Tesis Doctoral

Herramientas bioinformáticas para la predicción y análisis de estructuras de proteínas: de genomas a motivos estructurales

Lanzarotti, Esteban

2016-03-18

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

Herramientas bioinformáticas para la predicción y análisis de estructuras de proteínas: de genomas a motivos estructurales.

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

Esteban Lanzarotti

Director de tesis: Dr. Adrián Turjanski
Consejero de estudios: Dr. Marcelo Marti

Buenos Aires, 2016

Herramientas bioinformáticas para la predicción y análisis de estructuras de proteínas: de genomas a motivos estructurales.

Durante la última decada, varios estudios han mostrado que la anotación automática de genomas resuelve muy bien el problema de recuperar información a partir de una secuenciación. A su vez, esto favoreció el crecimiento desmedido en la cantidad de genomas depositados en bases de datos que poseen una anotación automática, para los cuales se volvería imposible realizar experimentos sobre cada uno de los genes presentes en éstos genomas de manera de mejorar el conocimiento disponible. Es por esto que, obtener una estructura molde para construir un modelo 3D de una determinada secuencia, es una herramienta poderosa para entender las funciones de las proteínas. Una vez modelada la estructura, las interacciones presentes en ella aportan información de la función proteica. En esta tesis desarrollamos un pipeline bioinformático que a partir de la secuencia de un genoma bacteriano puede identificar las regiones codificantes, anotar su función y computa diversas propiedades de las proteínas codificadas. En una segunda etapa desarrollamos un pipeline de predicción de estructura secundaria y terciaria de proteínas que se integró con el sistema anterior. Finalmente, mediante el desarrollo de una base de datos de interacciones de amino ácidos basada en el PDB, estudiamos en detalle las interacciones aromáticas. Los anillos aromáticos forman clusters de interacciones que tienen particularidades que difieren según hayan sido encontrados en el interior de las proteínas o regiones de interacción proteína-ligando o interfaces de interacción proteína-proteína.

Bioinformatic tools for prediction and analysis of protein structures: from genomes to structural motifs.

During the last decade, numerous studies have shown that automatic genome anotation performs quiet well in solving the problem of information retrieval from sequenced genomes. Also, this has favored an excesive growth of deposited (automatically) anotated genomes, which is imposible to design experiments over each of these genes present in each genome in order to improve the available knowledge. For this reason, obtaining a template structure to build a 3D model for a fixed sequence, is a powerful tool to understand protein functions. Once obtained the structure, the interactions involved in it, bring information about protein function. In this thesis we developed a bioinformatic pipeline which from a genomic sequence, identifies coding regions, anotates their functions and computes diverse properties. In a second stage, we developed a pipeline for prediction of protein secondary and tertiary structure, which was integrated to the previous system. Finally, by developing a structural database of amino acid interactions based ob PDB, we studied in detail aromatic interactions. Aromatics clusters have features that difer according they have been found i) in the core of the protein, ii) in small ligand binding regions and, iii) protein-protein interaction interfaces.

Índice general

Capítulo 1

Introducción

Es de común acuerdo que las técnologías de secuenciación de ADN que surgieron en la década del 2000 cambiaron la forma de obtener, acceder y analizar la información biológica, al punto que hoy día cualquier proyecto que tenga objetivos enmarcados en las áreas de la biotecnología suele incluir análisis bioinformáticos a partir de alguna/s secuenciación/es. Cada estudio da lugar a un conjunto de elementos que se desean secuenciar para ser analizardos posteriormente. Estos elementos pueden ir desde la obtención de secuencias de ADN producto de un aislamiento bacteriano hasta la secuenciación de una diversidad de elementos genómicos producto de extraer ADN de una muestra ambiental más compleja (*metagenómica*). Esto hizo que, durante mucho tiempo (tendencia que continúa hoy en día), las bases de datos biológicas se llenaran a velocidades gigantescas con información procedente de proyectos de secuenciación, alrededor de todo el planeta. Más aún, la velocidad a la que se producen secuencias biológicas supera ampliamente la velocidad con la que se pueden llevar a cabo experimentos que den información acerca de los elementos secuenciados, lo que da lugar a la necesidad de desarrollar metodologías y modelos bioinformáticos que permitan plantear hipótesis y responder preguntas sobre dichos elementos de manera de acotar el universo posible de experimentos a realizar. Para lograr estos fines, ya se han desarrollado varios sistemas de manejo y análisis de datos genómicos que se basan principalmente en deducir información biológica a partir de la comparación con (bases de) datos ya existentes (*genómica comparativa*).

Si bien biólogos alrededor del mundo entero han reconocido desde hace tiempo la importancia de estudiar la evolución para entender la organización de los organismos vivos, con el amplio desarrollo de las tecnologías de secuenciación, en los últimos años, la genómica comparativa ha convertido los

estudios evolutivos, en un "kit de herramientas de rutina". En general, colocar los genomas en un marco evolutivo ha demostrado ser útil para comprender el funcionamiento de los organismos, también ha aumentado considerablemente la comprensión de los procesos mediante los cuales los genomas evolucionan y ha llevado a una reevaluación de nuestra representación de la diversidad y la historia de la vida. En particular, hoy día, suele ser de interés el estudio de los *procesos biológicos* de los organismos, poniendo el foco en encontrar un conjuntos de genes que estén implicados en algún proceso de interés, de manera de estructurar un modelo que permita dar lugar al planteo de hipótesis científicas y/o al desarrollo biotecnológico. Por ejemplo, el análisis de los metabolismos subyacentes a la degradación de lípidos, puede ser de utilidad en la industría textil o alimenticia, el metabolismo de nitrogeno puede ser de utilidad en el entorno agropecuario, etc.

Varias plataformas de *anotación genómica* fueron desarrolladas para abordar el problema de interpretar la información producto de secuenciaciones, entre ellas, RAST, ISGA, SABIA, etc, que presentan dan buenos resultados anotando genomas chicos (bacterias, archeas, virus, etc) y serán brevemente comentadas en el Capítulo 1. Estas plataformas de software suelen partir de una secuencia genómica que procesan integrando varios conceptos: predicción de marcos de lectura (Glimmer, ORFinder, GeneMark, EasyGene, etc), búsquedas por similitud (principalmente usando BLAST sobre bases de datos de púbilco acceso como NR, Swissprot, PDB, etc), identificación de dominios funcionales (Usando HMMER sobre bases de datos de *perfiles* de secuencias como Pfam, Interpro, etc) e identificación de ortologías metabólicas (Basandose en bases de datos como COG, KEGG PATHWAYS). En cuanto a los métodos como Glimmer, Pfam o InterPro, estos son predictivos, están basados en usar datos previos para generar/entrenar modelos estadísticos de lo que sea de interés buscar y luego evaluar cada secuencia calculando un puntaje para sobre dichos modelos. En general usan estrategias basadas en modelos de Markov (IMM en el caso de Glimmer, HMM en el caso de Pfam) y se calculan los puntajes con variaciones del algoritmo de Viterbi. Por otro lado, los métodos de búsqueda por simlitud en el algoritmo alineamiento local de Smith-waterman, el cual usa programación dinámica. En parte, estas plataformas son posibles dado que todos estos enfoques tienen soluciones que escalan polinomialmente en función del tamaño del genoma, lo cual hace posible la aplicación a grandes volúmenes de datos como genomas enteros y además son trivialmente paralelizables dado que cada proceso se puede aplicar por separado sobre todos los genes, que a su vez, también pueden paralelizarse.

Lo que estas plataformas de anotación genómica proveen es una forma es-

tructurada (*pipeline*) de procesar los datos producto de secuenciaciones de manera de recuperar las secuencias proteicas subyacentes a los productos génicos que se puedan predecir de la secuencia. Cada secuencia se **anota** automáticamente a través criterios establecidos, sobre los puntajes producto de las comparaciones, dando lugar a una descripción general de cada producto génico. Si bien esta descripción puede ser de gran utilidad para comprender el rol que un determinado gen o conjunto de genes cumplen en un organismo, no da información específica de cada proteína, analizando lo que esa variante tiene de particular. Es por esto que luego de una anotación automática es necesario llevar adelante estudios de otras naturalezas de manera de poder comprender los mecanismos planteados por la variante de interés, y en particular, se ha visto que obteniendo información tridimiensional, acerca de las posibles estructuras involucradas, mejora bastante la calidad de los modelos que se puedan plantear en el transcurso de la investigación.

Gracias a todos estos desarrollos, en la actualidad, más de seis millones de secuencias de proteínas únicas han sido depositadas en bases de datos biológicas de público acceso, y este número sigue creciendo rápidamente. En lo que a las estructuras tridimensionales refiere, gracias a las iniciativas en genómica estructural de alto rendimiento, alrededor de cien mil estructuras de proteínas han sido hasta ahora determinadas experimentalmente, pero aún así, es enorme la diferencia que hay. Esta enorme disparidad entre los número de secuencias y estructuras ha conducido la investigación hacia métodos computacionales para predecir la estructura de las proteínas a partir de su secuencia, incluso a nivel de un genoma entero. Aún así, las técnicas que se han desarrollado para llevar adelante la predicción de estructuras, si bien son poderosas, no dejan de tener sus defectos. A pesar de que se pueden usar de manera totalmente automatizada, obteniendo el máximo provecho de ellas siempre se requiere la experiencia humana para poner el análisis de los resultados en un contexto biológico. Se puede ver claramente que hay una necesidad de acceder a la enorme cantidad de información, producto de predicciones computacionales, de forma sistemática y consistente, de manera tal que los trabajos de investigación sean más eficientes y relevantes. Para esto, se combinan varios métodos de predicción a partir de secuencias, en el caso de predicción de estructuras secundarias las soluciones al problema muestran buenos resultados, y lo mismo es para el caso de los predictores de accesibilidad al solvente y los predictores de desorden intríseco. Con respecto a predecir la estructura tridimencional, la técnica con mas éxito hasta el día de hoy, es conocida como modelado comparativo. Consiste en el modelado a partir de estructuras conocidas que funcionan como molde. Se usan algoritmos de busquedas por similitud en bases de datos de

secuencias biológicas que tengan estructura conocida y, una vez obtenida la estructura molde, se alinean en en el contexto de un conjunto de proteínas similares o incluso una familia de proteínas (en el mejor de los casos). Esta técnica se basa en el hecho de que hay una fuerte relación entre la secuencia de una proteína y su estructura. Más aún, la diversidad plegados que hoy día se encuentran depositados en bases de datos de público acceso es considerablemente menor que la diversidad de secuencias depositadas.

Hoy en día, no hay herramientas de libre acceso que permitan combinar ambas metodologías (genomica comparativa y predicción de estructuras) con el objeto de lograr una mejor anotación, ya que se contaría con la información estructural además de la anotación funcional. Además, usando la información tridimensional es importante destacar, que es posible determinar *sitios* vinculados a la función de la proteína, como ser: sitios catalíticos en el caso de enzimas o, interacciones entre proteína-proteína y/o proteína-ADN. En este sentido, uno de los objetivos del presente trabajo de doctorado, es desarrollar una heurística que permita encontrar los aminoácidos importantes para la función de las proteínas de un proteoma bacteriano, usando elementos de estadística, teoría de la información y data mining ampliamente difundidos para suministrar al experto una visión integral en el oscuro mundo del estudio de genomas bacterianos. En este trabajo de tesis comenzamos anotando genomas de bacterias, y depositando su anotación en bases de datos de público acceso. Si bien la anotación nos permitió obtener proteínas de interés tecnológico, como era la idea en un principio, no quisimos seguir desarrollando este tipo de metodologías de anotación porque nos pareció que el terreno estaba bastante cubierto. Sin embargo, las metodologías mencionadas previamente para anotar productos génicos, carecen de algo fundamental para el mejor entendimiento de la función proteica, que es el mapeo de información funcional sobre la estructura. Se sabe que pequeñas variantes en la secuencia de una proteína pueden producir cambios rotundos en su función, esto da a lugar a pensar que a pesar de que estos sistemas de anotación están en el estado del arte, en la asignación de función (*función biológica*) en términos generales, todavía es necesario poder anotar las particularidades de cada variante en cada genoma. Para esto, intentamos un enfoque estructural que será mencionado a lo largo de este trabajo, mostrando como es posible mapear información funcional, dada una estructura proteica conseguida a partir de modelos 3D contruidos por homología con estructuras conocidas.

Para esto, el paso a seguir es el diseño de un modelo que permita obtener información acerca de las particularidades estructurales con las que se analiza la función proteica, y los principales elementos de estos modelos son las

interacciones intermoleculares. A lo largo de las décadas se han descripto una inmensa cantidad de *interacciones débiles* distintas que pueden establacer la moléculas entre sí. Estas interacciones van desde las clásicas como puentes de hidrógeno, puentes salinos, apilamientos aromáticas, etc. hasta nuevos tipos de interacciones descriptos en los útlimos años como ser las que se forman entre cargas positivas y los anillos aromáticos (cation-π) o incluso, motivos de metioninas y anillos aromáticos que tienen roles importantes en la estabilización de complejos. Lo que todas estas interacciones tienen en común es que se dan entre interactores de tan diversas naturalezas que hace que se vuelva un reto la manera de interpretarlas. Hay numerosos casos donde se ha mostrado cómo, dada la estructura de una proteína, es posible modelar a grano grueso estos motivos interactores que pueden variar desde reducir un amino ácido entero a un sólo punto hasta tomar algunos de los átomos importantes según la naturaleza del amino ácido. Un ejemplo de esto es la interacción de puente hidrógeno que para modelarla, en general, se usan trés átomos: i) un átomo pesado (*aceptor*) con un par de electrones aceptores libres, ii) un átomo pesado (*donor*) que debería estar unido a un átomo de hidrógeno y iii) el átomo de hidrógeno unido al donor. Además, se ha mostrado cómo la distancia entre los átomos y el ángulo entre los átomos pesados centrado en el átomo de hidrógeno son dos buenos *descriptores* de la interacción de puente hidrógeno. Este es un ejemplo de como es posible identificar motivos de interés y definir parámetros en función de conocimiento previo que permita dar información acerca de las interacciones presentes en las proteínas, pero no es el único.

Otro caso interesante es el de la interacción aromática, que también está bien definida en la literatura y consiste en la interacción que hay entre dos anillos aromáticos. La estructura de un anillo aromático tiene la particularidad de que están restringidos los grados de libertad de los átomos que la componen a una geometría plana. Además, se sabe que la disposición planar de un anillo aromático frente a otro es un relevante descriptor de la interacción en el sentido en que permite diferenciar entre dos conformaciones que suelen adoptar dos anillos aromáticos cuando están interactuando en vacío: la forma-T que consiste en un anillo apilado perpendicularmente sobre el otro de manera que el ángulo entre las normales de los planos subyacentes a los anillos aromáticos es de 90°, y el apilamiento π que consiste en un anillo apilado sobre otro donde el ángulo entre las normales es de 0°. Además, se ha visto que en proteínas, las interacciones aromáticas son de gran relevancia para los procesos biológicos, más aún, en este trabajo de tesis hemos visto que lo anillos aromáticos forman clusters en las regiones internas de las proteínas, lo cuál las hace relevantes para el proceso de plegado, y además, que estos clusters de aromáticos también

aparecen en interacciones que establecen las proteínas con otras moléculas como pequeños ligandos o incluso otras proteínas. También, hemos visto que hay carácterísticas estructurales (como la *estructura secundaria* o la *accesibilidad al solvente*) de los residuos aromáticos que dan información acerca del tipo de cluster (interno, proteína-ligando, proteína-proteína) lo cual, en principio, nos permitiría poder asignar un rol funcional a cada cluster aromático que encontremos en la estructura de cada proteína modelada, y de esta manera obtener una anotación estructural particular de la variante que se está analizando.

Capítulo 2

Anotación de Genomas Bacterianos

2.1. Introducción

La secuenciación de genomas dió lugar a la aparición de muchas variantes de cada proteína de las cuales, al día de hoy, no se conoce su función específica. Este hecho trajo como principal consecuencia que se hizo prohibitivo realizar experimentos que evalúen la función de cada una de estas variantes, por lo que diseñar estrategias computacionales capaces de inferir dichas funciones, se convirtió en un tema en auge. Para entender el panorama en el que estamos situados, primero describiremos el proceso a través del cuál se obtiene información genómica a partir de una muestra biológica. Este proceso se divide en tres pasos fundamentales, como se puede ver en la Figura 2.1: Secuenciación, Ensamblado y Anotación. Estos tres pasos fundamentales pueden complicarse tanto como uno quiera y, a su vez, al día de hoy se dispone de varios sabores para aplicar en cada uno de ellos.

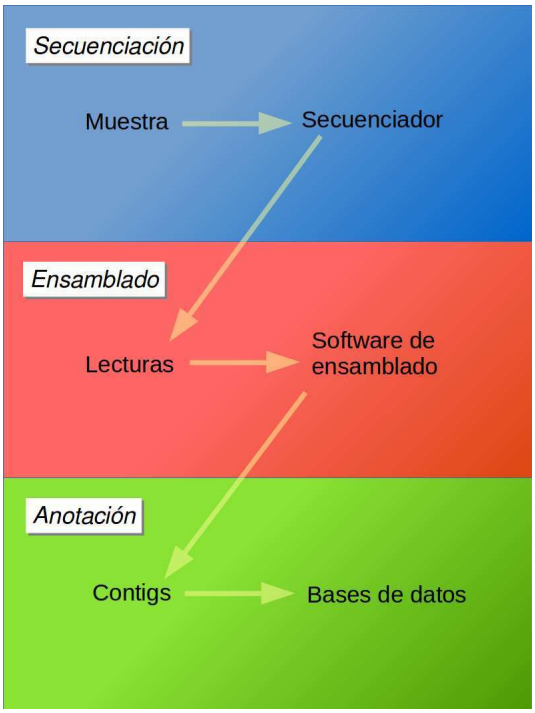

Figura 2.1: **Esquema global del proceso de secuenciación de ADN y posterior anotación.** Primero, durante la *secuenciación* se obtienen **lecturas** a partir de una **muestra**. Segundo, durante el proceso de *ensamblado*, usando esas **lecturas** se construyen los **contigs** de manera de cubrir lo mejor posible el genoma secuenciado. Tercero, durante la etapa de *anotación* usando los **contigs** se comparan las secuencias obtenidas con aquellas que se encuentran en las **bases de datos**.

Secuenciación La secuenciación de ADN cambió la manera de analizar muestras biológicas. Hoy día se dispone de una amplia gama de posibilidades que ofrecen las conocidas *plataformas de secuenciación*, que permiten **leer** secuencias de ADN presentes en una *muestra* biológica de interés. Estas secuenciaciones fueron evolucionando en su tipo y calidad al punto que a partir del 2008 comenzó una desviación con respecto a la *Ley de Moore* en los costos de secuenciación, induciendo a la generación de más datos en tiempos más cortos. A su vez, al irse llenando las bases de datos con información genómica y al ir avanzando el conocimiento funcional acerca de los genes de varios organismos, cada vez fué más necesario el uso de herramientas de genómica comparativa para caracterizar los nuevos genes.

Las características computacionales de las lecturas de ADN obtenidas a

partir de un experimento de secuenciación, dependen intrínsecamente del tipo de equipo y/o metodología que se haya usado para llevar adelante este proceso. La primer metodología que existió la inventó Sanger a principios de la década del 70. Esta técnica permitía leer del orden de las 500 pares de bases (500bp) de ADN en el orden de las pocas horas. Más adelante, con la puesta a punto de la secuenciación basada en fotoquímica (i.e.: *pyrosecuenciación*, etc) se alcanzan capacidades de lectura de ADN que hacen posible los estudios genómicos a gran escala llegando a capacidades de producción de datos del orden de los cientos de Megabases por hora.

En el caso de estudiar mediante una secuenciación, el genoma de una bacteria, se suele trabajar con dos tipos de datos genómicos: 1) genomas *completos* y 2) genomas en estado de borrador (*draft*). La diferencia entre ambos tipos consiste en que, en el primer caso, la secuencia de bases de ADN, cubre completamente el genoma de una bacteria, por lo cual se puede saber la cantidad de bases que separa cada par de genes en el genoma y se termina obteniendo el orden estricto de bases nuecleotídicas que componen dicho genoma. En cambio, en el caso de los genomas draft, sólo se trabaja con regiones no solapantes (denominadas *contigs*) que cubren parcialmente el genoma de una bacteria para las cuales, no necesariamente, se conoce el orden en el que ocuren en el genoma. Además, no se puede descartar el haber perdido genes del organismo con lo cual la información biológica que se obtiene es incompleta. De ambos tipos, el último es el más común, dado que es el que se obtiene con una sola secuenciación de la muestra (o corrida *Shotgun*), mientras que en el caso de los genomas completos, se requiere de sucesivos experimentos de secuenciación (i.e.: *Mate Pair*, *Paired end*, etc) que permitan relacionar los distintos contigs y descubrir huecos que no hayan sido cubiertos previamente. Estos contigs con los que se trabaja luego de la secuenciación, son el producto de un proceso conocido como *ensamblado*.

Ensamblado El proceso de ensamblado consiste en alinear entre sí, las lecturas provenientes de la salida de la secuenciación, de manera de obtener regiones de ADN contiguas (contigs) lo mas abarcativas posibles. El proceso de ensamblado es un proceso que requiere gran poder de cómputo y la solución óptima sería impracticable. Por esto existen varios programas que hacen uso de heurísticas basadas en estructuras de datos que se adecúan a este tipo de algoritmos (i.e.: *grafos de Debrujin*, etc), que se caracterizan por intentar resolver el problema con distintas filosofías, produciendo resultados distintos para un mismo conjunto de lecturas dado. Es por esto que la solución de compromiso que se suele adoptar en la mayoría de los casos consiste en aplicar más de un soft-

ware de ensamblado (i.e.: *Bowtie*, *Newbler*, *Celera*, etc) y, luego, elegir el mejor ensamblado usando algunas métricas de rendimiento.

Las difierencias entre los algoritmos de ensamblado se deben a que están optimizados para determinados tipos de secuencias. Algunos están diseñados para ensamblar muchos fragmentos pequeños y otros para pocos fragmentos grandes. Esto es así porque históricamente las estrategias de ensamblado fueron evolucionando a la par de las tecnologías de secuenciación. Los primeros algoritmos de ensamblado fueron desarrollados en la década del 80 para ordenar el espacio de fragmentos de secuencias producto de las secuenciaciones usando la técnica de Sanger. Pero a medida de la cantidad de lecturas que se podían leer fueron aumentando, las heurísticas de ensamblado se fueron adaptando y, al día de hoy, los mismos fabricantes de secuenciadores recomiendan los paquetes de software que producen los mejores resultados para las lecturas que se obtienen usando sus productos.

Anotación La anotación automática de genomas bacterianos consiste en la asignación de información funcional (*funcionalidad biológica*) a determinadas regiones de la secuencia genómica de un microorganismo. Los sistemas de anotación consisten en la aplicación de varios programas de predicción de patrones y búsquedas en bases de datos, con el objetivo de asignar *tags* correspondientes a lo que se desea saber de dichas regiones. Estos tags serían el producto completo de la anotación y suelen definirse en función de los objetivos de cada proyecto. A su vez, las regiones de interés que nos interesa anotar, en general, están relacionadas a elementos dentro de la secuencia genómica, que van desde sitios chicos (pocas pares de bases) como ser codones start, sitios de unión a ribosomas o protomores, hasta regiones que abarcan una mayor cantidad de pares de bases como secuencias que codifican para proteínas, rRNAs, tRNAs, etc. En el caso de las regiones que codifican distintos tipos de RNA, es posible usar programas de reconocimiento estadístico de patrones (fundamentalmente *modelos de markov*) de manera de clasificar cada región codificante según el tipo de RNA que sea (i.e.: rRNA-16s, tRNA-TYR, etc).

Por su parte, el proceso de anotación de secuencias que codifican proteínas, en general, requiere de detectar los marcos de lectura mediante el reconocimiento de patrones estadísticos producto de entrenar modelos predictivos usando regiones codificantes conocidas. Con estas regiones predichas, se traducen las secuencias de ADN de los marcos de lectura al alfabeto de las proteínas utilizando un código genético adecuado para el organismo que se está anotando. Una vez obtenida la secuencia de proteínas se realizan búsquedas de similaridad en bases de datos biológicas de manera de asignar información funcional

a partir de secuencias de proteínas similares previamente anotadas.

En resumen, los sistemas de anotación consisten en integrar varios programas de predicción, de manera de asignar funcionalidad biológica a cada región de interés dentro del genoma. Continuaremos profundizando en el proceso de anotación que formó parte fundamental de este trabajo de tesis.

2.2. Materiales y Métodos

En esta tesis trabajaremos usando las secuencias de ADN una vez ensambladas, ya sea de un genoma completo o de un genoma draft, el procedimiento de predicción de genes y posterior anotación de proteínas se divide fundamentalmente en 2 pasos como se muestra en la Figura 2.2:

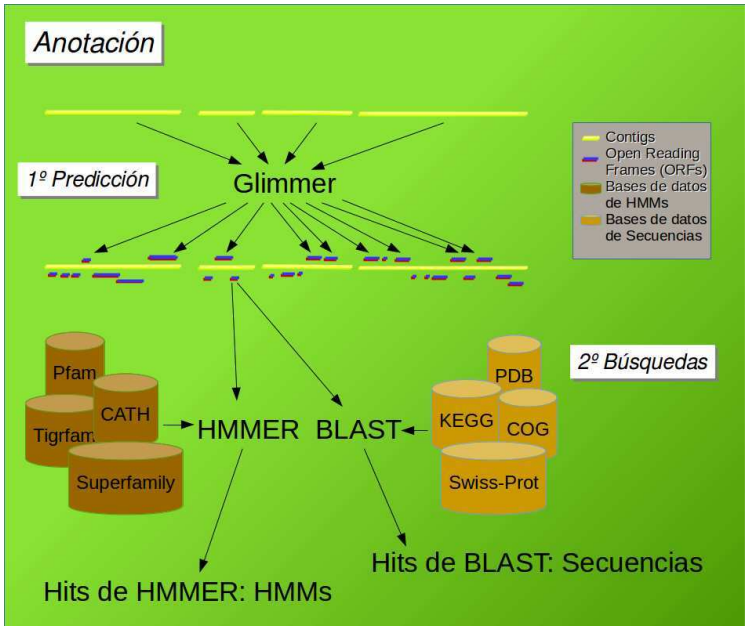

Figura 2.2: **Esquema global del proceso de anotación de secuencias que codifican proteínas.** Primero, se realiza una predicción de genes utilizando algoritmos que no usan comparaciones con secuencias que estan en las bases de datos. Segundo, una vez obtenidos los marcos de lectura, se comparan las secuencias obtenidas con secuencias que estan en las bases de datos.

- **Predicción de regiones codificantes.** Este paso se caracteriza por usar modelos estadísticos previamente construidos, o construidos en el momento previo a la predicción usando la propia secuencia genómica.

No usan búsquedas en bases de datos y tienen por objetivo tratar de predecir los *features* genómicos sín información de bases de datos. Para esto usamos Glimmer que será comentado a continuación.

- **Búsquedas en bases de datos.** Este paso suele ser posterior al anterior, dado que se realiza en función de las regiones que hayan sido predichas. En la mayoría de los casos, se traducen los productos génicos al alfabeto de proteínas y se realizan búsquedas por similaridad sobre bases de datos por entradas similares que hayan sido anotadas previamente. Para esto usamos BLAST y HMMER que también serán comentados, además de las bases de datos que usamos para obtener las anotaciones.

En general, este acercamiento dual que predice genes sin basarse en comparación con bases de datos permite determinar posibles genes que nunca hayan sido descubiertos antes, y por otro lado, permite asignar información funcional a un marco de lectura comparando con bases de datos los genes encontrados. La concatenación de estos dos procesos da lugar a una secuencia de procesamiento de datos a la que se le suele dar el nombre de *pipeline*. Estos pipelines se caracterizan por estar formados por una secuencia de procesos que funcionan como *piezas* intercambiables en los que el *output* de uno se convierte en el *input* del próximo proceso. Por lo tanto, describiremos las piezas que forman parte de estos procesos y luego, mostraremos como se concatenan para formar el *pipeline de anotación de genomas bacterianos*.

2.2.1. Predicción de regiones codificantes.

Dados los contigs, producto del ensamblado, se procede a identificar regiones del genoma objetivo con altas chances de ser genes. Para esta tarea se usó principalmente Glimmer3 [Delcher et al., 2007], una herramienta que debe ser entrenada con un genoma de referencia filogenéticamente cercano al genoma objetivo. También están disponibles otras herramientas como ser GeneMark [Besemer and Borodovsky, 2005] y ORF Finder [Rombel et al., 2002]. Entre estas tres herramientas Glimmer suele subestimar la cantidad de ORFs mientras que GeneMark suele sobreestimarlos. Por otro lado, ORF Finder sólo produce una salida de todo los marcos de lecturas posibles dando como resultado una gran cantidad de ORFs sobreestimados y se utiliza principalmente para encontrar todas las regiones codificantes y más tarde realizar búsquedas de similaridad para descartarlos manualmente. De entre estos, nosotros elegimos Glimmer para identificar las regiones que codifican proteínas y para identificar regiones que codifican tRNAs usamos una herramienta llamada tRNAScan-SE

[Lowe and Eddy, 1997]. Por último, una vez de predichos los elementos codificantes, en el caso de las regiones que codifican proteínas, es posible realizar predicciones de elementos proteicos. En nuestro caso, mencionamos los predicciones de localización celular, PSORT [Nancy et al., 2010] y SignalP [Petersen et al., 2011], que usamos para cumplir con objetivos específicos de los proyectos en los que nos vimos envueltos y los comentaremos brevemente.

tRNAScan-SE tRNAScan es una herramienta que está basada en modelos estadísticos de las secuencias de tRNAs conodidos. Implementa el algoritmo de Pavesi [Pavesi et al., 1994] que está basado en el reconocimiento de dos regiones de control intragénicas conocidas como cajas A y B, que son señales de terminación de la transcripción, y también tiene en cuenta la distancia entre estos elementos y la distancia a la RNA polimerasa III. En función de estos parámetros calcula un puntaje y si el puntaje está por encima de un determinado punto de corte, la secuencia analizada es considerada un tRNA.

Gene Locator and Interpolated Markov Modeler (GLIMMER) Es un sistema compuesto por una variedad de programas que brinda facilidades para predecir ORFs a partir de un conjunto de contigs. El proceso de predicción de ORFs consta de una etapa de entrenamiento y una etapa de predicción. Dado que Glimmer funciona identificando genes usando modelos de markov, necesita ser entrenado usando ORFs de referencia. Los ORFs que se deben utilizar para éste entrenamiento deben ser de un organismo filogenéticamente cercano al que se desea estudiar para que las predicciones den un mejor resultado. Otra forma de proveer ORFs es extrayendo las regiones mas largas donde no aparezcan codones stop con el comando `long-orfs` de Glimmer. La idea es la siguiente, si las secuencias de ADN fueran aleatorias y los codones estuvieran distribuídos aleatoriamente, los codones stop aparecerían con una frecuencia del orden de 3/64. Teniendo esto en cuenta, Glimmer permite buscar largas porciones dentro de los contigs donde no aparece un codon stop, tomando esto como un posible marco de lectura. De esta manera es posible conseguir un conjunto de ORFs de entrenamiento para suministrar a Glimmer y hacer la predicción, que usa información estadística de la composición de bases de las regiones codificantes y no codificantes. Luego, Glimmer 3.0 incorporó los sitios de unión a ribosoma para mejorar la predicción de los codones start.

Identificación sitios de unión a ribosoma. Para identificar sitios de unión a ribosoma (Ribosome Binding Site - RBS) está disponible una herramienta llamada RBSFinder. Encontrar sitios de unión a ribosoma es útil para el anotador

a la hora de corregir los codones start que el predictor de ORFs haya encontrado. Esto es importante debido a que, en el caso de los codones start que son codificantes (a diferencia de los codones stop), se puede dar que aparezca más de un codon start en el extremo $5'$ del gen, lo cuál suele dar lugar a que los predictores comentan errores en la posición del inicio del ORF. RBSFinder se alimenta con una secuencia contigua de ADN (i.e.: un contig) y una lista de coordenadas de ORFs producto de correr una predicción, usando Glimmer o GeneMark. Brinda la posibilidad de suministrar por la entrada un sitio consenso de unión a ribosoma propio del genoma que se esta anotando, y produce una salida con todos los RBSs posibles para los marcos de lectura suministrados. La manera de obtener este sitio consenso es tomando las últimas 50 bases de la subunidad 16s del ribosoma del organismo y con cada ventana de 5 bases de largo realizar una búsqueda de abundancia de cada una de ellas en la zonas upstream de cada ORF identificado.

Predicción de localización celular Existen modelos estadísticos que representan motivos lineales comunes como en las secuencias de proteínas como ser los pétidos señal que siguen la vía de secreción o las hélices alfa de paso transmembrana. Estos motivos se pueden modelar estadísticamente de la misma manera que las familias y dominios proteicos, sólo que en general, al ser motivos simples desde el punto de vista de su complejidad estadística, alcanza con pocos modelos para representarlos. En este trabajo usamos dos programas para realizar éstas predicciones sobre las secuencias traducidas a partir de los ORFs: i) **Signalp - Predicción de péptido señal**. Algunas proteínas presentan lo que se conoce como péptido señal, que controla la entrada a la *vía secretoria* de aquellas proteínas que lo presenten en el extremo N-terminal, tanto en procariotas como en eucariotas. ii) **PSORT - Consenso de predictores.** Consiste en un consenso basado en un conjunto de predictores que permite asignar una posible localización celular dada la secuencia de una proteína.

2.2.2. Búsquedas en bases de datos.

El paso de anotación es el encargado de asignar automáticamente información funcional biológica a cada ORF según su producto génico subyacente. En la mayoría de los casos se realizan búsquedas por similaridad en diversas bases de datos biológicas. Es por esto que cuando hablamos de anotación funcional, hablamos de búsquedas en bases de datos. Lo más importante para realizar búsquedas en bases de datos son los algoritmos de comparación. Estos algoritmos que permiten comparar secuencias biológicas, también son usados para

indexar bases de datos, dando lugar a herramientas denominadas **sistemas de búsqueda**.

Tanto los sistemas de búsqueda por comparación, como los tipos de los datos que se usan para comparar, pueden venir en varios sabores. En este trabajo mencionaremos los dos sistemas más difundidos en la comunidad que fueron los que usamos: *BLAST* y *HMMER*. Hicimos especial énfasis en estos dos programas dado que son de muy fácil manejo e instalación. (si bien las versiones en los repositorios de Debian pueden estar desactualizadas frente a la de los proveedores oficiales, se pueden instalar usando `apt-get`!). BLAST sirve para buscar en bases de datos de secuencias y HMMER sirve para realizar búsquedas en bases de datos de modelos ocultos de markov (HMMs). Están inspirados en dos algoritmos conocidos que son el algoritmo de alineamiento de secuencias de Smith-Waterman, en el caso de BLAST, y HMMER se basa en el algoritmo de Viterbi para encontrar la secuencia de estados de un HMM que maximiza la verosimilitud del modelo estadístico.

Aca es donde es importante recalcar que, si bien, un genoma puede estar anotado y disponible en una base de datos, es muy común que cada proyecto tenga interés en analizar distintos factores para los cuales la anotación previa no sea suficiente. Es por esto que cada proyecto genómico suele elegir distintos sistemas de búsqueda y distintas bases de datos dependiendo de las preguntas que se deseen hacer acerca del genoma secuenciado.

Basic Local Alignment Search Tool (BLAST)

BLAST [Altschul et al., 1990], [Altschul et al., 1997] es un sistema de programas que permite realizar búsquedas por similitud de secuencias biológicas. Es una de las herramientas mas útiles y gracias a éste programa es posible el análisis genómico comparativo. El mecanismo es el siguiente: dada una secuencia biológica a la que llamaremos query (i.e.: estructura primaria de una proteína) y una base de datos de secuencias que debe estar en formato fasta y previamente indexada usando el comando `formatdb`, BLAST devuelve un reporte de las similaridades de la secuencia query encontradas en esa base de secuencias. En ese reporte es posible ver el puntaje de alineamiento local (métrica que informa el parecido entre dos secuencias alineadas) de la proteína query con cada una de las secuencia con las que produjo un buen alineamiento. Entre otras cosas, del reporte de BLAST, es posible extraer, para cada secuencia que alinea con la query, su porcentaje de identidad y su cobertura, esto es de especial importancia a la hora de la anotación de genes, dado que si bien, puede haber un hit para un determinado ORF, éste puede ser descartado debido a que solamente cubre una porción de la secuencia que se desea ano-

tar. Además de los parámetros del alineamiento, BLAST provee un estadístico que, esencialmente, describe la probabilidad de que dos secuencias de ese largo produzcan al azar un alineamiento con ese puntaje. Este estadístico se calcula para cada base de datos en particular y recibió el nombre de $e-value$. Se calcula en función del puntaje del alineamiento y del tamaño de la base de datos en la que se está buscando. Más puntualmente, decrece exponencialmente con el puntaje obtenido.

Ahora bien, comparar amino ácidos no es trivial, debido a que hay amino ácidos que se parecen entre sí más que con otros. Por ejemplo, una fenilalanina es más parecida a una tirosina que a una cisteína, debido a su naturaleza aromática. Es por esto que, históricamente, se diseñaron dos tipos de matrices para realizar éstas comparaciones: PAM y BLOSUM. Las matrices tipo PAM se construyen a partir de los reemplazos de amino ácidos de proteínas filogenéticamente muy relacionadas. Por el contrario, las BLOSUM estan basadas en alineamientos locales, de regiones alineadas (i.e.: motivos dentro de las proteínas) de proteínas que no se parecen tanto. Ambas matrices fueron diseñadas para medir la distancia que hay entre los aminos ácidos de manera de poder puntuar los alineamientos. Por defecto, BLAST usa la matriz BLOSUM, más puntualmente: BLOSUM62.

Ademńas de BLAST, hay otra version del algoritmo denominada Position Specific Iterated BLAST (PSI-BLAST), que aplica sucesivas búsquedas para construir otra matriz de puntuación en función de los *hits* que se van encontrando en la base de datos. La primer iteración usa alguna de las matrices mencionadas previamente y de la segunda iteración en adelante, se construye un alineamiento múltiple con todas las secuencias que surgieron de la búsqueda y se calcula, por cada columna del alineamiento, la frecuencia con la que aparece cada amino ácido. De esta manera, la similitud no se evalúa en función de las dos secuencias alineadas, sino que se puntúa en función del perfil estadístico obtenido a través de las iteraciones. Este programa resuelve búsquedas en tiempos del mismo orden BLAST por cada una de sus iteraciones y es más sensible a similaridades de secuencia más bajas aunque igual de relevantes, se vió que permite encontrar relaciones de similitud más lejanas y obtener alineamientos más largos que logran una mayor cobertura sobre la secuencia query.

