Universidad Autónoma de Madrid

Departamento de Biología Molecular

Facultad de Ciencias

Desarrollo y Utilización de Métodos Computacionales en la Mejora del Proceso de Obtención de Nuevos Fármacos

Tesis Doctoral

Rubén Gil Redondo

Director de Tesis

Dr. Antonio Jesús Morreale de León

Tutor del Departamento

Dr. Joaquín Dopazo Blázquez

MADRID 2010

A mis padres, mi hermano, Olatz
y Ángel

Agradecimientos

El trabajo de esta Tesis ha sido realizado en la Unidad de Bioinformática del Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM) gracias a la dirección del Dr. Antonio Morreale, quien desde el principio siempre ha tenido un momento para ayudarme, enseñarme y orientarme, además de haber sido un amigo. Incluso en los dos últimos años, desde que perdimos a Ángel, ha sabido encontrar el tiempo necesario para dedicármelo a pesar de haberse visto involucrado de repente en multitud de nuevas responsabilidades. Sin su apoyo constante no habría podido lograrlo. También me gustaría agradecer al Dr. Joaquín Dopazo por haber aceptado ser el tutor de esta Tesis y por haberse mostrado siempre dispuesto a ayudar en lo que necesitase. Además, me habría sido imposible llevar a cabo este trabajo de no ser por los organismos que dieron soporte económico a los diferentes proyectos: MEC (BIO2001-3745, BIO2005-0576 y GEN2003-206420-C09-08), Comunidad de Madrid (GR/SAL/0306/2004, 200520M157 y proyecto BIPEDD – SBIO-0214-2006), el programa intramural del CSIC (PIF2005 y proyecto CAR), la Fundación Ramón Areces, la Fundación Genoma España y el Ministerio de Ciencia e Innovación a través de su Programa Personal Técnicos de Apoyo 2008. Y por supuesto al Barcelona Supercomputing Center (BSC) por su generosa donación de tiempo de computación, y al proyecto Ibercivis por permitir acercar nuestro trabajo a los ciudadanos.

También quiero dar las gracias a Ugo y David, con los que llevo en el laboratorio desde que llegué. Gracias a su esfuerzo y al de Antonio el grupo sigue adelante. Además, si no fuera por David que es multitarea, más de una vez me habría quedado atascado en algún tema hardware/software… A los otros compañeros de laboratorio (Alberto, Gonzalo, Raúl, Almudena, Javi, Alvaro, Alfonso y Helena) por su apoyo, por compartir su tiempo y sus conocimientos conmigo, por hacer más ameno el día a día y por hacerme sentir como en casa. Y no quiero olvidarme de los ex-compañeros: Mamen, siempre solucionándonos los problemas, incluso ahora que no está; Rubén, con el que compartía "rayaduras" y que siempre me echaba un cable cuando lo necesitaba; Sandrea, que además de fliparse también intentaba enseñarme inglés… que paciencia! :-D; Fernando, con quien podía compartir las alegrías y las penas por nuestro Madrid; y al resto que también contribuyó a hacer el trabajo más llevadero (Pedro, Sara, Guille, Sandra, Enrique, Jesús, Florian, Edu, Paulino, Virginia, Esther y Han), así como mis "otros" compañeros (JMendi, José Luís y Ana) en el laboratorio del Dr. Galo Ramírez, quien amablemente me proporcionó un sitio mientras se preparaba el mío. Muchas gracias también a nuestros colaboradores, en especial al Dr. Federico Gago, siempre apoyándonos en todo y dedicándonos su tiempo, ¿dónde estaríamos sin él? Y sus chic@s de Alcalá, nuestros "primos", Claire y Juan, es un placer poder contar con ellos. Y claro también a Ana, a quien conozco desde el Master y que sigue tan dispuesta a ayudar y tan divertida como siempre. Y gracias a nuestros visitantes, que tanto nos han aportado. Como David Pantoja, tan amable, tan metódico, ¡y tan madridista! Gracias a Jorge Estrada, de Zaragoza, por toda su ayuda en el desarrollo, por aportarnos nuevos puntos de vista, y por infundirnos su optimismo. Y a Almudena, de Málaga, con quien además de trabajo podía compartir las series más frikis que se pueden encontrar.

Gracias también a los compañeros del CBM y del Departamento. En especial a Mada, que ha tenido siempre una paciencia infinita para resolver todas mis dudas. A Reyes y sus chic@s, siempre atentos y preocupándose por nuestros temas laborales. A Jaime y su gente de Administración, arreglándonos continuamente todos los problemas de papeleos. A Muñoz y sus chicos de mantenimiento, siempre acudiendo al rescate.

Si he conseguido llegar hasta aquí también ha sido en parte gracias a mis compañeros de Universidad: Chema, Carles, Javi y David. Ellos me hicieron la carrera mucho más llevadera ("Genario", "usted cobra", las hormigas del bocadillo,…), y no solo eso, ¡cuantas prácticas les debo! (sobretodo al "Ingeniero" :-D).

Por supuesto mis amigos también han sido muy importantes estos años. Raul, cuya persistencia me permitió reencontrarme con viejos amigos (Rubén, Luís, y él mismo) así como conocer a otros nuevos estupendos (Alfredo, Rau, Javi, Alex y Encinas). Y ha sido para mí una alegría que finalmente todos hayan encontrado unas chicas tan fantásticas (Gema, Diana, Rocío, Adela, Sagrario, Noelia, Reme e Irina). Y como olvidarme de los amigos del "barrio", con los que he compartido tantas cosas: Paquirri, Ramón, Germán, Miguel y Pilar. Todo resultaba más llevadero sabiendo que a la tarde me encontraría con ellos. Y el resto de chic@s del grupo (Alvaro, Sergio, Ana, Iván, Mónica, Piri, María,...), que aunque ahora les vea menos, tanto a ellos como a los demás, siempre resulta agradable y divertido compartir juntos un rato y saber como les va.

Nunca podré agradecer lo suficiente a Javi, a Victor y a Toni y Maria José y su familia (José, Dani, Laura, Carlos,…) lo que han hecho por mi hermano. No podía imaginar que existiera gente tan buena. Gracias a ellos la tristeza y la preocupación se han convertido en alegría y orgullo, lo que en gran medida ha contribuido a que haya podido realizar el trabajo de esta Tesis.

Gracias también a mi familia, a los que no veo tanto como quisiera, pero con los que siempre puedo contar. A mis padres, que me han dado todo; a mi hermano, que me hace sentir orgulloso. Y gracias sobretodo a Olatz, por ser ella, por hacerme feliz y por haberme hecho volver a sentir lo que es vivir en una familia.

Y por supuesto gracias al Dr. Ángel R. Ortiz. Él fue quien confió en mí desde el principio y me dio la oportunidad de trabajar en algo tan interesante. Me hizo sentir valorado, tanto profesionalmente como personalmente. Y aunque en su momento no fui consciente, ahora sé que hizo todo lo que estuvo en su mano y más para que yo pudiera llegar a desarrollar el trabajo de esta Tesis y para que fuera feliz en su laboratorio. Desde que le conocí no dejó de sorprenderme con sus profundos conocimientos y sus fantásticas ideas. Me contagió de su entusiasmo por la ciencia, de su exigencia, y de su devoción por el trabajo bien hecho. Con él no había dudas, era alguien sincero, se sabía perfectamente cuando algo no le gustaba, pero más importante aún, también te hacía saber cuando sí le gustaba, lo que suponía una motivación extra. Ahora que no está, cuando tengo que hacer algo siempre me pregunto "¿qué pensaría Ángel sobre esto?", "¿cómo haría Ángel?". Tanto él como su mujer, Carme, fueron un ejemplo para todos nosotros. Ella, que ya de antes era nuestra colaboradora, llegó a convertirse en nuestra compañera y amiga. Su entereza nos daba fuerzas a todos; y a pesar de la situación, su ayuda en el trabajo, muy importante en esta Tesis, no nos ha faltado en ningún momento. Ángel, aún con las dificultades que tenía en los últimos meses de su vida, se preocupaba más por nosotros que por él mismo. Orgulloso como pocos, lo dio todo mientras pudo. Ojala le hubiera dicho cuanto le admiraba y le apreciaba, que me sentía honrado por haberle conocido, y que sin duda ha sido una de las personas que ha cambiado mi vida para bien. Cuando le perdimos, me di cuenta de que no solo había perdido a mi jefe, sino que también había perdido un amigo.

ÍNDICES

Índice General

Índice de Contenidos

Índice de Algoritmos

Índice de Tablas

Índice de Figuras

ABREVIATURAS

Abreviaturas

AChE: *Acetylcholinesterase*

ADME: Absorción Distribución Metabolismo Eliminación

ADN: Ácido Desoxirribonucleico

ADT: *AutoDock Tools*

API: *Application Programming Interface*

AUC: *Area Under the Curve*

BER: *Base Excision Repair*

BSC: Barcelona *Supercomputing Center*

CADD: *Computer-Aided Drug Design*

CBM-SO: Centro de Biología Molecular "Severo Ochoa"

CDK2: *Cyclic dependant kinase 2*

CoMFA: *Comparative Molecular Field Analysis*

CoMSIA: *Comparative Molecular Similarity Indices Analysis*

COMBINE: *COMparative BINding Energy*

DDC: Dopa Descarboxilasa

Era: *Estrogen receptor a*

FDA: *Food and Drug Administration*

FtsZ: *Filamenting Temperature-Sensitive mutant Z*

fXa: *Coagulation factor Xa*

GB: *Generalized Born*

GUI: *Graphical User Interface*

HDC: *Histidina Decarboxilasa*

HIV: *Human Immunodeficiency Virus*

IDE: *Integrated Development Environment*

ISM: *Implicit Solvent Model*

JFC: *Java Foundation Classes*

LCPO: *Linear Combinations of Pairwise Overlaps*

LOO: *Leave One Out*

LV: *Latent Variable*

MCSA: *Monte Carlo/Simmulated Annealing*

MGMT: O6-metilguanina-ADN-metiltransferasa

MLR: *Multiple Linear Regression*

MNDO: *Modified Neglect of Differential Overlap*

MVC: *Model-View-Controller*

NMA: *Normal Mode Analysis*

NMR: *Nuclear Magnetic Resonance*

NNRTI: *Non-Nucleoside HIV-1 RT Inhibitors*

p38MAP: *Mitogen-Activated Protein Kinase P38*

PB: *Poisson-Boltzmann*

PBS: *Portable Batch System*

PCNA: *Proliferating Cell Nuclear Antigen*

PDB: *Protein Data Bank*

PE: *Poisson Equation*

PIP: *PCNA Interacting Proteins*

PLP (1): *Pyridoxal-phosphate*

PLP (2): *Piecewise Linear Potential*

PLS: *Partial Least Squares*

QSAR: *Quantitative Structure-Activity Relationships*

RMN: Resonancia Magnética Nuclear

RMSD: *Root Mean Square Deviation*

RMS: abreviatura que se emplea en ocasiones para referirse al *RMSD*

ROC: *Receiver Operating Characteristic*

RT: *Reverse Transcriptase*

SASA: *Solvent Accessible Surface Area*

SDEP: *Standard Error Prediction*

STL: *Standard Template Library*

VS: *Virtual Screening*

VSDMIP: *Virtual Screening Data Management on an Integrated Platform*

RESUMEN

Resumen

El lanzamiento de un nuevo fármaco al mercado requiere un tremendo esfuerzo en investigación, desarrollo e inversión económica. Uno de los mayores cuellos de botella en este proceso es la búsqueda de nuevos compuestos que muestren actividad para una proteína con cierto interés terapéutico. Estos compuestos reciben comúnmente el nombre de *hits*, y tienen que pasar por un largo proceso de optimización para convertirse en *leads*, lo que serían nuevos fármacos en caso de superar con éxito las pruebas clínicas. Tras más de 30 años haciendo uso de la química combinatoria y el cribado de alto rendimiento (las técnicas que parecían la solución al cuello de botella), el ratio entre el número de nuevos fármacos obtenidos y el dinero invertido sigue estando por debajo de las expectativas. Esto ha impulsado el desarrollo de métodos teóricos para tratar de acelerar y mejorar las etapas de búsqueda de *hits* y su optimización. El más destacado es el cribado virtual de quimiotecas (equivalente *in silico* del cribado de alto rendimiento), cuyo núcleo principal es el *docking*, un método para calcular la mejor posición de un ligando en el centro activo de una proteína y estimar su afinidad. En un proceso de cribado virtual intervienen multitud de herramientas, datos, diferentes formatos para representar moléculas y resultados, etc. Por ello, aquí se presenta *VSDMIP*, una nueva plataforma para el cribado virtual de quimiotecas integrada con una base de datos relacional. *VSDMIP* facilita el almacenamiento y manejo de los datos generados, se adapta fácilmente a los requerimientos para cada método o herramienta usada, permite la comparación de diferentes protocolos de cribado virtual y es paralelizable dentro de arquitecturas de supercomputación. Además, se han desarrollado y estudiado mejoras en algunos puntos críticos del diseño de fármacos *in silico*, como la inclusión de un modelo de solvente implícito en el proceso de muestreo del *docking*, la aproximación a la flexibilidad del ligando mediante su análisis conformacional y la mejora de un método *QSAR-3D* mediante la implementación de una herramienta gráfica que permita la generación y análisis de modelos con alta capacidad predictiva para la optimización de *hits*. Finalmente, se realizan tests retrospectivos de los desarrollos para estudiar sus posibles puntos de mejora, y tests prospectivos en casos reales de cribado virtual para evaluar su aplicabilidad.

Abstract

Launching a new molecule to the market requires tremendous effort in research, development and money investment. One of the biggest bottlenecks in this process is the search for new compounds that exhibit activity for a protein with some therapeutic interest. These compounds are commonly called *hits*, and they have to go through a long process of optimization to become *leads*, which would be new drugs in case of passing the clinical trials. But after more than 30 years using combinatorial chemistry and high throughput screening (techniques that seemed the solution to the bottleneck), the ratio between the number of new drugs obtained and the money invested is still below the expectations. This has prompted the development of theoretical methods to try to accelerate and enhance the stages of search for *hits* and their optimization. The most notable is the *virtual screening* of compound libraries (the *in silico* equivalent of high throughput screening), whose nucleus is the *docking*, a method to calculate the best position of a ligand in the active site of a protein and to estimate their affinity. But *virtual screening* processes involve a multitude of tools, data, different formats for representating molecules and results, etc. So, here we present *VSDMIP*, a new platform for *virtual screening* of compound libraries integrated with a relational database. *VSDMIP* facilitates storage and handling of the generated data, easily adapts to the requirements for each used method or tool, allows comparison of different virtual screening protocols, and is parallelizable into supercomputing architectures. Improvements have also been developed and studied in some critical points of *in silico* drug design, like the inclusion of an implicit solvent model to guide the sampling process in *docking*, an approach to the flexibility of the ligand through conformational analysis, and improvement of a *3D-QSAR* method by implementing a graphical tool that allows the generation and analysis of models with high predictive capability for the optimization of *hits*. Finally, retrospective tests to examine possible areas of improvement and prospective tests of real cases are performed to assess *VSDMIP* applicability.

INTRODUCCIÓN

1. Introducción

La Química Médica se ocupa del descubrimiento, desarrollo, identificación e interpretación del modo de acción de compuestos biológicamente activos a nivel molecular. Dentro de estos compuestos se encuentran los fármacos, entes químicos que interaccionan con sistemas complejos (tanto humanos como animales) y son utilizados con propósitos medicinales. La idea es que la interacción de este compuesto, una molécula pequeña, con una diana de interés terapéutico produzca una modulación de un proceso biológico crítico o con un rol importante en cierta enfermedad. La creación de un fármaco es un proceso largo e iterativo, y la mayoría de los fármacos actuales han sido más bien descubiertos (en lugar de desarrollados) (Sneader, 2005). Esto se debe a que un gran número de ellos son productos naturales o derivados. Cuando se realiza una prueba con un compuesto y una diana terapéutica dada, se dice que se ha obtenido un *hit* si el compuesto muestra cierta actividad. Si tras comprobar y repetir las pruebas se sigue observando esta actividad, entonces el compuesto pasa a llamarse *hit* validado. A partir de este punto es cuando se trata de modificar dicho compuesto con el objetivo de convertirlo en un *lead*, es decir, que además de mejorar su actividad y selectividad por la diana tratada, también se debe lograr que cumpla una serie de condiciones indispensables para ser un nuevo fármaco, como son la originalidad, patentabilidad y accesibilidad (por extracción o síntesis). En esta etapa de optimización también se deben modificar los parámetros físico-químicos del compuesto para que cumpla los requisitos ADME (Absorción, Distribución, Metabolismo y Excreción). Una vez se ha obtenido un *lead* optimizado, comienzan una serie de pruebas pre-clínicas, en las que debe demostrar no ser tóxico, tanto en células o tejidos como en modelos animales. Si se superan con éxito, se pasaría a la etapa de pruebas clínicas, donde debe demostrar ser seguro y eficaz en humanos a través de diferentes fases: I) probar que el compuesto es inofensivo sobre un conjunto reducido de individuos sanos; II) probar la eficacia del compuesto en un conjunto medio de individuos con la enfermedad tratada; y III) estudio sobre un conjunto grande de individuos enfermos para probar la eficacia del compuesto y su carácter inofensivo. Si tras estas tres fases se demuestra que el compuesto es seguro y eficaz en humanos, y además es capaz de superar las barreras del mercado, se considera que se ha obtenido un nuevo fármaco. También existe una fase IV, que es la llamada farmacovigilancia, que consiste en el seguimiento del fármaco una vez ha sido

comercializado. El proceso general de la obtención de un nuevo fármaco se muestra de forma esquemática en la Figura 1-1.

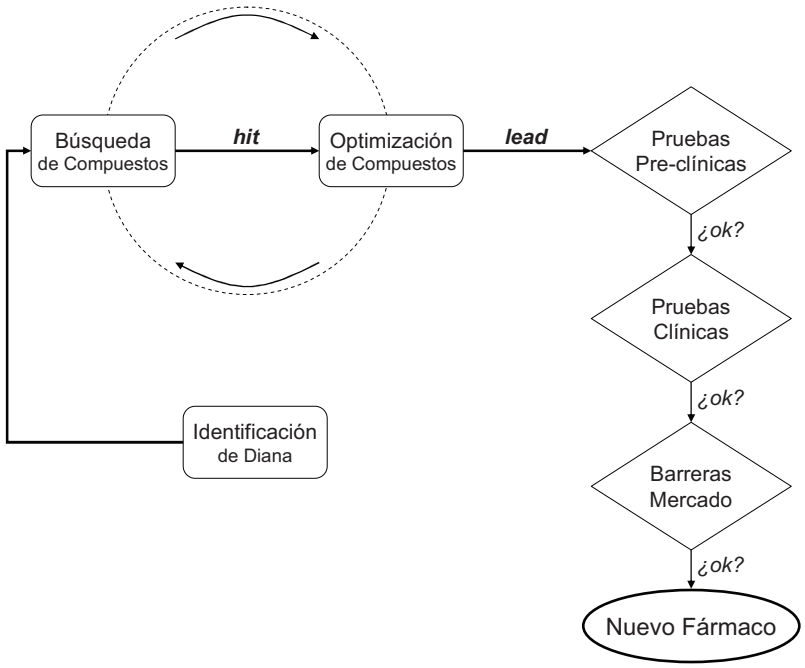


Figura 1-1. Esquema general del proceso de obtención de nuevos fármacos.

Lanzar un nuevo fármaco al mercado requiere un tremendo esfuerzo en investigación, desarrollo e inversión económica. Algunos estudios estiman que el tiempo medio invertido para lograr que una nueva molécula sea aprobada es de unos 15 años, con un coste que ronda los 800 millones de dólares (Munos, 2009; Smith, 2002). Tras la llegada del proyecto genoma humano y la genómica funcional ha habido un continuo incremento en el número de dianas terapéuticas disponibles para el diseño de fármacos. Durante este tiempo los avances en cristalografía y espectroscopía por resonancia magnética nuclear (RMN) han contribuido a la obtención de numerosos detalles estructurales de proteínas y complejos proteína-ligando. En respuesta a las nuevas dianas descubiertas y su determinación estructural, se ha convertido en una prioridad la obtención de compuestos bioactivos para esas dianas mediante un diseño y síntesis eficientes desde un punto de vista tanto económico como temporal. Debido a esto, los investigadores vieron en la química combinatoria y el cribado de alto rendimiento las soluciones al cuello de botella en el descubrimiento de fármacos. Pero después de más de 30 años usando estas técnicas, el ratio entre el número de nuevos fármacos obtenidos y los fondos invertidos en su generación está muy por debajo de las expectativas iniciales (Lahana, 1999; Ramesha, 2000). En cierto sentido esto ha impulsado el desarrollo de métodos teóricos que tratan de acelerar las etapas iniciales

del ciclo de diseño de fármacos, así como abaratar su coste y facilitar su organización. Dichos métodos teóricos, si han sido correctamente derivados, permiten la localización de *hits* de entre un conjunto de miles e incluso millones de moléculas (librerías químicas o quimiotecas). Este reducido conjunto de candidatos puede ser sometido a ensayos experimentales, y aquellos que muestren actividad pueden ser posteriormente optimizados para lograr el perfil farmacológico deseado que pueda convertirlo en un *lead*. Estos métodos teóricos son implementados computacionalmente (*in silico*) y se les encuadra dentro de lo que comúnmente se ha dado a conocer como diseño de fármacos asistido por ordenador (*CADD – Computer-Aided Drug Design*). Dicho término engloba las herramientas y recursos computacionales empleados para el almacenamiento, gestión, análisis y modelado de compuestos. Estas técnicas han contribuido al diseño de aproximadamente 50 compuestos (Jorgensen, 2004) que han entrado a la etapa de ensayos clínicos, y algunos de ellos ya han sido aprobados por la *FDA* (*Food and Drug Administration*) americana.

