Capítulo 16

Diseño de fármacos asistido por ordenador

Antonio Morreale, Almudena Perona, Javier Klett,
Álvaro Cortés-Cabrera y Helena G. Dos Santos

16.1. Introducción

De una manera muy general podemos decir que los sistemas biológicos comparten un lenguaje común de comunicación basado en las interacciones que se establecen entre distintos tipos de moléculas como proteínas, ácidos nucleicos y entidades químicas de menor tamaño denominadas genéricamente ligandos. Muchas enfermedades provienen precisamente de un mal funcionamiento de estas interacciones por lo que conocer cómo se producen nos va a permitir interferir en ellas a través del uso de ligandos que disminuyan, o incluso anulen, los efectos de la enfermedad. También se puede dar el caso de que lo que busquemos sean activadores de dichas interacciones.

La búsqueda y posterior comercialización de un nuevo fármaco es un proceso que requiere de un tremendo esfuerzo en investigación y desarrollo así como de una enorme inversión económica. Bajo el paradigma de que con el mayor número posible de candidatos a estudiar sería más probable encontrar nuevos fármacos, en los últimos 30 años se ha popularizado el empleo de técnicas experimentales de carácter masivo como la química combinatoria y el cribado de alto rendimiento (*high-throughput screening*). Sin embargo, el esfuerzo económico así invertido y el número de nuevos fármacos que han alcanzado el mercado no están en concordancia, y está claro que el número de estos está muy por debajo de las expectativas iniciales hechas en base a las técnicas masivas (Figura 16.1).

De alguna manera estos resultados tan inesperados han repercutido de forma positiva en el desarrollo paralelo de técnicas computacionales con el objetivo de servir como solución al cuello de botella existente en el actual modelo de búsqueda de nuevos fármacos. La idea subyacente es *intentar racionalizar y, en base al conocimiento, acelerar las etapas iniciales del diseño de fármacos.*

Los métodos teóricos, cuando están bien fundamentados e implementados, permiten seleccionar a partir de colecciones de millones de moléculas (*quimiotecas*) aquellos candidatos que tienen una mayor probabilidad de interaccionar con una *diana terapéutica* dada. Este reducido conjunto puede analizarse experimentalmente y aquellos compuesto que den señal de interacción con la diana, llamados *hits*, puede optimizarse hasta alcanzar los perfiles farmacocinéticos y farmacodinámicos adecuados y convertirse así en *leads*.

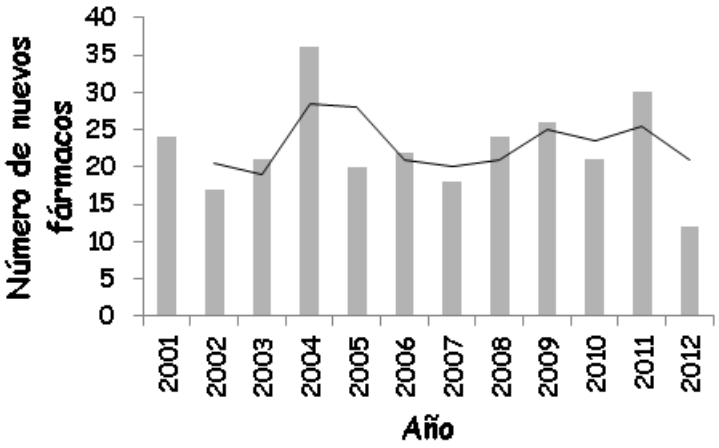


Figura 16.1: Evolución en el tiempo del número de nuevos fármacos que han alcanzado el mercado.

Desde una perspectiva teórica, y dependiendo de la información estructural que tengamos a nuestra disposición, se pueden presentar cuatro escenarios distintos que van desde el más favorable de todos (cuando tenemos información estructural de la diana y de alguno de sus ligandos, entonces empleamos *docking* y *cribado virtual*) pasando por dos casos intermedios (cuando conocemos o la estructura de algunos ligandos o de la diana, y entonces se usan los llamados *farmacóforos*) hasta el caso más desfavorable caracterizado por la ausencia de cualquier información estructural (donde es necesario llevar a cabo estudios experimentales que nos den idea sobre el tipo de estructuras involucradas en la interacción para poder aplicar algún método teórico).

En este capítulo revisaremos los fundamentos de la técnica denominada *docking* y su uso en el cribado virtual de quimiotecas, y repasaremos los tipos principales de interacciones que se producen entre un ligando y su diana y cómo cuantificarlas. De aquí en adelante usaremos indistintamente los términos *diana* y *receptor*, y en algunos casos proteína, si bien al final del capítulo hay dos apartados relacionados con el *docking* proteína-proteína y ligando-ácido nucleico, respectivamente. Terminaremos este estudio con un apartado donde se recogen las conclusiones principales así como las perspectivas futuras de estas técnicas.

16.2. Docking proteína-ligando

16.2.1. Definición del problema del *docking*

El problema del *docking* pude definirse de forma sencilla de la siguiente manera: dados a) la estructura 3D de una diana de interés, una proteína por ejemplo, y b) la estructura 3D de un ligando, encontrar cuál es la configuración 3D experimental que posee el complejo formado por la proteína y el ligando.

Dado que el *ligando* puede adoptar diversas posiciones dentro del sitio de unión de la proteína, la herramienta de *docking* consta, en primer lugar, de un *método de muestreo* (componente estructural) que enumera todas las configuraciones posibles ligando-receptor (*poses*) y, en segundo lugar, de una función matemática, llamada *función de puntuación o scoring*, que evalúa cómo de buenas son las interacciones existentes entre la proteína y el ligando en cada una de las poses (componente energética).

16.2.2. Componente estructural

El método de *docking* pertenece a una clase de técnicas más amplia generalmente conocida como *métodos basados en la estructura*, por lo que se asume un conocimiento previo de la estructura 3D tanto de la diana como del ligando. Por un lado, la estructura de la diana puede obtenerse por medio de diferentes métodos experimentales (principalmente cristalografía de rayos-X y espectroscopía de Resonancia Magnética Nuclear) o teóricos (modelado por homología) y se puede conseguir a través del Protein Data Bank (PDB)[4]. Por otro lado, la estructura experimental de muchos ligandos está accesible a través de la base de datos de Cambridge Structural Database (CSD) [1] o puede construirse de forma fácil con un programa de modelado molecular. Además, existen multitud de estructuras 3D de complejos receptor-ligando en el PDB, siendo ésta una fuente importante de estructuras para probar los algoritmos de *docking*, ya que conocemos a priori cuál es el resultado final.

Desde un punto de vista más técnico, para el *algoritmo de docking* la parte más importante de la estructura de la diana es la zona de unión del ligando, conocida como *bolsillo*, *sitio* o *centro activo*, o *cavidad de unión* (Figura 16.2). La unión del ligando al centro activo produce una modificación, activación o inhibición de la respuesta fisiológica de la diana. El centro activo puede estar localizado en la superficie de la diana o enterrado en su interior. La disposición 3D de las cadenas laterales y del esqueleto de los residuos de la diana en el centro activo determina la especificidad del ligando por esa diana en particular (Figura 16.2). Por último, algunos residuos del centro activo se han conservados a lo largo de la evolución, principalmente aquellos relacionados con la actividad de la diana. Si conocemos dónde está situado el centro activo (por ejemplo a través de la estructura 3D de complejos diana-ligando determinada experimentalmente) podremos guiar al algoritmo de *docking* a una región en particular de la diana, en lugar de buscar en toda su superficie (*docking ciego*).

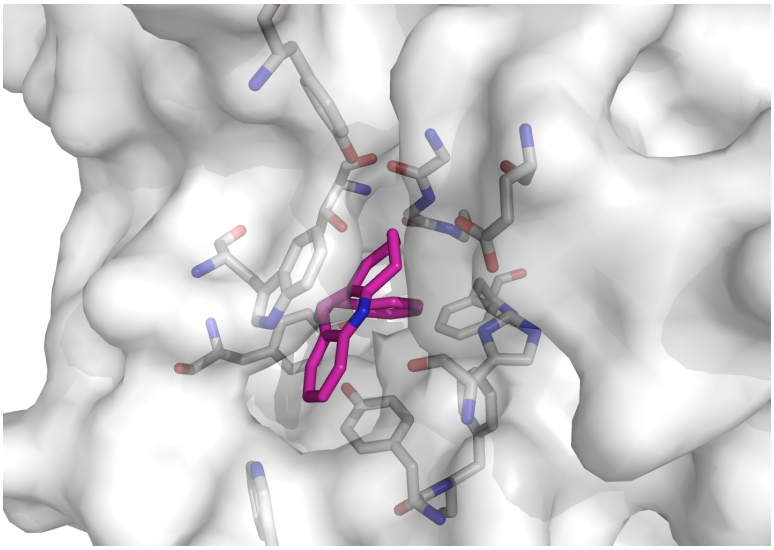


Figura 16.2: Complejo proteína-ligando mostrando detalles del centro activo y las interacciones entre el ligando y las cadenas laterales de los residuos que lo forman.

16.2.3. Componente energética

La estabilidad de un complejo molecular, así como la de cualquier molécula individual, se puede cuantificar recurriendo a los principios básicos de la física usando *modelos clásicos*, *mecánica de Newton*, o

cuánticos. Dado que los cálculos moleculares basados en mecánica cuántica son muy costosos y sobre todo para sistemas con un elevado número de átomos (como los complejos que nos ocupan), la mecánica clásica es el método de uso común en los cálculos de energías de interacción entre la diana y el ligando.