A los fines del desarrollo del pipeline de anotación, disponemos de listas de *hits* que presentan una determinada similaridad en secuencia para cada ORF. Usando el e-value pusimos un punto de corte para definir una noción computacional de homología que ha sido usada ampliamente por la comunidad bioinformática. En nuestro caso, usamos un valor de corte de 1e-05, por lo que

todos los hits que presenten un valor de corte superior a éste, no serán tenidos en cuenta. Luego, para cada uno de los hits, se evalúa la identidad de secuencia, calculada como la cantidad de residuos alineados idénticos dividido el largo de la región alineada y la cobertura de la región alineada sobre el largo total de la secuencia query. Usando estos dos parámetros se descartan aquellas secuencias que presenten una cobertura menor al 85 %.

HMMER

Este software (HMMER [Johnson et al., 2010]) también consiste en un sistema de programas que permite realizar búsquedas en bases de datos sólo que, en este caso, las bases de datos son de HMMs. Los HMMs son una herramienta matemática que tiene una gran variedad de aplicaciones. En el caso de la bioinformática de proteínas, uno de los principales usos que se le da es el de modelar, por posición, la frecuencia relativa con la que los amino ácidos ocurren en un conjunto de secuencias. HMMER permite realizar 2 tipos de búsquedas: 1) buscar en una base de datos de HMMs usando una secuencia query, y 2) buscar en una base de datos de secuencias usando un HMM query. El primer tipo de búsqueda es el que se usa para anotar ORFs en función de los hits obtenidos a partir de buscar en una o más bases de datos de HMMs.

El objetivo de los HMMs aplicados a secuencias biológicas es poder modelar un conjunto de proteínas usando una sola entidad matemática. La idea, es que un HMM permite representar lo que se conoce como un perfil estadístico, que se construye a partir de un alineamiento múltiple de dichas secuencias. De esta manera, es posible construir HMMs, que representan estadísticamente la información acerca de la cantidad de apariciones de cada amino ácido por cada columna del alineamiento, incluyendo, a su vez, información acerca de las probabilidades de inserción y deleción, por posición, que se calculan a partir de los gaps en las secuencias proteicas del alineamiento proporcionado. Una vez obtenido el HMM que representa una familia de proteínas, es posible calcular el puntaje (o la probabilidad) de que una determinada secuencia proteica sea "emitida" por el mismo, usando algoritmos como el de Viterbi [Viterbi, 1967]. Sin embargo, estos algoritmos que calculan la puntuación de una secuencia para un determinado HMM, son muy costosos como para realizar búsquedas en bases de datos grandes, por lo que HMMER implementa una serie de filtros para aplicar, previos a la puntuación, de manera de obtener un tiempo de búsqueda razonable.

HMMER hace dos suposiciones fuertes que constituyen una de las principales causas de problemas para detectar homología. Primero, ignora las correlaciones de a pares que puede haber entre las columnas del alineamiento

mútliple. Esto quiere decir que sólo modela la estructura primaria de las proteínas en el alineamiento, sin tener en cuenta que las posiciones de los amino ácidos en una familia de proteínas está fuertemente correlacionada debido a los contactos establecidos en la estructura terciaria. Segundo, los HMMs también suponen que las secuencias involucradas son entidades generadas independientemente desde un punto de vista probabilístico. Esto no es cierto debido a que las proteínas reales están relacionadas mediante su historia filogenética, mediante ancestros comunes, lo cual las hace altamente no-independientes. Estas dos suposiciones son uno de los principales problemas en la mayoría de los métodos para generar perfiles.

Los hits resultantes de HMMER también tienen un e-value asociado. Pero además, cada HMM tiene un valor de corte para el puntaje que le asigna el algoritmo de puntuación. Este valor se lo conoce como *Trusted Cutoff* y es un valor particular para cada modelo en la base de datos. Este valor indica si la proteína tiene alguna chance de pertenecer a la familia modelada por el HMM, y luego, es posible usar el e-value para evaluar que tan relevante es estadísticamente, esta relación de pertenencia. En nuestro caso, nos quedamos sólo con aquellas proteínas que cumplan que el puntaje asignado sea mayor al trusted cutoff del modelo que obtenemos como hit, y además, ponemos un punto de corte al e-value de 1e-05. Finalmente, en el caso de los dominios funcionales, se puede dar el caso que una secuencia presente una región dónde haya más de un hit representativo. En estos casos elegimos el dominio de mayor puntaje y todos aquellos dominios solapantes a éste y de menor puntaje son descartados.

Bidirectional Best Hit (BBH)

Como se mencionó previamente, el proceso de anotación consiste fundamentalmente en realizar búsquedas por similaridad en bases de datos, pero se ha visto que esta metodología mejora bastante su eficiencia cuando se realizan una comprobación recíproca contra el genoma que se está anotando. Dada una relación por similitud entre dos secuencias (BLAST por ejemplo) es posible definir los Bidirectional Best Hits (BBHs) entre ambos genomas. El proceso consiste en: dado un genoma, se toma un gen llamemosle genA y se realiza una búsqueda por similaridad contra los genes de otro genoma. Una vez obtenido el mejor resultado, se realiza otra búsqueda por similaridad contra el primer genoma y si el mejor resultado es el genA se dice que hay una relación de BBH entre ambos genes. El procedimiento se puede generalizar, se puede realizar la primer búsqueda contra una base de datos con muchas secuencias que no necesariamente sean del mismo organismo, (i.e.: KEGG, COG, Swis-

sProt, etc) dado que el mejor resultado en dicha base de datos, en particular, también será el mejor resultado en el genoma del cual se obtuvo. Luego, se continúa desde la segunda búsqueda sobre el genoma que se desea anotar como fue mencionado previamente. Esta técnica aumenta la rigurosidad en la anotación automática dado que asegura la relación de ortología. Esto es importante para definir la función de ciertas enzimas que a veces se transfiere erroneamente por similaridad con proteínas parálogas que tienen una función similar en tanto su actividad, pero pueden variar en su especificidad de sustrato.

2.2.3. Bases de datos

Las publicaciones científicas tradicionales en su conjunto, durante mucho tiempo, funcionaron como un gran repositorio de información donde los investigadores podían consultar información y conocimiento derivado de datos experimentales. Pero con la influencia del campo computacional y las tecnologías de la información a principios de la década del 70, el desarrollo de repositorios basados en sistemas computacionales comenzó a crecer, alcanzando su auge a principios de los 80s con la creación de los grandes consorcios que ocuparon el rol de mantener y distribuir los datos. Para la década del 90 las bases de datos continúan creciendo acaudaladamente y empiezan a llamar la atención de la comunidad científica en general, con las primeras secuenciaciones de genomas completos, tanto de procariotes como de eucariotes. Para la década del 2000, las bases de datos se vuelven bastante difíciles de manejar, no sólo por su tamaño, sino además, porque los distintos experimentos biológicos fueron dando lugar a distintos *tipos de datos*, al punto que en los últimos años fueron indispensables las iniciativas de *integración de datos* que permitan cruzar información entre varias bases de datos de forma unificada.

Hoy día, las bases de datos constituyen uno de los elementos básicos de cualquier investigación biológica dado que resultaron ser la forma más eficiente de almacenar información derivada de referencias bibliográficas, siendo evidencia de esto el hecho de que empezaron a constituirse importantes revistas científicas que publican artículos referidos exclusivamente a bases de datos. Se han llevado adelante numerosos esfuerzos por manejar toda esta información para que esté accesible y, dependiendo de los datos disponibles, los distintos tipos de repositorios ofrecen distintas funcionalidades para poder acceder a los datos. Pero, a pesar de todo esto, la integración de todos estos datos siempre se ha mostrado problemática y constituye uno de los principales elementos de análisis a la hora de poner a punto los datos en un proyecto de manera de poder trabajar con ellos. Más aún, debido a las diferencias, tanto en términos

técnicos como políticos, resulta inverosímil pensar que se puede resolver una consulta (*query*) biológica indexando una sola base de datos.

A pesar de la gran cantidad de bases de datos disponibles para realizar análisis biológicos, existen una serie de bases de datos que, con el correr de los años, mostraron ser de gran utilidad para la anotación de genomas de bacterias, y serán comentadas a continuación.

Kyoto Encyclopedia of Genes and Genomes (KEGG) [Kanehisa and Goto, 2000], [Kanehisa et al., 2006], [Kanehisa et al., 2007] KEGG es un proyecto que tiene por objetivo juntar todo el conocimiento referente a las reacciones químicas en organismos vivos. KEGG está dividido en varias entidades que, a su vez, están relacionadas. Las entidades más relevantes a nuestros propósitos serían: GENES, ORTHOLOGOUS y PATHWAYS. GENES junta todo el conocimiento referente a organismos. Sus genomas, los genes de cada genoma y la jerarquía taxonómica. En ORTHOLOGOUS se definen grupos de proteínas y/o genes ortólogos y se construyen clusters de estos mismos a los cuales se les asignan códigos que identifican enzimas. Estos códigos denominados *EC numbers*, consisten en una nomenclatura de cuatro dígitos que pretende especificar la función de una enzima usando una estructura de datos jerárquica de tipo arboreo. Por ejemplo, 2.1.1.4, denomina una enzima *Acetylserotonin methyltransferase*, mientras que 2.1.1.- denomina sólo *methyltransferase*. De esta manera, se puede especificar la función de una proteína de manera más general si es que no se conoce del todo su función más específica, como ser la especificidad de sustrato. Finalmente, PATHWAYS junta información de todos los caminos metabólicos. Está diseñada como un multigrafo, debido a que cada nodo puede estar conectado por más de un eje. Cada subgrafo del grafo principal hace referencia a un tipo de metabolismo (metabolismo de carbohidratos, metabolismo de aminoácidos, etc), con los nodos se representan los compuestos químicos que funcionan como sustratos y productos de las reacciones en las células (glucosa, prolina, etc) y los ejes vendrían a ser las enzimas que catalizan dichas reacciones (dehidrogenasa, aminotransferasa, etc). La anotación genómica usando KEGG consiste en asignar a cada gen del genoma estudiado un cluster de ortólogos, basandose en búsquedas por similaridad en los genomas presentes en la base de GENES. Con las ortologías asignadas se construyen los mapas metabólicos que definen los caminos de los compuestos químicos en el interior de los organismos y se reconstruyen los mapas del genoma que se desea estudiar.

UniprotKB: Swiss-Prot / TrEMBL [Consortium et al., 2014] UniprotKB es la unión de dos grandes bases de datos europeas: Swiss-Prot y TrEMBL. Swiss-Prot es una base de datos revisada y curada manualmente por especialistas que tiene como objetivo generar un repositorio con vocabulario controlado de todas las proteínas de todos los organismos. El método por el cual se deposita información en Swiss-Prot es el siguiente: Se usa como referencia la base de datos TrEMBL en donde hay una gran cantidad de secuencias de proteínas traducidas de genes secuenciados y depositados en EMBL. Varias de estas secuencias fueron depositadas por investigadores pero principalmente se producen a través de traducir automáticamente secuencias de nucleotidos previemente depositadas en otras bases de datos (i.e.:nt.) Cada una de estas secuencias es asignada a un investigador en algún lugar del mundo quien es el encargado de revisar la proteína comparando con varias bases de datos y sobre todo basandose en literatura. Una vez terminado el proceso de revisión se procede a quitar la secuencia de TrEMBL y ponerla en Swiss-Prot. Entonces, Swiss-Prot nos permite tener información de alta calidad sobre la función de una proteína. Un ejemplo de esto sería la proteína `A5U493`, de la cual extraeríamos: Nombre de la proteína: **Beta-lactamase**, Nombre del gen: **blaC**, Organismo: *Mycobacterium tuberculosis (strain ATCC 25177 / H37Ra)*.

COG: Clusters of orthologous groups [Tatusov et al., 2000] [Galperin et al., 2014] La base de datos COG es un intento por generar una clasificación filogenética de las proteínas que se encuentran en genomas completos, lo cual permite diferenciar entre proteínas ortólogas (entre genomas) y parálogas (dentro del mismo genoma), de manera de que cada grupo represente adecuadamente una ortología. COG consiste en grupos de genes aglomerados por una relación computacional de ortología conocida como Bidirectional best Hits (BBHs), que será explicada más adelante. O sea que, cada cluster de ortólogos, incluye proteínas para las cuales se infiere por similaridad que son ortólogas y dicha inferencia se cura manualmente. Esto permite reducir bastante la cantidad de errores de anotación, de la misma manera que lo hace el hecho de usar Swiss-Prot. Aunque, además, la base de datos COG está organizada de forma jerárquica de manera de agrupar cada ortología en una clase funcional, y así, facilitar una visualización "a vuelo de pájaro" de la generalidad de funcionalidades que encontramos en el genoma que estamos anotando. Un ejemplo puede sería la ortología **COG2367** de nombre **Beta-lactamase class A** de la cual extraemos la clase a la que pertenece: **V: Defense mechanisms**.

Protein Data Bank (PDB) [Berman et al., 2000] La base de datos PDB fue fundada en 1971, y fue la primer base de datos basada en computadoras, que ofreció datos en el campo de la biología molecular. Constituye uno de los repositorios más interesantes de entre todos los datos biológicos y tiene mirrors en Estados Unidos, Europa y Japón. Contiene información tridimensional de las estructuras que adoptan las proteínas. La forma mas común de indexar el PDB es usando el sistema BLAST para buscar sobre el conjunto de secuencias proteicas subyacente. Pero a lo largo de los años se han buscado otras estrategias más potentes para indexarla, dado que cuando se trata de entender la función de una proteína, su estructura, es de gran utilidad. Entraremos en más detalle sobre esta base de datos en el próximo capítulo.

Pfam [Bateman et al., 2004], [Finn et al., 2006] La base de datos Pfam es de familias y dominios de proteínas. Dado un conjunto de proteínas que tienen al menos un dominio en común, es posible agruparlas mediante un proceso de *alineamiento de secuencias*. Este proceso da lugar a un *alineamiento múltiple* de los dominio proteicos que se parecen en el cuál, a cada dominio, se le agregan *gaps* de manera de hacer coincidir las regiones que tienen más similitud. Estos *agrupamientos* de proteínas, que se representan mediante alineamientos múltiples, se usan para construir HMMs relacionados a cada familia en *Pfam* y, usando el software *HMMER*, es posible realizar búsquedas en *Pfam* teniendo como *query* la secuencia de una proteína. Pfam se caracteriza por tener una amplia cantidad y calidad de información funcional biológica. En Pfam, cada HMM puede representar una familia de proteínas o un dominio funcional que se ve replicado a lo largo de varias familias distintas. Lo interesante de Pfam es que permite *segmentar* una secuencia proteica en sus unidades funcionales: sus dominios. Esto es muy útil cuando se trata de describir la función de una proteína dado que es muy común cuando se trabaja con el genoma de una bacteria que se encuentren proteínas con una caracterización funcional muy general, o sea que si bien su función exacta es desconocida, presentan un arreglo de dominios que, cada uno de ellos, sí está bien caracterizado. Estas proteínas suelen ser anotadas con nombres como, por ejemplo "Hydrolase domain containing protein" y nada más. Es por esto que esta base de datos es de mucha utilidad a la hora de caracterizar la función de una proteína. Un posible ejemplo es la familia **PF00144** de nombre **Beta-lactamase**, de la cual se extrae el cubrimiento sobre el ORF que está siendo anotado.

TigrFam [Haft, 2003] Esta base de datos consiste de un conjunto bastante grande de HMMs en los que se pueden realizar búsquedas usando HMMER

como en el caso de Pfam. Aunque, a diferencia de Pfam, TigrFam tiene la particularidad de que los modelos de markov fueron creados usando conjuntos de *ortologías* que fueron seleccionadas a partir de muchos genomas de bacterias secuenciados. Las proteínas se agrupan en lo que TigrFam denominó *equivalogs*. Estos agrupamientos de proteínas tienen la particularidad de ser de tamaño completo o tan largo como sea posible. Esto da lugar a que, a diferencia de Pfam, los HMMs de TigrFam, describen más comunmente familias de proteínas que dominios funcionales aislados. Obviamente, hay casos donde esto no se puede lograr, lo cual da lugar a un HMM de un dominio funcional, pero usando el criterio de tamaño completo, no suele ser lo común. Esta base de datos es de las más útiles para determinar a la familia de proteínas a la que pertenece una secuencia proteica.

Superfamily (SCOP) [Wilson et al., 2009] Esta base de datos consiste en un conjunto de modelos ocultos de markov que modelan una clasificación de proteínas llamada SCOP y permite obtener una clasificación del plegado de la proteína. SCOP es una organización estructural y evolutiva, curada manualmente, de los plegados que se encuentran en la base de datos PDB. En esta base de datos las proteínas multi dominio se segmentan en dominios funcionales que las constituyen, que luego, son considerados como unidades separadas a la hora de construir la clasificación. La idea detrás de SCOP es generar una segmentación del espacio de estructuras según una jerarquía que desde la raiz, divide las estructuras a partir de la estructura secundaria, y cada una de ellas están agrupadas en grupos denominados *Superfamilias*. Una superfamilia se define como un conjunto de dominios para las cuales existe evidencia estructural y funcional de que decienden de un ancestro común. En esta base de datos, el nivel que sigue a superfamilia es *familia*, que agrupa dominios que tienen una clara similitud en secuencia. El nivel encima de superfamilia es *plegado*, que agrupa dominios que tienen mayormente la misma estructura secundaria y la cadena proteica presenta la misma topología global. Para construir cada HMM se usan como semilla las secuencias del nivel de Superfamilia filtradas a 95 % de identidad de secuencia. De esta manera, cada Superfamilia tiene uno o más modelos dependiendo de cuantas secuencias quedan en ella luego de aplicarse el filtro por identidad de secuencia. A modo de ejemplo, la proteína **Beta-lactamase, class A** esta organizada en la Familia: *beta-Lactamase/D-ala carboxypeptidase* que esta en la Superfamilia: *beta-lactamase/transpeptidase-like*, que le corresponde el Plegado: *beta-lactamase/transpeptidase-like*, y finalmente este plegado está clasificado en la Clase: *Multi-domain proteins (alpha and beta)*.

CATH [Sillitoe et al., 2015] Igual que Superfamily, esta base de datos contiene información estructural de plegados. Pero a diferencia de Superamily, la propuesta de CATH consiste en que a medida que más y más plegados se fueron resolviendo, se fue viendo que una división jerarquica de los plegados parece no alcanzar para clasificar las estructuras que existen, sino que se necesita definir un "espacio continuo" de plegados. Por esto, CATH clasifica las proteínas según sus plegados, usando la clasificación jerarquica claśica que consiste en 1. Clases, 2. Arquitecturas, 3. Topologías y 4. Superfamilias de homólogos, y además implementa, listas de punteros laterales entre las distintas entradas de la base de datos para representar el hecho de que el espacio de plegados no es fácil de segmentar. El criterio que se usa para crear la jerarquía es análogo al de SCOP, pero la forma de generar los modelos de markov varía levemente dando lugar a resultados ligeramente distintos a la hora de clasificar la estructura de una proteína. En CATH, los homólogos se buscan usando BLAST todos-contra-todos de manera de obtener conjuntos que se parezcan como mínimo en un 40 % de identidad y un 60 % de cobertura. Luego, con estos grupos se construyen los HMMs usando el software HMMER.

Así damos por concluída la explicación de los algoritmos de comparación y las bases de datos para poder explicar cómo hicimos para implementar un pipeline propio. Porque si bien nuestra intensión no fue la de reinventar la pólvora en el terreno de la anotación genómica, nos vimos en la necesidad de implementar ciertas herramientas del estado del arte localmente, de manera de no depender de sistemas ajenos dado que, al no disponer de los códigos fuente se tornaba difícil responder a las preguntas de interés que fueran surgiendo mientras se llevaba adelante cada proyecto. Por eso, en esta tesis diseñamos nuestro propio pipeline que comentaremos brevemente.

2.2.4. Pipeline para la anotación de genomas bacterianos.

Como muestra la Figura 2.3, el pipeline toma como entrada las secuencias de *contigs* ya ensamblados y, para cada uno de ellos, ejecuta los programas mencionados previamente de la siguiente manera: 1) se ejecuta tRNA-SCAN para detectar las regiones que codifican para tRNAs. 2) Glimmer se usa para predecir las regiones codificantes permitiendo sólo 30 pares de bases de solapamiento y un largo mínimo para cada marco de lectura de 150 pares de bases. 3) Los marcos de lectura se traducen a la secuencia proteica subyacente y se ejecutan búquedas BLAST contra KEGG, PDB, COG y SWISSPROT.

También, se realizan búsquedas usando el programa HMMER sobre PFAM, TIGRFAM, CATH y SUPERFAMILY como se describió previamente. 4) Por último, los programas PSORT y SIGNALP se usan para predecir localización celular.

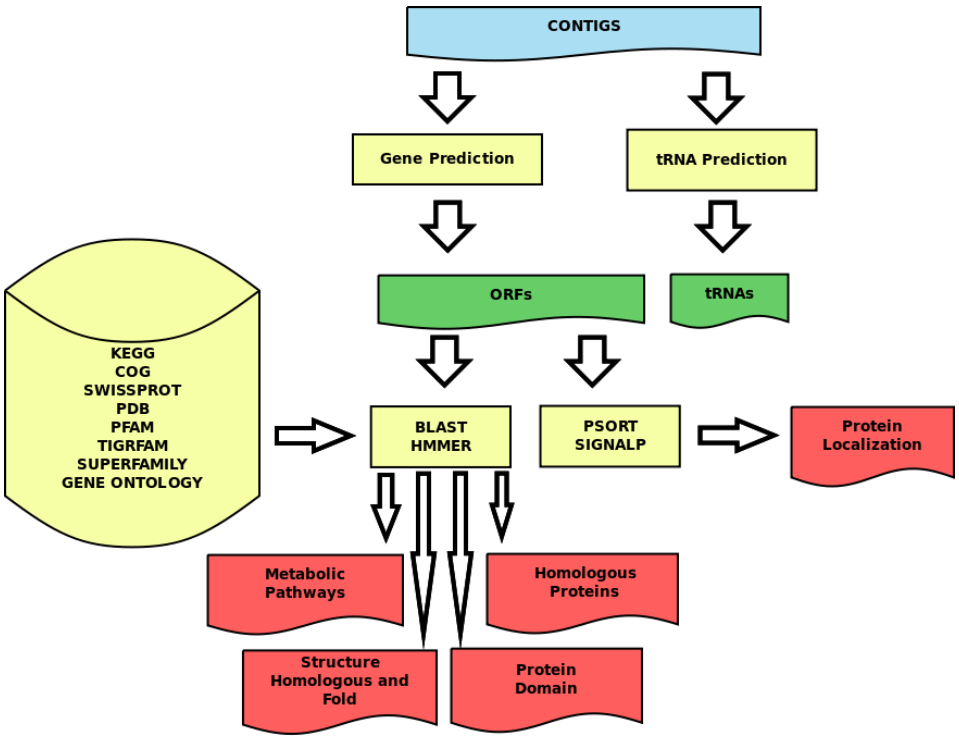

Figura 2.3: **Pipeline para la anotación funcional de genomas bacterianos.** A partir de los contigs se realiza la predicción de los genes y los tRNAs, usando Glimmer y tRNA-Scan respectivamente. Una evz obtenidos los genes, se procede a la anotación automática: Localización celular (Usando SignalP y PSORT), se asignan proteínas homólogas, vías metabólicas, plegados y dominios funcionales usando, BLAST y HMMER sobre varias bases de datos.

- **Proteínas homólogas y ortologías.** Usando BLAST contra COG, Swiss-Prot y KEGG se buscan proteínas homólogas a cada ORF. Obtener secuencias de proteínas homólogas que pertenezcan a otros organismos sirve para tener más confianza en el hecho de si el producto génico de un ORF que fue predicho es realmente una proteína funcional en el genoma que se estudia, porque al encontrarse en varios sistemas, las chances de que no sea funcional, son más bajas que si no lo tuvieramos más que

en el genoma que estamos anotando. Más puntualmente, usamos COG, para visualizar fácilmente la distribución de clases funcionales dentro de los genomas, y así, describir en términos generales el rol metabólico del organismo en su ecosistema. Usamos KEGG, para obtener información de vías metabólicas como explicaremos a continuación. Y, usamos Swiss-Prot para comparar con proteínas parecidas que hayan sido anotadas con alta calidad.

- **Vías metabólicas.** A las distintas especies de organismos bacterianos que pueden ser encontrados en los distintos ecosistemas del planeta, se las encuentra cumpliendo diferentes tipos de roles dependiendo de los nichos que habitan. Estos roles que adoptan dependen de las enzimas que se expresan en el interior de la célula, las cuales componen redes metabólicas que transforman los compuestos químicos que tengan biodisponibles. Las bases de datos que se encargan de definir los metabolismos suelen referenciar cada "paso metabólico" a una reacción catalizada por una enzima. Esto quiere decir que para poder mapear los metabolismos que hay en un genoma sobre los metabolismos que están disponibles en las bases de datos, primero debemos poder identificar las enzimas presentes en dicho genoma. Usando los resultados de búsquedas BLAST contra KEGG, asignamos un EC number a cada ORF, validando cada búsqueda usando BBHs. Así mapeamos cada ORF sobre los mapas metabólicos de KEGG, dando una visión del metaboloma subyacente al genoma que se desea analizar. Los mapas metabólicos ayudan al anotador para que pueda cerrar caminos metabólicos, si es que considera que puede haber enzimas que faltan y necesitan ser buscadas con más precisión en la secuencia genómica, o incluso, revisar la anotación automática por si una enzima faltante necesita ser reanotada.

- **Familias de proteínas y dominios funcionales.** Con los hits de HMMER contra las bases de datos Pfam y TigrFam, se obtienen la familia a las que pertenecen las proteínas deducidas de cada ORF y sus dominios funcionales. Los dominios funcionales, y las familias de proteínas que se representan usando HMMs, están curadas manualmente y tienen bien delimitados los extremos, lo cual es de gran utilidad, a la hora de seleccionar el codon start de cada ORF. A su vez, permiten visualizar la organización de dominios en una proteína que presenta una arquitectura multidominio, y detectar eventos de deleción o inserción fortuitos que hayan truncado el producto génico correspondiente dando lugar a una proteína que perdió su función original. Usando esta información es posible observar si los amino

ácidos en una secuencia corresponden con lo estipulado por la familia en lo referente a los residuos que se espera que estén conservados.

- **Clasificación de plegamiento y dominios estructurales.** Análogamente al paso anterior, para asignar anotaciones referentes a la estructura tridimensional a cada ORF, se realizan búsquedas usando HMMER contra CATH y Superfamily. Esta información permite saber si un determinado ORF es un buen candidato para un estudio estructural. El hecho de que, para un ORF, se obtengan hits en estas bases de datos permite asegurar que existe por lo menos una secuencia muy parecida para la cuál fue posible obtener información estructural, dado que las entradas de esta base de datos se hacen a partir de entradas en el PDB. A su vez, las jerarquías propuestas por CATH y Superfamily permiten visualizar a lo largo de todo el genoma la proporción de plegados presentes en el mismo, de la misma manera que con las categorías funcionales de COG.
- **Localización celular** Asignamos localización celular con SignalP y PSORT. Estos predictores se pueden aplicar a cada proteína por separado independientemente de las búsquedas en las bases de datos. Con signalP asignamos pétidos señal de exportación celular, y con PSORT asignamos las siguientes categorías de localización celular: Citoplasmáticas, Citoplásmicas de membrana interna, Citoplásmicas de membrana externa, Periplásmicas y Extracelulares.

2.3. Resultados

A continuación, mencionaremos dos casos de bacterias estudiadas usando las metodolgías mencionadas previamente. Un caso corresponde a una cepa del género Bacteroidete denominada *Bizionia argentinensis* JUB59(T), aislada en territorio antártico, y el otro corresponde a una cepa denominada *Exiguobacterium sp.* S17, aislada en lagunas de la región cuyana argentina, ambás bacterias fueron seleccionadas para estudiar la adaptación a ambientes extremos, de relevancia a la hora de descubrir enzimas de interés biotecnológico.

2.3.1. Bizionia argentinensis

La bacteria psicrotolerante *Bizionia argentinensis* JUB59(T) fue aislada en Territorio Nacional Antártico [Lanzarotti et al., 2011], [Bercovich et al., 2008] como parte del proyecto "Genoma Blanco". *Bizionia argentinensis* es un Bacteoroidete que pertenece al clado marino de la familia Flavobacteriaceae y suele

habitar aguas superficiales. Se ha visto que este tipo de bacteria es importante para el ciclo geoquímico del nitrógeno. Los miembros de este clado se encuentran repartidos ampliamente en ambientes marinos desde los trópicos hasta los oceanos polares, donde juegan un rol importante en la mineralización de la materia orgánica [Bauer et al., 2006]. El primer objetivo del proyecto consistió en entender mejor las adaptaciones a condiciones extremas en lecho marino y su rol ecológico, lo cual investigamos mediante la secuenciación y anotación del genoma del organismo. El segundo objetivo fue la búsqueda de proteínas que fueran novedosas desde el punto de vista funcional y estructural. En este sentido, usamos el genoma anotado para buscar proteínas cuya función y estructura no haya sido descripta previamente para luego caracterizarla estructuralmente por la técnica de Resonancia Magnética Nuclear (RMN). Para ello, en colaboración con el grupo de RMN del instituto Leloir se armó un pipeline para buscar proteínas en el genoma que cumplan con los siguientes requisitos: a) que tenga menos de 350 amino ácidos (Límite de la técnica de RMN), b) citosólicas o de exportación (no de membrana), c) con dominio estructural no conocido y d) con función desconocida. Entraremos en más detalle acerca de esto más adelante. Una vez seleccionadas bioinformáticamente las proteínas, fueron purificadas y caracterizadas por RMN. En esta tesis comentaremos el pipeline y un caso de éxito que dió lugar a 2 estructuras que fueron depositadas en el PDB.

Anotación de Bizionia argentinensis

Se realizó la secuenciación del genoma usando una tecnología *Roche 454 GS FLX system*. Se obtuvo un *shotgun* con 34.15X de covertura resultando en 405465 lecturas de alta calidad con un largo promedio de 277.1 de bases. Las secuencias fueron ensambladas usando Newbler y se obtuvieron 133 contigs a partir de los cuales llevamos adelante la anotación draft del genoma. La predicción de genes se hizo usando Glimmer 3.0. Los rRNA fueron predichos usando rRNAmmer y los tRNA usando tRNAscan-SE, mencionados previamente. Cada ORF fue anotado usando búsquedas BlastP contra Swissprot, COG y KEGG. También, hmmerscan se usó contra Pfam y Tigrfam para asignar dominios y familia a cada ORF. Finalmente, Signalp se usó para predecir péptidos señal. Los resultados se pueden ver resumidos en el Cuadro 2.1.

El análisis del genoma draft muestra que *B. argentinensis* tiene una densidad de codificación del 85.09 % y un contenido de GC del 33.77 %. Un total de 2940 ORFs fueron predichos, de los cuales, 979 fueron asignados como *hypothetical proteins* las cuales se definen como aquellas secuencias para las cuales no hay evidencia funcional en ninguna base de datos, porque o bien, no resultaron

en hits representativos o bien los hits también eran proteínas identifiacadas como *hypothetical proteins*. Además, 1583 fueron asignados a categorías de COG y 1195 a vías metabólicas en KEGG. Cabe aclarar que dentro de los ORFs anotados por COG, hay una gran cantidad que caen en dos categorías conocidas como *Function Unknown* y *General Function Prediction Only* como se puede ver en la Figura 2.4. O sea que, contando estos casos, más los mencionadas previamente, el genoma presenta más de un 40 % de proteínas de las cuales no se conoce su función específica, dejando en evidencia que todavía quedan "huecos" en el conocimiento que se pueda llegar a obtener acerca de la función de las proteínas de un genoma, usando sólo herramientas de comparación con bases de datos.

Ensamblado	
Número de contigs	77
Cobertura	34.15 X
Tamaño (pares de bases)	3293830
Contenido de G+C	33.78 %
Anotación	
Número de ORFs	2940
Densidad de codificación	85.1 %
tRNAs	35
Proteínas hipotéticas	979 (31.3 %)
ORFs asignados a clases de COG	1583 (53.8 %)
ORFs con dominios de Pfam	2122 (72.2 %)
ORFs con péptido señal	330 (11.3 %)
Localización	
Citoplasmáticas	1 333(45.3 %)
Citoplásmicas de membrana interna	542 (18.4 %)
Citoplásmicas de membrana externa	83 (2.8 %)
Periplásmicas	29 (1.0 %)
Extracelulares	22 (0.7 %)

Cuadro 2.1: **Estadísticas de la anotación de Bizionia argentinensis.**

En términos generales, el genoma de *B. argentinensis* contiene genes para las vías metabólicas del ciclo de Krebs, la glucólisis y las pentosas fosfato cubriendolas de forma completa. Consistente con los altos niveles de radiación UV-B fijada por la latitud a la que habita *B. argentinensis* e incluso por la presencia, del agujero de ozono de primavera, el genoma contiene 3 genes que codifican para fotoliasas y un sistema completo de reparación de DNA. Además, el genoma bacteriano presenta genes involucrados en la síntesis de carotenoides

y xantófilas. En particular, se encontró una copia de una β-caroteno hidroxilasa que cataliza la producción de β-criptoxantina y zeaxantina, pigmentos detectados previamente en otras flavobacterias antarticas [Humphry et al., 2001], pero no en miembros del género Bizionia [Bowman and Nichols, 2005], [Nedashkovskaya et al., 2005], [Nedashkovskaya et al., 2010]. De acuerdo con las permanentes bajas temperaturas del ambiente, el genoma también presenta genes codificando para la biosíntesis de ácidos grasos no saturados. Otro dato interesante de *B. argentinensis* es su capacidad para liberar peptidasas al ambiente para degradar compuestos circundantes. Hemos encontrado 24 genes pertenecientes a familias de peptidasas, en Pfam y/o Tigrfam, que contienen péptidos señal. Este hecho, podría servir para entender el rol ecológico de *B. argentinensis* en el ciclo natural del carbono atribuído a otros Bacteroidetes marinos.

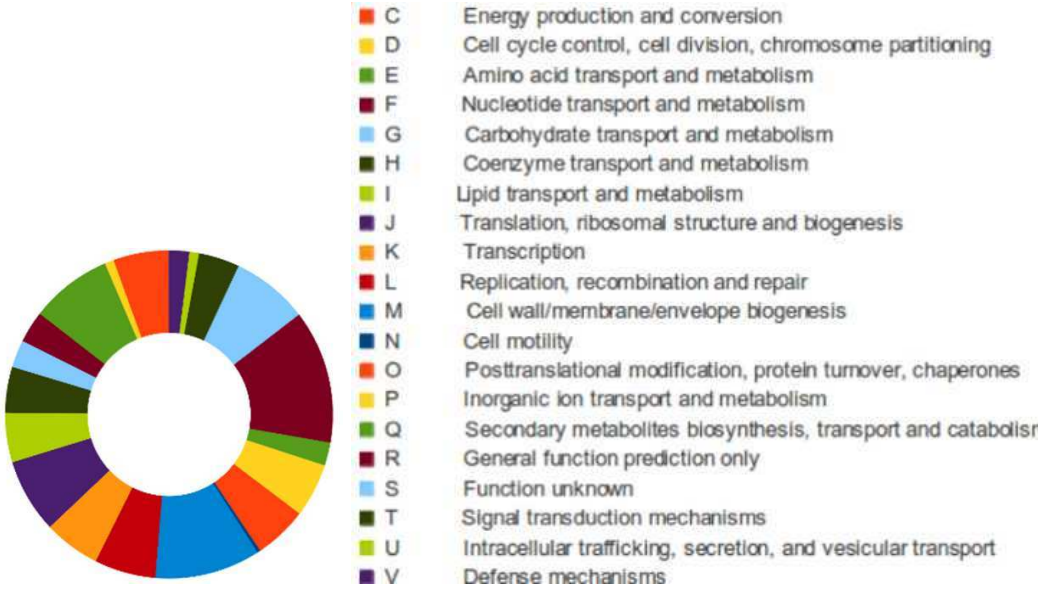

Figura 2.4: **Asignación de clases funcionales usando COG para Bizionia argentinensis.** En el panel izquierdo, gráfico de torta de la distribución de clases funcionales según COG. En el panel derecho, el código de colores para el gráfico de torta, los colores están en orden en sentido antihorario desde la parte superior.

Metabolismo del nitrogeno en Bizionia argentinensis

La secuenciación del genoma también reveló que *B. argentinensis* es una bacteria denitrificante anaeróbica facultativa capaz de reducir nitrito usandolo

como aceptor de electrones, capacidad que fue raramente reportada para los miembros del género Flavobacteriaceae [Jones et al., 2008]. Todos los genes relacionados al proceso de denitrificación fueron encontrados en el genoma, como muestra la Figura 2.5, con excepción de la nitrato reductasa. Por la vía asimilatoria, el NO_2- es reducido por otra nitrito reductasa que forma amoníaco (EC 1.7.2.2). El amoníaco, entonces, es usado como fuente de nitrogeno para, por ejemplo, la síntensis de la glutamina catalizada por la glutamina sintetasa (EC 6.3.1.2, también presente en el genoma). Por otro lado, *B. argentinensis* presenta enzimas para reducir nitrito hasta nitrogeno, completando el proceso de denitrificación desde el nitrito. En particular, la reducción de nitrito (NO_2-) a óxido nítrico (NO), es catalizada por una nitrito reductasa que contiene cobre (EC 1.7.2.1), luego, el óxido nítrico es reducido a óxido nitroso (N_2O) catalizado por una enzima que contiene hierro hémico y no-hémico dinuclear (EC 1.7.2.5), y finalmente, el óxido nitroso se reduce a nitrogeno gaseoso (N_2) catalizado por una reductasa (EC 1.7.2.4) con un núcleo mixto de cobre nuclear y un cluster de cobre-azufre, siendo el donor de electrones un citocromo c552.

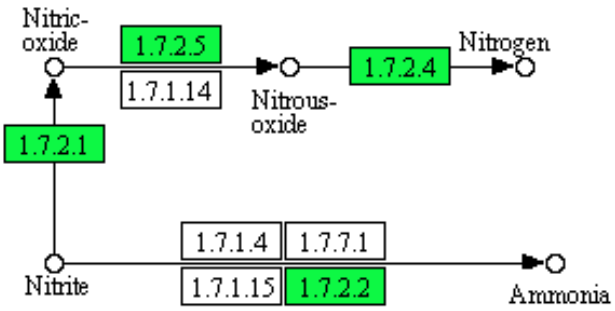

Figura 2.5: **Sistema de denitrificación de Bizionia argentinensis.** Por la vía asimilatoria, el nitrito es reducido a amoníaco mediante la catálisis de una nitrito reductasa (1.7.2.2). Por otro lado, la producción de nitrógeno gaseoso a partir de nitrito se realiza por la acción de tres enzimas: una nitrito reductasa que contiene cobre pasa el nitrito a óxido nítrico, luego el óxido nítrico es reducido a óxido nitroso por una reductasa de óxido nítrico (1.7.2.5), y por último, el óxido nitroso se reduce a nitrogeno gaseoso catalizado por una reductasa de óxido nitroso (1.7.2.4)

Búsquedas de enzimas con estructura y función desconocida: BA42 Como se mencionó previamente, este genoma fue utilizado para buscar enzimas con función desconocida que se encuentren adaptadas a bajas temperaturas. Para

esto, aplicamos un conjunto de filtros que permiten dejar como resultado una lista de ORFs con propiedades específicas, para realizar estudios estructurales por RMN. Como muestra la Figura 2.6, definimos filtros que buscan por genes de función y estructura desconocida.