Entre los métodos más utilizados destaca el cribado virtual (Shoichet, 2004) de quimiotecas para la búsqueda de *hits*, que viene a ser el equivalente *in silico* del cribado de alto rendimiento. La técnica más empleada en el cribado virtual es el *docking* (Warren et al., 2006), utilizado para predecir el modo de unión y la afinidad de un ligando por una proteína. En cuanto a la optimización de *hits*, los métodos basados en la cuantificación de la relación entre la estructura y la actividad de los complejos proteína ligando (*QSAR – Quantitative Structure Activity Relationships*) se emplean para guiar el desarrollo de *leads*. En la actualidad el diseño basado en fragmentos también se está convirtiendo en un método muy importante a la hora de descubrir nuevos *hits*. En los siguientes apartados se exponen los conceptos más importantes en cuanto a interacciones ligando-receptor se refiere, y se describen más en detalle las técnicas computacionales empleadas en la búsqueda de fármacos. Finalmente, se da una visión global de la hipótesis de trabajo adoptada en esta Tesis.

1.1. Interacciones Ligando-Receptor

Según los principios fundamentales de la termodinámica clásica (Gilson et al., 1997) toda asociación/disociación molecular (ver Figura 1-2) representa un equilibrio ($[R]+[L] \Leftrightarrow [R+L]$) que viene regido por una constante (afinidad) relacionada con la energía libre puesta en juego en el proceso bajo estudio.

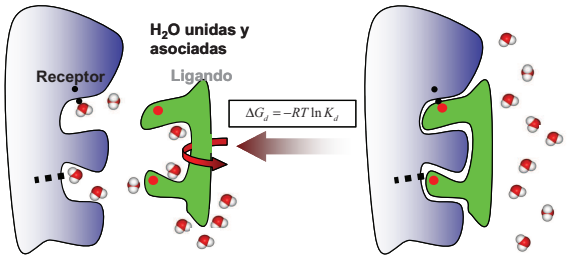


Figura 1-2. Esquema del proceso de disociación entre un ligando y su receptor en un medio acuoso.

Dicha energía libre de disociación se representa del siguiente modo:

$$\Delta G_d = -RT \ln K_d \qquad \textbf{[1-1]}$$

donde ΔG_d es la variación de energía libre de disociación del proceso, R es la constante universal de los gases ideales, T la temperatura, y K_d es la constante de disociación. En estudios computacionales normalmente se emplea la energía libre de unión (ΔG_b) en lugar de la de disociación. Obviamente $\Delta G_b = -\Delta G_d$. La energía libre de unión tiene una componente entálpica (ΔH_b) y una componente entrópica (ΔS_b):

$$\Delta G_b = \Delta H_b - T\Delta S_b \qquad \textbf{[1-2]}$$

La parte entálpica está formada por las interacciones de *van der Waals*, el potencial coulómbico, los enlaces por puentes de hidrógeno, las interacciones hidrofóbicas y las energías de desolvatación. Esta última tiene dos partes fundamentales: una polar y otra no polar. La componente polar se divide a su vez en desolvatación del receptor y desolvatación del ligando. La componente no polar incluye la parte de cavitación (el trabajo necesario para crear una cavidad dentro del seno del disolvente donde se alojará el soluto) y la parte correspondiente a las interacciones de *van der Waals* que se producirán entre el soluto y el solvente. Estas dos partes de la componente no polar se suelen considerar en conjunto y se estiman suponiendo una relación lineal con la superficie accesible al solvente, donde los parámetros se ajustan usando resultados experimentales. La parte entrópica vendría representada sobre todo por la flexibilidad en los enlaces rotables del ligando y los cambios conformacionales en el receptor como consecuencia de la unión. Todas ellas contribuyen a la hora de calcular la energía libre de unión del sistema, la cual quedaría finalmente representada del siguiente modo:

$$\Delta G_b = E_{vdw} + E_{elec} + E_{hb} + \Delta G^R_{desol} + \Delta G^L_{desol} + \Delta G_{hyd} + \Delta G_{conf} \qquad \textbf{[1-3]}$$

donde E_{vdw} y E_{elec} se corresponden con la energía de *van der Waals* y electrostática respectivamente, además la interacción por puentes de hidrógeno puede estar repartida entre estas componentes o bien aparecer como un término independiente (E_{hb}). ΔG^{R}_{desol} es la energía de desolvatación del receptor, ΔG^{L}_{desol} la desolvatación del ligando, ΔG_{hyd} la interacción hidrofóbica (componente no polar), y ΔG_{conf} representa la entropía originada por los cambios conformacionales en el ligando.

A continuación se describen las interacciones intermoleculares más importantes empleadas en la Ecuación [1-3].

1.1.1. Tipos de Interacciones

1.1.1.1. Energía de van der Waals (Potencial de Lennard-Jones)

Dependiendo de la distancia entre dos átomos esta interacción puede ser próxima a 0 si están muy alejados, o extremadamente alta si están demasiado cerca, como se muestra en la Figura 1-3, donde se tiene el potencial en el eje Y y la distancia entre los átomos en el eje X. ε representa el mínimo de la curva y σ es la distancia (finita) en la que el potencial es 0.

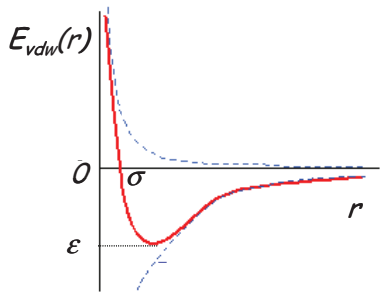


Figura 1-3. Energía en el potencial de *Lennard-Jones*.

Para calcularlo se puede utilizar la ecuación:

$$E_{vdw}(r) = 4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \qquad \textbf{[1-4]}$$

donde el primer término de la resta es la parte repulsiva cuando la distancia es pequeña (principio de exclusión de *Pauli*) y el segundo término es la parte atractiva cuando la distancia es grande (fuerzas dispersivas o de *London*).

1.1.1.2. Energía Coulómbica (Potencial de Coulomb)

Depende de las cargas y las distancias entre cada par de átomos en el receptor y el ligando. Se puede calcular mediante la ecuación:

$$E_{Coul}(r) = \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon r} \qquad \textbf{[1-5]}$$

donde q_i y q_j son las cargas para cada par de átomos *i* y *j*, *r* es la distancia entre ellos, ε_0 es la constante dieléctrica en el vacío y ε es la constante dielécrita en el medio (no confundir con el ε usando en el cálculo de la energía de *van der Waals* y que representaba el mínimo de la curva). En la Figura 1-4 se muestra la gráfica del comportamiento de esta interacción, tanto cuando es repulsiva (parte positiva) como atractiva (parte negativa), considerando una carga de 1 y una ε de 4.

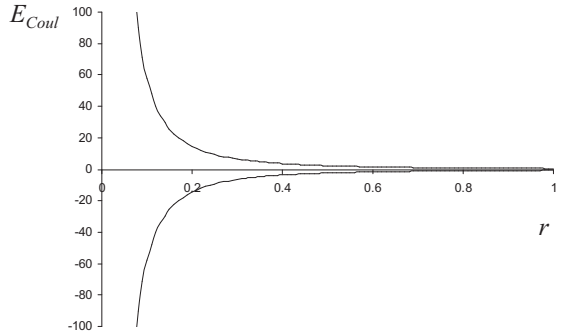


Figura 1-4. Gráfica del comportamiento del potencial de *Coulomb*.

1.1.1.3. Enlaces por Puentes de Hidrógeno

Son atracciones parcialmente electrostáticas (aunque con una componente cuántica importante) entre un átomo de hidrógeno unido a un átomo electronegativo (por ejemplo nitrógeno u oxígeno) y un átomo electronegativo adicional. En la Figura 1-5 se muestran un par de ejemplos de puentes de hidrógeno junto con las variables geométricas más importantes.

Figura 1-5. Ejemplos de puentes de hidrógeno y sus parámetros geométricos.

Son interacciones que en gran medida dependen de la orientación relativa de sus elementos constituyentes, es decir, son direccionales. Una de las formas para cuantificar

energéticamente el enlace por puente de hidrógeno es mediante la Ecuación [1-6], análoga de la Ecuación [1-4] pero añadiendo la componente direccional ($E(\Theta)$):

$$E_{hbond} = E(\Theta)4\varepsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{10}\right] \qquad \textbf{[1-6]}$$

1.1.1.4. Contribución Electrostática de la Desolvatación

El concepto de desolvatación está relacionado con el proceso de retirar moléculas de agua de la superficie de las proteínas y pequeñas moléculas como consecuencia de la interacción entre ellas. Es decir, parte de la superficie que antes estaba ocupada por agua se encuentra ahora ocluida por la otra molécula. El término de desolvatación tiene una componente puramente electrostática que tiene que ver con el desplazamiento dieléctrico del solvente producido por una molécula al acercarse a la otra, y viceversa, y otra no electrostática que está relacionada con la estructura del solvente, y por tanto con el cambio en la superficie accesible al solvente cuando las dos moléculas forman el complejo. La componente no electrostática se calcula a través de una relación lineal con el área de superficie accesible al solvente (*SASA – Solvent Accesible Surface Area*), por lo que en el resto de este apartado se abordará el término electrostático.

La forma clásica de calcular el potencial electrostático es a través de la ecuación de *Poisson* (*PE*) (Honig & Nicholls, 1995):

$$\nabla\left[\varepsilon(\mathbf{r})\nabla\phi(\mathbf{r})\right] = -4\pi\rho(\mathbf{r}) \qquad \textbf{[1-7]}$$

donde ρ es la distribución de carga, ε es la constante dieléctrica y ϕ es el potencial electrostático. Dicha ecuación se puede resolver analíticamente sólo para geometrías sencillas; en el caso de moléculas es necesario resolverla utilizando métodos numéricos. Conociendo el potencial electrostático es posible calcular la contribución electrostática a la desolvatación resolviendo la integral:

$$\Delta G_{ele} = \frac{1}{2}\int \rho(\mathbf{r})\phi(\mathbf{r})dv \qquad \textbf{[1-8]}$$

A partir de la ecuación [1-8] se puede calcular la energía de desolvatación aplicando ciclos termodinámicos. El proceso consiste en desolvatar primero las partes de las superficies moleculares que están en contacto, y después evaluar la interacción electrostática entre las cargas de ambas moléculas. El término coulómbico se obtiene calculando el producto entre las cargas del ligando y el potencial electrostático creado

por la distribución de cargas de la proteína en la posición de las cargas del ligando. Las desolvataciones individuales del receptor y el ligando se calculan en dos pasos sucesivos. El primer cálculo se realiza para el ligando y el receptor aislados, y el segundo para el receptor y el ligando en presencia del ligando y el receptor, respectivamente, pero sin cargas. El modelo completo se recoge en la Figura 1-6. En las imágenes, la esfera mayor representa el receptor y la menor el ligando, ambos rodeados por agua. Cuando uno de ellos aparece sin color (es decir, con el mismo color que el agua), significa que en ese paso no se están considerando sus cargas. Hay que hacer notar que en la figura el término $E_{LR}^{ele,s}$ se refiere a la interacción electrostática entre las cargas y no a la parte de desolvatación.

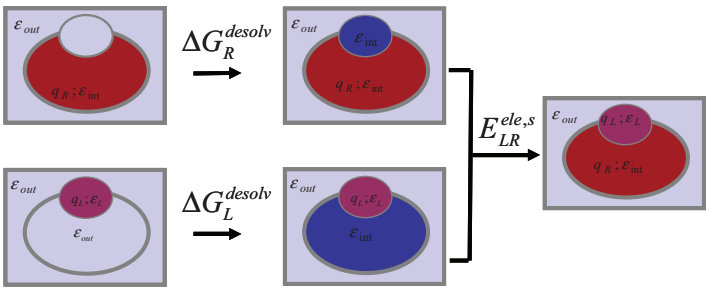


Figura 1-6. Esquema de desolvatación.

1.1.1.5. Contribución no Electrostática a la Desolvatación: Interacción Hidrofóbica

Es la estabilización adicional que se produce por el desplazamiento de moléculas de agua de la superficie de dos cadenas hidrofóbicas cuando éstas interaccionan entre sí. Este desplazamiento produce un aumento de entropía que se traduce en una disminución de la energía libre del sistema aumentando su estabilidad. En la Figura 1-7 se muestra esquemáticamente esta interacción.

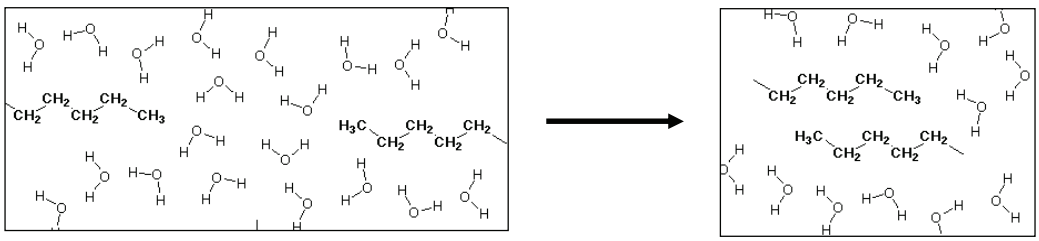


Figura 1-7. Esquema de la interacción hidrofóbica

La forma de cuantificar el efecto hidrofóbico es suponer una relación lineal con la superficie accesible al solvente (*SASA*) de la molécula como se muestra en la ecuación [1-9].

$$\Delta G_{hidroph} = a + b\Delta SASA \qquad \textbf{[1-9]}$$

donde $a = 0.0092\ kcal/mol$ y $b = 0.00542\ kcal/mol\text{Å}^2$. Cuando se quiere calcular la componente no polar para interacciones proteína-ligando lo que se hace es calcular la *SASA* tanto del complejo proteína-ligando como de la proteína y el ligando por separado. Para realizar un cálculo rápido y aproximado de la *SASA* se utiliza el método *LCPO* (*Linear Combinations of Pairwise Overlaps*) (Weiser et al., 1999) que está basado en combinaciones lineales de solapamientos de pares de átomos. Utilizando una sonda de 1.4 Å se calcula del siguiente modo:

$$A_i = P_1 S_1 + P_2 \sum_{j \in N(i)} A_{ij} + P_3 \sum_{\substack{j,k \in N(i) \\ k \in N(j) \\ k \neq j}} A_{jk} + P_4 \sum_{j \in N(i)} A_{ij} \left(\sum_{\substack{j,k \in N(i) \\ k \in N(j) \\ k \neq j}} A_{jk} \right) \qquad \textbf{[1-10]}$$

donde S_1 en el primer término es la superficie aislada de la esfera correspondiente al átomo *i*, A_{ij} es el área de esfera *i* cubierta por la esfera *j*, *N(i)* es la lista de vecinos de *i* (esferas que solapan con *i*). Así, el segundo término es la suma de los solapamientos dos a dos de la esfera *i* con sus vecinos. El tercer término es la suma de los solapamientos de los vecinos de *i* entre ellos. El cuarto término es una corrección para solapamientos múltiples. Los parámetros P_1-P_4 se derivan a partir de un conjunto de proteínas de entrenamiento. En el apartado a) de la Figura 1-8 se muestra un ejemplo de solapamiento entre esferas en la aproximación *LCPO*, mientras que en el apartado b) se muestra una representación del área de superficie accesible al solvente para un conjunto de esferas.

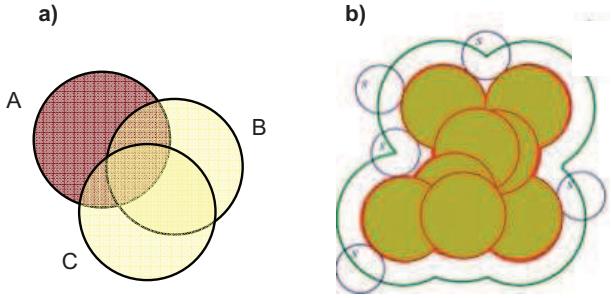


Figura 1-8. a) Solapamientos en la aproximación *LCPO*.
b) Representación de *SASA*.

1.1.1.6. Contribuciones Entrópicas

Básicamente se tienen seis contribuciones entrópicas principales (Zhou & Gilson, 2009): configuracional, translacional, rotacional, conformacional, vibracional y del solvente. La primera se refiere al hecho de elegir una determinada conformación de todas las posibles para una misma molécula, la del solvente a su reorganización interna, y el resto se refieren a la pérdida de grados de libertad que experimenta una molécula al unirse a otra.

1.1.1.7. Otras Interacciones

Otras interacciones que pueden darse son la formación de enlaces covalentes entre el ligando y el receptor (lo que produce inhibidores irreversibles), interacciones mediadas por moléculas de agua, o la presencia de átomos metálicos en el centro activo.

1.2. Docking

El *docking* molecular trata de predecir la posición, orientación y conformación nativas de una molécula pequeña (ligando) en el centro activo de una macromolécula de interés. El hecho de conocer las interacciones básicas que tienen lugar entre un ligando y su receptor permite la estimación de su afinidad y la utilización de técnicas de optimización. Los primeros trabajos sobre *docking* datan de principios de los años 80 (Kuntz et al., 1982), y hoy en día ha llegado a convertirse en un componente esencial de los programas de descubrimiento de fármacos *in silico* a pesar de que sigue siendo un área de investigación en desarrollo.

En términos generales, el *docking* puede describirse como la combinación de un algoritmo de búsqueda que trata de sugerir distintas disposiciones del ligando, y una función de puntuación cuyo objetivo es la identificación del modo de unión nativo. El número de posibles modos de unión de un ligando con la superficie de la proteína es virtualmente infinito. Por ello, el algoritmo de búsqueda debe ser lo suficientemente rápido y efectivo como para cubrir homogéneamente el espacio conformacional, incluyendo poses que estén muy próximas al modo de unión nativo. Por su parte, la función de puntuación necesita capturar adecuadamente la termodinámica de la interacción proteína-ligando para poder distinguir los modos de unión correctos (los cuales idealmente se deberían corresponder con el mínimo global de la función) del resto de los modos de unión propuestos por el algoritmo de búsqueda. También debe ser lo suficientemente rápida como para tratar con un gran número de potenciales

soluciones. A continuación se describen los diferentes métodos de muestreo en el *docking* y los principales tipos de funciones de puntuación.

1.2.1. Métodos de Muestreo

El caso ideal sería poder realizar una exploración exhaustiva de todos los grados de libertad de una molécula para poder encontrar su modo de unión nativo. Pero debido a la explosión combinatoria en el espacio de búsqueda, hacer esto es una tarea imposible. Y más aún en el caso de la proteína, donde el número de enlaces rotables es mucho mayor. Por ello la clasificación de los diferentes métodos de muestreo se basa en la aproximación que hace cada uno para tratar la flexibilidad del ligando, y en ocasiones también la de la proteína, ya que la mayoría de las veces se considera rígida. Aunque existen aproximaciones que tienen en cuenta su flexibilidad en diferentes maneras (ver apartado 1.2.2 de la página 13). Los siguientes apartados describen brevemente las principales categorías en que pueden dividirse los algoritmos de muestreo, aunque a veces la solución adoptada es el empleo conjunto de varios de ellos, y en otras ocasiones también algunos pueden admitir información adicional (como por ejemplo farmacóforos o conocimiento sobre la diana) para lograr mejores resultados.

1.2.1.1. Docking multi-confórmero

En este tipo de métodos se hace *docking* rígido utilizando una librería de confórmeros para un ligando dado, simulando de este modo su flexibilidad. A menudo se basan en complementariedad de forma con el centro activo o bien en algoritmos de mapeo de interacciones en el centro activo. Este tipo de técnicas resultan rápidas, pero su poder predictivo se ve reducido ya que las poses generadas no están refinadas, por ello su uso suele ir acompañado de una etapa de post-procesado con las poses seleccionadas. Debido a su rapidez, también suelen usarse como método de filtrado inicial de grandes quimiotecas.

1.2.1.2. Construcción Incremental

Son un tipo de algoritmos de búsqueda sistemática. Se basan en la reconstrucción del ligando dentro del centro activo, a menudo haciendo uso de librerías que tienen en cuenta las conformaciones más probables a la hora de conectar los fragmentos. Se suelen emplear dos aproximaciones: 1) partir el ligando en fragmentos y seleccionar uno como ancla sobre el que hacer *docking* rígido para posteriormente unir el resto de

fragmentos; ó 2) realizar el *docking* de todos los fragmentos para después emplear un algoritmo de re-conexión.

1.2.1.3. Métodos Estocásticos

En este tipo de métodos el ligando es considerado como un todo, y paso a paso se van aplicando cambios sobre una pose inicial o bien sobre una población de poses. Las nuevas poses generadas son puntuadas en función de las interacciones con la proteína de modo que la búsqueda pueda conducir a encontrar la pose nativa. En esta categoría se encuentran los algoritmos evolutivos y las simulaciones de *Monte Carlo*. Los primeros están basados en la teoría de evolución de Darwin. En estos se parte de un conjunto inicial de poses (población) y se tiene una función objetivo que puntúa cada modo de unión. Las poses menos buenas de la población van reemplazándose por las nuevas poses generadas a partir de las que tienen una puntuación mayor. Esta generación se realiza empleando operadores de mutación (pequeñas alteraciones en la pose) y cruzamiento (combinación de los parámetros de dos poses). El algoritmo termina tras cierto número de generaciones o si se ha convergido a una determinada solución. En cambio en los métodos basados en simulaciones de *Monte Carlo*, se parte de una pose inicial generada aleatoriamente y se le van aplicando cambios igualmente aleatorios. Tras cada modificación, la nueva pose es evaluada, y mediante el criterio de Metropolis (Metropolis et al., 1953) se decide si la nueva pose es utilizada como siguiente punto de partida o por el contrario se continúa con la última aceptada. Básicamente el criterio de Metropolis consiste en aceptar la nueva pose si su evaluación es más favorable que la última aceptada, o bien en aceptarla con una cierta probabilidad en caso de ser peor. El algoritmo termina de un modo similar a los evolutivos.

