamio que permite la activación del oxígeno por parte de las p450s. El hierro hémico se encuentra coordinado axialmente por una cisteína conservada como ligando proximal y típicamente por una molécula de agua como ligando distal (en lo que se conoce como estado de reposo de la p450).

Para llevar a cabo la catálisis, las CYPs deben unir a sus sustratos dentro del bolsillo del sitio activo, el cual se localiza en la parte superior del sitio distal del hemo, donde también se une el oxígeno. La comparación de las estructuras de CYPs disponibles muestra un plegado altamente conservado con una gran diversidad en la forma y el tamaño de los sitios de unión de sustratos, siendo estas características claves para definir la especificidad por el sustrato y de la reacción. (ver Figura 1.3). El sitio catalítico alrededor del hemo contiene unas pequeñas regiones llamadas sitios de reconocimiento de sustratos (SRS) que están involucradas en la unión de sustratos y en la catálisis. Estos SRSs muestran alta variación en su secuencia de aminoácidos, siendo la razón subyacente de la capacidad de las diferentes p450s para unir sustratos tan diversos. Los alineamientos estructurales de múltiples p450s diferentes revelan que existe flexibilidad conformacional e indican que son tres regiones las que causan

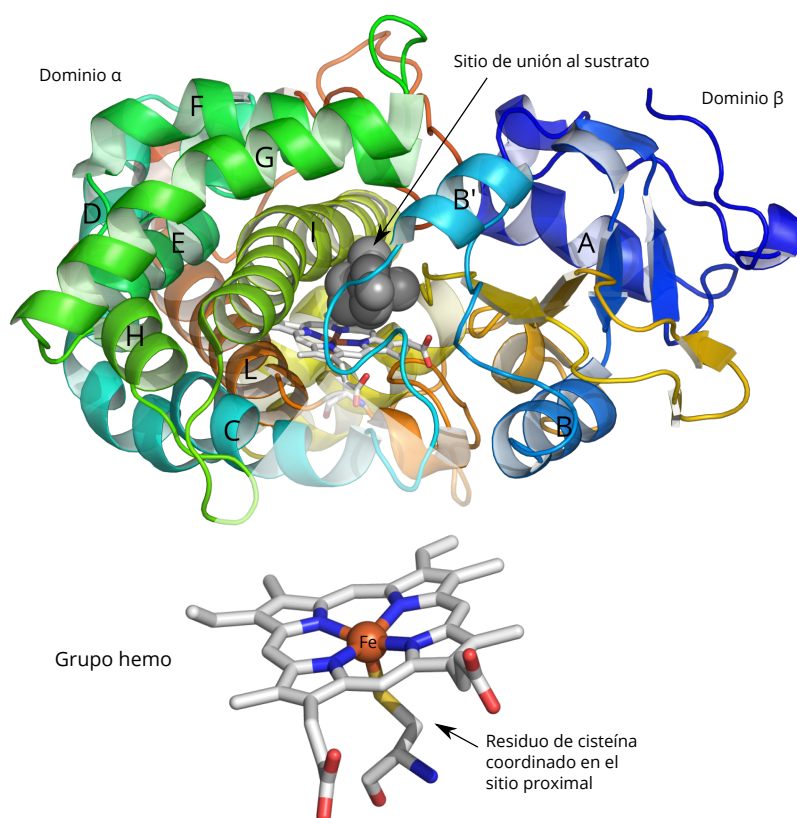


Figura 1.2: Estructura tridimensional típica de los citocromos p450 junto a su cofactor el grupo hemo coordinado a un residuo de cisteína. Se muestran señalados los dominios alfa y beta así como también el sitio de unión al sustrato. Las letras corresponden a la nomenclatura de las hélices alfa.

mayor variación en la estructura del esqueleto proteico y como consecuencia en el bolsillo de unión al sustrato: la región FG (el extremo C de la hélice F y el extremo N de la hélice G), la región B (los loops entre las hélices B/B', el loop BC y la hebra 5 de la hoja β 1) y dos regiones de loops que se muestran señaladas en la Figura 1.3[18]. A modo de ejemplo, la Figura 1.3 muestra en el panel A las estructuras de dos CYPs superpuestas dando cuenta de la alta conservación de la estructura tridimensional que presentan. Sin embargo, estos dos ejemplares son capaces de unir sustratos de tamaño y naturaleza química tan diversa como lo son la eritromicina y el alcanfor, cuyas estructuras se muestran en el panel B.

1.4.3 Ciclo catalítico típico

Comprender el mecanismo de reacción de un CYP requiere un conocimiento detallado de la naturaleza y estructura electrónica de sus especies oxidantes claves: el compuesto I (Cp I) y su ciclo de reacción asociado (el cual se muestra adaptado para *MtCYP121*, la proteína cuyo mecanismo de reacción fue objeto de estudio en la primera parte de esta tesis, en la figura 1.4). En la reacción de CYP arquetípica, el estado de reposo es generalmente un estado de hemo férrico hexacoordinado (6c), con la cisteína proximal y una molécula de agua (o anión hidróxido) unido al sitio distal (en donde se une el sustrato). Luego de la unión del sustrato específico, la reducción del hemo al estado ferroso y liberación de la molécula de agua (o hidróxido) distal, la enzima está preparada para unir oxígeno molecular (O_2) y así formar el complejo $FelI-O_2$, que luego se reduce para producir la especie ferroso-superóxido ($FelI-O_2^-$). Ambos electrones derivan en procariotas de sus *partners* redox (CYP reductasas), los cuales generalmente utilizan NADH/FADH como fuente final de electrones. Respecto a este punto, es interesante notar la considerable diversidad de *partners* redox observada entre los CYPs bacterianos [19]. Las especies ferroso-superóxido poseen una naturaleza intrínsecamente inestable, y luego de tomar dos protones del ambiente, se produce la ruptura de enlace O-O, con la consecuente liberación de una molécula de agua, y produciendo la for-

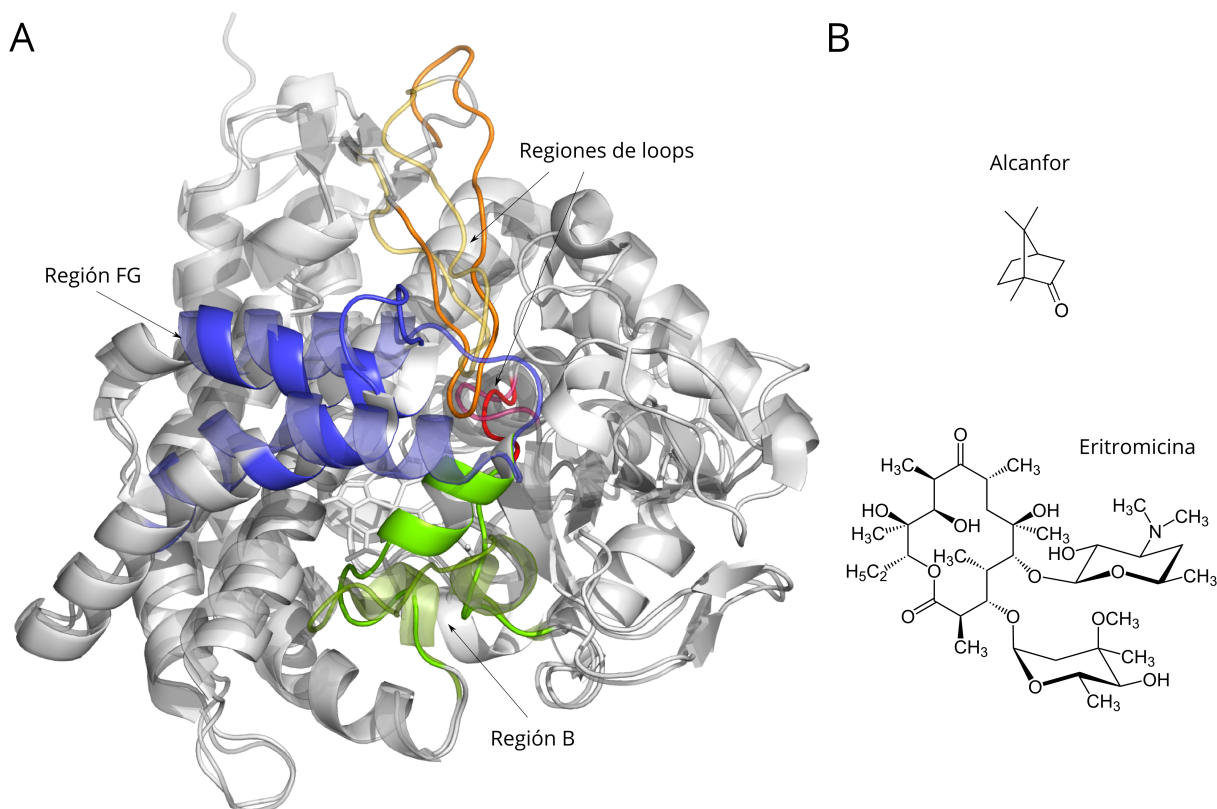


Figura 1.3: A. Superposición de las estructuras de dos p450. Una de ellas es la CYP113A1 también conocida como eryK (PDB-ID: 2JJO), y la otra es la famosa p450cam (PDB-ID: 1DZ4). B. Estructura de los correspondientes ligandos, eritromicina y alcanfor. Como puede observarse, el plegado está altamente conservado en ambas proteínas, pero aún siendo tan parecidas son capaces de unir sustratos muy diversos tanto en tamaño como en naturaleza química.

mación de la especie de oxo-ferrilo de valencia alta, llamada compuesto I (Cp I). La estructura electrónica del Cp I generalmente se describe como un estado $[P^{*+}-Fe_{IV}=O^{2-}]$, un hierro(IV) unido a un átomo de oxígeno dianiónico, y una porfirina catiónica. El Cp I es un oxidante muy fuerte y generalmente es capaz de abstraer un átomo de hidrógeno de carbonos alifáticos, generando el llamado Cp II y el radical del sustrato correspondiente. En el mecanismo típico de hidroxilación, el ligando OH del Cp II resultante (generalmente descrito como $[P^{*+}-Fe_{III}-OH^-]$ o $P-Fe_{IV}-OH^-$) se une nuevamente al radical alifático produciendo el alcohol correspondiente y devolviendo la enzima a su estado de reposo. Sin embargo, se ha demostrado experimentalmente que el Cp II también es capaz de catalizar la hidroxilación y otras reacciones oxidativas [19].

1.4.4 Aplicaciones

En los últimos años, ha habido un creciente interés en utilizar los CYPs como biocatalizadores para la producción de ciertos compuestos orgánicos, tales como drogas, agroquímicos e ingredientes alimentarios, cuya producción por síntesis química reviste una alta complejidad y/o un elevado costo. De este modo, se aprovecha la capacidad que tienen los microorganismos de modificar químicamente una amplia variedad de compuestos orgánicos de una forma muy específica [20, 21, 22]. Estos procesos catalizados por enzimas en la industria son cada día más numerosos, ya que presentan ventajas frente a los catalizadores no biológicos: la especificidad de sustrato, la regioespecificidad, la estereoespecificidad, y las condiciones suaves y poco agresivas en las que llevan a cabo las reacciones.

Los CYPs poseen además un tremendo potencial para la síntesis y descubrimiento de nuevas drogas, así co-

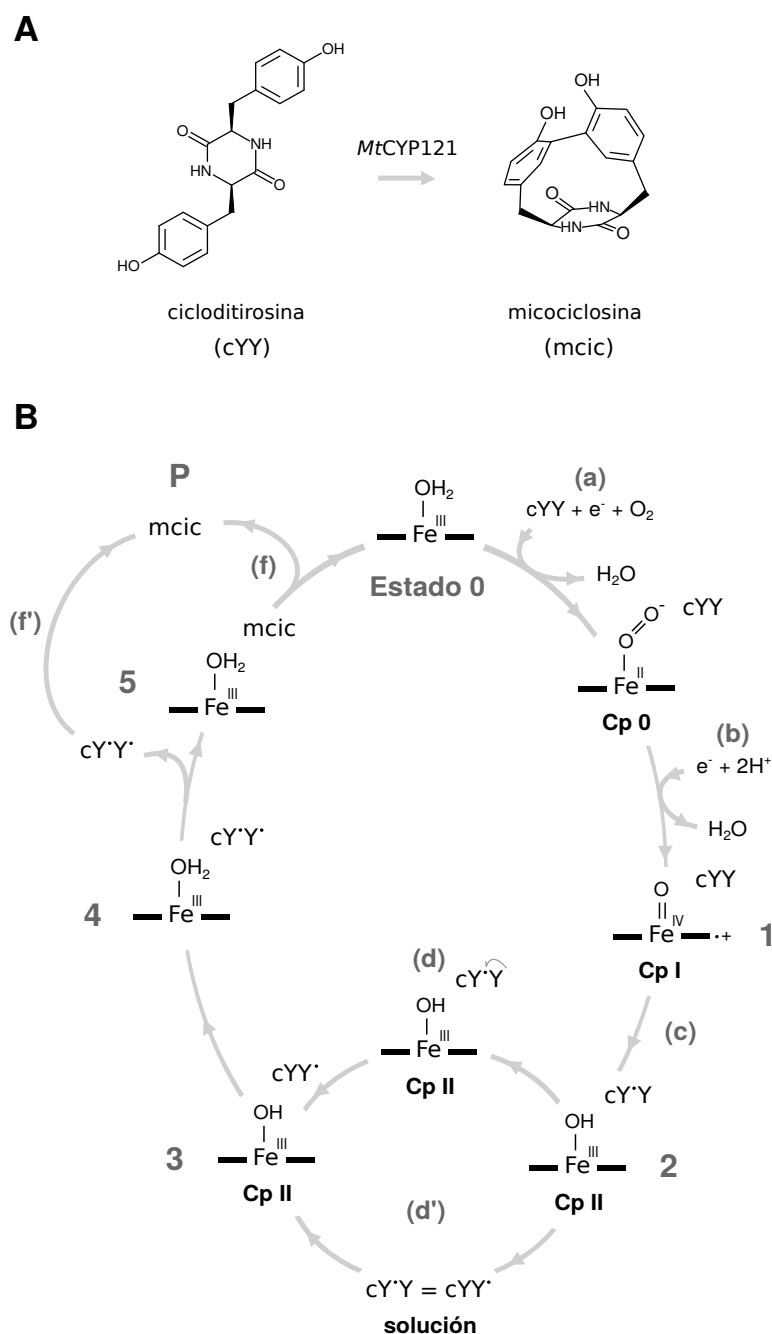


Figura 1.4: A. Reacción catalizada por el MtCYP121, mostrando la estructura del reactivo cicloditirosina (cYY) y el producto micociclosina (mcic). B. mecanismo de reacción propuesto para MtCYP121. Los estados de oxidación/coordinación del hemo relevantes se encuentran etiquetados con números: 0: Estado férrico de reposo con un agua como ligando axial, 1: Compuesto I con el sustrato unido, 2 y 3: Compuesto II con un radical tirosilo en las tirosinas cercana y lejana, respectivamente, 4: Proteína férrica unida al birradical, y 5: Proteína férrica unida al producto micociclosina (mcic). Las reacciones se designan con letras: (a) unión y reducción de oxígeno, (b) ruptura del enlace O-O y liberación de agua, (c) formación del primer radical en la tirosina cercana, (d) transferencia electrónica intramolecular tirosina-tirosina dentro de la proteína, (d') liberación del radical cYY y reingreso del mismo al sitio activo de MtCYP121 con el radical localizado en la tirosina lejana, (e) formación del segundo radical y liberación de agua, (f) formación del enlace C-C dentro de MtCYP121 y liberación de mcic, (f') formación del enlace C-C en solución.



Figura 1.5: Diferentes campos de investigación en p450s y sus aplicaciones.

mo también en su desarrollo. Aunque varios microorganismos han sido utilizados históricamente para la síntesis de drogas utilizando una reacción de hidroxilación, la aplicación comercial bien establecida de los CYPs es en la biotransformación de esteroides, tales como la hidroxilación de cortexolona a hidrocortisona por *Curvularia sp* [23] a una escala de ~100 toneladas por año. Adicionalmente, otro CYP expresado en *E. Coli* es utilizado para producir provastatina a partir de compactina [24]. Muchos pasos en la síntesis de antibióticos involucran p450s bacterianos, por ejemplo la síntesis de eritromicina y tetracenomicina requiere de la P450 eryF (CYP107) [25]. De modo similar, varios CYPs de plantas están involucrados en la producción de Taxol, una droga ampliamente utilizada en quimioterapia [26]. En otro ejemplo, la producción de la droga anticancerígena alcohol perílico a partir de limoneno se logró a través de la CYP153 expresada en *Pseudomonas putida* [27]. Estos ejemplos muestran que el verdadero potencial de los CYPs en la producción de drogas o intermediarios de drogas recién comienza a ser explorado [28].

Por otro lado, los CYPs juegan un rol fundamental en la farmacología. Los metabolitos de drogas que producen los CYPs presentes en el hígado y otros tejidos y órganos, en general son biológicamente activos, y entender sus efectos es crucial para evaluar la eficacia, la toxicidad y la farmacocinética de una droga. Sin embargo, tales estudios requieren grandes cantidades de metabolitos puros, los cuales como se mencionó anteriormente, pueden ser difíciles de sintetizar por métodos químicos. Una ruta alternativa para sintetizarlos es utilizar CYPs humanas directamente. Aunque este enfoque conlleva algunas limitaciones, como actividad reducida, estabilidad y factibilidad de expresión en *E. coli*, recientemente se ha desarrollado una estrategia de inmovilización con

varios CYPS para sintetizar metabolitos [29, 30]. Las enzimas inmovilizadas exhiben una mayor vida media. Esto abre la posibilidad de construir biorreactores utilizando la tecnología de biochips para la síntesis de metabolitos de drogas.

Los CYPs son empleados también como blancos para drogas. Existen ejemplos como el de la aromatasa p450 humana (CYP19A1), que oxida los andrógenos testosterona y androstenediona produciendo los estrógenos 17 β -estradiol y estrona, respectivamente, la cual ya es utilizada como target por la droga "letrozole" para el tratamiento del cáncer de mamas. En este sentido, los futuros avances en la biología estructural de la aromatasa podrían proveer las bases para el diseño racional de drogas más efectivas contra este tipo de cáncer [31]. Además, en las últimas décadas, también las CYPs microbianas han emergido como potenciales blancos para drogas, en particular desde el descubrimiento de que algunas de ellas, como la CYP51 (14-alfadesmetilasa de esteroides) en levadura [32] junto a otros CYPs, son esenciales para la viabilidad de *Mycobacterium tuberculosis*. Además, el éxito de los inhibidores azólicos [33], que actualmente se utilizan para tratar infecciones fúngicas ha reforzado la búsqueda de inhibidores para otros miembros de esta familia tan importante [34, 35].

Adicionalmente a la producción de drogas y metabolitos de las mismas, otros productos comerciales como tintas y pesticidas pueden ser sintetizados utilizando CYPs. Ejemplos de ello son la producción de índigo e indirrubina en cultivos bacterianos y plantas de tabaco transgénicas, lo cual posee un rol interesante para la industria de las tintas [36, 37]. Otro campo de aplicación de la producción de tintas por parte de CYPs es la horticultura, por ejemplo la producción de flores de colores exóticos, como ser las rosas azules se realiza transfiriendo genes de CYPs involucrados en la coloración de otras plantas [38].

Una aplicación adicional de estas enzimas en farmacología es su utilización como biosensores para monitorear niveles de droga en el plasma sanguíneo [39]. Monitorear consistentemente los niveles de droga es crítico, ya que la presencia de variantes genéticas o interacciones droga-droga o droga-alimento alteran la respuesta a drogas o generan toxicidad [40]. Es más, los CYPs como biosensores pueden ser utilizados para detectar contaminantes alimenticios, tales como carbamato y pesticidas organofosforados para mejorar la seguridad alimentaria [39]. Así, una enzima CYP inmovilizada sobre un electrodo, ya sea que se una a una droga o la convierta en su producto oxidado, puede actuar como un biosensor.

Pero las aplicaciones de estas enzimas tan versátiles no terminan aquí. También existen desarrollos que muestran su potencialidad para ser empleadas en la terapia génica con enzimas que activan prodrogas (GDEPT) para el tratamiento del cáncer [41, 42]. Las ciencias ambientales son otro de los campos en los que los CYPs son demandados, ya que algunos de ellos poseen la capacidad de detoxificar ciertos contaminantes ambientales [43, 44].

Pero además de aprovechar las habilidades naturales de biotransformación de compuestos orgánicos que poseen los microorganismos, desde que la ingeniería de proteínas se ha vuelto un proceso sencillo, existe un gran potencial en la adaptación de las actividades de las CYPs para producir metabolitos de alto valor. En este sentido, entender "la naturaleza de la bestia" es un requisito fundamental para poder modificar las propiedades catalíticas de las enzimas de manera predecible. En tal sentido, todas las aplicaciones aquí mencionadas que se muestran resumidas en la Figura 1.5, constituyen vastas razones por las cuales resulta imperioso avanzar en el estudio de las propiedades estructurales y mecanísticas de las p450s.

1.5 Objetivos de esta tesis

El objetivo general de esta tesis doctoral consistió en comprender la relación entre la estructura-dinámica y la función en la familia de los citocromos p450s bacterianos.

En primer lugar se comenzó estudiando en detalle el mecanismo de reacción de la enzima CYP121, un miembro singular dentro de la familia de los p450s por las características de la reacción que cataliza. Para ello se aplicaron diversas técnicas de simulación molecular basadas en dinámica molecular clásica y cálculos híbridos cuántico-clásicos. A partir de los resultados obtenidos se consiguió elaborar una propuesta mecanística para el rol de esta enzima.

A continuación se procedió a realizar un estudio más abarcativo de la diversidad estructural y funcional de la familia de los CYPs bacterianos con el objetivo de desarrollar un método que tomando como punto de partida la secuencia de un dado CYP y determinando una serie de propiedades físico-químicas, permita predecir el rango de sustratos y la reacción que éste cataliza. Este desafío requirió el empleo de diversas herramientas derivadas de la bioinformática, tales como análisis y alineamiento de secuencias y estructuras proteicas, construcción de árboles filogenéticos, modelado de estructuras proteicas, comparación de estructuras de ligandos, docking molecular e integración de datos provenientes de diversas bases de datos.

2

MÉTODOS

En este capítulo se presentarán de modo general los métodos que serán utilizados en la presente tesis, los cuales se dividen en métodos de simulación computacional y métodos bioinformáticos. Los detalles relativos a la implementación de los mismos para la resolución de problemáticas inherentes a cada capítulo, serán detallados en los capítulos correspondientes.

2.1 Simulación computacional

Al utilizar métodos de simulación computacional es importante tener presente que éstos se basan en modelos, es decir, en representaciones del mundo real que tienen características y limitaciones que los definen y diferencian entre sí, principalmente debido al nivel de aproximación de cada uno. Es preciso entonces identificar primero cuál de estos métodos es el indicado para dar respuesta a los interrogantes fenomenológicos que se planteen. Por un lado, los modelos más detallados, provenientes de la química cuántica (QM, por sus siglas en inglés, *Quantum Mechanics*) consideran de manera explícita la densidad electrónica de los átomos intervinientes en un sistema, por lo que permiten estudiar y resolver problemas referentes a mecanismos de reacciones químicas (formación y ruptura de enlaces covalentes), transferencia de cargas, estados electrónicamente excitados, espectroscopías IR, Raman, RMN, y UV-Vis. Por otro lado, los modelos de mecánica molecular o clásica (MM, por sus siglas en inglés, *Molecular Mechanics*), caracterizados por un nivel de detalle menor, consideran a los átomos como esferas rígidas con cargas puntuales, por lo que ya no permiten describir cambios en la distribución electrónica. Estos modelos MM son utilizados para resolver problemas referentes a la dinámica de sistemas de biomoléculas (proteínas, lípidos, carbohidratos, ácidos nucleicos) dependiente del tiempo y del entorno, cambios conformacionales pequeños, difusión de ligandos, determinación de factores moleculares que intervienen en una función biológica particular, interacciones proteína-proteína, proteína-ligandos, proteína-inhibidores, proteína-ADN. Dado que los modelos MM describen a los sistemas de manera mucho más aproximada que los modelos QM, el costo computacional asociado a la simulación de sistemas con un modelo u otro varía inmensamente. Mientras que en general suelen estudiarse sistemas QM de un tamaño aproximado de entre 50 y 150 átomos con tiempos de simulación de la escala de decenas a cientos de ps, los sistemas MM suelen abarcar hasta 3 órdenes de magnitud más (150.000 átomos), y tiempos de la escala de cientos de nanosegundos. En las secciones siguientes se hará una breve descripción sobre cómo funcionan los métodos cuánticos, clásicos e híbridos cuántico-clásicos, seguido de técnicas de muestreo de la superficie de energía potencial de los sistemas.

2.1.1 Métodos basados en la mecánica cuántica

Cuando se desea predecir las propiedades de sistemas moleculares microscópicos, considerando la distribución electrónica de un sistema, se deben resolver las ecuaciones de la mecánica cuántica, que tienen como

ecuación fundamental a la ecuación de Schrödinger dependiente del tiempo:

$$i\hbar \frac{\partial}{\partial t} \Psi(\mathbf{r}, t) = \left[\frac{-\hbar^2}{2\mu} \nabla^2 + V(\mathbf{r}, t) \right] \Psi(\mathbf{r}, t) \quad (2.1)$$

Esta ecuación, donde ∇ es el gradiente, es una ecuación diferencial cuya solución consiste en obtener la función de onda Ψ , que da la descripción matemática de un sistema de partículas cuánticas, sujetas al potencial dado (V). La función de onda contiene toda la información necesaria para calcular las propiedades del sistema. El objetivo básico de la química cuántica es resolver la ecuación de Schrödinger para un sistema molecular, pero dado que resulta imposible resolverla de forma exacta, se deben buscar soluciones aproximadas. El primer paso en el camino a la solución consiste en suponer que el potencial es independiente del tiempo. A partir de esto se deduce una ecuación independiente del tiempo, cuya solución solo depende de las coordenadas espaciales de las partículas, que además está unívocamente relacionada con la energía del sistema. Si se define ahora al operador hamiltoniano como:

$$\hat{H} = \frac{-\hbar^2}{2\mu} \nabla^2 + V \quad (2.2)$$

la ecuación de Schrödinger se reduce a:

$$\hat{H}\Psi(\mathbf{r}) = E\Psi(\mathbf{r}) \quad (2.3)$$

La ecuación de Schrödinger es, por lo tanto, una ecuación diferencial en derivadas parciales de autovalores, en donde un operador H actúa sobre la autofunción $\Psi(r)$, obteniendo como resultado a la función multiplicada por un escalar E , que es la energía del sistema. La segunda simplificación que se debe tener en cuenta es la denominada aproximación de Born-Oppenheimer, que consiste en desacoplar el movimiento de los núcleos del de los electrones. Este procedimiento se justifica ya que la velocidad relativa de los núcleos con respecto a la de los electrones resulta ser despreciable. Así, se asume que los electrones se mueven en un potencial de cargas positivas estáticas dadas por las posiciones nucleares. El problema queda ahora reducido a obtener la función de onda para el siguiente operador denominado hamiltoniano electrónico, para un sistema compuesto por M núcleos y N electrones:

$$\hat{H}_{el} = - \sum_{i=1}^N \frac{1}{2} \nabla_i^2 - \sum_{i=1}^N \sum_{A=1}^M \frac{Z_A}{r_{iA}} + \sum_{i=1}^N \sum_{j>i}^N \frac{1}{r_{ij}} \quad (2.4)$$

En este operador se pueden distinguir tres términos, el primero representa la energía cinética electrónica, el segundo es el potencial asociado a la interacción núcleo-electrón y el tercero es el potencial de la interacción electrón-electrón. Si bien con las aproximaciones mencionadas se ha simplificado el problema, aún no se obtuvo un método que permita resolver la ecuación de Schrödinger de manera exacta.

La clave para resolver de manera aproximada ecuaciones diferenciales de este tipo es utilizar el principio o teorema variacional. Éste postula que al evaluar la energía con una función de prueba $\psi'(r)$ (diferente de aquella que es solución exacta de la ecuación que no se conoce), la energía que se obtendrá será siempre mayor a la energía real (la correspondiente a la función de onda que es solución exacta) del sistema, o sea:

$$\langle \psi'(r) | \hat{H} | \psi'(r) \rangle \geq E_0 \quad (2.5)$$

donde E_0 representa la energía real del sistema que corresponde a evaluar:

$$\langle \psi(r) | \hat{H} | \psi(r) \rangle = E_0 \quad (2.6)$$

y la función de onda es la función de onda real del sistema, solución exacta de la ecuación de Schrödinger para el estado fundamental. Esto permite proponer como estrategia para encontrar soluciones aproximadas de la ecuación de Schrödinger, la búsqueda de aquellas funciones que minimicen la energía del sistema. Dado el

operador hamiltoniano, existe un conjunto infinito de soluciones (cada una representada por una autofunción $|\Psi_i\rangle$ y su autovalor ϵ_i) que se pueden ordenar de manera creciente según:

$$\hat{H}|\Psi_i\rangle = \epsilon_i|\Psi_i\rangle \quad i = 0, 1, 2, \dots \quad \epsilon_0 \leq \epsilon_1 \leq \dots \leq \epsilon_i \leq \dots \quad (2.7)$$

Aquí si los diferentes ϵ_i son los autovalores exactos, entonces ϵ_0 es la energía del estado fundamental, ϵ_1 es la energía del primer estado excitado, etc.

Teoría del Funcional de la Densidad (DFT)

Un formalismo alternativo para describir los sistemas atómico moleculares cuánticamente, que se ha popularizado en los últimos años debido a su eficiencia computacional y a la calidad de los resultados obtenidos, se conoce como teoría del funcional de la densidad (DFT)[45]. La teoría de los funcionales de la densidad se basa en los teoremas de Hohenberg y Kohn formulados en los años 60 que permiten describir el problema electrónico usando la densidad de carga electrónica (ρ). El primero de los teoremas establece que la función de onda, y por ende cualquier propiedad del sistema, se halla unívocamente determinada por la densidad electrónica del sistema. En el contexto de DFT entonces, la energía E en función de la densidad está dada por:

$$E[\rho] = T[\rho] + V_{ee}[\rho] + V_{ne}[\rho] \quad (2.8)$$

donde los términos corresponden a la energía cinética (T), la interacción electrón-electrón (V_{ee}) y la interacción núcleo-electrón (V_{ne}). Usando diferentes definiciones de cada uno de los términos anteriores, y utilizando la aproximación de Kohn y Sham que considera un sistema de referencia de electrones no interactuantes, la fórmula final para el funcional de la densidad puede escribirse como:

$$E[\rho] = T_s[\rho] + V_{ne}[\rho] + \frac{1}{2} \iint \frac{\rho(r_1)\rho(r_2)}{r_{12}} dr_1 dr_2 + E_{xc}[\rho] \quad (2.9)$$

Con esta definición se puede calcular la energía para todos los términos, excepto por la relativamente pequeña contribución del término de intercambio y correlación (E_{xc}), cuya forma funcional no es conocida. La teoría del funcional de la densidad, sumado a la aproximación de Kohn y Sham se basa en un funcional exacto por definición, donde los efectos debidos a la naturaleza cuántica del electrón, el intercambio y la correlación son descritos por el término E_{xc} . Sin embargo, como la dependencia funcional de E_{xc} con la densidad es desconocida, esto implica que la precisión y calidad de un cálculo basado en DFT estará íntimamente relacionada con la aproximación utilizada para darle forma al funcional de intercambio y correlación. Existen muchos funcionales de intercambio y correlación. Los más utilizados en sistemas bioquímicos son: BLYP, B3LYP, PBE [45, 46, 47].

Ya sea resolviendo la ecuación de Schrödinger o utilizando DFT, lo que se obtiene de un cálculo cuántico es, para esa configuración (o conformación) del sistema particular, la estructura electrónica (función de onda) y la energía.

2.1.2 Métodos basados en la mecánica clásica

Muchos sistemas de interés en química y biología requieren del estudio molecular de sistemas de gran cantidad de átomos, como por ejemplo las proteínas. Para este tipo de sistemas resulta, aún con el poder computacional existente hoy en día, prácticamente imposible su tratamiento completo con mecánica cuántica. Es por esto que en gran cantidad de problemas en los que no se requiere un detalle de la distribución electrónica, se utilizan métodos basados en la mecánica clásica, comúnmente denominados "mecánica molecular"(MM). En estos métodos se ignora el movimiento de los electrones, y se calcula la energía exclusivamente en función de las posiciones de los núcleos. Sin embargo, la mecánica cuántica se utiliza para el desarrollo de los distintos parámetros requeridos para el cálculo, que se explicarán en la próxima sección.

Un cálculo de dinámica molecular es la sucesión de una serie de pequeños pasos elementales que han de realizarse millones, o incluso billones de veces: En primer lugar se calcula la energía potencial asociada a una

conformación de la proteína. Para ello se utiliza un funcional clásico muy sencillo denominado campo de fuerzas (*force-field*) que ha sido parametrizado para reproducir resultados experimentales en sistemas modelos, o cálculos mecanocuánticos de alto nivel. Luego el cálculo de la derivada de la energía respecto a las posiciones atómicas permite obtener la fuerza que actúa sobre cada partícula del sistema. La segunda ley de Newton permite entonces asignar las aceleraciones que actuarían sobre cada partícula (de masa conocida) de la molécula. Escogiendo un incremento de tiempo tan pequeño que permita suponer que en él la aceleración es constante, es posible obtener numéricamente las velocidades que actúan sobre cada partícula del sistema por simple integración de las aceleraciones respectivas. Finalmente, la integración de las nuevas velocidades permite obtener las nuevas posiciones asignadas a cada partícula de la proteína. Al final de este proceso se ha conseguido realizar un paso de integración que corresponderá a la evolución del sistema en un pequeño elemento de tiempo. Repitiendo este proceso millones de veces, se simula la evolución temporal del sistema.

Campos de fuerza clásicos

En los métodos de mecánica molecular, la energía potencial viene dada por lo que se denomina campo de fuerza, que no es más que una expresión de la energía potencial dependiente de las coordenadas de los núcleos y una serie de parámetros. La expresión del campo de fuerza AMBER 99SB [48] - el cual fue utilizado para los cálculos realizados en este trabajo - contiene dos clases de contribuciones, las contribuciones de unión y las de no unión. Las de unión, se calculan para átomos formando un enlace o como máximo a 2 átomos de distancia, e incluyen penalidades para las desviaciones respecto de su valor de equilibrio de distancias de enlace, ángulos y ángulos diedros. Mientras que las dos primeras se representan a través de un potencial armónico centrado en la distancia de equilibrio, los ángulos diedros o torsiones están descriptos con una función periódica. Las contribuciones de no unión, se computan entre átomos que están en diferentes moléculas o a más de 4 enlaces de distancia, y se dividen en interacciones electrostáticas basadas en la ley de Coulomb, por un lado, y un potencial de Lennard-Jones, que describe en forma aproximada las interacciones dispersivas y la repulsión originada por el principio de exclusión de Pauli a distancias muy cortas. La expresión de la energía potencial resulta en el caso del campo de fuerzas AMBER utilizado en esta tesis en:

$$E(r^N) = \sum_{\text{enlaces}} \frac{k_{b,i}}{2} (r_i - r_{i0}) + \sum_{\text{angulos}} \frac{k_{\theta,i}}{2} (\theta_i - \theta_{i0}) + \sum_{\text{torsiones}} \frac{V_n}{2} \cos(n\omega - \gamma) + \sum_{i=1}^{N-1} \sum_{j=i+1}^N \left\{ \epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 2 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] + \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \right\} \quad (2.10)$$

Analicemos en detalle cada una de las contribuciones al campo de fuerza clásico. El primer término modela la contribución de estiramiento de los enlaces, el cual se describe a través de un potencial armónico. En éste, $k_{b,i}$ corresponde a la constante de fuerza asociada a la unión i y r_{i0} a la distancia de equilibrio de la misma unión. Si bien una mejor descripción de un enlace químico se obtendría con un potencial como por ejemplo el sugerido por Morse, este contiene un parámetro adicional y no es tan comúnmente utilizado en cálculos MM debido a su costo adicional. Por otra parte, el potencial armónico resulta una buena y sencilla aproximación al potencial real en la zona cercana al mínimo, donde no hay ruptura de enlaces. El segundo término corresponde a las contribuciones dadas por las flexiones angulares. Estas también se representan con un potencial armónico de constante $k_{\theta,i}$ y valor de equilibrio θ_0 . Normalmente, la energía requerida para modificar un ángulo es menor a la requerida para variar la distancia entre dos átomos, por lo que los valores de $k_{\theta,i}$ suelen ser menores que los de las constantes $k_{b,i}$. El término asociado a las torsiones o ángulos diedros presenta tres parámetros: V_n corresponde a la barrera energética asociada a la rotación entre dos mínimos sucesivos, n representa la multiplicidad del potencial, es decir indica el número de mínimos encontrados entre 0° y 360° , y por último γ , que corresponde a la fase de la función sinusoidal.

El último término de la ecuación 2.10 contiene las contribuciones de no unión, y depende de la distancia entre pares de átomos, r_{ij} . La primera parte, corresponde al potencial de Van der Waals, y contiene los parámetros

ϵ_{ij} , asociado a la profundidad del pozo de energía, y σ_{ij} , correspondiente a la distancia entre los átomos correspondiente al mínimo de energía. La parte electrostática contiene como parámetros las cargas parciales sobre los átomos q_i y q_j . En el campo de fuerzas de AMBER, la distribución de cargas es representada a través de cargas puntuales ubicadas en el centro de las coordenadas de cada átomo. Estas cargas se obtienen de manera de reproducir el potencial electrostático de la molécula, a partir de cálculos cuánticos.

Para disminuir el costo computacional asociado a los términos de no unión, se utiliza un radio de corte, tal que para los átomos que se encuentran a una distancia mayor que este radio de corte estas interacciones no se computan. En el caso de las interacciones de Lennard-Jones, al ser de corto alcance, esto no genera un problema siempre que el radio de corte sea lo suficientemente grande (usualmente entre 8 y 12 Å). En el caso de las interacciones coulómbicas, de mayor alcance, en los casos en que se utilizan condiciones periódicas de contorno en la simulación se utiliza una metodología denominada sumas de Ewald, un método originalmente desarrollado para el estudio de cristales iónicos, en su implementación denominada Particle Mesh Ewald (PME)[49, 50]. Esta metodología se utiliza para calcular las interacciones electrostáticas totales entre los átomos presentes en cada celda unidad, de forma muy eficiente. La energía electrostática se calcula en dos partes: la primera parte, en la cual a las interacciones electrostáticas de corto alcance se calculan en el espacio real mientras que la segunda parte (en el espacio recíproco) se calcula utilizando la transformada de Fourier. Este procedimiento tiene como ventaja una convergencia mucha más rápida que utilizando cálculos de interacciones directas. De esta manera se logra tener en cuenta las interacciones electrostáticas de largo alcance más allá del radio de corte de las interacciones de no unión, en forma altamente eficiente.

Parámetros asociados

Es importante notar que la mera expresión de la energía potencial no es suficiente para definir un campo de fuerza. El campo de fuerzas es constituido además por la gran cantidad de parámetros asociados a cada uno de los sumandos en la energía potencial. Para el caso de simulación de proteínas, el campo de fuerza de AMBER contiene los parámetros asociados a los aminoácidos naturales presentes en la mayoría de las proteínas, en sus distintos estados de oxidación o protonación, cuando corresponda. En este contexto, puede hablarse de la transferibilidad del campo de fuerza. Es decir, el hecho de que el mismo conjunto de parámetros puede utilizarse para moléculas asociadas. Por ejemplo, los parámetros para un residuo de triptofano serán los mismos en todas las proteínas estudiadas, y no será necesaria su parametrización en cada proteína que se desee estudiar. En el campo de fuerza de AMBER, los átomos dentro de cada residuo reciben un nombre de átomo y un tipo de átomo. Los tipos de átomo permiten que no sea necesario incluir parámetros para cada átomo incluido en cada uno de los aminoácidos. Distintas especies pueden poseer el mismo tipo de átomo, y los parámetros asociados a las uniones, los ángulos y los diedros (los denominados parámetros de unión) se dan por tipo de átomo, en lugar de para cada átomo de cada residuo en particular. Si bien en el campo de fuerza se incluyen parámetros específicos para todos los aminoácidos naturales y otras moléculas comúnmente encontradas en los sistemas biológicos, para incluir modificaciones post-traduccionales como la fosforilación, será necesario obtener los parámetros necesarios para realizar la simulación. El concepto de transferibilidad del campo de fuerza se utiliza a la hora de simular compuestos tipo droga y no tener que obtener de cero los parámetros de enlace (distancias de enlace, ángulos, diedros y ángulos impropios). En este trabajo utilizamos la estrategia desarrollada en el *Generalized AMBER Force Field* (GAFF). [51] sólo siendo necesario derivar las cargas parciales mediante cálculos cuánticos y el algoritmo RESP (*Restraint Electrostatic Potential*).

Modelos para el solvente

La gran mayoría de los fenómenos bioquímicos de interés ocurren en solución acuosa; no es de extrañar entonces el esfuerzo que químicos y físicos han hecho para desarrollar modelos que permitan simular esta vital sustancia. Dentro de los componentes utilizados para modelar al agua se encuentran fluidos dieléctricos, cargas puntuales fijas, cargas fluctuantes, dipolos puntuales, potenciales tipo Lennard-Jones, potenciales armónicos de

enlace, entre otros. La utilización de dichos componentes, así como la parametrización de los mismos, dependen en buena medida de qué propiedad interesa reproducir y con qué nivel de precisión en relación al costo computacional asociado. En los cálculos realizados en este trabajo, se utilizó un modelo de agua simple denominado TIP3P. En los modelos de agua simples, cada molécula de agua se mantiene en una geometría rígida, y la interacción entre moléculas de agua se describe a través de interacciones coulómbicas y de Lennard-Jones. En el modelo TIP3P se ubican tres cargas puntuales, una sobre cada átomo de la molécula. La carga negativa ubicada en el átomo de oxígeno ($q_O = -0.834$) es compensada por las cargas positivas en los átomos de hidrógeno ($q_H = +0.417$). La distancia entre el oxígeno y los átomos de hidrógeno ($r_{OH} = 0.96 \text{ \AA}$), así como el ángulo HOH (104.5°) se mantienen fijos (mediante el uso del algoritmo de SHAKE [52]). Otros modelos de agua simples usualmente utilizados son por ejemplo el modelo SPC/2 o su actualización SPC/2, y el modelo TIP4P, que utiliza cuatro puntos en vez de tres para describir la distribución de cargas. Los valores de los parámetros asociados a los modelos de agua simples se desarrollan de forma de reproducir las propiedades del agua medidas experimentalmente como la densidad, la función de distribución radial, y otras propiedades fisicoquímicas.

2.1.3 Métodos híbridos cuántico-clásicos (QM/MM)

Los métodos basados en la mecánica cuántica presentan información de la distribución electrónica del sistema y son imprescindibles cuando se desea estudiar procesos reactivos, pero traen aparejado un gran costo computacional. Por otro lado, los métodos llamados de mecánica molecular clásica (basados en campos de fuerza) presentan un costo computacional extremadamente menor, y permiten estudiar fenómenos estructurales asociados a moléculas de gran cantidad de átomos. Resulta entonces natural pensar que pueden combinarse ambas metodologías para el estudio de un sistema en el que sea necesaria una descripción de la densidad electrónica de una cierta región, que se debe describir con mecánica cuántica, y el resto del sistema puede tratarse con mecánica clásica. Un ejemplo de aplicación de este tipo de metodologías es el estudio de reacciones enzimáticas. [53, 54, 55, 56, 57, 58].

En ellas, el sitio reactivo debe tratarse necesariamente con mecánica cuántica, pero el resto de la enzima puede tratarse clásicamente. Esto constituye un avance significativo respecto de la utilización de sistemas modelo para tratar este tipo de problemáticas, en los que solo se incluía el sitio activo y algunos grupos pequeños relevantes. En las metodologías híbridas se considera el efecto de toda la proteína, en la mayoría de los casos crucial en el desarrollo del fenómeno reactivo. Para aplicar una metodología QM-MM, en primer lugar es necesario definir el sector del sistema que se tratará con mecánica cuántica, que llamaremos subsistema QM. El resto del sistema subsistema MM se tratará con mecánica clásica. Para la elección, debe tenerse en cuenta que cuanto mayor sea el subsistema QM, mayor calidad tendrá el cálculo, pero también su costo se incrementará significativamente. En el caso de que la simulación a realizar consista en el estudio de una molécula inmersa en un solvente, resulta natural elegir como subsistema QM a la molécula a estudiar y tratar al solvente con MM. En el caso del estudio de una reacción enzimática, la elección es más compleja. Resulta imprescindible incluir en el subsistema QM todos los átomos que incluyan enlaces que puedan romperse o formarse durante el proceso de interés. Sin embargo, para obtener una buena representación del problema de estudio, normalmente se requiere incluir una mayor cantidad de átomos que simplemente los involucrados en la reacción propiamente dicha.

Dentro de los métodos QM-MM, pueden distinguirse dos clases de metodologías: los esquemas aditivos y los esquemas sustractivos. La metodología utilizada en este trabajo corresponde a un esquema aditivo, por lo que nos concentraremos en este tipo de estrategias. En los métodos QM-MM aditivos, el Hamiltoniano consiste en la suma de la contribución del subsistema QM (H_{QM}), el subsistema MM (H_{MM}) y un término de acoplamiento QM-MM (H_{QM-MM}): -

$$H = H_{QM} + H_{MM} + H_{QM-MM} \quad (2.11)$$

En este trabajo de tesis, se utilizó el programa Hybrid-SIESTA que ha sido desarrollado por el Dr. Martí y colaboradores [55]. Este programa calcula la contribución cuántica (H_{QM}) a nivel de DFT, teniendo en cuenta sólo los átomos cuánticos, pero considerando a los electrones inmersos en el potencial generado no sólo por los

núcleos de éstos, sino además por el potencial generado debido a las cargas parciales de los átomos clásicos. La contribución clásica (H_{MM}) es calculada empleando el campo de fuerzas de AMBER sin modificaciones. El término crítico es el del acoplamiento cuántico-clásico (H_{QM-MM}). Este término también está compuesto por tres contribuciones y se calcula de acuerdo a:

$$E_{QM-MM} = \sum_{i=1}^C q_i \int \frac{\rho(r)}{|r - \tau_i|} dr + \sum_{i=1}^C \sum_{\alpha=1}^Q \frac{q_i Z_{\alpha}}{|R_{\alpha} - \tau_i|} + E_{QM-MM}^{LJ} \quad (2.12)$$

Siendo C el número de átomos clásicos con cargas parciales q_i y Q el número de átomos cuánticos con índice i , r es la coordenada electrónica, τ_i es la posición del núcleo cuántico y R_{α} es la posición del núcleo clásico. El primer término representa la interacción entre la densidad electrónica cuántica y las cargas parciales de los átomos clásicos. El segundo representa la interacción entre los núcleos de los átomos cuánticos y las cargas parciales clásicas. Estos dos términos representan la interacción electrostática entre los dos subsistemas. El tercer término representa la interacción de Van Der Waals y repulsiva a corto alcance entre los átomos de ambos sistemas y se calcula usando un potencial de Lennard-Jones idéntico al usado en el campo de fuerzas clásico.

El resultado de un cálculo híbrido consiste, de manera análoga a uno QM, en la obtención de la estructura electrónica del subsistema QM embebido en el entorno clásico, y la energía (y fuerzas) para todo el sistema QM+MM.

2.1.4 Exploración de la superficie de energía potencial (SEP)

La superficie de energía potencial (SEP), definida en el contexto de la aproximación de Born-Oppenheimer[45], es una función escalar que define el valor de la energía potencial del sistema en función de las coordenadas de los núcleos. En otras palabras, es la hipersuperficie definida por la energía calculada por un hamiltoniano QM o un campo de fuerzas MM para cada una de las posibles configuraciones (formas de ordenar los átomos) del sistema. Por ejemplo, en un sistema de N átomos depende de $3N$ variables, asociadas a las coordenadas cartesianas de cada uno de los núcleos.

El conocimiento de la superficie de energía potencial permite obtener información cinética y termodinámica de un sistema. De la superficie de energía potencial se puede deducir las fuerzas que sienten los átomos. Hay puntos en la superficie para los que la fuerza que sienten todos los átomos es cero (o la derivada de V con respecto a cada una de las coordenadas es cero). A estos puntos se los denomina puntos estacionarios, y pueden ser máximos, mínimos o puntos silla. Los mínimos son importantes porque nos indican las coordenadas para las que tenemos estructuras estables, mientras que las estructuras en los puntos silla corresponden a las de los estados de transición. Para pasar de reactivos a productos hay que pasar cerca de un punto silla y la diferencia de energía entre el mínimo y el punto silla es la barrera de potencial. El camino a través de la superficie de energía potencial que une dos mínimos y pasa por un punto silla, es decir que une los reactivos con los productos a través del estado de transición, se denomina camino de reacción.

Para sistemas grandes (como por ejemplo una proteína) la SEP resulta muy compleja y es imposible conocerla en su totalidad. En esta sección se describirán distintos métodos desarrollados para explorar la SEP de forma eficiente, que permiten concentrar la atención en el área de la SEP que se desea explorar. Si bien existe una gran cantidad de metodologías para realizar esta tarea, se desarrollarán solamente aquellas aplicadas en los cálculos realizados en este trabajo.

Minimización de la energía

En los mínimos de la SEP se encuentran las configuraciones o conformaciones estables. Para encontrarlos, debe aplicarse un algoritmo de minimización. Debido a la complejidad con la que varía la SEP con las coordenadas cartesianas, se utilizan métodos numéricos para obtener los mínimos de la función. Dentro de los diferentes métodos, pueden distinguirse los métodos que usan las derivadas de la función a minimizar, y aquellos que no. Dentro de los primeros, existen métodos que utilizan solo la derivada primera (que es nula en la posición del mínimo), llamados de primer orden, y los que utilizan además las derivadas segundas (deben ser positivas), llamados

de segundo orden. Los métodos utilizados en este trabajo corresponden a métodos de derivadas de primer orden, en particular los métodos llamados del gradiente conjugado y del gradiente descendente. La inclusión de la derivada primera en la determinación del mínimo de energía resulta de mucha utilidad, dado que el gradiente de la SEP indica la dirección hacia donde se encuentra el mínimo, y su valor absoluto otorga la pendiente en el punto. El gradiente de la SEP corresponde al opuesto de la fuerza, por lo que siguiendo la dirección de la fuerza sobre cada átomo puede gradualmente reducirse la energía del sistema. El método del gradiente descendente realiza movimientos en dirección paralela a la fuerza neta con pasos de una longitud arbitraria fijada al inicio del cálculo. En caso de que se realice un paso que produzca una reducción de la energía, la longitud del paso se incrementa por un factor. En caso de producirse un incremento de la energía se procede a reducir la longitud del paso. El método del gradiente conjugado, por su parte, se diferencia del método del gradiente descendente en que la dirección en la que se realiza el movimiento corresponde a la dirección de la fuerza neta más la dirección anterior multiplicada por un factor. Este factor se calcula como el cociente entre el módulo del gradiente del paso actual y el paso anterior. De esta manera, no se observa el fenómeno oscilatorio que se observa en el método del gradiente descendente en las zonas donde hay mínimos muy estrechos. Por último, debe tenerse en cuenta que estos métodos apuntan siempre a la reducción de la energía, por lo que el resultado final dependerá fuertemente de la configuración de partida. En otras palabras, simplemente a través de optimizaciones de geometría como las descritas en esta sección no se podrán atravesar barreras que conecten distintos mínimos, y el mínimo que se obtenga será el más cercano a la estructura de partida. Resulta entonces imprescindible la utilización de otras metodologías que permitan un muestreo más amplio de la SEP. En el contexto de esta tesis, se utilizó la minimización de energía para optimizar la geometría de las estructuras cristalográficas utilizadas como punto de partida para la dinámica molecular.

Simulaciones de dinámica molecular

En los casos de la presente tesis, la complejidad de la superficie de energía potencial involucra la existencia de gran cantidad de mínimos locales. Esto implica que no existe un solo mínimo representativo de cada sistema, sino que existen una serie de conformaciones, que estarán presentes en mayor o menor proporción según sus energías relativas. Cuando se realiza una medida experimental, todas las conformaciones relevantes estarán representadas en el ensamble de moléculas presente en el sistema. Una manera de representar este ensamble es a través de la reconstrucción de la evolución temporal de la dinámica del sistema, asumiendo que el sistema ha evolucionado un tiempo suficiente para que haya podido explorar las distintas conformaciones disponibles (convergencia de la simulación). La hipótesis ergódica [59] postula que el promedio temporal puede reemplazar el promedio sobre las configuraciones de un sistema dado, dando origen así a las simulaciones basadas en evolución temporal, diferenciándose de aquellas que buscan explorar el colectivo de configuraciones espaciales, como el método de Montecarlo. En una simulación de dinámica molecular (DM), se realiza la reconstrucción de la evolución temporal de la dinámica de un sistema a través de la resolución de las ecuaciones de movimiento de Newton. En otras palabras, la trayectoria se obtiene a través de la resolución de las ecuaciones diferenciales asociadas a:

$$\frac{d^2x_i}{dt^2} = \frac{F_{x_i}}{m_i} \quad (2.13)$$

donde se describe el movimiento de una partícula de masa m_i en la coordenada x_i , con F_{x_i} la fuerza en esa dirección. La fuerza sobre cada una de las partículas cambia cuando la partícula modifica su posición, y es función del movimiento de todas las otras partículas. La resolución analítica del problema no es posible, por lo que debe recurrirse a métodos numéricos. El método implementado en el programa AMBER se denomina algoritmo de Verlet[60], basado en el método de las diferencias finitas. En este método, la integración se parte en pequeños pasos, separados por un intervalo de tiempo fijo δt . Se calcula la fuerza en cada partícula como la suma de las interacciones con el resto de las partículas, y a partir de la fuerza puede calcularse la aceleración, y con ella la nueva posición de las partículas en el siguiente paso. Durante el δt , la fuerza es considerada constante. En el

algoritmo de Verlet se utilizan estratégicamente las expansiones de Taylor de las posiciones a diferentes tiempos, según:

$$r(t + \delta t) = r(t) + v(t)\delta t + \frac{a(t)}{2}\delta t^2 + \dots \quad (2.14)$$

$$r(t - \delta t) = r(t) - v(t)\delta t + \frac{a(t)}{2}\delta t^2 + \dots \quad (2.15)$$

Sumando miembro a miembro las expresiones 2.14 y 2.15, y despreciando los términos de orden superior al segundo, se obtiene:

$$r(t + \delta t) \approx 2r(t) - r(t - \delta t) + a(t)\delta t^2 \quad (2.16)$$

Las velocidades, necesarias para la estimación de la energía cinética, no aparecen explícitamente, pero pueden obtenerse de la resta de las mismas expresiones:

$$v(t) \approx \frac{r(t + \delta t) - r(t - \delta t)}{2\delta t} \quad (2.17)$$

Debe prestarse especial atención a la elección del intervalo de tiempo δt . Cuanto menor sea éste, mayor será la cantidad de pasos necesarios para realizar una simulación de una dada cantidad de tiempo, con lo que incrementará el costo computacional. Sin embargo, un intervalo de tiempo demasiado grande dará origen a inestabilidades en el algoritmo de integración. Por otra parte, debe elegirse un δt que sea menor a las escalas de tiempo asociadas a los movimientos que se desean observar, y que permita describir el movimiento de los átomos en la molécula. En la práctica se utiliza un intervalo de tiempo que sea aproximadamente un décimo del movimiento característico más rápido del sistema. En el caso de un sistema de átomos simulado de manera clásica, los movimientos más rápidos están asociados a las vibraciones de los enlaces, que se encuentran en el orden de los fs (1 fs equivale a 10^{-15} s). Por lo tanto, un intervalo de tiempo de 0,1-0,2 fs resulta apropiado. Las vibraciones más rápidas están asociadas a los enlaces en los que uno de los átomos involucrados en la unión es un átomo de hidrógeno. Una estrategia para aumentar el δt y así acelerar los cálculos consiste en utilizar el algoritmo SHAKE[61], que permite fijar estas uniones en sus posiciones de equilibrio.