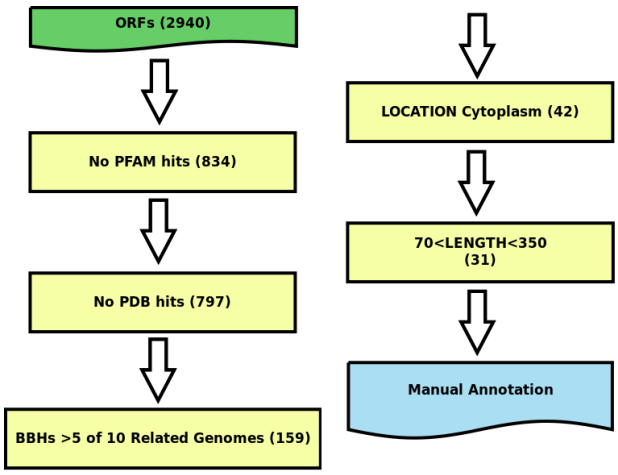

Figura 2.6: **Filtros para obtener enzimas para estudios mediantes técnicas de RMN.** Los filtros se aplican en forma consecutiva, los números entre paréntesis indican los ORFs que van quedando con la aplicación de cada uno de ellos. Los primeros dos filtros buscan conseguir proteínas con estructura y función desconocida. El tercer filtro es para verificar que los genes que expresan estas proteínas estén representados en varios genomas, para eliminar los falsos positivos de la predicción de genes. Los últimos 2 filtros son para que las proteínas cumplan con los requisitos experimentales: que sean solubles y no muy grandes.

Primero, se usaron los hits de Pfam para filtrar todos aquellos ORFs que no hayan conseguido hits significativos, de manera de contribuir en la definición de alguna familia aún no descripta. Cabe aclarar que aquellos dominios denominados *DUF (Domain of Unknown Function)* no fueron retenidos en el filtro, dejando pasar los ORFs con hits significativos en estos dominios a los pasos sucesivos. Segundo, se usaron los hits de PDB para filtrar todos aquellos que no tengan estructura conocida. Tercero, se computaron los BBHs contra 10 genomas relacionados para asegurar un grado extra de confianza en la predicción de los genes seleccionandose con 5 BBHs o más. Como cuarto filtro, seleccionamos sólo aquellas con predicción de localización citoplasmática o de exportación de manera de aumentar las chances de que sea soluble para facilitar la purificación de la proteína. Por último, se aplicó un filtro de máximo

350 amino ácidos para el largo de la secuencia, como requisito de la técnica de RMN, y cómo mínimo 70 amino ácidos para evitar polipéptidos de función muy general.

Las secuencias resultantes de este proceso de filtrado fueron curadas manualmente y de entre todas ellas la que cabe destacar es BA42. Con esta proteína se continuó estudiando mediante la técnica de RMN [Smal et al., 2012]. A su vez, el estudio experimental dió lugar a un PDB producto de los resultados de RMN, y además, la proteína siguió estudiandose y se llegó a obtener un cristal que se usó para resolver la estructura mediante rayos X [Aran et al., 2014]. Esta proteína tiene 145 amino ácidos y era miembro de una familia de Pfam de función desconocida (**DUF477** - PF04536), que gracias de estos estudios estructurales, se cambió el nombre de la familia a **TPM_phosphatase**.

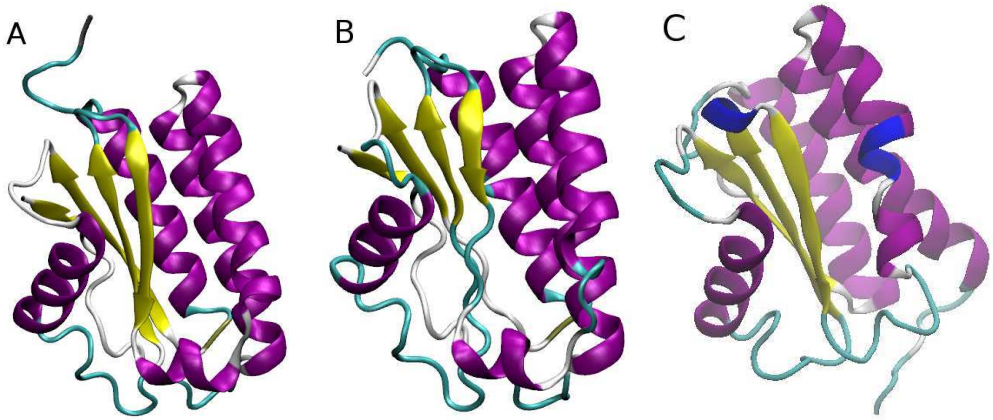

Figura 2.7: **Estructuras de la proteína BA42** A) Estructura resuelta por la técnica de RMN. B) Estructura resuelta mediante Rayos X. C) Estructura inferida mediante modelado comparativo.

Como se puede ver en la Figura 2.7A (RMN) y 2.7B (Rayos X), la estructura está compuesta por un dominio TPM y muestra una variante topológica nueva de 4 hojas beta, que constituyen la hoja beta central de un motivo alfa-beta-alfa. También posee dos sitios de unión a metales (no se muestra) que estabilizan un par de loops cruzados que no fue observada en miembros de esta familia previamente. Se vió que tanto la estructura como la dinámica de ésta proteína cambian con la concentración de metales divalentes, lo cual se observa en proteínas que modulan su actividad en función de la concentración de metales.

Como dato interesante, mientras se resolvía su estructura se publicó la es-

tructura de una *homóloga remota* que nos permitió realizar un modelo mediante técnicas de modelado comparativo que serán el tema del próximo capítulo. El modelado fue realizado usando una proteína relacionada, pero con identidad menor al 20 % de identidad que es considerado el límite del modelado comparativo. Esto nos dice que si bien se sabe que en identidades bajas la técnica llega al límite en términos del error con el que se modela cada átomo, la asignación de plegado, en términos generales, presenta un alto nivel de similitud. Se puede ver (Figura 2.7C) que a pesar de la baja identidad de secuencia con el templado que usamos para construir el modelo, obtuvimos practicamente el mismo plegado con alguna diferencias sobre todo de las conformaciones de los residuos según sus estructuras secundarias. Esto nos motivó a seguir adelante con las técnicas estructurales de manera de complementar las herramientas de anotación clásicas.

2.3.2. Exiguobacterium sp. S17

Los Lagos Andinos de Gran Altura (high-altitude Andean lakes, HAAL) consisten en varios lagos superficiales localizados en un desierto de gran altura sobre el nivel del mar conocido como La Puna. Están expuestos a condiciones ambientales *extremas* como altos niveles de radiación UV, salinidad elevada, y la presencia de metales pesados y metaloides principalmente arsénico [Dib et al., 2008], [R. Flores et al., 2009], [Dib et al., 2009], [Zenoff et al., 2006], [Ordoñez et al., 2009]. El género Exiguobacterium fue identificado como el taxón más dominante en HAAL. Exiguobacterium spp. ha sido hallado en una amplia variedad de habitats incluyendo frío y calor extremo, con temperaturas variando de -12° a 55° centígrados [Rodrigues and Tiedje, 2007], [Vishnivetskaya et al., 2006]. Este hecho confiere un interés sustancial en el género como un potencial sistema modelo para la investigación de los atributos que puedan explicar estas adaptaciones y los procesos evolutivos detrás de la adaptación a estos regímenes térmicos [Vishnivetskaya et al., 2009].

En colaboración con el Laboratorio de Investigaciones Microbiológicas de Lagunas Andinas en el PRIOMI de Tucumán, se llevó adelante el estudio de la resistencia a arsénico en estos ambientes extremos. Trabajamos con una cepa del genero Exiguobacterium denominada *Exiguobacterium sp.* S17 que fue aislada de la Laguna Socompa, que presenta concentraciones elevadas de arsénico. Esta cepa fue comparada con otra del mismo género denominada *Exiguobacterium sp.* N139, que fue aislada de otra laguna que no tiene arsénico en concentraciones tan elevadas como Socompa. Se llavaron adelante ensayos de tolerancia a arsénico, en presencia de arsenato y arsenito, y luego, com-

paramos los genomas de ambas cepas en cuanto a sus genes de resistencia a arsénico.

Aquí presentamos la secuencia del genoma borrador de *Exiguobacterium sp.* cepa S17 [Ordoñez et al., 2015], que fué aislado de un estromatolito ubicado en la Laguna Socompam, en el norte de Argentina, en los HAAL [Farías et al., 2013]. El genoma se obtuvo usando una estrategia de shotgun a genoma entero con un pirosecuenciador 454 GS Titanium en el instituto de agrobiotecnología de Rosario (INDEAR), Argentina. El ensamblado fue hecho usando Newbler version 2.5.3 usando la opción urt con una covertura genómica del 63X, obteniendose 193 contigs largos. El borrador del genoma tiene 3139227 bases de largo, con un contenido GC de 53.14 %. La anotación del genoma se realizó usando los procedimientos mencionados previamente, y además se anotaron usando ISGA [Hemmerich et al., 2010] y RAST [Aziz et al., 2008] de manera de mejorar la anotación comparando con otras dos anotaciones más.

Ensamblado	
Número de contigs	193
Cobertura	63X
Tamaño (pares de bases)	3139227
Contenido de G + C	53.14 %
Anotación	
Número de ORFs	3218
Densidad de codificación	86.12 %
tRNAs	48
Proteínas hipotéticas	1149 CDSs (36 %)
ORFs asignados a subsistemas de RAST	1381 (43 %)

Cuadro 2.2: textbfEstadísticas de la anotación de *Exiguobacterium sp. S17.*

La anotación devolvió un total de 3218 ORFs y fueron predichos 48 tRNAs. La anotación cubrió 360 subsistemas de RAST con 1381 (43 %) del total de los ORFs, y 1149 (36 %) fueron anotadas como proteínas hipotéticas. A su vez, *Exiguobacterium sp.* S17 presenta 102 genes relacionados a la respuesta a stress según RAST, una mayor cantidad de genes que los reportados para otros Exiguobacterium. S17 contiene un sistema completo de reparación de ADN, incluyendo UvrABC, MutL-MutS y fotoliasas bacterianas, y varios genes relacionados a resistencia a compuestos tóxicos, cómo antibióticos, arsénico, cadmio y mercurio. La alta resistencia a arsénico mencionada previamente en S17 puede ser explicada basandose en el número mayor de genes reportados para detoxificar éste compuesto, que serían 7 genes. Una diferencia importante entre

S17 y los otros Exiguobacterium spp. es la presencia del gen Acr3, que es una bomba conocida por contribuir a detoxificar la célula de arsenito, una de las especies más tóxicas del arsénico. En resumen, este genoma reveló adaptaciones esenciales para la supervivencia bajo condiciones extremas y es un modelo atractivo para estudiar los mecanismos de tolerancia de factores ambientales extremos, permitiendo la identificación de nuevos sistemas explotables para la bioremediación de metales y metaloides.

Resistenca a arsénico en Exiguobacterium sp. S17

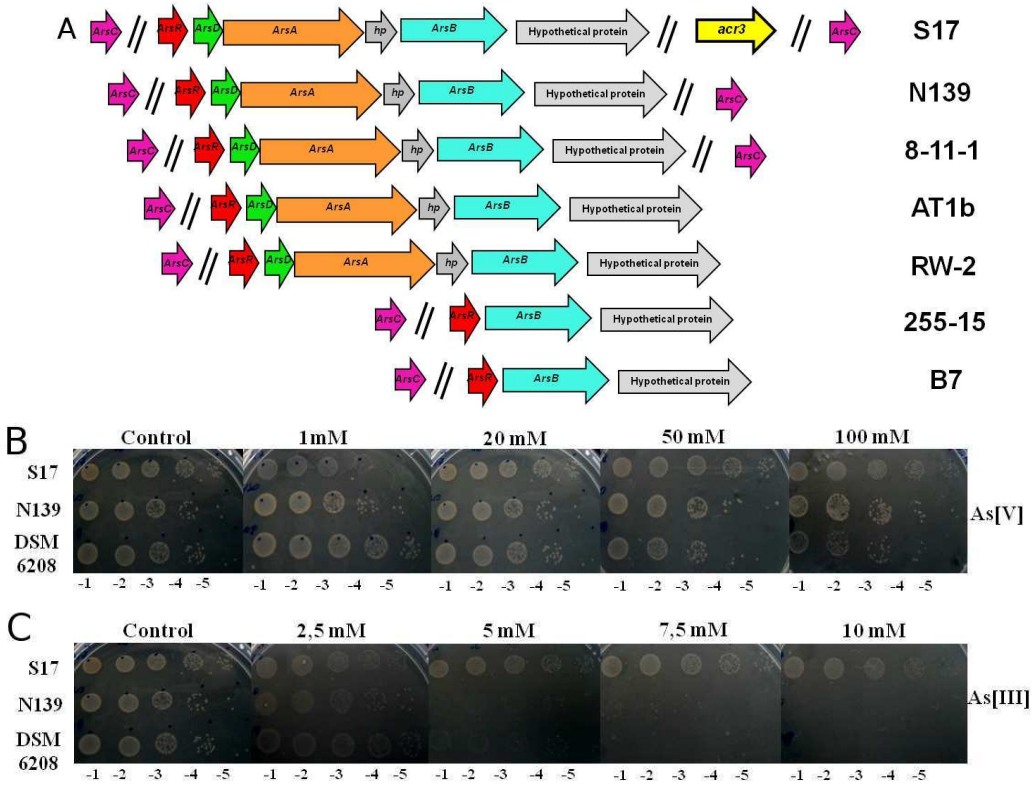

Figura 2.8: **Tolerancia y resistencia a arsénico en Exiguobacterium.** A) Ortoglogías de resistencia a arsénico obtenidas mediante BBHs entre los genomas secuenciados del género Exiguobacterium. B) Ensayo de tolerancia a arsenato para *Exiguobacterium sp.* S17, *Exiguobacterium sp.* N139 y una cepa control denominada *Exiguobacterium sp.* DSM 6208. C) Idem B) pero para tolerancia a arsenito.

Para analizar la resistencia a arsénico en S17, usamos la técnica de BBHs de manera de obtener las ortologías relacionadas a dicha resistencia, presentes en todos los genomas disponibles del género Exiguobacterium. Como muestra la Figura 2.8A, vimos que de entre todos los genomas de Exiguobacterium, S17 es la cepa que presenta más genes de resistencia a arsénico. De hecho, se hicieron ensayos de tolerancia comparando la respuesta entre S17 y N139, y resultaron en que tanto S17 como N139 toleran un cierto nivel de arsenato (As[V]), siendo capaces de crecer a concentraciones de hasta 100mM (Figura 2.8B). Pero, en presencia de arsenito (As[III]), sólo la cepa S17 fué capaz de crecer a concentraciones por encima de 5mM, mientras que N139, no creció a concentraciones mayores a 2.5mM (Figura 2.8B).

Teniendo en cuenta que el genoma de N139 está bastante relacionado con S17, en términos de las ortologías que tienen en común, 68.9 % y 73.6 % del total de los ORFs para S17 y N139, respectivamente, vimos que la única diferencia en términos de las resistencias conocidas al día de hoy, era la presencia del gen *acr3* que codifica para una bomba de extrusión de arsénico. Éste gen está presente en una enorme cantidad de organismos aislados y secuenciados en la Laguna Socompa. Hoy día, no hay información tridimensional de esta proteína, sólo hay trabajos acerca de su topología transmembrana. Por eso, utilizando herramientas de modelado estructural fue propuesto un modelo tridimensional de esta proteína [Ordoñez et al., 2015], usando servidores web especiales para el estudio de proteínas transmembrana: TMHMM, DAS, HMMTOP y TOPCONS, y otros de predicción de estructura secundaria: PSIPRED y PORTER. Además, usamos una evaluación de calidad para estos modelos de proteínas de membrana que se llama ProQM. Identificamos 10 segmentos transmembrana de tipo hélice alpha, separando aquellos residuos que miran a la región citosólica y aquellos que dán a la región extracelular. El templado que elegimos para usar como base en la confección del modelo 3D, es la estructura 3zuxA de una bomba de sodio/ácido bílico, que tiene 16 % de identidad y 93 % de cobertura en el alineamiento que obtuvimos. El análisis de ProQM arroja un puntaje de 0.714 (0-bajo, 1-alto), que representa un buen modelo a pesar de la baja identidad de secuencia. Además para validar el modelo, efectuamos un alineamineto con otras dos secuencias de Acr3: una de *Bacillus subtilis* (BsAcr3) [Aaltonen and Silow, 2008] y la otra de *Alkaliphilus metalliredigens* (AmAcr3) [Fu et al., 2009]. Fueron elegidas estas proteínas tanto por la disponibilidad de propuestas de topologías en la bibliografía como por la existencia de ensayos que corroboran la accesibilidad al solvente (citoplasmático o extracelular) de varios residuos que nosotros usamos para comparar con el modelo que obtuvimos. En el caso de BsAcr3, la accesibilidad fue analizada usando PhoA y

GFP, y en el caso de AmAcr3 se usó la técnica de escaneo de accesibilidad de cisteínas (SCAM). El alineamiento entre estas tres proteínas fué obtenido usando el software HMMER3 con la familia Acr3 de TIGRFAM (TIGR00832 - acr3: arsenical-resistance protein). Usando esta información, mapeamos los residuos que pertenecen a las regiones intra y extracelulares en la estructura y en el alineamiento.

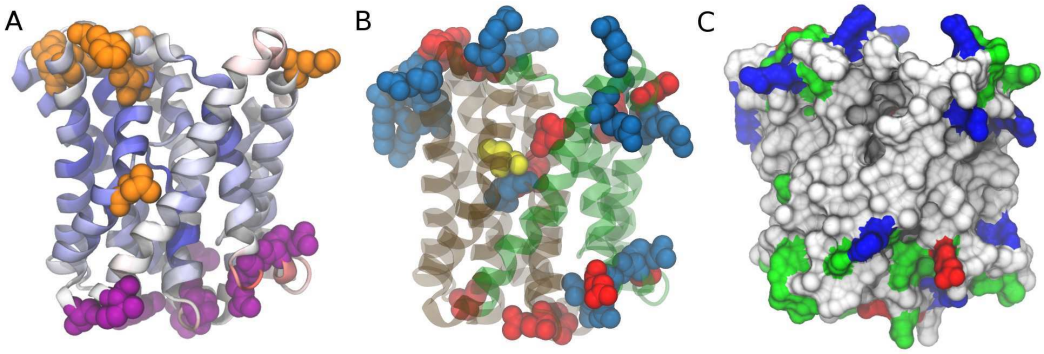

Figura 2.9: **Modelado de la bomba de extrusión de arsenito.** A) Evaluación de calidad del modelo de Acr3, el backbone está coloreado por residuo según la estimación de ProQM [Ray and Lindahl, 2010]: azul es buena calidad y rojo es calidad pobre. Los residuos en naranja corresponden a aquellos que presentan evidencia experimental de estar orientados al citosol en proteínas homólogas, y los residuos en violeta corresponden a aquellos expuestos al periplasma. B) En amarillo, una cisteína altamente conservada en esta familia, en rojo los residuos cargados negativamente y en azul los cargados positivamente. C) Visualización de superficie con las zonas hidrofóbicas en blanco, las zonas polares en verde, y en rojo y azul las zonas cargadas negativa y positivamente.

En la Figura 2.9A se puede ver que los residuos que coinciden con aquellos que fueron previamente mapeados al lado externo de la membrana, efectivamente, en el modelo, caen en loops del lado extracelular. Con la excepción de una leucina en la posición 137 (que Aaltonen et al. refieren a una metionina en la misma posición), que pareciera caer en la mitad de una hélice transmembrana. Si bien esto pareciera estar contradiciendo el modelo que se sigue de los predictores de segmentos transmembrana, en el modelo 3D se puede ver que esa posición cae en una región que se abre y se cierra para el transporte de los iones a través de la membrana, lo que puede explicar la accesibilidad al solvente. Además, se puede ver que hay una región donde las hélices alfa presentan una ruptura de su forma helicoidal formando unas pequeñas orquillas que están reportadas como un motivo común en transportadores de iones [Sciara and Mancia, 2012]. Estas orquillas están entre las posiciones 106 a 109

y las posiciones 263 a 266. Además, como se puede ver en la Figura 2.9B, hay una cisteína conservada (Cys107) que es *esencial* para la función de está proteína (demostrado mediante experimentos de mutagenesis en otras proteínas de la misma familia), situada justo entre estas dos orquillas que conforman el núcleo de la estructura, rodeado por los segmentos transmembrana que lo forman. Basandonos en la estructura resuelta por Hu et al. sobre la bomba de sodio/ácido bílico que usamos como templado para el modelo de Acr3, se pueden apreciar dos dominios, uno denominado *panel* y otro denominado *núcleo*, coloreados como verde y gris respectivamente en la Figura 2.9B. Finalmente, se aprecia una cavidad de entrada en la Figura 2.9C, que se puede contrastar con la Figura 2.9B donde queda claro como ésta entrada se sitúa justo en el medio de los dos paneles.

2.4. Conclusiones

Los sistemas de anotación automática de genomas bacterianos son posibles gracias a tres cosas: i) los métodos de predicción de regiones codificantes, ii) los métodos de búsqueda en bases de datos y, por supuesto, iii) las bases de datos.

Sin los métodos de predicción no podríamos conocer elementos genómicos nuevos, que no hayan sido caracterizados previamente. Sin los métodos de búsqueda no tendríamos *cómo* comparar lo que los métodos de predicción encontraron, y sin las bases de datos no tendríamos con *qué* comparar para poder caracterizar su función. De esta manera, la anotación genómica se ha convertido en una maquinaria que se retroaliementa. Las mismas bases de datos que se usan para anotar los nuevos genomas, se van incrementando con las anotaciones de los genomas anotados por ellas mismas. Esto tiene un problema como el lector se estará imaginando, que es que si los datos no se van actualizando y/o se van revisando de manera de quitar errores o se actualizan las entradas viejas con los datos nuevos, el sistema global de anotación puede ir rumbo a una pérdida de información importante. Es por esto que la anotación funcional como fué descripta en este capítulo y como está en el estado del arte, sólo es un paso intermedio en la caracterización funcional de una proteína, dado que siempre es necesaria la curación manual de las características anotadas.

En este trabajo, hemos desarrollado un protocolo que permite anotar genes basandose en búsquedas por similaridad en bases de datos, tanto de secuencias (FASTA) como de modelos ocultos de markov (HMM). Usando la información producto de la anotación diseñamos un conjunto de filtros que aplicados en forma de pipeline, permite seleccionar ORFs sín información estructural

disponible, que dieron lugar al estudio estructural y funcional de una proteína de Bizionia argentinensis. Además, usando el genoma de varios Exiguobacterium, caracterizamos la resistencia a arsénico en *Exiguobacterium sp. S17* y modelamos estructuralmente una bomba de extrusión de arsénico denominada Acr3, que está presente sólo en éste genoma.

Todas las metodologías mencionadas previamente, carecen de algo fundamental para el mejor entendimiento de la función proteica desde el punto de vista bioinformático: el mapeo de información funcional sobre la estructura. Se sabe que pequeñas variantes en la secuencia de una proteína pueden producir cambios rotundos en su estructura y su función. Esto da a lugar a pensar que los sistemas de anotación pueden inferir a nivel general la función proteica, pero les es imposible capturar dichos cambios en cada variante dentro de la misma familia. Para esto, intentamos un enfoque estructural que será mencionado en el próximo capítulo, mostrando como es posible mapear información funcional, dada una estructura proteica conseguida a partir de modelos 3D contruidos por homología con estructuras conocidas.

Capítulo 3

Modelado Estructural de Proteomas Bacterianos

3.1. Introducción

Predecir la estructura de una proteína consiste en definir la posición de los átomos que la conforman, partiendo de la secuencia (estructura primaria) de dicha proteína. Varias estrategias distintas existen para encarar estudios de estos tipos que van desde la parametrización de sistemas de simulación y posterior muestreo del espacio conformacional 3D que pueda adoptar el polipeptido, hasta herramientas de búsquedas en bases de datos que asignan estructura 3D por que existe un grado de similaridad entre proteínas.

No se puede mencionar la predicción de estructuras de proteínas sin mencionar la Critical Assessment of Structure Prediction (CASP). La CASP es una competencia que arrancó a mediados de la década del 90 y busca evaluar el *estado del arte* de los métodos de predicción de estructura proteica a partir de su secuencia. Los organizadores de la CASP proponen secuencias de proteínas para modelar, habiendo resuelto sus estructuras previamente pero dejándolas sin publicar. De esta manera, laboratorios de bioinformática estructural alrededor del mundo, forman equipos que intentan predecir la estructura que adoptan dichas proteínas para, luego, comparar los resultados obtenidos. La competencia ya va por su décima edición mostrando que a medida que fueron pasando los años se fué aprendiendo de las limitaciones de estas técnicas y de sus capacidades de progreso. Mencionamos esta competencia, porque centraliza el lenguaje científico acerca del área y, en particular, propone 3 categorías básicas sobre las cuales se desarrollaron los métodos de predicción de estructura. Estas categorías son:

- **Ab initio** Estos métodos son computacionalmente costosos. Pretenden resolver el problema usando como información sólo la secuencia proteica (denominada *target*), sin información de bases de datos. Parten de conocimiento previo modelando las interacciones fisicoquímicas presentes en el sistema y, usando sistemas de simulación, mediante dinámica molecular o muestreo por Monte Carlo, muestrean las conformaciones que puede adoptar la cadena de aminoácidos. Al principio, estos sistemas usaban sólo potenciales de interacción clásicos dónde los términos que definen los grados de libertad del sistema están derivados únicamente de los parámetros fisicoquímicos del mismo. Pero, a lo largo de las útlimas décadas fueron desarrollandose nuevos términos usando principios estadísticos, que modelan aspectos bioquímicos (globales como la topología y locales como la estructura secundaria) del proceso de plegado. Por ejemplo, usando información de fragmentos de pequeños péptidos con conformaciones conocidas que guíen la simulación al estado nativo. Estas metodologías han mostrado ser unas de las mejores maneras de conseguir información acerca de la estructura tridimensional de una determinada proteína, bajo condiciones en las que es difícil o hasta imposible obtener información experimental para una determinada secuencia de proteína y la secuencia no tiene una similitud razonable con otras proteínas cuya estructura es conocida.
- **Enhebrado (o reconocimiento de plegado)** Esta técnica busca identificar entre los plegados disponibles en las bases de datos, aquellos que puedan llegar a ajustarse a la secuencia porporcionada, a pesar de no haber una similitud detectable entre la secuencia que se desea modelar y la secuencia de la estructura candidata. Para esto, evalúan parámetros energéticos a medida que prueban como se adapta la secuencia proteica a cada plegado entre una base de datos de plegados conocidos. Esta prueba se realiza teniendo en cuenta parámetros energéticos de la estructura a medida que se computa un alineamiento. De manera que es posible usar estrategias parecidas a las mencionadas en el capítulo anterior para generar búsquedas en bases de datos, sólo que el sistema de puntuación usa un potencial energético estructural en lugar de la similitud en secuencia entre los residuos a travez de matrices tipo BLOSUM o PAM. Por eso, se desarrollaron (y se siguen desarrollando) estrategias de alineamiento (secuencia-estructura) y puntuación energética para mejorar las capacidades predictivas de estos algoritmos, aunque no suelen dar tan buenos resultados como los obtenidos si conociéramos la estructura de una proteína homóloga.

- **Modelado comparativo** Estos métodos son computacionalmente poco costosos comparado con los mencionados previamente. Consisten en utilizar el conocimiento disponible para comparar con estructuras conocidas pero a un nivel del ensamble conformacional completo, o sea incluyen muchos más grados de libertad en términos del largo de la secuencia. A diferencia de los otros métodos, el modelado comparativo utiliza estructuras completas de proteínas conocidas y las emplea como molde (usualmente denominado *templado*) para, posteriormente, estimar el grado de precisión con el que se realiza el modelado. Esta técnica permite muestrear el espacio conformacional sólo en la medida de disponer de dichas conformaciones entre los datos disponibles. A medida que los datos aumentan en las bases de datos también aumenta el cubrimiento del espacio conformacional de estructuras y por énde el poder predictivo de ésta técnica. Por último, cabe mencionar que hay proteínas para las cuales no se obtiene información alguna buscando en bases de datos, dejandonos sin otra posibilidad que usar estrategias ab initio. Es por esto que el modelado comparativo se considera de las formas más sencillas de obtener información acerca de la estructura de una proteína.

En este trabajo de tesis nos propusimos poner a punto un proceso lo suficientemente preciso como para obtener resultados confiables, pero lo suficientemente rápido como para procesar un genoma bacteriano en tiempos accesibles, puesto que nuestro objetivo es el de aumentar el espacio estructural de patógenos y/o bacterias de interés biotecnológico. Para ésto, la mejor opción es el modelado comparativo dado que, como primer acercamiento, permite obtener información tridimensional en poco tiempo y, además de eso, permite ordenar las secuencias de un proteoma desde aquellas que tienen alta identidad con proteínas de estructura conocida hasta las que carecen de posibles templados para realizar el propio modelado. Por eso comentaremos brevemente los pasos fundamentales que componen esta técnica para definir un proceso que cumpla con estos objetivos.

3.1.1. Modelado comparativo

El modelado comparativo no es nuevo, sino que surgió en 1969 con el trabajo de Browne et. al. [Browne et al., 1969] Éste método permite obtener información acerca de la estructura terciaria de una proteína basandose sólo en su secuencia. Esta técnica se basa en que existe una relación fuerte entre la secuencia y la estructura (o plegado) de una proteína. Entonces, si dos secuencias se parecen en términos de su similaridad en secuencia, se puede suponer que

sus estructuras serán similares también. Se ha probado, que esta técnica ampliamente usada, da buenos resultados para expandir el espacio estructural de proteomas en general [Pieper et al., 2014]. Una de las estrategias más famosas consiste en dividir el procedimiento en 4 pasos básicos que pueden complejizarse tanto como uno quiera. Basandose en el trabajo de [Marti-renom et al., 2003], los cuatro pasos principales son: i) **seleccionar el/los templado/s**, ii) **alinear la secuencia *target* con el/los templado/s**, iii) **construir modelos basado en los templados** y iv) **evaluar cada modelo de manera de elegir uno o mas modelos representativos**. Estos cuatro pasos, que se ven esquematizados en la Figura 3.1 y serán explicados en la próxima sección, constituyen cuatro estrategias generales que pueden ser resueltas con el nivel de precisión que uno desée, teniendo en cuenta, que si éste nivel de precisión aumenta, también lo harán el tiempo y/o los recursos de cómputo.

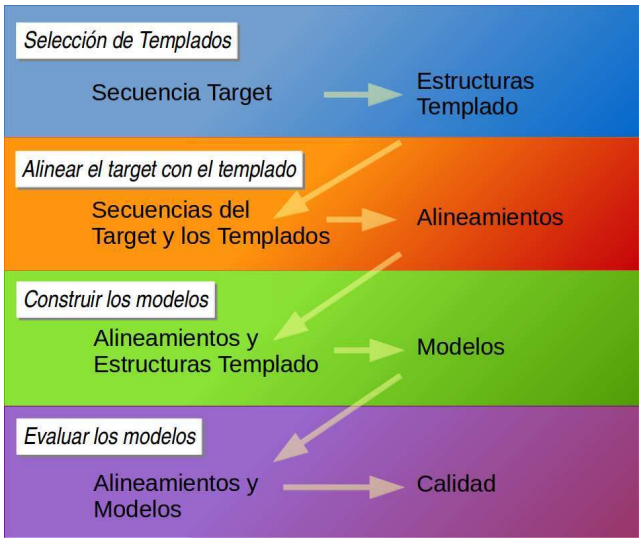

Figura 3.1: **Pasos del modelado comparativo de estructuras.** El proceso de modelado comparativo está dividido en 4 pasos fundamentales: 1) Se obtienen *templados* relacionados a la secuencia *target* que se desea modelar. 2) Se alinean la secuencia *target* con las secuencias de los templados, de manera de obtener los alineamientos de los cuales se deducirán las coordenadas espaciales del modelo. 3) Se producen varios modelos con las coordenadas 3D para los átomos de la estructura del *target*, en función de las estructuras templado y los alineamientos obtenidos del paso anterior. 4) Se evalúa la calidad de los modelos construidos para asignar un grado de calidad (bueno, malo, etc.)

Selección de templados

El proceso de modelado comparativo consiste en modelar la estructura de una proteína a partir de una estructura conocida que se le parezca, por eso, lo primero que hay que hacer es encontrar dicha estructura. Lo más fácil que se puede hacer para resolver este problema es usar las secuencias subyacentes a las estructuras que hay en el PDB, construir una base de datos de secuencias y realizar búsquedas por similaridad usando BLAST. Este primer acercamiento es bastante veloz en términos de procesamiento, pero se sabe que suele presentar ciertos problemas. Por otro lado, como indica la Figura 3.2, se sabe que hay un límite a partir del cuál se puede esperar cierta similitud estructural. Por debajo de una determinada cantidad de residuos alineados iguales, es poco aconsejable obtener una estructura [Rost, 1999], [Sander and Schneider, 1991].

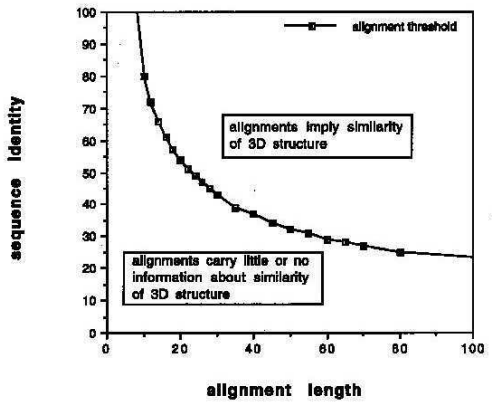

Figura 3.2: **Límite para la homología estructural.** *Esquema deducido del trabajo de [Sander and Schneider, 1991].* Dado un alineamiento, existe un límite a partir del cuál no es posible inferir homología estructural.

Sin embargo, es bien sabido que puede haber proteínas que tengan muy baja similaridad en secuencia, como ser entre 20 % y 30 % del total de residuos alineados iguales y, aún así, sus estructuras pueden estar relacionadas. Los hits que produce BLAST suelen bajar sólo hasta porcetajes de identidad de entre 30 % y 50 % cuando se pone un corte en el *e-value* de entre 10^{-3} y 10^{-5}, por lo que estaría dando lugar a errores de tipo II o falsos negativos. Estos serían casos en los que hubiera sido posible obtener un templado para la estructura que deseamos modelar pero no lo estamos encontrando. Cuando ésta búsqueda se pone difícil (i.e: con un BLAST no alcanza para encontrar

estructuras templado), se suele construir un *perfil* de la secuencia *target* usando la mayor cantidad de secuencias homólogas parecidas a ella. Existen muchas formas de realizar un perfil de una secuencia proteica y todas esas formas tienen en común que alinean todas las secuencias involucradas y luego modelan estadísticamente cada columna del alineamiento múltiple. Para esto, existe una variante de BLAST denominada PSI-BLAST, la cual permite realizar búsquedas iterativas sobre una determinada base de datos (preferiblemente una base de datos grande, con muchas secuencias) construyendo a cada paso (o sea, en cada nueva iteración) una matriz de puntuación de posición específica (PSSM). Esta matriz o PSSM, tiene por objetivo modelar la frecuencia con la que aparece cada amino ácido en cada posición de la secuencia *target* en el contexto de varias secuencias que se le parecen y que fueron seleccionadas a cada paso por el algoritmo. Luego de algunas iteraciones (usualmente entre 3 y 10) se detiene la búsqueda iterativa guardandose la útlima PSSM que constituye un perfil para la secuencia *target*. Finalmente, es posible reiniciar la búsqueda PSI-BLAST, pero esta vez sobre la base de datos de las secuencias del PDB, utilizando la PSSM construida previamente para obtener resultados con un esquema de puntuación basado en proteínas parecidas a la secuencia *target*. Este proceso se puede ver esquematizado en la Figura 3.3.

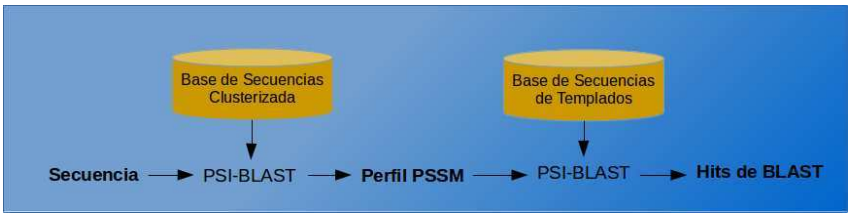

Figura 3.3: **Búsqueda de homólogos estructurales lejanos.** Dada la secuencia de una secuencia se construye un perfil estadístico por posición (en la secuencia) usando PSI-BLAST. El primer paso consiste en recaudar secuencias parecidas de una base de datos de muchas secuencias *clusterizada* por identidad de secuencia. Finalmente, se realiza una nueva búsqueda usando el perfil generado en el paso anterior, sobre una base de datos de secuencias de estructuras conocidas (templados), también *clusterizada* por identidad de secuencia.

A lo largo de los años, se vió que esta técnica aumenta bastante la cantidad de estructuras templado que se encuentran, dando lugar a una mayor cantidad de posibilidades para modelar. Más aún, como será explicado a continuación, resuelve un problema importante que es el de corregir la inserción de gaps en los alineamientos generados usando BLAST tradicional. Es por esto que esta estrategia se convirtió, con el correr del tiempo, en una de las formas

más comunes de realizar búsquedas de templados para realizar un modelado estructural comparativo y constituye una de las formas más fáciles de obtener resultados confiables. Cabe recalcar que hay estrategias más complejas que permiten obtener resultados más precisos, aunque, a medida que aumentan la precisión, aumentan consecuentemente los recursos computacionales que se requieren para resolver el problema. Finalmente, tienendo en cuenta que nos interesa modelar muchas secuencias pertenecientes a genomas enteros, usar PSI-BLAST constituye una de las mejores opciones a la hora de balancear, la calidad de los resultados obtenidos y los tiempos necesarios para obtenerlos.