Cuando se produce la unión entre el ligando y el centro activo de su diana se establecen una serie de interacciones específicas que son las responsables de la estabilidad total del complejo. El éxito de un método de *docking* radica principalmente en un conocimiento profundo y una implementación apropiada de las fuerzas directrices que rigen la unión entre la diana y el ligando. En concreto, a lo largo del protocolo de *docking* se hacen ciertas aproximaciones y se toman ciertas decisiones basadas en la cuantificación de la unión diana-ligando estableciendo un criterio de *bondad de ajuste* en virtud del cual se elegirá la mejor solución como resultado final del *docking*.

Los principales tipos de interacciones moleculares que se consideran fundamentales para entender y racionalizar la unión ligando-diana son las siguientes: *interacciones de van der Waals (vdW)*, *interacciones electrostáticas*, *interacciones por enlace de hidrógeno*, *interacciones con el disolvente*, *interacciones hidrofóbicas* y *contribuciones entrópicas*.

Interacciones de tipo *van der Waals* (vdW)

Cuando dos moléculas se aproximan y van entrando en contacto, las interacciones de vdW dan cuenta de dos tipos de fuerzas diferentes: a) *repulsión*, la cual actúa a corta distancia debido al solapamiento o superposición de las nubes electrónicas de los átomos que se acercan, y b) *atracción*, que se da a larga distancia y es debida a la correlación entre los electrones de los diferentes átomos (*fuerzas de dispersión de London*), y se debe más a la forma (o volumen) que propiamente al contenido electrostático. Ambas fuerzas dependen de la distancia entre los átomos (r) por lo que su representación es bastante directa. El modelo más usado es el del *potencial de Lennard-Jones*, donde el término repulsivo depende de r^{-12} y el atractivo, de r^{-6}.

Interacciones electrostáticas

Las interacciones electrostáticas están presentes en la mayor parte de los procesos de unión (interacciones carga-carga, enlaces de hidrógeno, apilamiento de nubes π o π-π *stacking*, interacciones hidrofóbicas, desolvatación...). Las fuerzas que rigen estas interacciones también se denominan de ajuste de la selectividad, un aspecto clave en el desarrollo de nuevos y más específicos. Sin embargo, su cálculo exacto sigue siendo uno de los mayores retos de nuestros días. La aproximación más simple es el *modelo Culómbico* (el producto de las cargas dividido por la distancia y una función dieléctrica sencilla que simula el apantallamiento ejercido por el disolvente). Otros métodos más complejos, y más costosos computacionalmente, están basados en el *modelo generalizado de Born* (GB) [36] o en la resolución de la *ecuación de Poisson-Boltzmann* (PB) [14].

Interacciones por enlace de hidrógeno

Se trata de una interacción muy selectiva y altamente dependiente de la orientación de sus constituyentes. Se establece entre un *átomo de hidrógeno* unido a un átomo electronegativo, llamado *donador* de enlace de hidrógeno, y otro átomo también electronegativo, llamado *aceptor* de enlace de hidrógeno. La fuerza del enlace depende de la posición relativa de los tres átomos implicados, es decir de las distancias y ángulos que haya entre ellos. Su papel en el reconocimiento molecular es importantísimo.

El efecto del disolvente

Todas las interacciones a tener en cuenta en biología tienen lugar en un entorno acuoso. Cuando las moléculas están aisladas en disolución, están completamente rodeadas de moléculas de agua. Sin embargo, cuando se produce la unión ligando-diana muchas de estas moléculas de agua son desplazadas. Este desplazamiento conlleva un gasto energético que debe ser contrarrestado por las nuevas interacciones formadas. Además, se produce una ganancia de entropía en las moléculas de agua liberadas. Desde un punto de vista teórico, hay dos modelos extremos para tener en cuenta los efectos del disolvente: a) *modelos de disolvente explícito*, donde las moléculas de agua están representadas con detalle atómico, y b) *modelos de disolvente implícito*, donde se construye una función matemática que trata de simular el comportamiento global del disolvente en función de su constante dieléctrica. También es posible considerar *modelos mixtos* en los cuales se tienen en cuenta explícitamente determinadas moléculas y el resto se consideran de manera implícita. En *docking* se suelen emplear modelos de disolvente implícitos ya que son lo suficientemente rápidos como para permitir un gran número de cálculos en un corto espacio de tiempo. Sin embargo, es necesario llegar a un compromiso entre exactitud y velocidad, cualidades que suelen estar inversamente relacionadas. Los métodos más populares son los ya citados GB y PB.

Interacciones hidrofóbicas

Algunos aminoácidos poseen cadenas laterales hidrófobas (leucina, valina, prolina...) al igual que muchos ligandos poseen partes hidrófobas en su estructura. Esto significa que no están bien, energéticamente, en un entorno acuoso. Si dos centros hidrófobos entran en contacto las moléculas de agua de su alrededor son liberadas y esta interacción (*efecto hidrofóbico*) produce una contribución positiva a la estabilización total de la unión.

Contribución entrópica

El concepto de *entropía* está íntimamente relacionado con la idea de orden. Una molécula aislada es libre de desplazarse, rotar y vibrar. Cuando se forma un complejo intermolecular, algunos de estos movimientos se pierden. Como consecuencia de ello, se establece un mayor orden en el sistema lo que lleva asociado una disminución de la entropía. Aparte de la entropía del disolvente, la *entropía del soluto* (o *entropía configuracional*) se suele dividir en dos partes: *conformacional* y *vibracional*. La parte conformacional tiene que ver con la reducción del número de pozos de energía que tanto el ligando como la proteína pueden visitar una vez que ha ocurrido la unión, mientras que la parte vibracional se refiere a los movimientos dentro de un pozo de energía en particular. La entropía es considerada como una propiedad difícil de calcular, y aunque tiene un papel importante en la estimación de la energía libre se suele ignorar.

Otras interacciones

En este apartado se incluyen las interacciones que dan lugar a la formación de *enlaces covalentes* entre el ligando y la diana produciendo inhibidores irreversibles, así como las *mediadas por moléculas de agua* específicas, *iones metálicos* o *átomos de halógenos*. Estas interacciones también son importantes y determinan en algunos casos la correcta predicción del modo de unión del ligando. Sin embargo, rara vez se tienen en cuenta o, si se hace, suele ser a un nivel teórico muy aproximado que está lejos de ser exacto.

16.2.4. *Docking*: consideraciones teóricas

En esta sección se trata el aspecto teórico de las herramientas comúnmente empleadas en *docking*. Sin embargo, antes de profundizar en este tema, es necesario comentar brevemente el modelo físico en el que se basa la unión entre el ligando y su diana.

Modelo físico de unión

La magnitud principal que determina la unión entre un ligando y su diana es la *energía libre de unión*, que se define como la diferencia de energías libres entre la correspondiente al complejo ligando-diana y la de sus respectivas especies aisladas (Ecuación 16.1) con las que se encuentra en equilibrio:

$$\Delta G_{unión} = -RT \ln K \tag{16.1}$$

siendo R la constante de los gases ideales, T la temperatura y K la constante de equilibrio. Esta constante se puede medir experimentalmente y compararla con la estimada por la Ecuación 16.1.

Como en cualquier otro equilibrio químico, la propiedad clave del sistema es el *potencial químico* (de donde deriva la energía libre de unión), que puede estimarse a través de la termodinámica estadística por medio de la función de partición. Una expresión común obtenida en términos de la energía potencial y el efecto del disolvente se muestra en la Ecuación 16.2 [10]:

$$\Delta G_{unión} = \langle U_{PL} \rangle - \langle U_P \rangle - \langle U_L \rangle + \langle W_{PL} \rangle - \langle W_P \rangle - \langle W_L \rangle - T\Delta S_{conf} \tag{16.2}$$

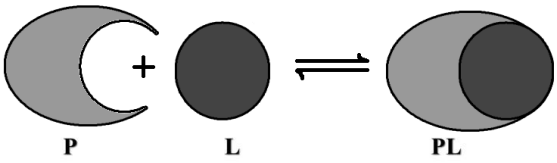
donde PL se refiere al complejo, P es la diana y L, el ligando. U es la energía potencial y W representa el efecto del disolvente. Todas estas magnitudes son energías promedio tipo Boltzmann, tal como indica el símbolo "$\langle\rangle$". El último término se refiere al cambio de entropía configuracional, donde normalmente se considera una sola configuración que representa al estado unido.

Actualmente hay tres modelos que representan la unión proteína-ligando. El primero de ellos data de 1894 y fue postulado por Emil Fischer [9]. Se trata del bien conocido modelo de *llave-cerradura* (*lock-and-key*): solo la llave correcta puede encajar en su cerradura. Se trata de una aproximación muy rígida, ya que no se consideran adaptaciones mutuas entre el ligando y la diana. Una aproximación más flexible, conocida como modelo de *acoplamiento inducido* (*induced fit*), fue posteriormente propuesta por Daniel Koshland en 1958 [24]. El acoplamiento inducido considera que la flexibilidad intrínseca de la diana se traduce en una reorganización de su centro activo para acomodar a los ligandos entrantes. Por último, el modelo de *selección conformacional* (*conformational selection*) (1999) [3], postula que es el ligando el que selecciona, de entre un conjunto de conformaciones accesibles de la diana, la más apropiada para su unión. En la Figura 16.3 puede verse una representación gráfica de los tres modelos.