Normalmente, es deseable realizar las simulaciones de DM a temperatura constante. De acuerdo al principio de equipartición de la energía, la temperatura está relacionada con la energía cinética (K) de la siguiente manera:

$$K = \sum_{i=1}^N \frac{m_i v_i^2}{2} = \frac{3k_B T}{2} N \quad (2.18)$$

donde m_i y v_i corresponden a la velocidad y la masa de la partícula i y N es el número de partículas del sistema. Por lo tanto, una manera simple de mantener la temperatura constante consiste en escalar las velocidades. Sin embargo, este método no refleja la naturaleza fluctuante de la temperatura alrededor de un promedio, y por ello no resulta ser el mejor de los enfoques de control térmico de las simulaciones. Uno de los métodos más utilizados para regular la temperatura en una simulación de DM es el algoritmo de Berendsen[62]. El mismo consiste en acoplar el sistema a un baño externo cuya temperatura se fija en la temperatura deseada, que remueve o entrega calor según sea necesario. Supongamos que a un tiempo dado t el sistema posee una temperatura $T(t)$. Puede verse de la ecuación 2.18 que si las velocidades se escalan por un factor λ , el cambio asociado en la temperatura resulta:

$$\Delta T = (\lambda^2 - 1)T(t) \quad (2.19)$$

En este método, las velocidades se escalan en cada paso, de manera que el cambio en la temperatura es proporcional a la diferencia entre la temperatura del sistema y la temperatura del baño:

$$\frac{dT(t)}{dT} = \frac{1}{\tau} [T_b - T(t)] \quad (2.20)$$

donde τ es un parámetro de acoplamiento que determina cuan fuertemente el baño está acoplado al sistema. El cambio de temperatura entre dos pasos sucesivos separados por δt , corresponde a:

$$\Delta T = \frac{\delta t}{\tau} [T_b - T(t)] \quad (2.21)$$

Con esta expresión, el factor de escalamiento de las velocidades λ resulta:

$$\lambda^2 = 1 + \frac{\delta t}{\tau} \left[\frac{T_b}{T(t)} - 1 \right] \quad (2.22)$$

Cuanto más grande sea este parámetro, más débil será el acoplamiento, por lo que la fluctuación en la temperatura será mayor. Normalmente, la utilización de un factor de acoplamiento entre 100 y 1000 veces más grande que el δt elegido otorga buenos resultados para los sistemas estudiados como en esta tesis. Por otra parte, es también común realizar simulaciones de DM en condiciones de presión constante. Para mantener la presión constante, el sistema debe variar el volumen de la celda en la que se realiza la simulación. Berendsen propuso entonces un esquema similar al planteado para la regulación de la temperatura, por el cual se acopla el sistema a un baño equilibrado a una cierta presión[62]. El volumen de la celda se escala entonces en un factor γ , lo que equivale a escalar las coordenadas atómicas en $\gamma^{1/3}$. Resulta entonces:

$$\gamma = 1 + \kappa \frac{\delta t}{\tau_P} [P - P_b] \quad (2.23)$$

donde $\kappa = -\frac{1}{V} \left(\frac{\partial V}{\partial P} \right)_T$ es la compresibilidad isotérmica, y τ_P corresponde al factor de acoplamiento del baño al sistema. Luego de realizar una simulación de DM, el resultado principal que se obtiene es la trayectoria de cada uno de los átomos que conforman el sistema a lo largo del tiempo simulado. La visualización de la misma a través de algún programa de representación gráfica aporta gran cantidad de información sobre el sistema. Con ella se pueden analizar la evolución de parámetros geométricos claves, como distancias, ángulos, ángulos diedros y otros parámetros de relevancia estructural como interacciones de puente hidrógeno e hidrofóbicas. Por otro lado, en el caso de una macromolécula, puede observarse la dinámica de la estructura terciaria, aunque los movimientos globales asociados a ella muchas veces requieren un largo tiempo de simulación para ser observados.

2.1.5 Estimación de energía libre: Integración termodinámica

El método de integración termodinámica (TI) es uno de los más utilizados para calcular diferencias de energía libre a partir de simulaciones de DM o MC. Consiste en comparar la diferencia en la energía libre entre dos estados dados (por ejemplo, A y B) cuyas energías potenciales U_A y U_B poseen diferentes dependencias de las coordenadas espaciales. Debido a que la energía libre de un sistema no es una simple función de las coordenadas del sistema en el espacio de fases, sino una integral pesada por Boltzmann sobre el espacio de fases (función de partición), la diferencia de energía libre entre dos estados no puede ser calculada directamente. En la integración termodinámica, la diferencia de energía libre se calcula definiendo un camino termodinámico entre los estados e integrando sobre los cambios de entalpía promediados por el ensamble a lo largo del camino. Es decir que se trata de cambiar sutilmente el sistema como está definido por los correspondientes parámetros clásicos del campo de fuerza, para pasar de la condición inicial a la final. El cambio se ejecuta representando el sistema como una combinación lineal de los estados inicial y final de acuerdo al siguiente Hamiltoniano, variando el parámetro de acoplamiento λ en once pasos discretos, desde $\lambda=0$ hasta $\lambda=1$.

$$H = H_A \lambda = (1 - \lambda) H_B \quad (2.24)$$

Donde H corresponde al Hamiltoniano actual (es decir, el campo de fuerzas utilizado para realizar la DM), H_A corresponde al campo de fuerzas clásico con los parámetros correspondientes al estado inicial y H_B lo equivalente para el estado final. El resultado de un cálculo de TI consiste por lo tanto en la obtención del valor correspondiente a la diferencia de energía libre entre los estados A y B (ΔG_{AB}).

2.2 Métodos bioinformáticos

Hasta aquí se han descrito los métodos que permiten estudiar en detalle atómico la estructura y función de una proteína. A continuación nos abocaremos a la descripción de los métodos que nos permiten obtener información funcional y arribar a conclusiones generales para un conjunto de proteínas de una misma familia.

La revolución de datos genéticos que se generó a partir del surgimiento de las técnicas de secuenciación (y continúa acentuándose en la medida en que éstas se vuelven cada vez más poderosas), derivó en la creación de numerosas bases de datos repletas de secuencias de nucleótidos y proteínas. Pero aunque el número de secuencias en las bases de datos está en el orden de varios millones, para la mayoría de ellas no se conoce ningún tipo de información adicional. Es aquí donde aparece el gran desafío de la bioinformática: convertir esa información en conocimiento biológico útil [63].

Para conocer más sobre una determinada secuencia, es preciso realizar diversos análisis, que dependerán del tipo de información que se desee obtener. Una de las etapas claves del análisis de secuencias es el alineamiento entre pares de secuencias para detectar homología y la comparación de una secuencia nueva contra una base de datos de secuencias conocidas para buscar similitudes que nos den alguna pista sobre la secuencia nueva. A su vez, las similitudes a nivel de secuencia pueden ayudar a realizar predicciones sobre la estructura proteica, y las secuencias de proteínas o ADN de diferentes organismos pueden ser comparadas para construir árboles filogenéticos, que representan las relaciones evolutivas entre especies o dentro de una misma familia de proteínas.

La idea principal que subyace al análisis de secuencias se basa en la evolución, dado que las biomoléculas importantes están conservadas entre organismos. Si la función proteica se conserva, entonces la secuencia también debería estar conservada, y esto es lo que permite que las lecciones aprendidas para una proteína puedan ser transferidas a otra si las secuencias en cuestión son lo suficientemente similares. Este concepto, descripto oportunamente en la Introducción, es lo que se denomina falacia por asociación.

2.2.1 Alineamiento de secuencias múltiples (ASM)

Los alineamientos múltiples pueden ser usados para encontrar patrones interesantes que caractericen a ciertas familias de proteínas, para construir árboles filogenéticos, para detectar homología entre secuencias nuevas y familias existentes y para ayudar a predecir la estructura secundaria y terciaria de nuevas secuencias. En general, el alineamiento de múltiples secuencias resulta ser una prueba más confiable de similitud que un alineamiento de a pares. Esto se debe a que la inclusión de un número mayor de secuencias tiende a resolver las ambigüedades que surgen en el alineamiento de a pares. Además, un alineamiento múltiple proporciona más información que un alineamiento de a pares sobre las posiciones de aminoácidos individuales, tales como la conservación general y las relaciones evolutivas. Esto cobra especial relevancia cuando se emplean métodos de comparación de secuencias para la construcción de árboles filogenéticos. El alineamiento múltiple es especialmente útil para ilustrar la conservación de la secuencia. Cuando la conservación se observa en muchas secuencias, puede ayudar a identificar aminoácidos que son importantes para la función o para la integridad estructural del plegado proteico.

Una forma muy conveniente de representar matemáticamente un AMS es a través de un modelo oculto de Markov (del inglés *HMM*). Los modelos ocultos de Markov son modelos probabilísticos que consisten en una serie de estados que representan las posibles letras en cada una de las columnas de un AMS. Este tipo de representación resulta muy útil por ejemplo cuando se cuenta con un perfil de un HMM que representa la secuencia de una familia de proteínas y se quieren buscar más homólogos que pertenezcan a la misma familia. De manera análoga, si se tiene una base de datos de HMMs pre-construidos, es posible determinar dada una única secuencia a cuál de los perfiles se parece más. Esta estrategia puede emplearse para la anotación y clasificación de secuencias. En este trabajo de tesis se construyeron HMMs para clasificar las secuencias de p450s en las etapas iniciales de la predicción de función.

2.2.2 Alineamiento de estructuras proteicas

Aunque es mucha la información que puede obtenerse a partir del análisis de las secuencias, existen numerosas limitaciones al utilizar información unidimensional solamente. Por ejemplo, no es posible determinar el modo de unión de las moléculas que interactúan con las proteínas, tales como sustratos, inhibidores, cofactores, etc. Tampoco es posible analizar estabilidad, plegado o efectos de mutaciones utilizando la secuencia solamente. Para realizar tales análisis necesitamos la estructura, es decir, las coordenadas tridimensionales de todos los átomos de la proteína.

El alineamiento estructural intenta establecer similitud entre dos o más estructuras proteicas en base a su forma y conformación tridimensional. Este proceso no requiere el conocimiento *a priori* de posiciones estructurales equivalentes, por lo que es una herramienta muy valiosa para la comparación de proteínas con baja similitud de secuencias, donde las relaciones evolutivas entre las proteínas no son detectadas fácilmente por las técnicas de alineamiento de secuencias estándar. Típicamente los resultados de un alineamiento estructural consisten en una superposición de los conjuntos de coordenadas atómicas y un valor para la desviación del valor cuadrático medio (RMSD) entre las estructuras, junto a un alineamiento de las secuencias correspondientes.

RMSD

La desviación del valor cuadrático medio (RMSD) es la medida de la distancia promedio entre los átomos (usualmente los del esqueleto proteico) de proteínas superpuestas. Esta métrica se utiliza comúnmente para medir la similitud entre estructuras tridimensionales de proteínas y se calcula como se muestra en la ecuación 2.25, donde a_i y b_i corresponden a los pares de átomos y N es el número total de átomos del sistema.

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N \|a_i - b_i\|^2} \quad (2.25)$$

Existen diversos programas que permiten comparar estructuras mediante el alineamiento estructural. En este trabajo de tesis, en particular se empleó la función *super* implementada en el programa PyMOL[64]. Este programa dispone de varias funciones para alinear estructuras, siendo las más comúnmente utilizadas *align* y *super*. La diferencia entre ambas radica en que la primera realiza un alineamiento de secuencias previo a la superposición de estructuras, lo que hace que falle cuando la identidad de secuencias es baja. *Super*, en cambio, realiza el alineamiento independientemente de la secuencia, es decir, solo se basa en la superposición estructural, por lo que es un método más robusto que el anterior para proteínas con baja identidad de secuencia.

2.2.3 Comparación de estructuras de ligandos

A los fines de poder comparar estructuras de ligandos computacionalmente, se debe contar en primer lugar con una representación adecuada de los mismos, dado que la representación tradicional de química orgánica (en dos dimensiones) que se utiliza para moléculas, si bien resulta ideal para el químico, no es válida para uso computacional ni especialmente para los algoritmos de búsqueda y almacenamiento informático. Por este motivo, es que en las bases de datos que almacenan información de estructura química, se utilizan otras representaciones para las moléculas pequeñas. Si bien existen varias de estas representaciones informáticamente amigables, una de las más extendidas es la denominada *SMILES* (acrónimo para *Simplified Molecular Input Line Entry System*) que es una notación lineal que permite representar moléculas y reacciones. A modo de ejemplo, el siguiente SMILES corresponde a la molécula de ácido acético: CC(=O)O. En la Figura 2.1 se ilustra el mecanismo por el cual se construye un SMILES para una molécula más compleja, como es el caso del antibiótico ciprofloxacina.

Los SMILES contienen la misma información que puede encontrarse en una tabla de conectividad, pero con la ventaja de ser mucho más compactos y más sencillos de manipular, dado que se trata de una construcción lingüística en vez de una estructura de datos computacional (como es el caso de las tablas de conectividades)[66]. Estas propiedades de los SMILES, entre otras, hacen que sean muy utilizados al manipular información química con herramientas computacionales.

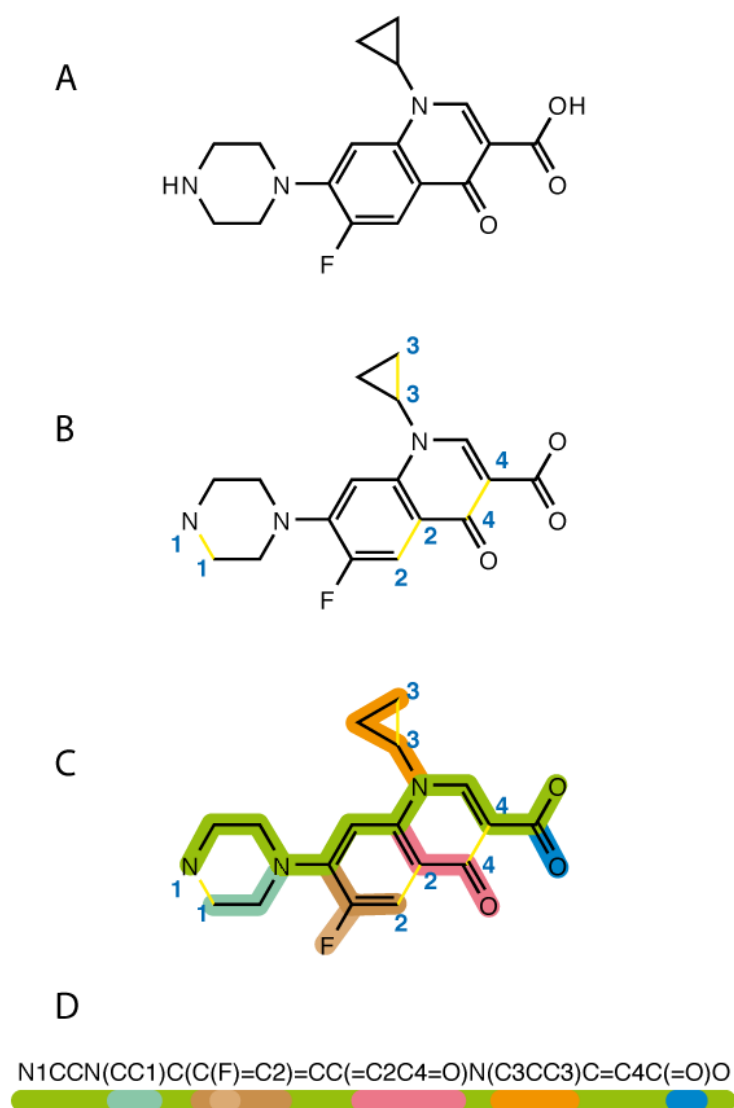


Figura 2.1: Ejemplo del mecanismo de generación de SMILES para la estructura del antibiótico ciprofloxacina. En primer lugar se abren los ciclos y luego se escriben como ramificaciones de un esqueleto principal.[65]

En segundo lugar, para poder cuantificar el grado de similitud entre dos moléculas, es necesario codificar esa representación de SMILES de modo de poder establecer comparaciones computacionalmente. Para este fin se utiliza lo que se denomina *fingerprints*, que consisten en una representación abstracta de las características estructurales de una molécula. Computacionalmente hablando, un *fingerprint* no es más que un arreglo booleano, es decir, un arreglo de ceros y unos, en el cual cada elemento es verdadero o falso. Los arreglos booleanos a su vez en general se representan como mapas de bits -arreglo de bytes o palabras en el cual cada bit representa una posición del arreglo booleano-. Como su nombre lo indica, una característica estructural es un mapa de bits en el cual cada bit representa la presencia o ausencia de un determinado patrón estructural.

En tercer lugar, para poder saber cuanto se parece una molécula a otra se requiere definir una medida de similitud. En tal sentido, la proporción de subestructuras en común entre las dos moléculas parecería ser una medida razonable de la similitud entre las mismas. En términos matemáticos, esto implica una comparación entre los bits que son "verdadero" en ambos *fingerprints*. Al describir las moléculas por la presencia o ausencia de características, los coeficientes de asociación binaria o medidas de similitud están basados en cuatro términos: *a*, *b*, *c*, *d*, los cuales se muestran en la tabla 2.1. En donde *a* es la cantidad de bits en el objeto A pero no en el objeto B, *b* es la cantidad de bits en el objeto B pero no en el objeto A, *c* es la cantidad de bits que son "verdadero" tanto en A como en B y *d* es la cantidad de bits que son "falso" tanto en A como en B.

		Objeto B	
		0	1
Objeto A	0	d	b
	1	a	c

Tabla 2.1: Términos que definen los coeficientes de asociación binaria para la comparación de *fingerprints*.

Existen diversos coeficientes para medir el grado de similitud entre dos moléculas [67]. Entre los más utilizados está el coeficiente de Tanimoto, el cual puede definirse como la proporción de bits "verdaderos" compartidos entre ambas moléculas [68, 69].

El coeficiente de Tanimoto, que fue el elegido para comparar estructuras de ligandos en este trabajo de tesis, se define entonces como el cociente:

$$\frac{c}{(a + b + c)} \quad (2.26)$$

Esta ecuación da como resultado para cada par de moléculas un número entre cero y uno, mientras más cercano a uno sea el valor, más parecidas son las moléculas.

A modo de ejemplo en la Figura 2.2 se muestran los valores del índice de Tanimoto que surgen de la comparación de una estructura química contra otras tres con diverso grado de similitud.

2.3 Modelado por homología

Debido a que la cantidad de secuencias disponibles crece a un ritmo mucho mayor que el número de estructuras, se hace necesaria la aplicación de métodos de modelado de la estructura proteica para compensar esta diferencia. El modelado por homología es el proceso por el cual una o más proteínas de estructura conocida, cuya secuencia es similar a la proteína de interés que carece de estructura conocida, se utilizan como moldes para modelar la estructura desconocida [70, 71, 72, 73]. En la actualidad, el modelado por homología (también llamado modelado comparativo) [74, 73, 75] es generalmente la manera más rápida de obtener un modelo aproximado de la estructura de una proteína cuando se conoce la estructura tridimensional de alguna proteína relacionada que pueda ser utilizada como molde.

El modelado por homología se basa en la observación de que las proteínas que tienen estructura primaria estrechamente relacionada (es decir, que han divergido de una proteína ancestral común durante la evolución) comparten segmentos de conformación similar. Se asume que si las secuencias aminoacídicas están cercana-

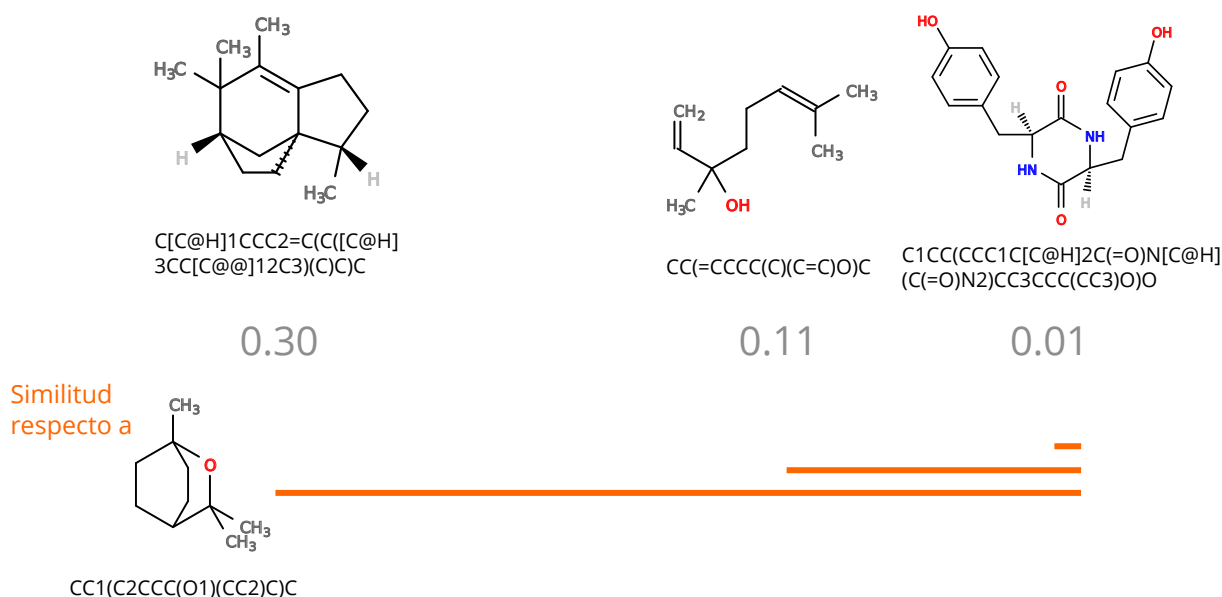


Figura 2.2: Ejemplos de estructuras químicas junto a sus correspondientes SMILES y el valor obtenido para el índice de Tanimoto entre cada par.

mente relacionadas, luego la estructura 3D de la proteína puede ser predicha partiendo de las estructuras 3D de otras proteínas que pertenezcan a la misma familia.

La metodología de modelado por homología puede dividirse en cuatro pasos principales, que se muestran en la Figura 2.3. A continuación se presenta una breve descripción de cada uno de ellos.

2.3.1 Selección de la estructura molde más adecuada

La calidad del modelado por homología es altamente dependiente de la elección de la estructura que se utilizará como molde. Como regla general, una estructura proteica puede proveer un modelo más o menos acertado para otra proteína si entre ellas comparten una identidad de secuencia superior al 50 % [9].

2.3.2 Alineamiento de las secuencias de la proteína a modelar y el molde

Es la etapa más importante y la más delicada, ya que la construcción del modelo se realizará conforme a este alineamiento. Existe una gran variedad de métodos de alineamiento de secuencias proteicas, y como se mencionó anteriormente, los AMS suelen ser más precisos que los de a pares. Por otro lado, cuando se contruye un alineamiento para modelar por homología, se tiende a ubicar los "gaps" fuera de las regiones de estructura secundaria de la proteína molde.

2.3.3 Generación de un modelo estructural basado en el alineamiento de secuencias

Existen varias aproximaciones para construir las coordenadas espaciales del modelo a partir del alineamiento realizado. El método original es el que se conoce como ensamble de cuerpo rígido. Otra familia de métodos, denominado modelado por coincidencia de segmentos, se basa en las posiciones aproximadas de los átomos conservados en los moldes. Y el tercer grupo de métodos, conocido como modelado por satisfacción de restricciones espaciales, utiliza un conjunto de criterios geométricos para crear una función de densidad de probabilidad para la localización de cada átomo en la proteína. La precisión asociada a los diversos métodos es similar cuando se los utiliza de manera óptima.

En este trabajo de tesis se utilizó el programa MODELLER [76], que es uno de los programas más ampliamente utilizados y pertenece al conjunto de métodos que implementa el modelado por homología por satisfacción de

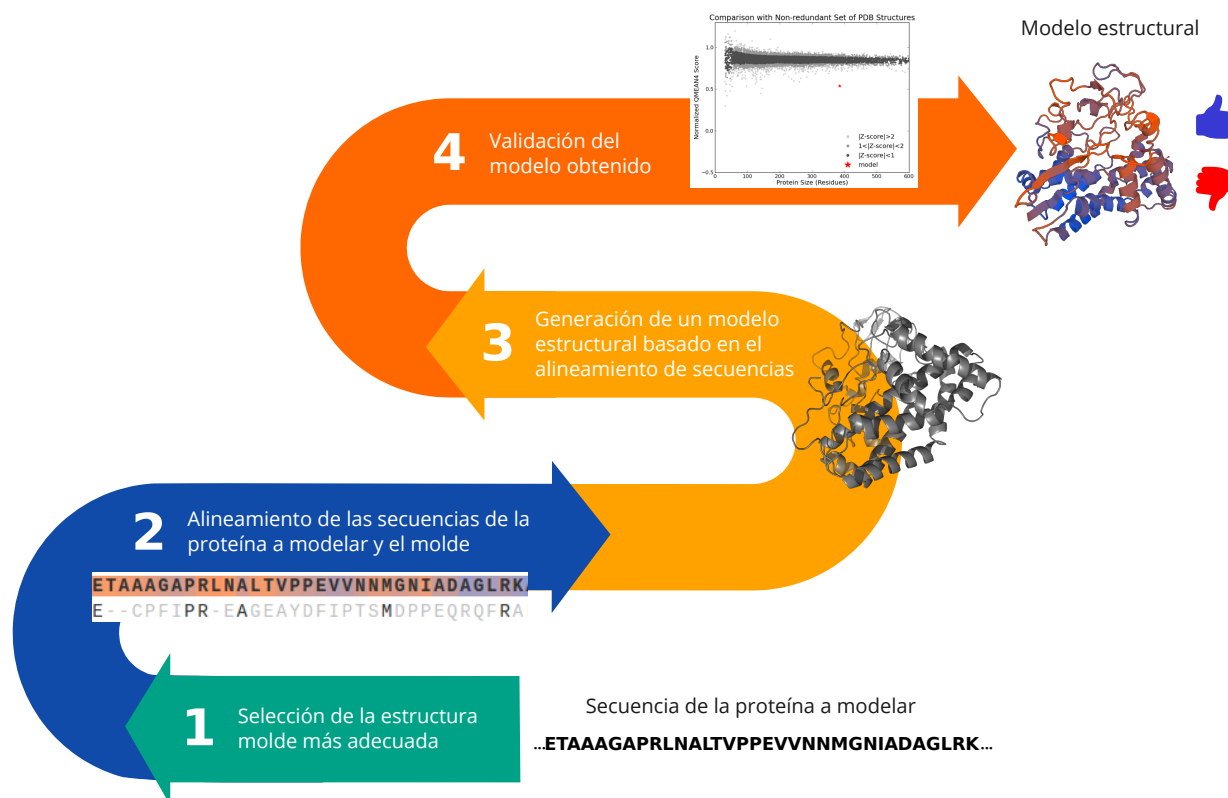


Figura 2.3: Etapas del proceso de modelado por homología

restricciones espaciales. Los métodos de esta clase generan un conjunto de restricciones en la estructura de la secuencia a modelar, basándose en su alineamiento con las estructuras molde. Estas restricciones en general se obtienen asumiendo que las correspondientes distancias y ángulos entre residuos alineados en las estructuras modelada y molde son similares. Estas restricciones derivadas de la homología típicamente se complementan con restricciones estereoquímicas sobre las longitudes de los enlaces, ángulos de enlace, ángulos diedros, y contactos de no-unión átomo-átomo obtenidas de un campo de fuerzas de mecánica molecular. El modelo se deriva luego minimizando las violaciones de todas las restricciones. Debido a que el modelado por satisfacción de restricciones espaciales permite utilizar muchos tipos diferentes de información sobre la secuencia a modelar, es quizás la más promisorio de todas las técnicas de modelado comparativo.

Modelado de loops

En cada dominio de plegamiento (usualmente asociado a una familia proteica), la variabilidad estructural es el resultado de sustituciones, inserciones y deleciones de residuos entre los distintos miembros de la familia. Tales cambios frecuentemente se corresponden con regiones de loops que conectan elementos de estructura secundaria en el plegado proteico. En consecuencia, los loops frecuentemente determinan la especificidad funcional del entorno proteico. Ellos son un elemento fundamental de los sitios activos y de unión, lo que hace que la precisión en el modelado de las regiones de loops sea un factor determinante de la usabilidad de los modelos comparativos en el estudio de las interacciones de las proteínas con sus ligandos. El modelado de los loops es un problema complejo, y puede ser visto como el análogo al problema del plegado proteico pero a menor escala. Se han descrito diversos enfoques para abordar este problema, entre los cuales se citan métodos *ab-initio*, búsqueda en bases de datos y procedimientos que combinan ambos. En esta tesis se empleó el protocolo implementado por defecto en el programa Modeller.

Modelado de las cadenas laterales de los aminoácidos

Análogamente que para el caso de los loops, las conformaciones de las cadenas laterales son predichas en

base a consideraciones energéticas o estéricas.

2.3.4 Validación del modelo obtenido

La información que se puede obtener del modelo depende de su calidad, de modo que es importante poder evaluarla. Para evaluar los modelos obtenidos por homología, se necesitan métodos que permitan testear la calidad de los mismos. Estos métodos deberían ser capaces de verificar la confiabilidad del modelo, es decir, distinguir un modelo apropiadamente plegado de uno inapropiadamente plegado y evaluar propiedades geométricas y estéricas de los modelos. Existen numerosos programas de testeo de calidad de los modelos con diferentes criterios. La mayoría de los métodos han sido desarrollados utilizando datos empíricos de proteínas globulares de estructura conocida. En esta tesis la calidad general de los modelos se testeó utilizando el método QMEAN [77, 78] y se trabajó con estructuras conocidas de modo de poder validar los resultados obtenidos.

2.4 Análisis del contexto génico

El análisis del contexto génico puede dar indicios de la función de una proteína desconocida debido a la forma en que están estructurados los genes bacterianos dentro del genoma. A los fines de reducir los recursos empleados al activar y desactivar la transcripción génica, los genes bacterianos frecuentemente están organizados en operones. Un único operón puede contener varios genes bajo la regulación transcripcional de un único promotor. Estos genes suelen estar relacionados de modo tal que sus productos génicos, en general enzimas, forman parte de una vía metabólica. Por ejemplo, el producto de la enzima A es luego el sustrato de la enzima B, la cual produce ahora otra molécula que es sustrato de la enzima C. Si las funciones de las enzimas A y C son conocidas, pero se desconoce la función de la enzima B, el hecho de saber que la enzima A y C se encuentran localizadas secuencialmente dentro del genoma da una pista acerca de la posible función de B: es muy probable que la enzima B lleve a cabo una reacción química que conecte los metabolitos producidos y consumidos por A y C respectivamente.

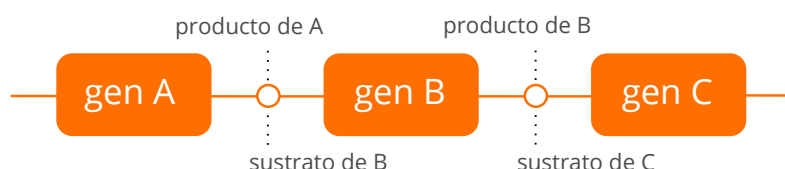


Figura 2.4: Fundamento para el uso de la información del contexto génico en la predicción de función de una proteína.

Partiendo de esta premisa, en este trabajo de tesis nos propusimos incluir el análisis de los compuestos generados y consumidos por los genes vecinos a los p450, a fines de encontrar posibles sustratos y/o productos de las enzimas en estudio.

2.4.1 Obtención de información del contexto génico

Para obtener la información de los genes adyacentes a los CYPs y sus correspondientes reactivos y productos, se diseñó y programó un pipeline bioinformático específico que conecta la información de los genes, enzimas, números EC, reacciones y compuestos disponible en la base de datos KEGG con los datos específicos de cada compuesto químico involucrado en las reacciones de interés (por ejemplo el SMILES), disponibles en la base de datos PubChem. El esquema de dicho pipeline se muestra en la Figura 2.5.

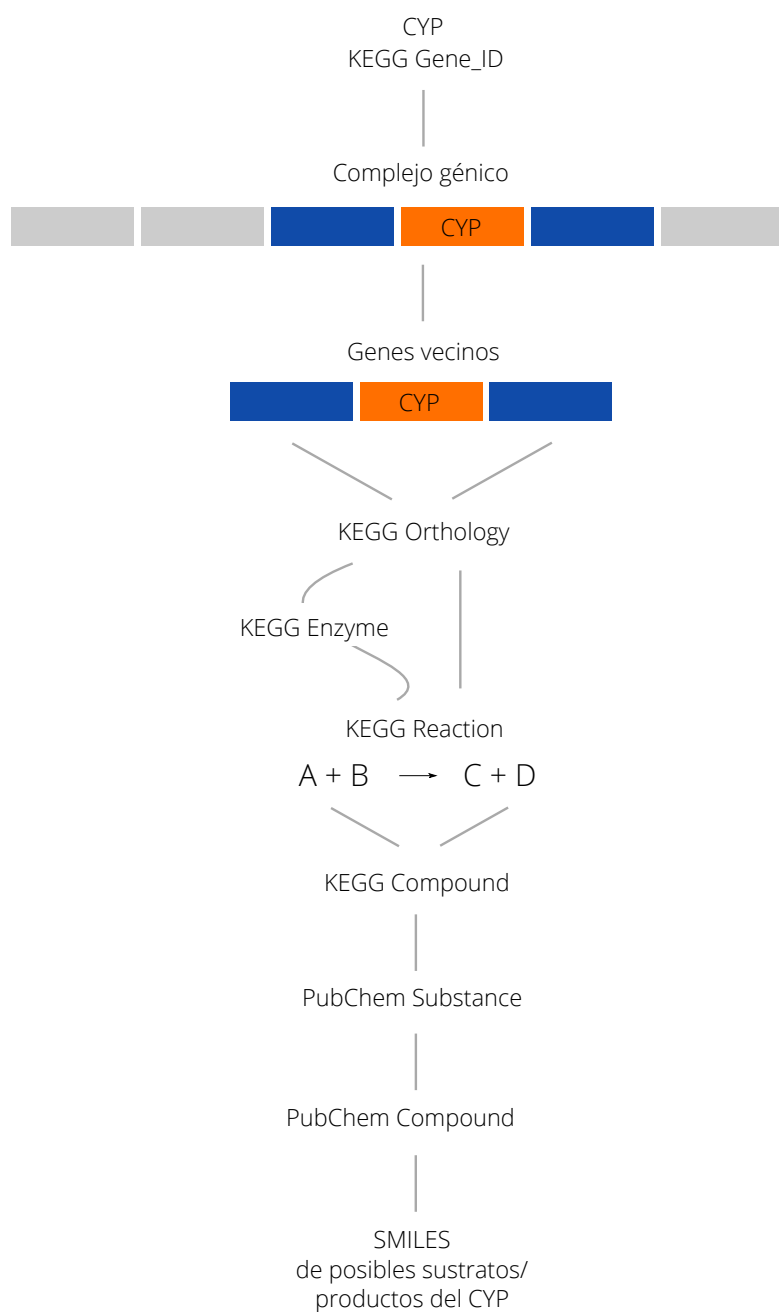


Figura 2.5: Esquema del flujo de datos implementado para la obtención de sustratos y productos candidatos por información del contexto génico.

2.5 Docking

Las estrategias de docking tienen como objetivo predecir la estructura del complejo proteína-ligando en base a la estructura del receptor y del ligando separados. Las técnicas comúnmente utilizadas para realizar docking automático se pueden agrupar principalmente en dos categorías: métodos de coincidencia geométrica y métodos de simulación de formación de complejos o de docking. Los primeros crean un modelo del sitio activo, donde típicamente se incluyen zonas de interacción puente de hidrógeno y zonas que son estéricamente accesibles y por lo tanto tratan de ubicar un determinado ligando en una determinada conformación dentro del modelo, considerándolo usualmente como un cuerpo rígido (por ejemplo DOCK[79], el cual es un programa muy eficiente para hacer un barrido rápido de bibliotecas de compuestos). La segunda clase de métodos de docking corresponde a aquellos que describen con mayor grado de detalle a la estructura del ligando y del receptor (por ej. el ligando se ubica de manera aleatoria fuera de la proteína, y luego explora traslaciones, orientaciones y conformaciones hasta que se encuentra el estado ideal del complejo proteína-ligando). Esta técnica es usualmente más lenta que los métodos de coincidencia geométrica, pero permite que la flexibilidad del ligando sea considerada y pueda ser utilizada con mayor grado de detalle para calcular la energía de unión del ligando en un sitio de unión plausible. En este punto, Autodock es uno de los programas más ampliamente utilizados que combina una rápida manera de evaluar la energía de interacción proteína-ligando, en base a una grilla de puntos que tiene precalculadas las energías de interacción de pares y que es usualmente utilizada durante toda la simulación con un algoritmo de búsqueda genético de tipo Lamarckiano[80].

Un algoritmo de docking se define por dos elementos fundamentales. El primero es un algoritmo de búsqueda, que permite a partir de una conformación inicial del ligando generar un conjunto de poses dentro del receptor que deben ser evaluadas. El segundo elemento es el que permite evaluar y clasificar dichas poses y es la denominada función de puntuación.

2.5.1 Algoritmo de búsqueda

El espacio de búsqueda en teoría consiste de todas las orientaciones y conformaciones posibles en que la proteína puede combinarse con el ligando. En la práctica, explorar exhaustivamente este espacio resulta imposible. Por esta razón es necesario utilizar heurísticas para poder aproximar una solución, inexacta pero lo suficientemente buena que permita comparar entre las distintas soluciones generadas por el mismo algoritmo. En esta tesis se utilizó el programa Autodock4, que emplea una heurística denominada algoritmos genéticos, cuyos fundamentos se explican a continuación.

Los algoritmos genéticos

Los algoritmos genéticos utilizan las ideas basadas en la genética y la biología evolutiva para explorar el espacio conformacional del ligando en el complejo. En el caso del docking molecular, el estado particular de una molécula se describe por un conjunto de valores que contienen información sobre la posición, la orientación y la conformación de un ligando con respecto a la proteína. Estos son los llamados estados variables del ligando y en los algoritmos genéticos cada estado variable es análogo a un gen. De este modo, el conjunto de estados del ligando conforman el genotipo, mientras que las coordenadas atómicas corresponden al fenotipo. En el docking molecular, el ajuste del ligando en una posición/conformación dada es la energía de interacción total del ligando con la proteína, la cual es evaluada utilizando la función de puntaje. De esta manera, la estrategia de búsqueda se puede plantear en base a un esquema híbrido de búsqueda global y luego local. Para abordar el primer punto se recurre a los algoritmos genéticos, los cuáles requieren generar una población inicial de ligandos con diferentes coordenadas (fenotipos), por lo general entre 150 a 300 individuos, dependiendo del tamaño del espacio de búsqueda, que van generando descendencia. Para ello, es relevante poder generar variabilidad entre las sucesivas generaciones y por lo tanto se imita un conocido proceso presente en la meiosis que es el llamado *crossing over* o intercambio cruzado, en el cual para el caso del docking molecular, se combinan pares aleatorios de individuos de una población entre los cuales se intercambian genes para dar lugar a un nuevo individuo, es decir, un ligando

con un dado estado posicional y conformacional, que conlleva una mezcla aleatoria de algunos genes parentales. Además, algunos individuos de la descendencia son sometidos a mutaciones aleatorias en las cuales un gen cambia su estado por un valor aleatorio (es decir, que en la práctica, en los estados hijos el ligando presenta conformaciones y posiciones que difieren de los parentales en algún aspecto). Por último, la selección de la descendencia de la siguiente generación se hace en base a la función de energía, es decir, que solo sobrevive aquel individuo que se encuentra en el entorno más favorable (en otras palabras, aquel que tiene el valor de energía de interacción proteína-ligando más negativo). Con respecto a la búsqueda local, se realiza una minimización de energía utilizando el algoritmo de Solis and Wet[81], el cual tiene la ventaja de no requerir la información del gradiente de energía provisto a través de la superficie de energía potencial, lo que facilita la búsqueda en el espacio torsional. Este método híbrido, que combina un algoritmo genético con un método adaptativo de búsqueda local es conocido como Algoritmo Genético Lamarkiano con búsqueda local y que a diferencia de otros métodos de exploración que han sido ampliamente utilizados como Monte Carlo o *Simulated Annealing*, permite abordar problemas en los se utilizan un número mayor de grados de libertad. Además, para hacer las minimizaciones, es posible utilizar el mismo campo de fuerzas que se utiliza para hacer el cálculo de la energía de interacción proteína-ligando.

2.5.2 Función de puntuación

La función de puntuación es la herramienta que permite comparar y clasificar rápidamente las distintas soluciones propuestas por el algoritmo de docking. Estas funciones de puntuación nacen de la necesidad de poder clasificar una serie de poses de una molécula para poder elegir aquellas que mejor representen el modo de unión de la molécula X con el receptor Y. Otro rol que cumplen las funciones de puntuación es permitir comparar entre distintas moléculas para saber cual es potencialmente el mejor compuesto que se une a un receptor dado.

La función de puntuación es en sí un estimador del ΔG_U , entonces la estrategia de búsqueda a utilizar lo que debe hacer es minimizar el valor de dicha función. A continuación se describe la función de puntuación utilizada en la presente tesis para estimar la energía de la pose del ligando, que es la implementada en Autodock.

La función de puntuación implementada en Autodock

Para modelar la energía libre de unión proteína-ligando suele utilizarse una expansión de la conocida expresión de energía libre de Gibbs agregando términos que compensan o estiman los efectos entrópicos y del solvente, como se muestra en la ecuación 2.27

$$\Delta G = \Delta G_{VDW} + \Delta G_{hbond} + \Delta G_{elec} + \Delta G_{conf} + \Delta G_{tor} + \Delta G_{sol} \quad (2.27)$$

en donde los 5 términos del campo de fuerza incluyen dispersión-repulsión, unión hidrógeno, electrostático y desviación de la geometría óptima en las uniones covalentes (términos de unión), respectivamente. Por su parte, ΔG_{tor} modela la restricción de las rotaciones internas, rotaciones globales y traslaciones. Por último, ΔG_{sol} modela la desolvatación luego de la unión del ligando y tiene en cuenta el efecto hidrofóbico (son los cambios entrópicos del solvente en la interfase proteína-ligando). En particular, Autodock implementa una aproximación similar utilizando el ciclo termodinámico de Wesson y Eisenberg[82]. La función incluye cinco términos (como puede verse en la ecuación 2.28)

$$\begin{aligned} \Delta G = \Delta G_{VDW} \sum_{i,j}^n \left(\frac{A_{ij}}{r_{ij}^{12}} + \frac{B_{ij}}{r_{ij}^6} \right) + \Delta G_{hbond} \sum_{i,j}^n E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} + \frac{D_{ij}}{r_{ij}^{10}} \right) + \\ \Delta G_{elec} \sum_{i,j}^n \frac{q_i q_j}{\epsilon_{ij} r_{ij}} + \Delta G_{tor} N_{tor} + \Delta G_{sol} \sum_{i,j}^n (S_i V_j + S_j V_i) e^{\frac{r_{ij}^2}{2\sigma^2}} \end{aligned} \quad (2.28)$$

en donde los cinco términos de ΔG de la derecha (ΔG_{VDW} , ΔG_{hbond} , ΔG_{elec} , ΔG_{tor} y ΔG_{sol}) son coeficientes que se ajustan contra datos experimentales de un conjunto de complejos cuyas constantes de disociación son

conocidas. Las sumatorias se llevan a cabo sobre todos los pares de átomos del ligando, i , y de la proteína, j , además de todos los pares de átomos del ligando que están separados por tres o más enlaces. Los tres primeros términos incluyen las contribuciones en vacío: Un término dispersivo 12-6 de Lennard-Jones, un término direccional 12-10 que incluye la interacción de puente hidrógeno, donde $E(t)$ es un término direccional que pesa el ángulo t entre el átomo de prueba y el átomo blanco, y por último un término que tiene en cuenta el potencial electrostático (el cuál utiliza cargas atómicas parciales sobre cada átomo estimadas usualmente con el método de Kollman o Gasteiger [83, 84]). El cuarto término es el que incluye la contribución a la energía libre dada por la desolvatación, el cual está basado en el método de Stouten et al.[85], que presenta la ventaja de funcionar correctamente con un modelo de cálculo de energía basado en el precálculo de una grilla de afinidad, como lo hace Autodock. En este método, para cada átomo en el ligando, un volumen parcial de los átomos de la proteína que lo rodean son pesados por una función exponencial y luego sumados, evaluando el porcentaje del volumen alrededor de los átomos del ligando que es ocupado por los átomos de la proteína. Este porcentaje es luego pesado por los parámetros de solvatación de los átomos del ligando para dar la energía de desolvatación. Como una medida de la penalidad entrópica dada por la restricción roto-traslacional que se produce al unirse el ligando al receptor, se incluye un término que es proporcional a N_{tor} (que es el número de enlaces en el ligando que pueden ser descritos por una hibridación de tipo sp_3). Este término es constante durante todo el cálculo de docking ya que depende principalmente de los grados de libertad rotacionales del ligando que fueran previamente definidos por el usuario (por lo general no más de 12). Finalmente, utilizando un método de interpolación trilineal se evalúa rápidamente cada término energético para cada tipo de átomo del ligando en una dada posición, utilizando los mapas de energía que fueran previamente calculados sobre el entorno proteico. El empleo de grillas precalculadas es clave, dado que permite reducir drásticamente el costo computacional que implica la evaluación energética de las conformaciones.

Para la construcción de los mapas de energía, Autodock genera una grilla tridimensional en torno a las coordenadas de los residuos de la proteína. Un átomo de prueba visita cada punto de la grilla, se calcula la energía de interacción y la guarda. De esta manera, el mapa de energía de cada tipo de átomo del ligando se guarda en tablas separadas, incluyendo las energías de dispersión/repulsión y de formación de puentes de hidrógeno. Además se calcula una grilla de interacción electrostática, utilizando una partícula de prueba con una carga positiva y una grilla que incluye las energías de desolvatación.

2.5.3 El algoritmo de docking utilizado en esta tesis

El procedimiento para hacer un cálculo de docking tal como se plantea en el esquema de AutoDock es el siguiente:

1. Construcción de grillas en la cavidad proteica. Se crea un mapa por cada tipo de átomo presente en el ligando (C, N, O, S, H) con información del campo electrostático y las interacciones no polares.
2. Definición de diedros rotables en el ligando.
3. Generación de la conformación del ligando en el sitio activo por medio, en este caso, de un algoritmo genético Lamarckiano.
4. Evaluación de la conformación (estimación de la energía libre de unión).
5. Vuelta al paso 3 hasta alcanzar el número de iteraciones predefinido.

EL MECANISMO DE REACCIÓN INUSUAL DE *Mt*CYP121

3.1 Introducción

La tuberculosis ha sido desde tiempos inmemoriales un problema de salud importante, especialmente en los países en vías de desarrollo, causando más de dos millones de muertes por año según reportes de la Organización Mundial de la Salud. Las principales causas de este impacto están relacionadas a la necesidad de tratamiento prolongado y a veces incompleto con las drogas de primera línea actuales (isoniazida y rifampicina), sumado a la emergencia de nuevas cepas resistentes a tales drogas y a la interacción negativa de las mismas con las suministradas a pacientes HIV positivos. Este escenario ha creado una necesidad urgente de nuevas drogas antituberculínicas [34]. Un gran progreso en nuestro conocimiento sobre *Mycobacterium Tuberculosis* (*Mt*) deriva de la secuenciación de su genoma completo, la cual entre otros hallazgos inesperados, reveló la presencia de 20 genes de CYPs diferentes [86, 87].

Entre los 20 CYPs de *Mycobacterium tuberculosis*, CYP121 ha despertado mayor interés, no solo debido a su esencialidad para la viabilidad del bacilo, sino porque se ha demostrado recientemente que cataliza una reacción de acoplamiento carbono-carbono inusual para una enzima de tipo CYP. *Mt*CYP121 convierte su sustrato, ciclo-ditirosina (cYY) a un producto llamado micociclosina (mccyc) (ver Figura 1.4-A), donde los dos anillos aromáticos de tirosina están unidos covalentemente a través de un enlace C-C en posición orto respecto al grupo OH de cada tirosina [88]. La estructura de *Mt*CYP121 ha sido resuelta tanto en ausencia [89] como en presencia de su sustrato [88], así como también en complejo con diferentes compuestos azólicos [89, 90, 91], y además se obtuvieron varias mutantes de un único residuo [92].

La enzima ha sido ampliamente caracterizada en términos bioquímicos a través de métodos espectroscópicos y mediciones cinéticas [93]. La estructura de rayos X de la enzima unida a cYY (ver Figura 3.1), muestra que una de las tirosinas del sustrato se ubica más cerca del hemo (razón por la cual la llamaremos tirosina cercana), con su grupo OH interactuando con dos moléculas de agua estructurales (W1 y W2 en la Figura 3.1), una de las cuales se encuentra también coordinada al hierro férrico; mientras que la otra tirosina (a la que llamaremos tirosina lejana) se encuentra apuntando hacia la superficie proteica, es decir, a la entrada del sitio de unión del sustrato.

Adyacente a la cYY, existen dos bolsillos llenos de moléculas de agua, lo cual sugiere que el sitio activo es fácilmente accesible para el solvente. Un dato interesante es que la comparación con la estructura libre no evidencia que ocurran cambios conformacionales importantes luego de la unión de cYY (el RMSD de los carbonos alfa entre ambas estructuras es 0.16 Å)[88]. El átomo de hierro está coordinado al agua W1, la cual también muestra interacciones de puente de hidrógeno con Ser237. Este residuo posiblemente reemplaza al residuo de treonina que se encuentra conservado en otras p450s, el cual se ha demostrado que está involucrado en la unión de oxígeno y actúa como donante de protón al superóxido unido al hierro [94, 95]. Una característica única de *Mt*CYP121 es la presencia de un residuo de arginina (Arg386) cerca del hierro, la cual se ha descrito que afecta fuertemente

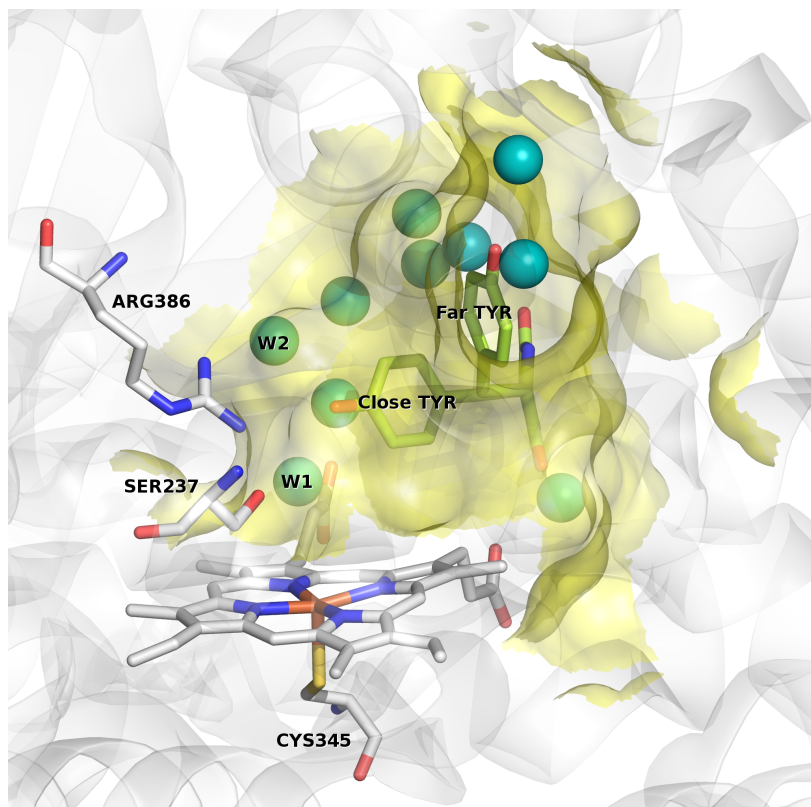


Figura 3.1: Estructura de la *MtCYP121* férrica unida a cYY. El sustrato cYY (sobre el cual se marcan la tirosina cercana y lejana), el grupo hemo, los residuos Cys345 (el ligando proximal), Arg386 y Ser237 se muestran representados con bastones. Las moléculas de agua del sitio activo se muestran como esferas en cian. El sitio de unión al sustrato se muestra como isosuperficie en amarillo.

el potencial redox del hemo [92, 89]. Por último, y no por ello menos importante, un trabajo reciente analizó la reactividad de *MtCYP121* hacia diferentes análogos de sustrato, mostrando la importancia tanto del anillo de dicetopiperazina como del correcto posicionamiento del hidroxilo de la tirosina para la conversión eficiente de cYY [96]. **Es importante notar en este punto, que aunque se ha aprendido mucho, el mecanismo preciso y las razones estructurales que permiten a *MtCYP121* realizar esta reacción inusual aún permanecen ampliamente desconocidos.**

Mecanismo propuesto

La propuesta más común para este mecanismo de reacción de acoplamiento C-C implica la generación consecutiva de dos radicales en sendos anillos aromáticos, que subsecuentemente se recombinan estableciendo el nuevo enlace C-C. Son claves en esta propuesta la naturaleza de las especies que generan el primer y segundo radical y el rol desempeñado por la matriz proteica en su generación y en el paso final de recombinación, como se muestra en la Figura 1.4-B. Como se mencionó anteriormente, en *MtCYP121*, cYY está unido de modo que un anillo de tirosina está cerca del hierro hémico (la tirosina cercana), mientras que la otra está lejos (la tirosina lejana) (como se muestra en la Figura 3.1). En el presente caso, para que la reacción de acoplamiento tirosina-tirosina pueda ocurrir, deben producirse dos radicales (uno en cada tirosina), que luego se combinarán para dar origen al producto cíclico. La formación del primer radical se propone que ocurre en la tirosina cercana, debido a la transferencia de un átomo de hidrógeno al Cp I [Pasos de reacción a-b-c en la Figura 1.4-B]. La formación del segundo radical debería ocurrir en la tirosina lejana, y para ello se han propuesto varios mecanismos. El primero asume que la cicloditirosina monoradicalaria (cY*Y) puede salir del sitio activo hacia la solución y reingresar en la orientación inversa [Paso de reacción d' en la Figura 1.4-B], produciéndose el segundo radical en un nuevo ciclo de la enzima (es decir, a través del Cp I) o en el mismo ciclo utilizando el Cp II como oxidante. Otra posibilidad interesante es que ocurra una transferencia electrónica interna (IET) desde la tirosina lejana hacia la cercana, resultando de este modo en que el primer radical se localice en la tirosina lejana [Paso de reacción d en la Figura

1.4-B], permitiendo así la formación del segundo radical por medio del hemo, otra vez utilizando el Cp I o el Cp II como oxidante. Finalmente el acoplamiento C-C podría ocurrir con o sin ayuda de la matriz proteica [Pasos de reacción f y f' en la Figura 1.4-B]. Poder discernir entre tales posibilidades y comprender el rol desempeñado por la matriz y el sitio activo proteico en la catálisis de la reacción es un asunto muy importante no solo debido a la relevancia de *MtCYP121* como potencial blanco para drogas sino también como aporte a la comprensión general de los mecanismos catalíticos de los CYPs. En este contexto hemos decidido abordar este problema utilizando métodos computacionales, específicamente mecánica cuántica (QM), dinámica molecular (MD) y métodos híbridos QM/MM.