Alinear el *target* con el templado

Una vez obtenido el templado con el que se desea realizar el modelado estructural, lo siguiente es construir un alineamiento entre la secuencia *target* y la secuencia del templado. En este caso, lo más eficiente es usar el alineamiento obtenido en el paso anterior. Además, usando PSI-BLAST, también mitigamos un error muy común que comete BLAST cuando las dos secuencias que se desean alinear no son tan parecidas. Se sabe que en los casos en que dos proteínas se parecen entre 40 % y 70 %, suelen apreciarse diferencias entre ambas secuencias, producto de eventos de *deleción* o *inserción* que puedan haber ocurrido en una o en ambas proteínas. Estas diferencias dan lugar a la inserción de *gaps* en el alineamiento óptimo resultante del proceso de alineamiento subyacente a la búsqueda realizada usando BLAST. En estos casos, BLAST, suele cometer errores cuando construye el alineamiento entre la secuencia *target* y la secuencia del templado, porque encontrar el mejor lugar para insertar un gap en un alineamiento, depende de elementos que no necesariamente están relacionados con la similaridad entre los amino ácidos, sino con las particularidades de la familia a la que pertenecen ambas secuencias, por ejemplo, estas inserciones o deleciones no suelen encontrarse en zonas estructuradas como hojas beta o hélices alfa, sino mas bien en Loops. En estos escenarios, PSI-BLAST, produce mejores alineamientos dado que estas regiones suelen estar conservadas a nivel de familia y se identifican mejor en alineamientos mútliples.

Entonces, el objetivo es corresponder cada amino ácido en la secuencia *target* con cada uno de los amino ácidos en el templado, teniendo en cuenta que lo principal es la estructura secundaria. La idea es alinear de manera tal que coincida la región conservada del plegado, introduciendo los gaps de manera de que se "quiebren" lo menos posible las zonas más estructuradas, y lograr esto de forma independiente a las similaridades entre amino ácidos que suponen los algoritmos que usan matrices como BLOSUM. Nuevamente, una de las maneras de abordar este problema es alinear ambas secuencias en el contexto

de una familia de proteínas, o al menos un conjunto de proteínas similares a las dos proteínas que deseamos alinear. El perfil generado por PSI-BLAST, luego de realizar varias iteraciones sobre una base de datos con muchas secuencias, consituye una representación estadística de dicho conjunto de secuencias. Por esto, al reiniciar la búsqueda PSI-BLAST sobre la base de secuencias del PDB, implícitamente se está utilizando información de varias secuencias para buscar, y además, para alinear los resultados obtenidos con la secuencia *target*.

Podemos pensar que el paso anterior (la selección de templados) se puede realizar al mismo tiempo con éste paso del proceso (alinear *target* y templado). Primero se construye una PSSM para la secuencia *target* usando búsquedas iteritivas de PSI-BLAST y luego, se reinica la búsqueda sobre las secuencias del PDB usando la PSSM construida. Así, en la misma búsqueda de templados, se obtiene el alineamiento entre la secuencia *target* y el resultado obtenido. Cada alineamiento resultante, puede presentar una determinada cantidad de residuos alineados idénticos, algunos similares y otros disímiles, y además, gaps insertados en distintas regiones de cada secuencia, tanto en la *target* como en el templado. Como comentaremos más adelante, estos parámetros suelen tenerse en cuenta a la hora de analizar la calidad de los resultados obtenidos luego de terminado el proceso global de modelado comparativo.

Construir los modelos

Una vez obtenida la estructura del templado y el alineamiento entre el *target* y el templado, es posible enumerar modelos para la proteína *target*. Dado un conjunto de proteínas en una familia, es posible identificar regiones estructuralmente conservadas, que pueden ser usadas para construir modelos. Para los residuos en las regiones variables de la proteína, que suelen coincidir con inserciones y deleciones en el alineamiento subyacente, la estructura de los templados no se puede usar para la construcción del modelo. La predicción de estas regiones suele encararse usando otras estrategias, como por ejemplo, mediante métodos de simulación ab initio, o bien, identificando fragmentos similares mediante búsquedas en librerías de loops conocidos. En ambos casos se obtienen buenos resultados prediciendo loops largos complementando el modelado comparativo de las regiones conservadas [Fiser et al., 2000] [Xiang et al., 2002] [Deane and Blundell, 2001]. Por último, se suele aplicar un paso de minimización energética a traves de técnicas que usan potenciales clásicos atomísticos y que tienen por objetivo relajar las tensiones estructurales y regularizar la estereoquímica del modelo [Flohil et al., 2002].

Las dos técnicas más usadas en modelado de estructura basada en homología son: i) **modelado por ensamblado de cuerpo rígido** (implemen-

tado, por ejemplo, por SWISS-MODEL [Guex and Peitsch, 1997]), que consiste en construir el modelo basado en algunas regiones que forman el núcleo de las proteínas, ensamblandolas con loops y confórmeros de cadenas laterales, que se obtienen a partir de fragmentar estructuras relacionadas, y ii) **modelado por satisfacción de restricciones espaciales**, (implementado por MODELLER [Šali and Blundell, 1993]), que usa, técnicas de optimización geométrica para satisfacer restricciones espaciales derivadas del alineamiento entre la secuencia *target* y las secuencias subyacentes a las estructuras templados. Estas técnicas de modelado funcionan muy bien para casos en los que la identidad de secuencia entre *target* y templado es elevada. Sin embargo, para casos en que la similaridad entre las dos secuencias es muy baja ($< 30\,\%$) se ha visto que es necesario alimentar el modelado usando información adicional de otras fuentes, como ser predicciones estructurales (Accesibilidad al solvente y Estructura secundaria) a partir de la secuencia que se desea modelar y alimentar con estos datos al proceso de evaluación del modelo.

Específicamente, MODELLER (Haremos énfasis en esta herramienta dado que es la que usamos en este trabajo), extrae las restricciones espaciales de dos fuentes. Primero, se derivan restricciones sobre las distancias y los ángulos dihedros de la secuencia *target* basadas en el alineamiento de dicha secuencia con los templados. Segundo, se obtienen las restricciones esteroquímicas, (como ser las distancias y los ángulos de los enlaces) de campos de fuerza de dinámica molecular CHARMM-22 [MacKerell et al., 1998] y, las preferencias estadísticas de ángulos dihedros y de distancias de interacción de no-unión, se derivan de un conjunto representativo de estructuras de proteínas conocidas. Luego, el modelo se calcula usando un método de optimización por gradiente conjugado y dinámica molecular, para minimizar los fallos en las restricciones espaciales. El procedimineto es conceptualmente similar al que se usa para determinar la estructura de proteínas usando restricciones derivadas de experimentos de RMN.

Como se aprecia en la Figura 3.4, MODELLER simula la estructura de la proteína agregando restricciones de manera incremental mientras busca el mínimo energético. En la línea inferior de la Figura 3.4 se puede ver una variable denominada Δr que representa, para cada residuo, la cantidad de vecinos tenidos en cuenta a la hora de agregar las restricciones. De esta manera, sólo las restricciones deducidas de los vecinos más cercanos son tenidas en cuenta cuando la simulación comienza, por ejemplo, los ángulos dihedros entre los residuos. A medida que la simulación avanza, cada vez que el sistema llega a un punto en el que no se producen cambios estructurales significativos (i.e.: en la Figura, el *average shift* por debajo de un determinado umbral), se agregan

nuevas restricciones y se continúa simulando a partir de este nuevo sistema. Al agregarse nuevas restricciones, la energía del sistema aumenta, y así, se itera nuevamente en la búsqueda del nuevo mínimo local. Se vuelven a agregar más restricciones cada vez que el sistema llega a un mínimo local y el proceso continúa hasta que se agregan todas las restricciones deducidas de la estructura global de la proteína. Al final del proceso se obtiene una estructura candidata, luego, este proceso se repite varias veces y se enumeran varias configuraciones candidatas del sistema muestreando el espacio conformacional. Finalmente, con todas estas conformaciones se pasa a la evaluación de calidad que será descripta a continuación.

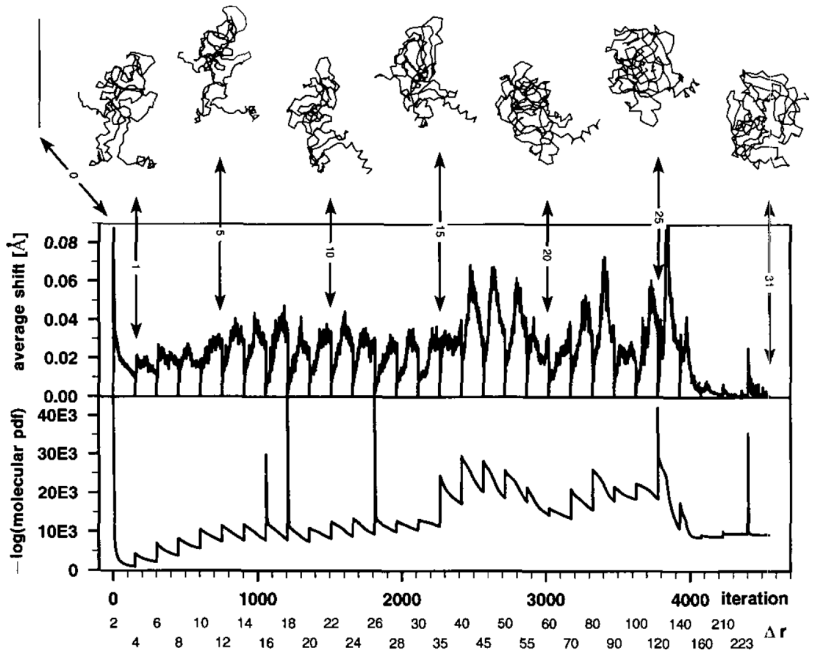

Figura 3.4: **Proceso de predicción de estructura usando MODELLER.** Inicialmente, se parte de un polipéptido desplegado y se van agregando restraints a medida converge el método del gradiente conjugado. Los restriants se agregan sucesivamente por cada residuo. La variable Δr define la cantidad de residuos vecinos consecutivos tenidos en cuenta a la hora de derivar los restraints. El método de gradiente conjugado converge cuando el desplazamiento promedio (*average shift*) alcanza un valor lo suficientemente bajo.

Evaluar los modelos

En el modelado comparativo, la calidad de los modelos está directamente relacionada con la similaridad en secuencia, de la secuencia *target* con los templados disponibles. Se sabe que a medida que la identidad de secuencia entre dos proteínas dismuniye, se puede apreciar un incremento en la divergencia estructural entre ambas (Figura 3.2) [Chothia and Lesk, 1986], [Rost, 1999], por lo que las estructuras que surjan luego del modelado, pueden presentar imperfecciones si fueron hechas basandose en templados muy distantes al *target* en términos de similaridad de secuencia. Las causas más comunes de estos errores pueden ser: cadenas laterales mal posicionadas, conformaciones incorrects de loops, distorsiones estereoquímicas en el backbone de la estructura proteica, errores en el alineamiento como los que fueron mencionados previamente, y por supuesto, la elección de un templado incorrecto [Baker and Sali, 2001].

La evaluación del modelo consiste, por un lado, en elegir entre todos los modelos del paso anterior, un modelo que represente lo mejor posible la estructura de la secuencia usada como *target*, y por otro lado, obtener un valor general de la calidad del modelo y otro valor local que nos indica qué regiones han sido modeladas con mayor o menor calidad. Para evaluar los modelos se usan potenciales que puntúan cada estructura teniendo en cuenta criterios energéticos. A veces vale la pena construir muchos modelos para diferentes alineamientos y elegir el mejor modelo basandose en la evaluación de calidad tridimensional más que basandose en parámetros del alineamiento. La evaluación de la estructura tridimensional estima la calidad de las coordenadas predichas en términos de su conformación e interacciones internas.

La idea es poder identificar modelos basados en los templados equivocados o las regiones del modelo que fueron construidas con alineamientos incorrectos. Además de funciones de puntuación basadas en principios energéticos, lo más común es que se usen estrategias estadísticas para detectar errores comparando ciertos aspectos de los modelos con sus distribuciones esperadas, que surgen de analizar estructuras conocidas resueltas usando complejos de alta resolución obtenidos mediante la técnica de rayos X [Hooft et al., 1996], [Eisenberg et al., 1997], [Melo and Feytmans, 1998].

Más puntualmente, los potenciales que se usan para evaluar la calidad de un modelo, se suelen dividir en dos tipos:

- **Internos:** Evalúan la autoconsistencia interna durante el proceso de construcción del modelo, a través de valuaciones energéticas (ab initio o estadísticas) para verificar si el modelo satisface o no los restraints impuestos.

- **Externos:** La evaluación radica en el uso de información que no fue usada en el cálculo del modelo, sino que se usan modelos estadísticos de propiedades estructurales como ser estructuras secundarias o accesibilidades al solvente, además de los términos clásicos de distancias y torsiones que analizan un conjunto de modelos y los ordenan según la *calidad* del ensamble [Manfred, 1993] [Lüthy R, 1992].

En ambos casos, los potenciales se calculan como sumatorias de términos, dónde cada uno de ellos puntúa un determinado aspecto de la calidad del modelo. Estos términos se construyen estadísticamente, como se mencionó previamente, a partir de estructuras conocidas resueltas con muy buena calidad, de manera que cada término puntúe la calidad por separado y luego se suman para obtener un puntaje global. Suelen tener en cuentas 3 aspectos de la estructura: la distancia entre los carbonos beta de los residuos (denominado `pairwise`), las torsiones de los residuos (`torsion`) y sus accesibilidades al solvente (`solvation`). Estos 3 aspectos deducidos a partir de las estructuras, fueron modelados de diferentes maneras a los largo de los útlimos 20 años dando lugar a diferentes potenciales estadísticos que evalúan la calidad de un modelo sín tener más datos que su estructura 3D. Además de estos potenciales, en la útlima década se pusieron a punto potenciales más complejos que usan información de alineamientos múltiples deducidos de hacer búsquedas en PSI-BLAST o predicciones de estructura secundaria y accesibilidad al solvente que, en general, usan información de alineamientos múltiples con muchas secuencias de proteínas relacionadas.

Por último, a los fines de este trabajo de tesis nos interesó probar con tres potenciales descriptos a continuación, que se usan con propositos distintos:

- **GA341:** Es un criterio que sirve para filtrar modelos que tienen una calidad inaceptable. GA341 es una función que depende de la identidad de secuencia del alineamiento obtenido y del Z-score que presenta el modelo con respecto a la distribución de un potencial estadístico basado en el grado de compactación de estructuras conocidas. Como se puede ver en la Figura 3.5, la forma de la función está determinada principalmente por un balance entre el porcentaje de identidad de secuencia y el potencial estadístico de compactación. La identidad de secuencia captura el hecho de que secuencias similares, generalmente tienen estructuras similares, y al combinar el potencial estadístico de compactación se modela el hecho de que a pesar de que esta identidad de secuencia entre dos proteínas puede ser baja, aún así, pueden tener estructuras parecidas [Melo and Sali, 2007].

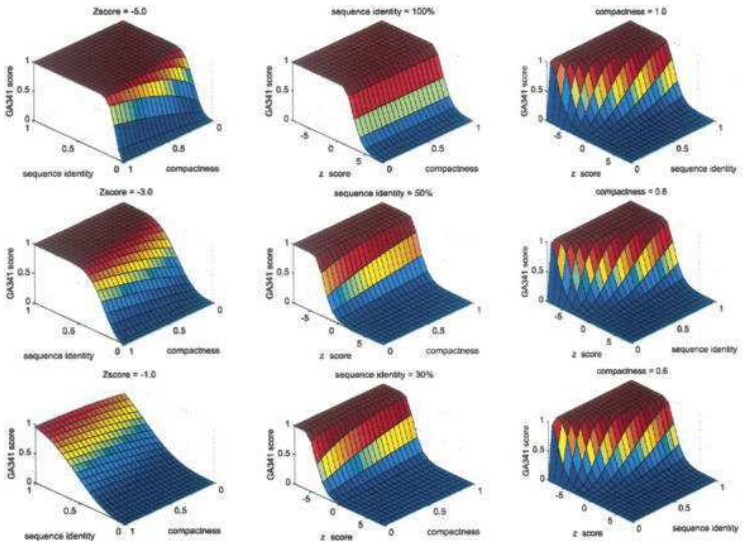

Figura 3.5: **Función discriminante GA341.** Sirve para filtrar modelos de mala calidad. La función está entrenada para tomar valores entre 0 y 1, cuanto más alto el valor, mejor se considera el modelo dependiendo de tres parámetros: su grado de compactación, el estadístico Z de un potencial energético y su identidad de secuencia. Tomamos que para valores de GA341 mayores a 0.7, los modelos se consideran aceptables. *Figura editada del trabajo de [Melo and Sali, 2007]*

- **QMEAN:** Este potencial sirve para evaluar los modelos utilizando subpotenciales que modelan los aspectos mencionados previamente. Usa 4 subpotenciales bien conocidos (torsion, pairwise, solvation, all_atom) que evalúan la calidad, cada término por separado, y se suman para obtener un valor total de energía. Además, agrega otros dos términos que tienen en cuenta la predicción de estructura secundaria y accesibilidad al solvente que mencionamos previamente. El potencial está probado sobre un set de estructuras cristalográficas y el programa calcula el Z-score (ZQMEAN) del potencial con respecto al evaluado en dichas estructuras como una medida de lo que se espera que sea una estructura resuelta usando rayos X. Esta distribución se puede ver en la Figura 3.6A donde se comparan los valores del potencial con las categorías de calidad para los modelos obtenidos en la competencia CASP8. Así, para cada modelo, se calcula el Zscore que se usa como parámetro de calidad como se ve en la Figura 3.6B, donde se aprecia que el valor de QMEAN está normalizado con respecto al largo de la secuencia proteica, y se muestra para un modelo en particular (cruz roja) los valores obtenidos para el puntaje global y para

cada subpotencial [Benkert et al., 2008], [Benkert et al., 2010].

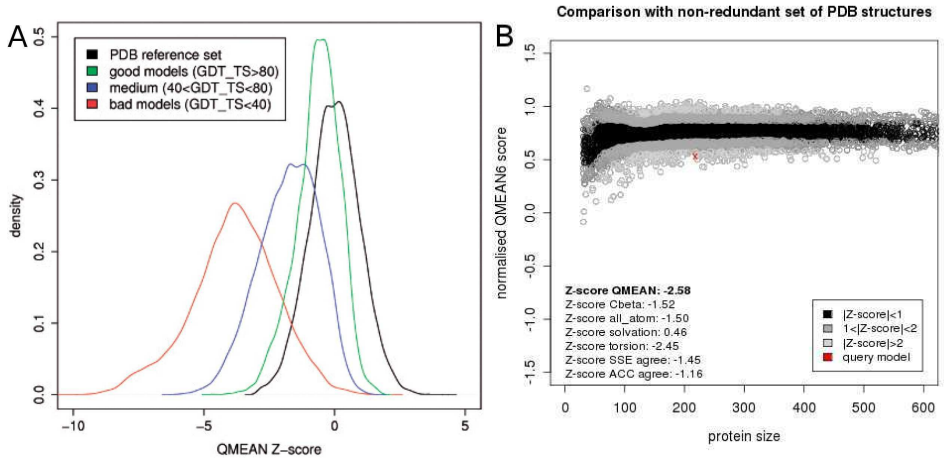

Figura 3.6: **Distribución de QMEAN para estructuras conocidas en el PDB.** A) Asignación de calidad de modelos según QMEAN basandose en el criterio de *Global Distant Test - Total Score* que se usó para evaluar el grado de error de los modelos en CASP8. *Figura tomada del trabajo de [Benkert et al., 2010]* B) Evaluación de la calidad en función de los valores del potencial para estructuras resueltas por Rayos X. El punto rojo marca dónde cae un modelo dado.

- **DOPE:** Este potencial estadístico es interno de MODELLER y se calcula por defecto cada vez que se construye un modelo. Provee un valor global de calidad de toda la estructura basado en la función energética que MODELLER usa para asignar las restricciones espaciales, y además, da un valor residuo por residuo, que evalúa la calidad local del modelado.

3.1.2. Proteínas en 3D: El formato de archivo PDB

Como mencionamos previamente, el PDB hizo su aparición a principios de los 70 como una de las primeras bases de datos computacionales que almacenan, mantienen y proveen datos biológicos. El éxito de esta base de datos radica en que se provee la información usando un formato de archivo unificado que permite almacenar información tridimensional de moléculas biológicas, independientemente del proceso que se usó para determinar dichas estructuras. De esta manera, ya sea que las estructuras fueran resueltas por cristalografías de rayos X, experimentos de NMR, microscopía electrónica, etc, la base de datos unifica el criterio con el que se almacenan los datos estructurales. Como

se puede ver en la Figura 3.7, en los útlimos 20 años, la cantidad de entradas en ésta base de datos fue aumentando hasta que, hoy día, ya se depositaron más de 100000 estructuras de moléculas biológicas.

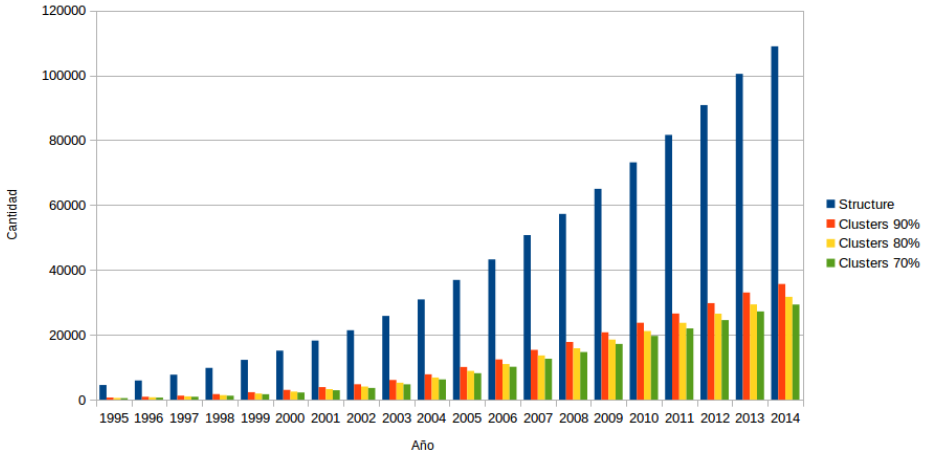

Figura 3.7: **Entradas depositadas en el PDB a lo largo de los últimos 20 años.** En azul, todas las entradas del PDB correspondientes a proteínas. En naranja, amarillo y verde la cantidad de clusters que se deducen según CD-HIT a 90 %, 80 % y 70 % de identidad de secuencia, respectivamente.

El formato de archivo PDB, no sólo permitió aumentar la cantidad de información estructural de manera ordenada, sino que además, fue posible ponerse de acuerdo acerca de la interpretación de un *ensamble biológico* para facilitar el análisis y/o el modelado de dichas estructuras. Esta interpretación consiste en la jerarquización de las moléculas y los átomos que la componen en una estructura de datos que tiene los siguientes niveles: **Estructura**, **Cadena**, **Residuo**, **Átomo**. A continuación, comentaremos acerca de estos niveles y mencionaremos ciertas propiedades que es posible calcular/asociar para cada uno de ellos y serán usadas como definiciones de esta tesis de acá en adelante:

- **Estructura:** Se llama estructura al ensamble entero que se haya en un archivo PDB. Este nivel comprende estructuras de proteínas, de ADN y de pequeños ligandos. Así mismo, también es posible encontrar complejos de interacción, como ser: complejos proteína-proteína, proteína-ADN y proteína-ligando.
    - **Resolución:** Solo en caso de ser una cristalografía de rayos X, el experimento es realizado con una determinada resolución medida en

Angstroms. Este dato es muy útil para filtrar los datos según su calidad.

- **Estructura cuaternaria:** Información acerca del estado de oligomerización de las cadenas polipeptídicas involucradas.
- **Tipo de estrucutra:** Este campo mantiene información acerca del experimento del cual se hizo el PDB (X-Ray, NMR) y del tipo de moléculas presentes en el PDB (prot: Sólo proteína, nuc: Sólo ADN y prot-nuc: proteína y ADN)

- **Cadena:** Las cadenas son la unidad que define a una determinada molecula protéica o de ADN, la cual a su vez puede contener pequeños ligandos en interacción con la misma. Este nivel de abstracción ayuda a separar conceptualmente las unidades biológicas de manera tal de poder distinguir diferentes subunidades en el ensamble biológico presente en cada PDB. Por ejemplo, en los casos donde hay estructuras cuaternarias, dímeros, trímeros, tetrámeros, etc.
  - **Identificador de proteína:** Información acerca de la/s proteína/s presente/s en el PDB. Se almacena una referencia al código de Uniprot correspondiente a la estructura protéica y los datos acerca de la cobertura sobre dicha secuencia, dado que es común que se resuelva sólo un dominio o fragmento de la proteína en cuestión.
  - **Dominios y Familias:** Información acerca de las familias y dominios funcionales involucrados contenidos en la cadena presente en el PDB, con referencia a los límites que denotan las regiones dentro de la misma. Es decir, desde y hasta qué amino ácido dentro de la cadena polipeptídica se encuentran los dominios.
- **Residuo:** El nivel de residuo es utilizado para varios fines distintos. Principalmente para agrupar los átomos pertenecientes al mismo amino ácido, pero tambien se encuentran agrupados en residuos los átomos pertenecientes a las bases nucleotídicas en las cadenas de ADN. Además, aquellos átomos de ligando pequeños unidos a traves de una red de enlaces covalentes, tambien son agrupados bajo el nivel de abstracción de Residuo.
  - **Superficie accesible al solvente:** Valores en Angstroms cuadrados denotando el grado de accesibilidad al solvente de cada residuo en cada estructura PDB. Disponible para cada cadena lateral de cada amino ácido y para cada pequeño ligando. Esta propiedad será ampliada a continuación.

    - **Estructura Secundaria:** Para las cadenas protéicas, se almacenan los valores de ángulos diheros (PHI y PSI) así como una asignación de estructura secundaria de tres niveles (HELIX, SHEET y LOOP) que obtenemos a partir de la definición de estructura secundaria propuesta por DSSP. Esta propiedad, también, será ampliada a continuación.

- **Átomo:** Este nivel de abstracción corresponde al tipo de átomo que compone cada residuo. Con esta entidad es posible caracterizar el tipo de átomo que contiene cada molécula, en términos de su rol en las interacciones intermoleculares. Por ejemplo, donores y aceptores de puente hidrógeno, carbonos aromáticos y alifáticos, etc.

    - **Tipo de átomo:** Cada átomo tiene su tipo en función de su rol como interactor entre moléculas. Estos tipos son Donores y Aceptors de puentes de hidrógeno, átomos con cargas de magnitud relevante (por ejemplo en los ácidos aspárticos y glutámicos o lisinas y argininas) y carbonos aromáticos o alifáticos. De esta manera es fácil, por ejemplo, identificar anillos aromáticos tanto en cadenas laterales de amino ácidos como en compuestos tipo droga o cofactores, lo cuál será de especial valor para el próximo capítulo.

Con esto damos por concluida la introducción que soporta las metodologías que implementamos para realizar predicciones estructurales, las cuales describiremos a continuación. Comentaremos de predicciones de propiedades estructurales como la estructura secundaria y la accesibilidad al solvente, a partir de la secuencia proteica, y cómo se calculan y se asignan dichos parámetros si es que se conoce la estructura. Luego, en la sección de resultados mostraremos cómo estudiar el grado de error en los modelos a partir de parámetros que se conocen de los modelos obtenidos luego de terminado el modelado como ser: los valores que asignan de los potenciales energéticos a cada modelo y el alineamiento que se usó para el modelado. Finalmente, usando dicho criterio estudiaremos con qué grado de precisión se están cubriendo los genomas de 3 organismos bacterianos.

3.2. Materiales y Métodos

3.2.1. Cálculo de propiedades estructurales

A continuación mencionaremos dos criterios que dan lugar a algoritmos de calculo/asignación de dos propiedades estructurales muy importantes que son: la superficie accesible al solvente y la estructura secundaria. Lo relevante de

estas dos propiedades es que se relacionan a nivel de residuo en la estructura de datos mencionada previamente, y además constituyen dos parámentros muy usados a la hora de evaluar la calidad de un modelo estructural.

Área de la superficie accesible al solvente

La superficie accesible al solvente (SAS o SASA para el Área) corresponde a la porción de cada amino ácido que no está oculta en la proteína contactando otros amino ácidos. Desde el punto de vista bioinformático es posible obtener el valor de SASA para cada amino ácido, utilizando un algoritmo que dispone esferas cercanas a los átomos del amino ácido en cuestión y, teniendo en cuenta los radios de Van der Walls, se filtran las esferas que producen choques (o *clashes*) y con las esferas que quedan se estima el área que está accesible a moléculas del solvente (se toma como solvente base el agua). Claramente, el cálculo es dependiente del tamaño de la esfera, por lo que se usa un valor de 1.4Å, que da los resultados más coincidentes con resultados experimentales. Para el cálculo usamos el algoritmo presente en el software VMD.

A veces es interesante calcular el área de la superficie de contacto que los amino ácidos establecen con pequeños ligandos o incluso con otras proteínas o péptidos. Para esto calculamos la accesibilidad de cada amino ácido en varias conidiciones, midiendo el área y quitando otros elementos interactores presentes en la estructura. Dado un cristal, es posible que la estructura de cada cadena polipeptídica venga con moléculas de agua, moléculas de ligandos o incluso que haya un estado de oligomerización (estructura cuaternaria) entre las distintas cadenas. Para saber las diferencias entre el área expuesta con y sín ligandos o con y sín interacciones proteína-proteína, calculamos el área accesible al solvente quitando moléculas según su tipo. Así, obtenemos varios valores de SASA para cada amino ácido dependiendo de si contamos las aguas, los ligandos, las otras cadenas polipeptídicas, etc. Se cuantificaron 4 tipos de superficies, enumeradas a continuación:

- **SASA sin aguas y/o otros solventes** ($SASA_{NOWAT}$) El área, sín las moléculas de agua que hayan sido cristalizadas con la/s proteína/s. Sólo para casos de PDB provenientes de experimentos de Rayos X.
- **SASA sin compuestos químicos heteroátomos** ($SASA_{NOHET}$) El área, sín las moléculas correspondientes a heteroátomos que no sean amino ácidos modificados. Se suele usar el mote de heteroátomo (`HETATM`) para los casos en los que los residuos a los que pertenece dicho átomo son de moléculas orgánicas pequeñas que hayan sido usadas en el proceso de

determinación de estructura. Estas moléculas pueden ser, tanto ligandos naturales de la proteínas en cuestión, como drogas o moléculas que hayan sido involucradas en el proceso de cristalización a la hora de resolver la estructura de forma experimental.

- **SASA sín interacciones proteínas-proteína** ($SASA_{NOPPI}$) El área, sín contar las interacciones proteína-proteína. O sea, sin tener en cuenta la estructura cuaternaria de la proteína, contando sólo el área como si cada subunidad estuviera aislada de las otras.
- **SASA total de amino ácido** ($SASA_{TOTAL}$): El área total del amino ácido sin el resto de la proteína a la que pertenece como si estuviera en vacío.

Estos valores nos permiten calcular el grado de exposición de los amino ácidos según sus interacciones con otras moléculas. Por ejemplo, para calcular cuánta área tiene en contacto un determinado amino ácido con los ligandos de la proteína, se restan $SASA_{NOHET} - SASA_{NOWAT}$. Esta resta siempre es positiva y mide el área compartida entre el amino ácido y el/los ligandos. A su vez, si nos interesa el área de contacto aportadada por un amino ácido en las interacciones proteína-proteína, hacemos la cuenta $SASA_{NOPPI} - SASA_{NOHET}$ que tambíen será siempre positiva por definición.

Asignación de Estructura Secundaria

Para asignar estructura secundaria, usamos el software DSSP que asigna elementos de estructura secundaria a cada uno de los amino ácidos, basandose en criterios de ángulos dihedros y puentes de hidrógeno entre los backbones de los amino ácidos involucrados. Así, con esta herramienta obtenemos los siguientes valores: Ángulo dihedro PHI, Ángulo dihedro PSI y el **Código de estructura secundaria**, para el cual usamos una codificación de tres niveles (HELIX: H, SHEET: E, LOOP: C), partiendo de la codificación que provee DSSP.

3.2.2. Predicción de propiedades estructurales

Cuando no es posible obtener estructuras de proteínas similares en las bases de datos, aún es posible predecir ciertas propiedades a nivel de residuo utilizando como entrada sólo la secuencia proteica. A continuación comentaremos acerca de 2 predicciones (entre otras) que estuvieron en auge durante la década del 2000: Estructura Secundaria y Accesibilidad al Solvente. A medida que

CASP fue avanzando, se pudo ver que la información filogenética aumentaba el poder predictivo de estos métodos. Por eso, todas estas técnicas, tienen en común que construyen un perfil estadístico a partir de la secuencia *target* con el cuál alimentan distintas estrategias de aprendizaje automático (*Redes neuronales*, *Máquinas de vectores de soporte*, etc). Usamos la predicción de estructura secundaria y de accesibilidad al solvente para alimentar la evaluación de calidad que realiza QMEAN, dado que mejora bastante el poder predictivo.

Predicción de accesibilidad al solvente

La predicción de accesibilidad al solvente se puso a punto usando el software SSPRO [Magnan and Baldi, 2014]. Este paquete de software usa redes neuronales para predecir a partir de un perfil de secuencias obtenido usando PSI-BLAST. Dado que en el pipeline de modelado comparativo calculamos un perfil, lo usamos como entrada de este software para después usarlos en la estimación de calidad.

Predicción de estructura secundaria

Del mismo modo que en el caso anterior, la predicción de estructura secundaria se realiza tomando como entrada el perfil de la secuencia que se desea predecir. Cabe mencionar aquellos predictores de estructura secundaria que usamos en este trabajo: PSIPred [Jones, 1999], JPred [Cuff et al., 1998], Porter [Pollastri and McLysaght, 2005]. A su vez, estos predictores fueron utilizados para obtener un consenso acerca de la estructura secundaria de la proteína, lo cuál fue demostrado exitoso en la competencia CASP [Albrecht et al., 2003]. Una distinción interesante para tener en cuenta en la predicción de estructura secundaria es el hecho de tener en cuenta el tipo de proteína que se desea estudiar. En el caso de las proteínas con pasos transmembrana, por ejemplo, la predicción de segmentos (hélices alfa fundamentalmente) dió lugar a varios predictores de estructura secundaria especializados en predecir este tipo particular de estructura.

3.2.3. Pipeline de Modelado Estructural

Pusimos a punto un pipeline de modelado estructural usando la técnica de modelado comparativo, porque la idea era procesar grandes datos en tiempos accesibles. Dividimos el proceso en subprocesos básicos inspirados en los pasos mencionados previamente como indica la Figura 3.8.

Los pasos de **búsquedas de templados** y **alineamiento entre el *target* y el templado**, se resumen en una búsqueda usando PSI-BLAST. La idea es usar el mismo hit de PSI-BLAST para extraer el alineamiento que se usará en el modelado. Pero, para poder buscar en la base de datos del PDB (las secuencias FASTA), primero es necesario construir una PSSM que permita obtener hits con menos similaridad de secuencia que usando sólo una búsqueda BLAST. Para esto, se usa una búsqueda PSI-BLAST para obtener una gran cantidad de secuencias de una base de datos. Es importante que esta base de datos esté agrupada por identidad de secuencia para no crear perfiles estadísticos sesgados en caso de que aparezcan varias proteínas muy parecidas. En nuestro caso implementamos la creación de la PSSM de la secuencia usando como base UniRef50, con 3 iteraciones de PSI-BLAST y un umbral de recaudación para el e-value de 0.00001. La base de datos UniRef50 es el resultado de un proceso de clustering entre las secuencias de la base de datos UniProt que, en particular, fue realizado con un punto de corte de manera que las secuencias de proteínas no tengan más del 50 % de identidad entre ellas. Dicha corrida de PSI-BLAST da como resultado una PSSM o *checkpoint* que contiene información acerca de las proporciones de cada amino ácido sobre cada posición (PSSM) de la secuencia *target*. Segundo, la búsqueda PSI-BLAST se reinicia usando el checkpoint mencionado previamente, pero esta vez, contra una base de datos de secuencias de estructuras conocidas (templados). Esta base de datos consiste en todas las secuencias de cada cadena proteica presente en el PDB, agrupadas usando CD-HIT [Li and Godzik, 2006] con un umbral del 95 % identidad entre cada una de ellas. A partir de esta última búsqueda, se obtienen los alineamientos locales calculados por el algoritmo de PSI-BLAST, quedandonos con aquellos que tengan un e-value por debajo de 0.00001 y que cubran por lo menos un 50 % de la secuencia *target*. Con respecto a la cobertura, cabe aclarar que algo que se podría hacer antes de realizar la búsqueda de templados, es fragmentar la secuencia en sus dominios usando una base de datos de dominios como Pfam, así el corte sobre la cobertura podría ser más alto para garantizar que el modelado se haga por unidad funcional. En nuestro caso, el punto de corte a la cobertura es sobre toda la secuencia proteica. Con estos alineamientos y la estructura del templado correspondiente se pasa a usar MODELLER [Šali and Blundell, 1993] para enumerar modelos estructurales a partir de la metodología mencionada previamente. Por cada templado se construyen 5 modelos diferentes y se evalúa su calidad usando los potenciales GA341 [Melo and Sali, 2007] y QMEAN [Benkert et al., 2008]. Para filtrar aquellos modelos inaceptables, nos quedamos sólo con aquellos modelos que presenten un puntaje de GA341 mayor a 0.7. Finalmente, como modelo rep-

resentativo de cada ORF, tomamos aquel que tiene el valor de QMEAN más alto.

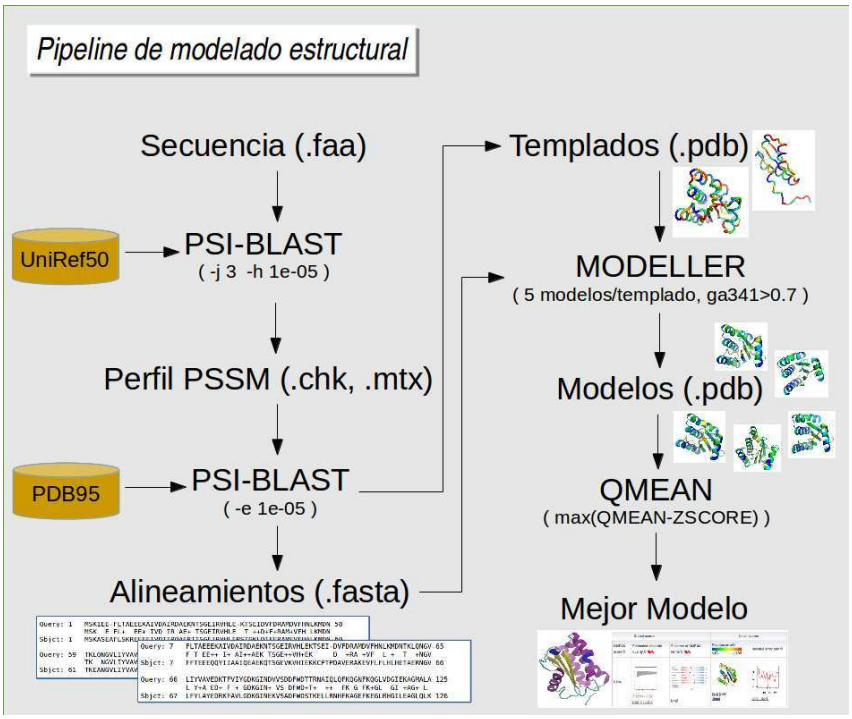

Figura 3.8: **Esquema del pipeline de modelado estructural.** El pipeline consiste en obtener un archivo PDB con un modelo representativo para una secuencia dada, que se provee como parámetro inicial. Primero, se construye un perfil estadístico por posición usando PSI-BLAST (en este caso el perfil se denomina PSSM) contra una base de datos grande y *clusterizada* por identidad de secuencia. Segundo, se reinicia la búsqueda sobre una base de datos de templados estructurales (en este caso, usamos las secuencias subyacentes a las estructuras en el PDB, y las clusterizamos por identidad de secuencia de manera de eliminar redundancia). Tercero, con las estructuras de los hits de la última búsqueda y los alineamientos de PSI-BLAST, se usa MODELLER para enumerar modelos mediante la técnica de modelado por satisfacción de restricciones espaciales. En este paso, nos quedamos sólo, con aquellos modelos que tengan un valor de GA341 menor a 0.7. Por último, se evalúa la calidad de los modelos que quedaron usando el potencial QMEAN que asigna valores de calidad según qué tan parecidos son los modelos a estructuras resueltas usando la técnica de Rayos X.