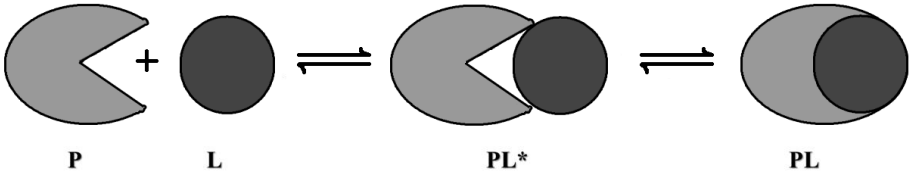
La función de puntuación o scoring

Una *función de puntuación*, o *scoring*, es una ecuación matemática que nos da un valor con el que cuantificar la fortaleza con la que un ligando se une a su diana. El paisaje energético de esta unión suele ser extremadamente complejo, con un gran número de *valles* (mínimos) y *montañas* (máximos) (Figura 16.3C). Se espera de la función de *scoring* que sea capaz de recorrer este paisaje y localizar los mínimos, ya que estos se corresponden con situaciones (configuraciones ligando-diana) plausibles en las que se tiene un complejo estable (Figura 16.3A y Figura 16.3B), siendo alguna de ellas similar a la

a) Modelo “Lock and Key”



b) Modelo “Induced Fit”



c) Modelo “Conformational selection”

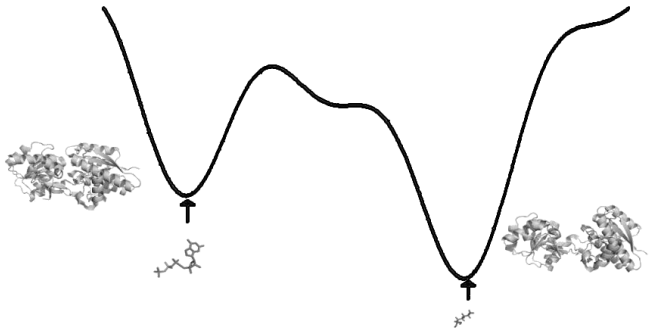


Figura 16.3: Representación gráfica de los tres modelos de unión más comunes.

experimental. Además, debe de hacerlo lo suficientemente rápido como para evaluar el elevado número de posibles configuraciones que se generan en los estudios de interacción ligando-diana, y seleccionar, de todos los mínimos posibles, aquel que corresponde a la estructura experimental como la mejor de las soluciones. Por tanto, una función de *scoring* debería ser tanto computacionalmente eficiente como fiable.

Las funciones de *scoring* se utilizan en distintas etapas del proceso de descubrimiento de nuevos fármacos, desde la identificación de un *hit* en cribado virtual hasta su optimización a *lead*, para evaluar las distintas poses en *docking*, para la identificación de ligandos de alta afinidad y para predecir afinidades de unión.

Hay tres tipos principales de funciones de *scoring* que se diferencian entre sí por los datos usados en su derivación: empíricos, basados en el conocimiento y basados en campos de fuerzas.

Las *funciones de scoring basadas en datos empíricos* son aquellas obtenidas a través de un análisis de regresión multilineal entre medidas experimentales de actividad y una serie de propiedades consideradas como fundamentales para que se produzca la unión, como interacciones por enlace de hidrógeno, interacciones iónicas, contactos polares y no polares. La contribución de cada una de estas interacciones a la energía de unión total viene ponderada por un coeficiente que se obtiene usando un conjunto de prueba (*training set*) de complejos ligando-diana. Es por ello que estas funciones son de limitada aplicación más allá del conjunto de prueba (Ecuación 16.3):

$$\Delta G_{unión} = \Delta G_0 + \Delta G_{hb} f(hb) + \Delta G_{ionic} f(ionic) + \Delta G_{lipo} f(lipo) + \Delta G_{rot} f(rot) \tag{16.3}$$

donde cada $f(int)$ es una función que representa a un tipo de interacción (int), y ΔG_{int} es el coeficiente obtenido de la ecuación de regresión y que da peso a la contribución relativa de cada tipo de interacción, siendo ΔG_0 una constante.

Las *funciones de scoring basadas en el conocimiento* hacen uso de las bases de datos de estructuras 3D para buscar qué tipos de interacción suceden más comúnmente entre ligandos y dianas y con qué frecuencia. Las frecuencias se convierten en energía libre usando la *fórmula inversa de Boltzmann* (Ecuación 16.4) y definiendo un *estado de referencia* que corresponde a aquél en el que no existe tal interacción. Este último punto es precisamente la principal debilidad de estos métodos, al no haber una manera única, ni sencilla, de definir el estado de referencia:

$$\Delta G_{unión} = \sum_{ij} A_{ij}(r) = -kT \sum_{ij} \ln \left(\frac{g_{ij}(r)}{g(r)} \right) \tag{16.4}$$

donde $A_{ij}(r)$ es una función que describe las interacciones entre los átomos tipo i y j (suma total de todos los átomos ligando-diana como una función de la distancia r), $g_{ij}(r)$ es la probabilidad de que los átomos i y j estén a una distancia r, $g(r)$ es la probabilidad del estado de referencia, k es la constante de Boltzmann y T la temperatura. Estas funciones también se conocen como *potenciales de fuerza media o potenciales estadísticos.*

Las *funciones de scoring basadas en campos de fuerzas* descomponen la energía de unión ligando-diana en la suma de una serie de términos de interacción individuales, tales como vdW, electrostático, enlace de hidrógeno, etc... que en su definición utilizan parámetros de mecánica molecular. Una función típica basada en un campo de fuerzas se muestra en la Ecuación 16.5, donde solo se han considerado las

interacciones de van der Waals y electrostáticas.

$$\Delta G_{unión} = \sum_{ij} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}} \right] \tag{16.5}$$

A_{ij} y B_{ij} son los parámetros de vdW que dependen del tipo de átomo asignado a i y j, r_{ij} es la distancia entre el i-ésimo átomo de la diana y el j-ésimo átomo del ligando, q_i y q_j son las cargas parciales de los átomos i y j respectivamente, y ε es la constante dieléctrica del solvente.

16.2.5. El proceso de docking

Para tener en cuenta el número tan elevado de grados de libertad que se manejan en *docking*, se han desarrollado diversas metodologías, que de manera general se pueden englobar en tres aproximaciones: *docking rígido*, *docking con proteína rígida y ligando flexible* y *docking flexible*.

Docking rígido

En la aproximación de *docking rígido* se considera que tanto el ligando como la diana son componentes rígidos, sin grados de libertad internos. Esto reduce la complejidad del problema a rotaciones y traslaciones del ligando en relación a la diana. Sin embargo, esta aproximación es demasiado simplista, ya que ambas especies son entidades de naturaleza flexible. No obstante, es un método muy usado como primera alternativa en programas de *docking* nuevos.

Docking de diana rígida y ligando flexible

Asumiendo que la aproximación de llave-cerradura es aceptable como modelo de unión ligando-diana, la flexibilidad de la proteína se puede omitir. De esta manera la *búsqueda conformacional del ligando* se convierte en la parte más importante del problema a resolver. En muchos casos la exploración sistemática no es siempre posible, debido al efecto de la explosión combinatoria que supone la enumeración de todas las posibles rotaciones de cada ángulo de torsional en el ligando, y es por tanto necesario recurrir a otras aproximaciones. Hay multitud de métodos o algoritmos implementados en los programas de *docking* aunque se pueden clasificar, de forma general, en tres categorías principales: construcción incremental, pre-cálculo de las conformaciones del ligando o su generación in situ. La *construcción incremental* implica el análisis conformacional *on the fly* (al vuelo) dentro de las limitaciones del sitio de unión, por división del ligando en fragmentos y su posterior unión de manera secuencial (éste es el método usado en los programas FlexX [31] o Surflex [18]). En el método de *conformaciones del ligando pre-calculadas*, las conformaciones del ligando son generadas antes de la operación de *docking* y guardadas en una base de datos para su uso posterior (así funcionan los programas CRDOCK [6] y GLIDE [32]). Por último, cuando las *conformaciones* son *generadas in situ*, la totalidad del ligando se adapta de forma continua al sitio activo de la diana. En este último caso, las técnicas más empleadas son: a) *complementariedad de forma*, basada en la evaluación de la concordancia entre la conformación del ligando y el sitio activo de la diana en términos geométricos (es el caso del programa DOCK [8]); b) *algoritmos genéticos*, basados en la teoría de la evolución de Darwin o en la teoría de la herencia de Lamarck (como los empleados en los programas GOLD [41] o AutoDock [11]); c) *algoritmo de Monte Carlo*, generación aleatoria de grupos de rotación, traslación y orientación de los ligandos y evaluándolos después con una función de scoring (como en el programa LigandFit [40]); d) *búsqueda tabú*, basada

en la generación de poses de forma aleatoria llevando una lista de los sitios o conformaciones que ya han sido probadas (es el caso del programa PRO_LEADS [28]); y e) *algoritmos bio-inspirados*, basados en la inteligencia de los enjambres o en las estrategias de las colonias de hormigas (como el programa PLANTS [23]).

Docking flexible

Con respecto a la diana, y siempre que se incluya la flexibilidad del ligando, los métodos se clasifican en función del grado de flexibilidad que esta incorpore. Una primera aproximación es la conocida como *soft docking*, donde se realiza una relajación de los potenciales de interacción, que trae como consecuencia una expansión en las dimensiones del centro activo, lo que simulan el efecto del ajuste o acoplamiento inducido (esta aproximación se usa en los programas Glide y GOLD). El siguiente paso es el uso, a través de un *algoritmo de Monte Carlo*, de una colección de rotámeros para probar los cambios en las cadenas laterales de la diana, como se hace en los programas GOLD, Glide, AutoDock, FlexX e ICM [37]. Si se dispone de varias estructuras de la misma diana, puede usarse el *esquema del complejo relajado*, que consiste en realizar diferentes experimentos de *docking* de forma individual y hacer luego una promedio con los resultados. Este proceso también se puede emplear con estructuras generadas por simulaciones de dinámica molecular (Capítulo 17) o análisis de modos normales (Capítulo 18). Ambas técnicas se han empleado también para tener en cuenta la flexibilidad total de la diana, mientras que al mismo tiempo se hace el *docking* del ligando. Los programas AutoDock, ICM, Glide o GOLD permiten realizar este tipo de *docking* usando diferentes conformaciones para la diana.