3.2 Resultados

Dada la complejidad de la reacción realizada por esta proteína, hemos examinado separadamente los siguientes tres pasos: (i) Generación del primer TYR* o cY*Y, (ii) Generación del segundo TYR* o cY*Y*, y (iii) Formación del enlace C-C o mcic, como se describe en la Figura 1.4-B. Asimismo los resultados se presentan separadamente para cada paso y luego son discutidos en el contexto de mecanismo completo de la reacción.

3.2.1 Generación del primer radical

Simulación por DM de cYY unido a *MtCYP121*

El oxidante más potente y común en los citocromos de tipo p450 es, como ya se mencionó, el Cp I. Una vez generado en el ciclo de reacción de la enzima, el Cp I es capaz de abstraer un átomo de hidrógeno e insertar un átomo de oxígeno en una amplia variedad de sustratos. La forma más común de determinar el producto inicial de la oxidación en un CYP dado es por lo tanto, observar los átomos del sustrato que están cerca del hierro [97]. Para determinar los posibles mecanismos para la generación del primer TYR*, partiendo de la estructura del cristal de *MtCYP121* unido al sustrato, hemos realizado 30 ns. de simulación de DM de cYY unido a *MtCYP121*, con el hemo en distintos estados (oxi, Cp I y Cp II), y posteriormente analizamos los correspondientes parámetros geométricos como se muestra en la Tabla 3.1. El Cp II y el hemo-oxi fueron analizados no solo con fines comparativos, sino además por el hecho de que se ha mostrado que el Cp II es capaz de llevar a cabo reacciones oxidativas, en tanto que al complejo oxi se le atribuye la capacidad de abstraer protones de los alrededores.

Parámetro ^a	Estado hémico de la proteína		
	Oxi	Cp I	Cp II
FeO1-TyrO _{OH}	3.61 ± 0.47	5.19 ± 0.29	5.31 ± 0.28
FeO1-TyrH _{OH}	2.83 ± 0.68	4.31 ± 0.28	4.43 ± 0.28
FeO1-TyrH _{CA}	2.68 ± 0.35	4.96 ± 0.40	5.12 ± 0.32
FeO2-TyrH _{OH}	2.21 ± 0.66	–	–
FeO2-TyrH _{CA}	2.86 ± 0.31	–	–
FeO1-H _{OH} -TyrO _{OH}	169°	158°	166°

Tabla 3.1: Interacciones entre el ligando hémico y la tirosina cercana. Los valores reportados fueron tomados de las respectivas DM y corresponden a distancias promedio (Å) ± DE

^aFeO1 corresponde al átomo de oxígeno unido al hierro, FeO2 corresponde al átomo de oxígeno en la molécula de O₂ que no está unido al hierro, TyrO_{OH} corresponde al átomo de oxígeno fenólico de la tirosina, TyrH_{OH} corresponde al átomo de hidrógeno fenólico de la tirosina y TyrH_{CA} corresponde al átomo de hidrógeno aromático en posición orto de la tirosina.

Los resultados evidencian que para el Cp I, TyrO no establece una interacción de puente de hidrógeno permanente con el oxígeno del ferrilo dada la presencia de varias (alrededor de cinco) moléculas de agua en el sitio activo compitiendo como aceptores de puente de hidrógeno, y en general una molécula de agua puentea la interacción entre el TyrOH y el oxígeno del ferrilo. Sin embargo, sí puede establecerse una interacción de puente

de hidrógeno transitoria entre el grupo fenol y el Cp I (la distancia más cercana FeO1-TyrH_{OH} es 1.57 Å) debido a la alta movilidad de la tirosina cercana y las moléculas de agua circundantes. Algo similar se observa para el Cp II. En el caso del complejo oxi, por otro lado, se establece una interacción de puente de hidrógeno permanente entre TyrOH y el ligando unido, permaneciendo así el hidrógeno del fenol cerca de ambos átomos de oxígeno. En todos los casos, el hidrógeno fenólico está más cerca del átomo de oxígeno reactivo unido al hemo que el hidrógeno unido al carbono orto de la tirosina. Por consiguiente, pareciera que la matriz proteica no ubica permanentemente al grupo fenólico de la tirosina a distancia reactiva de los átomos de oxígeno del Cp I y Cp II, pero sí permite que se establezca una interacción de puente de hidrógeno fuerte con el oxígeno molecular unido.

Abstracción del hidrógeno por parte del Cp I

Partiendo de una estructura representativa seleccionada de las simulaciones de DM de cYY unido a MtCYP121 en estado de Cp I, donde existe un puente de hidrógeno entre el oxígeno del ferrilo y el TyrH_{OH}, analizamos la reacción de abstracción del hidrógeno utilizando simulaciones QM/MM. El hidrógeno fenólico es más ácido y cercano al hemo que cualquiera de los hidrógenos aromáticos, razón por la cual solo consideramos la abstracción de ese hidrógeno. Finalmente, y debido a la conocida reactividad de multiestados [19] del Cp I analizamos ambos estados de espín, el doblete (D) y el cuarteto (C). El perfil de energía correspondiente para la abstracción del átomo de hidrógeno resultó ser como se muestra en la Figura 3.2, mientras que los parámetros estructurales y de estructura electrónica seleccionados se muestran en la Tabla 3.2. Las barreras energéticas obtenidas para los estados de espín D y C son similares y pequeños, <6 kcal/mol. El estado C corresponde al estado basal a lo largo del rango explorado de la coordenada de reacción. La reacción resulta en la formación de Tyr* y un hemo de tipo Cp II, como se evidencia del análisis de las poblaciones de Mulliken y espín, que muestran una tirosina casi neutra con un electrón desapareado (el espín total de la tirosina es cercano a la unidad). El Cp II muestra, como era esperado, un estado basal de triplete con dos electrones desapareados y una mezcla de caracteres [P-Fe^{III}-OH* <=> P-Fe^{IV}-OH⁻] (evidenciado por la carga y espín del grupo OH resultante). De este modo confirmamos que el Cp I es capaz de abstraer el primer hidrógeno de la tirosina cercana en cYY y con una barrera de activación baja.

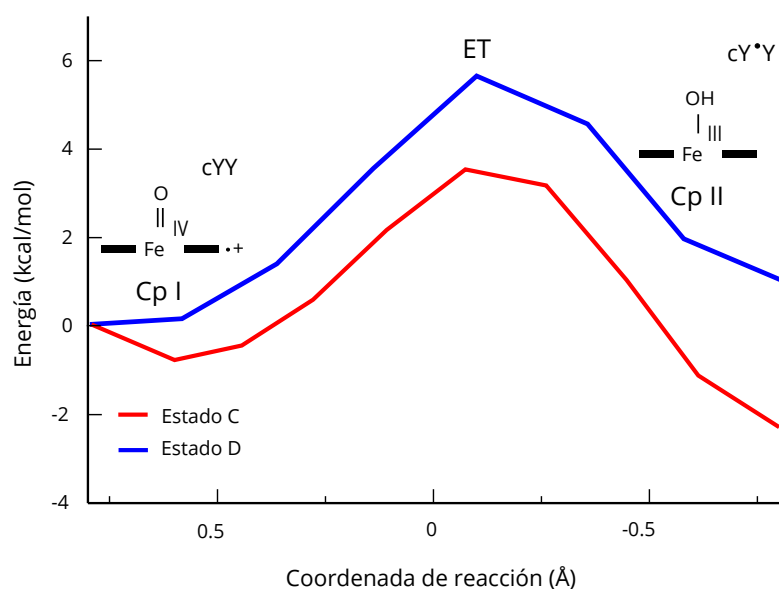


Figura 3.2: Perfil de energía para la abstracción de un átomo de hidrógeno y formación de Tyr* utilizando el Cp I como oxidante. Las líneas roja y azul corresponden a los estados de espín cuarteto y doblete respectivamente.

Especies oxigenadas alternativas como oxidantes

Como se describió en la introducción, durante el ciclo de reacción del CYP y el oxígeno, el dioxígeno se une en principio al hierro ferroso, y luego se reduce para formar las especies de hierro-superóxido (Fe^{III} -> O₂⁻). A fines comparativos y como una posible alternativa, evaluamos también si las especies de hierro-superóxido e hierro-dioxígeno son capaces de abstraer un hidrógeno de la tirosina cercana. Utilizamos el mismo protocolo y

Tabla 3.2: Distancias seleccionadas (d), poblaciones de Mulliken (q) y espín (s) relevantes para el paso de formación del primer radical en la tirosina cercana (cY*Y) en *MtCYP121*

Parámetro	Sistema					
	Cp I		ET		Cp II	
	D	C	D	C	D	C
Δ de espín	1.0	0	n.c	n.c	3.0	0
d Fe-O	1.68	1.69	1.74	1.74	1.83	1.83
d FeO-TyrH _{OH}	1.79	1.65	1.16	1.18	1.00	1.00
d TyrO-TyrH _{OH}	1.00	1.01	1.26	1.25	1.99	1.97
q Tyr cercana	-0.012	0.205	-0.048	-0.097	-0.075	-0.070
q OX(FeOX)	-0.337	-0.348	-0.284	-0.241	-0.189	-0.190
q TyrH _{OH}	-0.125	-0.133	-0.154	-0.172	-0.115	-0.115
q Hemo-Cys	0.131	0.136	0.329	0.330	0.263	0.262
s Tyr cercana	-0.118	0.164	-0.234	0.475	-0.981	0.990
s OX(FeOX)	0.697	0.762	0.656	0.605	0.261	0.270
s TyrH _{OH}	-0.001	-0.003	0.000	-0.006	0.005	-0.001
s Hemo-Cys	0.424	2.075	0.581	1.920	1.982	1.744

coordenada de reacción descritos anteriormente, y encontramos que el complejo oxi es incapaz de abstraer un protón/hidrógeno de la tirosina cercana. Sin embargo, el perfil energético para la reacción de transferencia con la especie de hierro superóxido, presentado en la Figura 3.3, sugiere que en este caso sí puede transferirse un protón/hidrógeno con una barrera muy pequeña. Las poblaciones de Mulliken y espín resultantes, sin embargo, muestran que la reacción corresponde a un transferencia de un protón, dado que el producto de Tyr se queda con una carga neta negativa cercana a la unidad. En consecuencia, la especie de hierro-superóxido es incapaz de generar el primer radical de Tyr, aunque la tirosina cercana puede participar como fuente de protones en los primeros pasos del ciclo de reacción del CYP para producir Cp I.

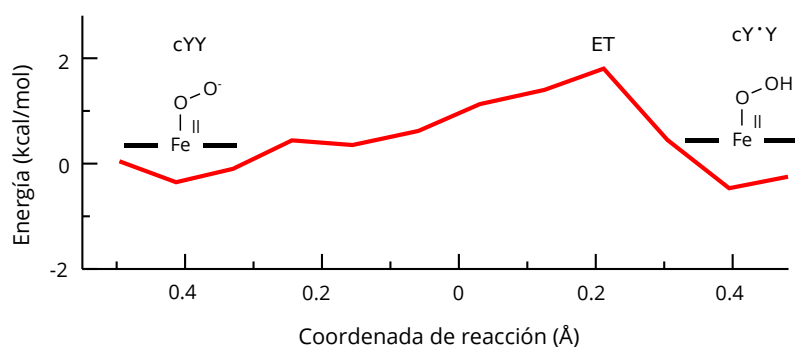


Figura 3.3: Perfil de energía de la reacción de transferencia de protón/hidrógeno de cYY a Fe^{II}-O₂⁻ en estado de espín cuarteto.

Evaluación del paso de re-unión del OH como reacción competitiva

El primer paso en las reacciones de hidroxilación comúnmente observadas en CYPs, en general consiste al igual que en este caso, en la generación de un radical alquilo y el Cp II. Este estado, generalmente procede a través del llamado paso de re-unión, que consiste en la adición del ligando radicalario OH unido al Cp II al sustrato radicalario, resultando en el producto hidroxilado deseado. En *MtCYP121* el mecanismo de re-unión llevaría a la formación de un catecol (orto dihidroxibenceno) en la tirosina cercana, en un ciclo catalítico completo. Para analizar la probabilidad de ocurrencia de esta reacción competitiva, calculamos el perfil energético para el paso de re-unión en ambos estados de espín, el doblete y el cuarteto. Los resultados presentados en la Figura 3.4 muestran que el paso de re-unión es exergónico, pero posee una barrera moderadamente alta (16-20 kcal/mol)

en ambos estados de espín estudiados. Por ello, a fines de evitar esta reacción indeseada, la generación del segundo radical debería tener una barrera más baja comparada con la reportada para el paso de re-unión.

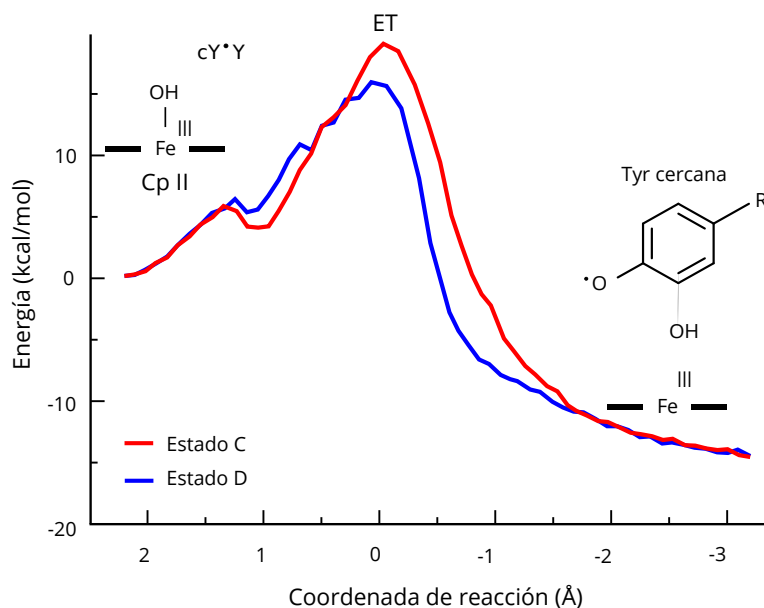


Figura 3.4: Perfil de energía para el paso de re-unión de OH. Las líneas roja y azul corresponden a los estados de espín cuarteto y doblete respectivamente.

3.2.2 Generación del segundo radical

Abstracción de átomos de hidrógeno por parte del Cp II

La formación del segundo radical de tirosina en *MtCYP121* es, como se describió anteriormente, el paso menos entendido y existen en torno a él varias posibilidades mecanísticas [ver pasos de reacción d, d' y e en la Figura 1.4-B]. La propuesta más simple que el intermediario monorradicalario cYY*, donde el radical se encuentra en la tirosina cercana, podría abandonar el sitio activo de la proteína y reingresar al mismo en la orientación inversa, quedando así el radical en la tirosina lejana y la tirosina cercana neutra. Esto permitiría que el segundo radical se genere otra vez en la tirosina cercana, lo cual podría ocurrir en un nuevo ciclo de la enzima (es decir, a través del Cp I) o en el mismo ciclo utilizando Cp II como oxidante. Anteriormente hemos mostrado que el Cp I puede realizar la abstracción de hidrógeno correspondiente, entonces analizaremos ahora la probabilidad de que el Cp II genere una tirosina radicalaria, tanto en el estado de espín S como en el T. A tal fin, seleccionamos una foto de la simulación de cYY con Cp II y realizamos las correspondientes optimizaciones y perfiles de energía de reacción QM/MM. Los resultados se muestran en la Figura 3.5 y en la Tabla 3.3. En el Cp II, el estado basal es el estado T, cuya curva yace alrededor de 3.5 kcal/mol por debajo del estado S. El mismo muestra una barrera de 3.5 kcal/mol para la abstracción del hidrógeno y la reacción es endergónica en una magnitud similar [ver Figura 3.5 - línea roja]. En el estado S la barrera es muy pequeña (alrededor de 1.5 kcal/mol) y también es sutilmente endergónica. El paso final para la reacción incluye la liberación de una molécula de agua luego de la transición de espín del hemo férrico resultante. Una vez más, el análisis de las poblaciones de Mulliken y espín muestra que el producto corresponde a una tirosina radicalaria. Los valores de carga y espín de la porción hémica resultante se corresponden con los de un hemo férrico unido a una molécula de agua. Esto permite concluir que, aunque con algún costo energético menor, el Cp II también es capaz de producir un radical en la tirosina cercana para formar cYY* en *MtCYP121*.

Transferencia electrónica interna Tyr-Tyr

Dado que tanto el Cp I como el Cp II son capaces de formar un radical en la tirosina cercana en *MtCYP121*, en principio ambos radicales podrían producirse en un único ciclo de oxidación del hemo. Sin embargo, para que

Parámetro	Sistema						
	Cp II		ET		Fe3-H ₂ O		
	S	T	S	T	S	T	C
Δ de espín	3.5	0	n.c	n.c	1.0	0	-
d Fe-O	1.93	1.91	2.00	2.02	2.01	2.02	2.32
d FeO-TyrH _{OH}	1.55	1.62	1.20	1.08	1.08	1.08	1.03
d TyrH _{OH} -TyrO	1.04	1.02	1.24	1.40	1.40	1.40	1.56
q Tyr	-0.041	-0.019	0.092	-0.146	-0.120	-0.144	-0.102
q L ^a	-0.393	-0.410	n.c	n.c	-0.093	-0.073	-0.003
q Hemo-Cys	0.435	0.414	0.299	0.262	0.213	0.299	0.183
s Tyr	-0.143	0.125	0.454	0.508	0.610	0.568	0.764
s L	0.049	0.088	n.c	n.c	-0.013	0.019	0.029
s Hemo-Cys	0.095	1.792	-0.435	1.472	-0.599	1.415	3.209

^aL simboliza al ligando unido al hemo, es decir, el OH en el Cp II y la molécula de agua en el producto.

Tabla 3.3: Distancias seleccionadas (d), poblaciones de Mulliken (q) y espín (s) relevantes para el paso de formación del segundo radical de tirosina en *MtCYP121*.

esto ocurra de manera eficiente, debe darse una reorganización del producto del primer radical, lo que implica ya sea la liberación y posterior reingreso del intermediario cYY*, o bien la migración de un electrón desde la tirosina lejana hacia la cercana en una reacción de transferencia electrónica intramolecular (IET, por sus siglas en inglés) [pasos de reacción d' y d en la Figura 1.4-B]. Dadas las potenciales complicaciones ya citadas que implicaría un mecanismo que involucre la liberación de cY*Y, pensamos en un posible mecanismo por el cual se genere el segundo radical pero reteniendo al intermediario cY*Y dentro de *MtCYP121*. Nuestra propuesta es que esto podría ocurrir a través de un proceso de IET, cuyo resultado sea una reorganización electrónica que establezca un estado con la densidad de espín localizada en la tirosina lejana y la densidad de carga negativa en la cercana o directamente en el grupo hemo. Como se mostró anteriormente, el proceso puede ocurrir a través de una serie de pasos elementales, que podrían o no involucrar un proceso de protonación/desprotonación de las especies involucradas. Además, resultados recientes de nuestro grupo [98] para péptidos que contienen radicales de tirosina en solución muestran que los pasos de IET pueden estar fuertemente acoplados a la toma y liberación de protones, conduciendo entonces a un proceso de transferencia electrónica acoplada a transferencia de protón (PCET, *proton coupled electron transfer*, por sus siglas en inglés). En este caso, la PCET se define en su sentido más amplio siguiendo el trabajo de Hammes-Shiffer y Soudackov [99], es decir que involucra el movimiento de un protón a través de diferentes enlaces ya sea de forma concertada o secuencial [98]. A continuación nos avocamos al análisis de esta última posibilidad.

Análisis clásicos de la transferencia electrónica Tyr-Tyr

Comenzamos nuestro análisis calculando el cambio en la energía libre (ΔG) asociado al proceso de transferencia electrónica de Tyr a Tyr* (o PCET) en *MtCYP121* utilizando un esquema de integración termodinámica como se describe en la sección métodos. Dividimos el proceso en tres pasos elementales: desprotonación de la tirosina lejana (Figura 3.6, paso 1), movimiento del electrón desde la tirosina lejana a la tirosina radicalaria cercana, que corresponde al proceso migración del radical hacia la tirosina lejana (Figura 3.6, paso 2), y protonación de la tirosina cercana (Figura 3.6, paso 3).

A pesar de que es sabido que los cambios absolutos en la energía libre del equilibrio de protonación calculados con campos de fuerzas clásicos utilizando TI arrojan resultados significativamente distantes de los valores experimentales, varios trabajos muestran que los valores relativos de energía libre de un residuo dado en el entorno proteico con respecto al valor de referencia en solución, son capaces de predecir correctamente el corri-

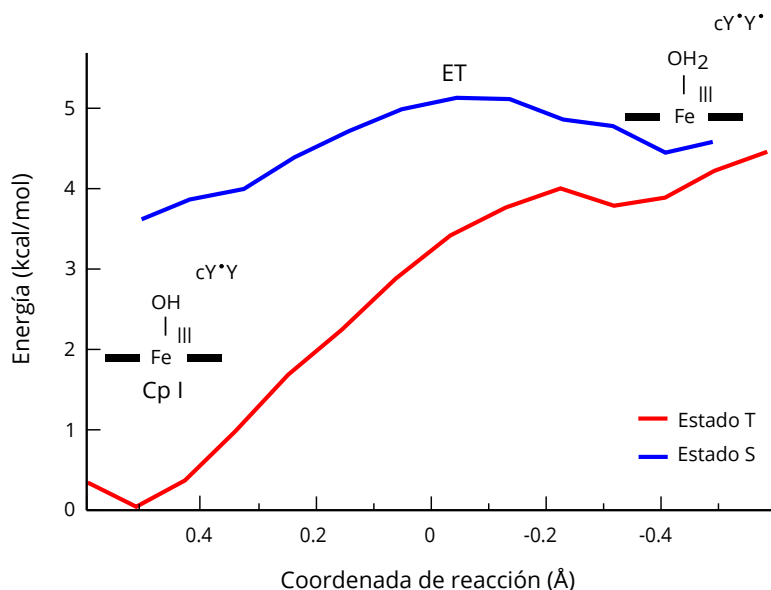


Figura 3.5: Perfil de energía para la abstracción de un átomo de hidrógeno y formación de Tyr* utilizando el Cp II como oxidante. Las líneas roja y azul corresponden a los estados de espín triplete y singlete respectivamente.

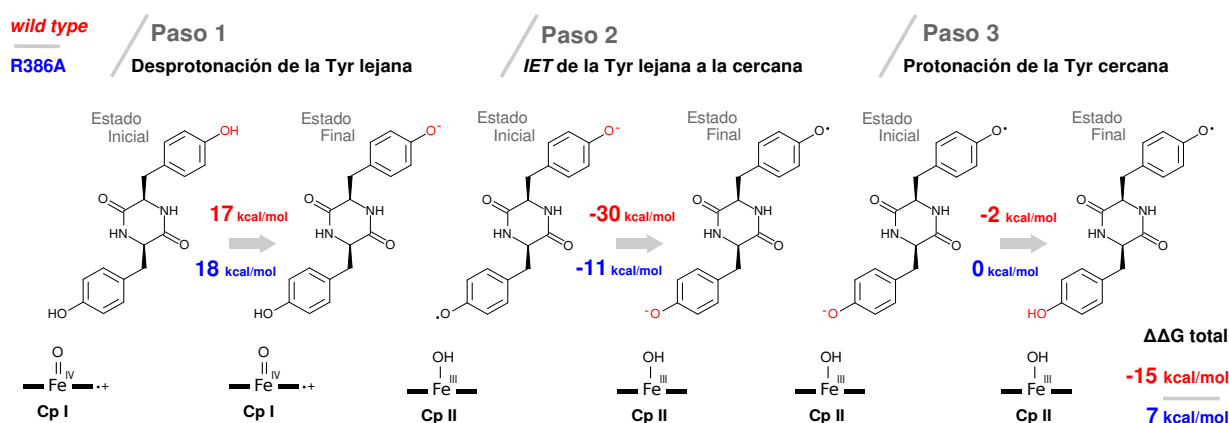


Figura 3.6: Análisis clásicos del proceso de PCET de Tyr* a Tyr utilizando TI. Los valores de $\Delta\Delta G$ corresponden a la diferencia de energía libre de cada proceso (paso) dentro de la proteína respecto al mismo proceso en agua.

miento o la tendencia en los valores de pKa [100, 101]. Por ello, en el presente caso, cada paso fue evaluado para el sustrato dentro de *MtCYP121* y en solución acuosa. Los valores reportados corresponden a la diferencia entre los valores de energía libre en la proteína y en agua para cada paso, y un valor negativo significa que el paso es más favorable que ocurra dentro de *MtCYP121* respecto a la solución.

Los resultados presentados en la Tabla 3.4 muestran que es improbable que la tirosina lejana se desprotona espontáneamente. El pKa de la tirosina en solución es de alrededor de 10.2 [102], pero dado que la energía libre del proceso de desprotonación dentro de *MtCYP121* es menor que en solución ($\Delta\Delta G = 17.3 \pm 1.0$ kcal/mol), su pKa debería ser aún más alto. Este no es el caso para la tirosina cercana donde el valor de energía libre de desprotonación es cercano al valor en solución y en consecuencia se espera un valor de pKa similar al de la tirosina en solución, lo que podría resultar en la protonación de la tirosina cercana luego de la IET. Un dato que es más importante, es que el paso de transferencia radicalaria es altamente exergónico, y compensa con creces la energía libre de la primer desprotonación, resultando en total en una reacción PCET espontánea ($\Delta\Delta G = -14 \pm 3.2$ kcal/mol). La naturaleza "PCET" del paso de IET se evidencia claramente por el hecho de que solo considerando los primeros dos pasos (o los tres) juntos se obtiene una reacción espontánea. Así, el entorno de

Paso	En agua	Dentro de la proteína		$\Delta\Delta G$ wt-agua	$\Delta\Delta G$ R386A-agua
		wild type	R386A		
1. Desprotonación de TYR lejana	-65.0 ± 0.2	-47.7 ± 0.8	-46.8 ± 0.6	17.3 ± 1.0	18.2 ± 0.8
2. IET de TYR lejana a cercana	-1.0 ± 0.1	-29.8 ± 1.5	-11.5 ± 0.5	-28.8 ± 1.6	-10.5 ± 0.6
3. Protonación de TYR cercana	60.9 ± 0.2	58.4 ± 0.4	60.6 ± 0.4	-2.5 ± 0.6	-0.3 ± 0.6
Total $\Delta\Delta G$				-14.0 ± 3.2	-7.6 ± 2.0

El error absoluto fue estimado como el desvío estándar de los valores obtenidos de dos simulaciones independientes.

Tabla 3.4: Cambios de energía libre asociados con la migración radicalaria determinados por integración termodinámica. Los valores de energía libre están expresados en kcal/mol.

la enzima *MtCYP121* favorece fuertemente una tirosina lejana radicalaria neutra y una tirosina cercana cargada negativamente.

Para comprender las bases moleculares de la localización radicalaria en la tirosina lejana, analizamos aquellos residuos que podrían promover la presencia de una tirosina lejana neutral y/o de una tirosina cercana cargada. Arg386 es un residuo clave localizado muy próximo a la tirosina cercana, y dada su carga positiva natural a pH fisiológico (Figura 3.1), decidimos analizar su influencia en el proceso de transferencia de carga. Para este propósito, calculamos la energía libre de la IET en la mutante R386A de *MtCYP121* (Figura 3.6, paso 2). Los análisis de la estructura y dinámica de la proteína mutante R386A muestran que no ocurren mayores cambios en la estructura proteica, un hecho que es consistente con la estructura de rayos X correspondiente [92]. Un punto interesante, y consistente con un rol importante de esta arginina como guía del proceso de transferencia radicalaria, es que el $\Delta\Delta G = -10.5 \pm 0.6$ kcal/mol para el paso de transferencia de carga (paso 2 en la Figura 3.6) es significativamente más pequeño que en la proteína wt (-28.8 ± 1.6 kcal/mol). Por consiguiente, Arg386 desempeña un rol clave para el proceso de IET propuesto.

Es importante señalar que la información presentada anteriormente solo mide el efecto de la proteína en el proceso de movimiento de la carga negativa de una tirosina a la otra, como se define en los parámetros clásicos. Por consiguiente, esto representa un caso extremo o ideal donde los residuos de tirosina o son neutros o tienen una carga de un electrón. Un cálculo más preciso requeriría el uso de métodos de energía libre QM/MM como lo hicieron y revisaron Kamerlin y colaboradores [103], lo cual es computacionalmente demasiado demandante para el sistema en cuestión.

Análisis QM/MM del proceso de IET de Tyr a Tyr*

Para entender más a fondo y con mayor detalle cómo la proteína puede modular o promover la IET, realizamos optimizaciones QM/MM del sustrato monoradicalario y negativamente cargado ($cYY^{\bullet-}$) relacionado directamente con el proceso de IET y analizamos las distribuciones de carga y espín resultantes (Tabla 3.5). En primer lugar calculamos la estructura electrónica de la especie $cYY^{\bullet-}$ dentro de la proteína, considerando la porción hémica clásicamente (es decir que no puede ocurrir transferencia de carga desde o hacia el grupo hemo). En concordancia con los resultados clásicos anteriormente descriptos, las poblaciones de Mulliken y espín resultantes muestran que la matriz proteica favorece la localización del radical (espín) en la tirosina lejana y de la carga negativa en la tirosina cercana. Adicionalmente, también calculamos la afinidad protónica de la tirosina lejana dentro de *MtCYP121* considerando una tirosina cercana neutra o radicalaria (es decir, después del primer paso de abstracción de hidrógeno). Los resultados muestran que la presencia de la tirosina radicalaria disminuye la afinidad protónica en 6 kcal/mol, favoreciendo así la liberación del protón en la tirosina lejana. Los resultados

obtenidos se vuelven aún más reveladores cuando se agrega la porción hémica en estado de Cp II al sistema QM. Como se muestra en la Tabla 3.5, para ambos estados de espín considerados (C y D), las tirosinas cercana y lejana muestran un carácter radicalario (carga pequeña y densidad de espín desapareado $>0.5e^-$) mientras que el hemo muestra un estado hidroxil-férrico ($Fe^{III}-OH^-$), que corresponde al estado basal de la *MtCYP121*. Además, la diferencia entre ambos estados de espín es que, mientras que el estado D presenta ambos residuos de tirosina acoplados antiferromagnéticamente y el único electrón desapareado permanece en el hemo, en el estado C, cYY evidencia el carácter de estado de triplete birradicalario esperado (ver más abajo).

Estos resultados sugieren que una vez que se formó el primer radical en la tirosina cercana, cuando la tirosina lejana libera su protón fenólico, el electrón es tomado por el Cp II, resultando en un cYY birradicalario y un hemo férrico unido a hidróxido. En consecuencia, una IET parece operar en *MtCYP121* para generar rápida y eficientemente el segundo radical. Es importante notar que en el presente trabajo no evaluamos los parámetros energéticos de acoplamiento y reorganización que en conjunto con el correspondiente cambio de energía libre controlan la velocidad de la IET. Sin embargo, notamos que la IET en residuos adyacentes y/o covalentemente unidos (como las dos tirosinas y el hemo en el presente sustrato) es muy rápida si la reacción es exergónica y por ello no se espera que la energía de acoplamiento y reorganización contribuyan de manera significativa a la velocidad de la reacción completa [104, 105].

Parámetro	Subsistema cuántico		
	[cYY]*	Cp II [cYY]*	
		D	C
q TYR cercana	-0.778	-0.007	-0.065
q TYR lejana	-0.222	-0.325	-0.299
s TYR cercana	0.430	-0.997	0.909
s TYR lejana	0.570	0.673	0.703
q Hemo-CYS	n.c	-0.188	-0.146
s Hemo-CYS	n.c	1.207	1.262
q L	n.c	-0.476	-0.486
s L	n.c	0.114	0.126

Tabla 3.5: Poblaciones de Mulliken y espín para el sustrato mono radicalario y negativamente cargado (cYY*)

3.2.3 cY*Y* a mcic: la formación del enlace C-C ocurre en solución

Simulaciones de DM de cY*Y* en solución y dentro de *MtCYP121*

Como se mencionó en la introducción, la reacción de acoplamiento C-C luego de la formación del birradical, puede ocurrir tanto dentro de la proteína como en solución luego de una posible liberación del sustrato birradicalario (cY*Y*). En cualquier caso, para que la reacción ocurra, los dos carbonos orto posibles deben acercarse lo suficiente para permitir que ocurra la formación del enlace (vide infra). Para analizar la probabilidad de este paso, inicialmente realizamos 30 ns de simulación de DM de cY*Y* en solución y dentro de *MtCYP121* y medimos las cuatro distancias relevantes (todas las combinaciones posibles de dos carbonos orto que puedan formar el enlace C-C). El histograma resultante para cY*Y* libre (Figura 3.7) muestra que todos los pares tienen una distribución bimodal con picos a 7 y 10.5 Å (los cuales corresponden a la distancia entre los carbonos orto más cercanos y los más lejanos en la conformación más estable). Un dato interesante es que se encuentra una significativa población de conformaciones en las cuales ambos átomos de carbono se encuentran a 5 Å, con una distancia mínima de 4.5 Å. Los resultados obtenidos de la simulación de DM de cY*Y* dentro de la proteína muestran que en este caso el par de carbonos orto más cercanos está más cerca que los otros durante la mayor parte del tiempo de simulación, y esta distancia es la misma que se obtuvo para cY*Y* en solución. Por consiguiente, se puede encontrar un valor de distancia similar y cercano de alrededor de 4.5 Å entre dos carbonos potencialmente reactivos

tanto en solución como dentro de MtCYP121.

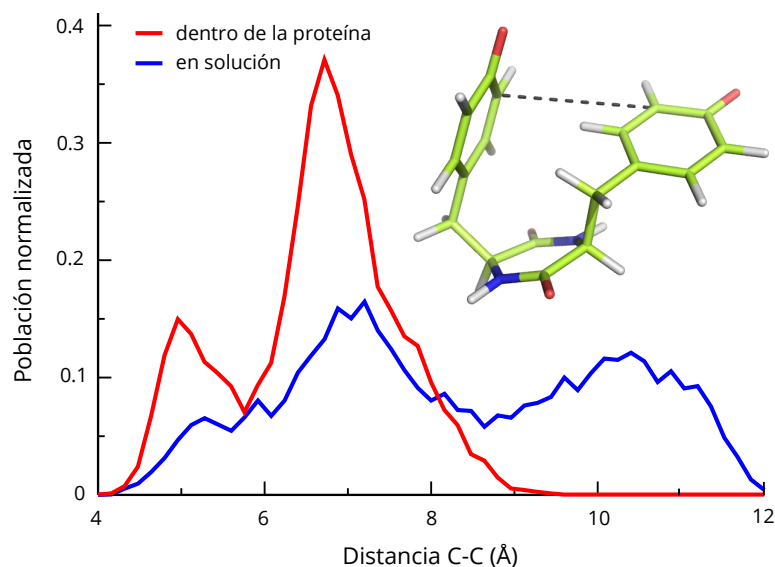


Figura 3.7: Histogramas de distancias C-C para los cuatro pares de carbonos relevantes obtenidas de la simulación de cY*Y* libre en solución (azul) y dentro de la proteína (rojo).

Formación del producto mcic en vacío

Comenzamos por el estudio del birradical cY*Y* aislado en vacío para tenerlo como reacción de referencia. El perfil de energía de reacción correspondiente se determinó utilizando la distancia $C\epsilon_{Tyr1}-C\epsilon_{Tyr2}$ como coordenada de reacción. Los resultados y estructuras clave a lo largo de la coordenada de reacción se muestran en la Figura 3.8 y Tabla 3.6.

Como se esperaba, el gráfico muestra que para distancias largas el estado basal corresponde al estado de espín de triplete (T). El análisis de las poblaciones de Mulliken y espín muestra un electrón desapareado localizado en cada tirosina y que ambas son neutras. En la especie reactiva en vacío, el estado singulete (S) se encuentra ~ 11 kcal/mol por encima del estado T, siendo aún ambas tirosinas neutras. La distancia más cercana entre dos carbonos orto en la estructura optimizada en vacío es 4.6 Å, similar al valor más cercano posible en solución. A medida que la distancia C-C comienza a decrecer, la energía del estado T aumenta continuamente, mientras que la energía del estado S también se incrementa pero con una pendiente menor. A una distancia de alrededor de 2.4 Å, ambas superficies se cruzan (llamaremos a esta estructura ET_A). Luego de este cruce, la curva del estado S cae a un primer mínimo correspondiente a un primer estado intermedio (I_A). La barrera efectiva partiendo del estado basal T es 16 kcal/mol. Un punto interesante es que la estructura del ET_A muestra un valor de ángulo diedro alrededor del enlace $C\alpha-C\beta$ de $\sim 137^\circ$, casi igual al valor más poblado entre ambos rotámeros de tirosina posibles en solución para cYY (64° y 183°). El estado I_A , que se encuentra 6 kcal/mol por debajo de cYY en el estado T, muestra un enlace C-C bien establecido (1.58 Å), con la conversión simultánea de ambos carbonos orto de una configuración puramente aromática sp^2 a sp^3 (Figura 3.8, estructura I_A). En el estado T del I_A , ambos espines desapareados se localizan en la misma tirosina, evidenciando claramente que ha ocurrido un cambio significativo en la estructura electrónica de la molécula. Para proceder desde el I_A hacia el producto final, los hidrógenos de los $C\epsilon$ (He) deberían ser transferidos a los oxígenos fenólicos. El análisis estructural del I_A muestra que ambos He están más cerca del átomo de oxígeno de la otra tirosina (2.43 Å) que del oxígeno fenólico de la misma tirosina (2.73 Å). En consecuencia analizamos la transferencia de hidrógenos cruzada y secuencial utilizando la diferencia entre las distancias $C\epsilon-He$ y $He-O\zeta$ como coordenada de reacción. El perfil de energía correspondiente muestra que el primer hidrógeno se transfiere con una barrera de ~ 20 kcal/mol pasando a través del ET_B y conduciendo hacia el I_B . En I_B una tirosina tiene un C orto planar mientras que la otra tiene un carbono orto tetraédrico (Figura 3.8, estructura I_B). La transferencia del segundo hidrógeno posee una barrera muy pequeña (~ 2 kcal/mol) y finalmente conduce al producto en el cual ambos carbonos orto adoptan una orientación planar (Figura 3.8, estructura mcic). Como se esperaba en base al análisis estructural, tratar de transferir el hidrógeno al oxígeno de

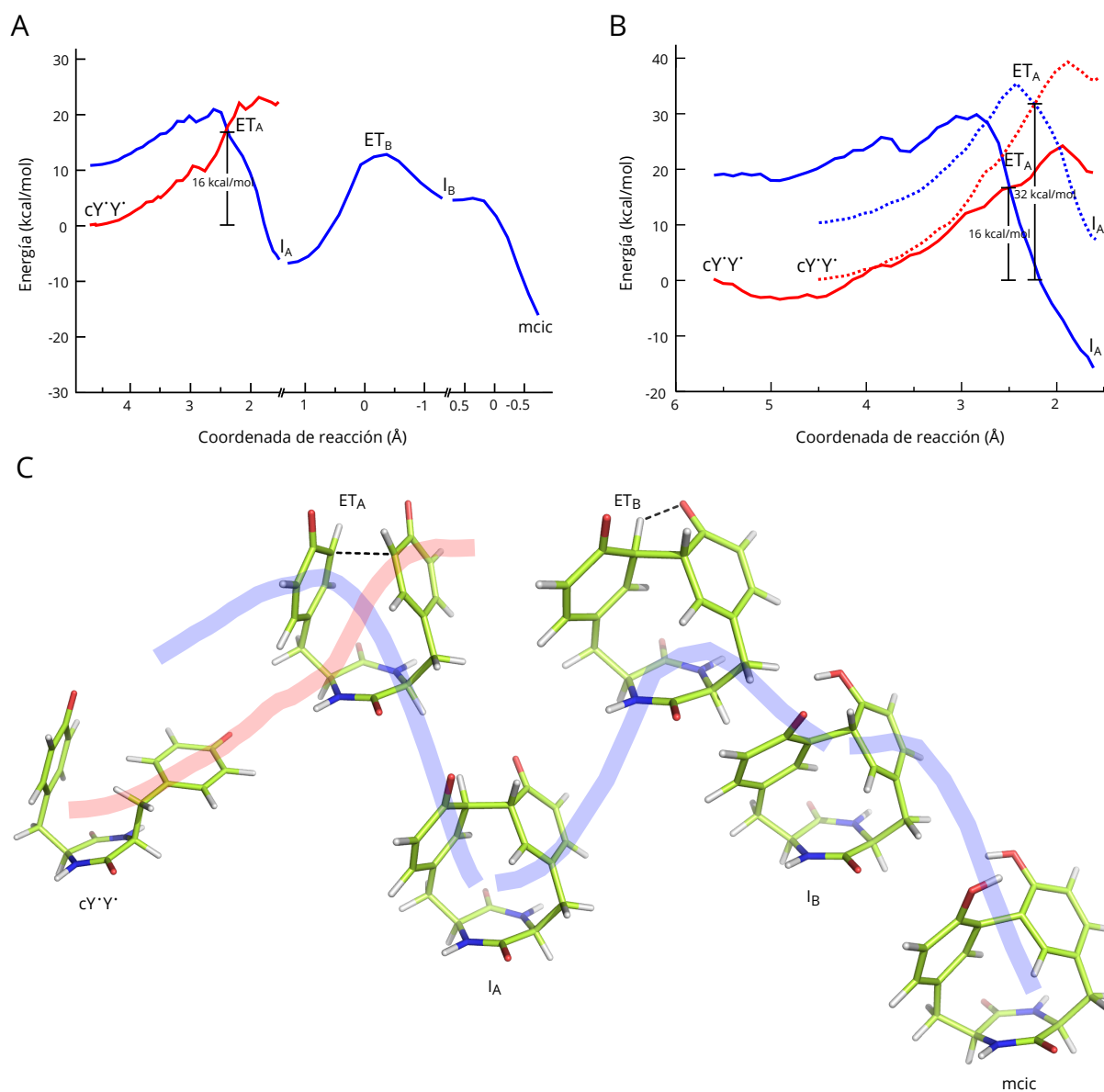


Figura 3.8: (A) Perfil de energía en función de la coordenada de reacción para la reacción completa en vacío. Para el primer paso de la reacción los cálculos fueron realizados para ambos estados de espín S (azul) y T (rojo). Para los pasos subsiguientes solo son relevantes los perfiles correspondientes al estado S. El símbolo // sobre el eje X indica el cambio en la coordenada de reacción. (B) Perfil de energía para la formación del enlace C-C en solución (línea continua) y dentro de la proteína (línea punteada) para los estados de espín S y T. (C) Estructuras claves en la reacción de formación de micociclosina.

Parámetro	Parámetros energéticos					
	Vacío		Solvente		Proteína	
	T	S	T	S	T	S
Δ de espín	0	10	0	19	0	11
$E_{act} cY^*Y^* \Rightarrow I$	22	10	23	10	40	25
E_{act} efectiva	16		16		32	
ΔE de reacción	-7		-15		6	

Parámetros estructurales, poblaciones de Mulliken y espín para la reacción completa en vacío						
	$cY^*Y^*(T)$	$ET_A(S/T)$	$I_A(S)$	$ET_B(S)$	$I_B(S)$	$mcic(S)$
d C ϵ -C ϵ	4.52	2.59	1.59	1.56	1.55	1.52
d C ϵ -H ϵ	1.11	1.11	1.13	1.51	2.42	2.38
d H ϵ -O ζ (misma TYR)	2.65	2.67	2.69	3.04	3.30	3.85
d H ϵ -O ζ (diferente TYR)	5.25	2.64	2.49	1.24	1.00	0.99
ángulo planar C ϵ	89.64°	65.60°	61.04°	58.32°	62.95°	68.42°
q TYR1	0.029	0.019	0.040	0.198	0.350	0.014
q TYR2	-0.032	-0.021	-0.038	-0.194	-0.350	-0.010
s TYR1	1.013	0.000	0.000	0.000	0.000	0.000
s TYR2	0.988	0.000	0.000	0.000	0.000	0.000

Tabla 3.6: Energía, distancias seleccionadas (d), poblaciones de Mulliken (q) y espín (s) relevantes para la formación del enlace C-C en vacío (referencia), solvente y dentro de la proteína.

la misma tirosina resulta en una barrera mucho más alta (~ 45 kcal/mol) que hacerlo al oxígeno de la otra tirosina (Figura 3.9 A). Dado que la barrera obtenida para el acoplamiento C-C es bastante elevada, decidimos explorar un mecanismo alternativo, en el cual primero el protón/hidrógeno migra desde el carbono orto al oxígeno fenólico y luego ocurre la formación del enlace C-C entre los radicales de tirosina localizados en ambos carbonos orto. La barrera obtenida para esta transferencia de protón/hidrógeno es de más de 50 kcal/mol (ver Figura 3.9 B, por consiguiente es significativamente más alta que la observada para el acoplamiento C-C de las tirosinas con el radical localizado en el oxígeno fenólico.

Formación del producto mcic en solución

Para estudiar la reacción de acoplamiento de tirosinas en solución, partimos de una foto tomada de la simulación de DM correspondiente en la cual los átomos de carbono estaban situados a 5 Å de distancia entre sí. El gráfico de energía correspondiente, presentado en línea continua en la Figura 3.9, presenta la misma forma que el descrito previamente para la reacción en vacío. En las estructuras de reactivos, el estado basal corresponde al estado T, con dos radicales de tirosina, y un estado S que se encuentra alrededor de ~ 19 kcal/mol por encima. Cuando la distancia C-C comienza a decrecer, la energía de ambos estados se incrementa continuamente. En la superficie del estado T, a una distancia C-C de aproximadamente 2 Å, se encuentra un ET (con una barrera de 23 kcal/mol). Sin embargo, a una distancia de alrededor de 2.5 Å, los perfiles de los estados T y S se cruzan, produciendo una estructura de transición efectiva (ET_A), la cual se encuentra unas 16 kcal/mol por encima de la energía del estado basal, un valor que es similar al obtenido en vacío. Luego de atravesar este ET_A , la energía desciende a lo largo del estado S hasta alcanzar el primer intermediario (I_A). La reacción procede luego de manera similar a la observada en vacío, con una primer transferencia de protón/hidrógeno asociada a una barrera alta, y una barrera muy pequeña para la transferencia del segundo protón/hidrógeno.

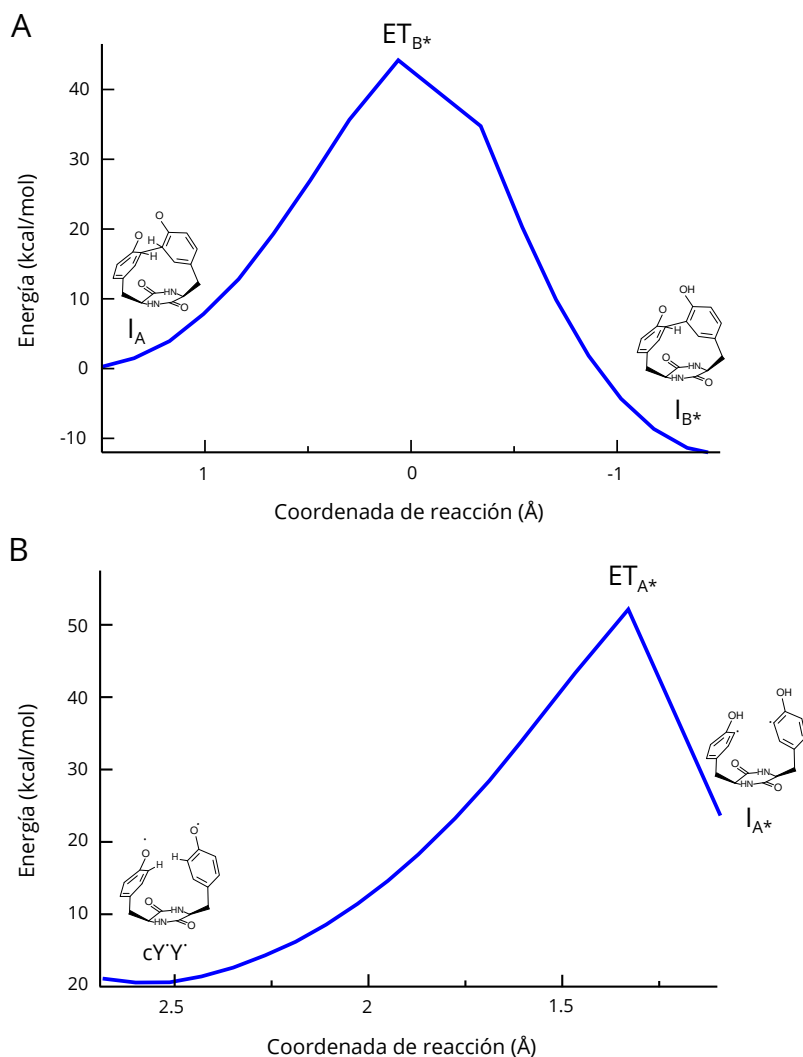


Figura 3.9: A. Perfil de energía para la transferencia de hidrógeno al oxígeno de la misma Tyr en la reacción de I_A a I_B . B. Perfil de energía para la transferencia del hidrógeno orto al oxígeno, como alternativa para generar el radical en el carbono orto previo al acoplamiento C-C.

Formación del producto mcic dentro de MtCYP121

Para estudiar la reacción de acoplamiento de las tirosinas dentro de *MtCYP121*, utilizamos la misma estrategia descrita anteriormente para la reacción en vacío y en solución. El gráfico de energía correspondiente, presentado en líneas punteadas en la Figura 3.9-B, posee la misma forma que los descritos previamente. En las estructuras de los reactivos, el estado basal corresponde al estado T, con dos radicales de tirosina y un estado S que se encuentra alrededor de ~ 11 kcal/mol por encima. A medida que la distancia C-C se reduce, la energía de ambos estados se incrementa continuamente. En la superficie correspondiente al estado T, a una distancia C-C de alrededor de 2 Å, se encuentra un ET (con una barrera de 40 kcal/mol), pero a una distancia de aproximadamente 2.2 Å las superficies de los estados T y S se cruzan (ET_A), produciendo una barrera efectiva de 32 kcal/mol respecto a la energía del estado basal. Luego de atravesar el ET, la energía desciende a lo largo del estado S hasta el primer intermediario I_A , el cual se ubica unas 5 kcal/mol por encima de los reactivos. La reacción prosigue de forma similar a lo observado en vacío y en solución, con una primera transferencia de protón/hidrógeno asociada a una barrera alta, y una barrera muy pequeña para la transferencia del segundo protón/hidrógeno.

En consecuencia, los resultados evidencian que en todos los casos (vacío, solución, proteína) la reacción de acoplamiento C-C ocurre siguiendo los mismos pasos elementales, siendo el paso limitante de la velocidad de la misma la formación del enlace C-C y la consecuente transición de espín. El hallazgo más importante, comparando los tres ambientes, es que dentro de *MtCYP121* la barrera energética para este primer paso es significativamente más alta que en vacío/solución, lo que posiblemente estaría reflejando el hecho de que el sitio de unión de la

proteína está preparado para alojar a cYY, pero no al producto mcic, y en consecuencia se requiere de energía adicional para permitir que ambas tirosinas se acerquen lo suficiente como para establecer el enlace C-C.

3.3 Discusión

En este capítulo hemos analizado en detalle el mecanismo de la reacción catalizada por *MtCYP121*, que comprende la reacción de acoplamiento C-C entre dos carbonos orto de los grupos fenólicos de cicloditirosina (cYY), para producir micociclosina (mcic), por medio de simulaciones de dinámica molecular y QM/MM. Para este fin dividimos la reacción en tres pasos elementales: (i) generación del primer radical de tirosina, (ii) generación del segundo radical de tirosina y (iii) reacción de acoplamiento C-C.

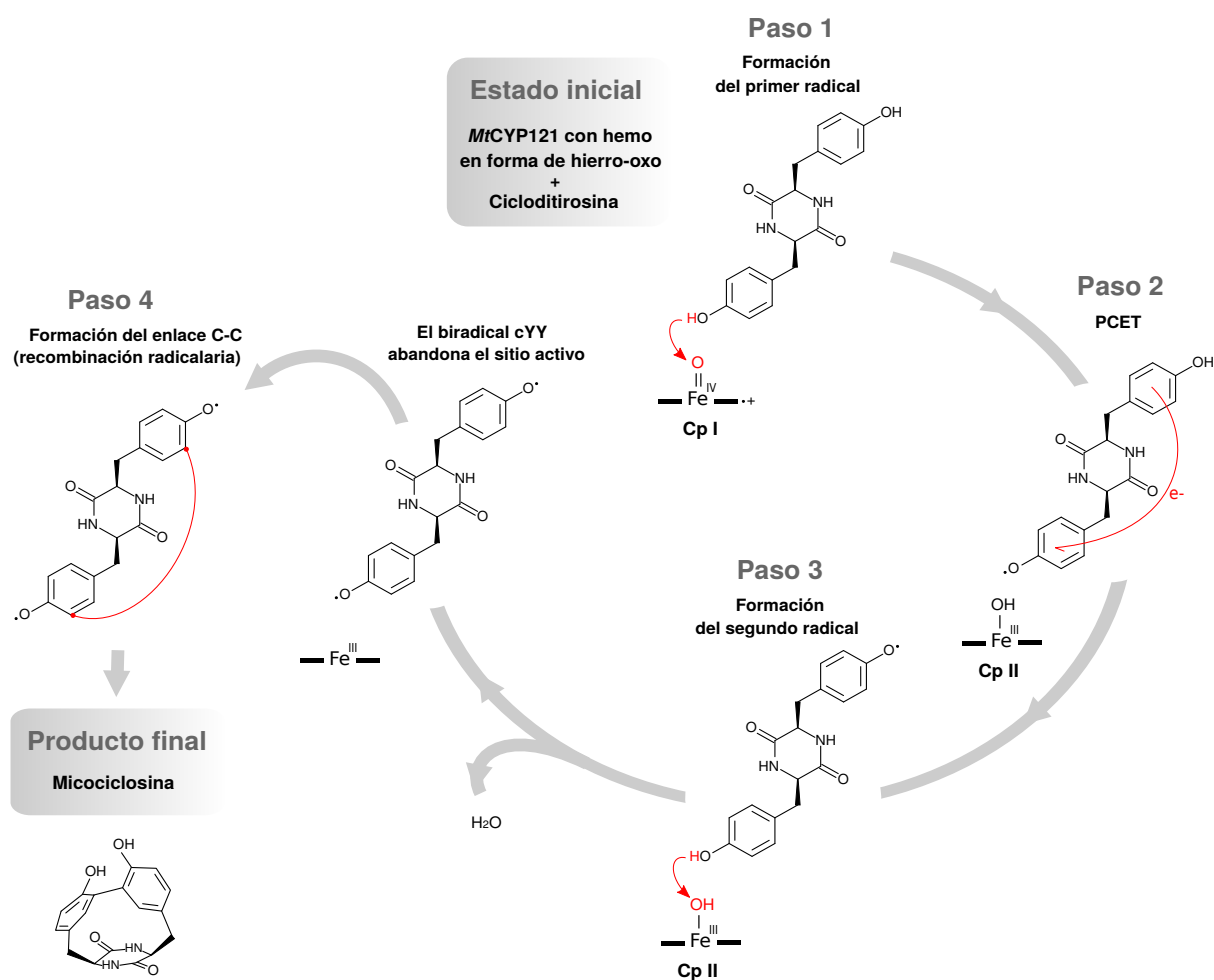


Figura 3.10: Mecanismo de reacción de cuatro pasos propuesto para *MtCYP121*.