3.3. Resultados

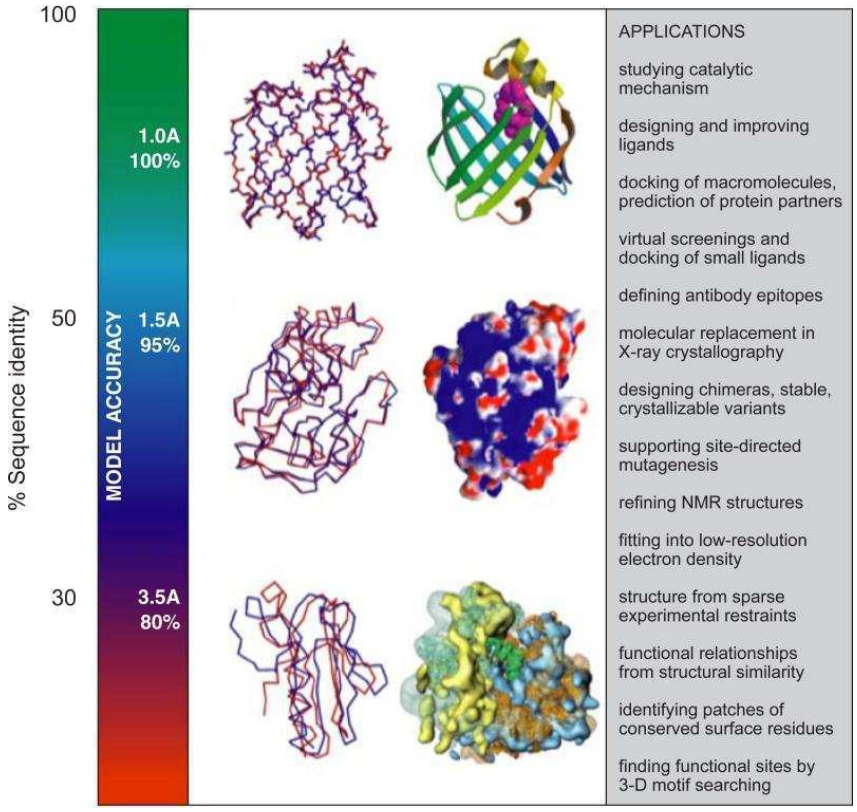

Figura 3.9: **Usos posibles para los modelos según su calidad estimada.** A medida que el grado de error disminuye es posible usar los modelos para análisis cada vez más detallados, desde identificar plegados hasta diseñar drogas. *Figura tomada del trabajo de [Eswar et al., 2006]*

Una pregunta que se puede hacer es ¿Con qué grado de error se están obteniendo las estructuras modeladas? De manera de poder tener una noción de que tipo de técnicas pueden aplicarse sobre los modelos. Una vez terminado el proceso de modelado, hay parámetros que pueden ser usados para estimar la calidad del modelado en general. Estos son: las **Características del Alineamiento** y los valores de los **Potenciales de Calidad**. Por un lado, los alineamientos resultantes pueden presentar identidades de secuencia muy bajas o mismo, haber resultado en la inserción de muchos gaps que inducen a pensar que la relación de homología entre ambas secuencias (*target* y templado) está

en los límites de lo aceptable. Esto sumado al hecho de que las aplicaciones que se le puedan dar a cada modelo dependen del grado de presición con el que se haya logrado predecir la estructura 3D, como indica la Figura 3.9. Según el grado de error vaya disminuyendo (usualmente medido en término de RMSD), es posible usar los modelos para análisis cada vez más detallados.

3.3.1. Análisis del pipeline de modelado estructural

Para analizar el rendimiento del pipeline y obtener una estimación del grado de error con el que se obtienen los modelos, construimos un conjunto de datos de estructuras conocidas basandonos en el PDB, con el objetivo de realizar predicciones a partir de las secuencias subyacentes y comparar las diferencias que hay entre el modelo y la estructura resuelta. Fue seleccionado un conjunto de estructuras de prueba de entre las estructuras depositadas en el PDB, seleccionando sólo aquellas estructuras monoméricas que hayan sido resueltas usando la técnica de rayos X y que todos los residuos hayan sido resueltos. A su vez, aplicamos un filtro de similitud *clusterizando* las secuencias CD-HIT a 70 % de identidad y quedandonos con la proteína centroide. Una vez filtradas estas estructuras, modelamos usando el pipeline, teniendo en cuenta que no se usen templados de más de 95 % de manera de evitar los modelos realizados con proteínas practicamente iguales. Para comparar las diferencias entre los modelos y las estructuras reales, calculamos las siguientes medidas que nos dicen qué tan distintas son ambas estructras:

RMSD: Es la desviación cuadrática media de las distancias entre los carbonos α luego de obtener la superposición óptima de ambas estructuras. Aunque es bien sabido que el RMSD aumenta comforme aumente el tamaño de la proteína, este valor permite darse una idea del grado de error que se produce en el modelo, en términos de Angstroms. Lo cuál será de valor a la hora de estimar el error esperado en el modelado de las proteínas en genomas secuenciados de bacterias (calculado usando el software de TMscore).

TMscore: [Zhang and Skolnick, 2004] Este puntaje pretende mejorar el cálculo del RMSD de manera que sea independiente del tamaño de la proteína. A su vez, los autores vieron que usando como base los criterios de plegado definidos en SCOP y CATH, para la mayoría de los casos, un TMscore de mas de 0.5 indica algún grado de similaridad estructural y valores por encima de 0.7 indican que las estructuras que se están comparando presentan practicamente el mismo plegado [Xu and Zhang, 2010]. En este caso,

lo usamos para estimar el grado de error en términos de haber obtenido el plegado correcto. Elejimos este método dado que se calcula rápidamente, y además, es uno de los últimos puntajes que adoptó CASP en los últimos años para evaluar la similaridad estructural entre los modelos y la estructura resuelta.

Utilizando estas dos medidas, analizaremos a nivel global cómo influyen en el modelado, los parámetros del alineamiento y las evaluaciones de los potenciales de calidad. Estos parámetros son:

- **Identidad de secuencia** calculada como la cantidad de residuos alineados iguales sobre el largo del alineamiento.
- **Largo del alineamiento** para el alineamiento local producto de la búsqueda usando PSI-BLAST.
- **Cobertura** del alineamiento sobre la secuencia *target*.
- **Gaps**, tanto en el *target* como en el templado.
- **QMEAN** para evaluar la calidad como potencial externo a MODELLER.
- **DOPE** para evaluar la calidad como potencial interno a MODELLER.

Las figuras muestran, para cada parámetro, la distribución de resultados obtenidos comparando las estructuras reales con los modelos. Tanto el RMSD, como el TMscore están sujetos a las propiedades del alineamiento, que a medida que se detecta una mayor divergencia en términos evolutivos (las secuencias alineadas tienen baja identidad y/o fueron insertados muchos gaps), mayor es el grado de error en la predicción de la estructura.

Con respecto a la identidad de secuencia, podemos apreciar en la Figura 3.10, que la mayoría de las estructuras modeladas obtuvieron hits entre 20 % y 60 %. Aún así logramos un rango casi completo de las identidades de secuencia. Se puede ver que no se producen hits por debajo de 10 % de identidad, aunque hay varios hits que llegan hasta 20 %. Queda claro que cuanto menor es la identidad, más dificil es que el modelado consiga un RMSD lo sufienciente bajo como para que el modelo se pueda usar en análisis estructurales. Más aún, por debajo de 20 % de identidad es imposible dar con modelos de menos de 2.0Å de RMSD, lo cuál es el rango en el que se encuentran las interacciones de puente hidrógeno [Legon and Millen, 1987]. En términos del plegado que se consigue, vemos que se obtienen pocos valores por debajo de 0.7 y que estos valores se dan principalmente cuando la identidad cae por debajo del 40 %. A pesar de esto, la mayoría de los modelos que fueron producidos con alineamientos

que tienen menos de 40 % presentan valores de TMscore por encima de 0.8. El hecho de que se obtengan identidades de secuencia en valores bajos, es evidencia de que las búsquedas usando PSI-BLAST permiten llegar a valores de identidad en zonas de *homología remota*. Ya sea porque las iteraciones son suficientes para que converja el perfil, o porque la relación secuencia/plegado que logra cubrir el PDB clusterizado al 95 % es suficiente para cubrir estas zonas en las búsquedas por similitud, lo cuál ha sido evidenciado desde hace unos años [Kihara and Skolnick, 2003]. Finalmente, cabe recalcar que en todo el rango de identidades de secuencia (tanto para valores bajos como altos) se aprecian casos que no fueron predichos bien. Más puntualmente, se puede ver que obtenemos unos pocos casos con valores de identidad de secuencia mayores a 90 % y que a pesar de eso, dan valores de RMSD mayores a 3.0Å y valores de TMscore por debajo de 0.7. Más adelante veremos algunos de estos casos donde se puede ver que corresponden a variabilidades estructurales que hacen que los valores de RMSD y TMscore den valores elevados debido a cambios conformacionales entre el modelo y la estructura real, a pesar de haber conseguido practicamente el mismo plegado.

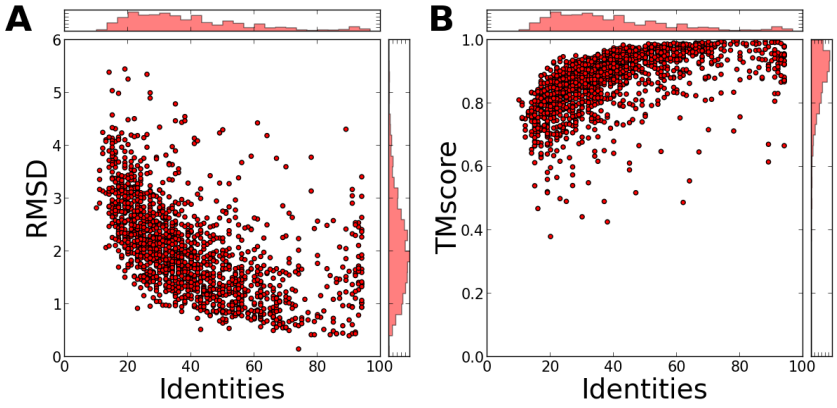

Figura 3.10: **Efecto de la identidad de secuencia en el modelado.** A) RMSD entre el modelo y la estructura proveniente del PDB, en función de la identidad de secuencia para las estructuras del conjunto de prueba. B) TMscore en función de la identidad de secuencia, valores por encima de 0.7 indican que el modelo presenta practicamente el mismo plegado que la estructura proveniente del PDB.

Con respecto al largo del alineamiento, vemos en la Figura 3.11A, que no se aprecia una correlación clara con el grado de error en el plegado. Aunque, como fue previsto, parece tener injerencia en el RMSD que se obtiene, cuanto más

largo es el dominio modelado, más grandes son los valores que puede obtener esta medida. Esto es esperable, dado que el modelado no se está haciendo sobre dominios, sino que, en este caso, modelamos sobre la secuencia completa. Por el contrario, para los valores de TMscore (Figura 3.11B), vemos que a medida que el largo del alineamiento aumenta, es más difícil obtener casos por debajo de 0.8. Mientras que modelos realizados con alineamientos de largo alrededor de 100, dan lugar a varios casos con TMscore por debajo de 0.7, lo cual está de acuerdo con el hecho de que a medida que el largo del alineamiento disminuye, hacen falta relaciones de similitud de secuencia más altas para lograr homología estructural [Sander and Schneider, 1991].

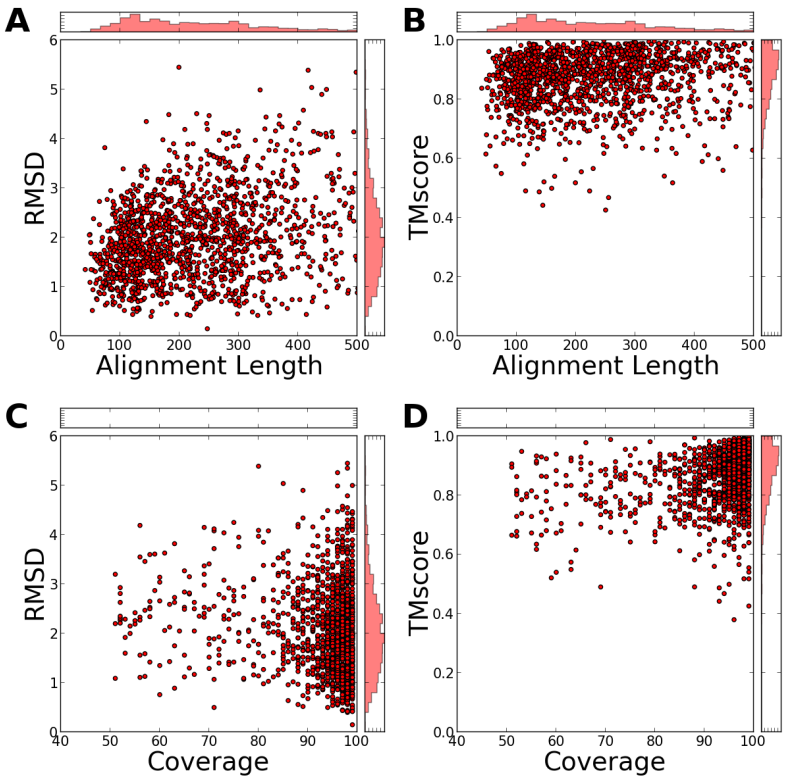

Figura 3.11: **Efecto del largo del alineamiento y de la cobertura en el modelado.** A) Efecto del largo del alineamiento sobre el RMSD obtenido comparando el modelo con la estructura de la secuencia que se usó para modelar. B) Efecto del largo del alineamiento sobre el TMscore. C) Efecto de la cobertura sobre el RMSD. C) Efecto de la cobertura sobre el TMscore.

Analizando la cobertura (Figura 3.11C y D), vemos que a pesar de haber puesto una cota inferior para elegir los templados, la mayoría de los alineamientos presentan valores elevados, casi todos están por encima del 85 %. Aunque a diferencia de la identidad de secuencia, no hay una correlación clara entre este parámetro del alineamiento y las medidas de error obtenidas, debido a que aparecen muchos modelos con alto grado de cobertura sobre la secuencia *target* que presentan valores de bajos de TMscore y valores altos de RMSD.

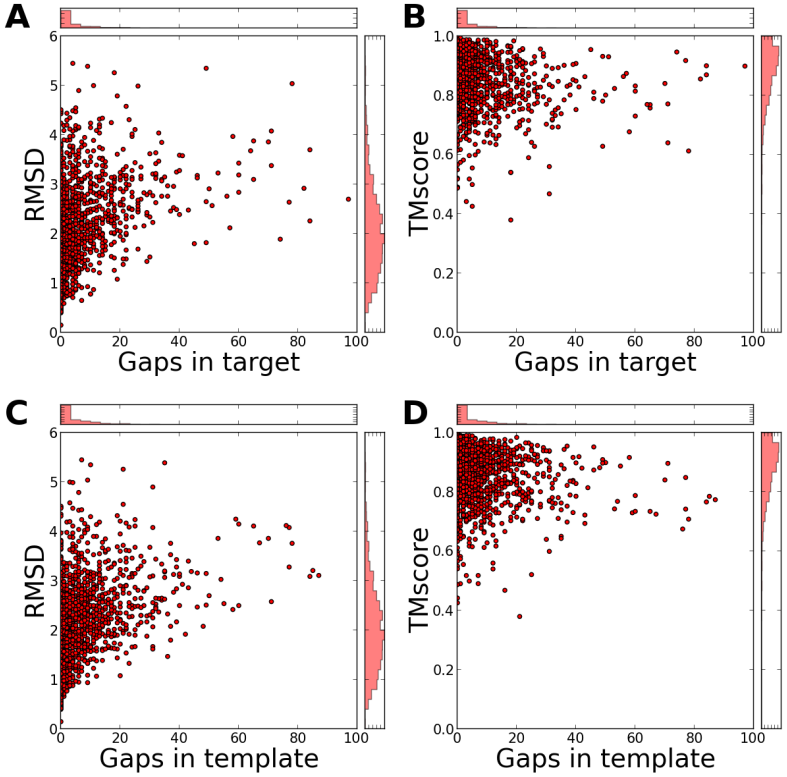

Figura 3.12: **Efecto de la inserción de gaps en el modelado.** A) Efecto de la inserción de gaps en el *target* sobre el RMSD. B) Efecto de la inserción de gaps en el *target* sobre el TMscore. C) Efecto de la inserción de gaps en el templado sobre el RMSD. D) Efecto de la inserción de gaps en el templado sobre el TMscore.

Las inserciones de gaps, tanto en el *target* como en el templado (Figura 3.12), también son importantes para la evaluación final del modelo, obtuvimos que la cantidad de gaps que se insertan, gobierna el mínimo RMSD que se puede lograr, lo cuál es consistente con el hecho de que los gaps suelen agre-

garse en las regiones que presentan Loops, que son regiones que suelen tener una mobilidad que predispone el valor de RMSD hacia valores levemente más elevados que en los casos con menos gaps. Sin embargo, esta tendencia que gobierna el RMSD, no se aprecia tanto en el puntaje obtenido por TMscore, sino que, al contrario, hay casos con fallos fuertes en la predicción del plegado que aparecen en alineamientos libres de gaps. Esto sustenta que el modelado del plegado no se ve afectado por las inserciones de gaps, como en el caso del RMSD. Por último, cabe destacar que no se aprecia una diferencia clara al respecto de si son más dañinos para el modelado las inserciones de gaps en el *target* o en el templado, en ambos casos la tendencia parece estar distribuída de igual manera.

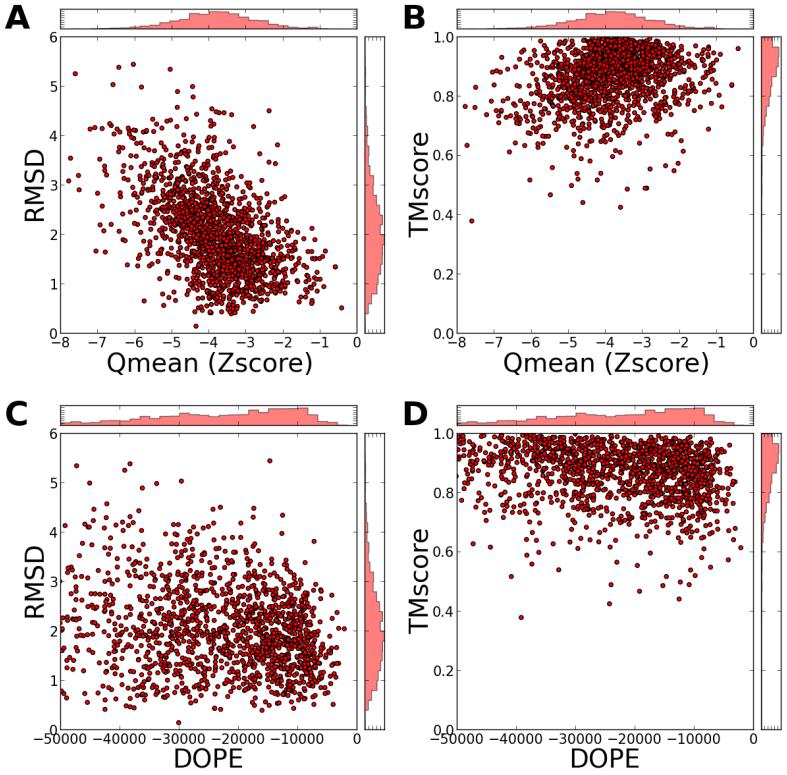

Figura 3.13: **Evaluación de la calidad para los modelos realizados sobre el conjunto de estructuras conocidas.** A) RMSD entre el modelo y la estructura conocida en función del Zscore de QMEAN. B) TMscore en función del Zscore de QMEAN. C) RMSD en función del potencial DOPE. D) TMscore en función del potencial DOPE.

En el caso de los potenciales de calidad, primero se ve que el potencial QMEAN presenta una correlación fuerte con el RMSD obtenido (Figura 3.13A), aunque no pareciera presentar sensibilidad ante los errores en el plegado (Figura 3.13B). Con respecto al DOPE, hay una aparente correlación con el RMSD (Figura 3.13C), se sabe que ambos valores dependen del tamaño de la proteína, lo cuál implica que esta dependencia entre el DOPE y el RMSD perfectamente podría haberse obtenido por azar. Lo mismo para la relación entre el TMscore y el DOPE (Figura 3.13D). Sin embargo, tanto QMEAN como DOPE, en algunos casos detectan ciertos errores de plegado, dado que para valores altos de DOPE y bajos de QMEAN, aparecen valores de TMscore bajos mientras que para DOPE más bajos y QMEAN más altos, no.

3.3.2. Estimación de Calidad

Vimos que la mayoría de los resultados obtenidos, están en el rango de entre 1.0Å y 3.0Å de error y que, muy rara vez, se obtienen plegados mal predichos con TMscore por debajo de 0.5, de hecho la mayoría presentan valores por encima de 0.8, asegurando una alta similitud estructural general de entre todos lo modelos obtenidos. Ahora, en la Figura 3.14, se ve que la identidad de secuencia y el estadístico Zscore de QMEAN correlacionan, como es de esperarse, y que, a su vez, cuanto más altos sean los valores de ambos parámetros, más precisos serán los modelos, y viceversa, cuánto menores sean estos valores, peor será la calidad del modelo. La **identidad de secuencia** y el **Zscore del potencial QMEAN** combinados permiten tener una estimación general del grado de error con el que se producen los modelos medido en términos de RMSD y TMscore. O sea, que la distribución general de los fallos en el plegado y el error en Angstroms de los carbonos alfa, se puede estimar a partir de datos que se conocen luego de obtener el modelo.

Como mencinamos antes, descartamos todos aquellos modelos con $GA341 <$ 0,7 y para todos los modelos que pasen dicho filtro, adoptamos tres niveles de calidad para asignar a los modelos de manera de saber el grado de error general con el que fueron predichos. La primer categoría, *Good*, consiste en modelos que se consiguen modelando con 40 % de identidad para arriba, para los que rara vez se obtienen valores de RMSD por arriba de 2.0Å. Estos modelos son los de mejor calidad aunque pocas veces alcanzan valores menores a 0.5Å. Luego, la próxima categoría, la denominada *Twilight* de los modelos obtenidos que son aquellos con identidad de secuencia entre 20 % y 40 %. En esta zona, los modelos dan buenos resultados sólo cuando el puntaje de QMEAN está por encima de -3. En los casos en que no es así (QMEAN<-3) se empiezan

a ver fallos en la predicción del plegado y el RMSD alcanza valores de entre 3.0Å y 5.0Å. Finalmente, la categoría *Bad*, modelos con menos de 20 % de identidad que son modelos que no se pueden usar en aplicaciones que requieran una precisión de los pocos Angstroms, dado que los valores de RMSD pueden llegar a superar los 5.0Å.

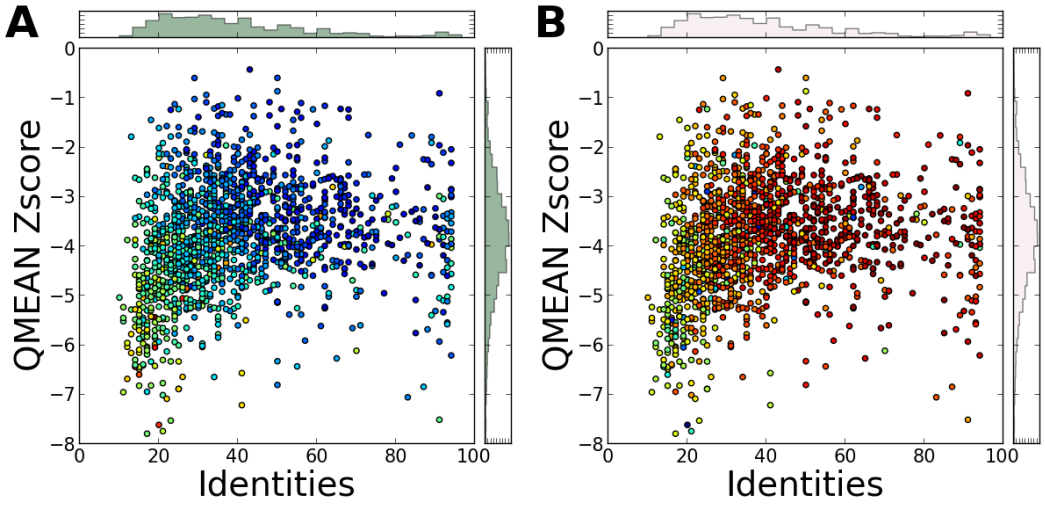

Figura 3.14: **Análisis del rendimiento del pipeline de predicción de estructura.** A) QMEAN Zscore en función de la identidad de secuencia, en colores se aprecia el RMSD, en azul valores menores a 2.0Å, y el amarillo los valores por encima de 4.0Å. B) QMEAN Zscore en función de la identidad de secuencia, en colores se aprecia el TMscore, rojo para valores por encima de 0.9, y verde para valores por debajo de 0.7.

Finalmente, cabe analizar aquellos casos que caen en la categoría *Good*, pero presentan fallos que no pueden ser detectados por los parámetros que se obtienen luego de modelar. Uno de estos es el caso de 3ckfA como se ve en la Figura 3.15A, que tiene una identidad de secuencia bastante elevada (89 % de identidad y 99 % de cobertura), y presenta valores de RMSD=4.31 y TMscore=0.67, al borde del fallo por plegado incorrecto, aunque con un valor de QMEAN=-3.96, lo cuál permitiría sospechar de un modelo con una calidad no tan buena. Pero en el caso de 3gmtA (Figura 3.15B), que presentó un alineamiento con 64 % de identidad y 99 % de cobertura y un puntaje de QMEAN de -2.8, y aún así presenta RMSD de 4.19 y un TMscore de 0.55, casi al límite del parecido de plegado al azar. Este caso hubiera sido imposible de detectar usando los parámetros de alineamiento y los potenciales de calidad.

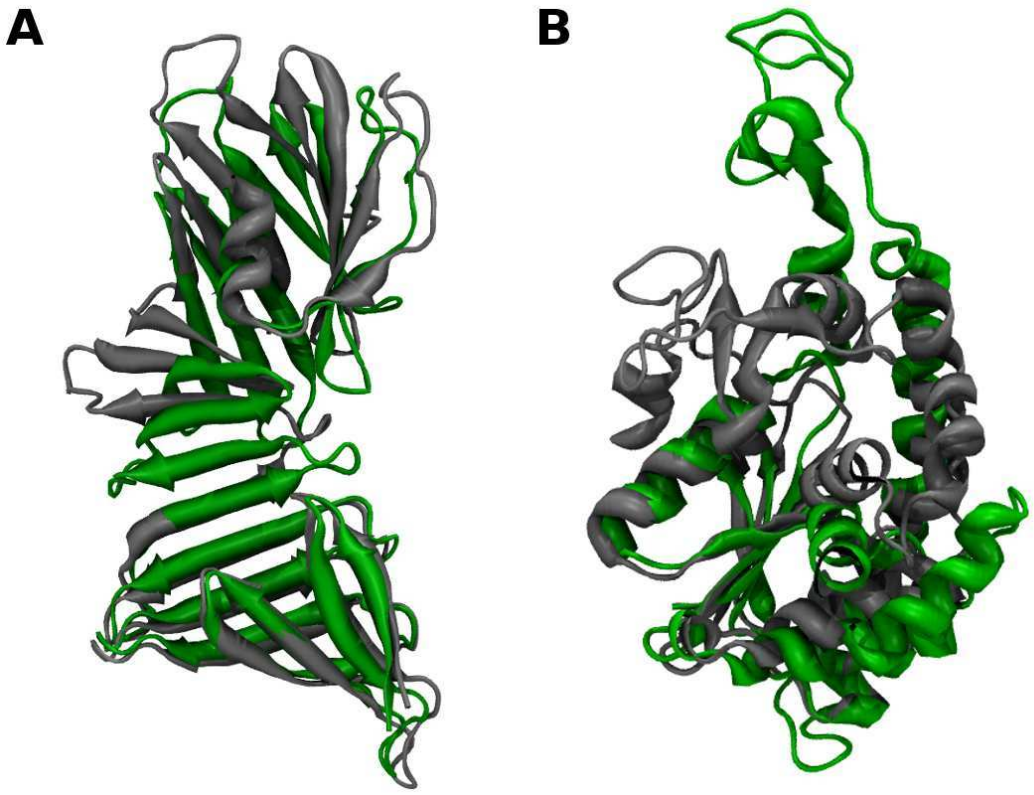

Figura 3.15: **Errores de modelado que no fueron capturados.** Modelos que fueron predichos con errores y no fueron detectados por los parámetros del alineamiento ni por los potenciales de calidad. A) Modelo de 3ckfA con errores críticos y alta identidad de secuencia, tiene 89 % de identidad de secuencia y presenta un RMSD=4.31 y un TMscore=0.67. B) Modelo de 3gmtA con RMSD=4.19 y TMscore=0.55 a pesar de haber sido predichas de buena calidad (ident=64 %, QMEAN=-2.8). *En verde las estructuras conocidas y los modelos en gris*

Teniendo en cuenta lo visto hasta recién, casos como los mencionados previamente, no son más que modelos que predijeron una estructura con el mismo plegado pero en otra conformación. Casos en los que la estructura nativa tiene uno o más dominios que manifiestan movimientos no concertados y, si bien cada uno de ellos presentaría valores bajos de RMSD por separado, sus orientaciones mutuas dan lugar a un modelo en una conformación que difiere de la estructura conocida que se usa para comparar. Estas son consideradas proteínas difíciles, en las que el pipeline falla en encontrar la conformación adecuada, principalmente por una cuestión de la variabilidad conformacional.

Finalmente, pudimos ver que los modelos que produce el pipeline son lo suficientemente buenos como para no cometer errores fuertes, y además tenemos una manera de clasificar los modelos según su nivel de calidad en función del RMSD y el TMscore que pueden llegar a adoptar.

3.3.3. Aplicación del pipeline a genomas de bacterias

El pipeline de predicción de estructura fue aplicado al estudio de grandes conjuntos de secuencias de proteínas provenientes de genomas de patógenos bacterianos secuenciados, con el objetivo de priorizar blancos terapéuticos. Predijimos las estructuras de 3 genomas: *Mycobacterium tuberculosis*, *Klebsiella pneumoniae* y *Corynebacterium pseudotuberculosis*.

En el caso de *M. tuberculosis* [Radusky et al., 2014], primero se filtraron aquellas secuencias proteicas para las cuales haya sido depositada una estructura en el PDB. Este primer filtro, dejó afuera de la construcción de modelos a 441 estructuras únicas qué están disponibles. Para todo el resto de los ORFs, intentamos predecir su estructura, usando el pipeline mencionado previamente. Para *C. pseudotuberculosis* [Radusky et al., 2015] se analizaron varias cepas de la especie, y una de ellas (la cepa 1002) fue usada como base para el estudio estructural. Por cada secuencia en 1002, se construyeron varios modelos usando los protocolos mencionados previamente. Cada secuencia perteneciente a las otras 14 cepas fue comparada usando BLAST, y para cada secuencia que diera un hit significativo con identidad arriba del 85 %, se usó una metodología de mutaciónes puntuales denominada *Single Amino Acid Sustitution (SAAS)*, en cada amino ácido diferente entre la secuencia en dicha cepa y la secuencia en 1002, cambiando las estructuras modeladas obtenidas previamente. Al los fines de este trabajo de tesis, usamos para el análisis sólo los modelos obtenidos para 1002. Por útlimo, en el caso de *K. pneumoniae*, hicimos un modelado para cada ORF en el genoma, de la misma manera que se mencionó para los otros dos patógenos.

Ahora bien, usando los parámetros previamente mencionados para puntuar la calidad de los modelos, nos preguntamos con qué grado de error estamos generando modelos usando el pipeline mencionado previamente. Nos interesó, analizar particularmente: i) la proporción de ORFs para los que obtuvimos templados, ii) de los modelos generados, cuántos ORFs obtuvieron al menos un modelo aceptable, usando como criterio sólo el valor de la función GA341>0.7, iii) según los criterios de corte que mencionamos previamente, con qué grado de error estamos abarcando el genoma del patógeno en cuestión. Estos tres puntos a analizar pueden ser deducidos de los alineamientos y los valores de

los potenciales como explicamos anteriormente.

Genoma	**Total de ORFs**	**ORFs con templados**	**Modelos confiables**
Corybacterium pseudotuberculosis (Cp)	2090	1187 (56.8 %)	1051 (50.2 %)
Mycobacterium tuberculosis (Mt)	3982	2722 (68.9 %)	2337 (58.6 %)
Klebsiella pneummonieae (Kp)	5736	3539 (61.7 %)	3348 (58.3 %)

Cuadro 3.1: **Resultados del modelado estructural para los genomas de 3 patógenos.** Los **ORFs con templados** corresponden a aquellos que tienen modelos pero tienen un GA341<0.7, por lo que no se los considera confiables. Los **Modelos confiables** son los ORFs que tienen por lo menos un modelo que cumple con el criterio de GA341. *Los porcetajes se calculan con respecto al total de ORFs.*

Como se puede ver en el Cuadro 3.1, los resultados de los distintos genomas pueden variar bastante. En el caso de *C. pseudotuberculosis* la cantidad de ORFs predichos es bastante menor que en el caso de *M. tuberculosis* y *K. pneumoniae*, debido al tamaño reducido del genoma. *M. tuberculosis* presenta el menor cubrimiento sobre el total de ORFs, así como la proporción de modelos aceptables. *K. pneumoniae* es el patógeno con más similitud estructural contra las secuencias en la base de datos de templados, llegando al 58.3 % del total de ORFs que tienen un modelo confiable. *K. pneumoniae* es para el que menos modelos se tuvo que descartar por confiabilidad reduciendo sólo un 4 % de los ORFs por no cumplir con el criterio de $GA341 > 0{,}7$. Aunque, para los otros dos sólo se descarta alrededor de un 6 % lo cuál tampoco es demasiado elevado. Esto indica que el pipeline de modelado comparativo logra producir estructuras de proteínas de patógenos bacterianos que son en su mayoría confiables.

Genoma	**Good**	**Twilight**	**Bad**
Cp	547(52.0 %)/-0.8/40.6	203(19.31 %)/-2.8/27.3	301(28.6 %)/-5.4/14.7
Mt	897(38.3 %)/-1.0/40.8	552(23.6 %)/-2.8/27.0	888(37.9 %)/-5.3/14.7
Kp	2247(67.1 %)/-0.7/60.2	508(15.1 %)/-2.8/27.2	593(17.7 %)/-5.8/15.6

Cuadro 3.2: **Estadísticas generales del modelado estructural según las 3 categorías propuestas: Good, Twilight, Bad**. En cada celda, entre barras, se presentan de izquierda a derecha, los valores de: Cantidad de ORFs en la categoría, Zscore de QMEAN promedio e identidad de secuencia promedio. *Los porcetajes se calculan con respecto al total de ORFs.*

Con respecto a las categorías de calidad para aquellos modelos aceptables (Cuadro 3.2), en la categoría *Good*, la identidad de secuencia promedio en el

caso de *K. pneumoniae* indica que sus modelos son considerablemente mejores que los de *C. pseudotuberculosis* y *M. tuberculosis*, y también el valor del Zscore de QMEAN. O sea, que no sólo a nivel de cubrimiento sobre el total de ORFs, sino que, en general, *K. pneumoniae* tiene ORFs que son más fáciles de modelar. Los modelos de buena calidad llegan hasta casi un 40 % del total de ORFs, con una identidad de secuencia promedio del 60 %, mientras que para los otros dos patógenos (*M. tuberculosis* y *C. pseudotuberculosis*) sólo se llega a un promedio del 40 %. Por último, en las otras dos categorías, todos los valores (cantidad/QMEAN/identidad) son parecidos entre los 3 genomas, lo cuál sugiere que las proteínas difíciles de modelar podrían llegar a ser un reto para cualquier patógeno con el que se esté trabajando.

3.4. Conclusiones

Al día de hoy las técnicas de modelado comparativo, se han mostrado de gran utilidad para varias aplicaciones. En esta instancia del trabajo, pusimos a punto un pipeline de modelado estructural que da resultados aceptables. Tambien fuimos capaces de estimar el grado de error que se produce, en promedio en los modelos, en función de parámetros que pueden observarse, una vez terminado el proceso de modelado. Este tipo de pipelines es considerado una de las formas más sencillas de obtener estructuras usando pocos recursos computacionales, comparado con otras técnicas como Threading o estrategias Ab initio.

Se puede decir que lo más importante a la hora de modelar es tener templados que se parezcan lo más posible, lo cuál no es noticia nueva. Sin embargo, vemos que a pesar de que la técnica de modelado comparativo comete pequeños errores, estos rara vez son más allá de 5.0Å. Es más, para los tres estructuromas que vimos, más de la mitad de los modelos obtenidos fueron de buena calidad, presentando un error promedio de menos de 2.0Å. Sin embargo, los modelos que presentan una calidad baja no pueden ser usados para estudios en las que estén involucradas, por ejemplo, interacciones de corto alcance como los puentes de hidrógeno.

En el caso de los genomas que modelamos, vimos que *M. tuberculosis* es bastante más dificil de modelar que *K. pneumoniae*, no sólo en el sentido que *M. tuberculosis* presentó una proporción menor de templados con los cuales construir los modelos, sino que además, la distribución de identidades de secuencia, y de puntajes QMEAN, supone una proporción más baja de modelos de buena calidad. Esto es bastante curioso dado que en el PDB, hoy día es posible encontrar más de 1000 entradas que refieren a proteínas de *M. tu-*

berculosis mientras que para *K. pneumoniae*, todavía no superan las pocas decenas. Lo cuál abre la pregunta: ¿qué genoma tiene más información estructural disponible? ¿el que tiene más estructuras resueltas o el que tiene más modelos confiables?