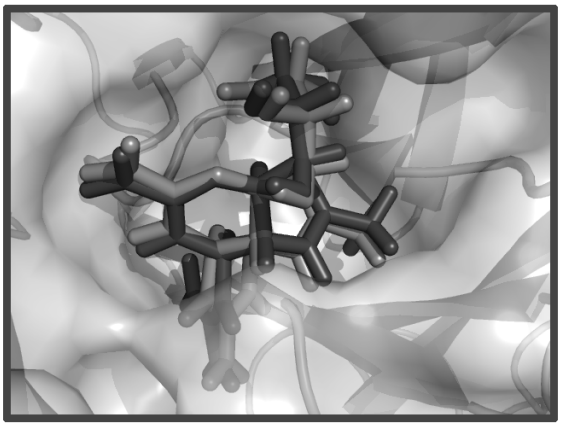
La etapa de re-scoring

Es bien conocido que la clasificación de las poses de *docking* atendiendo a las funciones de *scoring* no garantiza siempre que la mejor solución de la lista sea la correcta. Es por ello aconsejable considerar una descripción más detallada del proceso de unión incluyendo funciones de *scoring* más precisas y complejas en etapas posteriores a las aproximaciones iniciales. Algunos métodos usan una aproximación basada en *factores de ponderación*, de manera que a la puntuación original es escalada por un factor que depende de aspectos geométricos o propiedades *drug-like*.

16.2.6. El problema de docking: cómo evaluar la validez de los resultados

La evaluación y ordenación de las múltiples soluciones predichas por el algoritmo de búsqueda son los aspectos más críticos de los protocolos de *docking*. La función de *scoring* debería representar de la forma más adecuada posible la termodinámica de la unión ligando-diana para ser capaz de diferenciar el verdadero modo de unión entre todos los demás.

Al menos son necesarias dos cosas para probar la precisión de un nuevo programa de *docking* y su función (o funciones) de *scoring*: un conjunto de complejos con estructura 3D conocida y una medida para cuantificar cómo se parecen los resultados de *docking* a las estructuras experimentales. Aquí, el uso de la *desviación cuadrática media*, o *RMSD* de sus siglas en inglés *root-mean-square deviation*, es el más extendido (Figura 16.4). Con esto nos referimos a la parte estructural del *docking*. Además, si se tienen datos de afinidad/actividad experimental del ligando por la diana, se pueden evaluar cómo de bien las funciones de *scoring* reproducen estos valores. Esta es la parte energética del *docking*.



Negro = Estructura Obtenida Experimentalmente

Gris = Estructura Predicha por Docking

$$RMSD = \sqrt{\frac{\sum_{i=1}^{n} (x_{1,i} - x_{2,i})^2}{n}} = 0.66$$

$x_{1,i} = \{número\ de\ átomo\ "i",\ estructura\ experimental\ \}$

$x_{2,i} = \{número\ de\ átomo\ "i",\ estructura\ predicha\ \}$

$n = \{número\ total\ de\ átomos\}$

Figura 16.4: Representación gráfica del RMSD.

Evaluación estructural

En un estudio de *docking* típico, y con el fin de ver si nuestro algoritmo es adecuado, dado un conjunto de complejos ligando-diana, los ligandos se extraen de sus complejos y se hace *docking* de cada uno de ellos en su propia diana. Después, se calcula el RMSD para cada pose usando como referencia la estructura experimental. Si el valor del RMSD de la pose elegida como la mejor de las soluciones está por debajo de 1.0 Å se considera que el resultado es aceptable, aunque muchos autores aumentan este valor de corte hasta 1.5 Å o incluso hasta 2.0 Å (Figura 16.4). Por encima de este valor límite los resultados se consideran incorrectos. Otro parámetro para probar la eficacia del método de *docking* es el porcentaje de acierto o éxito, definido como el porcentaje de estructuras predichas con valores de RMSD por debajo de 2.0 Å. El valor medio para la mayoría de los programas de *docking* está en torno al 70-75 %, aunque pueden encontrarse porcentajes mayores dependiendo del programa y del conjunto de complejos ligando-diana utilizado en su evaluación.

Evaluación energética

Esta es, con diferencia, la evaluación más complicada en los cálculos de *docking*. La función de *scoring* debe ser capaz de distinguir y elegir como mejor solución, de entre todas las posibles, aquella con el menor valor de RMSD con respecto a la estructura experimental. Generalmente se asume que el problema de muestreo (encontrar la pose correcta) está más o menos resuelto, pero la baja correlación encontrada entre los valores de energía calculados con la función de *scoring* y el RMSD indican lo contrario.

Tratar de predecir valores de afinidad/actividad es algo incluso más exigente. Un estudio bastante extensivo empleando diferentes dianas y combinaciones de programas de *docking* y funciones de *scoring* demuestra la falta de una correlación significativa entre valores de afinidad/actividad y las energías libres de unión calculadas. Esto significa que la función de *scoring*, aun siendo capaz de reproducir estructuras experimentales de forma precisa, no es suficientemente buena para la predicción de afinidad/actividad, principalmente debido al hecho de que sacrifica precisión en favor de la velocidad de cálculo. En otras palabras, los principios físicos subyacentes que gobiernan las interacciones ligando-diana no están debidamente implementados. De hecho, la entropía, los efectos del disolvente y la flexibilidad de la proteína raramente se tienen en cuenta o se usan a un nivel teórico muy bajo, a pesar del aumento en el poder de computación. Sin embargo, en un futuro cercano se esperan avances significativos en esta área.

16.3. Cribado virtual

El objetivo principal del *cribado virtual* (VS, siglas que en inglés hacen referencia a *Virtual Screening*) es diferenciar, de entre un conjunto de pequeñas moléculas o ligandos (*quimioteca*), aquellas que teóricamente pueden encajar en una cavidad para bloquear/activar la función de una determinada diana (*ligandos verdaderos*) de las que no (*falsos ligandos*) [34]. Se trata de una alternativa teórica al método experimental conocido como *High-Throughput Screening* (HTS) o *cribado farmacológico de alto rendimiento*, pero con un coste y una necesidad de recursos mucho más reducidos.

Dado que los métodos teóricos más fiables son a su vez lo más costosos computacionalmente, y que el número de moléculas con el que estamos tratando es bastante elevado (del orden de millones), el protocolo de VS se configura como una serie de filtros sucesivos donde la complejidad del filtro aumenta a medida que el número de moléculas va disminuyendo. Normalmente se representa como un embudo

con la parte ancha hacia arriba y la estrecha hacia abajo, lo que da una idea de la reducción en el número de moléculas a medida que se avanza en el protocolo (Figura 16.5).

El VS está ganando aceptación dentro del campo del diseño de fármacos y cada vez contribuye con más moléculas como nuevos candidatos a fármaco. Sin embargo, esta técnica aún no está en un grado de desarrollo que podamos considerar maduro. Aun así, se están produciendo avances muy prometedores en la metodología que ya están produciendo buenos resultados.

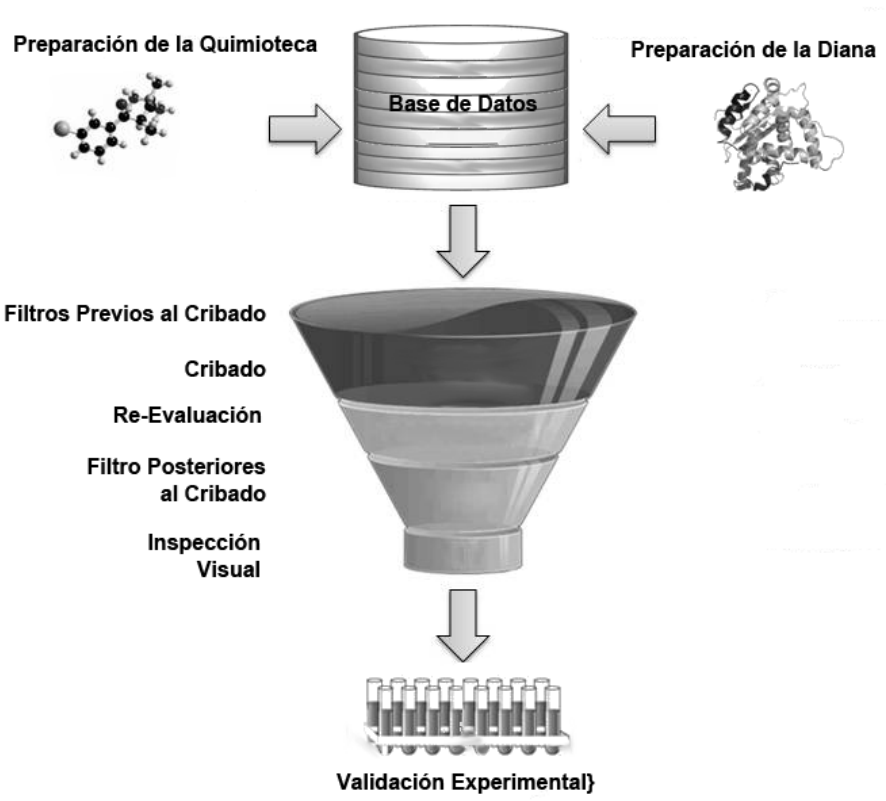


Figura 16.5: Representación de un protocolo general de cribado virtual.