1.2.1.4. Simulación

En este apartado se encuentran los métodos basados en dinámica molecular y minimización energética, que quizá son los que mejor tienen en cuenta la flexibilidad del ligando y del receptor. Aunque estas técnicas no suelen considerarse en sí mismas como algoritmos de muestreo en el *docking*, ya que a menudo son incapaces de superar barreras energéticas elevadas en tiempos de simulación razonables (Sousa et al., 2006), lo que implica que las poses queden atrapadas en mínimos locales de la superficie energética. Además, necesita de más recursos computacionales que los otros algoritmos. Debido a esto y a su alta dependencia en la estructura de partida hacen que estos

métodos no sean aconsejables en etapas iniciales de *docking* (Brooks et al., 1983; Cornell et al., 1995; Weiner et al., 1984). Por ello suelen ser empleados como optimización de complejos receptor-ligando ya identificados.

1.2.2. Flexibilidad de la Proteína

Cuando el ligando y la proteína se unen adaptan mutuamente sus conformaciones. A este fenómeno se le conoce como ajuste inducido, y debería ser tenido en cuenta en los algoritmos de *docking* (Teague, 2003). Sin embargo el gran número de grados de libertad hace que se produzca una explosión combinatoria del espacio conformacional. Por ello, la mayoría de los algoritmos de *docking* sólo tratan la flexibilidad del ligando y mantienen la proteína rígida. Esto puede no ser un problema en ciertos sistemas, donde sólo se observan pequeños cambios tras la unión. Además, está ampliamente documentado el hecho de que el movimiento de la cadena principal en las proteínas está habitualmente restringido a menos de 1 Å. Pero cuando no se tiene información acerca de los efectos que sufre una determinada proteína tras la unión, es necesario utilizar un método que tenga en cuenta su flexibilidad. En los últimos años se están explorando diferentes alternativas para tratar la flexibilidad de la proteína. Una de ellas consiste en la realización de simulaciones por dinámica molecular para un conjunto de complejos proteína-ligando, habitualmente obtenidos de poses generadas y evaluadas en un *docking* previo. Similar a esta aproximación es la usada en el programa de *docking* *Glide* de la compañía *Schrödinger* (http://www.schrodinger.com/) en combinación con métodos de predicción o modelado por homología. En su método realizan el *docking* reduciendo el radio de *van der Waals* e incrementando el valor de corte para las energías coulómbica y de *van der Waals*. Posteriormente, con la herramienta de predicción del programa *Prime* acomodan el ligando reorientando las cadenas laterales cercanas. Otros métodos que han sido muy empleados son el uso de librerías de rotámeros (Bower et al., 1997; Canutescu et al., 2003; Dunbrack, 1999; Ponder & Richards, 1987), la utilización de un conjunto de conformaciones de la proteína (Claussen et al., 2001; Lorber & Shoichet, 1998), o el empleo de modos normales (*NMA – Normal Mode Analysis*) (Cavasotto et al., 2005). Pero a pesar de todas estas aproximaciones, la flexibilidad de las proteínas continúa siendo uno de los mayores retos en el *docking*.

1.2.3. Funciones de Puntuación

Los objetivos principales de las funciones de puntuación son dirigir el *docking* hacia el modo de unión nativo y predecir la afinidad de la pose final. Se pueden agrupar en cuatro categorías, las cuales se comentan en los próximos apartados.

1.2.3.1. Empíricas

Se basan en la idea de que las energías libres de unión pueden ser descritas como la suma ponderada de una serie de términos no correlacionados, como son los enlaces por puente de hidrógeno, los contactos no polares y aromáticos, o las penalizaciones por entropía. Los pesos relativos de esos términos se obtienen mediante análisis de regresión haciendo uso de complejos proteína-ligando cuyas energías libres de unión y estructuras son conocidas (Bohm, 1998; Eldridge et al., 1997; Rarey et al., 1996). Aunque este tipo de funciones resultan rápidas, sufren de una limitada descripción de los aspectos físicos del proceso de unión y de una dependencia del conjunto de datos usados en la parametrización.

1.2.3.2. Basadas en Campos de Fuerzas

En este caso se utilizan funciones de energía más universales y con un sentido más físico, como por ejemplo las energías de interacción de *van der Waals* y electrostática y las energías intramoleculares (Huey et al., 2007; Verdonk et al., 2003). Además, se están comenzando a introducir modelos de solvente implícito para tener en cuenta los efectos del solvente en la asociación receptor-ligando (Grosdidier et al., 2007; Huey et al., 2007; Zou et al., 1999). Los programas de docking suelen emplear mallas tridimensionales de interacción (conocidas como *grids*), divididas en puntos equidistantes, para aproximar la energía exacta del campo de fuerzas. Esto permite acelerar el cálculo, ya que la energía libre para un ligando puede calcularse como una suma del valor de los puntos de malla ocupados por los átomos de la molécula (o una interpolación de sus valores), teniendo en cuenta su carga y el tipo de átomo. También es habitual completar una función basada en campos de fuerzas con algunos términos empíricos.

1.2.3.3. Basadas en Conocimiento

Son funciones derivadas a partir de análisis estadísticos sobre estructuras cristalográficas de complejos proteína-ligando (Gohlke et al., 2000; Muegge & Martin,

1999). De estos análisis se obtienen distribuciones de parejas de tipos de átomos procedentes de la proteína y el ligando, y en función de estas distribuciones de frecuencias se obtienen las energías libres usando la ley inversa de Boltzmann y definiendo estados de referencia donde no se tienen interacciones. Así pues, la puntuación es calculada como la suma de todos los pares de interacción entre cada átomo del ligando y la proteína dentro de una esfera de cierto radio. Este tipo de aproximaciones tiene un problema similar al que se tenía en las funciones empíricas, es decir, son muy dependientes del conjunto de datos empleado para derivar el modelo.

1.2.3.4. Consenso

En esta aproximación se intenta capturar las bondades de varias funciones de puntuación combinando linealmente los diferentes resultados. Se basa en el hecho de que, aunque se puede decir que en general las funciones de puntuación funcionan relativamente bien, es cierto que no se ha podido identificar ninguna de ellas que sea superior a las demás en todos los tipos de dianas. Por ello es difícil elegir a priori cual sería la función apropiada a utilizar para una diana dada. La solución sería tomar un conjunto de funciones, de las que se conozca que tengan una alta precisión (Charifson et al., 1999), y combinar los valores obtenidos con cada una. Esta técnica suele ser muy empleada en cribado virtual pues se ha comprobado que se obtienen buenos resultados (Bissantz et al., 2000; Klon et al., 2004; Kontoyianni et al., 2004; Stahl & Rarey, 2001; Wang et al., 2003).

1.3. Cribado Virtual

El cribado virtual es la aproximación *in silico* a la técnica experimental del cribado de alto rendimiento. Consiste en la búsqueda, dentro de una base de datos, de compuestos que puedan interaccionar con una proteína dada. Dicha búsqueda puede realizarse midiendo el grado de ajuste a una serie de parámetros o condiciones provenientes de un farmacóforo, la estructura de otro ligando activo, o la estructura del receptor. La elección de un método u otro depende de la información con la que se cuenta y de las restricciones relacionadas con los recursos disponibles (económicos, de computación y de tiempo). Mientras que las dos primeras aproximaciones identifican compuestos similares a los usados como consulta, la aproximación basada en la estructura del receptor devuelve compuestos complementarios a la forma (estructural y energética) de su centro activo. Por ello cuando se habla de cribado virtual se pueden

distinguir dos tipos: 1) basados en la estructura del ligando; ó 2) basados en la estructura del receptor. Este último es el caso más favorable en la etapa de identificación de *hits*, aunque también es el más costoso. A continuación se describirán ambas aproximaciones, aunque el desarrollo de esta Tesis se centra principalmente en el cribado basado en la estructura del receptor.

1.3.1. Basado en la Estructura del Ligando

El cribado en este tipo de aproximaciones se basa en el uso de compuestos activos como plantillas para la búsqueda de otros nuevos. Resultan útiles cuando no se dispone de la estructura del receptor, aunque en ocasiones, gracias a que resultan computacionalmente bastante eficientes, son utilizados como un filtrado previo al cribado basado en la estructura del receptor para eliminar compuestos que no cumplan una serie de características que se sepan que son esenciales para la unión. Los métodos empleados en este tipo de búsqueda se pueden agrupar en tres clases: 1) búsqueda por similitud, incluyendo huellas dactilares moleculares (*fingerprints*) 2D, basada en el Principio de Similitud de Propiedades (Johnson & Maggiora, 1990), el cual establece que los compuestos similares presentan también propiedades similares; 2) métodos farmacofóricos, que consisten en definir una serie de restricciones (características o subestructuras) tridimensionales que deben cumplir los ligandos; el patrón farmacofórico suele obtenerse a partir de un conjunto de ligandos activos, y dicho patrón es usado en una búsqueda de subestructura; y 3) métodos basados en aprendizaje automático (*machine learning*), en los cuales se deriva una regla de clasificación a partir de un conjunto de ligandos activos e inactivos.

1.3.2. Basado en la Estructura del Receptor

En este tipo de aproximación se suele hacer un uso intensivo de los algoritmos de *docking* proteína-ligando para determinar el modo de unión de una proteína con todos los compuestos existentes en una base de datos (Alvarez, 2004; Klebe, 2006; McInnes, 2007). Las conformaciones obtenidas son usadas para aproximar la energía libre de unión o la afinidad relativa de cada compuesto. Los compuestos más prometedores en este sentido serían seleccionados para ser probados experimentalmente. Pero lo primero que se necesita son las estructuras moleculares del receptor, determinadas por cristalografía de rayos-X, resonancia magnética nuclear o modelado por homología (Davis et al., 2003; Kasimova et al., 2002; Oshiro et al., 2004). También hay otros

factores importantes a tener en cuenta y que pueden ayudar a obtener un mejor resultado en el cribado virtual, como son la correcta asignación de los estados de protonación o la consideración de moléculas de agua en el sitio de unión (Klebe, 2006). Otro apartado importante es el diseño de la base de datos de los compuestos a ser cribados. Existen multitud de bases de datos propietarias en las compañías farmacéuticas y además hay otros cientos o miles de compuestos disponibles comercialmente y que pueden ser encontrados en bases de datos no comerciales o colecciones propias de compuestos naturales o sintéticos (Irwin & Shoichet, 2005; Klebe, 2006; Leach et al., 2006).

1.4. Métodos QSAR

Las técnicas *QSAR* (*Quantitative Structure-Activity Relationships*) (Hansch et al., 1996) generan descriptores basados en la estructura molecular, y mediante una serie de algoritmos son capaces de encontrar una relación entre éstos y cierta propiedad de interés, como la actividad. Estas relaciones pueden utilizarse para predecir la actividad de nuevas moléculas. Los descriptores físico-químicos (hidrofobicidad, topología, enlaces por puentes de hidrógeno, campos de fuerza electrostáticos y efectos estéricos) son determinados empíricamente. Las actividades que se emplean en *QSAR* incluyen mediciones químicas y datos biológicos o bioquímicos. *QSAR* se utiliza en muchas disciplinas, entre las que destacan el diseño de fármacos (incluida farmacocinética) y la predicción de toxicidad. Por ello resulta una técnica muy útil a la hora de desarrollar un fármaco de mayor eficacia, mejor solubilidad, menor toxicidad, o simplemente para romper una patente existente. En la investigación biomédica resulta de especial interés el empleo de modelos *QSAR-3D*. Existen varios tipos de aproximaciones para la generación de estos modelos: aquellas que únicamente tienen en cuenta los parámetros físico-químicos de los sustituyentes en una serie congenérica de compuestos (Hansch et al., 2002), o las que calculan campos de interacción molecular en puntos discretos de una malla tridimensional que incluye los ligandos alineados en el espacio, como los populares *CoMFA* (*Comparative Molecular Field Analysis*) (Cramer et al., 1988) y *CoMSIA* (*Comparative Molecular Similarity Indices Analysis*) (Klebe et al., 1994). Además, cuando tras un cribado virtual se tienen unos cientos de compuestos para los que se conoce su actividad biológica, como es habitual en los proyectos de química médica, sería posible (empleando una función de puntuación lo suficientemente precisa como para describir las interacciones proteína-ligando) derivar un modelo *QSAR* con

alta capacidad predictiva usando un método como el análisis comparativo de energías de unión de *COMBINE* (*COMparative BINding Energy*) (Ortiz et al., 1995).

1.5. Diseño Basado en Fragmentos

Se trata de una aproximación similar al cribado virtual y que ha ido cobrando fuerza los últimos años. Mientras que en el cribado la búsqueda se hace sobre moléculas completas, el diseño basado en fragmentos se basa en la construcción de nuevos ligandos a partir del ensamblaje de pequeños compuestos que se unen a diferentes regiones del centro activo, suficientemente cercanas como para que se puedan mantener las interacciones favorables de cada fragmento tras su unión (Erlanson, 2006). Aunque existen equivalentes experimentales de esta técnica (Congreve et al., 2008), la aproximación *in silico* representa una potente alternativa (Schneider & Fechner, 2005). Tiene principalmente tres ventajas sobre el cribado virtual: 1) la fracción del espacio químico (Dobson, 2004) que puede ser probada es mucho mayor que con el cribado debido al gran número de combinaciones que se pueden realizar; además, sólo es necesario probar los fragmentos por separado para posteriormente tratar de combinar los más prometedores; 2) se consigue una alta tasa de acierto, ya que las moléculas pequeñas tienen más facilidad que las grandes para encontrar un modo de unirse (Hann et al., 2001); y 3) se consigue una eficacia del ligando mayor (Hopkins et al., 2004), es decir, que al tratarse de moléculas pequeñas la mayor parte de los átomos están involucrados en la interacción con la proteína. Al obtenerse ligandos más eficientes y pequeños también se favorecen las propiedades fármaco-cinéticas (Hann et al., 2001). Otra característica importante a tener en cuenta es que la energía libre de unión resultante de la conjugación óptima de dos fragmentos es más favorable que la suma de las energías de unión de estos por separado. Esto es debido, en parte, a que en el caso de fragmentos hay que considerar la pérdida entrópica de fijar el cuerpo rígido de cada uno, pero cuando se tiene una molécula completa sólo se tiene una pérdida entrópica. Es cierto que también se generarían nuevos enlaces rotables, pero la pérdida entrópica por fijar estos enlaces es mucho menor en comparación con la pérdida entrópica por fijar el cuerpo rígido del ligando.

Al la hora de juntar los diferentes fragmentos se pueden seguir dos enfoques diferentes. Uno se basa en la unión de estos ya posicionados en sitios de interacción claves del receptor. Dicha unión se realiza automáticamente a través de conectores, que son pequeños fragmentos lo suficientemente flexibles como para no alterar la eficacia

de los fragmentos que se quieren unir. Como resultado se obtiene una molécula completa que satisface todas las interacciones claves que tenían los componentes originales. El otro enfoque parte de un fragmento colocado en un sitio de interacción clave dentro del receptor; a partir de ahí se le van uniendo otros nuevos de modo que también tengan buenas interacciones con el receptor y además se sigan manteniendo las del fragmento original. Por supuesto ambos enfoques tienen sus puntos fuertes y sus puntos débiles (Schneider & Fechner, 2005). Por ejemplo, la estrategia de crecimiento tiene problemas cuando los bolsillos de unión del sitio activo están separados por amplias zonas en las que la interacción entre el ligando y la proteína es complicada. En cambio, la estrategia de unión tiene el problema de que necesita que los fragmentos estén correctamente orientados para lograr hacer las conexiones óptimas.

Pero estos problemas no son los únicos en esta técnica. Uno de los puntos más críticos es la accesibilidad a la hora de sintetizar las moléculas propuestas. Por ello se trata de derivar reglas de conexión basadas en la existencia de ciertos enlaces en compuestos orgánicos o de reacciones de síntesis orgánica. Aunque también se pueden seguir otras estrategias, como por ejemplo buscar moléculas comerciales que tengan una alta similitud con la propuesta y ver si es posible modificarla para obtener la deseada.

1.6. Hipótesis de Trabajo

Durante los últimos años se han desarrollado una gran variedad de métodos relacionados con el proceso de obtención de nuevos fármacos (búsqueda y optimización de *hits*). La integración de estos métodos no resulta sencilla, ya sea por carecer de un conocimiento profundo sobre ellos o por el hecho de no tener un modo apropiado de tratar las diferentes representaciones de los datos que utilizan y se generan. Además, aunque cada vez son más precisos, también es cierto que cada vez se necesita disponer de mayor potencia de computación, por lo que se deben poder integrar en grandes arquitecturas computacionales para cálculo intensivo en paralelo. Por ello sería importante contar con una plataforma de software capaz de unir los diferentes métodos a través de flujos de trabajo, que se apoye sobre una base de datos sólida donde almacenar y gestionar la información generada, y que sea capaz de adaptarse a arquitecturas de computación paralelas. También sería importante que tuviera la suficiente flexibilidad para integrar fácilmente nuevas funcionalidades, y que estuviera libremente disponible para la comunidad científica, de modo que pueda seguir creciendo y adaptándose a sus necesidades.

Pero resulta fundamental conocer el funcionamiento interno y la problemática de las bases del diseño de fármacos *in silico*. Y en el centro de todo se encuentra el *docking*, que es el núcleo del cribado virtual, junto con sus componentes más importantes: muestreo, evaluación y flexibilidad. La experimentación en el desarrollo y estudio detallado de algoritmos para el análisis conformacional de ligandos, la exploración del sitio activo, y la aproximación, mediante funciones de puntuación, de la afinidad con la proteína de la pose generada, permitirá sentar las bases para abordar futuros desafíos en el campo, como son la flexibilidad de la proteína o el diseño basado en fragmentos.

Finalmente, la etapa de optimización de *hits* resulta crítica para llevar a buen término un proyecto de diseño de fármacos. Los métodos *QSAR-3D* han demostrado su utilidad en este aspecto. Uno de ellos es *COMBINE* (Ortiz et al., 1995), basado en la idea de que es posible derivar, mediante métodos estadísticos, un modelo relativamente simple que resuma las diferencias en la afinidad de unión de una serie de complejos receptor-ligando. Este modelo se obtiene a partir de correlacionar datos experimentales de afinidades de unión con componentes de las energías de interacción receptor-ligando calculadas a partir de las estructuras tridimensionales. Este método, a pesar de haber obtenido excelentes resultados en numerosos trabajos, ha visto como su uso ha sido relegado a un segundo plano al no disponer de un interfaz sencillo e intuitivo donde poder desarrollar y analizar los modelos. Disponer de una herramienta gráfica para manejar las funcionalidades que ofrece *COMBINE* facilitaría la generación de modelos con alta capacidad predictiva.

Todo el trabajo de esta Tesis, en su conjunto, permitirá abordar proyectos de búsqueda de nuevos *hits*, pudiendo posteriormente conocer la actividad de las moléculas obtenidas gracias a la colaboración con diferentes grupos experimentales. Además servirá como punto de partida a la hora de acometer otros desafíos en el campo del diseño de fármacos por computador.

OBJETIVOS

2. Objetivos

La presente Tesis se centra en la mejora, desde un punto de vista computacional, del proceso de obtención de nuevos fármacos. En concreto de las técnicas empleadas en la búsqueda *in silico* de nuevos *hits* y el desarrollo de *leads*. De su grado de éxito dependen en gran medida la cantidad de tiempo e inversión necesarios para llevar un nuevo fármaco al mercado. Los principales objetivos a lograr en este trabajo son:

1. Desarrollo de una plataforma computacional en la que puedan implementarse e integrarse diferentes protocolos y herramientas para ayudar en la búsqueda de nuevos *hits* y su posterior desarrollo en *leads*. Dicha plataforma deberá apoyarse en una arquitectura flexible, escalable, paralelizable y que pueda almacenar la información de un modo ordenado y fácilmente accesible.

2. Desarrollo de un algoritmo de *docking* proteína-ligando con una función de puntuación precisa, capaz de tener en cuenta los efectos de la desolvatación, y con una aproximación a la flexibilidad del ligando basada en un conjunto de conformaciones pre-calculadas.

3. Implementación de una interfaz gráfica de usuario para facilitar la realización de estudios *QSAR-3D* en la etapa de optimización de *hits*.

Finalmente, además de realizar validaciones retrospectivas en cada desarrollo, también se llevarán a cabo validaciones prospectivas en casos reales de cribado virtual donde las moléculas seleccionadas se comprarán y probarán experimentalmente gracias a los laboratorios experimentales con lo que colaboramos.

MATERIALES Y MÉTODOS

3. Materiales y Métodos

3.1. Materiales

3.1.1. Conjuntos de Datos

3.1.1.1. Análisis Conformacional de Ligandos para Generar Confórmeros Bioactivos

El objetivo del análisis conformacional de un ligando es obtener un conjunto de confórmeros que sean representativos de su espacio conformacional. Dentro de este espacio se supone que deben encontrarse las conformaciones bioactivas, es decir, aquellas que adopta el ligando cuando se une a una proteína. Por ello para evaluar la eficacia de un método de análisis conformacional se suelen emplear ligandos con conformaciones bioactivas conocidas (usualmente a través de rayos-X) de modo que pueda comprobarse si el conjunto de confórmeros generados contiene estructuras similares a éstas o no. A continuación se comentan los conjuntos de datos seleccionados para este propósito.