Capítulo 4

Clusters de Aromáticos en proteínas: Plegado, Interacción con Drogas y en Complejos Proteína-Protéina

4.1. Introducción

El rápido incremento de información estructural de proteínas acumulandose en el PDB, ofrece la oportunidad de mirar el espacio estructural de las biomoléculas como un todo, para explorar/enteder los principios que regulan la relación estructura función. Varios estudios sobre proteínas han mostrado que las interacciones que establecen los residuos aromáticos, son importantes para la estabilización de la estructura proteica, para el proceso de plegado, para el reconocimiento en las interacciones proteína-proteína y para la unión a ligandos [Burley and Petsko, 1985], [Bashford et al., 1987], [Serrano et al., 1991], [Mitchell et al., 1994], [Lesk and Fordham, 1996], [Chothia et al., 1998], [Michnick and Shakhnovich, 1998], [de Araujo et al., 1999], [Samanta et al., 1999], [Kannan and Vishveshwara, 2000], [Larson and Davidson, 2000], [Bhattacharyya et al., 2002], [Chelli et al., 2002], [Aravinda et al., 2003], [Johnson et al., 2007], [Espinoza-Fonseca and García-Machorro, 2008], [Eidenschink et al., 2009]. Al día de hoy, las interacciones diméricas de anillos aromáticos han sido caracterizadas en profundidad y se han clasificado en 3 conformaciones dependiendo de la orientación entre los ángulos planares definidos por los anillos aromáticos: cara a cara (o $\pi - stacking$), cara a borde (o $T - shape$) o borde a borde, teniendo en cuenta que los ensambles también suelen adquirir orienta-

ciones intermedias. Los estudios previos realizados en estructuras de proteínas muestran que, de las tres mencionadas, la interacción aromática está favorecida en las conformaciones de tipo *T-shape* [Burley and Petsko, 1985], [Hunter et al., 1991]. A su vez, las moléculas aromáticas tienden a formar *clusters* de más tamaño con una naturaleza energética aditiva, que también adoptan conformaciones específicas como se vió en el caso de los *clusters* de benceno en vacío [Kannan and Vishveshwara, 2000], [Easter et al., 2005], [Morimoto et al., 2007], [Gonzalez and Lim, 2001], [Dang, 2000], [Engkvist et al., 1999], [Krause et al., 1991], [De Meijere and Huisken, 1990], [Tauer and Sherrill, 2005]. Estos *clusters*, han sido analizados usando métodos tanto experimentales como computacionales. Teniendo esto en cuenta, pensamos que estos *clusters* de mayor tamaño, deberían aparecer también en las estructuras proteicas y que podrían ser importantes para la estructura y la función de las proteínas. Para responder a esta pregunta, identificamos y caracterizamos los *clusters* formados por anillos aromáticos, basandonos en las estructuras disponibles en el PDB.

Desde el punto de vista de la comunidad química, los anillos aromáticos se usan con frecuencia en el diseño de drogas, dado que la introducción de interacciones aromáticas entre una droga y su *target* puede contruibuir a optimizar tanto su afinidad como su especificidad [Li et al., 2013]. El principal objetivo del diseño de drogas es desarrollar compuestos novedosos que puedan prevenir o curar clínicamente enfermedades importantes mediante tres maneras: inhibir las funciones del *target*, compitiendo con sus sustratos naturales, inhibir las interacciones proteína-proteína uniendo contra las interfaces de interacción y, en algunas enfermedades como el cancer, activar proteínas que pierden su función debido a mutaciones [Mandal et al., 2009]. Pero, aparte de la búsqueda por encontrar compuestos que se unen muy fuertemente a su *target*, se ha hecho un considerable esfuerzo analizando las interacciones que dichos compuestos establecen con otras proteínas que puedan dar lugar a efectos laterales no deseados [Huggins et al., 2012]. Un caso paradigmático en éste área es la familia de las proteínas quinasas, que han sido el objetivo de muchos compuestos químicos que deben unir a una proteína específica evitando generar interferencia con otras proteínas estructuralmente similares. Entre todos estos compuestos, la pirimidina es el motivo más común entre ellos [Ghose et al., 2008], lo cuál nos muestra que los sistemas aromáticos parecen ser una característica útil usada en muchos compuestos líderes que son modificados posteriormente para optimizar la selectividad. Por otro lado, si se requiere una selectividad amplia (*promiscuidad*, cubrir más de un *target*), como es el caso de las drogas de tienen como *target* la proteasa de HIV debido a su alta tasa de mutación, los sistemas multi-anillo han mostrado ser inhibidores exitosos [Jayaraman and

Shah, 2008]. Por eso, la popularidad de los anillos aromáticos se se debe en parte al hecho de que permiten la generación de *esqueletos* comunes en los compuestos líderes que pueden ser optimizados de manera de lograr mejorar la selectividad según los requerimientos del *target* [Ritchie and Macdonald, 2014]. Por otro lado, cuanto menor es la cantidad de anillos aromáticos que contiene una droga de vía oral, más desarrollable es la droga en términos de los candidatos que se conocen en el mercado, más específicamente, tener más de 3 anillos aromáticos hace que el compuesto tenga muchas chances de no lograr ser una buena droga incrementando el riesgo de desarrollar dicho compuesto [Ritchie and Macdonald, 2009]. También, si se sabe que un compuesto en proceso de optimización tiene baja solubilidad, disminuir la cantidad de anillos aromáticos suele ser beneficioso. Esto sugiere que los anillos aromáticos son un recurso importante que debe ser tenido en cuenta y racionalizado. Entonces, la localización de los anillos aromáticos en una droga debería ser optimizada de acuerdo a las interacciones más importantes con su *target*.

Con respecto a las interfaces de interacción entre proteínas, los residuos aromáticos, son importantes en el proceso de reconocimiento y unión jugando el rol de residuos *ancla* [Rajamani et al., 2004]. Las interfaces entre proteínas presentan residuos polares y aromáticos cerca del medio de la región de contacto y, sólo en el 20 % de los complejos, residuos alifáticos cumplen el rol de ancla. Además, otros estudios [Ma et al., 2003] han visto que residuos aromáticos conservados, principalmente Trp, y en menor medida Phe y His, en la superficie de las proteínas, indican en gran medida una interfaz de interacción proteína-proteína. Las interfaces de interacción entre proteínas no sólo son importantes desde el punto de vista biológico. Desde el punto de vista químico, en los últimos años, hubo un incremento en el desarrollo de drogas diseñadas para unir interfaces de unión entre proteínas. Debido a que en la optimización de un inhibidor competitivo, puede ser que sea más fácil competir con un interactor proteico que con un pequeño ligando. Este es el caso, de la proteína CDK2 de la familia de las quinasas, para la cuál, la búsqueda de la especificidad en drogas que compiten por el ATP es todo un reto, debido a la similaridad estructural en el bolsillo donde se une. Sin embargo, la unión a la proteína que fosforila dicha quinasa es diferente a las demás quinasa, lo cuál es prometedor para la optimización de su especificidad [Chohan et al., 2015].

A continuación, comentaremos acerca de cómo construimos conjuntos de datos basandonos en el PDB, para estudiar los *clusters* de aromáticos y cómo los definimos a partir de la detección de interacciones aromáticas, que nos permitirá detectarlos en todas las estructuras analizadas.

4.2. Materiales y Métodos

4.2.1. Generación de conjuntos de datos.

Con el objetivo de caracterizar geométricamente y entender el rol de los *clusters* de aromáticos en proteínas, decidimos separarlos en: i) aquellos involucrados sólo en el plegado de la proteína, o sea teniendo en cuenta las interacciones aromáticas que se dan entre los residuos de la misma cadena proteica, ii) aquellos presentes en sitios donde se encuentren unidos ligandos tipo droga y iii) los *clusters* involucrados en las interfaces de interacción proteína-proteína. Para esto, seleccionamos de entre todas las estructuras disponibles en el PDB, 3 conjuntos de datos que comentaremos a continuación:

Selección de PDBs para detección de interacciones aromáticas intraproteína (IP) Para estudiar las interacciones aromáticas entre los residuos de una misma cadena, seleccionamos estructuras a partir del PDB. Fueron consideradas aquellas estructuras de cadenas proteicas que cumplen con ser proteínas únicas definidas según el identificador de Uniprot, y además, pasamos un filtro por identidad de secuencia, *clusterizando* las secuencias de dichas cadenas usando CD-HIT al 95 % de identidad de secuencia y quedandonos con la secuencia centroide. Este proceso dió lugar a 18547 casos para analizar.

Selección de PDBs para complejos proteína-droga (PD) Para estudiar las interacciones aromáticas que establecen los anillos aromáticos de las drogas con los anillos aromáticos de las proteínas, primero filtramos todos aquellos compuestos que figuren en las lineas HETATM de cada archivo PDB, que no sean residuos modificados. Segundo, filtramos aquellos que tienen menos de 100Da de peso molecular. Tercero, seleccionamos compuestos tipo droga. Ordenamos todos los nombres de los compuestos químicos (los códigos de tres letras) en el PDB por su cantidad de apariciones en toda la base de datos, de mayor a menor. Filtramos manualmente aquellos que aparecían muchas veces debido a que correspondían a metabolitos conocidos hasta que llegamos a compuestos que aparecen menos de 100 veces. Esto es para reducir el sesgo que hay debido a que algunos compuestos químicos aparecen diferencialmente más que otros. Obtuvimos 13200 compuestos únicos, a partir de los cuáles elegimos todas las entradas del PDB que los contengan (52926 entradas PDB). Finalmente, aplicamos un filtrados por 95 % de identidad de secuencia como se explico en el caso del conjunto de datos para las interacciones intracatenarias. Este proceso dió lugar a 17352 casos para analizar.

Selección de PDBs para complejos proteína-proteína (PP) En este caso, decidimos enfocarnos en las interacciones entre oligómeros diméricos tanto de homodímeros como heterodímeros. Incluir el aporte de cada cadena en el análisis de los *clusters* aporta una coordenada más que quisimos involucrar. Elegimos de entre las entradas del PDB, aquellas que contengan sólo 2 cadenas y que mencionen explícitamente (en el encabezado del archivo PDB) que la estructura corresponde a un dímero. Sobre éstas entradas aplicamos la misma estrategia de *clusterizado* entre las secuencias que mencionamos para los dos casos anteriores (a 95 % de identidad de secuencia), de las cuales nos quedamos una entrada por cada par centroide-centroide, obteniendo tanto homodímeros como heterodímeros. Este proceso dió lugar a 8279 casos para analizar.

Desarrollo de una base de datos relacional de interacciones aromáticas Para detectar interacciones aromáticas como comentaremos a continuación, construimos una base de datos que integra la información de los 3 conjuntos de datos mencionados previamente. Mediante scripts escritos en python, calculamos las propiedades que serán desarrolladas a lo largo de este capítulo. Estos scripts utilizan principalmente 2 bibliotecas de software: Biopython, Pybel (OpenBabel). Con Biopython parseamos los PDBs, levantandolos a memoria en una estructura de datos como la mencionada en el Capítulo 3 (Estructura → Cadena → Residuo → Átomo), y con Pybel se parsea cada compuesto tipo droga y cada residuo en las proteínas de manera de detectar los anillos aromáticos. Una vez obtenidas las interacciones se identifican los *clusters* y con los centros geométricos de los anillos se calculan *ángulos conformacionales* que describen su configuación. Además, se calculan parámetros que describen su entorno y posición relativa al centro de la proteína. Por útlimo, para cada residuo formando cada cluster se calcula su estructura secundaria y su accesibilidad al solvente usando DSSP y VMD como fue mencionado en el Capítulo 3.

4.2.2. Detección de interacciones aromáticas.

Para detectar interacciones aromáticas necesitamos detectar los anillos, que según la biblioteca Pybel, que es la que usamos, se trata a cada anillo como grupos de 5 o 6 átomos, por lo que sistemas fusionados como el indol, se toman como 2 anillos: un benceno y un pirrol. Una vez detectados los anillos, se conoce que las interacciones aromáticas pueden ser descriptas principalmente por la distancia entre los centros de cada uno de ellos [Chelli et al., 2002]. Sin embargo, también fueron propuestos algunos ángulos que describen, por ejemplo, la posición relativa de los anillos entre sí, o bien el ángulo entre la

normal de los planos [Brocchieri and Karlin, 1994], [McGaughey et al., 1998], [Marsili et al., 2008]. En este trabajo, nos enfocamos en la distancia (d) y el ánguilo planar (α) entre los planos aromáticos, como está definido en la Figura 4.1 A.

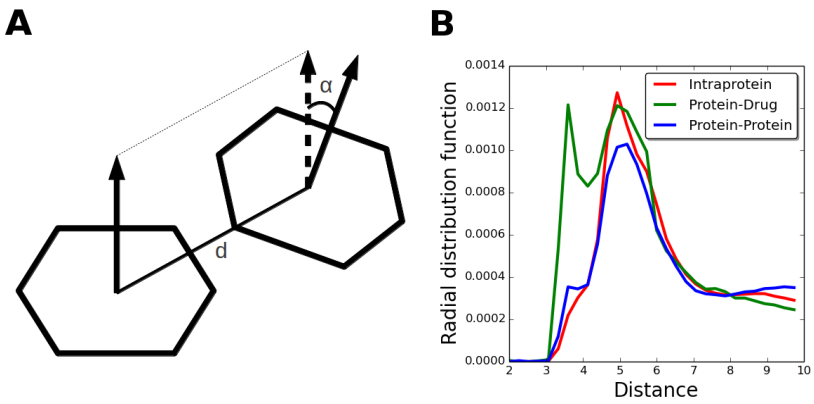

Figura 4.1: **Interacciones aromáticas presentes en proteínas.** A) Esquema de la distancia (d) y el ángulo planar (α) en una interacción aromática. B) Función de distribución radial para la distancia entre los centros de anillos aromáticos, IP (rojo), PD (verde) y PP (azul).

Estos dos parámetros son muy útiles para visualizar las diferentes conformaciones en las que dos anillos aromáticos pueden estar interactuando: $\pi - stacking$ y $T - shape$. Usando estos parámetros, buscamos entre las entradas de los conjuntos de datos mencionados previamente para identificar y analizar las interacciones basandonos en un umbral de distancia. Se puede hacer una distinción entre las distribuciones radiales ($g(r)$, Figura 4.1 B) de las distancias entre los 3 conjuntos de datos.

En el caso de las interacciones que se establecen entre anillos de residuos (R) entre sí, se puede apreciar un sólo máximo local en toda la distribución, tanto para IP como para PP. En cambio, en el caso de PD, se aprecian, 2 máximos. Como veremos más adelate, ambos máximos corresponden a las configuraciones mencionadas previamente (π y T). Esto se da, debido a que en el caso de PD las interacciones están más libres, lo cuál permite que las interacciones en su conjunto visiten más ambos mínimos del terreno energético, y da lugar a interacciones aromáticas más compactas. Para el caso de IP y PP, la extistencia de un sólo mínimo, significa que las interacciones están restringidas en sus grados de libertad, probablemente por las restricciones

impuestas por el backbone de la proteína sobre los residuos aromáticos. En función de esto fijamos un umbral a 7.0Å para las interacciones aromáticas entre los centros geométricos de los anillos aromáticos. A las interacciones entre residuos tanto en IP como en PP, las llamamos R-R y las que se dan entre anillos en las drogas y anillos en los residuos, interacciones D-R.

4.2.3. Definición de *cluster* de aromáticos

La interacción aromática definida arriba define una relación entre dos anillos aromáticos. Podemos pensar en estas interacciones como los ejes de un *grafo* en el cual los nodos serían los anillos aromáticos. Para caracterizar tres o más anillos que interactúan formando *clusters* de mayor tamaño, definimos cada *cluster de aromáticos* como cada una de las *componentes conexas* del grafo mencionado previamente. Usando este criterio buscamos caracterizar geométricamente los *clusters* aromáticos presentes en los tres conjuntos de datos (IP, PD y PP). De los complejos PD, contamos sólo aquellas interacciones D-R, dadas entre los anillos aromáticos en los residuos y los anillos aromáticos en las drogas. En el caso de PP, nos quedamos con aquellas interacciones R-R que involucran un residuo en cada cadena y para las interacciones en IP, sólo aquellas interacciones R-R entre la misma cadena. Cabe aclarar que tanto para PD como para PP, para contar la cantidad de interacciones que tienen los *clusters*, también contamos las interacciones R-R que establecen los anillos de la misma cadena proteica, de manera de no subestimar los *clusters* más compactos.

4.3. Resultados

4.3.1. Las interacciones aromáticas en el PDB: IP, PD y PP

Como se puede ver en el Cuadro 4.1, para PD y PP, obtuvimos 22676 y 21603 interacciones aromáticas, respectivamente, mientras que en el caso de IP fueron 467775. En cuanto a los *clusters*, la tendencia es parecida, sólo que dadas las cantidades que encontramos obtenemos un promedio de 1.82 y 2.19 interacciones por *cluster* para PP y PD respectivamente, mientras que para IP tenemos 3.93 interacciones por complejos, lo cuál indica que los *clusters* IP están más empaquetados que los *clusters* PD y PP. También se dió que en los 3 conjuntos de datos hay casos donde no encontramos ni una sola interacción aromática y la proporción de aquellos que tienen por lo menos una interacción aromática presenta diferencias, en IP llega hasta el 88 %, mientras que para

PD y PP sólo 49 % y 53 %.

	Intraproteína	**Proteína-Droga**	**Proteína-Proteína**
Cantidad de entradas	18547	17352	8283
Entradas con al menos un *cluster* detectado	16430 (88 %)	8506 (49 %)	4422 (53 %)
Cantidad de anillos aromáticos detectados	118821	10328	11827
Cantidad de interacciones aromáticas detectadas	467775	22676	21603

Cuadro 4.1: **Estadísticas de los *clusters* de aromáticos.** De la segunda a la cuarta columna, en orden: Intraproteína (IP), Proteína-Droga (PD) y Proteína-Proteína (PP). *Los porcentajes en la segunda fila son con respecto a las entradas totales de la primer fila.*

En el caso de los complejos PD, las familias de proteínas más abudantes según Pfam, son las denominadas quinasas (Pkinase:PF00069 y Pkinase_tyr:PF07714) con el 10.09 % del conjunto de datos, seguidas por las tripsinas (Trypsin:PF00089), los receptores nucleares (Hormone_recep:PF00104) y las aspartil-proteasas retrovirales (RVP:PF00077), con el 4.66 %, 3.58 % y 2.24 % respectivamente. Más del 65 % entre todas las drogas del conjunto de datos, tienen por lo menos un anillo aromático, siendo más abudantes aquellas con 2 anillos aromáticos. También, más del 86 % de los compuestos que contienen anillos aromáticos están involucrados en por lo menos un *cluster* aromático y aquellos que están formando un sólo *cluster* comprenden más del 50 % de los casos. En cuanto a la cantidad de anillos aromáticos distintos presentes entre cada una de las drogas, resultó que el anillo más abundante es el benceno alcanzando un 55.5 % de todos los anillos pertenecientes a compuestos, seguido por pirimidinas (11.6 %), piridinas (8.1 %). Las familias más abudantes en los complejos PP son las inmunoglobulinas (V-set:PF07686 and C1-set:PF07654), las proteínas quinasas (Pkinase:PF00069), las proteínas Ras (Ras:PF00071) y las deshidrogenasas de cadena corta (adh_short:PF00106), con 2.1 %, 1.04 %, 1.03 % y 0.91 % respectivamente.

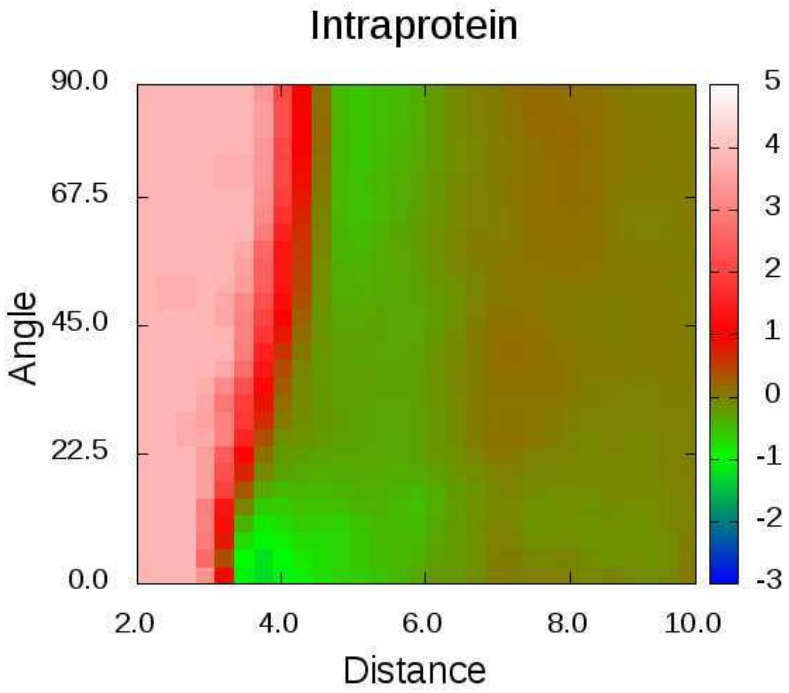

Figura 4.2: **Distribución de interacciones aromáticas intraproteína (IP).** Relación entre el ángulo planar aromático y la distancia entre los centros de masa de los anillos. Se pueden diferenciar los dos tipos de interacciones: la $T-shape$ con ángulos superiores a 60° y el $\pi-stacking$ con ángulos por debajo de 10°

Usando los 3 tipos de *clusters* detectados, analizamos las interacciones aromáticas involucradas, comparándolas en términos de las distribuciones de distancia y ángulo planar. Como se puede ver en la Figura 4.2, usando el ángulo planar podemos distinguir los 2 mínimos energéticos presentes en IP: $forma\text{-}T$, con un águlo planar mayor a 60° y $\pi-stacking$ con un ángulo planar menor a 15°). En este caso, se aprecia un muy leve enriquecimiento para el caso de $\pi-stacking$, pero en general, no hay una abundancia diferencial entre alguna de las dos conformaciones y, más aún, la distribución del ángulo planar está distribuída más uniformemente que en los otros dos casos (PP y PD).

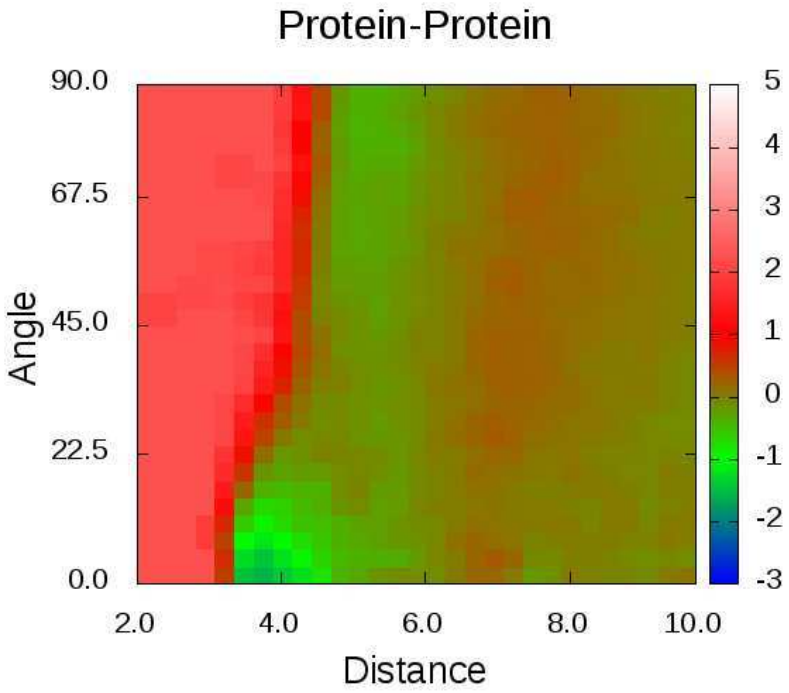

Figura 4.3: **Distribución de interacciones aromáticas entre proteínas (PP).**

En el caso de las interacciones aromáticas en las interfaces de interacción entre proteínas, la distribución de las interacciones (Figura 4.3) presenta una superficie con las mismas características que en el caso anterior (IP). Tampoco encontramos una conformación preferencial entre las dos conformaciones de interacción. Ambas distribuciones coinciden y no se observan diferencias significativas, salvo por un leve pico en 3.5Å en el caso de PP, que se aprecia en éste gráfico como un aumento muy leve en la región de $\pi - stacking$. Además, la distribución en distancia (Figura 4.1B) presenta un valor máximo alrededor de 5.3Å, y el primer mínimo se encuentra a 7.0Å. Sin embargo, en interacciones PD, si bien encontramos el mismo máximo alrededor de 5.3Å, también encontramos otro a 3.5Å de distancia, que corresponde al valor óptimo de distancia de interacción en la conformación de $\pi - stacking$ según [McGaughey et al., 1998]. Las interacciones aromáticas entre drogas y proteí-

nas parecen ser más compactas, en términos de la distancia entre los centros de masa de los anillos, que en el caso de las interacciones entre e intra proteínas.

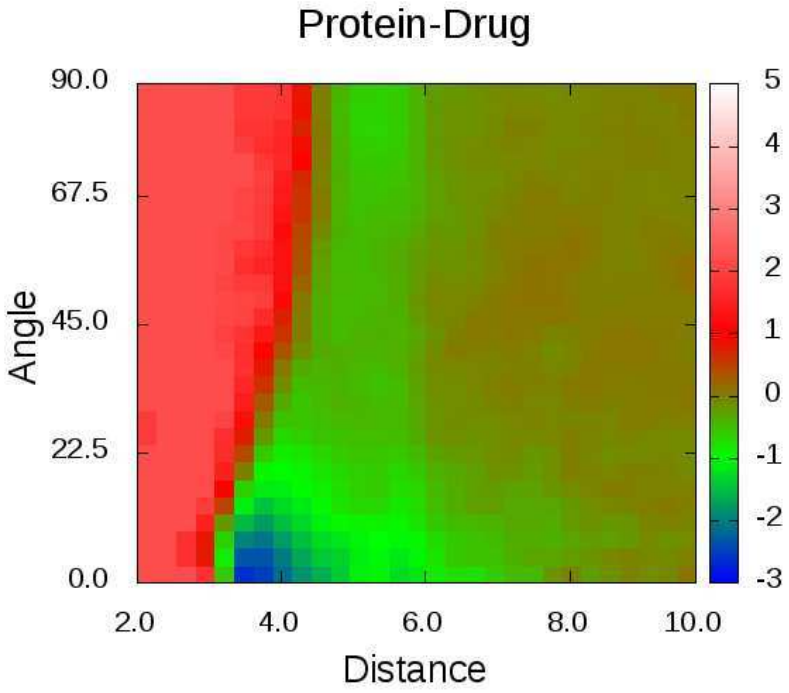

Figura 4.4: **Distribución de interacciones aromáticas entre drogas y proteínas (PD).**

Como se puede ver en la Figura 4.4, en las interacciones en PD, la conformación de $\pi - stacking$, es la más abudante con una distancia de 3.5Å (el mismo máximo mencionado previamente) y el ángulo planar por debajo de 10°. Esto es consistente con los resultados obtenidos por [Chelli et al., 2002] donde se aprecia un aumento en la variación de energía libre de la conformación π por sobre la T en ambientes hidrofílicos. En el caso de IP, esto no es así debido a la naturaleza hidrofóbica de los núcleos preoteicos. Mientras que para los *clusters* en PP, las interacciones parecieran estar en un ambiente casi intermedio, donde se aprecia una 2da componente en el $g(r)$ a 3.5Å que es

concecuencia de que las interfaces entre proteínas son levemente más hidrofílicas que los núcleos proteicos pero no tanto como los *bolsillos* dónde se unen ligandos tipo droga.

4.3.2. Los *clusters* de anillos aromáticos

En términos generales, se puede ver en el Cuadro 4.2, cómo la abundancia de éstos *clusters* decrece conforme aumenta el tamaño del *cluster*.

Tamaño del *Cluster*	Proteína-Droga	Proteína-Proteína	Intraproteina
2	4900 (47.5)	6927 (58.6)	49565 (41.7)
3	2596 (25.2)	2750 (23.3)	25983 (21.9)
4	1379 (13.4)	1201 (10.1)	13461 (11.3)
5	664 (6.4)	452 (3.8)	8121 (6.8)
6	398 (3.9)	235 (2.0)	6026 (5.1)
7	193 (1.9)	115 (1.0)	3905 (3.3)
8	186 (1.8)	147 (1.2)	11760 (9.9)

Cuadro 4.2: **Cantidad de *clusters* detectados según su tamaño, para los 3 conjuntos de datos (PD, PP y IP).** Los números entre paréntesis corresponden al porcentaje sobre el total de cada columna.

Tamaño del *Cluster*	Proteína-Droga	Proteína-Proteína	Intraproteina
2	1.00	1.00	1.00
3	2.26	2.41	2.14
4	3.82	4.07	3.50
5	5.45	5.92	4.90
6	7.31	7.57	6.41
7	9.15	9.58	8.03
8	11.17	11.18	9.71

Cuadro 4.3: **Número promedio de interacciones según el tamaño del *cluster*, para los 3 conjuntos de datos (PD, PP y IP).**

En el caso de IP, esta relación pareciera ser exponencial dado que la cantidad de *clusters* de un tamaño dado es alrededor de la mitad que los *clusters* con un anillo menos. Lo mismo sucede en el caso de PD. Pero para PP, esta relación se pierde ligeramente, apreciandose un decremento más pronunciado a medida que el tamaño del *cluster* aumenta, indicando que los *clusters* de mayor tamaño

en PP son menos frecuentes que para IP y PD. Los *clusters* diméricos son mas abundantes en PP que PD, llegando al 58.6 % y 47.5 %, respectivamente. Por otro lado, los *clusters* triméricos, llegan al 25.2 % en PD y 23.3 % en PP, tetrámeros 13.4 % y 10.1 %, y así, los *clusters* más grandes, son más en PD que en PP. Otra diferencia es que el número promedio de interacciones (Cuadro 4.3) para cada *cluster* es mayor en el caso de PP que en PD. Esto es una tendencia general, a los largo de los diferentes tamaños de *cluster*, la cantidad promedio de interacciones es mayor en PP que en PD. Lo que implica que los complejos PP tienen una tendencia a formar *clusters* más compactos.

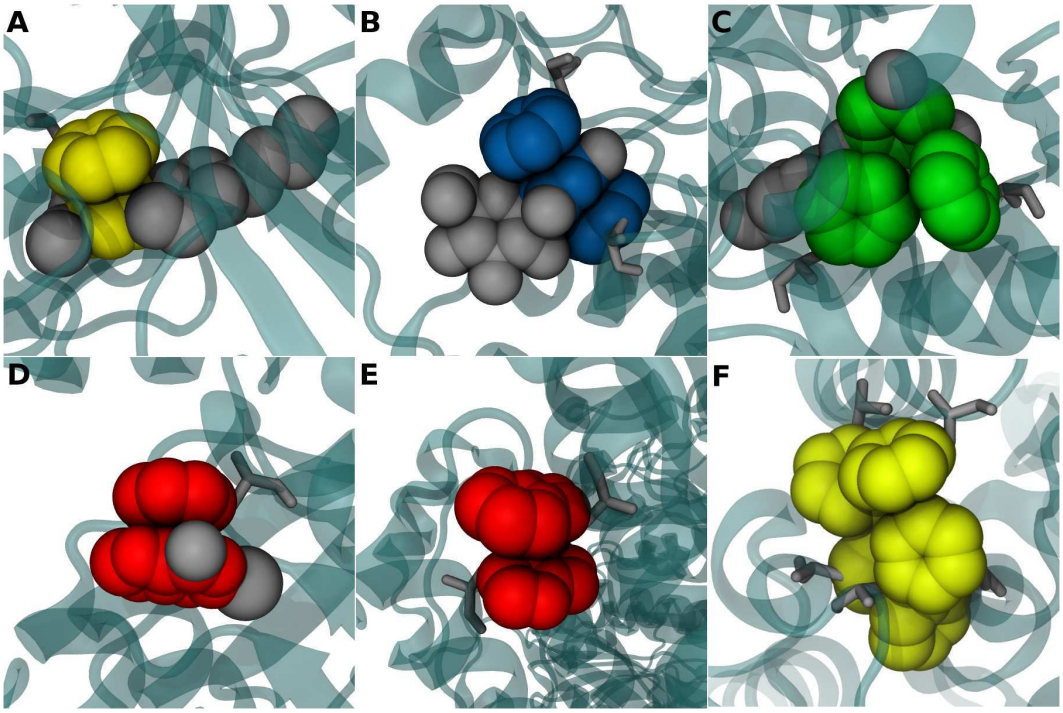

Figura 4.5: ***Clusters* de aromáticos en regiones de interacción Proteína-Droga y Proteína-Proteína.** A) Un *cluster* dimérico formado por una fenilalanina y un anillo aromático en una droga, `PDB 3oc0`. B) Un trímero con 2 interacciones formado por dos anillos aromáticos en residuos de la proteína "ensanguchando" un anillo en una droga, `PDB 3bda`. C) Trímero con 3 interacciones formado por dos anillos en residuos interactuando entre sí y un tercer anillo con una droga, `PDB 1cgk`. D) Un *cluster* de 3 anillos formados por un anillo en un residuo de la proteína y dos anillos aromáticos que conforman un sistema fusionado en una droga, `PDB 1m67`. E) Interacción entre los triptofanos en una interfaz de interacción proteína-proteína, `PDB 3pgx`. F) Un pentámero en una interfaz proteína-proteína en dónde hay 2 anillos de una cadena y 3 anillos de la otra, `PDB 2aml`.

Vimos que los *clusters* de aromáticos se encuentran en gran abundancia formando parte de los plegados de las proteínas. Pero, como mostramos en el Cuadro 4.2, los anillos aromáticos también pueden formar *clusters* en las regiones de interacción con otras proteínas y con ligandos pequeños tipo droga, como muestra la Figura 4.5. En estos casos los anillos aromáticos establecen interacciones aromáticas con las otras moléculas como se ve en las Figuras 4.5A y 4.5B donde se aprecian un dímero y un trímero con 2 interacciones. Además, hay *clusters* en donde los anillos que interactúan con otras moléculas también lo hacen con otros anillos de la misma proteína como se ve en la Figura 4.5C, un trímero con 3 interacciones. También, pueden estar definidos por sistemas de anillos fusionados en drogas (Figuras 4.5D) y formando parte de los sitios de interacción entre proteínas de varios tamaños (4.5E y 4.5F).

Composición de los *clusters* de aromáticos En el caso de IP, el análisis de la composición aminoacídica de los trímeros muestra que el porcentaje de Phe esta entre 45 % y 50 %, para Tyr es entre 31 % y 39 % y para Trp, entre 15 % y 17 %, que son proporciones acorde a las percibidas por las abundacias naturales de estos amino ácidos. A la hora de analizar los tipos de interacciones en los diferentes *clusters*, calculamos cuantos dímeros se forman con cada par de amino ácidos. Los resultados correspondientes son: Phe-Tyr, 31 %; Phe-Phe, 26 %; Phe-Trp, 15 %; Tyr-Tyr, 11 %; Tyr-Trp, 9 %; and Trp-Trp, 3 %. Las interacciones mas abundantes son entre Phe y Tyr, seguidas de cerca por aquellas entre Phe y Phe. Curiosamente, Phe y Trp interactúan con más frecuencia que Tyr y Tyr, y que Tyr y Trp, esto probablemente se deba al ligero grado de hidrofilicidad que presenta Tyr dado su grupo polar OH. Las interacciones entre Trp y Trp ocurren con muy baja frecuencia, probablemente como consecuencia de la baja abundancia natural de los Trp en proteínas, no más del 1.5 % basandose en Uniprot.

Para PD y PP, como se ve en los Cuadros 4.4 y 4.5, la distribución de amino ácidos por *cluster*, también parece seguir las abundancias naturales, sólo que se puede apreciar una ligera preferencia por los Trps en PP con respecto a PD. Esto está en acuerdo con lo evidenciado por [Ma et al., 2003], que los Trps expuestos suelen ser una fuerte evidencia de una interfaz proteína-proteína. Otro hecho interesante es que a medida que el tamaño del *cluster* aumenta, esta preferencia de los *clusters* en PP por tener más Trps, se vuelve cada vez más marcada, lo cuál se debe a que cada Trp es más grande que el resto de los amino ácidos, como para generar más superficie de contacto, y por ende, participar de *clusters* de más tamaño. Mientras que la aparición de un Trp o más en bolsillos de unión a ligandos tipo droga (PD), puede no estar favorecida

dado que presetan superficies de interacción, en promedio, más chicas.

Residuo	Dímeros	Trímeros	Tetrámeros	Pentámeros
Phe	2063 (42.1)	1527 (37.2)	1135 (36.0)	642 (32.6)
Tyr	1541 (31.5)	1207 (29.4)	809 (25.7)	462 (23.5)
His	989 (20.2)	622 (15.2)	493 (15.6)	262 (13.3)
Trp Bnz	215 (4.4)	465 (11.3)	425 (13.5)	346 (17.6)
Trp Prl	89 (1.8)	283 (6.9)	290 (9.2)	255 (13.0)

Cuadro 4.4: **Cantidad de apariciones de los 4 residuos en los *clusters* detectados en regiones de interacción Proteína-Droga.** Trp Bzn y Prl corresponde al benceno y al pyrrol que conforman el sistema indole del Triptofano. *Los números entre paréntesis corresponden al porcentaje sobre el total de cada columna.*

Residuo	Dímeros	Trímeros	Tetrámeros	Pentámeros
PHE	5428 (39.2)	2888 (35.0)	1515 (31.5)	686 (30.3)
TYR	4470 (32.3)	2030 (24.6)	1050 (21.9)	457 (20.2)
HIS	3022 (21.8)	1225 (14.9)	692 (14.4)	286 (12.6)
TRP Bnz	682 (4.9)	1130 (13.7)	825 (17.2)	438 (19.4)
TRP Prl	249 (1.8)	975 (11.8)	716 (14.9)	393 (17.4)

Cuadro 4.5: **Cantidad de apariciones de los 4 residuos en los *clusters* detectados en regiones de interacción Proteína-Proteína.** Trp Bzn y Prl corresponde al benceno y al pyrrol que conforman el sistema indole del Triptofano. *Los números entre paréntesis corresponden al porcentaje sobre el total de cada columna.*

Análisis del entorno de los *clusters* Estudiamos las características del entorno estructural en el que encontramos cada *cluster* en el interior de las proteínas. Para cada cluster en IP, calculamos el grado de hidrofobicidad medio del entorno y la posición relativa con respecto al centro de la proteína. Luego, para PD y PP, analizamos el porcentaje de area expuesta y el porcentaje de contacto que presentan los anillos de los residuos que conforman cada *cluster*. Los valores calculados son:

- **% Expuesto** y **% en contacto**: Por un lado, como fue mencionado en el Capítulo 3, para cada residuo, es posible calcular el porcentaje de área expuesta, teniendo en cuenta el resto de los átomos involucrados en la proteína. Se calcula dividiendo el área accesible al solvente de cada residuo en el contexto de la proteína por el área total de superficie del residuo aislado. Por otro lado, para los casos de complejos PD y PP, calculamos el porcentaje en contacto de cada residuo como el valor de

superficie que queda oculto en presencia del otro interactor (la droga o la otra proteína) con respecto al área expuesta del residuo si no estuviera el interactor.