16.3.1. Posibles escenarios para el VS

Dependiendo de la información estructural de la que dispongamos, el VS se puede dividir en dos grandes grupos: el *VS basado en la estructura* (SBVS, siglas que en inglés corresponden a *Structure-Based Virtual Screening*), cuando la estructura 3D de la diana es conocida, y b) el *VS basado en el ligando* (LBVS, siglas que en inglés corresponden a *Ligand-Based Virtual Screening*), cuando lo que se conoce es la estructura de un grupo de ligandos activos (o no) frente a la diana de interés. En el primer caso la técnica más utilizada es *docking*, mientras que el segundo caso la estrategia dominante está basada en seleccionar motivos comunes entre las moléculas activas (o las inactivas) para definir lo que se denomina un *farmacóforo*. El farmacóforo se emplea entonces como un molde o plantilla sobre la que se realizan cálculos de semejanza molecular con los compuestos de la quimioteca con el fin de recuperar aquellos que más se parezcan a la plantilla.

En comparación, el número de aplicaciones publicadas donde se aplica el SBVS es aproximadamente el

triple al correspondiente a LBVS, aunque este último ha demostrado ser más eficiente en la identificación de nuevas moléculas que el SBVS. Sin embargo, quizás el mejor enfoque sería combinar, en la medida de lo posible, las dos técnicas, de manera que se pueda aprovechar toda la información disponible relativa al problema que se está tratando. Esta estrategia debería sin duda aumentar nuestras probabilidades de éxito. De hecho, en muchos de los estudios de VS en los cuales se han combinado ambos métodos, los candidatos seleccionados han resultado ser los mejores en términos de actividad. En la Figura 16.5 se describe un protocolo general de VS.

16.3.2. Estudios de VS retrospectivos y prospectivos

La palabra *retrospectivo* hace aquí alusión a aquellos estudios de VS donde los resultados ya se conocen de antemano, y por lo tanto son muy útiles en la validación de nuevos protocolos. Por lo general, el conjunto de estudio consiste en unas cuantas moléculas cuya actividad frente a la diana de interés ya ha sido confirmada, y una serie de señuelos (*decoys*) que se suponen sin actividad frente a la misma diana. Como se ha mencionado antes, el experimento consiste en ver si el método elegido para realizar el VS es capaz de distinguir entre ambos grupos de moléculas. La elección de los señuelos es un punto muy delicado y en gran medida el factor determinante de que los resultados sean de alguna manera significativos y el método pueda aplicarse de manera prospectiva (ver más adelante) en estudios reales de búsqueda de nuevos fármacos. En particular, los señuelos deben de ser estructuralmente distintos a los compuestos activos pero con las mismas propiedades físico-químicas. De esta manera se evita un posible sesgo de que las moléculas sean seleccionadas simplemente porque se parezcan a las activas. Una fuente bien establecida de conjuntos de datos para estudios de VS es DUD [15] (acrónimo del inglés *Directory of Useful Decoys*) y su versión mejorada DUD-E [29] (la E corresponde a *Enhanced*).

Al igual que en docking, para hacer una valoración retrospectiva de un protocolo de VS además de un conjunto adecuado de datos (activos + señuelos) se necesita una medida para evaluar la eficiencia del método. En general, una medida fiable debería: a) ser independiente de variables extensivas (es decir, aquellas que dependan del número de moléculas activas, señuelos, o incluso de las dianas); b) ser suficientemente robusta; c) plantear un modo directo de evaluar el error; d) no contener ningún parámetro libre; y e) ser interpretable y fácilmente entendible. Las medidas más comunes son el *factor de enriquecimiento* (EF, acrónimo que en inglés corresponden a *Enrichment Factor*) y el *área bajo la curva* (AUC, acrónimo que en inglés corresponden a *Area Under the Curve*) de una representación ROC (acrónimo de *Receiver Operating Characteristic*). El EF mide, para un porcentaje determinado del conjunto de estudio, la relación entre la cantidad de compuestos activos recuperados y la cantidad que de estos se habrían seleccionado si se hubiera hecho al azar. El AUC de una curva ROC representa la fracción de veces que una molécula activa seleccionada al azar poseerá una mayor puntuación que una inactiva también seleccionada al azar. En la Figura 16.6 se muestra un ejemplo de cada una de las curvas así como las fórmulas necesarias para su cálculo.

Por otro lado, en los *estudios prospectivos* de VS no tenemos un conocimiento previo sobre el tipo de compuestos que podrían unirse o no a una diana en particular. Su objetivo es por tanto intentar encontrar candidatos mediante el cribado de quimiotecas. Estos estudios son la prueba final de la valía de cualquier protocolo de VS. La técnica claramente dominante es el *docking*, pero debido a su carga computacional, se usan filtros basados, por ejemplo, en puntos farmacofóricos obtenidos de la estructura de la diana, con el objetivo de reducir el tamaño de la quimioteca. Es necesario señalar de nuevo que en un experimento de VS basado en *docking* la función de *scoring* se evalúa no sólo por su habilidad en diferenciar entre ligandos verdaderos y falsos (compuestos activos e inactivos) a través de la asignación correcta de puntuaciones a cada una de las poses de docking, sino también por su habilidad para recuperar tantos quimiotipos (diferentes fragmentos o bloques químicos) como sea posible con el fin de

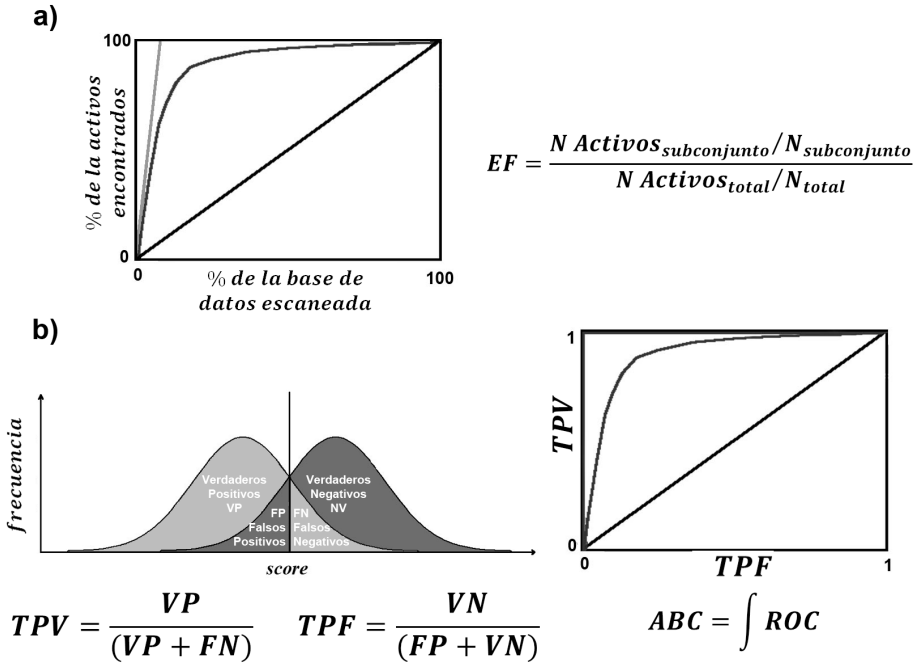


Figura 16.6: Curvas EF y ROC con sus ecuaciones.

lograr una mayor diversidad.

16.3.3. Realización de estudios de VS: herramientas e infraestructuras

La cantidad de datos que hay que manejar en los estudios de VS es enorme. Por lo tanto, el método tradicional de trabajar con una diana, unos pocos ligandos, un programa de *docking* y varios ficheros es inapropiado. El desafío del VS, al menos de una manera genérica, es cómo cribar quimiotecas que contienen millones de compuestos en un periodo de tiempo razonable con un cierto grado de confianza. Hoy en día es factible realizar LBVS en ordenadores personales, ya que es posible comparar varios millones de moléculas por CPU al día sin mayor problema. Sin embargo, para protocolos más elaborados es necesario recurrir a la *computación de alto rendimiento* (HPC, acrónimo del inglés *High Performance Computing*), a la *computación en grid* (*grid computing*), o a la *computación en la nube* (*cloud computing*).

En HPC la aproximación más directa es procesar los ligandos usando un esquema en paralelo donde la quimioteca se divide en N particiones iguales correspondientes al número de CPUs disponibles. *Grid computing* hace referencia a un conjunto heterogéneo de dispositivos de cálculo conectados entre sí a través de la red y que pertenecen y son administradas por distintas instituciones. La *computación voluntaria* (VC, acrónimo del inglés *Volunteer Computing*) es un caso extremo de computación grid. La diferencia entre computación grid y HPC estriba principalmente en que la primera tiende a ser más heterogénea y a estar geográficamente más dispersa que la segunda, aprovechando la gran cantidad de ordenadores conectados a través de Internet y proporcionando un poder de cálculo virtualmente infinito, mucho más allá de cualquier centro de supercomputación. Algunos de los proyectos de VC más

conocidos son SETI@home [1] y FOLDING@home [2]. También hay ejemplos donde se usa *docking* y VS como el proyecto Screensaver-Lifesaver (proyecto donde se emplearon más de 1.5 millones de PCs y se cribaron del orden de 3500 millones de compuestos frente a dianas contra el cáncer) [3], Docking@Home (enfocado en la búsqueda de nuevos fármacos contra el virus del sida) [4], WISDOM (para encontrar nuevos inhibidores contra la malaria) [20] o Ibercivis (para encontrar nuevos inhibidores contra el cáncer o la enfermedad de Alzheimer) [5]. Por último, cloud computing es una tecnología emergente que se refiere a la provisión de recursos computacionales bajo demanda por medio de una red de ordenadores, y la principal ventaja es que libera a los usuarios de la dependencia de cierto hardware y software.