3.1.1.1.1 Conjunto de Boström et al.

En su publicación (Boström et al., 2003) realizan un estudio de la eficacia del programa *OMEGA* (de la compañía *OpenEye*) a la hora de encontrar conformaciones bioactivas para un conjunto de 36 ligandos procedentes de complejos proteína-ligando. Estos ligandos serán utilizados para las pruebas de análisis conformacional.

3.1.1.1.2 Conjunto de Good et al.

En esta otra publicación (Good & Cheney, 2003) utilizan un conjunto de 30 ligandos para realizar una comparación de diversos programas que pueden ser usados para realizar análisis conformacional de ligandos (*CONFORT*, *CONFIRM*, *DOCK*, *OMEGA*), un programa para la generación de estructuras tridimensionales (*CONCORD*) y la minimización de la estructura de rayos-X mediante el campo de fuerzas de *Tripos* implementado en el programa *Sybyl* (http://www.tripos.com).

3.1.1.1.3 Conjunto de Astex Therapeutics

Hartshorn et al., de la compañía *Astex Therapeutics*, publicaron un estudio (Hartshorn et al., 2007) a través del cual obtuvieron un conjunto de 85 complejos

cristalográficos proteína-ligando de calidad y lo suficientemente representativos teniendo en cuenta su interés para las industrias farmacéuticas o agroquímicas.

3.1.1.2. Estudios de Docking Rígido/Flexible

Para este propósito se utilizan, al igual que para probar el análisis conformacional, las estructuras cristalográficas de los complejos del conjunto de *Astex Therapeutics*.

3.1.1.3. Modelo de Solvente Implícito

Para la validación del modelo de solvente implícito se ha utilizado un conjunto de 826 complejos (ver Tabla 3-I) proteína-ligando procedentes de 23 proteínas diferentes. Algunos de ellos han sido tomados directamente de la base de datos de *LPDB* (Roche et al., 2001) y otros han sido generados con el programa *CDOCK* (ver apartado 3.1.2.1.6 de la página 36). Los tipos de átomos, los radios y las cargas son fijados a partir de los parámetros del campo de fuerzas de *AMBER* (ver apartado 3.1.2.1.2 de la página 35). Del mismo modo se utiliza el programa *protonate* de este paquete para establecer la posición de los átomos de hidrógeno. En el caso de los ligandos las cargas son calculadas con el programa *MOPAC* (ver apartado 3.1.2.1.15 de la página 42). Se utiliza la solución numérica de la ecuación de *Poisson* calculada con el programa *DelPhi* (ver apartado 3.1.2.1.10 de la página 39) para comparar con los resultados obtenidos por el modelo *ISM* (ver apartado 3.2.2 de la página 60).

PDB ID	Descripción	# Complejos
1HVI	*HIV-1 protease*	43
1HVJ	*HIV-1 protease*	65
1HVK	*HIV-1 protease*	50
1HIH	*HIV-1 protease*	57
1HPX	*HIV-1 protease*	85
1MCJ	*Inmunoglobulin*	44
1RBP	*Retinol binding protein*	40
2UPJ	*HIV-1 protease*	48
1ABE	*L-arabinose binding protein*	60
1AJX	*HIV-1 protease*	64
5ABP	*L-arabinose binding protein*	34
1DBB	*Inmunoglobulin*	3
1FKG	*FK506 binding protein*	20
1FKH	*FK506 binding protein*	20
1MRK	*α-Trichosanthin*	9
1STP	*Biotin binding protein*	78
1B9V	*Influenza virus neuraminidase*	20
1DBM	*Inmunoglobulin*	20
1TNG	*Trypsin*	3
1TNI	*Trypsin*	20
1TNK	*Trypsin*	20
1TNL	*Trypsin*	3
1BMA	*Trypsin*	20

Tabla 3-I. Complejos utilizados para la validación de *ISM*.

3.1.1.4. Tests de Herramienta Gráfica para QSAR-3D

El principal objetivo de estos tests es validar que los resultados obtenidos por *COMBINE* (ver apartado 3.1.2.1.8 de la página 38) a través de su implementación gráfica son los mismos que se obtenían en las publicaciones donde se han realizado análisis *COMBINE*. Se han seleccionado dos de ellas para realizar los tests. En la primera se empleaban 48 inhibidores (32 para entrenamiento y 16 para test) de la proteasa tipo 1 del virus de inmunodeficiencia humana (*HIV-1*) (Perez et al., 1998). Este conjunto es además el que acompañaba la distribución original de *COMBINE* como ejemplo. En la segunda publicación se utilizaban 27 inhibidores no nucleosídicos de la transcriptasa inversa (*RT*) *HIV-1* (Rodriguez-Barrios & Gago, 2004), otra diana relevante desde el punto de vista farmacológico por su implicación en el virus del sida.

3.1.1.5. Estudios de Validación de Cribado Virtual

Para realizar los estudios de validación de cribado virtual con diferentes protocolos se necesitan una serie de proteínas para las cuales se conozcan conjuntos de ligandos activos experimentalmente. Dichos compuestos activos se han obtenido de un conjunto de datos binarios (activo/inactivo) publicados en la página *web CHEMOINFORMATICS.ORG*. Se toman un total de 9 conjuntos para 6 proteínas diferentes: 3 conjuntos para el factor de coagulación *Xa* (*fXa*) (Maignan et al., 2000; Murcia & Ortiz, 2004), que es una serin proteasa que ha recibido gran interés farmacológico debido a que puede servir como diana para desarrollar medicamentos antitrombóticos; 1 conjunto para acetilcolinesterasa (*AChE*) (Jacobsson et al., 2003; Kryger et al., 1999), una de las dianas en la enfermedad de Alzheimer; 1 conjunto para la kinasa dependiente de ciclina 2 (*CDK2*) (Arris et al., 2000; Thomas et al., 2006), una proteína involucrada en la regulación del ciclo celular; 2 conjuntos para el receptor de estrógeno alpha (*Era*) (Bissantz et al., 2000; Shiau et al., 1998), relacionada con cáncer, envejecimiento y obesidad; 1 conjunto para neuraminidasa (Burmeister et al., 1993; Murray et al., 1999), una de las glicoproteínas de superficie en el virus de la gripe; y 1 conjunto para *MAP* kinasa p38 (Cavasotto & Abagyan, 2004; Wang et al., 1997), envuelta en procesos de diferenciación celular y apoptosis. Cada uno de los conjuntos se completa añadiendo una serie de ligandos considerados inactivos, y que han sido seleccionados aleatoriamente de un subconjunto de 9862 compuestos de la colección *Maybridge Hit-Finder* (la mayoría de estos compuestos cumple la regla del 5 de

Lipinski (Lipinski et al., 2001), como puede verse en la Tabla 3-II). El resumen de los conjuntos utilizados para los experimentos de cribado virtual se muestra en la Tabla 3-III.

	MW[a]	HBA[b]	HBD[c]	RB[d]	logP[e]
Media	311	4.82	1.12	4.58	2.65
Desviación estándar	75.7	1.87	1.06	2.39	1.65

Tabla 3-II. Media y desviación estándar de las propiedades de Lipinski y el número de enlaces rotables para la base de datos de ligandos de *Maybridge*. Datos calculados con *Filter 2.0* (ver apartado 3.1.2.1.12 de la página 40). Usando las reglas *MOL_WT [0, 500]*, *ROT_BONDS [0, 20]*, *LIPINSKI_DONORS [0, 5]*, *LIPINSKI_ACCEPTORS [0, 10]* y *XLOGP [-5.0, 5.0]*, 743 moléculas no cumplen las reglas de Lipinski. [a] Peso molecular (en Da); [b] Número de aceptores de puente de hidrógeno; [c] Número de donadores de puente de hidrógeno; [d] Número de enlaces rotables; [e] log del coeficiente de partición octanol/agua.

Experimentos de cribado virtual	# activos	# inactivos	Total
fXa (*Fontaine*)	432	500	932
fXa (*Jacobsson*)	127	500	627
fXa (*Jorissen-Gilson*)	50	500	550
AChE (*Jacobsson*)	54	1000	1054
CDK2 (*Jorissen-Gilson*)	50	1000	1050
ERa (*Jacobsson*)	142	1000	1142
ERa (*Stahl*)	50	1000	1050
Neuraminidasa (*Stahl*)	17	1000	1017
p38MAP (*Stahl*)	22	1000	1022

Tabla 3-III. Información de los conjuntos de datos utilizados en los experimentos de cribado virtual.

3.1.1.6. Aplicaciones Reales de Cribado Virtual

Se han realizado diversas aplicaciones en casos reales de las técnicas desarrolladas en esta Tesis. A continuación se describen los materiales utilizados en varias de estas aplicaciones.

3.1.1.6.1 Proteína MGMT

Se trata de la proteína humana O6-metilguanina-ADN-metiltransferasa (*MGMT*), también llamada O6-alquilguanina-ADN-alquiltransferasa (*hAGT*). Dicha proteína previene las mutaciones y apoptosis resultantes de daños alquílicos en las guaninas (Brent et al., 1985; Mattern et al., 1998; Pepponi et al., 2003; Tagliabue et al., 1992). Debido a ello, representa un mecanismo de resistencia (Gerson, 2002; Gerson, 2004; Margison et al., 2003) a los tratamientos quimioterapéuticos basados en este tipo de daños. Esto hace que sea una diana interesante para buscar moléculas que puedan inhibir su función, de modo que se potencie la efectividad de ciertos tratamientos

quimioterapéuticos (Esteller et al., 2000; Hegi et al., 2005). Para encontrar estas moléculas mediante cribado virtual se utiliza una de las estructuras cristalográficas de la proteína depositada en el *PDB* (Protein Data Bank, http://www.rcsb.org/pdb/), en concreto la *1t39* (Daniels et al., 2004), la cual no presenta diferencias substanciales en el sitio activo con otras estructuras similares depositadas. La quimioteca para el cribado se obtiene de la base de datos *ZINC* (Irwin & Shoichet, 2005) (versión 6), que está disponible públicamente (http://zinc.docking.org/). En total se seleccionan ~2.3 millones de ligandos, en formato *SMILES* (de la compañía *Daylight*, http://www.daylight.com/), con un máximo de 7 enlaces rotables y más o menos cumpliendo la regla del 5 de Lipinski.

3.1.1.6.2 Proteína Ape1

La endonucleasa humana apurínica/apirimidínica *Ape1* es una enzima multifuncional esencial en la reparación del ADN. Se encarga de iniciar la eliminación de los sitios apurínicos/apirimidínicos del ADN, suprime los motivos 3´ de bloqueo de replicación, y modula la actividad de unión del ADN a varios reguladores de transcripción. Todas estas funciones, relacionadas con las rutas *BER* (*Base Excision Repair*) de reparación del ADN, convierten a *Ape1* en una diana terapéutica interesante (Madhusudan & Hickson, 2005) para potenciar, mediante su inhibición, la actividad de los agentes quimioterapéuticos alquilantes, tal como sucedía con *MGMT*. Para encontrar inhibidores mediante cribado virtual se utiliza la estructura cristalográfica cuyo código *PDB* es *1hd7* (Beernink et al., 2001). La quimioteca se vuelve a seleccionar de la base de datos de *ZINC*, pero en este caso la versión 7 y sin aplicar restricciones en cuanto a su número de torsionales. Tras eliminar los compuestos redundantes, el tamaño total de la quimioteca es de aproximadamente 4 millones de moléculas.

3.1.1.6.3 Proteína HDC

El enzima histidina decarboxilasa (*HDC, EC 4.1.1.22*) (Moya-Garcia et al., 2009) se encarga de catalizar la α-decarboxilación de la histidina, proceso en el que se produce histamina. La histamina juega un importante papel en la fisiopatología de los mamíferos. Está envuelta en procesos alérgicos y otras respuestas inflamatorias, secreción de ácido gástrico, pérdida de masa ósea, control del sueño y de la ingesta de comida, esquizofrenia, proliferación celular, cáncer, etc. Por lo tanto el control de la producción de histamina a través de la modulación de la actividad del enzima *HDC* ayudaría en el tratamiento de ciertas enfermedades. Pero para realizar un cribado virtual

con el que encontrar moléculas que inhiban la actividad de *HDC*, y por lo tanto la producción de histamina, es necesario disponer de su estructura tridimensional. Debido a la naturaleza inestable de la enzima ha sido imposible determinar su estructura mediante técnicas experimentales. Moya-García et al. (Moya-Garcia et al., 2006) han generado un modelo de la estructura del *HDC* de mamíferos utilizando para ello la aproximación computacional de modelado por homología (Baker & Sali, 2001), basándose en la estructura cristalográfica de la proteína dopa descarboxilasa de cerdo (*DDC*, código *PDB 1js3*, *EC 4.1.1.28*). Ambas proteínas son enzimas descarboxilasas dependientes de *PLP* pertenecientes a la misma familia y además comparten una alta identidad de secuencia, lo que facilitó la resolución de un modelo de alta calidad. Además, dicho modelo ha sido refinado y validado a través de ensayos experimentales de mutagénesis dirigida, y al estudio de la propia reacción de decarboxilación mediante simulaciones por dinámica molecular empleando una estrategia combinada de mecánica cuántica y mecánica molecular (Moya-Garcia et al., 2008). La especificidad (con respecto a la proteína humana *DDC* y a la *HDC* bacteriana de gram-negativa, presente en la microbiota intestinal) es muy importante (Moya-Garcia et al., 2006), de ahí la necesidad de que la zona de reacción de la enzima esté perfectamente definida de modo que se facilite la búsqueda de inhibidores específicos a través del cribado virtual. Además, está demostrado que resulta más complicado obtener buenos resultados en cribado virtual empleando estructuras modeladas (McGovern & Shoichet, 2003), por lo que es de vital importancia definir la estructura tanto como sea posible. Para la realización del cribado se emplea de nuevo la quimioteca de 4 millones de moléculas utilizada también en el cribado de *Ape1*.

3.1.1.6.4 Proteína PCNA

La proteína humana *PCNA* (*Proliferating Cell Nuclear Antigen*) es una proteína esencial para el metabolismo del ADN de la célula, así como en el control del ciclo celular, ya que interacciona con diversas proteínas involucradas en estos procesos mediante la llamada "caja *PIP*" (*PCNA Interacting Proteins*). Representa una diana antitumoral apropiada dada su esencialidad en numerosos procesos celulares, y especialmente en los proliferativos (Maga & Hubscher, 2003). Concretamente, la supresión de la expresión de *PCNA* con oligonucleótidos antisentido inhibe selectivamente la proliferación celular en cánceres gástricos *in vitro* e *in vivo* (Sakakura et al., 1994), y la supresión de *PCNA* en células bajo estímulo proliferativo se traduce

en la muerte celular programada o apoptosis (Paunesku et al., 2001; Prasanth et al., 2004), lo que indica que la inhibición de *PCNA* podría constituir una estrategia farmacológica efectiva en el tratamiento del cáncer. Para la realización del cribado virtual se utiliza la estructura cristalográfica cuyo código *PDB* es *1u7b* (Bruning & Shamoo, 2004). Se opta por esta estructura pues es la de mejor resolución (de entre todas las resueltas), y además *PCNA* se encuentra unida al péptido *FEN1*, por lo que el hueco del centro activo está definido. La quimioteca utilizada es un subconjunto de la base de datos de *ZINC*, en concreto las moléculas pertenecientes a los proveedores de *ChemBridge*, *Asinex* e *IBScreen*. En total se tienen aproximadamente 1.1 millones de moléculas.

3.1.1.6.5 Proteína FtsZ

FtsZ (Oliva et al., 2004) (*Filamenting Temperature-Sensitive mutant Z*) es una proteína del citoesqueleto de las bacterias que se ensambla en un anillo para mediar durante la división celular bacteriana. Representa el equivalente procariótico a la tubulina de las células eucarióticas. Por lo tanto *FtsZ* es una interesante diana terapéutica para la que buscar agentes antibacterianos. El código *PDB* de la estructura cristalográfica utilizada para el cribado virtual es *1ofu* (Cordell et al., 2003). La quimioteca utilizada inicialmente es la perteneciente al proveedor *ChemBridge* obtenida de *ZINC*. Aproximadamente tiene medio millón de moléculas.

3.1.2. Software

El software utilizado puede agruparse en tres categorías: programas, librerías y servidores *web*. Los primeros se refieren a aquél software que funciona como un ente completo e independiente, es decir, no necesitan integrarse dentro de otro código fuente para funcionar, sólo necesitan unos datos de entrada para producir su resultado. Las librerías en cambio son un tipo de software que ofrecen una serie de funcionalidades para ser integradas dentro de programas independientes, o bien proporcionan una plataforma de ejecución. Y en cuanto a los servidores *web*, su filosofía es similar a la de los programas independientes pero en este caso la ejecución se solicita vía *web* y se realiza en un computador externo, que habitualmente suele interactuar con una base de datos. En los siguientes apartados se comentan los principales programas, librerías y servidores *web* utilizados.

3.1.2.1. Programas

Primero hay que distinguir entre los programas desarrollados en el propio laboratorio y los que son externos. Esta distinción es importante, ya que en los programas internos no sólo hay que conocer su funcionamiento, sino que también hay que conocer su código fuente con el fin de llevar a cabo las modificaciones y mejoras necesarias. De algunos programas externos también se dispone del código fuente, pero resulta más difícil su modificación y en ocasiones es más adecuado elegir el programa que mejor se adapte a las necesidades. A continuación se irán comentando cada uno de los programas.

3.1.2.1.1 ALFA (interno)

Es un programa que sirve para generar los confórmeros de un ligando dado para aportarle flexibilidad en cuanto al movimiento de sus enlaces rotables se refiere. En primer lugar lee una molécula en formato *PDB* y a continuación la transforma en un grafo no dirigido, donde los nodos son los átomos y los arcos corresponden a los enlaces atómicos. Se dice que el grafo es no dirigido porque los arcos no tienen dirección asignada. Gracias a esta representación se pueden tener los datos de la estructura de la molécula en el programa además de asignar tipos a cada átomo basándose en los utilizados en el campo de fuerzas de *AMBER* (ver Tabla 3-IV).

Identificador	Radio (Å)	Descripción
1	1.91	Carbono sp^3
2	1.91	Carbono sp^2
3	1.82	Nitrógeno con 2 o menos enlaces
4	1.72	Oxígeno de grupo hidroxilo
5	1.66	Oxígeno de grupo carbonilo
6	1.68	Oxígeno de ésteres y éteres
7	2.00	Azufre
8	1.49	Hidrógeno alifático genérico
9	1.46	Hidrógeno aromático o unido a un carbono sp^2
10	0.60	Hidrógeno unido a nitrógeno o azufre
11	0.00	Hidrógeno unido a oxígeno
12	2.10	Fósforo
13	2.35	Iodo
14	2.22	Bromo
15	1.95	Cloro
16	1.75	Flúor
17	1.39	Hidrógeno alifático unido a un carbono con 1 grupo electronegativo
18	1.29	Hidrógeno alifático unido a un carbono con 2 grupos electronegativos
19	1.19	Hidrógeno alifático unido a un carbono con 3 grupos electronegativos
20	1.41	Hidrógeno aromático unido a un carbono con 1 grupo electronegativo
21	1.36	Hidrógeno aromático unido a un carbono con 2 grupos electronegativo
22	1.10	Hidrógeno alifático unido a un carbono con un grupo cargado positivamente
23	1.82	Nitrógeno con 3 o más enlaces

Tabla 3-IV. Tipos de átomos usados en *ALFA*.

Después se generan las conformaciones según los ángulos diedros fijados en un fichero de configuración. De este modo se irán obteniendo cada uno de los posibles

confórmeros realizando combinaciones de ángulos rotables para cada torsional. El programa también calcula la energía de *van der Waals* de las interacciones 1-4 para cada nuevo confórmero. Esto significa que sólo se calculará para interacciones entre átomos que estén separados por 3 enlaces o más. Al igual que en *CGRID*, se utiliza el campo de fuerzas de *AMBER* para calcularlas. Opcionalmente también se puede calcular un valor de *RMSD* (ver apartado 3.2.7.1.1 de la página 79) con respecto a una conformación indicada como referencia. Habitualmente se utiliza como referencia la conformación bioactiva y poder así comprobar si el programa obtiene alguna similar o suficientemente parecida. La salida de este programa es o bien un fichero multi-*PDB* o bien un fichero de coordenadas conteniendo las de todas las conformaciones. Después hay que utilizar un segundo programa que obtiene las mejores rotaciones aplicando un valor de corte de 30 *kcal/mol* con respecto a la conformación de mínima energía encontrada. La entrada para este segundo programa es un fichero de coordenadas y la salida un fichero multi-*PDB*. Así pues, la entrada y salida para todo este proceso de generación de confórmeros están limitadas al formato *PDB*.

Para generar el fichero de configuración en el que se indican los enlaces rotables y los valores permitidos para cada ángulo diedro es necesario ver la molécula en un programa de visualización y analizarla con el fin de encontrar los enlaces rotables y determinar los ángulos diedros en función de la topología. Por eso es imposible utilizar el programa en un proceso de cribado virtual ya que normalmente se tiene en cuenta un número muy elevado de ligandos (del orden de millones) como para realizar una inspección visual de todos.

3.1.2.1.2 AMBER (externo)

AMBER (Case et al., 2005) es un paquete de programas para realizar simulaciones de biomoléculas basándose en un campo de fuerzas de mecánica molecular. En concreto para el protocolo de cribado se utiliza el programa *protonate* para añadir las posiciones de los átomos de hidrógeno en una proteína, además de los parámetros del campo de fuerzas usados en *AMBER*. También suele utilizarse el paquete para realizar dinámicas del receptor con el objetivo de relajar su estructura, o bien dinámicas de los resultados del cribado para minimizar su energía. Además aporta otra serie de herramientas para la preparación y parametrización de ficheros de moléculas.