- **CCDR**: Es la posición relativa del *cluster* con respecto al centro de la proteína en relación al tamaño general de la proteína. Se calcula dividiendo la distancia del centro de masa del *cluster* al centro de masa de la proteína por el *radio de giro* de la misma. Valores cercanos a 0 indican que el *cluster* se encuentra cerca del centro de la proteína, valores menores 1 indican que el *cluster* está en el interior y valores mayores a 1 indican que el *cluster* está posicionado en la superficie de la proteína.

- **EHI**: Es el índice de hidrofobicidad del entorno, se calcula obteniendo el promedio del valor del índice de hidrofobicidad ([Kyte and Doolittle, 1982]) sobre todos los residuos interactuando con el *cluster*. Los residuos interactuando se definen como aquellos residuos con el centro de masa de la cadena lateral a menos de 8Å de la cadena lateral de cualquier residuo aromático entre los residuos del *cluster*.

Con respecto al % expuesto (Figura 4.6), los resultados muestran que los clusters en IP están bastante poco expuestos o mejor dicho, enterrados dentro de las estructuras proteicas dado que la gran mayoría de ellos están expuestos como mucho un 10 % y un muy pequeño porcentaje tiene valores mayores al 20 %. Los anillos aromáticos en *clusters* PD tienen una superficie expuesta promedio de mayor tamaño que los que se encuentran en IP alcanzando su valor máximo en la distribución alrededor del 10 % y superando a la distribución de los anillos en IP a partir del 5 % de exposición. Esto quiere decir que un residuo aromático tomado al azar con un % expuesto mayor al 5 %-10 % tiene más chances de ser un residuo interactor (tanto con drogas como otras proteínas) que un residuo involucrado solamente en el plegado de la estructura proteica. Los *clusters* en PP, tienen esta tendencia aún más marcada presentando pocos valores menores al 10 % de exposición y superando a la distribución de PD a partir del 20 % de exposición. Esto quiere decir que el porcentaje de exposición de los anillos aromáticos es un parámetro que explica gran parte de la función que cumple dicho anillo en la estructura proteica.

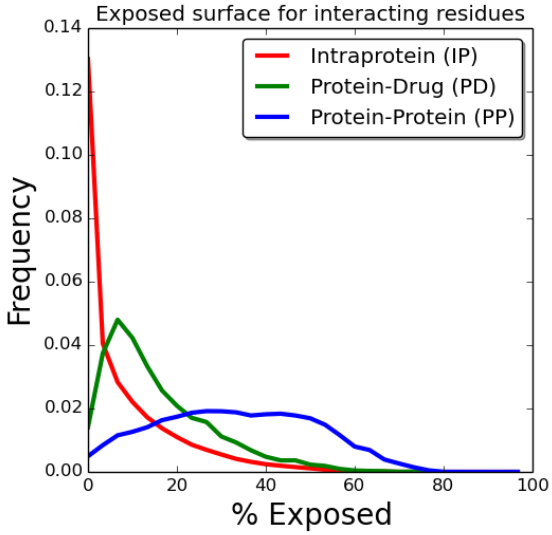

Figura 4.6: **Accesibilidad al solvente en *clusters* de aromáticos.** Distribución de porcentaje de superficie expuesta según el tipo de *cluster*.

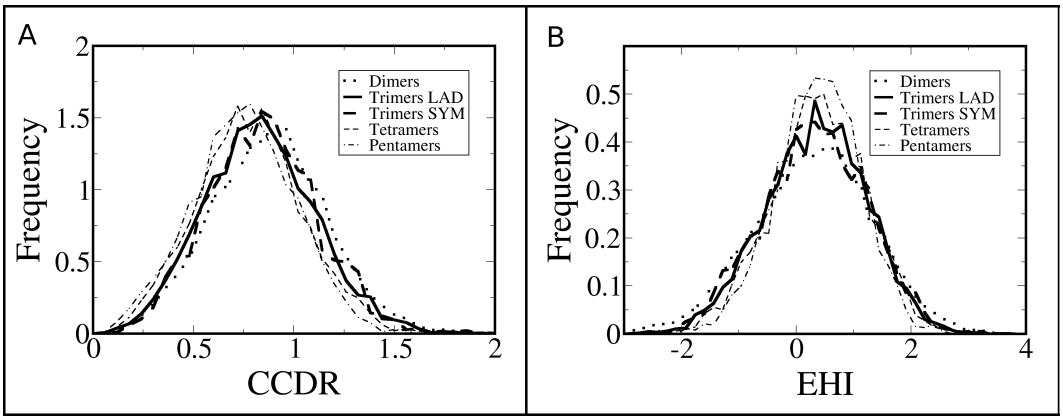

Figura 4.7: **Análisis del entorno de los *clusters* aromáticos en IP.** A) CCDR expresa la posición relativa de cada *cluster* con respecto al centro de la proteína. B) EHI expresa el promedio de hidrofobicidad de los residuos del entorno de cada *cluster*, basado en el índice de hidrofobicidad de [Kyte and Doolittle, 1982].

Como se puede ver en la Figura 4.7A, los resultados indican que la mayoría de los *clusters* se encuentran en el interior de la proteína (valores de $CCDR < 1$), mostrando que los *clusters* en IP tienden a formar parte del núcleo hidrofóbico de la proteína. Estos resultados son los mismos para trímeros LAD y SYM, que serán explicados a continuación. En resumen, ambos resultados muestran que los *clusters*, como era de esperarse, están enterrados en el interior de las proteínas formando parte de su núcleo. Sin embargo, el resultado interesante es que los clusters en PD se encuentran mas hacia el interior de la proteína que en PP. Si bien este resultado se podría esperar porque los anillos en las drogas sueltas pueden encajar mejor hacia el interior de la proteína que lo que lo hacen los anillos en otra proteína, es interesante ya que la tendencia es muy marcada.

Para analizar la naturaleza del entorno de cada *cluster* en IP, computamos para cada uno de ellos, el EHI como lo definimos previamente para medir el grado de hidrofobicidad promedio de los residuos vecinos circundantes. Valores muy negativos corresponden a ambientes hidrofílicos (el índice de la ARG es $-4{,}5$) y valores muy positivos corresponden a residuos muy hidrofóbicos (el correspondiente para Leu es 4,5). Otro dato a tener cuenta, es que la hidrofobicidad promedio ponderada por la abundacia de los amino ácidos en el PDB es $-0{,}26$. Los resultados muestran que en general el entorno de los clusters de aromáticos es mas bien hidrofóbico, alcanzando un valor promedio de EHI de alrededor de 0,3, como se puede ver en la Figura 4.7B.

Ahora analizando los clusters en PD y PP, en la Figura 4.8 vemos la relación entre la superficie expuesta de cada residuo que forma parte de los *clusters* y la superficie de contacto con la molécula interactora, la otra cadena proteica en el caso de PP y la droga en el caso de PP. Se puede ver que los residuos invulocrados en los *clusters* PP están, en promedio, más expuestos que los *clusters* en PD. Esto quiere decir que, dado un residuo aromático, la superficie expuesta brinda información acerca del rol funcional que está jugando dicho residuo. Además, el porcetaje promedio de contacto también es más grande en PP que en PD. Teniendo en cuenta que la superficie de contacto es uno de los principales parámetros que permiten modelar las interacciones hidrofóbicas (a mayor superficie en contacto, mayor es la exclusión de las moléculas de agua), estaría indicando que los *clusters* de aromáticos en PP son más estables energéticamente que los que encontramos en PD.

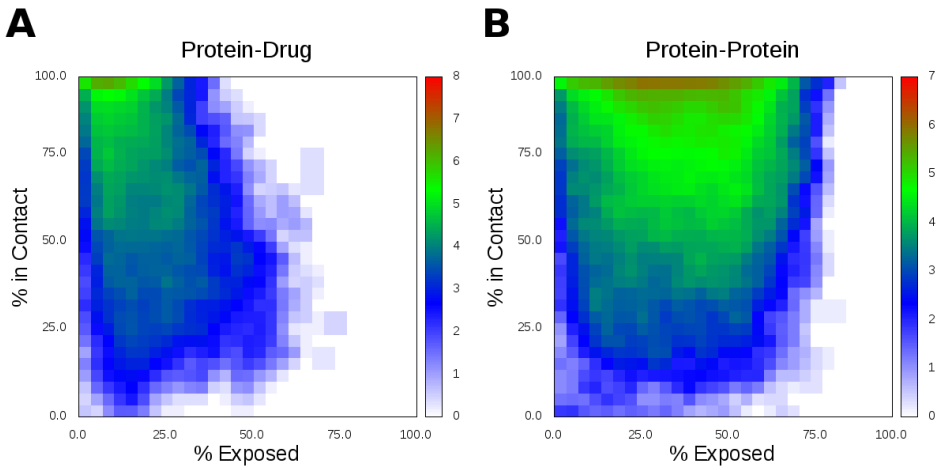

Figura 4.8: **Porcentaje de superficie expuesta al solvente y porcentaje de superficie de contacto.** A) Porcentajes de superficie expuesta y de contacto para *clusters* en Proteína-Droga. B) Porcentajes de superficie expuesta y de contacto para *clusters* en Proteína-Proteína.

4.3.3. Los trímeros de aromáticos en estructuras de proteínas y sus propiedades geométricas.

El primer *cluster* posible (más allá del dímero) es el trímero aromático. Optimizaciones ab initio en vacío muestran que los trímeros de benzenos prefieren conformaciones principalmente de 2 tipos geométricos: uno triangular simétrico (SYM, i.e.: triángulos equilateros) con las tres distancias de interacción similares entre los centros de masa de los anillos aromáticos así como tres ángulos conformacionales similares (alrededor de los 60°), y otro de tipo *apilado* (LAD, i.e.: como los escalones de una escalera) con una de las distancias entre los centros de masa mas grande que las otras dos. [Gonzalez and Lim, 2001], [Tauer and Sherrill, 2005], [Engkvist et al., 1999]. Buscamos todos los trímeros en el conjunto de estructuras seleccionadas computando las propiedades geométricas de todos los *clusters* de aromáticos, como fue descripto en la sección de métodos. Curiosamente, encontramos que los trímeros de anillos aromáticos en proteínas repiten fuertemente las estructuras previamente mencionadas, lo que permite asignarlos en dos categorías:

LAD Trimeros que presentan un anillo central que interactúa con los otros dos anillos que no están interactuando entre sí porque están demasiado lejos y por esto tienen un ángulo conformacional y una distancia entre los centros

de masa que son más grandes. Estos trímeros tienen 2 interacciones, como se muestra en la Figura 4.9A.

SYM Aquellos trímeros donde los tres anillos están formando un triángulo con simetría circular donde cada anillo aromático está en contacto con los otros dos y, por esto tiene, los tres ángulos conformacionales y las tres distancias entre los centros de masa, similares. Estos trímeros tienen 3 interacciones, como se muestra en la Figura 4.9B.

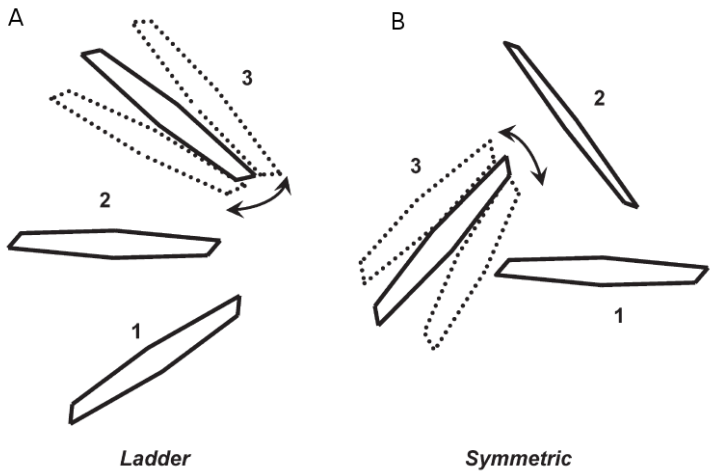

Figura 4.9: **Trímeros de aromáticos: El primer *cluster* más allá del dímero.** A) Trímero tipo LAD: con sólo 2 interacciones establecidas entre los anillos aromáticos que lo conforman. B) Trímero tipo SYM: Con 3 interacciones establecidas, da lugar a un *cluster* más compacto.

En el contexto de los complejos proteína-droga, podemos encontrar dos tipos de trímeros arómaticos: los DRR que están formados por un anillo de la droga y 2 de la proteína y los DDR que tienen 2 anillos de la droga y 1 de la proteína. En este trabajo, nos dedicaremos sólo al caso de los DRR para entender como es que los anillos de las drogas encajan entre los anillos de las proteínas.

Análisis estructural de trímeros en IP Teniendo en cuenta la clasificación de trímeros por su estructura geométrica, analizamos su impacto en las estructuras de las proteínas y clasificar sus propiedades. Para determinar el impacto

de los trímeros en las estructuras disponibles, calculamos el porcentaje de estructuras de proteínas que tengan un trímero de cada tipo. Si tomamos que los clusters de tamaño mayor a 3 están compuestos por subtrímeros, más de la mitad de las estructuras tienen como mínimo un trímero. Más importante aún, 80 % de todos los dímeros son parte de un trímero, y casi la mitad de los trímeros son parte de un tetrámero, mostrando la clara tendencia a incrementar su tamaño. Con respecto a cada tipo de trímero, los de tipo LAD son claramente el tipo más abundante de trímero llegando a un 78 % de los mismos. Estos resultados son llamativos ya que hasta ahora no se habian caracterizado *clusters* en estructuras de proteínas salvo dímeros.

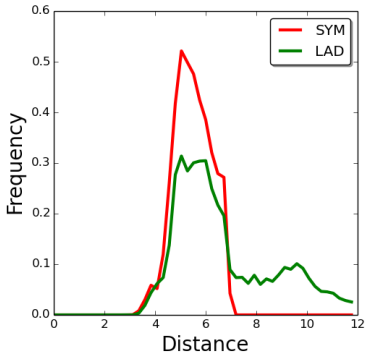

Figura 4.10: **Distribución de distancias en trímeros aromáticos en IP.** En rojo, las distancias entre los centros de masa de los anillos aromáticos pertenecientes a trímeros SYM. En verde, la distribución para las distancias en trímeros LAD.

Para analizar las configuración en las que se encuentran los trímeros, analizamos primero la distribución de distancias que hay entre los centros de masa de los anillos aromáticos (Figura 4.10). Para ambos tipos de trímeros (SYM y LAD) se puede ver que las distribuciones tiene una componente importante que corresponde a las distancias entre aquellos residuos que están interactuando. En el caso de los trímeros SYM, serían las tres interacciones presentes, y en el caso de los trímeros LAD, corresponde a las únicas dos interacciones. Además, los trímeros LAD, presentan una componente extra que correspende a la distancia de los dos residuos que no interactúan, la cuál alcanza su valor máximo entre 9Å y 10Å.

Para los SYM, la distribución de distancias de éstos trímeros (Figura 4.10), presenta un valor promedio de 5.8Å y un desvío estandard de 1.05Å. Sepa-

ramos las tres distancias en la mínima, media y máxima distancias, y las tres distribuciones se solapan, pero presentan picos separados en 5.0Å, 6.0Å y 6.5Å, mostrando la tendencia de los trímeros SYM a ser mas compactos pero no del todo simétricos (Figura 4.11A). Cabe destacar, que la falta de simetría observada para los trímeros SYM en proteínas probablemente se pueda explicar por las restricciones en los grados de libertad impuestos por el resto de la estructura proteica, pero, a pesar de esto, de todas formas adoptan las mismas conformaciones que están descriptas como los mínimos energéticos para los sistemas de trímeros de benzenos en vacío.

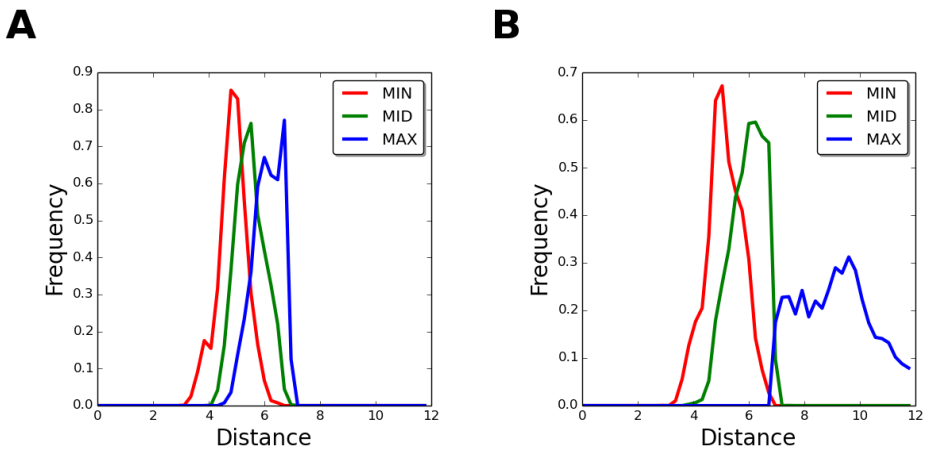

Figura 4.11: **Diferencias entre las tres distancias en trímeros.** A) Distancias mínima (rojo), media (verde) y máxima (azul) para trímeros SYM. B) Idem trímeros LAD.

En el caso de los trímeros LAD (Figura 4.11B), las distancias entre los centros de masa alcanzan, como era de esperarse, una distribución de dos componentes. Una similar a la reportada previamente para los trímeros SYM con un promedio en 5.7Å correspondiente a la distribución de distancias del anillo central contra los otros dos anillos con los que está interactuando, y la otra que representa la distancia entre los residuos que no interactúan con un valor promedio de 10Å.

Caracterizamos la distribución de los ángulos conformacionales definidos como los ángulos formados por los tres centros de masa de cada anillo aromático, dando lugar a tres ángulos por cada trímero como indica la Figura 4.12A, dónde se esquematiza un anillo en una droga (D) interactuando con otros dos anillos en la proteína (R).

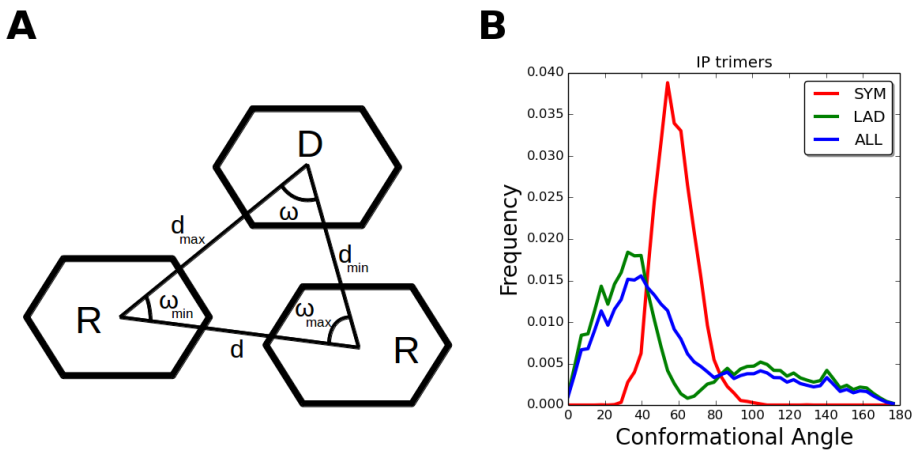

Figura 4.12: **Ángulos conformacionales de trímeros en estructuras de proteínas.** A) Esquema de los ángulos dados tres anillos. Fijando un anillo, por un lado se define una distancia entre los otros dos anillos y su correspondiente ángulo conformacional y por otro lado, quedan fijados dos ángulos (uno más chico ω_{min} y otro más grande ω_{max}) con sus distancias correspondientes (d_{min} y d_{max}). B) Distribución de los ángulos conformacionales de trímeros en IP

La distribución de dichos ángulos, para el caso de los trímeros intraproteína (Figura 4.12B) muestra que los trímeros SYM presentan una componente alrededor de los 60°, desde 40° hasta 80°, corroborando la naturaleza simétrica de los trímeros SYM mencionada previamente. Los ángulos conformacionales para el caso de los trímeros LAD, presentan una distribución ancha de dos componentes lo cuál coincide con las dos componentes de la distribución de distancia. La primer componente que va desde 0° hasta 60° corresponde con los 2 anillos que están flanqueando al central y no interactúan entre sí. La segunda componente, que corresponde al anillo central es más amplia y barre desde 60° hasta 180°. Curiosamente, ningún LAD presenta una conformación *perfecta* (el anillo central con 180° o los dos anillos que flanquean con 0°). Comparativamente, el ángulo conformacional para el caso de los dos anillos exteriores en éstos trímeros presentan valores más bajos que en los trímeros SYM.

Análisis estructural en trímeros en PD y PP Comparando los *clusters* en PD y PP, los ángulos conformacionales de los trímeros SYM entre los distintos tipos (PD-SYM y PP-SYM), presentan una distribución de una sola componente alrededor de 60° corroborando la naturaleza simétrica de los trímeros SYM co-

mo en el caso de IP. En el caso de los trímeros LAD (PD-LAD y PP-LAD), los ángulos conformacionales están distribuidos en dos componentes una alrededor de 20° que corresponde a los 2 anillos presentes en los residuos de la proteína que están interactuando con el anillo en la droga que esta representado por la segunda componente que toma valores entre 100° y 160° con el máximo en 140°. En este caso, las distribuciones de PD-LAD y PP-LAD son ligeramente diferentes, se puede apreciar un desplazamiento en ambas componentes de la distribución. En los trímeros PP-LAD (Figura 4.13A), la distribución tiene los dos máximos alrededor de 40° y 120°, mientras que en el caso de los PD-LAD (Figura 4.13B), estos dos máximos se encuentran alrededor de 20° y 140°, respectivamente. Esto significa que, en PD-LAD, los dos anillos en los residuos que interactuan con el tercero en la droga, están mas separados entre ellos comparado con los presentes en PP-LAD, donde los dos anillos pertenecientes a la misma cadena proteica, estarían mas cercanos entre sí mientras hacen contacto con el tercer anillo en la otra cadena.

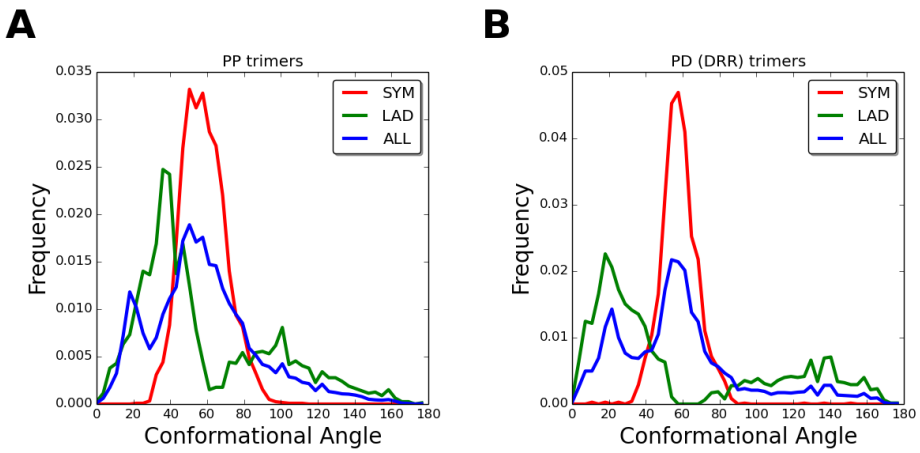

Figura 4.13: **Ângulos conformacionales en trímeros presentes en regiones de interacción.** A) Para trímeros en Proteína-Proteína. B) Para trímeros en Proteína-Droga.

Hemos analizado las propiedades geométricas de los trímeros porque permiten entender cómo se dan las interacciones entre *clusters* más allá del dímero. En particular los ángulos conformacionales de los trímeros dan información acerca de cómo un anillo aromático se ancla contra otros dos anillos dispuestos para interactuar con él. Un primer paso a la hora de entender este fenómeno es definir cuál de los tres anillos es el foráneo. Sería aquel que interactúa con los

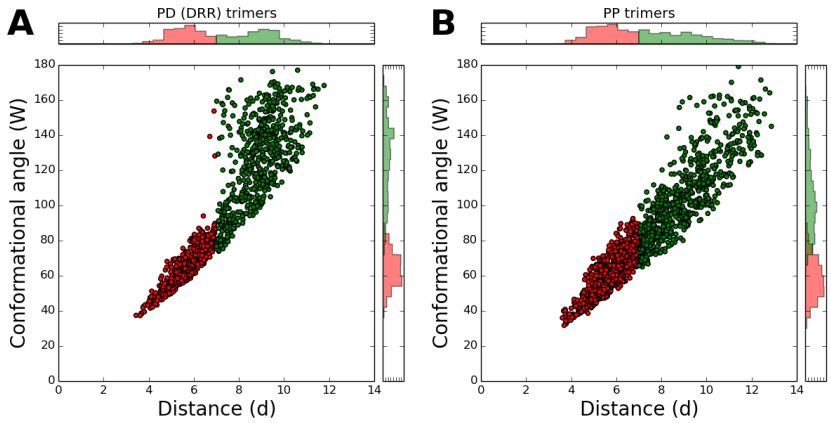

Figura 4.14: **Relación entre las distancias y los ángulos conformacionales.** Dependencia del ángulo conformacional del anillo foráneo dada la distancia entre los dos residuos en la misma cadena proteica. A) Para trímeros en Proteína-Droga. B) Para trímeros en Proteína-Proteína.

dos que pertencen a la misma molécula. Para los trímeros en PD, analizamos el caso de los DRR en los que uno de los anillos es de la droga y los otros dos de la proteína. En el caso de los trímeros PP distinguimos el anillo foraneo de los otros dos por el código de la cadena a la que pertenece. Tomando como ejemplo el caso de las drogas, centradas en el anillo D, definimos su ángulo conformacional (ω) y la distancia (d) entre los dos anillos R de los residuos, como se muestra en la Figura 4.12A. Encontramos que hay una correspondencia entre ω y d que depende del tipo de trímero.

Tanto para los trímeros LAD como SYM, se puede ver que ω aumenta a medida que aumenta d. Pero, en el caso de los trímeros PP (Figura 4.14B) se puede ver la misma relación, con la diferencia de que ésta, está más restringida en los valores que puede adoptar el ángulo para una determinada distancia. Además, en el caso de PD, ω llega a valores de 160° con d entre 7Å y 8Å, mientras que, en el caso de PP, los valores de ω llegan por encima de 160° recién para valores de d por encima de 8Å. Analizando el ángulo conformacional del anillo foráneo, hemos visto que tiene una distribucion con dos máximos para el caso de PD, lo cuál no es así para el caso de PP. Como se mencionó previamente, las interacciones aromáticas en los complejos PP tienen restricciones en sus grados de libertad, probablemente asociadas, al hecho de que los anillos no están libres sino unidos covalentemente al backbone de la proteína. Sin embargo, en el caso de PD, donde éstas restricciones sobre el anillo forá-

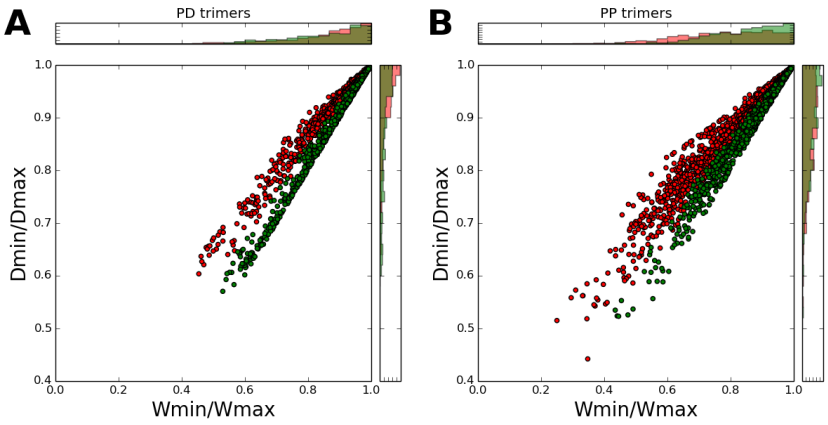

Figura 4.15: **Relación entre los ratios de las distancias máximas y mínimas, y los ángulos conformacionales máximos y mínimos (permite visualizar la preferencia del anillo foráneo por los otros dos anillos).** A) Para trímeros en Proteína-Droga. B) Para trímeros en Proteína-Proteína.

neo no están presentes, aparecen los dos máximos en la distribución de ω que corresponden con las dos conformaciones (LAD y SYM) que están reportadas como las más favorables para el sistema en vacío.

En la Figura 4.15, vemos las relaciones entre la máxima y la mínima distancia establecidas por el anillo foráneo contra los otros dos anillos en la misma cadena proteica. Estas relaciones indican que estos trímeros están principalmente distribuídos equidistantes entre sí, aunque hay algunos casos en los que esta relación es preferencial por uno de los anillos. Este efecto es menor en los trímeros PP que en los PD. Esto probablemente se deba al hecho de que el mínimo energético de estos sistemas se encuentra en el estado equidistante y debido a que las configuraciones en las que se encuentran los trímeros PP tienen más restricciones en los grados de libertad, estos estados son, en promedio, más difíciles de alcanzar. Por útlimo y para pasar a hablar de clusters en general, nos preguntamos: ¿podemos definir la posición de un anillo en una droga? Utilizando la relación entre la distancia y el ángulo conformacional es posible ubicar, entre dos anillos aromáticos dados, el sitio donde encajaría un tercer anillo. Por lo tanto, es posible definir un sitio de posicionamiento ideal de anillos aromáticos, usando sólo la distancia entre los anillos que flanquean el sitio, debido a que el ángulo en el cual se posicionaría un potencial anillo depende de ésta. A su vez, si se da el caso de que más de dos anillos definen un mismo sitio, aumentan las chances de que ahí se estabilice un anillo foraneo.

Así, este razonamiento se puede extrapolar a más de dos, si más anillos están dispuestos de forma tal que definen un sitio, entonces este sitio tiene aún más chances de unir y estabilizar un anillo foraneo.

Estructura secundaria en IP Usamos los trímeros en IP para estudiar cómo están involucrados en motivos estructurales, que a menudo están relacionados a elementos específicos de estructura secundaria. Usamos DSSP [Kabsch and Sander, 1983] para asignar tres tipos de estructura secundaria (E: hoja Beta, H: Hélice alfa y L: Loop) a los que pertenecen los interactores involucrados en cada trímero. Clasificamos cada trímero en IP en grupos según estos tres elementos estructurales, como muestra la Figura 4.16A.

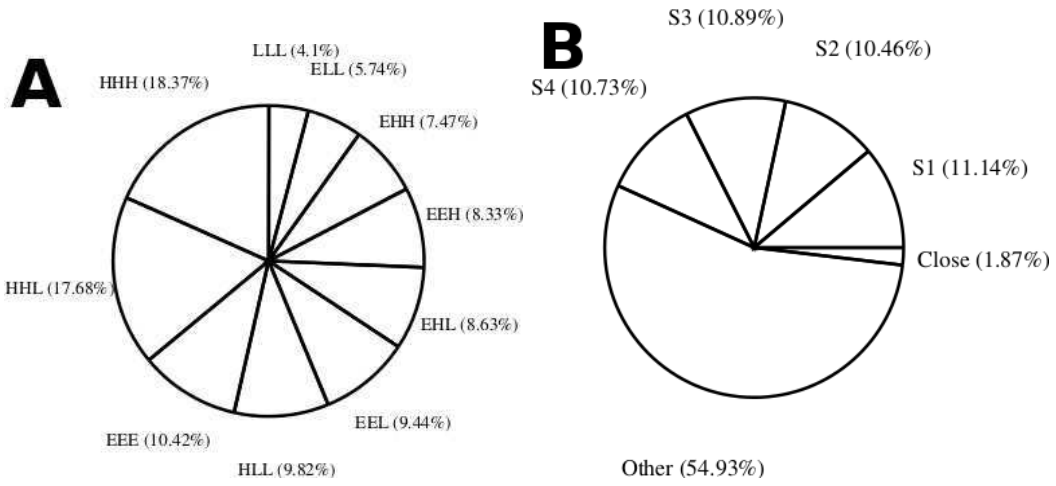

Figura 4.16: **Proporción de trímeros según estructura secundaria y distancia en secuencia.** A) Clasificación de trímeros según su estructura secundaria representando con cada letra la combinación de posibilidades para cada trímero: *H = Hélice alpha, E = Hoja beta, L = Loop.* B) Clasificación de trímeros según su distancia en la estructura primaria (secuencia) de la proteína: *close* con los tres residuos del trímero separados no más de 5 posiciones, *s1-s4* con dos residuos separados por 1 a 4 posiciones en secuencia, y *other* aquellos que los tres residuos están separados más de 5 posiciones de cada uno de los otros dos.

Los datos muestran que el grupo más abundante tiene tres residuos en hélice alfa (HHH, con 18 %) y si tomamos aquellos agrupados en por lo menos dos de los residuos interactores perteneciendo a hélice alfa (HHX, con X siendo cualquiera de las tres clases) alcanzamos a cubrir un 43.5 % de los casos detectados. Por otro lado, los trímeros clasificados como EEX representan todos

juntos, alrededor de un 28 %. Estos resultados muestran la clara preferencia por los trímeros de pertenecer a estructuras helicoldales. Esto es todavía más interesante si se tiene en cuenta en la preferencia de los amino ácidos aromáticos por aparecer más comunmente en hojas beta. [Malkov et al., 2008]

Otra pregunta interesante que nos podemos hacer sobre los trímeros en IP, es si la distribución en secuencia de amino ácidos que forman trímeros de aromáticos tiene alguna tendencia en particular. Para analizar esto, calculamos las distancias en secuencia que están dadas por las posiciones de los residuos aromáticos en la cadena proteica. Clasificamos los trímeros según el patrón de la distribución de las distancias en secuencia, en las siguientes categorías: *cercanos* aquellos que tienen los tres residuos a no más de 5 posiciones en secuencia entre los tres; *s1, s2, s3 y s4*, aquellos que tienen por lo menos 2 de los residuos participantes del *cluster* a 1-4 de distancia en secuencia y que no pertenecen a los *cercanos*; y por último, aquellos que tienen los tres residuos separados por más de 5 posiciones en la secuencia de la proteína, fueron denominados *otros*.

La distribución de los trímeros en estas categorías se muestran en la Figura 4.16B. No se encontraron diferencias significativas entre ambos tipos de trímeros (LAD y SYM). Más de la mitad de los trímeros pertenecientes a ambos tipos geométricos, caen en la categoría de *otros*, mientras que cada uno de los grupos *s1-s4* alcanza cada uno un 10 % en promedio. Queda claro que hay una gran cantidad (casi la mitad) de trímeros aromáticos que están compuestos por subdímeros cercanos en secuencia y un tercero no cercano (a distancia 5 o más). Finalmente, aquellos clasificados como *cercanos*, alcanzan un 2 % del total de los trímeros, marcando la tendiencia de los trímeros a ser no locales, a diferencia de los dímeros, donde el grupo *cercanos* llega hasta el 21 %. Hemos encontrado que los trímeros son no locales en secuencia, lo que significa que tienden a integrar diferentes elementos estructurales distantes siendo relevantes para la estabilidad y el plegado de las proteínas. También clasificamos los tetrámeros según su distribución de distancias en secuencia de los residuos que conforman el *cluster*. Como con los trímeros, los agrupamos en *cercanos* siempre que los 4 residuos estén separados todos entre sí por distancias en secuencia menores a 4, y como *s2*, *s3*, y *s4*, cuando los residuos pertenecen a 2, 3 o 4 regiones diferentes, definiendo que 2 residuos pertenecen a la misma región de la secuencia de la proteína cuando están separados a lo sumo 4 posiciones entre sí. Notablemente, la mayoría de los *clusters*, juntan 3 o hasta 4 regiones de la proteína, denotando la naturaleza no local de estos *clusters* en la estabilización de la estructura terciaria de las proteínas.

SYM

$HHX - close$	2,0	$EEX - close$	0,0
$HHX - i + (1,3,4)$	$1{,}5 - 1{,}7$	$EEX - i + (1,3,4)$	0,2
$HHX - i + 2$	0,1	$EEX - i + 2$	2,2

LAD

$HHX - i + (1,3,4)$	$1{,}4 - 1{,}7$	$EEX - i + (1,3,4)$	$0{,}2 - 0{,}6$
$HHX - i + 2$	0,1	$EEX - i + 2$	2,1

Cuadro 4.6: **Puntajes de disparidad para la relación entre la estructura secundaria y la distancia en secuencia.** Cada puntaje se calcula como la división entre la probabilidad de que un trímero pertenezca a los dos grupos a la vez, y las probabilidades de que ambas clasificaciones se hayan dado de forma independiente (i.e.: $HHX - close = P(HHX \cap close)/(P(HXX) * P(close))$). Valores mayores que uno implican una preferencia por pertenecer a ambas clasificaciones, mientras valores menores que uno sugieren que ambas clasificaciones se evitan. *Sólo se muestran los casos en los que se ven diferencias significativas.*

Nos propusimos analizar las relaciones que hay entre las clasificaciones de estructura secundaria y primaria, como se muestra en el Cuadro 4.6 calculando los puntajes de disparidad para los dos grupos de clasificaciones de trímeros. Se puede ver que en ambos tipos de trímeros, LAD y SYM, el grupo HHX muestra una clara preferencia por los grupos s1, s3 y s4, mientras que rara vez se lo ve en s2. Esto probablemente refleja el hecho de que los anillos cercanos en secuencia están en estructura helicoidal y para estar en contacto deben aparecer consecutivos (s1) o separados por lo menos por 3 residuos, dado que un giro de alfa hélice se da cada 3.6 residuos. Por el contrario, los trímeros EEX muestran una clara preferencia por el grupo s2 sobre los demás grupos, lo cuál refleja el hecho de que en hojas beta las cadenas laterales se alternan apuntando en la misma dirección cada 2 residuos. Los trímeros SYM que pertenecen al grupo *cercanos* se encuentran principalmente en HHX y muy rara vez en EEX. Estos resultados muestran que los grupos de estructura primaria y secundaria están sesgados debido a las propiedades estructurales intrínsecas de los elementos de estructura secundaria. Todo este análisis puede ser de utilidad para clasificar familias de proteínas que tengan *clusters* de aromáticos como motivos estructurales relevantes para su función.