Además de estas infraestructuras, hay una tendencia natural hacia la automatización de las tareas relacionadas con VS proporcionando una *Interfaz Gráfica de Usuario* (GUI, acrónimo del inglés *Graphic User Interface*) que facilite la definición de los diferentes pasos del protocolo: preparación de ligando y diana, *docking* y visualización de los resultados. La GUI da acceso a las principales capacidades implementadas en el programa de *docking* de una forma sencilla, y la mayoría de estos programas poseen ya una GUI (como BDT [39] y DOVIS [19] para AutoDock). Otra alternativa consiste en desarrollar *conectores* (*plugins* en inglés, un componente de software que añade cierta funcionalidad a una aplicación) que pueden implementarse en una interfaz tipo PyMOL [6] y parece ser la tendencia actual dada la cantidad de estos que han visto la luz últimamente (AutoDock/Vina [33], AMBER/AutoDock/SLIDE [26]). Un paso más allá son las llamadas plataformas de VS, es decir, sistemas más sofisticados que integran multitud de piezas que hacen posible de manera sencilla configurar protocolos de VS más complejos (Pipeline Pilot [13], DVSDMS [42], SOMA [25] y VSDMIP [5]). Una plataforma de este tipo debe hacer frente a la gran cantidad de datos de la forma más eficiente posible, lo que implica incluir un motor de base de datos por debajo de cada operación de VS. Aunque este tema parece ser de conocimiento e interés general, muy pocas plataformas la incluyen. Por último, otro ejemplo interesante es DOCK Blaster [17], donde se puede realizar un estudio completo de VS usando como entrada solamente el código PDB de la diana deseada, con todas las herramientas necesarias implementadas en una plataforma vía web.

16.4. Docking proteína-proteína

Al igual que en el estudio de las interacciones entre ligandos y sus dianas, entender la manera en la que se producen las interacciones entre proteínas aporta información sobre si esa unión está relacionada con su función y cómo manejarla con fines terapéuticos.

El punto de partida para estudiar el modo de unión entre dos proteínas mediante herramientas de *docking* es el conocimiento de las correspondientes estructuras tridimensionales de ambos sistemas, procedentes de la cristalografía de rayos-X, la espectroscopía de RMN, la criomicroscopía electrónica o bien modeladas de forma teórica.

Los diferentes algoritmos que se emplean en los estudios de *docking* proteína-proteína tienen como objetivo obtener una lista ordenada de todas las posibles *soluciones* (*poses*), entre las cuales y al igual que en el *docking* ligando-diana, debe de haber una lo más semejante posible a la estructura nativa (experimental) del complejo. Obviamente, en la mayoría de los casos, la estructura del complejo formado entre las proteínas a estudiar es desconocida y es ahí donde juega un papel importante cómo evaluamos

[1] Proyecto SETI@home project. `http://setiathome.berkeley.edu`
[2] Proyecto FOLDING@home. `http://folding.stanford.edu`
[3] Proyecto Screensaver-Lifesaver. `http://www.chem.ox.ac.uk/curecancer.html`
[4] Proyecto Docking@Home. `http://docking.cis.udel.edu`
[5] Proyecto Ibercivis. `http://www.ibercivis.es`
[6] The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC. `http://www.pymol.org`

la calidad de las diferentes poses de *docking*. Esto se suele llevar a cabo a través de una función de *puntuación* (*scoring*), la cual nos ayuda a producir una *priorización* (*ranking*) de las soluciones.

La fuerza que conduce a la unión entre proteínas se corresponde con el cambio de energía libre asociado a dicho proceso, lo que a su vez depende de las propiedades físico-químicas y estructurales de las proteínas implicadas tanto de una manera global (superficie polar/apolar) como de una manera más local a través de la interacción entre determinados residuos que pueden ser claves para la estabilidad del complejo que se forma.

La predicción de la unión entre dos proteínas representa un problema de mayor complejidad en comparación con el que hemos visto anteriormente sobre interacciones proteína-ligando, ya que el número de grados de libertad en este caso es mucho mayor. Tratar de reproducir de forma detallada las interacciones que se producen en cada una de las posibles poses usando métodos como dinámica molecular puede convertirse en algo totalmente inabordable con los medios computacionales actuales. En cualquier caso, si se conocen las zonas de unión en la superficie de las proteínas, el número de posibilidades del problema se reduce drásticamente, aumentando así las posibilidades de éxito. Para añadir un punto extra a la complejidad del problema, hay que tener en cuenta que las proteínas no son en absoluto estáticas (tal y como se pone de manifiesto en diferentes capítulos de este libro), y que las estructuras de una proteína antes y después de unirse pueden ser significativamente diferentes. Esto implica que para predecir de manera precisa dicha unión, en muchos casos debemos tener en cuenta la flexibilidad de las estructuras. Este problema se ha abordado desde diferentes puntos de vista y nos puede ayudar a entender las capacidades y limitaciones de los diferentes algoritmos de *docking* que vamos a tratar a continuación.

16.4.1. El proceso de docking

Existen muchos métodos computacionales capaces de generar cientos de posibles poses entre las dos proteínas sometidas a estudio. El primer paso suele consistir en una búsqueda sistemática de las posibles geometrías de unión, seguido de una evaluación a través de una función de puntuación, para finalizar con una etapa de refinado de las estructuras más prometedoras, es decir, de aquellas que hayan obtenido una mejor puntuación. Este proceso se puede repetir varias veces, dependiendo del protocolo utilizado por los diferentes programas de *docking*.

Docking rígido

Está demostrado que un porcentaje importante de las proteínas estudiadas hasta nuestros días presentan movimientos relativamente pequeños una vez que se produce su unión a otras. Por tanto, en estos casos podemos considerar el *docking rígido* como una buena aproximación. Como veremos más adelante, es importante refinar las soluciones más prometedoras para tener en cuenta estos movimientos (por pequeños que sean) y así reproducir las interacciones nativas presentes en el complejo con un mayor acierto.

Dentro de las estrategias más comunes para realizar *docking* rígido entre proteínas están los métodos de *indexado geométrico*, donde la superficie de la proteína se reduce a una serie de descriptores y se buscan las partes de ambas proteínas que puedan encajar entre sí [30], o los métodos basados en *correlaciones entre las Transformadas Rápidas de Fourier* (TRF), que localizan de forma muy eficiente los solapamientos entre las superficies de las proteínas [21].

La aproximación basada en el método de *indexado geométrico* se usó originalmente como una técnica de visualización computacional, para ajustar uno o más conjuntos de datos. En este tipo de algorit-

mos se reduce cada proteína a un conjunto de triángulos que se almacenan en tablas indexadas, las cuales permiten la búsqueda de triángulos de manera rápida. Dado que estos triángulos representan puntos en la superficie con ciertas propiedades geométricas (concavidad/convexidad) y físico-químicas (hidrofobicidad/hidrofilicidad) podemos buscar triángulos coincidentes con propiedades geométricas o físico-químicas complementarias. Mediante este proceso, podemos evaluar las diferentes poses generadas a través de rotaciones y traslaciones de una de las proteínas (considerada como ligando) alrededor de la otra (considerada como receptor).

Desde otro punto de vista, en la *aproximación basada en TRF*, cada proteína se representa como una malla cúbica donde a cada punto de la malla se le asigna un identificador dependiendo de si pertenece al interior de la proteína, a su superficie o al exterior. Utilizando métodos geométricos se pueden superponer las superficies de las mallas del receptor y del ligando y calcular la bondad del ajuste. El problema de esta aproximación es de nuevo la capacidad de cálculo de los ordenadores disponibles hoy en día, por lo que se han desarrollado aproximaciones más eficientes basadas en el teorema de correlación de Fourier. La *transformación discreta de Fourier* de las mallas correspondientes al receptor sólo se calculan una vez, y una vez para cada orientación del ligando. El problema de tener que pre-calcular todas las rotaciones deseadas del ligando en el espacio de coordenadas cartesianas para después transformarlo a sus equivalentes en el espacio de Fourier, se puede evitar correlacionando bases de funciones polares esféricas que representen la forma de la superficie de la proteína.

Docking flexible

Si las proteínas que forman el complejo experimentan cambios conformacionales apreciables una vez que este se ha formado, es difícil que obtengamos alguna solución parecida a la nativa usando *docking* rígido incluso con un refinado posterior de las mejores soluciones. Predecir los cambios de conformación de las proteínas es una tarea costosa en términos computacionales, así como compleja desde el punto de vista de la descripción física de la unión, y por ello, hasta el momento, no se ha conseguido desarrollar una estrategia directa que nos permita confiar en las soluciones obtenidas. En cualquier caso, ya que este es un problema de gran interés para entender cómo funciona la maquinaria celular, la comunidad científica está desarrollando múltiples estrategias que, en algunos casos, ya se ha conseguido aplicar con éxito. Alguna de ellas se revisa a continuación.

Si sabemos que el cambio conformacional es grande sólo en una de las proteínas implicadas en la formación del complejo, podemos intentar reproducir sus conformaciones mediante técnicas de dinámica molecular [12] (Capítulo 17), modos normales (Capítulo 18) o incluso RMN (Sección 7.7), generando así un conjunto de estructuras iniciales. En algunos casos encontraremos que la estructura nativa de esta proteína está dentro del conjunto de estructuras pre-generadas, por lo que es razonable pasar a la aproximación de *docking* rígido. Todas las poses que se obtengan serán evaluadas seleccionando aquellas que resulten con las mejores puntuaciones. Al usar esta aproximación debemos tener en cuenta que se puede producir un número significativo de falsos positivos ya que en algunos casos podemos encontrar poses con una buena puntuación debido a un buen ajuste de las superficies de interacción, pero que no sean similares a la estructura nativa.

En otros casos podemos tener ciertas evidencias sobre la localización del sitio de unión entre ambas proteínas o incluso sobre algunas de las interacciones que se establecen cuando se forma el complejo. De esta manera podemos reducir de forma importante el número de poses posibles. Aquí es más factible el uso de *docking* y dinámica molecular de forma simultánea. En este modelo es posible incluir la flexibilidad de toda la proteína o de una cierta región durante el *docking*, por lo que, en principio, es un método parecido al escenario real en el que se produce la formación de complejos entre proteínas.