3.1.2.1.3 Apolar (interno)

Es un programa utilizado para calcular la componente no electrostática de la energía libre de solvatación. Se basa en la pérdida de área de superficie accesible al solvente (*SASA – Solvent Access Surface Area*) debida a la formación del complejo receptor-ligando. El modo de realizar este cálculo se explica más detalladamente en el apartado 3.2.3.2.4 de la página 68.

3.1.2.1.4 AutoDock (externo)

AutoDock (Morris et al., 1998) es un conjunto de herramientas para realizar *docking* proteína-ligando basándose en un algoritmo genético. Incluye el programa *AutoGrid* para el cálculo de *grids* de interacción (ver explicación en apartado 3.1.2.1.7), y el programa *AutoTors* para la configuración de los enlaces rotables del ligando. El uso de *AutoDock* está muy extendido gracias a que se trata de *software* libre (licencia *GNU*). Además existen numerosas interfaces gráficas de usuario para manejarlo, entre ellas las pertenecientes al *AutoDock Tools* (*ADT*).

3.1.2.1.5 BABEL (externo)

Es una herramienta (http://openbabel.org/) que sirve para inter-convertir diversos formatos de ficheros de uso común en química computacional. También tiene otras funcionalidades extra (como son el cálculo de cargas o la adición de hidrógenos en una molécula), pero lo que realmente interesa es la capacidad de convertir unos formatos en otros.

3.1.2.1.6 CDOCK (interno)

Es el programa (Perez & Ortiz, 2001) de *docking* que hace uso de las *grids* obtenidas de *CGRID* (ver apartado 3.1.2.1.7) para obtener la mejor pose de un ligando dentro del centro activo de una proteína además de la energía de *van der Waals* y electrostática para dicha pose.

Este programa utiliza un algoritmo de búsqueda exhaustiva manteniendo rígidos tanto el ligando como la proteína, es decir, que no se generan confórmeros (la flexibilidad del ligando se simula realizando *dockings* para sus diferentes confórmeros). Se va trasladando el ligando moviendo su centro de masas a cada punto de *grid* donde se realiza un muestreo completo del espacio rotacional a través de los ángulos de Euler (Lattman, 1972) con una resolución de 27º. Las 512 mejores poses, en función de su energía, son sometidas a un proceso de minimización haciendo uso del algoritmo

SIMPLEX (Nelder, 1965). La pose de más baja energía obtenida en este proceso es la dada como resultado del *docking* para ese confórmero particular. Esta energía debe servir para distinguir la mejor pose para una molécula y para distinguir entre las mejores poses de diferentes moléculas, es decir, para poder ordenarlas en un proceso de cribado dependiendo de su afinidad con la proteína.

A la hora de calcular la energía de una pose, se calcula la interacción de cada átomo del confórmero con la proteína haciendo uso de la *grid* apropiada para el tipo de átomo obtenida de *CGRID*. Las *grids* de átomos concretos se utilizan para calcular la parte de *van der Waals* en la energía, y la *grid* electrostática se utiliza como su propio nombre indica para la parte electrostática. Ya que la posición de un ligando no suele coincidir exactamente con un punto de *grid*, lo que se hace es una interpolación trilinear (Bliznyuk & E.Gready, 1999) con los 8 puntos de *grid* más cercanos al átomo.

3.1.2.1.7 CGRID (interno)

Es un programa (Perez & Ortiz, 2001) que se utiliza para precalcular dentro del centro activo de una proteína las energías de interacción con un posible ligando. El objetivo es obtener una serie de mallas tridimensionales (*grids*) que contengan la información de estas interacciones para poder utilizarla posteriormente en el programa de *docking* (*CDOCK*). Sólo será necesario calcular las *grids* una vez para cada centro activo con el que se quieran realizar diversos *dockings*.

Lo primero que hace es crear una caja tridimensional alrededor del centro activo de la proteína. Para saber donde se encuentra dicho centro activo se toma como referencia un ligando ya conocido para la proteína, generando la caja tomando como esquina inferior y superior las coordenadas mínimas y máximas del ligando y ampliándola en cada dirección en la cantidad deseada (habitualmente 5 Å). En caso de no disponer de un ligando conocido para esa proteína se puede colocar uno en el lugar donde se encuentra el centro activo determinado teóricamente (programas de búsqueda de cavidades (Ho & Marshall, 1990), información evolutiva (López-Romero, 2004), etc) o bien en función al conocimiento experimental que se tenga de la proteína.

Una vez creada la caja se divide en puntos de grid con un espaciado determinado entre ellos (suele estar entre 0.375 Å y 0.5 Å). Las interacciones se calcularán para cada uno de esos puntos. El modo de hacerlo es ir colocando en cada punto de *grid* uno de los átomos principales en moléculas orgánicas (carbono, nitrógeno, oxígeno, azufre, hidrógeno, fósforo, flúor, cloro, bromo y iodo) y calcular la interacción de ese átomo en

dicho punto de *grid* con todos los átomos de la proteína. Esta información se almacena en una matriz tridimensional para cada tipo de átomo usado. Las interacciones calculadas son *van der Waals* utilizando el campo de fuerzas de *AMBER* (ver apartado 3.1.2.1.2 de la página 35), y la interacción coulómbica, ambos calculados para cada par de átomos del complejo proteína-ligando.

En la Figura 3-1 se muestra un ejemplo de la representación de la *grid* de carbono, que como se puede apreciar es muy útil a la hora de delimitar la cavidad dentro del centro activo. Esta *grid* está coloreada con un isocontorno de -1 *kcal/mol*, es decir, que sólo se muestran los puntos de *grid* cuyo valor para la interacción con un átomo de carbono es menor o igual a -1 *kcal/mol*.

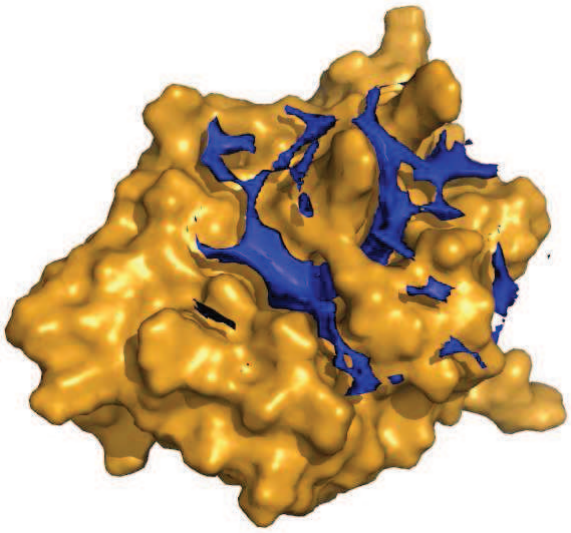


Figura 3-1. Ejemplo de *grid* de carbono a -1 *kcal/mol* para el centro activo de la proteína *MGMT*.

3.1.2.1.8 COMBINE (interno)

El análisis *COMBINE* (Ortiz et al., 1995) (*COMparative BINding Energy*) es un método *3D-QSAR* basado en complejos ligando-receptor. Resulta muy útil a la hora de correlacionar actividades farmacológicas con energías de interacción en procesos de optimización de potenciales fármacos. Primero, se calcula la interacción del complejo proteína-ligando para un conjunto de ligandos utilizando una función de energía basada en mecánica molecular. Después se seleccionan aquellos componentes de la energía de interacción que muestran una mayor capacidad predictiva de modo que se obtiene una ecuación de regresión en la cuál se correlaciona la actividad farmacológica con las energías de interacción de regiones clave en los ligandos y la proteína. En la Figura 3-2 puede verse un esquema del flujo de ejecución de un análisis *COMBINE*.

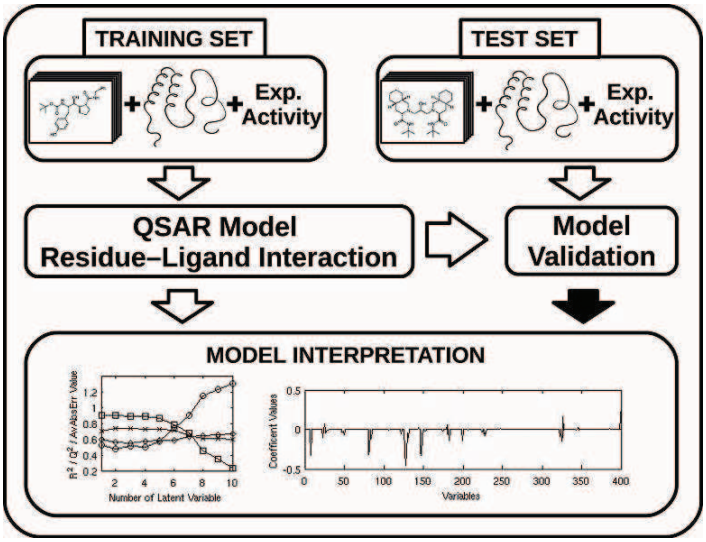


Figura 3-2. Flujo de ejecución de un análisis *COMBINE*.

3.1.2.1.9 CORINA (externo)

El programa *CORINA*, de la compañía Molecular Networks (http://www.molecular-networks.com/), sirve principalmente para obtener coordenadas tridimensionales para los átomos de moléculas en 2D. Para ello tiene en cuenta distancias de enlaces y ángulos teóricos para cada tipo de átomo de modo que se obtiene la configuración de mínima energía, es decir, la más estable y que en teoría podría ser la bioactiva.

Además de generar diferentes conformaciones tridimensionales también es capaz de tener en cuenta la estereoquímica de las moléculas así como de añadir los átomos de hidrógeno necesarios. Realiza las conversiones de un modo relativamente rápido y además admite diversos formatos de entrada y salida. Hay que hacer notar que *CORINA* genera una única estructura tridimensional, es decir, no tiene en cuenta los posibles confórmeros originados por rotaciones en los torsionales de las moléculas como hace el programa *ALFA* (ver apartado 3.1.2.1.1 de la página 34).

3.1.2.1.10 DelPhi (externo)

DelPhi es un programa de cálculo de potencial electrostático mediante la resolución de la ecuación de *Poisson* a través del método de las diferencias finitas sobre una *grid* cúbica tridimensional. Se utiliza para el cálculo de interacciones, cambios en el *pKa*, dieléctricos, y principalmente para energías de desolvatación. Puede obtenerse en la página: http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:DelPhi.

3.1.2.1.11 DOCK (externo)

Este programa (Kuntz et al., 1982) fue uno de los primeros métodos que se utilizó para analizar centros activos y sugerir ligandos que presentasen una complementariedad estérica con dicho centro activo. El programa lo rellena con una serie de esferas de distintos tamaños que se solapan entre sí, de tal manera que la unión de todas las esferas representa, de forma razonable, el volumen de la cavidad. A continuación, el programa trata de emparejar distancias entre los centros de las esferas con distancias entre los átomos de la molécula. Grupos de cuatro esferas y cuatro átomos se emparejan al mismo tiempo produciendo un conjunto de seis distancias que han de cumplir un rango de precisión determinado. Una vez que las seis distancias están dentro del rango, la molécula se traslada a los puntos definidos por los centros de las esferas, teniendo cuidado de que el resto de la molécula no de lugar a interacciones desfavorables con el centro activo. Para evaluar la orientación de la molécula dentro del centro activo *DOCK* pueden utilizar tres funciones de tanteo: dos basadas en cálculos en *grid* de campos de fuerzas, y una tercera basada en contactos, que se suele utilizar como filtro rápido de modo que se ve si una molécula "cabe" o no en el sitio activo. Además, *DOCK* puede trabajar con sus propias esferas o se le pueden proporcionar desde fuera, por ejemplo las funciones gaussianas, calculadas con *GAGA* (ver apartado 3.1.2.1.14). En la Figura 3-3 se tiene un esquema del funcionamiento de *DOCK*.

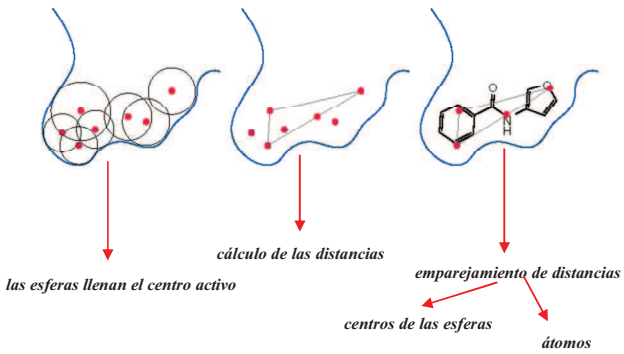


Figura 3-3. Ajuste entre esferas y átomos en *DOCK*.

3.1.2.1.12 Filter (externo)

Filter es un software de la compañía *OpenEye* (http://www.eyesopen.com/) que sirve para filtrar y seleccionar moléculas rápidamente basándose en el cálculo de sus propiedades físico-químicas y sus grupos funcionales.

3.1.2.1.13 FRED (externo)

FRED es un programa de *docking* proteína-ligando de la compañía *OpenEye*. Dicho programa examina exhaustivamente todas las posibles poses del ligando en el sitio activo de la proteína, filtrándolas por complementariedad de forma o por propiedades farmacofóricas. Finalmente selecciona una solución basándose en una o varias funciones de puntuación que lleva incorporadas.

3.1.2.1.14 GAGA (interno)

Es un programa (Wang et al., 2004) capaz de definir automáticamente un conjunto mínimo de puntos farmacofóricos que mapean las propiedades de interacción del sitio activo de una proteína; este conjunto es quasi-óptimo. Se basa en aproximar la *grid* de interacción (calculada por *CDOCK* para una sonda) mediante una expansión lineal de funciones gaussianas. Una función heurística se encarga de seleccionar de esa expansión el conjunto de gaussianas necesario para aproximar la *grid* con una determinada precisión. La *grid* de energías es transformada en una función discreta $f(\mathbf{x})$. El objetivo es encontrar el conjunto mínimo de gaussianas del tipo:

$$\left\{ g_j(\mathbf{x}) = a_j e^{-\frac{(\mathbf{x}_j - \mathbf{x})^2}{2\alpha_j^2}} \right\}_{j=1}^{M} \qquad \textbf{[3-1]}$$

de modo que el error de la aproximación sea menor a cierto epsilon:

$$\left\| f(\mathbf{x}) - \sum_{j}^{M} c_j g_j(\mathbf{x}) \right\| < \varepsilon \qquad \textbf{[3-2]}$$

donde M es el número de gaussianas, c son los coeficientes, x los centros y α las amplitudes. La sonda que se suele utilizar para calcular las *grids* de interacción es la del benceno ya que al ser hidrofóbica representa muy bien el centro activo. Otras sondas son el agua y el metanol, ya que proporcionan zonas de interacción por puente de hidrógeno. Los centros de las esferas generadas por *GAGA* pueden ser utilizados en *DOCK* como puntos de anclaje en sustitución de los que podría crear él mismo. En la Figura 3-4 se muestra un ejemplo.

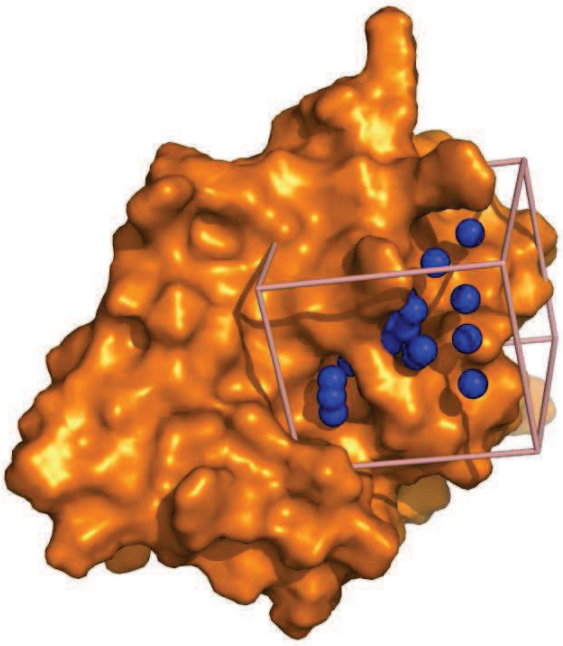


Figura 3-4. Ejemplo de esferas generadas con *GAGA* para el centro activo de la proteína *MGMT*.

3.1.2.1.15 MOPAC (externo)

Son las siglas de *Molecular Orbital PACkage*. Es un paquete de propósito general (Stewart, 1990) para el cálculo de orbitales moleculares semiempíricos usado para el estudio del estado sólido, estructuras moleculares y reacciones químicas. Dispone de una serie de *Hamiltonianos* semiempíricos que se usan en la parte electrónica del cálculo para obtener orbitales moleculares, el calor de formación y su derivada con respecto a la geometría molecular. Usando estos resultados *MOPAC* es capaz de calcular el espectro vibracional, magnitudes termodinámicas, los efectos de las sustituciones isotópicas y constantes de fuerza para moléculas, radicales, iones y polímeros.

En el desarrollo de un protocolo para el cribado virtual de quimiotecas resulta muy útil a la hora de calcular las cargas parciales de los átomos en los ligandos. Además también se usa para la parametrización de torsionales.

3.1.2.1.16 ROCS (externo)

ROCS es un otro programa de la compañía *OpenEye* y se utiliza para realizar comparaciones rápidas de ligandos en cuanto a su forma basándose en su volumen.

3.1.2.1.17 PyMOL (externo)

PyMOL (http://pymol.org/) es un programa gratuito que sirve para la visualización de moléculas a partir de diferentes tipos de ficheros. Además incorpora diversas utilidades para trabajar con ellas, y tiene la posibilidad de añadir nuevos *plug-ins* y usar *scripts*.

3.1.2.1.18 VMD (externo)

VMD (*Visual Molecular Dynamics*) (http://www.ks.uiuc.edu/Research/vmd/) es un programa de visualización molecular para poder ver, animar y analizar grandes sistemas biomoleculares usando gráficos 3D. Permite la incorporación de *scripts*, funciona en diferentes sistemas operativos y es de libre distribución.

3.1.2.2. Librerías

Son librerías software de clases y funciones que resultan útiles en el desarrollo de otros programas, ya que incorporan la implementación de una serie de funcionalidades que sirven de base para la creación de otras más complejas y específicas. A continuación se describen brevemente las utilizadas en esta Tesis.

3.1.2.2.1 BOINC

BOINC (http://boinc.berkeley.edu/) (*Berkeley Open Infrastructure for Network Computing*) es una plataforma software cuya misión es facilitar a la comunidad científica la creación y manipulación de proyectos de computación global.

3.1.2.2.2 GRID Superscalar

Es una librería (Badia et al., 2003) destinada a convertir automáticamente aplicaciones secuenciales compuestas de tareas en una aplicación paralela en la que dichas tareas serán ejecutadas en diferentes servidores de un *GRID* de computación.

3.1.2.2.3 JFC y Swing Application Framework

JFC (*Java Foundation Classes*) y *Swing Application Framework* (https://appframework.dev.java.net/) son librerías *Java* para añadir funcionalidad gráfica e interactividad a las aplicaciones desarrolladas.

3.1.2.2.4 JFreeChart y JCommon

Las librerías *Java JFreeChart* (http://www.jfree.org/jfreechart/) y *JCommon* (http://www.jfree.org/jcommon/) permiten generar gráficas interactivas para la

manipulación y análisis de los datos dentro de aplicaciones de interfaces gráficas de usuario (*GUI – Graphical User Interface*). Ambas librerías se distribuyen bajo licencia *GNU LGPL* (*Lesser General Public License*) (http://www.gnu.org/licenses/lgpl.html).

3.1.2.2.5 OpenPBS

OpenPBS (http://www.openpbs.org/) es la versión de código abierto de *PBS* (*Portable Batch System*), un *software* para la gestión de trabajos en un *cluster* de procesadores. Sirve para distribuir las diferentes tareas entre los recursos disponibles de cálculo.

3.1.2.2.6 MySQL++

MySQL++ (http://tangentsoft.net/mysql++/) es una librería escrita en lenguaje *C++* y sirve como envoltorio de las funciones *C* del *API* de *MySQL*. Permite trabajar con la base de datos de un modo tan simple como tratar con contenedores *STL* de la librería estándar de *C++*.

3.1.2.2.7 OEChem TK

OEChem TK es un conjunto de utilidades (*toolkit*) desarrollado por la compañía *OpenEye*. Cuenta con numerosas funciones, principalmente para el manejo de moléculas pequeñas y sus componentes. Está disponible en lenguaje *C++*, aunque también posee interfaces para los lenguajes *Python* y *Java*.

3.1.2.3. Servidores Web

Existen algunas herramientas o bases de datos que están accesibles a través de servidores *web* públicos. A continuación se comentan algunas de los utilizados en esta Tesis.

3.1.2.3.1 H++

Es un servidor (http://biophysics.cs.vt.edu/H++/) que permite realizar predicciones de estados de protonación y valores *pK* de grupos ionizables en macromoléculas. También permite añadir átomos de hidrógeno de acuerdo al *pH* especificado del entorno, y realizar cálculos relacionados con propiedades electrostáticas. Hace uso del campo de fuerzas de *AMBER*.

3.1.2.3.2 ZINC

ZINC (http://zinc.docking.org/) es una base de datos de moléculas pequeñas que agrupa el catálogo de un gran número de proveedores. Para cada molécula, también se

almacenan una serie de características físico-químicas interesantes además de información estructural. Las moléculas están agrupadas por conjuntos de interés (por vendedores, propiedades,…), aunque también se pueden especificar parámetros de búsqueda para obtener una nueva agrupación.