[p]

Una propuesta detallada y completa del mecanismo de reacción derivada de nuestros resultados en la cual se resalta el rol clave de la matriz proteica de *MtCYP121* se muestra en la Figura 3.10. Los valores energéticos claves son reportados en la Tabla 3.7 y se discuten a continuación.

Nuestros resultados para el primer paso muestran que lo más probable que el Cp I sea el responsable de la generación del primer radical de tirosina en la tirosina cercana del sustrato cYY. La reacción de abstracción de hidrógeno (del oxígeno fenólico) posee una barrera relativamente pequeña, si se la compara por ejemplo con las barreras de abstracción de hidrógeno en CYPs que hidroxilan alquilo, que están en el rango de 10-20

Paso	Estado de espín ^a	ΔE	$\Delta E^\ddagger$	Método
Formación del primer radical por medio del Cp I	Q	-	$4 \pm$	QM-MM
re-unión del OH	Q	-	16 ± 0.2	QM-MM
Contribución clásica a la PCET ^b	D	-14.0 ± 3.2	n.c ^c	TI
Formación del segundo radical por medio del Cp II	T	3.5 ± 0.1	3.5 ± 0.1	QM-MM
Formación del enlace C-C en vacío y en solución ^d	T=>S	-7 ± 0.1 (-15 ± 0.1)	16 ± 0.1 (16 ± 0.1)	QM y (QM-MM)

^aEl estado de espín de referencia corresponde al estado basal donde ocurre la reacción. Si la reacción involucra un cambio de espín se mencionan ambos estados.

^bEn este caso el valor corresponde a la energía libre (es decir, ΔG) en vez de energía potencial.

^cNo calculado.

^dLos valores entre paréntesis corresponden a los obtenidos en solución.

Tabla 3.7: Cambios de energía libre asociados con la migración radicalaria determinados por integración termodinámica. Los valores de energía libre están expresados en kcal/mol.

kcal/mol [97] y es iso o levemente exergónica. Lo que parece interesante es el hecho de que la interacción de puente de hidrógeno entre el oxo del Cp I y el OH_{Tyr} se establece solo transitoriamente, mientras que cuando el dioxígeno está unido al hemo el puente de hidrógeno que se establece es permanente. Además mostramos que ni el dioxígeno ni el peróxido son capaces de producir el primer radical de tirosina, la primer reacción es termodinámicamente inviable, mientras que en el último caso se produce la abstracción de un protón (no un hidrógeno), lo que produce una tirosina negativa (que podría ser protonada nuevamente por las moléculas de agua presentes en el sitio activo, ver más abajo), y eventualmente participar como fuente de protón para la formación del Cp II. Por consiguiente, parece que la tirosina cercana del sustrato cYY se encuentra posicionada en una posición subóptima con respecto al hierro hémico para ser oxidada. Aunque podríamos no dar mayor importancia a esta observación, las proteínas casi nunca hacen algo sin un propósito específico.

La pregunta que surge entonces es por qué el grupo fenol está posicionado en una posición subóptima. Para responder a esta pregunta proponemos que el posicionamiento del fenol conlleva las siguientes dos consecuencias claves. La primera es que una vez que se producen el primer radical y el Cp II, pueden ocurrir dos reacciones: la PCET propuesta que permite que ocurra la reacción de MtCYP121 pero también el paso de re-unión del OH observado para muchos CYPs que catalizan reacciones simples de hidroxilación. Las reacciones de hidroxilación en fenoles típicamente ocurren en el carbono orto, pero para que puedan tener lugar, el átomo de carbono debe estar cerca del OH_{Cp II}, un hecho que en MtCYP121 ocurre solo transitoriamente. Aún para esta estructura más reactiva calculamos la barrera energética de la reacción de re-unión, lo que produjo una barrera de entre 16 y 20 kcal/mol [Figura 3.4]. Las barreras obtenidas para el paso de re-unión del OH en otros CYPs son mucho más pequeñas, en el rango de 3-4 kcal/mol, y pueden ser aún menores para algunos estados de espín. Esta diferencia puede ser explicada simplemente por la posición relativa del carbono reactivo, que en cYY se encuentra a aproximadamente 5 Å del oxígeno del Cp II, mientras que en los CYPs que hidroxilan la distancia óptima es <3 Å [97].

Otra explicación posible para la elevada barrera observada para el paso de re-unión, es la deslocalización del espín del radical de tirosina en todo el grupo fenol. En los radicales alquilo, que son los sustratos usuales hidroxilados por CYPs, el espín está altamente localizado en el carbono reactivo (>0.8e-) [106], mientras que en el radical de tirosina estudiado en este trabajo la densidad de espín en el carbono orto es pequeña (<0.3e-). La baja

densidad de espín sobre el carbono orto además podría contribuir a explicar por qué el paso de acoplamiento C-C también exhibe una barrera alta (ver más abajo). En cualquier caso, la alta barrera para la hidroxilación es consistente con los hallazgos recientes realizados utilizando espectrometría de masas que muestran que no se detectó hidroxilación de cYY, ni siquiera en los minoritarios productos secundarios (<2 %) diferentes de mcic [96].

En segundo lugar, es importante señalar que en cada ciclo de reacción *MtCYP121* requiere dos electrones para reducir el dioxígeno a Cp I. Es importante que cada ciclo de reducción de oxígeno resulte en oxidación de cYY, y no en reducción improductiva de oxígeno a agua. Para tal fin, es necesario un acoplamiento fuerte de los cosustratos (cYY y oxígeno molecular). En muchas hemoproteínas, un residuo de tirosina localizado a distancia de puente de hidrógeno del dioxígeno unido al hemo es el principal responsable de esta estabilización y en general de la alta afinidad por oxígeno. Esta alta afinidad es un elemento clave para asegurar la eficiencia de las reacciones oxidativas subsiguientes [56, 107]. Dadas las observaciones antes mencionadas en relación a la posición de la tirosina cercana del sustrato cYY y sus interacciones con el ligando hémico dioxígeno, calculamos la energía de unión del oxígeno a *MtCYP121* reducida en presencia y en ausencia de cYY. Nuestros resultados muestran que la presencia de cYY incrementa la energía de unión del oxígeno en alrededor de 7 kcal/mol. Este valor es similar al reportado en un trabajo previo de nuestro grupo en el que se compara la hemoglobina truncada *N wt* (cuyo residuo TyrB10 se encuentra formando un puente de hidrógeno con el ligando dioxígeno) con la mutante TyrB10-Ala, que evidencia una disminución de más de 100 veces la afinidad de oxígeno [56, 107]. Por consiguiente, está claro que la presencia de cYY incrementa significativamente la afinidad por oxígeno en *MtCYP121*, promoviendo de esta manera su reducción solo cuando el sustrato de la enzima está presente, y decrementando así el consumo de no productivo del poder reductor. La relevancia del OH_{Tyr} para la reacción de oxidación es consistente con observaciones recientes que muestran que cuando la tirosina es reemplazada por fenilalanina, puede encontrarse en su sitio cerca del hemo, pero es oxidada a una velocidad muy baja. Además, en la estructura cristalina correspondiente no se encuentra la red de moléculas de agua cercanas al hierro.

Respecto al paso de formación del segundo radical, los resultados presentados aquí favorecen claramente un proceso de transferencia electrónica interna (o rearrreglo de carga), que puede ser descripto como PCET. Específicamente la presencia de una tirosina cercana radicalaria disminuye ligeramente la afinidad protónica de la tirosina lejana (favoreciendo su liberación al ambiente acuoso), mientras que la matriz proteica, en parte debido a la presencia de Arg386, favorece la localización de la carga negativa en la tirosina cercana. Esto resulta en una fuerza impulsora neta para la formación de una tirosina lejana radicalaria y una tirosina cercana cargada negativamente. Adicionalmente, aunque el Cp II es capaz de abstraer un hidrógeno de la tirosina cercana para generar una tirosina radicalaria, lo hace con un costo energético de alrededor de 4 kcal/mol (aunque debería tenerse en cuenta que la barrera y el costo energético es menor que el observado para la abstracción de hidrógeno de sustratos alquiles por parte del Cp II, de 20 y 13 kcal/mol respectivamente [19, 97]). Un dato sorprendente es que puede actuar eficientemente como trampa electrónica, produciendo el birradical cY*Y* y regenerando el estado basal de la enzima hidroxio férrica. La propuesta alternativa (sin IET) requiere que cY*Y abandone la proteína y reingrese en la dirección opuesta. Esta propuesta sin embargo, no solo disminuiría la velocidad total de la enzima sino que además llevaría a la presencia de una cantidad significativa de cY*Y en solución, lo cual debido a su naturaleza radicalaria podría conducir a la formación de productos secundarios no deseados, como por ejemplo dímeros de cYY (debido a la formación de enlaces C-C entre radicales de tirosina de diferentes moléculas de cY*Y).

Observando ambos pasos de generación de radicales de tirosina en conjunto, está claro que el paso limitante debería ser la formación del oxidante fuerte denominado Cp I. Una vez que el Cp I está presente, éste abstrae el hidrógeno fenólico produciendo un radical en la tirosina cercana, que luego procede rápida y directamente hacia el birradical a través de la PCET. Por consiguiente, ambos radicales de tirosina se forman de manera casi concertada en un único ciclo de reducción del hemo. Una vez que se establece el birradical cY*Y*, el último paso de la reacción es el acoplamiento Tyr-Tyr o formación del enlace C-C. Nuestros resultados muestran que los pasos mecanísticos de esta reacción son los mismos para la reacción en vacío, en agua o dentro de *MtCYP121*. En todos los casos el paso limitante es la formación del enlace C-C, que conduce hacia el primer intermediario I_A donde el enlace C-C ya está establecido, y en el cual los dos carbonos orto adoptan transitoriamente una conformación

sp3. El primer paso de la reacción se dificulta más por el hecho de que requiere una transición de espín de triplete a singulete, dado que en cY*Y* (que posee un electrón desapareado localizado en cada anillo aromático) el estado basal es el triplete. Respecto a la energía y geometría de los diferentes estados, nuestros resultados coinciden perfectamente con cálculos previos de la reacción de acoplamiento Tyr-Tyr en vacío utilizando B3LYP, donde se calculó la energía relativa de los intermediarios estables a lo largo de la reacción, mostrando que un estado similar al I_A se encuentra alrededor de 12 kcal/mol por debajo de cY*Y*. En ese caso no se reportaron barreras [108].

Más allá del mecanismo, cuando se compara la energética de la reacción en los tres ambientes estudiados está claro que la matriz proteica no la promueve. De hecho, aunque la simulación de DM muestra que los carbonos reactivos pueden encontrarse cerca uno del otro, la barrera energética para establecer el enlace es significativamente mayor dentro de la proteína comparada con la reacción en vacío/agua. En otras palabras, el sitio activo de la proteína parece estar preparado para alojar apropiadamente al sustrato cYY pero no al producto, y en consecuencia es posible que el paso de formación del enlace C-C ocurra fuera de la proteína en solución. El hecho de que el sitio activo de MtCYP121 sea incapaz de alojar apropiadamente el producto es consistente con resultados recientes que muestran que el producto mcic no inhibe a la proteína posiblemente debido a una incapacidad de unirse (no se observó unión a 80 μM) [96].

La ocurrencia del paso de formación del enlace C-C en solución tiene varias implicancias interesantes. En primer lugar, de forma similar a lo planteado para cY*Y, podría argumentarse que cY*Y* en solución podría dar lugar a productos secundarios no deseados. Sin embargo es importante notar que dado que ambos radicales de tirosina pertenecen a la misma molécula, están unidos por el anillo de dicetopiperazina que forman los átomos del backbone de ambas tirosinas, la chance de que se forme un enlace intramolecular es mucho más probable que la formación intermolecular de dímeros de cYY, especialmente si la concentración del birradical es pequeña. Otro dato importante es que los radicales de tirosina son bastante estables frente a la oxidación por oxígeno molecular si se los compara con otros residuos, tales como Trp o Cys [109], y en consecuencia pueden vivir lo suficiente en solución como para dar origen al producto deseado. Otra implicancia de la formación del enlace C-C no catalizada está relacionada con el desarrollo de inhibidores selectivo basados en análogos de sustrato y/o producto. Nuestros resultados sugieren que los análogos de sustrato serían mejores candidatos a unirse con MtCYP121 comparados con los análogos de producto, dado que el producto se forma fuera de la matriz proteica.

Finalmente, un tema importante acerca de la presente propuesta mecanística es que involucra varias predicciones que podrían ser confirmadas por medios experimentales. Por ejemplo el agregado de una trampa de espín externa a la enzima en presencia del sustrato podría conducir a la identificación de cY*Y* pero no cY*Y. Además, la mutación de la Arg386 por un residuo neutro (como Gln o Leu), podría retardar la IET conduciendo a la acumulación de cY*Y y/o cambios significativos en la velocidad total. El mismo o un efecto adicional podría conseguirse realizando mutaciones que estabilicen la carga negativa en la tirosina lejana, como por ejemplo Gln385Arg y/o Phe76Arg/Lys.

3.4 Conclusión

Los resultados presentados contribuyen a un entendimiento más profundo de los mecanismos generales de las enzimas CYPs, en particular de aquellas que realizan reacciones de acoplamiento C-C. En este contexto, puede realizarse una comparación interesante con estudios previos de QM/MM de la CYP StaP que cataliza la formación de un enlace intramolecular en el ácido cromopirrólico en un paso clave de la síntesis de estaurosporina [110]. En el trabajo mencionado, los autores fueron capaces de mostrar que dado que el carbono reactivo está demasiado lejos del oxígeno ferrílico del Cp I, el primer radical es formado por el Cp I a través de un proceso de PCET, en el cual un protón del sustrato (derivado de un grupo NH adyacente al carbono reactivo) es transferido al Cp I a través de una red de puentes de hidrógenos mediada por agua, simultáneamente con un proceso de transferencia electrónica. La barrera para este proceso es de 16.3 kcal/mol. El segundo radical en este caso se forma de manera concertada con la formación del enlace C-C y una PCET hacia el Cp II, produciendo una barrera

de 20 kcal/mol. Por consiguiente, en este caso, dado que ambos carbonos reactivos están alejados del grupo ferrilo reactivo, el proceso de PCET es clave, como lo es en el paso de generación del segundo radical presentado aquí para *MtCYP121*. También es interesante el hecho de que en este caso el paso que implica la formación del enlace C-C muestra la barrera más alta. En consecuencia, pareciera ser que para los CYPs involucrados en reacciones de acoplamiento C-C, el Cp I es retenido como el principal oxidante e iniciador del ciclo de reacción y que el Cp II actúa como oxidante secundario, permitiendo que la formación del enlace C-C ocurra en un solo ciclo de la enzima. Sin embargo, dado que en uno o ambos pasos de generación del radical el carbono reactivo puede estar posicionado lejos del hierro (posiblemente para evitar la hidroxilación improductiva), el mecanismo opera a través de un proceso de PCET posiblemente potenciado por la matriz proteica [110].

3.5 Metodología específica del capítulo

La estructura de partida para los cálculos fue la estructura cristalográfica del complejo formado por *MtCYP121* y su sustrato cicloditrosina (PDB ID: 3G5H). El ligando hémico (oxo, peroxo, oxígeno) fue agregado *in silico* para generar el complejo ternario deseado. Dada la importancia de las moléculas de agua en el sitio activo [88], cinco de las mismas que estaban presentes dentro del sitio activo de *MtCYP121* en la estructura cristalina de partida fueron conservadas. Las simulaciones clásicas fueron realizadas utilizando el módulo PMEMD del paquete de programas Amber11 [111], junto a los parámetros del campo de fuerzas ff99SB [48] para los aminoácidos. Los parámetros del grupo hemo en los diferentes estados de coordinación empleados fueron desarrollados y testeados exhaustivamente en trabajos anteriores [112, 113, 114, 115, 116, 117]. Las cargas parciales para los diferentes estados de cYY fueron determinadas utilizando el protocolo sugerido para que sean compatibles con el campo de fuerzas de Amber. Las estructuras electrónicas correspondientes a cada sistema fueron determinadas a nivel HF/6-31G* y las cargas fueron derivadas utilizando el método RESP (Restrained Electrostatic Potential) [118]. Todos los demás parámetros fueron asignados por analogía a los parámetros de tirosina en el campo de fuerzas ff99SB o GAFF. Esta metodología a igual que los parámetros de tirosina radicalaria y tirosinato fueron ampliamente testeados en un trabajo previo [98]. Las estructuras iniciales fueron inmersas en una caja octaédrica preequilibrada rellena con moléculas de agua TIP3P hasta una distancia de 10 Å de la superficie proteica, resultando en un sistema de alrededor de 32.900 átomos. Las simulaciones fueron realizadas a 1 atm de presión y 300 K de temperatura, y mantenidas respectivamente con el barostato y termostato de Berendsen [62]. Se utilizaron condiciones periódicas de contorno y sumas de Ewald (con un espaciado de grillas de 1 Å) para las interacciones electrostáticas de largo alcance; y para calcular las interacciones directas se utilizó una distancia de corte de 10 Å.

3.5.1 Integración termodinámica

Se utilizó el esquema de integración termodinámica (TI) tal como está implementado en el paquete Amber[54] para calcular las diferencias de energía libre de los siguientes pasos: (i) Protonación/Desprotonación de la tirosina [$\text{cY*Y-OH} \rightarrow \text{cY*Y-O}^-$] (Figura 3.6, pasos 1 y 3) y (ii) Transferencia/reorganización de carga en el sustrato monoradicalario negativamente cargado [$\text{cY*Y} \rightarrow \text{cYY}^*$] (Figura 3.6, paso 2). Para cada reacción se simuló once ventanas correspondientes cada una de ellas a un valor distinto de λ : 0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.99. El tiempo de cada simulación de DM fue de al menos 2 ns o hasta lograr la convergencia del valor correspondiente de $\langle dE/d\lambda \rangle$ para cada ventana. Todas las demás condiciones de simulación fueron idénticas a las explicadas anteriormente. Finalmente se obtuvo la energía libre de Gibbs integrando numéricamente $\langle dE/d\lambda \rangle$ respecto a la curva de λ . Para tener una estimación del error de los valores de energía libre obtenidos, se realizaron dos simulaciones independientes (cada una utilizando un conjunto separado de once ventanas). Los valores reportados corresponden al promedio $\pm$ la desviación estándar (DE) de los dos valores obtenidos para un mismo λ en ambos conjuntos de simulaciones.

3.5.2 Métodos cuánticos e híbridos cuántico-clásicos (QM/MM)

Las estructuras de partida para las simulaciones QM/MM fueron tomadas de las correspondientes simulaciones clásicas, previo enfriamiento lento de las fotos seleccionadas a 0K. Los cálculos cuánticos fueron realizados a nivel DFT utilizando el código SIESTA con nuestra implementación propia de QM/MM llamada Hybrid [55]. Se utilizó la aproximación de espín irrestricto en todos los sistemas estudiados.

Los diferentes subsistemas QM incluyeron: el grupo hemo (sin sus cadenas laterales), el ligando hémico (dependiendo de la especie a simular en cada caso) la cadena lateral de la tirosina cercana del sustrato (para la formación del primer y segundo radical), el sustrato completo (para el análisis del paso de transferencia electrónica interna), y el grupo tiol de la cisteína proximal. Para el estudio de la reacción de acoplamiento C-C sólo se trató a nivel cuántico el sustrato. El resto de la proteína y las moléculas de agua fueron tratadas clásicamente.

ANÁLISIS DE LA DIVERSIDAD ESTRUCTURAL Y FUNCIONAL DE LOS p450s

4.1 Introducción

Cómo ya se mencionó en la introducción general, los citocromos de tipo p450 bacterianos poseen una enorme diversidad funcional, y una amplia capacidad para llevar a cabo diferentes reacciones de oxidación sobre sus sustratos, contribuyendo de manera significativa a la generación de una gran variedad de productos naturales, muchos de ellos con un enorme potencial biotecnológico. Determinar el rol de los CYPs en la producción de estos metabolitos es esencial para comprender su biosíntesis y potencial rol biológico en el organismo. Por otro lado, el reciente crecimiento exponencial en la secuenciación de diferentes microorganismos ha resultado en la anotación de más de 15000 secuencias a este grupo de proteínas, anotación que, salvo en casos excepcionales, no va más allá de una asignación a la familia correspondiente. En este contexto predecir, o al menos acotar, *in-silico* el rango de sustratos y reacciones catalizadas por cada una de éstas, es un interesante pero difícil desafío.

Determinar propiedades o aspectos funcionales específicos, como la especificidad/afinidad por cierto sustrato o la eficiencia catalítica de una proteína con precisión y detalle a nivel aminoacídico, no es una tarea sencilla. Sin embargo, si la función proteica está determinada por la estructura de la proteína y ésta por su secuencia de aminoácidos, debería ser posible, en principio, predecir la función de las mismas a partir de estructuras y secuencias proteicas de función conocida. Todo esto sería válido únicamente si se conoce la variabilidad de todas las propiedades relevantes en estas proteínas. En este sentido, considerando que los miembros de la familia de citocromos p450 comparten un mismo dominio de plegamiento y aminoácidos críticos conservados, se espera poder predecir adecuadamente sus posibles funciones.

Con el objetivo de desarrollar un método que tomando como punto de partida la secuencia de un dado CYP y determinando una serie de propiedades físico-químicas, permita predecir el rango de sustratos y la reacción que ésta cataliza, en este capítulo se presenta un estudio evolutivo-molecular de toda la familia de p450s bacterianos que mediante la determinación de diferentes propiedades estructurales -estructura 3D, análisis de residuos del sitio activo, docking molecular- y un análisis del contexto génico, permite el desarrollo de un método predictivo.

4.2 Resultados

Los resultados de este capítulo se organizan de la siguiente manera: en primer término se describe la diversidad de las secuencias de los CYPs para luego analizar que tan bien representado está este universo de secuencias por las estructuras disponibles. En segundo término se presenta un análisis de nuestra capacidad de poder ge-

nerar las estructuras de aquellos CYPs para los que solo poseemos información de secuencia y evaluar el grado de confianza de las estructuras generadas. Luego se procede a analizar la diversidad de sustratos conocidos, sus similitudes químicas y relación entre estas y la distancia filogenética entre los CYPs correspondientes. En cuarto lugar se analiza el contexto génico del CYP en relación con la reacción que cataliza, para finalmente analizar la relación entre la reacción catalizada y la estructura del complejo enzima-sustrato.

4.2.1 Análisis filogenético de las secuencias

Para comenzar el estudio de la diversidad de los p450 bacterianos, se realizó un alineamiento múltiple (AMS) de las 12303 secuencias del conjunto 1*, incorporando información estructural correspondiente a un subconjunto de 23 PDBs que comparten como máximo un 30 % de identidad de secuencia. Para ello se utilizó el programa PROMALS3D (ver sección 4.4). En un paso posterior, se analizó el AMS y se quitaron todas las columnas del alineamiento con alta entropía de secuencias a fines de poder generar un árbol filogenético de los p450s que evidencie las características claves de esta familia. Para esta tarea se utilizó el programa BMGE (Block Mapping Gathering with Entropy, ver sección 4.4).

Se obtuvo así tanto el modelo oculto de Markov como el árbol filogenético correspondiente (Figura 4.1). La construcción del árbol se hizo con el programa RaXML y para la visualización se utilizó Dendroscope e Inkscape (ver sección 4.4). El árbol de la Figura 4.1 muestra la diversidad de p450s bacterianos, en donde pueden identificarse 15 grupos con características distintivas, los cuales se muestran resaltados con colores y etiquetados con números crecientes en sentido horario. Solo dos grupos asignados son parafiléticos, el 6 y el 15, mientras que los trece restantes resultan ser monofiléticos. Hay además algunos pequeños grupos que representan menos del 2 % de las secuencias que no han sido estudiados con mayor detalle.

La obtención de los ASM y HMM correspondientes, permite analizar en el contexto estructural la conservación y varianza de aquellos residuos que conforman el sitio activo de los CYPs y por ende definen la especificidad de sustrato. La Figura 4.2 A y B muestra secciones del logo derivado del HMM (recortadas con el método de BMGE) para todos los CYPs analizados y para cada grupo por separado. El modelo general arroja una longitud óptima del dominio CYP de 400 residuos, de estos unos 32 residuos corresponden a posiciones con alta presión de selección (columnas grises en la Figura 4.2 A y B). Entre los más destacados se encuentran:

- i). el loop de unión al hemo, caracterizado según la literatura por el motivo F-X-X-G-X-X-C-X-G (posiciones 351-360), localizado en la cara proximal del hemo justo antes de la hélice L, el cual contiene a la cisteína que sirve como quinto ligando axial del hierro. Según se observa en el logo, este motivo estaría aún más definido para los p450s bacterianos, ya que la posición 352 queda definida por glicina o serina (un residuo chico) y la posición 356 muestra casi exclusivamente histidina o arginina, transformándose el motivo en F-[G/S]-X-G-X-[H/R]-X-C-X-G. La identidad de la posición 356 se encuentra repartida entre los distintos grupos de p450s bacterianos, mientras que gran parte de los grupos tienen mayormente histidina, los grupos 2 y 15 presentan altamente conservada una arginina, exceptuando al grupo 1 en el cual la conservación de ambos aminoácidos es similar y el grupo 3, que presenta una mayor variabilidad.
- ii). En segundo lugar, en torno a las posiciones 288-291 se evidencia la presencia del motivo EXXR, situado en la hélice K, altamente conservado en todos los grupos a excepción del 10, que posee mayormente triptófano en lugar de arginina y algunas secuencias presentan glutamina en vez de glutámico, como es el caso de la CYP157C1 de *Streptomyces* sp. que presenta el motivo QXXW [119].
- iii). Una tercer subsecuencia consenso altamente conservada en esta familia de proteínas que se observa claramente en el logo, es la definida por el motivo [A/G]-GX-[D/E]-T-[T/S] (posiciones 249-254), el cual está situado sobre la hélice I y constituye el canal de transferencia de protones [120]. En cuanto a este último motivo, se observa muy conservado en todos los grupos a excepción del grupo 3 que evidencia una mayor variabilidad y en particular en las dos posiciones finales (T-[T/S]) si bien treonina sigue dentro de los aminoácidos posibles, son otros los predominantes.
- iv). Finalmente, otra secuencia interesante, es [N/H]-R-D en las posiciones 327-329, porque presenta variantes

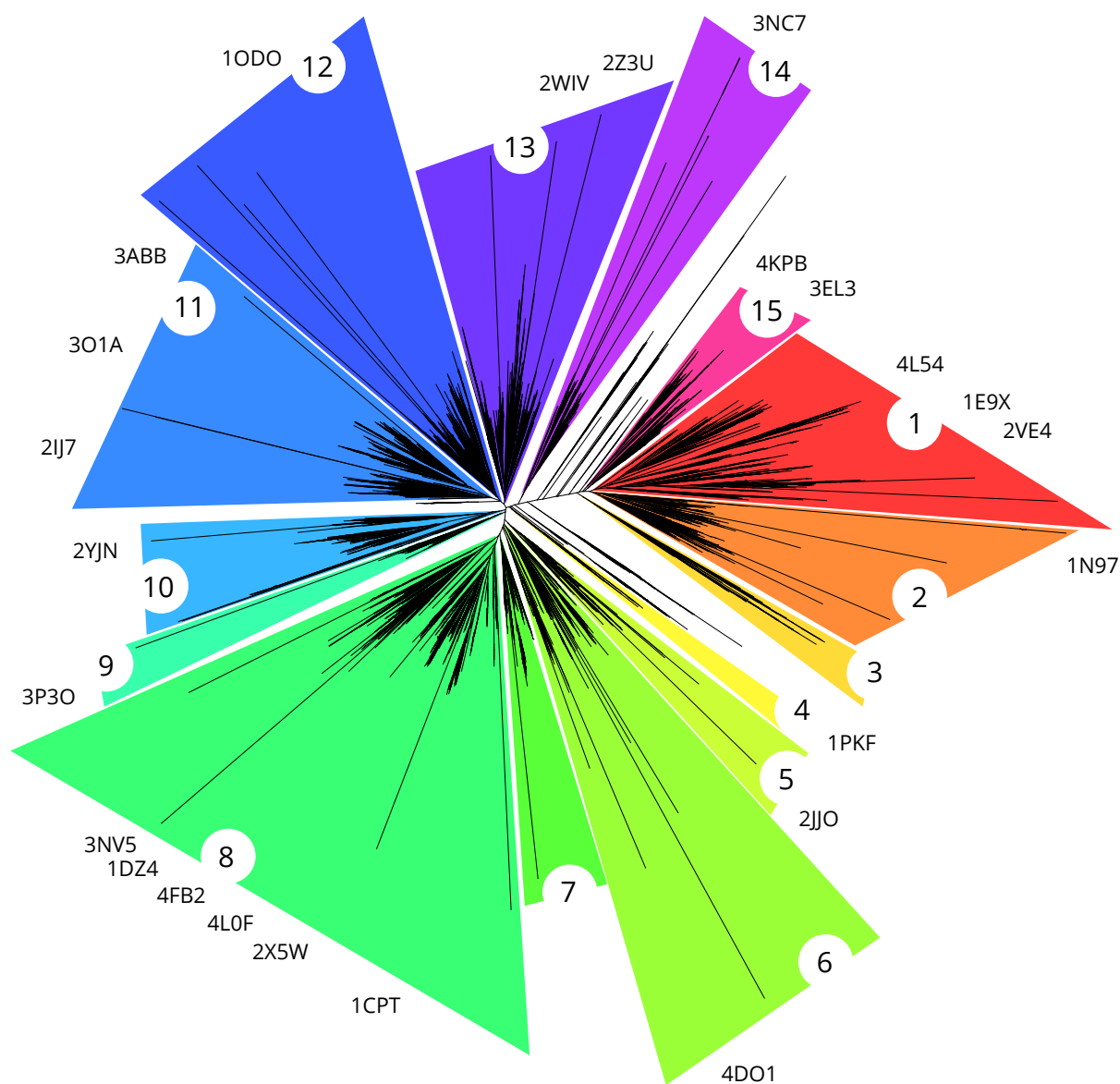
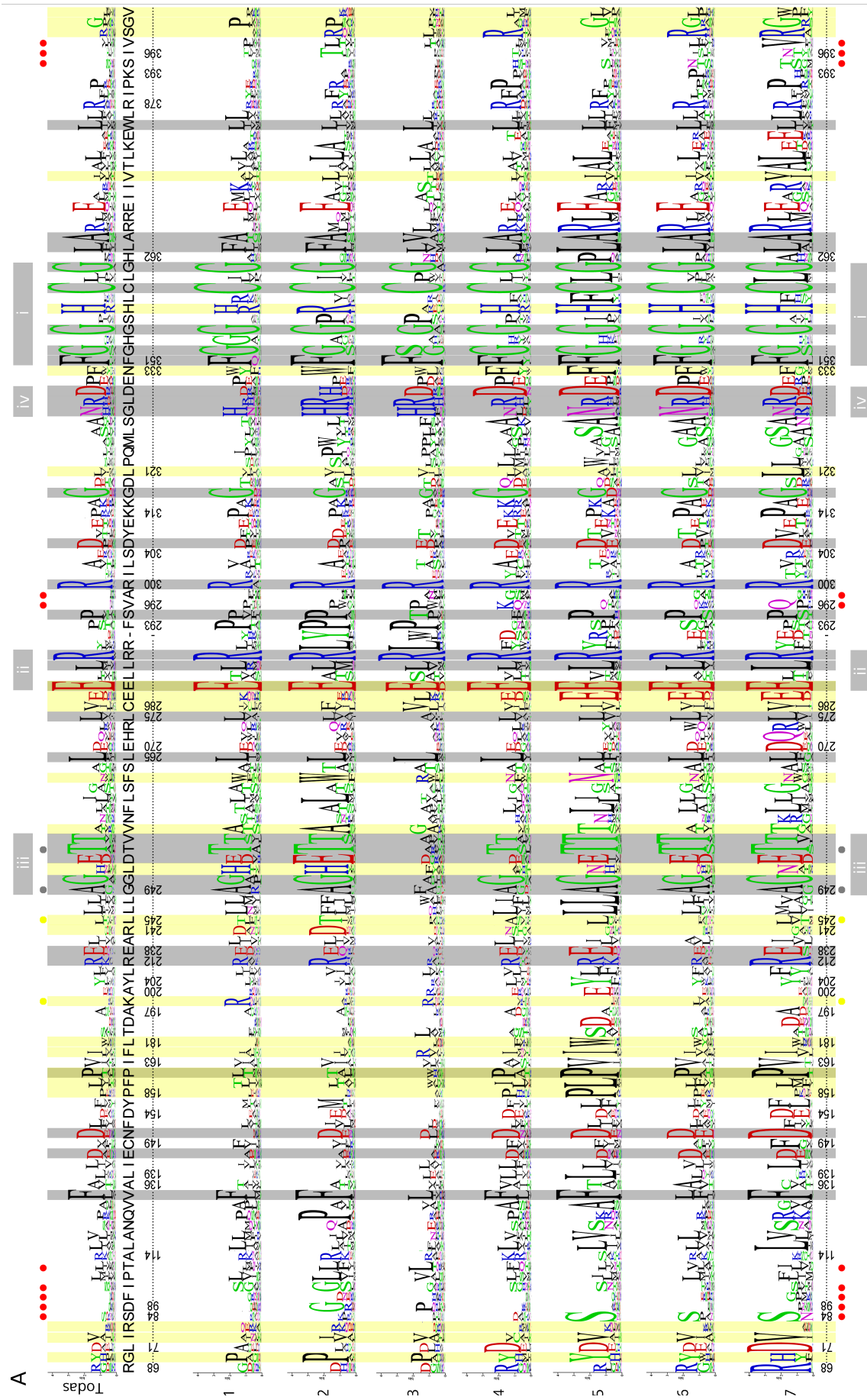


Figura 4.1: Árbol filogenético de todas las secuencias de p450s que componen el conjunto 1* presentado como un filograma radial. Los distintos grupos identificados se muestran señalados con colores y numerados para poder referenciarlos a posteriori. Los códigos de PDBs corresponden a las estructuras que fueron tenidas en cuenta en el AMS.



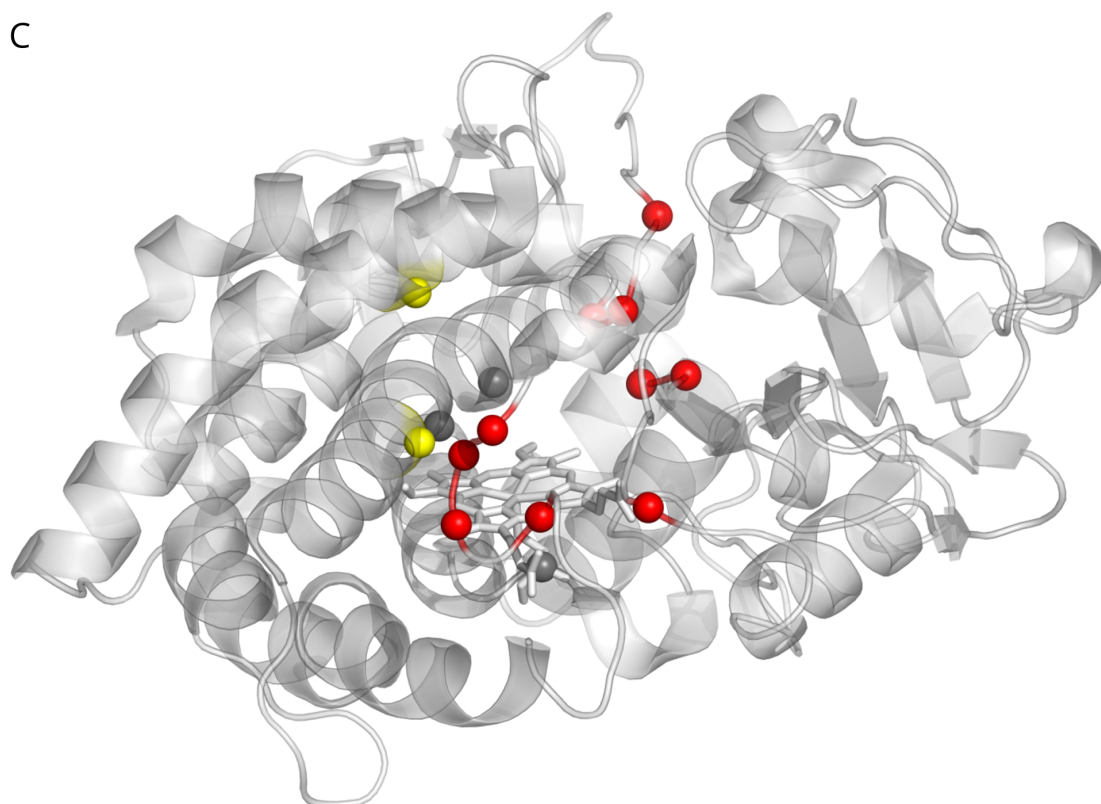


Figura 4.2: A y B. Secuencias logo para todas las secuencias de p450s (primera línea en la parte superior) y para cada grupo por separado. También se muestra una secuencia y un índice, ambos correspondientes a la secuencia de la p450cam (UniprotID P00183), a modo de referencia. Las columnas señaladas con puntos corresponden a posiciones estructurales conocidas en sitios activos de los p450s. Las posiciones resaltadas con gris corresponden a aminoácidos conservados en la gran mayoría de los miembros (ver primer logo); mientras que las resaltadas con amarillo permiten diferenciar a los distintos grupos entre sí. C. Sobre la estructura de p450cam (PDB_ID: 1DZ4) se muestran los residuos que componen el sitio activo (identificados con puntos en el panel A) coloreados de acuerdo al grado de conservación: los grises son los ampliamente conservados, que corresponden a las columnas grises, los amarillos son los que varían entre grupos (columnas amarillas), y los rojos son aquellos que varían mucho (columnas blancas).

diferentes en los grupos del lado derecho del árbol (1, 2, 3 y 15), que tienen bastante conservada una histidina, respecto a los del lado izquierdo del árbol (4, 5, 6, 7, 8, 9, 10, 11, 12, 13 y 14), en los que no solo no hay histidina sino que en varios existe una asparagina muy conservada.

Análisis de conservación de los residuos del sitio activo

El sitio activo –de unión al sustrato– está formado por unos 14 residuos, 5 de los cuáles son altamente importantes. En la Figura 4.2 C se muestra la ubicación de los mismos en su disposición estructural en torno al sitio de unión ubicado sobre el grupo hemo coloreados de acuerdo al grado de conservación. Los grises son los que se encuentran altamente conservados, uno de ellos corresponde a la cisteína coordinada al hemo y los otros dos (G248 y T252) forman parte del motivo [A/G]-GX-[D/E]-T-[T/S] antes mencionado. Los amarillos corresponden a posiciones en las cuales la identidad de los aminoácidos es variable entre los diferentes grupos, por ejemplo en la posición 198 los grupos 1 y 2 presentan preponderantemente arginina, mientras que el grupo 9 presenta triptófano/serina y de forma análoga, en la posición 245, mientras que los grupos 2 y 15 tienen alta presencia de treonina, la mayoría del resto presenta mayoritariamente leucina o glicina. Por último existen nueve posiciones (marcadas en rojo) que evidencian una variabilidad muy grande, como puede observarse en la estructura se encuentran formando parte de loops y esto se condice con que resulten claves para alojar sustratos de tipos y tamaños tan diversos.

Análisis de conservación grupo-específica

El análisis de los logos correspondientes a los 15 grupos, permitió detectar algunas peculiaridades de cada uno de ellos, algunas de las cuales se citan a continuación: se observó que la R198 solo está conservada en el grupo 1; en tanto que D241, A255, W260 y W333 son característicos del grupo 2; la R164 se encuentra parcialmente conservada en el grupo 3; L321 y V397 son distintivos del grupo 7; el residuo E370 es una singularidad del grupo 9; en tanto que el W291 lo es para el grupo 10 y la R73 para el grupo 11. Por último hay un pequeño motivo R398-[G-S]-P400 muy conservado en el grupo 14 y el grupo 15 se distingue por los residuos L71 y T245. Existen además residuos parcialmente conservados, es decir, que son compartidos por varios grupos. Estos resultados se resumen en la Tabla 4.1.

Representatividad de las estructuras de CYPs disponibles

Con el objetivo de avanzar sobre la caracterización estructural, para determinar la representatividad de las estructuras de CYPs disponibles en relación con la diversidad total realizamos un segundo árbol pero sólo con aquellas secuencias representativas que poseen estructura conocida. Dicho árbol muestra la presencia de 14 grupos, es decir que solo un grupo de los 15 (el denominado grupo 3) no posee una estructura que lo caracterice. Para el resto se evidencia que hay al menos una estructura disponible para cada uno. El RMSD de los carbonos alfa dentro de un mismo grupo varía entre 0.3 y 4.5 Å, mientras que entre grupos varía entre 1.189 y 8.387 Å. Esto muestra que incluso dentro de un mismo grupo, existe una considerable diversidad estructural, lo que sugiere nuevamente, al igual que la falta de conservación de los residuos del sitio activo mencionada en el punto anterior, que un mismo grupo evolutivo de CYPs es capaz de unir una amplia variedad de sustratos.

4.2.2 Generación de estructuras por modelado por homología

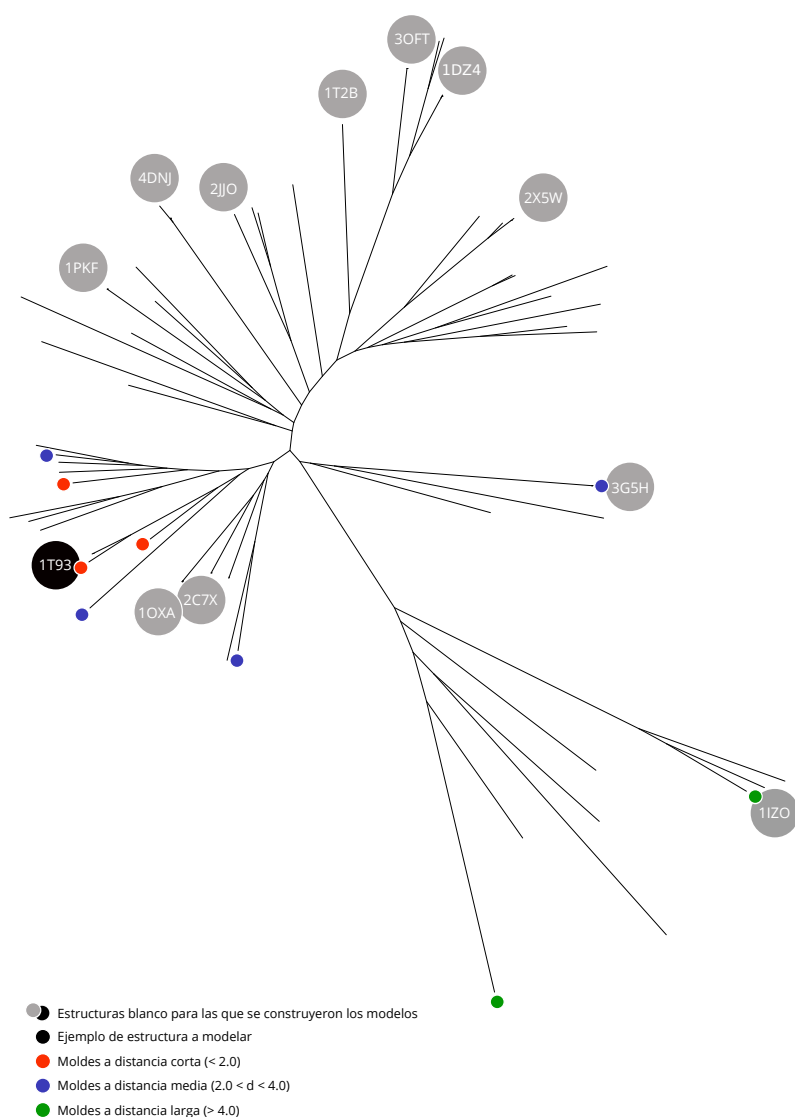
Habiendo caracterizado el conjunto de secuencias, la conservación y diversidad estructural de los CYPs, el siguiente paso es analizar nuestra capacidad de poder generar las estructuras de aquellos CYPs para los que solo poseemos información de secuencia y evaluar el grado de confianza de las estructuras generadas. Para ello, una vez determinado el grupo filogenético al que pertenece una secuencia dada, es preciso encontrar la proteína más cercana que posee información estructural. Esto constituye el primer paso para obtener su estructura modelándola por homología: la elección del molde. Para probar qué tanto influye la elección de la estructura molde en la calidad del modelo obtenido al modelar por homología, se construyeron nueve modelos para cada una de las doce estructuras que conforman el conjunto 3 (definido en la sección 4.4). Dichas estructuras se encuentran

Grupo	Residuos conservados																				
	L71	R73	S84	P158-L-P160	R164	W181	R198	D241	T245	A249-GHET-T254	A255	W260	N260	E370	V286-E-E288	W291	L321	W333	W397	R398-[G-S]-P400	
1										*											
2										*											
3																					
4																					
5			*	*		*							*								
6			*																		
7			*																		
8			*			*															
9																					
10																					
11										*			*								
12				*		*				*			*		*						
13						*				*			*		*						
14			*			*				*			*		*						
15																					

Tabla 4.1: Presencia de residuos que definen los grupos filogenéticos. En negrita y con círculos se indican los residuos de conservación grupo-específica y los asteriscos indican residuos conservados en más de un grupo.

identificadas con círculos de color gris-negro en la Figura 3A. Para cada uno de los modelos se empleó una estructura molde diferente (el protocolo para el modelado se describe en detalle en el capítulo 2). El criterio utilizado para la elección de los moldes fue tomar en cada caso tres proteínas que se hallen a distancia evolutiva corta (es decir, lo más cerca posible en el árbol filogenético, usualmente en el mismo grupo), tres a distancia media y las últimas tres muy distantes (pertenecientes sí o sí a otro grupo). De este modo, se obtuvieron 108 modelos, los cuales fueron clasificados arbitrariamente en tres conjuntos de acuerdo con la distancia evolutiva entre el molde y el blanco: distancia corta ($d < 2.0$), distancia media ($2.0 < d < 4.0$) y distancia larga ($d > 4.0$). (Para una definición precisa de la distancia evolutiva “d” vease la sección 4.4)

A Árbol filogenético del conjunto de p450s para las que se conoce su estructura



B Calidad del modelo según la distancia evolutiva entre la proteína a modelar y el molde

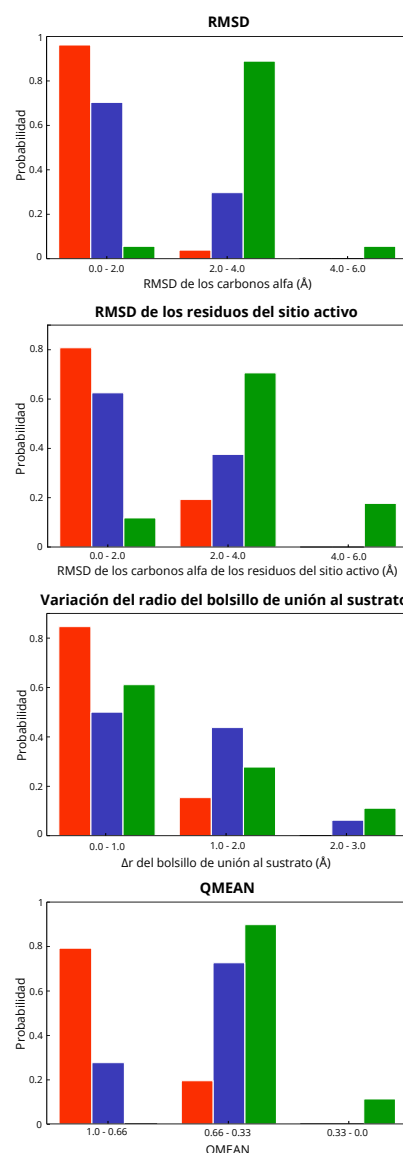


Figura 4.3: A. Árbol filogenético de todas las p450s para las que se conoce su estructura. Los círculos en gris y negro identifican las doce estructuras que se apuntó a recrear a través de los modelos por homología. B. Evaluación de la calidad de los modelos por RMSD, RMSD de los residuos del sitio activo, Δr del bolsillo de unión a sustrato y QMEAN.

A modo de ejemplo, para modelar la estructura indicada en negro en la Figura 4.3 A se utilizaron tres moldes a distancia corta (círculos de color rojo), cuatro a distancia media (indicados en azul) y dos a distancia larga (indicados en verde).

Evaluación de los modelos

Los modelos generados fueron evaluados por cuatro métodos diferentes: i) determinando el RMSD-CA de cada uno de ellos respecto a la estructura cristalográfica real de la proteína modelada (blanco), ii) determinando el RMSD-CA de los residuos del sitio activo, iii) determinando la variación del radio del bolsillo de unión al sustrato y iv) determinando el QMEAN (una función de puntuación que tiene en cuenta los principales aspectos geométricos de las estructuras proteicas, ver sección 4.4 para más detalles). Las distribuciones de probabilidad obtenidas - agrupadas en 3 niveles de calidad del resultado- para cada uno de los 3 conjuntos de modelos (distancia corta, media y larga) se muestran en la Figura 4.3 B. Los datos arrojados por RMSD muestran que si la distancia es corta (rojo) puede determinarse casi la totalidad (> 95 %) de las estructuras con un RMSD < 2Å (lo cual puede considerarse una excelente predicción), para distancia media (azul) este porcentaje se reduce a un 65 % y si el molde se halla muy alejado del blanco la predicción sólo es correcta en 5 % de los casos. En el caso de los resultados obtenidos por QMEAN, se evidencia la misma tendencia, pero aún más enfatizada que en el caso anterior: más del 80 % de los casos que componen el conjunto de distancia corta obtuvieron un valor de QMEAN entre 0.66 y 1 frente al 27 % de los de distancia media. Lo que es interesante, es que si bien la tendencia observada para variación del radio del bolsillo es similar, un número significativamente mayor de estructuras a distancias evolutivas intermedias y largas, son modeladas con tamaños del sitio activo similares (e incluso muy) a los de la estructura de referencia.

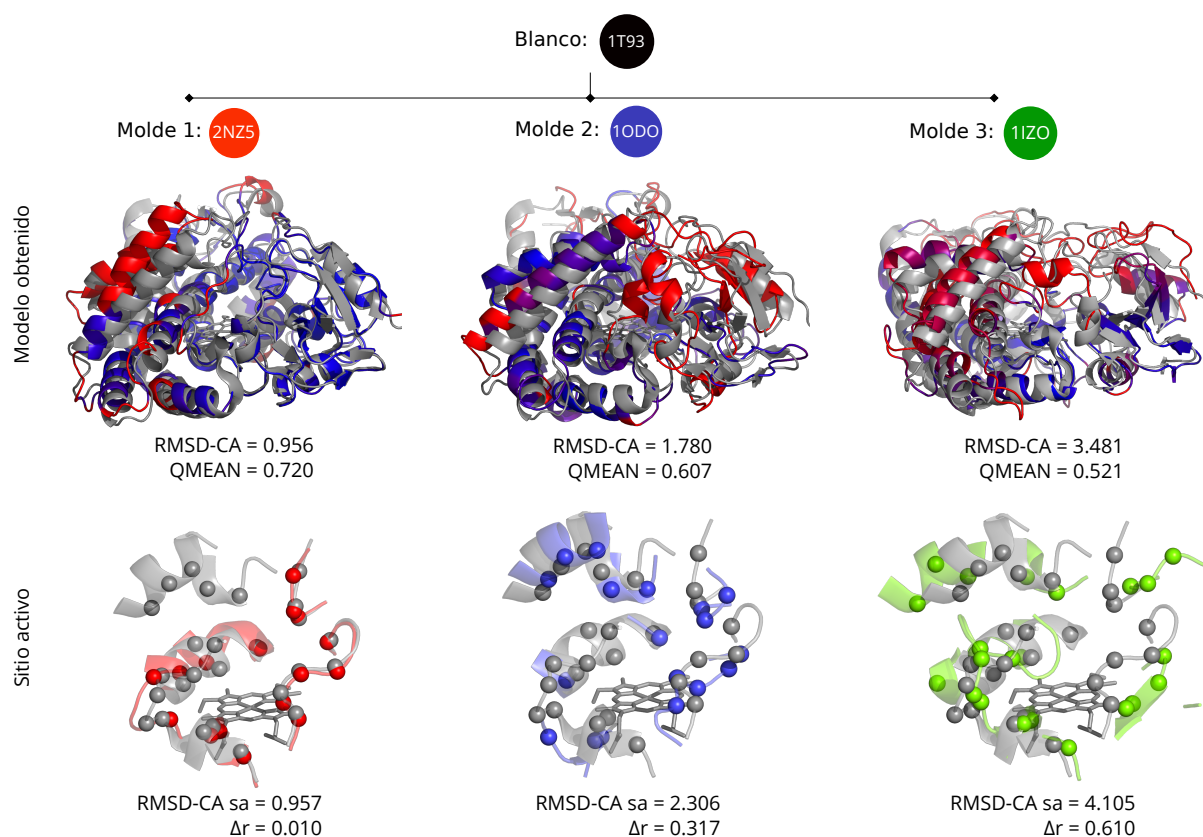


Figura 4.4: Tres ejemplos de modelos para un mismo blanco, cada uno obtenido a partir de un molde correspondiente a un rango de distancia diferente (corta, mediana y larga). En el panel superior se muestra la superposición estructural de la estructura de referencia (gris) y los modelos en escala de color desde rojo, que indica zonas pobremente alineadas, a azul que indica zonas de buena coincidencia entre las estructuras, junto a los valores de RMSD y QMEAN. En el panel inferior se muestran los correspondientes sitios activos, en gris el de referencia y en color el del modelo junto a los valores de RMSD y variación del radio del sitio activo. Las esferas representan los carbonos alfa de los aminoácidos del sitio activo.