16.4.2. Clasificación y post-procesado de las soluciones

Una función ideal de evaluación de poses de *docking* debería ser capaz de reconocer contactos similares a los que presenta la estructura nativa, y a la vez, de diferenciarlos de aquellos que no lo son. Las *funciones de evaluación* pueden estar basadas en *campos de fuerza*, en las que diferentes contribuciones energéticas se encuentran ponderadas, o pueden desarrollar *potenciales estadísticos* basados en la información estructural que se encuentra en las bases de datos que contienen las estructuras experimentales de los complejos. Normalmente, un único *descriptor* (como la complementariedad de superficie) o un *término energético* (vdW o electrostático) no suele ser capaz de distinguir entre poses nativas y no nativas, por lo que normalmente se utiliza una combinación de varios términos.

Las *funciones de energía basadas en campos de fuerza* son semejantes a las descritas anteriormente para *docking* proteína-ligando (Sección 16.2), aunque algunos de sus parámetros pueden estar ajustados a tipos de átomo específicos pertenecientes a los aminoácidos que forman parte de la estructura de las proteínas [2]. Las *funciones empíricas*, al estar basadas en un modelo de regresión estadística entre la actividad experimental y ciertas propiedades de los modos de unión nativos, se entrenan específicamente con un conjunto determinado de complejos proteína-proteína, por lo que su aplicabilidad es reducida [7]. Finalmente, las funciones de energía desarrolladas como *potenciales estadísticos* se basan en el análisis de la frecuencia con la que se observan ciertas interacciones entre determinados residuos en las estructuras experimentales de los complejos proteína-proteína. Este análisis estadístico se puede llevar a cabo tanto a nivel de contactos entre residuos, o de forma más detallada, contactos entre átomos. Basándonos en esta información, podemos construir una función de evaluación basada en el conocimiento (*knowledge-based*) donde se evalúa cada pose comparando sus contactos con la frecuencia con que estos se producen en la base de datos de estructuras [16].

Las poses obtenidas por un algoritmo de *docking* requieren siempre de un *refinado* posterior mediante el ajuste fino de las posiciones atómicas, lo que normalmente implica la reorientación de las cadenas laterales de los residuos, así como ciertos movimientos en las asas que conectan otros elementos de estructura secundaria. El éxito del refinado de las soluciones depende de que el conjunto de estructuras seleccionadas contenga alguna solución lo suficientemente parecida a la estructura nativa. Por lo tanto, en la clasificación de soluciones no sólo se necesita reconocer y preseleccionar modos de unión cercanos a la estructura nativa, sino que a la vez tiene que tolerar ciertas imprecisiones en la interfaz de contacto proteína-proteína para dar la oportunidad al refinado de reconstruir adecuadamente las interacciones nativas.

Con anterioridad al refinado de las soluciones se suelen aplicar *algoritmos de agrupamiento* (*clustering*) para reducir el número de candidatos. Estos algoritmos producen subconjuntos de soluciones de los que se escoge como representante de cada subconjunto aquella pose que tiene la mejor puntuación.

El *refinado de soluciones* puede llevarse a cabo mediante técnicas de *minimización de la energía*, donde la descripción de la estructura del complejo está basada en un campo de fuerzas de mecánica molecular (ver). Estas minimizaciones suelen converger a mínimos de energía locales cercanos a la posición inicial donde se producen reorientaciones de las cadenas laterales de ciertos residuos, pequeños movimientos que eliminen choques entre átomos, o que permiten adoptar posiciones idóneas para establecer redes de enlaces de hidrógeno. Otra alternativa es el uso de la *dinámica molecular* mediante la cual es posible lograr mayores cambios conformacionales en comparación con las técnicas basadas en la minimización de la energía. De este modo podemos reorientar la solución de *docking* a mínimos de energía cercanos a la posición inicial pero que mejoren sustancialmente las interacciones entre las superficies de contacto. Para reproducir de forma realista la interacción, en la mayoría de los casos es necesaria la *incorporación de moléculas de agua e iones* durante la simulación, ya que es normal que se encuentren mediando ciertas interacciones entre las proteínas. Por supuesto el uso de la dinámica molecular para el refinado

incrementa notablemente el coste computacional, por lo que debemos tener en cuenta el número de soluciones que queremos refinar respecto al coste computacional que podemos permitirnos. Finalmente, la aplicación de *modos normales* en el refinado de soluciones de *docking* se aplica también con cierto éxito, si bien su uso está menos extendido [27] (ver Capítulo 18).

Por último, una vez que hemos obtenido el conjunto de soluciones refinadas, es habitual realizar una nueva evaluación, ya que tras el refinado muchas poses habrán mejorado sus energías de interacción y por tanto, el *ranking* habrá cambiado con respecto a los resultados iniciales del *docking*.

16.5. Docking proteína-ácido nucleico

Aunque en los pasados 25 años se han producido grandes avances en los métodos de *docking* proteína-ligando y proteína-proteína, hasta el momento, existen muy pocos diseñados específicamente para el modelado de las interacciones en las que están involucrados los ácidos nucleicos [22]. La falta de programas de este tipo se debe principalmente a la dificultad del problema. Por un lado, el número de estructuras tridimensionales de moléculas que contienen ácidos nucleicos es muy bajo en comparación con el número de estructuras que se pueden encontrar de otro tipo de moléculas biológicas, dada la dificultad que conlleva su cristalización debido a su gran flexibilidad y su naturaleza de polianiones. Otras dificultades asociadas tienen que ver con la identificación de las superficies de interacción con otras moléculas. Esto ha motivado que el estudio de estructuras proteína-DNA y proteína-RNA se suele hacer a través de los mismos métodos que se usan en *docking* proteína-proteína, haciendo ciertas adaptaciones que no van más allá de incluir los tipos de átomos específicos que nos encontrados en las moléculas de DNA y RNA. En este sentido, el número de tipos de átomos que nos podemos encontrar en el RNA es considerablemente mayor que en el DNA, ya que debido a las modificaciones postraduccionales el RNA puede estar formado por más de 100 tipos de nucleótidos diferentes. Es importante remarcar que las adaptaciones de estos métodos no sólo se refieren a la generación de poses de *docking* [35], sino que puede corresponder también a otras a fases del modelado como el refinado de soluciones más prometedoras [27].

Finalmente, y debido a la ambigüedad de las soluciones obtenidas con métodos que no han sido diseñados para ese propósito, existen bases de datos con estructuras tridimensionales de complejos proteína-DNA obtenidas de forma experimental que permiten validar si un método es mejor que otro al generar las poses [38].

16.6. Conclusiones generales y perspectivas de futuro

A pesar de su notable éxito, la predicción rápida y eficaz de las interacciones ligando-diana sigue siendo el mayor desafío en *docking*. De la bibliografía de *docking* se deduce que tener en cuenta la flexibilidad de la diana y una buena función de *scoring* siguen siendo las principales preocupaciones, así como, en menor medida, la flexibilidad del ligando.

En lo que se refiere a la flexibilidad del ligando, los mejores resultados en términos de eficacia se obtienen normalmente cuando se usan como entrada múltiples conformaciones, en lugar de sólo una, siempre que éstas representen de forma adecuada el *espacio conformacional del ligando*. Por otro lado, como la flexibilidad de la diana depende en gran medida de la extensión de los cambios conformacionales que suceden durante el proceso de unión, el grado de movimiento permitido puede variar de una diana a otra. Cuando están involucrados grandes movimientos sigue siendo virtualmente imposible tenerlos en

cuenta, principalmente debido al coste computacional que requieren, aunque actualmente se han hecho algunos progresos en este sentido.

Con respecto al *scoring*, aunque se han desarrollado un gran número de funciones en las últimas décadas, el proceso de *docking* sigue siendo muy dependiente de las características específicas del sitio activo y del ligando. Por esta razón, la meta de obtener una función de *scoring* universal puede que sea demasiado ambiciosa, ya que ninguna de ellas podría funcionar bien en todos los casos. No obstante, las funciones de *scoring* hechas a medida para una determinada diana son una alternativa muy prometedora, sobre todo cuando se posee suficiente información acerca de ésta y algunos ligandos.

Como se ha demostrado en varios estudios comparativos, los métodos de *docking* actuales son, por lo general, más capaces de predecir modos de unión y no afinidades. Así, la mejora del rendimiento de la función de *scoring* en la *predicción de afinidad* parece ser una meta más urgente para desarrollos futuros. Esto requiere esfuerzos continuos en el diseño de algoritmos mejorados para las interacciones polares, las energías de solvatación/desolvatación, la entropía configuracional, etc. . . , sin comprometer la eficiencia en términos de tiempo de ejecución.

Si bien es cierto que los métodos de VS son muy populares, siempre se han caracterizado por una tasa muy alta de *falsos positivos* y un rendimiento muy desigual entre dianas consideradas asequibles, y aquellas imposibles de predecir sin tener información previa. Estos problemas, como se ha comentado antes, derivan claramente de la imposibilidad de predecir afinidades de una manera consistente y con un límite de error aceptable.

A día de hoy, a pesar de las continuas mejoras metodológicas en la interpretación de la unión entre moléculas pequeñas y dianas, como son la incorporación rutinaria de la desolvatación o la flexibilidad en los cálculos, el problema sigue vigente y está relacionado con la enorme complejidad del evento que se pretende simular y las aproximaciones que tienen que aplicarse para que el cálculo de la afinidad se pueda realizar en un tiempo adecuado con las tecnologías existentes.

Por otro lado, estos cálculos relacionados con efectos como la flexibilidad de las dianas, la desolvatación y la entropía, que intentan corregir este problema fundamental, conllevan un considerable aumento del tiempo de computación, por lo que seguramente serán también necesarias mejoras no solo metodológicas sino también de implementación incremento de la velocidad de computación, siendo este problema aún más acuciante al aumentar año tras año el número de moléculas disponibles en las *quimiotecas*.