3.1.3. Entornos de Computación

3.1.3.1. Cluster de la Unidad de Bionformática del CBM-SO

El *cluster* de la Unidad de Bionformática cuenta con 41 máquinas biprocesador todas ellas con sistema operativo *Red Hat Enterprise Linux* 4. Se dividen del siguiente modo: 16 *SUN Xeon* biprocesador a 3.0 *GHz* con 1 *GB* de *RAM* y 100 *GB* de disco duro; 11 *HP Xeon* biprocesador a 2.4 *GHz* con 1 *GB* de *RAM* y 70 *GB* de disco duro; 14 *HP Xeon* biprocesador a 3.0 *GHz* con 2 *GB* de *RAM* y 70 *GB* de disco duro. Además tiene otras 5 máquinas *Xeon* con 8 *cores* cada una a 2.3 *GHz* con 6 *GB* de *RAM* y 150 *GB* de disco duro, y 12 máquinas también *Xeon* con 8 *cores* cada una pero a 2.6 *GHz* con 8 *GB* de *RAM* y 150 *GB* de disco duro. Un *HP Xeon* biprocesador a 3.0 *GHz* con 2 *GB* de *RAM* y 70 *GB* de disco duro actúa como frontal para la conexión y operación en el *cluster*. También usa el sistema operativo *Red Hat Enterprise Linux* 4. Todas las máquinas tienen montadas 3 unidades (*/data*, */scratch*, */home*) exportadas a través de *GPFS* desde una *SAN* de *IBM* con aproximadamente 6 *TB* de espacio por 2 servidores con doble tarjeta *GIGABIT* y *FC*. La red es de tipo *GIGABIT* fundamentalmente, aunque también hay una parte con *INFINIBAND* de 20 *GB*. Hay 2 *switches GIGABIT* de 48 conectores cada uno que van a cada nodo. Hay un conector en el switch dedicado a conectar la red interna con el laboratorio de trabajo donde cada persona dispone de una máquina con sistema operativo *Linux*, desde la cual además puede abrir una conexión por escritorio remoto a una máquina de 4 *cores* y 4 *GB* de *RAM* con *Windows Server 2008*. Este *cluster* tiene instalado el *software Portable Batch System* (*PBS*) de *OpenPBS* (ver apartado 3.1.2.2.5 de la página 44) para gestionar el uso de sus procesadores por los diferentes usuarios mediante un sistema de colas de trabajos. En la Figura 3-5 se muestra el *cluster* de la Unidad de Bioinformática.

Figura 3-5. *Cluster* de la Unidad de Bioinformática.

También cuenta con dos servidores de base de datos *MySQL* instalados en dos máquinas monoprocesador. Una es para datos de producción y otra para datos de pruebas. La máquina de producción es un *Pentium IV* a 2.8 *GHz* con 1 *GB* de *RAM*, 1 *TB* de disco duro y sistema operativo *Red Hat Enterprise Linux* 4; mientras que la de pruebas es un *Pentium IV* a 2.8 *GHz* con 1 *GB* de *RAM*, 125 *GB* de disco duro y sistema operativo *Fedora Core* 8.

3.1.3.2. MareNostrum en el BSC-CNS

MareNostrum es uno de los supercomputadores más potentes de Europa (ocupó la posición 40 del ranking mundial en Noviembre del 2008). Está alojado en el BSC-CNS (http://www.bsc.es/) (*Barcelona Supercomputing Center* – Centro Nacional de Supercomputación). Dispone de más de 10200 procesadores *PowerPC* con una capacidad total de cálculo de 94.21 *Teraflops*. Cuenta con un total de 20 *TB* de memoria *RAM* y con ~500 *TB* de disco duro. Trabaja bajo sistema operativo *Linux* en su última distribución *SUSE*.

3.1.3.3. Ibercivis

Ibercivis (http://www.ibercivis.es/) es una plataforma creada sobre *CIVICO* (Antolí et al., 2008), una infraestructura abierta destinada a la creación e implementación de nuevas plataformas de computación voluntaria basadas en la librería *BOINC* (ver apartado 3.1.2.2.1 de la página 43). Ibercivis trata de integrar la creciente necesidad científica de potencia de cálculo con la difícil tarea de divulgar los avances científicos. Hoy en día Ibercivis cuenta con más de 10000 núcleos de computación voluntarios.

3.2. Métodos

3.2.1. Automatización del Análisis Conformacional de Ligandos

La limitación más importante del programa *ALFA* es que no funciona de un modo automático, es decir, no es capaz de generar confórmeros por sí solo. Necesita de un fichero externo en el que el usuario le indicará cuáles son los enlaces que pueden rotar en la molécula y qué valores (en grados) pueden adoptar los ángulos diedros que forman. Además, tras la generación exhaustiva de los posibles confórmeros, necesita de un paso de postprocesado para seleccionar sólo aquellos de mejor energía. Debido a todo esto, resulta imposible integrar este programa dentro de un proceso automático de cribado que persigue la generación de confórmeros dentro de un programa de *docking*: habría que preparar manualmente cada fichero tras haber realizado un estudio de la molécula, de las que generalmente se tienen millones. Por ello se opta por realizar una nueva implementación tomando como base la idea del antiguo ALFA. Esta nueva implementación se hace en un lenguaje de programación más potente y flexible como es *C++*, apoyándose así en el paradigma de programación orientada a objetos. Además hace uso de la librería *OEChem TK* (ver apartado 3.1.2.2.7 de la página 44) para el manejo de diferentes formatos de ficheros, captura de parámetros de entrada, búsqueda de enlaces rotables y manipulación de patrones *SMARTS* (formato desarrollado por la compañía *Dayligh*, al igual que *SMILES*). Así pues, el nuevo *ALFA* presenta como principales ventajas, entre otras, sobre el anterior: la detección y clasificación automática de enlaces rotables (junto a la asignación de sus posibles valores de rotación) y la obtención de una lista final de confórmeros. Todas ellas se describirán en los siguientes apartados.

3.2.1.1. Detección Automática de Enlaces Rotables

La libertad de movimiento de los átomos dentro de una molécula está determinada inicialmente por aquellos enlaces que pueden girar, de modo que actúan como ejes internos para los átomos a ambos lados de dicho enlace y los que están unidos a ellos. Para determinar cuáles son los enlaces que pueden rotar, puede representarse la topología de la molécula en forma de grafo no dirigido, de modo que pueda aplicársele algoritmos pertenecientes a la teoría de grafos. En concreto es necesario un algoritmo que permita calcular sus puentes. Un puente es un enlace en el grafo que cuando se elimina deja dicho grafo desconectado, es decir, que no existe otro camino para llegar

desde uno de los átomos que formaban el enlace hasta el otro. Este puente equivaldría a un enlace rotable en la molécula, puesto que al no haber otro camino alternativo (no encontrarse en un ciclo) tendría libertad de giro. El cálculo de los puentes de un grafo se basa primeramente en calcular sus puntos de articulación, que son aquellos nodos del grafo que cuando se eliminan (junto con los enlaces que parten de él) dejan igualmente el grafo desconectado.

Pero en el caso de moléculas, no basta simplemente con tratar sus átomos como un grafo y encontrar sus puentes, ya que también habrá que tener en cuenta aspectos como: i) la cardinalidad de los enlaces, puesto que los enlaces con orden igual o superior a 2 no son rotables; ii) los átomos terminales, ya que aunque el enlace que lo une a otro átomo sea un puente, la rotación de dicho átomo terminal no produciría ningún cambio importante en la molécula. Así pues, para simplificar la búsqueda automática de torsionales en ligandos se ha empleado la librería de objetos y funciones *OEChem TK*. La molécula se lee directamente de un fichero en formato *mol2* (de la compañía *Tripos*) o *PDB* (mediante *OEReadMol2File* u *OEReadPDBFile* respectivamente) y se guarda dentro de un objeto tipo *OEMol*. A continuación se determinan los anillos (*OEFinRingAtomsAndBonds*), la aromaticidad (*OEAssignAromaticsFlags*) y la hibridación (*OEAssignHybridization*). En el caso de ficheros *PDB* también se necesita determinar los órdenes de enlace (algo ya explícito en ficheros tipo *mol2*) mediante la función *OEPerceiveBondOrders*. Con esta información, ya es posible determinar cuáles son los enlaces rotables mediante la llamada a la función *GetBonds* del objeto *OEMol* e indicando como condición *OEIsRotor*. Esto generará una lista (*OEIter*) de objetos tipo enlace (*OEBondBase*), los cuales contienen el átomo de inicio y fin de cada enlace.

3.2.1.2. Asignación Automática de Estados Rotaméricos

Una vez se ha solucionado el problema de obtener los enlaces que pueden rotar, lo siguiente que se necesita conocer es cuánto pueden hacerlo, o más bien en qué posiciones se pueden colocar de modo que se reduzca la probabilidad de obtener choques estéricos entre los átomos de la molécula. Para ello lo primero será definir el modo en que se medirá la posición de un torsional. Esto se hará mediante el ángulo diedro, es decir, el ángulo que forman los vectores característicos de los planos formados por los átomos de los enlaces rotables. De este modo, se define un torsional mediante 4 átomos: si por ejemplo los llamásemos *ABCD*, significaría que hay un

enlace rotable entre *B* y *C*, y que la posición de ese enlace (los grados a los que se encuentra) se mide mediante el ángulo que forman los vectores característicos de los planos *ABC* y *BCD*; *A* sería un átomo antecesor de *B* y *D* un antecesor de *C*.

Combinando diferentes ángulos en cada torsional de la molécula se obtienen los confórmeros de la misma, es decir, las diferentes estructuras que puede presentar. Aunque se tiene el problema de la explosión combinatoria. Si por ejemplo se considerase que cada ángulo diedro se puede mover en pasos de 1 grado, cada uno tendría 360 posiciones diferentes. Así pues una molécula pequeña con sólo 5 enlaces rotables tendría 360^5 combinaciones diferentes, algo inabordable. Por ello se busca restringir el movimiento de cada torsional a los ángulos más probables. Para hacerlo se estudian los grupos químicos de la molécula, de modo que en función de aquéllos que se encuentran a ambos lados de los átomos que delimitan un enlace se pueden conocer cuáles son los grados que puede tomar el diedro a partir de estudios que se han hecho ya sea experimental o teóricamente. Así los torsionales pueden clasificarse en tipos y según su tipo se les asignan una serie de ángulos que pueden rotar. Este conjunto de ángulos suele tener un tamaño medio de 4 por torsional, reduciéndose notablemente el problema de la explosión combinatoria.

Para la selección de ángulos diedros se ha partido de un conjunto de reglas en formato *SMARTS* que especifican diferentes alternativas de grupos funcionales junto con los átomos que definen los torsionales y los ángulos posibles. *SMARTS* es un lenguaje de patrones basado en *SMILES*, que a su vez es un lenguaje para representar moléculas en *2D* de un modo extremadamente comprimido. Lo primero que se hace es determinar la regla general (ver Tabla 3-V) para aplicar al torsional. Esto consiste en decidir cuáles son los ángulos que puede adquirir su diedro en función de la hibridación (determinada con el *OEChem TK*) de los dos átomos centrales del mismo.

Hibridación	Ángulos (grados)
sp^3-sp^3	-60 60 180
sp^2-sp^2	-120 -60 0 60 120 180
sp^3-sp^2	0 180

Tabla 3-V. Reglas generales de asignación de ángulos.

Tras seleccionar la regla general lo que se hace es comprobar si ésta debe sobrescribirse con alguna regla particular de grupo funcional. El modo de realizar esto es teniendo el conjunto de reglas en forma de lista ordenada de menor a mayor prioridad, de tal manera que, por ejemplo, las primeras reglas que aparecen son las

generales y según se avanza en la lista estas son más específicas o diferenciadoras (particulares). Así se asegura que siempre se selecciona una regla para cada torsional.

En el programa, las reglas se almacenan como objetos tipo *TorsionRule*, los cuales tienen como atributos el nombre de la regla, el patrón *SMARTS*, el índice dentro del patrón de los cuatro puntos que determinan el diedro, y la lista de los posibles ángulos que puede adoptar. Para torsionales que se consideren rígidos no se especificará ningún ángulo (como por ejemplo para los trifluorometilos). En la Tabla 3-VI se muestran las reglas que incorpora el nuevo *ALFA* por defecto, aunque este conjunto puede ampliarse/sustituirse mediante un fichero externo.

Tabla 3-VI. Tipos de torsionales, patrones y conjuntos de ángulos.

Nombre	Patrón	Átomos diedro	Ángulos (grados)
SP3-SP3	*~[*^3]~[*^3]~*	1 2 3 4	-60 60 180
SP3-SP2	*~[*^3]~[*^2]~*	1 2 3 4	-120 -60 0 60 120 180
SP2-SP2	*~[*^2]~[*^2]~*	1 2 3 4	0 180
polysaccarideBridges1	O@[CD3]O[CD3]([#1])@C	2 3 4 5	0 30 -30 180
polysaccarideBridges2	O@[CD3]([#1])O[CD3]([#1])@C	3 2 4 5	0 30 -30 180
acids1	[OD1]~C(~[OD1])[CX4](*)*	1 2 4 5	30 -30 -60 60 90 -90 0 180
acids2	[a]cC([OD1])=O	1 2 3 4	0 180 60 -60 90 -90 20 -20
acids3	[OD1]C(=O)[CD2]C	1 2 4 5	0 45 90
sulfonamides1	NS(=O)(=O)c1[cD2][cD2]a[cD2][cD2]1	1 2 5 6	90
sulfonamides2	c([aD2])S(=O)(=O)[ND2][CD2]	1 3 6 7	60 -60
sulfonamides3	O=S(=O)N[CX4D3]*	2 4 5 6	-90 90 120 -120
sulfonamides4	O=S(=O)N[CX4D2]*	2 4 5 6	-90 90 120 -120
sulfonamides5	[c]S(=O)(=O)NC	1 2 5 6	-70 70 90 -90 50 -50
sulfonamides6	*=*-S(=O)(=O)C	1 2 3 6	90 -90 60 -60
sulfonamides7	O=S(=O)N[CH2]	1 2 4 5	-60 60 180 0 30 -30
sulfonamides8	[aD2]c([aD2])S(=O)(=O)[ND2^3]	1 2 4 7	90 -90 120 -120 60 -60
sulfonamides9	[aD2]c([aD3])S(=O)(=O)[ND2^3]	1 2 4 7	80 -80 110 -110
sulfonamides10	[aD3]c([aD3])S(=O)(=O)[ND2^3]	1 2 4 7	70 -70 110 -110
sulfonamides11	[aD2]c([aD2])S(=O)(=O)[CD2^3]	1 2 4 7	90 -90 110 -110 70 -70
sulfonamides12	[a]cS(=O)(=O)[C,N]	1 2 3 6	90 -90 60 -60 30 -30 0
sulfone1	O=S(=O)[CD2][CD3][#1]	2 4 5 6	30 -30
hydrazides1	[O,S]=C[ND2][ND2]	1 2 3 4	0 180
hydrazides2	[O,S]=C[ND2][ND2]-,=*	2 3 4 5	180 90 -90
cyclopropyl-ketones1	O=CC1([#1])[CD2][CD2]1	1 2 3 4	180
cyclopropyl-ketones2	O=CC1([#1])CC1	1 2 3 4	180 160 -160 0 20 -20
cyclopropyl-ketones3	O=CC1([*])CC1	1 2 3 4	180 160 -160 0 120 -120 90 -90 30 -30
epoxy-ketone1	O=C([*D2])C1([#1])O[CD2,CD3]1	1 2 4 5	0 180
oppositeEndOfTertAmide1	O=C([ND3])[CD2]*	1 2 4 5	0 30 -30 100 -100 80 -80
oppositeEndOfTertAmide2	O=C([CD3^3])[CD2]*	1 2 4 5	0 30 -30
oppositeEndOfTertAmide3	O=C([ND3])[CD3][#1]	1 2 4 5	180 150 -150 120 -120
misc1	[CD2]C(=O)[ND2]-!@[CD3][#1]	2 4 5 6	0 30 -30 60 -60 180
misc2	[cD2]c([cD2])-!@[CD2^3][CD3^3]	1 2 4 5	90 -90 70 -70 110 -110
misc3	c[CD2][ND3](C)c	1 2 3 4	90 -90 60 -60 120 -120
carbonyls1	O=CC=O	1 2 3 4	180 0 120 -120 90 -90
carbonyls2	C=CC=O	1 2 3 4	0 180 20 -20 160 -160
carbonyls3	O=C[CD2][ND2]	1 2 3 4	0 -30 30 150 -150 180

Nombre	Patrón	Átomos diedro	Ángulos (grados)
carbonyls4	O=C[CD2]C=O	1 2 3 4	0 -30 30 60 -60 130 -130
carbonyls5	O=C(c)[ND2][CD3][#1]	2 4 5 6	0 -30 30
carbonyls6	O=C[ND2][CD3]*	2 3 4 5	20 -20 120 -120 60 -60 0
carbonyls7	O=CN[CD2]*	2 3 4 5	180 150 -150 -120 120 0 30 -30
carbonyls8	O=Ccc[OD1]	1 2 3 4	0 180 90 -90 30 -30
carbonyls9	O=C[CD4][CD1]	1 2 3 4	0 30 -30 60 -60 120 -120
carbonyls10	O=C[CD3][OD1]	1 2 3 4	0 30 -30 60 -60 120 -120
carbonyls11	O=C[CD2][CD1]	1 2 3 4	0 30 -30 60 -60 90 -90 120 -120
carbonyls12	O=C[CD3][#1]	1 2 3 4	0 30 -30 180
amideneAndGuanidine1	[aD3]cC(~[ND1])~[ND1]	1 2 3 4	0 30
amideneAndGuanidine2	[a]cC(~[ND1])~[ND1]	1 2 3 4	0 30
amideneAndGuanidine3	*[ND2]~C(~[ND1])~[ND1]	1 2 3 4	0 30
amideneAndGuanidine4	[CD2][CD2][ND2]~C(~[ND1])~[ND1]	1 2 3 4	-70 70 90 -90 110 -110
ether1	aCO[CD2]	1 2 3 4	180 100 -100
isoprene1	C=C[CX4D2]*	1 2 3 4	0 180 90 -90 60 -60 30 -30
isoprene2	C=Cc[a]	1 2 3 4	0 90 -90 180 30 -30 150 -150
arylSecondaryAmines1	[aD2]c([aD2])[ND2][CD2]	1 2 4 5	0 180
arylSecondaryAmines2	[aD2]c([aD3])[ND2][CD2]	1 2 4 5	0
arylSecondaryAmines3	[aD2]c([aD2])[ND2][CD1]	1 2 4 5	0 90 -90 180
arylSecondaryAmines4	ac[ND2][CD2]	1 2 3 4	90 -90 160 -160 20 -20
aromaticSubtituents1	[aD3]c([aD3])[CD2]C	1 2 4 5	90 -90 60 -60 120 -120
aromaticSubtituents2	[aD2]c([aD2])[ND3]([CD1])[CD2]	1 2 4 5	0 180
aromaticSubtituents3	[aD3][c,n]([aD2])[C^3D3][#1]	1 2 4 5	0 -30 30 60 -60 160 -160
aromaticSubtituents4	a[CD2X4][ND3^3]*	1 2 3 4	60 -60 180 160 -160 90 -90 120 -120
aromaticSubtituents5	an[CD2X4][CD1]	1 2 3 4	90 -90
aromaticSubtituents6	[aD3]c([aD2])C(=O)[C^3]	1 2 4 5	0 20 -20 150 -150 180
aromaticSubtituents7	[aD3]c([aD2])O[CD2]	1 2 3 4	180
aromaticSubtituents8	a[ND2][CD2X4][CD2X4]	1 2 3 4	180 160 -160 80 -80 60 -60
aromaticSubtituents9	[ND1]C(=O)c([aD3])	1 2 4 5	0 180 30 -30 150 -150
aromaticSubtituents10	[aD2]c([aD2])c([aD2])[aD2]	1 2 4 5	-150 -30 30 150
aromaticSubtituents11	[a]c[CD2][*D2]	1 2 3 4	-90 90 180 0 30 -30 150 -150
aromaticSubtituents12	[a]cC(=O)c[a]	1 2 3 4	-150 -30 0 30 150 180
aromaticSubtituents13	[a]cC(=O)[*D2]	1 2 3 4	0 180 30 -30 150 -150
aromaticSubtituents14	[a]cOC	1 2 3 4	0 180 30 -30 150 -150
borderlineLow-res1	[CD2]C(=O)[ND2][CD3][#1]	2 4 5 6	90 -90 60 -60 120 -120
conjugatedSubstituents1	a[CD2]C=*	1 2 3 4	150 -150 180 30 -30 0
conjugatedSubstituents2	C=CC=C	1 2 3 4	0 180 30 -30 150 -150 60 -60 120 -120
conjugatedSubstituents3	cO[CD2]*	1 2 3 4	0 30 -30 60 -60 90 -90 180
conjugatedSubstituents4	C=N[ND2]*=,:*	2 3 4 5	0 30 -30 150 -150 180
conjugatedSubstituents5	c[CD2][ND2]c	1 2 3 4	60 -60 80 -80 180
conjugatedSubstituents6	C=[CD3][ND3]*	1 2 3 4	30 -30 60 -60 90 -90 0 180
ureas1	[ND2]C(=O)Nc[nD2]	2 4 5 6	0 180
ureas2	[ND2]C(=O)[ND2]*	1 2 4 5	0 180
carbamates1	C[ND2]C(=O)O	1 2 3 4	0 180
carbamates2	[ND2]C(=O)OC	3 2 4 5	0
carbamates3	OC(=O)N*	3 2 4 5	0 20 -20 120 -120 160 -160 180