A modo de ejemplo, si se observan los resultados obtenidos en el modelado de la estructura de la proteína CYP158A2 de *Streptomyces Coelicolor* (PDB ID: 1T93) partiendo de tres moldes diferentes (uno en cada rango de

distancia evolutiva, PDB IDs: 2NZ5, 1ODO y 1IZO), puede verse la tendencia creciente de los valores de $RMSD_{CA}$ desde 0.96 Å con el molde a distancia corta, pasando a 1.8 Å con el molde a distancia media y 3.5 Å con el molde a distancia larga. La Figura 4.4 muestra en el panel superior la superposición estructural de los modelos obtenidos sobre la estructura de referencia (gris) coloreados por $RMSD_{CA}$. Se observa que particularmente en las zonas de loops (cuarto superior derecho en la representación de la estructura proteica) la diferencia en la calidad de los modelos es muy notoria y decrece cuando la distancia evolutiva entre el blanco y el molde es mayor. En el panel inferior, se muestra la superposición de las estructuras a nivel del sitio activo, nuevamente en gris la estructura de referencia y en color el modelo, en donde las esferas representan los carbonos alfa de los aminoácidos considerados parte del sitio activo. Puede verse que en el modelo de distancia corta (izquierda) las posiciones de los carbonos alfa coinciden en gran medida, traduciéndose en un valor de Δr pequeño (0.01 Å); si nos remitimos al modelo de distancia media (centro) los carbonos alfa no coinciden tan bien ($RMSD_{CASA} > 2$) pero sin embargo el volumen del sitio no presenta una variación importante (0.32 Å); ya en el modelo a distancia larga (derecha), se observa que el $RMSD-CASA$ es bastante elevado (> 4 Å) pero aún así el volumen del sitio activo es similar al de la estructura de referencia.

En resumen, podemos suponer que si hay disponibles estructuras a distancias evolutivas cercanas, los modelos serán de excelente calidad, para distancias evolutivas medias, si bien la calidad general del modelo (QMEAN) puede no ser óptima, tanto el RMSD como el tamaño del sitio activo se espera sean razonables. Finalmente, incluso para modelos generados con distancias evolutivas largas, es altamente probable que el volumen del sitio activo sea predicho correctamente. Habiendo concluido el análisis de las estructuras, pasaremos ahora a analizar la diversidad de sustratos.

4.2.3 Análisis de la diversidad de sustratos de los CYPs

A los fines de mostrar la diversidad química de los sustratos o ligandos de los p450s, se calculó el grado de similitud entre cada par de ligandos utilizando como métrica el índice de Tanimoto como se describe en el Capítulo 2.

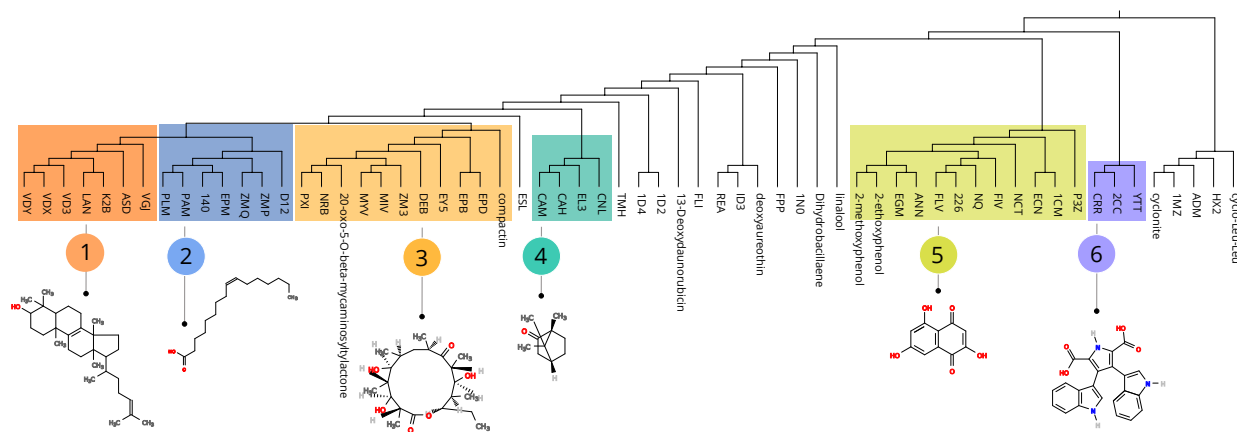


Figura 4.5: Dendrograma de sustratos de los citocromos p450 agrupados de acuerdo a su similitud química. La métrica utilizada para medir la similitud química fue el índice de Tanimoto entre cada par de estructuras. Por cada grupo se muestra una estructura representativa.

Los resultados obtenidos se agruparon jerárquicamente de acuerdo a la distancia entre ellos y se representaron de forma gráfica en una estructura tipo dendrograma (ver Figura 4.5). A partir del mismo se definieron seis grupos, los cuales fueron enumerados del 1 al 6 e identificados cada uno con un color diferente. A los fines de dar una idea del tipo de estructuras que engloba cada grupo, se mapeó sobre el dendrograma una estructura representativa de cada uno. De esta manera se consiguió describir el universo de posibles sustratos conocidos hasta el momento para esta familia de proteínas.

El dendrograma de sustratos de los CYPs (Figura 4.5) muestra que existe una moderada variedad de estructuras de complejos, que alojan sustratos de diverso tamaño y forma. Existen grupos claramente reconocibles como ser: 1) el de los tipo-esteroideo (por ejemplo el lanosterol), 2) el de los ácidos grasos (como el ácido palmitoleico), 3) el de los macrólidos (como la eritromicina), 4) el grupo de los terpenoides (como el alcanfor), 5) un grupo con uno y dos anillos tipo fenólicos, 6) un grupo de ligandos complejos con varios anillos fenólicos conectados por cadenas carbonadas con grupos salientes de tipo ceto o metilo. Asimismo pueden verse varios ligandos que no pertenecen a ninguno de los grupos antes mencionados, entre ellos se encuentra por ejemplo el antibiótico filipin (FLI), cuya estructura posee un anillo enorme de 28 carbonos. Esto es una muestra clara de la gran diversidad química existente entre las moléculas capaces de unirse a los CYPs.

¿Es posible determinar cuál es el ligando analizando sólo la secuencia?

Si se apunta a predecir el (o los) posibles sustratos de un dado CYP partiendo solamente de la información de la secuencia, la primer pregunta que surge es cuál es el grado de correlación entre la similitud de secuencias entre dos CYPs y la similitud de sus sustratos. Dicho de otro modo, si es posible a partir de la similitud de secuencia inferir el sustrato.

Para ello en la Figura 4.6 se muestra el esquema de los sustratos mapeados sobre el árbol filogenético para todos aquellos CYPs con sustrato conocido. La figura manifiesta la amplia diversidad de ligandos y pareciera mostrar que no existe una relación estrecha entre la identidad de secuencia y la similitud de los ligandos, más allá de casos puntuales. A modo de ejemplo, dentro de uno de los grupos monofiléticos (grupo 8) se evidencia la presencia de sustratos que van desde el tipo esteroideo, los similares al alcanfor y hasta ácidos carboxílicos de cadena larga; en tanto que en otro de los grupos (grupo 11) puede verse tanto un compuesto pequeño con dos anillos fenólicos (como la flaviolina) como también compuestos de mayor tamaño como la vitamina D3 y hasta macrólidos (filipin).

Para obtener una medida alternativa, más cualitativa, de la falta de correlación entre la similitud de secuencia y la de sustrato, y que además se centre en los pares cercanos, en la Figura 4.7 se muestra la distribución de probabilidades para la similitud química del sustrato (computada con el índice de Tanimoto) entre pares de CYPs, para todos aquellos pares de proteínas que son vecinas en el árbol filogenético (en naranja). Esto es, para cada CYP su par corresponde a otra CYP cuya similitud en secuencia es la mayor de entre todas las CYPs que componen el conjunto. La distribución claramente se encuentra sesgada a valores de similitud bajos (< 0.5) y solo hay un pequeño 4 % en donde proteínas similares poseen ligandos altamente similares ($IS > 0.9$). Como confirmación de que la distribución de similitud de pares de sustratos entre proteínas vecinas es equivalente a la que se obtendría por azar, calculamos también lo que sería la distribución de fondo (histograma azul) tomando pares de CYPs al azar del árbol filogenético y determinando el IS de sus respectivos sustratos. Como puede verse en la Figura 4.7 la distribución de fondo y la de los pares más cercanos son equivalentes, lo que confirma que no hay correlación significativa entre similitud de secuencia y similitud de sustratos.

Por lo expuesto anteriormente, podemos decir que es imposible utilizar la similitud en secuencia para asignar el sustrato y por lo tanto la función a un dado CYP. En este contexto es que cómo próximo objetivo buscamos analizar la relación entre el contexto génico del CYP en cuestión y sus potenciales sustratos-productos.

4.2.4 Contexto génico

En bacterias, muchas veces, los genes asociados a una misma vía metabólica se encuentran asociados en operones -unidades de expresión conjunta- ubicados secuencialmente en el mismo orden en que sus productos enzimáticos procesan los metabolitos. En otras palabras, es común que el gen río abajo (o río arriba) del gen de interés, sea el que genera el sustrato (o utiliza como su sustrato el producto) del gen de interés. Tomando esto como punto de partida, nuestra hipótesis es que si se conocen los sustratos/productos de los genes vecinos (río arriba/abajo) del gen de un dado CYP particular, estos serían potenciales candidatos a ser sustratos/productos del mismo. Para determinar el contexto génico de los CYPs, y por lo tanto sus vecinos, se evaluaron distintas estrategias. La primera de ellas consistió en tomarlo de la base de datos Gene, del NCBI, la segunda de la base de datos KEGG y la tercera de la base de datos BioCyc.

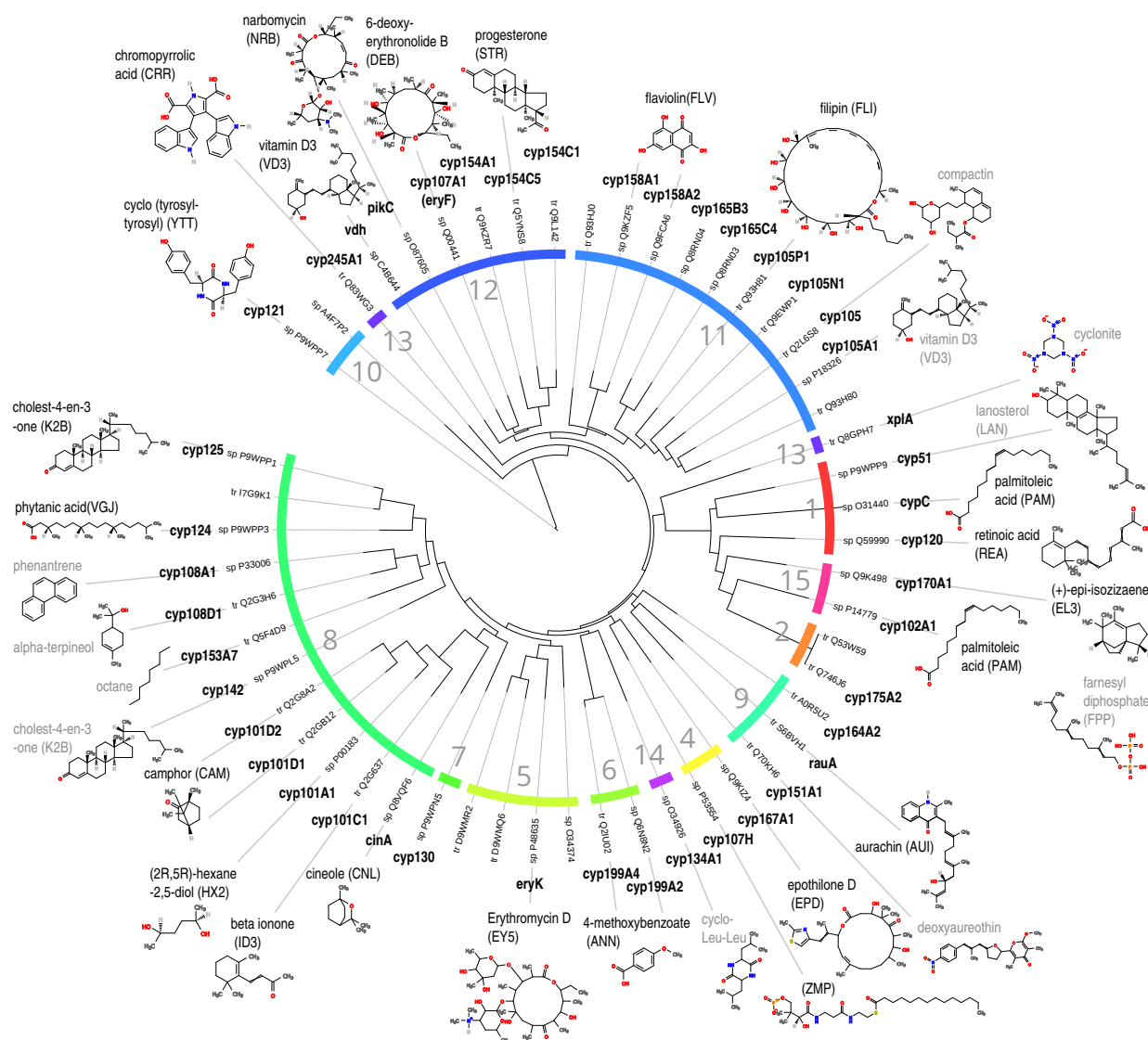


Figura 4.6: Sustratos de los p450s mapeados sobre el árbol filogenético de los que poseen estructura conocida. Los nombres de los sustratos que han sido cristalizados se muestran en negro y los no cristalizados en gris. Como puede observarse, la diversidad de ligandos es muy amplia y la identidad química de los mismos no parece estar determinada por la pertenencia a un clado.

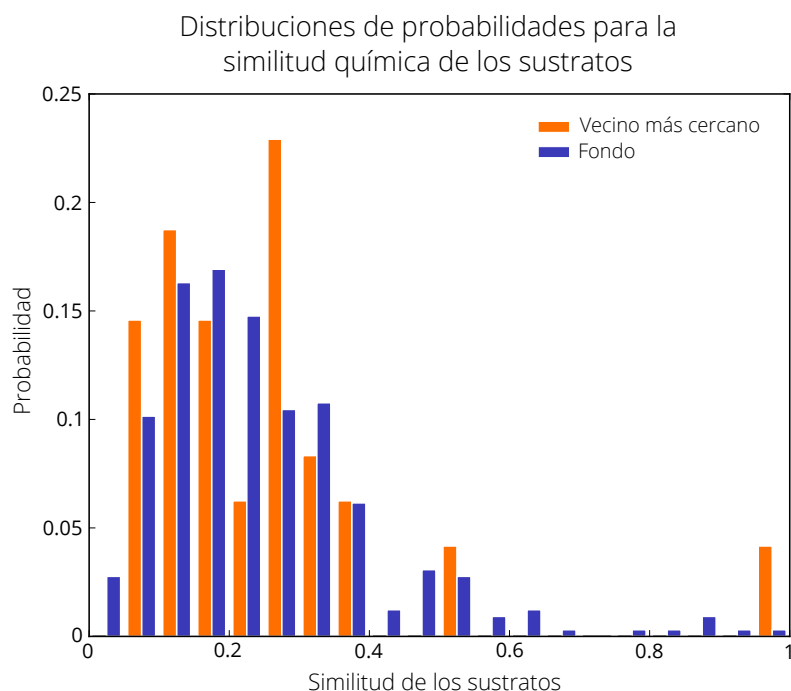


Figura 4.7: Distribución de similitud entre pares de sustratos de proteínas vecinas en el árbol. Para cada proteína de sustrato conocido se evaluó el parecido químico entre su sustrato y el correspondiente a la proteína más cercana (naranja) y se lo comparó con la distribución de similitud de todos los pares de ligandos entre sí (azul).

Dado que cada base de datos tiene una forma diferente de almacenar la información, no es una tarea sencilla desarrollar un protocolo general para acceder a información del mismo tipo, en este caso del contexto génico, almacenada en bases de datos diferentes. Esto se suma a que en muchos casos no está debidamente documentada la forma de acceder a los datos de forma programática. Para decidir cuál estrategia era más promisoría, lo que se hizo fue evaluar para todas las proteínas en el conjunto 1, la existencia o no de los enlaces correspondientes a las bases de datos Gene, KEGG y BioCyc. Esta evaluación mostró que un 29 % de las secuencias del conjunto 1 tienen información en KEGG, un 17 % tiene información en BioCyc y solo un 5 % tiene información en Gene. Esto da la pauta de que la base de datos KEGG tiene información para un mayor número de los CYPs en estudio. Pero lo más interesante es que a su vez, la mayoría de las secuencias con enlace a Gene (76 %) o BioCyc (97 %) tienen también su enlace correspondiente a KEGG (ver Figura 4.8), es decir, de los tres recursos, KEGG es el más abarcativo, y los otros parecieran ser redundantes. Por estos motivos se utilizó KEGG como fuente principal de datos para la determinación del contexto génico y la información asociada.

A pesar de que 4538 CYPs poseen un KEGG_ID, es decir, se encuentran reportados en la base de datos, esto no significa que tengan necesariamente una función asignada, ya que muchas veces simplemente se reporta la presencia del gen para el CYP en el genoma. Por lo tanto, a continuación determinamos cuántos de ellos tienen un EC number o una reacción asignada, lo que significa que conocemos cuál es su sustrato. Como se muestra en la Tabla 4.2, sólo para 414 CYPs tenemos certeza sobre el sustrato. Este conjunto define entonces nuestro control positivo para la predicción de potenciales sustratos a partir de la información del contexto génico. El mismo análisis lo realizamos para los vecinos de cada uno de los CYPs disponibles en KEGG (Tabla 4.2) y cómo puede apreciarse la información es más bien escasa. De los 414 CYPs con sustrato conocido, solamente se conoce la función de uno de los vecinos para 61 CYPs y de ambos para 4. Este pequeño conjunto fue utilizado a continuación para analizar la hipótesis propuesta.

Validación del uso del contexto génico para la predicción de sustrato de los CYPs

Para validar la utilización del contexto génico de los CYPs como método para inferir posibles sustratos, analizamos entonces la relación -utilizando el índice de similitud descrito anteriormente- entre los sustratos/productos reportados para uno o ambos genes vecinos, y el sustrato del CYP a predecir. Vale señalar que en muchos casos

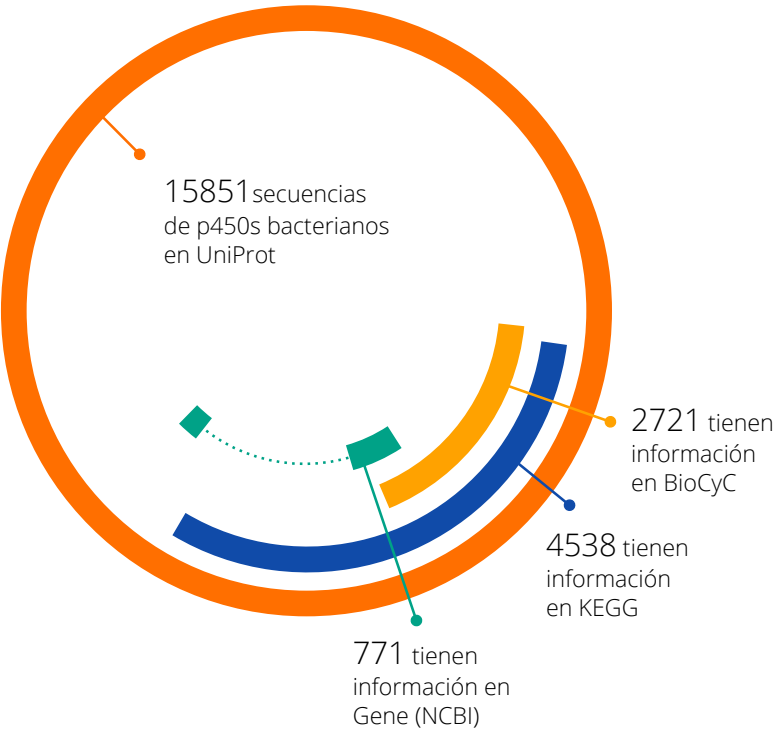


Figura 4.8: Representación de las secuencias de p450s bacterianos en UniProt en las bases de datos KEGG, Gene y BioCyc. Puede observarse que las secuencias que presentan información en KEGG abarcan a la gran mayoría de las que presentan información en Biocyc y Gene.

	CYPs con KEGG_ID	CYPs con sustrato conocido
Total	4538	414
Hay información de la función de solo uno de los genes vecinos	674 (14.9 %)	61(0.15 %)
Hay información de la función de ambos vecinos	41 (0.9 %)	4(0.01 %)

Tabla 4.2: Contenido de información para los genes vecinos de los CYPs en general y para el subconjunto de las cuales se conoce el sustrato.

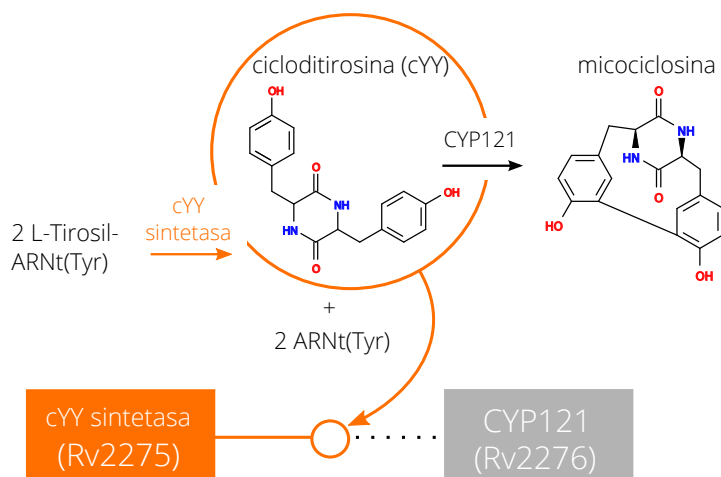
la información obtenida de los sustratos de los vecinos corresponden a compuestos secundarios o metabolitos comunes, es decir, que no son el reactivo principal de la reacción que nos interesa, por ejemplo H₂O, coA, NAD, H₂O₂, H₂PO₄, SH₂, H⁺, etc. Estos casos fueron excluidos del análisis.

El análisis global de cada uno de los 61 casos donde hay información disponible del contexto génico, muestra que para 21 CYPs el sustrato correcto es uno de los sustratos/productos de sus genes vecinos. A modo de ejemplo, en la Figura 4.9 se muestran dos casos en los que los sustratos de CYPs han sido predichos exitosamente a través del análisis de los productos del gen anterior. En el primer caso corresponde al sustrato de la CYP121 de *Mycobacterium tuberculosis* (cicloditirosina) -estudiado en el capítulo 3 esta tesis-, el mismo, es directamente el producto de la enzima ubicada río arriba en el genoma, en este caso la cicloditirosina sintetasa. De forma análoga, en el segundo caso, el sustrato de la CYP170A1 de *Streptomyces coelicolor* ((+)epi-isozizaeno) es sintetizado por la enzima epi-isozizaeno sintetasa codificada por el gen SCO5222 también ubicado río arriba. Si bien estos casos representan casos ideales, su objetivo es demostrar conceptualmente la estrategia. En el capítulo 6, mostraremos cómo, incluso cuando la información obtenida por los sustratos/productos de los genes vecinos no corresponde directamente al sustrato del CYP, es posible utilizarla en combinación con otros enfoques que

abordaremos en esta tesis para determinar o acotar el sustrato.

Predicción de sustrato por contexto génico

1. CYP121 (Micociclosina sintetasa)



2. CYP170A1 (Epi-isozizaeno 5-monooxigenasa)

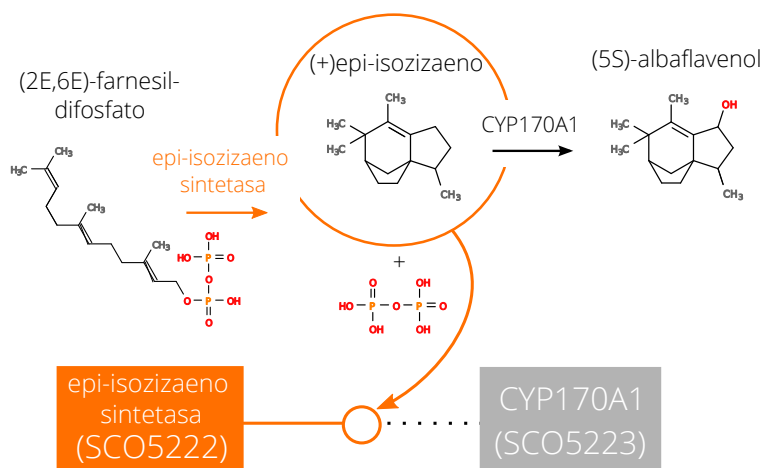


Figura 4.9: Dos ejemplos de sustratos de CYPs predichos exitosamente a través del análisis de los productos del gen anterior. En el caso 1, el sustrato de la CYP121 de *Mycobacterium tuberculosis* (cicloditirosina), es producido por la enzima cicloditirosina sintetasa, codificada por el gen inmediato anterior al que codifica al CYP. De forma análoga, en el segundo caso, el sustrato de la CYP170A1 de *Streptomyces coelicolor* ((+)epi-isozizaeno) es sintetizado por la enzima epi-isozizaeno sintetasa codificada por el gen SCO5222.

Los resultados presentados en el párrafo anterior, muestran que el análisis del contexto genico tiene un enorme potencial para determinar o contribuir a la determinación del sustrato de un dado CYP. Sin embargo, no es posible tener una mayor estadística del potencial de la estrategia dada la limitada cantidad de datos de contexto génico disponibles. Si bien esto constituye una barrera hoy en día, en la medida que se vayan secuenciando y anotando más genomas, o se mejore la anotación de los existentes, la estrategia propuesta se espera incremente su alcance y potencial.

4.2.5 Reacción

Finalmente, el último punto que analizaremos es aquel directamente relacionado con el capítulo 5, que tiene que ver específicamente con el tipo de reacción realizada por el CYP una vez que el sustrato ya se encuentra alojado en el sitio activo. Como ya se mencionó en la introducción, los CYPs catalizan una amplia variedad de

tipos de reacciones: oxidación, hidroxilación, epoxidación, desalquilación, entre otras y en todos los casos es clave determinar cuál es el (o los) átomos del sustrato que se encuentran a distancia de reacción por parte del oxo ferrilo, y el grupo funcional de los mismos y sus vecinos. En este contexto decidimos analizar para aquellos CYPs cuya reacción y complejo enzima-sustrato es conocida, diferentes parámetros estructurales relacionados con el mecanismo de reacción.

Partiendo de las estructuras disponibles, se analizó en cada caso la distancia de cada uno de los carbonos del sustrato al átomo de hierro hémico. Se concluyó que en el 50 % de los casos el átomo reactivo es el más cercano al hierro, en otro 25 % de los casos es el carbono adyacente al más cercano y en el 25 % de los casos restante no es ninguno de los anteriores. Por otra parte, se analizó la distribución de distancias al carbono reactivo en comparación con la distribución de distancias al carbono más cercano (Figura 4.10), observándose que los carbonos reactivos, están en su gran mayoría a menos de 4 Å.

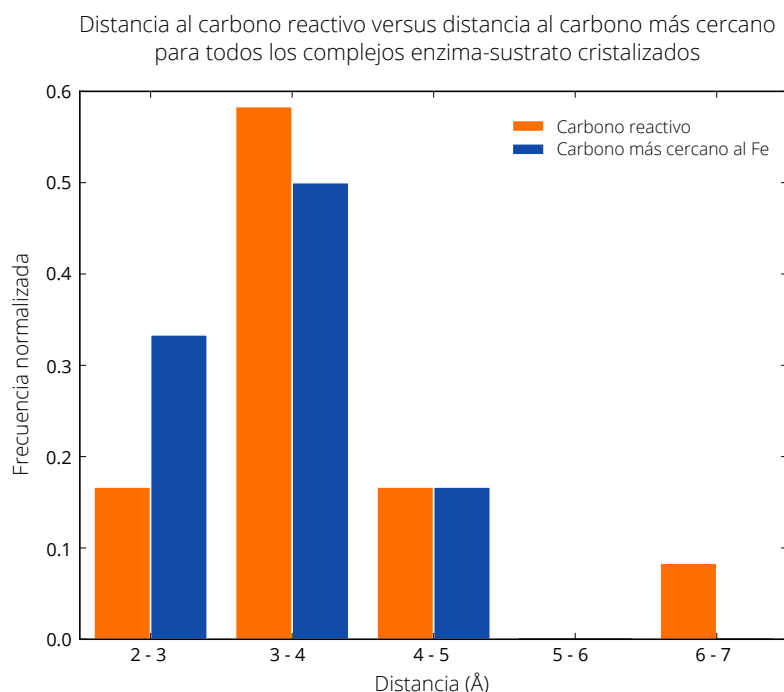


Figura 4.10: Distribución de distancias del hierro al carbono reactivo (naranja) en comparación con la distribución de distancias del hierro al carbono más cercano (azul).

4.3 Conclusiones

El análisis de la diversidad de secuencias y estructuras de los p450s muestra que existe una buena variedad de estructuras disponibles, es decir que la diversidad de secuencias está bastante bien representada por las estructuras, lo cual se traduce en una buena capacidad para obtener las estructuras faltantes modelándolas por homología.

En cuanto a la posibilidad de predecir posibles sustratos de un dado CYP partiendo solamente de la información de su secuencia, el análisis de la diversidad química de los ligandos en relación con la distancia evolutiva entre las proteínas indica que no hay correlación significativa entre similitud de secuencia y similitud de sustratos, por lo que resulta imposible utilizar la similitud en secuencia como único criterio para asignar el sustrato, y por lo tanto la función a un dado CYP.

Por su parte, la utilización de la información del contexto génico como método para inferir posibles sustratos, si bien en la actualidad aún posee aplicabilidad limitada debido a la carencia de anotación de muchas secuencias,

promete ser de gran ayuda en la medida que se vayan anotando más secuencias y se incorpore información a las bases de datos.

Finalmente, respecto a la inferencia del tipo de reacción que podría catalizar un CYP en función del modo de unión enzima-sustrato, el análisis de la estructura de los complejos disponibles cuya actividad catalítica es conocida muestra que en la gran mayoría de los casos el átomo reactivo -usualmente un carbono- está entre los más cercanos al hierro y típicamente a una distancia menor a 4 Å. Estos datos refuerzan la importancia de poder predecir el modo de unión enzima-sustrato y a partir de ello determinar cuáles serían los átomos candidatos a estar involucrados en la reacción, ya que con esta información en mano, el abanico de posibilidades se acota muchísimo, de modo que en algunos casos solo haría falta un poco de intuición química para formular hipótesis que permitan echar luz a la reactividad. Con esta idea es que en el próximo capítulo nos proponemos evaluar nuestra capacidad para predecir el modo de interacción enzima-sustrato y si es posible dada una enzima encontrar cuál es el sustrato correcto entre una serie de sustratos potenciales.

4.4 Métodos específicos del capítulo

4.4.1 Obtención de los conjuntos de secuencias de partida

El conjunto de secuencias de citocromos p450s bacterianos se obtuvo de la base de datos UniProt, utilizando la siguiente clave de búsqueda: "name:"cytochrome p450" taxonomy:"Bacteria [2]"". Esto arrojó un total de 15851 secuencias. De aquí en adelante llamaremos a este conjunto: conjunto 1. Del conjunto 1 se seleccionaron aquellas secuencias que tienen al menos una estructura en la base de datos PDB, originándose, de este modo, el conjunto 2, compuesto por todas las secuencias de p450s bacterianos que tienen estructura tridimensional conocida (53). Finalmente, el último conjunto, denominado conjunto 3, es un subconjunto del conjunto 2 compuesto por doce estructuras para las cuales se conoce la estructura del complejo enzima-sustrato natural, así como también la reacción catalizada.

4.4.2 Filtrado para seleccionar solo las secuencias representativas

Para el conjunto de secuencias denominado conjunto 1 es preciso considerar solo las más representativas a los fines de conseguir un conjunto lo más acotado posible que conserve la mayor variedad posible de secuencias y a la vez reducir, al menos parcialmente, el sesgo hacia las secuencias más estudiadas, eliminando de esta manera la enorme redundancia. Para conseguir esto se utilizó la técnica de agrupamiento por porcentaje de identidad de secuencia, la cual consiste en establecer un máximo de identidad de secuencia permitido y agrupar aquellas cuya identidad supere un umbral determinado. Esto permite reducir el número de secuencias del conjunto considerando un solo ejemplar entre aquellos que compartan mayor identidad que el corte. A los fines de evaluar cual sería un porcentaje (umbral) de corte apropiado se evaluaron valores de 99 % y luego de 95 % a 45 % en intervalos de 5 %. Para esto se utilizó el programa CD-HIT[121]. Con el mismo se realizó un análisis de cómo se reduce el número de secuencias representativas a medida que disminuye el porcentaje de identidad de secuencia máximo entre cualquier par del conjunto que se establece como corte. Los resultados indican que un *cut-off* de 99 % de identidad de secuencia sería el más apropiado, dado que permite reducir considerablemente el tamaño conjunto de datos sin perder información significativa. El nuevo conjunto de secuencias generado (conjunto 1*) posee 12303 ejemplares, lo que representa una reducción del 22 % del volumen de datos a tratar.

4.4.3 Alineamiento de secuencias teniendo en cuenta la información estructural

Partiendo del conjunto de secuencias (conjunto 1*) y el subconjunto que posee estructura, se realizó un alineamiento de secuencias para cada conjunto, incluyendo información estructural. Para ello se utilizó el programa PROMALS3D[122] Dado que el número de estructuras que el programa soporta es limitado (30) se decidió utilizar el conjunto más diverso en cuanto a secuencia que a su vez incluyera la mayor cantidad de estructuras posible

dentro del límite establecido por el programa. Tal conjunto resultó ser el que comparte como máximo un 30 % de identidad de secuencia y está integrado por 23 estructuras.

Análisis del alineamiento

Dado que las secuencias poseen regiones muy divergentes, el alineamiento múltiple de secuencias (AMS) contiene bloques de subsecuencias pobremente alineadas. Estos bloques se removieron con el algoritmo y programa Block Mapping and Gathering with Entropy (BMGE), el cual permite seleccionar bloques del alineamiento que son adecuados para una inferencia filogenética apropiada. El recorte para la filogenia se realizó con una matriz de sustitución Blosum62, una frecuencia de gaps a 0.2 y entropía a 0.9, resultando en un recorte de 356 a 143 columnas. El AMS recortado se usó para construir una filogenia por máxima verosimilitud (ML, del inglés, *maximum likelihood*).

4.4.4 Construcción de los árboles filogenéticos

Para construir el árbol del conjunto 2 se utilizó el programa PhyML 3.0[123], partiéndose de un modelo de sustitución JTT, un árbol de inicio de BIONJ[124] optimizado y con medidas de soporte SH-aLRT [125]. Por otra parte, para el árbol de todas las secuencias (conjunto 1*) se recurrió a otra herramienta, debido a que el tamaño del conjunto de datos excedió la capacidad de procesamiento de PhyML. La herramienta elegida entonces fue RAxML 8.2.4 [126]. Lo mismo ocurrió con el visualizador, mientras que para el primer caso se utilizó el programa iTol [127], para el árbol de todas las secuencias se optó por Dendroscope 3 [128]. Los árboles resultantes se observaron y editaron con Inkscape 0.48.4. A modo de control, puesto que se usaron programas distintos para las filogenias de los conjuntos 1* y 2, se repitió la filogenia del conjunto 2 con RAxML, alcanzándose resultados equivalentes (datos no mostrados).

4.4.5 Construcción de las secuencias logo del alineamiento

Se construyeron perfiles de HMMER para todo el conjunto 1* y también para cada uno de los grupos de secuencias previamente identificados partiendo de la filogenia. Para esto se utilizó el programa hmmbuild con las configuraciones por defecto (HMMER 3.0 [129]). Una vez obtenidos estos perfiles, se generaron los logos de cada uno de los grupos usando el programa WebLogo 3 [130].

4.4.6 Obtención de la distancia evolutiva entre todos los pares de secuencias del árbol filogenético

Para obtener un valor numérico que describa la distancia filogenética de un par específico de secuencias dentro del árbol filogenético, se calculó la distancia copenética, que se define como la altura del dendrograma donde las dos ramas que incluyen a ambos nodos hojas (que representan a las secuencias) se unen en una única rama. Para computar esta distancia se utilizó la función "cophenetic.phylo", que calcula la distancia entre todos los pares de secuencias de un árbol filogenético en base a la longitud de las ramas. Esta función forma parte del paquete de análisis de árboles filogenéticos "ape" (Analyses of Phylogenetics and Evolution) para R [131].

4.4.7 Generación de estructuras por modelado por homología y evaluación de los modelos

En el proceso de predicción de estructura proteica, una vez que se generan los modelos, se hace necesario contar con algún/os método/s que permita/n determinar la calidad de los mismos.

Alineamiento estructural y determinación de los parametros estructurales

El alineamiento estructural entre el modelo obtenido y la estructura conocida suele ser la primera aproximación empleada al evaluar la calidad de un modelo. Como ya se describió en el capítulo Métodos, típicamente los resultados de un alineamiento estructural consisten en una superposición de los conjuntos de coordenadas

atómicas y un valor para la desviación del valor cuadrático medio (RMSD) entre pares de estructuras. Es posible realizar el alineamiento estructural en base a todos los átomos de las proteínas o a un subconjunto.

Cálculo del RMSD de los carbonos alfa

Cuando se alinean estructuras con secuencias muy diferentes, en general no se toman en cuenta los átomos de las cadenas laterales, ya que sus identidades difieren entre los residuos alineados. Por esta razón, si se quiere tener una idea de cuánto ha variado la estructura en general, es común utilizar en el alineamiento estructural sólo los aminoácidos del esqueleto proteico, y aún más, dada la conformación plana del enlace peptídico, por una cuestión de simplicidad y eficiencia frecuentemente sólo se consideran las posiciones de los carbonos alfa. Las estructuras obtenidas por modelado por homología fueron alineadas respecto a la estructura conocida de cada una de las proteínas utilizando el programa PyMOL, como se describió en el capítulo Métodos.

Cálculo del RMSD de los carbonos alfa de los residuos del sitio activo

Dado que en particular nos interesa la capacidad de un dado CYP de unir a su sustrato, y esta depende de la estructura de su sitio activo, es interesante evaluar específicamente la conservación en las posiciones de los aminoácidos que componen dicho sitio. El criterio para definir los residuos que se consideran parte del sitio activo fue incluir todos los aminoácidos que tienen al menos uno de sus átomos a una distancia inferior a 3.5 Å del ligando en el cristal del complejo enzima-sustrato correspondiente. En base a esta selección de átomos se calculó el RMSD de los carbonos alfa de los residuos del sitio activo de manera análoga a lo realizado para la proteína completa.

Cálculo de la variación del radio del bolsillo de unión al sustrato (Δr)

Adicionalmente al cálculo de la variación de las posiciones de los residuos del sitio activo, un dato interesante para evaluar qué tan bien logra el modelo reproducir las dimensiones del bolsillo de unión consiste en determinar la variación en el tamaño de dicha cavidad. Para ello lo que se requiere es tener una medida simple de cómo varía el tamaño de la cavidad del modelo respecto de la estructura de referencia. En este contexto, una variación pequeña en el tamaño de la cavidad se asociaría con una mejor calidad del modelo, y viceversa. Para obtener un parámetro fácil de calcular que describa el tamaño de la cavidad decidimos determinar el radio medio del bolsillo de unión al sustrato (r_p). Para determinar r_p se calcula primero el centro de masa de la cavidad (definida por los carbonos alfa de los residuos que componen el sitio activo), de acuerdo a:

$$r_{CM} = \frac{\sum_i m_i r_i}{\sum_i m_i} = \frac{1}{M} \sum m_i r_i \quad (4.1)$$

Y luego se determina el promedio de las distancias de cada uno de los carbonos alfa del bolsillo de unión a dicho centro de masa

$$r_p = \frac{1}{N} \sum_1^N \sqrt{(x - x_{CM})^2 + (y - y_{CM})^2 + (z - z_{CM})^2} \quad (4.2)$$

En otras palabras el r_p es cómo el momento de inercia del “cuerpo” representado por el bolsillo de unión a sustrato. Una vez obtenido este valor de r_p para cada modelo y para la estructura de referencia correspondiente, se computó la diferencia entre ambos, obteniéndose de este modo la variación del radio promedio del bolsillo de unión (Δr_p).

QMEAN (Qualitative Model Energy Analysis)

QMEAN[77] es una función de puntuación compuesta que describe los principales aspectos geométricos de las estructuras proteicas a través de cinco descriptores estructurales diferentes. La geometría local es analizada por un potencial de ángulo de torsión sobre tres aminoácidos consecutivos. Un potencial específico de la estructura secundaria se utiliza para testear las interacciones de largo alcance. Un potencial de solvatación describe el grado de exposición al solvente de los residuos. Dos términos sencillos describen el grado de coincidencia entre

la estructura secundaria predicha y la calculada y la accesibilidad al solvente. En este trabajo de tesis se utilizó esta función para la evaluación de los modelos, tal como está implementada en el servidor de QMEAN [78].

PREDICCIÓN DE LA ESTRUCTURA DEL COMPLEJO CYP-SUSTRATO

5.1 Introducción

Determinar la estructura del complejo enzima-sustrato constituye un paso clave cuando se desea realizar inferencias sobre la función de una enzima. Dado que en los CYPs la reactividad está asociada con la proximidad al átomo de hierro hémico (como se mostró en el capítulo 4), es fundamental poder determinar la posición que adopta el ligando en el sitio activo de la enzima para luego hacer foco en los átomos que se ubican en la cercanía del hierro. Ésta información es crucial porque al considerar sólo ese conjunto reducido de átomos como candidatos a estar involucrados en la reactividad, el número de reacciones posibles se reduce drásticamente. Por otro lado, si para un caso existen varios sustratos posibles (o potenciales), es razonable suponer que sólo el sustrato correcto encajara correctamente en el sitio activo, y entonces determinar el potencial modo de unión para un par enzima-sustrato permite determinar qué sustratos encajan y cuáles no. En otras palabras, lograr predecir el modo en que el ligando se unirá a la enzima determina en gran parte el éxito de la predicción de su función, porque permite acotar el abanico de posibles reacciones. Cuando se trata de predecir la estructura de un complejo enzima-sustrato, una de las metodologías que más se utilizan son los denominados métodos de docking. En este capítulo nos proponemos por un lado evaluar la capacidad de tales métodos de docking -en particular el implementado en el programa Autodock-, para predecir el modo de interacción enzima-sustrato y adicionalmente -de manera análoga a un ensayo de *virtual screening*- nos proponemos evaluar si es posible dada una enzima encontrar cuál es el sustrato correcto entre una serie de sustratos potenciales. Como variables específicas asociadas a su implementación evaluaremos diferentes parámetros asociados con la estructura del receptor, cómo ser el uso de modelos obtenidos por homología y la flexibilidad del receptor, y aquellos asociados a la función de puntaje relacionados con el uso de sesgo por sitios farmacofóricos.

5.2 Resultados

5.2.1 Evaluación de la capacidad para reproducir la interacción receptor-ligando observada en los cristales

La primer prueba que se realizó consistió en evaluar la capacidad del Autodock4 (el programa elegido) de predecir correctamente el modo de unión del sustrato dentro del receptor en los casos en los que la estructura del complejo es conocida. Este ensayo se realizó a través de un experimento de re-docking sobre los 17 complejos enzima-sustrato representativos que integran el set 4 (ver sección 5.4).

A modo de ejemplo, en la Figura 5.1 se muestran los resultados obtenidos para dos complejos enzima-

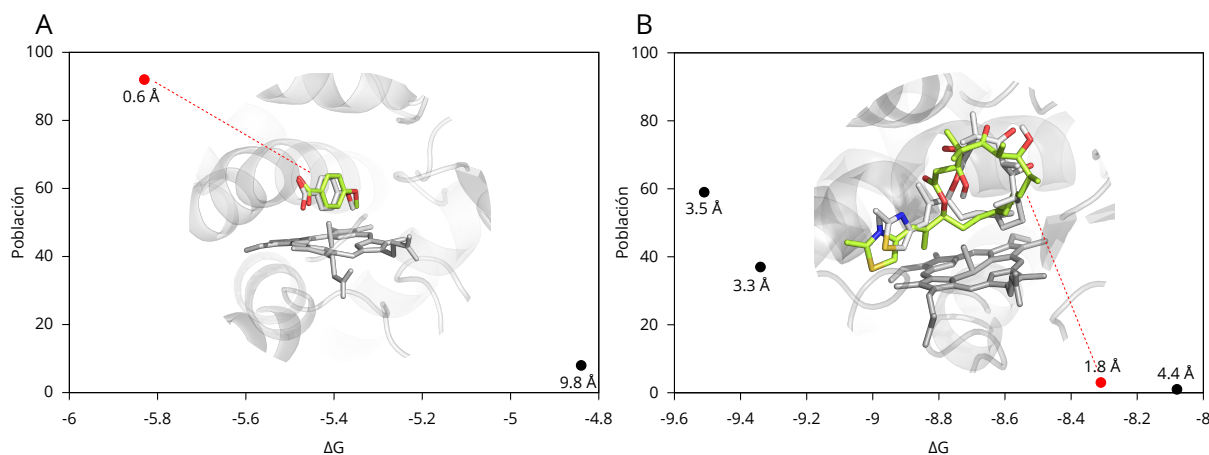


Figura 5.1: A. Gráfica de ΔG vs Población para el experimento de re-docking del complejo CYP199A2 con su ligando 4-metoxibenzoato. B) Gráfica de ΔG vs Población para el experimento de re-docking del complejo p450epoK con su ligando epotilona D. Los puntos correspondientes a la estructura correcta (o más cercana a la del cristal) se muestran en rojo. Las etiquetas sobre los puntos de ambas figuras representan el valor de RMSD entre los átomos de carbono del ligando del complejo predicho por el programa de docking respecto a la estructura cristalográfica del complejo de referencia (PDB IDs: 4DNJ en el caso A y 1PKF en el caso B). Adicionalmente se muestra para cada complejo la estructura predicha superpuesta sobre la estructura de referencia.

sustrato. En el panel A los resultados corresponden al redocking del complejo CYP199A2 con su ligando 4-metoxibenzoato (PDB ID: 4DNJ). Como puede verse el histograma de energía (ΔG) versus población solo muestra dos poses para el ligando dentro del sitio de unión, una de las cuales, es sin dudas la más representativa dado que no solo tiene asociado un menor valor de energía sino que además posee una población de 92 %. La superposición de la estructura cristalográfica de referencia y el resultado del docking y el subsiguiente cálculo del RMSD de los átomos del ligando en ambas estructuras, arrojó un valor de 0.6 Å, lo cual confirma que el docking logra reproducir la pose del ligando en el cristal. Un segundo ejemplo lo presenta el caso del complejo p450epoK con su ligando epotilona D (PDB ID: 1PKF (panel B de la Figura 5.1)). Para el mismo, el resultado del docking que más se acerca a la estructura observada en el cristal (RMSD 1.8 Å) no es el predicho con mayor probabilidad de ocurrencia, sino que se encuentra en una región del gráfico en donde es muy difícil distinguirlo del resto de los complejos. El complejo predicho como el mejor (es decir, el de mayor población y menor energía de unión) presenta un RMSD de 3.5 Å, ya que en el mismo el ligando se encuentra rotado casi 180 grados. Este es un caso en el cual el docking no logra reproducir la pose del ligando en el cristal.

Estos resultados ejemplifican, por un lado, como se analizan los resultados de un docking, y por otro la variabilidad esperada en relación con la capacidad predictiva del método para determinar la estructura correcta del complejo proteína-ligando. Para evaluar entonces la capacidad del Autodock4 de manera conjunta para los resultados para los 17 complejos sobre los que se realizó el re-docking, se utilizó una variante de los gráficos de ΔG vs población mostrados en la figura anterior, que consiste en graficar la diferencia en la energía de unión predicha ($\Delta\Delta G$) y la diferencia en la población (ΔP) entre el mejor complejo (aquel en el cual el ligando está más cerca de la pose observada en el cristal) y el complejo mejor rankeado de los predichos, es decir, aquel que presenta menor valor de energía de unión. En caso de que ambos coincidan, se toma la diferencia con el segundo complejo de menor energía. Para determinar cuál es la pose más cercana a la del cristal, se calculó el RMSD corregido por simetría (ver sección 5.4) del ligando en cada pose respecto a la conformación del cristal. En el caso de ligandos muy flexibles, como K2B y PAM, solo se tuvieron en cuenta los 8 átomos más cercanos al hierro en el cálculo del RMSD, ya que es la parte que más interesa predecir correctamente.

En la Figura 5.2 se muestra el gráfico de $\Delta\Delta G$ vs ΔP obtenido. La ventaja de este tipo de gráfico es que nos permite comparar resultados de diferentes experimentos de docking en simultáneo. Aquí, un valor de $\Delta\Delta G$ menor a cero, implica que el complejo de menor valor de RMSD tiene un valor de energía libre de unión más bajo que el resto de los complejos predichos y la magnitud del valor de $\Delta\Delta G$ refleja la diferencia entre el complejo

mejor predicho y el primer falso positivo. Por otro lado, un valor positivo de $\Delta\Delta G$ significa que el mejor complejo tiene una energía de interacción más alta respecto de aquel rankeado primero, indicando que la mejor predicción es incorrecta. De manera análoga, un valor de ΔP positivo, significa que la mejor predicción tiene un valor de población mayor comparado con los otros complejos predichos, mientras que un valor negativo de ΔP significa que la mejor predicción tiene un valor de población más bajo con respecto a las otras predicciones incorrectas. De este modo, los resultados localizados en la esquina superior izquierda del gráfico mostrado en la Figura 5.2 corresponden a aquellos casos en los que el complejo correcto es rankeado correctamente (es decir, tiene el menor valor de energía de unión y el mayor valor de población).

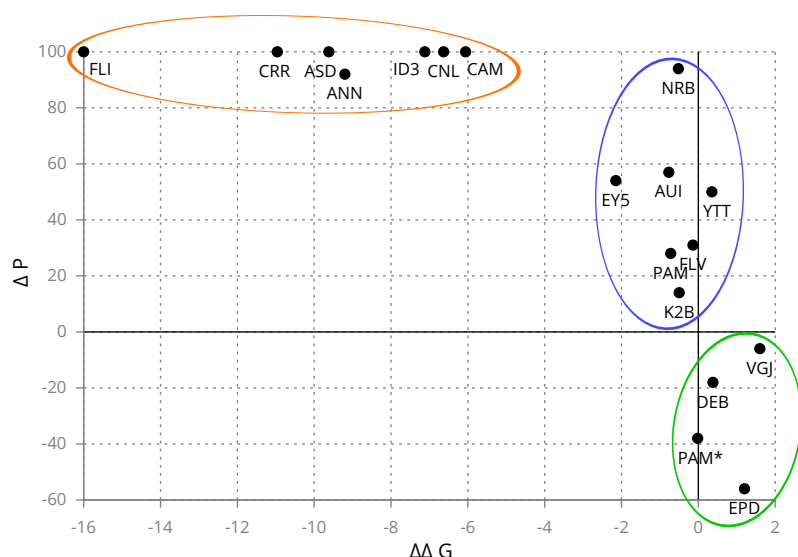


Figura 5.2: Gráfico de $\Delta\Delta G$ vs. ΔP para el redocking de todos los complejos. Ver el texto para la definición de $\Delta\Delta G$ y ΔP .

La Figura 5.2 muestra que los resultados se pueden clasificar en tres grupos (marcados con elipses de color), los primeros dos (naranja y azul) se encuentran en el cuadrante superior izquierdo, el grupo naranja posee valores altos de ΔP y bajos de $\Delta\Delta G$ y corresponde a los casos de éxito claro, o sea donde el docking es capaz, sin duda, de “predecir” la estructura del complejo. El segundo grupo, señalado en color azul, corresponde a aquellos complejos donde si bien el ΔP es positivo, el valor del $\Delta\Delta G$ es muy pequeño por lo que si bien la predicción es correcta, es posible que en un experimento “real” sea difícil distinguir el complejo correcto de falsos positivos. Por último el tercer grupo, marcado en color verde, comprende aquellos complejos donde el Autodock es incapaz de predecir la estructura correcta del complejo proteína-ligando, y existe una estructura con una pose incorrecta cuya energía y población son mejores que las de la pose más cercana al cristal. En resumen, el Autodock posee una capacidad aceptable para predecir la estructura de complejos CYPs-sustrato, si bien falla en algunos, y en otros el resultado no es tan claro. Por ello si bien avanzaremos por un lado con los casos positivos, en el apartado siguiente buscaremos mejorar la performance del docking utilizando la información analizada en el capítulo anterior.

Implementación de docking sesgado

Con la idea de introducir un sesgo que permita mejorar el desempeño del docking en aquellos casos en los que resulta difícil predecir la estructura del complejo, se realizó un análisis de las características compartidas -y diferenciales- si las hubiera, entre los CYPs que unen farmacóforos similares. El objetivo de este estudio consistió en determinar interacciones importantes en la unión de ligandos que comparten un determinado grupo funcional, para luego aplicar un sesgo que ayude al correcto posicionamiento del ligando en el sitio de unión. A tal fin, en primer término se procedió a alinear las estructuras de complejos que unen ligandos de características químicas similares, y luego se analizaron las interacciones que éstos establecen con los aminoácidos circundantes. Un caso interesante es el de los CYPs que unen ácidos grasos, cuyo ejemplo representativo dentro de los

complejos evaluados en la sección anterior, es el de la p450BS unida a ácido palmitoleico (PAM). Como se describió en el capítulo 4, existen CYPs que unen ácidos grasos de cadena larga en diversos grupos filogenéticos de p450s, puntualmente los hay en el grupo 1, 8 y 15. Inclusive un mismo ácido graso puede unirse a CYPs de distintos grupos en una orientación totalmente diferente propiciando reacciones enzimáticas diversas. Tal es el caso del ácido palmitoleico que se une tanto a la p450BS (cyp152B1), responsable de la hidroxilación de ácidos grasos en las posiciones α y β , como también a la p450BM-3 (cyp102A1) que hidroxila ácidos grasos pero en el extremo opuesto, es decir, en las posiciones ω_1 , ω_2 y ω_3 . La pregunta clave aquí es si existe alguna interacción que determine la orientación en la que se unirá el ácido graso.