Por lo que respecta al *docking proteína-proteína*, y aunque recientemente se han realizado grandes avances en la incorporación de la flexibilidad, el desarrollo de herramientas capaces de incorporar de forma precisa tanto los movimientos intrínsecos de la proteína como los movimientos asociados a la unión a otras moléculas sigue siendo una tarea pendiente, ya que existe un cuello de botella en el desarrollo de nuevos métodos por la falta de precisión en la estimación de estos movimientos.

También es de esperar que en los próximos años se produzcan mejoras en la evaluación de las *poses de docking*, ya que los métodos actuales presentan problemas asociadas con la rigidez. Otras mejoras que aún están lejos de poder alcanzarse debido a su complejidad, pero que son de gran interés para la comunidad científica, tienen que ver con la predicción de posibles plegamientos relacionado con la unión.

Finalmente, en lo relativo al modelado de *interacciones proteína-ácido nucleico*, queda aún mucho trabajo por delante. Es de esperar que se produzcan mejoras notables en un futuro no muy lejano, sobre todo teniendo en cuenta el incremento que se está produciendo en el número de estructuras depositadas en las bases de datos. Por ejemplo, en el año 2000 sólo existían 551 estructuras en el PDB que contienen fragmentos de DNA, mientas que en el año 2012 esta cifra ascendía a 3956. Sin lugar a dudas, el hecho de que las técnicas aquí tratadas vayan por delante en el estudio de las interacciones

entre proteína-ligando y proteína-proteína allanan en buena medida el camino para cuando llegue el momento en el que se disponga de la suficiente información estructural como para abordar directamente el problema del *docking* proteína-ácido nucleico.

16.7. Bibliografía

[1] F. H. Allen. The cambridge structural database: a quarter of a million crystal structures and rising. *Acta crystallographica. Section B, Structural science*, 58(Pt 3 Pt 1):380–8, 2002.

[2] J. Audie. Development and validation of an empirical free energy function for calculating protein-protein binding free energy surfaces. *Biophysical chemistry*, 139(2-3):84–91, 2009.

[3] C. Berger, S. Weber-Bornhauser, J. Eggenberger, J. Hanes, A. Pluckthun, and H. R. Bosshard. Antigen recognition by conformational selection. *FEBS letters*, 450(1-2):149–53, 1999.

[4] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–42, 2000.

[5] A. C. Cabrera, R. Gil-Redondo, A. Perona, F. Gago, and A. Morreale. Vsdmip 1.5: an automated structure- and ligand-based virtual screening platform with a pymol graphical user interface. *Journal of computer-aided molecular design*, 25(9):813–24, 2011.

[6] A. Cortes Cabrera, J. Klett, H. G. Dos Santos, A. Perona, R. Gil-Redondo, S. M. Francis, E. M. Priego, F. Gago, and A. Morreale. Crdock: an ultrafast multipurpose protein-ligand docking tool. *Journal of chemical information and modeling*, 52(8):2300–9, 2012.

[7] M. D. Eldridge, C. W. Murray, T. R. Auton, G. V. Paolini, and R. P. Mee. Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *Journal of computer-aided molecular design*, 11(5):425–45, 1997.

[8] T. J. Ewing, S. Makino, A. G. Skillman, and I. D. Kuntz. Dock 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of computer-aided molecular design*, 15(5):411–28, 2001.

[9] E. Fischer. Synthesen in der zuckergruppe. *Berichte der deutschen chemischen gesellschaft*, 23(2):2114–2141, 1890.

[10] M. K. Gilson, J. A. Given, B. L. Bush, and J. A. McCammon. The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophysical journal*, 72(3):1047–69, 1997.

[11] D. S. Goodsell, G. M. Morris, and A. J. Olson. Automated docking of flexible ligands: applications of autodock. *Journal of molecular recognition : JMR*, 9(1):1–5, 1996.

[12] R. Grunberg, J. Leckner, and M. Nilges. Complementarity of structure ensembles in protein-protein binding. *Structure*, 12(12):2125–36, 2004.

[13] M. Hassan, R. D. Brown, S. Varma-O'brien, and D. Rogers. Cheminformatics analysis and learning in a data pipelining environment. *Molecular diversity*, 10(3):283–99, 2006.

[14] B. Honig and A. Nicholls. Classical electrostatics in biology and chemistry. *Science*, 268(5214):1144–9, 1995.

[15] N. Huang, B. K. Shoichet, and J. J. Irwin. Benchmarking sets for molecular docking. *Journal of medicinal chemistry*, 49(23):6789–801, 2006.

[16] S. Y. Huang and X. Zou. An iterative knowledge-based scoring function for protein-protein recognition. *Proteins*, 72(2):557–79, 2008.

[17] J. J. Irwin, B. K. Shoichet, M. M. Mysinger, N. Huang, F. Colizzi, P. Wassam, and Y. Cao. Automated docking screens: a feasibility study. *Journal of medicinal chemistry*, 52(18):5712–20, 2009.

[18] A. N. Jain. Surflex: fully automatic flexible molecular docking using a molecular similarity-based search engine. *Journal of medicinal chemistry*, 46(4):499–511, 2003.

[19] X. Jiang, K. Kumar, X. Hu, A. Wallqvist, and J. Reifman. Dovis 2.0: an efficient and easy to use parallel virtual screening tool based on autodock 4.0. *Chemistry Central journal*, 2:18, 2008.

[20] V. Kasam, J. Salzemann, M. Botha, A. Dacosta, G. Degliesposti, R. Isea, D. Kim, A. Maass, C. Kenyon, G. Rastelli, M. Hofmann-Apitius, and V. Breton. Wisdom-ii: screening against multiple targets implicated in malaria using computational grid infrastructures. *Malar J*, 8:88, 2009.

[21] E. Katchalski-Katzir, I. Shariv, M. Eisenstein, A. A. Friesem, C. Aflalo, and I. A. Vakser. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proceedings of the National Academy of Sciences of the United States of America*, 89(6):2195–9, 1992.

[22] R. M. Knegtel, J. Antoon, C. Rullmann, R. Boelens, and R. Kaptein. Monty: a monte carlo approach to protein-dna recognition. *Journal of molecular biology*, 235(1):318–24, 1994.

[23] O. Korb, T. Stutzle, and T. E. Exner. Empirical scoring functions for advanced protein-ligand docking with plants. *Journal of chemical information and modeling*, 49(1):84–96, 2009.

[24] D. E. Koshland. Application of a theory of enzyme specificity to protein synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2):98–104, 1958.

[25] P. T. Lehtovuori and T. H. Nyronen. Soma–workflow for small molecule property calculations on a multiplatform computing grid. *Journal of chemical information and modeling*, 46(2):620–5, 2006.

[26] M. A. Lill and M. L. Danielson. Computer-aided drug design platform using pymol. *Journal of computer-aided molecular design*, 25(1):13–9, 2011.

[27] E. Lindahl and M. Delarue. Refinement of docked protein-ligand and protein-dna structures using low frequency normal mode amplitude optimization. *Nucleic acids research*, 33(14):4496–506, 2005.

[28] C. W. Murray, C. A. Baxter, and A. D. Frenkel. The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. *Journal of computer-aided molecular design*, 13(6):547–62, 1999.

[29] M. M. Mysinger, M. Carchia, J. J. Irwin, and B. K. Shoichet. Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. *Journal of medicinal chemistry*, 55(14):6582–94, 2012.

[30] R. Norel, D. Fischer, H. J. Wolfson, and R. Nussinov. Molecular surface recognition by a computer vision-based technique. *Protein engineering*, 7(1):39–46, 1994.

[31] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe. A fast flexible docking method using an incremental construction algorithm. *Journal of molecular biology*, 261(3):470–89, 1996.

[32] M. P. Repasky, M. Shelley, and R. A. Friesner. Flexible ligand docking with glide. *Curr Protoc Bioinformatics*, Chapter 8:Unit 8 12, 2007.

[33] D. Seeliger and B. L. de Groot. Ligand docking and binding site analysis with pymol and autodock/vina. *Journal of computer-aided molecular design*, 24(5):417–22, 2010.

[34] B. K. Shoichet. Virtual screening of chemical libraries. *Nature*, 432(7019):862–5, 2004.

[35] M. J. Sternberg, H. A. Gabb, and R. M. Jackson. Predictive docking of protein-protein and protein-dna complexes. *Current opinion in structural biology*, 8(2):250–6, 1998.

[36] W. Still, A. Tempczyk, R. Hawley, and T. Hendrickson. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc*, 112:6127–6129, 1990.

[37] M. Totrov and R. Abagyan. Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins*, Suppl 1:215–20, 1997.

[38] M. van Dijk and A. M. Bonvin. A protein-dna docking benchmark. *Nucleic acids research*, 36(14):e88, 2008.

[39] M. Vaque, A. Arola, C. Aliagas, and G. Pujadas. Bdt: an easy-to-use front-end application for automation of massive docking tasks and complex docking strategies with autodock. *Bioinformatics*, 22(14):1803–4, 2006.

[40] C. M. Venkatachalam, X. Jiang, T. Oldfield, and M. Waldman. Ligandfit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *Journal of molecular graphics & modelling*, 21(4):289–307, 2003.

[41] M. L. Verdonk, J. C. Cole, M. J. Hartshorn, C. W. Murray, and R. D. Taylor. Improved protein-ligand docking using gold. *Proteins*, 52(4):609–23, 2003.

[42] T. Zhou and A. Caflisch. Data management system for distributed virtual screening. *Journal of chemical information and modeling*, 49(1):145–52, 2009.