Nombre	Patrón	Átomos diedro	Ángulos (grados)
piperidineAmide1	O=CN1[CD2][CD2][CD2][CD2][CD2]1	1 2 3 4	0
amidesAndEsters1	[*D2]C(=O)O[CD3][#1]	2 4 5 6	0 30 -30
amidesAndEsters2	[OD2]C(=O)[CD2][CD2^3]	3 2 4 5	0 30 -30 120 -120 180
amidesAndEsters3	[O,SD1]=C(C)[ND2]C=[O,S]	1 2 4 5	0 180
amidesAndEsters4	[O,SD1]=C(C)[ND2][#7,#8]=*	1 2 4 5	0 180
amidesAndEsters5	[O,SD1]=C(C)[ND2]N	1 2 4 5	0 180
amidesAndEsters6	[O,SD1]=C(C)cn	1 2 4 5	0 180
amidesAndEsters7	[O,SD1]=C([#6])[ND2]*	1 2 4 5	0 20 -20
amidesAndEsters8	[O,SD1]=C[ND2]*	1 2 3 4	0 20 -20 180
amidesAndEsters9	O=C[ND3][CD3X4][#1]	2 3 4 5	0 180 20 -20
amidesAndEsters10	O=CNc([aD2,aD3])[aD3]	2 3 4 5	20 -20 -90 90 60 -60 120 -120 0
amidesAndEsters11	O=CNc[a]	2 3 4 5	-20 20 90 -90 -160 160
amidesAndEsters12	O=C([CD2,CD3])O[CD2]	1 2 4 5	0
amidesAndEsters13	O=C([CD1])O[CD1]	1 2 4 5	0
amidesAndEsters14	[O,S]=CO[CD1]	1 2 3 4	0 20 -20 180
amidesAndEsters15	O=CO[CD2][CD1]	2 3 4 5	180
amidesAndEsters16	O=CO[CD2]*	2 3 4 5	180 60 -60 90 -90
amidesAndEsters17	O=CO[CD3]*	2 3 4 5	120 -120 180 0 60 -60
amidesAndEsters18	O=CO[CD4]*	2 3 4 5	-60 60 120 80 -80
amidesAndEsters19	O=CO[CD3,CD4]	1 2 3 4	0 30 -30 60 -60
amidesAndEsters20	O=CO*	1 2 3 4	0 30 -30 60 -60
amidesAndEsters21	O=C[ND3]([*D3])[*D3]	1 2 3 4	20 -20 0 180 150 -150
amidesAndEsters22	O=C[ND3]*	1 2 3 4	0 180
amidesAndEsters23	CC[ND3](CC)[CD2,CD3]*	2 3 6 7	180 60 -60 120 -120 0 180 30 -30
amidesAndEsters24	[a][CD2][CD2][ND3]	1 2 3 4	90 -90 180 60 -60
amidesAndEsters25	[ND3]C(=O)[nD3]*	1 2 4 5	90 -90 60 -60 120 -120
amidesAndEsters26	[CD2]OC(=O)[CD2][CD3]	4 3 5 6	0 150 -150
t-butyl1	C([CD1])([CD1])([CD1])c[a]	2 1 5 6	90 30
t-butyl2	**C([CD1])([CD1])[CD1]	1 2 3 4	180 150
propyl1	[CD1]C([CD1])([#1])[CD2]*	4 2 5 6	60 -60
highlySubstitutedAlkane1	*[CD2X4][CD3X4]([#1])[CD3]	1 2 3 4	180 60 -60 40 -40
highlySubstitutedAlkane2	c[CD2^3][CD3^3][#1]	1 2 3 4	180 60 -60
highlySubstitutedAlkane3	[CD2^3][CD2^3][CD3^3][#1]	1 2 3 4	60 -60 180 30 -30 0 160 -160 120 -120
highlySubstitutedAlkane4	[*D2][CD2][CRH]([*R])[*R]	1 2 3 4	30 -30 120 -120 150 -150 60 -60 180 0
highlySubstitutedAlkane5	[*D2][CD2][CX4D3][*D2]	1 2 3 4	30 -30 120 -120 150 -150 60 -60 180
highlySubstitutedAlkane6	*[CHD3][CH2D2]*	1 2 3 4	150 -150 60 -60 180 -90 90 0 30 -30
highlySubstitutedAlkane7	[CD1]C([CD1])[CD2]*	1 2 4 5	60 -60 180 80 -80 30 -30
nitro1	[aD3]cN(~[OD1])~[OD1]	1 2 3 4	0 60 -60
nitro2	**N(~[OD1])~[OD1]	1 2 3 4	-
trifluoromethyl1	**C(F)(F)F	1 2 3 4	-
trichloromethyl1	[a]cC(Cl)(Cl)Cl	1 2 3 4	-
CSDSPECIFICRULES1	a[PD3](a)-[PD3](a)a	1 2 4 5	180 60 -60
CSDSPECIFICRULES2	PPcc	1 2 3 4	60 -60
phosphorusContainingGroups1	[OD1]~PO*	1 2 3 4	0 -30 30 -60 60 120 -120
phosphorusContainingGroups1	[OD1]~P(~[OD1])(~[OD1])[OD2][CD2]*	2 5 6 7	0 60 120 180 -120 -60
phosphorusContainingGroups1	S=POc	1 2 3 4	0 -60 60 90 -90
phosphorusContainingGroups1	[a]cCP(c)(c)c	1 2 3 4	90 -90
interRingPair1	[R][R][C^2,N^2;!R][R]	1 2 3 4	-150 30
AE11M	O=C[ND3]c[a]	2 3 4 5	-20 20 90 -90 -160 160 0 180
Ar-Ar	[a][a][a][a]	1 2 3 4	15 -15 165 -165
Alr-Ar	[R!a][R!a][a][a]	1 2 3 4	60 -60 120 -120

Nombre	Patrón	Átomos diedro	Ángulos (grados)
interRingPair2	`[a][a][C^2,N^2,O^3;!R][R]`	1 2 3 4	-30 30 -90 90 -150 150 180
IRP2	`[R!a][R!a]C(=O)[R]`	1 2 3 5	0 150 -150
IRP3	`[a][a][C^3][R!a]`	1 2 3 4	-60 -120 60 120
IRP4	`[R][ND3;R]C(=O)[R]`	1 2 3 5	0 180
sulfonamides13	`[a]S(=O)(=O)N[H]`	1 2 5 6	30 -30

El proceso de asignación de reglas a torsionales hace uso de objetos *OESubSearch* para buscar coincidencias (objetos *OEMatchBase*) de un patrón *SMARTS* dentro de la estructura topológica del ligando. Para cada coincidencia se encuentra el torsional al que corresponde, se actualizan los puntos de referencia (los llamados átomos *A* y *D*), y todo ello junto con el tipo de torsional y sus ángulos se almacena en un objeto *RotatableBond*. Los torsionales que finalmente quedan sin ángulos son eliminados (no rotables). El Algoritmo 3-1 muestra este proceso.

```
para cada regla hacer
   subBusqueda <- crearSubBusqueda(regla.patron)
   coincidencias <- aplicarSubBusqueda(subBusqueda, molécula)
   para cada torsional
      para cada coincidencia
         si coincidencia pertenece al torsional
            tomar referencias, tipo y ángulos
         fin si
     fin para
   fin para
fin para
para cada torsional
   si numero de ángulos es 0
      borrar(torsional)
   fin si
fin para
```

Algoritmo 3-1. Asignación de reglas de torsional en el nuevo *ALFA*.

3.2.1.2.1.1 Uso de la Información 3D Original

Con el objetivo de mejorar el conjunto de ángulos que se asigna a cada torsional, es posible indicar al programa que añada a dicho conjunto el ángulo que ya tenía la estructura *3D* de partida del ligando, ya que los ángulos iniciales suelen pertenecer a una estructura válida. En caso de añadir para un torsional el ángulo original, éste reemplazará a cualquiera de los ángulos de su regla asignada siempre que la diferencia entre ambos sea menor de 30º, evitando así la redundancia.

3.2.1.2.2 Eliminación de Ángulos Equivalentes

En ocasiones dos ángulos diferentes de un torsional pueden dar origen a la misma estructura (debido a la simetría). Para evitar hacer cálculos innecesarios generando estructuras repetidas, uno de los ángulos es eliminado. Para detectar estos ángulos lo que se hace es dejar fijos todos los torsionales excepto el que se quiere comprobar, y

después generar las dos estructuras a las que darían lugar los dos ángulos que se quieren evaluar. Si el *RMSD* (calculado con la función *OERMSD* del *OEChem TK*) de la estructura generada es menor a un cierto valor *delta* (un número real muy pequeño), se considera que los ángulos son equivalentes y se elimina uno de ellos. El proceso se resume en el Algoritmo 3-2.

```
delta = 0.003
para cada torsional
   para cada ángulo a1
      estructura1 <- generarEstructura(a1)
      para cada ángulo restante a2
         estructura2 <- generarEstructura(a2)
         si RMSD(estructura1, estructura2) menor que delta
            eliminar(a2)
         fin si
      fin para
   fin para
fin para
```

Algoritmo 3-2. Eliminación de ángulos equivalentes

3.2.1.3. Generación de Confórmeros

Una vez se conocen los torsionales del ligando y sus conjuntos de ángulos posibles, el siguiente paso es la generación de confórmeros formando todas las combinaciones posibles de ángulos. Pero a pesar de haber introducido la funcionalidad para la asignación automática de ángulos aún se tiene el problema de la explosión combinatoria cuando se está tratando una molécula grande con muchos torsionales. Por ello se introduce un algoritmo que es capaz de explorar el abanico de estados conformacionales sin necesidad de generarlos exhaustivamente. Dicho algoritmo hace uso de un generador de números aleatorios para producir los nuevos confórmeros y utiliza un proceso llamado de "enfriamiento simulado" (*simmulated annealing*) (Wilson & Cui, 1990) para dirigir la amplitud de los saltos en el espacio conformacional. Se le da este nombre por su analogía con los sistemas físicos en los que haciendo descender poco a poco la temperatura se llega al estado más estable, que en este caso particular correspondería al mejor confórmero. Por lo tanto es un proceso en el que se pretende buscar el mínimo global, ya que se desean confórmeros lo más estables posibles y esto se logra con energías bajas. El hecho de que las energías que no son mejores a las últimas aceptadas puedan llegar a aceptarse en función de cierta probabilidad evita que el proceso se quede atrapado en mínimos locales. Así pues a este proceso completo de búsqueda estocástica se le denomina *Monte Carlo/Simmulated Annealing* (*MCSA*). El nombre *Monte Carlo* hace referencia al uso de número aleatorios.

Como visión general del proceso de *MCSA* se puede decir que se tiene una temperatura inicial y una serie de rondas de generación de confórmeros a una temperatura dada tomando como base el último aceptado. Un confórmero se acepta o no aplicando el criterio de Metropolis por el que se aceptan todos aquellos cuya energía sea mejor que la última aceptada, o bien en caso de que no lo sea, se aceptarían con una cierta probabilidad dependiente de la temperatura. Así pues, en cada ronda se tiene un número máximo de confórmeros generados o aceptados, y cuando se alcanza uno de estos números se pasa a la siguiente ronda descendiendo a su vez la temperatura. En el Algoritmo 3-3 se muestra el funcionamiento general del nuevo modo de generar los confórmeros en *ALFA*. En dicho algoritmo hay 6 puntos numerados que serán comentados más en detalle en los siguientes apartados.

```
guardarConformeroEntrada
possibleConformers <- resulado de Ecuación [3-3]
maxCombinations <- parámetro entrada (defecto 300000, 0 = ilimitado)
lastAcceptedConformer <- (0,...,0)    | Punto 1

si (maxCombinations != 0) y (possibleConformers > maxCombinations)
   MCSA <- true
   inicializarGeneradorNumerosAleatorios
   temperature <- temperatura inicial
   maxGeneratedPerRoundMCSA <- MAX_GENERATED_PER_MCSA_ROUND * numEnlacesRotables
   maxAcceptedperRoundMCSA <- maxGeneratedPerRoundMCSA * MAX_ACCEPTED_PER_MCSA_ROUND
fin si

repetir
   si MCSA
      conformero <- obtenerConformeroMCSA(lastAcceptedConformer)    | Punto 2
      conformerosGeneradosPorRonda++
   si no
      conformero <- obtenerSiguienteConformero(lastAcceptedConformer)    | Punto 3
   fin si

   hacerRotaciones(conformero)
   energia <- evaluarEnergia(conformero)

   si MCSA
      aceptarConformero <- criterioMetropolis(energia,lastAcceptedconformer)    | Punto 4
      conformerosAceptadosPorRonda++
   si no
      aceptarConformero <- true
   fin si

   si aceptarConformero
      lastAcceptedConformer = conformero
      evaluarInsercionEnLista(conformero)    | Punto 6 (a)
   fin si
hasta (cumpleCriterioSalida)    | Punto 5

limpiarRepetidos(lista)    | Punto
aplicarCutoff(lista)       |  6 (b)
```

Algoritmo 3-3. Generación de confórmeros en *ALFA*.

En primer lugar se toma la estructura de entrada como primer confórmero, para a continuación calcular el número máximo de confórmeros que se pueden formar mediante la Ecuación [3-3]:

$$possibleConformers = \prod_{i=1}^{n} numAng(i) \qquad \textbf{[3-3]}$$

donde *n* representa el número de torsionales de la molécula y *numAng(i)* el número de ángulos para el torsional *i*. Después se determina el máximo número de combinaciones que se van a generar. Este parámetro se toma de la línea de comandos y por defecto tiene un valor 300000. En caso de que se deseen generar todas las combinaciones posibles se le puede asignar el valor 0.

También hay que inicializar la variable que guarda el último confórmero aceptado (*lastAcceptedConformer*) en un vector cuyo tamaño es igual al número de torsionales de la molécula. Los detalles de la codificación de los confórmeros pueden verse en el apartado 3.2.1.3.1 de la página 57.

Antes de entrar en el proceso de generación de confórmeros, se decide en base a la variable *maxCombinations* si se hará de modo exhaustivo o por *MCSA*. Cuando *maxCombinations* sea distinto de 0 y menor al número de confórmeros posibles, se utilizará *MCSA* para explorar el espacio conformacional. Para ello en primer lugar se inicializa el generador de números aleatorios, la temperatura inicial, el número máximo de confórmeros generados y aceptados por ronda de *MCSA*. A continuación se entra en el bucle de generación de confórmeros, que se hace de modo diferente en los casos *MCSA* (ver apartado 3.2.1.3.2 de la página 58) y exhaustivo (ver apartado 3.2.1.3.3 de la página 58). Tras la generación de los parámetros de un nuevo confórmero, se efectúan las rotaciones (haciendo uso de la clase *OESetTorsion* del *OEChem TK*) y se calcula la energía interna de la estructura generada. El nuevo confórmero es aceptado como válido en el caso exhaustivo, pero en el caso *MCSA* será el criterio de Metropolis el que determine si se acepta o no (ver apartado 3.2.1.3.4 de la página 59). Si el confórmero es aceptado, deberá determinarse si se inserta o no en la lista de soluciones provisionales dependiendo de su energía y del espacio libre en dicha lista (ver apartado 3.2.1.3.6 de la página 60). Tras la generación de cada confórmero también se evalúa si hay que continuar o no con el proceso. Este proceso es también algo diferente para los modos exhaustivo y *MCSA* (ver apartado 3.2.1.3.5 de la página 59). Finalmente, tras terminar el proceso de generación de confórmeros, se realiza una limpieza en la lista final de soluciones (ver apartado 3.2.1.3.6 de la página 60).

3.2.1.3.1 Codificación de Confórmeros (Punto 1)

Cada confórmero se codifica en forma de vector cuyo tamaño es igual al número de torsionales de la molécula. Cada uno de los campos de dicho vector contiene el índice del ángulo usado dentro del conjunto de ángulos de cada torsional. El valor 0 se utiliza para indicar que aún no se ha asignado ningún ángulo.

A partir de un vector de confórmero, se puede obtener a su vez un número que representa la combinación de ángulos que codifica dicho vector. Este número, llamado índice del confórmero, es único, y sirve para identificar unívocamente un confórmero dado dentro de una molécula. Para obtenerlo, en primer lugar hay que construir otro vector del mismo tamaño en el que se almacenará en cada celda el peso para el torsional que representa, es decir, el número que indicaría cuántas combinaciones se han generado en los torsionales de su parte izquierda. La fórmula general para rellenarlo sería:

$$\begin{aligned} vector(1) &= 1 && Para\ i = 1 \\ vector(i) &= numAng(i-1) * vector(i-1) && Para\ i > 1 \end{aligned} \qquad \textbf{[3-4]}$$

donde *i* representa el número de torsional para la posición, y *numAng* es el número de ángulos para el número de torsional dado entre paréntesis. Con los valores de este vector podría calcularse el índice del confórmero que representa un vector de codificación de confórmero de la siguiente manera:

$$indice = \sum_{i=1}^{n} (conformer(i) - 1) * vector(i) \qquad \textbf{[3-5]}$$

donde *n* representa el número de torsionales, *conformer* el vector que contiene la combinación de índices de ángulos para cada torsional, y *vector* es el vector de tamaños previamente calculado. A continuación se muestra un ejemplo:

Una molécula con 7 torsionales teniendo:	`numAng = [3, 2, 4, 3, 1, 6, 2]`
generaría el vector de tamaños:	`vector = [1, 3, 6, 24, 72, 72, 432]`
Con ello el confórmero:	`conformer = [2, 1, 2, 3, 1, 4, 2]`
obtendría el identificador:	
`(2-1)*1+(1-1)*2+(2-1)*6+(3-1)*24+(1-1)*72+(4-1)*72+(2-1)*432 =` **`703`**	
de un total de `864 (3*2*4*3*1*6*2)` *posibles confórmeros.*	

3.2.1.3.2 Generar Confórmero en Modo MCSA (Punto 2)

Para generar un nuevo confórmero en el proceso *MCSA* se realizan modificaciones aleatorias sobre el último confórmero aceptado por el criterio de Metropolis (ver 3.2.1.3.4 de la página 59). En caso de que aún no haya ninguno, se inicializan aleatoriamente todos los valores.

A la hora de modificar un confórmero, se decide primeramente cuantos cambios se van a realizar (el número de torsionales cuyo ángulo se va a cambiar). Por defecto este valor es 1 (constante *CHANGES_IN_MCSA = 1*), produciéndose en el 30% de las ocasiones (constante *PROB_EXTRA_CHANGE = 0.30*) otro cambio extra. A continuación se decide aleatoriamente cuál o cuáles serán los torsionales a modificar y se les asigna un valor aleatorio diferente al que ya tuvieran (en el caso de ser posible). Dicho valor representa un nuevo índice de ángulo dentro del conjunto de los posibles para ese torsional en cuestión.

3.2.1.3.3 Generar Confórmeros en Modo Exhaustivo (Punto 3)

En este caso simplemente habrá que modificar el último confórmero aceptado (en este modo se aceptan todos) sumando 1 a la posición más a la izquierda (primer torsional) del vector que lo representa. Si tras la suma en una celda el número obtenido es mayor al número de ángulos para ese torsional, se volverá a poner en 1 (primer ángulo) acarreando un 1 que se sumará con el valor del siguiente torsional. Este proceso se repite hasta que no se producen acarreos o bien se han recorrido todos los torsionales. Un ejemplo sería:

Para una molécula con 7 torsionales teniendo:

`numAng = [3, 2, 4, 3, 1, 6, 2]`

`conformer.i = [3, 2, 4, 2, 1, 4, 2]`

el `conformer.(i+1)` *sería:*

posición 1: `3 + 1 = 4` *-> se pone 1 y se acarrea 1 para sumar a la siguiente posición*

posición 2: `2 + 1 (acarreo) = 3` *-> se pone 1 y se acarrea 1 a la siguiente posición*

posición 3: `4 + 1 (acarreo) = 5` *-> se pone 1 y se acarrea 1 a la siguiente posición*

posición 4: `2 + 1 (acarreo) = 3` *-> como 3 es <= 3 (número de ángulos para la posición 4) se mantiene este valor y no se produce acarreo, por lo que termina el proceso dejándose igual el resto de valores, obteniéndose:*

`conformer.(i+1) =` **`[1, 1, 1, 3, 1, 4, 2]`**

donde *numAng* es el número de ángulos disponibles para el torsional dado entre paréntesis, y *conformer.i* y *conformer.(i+1)* es un confórmero dado y su sucesivo.

3.2.1.3.4 Criterio de Aceptación de Metropolis (Punto 4)

Consiste en aceptar el nuevo confórmero siempre que su energía sea menor que la del último confórmero aceptado, o bien en caso de de no serlo se aceptaría con cierta probabilidad dada por:

$$p = e^{-delta\,/\,temperatura} \qquad \textbf{[3-6]}$$

dónde *e* es aproximadamente 2.7182818 y *delta* es:

$$delta = newEnergy - lastEnergy \qquad \textbf{[3-7]}$$

siendo *newEnergy* la energía del nuevo confórmero a evaluar y *lastEnergy* la del último aceptado. De este modo se irá dirigiendo la búsqueda hacia el mínimo global descendiendo la temperatura en cada ronda en un factor determinado (en este caso, *ANNEALING_SCHEDULE = 0.85*). El hecho de aceptar confórmeros peores que el último en función de cierta probabilidad ayuda a evitar quedar atrapados en mínimos locales además de favorecer la exploración homogénea del espacio conformacional de la molécula.