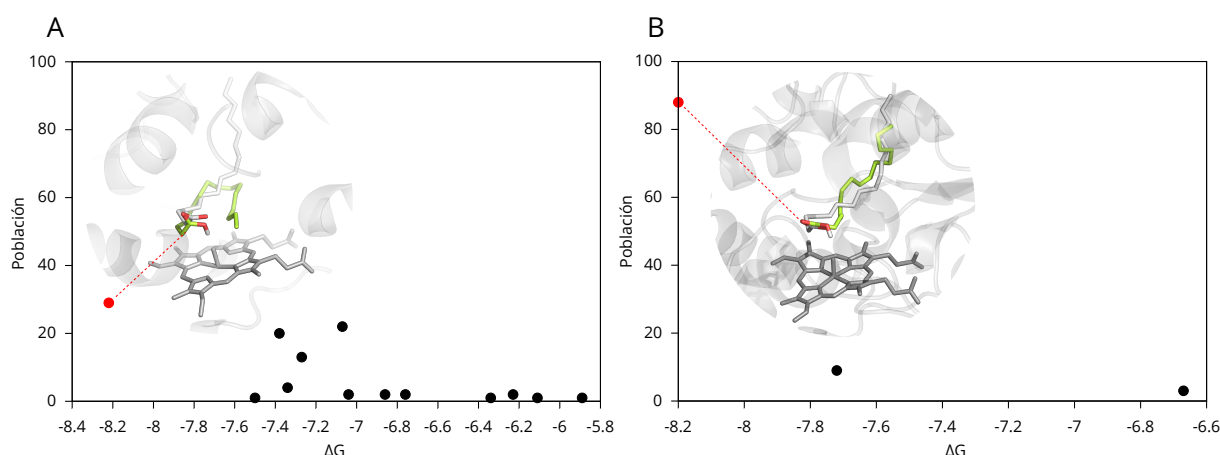


Figura 5.3: Histogramas de ΔG vs. Población para el redocking del ácido palmitoleico (PAM) sobre la p450BS. En el panel A se muestra el resultado del docking tradicional y en el panel B el correspondiente al docking sesgado junto a la estructura mejor rankeada en cada caso (carbonos en verde) superpuesta sobre la estructura del cristal (carbonos en blanco).

El análisis realizado muestra que para el caso de la p450BS, en la que el extremo del carboxilato es el que se ubica en la zona reactiva, existe una interacción de este carboxilato con el grupo guanidinio de la R242[132]. Asimismo, se observó que para la p450BM-3 y todas las demás que unen ácidos grasos en la orientación contraria, es decir, con el carboxilato hacia afuera de la zona reactiva, este residuo de arginina se encuentra reemplazado por un aminoácido hidrofóbico (ILE, VAL, PHE), lo que se condice con que sea la cola hidrofóbica y no el carboxilato la que ocupe el lugar cercano al hemo. Dado que en las tres estructuras que se han podido comparar el carboxilato se encuentra anclado siempre en el mismo lugar al lado de la R242, es que se pensó en aplicar un pozo de potencial en esa coordenada que favorezca el correcto posicionamiento del carboxilato, siempre que el análisis de la secuencia indique que está presente la R242. Este sesgo logró mejorar notoriamente el ΔP , y el posicionamiento del ligando en general, ya que la pose más cercana a la del cristal resultó ser ahora la mejor rankeada aún considerando todos los átomos del ligando en el cálculo del RMSD (Figura 5.3).

5.2.2 Docking sobre modelos por homología

Habiéndose analizado ya la capacidad del docking para predecir la estructura de complejos CYP-sustrato, en esta sección se analizará el desempeño del docking en los casos en que no se conoce la estructura proteica y por ende se hace necesario recurrir a un modelo por homología para representarla.

Para ensayar esto, se seleccionaron seis complejos CYP-sustrato en los que el docking es capaz de reproducir adecuadamente la estructura del cristal y se construyeron nueve modelos para cada receptor, obteniéndose cada uno de ellos a partir de un molde diferente como se describió en detalle en el capítulo 4. Luego se evaluaron los resultados del docking de cada sustrato sobre la serie de nueve modelos del correspondiente receptor. A modo de ejemplo en la Figura 5.4 se muestran tres estructuras predichas para el complejo p450cinA con su ligando eucaliptol (carbonos en verde) superpuestas sobre la estructura cristalográfica del complejo (PDB ID: 1T2B) (carbonos en blanco). En los tres casos se utilizó un modelo diferente del receptor obtenido cada uno de ellos en

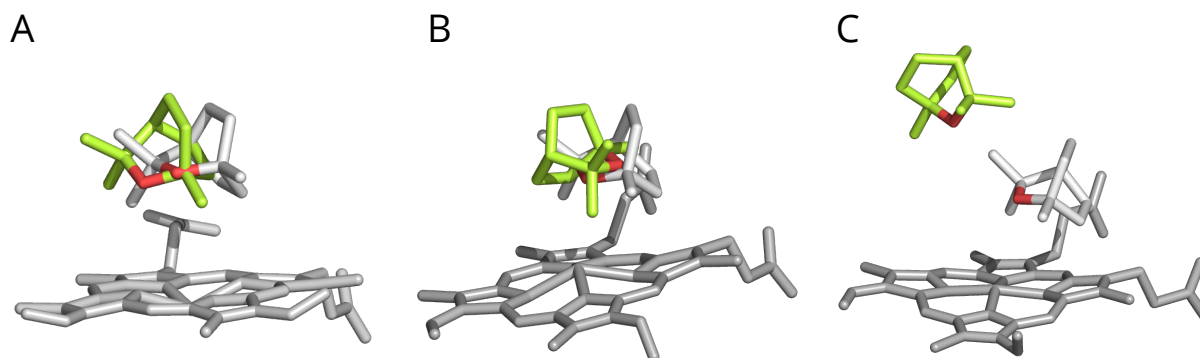


Figura 5.4: Tres estructuras predichas para el complejo p450cinA con su ligando eucaliptol (carbonos en verde) superpuestas sobre la estructura cristalográfica del complejo (PDB ID: 1T2B) (carbonos en blanco). En el panel A se muestra el resultado del docking sobre el modelo generado a partir de la estructura de PDB ID: 3NV5. En el panel B se muestra el complejo resultante para el modelo generado en base a una estructura un poco más alejada evolutivamente de la p450cinA (PDB ID: 3O1A). En el panel C se muestra un caso en el cual el complejo fue predicho erróneamente si bien la distancia evolutiva respecto a la proteína molde (PDB ID: 2Z36) es similar a la del panel A.

base a una proteína a distancia media ($2 < d < 4$) de la p450cinA. En el panel A se muestra el resultado del docking sobre el modelo generado a partir de la estructura de PDB ID: 3NV5, que constituye el resultado más cercano al del cristal ($\text{RMSD}_{\text{ligando}} = 1.4 \text{ \AA}$). En el panel B se muestra el complejo resultante para el modelo generado en base a una estructura un poco más alejada evolutivamente de la p450cinA (PDB ID: 3O1A). Puede verse que si bien la pose del ligando no es tan acertada como la del panel A, el resultado es aún bastante bueno ($\text{RMSD}_{\text{ligando}} = 1.6 \text{ \AA}$). Por último en el panel C se muestra un caso en el cual el complejo fue predicho erróneamente ($\text{RMSD}_{\text{ligando}} = 5.4 \text{ \AA}$) si bien la distancia evolutiva respecto a la proteína molde (PDB ID: 2Z36) es similar a la del panel A.

La figura muestra cómo en el caso de poseer un modelo de buena calidad (Figura 5.4 A) el método es capaz, como era de esperarse, de determinar la estructura del complejo. Este caso representaría un resultado exitoso. Incluso para un modelo de calidad media, como el que se muestra en la Figura 5.4 B, el docking consigue aproximarse a la estructura del complejo. En cambio, si el modelo estructural es de baja calidad, el Autodock4 no es capaz de determinar la estructura correcta del complejo, por lo que el resultado se considera fallido.

El mismo análisis fue realizado para los nueve modelos de cada uno de los seis complejos disponibles. Los resultados muestran para los modelos obtenidos a partir de moldes a distancia corta (< 2) y mediana ($2 < d < 4$), se logra predecir acertadamente la ubicación del ligando en el sitio activo en aproximadamente en la mitad de los casos, aunque varíe la orientación del mismo (Figura 5.4 A). En cambio cuando la distancia evolutiva es grande la gran mayoría ($> 80 \%$) de las predicciones son fallidas. Este resultado es más que aceptable si se tiene en cuenta que se está trabajando sobre estructuras modeladas por homología, lo que implica que no puede esperarse una precisión muy alta en los resultados del docking por la imprecisión asociada a los modelos en sí mismos.

Un análisis más detallado muestra, sin embargo, que existe una marcada diferencia entre los resultados obtenidos para proteínas que unen ligandos “pequeños”, y aquellas que unen ligandos “grandes”, como puede evidenciarse en las distribuciones de resultados exitosos y fallidos para cada una de estas categorías (Figura 5.4 B y C). La tendencia observada indica que para las proteínas que unen sustratos pequeños, como el alcanfor, eucaliptol y metoxibenzoato, existe una probabilidad muy alta o moderadamente alta (0.9, 0.6 y 0.6, respectivamente) de predecir acertadamente el modo de unión del ligando utilizando un modelo generado a partir de un molde a distancia corta o mediana. En tanto que para las proteínas que unen sustratos de gran tamaño, como es el caso de los macrólidos, para conseguir un modelo que reproduzca adecuadamente las dimensiones del sitio activo es un requisito fundamental contar con un molde muy cercano evolutivamente, y aún así el éxito no está garantizado. El origen de esta diferencia, radica posiblemente en el hecho de que al ser tan extensa la superficie de contacto entre el ligando y el receptor, pequeñas diferencias de posición en los loops que conforman el sitio activo, el resultado será una cavidad de menor tamaño en la cual el ligando no entra, lo cual se ve reflejado en el resultado de docking con una energía de interacción muy desfavorable. En función de esta observación, se

planteó como posible alternativa para mejorar los modelos por homología que potencialmente unen ligandos grandes, utilizar como molde estructuras de complejos (enzima-sustrato) para la generación del modelo estructural en vez de estructuras sin ligando. La idea subyacente a esta propuesta, es que la presencia del ligando “grande” en el molde, contribuiría a conservar el tamaño de la cavidad, evitando que ciertas cadenas laterales se ubiquen bloqueando la posición de un posible ligando.

Docking sobre modelos realizados con moldes en presencia de sus ligandos

Para los casos de proteínas que unen ligandos grandes, en los que como se comentó anteriormente el docking en general no arroja buenos resultados, se realizaron entonces modelos por homología en base a moldes cocrystalizados con algún ligando grande en su interior. En el proceso de modelado se tuvo en cuenta el espacio ocupado por el ligando de modo de preservar lo máximo posible las dimensiones de la cavidad.

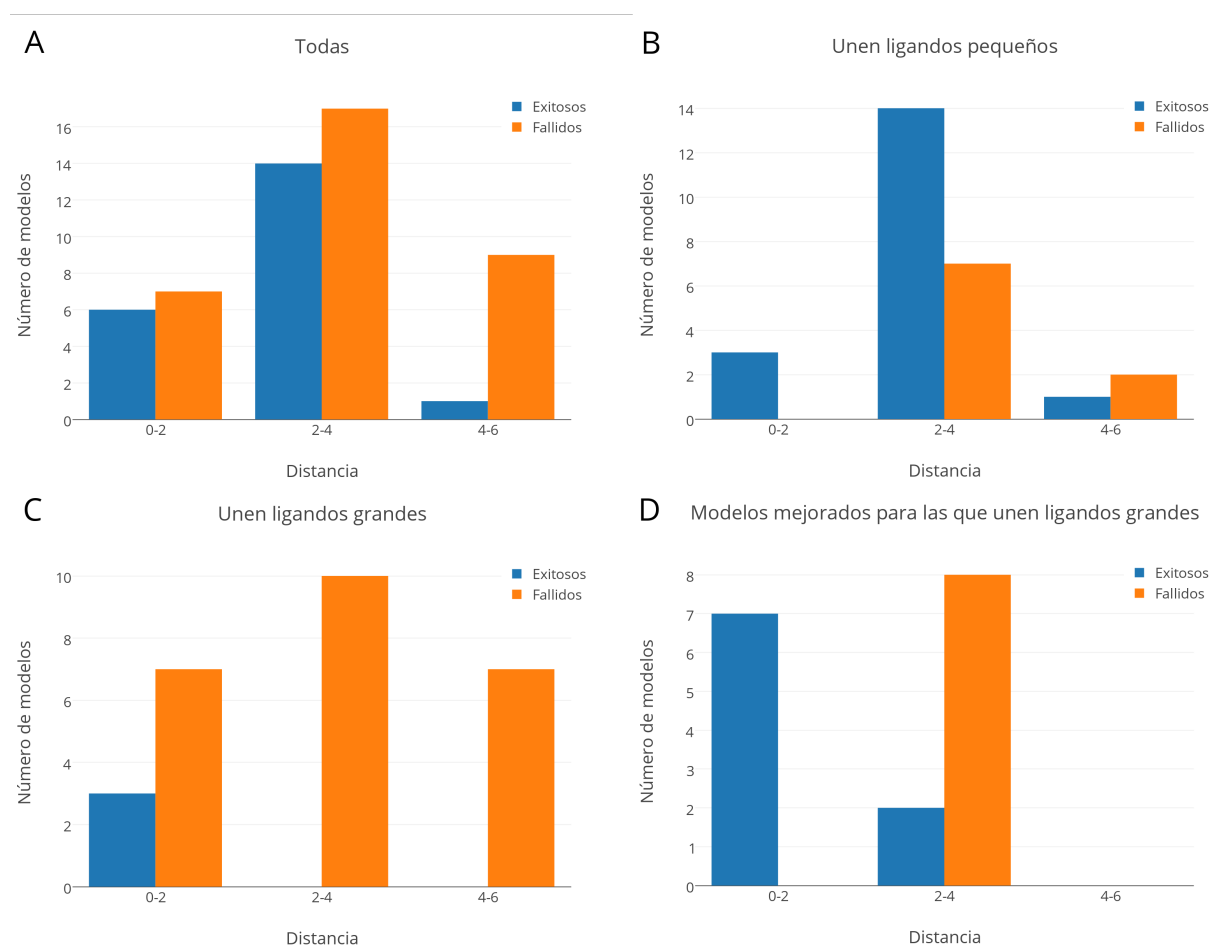


Figura 5.5: Distribución de resultados del docking sobre 54 modelos de seis proteínas según la distancia evolutiva entre el molde utilizado para generar el modelo y la proteína a modelar. En el panel A se muestran los resultados generales clasificados en exitosos (RMSD<3) en color azul, o fallidos (RMSD>3) en color rojo. En el panel B se muestran los resultados correspondientes a las tres proteínas del conjunto que unen sustratos pequeños. En el panel C se muestra lo análogo para las tres que unen sustratos de mayor tamaño. En el panel D se observan los resultados del docking de ligandos de gran tamaño sobre los modelos mejorados.

Los resultados que se resumen en la Figura 5.5 D muestran que esta modificación mejora de manera significativa el desempeño del docking de ligandos de gran tamaño. No solo se observó en todos los casos que la energía de interacción mejora sustancialmente, sino que también se consiguió predecir acertadamente el modo de unión en el 70 % de los casos para uno de los complejos evaluados (el de la p450 PikC unida a narbomicina, PDB ID: 2C7X). En el caso del otro complejo evaluado (correspondiente a la p450 eryK unida a eritromicina, PDB ID: 2JJO) si bien no se logró la predicción del modo de unión, la energía de interacción sí tomó valores más razonables.

Vale aclarar que una dificultad adicional para este último caso es que no hay disponible ninguna estructura a distancia corta cocrystalizada con un ligando de tamaño similar, por lo cual todos los modelos se hicieron partiendo de estructuras a mediana distancia. De todos modos, cabe resaltar que se trata de ligandos que cuentan respectivamente con 38 y 49 átomos (sin contar hidrógenos), por lo cual el desafío de encontrar el modo de unión no es para nada sencillo.

En conjunto los resultados para el docking ya sea sobre estructuras cristalográficas o modelos por homología muestra que la performance del Autodock es adecuada pudiendo determinar el modelo de unión correcto en un número significativo de casos. Esto se traduce en una capacidad de predicción equivalente en modelos de calidad alta o media, particularmente para ligandos de tamaño chico o medio. Por otro lado los resultados también muestran que el análisis detallado de las interacciones y residuos conservados, permite en ciertos casos particulares mejorar los resultados de manera significativa mediante el uso de docking sesgado. Finalmente, también es posible mejorar la performance del docking de ligandos de mayor tamaño para modelos por homología cuando los mismos son contruidos a partir de estructuras con ligando. Habiendo entonces analizado la performance del docking en relación con el tipo de ligando y la calidad de la estructura, procederemos ahora a analizar la capacidad del mismo de seleccionar el ligando correcto (o el tipo de ligando) del total de ligandos posibles.

5.2.3 Selectividad

Cómo primer abordaje para determinar la capacidad del método de docking para seleccionar el ligando correcto, se seleccionó un complejo representativo por cada tipo de ligando definido en el capítulo 4 y se realizó un primer análisis cruzado entre todos ellos.

Este análisis consistió en evaluar el resultado del docking de cada uno de estos ligandos representativos (cuyas estructuras se muestran en la Figura 5.6 A junto a sus PDB IDs) sobre cada uno de los receptores y luego analizarlos conjuntamente mediante la superposición de los histogramas de ΔG vs P obtenidos. Los resultados se muestran en la parte inferior de la Figura 5.6, señalándose en color rojo los resultados que corresponden al ligando correcto en cada caso. En el panel B, a modo de ejemplo, se presentan las gráficas de ΔG vs P para el docking de los seis ligandos sobre la estructura de la CYP154C5 (PDB ID: 4JBT). Como puede verse, hay un punto rojo arriba a la izquierda, que es el que corresponde al ligando correcto (androstenediona, ASD), que en este caso sobresale claramente, ya que si bien hay otros resultados con alta población, éste presenta una energía bastante más baja. Es notorio además que los otros ligandos que presentan un resultado con una población significativa (CRR, CAM, ANN) todos presentan anillos aromáticos, quedando claramente excluidos como potenciales ligandos el ácido graso (PAM) y el macrólido. Resultados equivalentes se observan en los paneles C y D que muestran los resultados del mismo test de ligandos para la CYP245A1 (PDB ID: 2Z3U) que une ácido cromopirrólico (PDB ID: CRR) y la cyp199 (PDB ID: 4DNJ) cuyo ligando es metoxybenzoato (PDB ID: ANN) respectivamente. En ambos casos nuevamente el ligando correcto se observa como un claro outlier en el gráfico, lo que hace que pueda identificarse fácilmente, el tipo de ligando que une. Por otro lado, el panel E muestra el resultado para la p450 BS (PDB ID: 1IZO) que une ácidos grasos como el palmitoleico (PAM). En este caso se observa que hay dos ligandos que presentan una pose con menor energía de unión y mayor población, con lo cual aquí no sería posible detectar el ligando correcto con el docking convencional.

En conjunto lo que se pudo observar es que, a excepción de los ácidos grasos (aquí representados por el ácido palmitoleico, PAM), que vale decir que son moléculas difíciles para el docking por su gran flexibilidad, en todos los demás casos evaluados o bien existe un punto outlier claro, como en el caso de la CYP154C5, o existen dos o tres resultados que podrían ser correctos, lo cual ya implica una reducción enorme en el abanico de posibilidades. Por otro lado un dato interesante adicional, surge del análisis detallado que muestra que, en general, si hay más de un ligando candidato, es decir que cae en la zona de baja energía y alta población, éstos comparten cierta semejanza estructural, por ejemplo, son de tamaño similar. En otras palabras, los resultados para la evaluación de la capacidad del método de docking para seleccionar el ligando correcto de una dada CYP, muestran que realizando el docking con una serie de ligandos lo suficientemente diversos o bien es posible identificar un ligando con altas chances de ser el correcto (un outlier claro), o al menos se puede tener una idea del tipo de ligando que

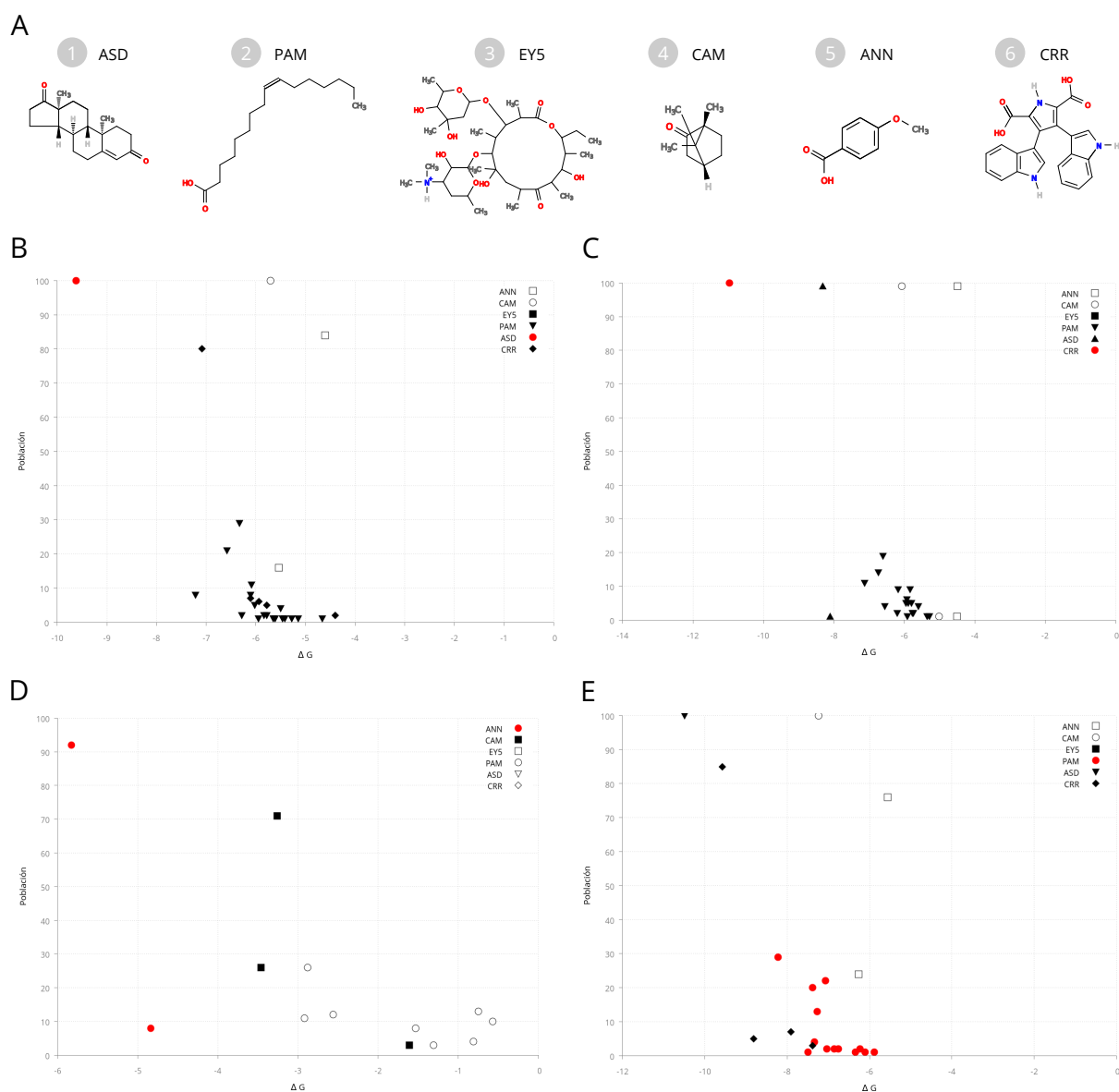


Figura 5.6: Evaluación de la selectividad de los CYPs frente a diversos tipos de ligandos. A. Estructuras de ligandos representativos. B, C, D y E. Gráficas de ΔG vs P correspondientes al docking de los seis ligandos sobre la estructura de la CYP154C5 (PDB ID: 4JBT) en B, la CYP245A1 (PDBID: 2Z3U) en C, la cyp199 (PDB ID: 4DNJ) en D y la p450BS (PDB ID: 1IZO) en E. Los puntos rojos corresponden al resultado del docking para el sustrato correcto en cada receptor. El docking de algunos de los ligandos sobre determinados receptores arroja resultados con energía positiva, pero dado que esto implica una energía de unión muy desfavorable, solo se muestran los resultados con energía inferior a cero.

podría unir la proteína.

Selectividad entre ligandos del mismo grupo

Una vez probado que el docking permite seleccionar el ligando correcto en la mayoría de los casos, el próximo interrogante planteado tiene que ver con evaluar dónde -en términos de similitud química- está el límite de esa selectividad.

Concretamente, ¿puede el docking diferenciar entre ligandos muy parecidos? Para responder este interrogante, se amplió el set de ligandos de testeo a 13 ligandos dentro de los cuales varios pertenecen al mismo grupo y se evaluaron los histogramas de ΔG vs P resultantes. En la Figura 5.7 A se muestran los resultados para la p450eryK que une eritromicina (EY5), un ligando de gran tamaño perteneciente al mismo grupo que DEB, NRB y EPD y en la Figura 5.7 B lo propio para la p450cam que une alcanfor (CAM), un ligando pequeño que comparte grupo con CNL. Como se observa, ligandos de un mismo grupo suelen arrojar resultados similares con la presencia de varios outliers. Esto muestra que si bien el docking no tendría la capacidad para detectar el ligando correcto entre varios similares, sí permite al menos reducir el espacio de posibilidades.

Habiendo determinado que el docking permite seleccionar el correcto entre ligandos candidatos (o al menos acotar el rango de ligandos posibles), la siguiente pregunta que se intentó responder es cuán selectivo puede ser el docking, si en vez de trabajar con estructuras cristalográficas del receptor, lo que se tiene es un modelo obtenido por homología.

Selectividad del docking sobre modelos de un receptor

La pregunta que se apunta a responder en esta sección es si es posible conservar la selectividad observada para el método de docking sobre modelos por homología. Para ello se seleccionaron aquellos modelos que resultaron exitosos en la predicción del modo de unión ligando-receptor y sobre ellos se probaron los seis ligandos representativos.

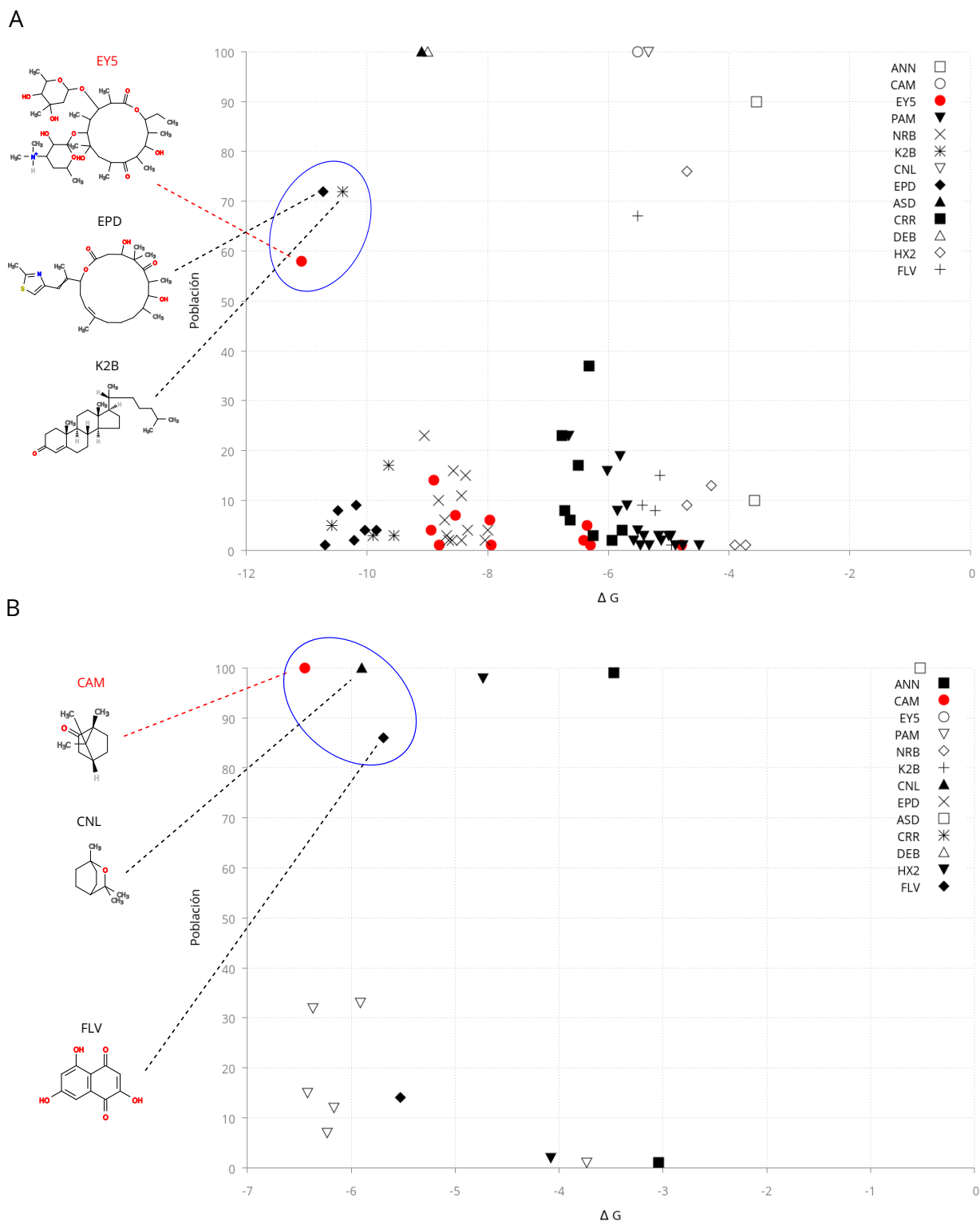
A modo de ejemplo, en la Figura 5.8 B, C y D se muestran los resultados obtenidos para tres modelos de la p450 cam (denominados I, II y III) los cuales fueron generados a partir de moldes a distinta distancia evolutiva, siendo el modelo I el correspondiente al molde más cercano y el modelo III el correspondiente al más lejano. Como puede observarse, al utilizar modelos se pierde un poco la selectividad en comparación al docking sobre la estructura real (Figura 5.8 A). Sin embargo, y consistentemente con los resultados observados para la capacidad predictiva en ligandos individuales, en los modelos I y II aún es posible encontrar al ligando correcto en la región superior izquierda del gráfico.

Análogamente, la Figura 5.9 muestra la selectividad sobre tres modelos de otra proteína, la p450 PikC que une nabomicina. Este ligando es bien distinto al del ejemplo anterior. En este caso la selectividad se mantiene bastante bien en los dos primeros modelos y decae recién en el tercero, es decir, en el más lejano. Estos resultados sugieren que aún sin tener estructura resuelta, si se consigue tener un modelo de buena calidad (utilizando moldes a distancia evolutiva cercana o media), hay altas chances de poder predecir el tipo de sustrato que se unirá a un determinado CYP.

5.3 Conclusiones

El análisis de los experimentos de redocking de los complejos disponibles evidencia que el programa Auto-dock4 es capaz de predecir de manera adecuada la estructura de complejos CYPs-sustrato, en la mayoría de los casos. Más aún, en determinados casos, el reconocimiento de patrones de interacciones compartidos entre proteínas que unen sustratos similares permite introducir sesgos que aumenten la chance de encontrar el modo de unión del ligando a través del docking. Sin embargo, dado que existen aún pocas estructuras de complejos para cada grupo de ligandos, esto solo puede hacerse para algunos farmacóforos particulares.

Por otra parte, el uso de modelos de la estructura del receptor para la predicción de la estructura del complejo enzima-sustrato, requiere en primer lugar que el modelo sea de alta calidad. Esto como se ha visto en el capítulo



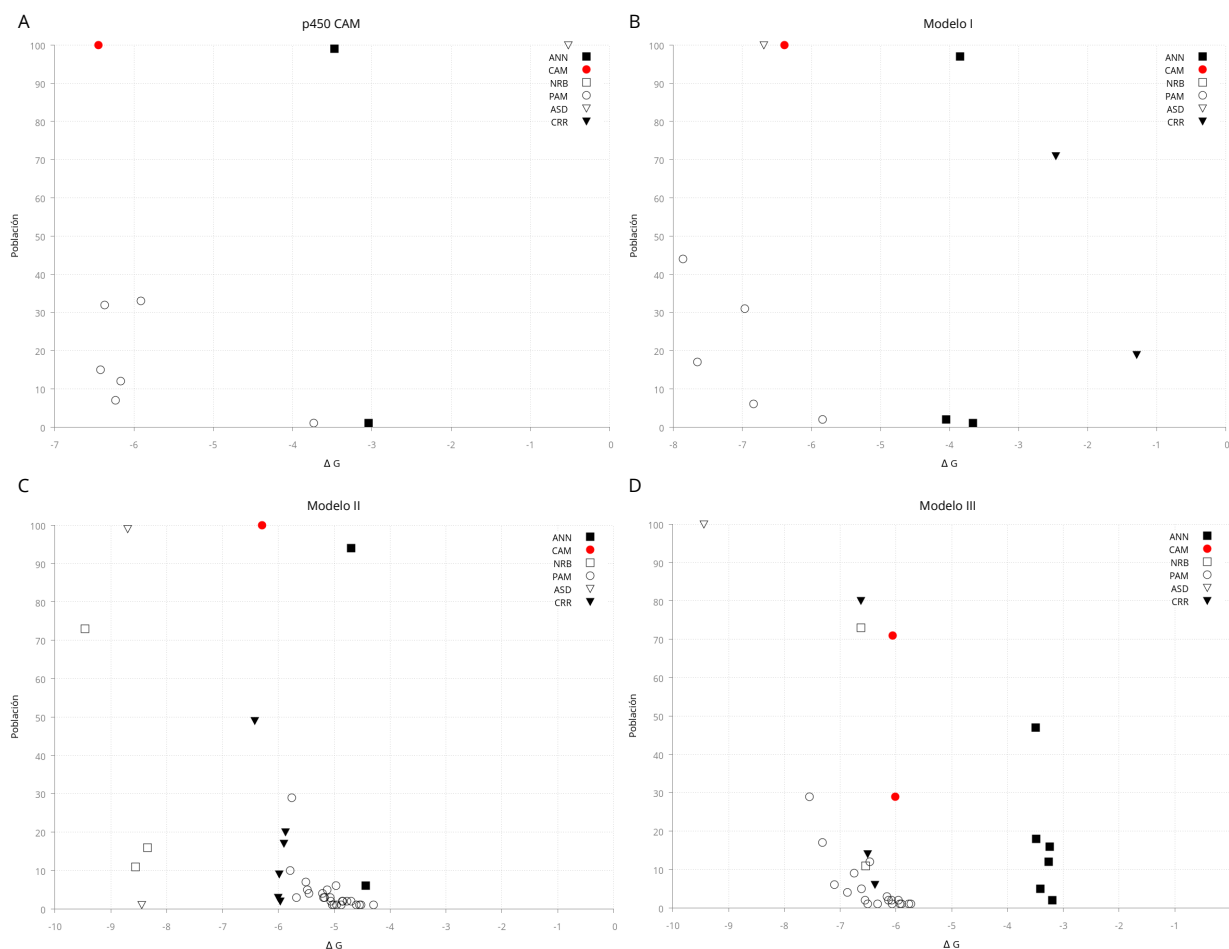


Figura 5.8: Ensayo de selectividad del docking sobre tres modelos de la p450 cam. En el panel A se muestra la selectividad obtenida a partir del docking sobre la estructura real de p450cam. En los paneles B, C y D se muestra lo análogo para cada uno de los tres modelos.

anterior, es mucho más factible de conseguir si se parte de un molde a una distancia evolutiva cercana o mediana. Adicionalmente, en el caso de enzimas que unen ligandos de gran tamaño, es un requisito adicional que el modelo logre reproducir el tamaño de la cavidad de unión del ligando. Para esto es clave elegir como molde la estructura de un complejo cuyo ligando sea grande y considerar el espacio ocupado por el ligando al construir el modelo, o sea utilizar como molde no solo la estructura de la enzima sino el complejo enzima-ligando.

En cuanto a la selectividad, se ha visto que realizando un conjunto de experimentos de docking con una serie de ligandos lo suficientemente diversos o bien es posible identificar un ligando con altas chances de ser el correcto (un outlier claro), o al menos se puede tener una idea del tipo de ligando que podría unir la proteína. Cuando se trabaja con modelos de un receptor, se ha visto que lógicamente la selectividad disminuye, pero sin embargo para modelos de buena calidad aún es posible encontrar el ligando correcto dentro del conjunto de 2-3 ligandos candidatos. Esto significa que aún sin tener estructura resuelta, si se consigue tener un modelo de buena calidad, hay altas chances de poder predecir el tipo de sustrato que se unirá a un determinado CYP.

En resumen habiendo evaluado la capacidad y limitaciones de los métodos de docking para determinar el ligando y su modo de unión a un dado CYP, pasaremos a integrar esta información con los análisis realizados en el capítulo anterior, para obtener un modelo bioinformático predictivo del sustrato y reacción de un dado CYP a partir del conocimiento de su secuencia.

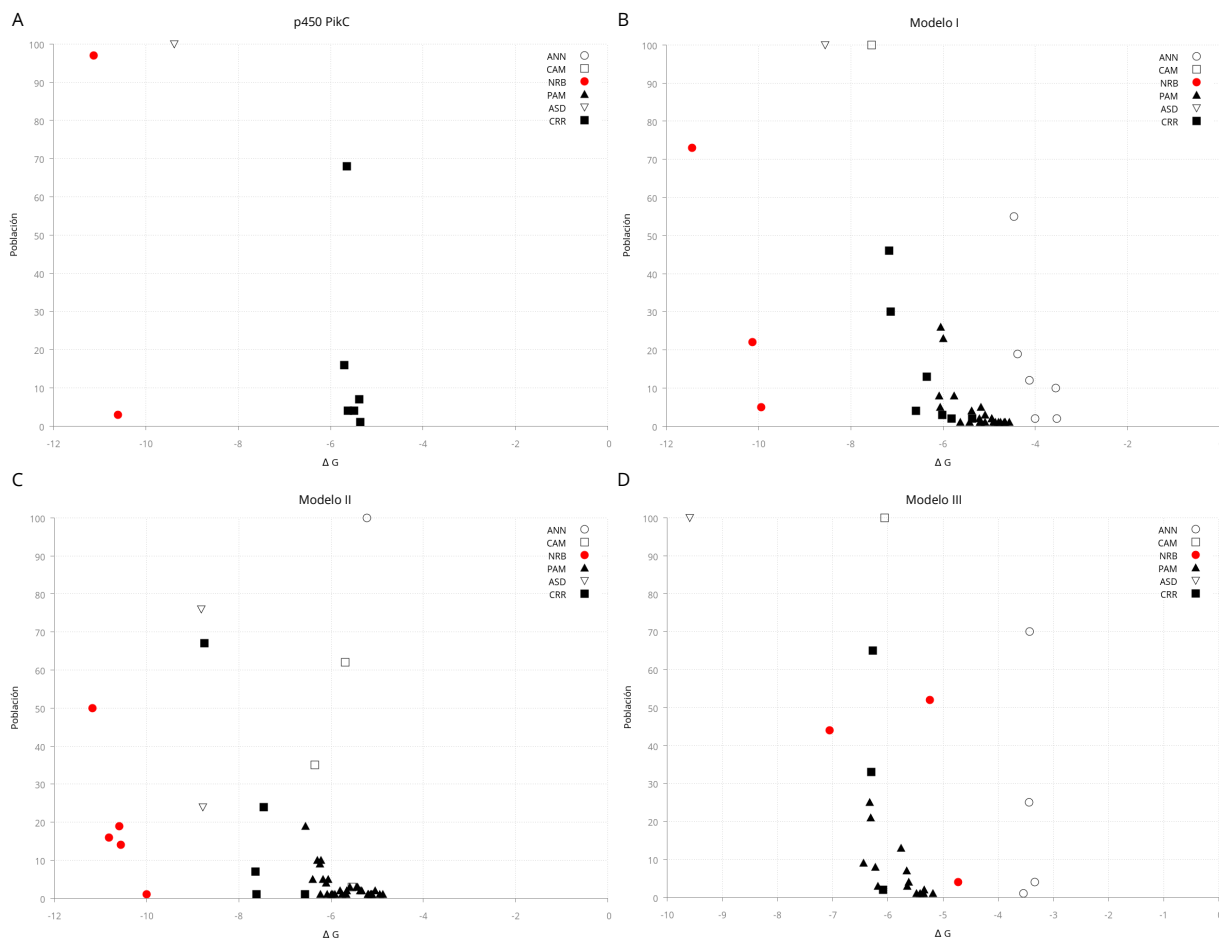


Figura 5.9: Ensayo de selectividad del docking sobre tres modelos de la p450 pikC. En el panel A se muestra la selectividad obtenida a partir del docking sobre la estructura real de la p450 pikC. En los paneles B, C y D se muestra lo análogo para cada uno de los tres modelos.

5.4 Métodos específicos del capítulo

5.4.1 Obtención de las estructuras de complejos CYP-sustrato

El conjunto de complejos enzima-sustrato que se utilizó en este capítulo estuvo conformado por 17 complejos representativos (que en adelante denominaremos set 4). Para lograr que el conjunto sea lo más representativo de la diversidad estructural de estas enzimas, se tomó uno de cada grupo filogenético (ver capítulo 4) para los 11 grupos que tienen al menos un complejo cristalizado y luego se completó el set con 6 complejos más para incluir una mayor variedad química de sustratos y poder realizar también análisis intragrupo.

5.4.2 Preparación de la estructura de los receptores

Las estructuras de los receptores proteicos utilizadas en los cálculos de docking fueron obtenidas a partir de los complejos que conforman el set 4. Dado que no todos los cristales contaban con la presencia del átomo de oxígeno que ocupa la sexta posición de coordinación en el sitio distal del hemo en el estado activo de estas proteínas, una vez extraída la porción proteica del complejo junto a su cofactor se procedió a incorporar el átomo de oxígeno faltante. Tal incorporación se realizó a través de un alineamiento de los átomos de hierro del hemo de todos los receptores a una estructura que sí ha sido cristalizada con el átomo de oxígeno en cuestión (PDB ID: 1DZ8). Esto permitió obtener las coordenadas del átomo de oxígeno para todas las demás estructuras.

5.4.3 Preparación de las estructuras de los ligandos

Las estructuras de los ligandos fueron preparadas partiendo de su correspondiente SMILES (obtenido de la base de datos PubChem), que fue luego procesado con OpenBabel[133] para generar la estructura tridimensional. Esto permitió que la estructura del ligando generada no tenga el sesgo del confórmer presente en el cristal obtenido del PDB, lo cual sí ocurriría si se utilizara directamente las coordenadas del ligando en el cristal del complejo.

5.4.4 Cálculo del RMSD de la estructura del ligando

El cálculo del RMSD de la estructura del ligando obtenida en la simulación de docking, respecto a la estructura del ligando en el cristal es la métrica más ampliamente utilizada para evaluar los resultados del docking. Sin embargo, si se utiliza la fórmula estándar para calcular el RMSD (ecuación 5.1), que compara pares de átomos que llevan el mismo nombre, se observa que en algunos casos se sobreestima el valor del RMSD. Esto ocurre en general cuando se trata de moléculas que tienen alguna porción simétrica, por ejemplo un anillo bencénico, en donde un giro de 180° del anillo resulta en una estructura idéntica que correspondería a un RMSD cercano a cero, pero sin embargo como los átomos que llevan el mismo nombre se encuentran del lado opuesto del anillo, se obtiene un RMSD mucho mayor.

$$HA_RMSD_s = \sqrt{\frac{1}{N} \sum_{i=1}^N \|a_i - b_i\|^2} \quad (5.1)$$

Como solución a este problema, existen distintas alternativas: una de ellas se denomina RMSD de distancia mínima (ecuación 5.2) y es la implementada por el programa Autodock Vina [134]. Esta métrica determina los pares de átomos entre la estructura de referencia y la pose del ligando obtenida del docking asociándolos al átomo del mismo elemento más cercano. Esta aproximación, si bien corrige el valor del RMSD en algunos casos, en otros conduce a una subestimación del mismo.

$$RMSD_{\min}(A, B) = \sqrt{\frac{1}{N} \sum_{i=1}^N (\min_j \|a_i - b_j\|)^2} \quad (5.2)$$

Otra alternativa es calcular el RMSD corregido por simetría, basada en una implementación del algoritmo húngaro[135]. Este algoritmo resuelve el problema de asignación óptima entre un set de átomos del ligando a y un set de átomos de la pose del docking b del mismo tamaño. Para todos los grupos de átomos del mismo tipo, se calcula el cuadrado de la distancia entre cada átomo de la referencia a_i y cada átomo de la pose b_j . El algoritmo húngaro se utiliza para determinar las asignaciones uno a uno entre los átomos de la pose y la referencia de modo tal que la distancia total entre átomos sea mínima. En base a los pares de átomos obtenidos se calcula el RMSD utilizando la fórmula estándar.

$$HA_RMSD_h = \sqrt{\frac{1}{N} \sum_{i=1}^N \|a_i - b_{c(i)}\|^2} \quad (5.3)$$

Los resultados del docking expuestos en este capítulo fueron evaluados utilizando el RMSD corregido por simetría tal como se encuentra implementado en el programa DOCK6[79].

5.4.5 Protocolo de docking

Para los cálculos de docking planteados en el presente capítulo, se empleó la versión 4.2.5 del programa AutoDock. Para cada par proteína-ligando, se creó una grilla de forma cúbica de 22,5 Å de lado con una resolución de 0,375 Å que abarcó en todos los casos la totalidad de los aminoácidos presentes en el sitio de unión del ligando. Utilizando el programa Autogrid se calcularon los mapas de energía para cada tipo de átomo de cada

ligando. Las coordenadas del centro de la grilla se obtuvieron a partir del cálculo del centro de masa del ligando. Los parámetros del algoritmo genético utilizado para la búsqueda conformacional se han mantenido en sus valores por defecto (150 para el tamaño inicial de la población, 2.5×10^6 como el número máximo de evaluaciones de energía y 2.7×10^4 como el número máximo de generaciones). Para cada par enzima-ligando se realizaron 100 corridas de docking diferentes y los resultados se agruparon utilizando un valor de RMSD de 2 Å.

6

INTEGRACIÓN DEL CONOCIMIENTO GENERADO EN UN ESQUEMA PREDICTIVO

6.1 Introducción

Como se mencionó al comienzo del capítulo anterior, uno de los objetivos de la presente tesis consiste en desarrollar una estrategia que, tomando cómo punto de partida la secuencia de un dado CYP y las propiedades descritas en los capítulos anteriores, permita predecir (o acotar) su posible rango de sustratos y la reacción asociada. Habiendo analizado entonces la relación y poder predictivo de diferentes propiedades, como ser la relación filogenética, el modelado por homología, el contexto génico y la estructura del complejo enzima-sustrato, desarrollamos ahora un esquema de flujo de trabajo que tomando como punto de partida la secuencia, permita inferir el posible sustrato y la reacción catalizada. Es importante notar que dada la diversidad de las propiedades determinadas y el proceso subyacente, si bien es posible diagramar un flujo de información y toma de decisiones (ver resultados), la integración final debe ser hecha caso por caso y analizando personalmente y manualmente los resultados obtenidos. En este sentido el objetivo no es desarrollar un método automático de predicción de sustrato para un dado CYP, sino un flujo de trabajo e integración de la información que acote el universo de posibilidades y/o genere sustratos potenciales con una estimación de la confianza en la predicción.

6.2 Resultados

La estrategia de trabajo desarrollada para el esquema predictivo se muestra en la Figura 6.1. Tomando como punto de partida una secuencia de interés, en primer término se analiza si existe (o no) una estructura disponible, en caso afirmativo se chequea si hay una propuesta de sustrato y si la hay se procede a realizar el docking con el mismo (ver abajo). En el caso de no contar con una estructura, que constituye el caso más probable dado el bajo número de estructuras de CYPs que han sido resueltas (sólo 53), se alinea la secuencia contra los HMMs de los 15 grupos definidos en el capítulo 4, y se la asigna al grupo más probable. Esto permite determinar cuál es la estructura más cercana para usar como molde para el modelado por homología priorizando las estructuras co-cristalizadas con sustratos, ya que como se concluyó en el capítulo 5, esta consideración resulta crucial cuando se trata de modelar proteínas que unen ligandos de gran tamaño. Tanto para aquellos casos donde exista una estructura, como aquellos que han sido modelados, se analiza si existe o no una propuesta funcional que permita inferir el posible (o potencial) sustrato. En los casos en los que se cuenta con un sustrato probable, se procede a realizar el docking. Si el docking da un resultado satisfactorio, o sea que existe un complejo enzima-sustrato que es un claro outlier en el gráfico de energía de unión vs. población y cuya estructura es consistente con las interacciones claves observadas en el grupo filogenético (del CYP y de los sustratos), se considera al sustrato propuesto como altamente probable. En paralelo, y principalmente en aquellos casos donde no se cuenta

con una propuesta de posible sustrato, se utiliza la información del contexto génico para determinar familias de posibles sustratos. Con este set de sustratos se procede a realizar el análisis de docking enzima-sustrato y evaluar aquel (o aquellos) sustrato (o grupo) que resulta en complejos enzima-sustrato confiables. Finalmente, en aquellos casos en los que no hay información del contexto génico pero se cuenta con un modelo estructural de calidad, es posible realizar un docking con un sustrato de cada uno de los grupos y de este modo descartar aquellos que no encajan en el sitio activo. Para validar el esquema desarrollado, a continuación se describe la estrategia y resultados asociados para tres casos específicos.

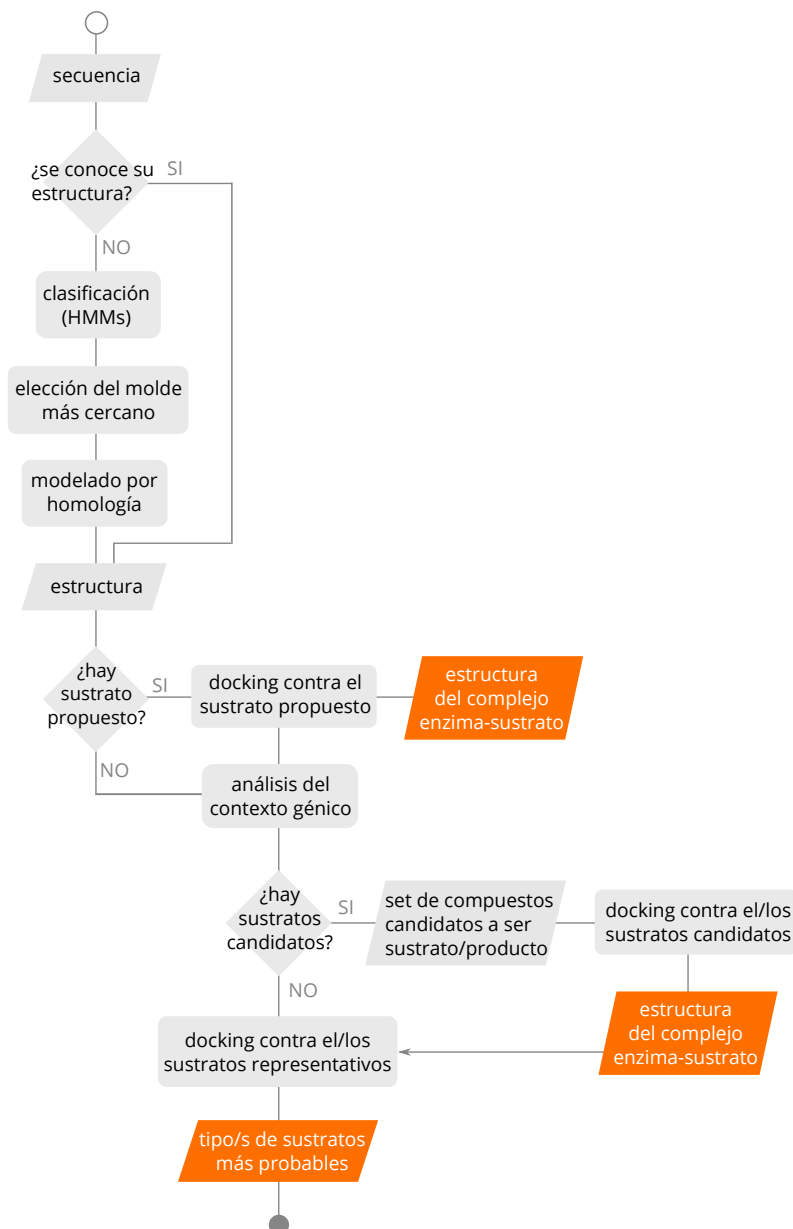


Figura 6.1: Estrategia para la toma de decisiones y flujo de información del modelo predictivo.

Los ejemplos seleccionados comprenden tres CYPs para los cuales existe información del contexto génico, es decir que podemos obtener una lista de compuestos candidatos a ser sustratos a partir del análisis de los productos y sustratos de sus genes adyacentes, y a que a su vez se conoce el sustrato real como también la correspondiente estructura del complejo enzima-sustrato, dato estrictamente necesario para que sea posible validar las predicciones. Para poner a prueba el pipeline completo, en todos los casos supondremos que no

conocemos la estructura del CYP en cuestión, por lo cual procederemos a obtener un modelo utilizando como molde una CYP cercana, ya que como pudimos concluir en los capítulos anteriores, de este modo es mucho más probable arribar a un modelo de alta calidad.