Estructura secundaria en PD y PP Una vez analizadas las tendencias en estructura secundaria de los *clusters* en IP, nos queda preguntarnos como son en PD y PP.

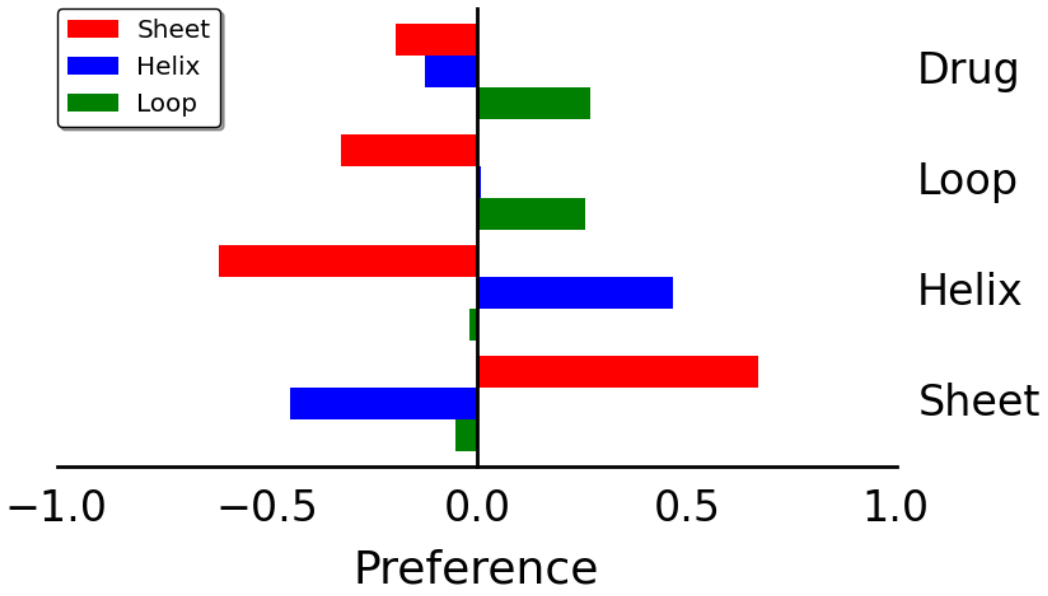

Figura 4.17: **Preferencias por estructura secundaria en *clusters* PD y PP.** El valor de preferencia se calcula dividiendo la probabilidad de encontrar un determinado residuo aromático interactuando con otro en una estructura secundaria, por la probabilidad de que ese anillo esté en esa estructura secundaria en el conjunto de datos. Por ejemplo, la preferencia de un anillo en un residuo en un Loop por un anillo en una droga se calcula como $Pref_{Loop}^{Drug} = P(Residuo\ en\ Loop\ interactuando\ con\ droga)/P(Loop) - 1$. *Se resta 1 sólo para trasladar la escala.*

Con respecto a la distancia en secuencia, esta tendencia se mantiene (aparecen dos de los anillos involucrados cercanos en secuencia), pero con respecto a la estructura secundaria, podemos apreciar diferencias. En la Figura 4.17 se puede ver la diferencia significativa que hay en términos de la estructura secundaria en la que se encuentran los residuos aromáticos. Los aromáticos interactuando con drogas tienen una clara preferencia por las regiones de tipo loop, mientras que los otros dos tipos de estructura (alfa y beta) parecen relegados. Por otro lado, aquellos aromáticos interactuando en interfaces proteína-proteína tienen una tendencia a estructuras de tipo alfa y beta y, más aún, las estructuras secundarias en las que están los residuos de ambas cadenas proteicas coinciden. Los que se encuentran en hélices alfa, tienen una preferencia por otros residuos aromáticos en hélice alfa y, análogamente, aquellos en hojas beta, interactúan con otros residuos aromáticos en hojas beta. Cuando un residuo en un *cluster* de aromáticos en interfaces PP, se encuentra en un

loop, es común encontralo interactuando con otro/s aromáticos en loops. Finalmente, pareciera que, dado un anillo aromático en la superficie de la estructura de una proteína, podríamos distinguir aquellos que tienen una tendencia a ser encontrados formando *clusters* proteína-proteína de aquellos formando *clusters* proteína-droga.

4.4. Conclusiones

Nuestros resultados, basados en el análisis de todos los *clusters* aromáticos en proteínas conocidas, muestran que los clusters IP: i) más allá del dímero, los *clusters* de aromáticos se pueden encontrar en más de la mitad de las proteínas estudiadas (proteínas únicas); ii) los *cluster* se hayan adoptando las mismas conformaciones que adoptan los motivos triméricos de bencenos aislados (SYM y LAD); y iii) estos *clusters* unen estructuras no locales que en principio estarían alejadas dada su lejanía en la estructura primaria de las proteínas.

Las interacciones aromáticas en proteínas han sido el objeto de muchos estudios desde las primeras determinaciones estructurales. En los años 80, se vió que las interaccciones entre anillos aromáticos en proteínas presentan distribuciones configuracionales esecialmente al azar. En este trabajo extendemos la idea de trabajar con las interacciones aromáticas, más allá del dímero en una exploración extensiva de los parámetros involucrados. Vimos que el impacto de los trímeros de aromáticos en proteínas es extremadamente alto, apareciendo en un altísimo porcentaje junto con los tetrámeros y *clusters* de más tamaño. Los trímeros SYM son los *clusters* más chicos con una intertacción sinérgica, estableciendo 3 interacciones con tres anillos aromáticos, lo cuál es de especial valor desde un punto de vista de la energía de interacción. Hemos mostrado que los *clusters* más grandes pueden entenderse en términos de los trímeros que los componen, dado que se observan los mismos ángulos conformacionales tomando los dos residuos más cercanos para calcularlo sobre cada anillo del *cluster*.

Los análisis de estructura secundaria mostraron una preferencia por las hélices alfa, en contraposición a la preferencia natural de los anillos aromáticos por las hojas beta y, especialmente, en juntar elementos de estructura secundaria lejanos en la secuencia proteica, lo cuál también se aprecia en el análisis de las distancias en secuencia, que mostró que los *clusters* tienden a ser no locales. Hecho que es validado por el análisis del entorno de los *clusters* que mostró que, en promedio, forman parte del núcleo de las proteínas.

A su vez, es interesante el hecho de que los trímeros y los tetrámeros ob-

servados coincidan, en promedio, con las estructuras optimizadas de *clusters* de benzenos y toluenos en vacío, y que a medida que se relajan los grados de libertad (por ejemplo en PD con respecto a IP y PP) estos mínimos energéticos se aprecian más todavía. Con respecto a las observaciones realizadas en los *clusters* PD y PP, vimos que es posible diferenciarlos usando varios parámetros. Esto quiere decir, por ejemplo, que sabiendo el área accesible al solvente de un residuo o la estructura secundaria de los residuos involucrados en los *clusters*, es posible tener una idea de la posible función de un *cluster* de aromáticos.

Capítulo 5

Conclusiones

Resumiendo lo expuesto hasta ahora, primero desarrollamos un pipeline bioinformático para anotar genomas de bacterias con el cual anotamos y depositamos en bases de datos de público acceso 2 genomas aislados en territorio Argentino: *Bizionia argentinesis* y *Exiguobacterium sp. S17*. Segundo, desarrollamos otro pipeline para realizar modelado comparativo de estructuras de proteínas y analizamos los genomas de 3 patógenos: *Mycobacterium tuberculosis*, *Klebsiella pneumoniae* y *Corynebacterium pseudotuberculosis*. Por útlimo, analizamos una gran cantidad de estructuras conocidas de proteínas para caracterizar motivos estructurales relevantes: los *clusters* de anillos aromáticos.

Con respecto al primer capítulo, vimos que los sistemas de anotación son posibles gracias a la acción combinada de bases de datos y algoritmos de búsqueda, y comparando con los elementos de dichas bases de datos, asignan función basandose en similaridad con otras secuencias de las que tenemos información acerca de su función. Este tipo de estrategias nos permiten detectar características conservadas, como por ejemplo, dominios, familias, etc. Sin embargo, queda claro que un alto porcentaje de las proteínas quedan anotadas como "de función desconocida", y hace falta emprender nuevos proyectos científicos que permitan una caracterización a gran escala de funciones proteicas para mejorar la anotación de la gran cantidad de información que se genera con las técnicas de nueva generación en secuenciación masiva.

En términos de lo visto en el segundo capítulo, vimos que las técnicas de predicción de estructura basadas en modelado comparativo, dan la posibilidad de analizar la función de cada variante proteica a partir de estructuras de proteínas similares, lo cual ayuda al trabajo de anotación de proteínas y le permite al curador manual mejorar la anotación automática. Hemos visto que se pueden realizar proyectos de predicción de estructura a escala genómica

cubriendo hoy día, entre un 30 % y un 50 % de los genes de una bacteria con una calidad razonable. Este tipo de proyectos cobra gran interés en la búsqueda de nuevas proteínas blanco en patógenos o en la búsqueda de proteínas con nuevas actividades de interés biotecnológico.

En el tercer capítulo, hemos visto que los clusters de aromáticos, más allá del dímero se encuentran en gran número de las proteínas y que, interesantemente, adoptan una conformación similiar a los clusters de bencenos aislados. Esto implicaría que las secuencias han evolucionado para optimizar estas interacciones y estabilizar la estructura de las proteínas. También observamos que los *clusters* de anillos aromáticos permiten obtener información funcional acerca de ciertas regiones de una proteína. Estas redes de interacciones de residuos aromáticos, permiten caraterizar las interacciones que pueden establecer las proteínas. Vimos que analizando las propiedades estructurales, en particular, la estructura secundaria y la accesibilidad al solvente, de los *clusters* de aromáticos podemos predecir qué rol tendrán en la proteína, ya sea uniendo ligandos o participando de la interacción entre proteínas. Analizando una gran cantidad de estructuras de proteínas depositadas en el PDB, obtuvimos como resultado que, los amino ácidos aromáticos que interactúan con otras moléculas, presentan un grado de exposición al solvente más grande que el promedio de los amino ácidos aromáticos en general. Más aún, aquellos que están interactuando con ligandos tipo droga suelen aparecer en Loops, a diferencia de la propensión natural de los mismos por las hojas beta.

Finalmente, como comentamos antes, los anotadores automáticos de genomas resuelven muy bien el problema de obtener información relevante a partir de una secuenciación, y que, a su vez, esto favoreció el creciemiento desmedido en la cantidad de genomas depositados en bases de datos que poseen una anotación automática, de los cuales se volvería prohibitivo realizar experimentos sobre cada uno de los genes presentes en éstos genomas de manera de mejorar el conocimiento disponible. Es por esto que, obtener una estructura molde para modelar estructuralmente una determinada secuencia, es una herramienta poderosa para entender las funciones de aquellas proteínas para las cuales esto fuera posible. Una vez modelada la estructura, las interacciones presentes en ella aportan bastante a la comprensión de la función proteica, de las cuales, en esta tesis nosotros hicimos foco en los *clusters* de anillos aromáticos como una forma de integrar la información de las redes de interacciones aromáticas.

Bibliografía

[Aaltonen and Silow, 2008] Aaltonen, E. K. and Silow, M. (2008). Transmembrane topology of the acr3 family arsenite transporter from bacillus subtilis. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, 1778(4):963–973.

[Albrecht et al., 2003] Albrecht, M., Tosatto, S. C., Lengauer, T., and Valle, G. (2003). Simple consensus procedures are effective and sufficient in secondary structure prediction. *Protein Engineering Design and Selection*, 16(7):459–462.

[Altschul et al., 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

[Altschul et al., 1997] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST : a new generation of protein database search programs. 25(17):3389–3402.

[Aran et al., 2014] Aran, M., Smal, C., Pellizza, L., Gallo, M., Otero, L. H., Klinke, S., Goldbaum, F. A., Ithurralde, E. R., Bercovich, A., Mac Cormack, W. P., et al. (2014). Solution and crystal structure of ba42, a protein from the antarctic bacterium bizionia argentinensis comprised of a stand-alone tpm domain. *Proteins: Structure, Function, and Bioinformatics*, 82(11):3062–3078.

[Aravinda et al., 2003] Aravinda, S., Shamala, N., Das, C., Sriranjini, A., Karle, I. L., and Balaram, P. (2003). Aromatic-aromatic interactions in crystal structures of helical peptide scaffolds containing projecting phenylalanine residues. *Journal of the American Chemical Society*, 125(18):5308–5315.

[Aziz et al., 2008] Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., et al.

(2008). The rast server: rapid annotations using subsystems technology. *BMC genomics*, 9(1):75.

[Baker and Sali, 2001] Baker, D. and Sali, a. (2001). Protein structure prediction and structural genomics. *Science (New York, N.Y.)*, 294(5540):93–6.

[Bashford et al., 1987] Bashford, D., Chothia, C., and Lesk, A. M. (1987). Determinants of a protein fold: Unique features of the globin amino acid sequences. *Journal of molecular biology*, 196(1):199–216.

[Bateman et al., 2004] Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004). The Pfam protein families database. *Nucleic acids research*, 32(Database issue):D138–41.

[Bauer et al., 2006] Bauer, M., Kube, M., Teeling, H., Richter, M., Lombardot, T., Allers, E., Würdemann, C. A., Quast, C., Kuhl, H., Knaust, F., et al. (2006). Whole genome analysis of the marine bacteroidetes 'gramella forsetii'reveals adaptations to degradation of polymeric organic matter. *Environmental Microbiology*, 8(12):2201–2213.

[Benkert et al., 2010] Benkert, P., Biasini, M., and Schwede, T. (2010). Toward the estimation of the absolute quality of individual protein structure models. 1:1–8.

[Benkert et al., 2008] Benkert, P., Tosatto, S. C. E., and Schomburg, D. (2008). QMEAN: A comprehensive scoring function for model quality assessment. *Proteins*, 71(1):261–77.

[Bercovich et al., 2008] Bercovich, A., Vazquez, S. C., Yankilevich, P., Coria, S. H., Foti, M., Hernández, E., Vidal, A., Ruberto, L., Melo, C., Marenssi, S., Criscuolo, M., Memoli, M., Arguelles, M., and Mac Cormack, W. P. (2008). Bizionia argentinensis sp. nov., isolated from surface marine water in Antarctica. *International journal of systematic and evolutionary microbiology*, 58(Pt 10):2363–7.

[Berman et al., 2000] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1):235–242.

[Besemer and Borodovsky, 2005] Besemer, J. and Borodovsky, M. (2005). Genemark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic acids research*, 33(suppl 2):W451–W454.

[Bhattacharyya et al., 2002] Bhattacharyya, R., Samanta, U., and Chakrabarti, P. (2002). Aromatic–aromatic interactions in and around α-helices. *Protein engineering*, 15(2):91–100.

[Bowman and Nichols, 2005] Bowman, J. P. and Nichols, D. S. (2005). Novel members of the family flavobacteriaceae from antarctic maritime habitats including subsaximicrobium wynnwilliamsii gen. nov., sp. nov., subsaximicrobium saxinquilinus sp. nov., subsaxibacter broadyi gen. nov., sp. nov., lacinutrix copepodicola gen. nov., sp. nov., and novel species of the genera bizionia, gelidibacter and gillisia. *International journal of systematic and evolutionary microbiology*, 55(4):1471–1486.

[Brocchieri and Karlin, 1994] Brocchieri, L. and Karlin, S. (1994). Geometry of interplanar residue contacts in protein structures. *Proceedings of the National Academy of Sciences*, 91(20):9297–9301.

[Browne et al., 1969] Browne, W. J., North, A., Phillips, D., Brew, K., Vanaman, T. C., and Hill, R. L. (1969). A possible three-dimensional structure of bovine α-lactalbumin based on that of hen's egg-white lysozyme. *Journal of molecular biology*, 42(1):65–86.

[Burley and Petsko, 1985] Burley, S. and Petsko, G. (1985). Aromatic-aromatic interaction: a mechanism of protein structure stabilization. *Science*, 229(4708):23–28.

[Chelli et al., 2002] Chelli, R., Gervasio, F. L., Procacci, P., and Schettino, V. (2002). Stacking and T-shape Competition in Aromatic-Aromatic Amino Acid Interactions. (12):6133–6143.

[Chohan et al., 2015] Chohan, T. A., Qian, H., Pan, Y., and Chen, J.-Z. (2015). Cyclin-dependent kinase-2 as a target for cancer therapy: Progress in the development of cdk2 inhibitors as anti-cancer agents. *Current medicinal chemistry*, 22(2):237–263.

[Chothia et al., 1998] Chothia, C., Gelfand, I., and Kister, A. (1998). Structural determinants in the sequences of immunoglobulin variable domain. *Journal of molecular biology*, 278(2):457–479.

[Chothia and Lesk, 1986] Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO journal*, 5(4):823.

[Consortium et al., 2014] Consortium, U. et al. (2014). Uniprot: a hub for protein information. *Nucleic acids research*, page gku989.

[Cuff et al., 1998] Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M., and Barton, G. J. (1998). Jpred: a consensus secondary structure prediction server. *Bioinformatics*, 14(10):892–893.

[Dang, 2000] Dang, L. X. (2000). Molecular dynamics study of benzene–benzene and benzene–potassium ion interactions using polarizable potential models. *The Journal of Chemical Physics*, 113(1):266–273.

[de Araujo et al., 1999] de Araujo, A. F. P., Pochapsky, T. C., and Joughin, B. (1999). Thermodynamics of interactions between amino acid side chains: experimental differentiation of aromatic-aromatic, aromatic-aliphatic, and aliphatic-aliphatic side-chain interactions in water. *Biophysical journal*, 76(5):2319–2328.

[De Meijere and Huisken, 1990] De Meijere, A. and Huisken, F. (1990). Co2-laser induced photodissociation studies of size-selected small benzene clusters. *The Journal of Chemical Physics*, 92(10):5826–5834.

[Deane and Blundell, 2001] Deane, C. M. and Blundell, T. L. (2001). Coda: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Science*, 10(3):599–612.

[Delcher et al., 2007] Delcher, A. L., Bratke, K. a., Powers, E. C., and Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics (Oxford, England)*, 23(6):673–9.

[Dib et al., 2008] Dib, J., Motok, J., Zenoff, V. F., Ordoñez, O., and Farías, M. E. (2008). Occurrence of resistance to antibiotics, uv-b, and arsenic in bacteria isolated from extreme environments in high-altitude (above 4400 m) andean wetlands. *Current microbiology*, 56(5):510–517.

[Dib et al., 2009] Dib, J. R., Weiss, A., Neumann, A., Ordoñez, O., Estévez, M. C., and Farías, M. E. (2009). Isolation of bacteria from remote high altitude andean lakes able to grow in the presence of antibiotics. *Recent patents on anti-infective drug discovery*, 4(1):66–76.

[Easter et al., 2005] Easter, D. C., Terrell, D. A., and Roof, J. A. (2005). Monte carlo studies of isomers, structures, and properties in benzene-cyclohexane clusters: Computation strategy and application to the dimer and trimer,(c6h6)(c6h12) n, n= 1-2. *The Journal of Physical Chemistry A*, 109(4):673–689.

[Eidenschink et al., 2009] Eidenschink, L. A., Kier, B. L., and Andersen, N. H. (2009). Determinants of fold stabilizing aromatic-aromatic interactions in short peptides. In *Peptides for Youth*, pages 73–74. Springer.

[Eisenberg et al., 1997] Eisenberg, D., Lüthy, R., and Bowie, J. U. (1997). Verify3d: assessment of protein models with three-dimensional profiles. *Methods in enzymology*, (277):396–404.

[Engkvist et al., 1999] Engkvist, O., Hobza, P., Selzle, H. L., and Schlag, E. W. (1999). Benzene trimer and benzene tetramer: Structures and properties determined by the nonempirical model (NEMO) potential calibrated from the CCSD(T) benzene dimer energies. *The Journal of Chemical Physics*, 110(12):5758.

[Espinoza-Fonseca and García-Machorro, 2008] Espinoza-Fonseca, L. M. and García-Machorro, J. (2008). Aromatic–aromatic interactions in the formation of the mdm2–p53 complex. *Biochemical and biophysical research communications*, 370(4):547–551.

[Eswar et al., 2006] Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M., Eramian, D., Shen, M.-y., Pieper, U., and Sali, A. (2006). Comparative protein structure modeling using modeller. *Current protocols in bioinformatics*, pages 5–6.

[Farías et al., 2013] Farías, M. E., Rascovan, N., Toneatti, D. M., Albarracín, V. H., Flores, M. R., Poiré, D. G., Collavino, M. M., Aguilar, O. M., Vazquez, M. P., and Polerecky, L. (2013). The discovery of stromatolites developing at 3570 m above sea level in a high-altitude volcanic lake socompa, argentinean andes. *PloS one*, 8(1):e53497.

[Finn et al., 2006] Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. L., and Bateman, A. (2006). Pfam: clans, web tools and services. *Nucleic acids research*, 34(Database issue):D247–51.

[Fiser et al., 2000] Fiser, A., Do, R. K. G., and Šali, A. (2000). Modeling of loops in protein structures. *Protein science*, 9(09):1753–1773.

[Flohil et al., 2002] Flohil, J. a., Vriend, G., and Berendsen, H. J. C. (2002). Completion and refinement of 3-D homology models with restricted molecular dynamics: application to targets 47, 58, and 111 in the CASP modeling competition and posterior analysis. *Proteins*, 48(4):593–604.

[Fu et al., 2009] Fu, H.-L., Meng, Y., Ordóñez, E., Villadangos, A. F., Bhattacharjee, H., Gil, J. A., Mateos, L. M., and Rosen, B. P. (2009). Properties of arsenite efflux permeases (acr3) from alkaliphilus metalliredigens and corynebacterium glutamicum. *Journal of Biological Chemistry*, 284(30):19887–19895.

[Galperin et al., 2014] Galperin, M. Y., Makarova, K. S., Wolf, Y. I., and Koonin, E. V. (2014). Expanded microbial genome coverage and improved protein family annotation in the cog database. *Nucleic acids research*, page gku1223.

[Ghose et al., 2008] Ghose, A. K., Herbertz, T., Pippin, D. A., Salvino, J. M., and Mallamo, J. P. (2008). Knowledge based prediction of ligand binding modes and rational inhibitor design for kinase drug discovery. *Journal of medicinal chemistry*, 51(17):5149–5171.

[Gonzalez and Lim, 2001] Gonzalez, C. and Lim, E. C. (2001). Ab Initio Study of the Intermolecular Interactions in Small Benzene Clusters : The Equilibrium Structures of Trimer , Tetramer , and Pentamer. pages 1904–1908.

[Guex and Peitsch, 1997] Guex, N. and Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, 18(15):2714–23.

[Haft, 2003] Haft, D. H. (2003). The TIGRFAMs database of protein families. *Nucleic Acids Research*, 31(1):371–373.

[Hemmerich et al., 2010] Hemmerich, C., Buechlein, A., Podicheti, R., Revanna, K. V., and Dong, Q. (2010). An ergatis-based prokaryotic genome annotation web server. *Bioinformatics*, 26(8):1122–1124.

[Hooft et al., 1996] Hooft, R., Vriend, G., Sander, C., and Abola, E. E. (1996). Errors in protein structures. *Nature*, 381(6580):272–272.

[Huggins et al., 2012] Huggins, D. J., Sherman, W., and Tidor, B. (2012). Rational approaches to improving selectivity in drug design. *Journal of medicinal chemistry*, 55(4):1424–1444.

[Humphry et al., 2001] Humphry, D. R., George, A., Black, G. W., and Cummings, S. P. (2001). Flavobacterium frigidarium sp. nov., an aerobic, psychrophilic, xylanolytic and laminarinolytic bacterium from antarctica. *International journal of systematic and evolutionary microbiology*, 51(4):1235–1243.

[Hunter et al., 1991] Hunter, C. A., Singh, J., and Thornton, J. M. (1991). π-π interactions: The geometry and energetics of phenylalanine-phenylalanine interactions in proteins. *Journal of molecular biology*, 218(4):837–846.

[Jayaraman and Shah, 2008] Jayaraman, S. and Shah, K. (2008). Comparative studies on inhibitors of hiv protease: a target for drug design. *In silico biology*, 8(5-6):427–447.

[Johnson et al., 2010] Johnson, L. S., Eddy, S. R., and Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure.

[Johnson et al., 2007] Johnson, R. M., Hecht, K., and Deber, C. M. (2007). Aromatic and cation-π interactions enhance helix-helix association in a membrane environment. *Biochemistry*, 46(32):9208–9214.

[Jones et al., 2008] Jones, C. M., Stres, B., Rosenquist, M., and Hallin, S. (2008). Phylogenetic analysis of nitrite, nitric oxide, and nitrous oxide respiratory enzymes reveal a complex evolutionary history for denitrification. *Molecular biology and evolution*, 25(9):1955–1966.

[Jones, 1999] Jones, D. T. (1999). Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. pages 195–202.

[Kabsch and Sander, 1983] Kabsch, W. and Sander, C. (1983). Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. 22:2577–2637.

[Kanehisa et al., 2007] Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., and Yamanishi, Y. (2007). KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, 36(Database):D480–D484.

[Kanehisa and Goto, 2000] Kanehisa, M. and Goto, S. (2000). KEGG : Kyoto Encyclopedia of Genes and Genomes. 28(1):27–30.

[Kanehisa et al., 2006] Kanehisa, M., Goto, S., Hattori, M., Aoki-kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., and Araki, M. (2006). From genomics to chemical genomics : new developments in KEGG. 34:354–357.

[Kannan and Vishveshwara, 2000] Kannan, N. and Vishveshwara, S. (2000). Aromatic clusters : a determinant of thermal stability of thermophilic proteins. 13(11):753–761.

[Kihara and Skolnick, 2003] Kihara, D. and Skolnick, J. (2003). The pdb is a covering set of small protein structures. *Journal of molecular biology*, 334(4):793–802.

[Krause et al., 1991] Krause, H., Ernstberger, B., and Neusser, H. (1991). Binding energies of small benzene clusters. *Chemical physics letters*, 184(5):411–417.

[Kyte and Doolittle, 1982] Kyte, J. and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105 – 132.

[Lanzarotti et al., 2011] Lanzarotti, E., Pellizza, L., Bercovich, A., Foti, M., Coria, S. H., Vazquez, S. C., Ruberto, L., Herna, E. A., Dias, R. L., Cormack, W. P. M., Cicero, D. O., Smal, C., Nicolas, M. F., Tereza, A., Vasconcelos, R., Marti, M. A., and Turjanski, A. G. (2011). Draft Genome Sequence of Bizionia argentinensis , Isolated from Antarctic Surface Water. 193(23):6797–6798.

[Larson and Davidson, 2000] Larson, S. M. and Davidson, A. R. (2000). The identification of conserved interactions within the sh3 domain by alignment of sequences and structures. *Protein Science*, 9(11):2170–2180.

[Legon and Millen, 1987] Legon, A. and Millen, D. (1987). Directional character, strength, and nature of the hydrogen bond in gas-phase dimers. *Accounts of Chemical Research*, 20(1):39–46.

[Lesk and Fordham, 1996] Lesk, A. M. and Fordham, W. D. (1996). Conservation and variability in the structures of serine proteinases of the chymotrypsin family. *Journal of molecular biology*, 258(3):501–537.

[Li et al., 2013] Li, S., Xu, Y., Shen, Q., Liu, X., Lu, J., Chen, Y., Lu, T., Luo, C., Luo, X., Zheng, M., et al. (2013). Non-covalent interactions with aromatic rings: Current understanding and implications for rational drug design. *Current pharmaceutical design*, 19(36):6522–6533.

[Li and Godzik, 2006] Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22(13):1658–9.

[Lowe and Eddy, 1997] Lowe, T. M. and Eddy, S. R. (1997). trnascan-se: a program for improved detection of transfer rna genes in genomic sequence. *Nucleic acids research*, 25(5):955–964.

[Lüthy R, 1992] Lüthy R, Bowie JU, E. D. (1992). Assessment of protein models with three-dimensional profiles. *Nature*.

[Ma et al., 2003] Ma, B., Elkayam, T., Wolfson, H., and Nussinov, R. (2003). Protein – protein interactions : Structurally conserved residues distinguish between binding sites and exposed protein surfaces. (Track II).

[MacKerell et al., 1998] MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R., Evanseck, J., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S. a., et al. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry B*, 102(18):3586–3616.

[Magnan and Baldi, 2014] Magnan, C. N. and Baldi, P. (2014). Sspro/accpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*, 30(18):2592–2597.

[Malkov et al., 2008] Malkov, S. N., Zivković, M. V., Beljanski, M. V., Hall, M. B., and Zarić, S. D. (2008). A reexamination of the propensities of amino acids towards a particular secondary structure: classification of amino acids based on their chemical structure. *Journal of molecular modeling*, 14(8):769–75.

[Mandal et al., 2009] Mandal, S., Moudgil, M., and Mandal, S. K. (2009). Rational drug design. *European journal of pharmacology*, 625(1-3):90–100.

[Manfred, 1993] Manfred, S. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins*.

[Marsili et al., 2008] Marsili, S., Chelli, R., Schettino, V., and Procacci, P. (2008). Thermodynamics of stacking interactions in proteins. *Physical Chemistry Chemical Physics*, 10(19):2673–2685.

[Marti-renom et al., 2003] Marti-renom, C. M. A., Fiser, A., Madhusudhan, M. S., John, B., Stuart, A., Eswar, N., Pieper, U., Mirkovic, N., Shen, M.-y., and Sali, A. (2003). Modeling Protein Structure from Its Sequence. pages 1–33.

[McGaughey et al., 1998] McGaughey, G. B., Gagné, M., and Rappé, A. K. (1998). pi-stacking interactions alive and well in proteins. *Journal of Biological Chemistry*, 273(25):15458–15463.

[Melo and Feytmans, 1998] Melo, F. and Feytmans, E. (1998). Assessing protein structures with a non-local atomic interaction energy. *Journal of molecular biology*, 277(5):1141–1152.

[Melo and Sali, 2007] Melo, F. and Sali, A. (2007). Fold assessment for comparative protein structure modeling. *Protein science : a publication of the Protein Society*, 16(11):2412–26.

[Michnick and Shakhnovich, 1998] Michnick, S. W. and Shakhnovich, E. (1998). A strategy for detecting the conservation of folding-nucleus residues in protein superfamilies. *Folding and Design*, 3(4):239–251.

[Mitchell et al., 1994] Mitchell, J. B., Nandi, C. L., McDonald, I. K., Thornton, J. M., and Price, S. L. (1994). Amino/aromatic interactions in proteins: is the evidence stacked against hydrogen bonding? *Journal of molecular biology*, 239(2):315–331.

[Morimoto et al., 2007] Morimoto, T., Uno, H., and Furuta, H. (2007). Benzene ring trimer interactions modulate supramolecular structures. *Angewandte Chemie International Edition*, 46(20):3672–3675.

[Nancy et al., 2010] Nancy, Y. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S. C., Ester, M., Foster, L. J., et al. (2010). Psortb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, 26(13):1608–1615.

[Nedashkovskaya et al., 2005] Nedashkovskaya, O. I., Kim, S. B., Lysenko, A. M., Frolova, G. M., Mikhailov, V. V., and Bae, K. S. (2005). Bizionia paragorgiae gen. nov., sp. nov., a novel member of the family flavobacteriaceae isolated from the soft coral paragorgia arborea. *International journal of systematic and evolutionary microbiology*, 55(1):375–378.

[Nedashkovskaya et al., 2010] Nedashkovskaya, O. I., Vancanneyt, M., and Kim, S. B. (2010). Bizionia echini sp. nov., isolated from a sea urchin. *International journal of systematic and evolutionary microbiology*, 60(4):928–931.

[Ordoñez et al., 2009] Ordoñez, O. F., Flores, M. R., Dib, J. R., Paz, A., and Farías, M. E. (2009). Extremophile culture collection from andean lakes: extreme pristine environments that host a wide diversity of microorganisms with tolerance to uv radiation. *Microbial ecology*, 58(3):461–473.

[Ordoñez et al., 2015] Ordoñez, O. F., Lanzarotti, E. O., Kurth, D. G., Cortez, N., Farias, M. E., and Turjanski, A. G. (2015). Genome comparison of

two exiguobacterium strains from high altitude andean lakes with different arsenic resistance: Identification and 3d modeling of the acr3 efflux pump. *Frontiers in Environmental Science*, 3:50.

[Pavesi et al., 1994] Pavesi, A., Conterio, F., Bolchi, A., Dieci, G., and Ottonello, S. (1994). Identification of new eukaryotic trna genes in genomic dna databases by a multistep weight matrix anaylsis of transcriptional control regions. *Nucleic acids research*, 22(7):1247–1256.

[Petersen et al., 2011] Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nature methods*, 8(10):785–786.

[Pieper et al., 2014] Pieper, U., Webb, B. M., Dong, G. Q., Schneidman-Duhovny, D., Fan, H., Kim, S. J., Khuri, N., Spill, Y. G., Weinkam, P., Hammel, M., et al. (2014). Modbase, a database of annotated comparative protein structure models and associated resources. *Nucleic acids research*, 42(D1):D336–D346.

[Pollastri and McLysaght, 2005] Pollastri, G. and McLysaght, A. (2005). Porter: a new, accurate server for protein secondary structure prediction. *Bioinformatics (Oxford, England)*, 21(8):1719–20.

[R. Flores et al., 2009] R. Flores, M., F. Ordoñez, O., J. Maldonado, M., and E. Farías, M. (2009). Isolation of uv-b resistant bacteria from two high altitude andean lakes (4,400 m) with saline and non saline conditions. *The Journal of general and applied microbiology*, 55(6):447–458.

[Radusky et al., 2014] Radusky, L., Defelipe, L. A., Lanzarotti, E., Luque, J., Barril, X., Marti, M. A., and Turjanski, A. G. (2014). Tuberq: a mycobacterium tuberculosis protein druggability database. *Database*, 2014:bau035.

[Radusky et al., 2015] Radusky, L. G., Hassan, S. S., Lanzarotti, E., Tiwari, S., Jamal, S. B., Ali, J., Ali, A., Ferreira, R. S., Barh, D., Silva, A., et al. (2015). An integrated structural proteomics approach along the druggable genome of corynebacterium pseudotuberculosis species for putative druggable targets. *BMC Genomics*, 16(Suppl 5):S9.

[Rajamani et al., 2004] Rajamani, D., Thiel, S., Vajda, S., and Camacho, C. J. (2004). Anchor residues in protein – protein interactions. 101(31):11287–11292.

[Ray and Lindahl, 2010] Ray, A. and Lindahl, E. (2010). Model Quality Assessment for Membrane Proteins. pages 1–9.

[Ritchie and Macdonald, 2014] Ritchie, T. J. and Macdonald, S. J. (2014). Physicochemical descriptors of aromatic character and their use in drug discovery: Miniperspective. *Journal of medicinal chemistry*, 57(17):7206–7215.

[Ritchie and Macdonald, 2009] Ritchie, T. J. and Macdonald, S. J. F. (2009). The impact of aromatic ring count on compound developability–are too many aromatic rings a liability in drug design? *Drug discovery today*, 14(21-22):1011–20.

[Rodrigues and Tiedje, 2007] Rodrigues, D. F. and Tiedje, J. M. (2007). Multi-locus real-time pcr for quantitation of bacteria in the environment reveals exiguobacterium to be prevalent in permafrost. *FEMS microbiology ecology*, 59(2):489–499.

[Rombel et al., 2002] Rombel, I. T., Sykes, K. F., Rayner, S., and Johnston, S. A. (2002). Orf-finder: a vector for high-throughput gene identification. *Gene*, 282(1):33–41.

[Rost, 1999] Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein engineering*, 12(2):85–94.

[Šali and Blundell, 1993] Šali, A. and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal of molecular biology*, 234(3):779–815.

[Samanta et al., 1999] Samanta, U., Pal, D., and Chakrabarti, P. (1999). Packing of aromatic rings against tryptophan residues in proteins. *Acta Crystallographica Section D: Biological Crystallography*, 55(8):1421–1427.

[Sander and Schneider, 1991] Sander, C. and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9(1):56–68.

[Sciara and Mancia, 2012] Sciara, G. and Mancia, F. (2012). Highlights from recently determined structures of membrane proteins: a focus on channels and transporters. *Current opinion in structural biology*, 22(4):476–481.

[Serrano et al., 1991] Serrano, L., Bycroft, M., and Fersht, A. R. (1991). Aromatic-aromatic interactions and protein stability: investigation by double-mutant cycles. *Journal of molecular biology*, 218(2):465–475.

[Sillitoe et al., 2015] Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., Furnham, N., Laskowski, R. A., Lee, D., Lees, J. G., et al. (2015). Cath: comprehensive structural and functional annotations for genome sequences. *Nucleic acids research*, 43(D1):D376–D381.

[Smal et al., 2012] Smal, C., Aran, M., Lanzarotti, E., Papouchado, M., Foti, M., Marti, M. A., Coria, S. H., Vazquez, S. C., Bercovich, A., Mac Cormack, W. P., et al. (2012). 1h, 15n and 13c chemical shift assignments of the ba42 protein of the psychrophilic bacteria bizionia argentinensis sp. nov. *Biomolecular NMR assignments*, 6(2):181–183.

[Tatusov et al., 2000] Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000). The COG database : a tool for genome-scale analysis of protein functions and evolution. 28(1):33–36.

[Tauer and Sherrill, 2005] Tauer, T. P. and Sherrill, C. D. (2005). Beyond the Benzene Dimer : An Investigation of the Additivity of π-π Interactions. pages 10475–10478.

[Vishnivetskaya et al., 2009] Vishnivetskaya, T. A., Kathariou, S., and Tiedje, J. M. (2009). The exiguobacterium genus: biodiversity and biogeography. *Extremophiles*, 13(3):541–555.

[Vishnivetskaya et al., 2006] Vishnivetskaya, T. A., Petrova, M. A., Urbance, J., Ponder, M., Moyer, C. L., Gilichinsky, D. A., and Tiedje, J. M. (2006). Bacterial community in ancient siberian permafrost as characterized by culture and culture-independent methods. *Astrobiology*, 6(3):400–414.

[Viterbi, 1967] Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2):260–269.

[Wilson et al., 2009] Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C., and Gough, J. (2009). Superfamily—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic acids research*, 37(suppl 1):D380–D386.

[Xiang et al., 2002] Xiang, Z., Soto, C. S., and Honig, B. (2002). Evaluating conformational free energies: The colony energy and its application to the problem of loop prediction.

[Xu and Zhang, 2010] Xu, J. and Zhang, Y. (2010). How significant is a protein structure similarity with tm-score= 0.5? *Bioinformatics*, 26(7):889–895.

[Zenoff et al., 2006] Zenoff, V. F., Sineriz, F., and Farías, M. (2006). Diverse responses to uv-b radiation and repair mechanisms of bacteria isolated from high-altitude aquatic environments. *Applied and environmental microbiology*, 72(12):7857–7863.

[Zhang and Skolnick, 2004] Zhang, Y. and Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins*, 57(4):702–10.