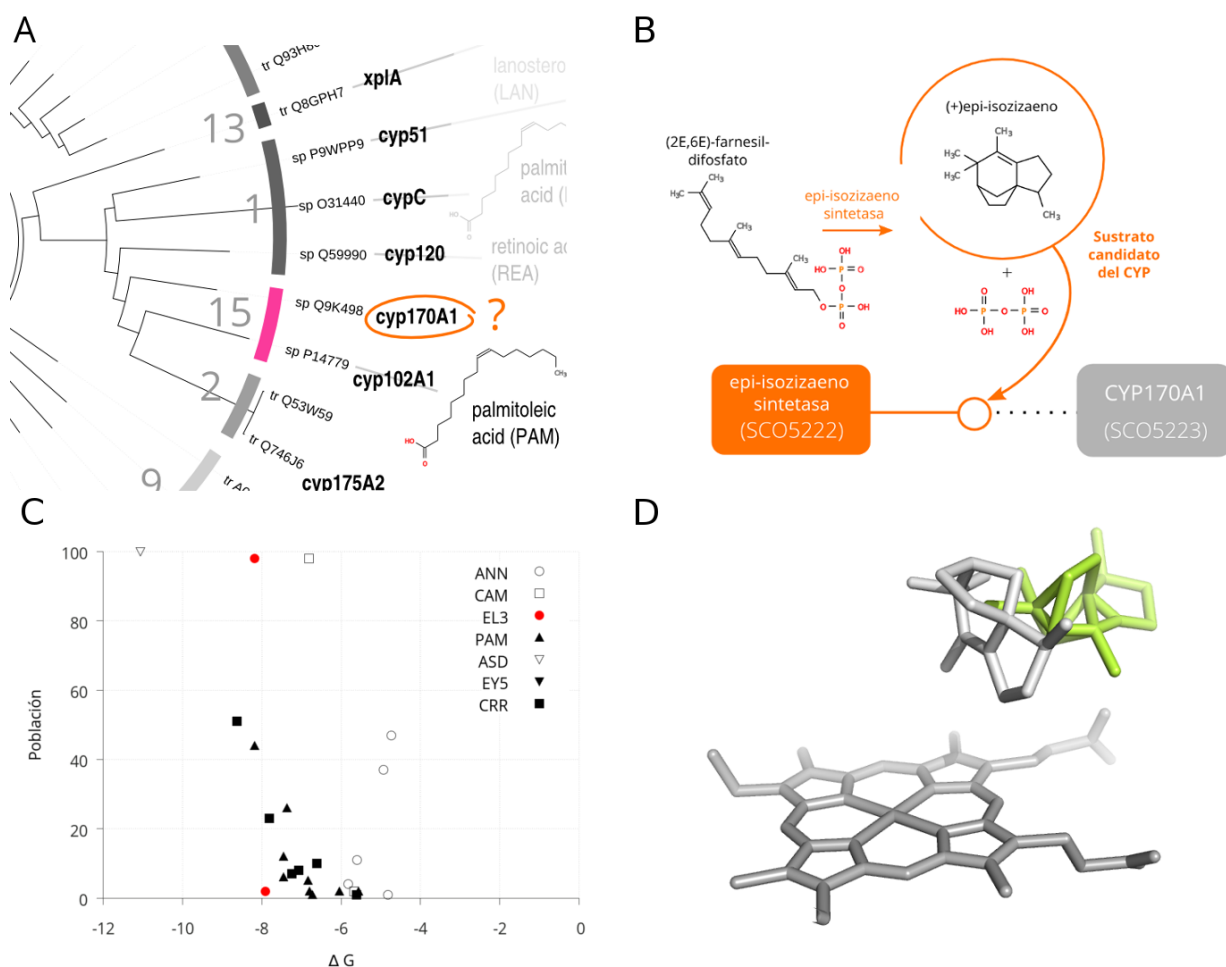


Figura 6.2: A. La comparación de la secuencia del CYP170 contra la base de datos de HMMs sugiere que pertenece al grupo 15. B. El análisis del contexto génico muestra que la enzima codificada por el gen río arriba del SCO5223 sintetiza un compuesto denominado (+)epi-isozizaeno (PDB ID: EL3). C. Histogramas resultantes del docking del compuesto predicho y de los seis ligandos representativos de CYPs. D. Resultado del docking del compuesto EL3 sobre la estructura del complejo de referencia.

6.2.1 Caso 1: CYP170A1

El primer caso elegido es el del CYP170A1 (UniProt ID: Q9K498), cuyo flujo de información asociado y resultados se muestran en la Figura 6.2. La comparación de la secuencia de este citocromo contra la base de datos de HMMs sugiere asignarlo al grupo 15 (ver Figura 6.2A). En este grupo se cuenta con una única estructura co-cristalizada con su ligando, que es un ácido graso (PDB ID: 1FAG), por lo cual se la tomó como molde para construir un modelo estructural del CYP170A1. El análisis del contexto génico (Figura 6.2 B), por otro lado permitió identificar como un potencial sustrato al compuesto (+)epi-isozizaeno, generado por la enzima que codifica el gen río arriba del CYP en cuestión. Para evaluar la factibilidad de que el sustrato predicho por el contexto génico sea el probable, se procedió a realizar el protocolo de docking. El resultado obtenido muestra solo dos conformaciones, una de las cuales posee un valor de energía de interacción muy razonable ($\Delta G < -8$) junto a un valor de población muy alto ($> 95\%$), por lo que el docking estaría respaldando la predicción realizada por medio del contexto génico. Adicionalmente, a los fines de evaluar la selectividad de la enzima, se realizó el docking de un compuesto representativo de cada tipo de ligando (definidos en el capítulo 4) y se lo comparó con el resultado

obtenido para el (+)epi-isozizaeno. Como se muestra en la Figura 6.2C, esta enzima pareciera ser más afín por ligandos pequeños similares al predicho, como por ejemplo el alcanfor (CAM). Esto nos da la pauta de que si no tuviésemos información del contexto génico, el docking contra la serie de ligandos representativos nos sugeriría de igual modo que esta enzima une un ligando pequeño con anillos aromático/alifáticos. Otro tipo de ligando que presenta una energía de interacción favorable y un valor alto de población es el de las hormonas esteroideas, aquí representadas por ASD. Sin embargo se observó que este ligando suele presentar resultados similares con diversas enzimas, por lo que no resulta particularmente llamativo.

El presente caso, de acuerdo a los resultados descritos en el párrafo anterior puede constituir un caso de éxito, ya que el ligando predicho es efectivamente el correcto y la estrategia desarrollada permite arribar a una estructura del complejo enzima-sustrato muy cercana a la real (observada en el cristal) como se muestra en la Figura 6.2D.

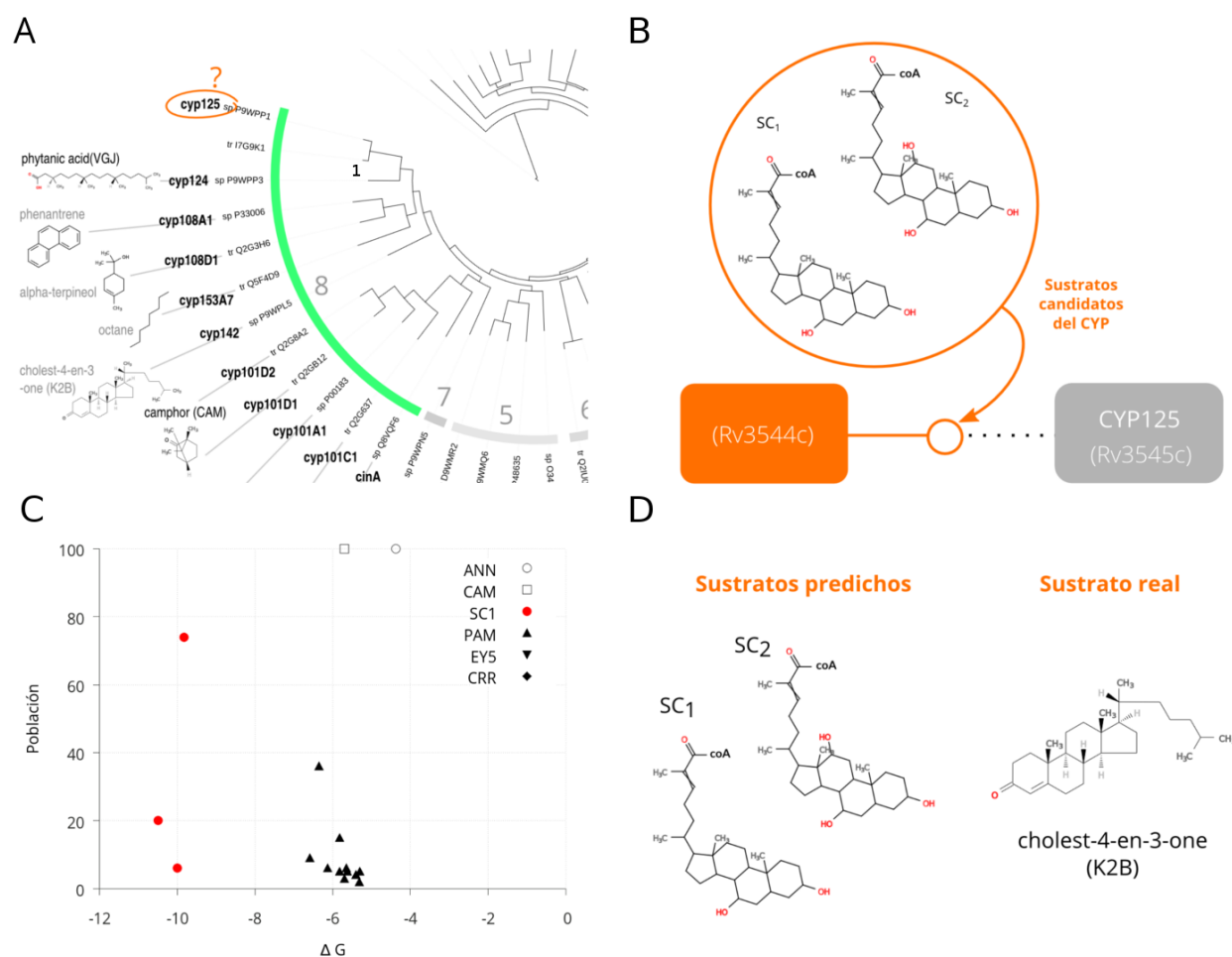


Figura 6.3: A. La comparación de la secuencia del CYP125 contra la base de datos de HMMs sugiere que pertenece al grupo 8. B. El análisis del contexto génico muestra dos compuestos (aquí denominados SC1 y SC2) como posibles sustratos. C. Histogramas resultantes del docking de uno de los compuestos predichos y de los ligandos representativos de CYPs. D. Comparación de las estructuras químicas de los compuestos predichos por el pipeline con el sustrato real (colest-4-en-3-one).

6.2.2 Caso 2: CYP125

El segundo ejemplo, comprende el caso del CYP125 (Uniprot ID: P9WPP1). La comparación de la secuencia frente a los modelos de Markov permite asignarla al grupo 8 (ver Figura 6.3A). Dentro este grupo hay varias estructuras cocrystalizadas con su correspondiente sustrato pero la más cercana es la correspondiente al CYP124

(Uniprot ID: P9WPP3) que tiene como sustrato al ácido fitánico, por lo que se realizó el modelo por homología a partir de este molde (PDB ID: 2WM4). La evaluación del contexto génico arrojó dos compuestos similares al colesterol como posibles sustratos del CYP125 y cuyo nombre es demasiado extenso, por lo que aquí se los denominará SC1 y SC2 (ver Figura 6.3 B). Ambos compuestos contienen coenzima-A como parte de su estructura, lo cual hace sospechar que la predicción no fue acertada porque ninguno de los CYPs une tal molécula, pero si nos enfocamos en el resto de la molécula (el fragmento que se parece al colesterol), sí tenemos una estructura química parecida a algunos de los sustratos conocidos de los CYPs. En este caso se realizó el docking solo para este fragmento de los compuestos candidatos SC1 y SC2, obteniéndose el resultado que se observa en la Figura 6.3 C (solo se muestra el correspondiente al SC1). Al comparar con el desempeño del docking obtenido por los demás compuestos representativos, se observa en este caso que el SC1 presenta valores de energía y población que lo hacen diferenciarse bastante del resto. Este resultado sugiere que es posible que el CYP125 efectivamente una una molécula similar al colesterol, y de hecho si la comparamos con el sustrato real (colest-4-en-3-one) como se muestra en el panel B, se ve que claramente se parecen. Este es un caso en el cual, si bien el pipeline no lleva a encontrar el sustrato correcto directamente, sí aporta claros indicios de la naturaleza química del mismo.

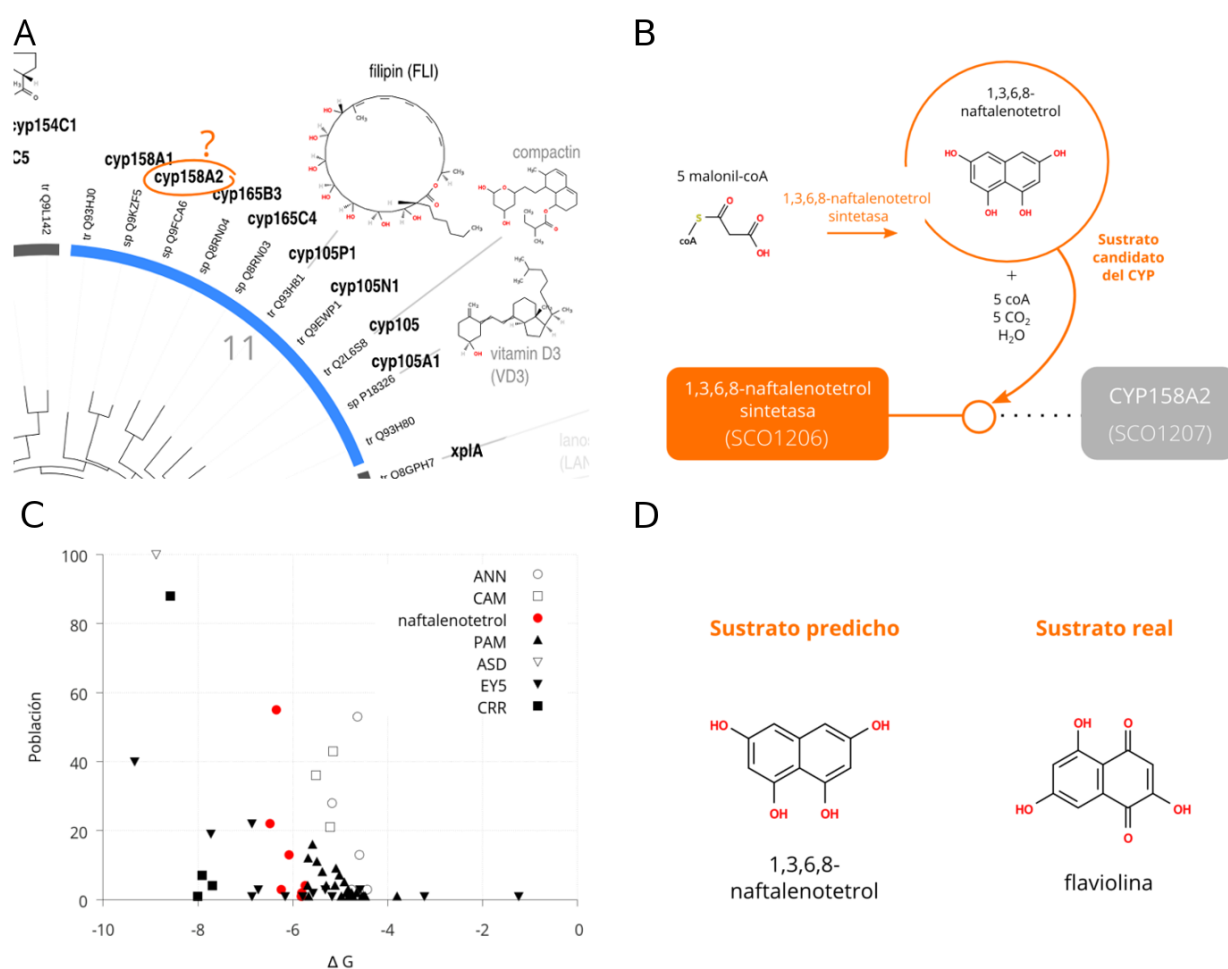


Figura 6.4: A. La comparación de la secuencia del CYP158A2 contra la base de datos de HMMs sugiere que pertenece al grupo 11. B. El análisis del contexto génico muestra que la enzima codificada por el gen río arriba del SCO1207 sintetiza un compuesto denominado 1,3,6,8-naftalenotetrol. C. Histogramas resultantes del docking del compuesto predicho y de los seis ligandos representativos de CYPs. D. Comparación de las estructuras químicas del compuesto predicho por el pipeline (1,3,6,8-naftalenotetrol) con el sustrato real (flaviolina).

6.2.3 Caso 3: CYP158A2

El tercer y último ejemplo tiene como protagonista al CYP158A2 (Uniprot ID: Q9FCA6). Nuevamente se partió de la secuencia, se la comparó contra la base de datos de modelos ocultos de Markov (HMMs) de los 15 grupos y en base a los resultados obtenidos se la asignó al grupo 11 (ver Figura 6.4 A). En este grupo existen dos estructuras cocrystalizadas con su correspondiente ligando: la del CYP105P1 con su sustrato filipin (PDB ID: 3ABA) y la del CYP105A1 con el producto hidroxivitamina D (PDB ID: 2ZBZ). El blast de la secuencia del CYP158A2 respecto a estas dos muestra que la más cercana es la que hidroliza vitamina D, por lo que se utilizó ésta como molde para el modelado por homología. Seguidamente se analizó el contexto génico y se obtuvo un compuesto candidato denominado 1,3,6,8-naftalenotetrol, el cual es producto de la reacción de la enzima 1,3,6,8-naftalenotetrol sintetasa, codificada por el gen río arriba del SCO1207 (panel B en la Figura 6.4). El docking de este compuesto candidato sobre el modelo estructural del CYP158A2 (ver panel C de la Figura 6.4) muestra que el compuesto predicho presenta una interacción favorable en el sitio de unión de la enzima ($\Delta G < -6$) y una población moderadamente alta ($P \approx 55$). Si se lo compara con el desempeño de los compuestos representativos de los diferentes tipos de ligandos de p450s, se puede decir que el 1,3,6,8-naftalenotetrol se encuentra entre los tres o cuatro ligandos más probables. Vale decir que los que lo acompañan en este set son también ligandos con anillos aromáticos, pero que presentan un tamaño bastante mayor, lo cual podría poner en duda si esta enzima une un ligando grande o uno pequeño. Sin embargo, este es un caso bien difícil, ya que el CYP158A2 une no solo una sino dos y hasta tres moléculas del mismo sustrato en simultáneo, dado que la reacción que cataliza consiste precisamente en la formación de dímeros y trímeros de su sustrato flaviolina (ver panel D en la Figura 6.4).

Los casos descriptos representan una prueba de concepto, de que cuando la información disponible es adecuada (estructura cercana y contexto génico) la estrategia desarrollada tiene una excelente capacidad predictiva.

6.2.4 Aplicación a gran escala

Habiendo validado el esquema predictivo, se procedió a analizar su potencial para el conjunto total de secuencias de CYPs disponibles. En la Figura 6.5 se muestra como evoluciona el número de secuencias de acuerdo al esquema de trabajo de la Figura 6.1.

Como puede verse, del total de 15831 CYPs solo 53 poseen estructura y entre ellas son menos de la mitad las que presentan estructura del complejo enzima-sustrato. El 70 % de este set de estructura conocida tiene una propuesta de sustrato en la bibliografía y para el tercio de las que no la tienen, existe un sustrato candidato que puede ser inferido por análisis del contexto génico. Esto significa que hay aproximadamente 17 casos con estructura conocida para los que no se conoce la estructura del complejo, pero que el mismo podría inferirse utilizando docking y complementando con un análisis del contexto génico (círculos verdes en la Figura 6.5). Hay unos 12 restantes, para los cuáles es imposible tener un sustrato candidato por lo que la única posibilidad es dockear un representante de cada grupo y en función de los resultados acotar el rango de posibles grupos de sustratos. Respecto al conjunto para el cual no se conoce la estructura, suman 937 secuencias entre las que tienen un sustrato propuesto (396) y las que tienen algún sustrato candidato (541) (ambas señaladas en color naranja en la Figura 6.5), pero dado que para éstas es necesario modelar la estructura, es preciso considerar solo aquellas que permitirían obtener un modelo de calidad. Con este requisito, este número se reduce a 473 (círculos azules en la Figura 6.5). En otras palabras para aproximadamente 3 % de los CYPs conocidos es posible aplicar nuestra estrategia en todo su potencial combinando el docking con una propuesta de sustrato, que permita inferir un posible sustrato con alta probabilidad.

Finalmente, es importante mencionar que es posible generar un modelo de buena calidad para aproximadamente la mitad de los CYPs conocidos (señalados en amarillo en la Figura 6.5). Esto abre las puertas para analizar posibles tipos de sustratos por docking y en caso de que aparezcan candidatos surgidos del análisis del contexto genico y/o de alguna propuesta funcional verificar la misma.

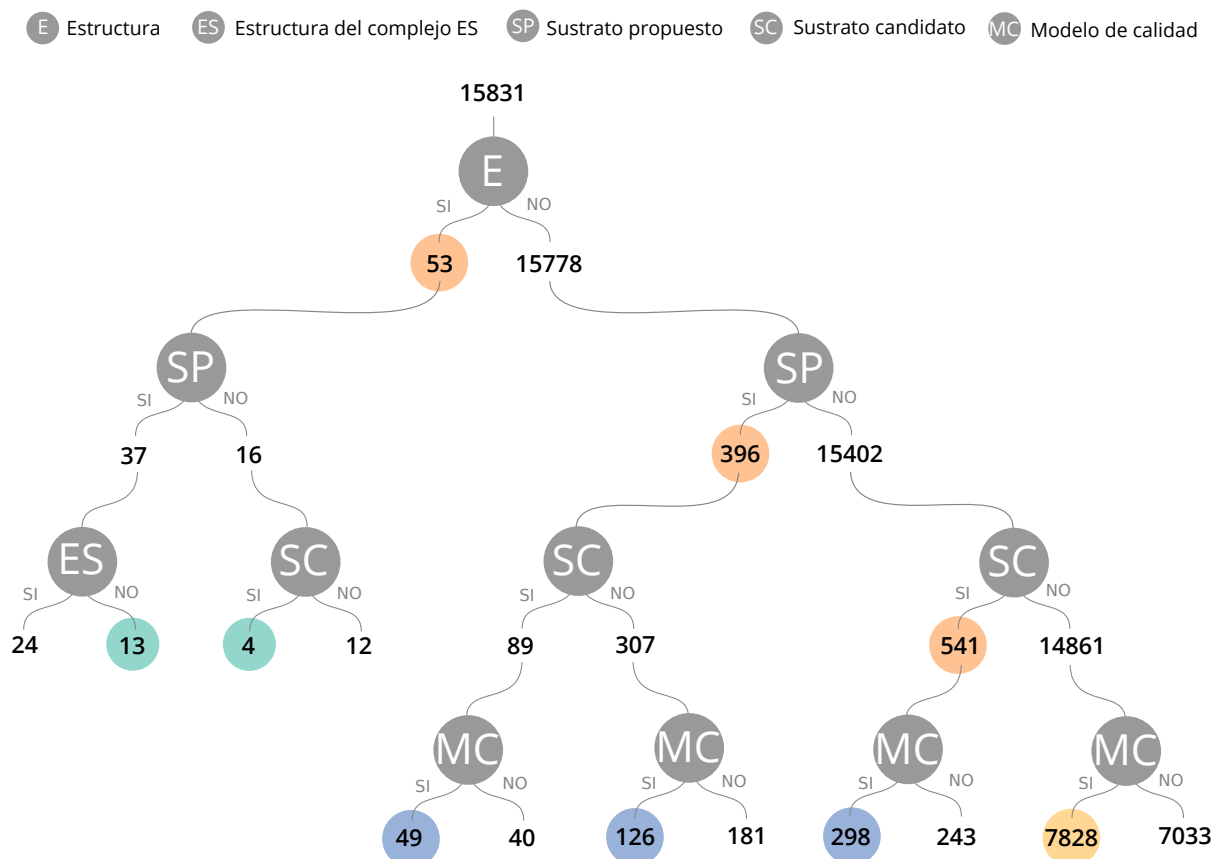


Figura 6.5: Evolución del número de secuencias según la disponibilidad de información de: estructura, estructura del complejo enzima-sustrato, sustrato propuesto, sustrato candidato (inferido por el contexto génico) y posibilidad de generar un modelo de calidad (en función de la disponibilidad de una secuencia cercana que tenga estructura).

6.3 Conclusiones

En este capítulo nos propusimos integrar los conocimientos y capacidades descriptas en los capítulos 4 y 5 en un esquema predictivo que permitiera encontrar con un cierto grado de confianza el posible sustrato (o tipo de sustrato) para un dado CYP, tomando como punto de partida solamente la información de su secuencia.

El primer paso clave en el esquema propuesto es la comparación de la secuencia contra la base de datos de HMMs de los 15 grupos filogenéticos definidos en el capítulo 4, en tanto que, como se ha mostrado en los ejemplos descriptos, permite determinar el grupo al que pertenece el CYP y en función de ello escoger el molde más adecuado para el modelado por homología. Otro punto importante en el esquema, en particular cuando no se cuenta con una propuesta de posible sustrato, es el análisis del contexto génico para inferir sustratos potenciales. Esta estrategia ha mostrado ser efectiva para dar indicios del tipo de sustrato, e inclusive en algunos casos ha llegado a revelar el sustrato exacto del CYP en cuestión. Por otra parte, el docking manifestó ser una buena herramienta para seleccionar los sustratos más probables, permitiendo descartar aquellos que no encajan en el sitio activo y sirviendo además como validación de la predicción realizada por información del contexto génico.

En este contexto, los tres ejemplos presentados representan una prueba de concepto, y constituyen una evidencia clara de que cuando la información disponible es adecuada, es decir, que se cuenta con una estructura cercana que permita obtener un modelo estructural de calidad y a su vez existe información de sustrato posible en el contexto génico, la estrategia desarrollada presenta una excelente capacidad predictiva.

En última instancia, evaluamos la capacidad de aplicar la estrategia desarrollada a gran escala. En este sentido, la evaluación del conjunto total de secuencias de CYPs disponibles respecto a: contenido de información de estructura, estructura del complejo enzima-sustrato, sustrato propuesto y sustrato candidato y posibilidad de generar un modelo de calidad (cuyos resultados se resumen en la Figura 6.5) reveló lo siguiente:

- i). para aproximadamente el 3 % de los CYPs conocidos (es decir, unos 490 -que es el número que suman los valores resaltados en círculos verdes y azules en la Figura 6.5-) es posible aplicar nuestra estrategia en todo su potencial combinando el docking con una propuesta de sustrato, que permita inferir un posible sustrato con alta probabilidad.
- ii). para aproximadamente la mitad de los CYPs conocidos es posible generar un modelo de buena calidad, lo cual permite analizar posibles tipos de sustratos por docking y en caso de que aparezcan candidatos surgidos del análisis del contexto génico y/o de alguna propuesta funcional verificar la misma.

Desde una visión general, los resultados muestran que la principal limitación que se encuentra hoy en día es la baja disponibilidad de datos de contexto génico. Sin embargo, esta situación podría revertirse en la medida que se vaya mejorando el nivel de anotación de las secuencias. Contando con tal mejora en la disponibilidad de sustratos candidatos, el potencial del método podrá ser ampliamente aprovechado.

DISCUSIÓN Y PERSPECTIVAS FUTURAS

Predecir la función de una proteína a partir de la información contenida en su secuencia y/o estructura constituye uno de los objetivos más importantes de la biología estructural, en especial si se considera el enorme crecimiento en el número de secuencias disponibles surgidas de múltiples proyectos de secuenciación genómicos. Realizar una asignación o anotación funcional general, típicamente en base a la similitud con secuencias de función bioquímica conocida por medio de BLAST o bien por asignación a una determinada familia a través de la comparación con perfiles de HMMER, es en la actualidad una práctica común. No obstante, determinar aspectos funcionales específicos como la especificidad/afinidad por cierto sustrato o la eficiencia catalítica de una proteína dada con precisión y detalle a nivel de residuo, no es una tarea sencilla. Sin embargo, si se asume que la función proteica está determinada por la estructura y las propiedades fisicoquímicas de sus residuos, lo cual es información que se encuentra codificada en la estructura primaria de la proteína, debería ser posible, en principio, predecir las propiedades funcionales en base a la secuencia y estructura.

Como se ha mencionado en la introducción de esta tesis, las enzimas pertenecientes a la familia de los citocromos p450s catalizan, en su conjunto, miles de reacciones sobre sustratos diferentes. Estas reacciones -que desde el punto de vista químico pueden considerarse oxidaciones- comprenden desde la inserción de grupos -OH (u -O⁻) en carbonos alifático/aromáticos, hasta el clivaje (por hidrólisis) y formación de enlaces, incluso aquellos de tipo C-C. Es más, aún cuando sus sustratos sean muy similares o hasta la misma molécula, las reacciones pueden diferir según se trate de uno u otro CYP. En este sentido, y si bien en una primera aproximación la inclusión de un gen/proteína en la familia de los CYPs permite acotar su función a "posible oxidasa" esto dice poco sobre a quién oxida (o sea cuál es (o son) su/s sustrato/s), y menos aún cuál de todas las oxidaciones posibles realiza sobre el mismo. Por esta razón, asignar una función específica a un determinado CYP, o en otras palabras determinar qué oxidación realiza sobre qué sustrato, constituye un enorme desafío.

Contribuir a la resolución de este desafío es de enorme relevancia en el contexto de la era genómica. Hoy el número de secuencias crece a una velocidad tal que es imposible que todos los productos génicos puedan ser caracterizados experimentalmente debido al costo que esto implica. Es aquí donde la bioinformática cobra protagonismo. Resulta imperioso desarrollar métodos que permitan la anotación precisa de la función proteica, que constituye la clave para entender la vida a nivel molecular, y es la base para que numerosas aplicaciones de los CYPs en campos como la biotecnología, la medicina y la biorremediación, puedan realmente concretarse.

Abordando entonces este desafío, en el presente trabajo de tesis, se emplearon diversos enfoques para dar respuesta a los múltiples interrogantes que se requiere responder para definir precisamente la función de cada una de estas enzimas.

Para abordar el problema en cuestión utilizamos, por un lado, métodos de simulación computacional -tanto clásicos como cuánticos- que permiten entender la fisicoquímica del proceso de catálisis con detalle atómico y proponer, entre otras cosas, el mecanismo de reacción para un dado CYP y la relevancia que tienen para el mismo, los diferentes elementos estructurales de la proteína, específicamente la presencia y posición de residuos específicos en el andamiaje del dominio CYP bacteriano. Por otro lado, utilizamos una amplia batería de métodos bioinformáticos para obtener un panorama global de la función de estas enzimas y desarrollar una estrategia

que permita predecir -o al menos acotar- la función específica de un dado CYP tomando cómo punto de partida solamente la información de su secuencia.

Entre los métodos bioinformáticos destacan por un lado el uso de herramientas de docking, que cómo demuestran los resultados pueden ser combinadas con análisis de conservación para sesgar cierto tipo de interacciones claves en la determinación de la especificidad de sustratos para ciertos subgrupos de CYPs, y también con modelos por homología, que si bien poseen una menor capacidad para la determinación de la estructura del complejo CYP-sustrato de manera precisa, pueden funcionar como un tamiz, posibilitando descartar aquellos tipos de sustratos que no caben en el sitio de unión. En este sentido, los resultados presentados en esta tesis, muestran que los métodos de docking poseen una capacidad que excede su objetivo de diseño inicial -predecir la estructura de complejos proteína-ligando y cribado virtual de compuestos para la búsqueda de nuevos fármacos- en un contexto de análisis bioinformático integral para la predicción de la relación estructura-función.

El otro método destacado comprende la utilización de la información genómica, en este caso el contexto génico, que combinado con un enfoque de integración de datos con diferentes naturalezas ontológicas (similitud de estructura química, organización de los genes en el genoma, vías metabólicas) derivados de bases de datos públicas mostró un excelente desempeño en cuanto a su contribución para la propuesta y asignación de posibles sustratos de un dado CYP.

Finalmente es importantísimo destacar, que en todas las etapas del desarrollo de este trabajo, el uso de herramientas de visualización fue clave para poder integrar la información, por ejemplo permitió mapear la diversidad de sustratos de los CYPs sobre la diversidad de secuencias y estructuras como se muestra en la Figura 4.6.

7.1 La predicción de función de los CYPs bacterianos

Más específicamente, el hilo conductor de esta tesis se centró entonces en predecir y entender la función de los CYPs bacterianos. Al encarar dicho objetivo, una de las primeras preguntas que nos propusimos responder fue acerca de la diversidad de los citocromos bacterianos conocidos. Siendo que nuestro objeto de estudio se basa en la estructura proteica, la pregunta obligada que viene de la mano tiene que ver con la disponibilidad de estructuras. Específicamente nos interesa saber ¿qué tan diversas son las secuencias de los CYPs disponibles? y por supuesto, ¿cuántas (y cuáles) de ellas tienen estructura?. La búsqueda de respuesta a estas preguntas, nos condujo en primer lugar a realizar un estudio filogenético que permitió establecer una clasificación de los CYPs bacterianos en 15 grupos, 14 de los cuales poseen al menos un miembro de estructura tridimensional conocida. Este estudio mostró que si bien el número de estructuras conocidas es bajo (<0.4 %), la diversidad de secuencias se encuentra bastante bien representada por las mismas, lo cual se traduce en que se cuenta con la capacidad de generar un modelo estructural de muy buena calidad (utilizando moldes que se encuentran a distancias filogenéticas cortas) para alrededor del 50 % de los CYPs conocidos. Profundizando el análisis, también es importante notar, que si consideramos aquellos CYPs para los cuales existe una propuesta funcional (no verificada) o es posible inferirla a través del análisis del contexto génico, arribamos a la conclusión de que para aproximadamente el 3 % de los CYPs conocidos es posible por medio de la estrategia de docking desarrollada en este trabajo, arribar a un modelo del potencial complejo enzima-sustrato, y por lo tanto acotar la especificidad de su función.

Una vez caracterizada la diversidad de CYPs disponibles, nos abocamos a dar respuesta a otro interrogante: ¿es posible predecir el (o los) posible(s) sustratos de un dado CYP partiendo solamente de la información de la secuencia?. Para ello primero abordamos la siguiente pregunta ¿cuál es la relación entre la similitud de las secuencias de dos CYPs y la similitud química de sus sustratos?. Para evaluarla, realizamos un análisis de la diversidad química de los ligandos en relación con la distancia evolutiva entre las proteínas, el cual arrojó como resultado que no hay correlación significativa entre similitud de secuencia y similitud de sustratos, por lo que resulta imposible utilizar "solamente" la similitud en secuencia para asignar el sustrato y por lo tanto la función a un dado CYP. Esto se ve hermosamente reflejado al mapear las estructuras químicas de los sustratos sobre el árbol filogenético (Figura 4.6), donde se observa que sustratos de naturaleza química similar -o aún idénticos- se

unen a CYPs pertenecientes a grupos filogenéticos bien diferentes alejados evolutivamente.

Acto seguido nos planteamos como estrategia alternativa evaluar la información del contexto génico a los fines de conseguir algún indicio sobre la naturaleza química del sustrato que une. Este enfoque demostró ser exitoso en varios casos de testeo, como se describió en el capítulo 6. Uno de los ejemplos "estrella" es el caso de la CYP170, cuyo sustrato es producido directamente por la enzima codificada por el gen río arriba del que codifica para la CYP170, por lo que aquí el análisis del contexto génico nos conduce directamente al sustrato. Si bien existen otras situaciones en las que los indicios son un poco más indirectos, como se describe en el caso del CYP125, en el cual una porción de la estructura del sustrato predicho se parece al real, estos, junto a otros resultados similares evidencian que el análisis del contexto génico constituye un recurso muy valioso para echar luz sobre los posibles sustratos de una enzima. Su limitación actualmente está dada únicamente por el bajo nivel de anotación que presentan los genomas en general, situación que seguramente se irá revirtiendo con el transcurso del tiempo.

Ya exploradas las diversas vías que nos permiten contar con una estructura del receptor y del correspondiente sustrato potencial, el próximo paso fue explorar la capacidad de los métodos que permiten arribar a la estructura del complejo enzima-sustrato, dado que la determinación de ésta constituye un paso clave cuando se desea realizar inferencias sobre la reacción que cataliza una enzima. En esta línea, se evaluaron los métodos de docking, -que constituyen una de las metodologías más utilizadas para predecir la estructura de un complejo enzima-sustrato- en dos sentidos: por un lado el clásico, que tiene que ver con predecir el modo de interacción enzima-sustrato y adicionalmente -de manera análoga a un ensayo de virtual screening- se evaluó su capacidad para encontrar dada una enzima cuál es el sustrato correcto entre una serie de sustratos potenciales, es decir, su selectividad. Los resultados obtenidos del redocking de los complejos disponibles, muestran que el programa Autodock4 es capaz de predecir de manera adecuada la estructura de complejos CYP-sustrato, en la mayoría de los casos.

En el caso de utilizarse un modelo de la estructura del receptor para la predicción de la estructura del complejo enzima-sustrato, se observó que es un requisito fundamental que el modelo sea de alta calidad, lo cual está correlacionado con que el molde empleado para generarlo esté a una distancia evolutiva cercana o mediana. Por esta razón, como se mencionó anteriormente, solo consideramos que podemos obtener un modelo estructural para aproximadamente la mitad de los CYPs conocidos. Cuando se trata de enzimas que unen ligandos de gran tamaño, existe un requisito adicional para arribar a un modelo de calidad, que es que el modelo logre reproducir el tamaño de la cavidad de unión del ligando. Para esto es clave elegir como molde la estructura de un complejo cuyo ligando sea grande y considerar el espacio ocupado por el ligando al construir el modelo, o sea utilizar como molde no solo la estructura de la enzima sino el complejo enzima-ligando. En cuanto a la selectividad por el tipo de ligando correcto, se ha visto que realizando un conjunto de experimentos de docking con una serie de ligandos lo suficientemente diversos, o bien es posible identificar un ligando con altas chances de ser el correcto (un outlier claro), como se mostró para el caso de la p450cam; o al menos se puede tener una idea del tipo de ligando que podría unir la proteína, como se evidenció en el caso de la p450eryK. Viendo estos resultados, la siguiente pregunta tiene que ver con qué ocurre con la selectividad cuando se trabaja con modelos de un receptor. Aquí nuestros resultados muestran que lógicamente la selectividad disminuye, pero sin embargo para modelos de buena calidad aún es posible encontrar el ligando correcto dentro del conjunto de 2-3 ligandos candidatos. Esto significa que aún sin tener estructura resuelta, si se consigue tener un modelo de buena calidad, hay altas chances de poder predecir el tipo de sustrato que se unirá a un determinado CYP. En conjunto, los resultados obtenidos en los capítulos 4, 5 y 6 de ésta tesis, constituyen una muestra clara de lo que se puede conseguir integrando diversos métodos bioinformáticos, cuya sinergia resulta fundamental para poder responder a interrogantes como los planteados.

Finalmente, y una vez obtenido el complejo enzima-sustrato, resulta más asequible la tarea de plantear hipótesis acerca de la reacción que podría catalizar un CYP en cuestión. El fundamento que respalda esta afirmación, es que en las hemoproteínas los átomos involucrados en la reacción son los que se encuentran más próximos al átomo de hierro hémico, entonces al lograr determinar cuál es la posición en la que se une el sustrato a la enzima, se acota muchísimo el rango de posibles reacciones. Conociendo cuáles son los átomos que caen dentro

del radio de distancia al hierro hémico en el que típicamente se encuentra el átomo reactivo en los CYPs, se plantean las diversas hipótesis funcionales. Luego es preciso evaluar la factibilidad de cada una de ellas, y para ello empleamos métodos de simulación computacional tanto clásicos como híbridos cuántico-clásicos. En esta tesis, nos propusimos estudiar un caso difícil, el mecanismo de reacción del CYP121 de *Mt*, una enzima cuya reacción de acoplamiento C-C es muy poco común dentro de la familia. Esta reacción, por su complejidad se evaluó en tres etapas, y para cada una de ellas se testearon varias hipótesis alternativas hasta que finalmente pudimos arribar a una propuesta mecanística detallada y completa del mecanismo de reacción en la cual se resalta el rol clave de la matriz proteica de *MtCYP121*.

En resumen, la tarea de asignar función a los productos génicos es un tópico de importancia primordial para la biología. Aún con la abundancia de datos de secuencia con que se cuenta en la era post-genómica, sólo una pequeña fracción de los productos de los genes se han anotado experimentalmente. Por lo tanto, constituye un reto para la biología computacional poder predecir la función -más allá de la asignación de la proteína en una dada familia particular- en base a la secuencia, estructura y otros datos cuando las anotaciones experimentales no están disponibles. En este sentido, la relevancia de los resultados obtenidos en este trabajo de tesis, es clave ya que constituye un aporte a la difícil tarea de descifrar la función de los genes.

7.2 Perspectivas futuras

Durante la realización de este trabajo de tesis muchas veces sentí que las preguntas que se iban abriendo eran muchas más de las que iba a ser capaz de responder en el tiempo disponible, es así que en un momento me di cuenta de que para conseguir llegar a buen puerto, debía aceptar tal limitación y acotar mi trabajo a lo realizable. En esa instancia pude ver que aún con todo eso que quedaba por hacer, si me detenía a mirar el vaso lleno, veía que todo lo que sí logré hacer es un montón, significó para mí un aprendizaje enorme y me enriqueció muchísimo. Es entonces en esta sección, en donde dejaré plasmadas algunas de esas ideas que fueron surgiendo como extensiones de este trabajo y que por una simple cuestión de tiempo no son parte de esta tesis.

Comenzaré mencionando uno de los puntos (o aspectos) que considero podrían profundizarse o mejorarse, es el de la predicción de función por información del contexto génico. En este trabajo la información de las enzimas, vecinas de los CYPs, y sus reacciones asociadas se obtuvo "solamente" de la base de datos KEGG, por la simple razón de que en ella fue posible acceder de manera programática a la información de los genes vecinos, algo que en MetaCyc por ejemplo, no se consiguió realizar. Esto se debió en gran parte a la falta de documentación de los métodos que permiten hacer consultas en tales bases de datos. Sin embargo, considero que a futuro sería interesante **incorporar información de otras bases de datos**, como MetaCyc o ENZYME así como también **desarrollar algún método alternativo más robusto para obtener cuáles son los genes vecinos de un gen dado**, para de este modo independizarnos de que tal información esté disponible en una base de datos particular. Esto si bien suena interesante, hay que tener en cuenta que aumenta mucho la complejidad de diseño de la herramienta, debido a que se deben tomar decisiones sobre qué hacer en caso de que la información no esté disponible en el campo de la base de datos que se espera, o en caso de que la información proveniente de distintas bases de datos no coincida, hay que decidir cuál se elige, etc. De cualquier modo la creciente cantidad de información genómica y el uso complementario de diferentes fuentes de datos se espera aumenten el poder predictivo de nuestra estrategia.

Otro de los puntos en los que podrían introducirse mejoras tiene que ver con los métodos empleados para la predicción del complejo enzima-sustrato (docking). En este trabajo vimos que si bien el Autodock otorga resultados satisfactorios, existe un amplio espacio para mejoras. Ejemplo de estas ideas fue la implementación de un **sesgo en base a las interacciones enzima-ligando** observadas en complejos cuyos ligandos comparten un determinado grupo funcional. Este enfoque mostró una mejora notable en los resultados del docking, podría ser extendido a otros casos. Adicionalmente, otra mejora a implementar podría ser la **utilización de docking flexible**, o la aproximación al mismo implementada en colaboración con otros miembros del grupo de investigación, que permite considerar un ensamble de conformaciones del receptor (obtenidas por dinámica molecular) en vez

de una sola.

Por otro lado, y mas allá de las mejoras propuestas, un punto importante del desarrollo futuro reside en la aplicabilidad y disponibilidad de la estrategia desarrollada para su uso por parte de terceros. En este sentido y para ampliar la usabilidad e impacto de la misma en la comunidad científica, pensamos que sería útil **implementar una aplicación web para permitir y facilitar el uso de las herramientas y análisis realizados a otros investigadores.**

Finalmente a continuación proponemos algunas aplicaciones basadas en el uso de la estrategia desarrollada.

7.3 Aplicaciones

Como se ha mencionado en la introducción, un organismo típicamente cuenta con decenas de CYPs diferentes, en tal sentido los desarrollos materializados en esta tesis, serían de gran utilidad por ejemplo en aquellos estudios que buscan determinar la correspondencia entre una determinada función (que se conoce que realiza el organismo y que además se propone está asociada a un CYP), con alguno de los CYPs presentes en el mismo. Dicho de otro modo, esta herramienta permitiría **decidir, o al menos orientar, acerca de cuál/es de los CYPs con los que cuenta el organismo, más probablemente realice/n la función observada.**

En esta misma línea, una de las extensiones o aplicaciones naturales de esta tesis, sería la **validación in-vitro de la predicción realizada para uno (o un par) de CYPs.** Lo que se buscaría es tomando como punto de partida el (o los) potenciales sustratos predichos para un dado CYP de interés, su expresión-purificación in-vitro, y luego la determinación de la capacidad del mismo de unir y modificar químicamente alguno de los sustratos propuestos -posiblemente por técnicas de química analítica-. En este punto sería importante considerar factores adicionales como la disponibilidad comercial del sustrato y la factibilidad de producción in-vitro del CYP.

Otra de las potenciales aplicaciones tiene que ver con la **incorporación de las herramientas generadas en este trabajo a la mejora en la predicción de las vías metabólicas presentes en un organismo.** Esta tarea comprende dos subtareas: por un lado la predicción del reactoma, es decir, dado el genoma anotado de un organismo, predecir todas las asociaciones entre las enzimas y las reacciones que metabólicas catalizan; y en segundo término el problema de predicción de vías metabólicas: dado el reactoma de un organismo y su genoma anotado, predecir el conjunto de vías metabólicas presentes en el organismo. La herramienta aquí descrita, aportaría fundamentalmente a la etapa de predicción del reactoma, la cual es fundamental para proceder luego con la predicción de las rutas metabólicas.

Finalmente, es importante destacar que en cuanto a la predicción de función en general, este trabajo de tesis constituye una prueba de concepto de la hipótesis que asevera que la función puede ser inferida partiendo solamente de la información contenida en la secuencia, a través de la determinación de propiedades estructurales claves para la reactividad química. Es por ello que considero que este trabajo deja abierta la posibilidad a futuro de **extender la metodología aquí empleada a otras familias de proteínas.**

PUBLICACIONES

Durante el tiempo de realización de esta tesis doctoral se originaron las siguientes publicaciones:

- Lopez, E. D., Arcon, J. P., Gauto, D. F., Petruk, A. A., Modenutti, C. P., **Dumas, V. G.**, Marti, M. A., and Turjanski, A. G. *Watclust: a tool for improving the design of drugs based on protein-water interactions*. Bioinformatics 31, 22 (2015), 3697-3699.
- Romero, J. M., Martin, M., Ramirez, C. L., **Dumas, V. G.**, and Marti, M. A. *Efficient calculation of enzyme reaction free energy profiles using a hybrid differential relaxation algorithm*. Combined Quantum Mechanical and Molecular Mechanical Modelling of Biomolecular Interactions (2015), 33-65.
- **Dumas, V. G.**, Defelipe, L. A., Petruk, A. A., Turjanski, A. G., and Marti, M. A. *QM/MM study of the C-C coupling reaction mechanism of cyp121, an essential cytochrome p450 of Mycobacterium tuberculosis*. Proteins 82, 6 (2013), 1004-1021.
- Gauto, D., Modenutti, C., **Dumas, V.**, Álvarez, L., Bustamante, J. P., Turjanski, A., and Marti, M. *Determining free energies of protein-ligand binding, and association/dissociation processes using computer simulations*. World Research Journal of Peptide and Protein 1, 1 (2012), 21-32.

AGRADECIMIENTOS

La realización de este trabajo de tesis no hubiese sido posible sin el apoyo de un gran número de personas e instituciones, es por eso que deseo dedicar las próximas líneas a expresarles mi más sincero agradecimiento.

En primer lugar quiero agradecer a la Universidad Pública, que me ha dado la posibilidad de formarme profesionalmente en esta carrera que elegí hace ya diez años, y a los que trabajan día a día para sostenerla, para que nunca deje de ser pública y gratuita, para todos.

Al CONICET, que me otorgó el sustento durante el desarrollo de mi doctorado.

A los miembros del jurado, que muy amablemente han tomado la responsabilidad de evaluar mi trabajo.

A mi director, Marcelo, en primer lugar por haber confiado en mí para la realización de este trabajo de tesis, por la motivación constante que supiste transmitirme, aún en los momentos más críticos. Gracias por guiarme y también por darme la libertad para explorar diversos caminos hacia la resolución de un problema. Por darme la posibilidad de viajar y enriquecerme en el trabajo con pares. Quiero agradecerle muy especialmente por tu comprensión y por ayudarme a compatibilizar el trabajo con mis tareas de madre, aceptando que a veces trabaje desde casa en horas de la madrugada en vez del horario convencional. No solo sos un gran científico sino fundamentalmente una gran persona, siempre dispuesto a aconsejarme cuando lo necesité. Por todo esto, ha sido un gusto para mi trabajar con vos.

A todo el grupo ya multitudinario de simuladores y bioinformáticos con los que tuve la suerte de compartir el trabajo diario en estos cinco años y medio, gracias por estar siempre dispuestos a colaborar en lo que sea, por la buena onda de todos los días, por sus chistes y ocurrencias de las que tanto nos reímos, en síntesis, gracias por hacer que ir a trabajar sea tan divertido. Dentro de este gran grupo, quiero hacer mención especial a tres personas con las que más allá del trabajo pude construir una hermosa amistad. A Clau, Lucía y Noosh. Realmente podría escribir un largo párrafo en agradecimiento a cada una, pero lo voy a resumir diciéndoles: gracias por estar en todas, por ser mis psicólogas, por el aguante de siempre. Noosh, no me alcanzan las palabras para agradecer tu ayuda durante nuestros días en el improvisado jardín Encalada. No solo cuidaste a León como si fuese tu propio hijo sino que además me aconsejaste en mi trabajo. Sos una gran madre, una gran científica y una excelente amiga, es un lujo haber podido compartir este tiempo con vos.

Quiero agradecer a mis padres, por ser mi modelo de lucha y sacrificio. A mis hermanos, Flo, Joaco y Pau, gracias por apoyarme siempre.

A mis amigas de Franck, las de la escuela, las de toda la vida. Gracias porque aún a cientos de kilómetros nunca dejan de estar presentes.

A mis suegros, Ana y Sergio, y a mi cuñada Belu, gracias por el aguante.

Y por último, el más especial de los agradecimientos es para las dos personas con las que comparto todos los días de mi vida: A mi compañero Juan, infinitas gracias por acompañarme, aconsejarme, por ser mi sostén y por ayudarme a levantarme en mis tropiezos. Sin tu ayuda, realmente no hubiese sido posible llegar hasta aquí. Y a mi pequeño hijo León, mi principal motivación, el que hace que me esfuerce todos los días para dar lo mejor de mí, gracias por regalarme siempre esa sonrisa hermosa y mágica capaz de transformarlo todo.

A todos, ¡muchas gracias!

BIBLIOGRAFÍA

- [1] A D Miranker. Mass spectrometry of proteins of known mass. *Proceedings of the National Academy of Sciences of the United States of America*, 97(26):14025–7, dec 2000. ISSN 0027-8424. doi: 10.1073/pnas.011526498. URL <http://www.ncbi.nlm.nih.gov/pubmed/11114170><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC34088>.
- [2] Russ B Altman and Soumya Raychaudhuri. Whole-genome expression analysis: challenges beyond clustering. *Current Opinion in Structural Biology*, 11(3):340–347, 2001. ISSN 0959440X. doi: 10.1016/S0959-440X(00)00212-8.
- [3] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick, N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhuetter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigó, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, X Zhu, R. L. Sinsheimer, F. Sanger, P. H. Seeburg, E. C. Strauss, J. A. Kobori, G. Siu, L. E. Hood, J. Gocayne, A. Martin-Gallardo, W. R. McCombie, M. A. Jensen, M. D. Adams, M. D. Adams, M. D. Adams, A. R. Kerlavage, C. Fields, J. C. Venter, M. D. Adams, M. B. Soares, A. R. Kerlavage, C. Fields, J. C. Venter, M. H. Polymeropoulos, M. Marra, M. D. Adams, O. White, F. Sanger, A. R. Coulson, G. F. Hong, D. F. Hill, G. B. Petersen, B. W. J. Mahy, J. J. Esposito, J. C. Venter, R. D. Fleischmann, C. M. Fraser, C. J. Bult, J. F. Tomb, H. P. Klenk, J. C. Venter, H. O. Smith, L. Hood, H. Schmitt, S. Zhao, X. Lin, J. L. Weber, E. W. Myers, P. Green, E. Pennisi, J. C. Venter, M. D. Adams, E. Marshall, E. Pennisi, M. D. Adams, G. M. Rubin, E. W. Myers, F. S. Collins, F. Sanger, S. Nicklen, A. R. Coulson, J. M. Prober, G. Myers, S. Selznick, Z. Zhang, W. Miller, M. Hattori, I. Dunham, A. B. Carvalho, B. P. Lazzaro, A. G. Clark, G. D. Schuler, S. F. Altschul, W. Gish, W. Mi-

ller, E. W. Myers, D. J. Lipman, M. Olivier, N. Chaudhari, W. E. Hahn, R. J. Milner, J. G. Sutcliffe, D. Dickson, B. Ewing, P. Green, H. Roest Crollius, K. D. Pruitt, K. S. Katz, H. Sicotte, D. R. Maglott, E. C. Uberbacher, Y. Xu, R. J. Mural, C. Burge, S. Karlin, R. J. Mural, A. A. Salamov, V. V. Solovyev, G. L. Miklos, B. John, U. Francke, J. E. Horvath, S. Schwartz, E. E. Eichler, W. A. Bickmore, A. T. Sumner, G. P. Holmquist, G. Bernardi, S. Zoubak, O. Clay, G. Bernardi, S. Ohno, K. W. Broman, J. C. Murray, V. C. Sheffield, R. L. White, J. L. Weber, M. J. McEa-
chern, A. Krauskopf, E. H. Blackburn, A. Bird, M. Gardiner-Garden, M. Frommer, F. Larsen, G. Gundersen, R. Lopez, H. Prydz, S. H. Cross, A. Bird, C. Grunau, W. Hindermann, A. Rosenthal, F. Antequera, A. Bird, S. H. Cross, D. Slavov, A. F. Smit, A. D. Riggs, D. J. Elliott, A. V. Makeyev, A. N. Chkheidze, S. A. Lievhäber, Y. Pan, W. K. Decker, A. H. H. M. Huq, W. J. Craig, P. Nouvel, I. Goncalves, L. Duret, D. Mouchiroud, T. F. Smith, M. S. Waterman, A. L. Delcher, B. J. Trask, D. Sharon, W. B. Barbazuk, A. McLysaght, A. J. Enright, L. Skra-
banek, K. H. Wolfe, D. W. Burt, L. Skrabanek, K. H. Wolfe, P. Taillon-Miller, Z. Gu, Q. Li, L. Hillier, P. Y. Kwok, P. Taillon-Miller, E. E. Piernot, P. Y. Kwok, D. Altshuler, G. T. Marth, M. Cargill, M. K. Halushka, J. Zhang, T. L. Madden, M. W. Nachman, V. L. Bauer, S. L. Crowell, C. F. Aquadro, D. A. Nickerson, L. Jorde, D. G. Wang, M. Przeworski, R. R. Hudson, A. Di Rienzo, S. Tavare, A. G. Clark, H. Kaessmann, F. Heissig, A. von Haeseler, S. Paabo, E. L. Sonnhammer, S. R. Eddy, R. Durbin, A. Bateman, C. P. Ponting, J. Schultz, F. Milpetz, P. Bork, D. A. Goodenough, J. A. Goliger, D. L. Paul, D. G. Wilkinson, F. Nakamura, R. G. Kalb, S. M. Strittmatter, P. J. Horner, F. H. Gage, P. Casaccia-Bonnel, C. Gu, M. V. Chao, S. Wang, B. A. Barres, M. Geppert, T. C. Sudhof, J. T. Littleton, H. J. Bellen, A. Maximov, T. C. Sudhof, I. Bezprozvanny, G. Lemke, N. Perrimon, M. Bernfield, U. Lindahl, M. Kusche-Gullberg, L. Kjellen, T. L. Hurskainen, S. Hirohata, M. F. Seldin, S. S. Apte, R. A. Black, J. M. White, L. Aravind, V. M. Dixit, E. V. Koonin, P. Garcia-Meunier, M. Etienne-Julan, P. Fort, M. Piechaczyk, F. Bonhomme, N. R. Mansur, K. Meyer-Siegler, J. C. Wurzer, M. A. Sirover, N. A. Tatton, N. Kenmochi, F. W. Chen, Y. A. Ioannou, H. O. Madsen, K. Poulsen, O. Dahl, B. F. Clark, J. P. Hjorth, D. M. Chambers, J. Peters, C. M. Abbott, A. Khalyfa, B. M. Carlson, J. A. Carlson, E. Wang, D. Aeschlimann, V. Thomazy, P. Munroe, S. M. Wu, W. F. Cheung, D. Frazier, D. W. Stafford, B. Furie, J. W. Kehoe, C. R. Bertozzi, T. Pawson, P. Nash, A. W. van der Velden, A. A. Thomas, C. M. Fraser, H. Tettelin, D. Brett, H. J. Muller, H. Kern, A. P. Feinberg, C. A. Collins, C. Guthrie, S. R. Eddy, Q. Wang, J. Khillan, P. Gadue, K. Nishikura, M. Holcik, N. Sonenberg, R. G. Korneluk, T. A. McKinsey, C. L. Zhang, J. Lu, E. N. Olson, E. Capanna, M. G. M. Romanini, J. Maynard Smith, D. Charlesworth, B. Charlesworth, M. T. Morgan, J. E. Bailey, R. Maleszka, H. G. de Couet, G. L. Miklos, G. L. Miklos, J. P. Crutchfield, K. Young, M. Gell-Mann, S. Lloyd, A. L. Barabasi, R. Albert, E. Colucci-Guyon, B. Ewing, P. Green, B. Ewing, L. Hillier, M. C. Wendl, P. Green, E. S. Lander, M. S. Waterman, A. Krogh, K. Sjö-
lander, K. Sjölander, A. Bairoch, R. Apweiler, R. L. Tatusov, M. Y. Galperin, D. A. Natale, and E. V. Koonin. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–51, feb 2001. ISSN 0036-8075. doi: 10.1126/science.1058040. URL <http://www.ncbi.nlm.nih.gov/pubmed/11181995>.

- [4] Philip E. Bourne and Helge Weissig, editors. *Structural Bioinformatics*. Methods of Biochemical Analysis. John Wiley Sons, Inc., Hoboken, NJ, USA, feb 2003. ISBN 9780471721208. doi: 10.1002/0471721204. URL <http://doi.wiley.com/10.1002/0471721204>.
- [5] Lawrence Shapiro and Tim Harris. Finding function through structural genomics. *Current Opinion in Biotechnology*, 11(1):31–35, 2000. ISSN 09581669. doi: 10.1016/S0958-1669(99)00064-6.
- [6] J M Thornton, a E Todd, D Milburn, N Borkakoti, and C a Orengo. From structure to function: approaches and limitations. *Nature structural biology*, 7 Suppl(november):991–994, nov 2000. ISSN 1072-8368. doi: 10.1038/80784. URL <http://www.nature.com/doifinder/10.1038/80784>.
- [7] K. O. Wrzeszczynski, Y. Ofran, B. Rost, R. Nair, and J. Liu. Automatic prediction of protein function. *Cellular and Molecular Life Sciences (CMLS)*, 60(12):2637–2650, dec 2003. ISSN 1420-682X. doi: 10.1007/s00018-003-3114-8. URL <http://link.springer.com/10.1007/s00018-003-3114-8>.
- [8] Iddo Friedberg. Automated protein function prediction—the genomic challenge. *Briefings in bioinformatics*, 7(3):225–42, sep 2006. ISSN 1467-5463. doi: 10.1093/bib/bbl004. URL <http://www.ncbi.nlm.nih.gov/pubmed/16772267>.

- [9] C Chothia and A M Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, 5(4):823–6, apr 1986. ISSN 0261-4189. URL <http://www.ncbi.nlm.nih.gov/pubmed/3709526><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1166865>.
- [10] P. Bork. Powers and Pitfalls in Sequence Analysis: The 70 *Genome Research*, 10(4):398–400, apr 2000. ISSN 10889051. doi: 10.1101/gr.10.4.398. URL <http://www.genome.org/cgi/doi/10.1101/gr.10.4.398>.
- [11] Walter R. Gilks, Benjamin Audit, Daniela de Angelis, Sophia Tsoka, and Christos A. Ouzounis. Percolation of annotation errors through hierarchically structured protein sequence databases. *Mathematical Biosciences*, 193(2):223–234, 2005. ISSN 00255564. doi: 10.1016/j.mbs.2004.08.001.
- [12] R. W. Estabrook. A PASSION FOR P450s (REMEMBRANCES OF THE EARLY HISTORY OF RESEARCH ON CYTOCHROME P450). *Drug Metabolism and Disposition*, 31(12):1461–1473, dec 2003. ISSN 0090-9556. doi: 10.1124/dmd.31.12.1461. URL <http://dmd.aspetjournals.org/cgi/doi/10.1124/dmd.31.12.1461>.
- [13] Tsuneo Omura and Ryo Sato. The Carbon Monoxide-binding Pigment of Liver Microsomes I. EVIDENCE FOR ITS HEMOPROTEIN NATURE. *THE JOURNAL OF BIOLOGICAL CHEMISTRY*, 239(7), 1964.
- [14] P Stasiecki, F Oesch, G Bruder, E D Jarasch, and W W Franke. Distribution of enzymes involved in metabolism of polycyclic aromatic hydrocarbons among rat liver endomembranes and plasma membranes. *European journal of cell biology*, 21(1):79–92, apr 1980. ISSN 0171-9335. URL <http://www.ncbi.nlm.nih.gov/pubmed/6769676>.
- [15] F Oesch, P Bentley, M Golan, and P Stasiecki. Metabolism of benzo(a)pyrene by subcellular fractions of rat liver: evidence for similar patterns of cytochrome P-450 in rough and smooth endoplasmic reticulum but not in nuclei and plasma membrane. *Cancer research*, 45(10):4838–43, oct 1985. ISSN 0008-5472. URL <http://www.ncbi.nlm.nih.gov/pubmed/3875401>.
- [16] I C Gunsalus and G C Wagner. Bacterial P-450cam methylene monooxygenase components: cytochrome m, putidaredoxin, and putidaredoxin reductase. *Methods in enzymology*, 52:166–88, 1978. ISSN 0076-6879. URL <http://www.ncbi.nlm.nih.gov/pubmed/672627>.
- [17] D W Nebert, M Adesnik, M J Coon, R W Estabrook, F J Gonzalez, F P Guengerich, I C Gunsalus, E F Johnson, B Kemper, and W Levin. The P450 gene superfamily: recommended nomenclature. *DNA (Mary Ann Liebert, Inc.)*, 6(1):1–11, feb 1987. ISSN 0198-0238. doi: 10.1089/dna.1987.6.1. URL <http://www.ncbi.nlm.nih.gov/pubmed/3829886>.
- [18] Mary A. Schuler and Stephen G. Sligar. Diversities and similarities in P450 systems. *Metal Ions in Life Sciences*, 3:1–26, 2007. ISSN 1559-0836.
- [19] Sason Shaik, Hajime Hirao, and Devesh Kumar. Reactivity patterns of cytochrome P450 enzymes: multifunctionality of the active species, and the two states-two oxidants conundrum. *Natural product reports*, 24(3):533–52, jun 2007. ISSN 0265-0568. doi: 10.1039/b604192m. URL <http://www.ncbi.nlm.nih.gov/pubmed/17534529>.
- [20] F. Peter Guengerich. Cytochrome p450 enzymes in the generation of commercial products. *Nature Reviews Drug Discovery*, 1(5):359–366, may 2002. ISSN 1474-1776. doi: 10.1038/nrd792. URL <http://www.nature.com/doifinder/10.1038/nrd792>.
- [21] V. B. Urlacher, S. Lutz-Wahl, and R. D. Schmid. Microbial P450 enzymes in biotechnology. *Applied Microbiology and Biotechnology*, 64(3):317–325, apr 2004. ISSN 0175-7598. doi: 10.1007/s00253-003-1514-1. URL <http://link.springer.com/10.1007/s00253-003-1514-1>.
- [22] Mattijs K Julsing, Sjef Cornelissen, Bruno Bühler, and Andreas Schmid. Heme-iron oxygenases: powerful industrial biocatalysts? *Current Opinion in Chemical Biology*, 12(2):177–186, 2008. ISSN 13675931. doi: 10.1016/j.cbpa.2008.01.029.

- [23] K Sonomoto, M M Hoq, A Tanaka, and S Fukui. 11beta-Hydroxylation of Cortexolone (Reichstein Compound S) to Hydrocortisone by *Curvularia lunata* Entrapped in Photo-Cross-Linked Resin Gels. *Applied and environmental microbiology*, 45(2):436–43, feb 1983. ISSN 0099-2240. URL <http://www.ncbi.nlm.nih.gov/pubmed/16346194><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC242304>.
- [24] Tadashi FUJII, Yoshikazu FUJII, Kazuhiro MACHIDA, Atsushi OCHIAI, and Masashi ITO. Efficient Biotransformations Using *Escherichia coli* with *tolC* *acrAB* Mutations Expressing Cytochrome P450 Genes. *Bioscience, Biotechnology, and Biochemistry*, 73(4):805–810, apr 2009. ISSN 0916-8451. doi: 10.1271/bbb.80627. URL <http://www.tandfonline.com/doi/full/10.1271/bbb.80627>.
- [25] A Shafiee and C R Hutchinson. Macrolide antibiotic biosynthesis: isolation and properties of two forms of 6-deoxyerythronolide B hydroxylase from *Saccharopolyspora erythraea* (*Streptomyces erythraeus*). *Biochemistry*, 26(19):6204–10, sep 1987. ISSN 0006-2960. URL <http://www.ncbi.nlm.nih.gov/pubmed/2446657>.
- [26] S Jennewein, C D Rithner, R M Williams, and R B Croteau. Taxol biosynthesis: taxane 13 alpha-hydroxylase is a cytochrome P450-dependent monooxygenase. *Proceedings of the National Academy of Sciences of the United States of America*, 98(24):13595–600, nov 2001. ISSN 0027-8424. doi: 10.1073/pnas.251539398. URL <http://www.ncbi.nlm.nih.gov/pubmed/11707604><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC61086>.
- [27] Jan B van Beilen, René Holtackers, Daniel Lüscher, Ulrich Bauer, Bernard Witholt, and Wouter A Duetz. Biocatalytic production of perillyl alcohol from limonene by using a novel *Mycobacterium* sp. cytochrome P450 alkane hydroxylase expressed in *Pseudomonas putida*. *Applied and environmental microbiology*, 71(4):1737–44, apr 2005. ISSN 0099-2240. doi: 10.1128/AEM.71.4.1737-1744.2005. URL <http://www.ncbi.nlm.nih.gov/pubmed/15811996><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1082528>.
- [28] Santosh Kumar. Engineering cytochrome P450 biocatalysts for biotechnology, medicine and bioremediation. *Expert Opinion on Drug Metabolism Toxicology*, 6(2):115–131, feb 2010. ISSN 1742-5255. doi: 10.1517/17425250903431040. URL <http://www.tandfonline.com/doi/full/10.1517/17425250903431040>.
- [29] Stephen B. Lamb, David C. Lamb, Steven L. Kelly, and David C. Stuckey. Cytochrome P450 immobilisation as a route to bioremediation/biocatalysis. *FEBS Letters*, 431(3):343–346, jul 1998. ISSN 00145793. doi: 10.1016/S0014-5793(98)00771-6. URL <http://doi.wiley.com/10.1016/S0014-5793%2898%2900771-6>.
- [30] R. Nicoli, M. Bartolini, S. Rudaz, V. Andrisano, and J.-L. Veuthey. Development of immobilized enzyme reactors based on human recombinant cytochrome P450 enzymes for phase I drug metabolism studies. *Journal of Chromatography A*, 1206(1):2–10, 2008. ISSN 00219673. doi: 10.1016/j.chroma.2008.05.080.
- [31] Debashis Ghosh, Jennifer Griswold, Mary Erman, and Walter Pangborn. Structural basis for androgen specificity and oestrogen synthesis in human aromatase. *Nature*, 457(7226):219–223, jan 2009. ISSN 0028-0836. doi: 10.1038/nature07614. URL <http://www.nature.com/doifinder/10.1038/nature07614>.
- [32] V F Kalb, C W Woods, T G Turi, C R Dey, T R Sutter, and J C Loper. Primary structure of the P450 lanosterol demethylase gene from *Saccharomyces cerevisiae*. *DNA (Mary Ann Liebert, Inc.)*, 6(6):529–37, dec 1987. ISSN 0198-0238. doi: 10.1089/dna.1987.6.529. URL <http://www.ncbi.nlm.nih.gov/pubmed/3322742>.
- [33] Sudagar S. Gurucha, Kay Fowler, Gurydal S. Besra, Tobias Kieser, Andrew W. Munro, Alison Richmond, Iain S. Hunter, Kirsty J. McLean, and Ker R. Marshall. Azole antifungals are potent inhibitors of cytochrome P450 mono-oxygenases and bacterial growth in mycobacteria and streptomycetes. *Microbiology*, 148(10):2937–2949, oct 2002. ISSN 1350-0872. doi: 10.1099/00221287-148-10-2937. URL <http://mic.microbiologyresearch.org/content/journal/micro/10.1099/00221287-148-10-2937>.

- [34] Anil Koul, Eric Arnoult, Nacer Lounis, Jerome Guillemont, and Koen Andries. The challenge of new drug discovery for tuberculosis. *Nature*, 469(7331):483–490, jan 2011. ISSN 1476-4687. doi: 10.1038/nature09657. URL <http://www.ncbi.nlm.nih.gov/pubmed/21270886>.
- [35] Hugues Ouellet, Jonathan B Johnston, and Paul R Ortiz de Montellano. The Mycobacterium tuberculosis cytochrome P450 system. *Archives of biochemistry and biophysics*, 493(1):82–95, jan 2010. ISSN 1096-0384. doi: 10.1016/j.abb.2009.07.011. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2812603&tool=pmcentrez&rendertype=abstract>.
- [36] Elizabeth M.J. Gillam, Anna Marie A. Aguinaldo, Lisa M. Notley, Donghak Kim, Ralf G. Mundkowsky, Alexander A. Volkov, Frances H. Arnold, Pavel Souček, James J. DeVoss, and F.Peter Guengerich. Formation of Indigo by Recombinant Mammalian Cytochrome P450. *Biochemical and Biophysical Research Communications*, 265(2):469–472, nov 1999. ISSN 0006291X. doi: 10.1006/bbrc.1999.1702. URL <http://linkinghub.elsevier.com/retrieve/pii/S0006291X99917023>.
- [37] Heribert Warzecha, Andreas Frank, Markus Peer, Elizabeth M.J. Gillam, F. Peter Guengerich, and Matthias Unger. Formation of the indigo precursor indican in genetically engineered tobacco plants and cell cultures. *Plant Biotechnology Journal*, 5(1):185–191, jan 2007. ISSN 1467-7644. doi: 10.1111/j.1467-7652.2006.00231.x. URL <http://doi.wiley.com/10.1111/j.1467-7652.2006.00231.x>.
- [38] Elizabeth M. J. Gillam and F. Peter Guengerich. Exploiting the Versatility of Human Cytochrome P450 Enzymes: The Promise of Blue Roses From Biotechnology. *IUBMB Life (International Union of Biochemistry and Molecular Biology: Life)*, 52(6):271–277, dec 2001. ISSN 1521-6543. doi: 10.1080/152165401317291110. URL <http://doi.wiley.com/10.1080/152165401317291110>.
- [39] Nikitas Bistolas, Ulla Wollenberger, Christiane Jung, and Frieder W. Scheller. Cytochrome P450 biosensors—a review. *Biosensors and Bioelectronics*, 20(12):2408–2423, 2005. ISSN 09565663. doi: 10.1016/j.bios.2004.11.023.
- [40] Nada Bozina, Vlasta Bradamante, and Mila Lovrić. Genetic polymorphism of metabolic enzymes P450 (CYP) as a susceptibility factor for drug response, toxicity, and cancer risk. *Arhiv za higijenu rada i toksikologiju*, 60(2):217–42, jun 2009. ISSN 0004-1254. doi: 10.2478/10004-1254-60-2009-1885. URL <http://www.ncbi.nlm.nih.gov/pubmed/19581216>.
- [41] Partha Roy and David J. Waxman. Activation of oxazaphosphorines by cytochrome P450: Application to gene-directed enzyme prodrug therapy for cancer. *Toxicology in Vitro*, 20(2):176–186, 2006. ISSN 08872333. doi: 10.1016/j.tiv.2005.06.046.
- [42] Ling Chen and David Waxman. Cytochrome P450 Gene-directed Enzyme Prodrug Therapy (GDEPT) for Cancer. *Current Pharmaceutical Design*, 8(15):1405–1416, jul 2002. ISSN 13816128. doi: 10.2174/1381612023394566. URL <http://www.eurkaselect.com/openurl/content.php?genre=article&issn=1381-6128&volume=8&issue=15&page=1405>.
- [43] Marc Morant, Søren Bak, Birger Lindberg Møller, and Danièle Werck-Reichhart. Plant cytochromes P450: tools for pharmacology, plant protection and phytoremediation. *Current Opinion in Biotechnology*, 14(2):151–162, 2003. ISSN 09581669. doi: 10.1016/S0958-1669(03)00024-7.
- [44] P.C. Abhilash, Sarah Jamil, and Nandita Singh. Transgenic plants for enhanced biodegradation and phytoremediation of organic xenobiotics. *Biotechnology Advances*, 27(4):474–488, 2009. ISSN 07349750. doi: 10.1016/j.biotechadv.2009.04.002.
- [45] Ira N. Levine. *Quantum chemistry*. Prentice Hall, 2000. ISBN 9780136855125.

- [46] Axel D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648, 1993. ISSN 00219606. doi: 10.1063/1.464913. URL <http://scitation.aip.org/content/aip/journal/jcp/98/7/10.1063/1.464913>.
- [47] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized Gradient Approximation Made Simple. *Physical Review Letters*, 77(18):3865–3868, oct 1996. ISSN 0031-9007. doi: 10.1103/PhysRevLett.77.3865. URL <http://link.aps.org/doi/10.1103/PhysRevLett.77.3865>.
- [48] Viktor Hornak, Robert Abel, Asim Okur, Bentley Strockbine, Adrian Roitberg, and Carlos Simmerling. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins: Structure, Function, and Bioinformatics*, 65(3):712–725, 2006.
- [49] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An N log (N) method for Ewald sums in large systems. *The Journal of chemical physics*, 98(12):10089–10092, 1993.
- [50] Romelia Salomon-Ferrer, Andreas W Gotz, Duncan Poole, Scott Le Grand, and Ross C Walker. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *Journal of Chemical Theory and Computation*, 9(9):3878–3888, 2013.
- [51] Junmei Wang, Romain M Wolf, James W Caldwell, Peter A Kollman, and David A Case. Development and testing of a general amber force field. *Journal of computational chemistry*, 25(9):1157–1174, 2004.
- [52] Miyamoto S Kollman P A SETTLE. An analytical version of the SHAKE and RATTLE algorithm for rigid water molecules. *J. Comput. Chem*, 13:952–962, 1992.
- [53] Victoria G. Dumas, Lucas A. Defelipe, Ariel A. Petruk, Adrian G. Turjanski, and Marcelo A. Marti. QM/MM study of the C-C coupling reaction mechanism of CYP121, an essential Cytochrome p450 of Mycobacterium tuberculosis. *Proteins: Structure, Function, and Bioinformatics*, 82(6):1004–1021, jun 2014. ISSN 08873585. doi: 10.1002/prot.24474. URL <http://doi.wiley.com/10.1002/prot.24474>.
- [54] Lucas A Defelipe, Esteban Lanzarotti, Diego Gauto, Marcelo A Marti, and Adrián G Turjanski. Protein Topology Determines Cysteine Oxidation Fate: The Case of Sulfenyl Amide Formation among Protein Families. *PLoS computational biology*, 11(3):e1004051—e1004051, 2015.
- [55] Alejandro Crespo, Damián a. Scherlis, Marcelo a. Martí, Pablo Ordejón, Adrián E. Roitberg, and Darío a. Estrin. A DFT-based QM-MM approach designed for the treatment of large molecular systems: Application to chorismate mutase. *The Journal of Physical Chemistry B*, 107(49):13728–13736, dec 2003. ISSN 1520-6106. doi: 10.1021/jp036236h. URL <http://pubs.acs.org/doi/abs/10.1021/jp036236h>.
- [56] Alejandro Crespo, Marcelo A Martí, Susana G Kalko, Antonio Morreale, Modesto Orozco, Jose L Gelpi, F Javier Luque, and Darío A Estrin. Theoretical study of the truncated hemoglobin HbN: exploring the molecular basis of the NO detoxification mechanism. *Journal of the American Chemical Society*, 127(12):4433–4444, 2005.
- [57] Adrian Gustavo Turjanski, Gerhard Hummer, and J Silvio Gutkind. How mitogen-activated protein kinases recognize and phosphorylate their targets: A QM/MM study. *Journal of the American Chemical Society*, 131(17):6141–6148, 2009.
- [58] Juan Pablo Arcon, P Rosi, Ariel Alcides Petruk, Marcelo Adrian Marti, and Dario A Estrin. Molecular Mechanism of Myoglobin Autoxidation: Insights from Computer Simulations. *The Journal of Physical Chemistry B*, 119(5):1802–1813, 2015.
- [59] Andrew R. Leach. *Molecular modelling : principles and applications*. Longman, 1996. ISBN 9780582239333.

- [60] Loup Verlet. Computer Experiments on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules. *Physical Review*, 159(1):98–103, jul 1967. ISSN 0031-899X. doi: 10.1103/PhysRev.159.98. URL <http://link.aps.org/doi/10.1103/PhysRev.159.98>.
- [61] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman J.C Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, mar 1977. ISSN 00219991. doi: 10.1016/0021-9991(77)90098-5. URL <http://linkinghub.elsevier.com/retrieve/pii/0021999177900985>.
- [62] Herman J. C. Berendsen, J. P. M. Postma, Wilfred F. van Gunsteren, ARHJ DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *The Journal of chemical physics*, 81(8):3684–3690, 1984. ISSN 00219606. doi: 10.1063/1.448118. URL <http://scitation.aip.org/content/aip/journal/jcp/81/8/10.1063/1.448118>.
- [63] Marketa J. Zvelebil and Jeremy O. Baum. *Understanding bioinformatics*. Garland Science, 2008. ISBN 0815340249.
- [64] Warren L DeLano. The PyMOL molecular graphics system. 2002.
- [65] F Dardel. Mecanismo de generación de SMILES (By Original by Fdardel, slight edit by DMacks - Own work, CC BY-SA 3.0). URL <https://commons.wikimedia.org/w/index.php?curid=2556784>.
- [66] Inc DAYLIGHT Chemical Information Systems. SMILES - A Simplified Chemical Language. URL <http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>.
- [67] J.D. Holliday, C-Y. Hu, and P. Willett. Grouping of Coefficients for the Calculation of Inter-Molecular Similarity and Dissimilarity using 2D Fragment Bit-Strings.
- [68] T.T. Tanimoto. No Title. *IBM Internal Report*, 1957.
- [69] D J Rogers and T T Tanimoto. A Computer Program for Classifying Plants. *Science (New York, N.Y.)*, 132(3434): 1115–8, oct 1960. ISSN 0036-8075. doi: 10.1126/science.132.3434.1115. URL <http://www.ncbi.nlm.nih.gov/pubmed/17790723>.
- [70] Frank Eisenhaber, Bengt Persson, and Patrick Argos. Protein Structure Prediction: Recognition of Primary, Secondary, and Tertiary Structural Features from Amino Acid Sequence. <http://dx.doi.org/10.3109/10409239509085139>, 2008.
- [71] Sharon Goldsmith-Fischman and Barry Honig. Structural genomics: computational methods for structure analysis. *Protein science : a publication of the Protein Society*, 12(9):1813–21, sep 2003. ISSN 0961-8368. doi: 10.1110/ps.0242903. URL <http://www.ncbi.nlm.nih.gov/pubmed/12930981><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2323979>.
- [72] P N Lipke, M H Chen, H de Nobel, J Kurjan, and P C Kahn. Homology modeling of an immunoglobulin-like domain in the *Saccharomyces cerevisiae* adhesion protein alpha-agglutinin. *Protein science : a publication of the Protein Society*, 4(10):2168–78, oct 1995. ISSN 0961-8368. doi: 10.1002/pro.5560041023. URL <http://www.ncbi.nlm.nih.gov/pubmed/8535254><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2142996>.
- [73] Roberto Sánchez and Andrej Šali. Advances in comparative protein-structure modelling. *Current Opinion in Structural Biology*, 7(2):206–214, apr 1997. ISSN 0959440X. doi: 10.1016/S0959-440X(97)80027-9. URL <http://linkinghub.elsevier.com/retrieve/pii/S0959440X97800279>.
- [74] Jürgen Bajorath, Ronald Stenkamp, and Alejandro Aruffo. Knowledge-based model building of proteins: Concepts and examples. *Protein Science*, 2(11):1798–1810, nov 1993. ISSN 09618368. doi: 10.1002/pro.5560021103. URL <http://doi.wiley.com/10.1002/pro.5560021103>.

- [75] Marc A. Martí-Renom, Ashley C. Stuart, András Fiser, Roberto Sánchez, Francisco Melo And, and Andrej Šali. Comparative Protein Structure Modeling of Genes and Genomes. <http://dx.doi.org/10.1146/annurev.biophys.29.1.291>, 2003.
- [76] Andrej Šali and Tom L. Blundell. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *Journal of Molecular Biology*, 234(3):779–815, dec 1993. ISSN 00222836. doi: 10.1006/jmbi.1993.1626. URL <http://linkinghub.elsevier.com/retrieve/pii/S0022283683716268>.
- [77] Pascal Benkert, Silvio C E Tosatto, and Dietmar Schomburg. QMEAN: A comprehensive scoring function for model quality assessment. *Proteins: Structure, Function, and Bioinformatics*, 71(1):261–277, 2008.
- [78] Pascal Benkert, Michael Künzli, and Torsten Schwede. QMEAN server for protein model quality estimation. *Nucleic acids research*, 37(Web Server issue):W510–4, jul 2009. ISSN 1362-4962. doi: 10.1093/nar/gkp322. URL <http://www.ncbi.nlm.nih.gov/pubmed/19429685><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2703985>.
- [79] P. T. Lang, S. R. Brozell, S. Mukherjee, E. F. Pettersen, E. C. Meng, V. Thomas, R. C. Rizzo, D. A. Case, T. L. James, and I. D. Kuntz. DOCK 6: Combining techniques to model RNA-small molecule complexes. *RNA*, 15(6):1219–1230, jun 2009. ISSN 1355-8382. doi: 10.1261/rna.1563609. URL <http://rnajournal.cshlp.org/cgi/doi/10.1261/rna.1563609>.
- [80] Garrett M. Morris, David S. Goodsell, Robert S. Halliday, Ruth Huey, William E. Hart, Richard K. Belew, and Arthur J. Olson. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, 19(14):1639–1662, nov 1998. ISSN 0192-8651. doi: 10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B. URL <http://doi.wiley.com/10.1002/%28SICI%291096-987X%2819981115%2919%3A14%3C1639%3A%3AAID-JCC10%3E3.0.CO%3B2-B>.
- [81] Francisco J Solis, Roger, , and J-B Wets. MINIMIZATION BY RANDOM SEARCH TECHNIQUES*t. *MATHEMATICS OF OPERATIONS RESEARCH*, 6(1), 1981.
- [82] Laura Wesson and David Eisenberg. Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Science*, 1(2):227–235, dec 2008. ISSN 09618368. doi: 10.1002/pro.5560010204. URL <http://doi.wiley.com/10.1002/pro.5560010204>.
- [83] Mario Marsili Johann Gasteiger. Iterative partial equalization of orbital electronegativity—a rapid access to atomic charges. *Tetrahedron*, 36(22):3219–3228, 1980.
- [84] Marsili M Gasteiger J. pi-charge distribution from molecular topology and pi-orbital electronegativity. *Croat. Chem. Acta*, (53):601–614, 1981.
- [85] Pieter F. W. Stouten, Cornelius Frömmel, Haruki Nakamura, and Chris Sander. An Effective Solvation Term Based on Atomic Occupancies for Use in Protein Simulations. <http://dx.doi.org/10.1080/08927029308022161>, 2006.
- [86] STea Cole, R Brosch, J Parkhill, T Garnier, C Churcher, D Harris, S V Gordon, K Eiglmeier, S Gas, C E 3rd Barry, and Others. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, 393(6685):537–544, 1998.
- [87] Kirsty J. McLean, Daniel Clift, D. Geraint Lewis, Muna Sabri, Philip R. Balding, Michael J. Sutcliffe, David Leys, and Andrew W. Munro. The preponderance of P450s in the Mycobacterium tuberculosis genome. *Trends in Microbiology*, 14(5):220–228, 2006. ISSN 0966842X. doi: 10.1016/j.tim.2006.03.002.
- [88] Pascal Belin, Marie Hélène Le Du, Alistair Fielding, Olivier Lequin, Mickaël Jacquet, Jean-Baptiste Charbonnier, Alain Lecoq, Robert Thai, Marie Courçon, Cédric Masson, Christophe Dugave, Roger Genet, Jean-Luc Pernodet, and Muriel Gondry. Identification and structural basis of the reaction catalyzed by

- CYP121, an essential cytochrome P450 in *Mycobacterium tuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America*, 106(18):7426–31, may 2009. ISSN 1091-6490. doi: 10.1073/pnas.0812191106. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2678619&tool=pmcentrez&rendertype=abstract>.
- [89] David Leys, Christopher G Mowat, Kirsty J McLean, Alison Richmond, Stephen K Chapman, Malcolm D Walkinshaw, and Andrew W Munro. Atomic Structure of *Mycobacterium tuberculosis* CYP121 to 1.06 Å Reveals Novel Features of Cytochrome P450. *Journal of Biological Chemistry*, 278(7):5141–5147, feb 2003. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M209928200. URL <http://www.jbc.org/content/278/7/5141>.
- [90] H. E. Seward, A. Roujeinikova, K. J. McLean, A. W. Munro, and D. Leys. Crystal Structure of the *Mycobacterium tuberculosis* P450 CYP121-Fluconazole Complex Reveals New Azole Drug-P450 Binding Mode. *Journal of Biological Chemistry*, 281(51):39437–39443, dec 2006. ISSN 0021-9258. doi: 10.1074/jbc.M607665200. URL <http://www.jbc.org/cgi/doi/10.1074/jbc.M607665200>.
- [91] Sean A. Hudson, Kirsty J. McLean, Sachin Surade, Yong-Qing Yang, David Leys, Alessio Ciulli, Andrew W. Munro, and Chris Abell. Application of Fragment Screening and Merging to the Discovery of Inhibitors of the *Mycobacterium tuberculosis* Cytochrome P450 CYP121. *Angewandte Chemie International Edition*, 51(37):9311–9316, sep 2012. ISSN 14337851. doi: 10.1002/anie.201202544. URL <http://doi.wiley.com/10.1002/anie.201202544>.
- [92] Kirsty J McLean, Paul Carroll, D Geraint Lewis, Adrian J Dunford, Harriet E Seward, Rajasekhar Neeli, Myles R Cheesman, Laurent Marsollier, Philip Douglas, W Ewen Smith, Ida Rosenkrands, Stewart T Cole, David Leys, Tanya Parish, and Andrew W Munro. Characterization of active site structure in CYP121. A cytochrome P450 essential for viability of *Mycobacterium tuberculosis* H37Rv. *The Journal of biological chemistry*, 283(48):33406–16, nov 2008. ISSN 0021-9258. doi: 10.1074/jbc.M802115200. URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2662266&tool=pmcentrez&rendertype=abstract>.
- [93] Kirsty J McLean, Myles R Cheesman, Stuart L Rivers, Alison Richmond, David Leys, Stephen K Chapman, Graeme A Reid, Nicholas C Price, Sharon M Kelly, John Clarkson, W.Ewen Smith, and Andrew W Munro. Expression, purification and spectroscopic characterization of the cytochrome P450 CYP121 from *Mycobacterium tuberculosis*. *Journal of Inorganic Biochemistry*, 91(4):527–541, 2002. ISSN 01620134. doi: 10.1016/S0162-0134(02)00479-8.
- [94] K G Ravichandran, S S Boddupalli, C A Hasermann, J A Peterson, and J Deisenhofer. Crystal structure of hemoprotein domain of P450BM-3, a prototype for microsomal P450's. *Science (New York, N.Y.)*, 261(5122):731–6, aug 1993. ISSN 0036-8075. doi: 10.1126/science.8342039. URL <http://www.ncbi.nlm.nih.gov/pubmed/8342039>.
- [95] J. K. Yano, L. S. Koo, D. J. Schuller, H. Li, P. R. Ortiz de Montellano, and T. L. Poulos. Crystal Structure of a Thermophilic Cytochrome P450 from the Archaeon *Sulfolobus solfataricus*. *Journal of Biological Chemistry*, 275(40):31086–31092, oct 2000. ISSN 0021-9258. doi: 10.1074/jbc.M004281200. URL <http://www.jbc.org/cgi/doi/10.1074/jbc.M004281200>.
- [96] Matthieu Fonvielle, Marie Hélène Le Du, Olivier Lequin, Alain Lecoq, Mickaël Jacquet, Robert Thai, Steven Dubois, Guillaume Grach, Muriel Gondry, and Pascal Belin. Substrate and reaction specificity of *Mycobacterium tuberculosis* cytochrome P450 CYP121: Insights from biochemical studies and crystal structures. *Journal of Biological Chemistry*, 288(24), 2013. ISSN 00219258. doi: 10.1074/jbc.M112.443853.
- [97] Sason Shaik, Shimrit Cohen, Yong Wang, Hui Chen, Devesh Kumar, and Walter Thiel. P450 Enzymes: Their Structure, Reactivity, and Selectivity—Modeled by QM/MM Calculations. 2009.
- [98] Ariel A. Petruk, Silvina Bartesaghi, Madia Trujillo, Darío A. Estrin, Daniel Murgida, Balaraman Kalyanaraman, Marcelo A. Marti, and Rafael Radi. Molecular basis of intramolecular electron transfer in proteins during

- radical-mediated oxidations: Computer simulation studies in model tyrosine–cysteine peptides in solution. *Archives of Biochemistry and Biophysics*, 525(1):82–91, 2012. ISSN 00039861. doi: 10.1016/j.abb.2012.05.012.
- [99] Sharon Hammes-Schiffer and Alexander V. Soudackov. Proton-Coupled Electron Transfer in Solution, Proteins, and Electrochemistry†. 2008.
- [100] John Mongan, David A. Case, and J. Andrew McCammon. Constant pH molecular dynamics in generalized Born implicit solvent. *Journal of Computational Chemistry*, 25(16):2038–2048, dec 2004. ISSN 01928651. doi: 10.1002/jcc.20139. URL <http://doi.wiley.com/10.1002/jcc.20139>.
- [101] † Thomas Simonson, *, , § Jens Carlsson, ‡, and ‡ David A. Case*. Proton Binding to Proteins: pKa Calculations with Explicit and Implicit Solvent Models. 2004.
- [102] EJ Wood. Data for Biochemical Research (third edition). *Biochemical Education*, 15(2):97, apr 1987. ISSN 03074412. doi: 10.1016/0307-4412(87)90110-5. URL <http://linkinghub.elsevier.com/retrieve/pii/0307441287901105>.
- [103] Shina C. L. Kamerlin, Maciej Haranczyk, and Arieh Warshel. Progress in Ab Initio QM/MM Free-Energy Simulations of Electrostatic Energies in Proteins: Accelerated QM/MM Studies of pKa, Redox Reactions and Solvation Free Energies†. 2008.
- [104] Bernd Giese, Michael Graber, and Meike Cordes. Electron transfer in peptides and proteins. *Current Opinion in Chemical Biology*, 12(6):755–759, 2008. ISSN 13675931. doi: 10.1016/j.cbpa.2008.08.026.
- [105] Jian Gao, Pavel Müller, Min Wang, Sonja Eckhardt, Miriam Lauz, Katharina M. Fromm, and Bernd Giese. Electron Transfer in Peptides: The Influence of Charged Amino Acids. *Angewandte Chemie International Edition*, 50(8):1926–1930, feb 2011. ISSN 14337851. doi: 10.1002/anie.201003389. URL <http://doi.wiley.com/10.1002/anie.201003389>.
- [106] † Jan C. Schöneboom, ‡ Shimrit Cohen, † Hai Lin, , ‡ Sason Shaik, *, and † Walter Thiel*. Quantum Mechanical/Molecular Mechanical Investigation of the Mechanism of CH Hydroxylation of Camphor by Cytochrome P450cam: Theory Supports a Two-State Rebound Mechanism. 2004.
- [107] Dardo N Ferreiro, Leonardo Boechi, Darío A Estrin, and Marcelo A Martí. The key role of water in the dioxygenase function of Escherichia coli flavohemoglobin. *Journal of inorganic biochemistry*, 119:75–84, feb 2013. ISSN 1873-3344. doi: 10.1016/j.jinorgbio.2012.10.015. URL <http://www.ncbi.nlm.nih.gov/pubmed/23220591>.
- [108] † Igor L. Shamovsky, *, , † Richard J. Riopelle, and Gregory M. Ross‡. Ab Initio Studies on the Mechanism of Tyrosine Coupling. 2001.
- [109] E P Hunter, M F Desrosiers, and M G Simic. The effect of oxygen, antioxidants, and superoxide radical on tyrosine phenoxyl radical dimerization. *Free radical biology medicine*, 6(6):581–5, 1989. ISSN 0891-5849. URL <http://www.ncbi.nlm.nih.gov/pubmed/2546863>.
- [110] Yong Wang, Hui Chen, Masatomo Makino, Yoshitsugu Shiro, Shingo Nagano, Shumpei Asamizu, Hiroyasu Onaka, and Sason Shaik. Theoretical and Experimental Studies of the Conversion of Chromopyrrolic Acid to an Antitumor Derivative by Cytochrome P450 StaP: The Catalytic Role of Water Molecules. 2009.
- [111] David A. Case, Thomas E. Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M. Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J. Woods. The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688, dec 2005. ISSN 0192-8651. doi: 10.1002/jcc.20290. URL <http://doi.wiley.com/10.1002/jcc.20290>.

- [112] Marcelo A. Marti, Alejandro Crespo, Luciana Capece, Leonardo Boechi, Damián E. Bikiel, Damián A. Scherlis, and Dario A. Estrin. Dioxygen affinity in heme proteins investigated by computer simulation. *Journal of inorganic biochemistry*, 100(4):761–70, apr 2006. ISSN 0162-0134. doi: 10.1016/j.jinorgbio.2005.12.009. URL <http://www.ncbi.nlm.nih.gov/pubmed/16442625>.
- [113] L. Capece, M. A. Marti, A. Bidon-Chanal, A. Nadra, F. J. Luque, and D. A. Estrin. High pressure reveals structural determinants for globin hexacoordination: Neuroglobin and myoglobin cases. *Proteins: Structure, Function, and Bioinformatics*, 75(4):885–894, jun 2009. ISSN 08873585. doi: 10.1002/prot.22297. URL <http://doi.wiley.com/10.1002/prot.22297>.
- [114] Leonardo Boechi, Marcelo A. Martí, Mario Milani, Martino Bolognesi, F. Javier Luque, and Darío A. Estrin. Structural determinants of ligand migration in Mycobacterium tuberculosis truncated hemoglobin O. *Proteins: Structure, Function, and Bioinformatics*, 73(2):372–379, apr 2008. ISSN 08873585. doi: 10.1002/prot.22072. URL <http://doi.wiley.com/10.1002/prot.22072>.
- [115] Luciana Capece, Dario A. Estrin, and Marcelo A. Marti. Dynamical Characterization of the Heme NO Oxygen Binding (HNOX) Domain. Insight into Soluble Guanylate Cyclase Allosteric Transition†. 2008.
- [116] Alejandro D. Nadra, Marcelo A. Martí, Alessandra Pesce, Martino Bolognesi, and Darío A. Estrin. Exploring the molecular basis of heme coordination in human neuroglobin. *Proteins: Structure, Function, and Bioinformatics*, 71(2):695–705, may 2008. ISSN 08873585. doi: 10.1002/prot.21814. URL <http://doi.wiley.com/10.1002/prot.21814>.
- [117] Luciana Capece, Leonardo Boechi, Laura L. Perissinotti, Pau Arroyo-Mañez, Damián E. Bikiel, Giulietta Smulevich, Marcelo A. Marti, and Dario A. Estrin. Small ligand-globin interactions: Reviewing lessons derived from computer simulation. *Biochimica et Biophysica Acta - Proteins and Proteomics*, 1834(9), 2013. ISSN 15709639. doi: 10.1016/j.bbapap.2013.02.038.
- [118] Christopher I. Bayly, Piotr Cieplak, Wendy Cornell, and Peter A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *The Journal of Physical Chemistry*, 97(40):10269–10280, 1993.
- [119] Sanjeewa Rupasinghe, Mary A. Schuler, Norio Kagawa, Hang Yuan, Li Lei, Bin Zhao, Steven L. Kelly, Michael R. Waterman, and David C. Lamb. The cytochrome P450 gene family CYP157 does not contain EXXR in the K-helix reducing the absolute conserved P450 residues to a single cysteine. *FEBS Letters*, 580(27): 6338–6342, nov 2006. ISSN 00145793. doi: 10.1016/j.febslet.2006.10.043. URL <http://doi.wiley.com/10.1016/j.febslet.2006.10.043>.
- [120] Danièle Werck-Reichhart, René Feyereisen, DR Nelson, L. Koymans, T. Kamataki, JJ Stegeman, R Feyereisen, DJ Waxman, MR Waterman, O Gotoh, MJ Coon, RW Estabrook, DR Nelson, M Ingelman-Sundberg, AK Daly, M Oscarson, D Nebert, M Frey, P Chomet, E Glawischnig, C Stettner, S Grün, A Winklmair, W Eisenreich, A Bacher, RB Meeley, SB Briggs, D Rozman, M Stromstedt, LC Tsui, SW Scherer, MR Waterman, SM Paquette, S Bak, R Feyereisen, N Tijet, C Helvig, R Feyereisen, C Chapple, O Gotoh, SE Graham, JA Peterson, O Gotoh, PA Williams, J Cosme, V Sridhar, E Johnson, DE McRee, TL Poulos, BC Finzel, IC Gunsalus, GC Wagner, J Kraut, I Schlichting, J Berendzen, K Chu, AM Stock, SA Maves, DE Benson, RM Sweet, D Ringe, GA Petsko, SG Sligar, H Li, TL Poulos, SY Park, H Shimizu, S Adachi, A Nakagawa, I Tanaka, K Nakahara, H Shoun, E Obayashi, H Nakamura, T Iizuka, Y Shiro, JA Hasler, R Estabrook, M Murray, I Pikuleva, M Waterman, J Capdevila, V Holla, C Helvig, JR Falck, G Farrell, R Feyereisen, R Kahn, F Durst, FJ Gonzalez, S Kimura, D Werck-Reichhart, H Hehn, L Didierjean, DJ Waxman, P Honkakoski, M Negishi, D Mansuy, M Klingenberg, T Omura, R Sato, M Newcomb, R Shen, S-Y Choi, PH Toy, PF Hollenberg, Vaz, , MJ Coon, CR Jefcoate, PA Williams, J Cosme, V Sridhar, EF Johnson, DE McRee, I Stoilov, AN Akarsu, M Sarfarazi, DH Geller, RJ Auchus, BB Mendonça, W Miller, CA Stratakis, A Vottero, A Brodie, LS Kirschner, D DeAtkine, Q Lu, W Yue, CS Mitsiadis, AW Flor, GP Chrousos, C Delye, F Laigret, MF Corio-Costet, CCS Chapple, T Vogt, BE Ellis, CR Somerville,

- M Szekeres, K Nemeth, Z Koncs-Kalman, J Mathur, A Kauschmann, T Altmann, GP Redei, F Nagy, J Schell, C Koncs, MM Neff, SM Nguyen, EJ Malancharu, S Fujioka, T Noguchi, H Seto, M Tsubuki, T Honda, S Takatsuto, J Chory, TA Holton, F Brugliera, DR Lester, Y Tanaka, CD Hyland, JGT Menting, C-Y Lu, E Farcy, TW Stevenson, EC Cornish, VM Chavez, G Marques, JP Delbecque, K Kobayashi, M Hollingsworth, J Burr, JE Natzle, MB O, JM Mingot, MA Penalva, and JM Fernandez-Canon. Cytochromes P450: a success story. *Genome Biology*, 1(6):reviews3003.1, 2000. ISSN 14656906. doi: 10.1186/gb-2000-1-6-reviews3003. URL <http://genomebiology.biomedcentral.com/articles/10.1186/gb-2000-1-6-reviews3003>.
- [121] Ying Huang, Beifang Niu, Ying Gao, Limin Fu, and Weizhong Li. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics (Oxford, England)*, 26(5):680–2, mar 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq003. URL <http://www.ncbi.nlm.nih.gov/pubmed/20053844><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2828112>.
- [122] Jimin Pei, Bong-Hyun Kim, and Nick V Grishin. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic acids research*, 36(7):2295–300, apr 2008. ISSN 1362-4962. doi: 10.1093/nar/gkn072. URL <http://www.ncbi.nlm.nih.gov/pubmed/18287115><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2367709>.
- [123] S Ephane Guindon and Olivier Gascuel. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Syst. Biol.*, 52(5):696–704, 2003. ISSN 1063-5157. doi: 10.1080/10635150390235520.
- [124] O Gascuel. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7):685–695, 1997. ISSN 0737-4038.
- [125] Maria Anisimova and Olivier Gascuel. Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative. *Systematic Biology*, 55(4):539–552, aug 2006. ISSN 1063-5157. doi: 10.1080/10635150600755453. URL <http://sysbio.oxfordjournals.org/cgi/doi/10.1080/10635150600755453>.
- [126] Alexandros Stamatakis. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics (Oxford, England)*, 30(9):1312–3, may 2014. ISSN 1367-4811. doi: 10.1093/bioinformatics/btu033. URL <http://www.ncbi.nlm.nih.gov/pubmed/24451623><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3998144>.
- [127] Ivica Letunic and Peer Bork. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics (Oxford, England)*, 23(1):127–8, jan 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btl529. URL <http://www.ncbi.nlm.nih.gov/pubmed/17050570>.
- [128] Daniel H Huson and Celine Scornavacca. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic biology*, 61(6):1061–7, dec 2012. ISSN 1076-836X. doi: 10.1093/sysbio/sys062. URL <http://www.ncbi.nlm.nih.gov/pubmed/22780991>.
- [129] Robert D Finn, Jody Clements, and Sean R Eddy. HMMER web server: interactive sequence similarity searching. *Nucleic acids research*, 39(Web Server issue):W29–37, jul 2011. ISSN 1362-4962. doi: 10.1093/nar/gkr367. URL <http://www.ncbi.nlm.nih.gov/pubmed/21593126><http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3125773>.
- [130] G. E. Crooks, Gary Hon, John-Marc Chandonia, and Steven E. Brenner. WebLogo: A Sequence Logo Generator. *Genome Research*, 14(6):1188–1190, may 2004. ISSN 1088-9051. doi: 10.1101/gr.849004. URL <http://www.genome.org/cgi/doi/10.1101/gr.849004>.
- [131] Emmanuel Paradis, Julien Claude, and Korbinian Strimmer. APE: Analyses of Phylogenetics and Evolution in R language. *BIOINFORMATICS APPLICATIONS NOTE*, 20(2):289–290, 2004. doi: 10.1093/bioinformatics/btg412.

- [132] Dong-Sun Lee, Akari Yamada, Hiroshi Sugimoto, Isamu Matsunaga, Hisashi Ogura, Kosuke Ichihara, Shin-Ichi Adachi, Sam-Yong Park, and Yoshitsugu Shiro. Substrate Recognition and Molecular Mechanism of Fatty Acid Hydroxylation by Cytochrome P450 from *Bacillus subtilis* CRYSTALLOGRAPHIC, SPECTROSCOPIC, AND MUTATIONAL STUDIES*. 2003. doi: 10.1074/jbc.M211575200.
- [133] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, Geoffrey R Hutchison, D Weininger, P Murray-Rust, HS Rzepa, P Murray-Rust, HS Rzepa, P Murray-Rust, HS Rzepa, M Wright, P Murray-Rust, HS Rzepa, GL Holliday, P Murray-Rust, HS Rzepa, K Fogel, A Bender, HY Mussa, RC Glen, S Reiling, D Filimonov, V Poroikov, Y Borodina, T Glorizova, HL Morgan, BD McKay, AA Gakh, MN Burnett, SV Trepalin, AV Yarkov, IV Pletnev, AA Gakh, AA Gakh, MN Burnett, SV Trepalin, AV Yarkov, T Halgren, T Halgren, T Halgren, R Nachbar, T Halgren, A Andronico, A Randall, RW Benz, P Baldi, A Rappe, C Casewit, K Colwell, W Goddard, WM Skiff, J Wang, RM Wolf, JW Caldwell, PA Kollman, DA Case, J Wang, W Wang, PA Kollman, DA Case, NM O, K Martin, B Hoffman, NM O, J Myers, T Allison, S Bittner, B Didier, M Frenklach, William H Green, Y Ho, J Hewson, W Koegler, C Lansing, P Lind, M Alm, A Amini, PJ Shrimpton, SH Muggleton, MJE Sternberg, S Arbor, GR Marshall, Z Huang, CF Wong, AD Hill, PJ Reilly, RS Armen, JH Chen, CL Brooks, L Liu, H Ma, N Yang, Y Tang, J Guo, W Tao, Duan Jaa, I Wallach, N Jaitly, R Lilien, YD Paila, S Tiwari, D Sengupta, A Chattopadhyay, JL Melville, JD Hirst, T Pencheva, D Lagorce, I Pajeva, BO Villoutreix, MA Miteva, L Schietgat, J Ramon, M Bruynooghe, M Krier, MC Hutter, X Wang, J Huan, A Smalter, GH Lushington, T Cheng, Q Li, Y Wang, SH Bryant, VV Mihaleva, HA Verhoeven, RCH de Vos, RD Hall, RCHJ van Ham, DC Bas, DM Rogers, JH Jensen, L Fabian, CP Brock, M Dehmer, N Barbarini, K Varmuza, A Graber, JJ Langham, AN Jain, F Fontaine, M Pastor, I Zamora, M Konyk, A De Leon, M Dumontier, T Kogej, O Engkvist, N Blomberg, S Muresan, C Reynès, H Host, A-C Camproux, G Laconde, F Leroux, A Mazars, B Deprez, R Fahraeus, BO Villoutreix, O Sperandio, D Lagorce, T Pencheva, BO Villoutreix, MA Miteva, MJ Gómez, F Pazos, FJ Guijarro, V de Lorenzo, A Valencia, J Kazius, S Nijssen, J Kok, T Bäck, AP IJzerman, NM O, M Brüstle, M Buehler, J Dodson, A van Duin, CW Bullock, RB Jacob, OM McDougal, G Hampikian, T Andersen, X Jiang, K Kumar, X Hu, A Wallqvist, J Reifman, D Lagorce, O Sperandio, H Galons, MA Miteva, BO Villoutreix, A Maunz, C Helma, S Kramer, A Maunz, C Helma, S Kramer, C Helma, MA Meineke, CF Vardeman, T Lin, CJ Fennell, JD Gezelter, P Tosco, T Balle, P Tosco, T Balle, IV Filippov, MC Nicklaus, DR Koes, CJ Camacho, CR Jacob, SM Beyhan, RE Buló, ASP Gomes, AW Götz, K Kiewisch, J Sikkema, L Visscher, William H Green, Joshua W Allen, Robert W Ashcraft, Gregory J Beran, Caleb A Class, Connie Gao, C Franklin Goldsmith, Michael R Harper, Amrit Jalan, Gregory R Magoon, David M Matheu, Shamel S Merchant, Jeffrey D Mo, Sarah Petway, Sumathy Raman, Sandeep Sharma, Jing Song, Kevin M Van Geem, John Wen, Richard H West, Andrew Wong, Hsi-Wu Wong, Paul E Yelvington, Joanna Yu, A Karwath, L De Raedt, DC Lonie, E Zurek, N Zonta, IJ Grimstead, NJ Avis, A Brancale, JH Chen, E Linstead, SJ Swamidass, D Wang, P Baldi, TWH Backman, Y Cao, T Girke, J Ahmed, CL Worth, PF Thaben, C Matzig, C Blasse, M Dunkel, R Preissner, MA Miteva, F Guyon, P Tuffery, JL Sharman, CP Mpamhanga, M Spedding, P Germain, B Staels, C Dacquet, V Laudet, AJ Harmar, R Esposito, G Ermondi, G Caron, I Wallach, R Lilien, A Poater, B Cosenza, A Correa, S Giudice, F Ragone, V Scarano, L Cavallo, B-b Yan, M-z Xue, B Xiong, K Liu, D-y Hu, J-k Shen, P Rydberg, DE Gloriam, L Olsen, S Ingsriswang, E Pacharawongsakda, RA Bauer, PE Bourne, A Formella, C Frommel, C Gille, A Goede, A Guerler, A Hoppe, EW Knapp, T Poschel, U Schmidt, S Struck, B Gruening, J Hossbach, IS Jaeger, R Parol, U Lindequist, E Teuscher, R Preissner, RA Bauer, S Gunther, D Jansen, C Heeger, PF Thaben, R Preissner, J Ahmed, S Preissner, M Dunkel, CL Worth, A Eckert, R Preissner, M Kuhn, D Szklarczyk, A Franceschini, M Campillos, C von Mering, LJ Jensen, A Beyer, P Bork, IV Tetko, J Gasteiger, R Todeschini, A Mauri, D Livingstone, P Ertl, VA Palyulin, EV Radchenko, NS Zefirov, AS Makarenko, O Sperandio, M Petitjean, P Tuffery, Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open Babel: An open chemical toolbox. *J Cheminf*, 3(1):33, 2011. ISSN 1758-2946. doi: 10.1186/1758-2946-3-33. URL <http://jcheminf.springeropen.com/articles/10.1186/1758-2946-3-33>.
- [134] Oleg Trott and Arthur J. Olson. AutoDock Vina: Improving the speed and accuracy of docking with a new

scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):NA-NA, 2009. ISSN 01928651. doi: 10.1002/jcc.21334. URL <http://doi.wiley.com/10.1002/jcc.21334>.

- [135] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2): 83-97, mar 1955. ISSN 00281441. doi: 10.1002/nav.3800020109. URL <http://doi.wiley.com/10.1002/nav.3800020109>.