3.2.1.3.5 Condición de Salida (Punto 5)

El criterio de salida para el caso de la generación exhaustiva de confórmeros es sencillo, simplemente se comprueba que se hayan generado ya el número total de confórmeros posibles. En el caso de que se esté haciendo *MCSA* se tienen dos criterios de salida: 1) que el número de confórmeros generados sea igual al máximo de confórmeros a generar (*maxCombinations*), ó 2) que finalice una ronda sin haber aceptado ningún confórmero. Se considera que una ronda *MCSA* ha finalizado en los siguientes casos: i) que se haya alcanzado el máximo número de confórmeros generados por ronda (*maxGeneratedPerRoundMCSA*), o ii) que se haya alcanzado el máximo número de confórmeros aceptados por ronda (*maxAcceptedperRoundMCSA*). Al finalizar una ronda se pondrán a 0 sus contadores de confórmeros aceptados y generados, y se decrementará la temperatura multiplicándola por el parámetro *ANNEALING_SCHEDULE* (0.85).

3.2.1.3.6 Lista de Soluciones (Punto 6)

Para mantener las soluciones generadas (confórmeros) se crea una lista dinámica cuyo tamaño lo determina el parámetro que indica el número máximo de soluciones que se quieren obtener (*-howManySelect*, que por defecto es 100). En el caso de *MCSA*, el tamaño máximo de esta lista se incrementa en un 50% ya que en ocasiones, debido a la generación aleatoria, se producen resultados repetidos que posteriormente habrá que eliminar. La lista se encuentra ordenada de menor a mayor energía interna del confórmero, de modo que al principio de la lista se tienen los teóricamente más estables. El Algoritmo 3-4 describe como sería la inserción de un nuevo elemento en la lista.

```
sea L la lista dinámica de confórmeros
sea e el nuevo elemento a insertar
insertar <- falso
si (tamaño(L) >= MAXIMO) entonces
   si (energía(ultimo(L)) > energía(e) entonces
      insertar <- verdadero
   fin si
fin si
si (insertar) entonces
   para cada conformero c en L
      si energía(e) < energía(c)
         insertarAntes(L, e, c)
         yaInsertado = verdadero
         terminar bucle
      fin si
   fin para
   si no yaInsertado
        insertarUltimo(L, e)
   fin si
fin si
```

Algoritmo 3-4. Inserción de un confórmero en la lista dinámica.

Una vez se ha completado el proceso de generación de confórmeros, la lista final se trunca hasta el valor del parámetro *howManySelect*, tras haber eliminado los posibles repetidos (en el caso *MCSA*) basándose en su índice de confórmero. Después se eliminan también aquellos confórmeros cuya energía interna sea menor a *[mínima encontrada + cut-off]*. El valor por defecto del parámetro *cut-off* es 0.0 y significa que no se aplique dicho corte. El objetivo de aplicar un *cut-off* es eliminar estructuras que al ser muy diferentes en energía a las mejores es posible que se deba a que puedan estar presentado algún choque estérico.

3.2.2. Nuevo Modelo de Solvente Implícito

Como ya se ha comentado, la desolvatación es una componente importante de la energía libre de unión en complejos receptor-ligando. Su cálculo requiere de tiempos de computación más elevados que otras componentes y esto hace que, por ejemplo, su uso dentro de un proceso de cribado virtual esté restringido a etapas finales de

postprocesado y reordenación. Por lo tanto, sería conveniente disponer de un modelo electrostático menos costoso pero tan riguroso como los comúnmente utilizados (resolución de la ecuación de *Poisson-Boltzmann*, por ejemplo). El nuevo modelo que aquí se propone está basado en el trabajo de Hassan et al., cuyos detalles pueden consultarse en las publicaciones originales (Hassan et al., 2000a; Hassan et al., 2000b; Hassan & Mehler, 2002). Parte de la teoría de líquidos polares de *Lorentz-Debye-Sack* que establece que el efecto atenuación (*screening*) debido al solvente puede representarse a través de una función dieléctrica que depende de forma sigmoidal con la distancia, tal y como se muestra en la siguiente ecuación:

$$D(r) = \frac{\varepsilon + 1}{1 + k\exp[-\lambda(\varepsilon + 1)r]} - 1 \qquad \textbf{[3-8]}$$

dónde r es la distancia entre átomos, ε es la constante dieléctrica del solvente, $k=(\varepsilon-1)/2$ y λ es un parámetro que controla la tasa de cambio de $D(r)$. En el apartado a) de la Figura 3-6 pueden verse ejemplos de la curva para $D(r)$ en función de r y para diferentes valores de λ.

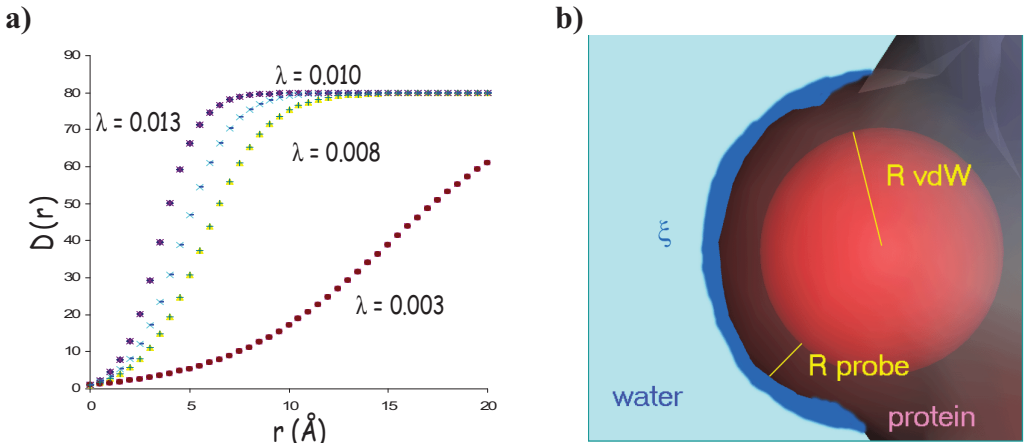


Figura 3-6. a) Función de *screening* sigmoidal para diferentes valores de λ; b) representación del radio de *Born*

El segundo aspecto clave en el modelo es la asunción de que la principal contribución a la desolvatación de un átomo tiene su origen en el desplazamiento de la primera capa de moléculas de agua que rodean a dicho átomo. Esta primera capa se tiene en cuenta considerando el denominado radio de *Born* ($R_{i,Bv(s)}$), expresado únicamente como función de la fracción de superficie expuesta al solvente del átomo ($\xi_i=A_i/4\pi(R_{vdw}+R_{probe})^2$), y representa los procesos de transferencia de un átomo desde el vacío ($R_{i,v}$) hasta el interior de una proteína ($R_{i,p}$), rodeados de solvente ($R_{i,w}$) o vacío

respectivamente (ver apartado b de la Figura 3-6). Así pues, el radio de *Born* se calcula asumiendo la siguiente relación lineal con la superficie expuesta:

$$R_{i,Bv} = R_{i,v}\xi_i + R_{i,p}(1-\xi_i)$$
$$R_{i,Bs} = R_{i,w}\xi_i + R_{i,p}(1-\xi_i) \qquad \textbf{[3-9]}$$

dónde $R_{i,w}=R_{i,cov}+h_{(+,-)}$, $R_{i,p}=R_{i,cov}+g$ y $R_{i,v}=R_{cov}$, siendo R_{cov} el radio covalente. y $h_{(+,-)}$ y g son cantidades positivas que tienen en cuenta el aumento de la cavidad debido a los efectos de la carga. En particular, $h_{(+,-)}$ depende de la carga atómica. Aplicando esta función al proceso de solvatación se obtiene la siguiente ecuación:

$$\Delta G_{elec} = \sum_{i<j}^{N} \frac{q_i q_j}{r_{ij}} \left[\frac{1}{D_S(r_{ij})} - \frac{1}{D_V(r_{ij})} \right]$$
$$+ \frac{1}{2} \sum_{i=1}^{N} q_i^2 \left\{ \frac{1}{R_{i,B_S}} \left[\frac{1}{D_S(R_{i,B_s})} - 1 \right] - \frac{1}{R_{i,B_V}} \left[\frac{1}{D_V(R_{i,B_V})} - 1 \right] \right\} \qquad \textbf{[3-10]}$$

dónde N es el número de cargas, q_i las cargas atómicas, r_{ij} las distancias interatómicas, $D(r)$ la función de *screening* y R los radios de *Born*. El primer término representa las interacciones coulómbicas entre las partículas cargadas atenuada por la función dieléctrica sigmoidal de la Ecuación [3-8]. En dicho término s se refiere al solvente y v al vacío.

Basándose en el ciclo termodinámico mostrado en la Figura 3-7, es posible derivar una adaptación de la Ecuación [3-10] al problema de los complejos proteína-ligando.

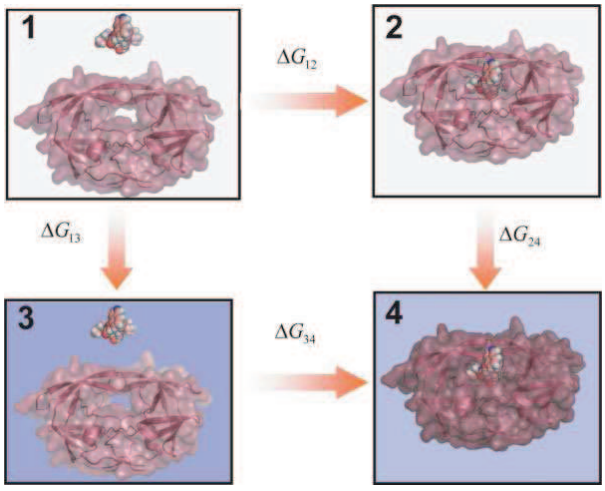


Figura 3-7. Ciclo termodinámico empleado en el cálculo de la ΔG_{elec}.

En la Figura 3-7, las cajas representan los estados de vacío (1 y 2) y solución (3 y 4), tanto de proteína y ligando aislados (1 y 3) como en complejo (2 y 4). De esta manera ΔG_{elec} puede calcularse:

$$\Delta G_{elec} = \Delta G_{34} = \Delta G_{12} + (\Delta G_{24} - \Delta G_{13}) \qquad \textbf{[3-11]}$$

ΔG_{13}, ΔG_{24} y ΔG_{12} se obtienen directamente de la Ecuación [3-10], y asumiendo que la unión proteína-ligando fuera rígida, tras reorganizar los términos, se obtiene:

$$\Delta G_{elec} = \sum_{i=1}^{N_L}\sum_{j=1}^{N_R}\frac{q_i q_j}{D_s(r_{ij})r_{ij}} + \frac{1}{2}\sum_{i=1}^{N_L+N_R} q_i^2\left[\left(\frac{1}{D_S(R_S^C)R_S^C} - \frac{1}{D_S(R_S^U)R_S^U}\right) + \left(\frac{1}{R_S^U} - \frac{1}{R_S^C}\right)\right] \qquad \textbf{[3-12]}$$

dónde N_L y N_R son el número de átomos en ligando y receptor respectivamente, R_S^C y R_S^U son los radios de *Born* para ligando y receptor tanto en complejo como separados. La Ecuación [3-12] representa la base del nuevo modelo de solvente implícito para complejos proteína-ligando. Dicho modelo recibe el nombre de *ISM* (*Implicit Solvent Model*) (Morreale et al., 2007).

3.2.2.1. Modelo de Superficie

Los valores *SASA* necesarios en la Ecuación [3-9] se obtienen a través de la aproximación *LCPO* (ver final del apartado 1.1.1.5 de la página 8) usando una sonda con un radio (r_{probe}) de 1.4 Å. Los cálculos se realizan mediante la Ecuación [1-10] de la página 9. Los parámetros P_1 a P_4 se han derivado a partir del conjunto de datos de prueba de *ISM* (ver apartado 3.1.1.3 de la página 28).

3.2.2.2. Corrección de Enlace Puente de Hidrógeno

Con el objetivo de tener en cuenta los efectos de las interacciones por puentes de hidrógeno sobre la energía de desolvatación del ligando, se introduce un término de corrección en la Ecuación [3-12]. El motivo de esta corrección se explica en el apartado 4.2 de la página 104 (sección de Resultados). Para obtener el término de corrección, se realiza un análisis de regresión entre la diferencia en las energías de desolvatación calculadas con los métodos de *Poisson* e *ISM* y el número y tipo de enlaces por puente de hidrógeno del ligando. A la hora de obtener los puentes de hidrógeno de cada pose, primero se determinan los átomos aceptores y donadores (o ambos) tanto en la proteína como en el ligando. Para la proteína se utilizan los tipos usados en el programa *HBPLUS* (McDonald & Thornton, 1994). En el caso del ligando se consideran

aceptores los nitrógenos neutros y los oxígenos que no sean éteres. Los donadores serían los nitrógenos u oxígenos que tengan al menos un átomo de hidrógeno unido. Para determinar si hay o no puente de hidrógeno, se realizan parejas de aceptores-donadores (uno del ligando y otro de la proteína) y se comprueba si cumplen unos criterios de distancia y ángulo entre el donador, el aceptor y el antecedente del aceptor. El rango de distancia es 2.60-3.40 Å cuando aceptor o donador tienen hibridación sp^2, y 3.10-3.95 Å en el resto de los casos. Para el ángulo el rango es 120-180º. Una vez se tienen los puentes de hidrógeno, estos se clasifican por el tipo de interacción en: cargado-cargado (*cc*), neutro-cargado (*nc*) y neutro-neutro (*nn*). Se observó que también era conveniente tener en cuenta la presencia o ausencia de un grupo amino protonado en los ligandos, por lo que se introdujo una variable adicional para indicarlo (*npn*). Esto se hace para compensar que en la derivación de los parámetros *LCPO* se tiene poca representatividad de grupos amino protonados, encontrándose la mayoría enterrados. Así pues, finalmente el término de corrección queda:

$$\Delta G_{corr} = a + b \cdot hb_{cc} + c \cdot hb_{cn} + d \cdot hb_{nn} + e \cdot npn \qquad \textbf{[3-13]}$$

dónde hb_{xx} es el número de puentes de hidrógeno del tipo *xx*. La energía electrostática de unión final mediante el método ISM quedaría:

$$\Delta G_{ISM} = \Delta G_{elec} + \Delta G_{corr} \qquad \textbf{[3-14]}$$

3.2.2.3. Parametrización del Modelo

Como se ha visto en los apartados previos, el modelo *ISM* hace uso de una serie de parámetros cuyo valor puede verse en la Tabla 3-VII. El parámetro *Escala* sirve para reducir el radio inicial obtenido de *AMBER*. $h_{(+,-)}$ y g tienen en cuenta la ampliación del radio dentro del solvente o la proteína. El primero depende del tipo de carga (0.85 Å para las positivas y 0.35 Å para las negativas). Los parámetros relacionados con el solvente son la pendiente de la función dieléctrico sigmoidal $\lambda_{(+,-)}$ (0.013 para todos los átomos excepto aquellos que tienen una carga formal positiva, en cuyo caso sería 0.007), ε es la constante dieléctrica del solvente y r_{probe} es el radio (en Å) de la molécula de agua empleada como sonda para calcular la superficie accesible al solvente. Hay otra serie de parámetros relacionados con los puentes de hidrógeno: r y α son el mínimo radio (en Å) y ángulo (en grados) entre el donador y el aceptor, y donador-aceptor-antecedente aceptor respectivamente. *a*, *b*, *c*, y *d* se corresponden con los parámetros de ajuste para tener en cuenta la corrección por puentes de hidrógeno. Por último, los

parámetros de la sección *PE/SCP-ISM* son los obtenidos para el ajuste en la comparación de *ISM* y la ecuación de *Poisson*.

Parámetro	Valor
Radio Atómico	
Escala (*Scale*)	0.6
$h_{(+,-)}$	0.85 (0.35)
g	0.5
Solvente	
$\lambda_{(+,-)}$	0.013 (0.007)
ε	78.39
r_{probe}	1.4
Puentes de Hidrógeno	
r	3.4
α	120
a	-0.29
b	1.02
c	-0.25
d	0.02
e	10.52
PE/SCP-ISM	
A	5.3
B	0.09
C	1.06
D	0.97

Tabla 3-VII. Parámetros más relevantes del modelo *ISM*.

Los parámetros de la Tabla 3-VII se han obtenido realizando búsquedas exhaustivas sobre un subconjunto del espacio de valores para los parámetros, utilizando como función de ajuste el error cuadrático entre los resultados de *ISM* y los correspondientes a la ecuación de *Poisson*. Los siguientes parámetros se modificaron sistemáticamente: factor de escala para el radio, de 0.3 a 1.3 Å en intervalos de 0.1 Å; el factor de ampliación $h_{(+,-)}$, de 0.35 a 0.85 Å en intervalos de 0.1 Å; λ, de 0.001 a 0.020 en intervalos de 0.001; y valores fijos para ε (78.39) y el radio de la sonda del solvente (1.4 Å).

3.2.3. Ampliación de CDOCK

3.2.3.1. Simulación de la Flexibilidad del Ligando

Para simular que el ligando es flexible dentro de *CDOCK*, se introducen modificaciones para que éste sea capaz de aceptar un fichero *multi-PDB* que contenga un conjunto de confórmeros para dicho ligando. Este fichero puede generarse mediante el programa *ALFA*. *CDOCK* leerá la lista de confórmeros y tomará tanto la información común a todos ellos (por ejemplo los tipos de átomos) como la información particular de cada uno (coordenadas, centro de masas y cargas atómicas). A partir de este punto se procede de un modo similar a como funciona el nuevo *ALFA* cuando utiliza *MCSA*. La principal diferencia en este caso es la representación de la pose, que se hace a través de

un vector de tres elementos: 1) punto de *grid* donde se coloca el centro de masas del confórmero, 2) identificador de la matriz de rotación usada para generar la orientación, y 3) el identificador del confórmero utilizado. Así pues, los valores que codifican una pose de esta manera son: identificador de confórmero, translación y rotación. Además, con el objetivo de evitar la exploración de poses susceptibles de tener choques estéricos, se eliminan como posibles translaciones todos aquellos puntos de *grid* que estén a una distancia *d* de la proteína. Dicho valor *d* se obtiene calculando, entre todos los confórmeros, la distancia más pequeña desde su centro de masas a cualquiera de las paredes del rectángulo mínimo que contiene al confórmero. El método *MCSA* sustituye a la exploración exhaustiva que se hacía en *CDOCK* con el fin de reducir el tiempo empleado en la exploración del sitio activo, pero manteniendo ésta homogénea. Para promover la variabilidad conformacional de las poses seleccionadas para la minimización por *SIMPLEX*, lo que se hace es mantener una lista de las mejores poses seleccionadas para cada confórmero, y serán las soluciones de estas listas las que pasen a la etapa de minimización. El resultado final será la mejor o mejores soluciones (en términos energéticos) de entre todas las listas.

3.2.3.2. Completando la Función de Puntuación

CDOCK contaba originalmente con una función de puntuación compuesta de dos términos de interacción (*van der Waals* y coulómbico) cuyos valores en la *grid* son pre-calculados a través del programa *CGRID*. Para completar y mejorar su función de puntuación se hace uso de las técnicas empleadas en el modelo *ISM*, de modo que se añade: atenuación sigmoidal para el término coulómbico, desolvataciones para ligando y receptor, enlaces por puente de hidrógeno, y cálculo de la componente no polar. Con el objetivo de acelerar su cálculo en el *docking*, el programa *CGRID* precalcula todo lo necesario para el cálculo de estos términos y aquello que sea independiente del ligando, dejando al programa *CDOCK* la responsabilidad de completar el cálculo para el ligando y pose específicos. A continuación se describen los cuatro nuevos términos.

3.2.3.2.1 Coulómbico Atenuado

A la fórmula del cálculo del potencial coulómbico se le añade un efecto de atenuación (*screening*) debido al solvente. Este efecto se calcula mediante la Ecuación [3-8] usando los valores $\varepsilon=78.39$ y $\lambda=0.013$. Así pues, la nueva fórmula para el cálculo del término coulómbico en *CGRID* para cada punto de *grid* es:

$$coulombico = \sum_{i=1}^{n} \frac{q_i}{r_i \cdot D(r_i)} \qquad \textbf{[3-15]}$$

dónde n es el número de átomos del receptor, q_i es la carga del átomo i, r_i es la distancia del átomo i al punto de *grid*, y $D(r_i)$ la función dieléctrica.

3.2.3.2.2 Desolvataciones

Se emplea el modelo *ISM* (ver apartado 3.2.2 de la página 60) para calcular las energías de desolvatación de receptor y ligando. Su implementación es modular, de tal manera que es posible integrarla entre los programas *CGRID* y *CDOCK*. En el primero se harían todos los cálculos que no dependan del ligando, y en el segundo los restantes para obtener la energía de desolvatación del ligando y pose específicos. El reparto de tareas entre ambos programas queda como se indica en la Tabla 3-VIII.

CGRID	***CDOCK***
Átomos vecinos de cada punto de *grid*	-
Distancias interatómicas del receptor	Distancias interatómicas del ligando
Parámetros específicos del receptor	Parámetros específicos del ligando
Solapamientos entre átomos del receptor	Solapamientos entre átomos del ligando
Superficies totales de cada átomo del receptor	Superficies totales de cada átomo del ligando
SASA de cada átomo del receptor libre	*SASA* de cada átomo del ligando libre
-	Solapamientos entre átomos del ligando y el receptor
-	*SASA* de cada átomo en el complejo

Tabla 3-VIII. Reparto del modelo *ISM* entre los programa *CGRID* y *CDOCK*.

Como puede verse, una de las tareas de *CGRID* consiste en el cálculo de los átomos vecinos de cada punto de *grid*. Dicha tarea es la única que no pertenece propiamente al modelo *ISM*, ya que se ha implementado con el objetivo de acelerar la búsqueda de solapamientos proteína-ligando en *CDOCK*. Esta aceleración consiste en que, para cada átomo del ligando, sólo será necesario buscar átomos solapantes del receptor entre los vecinos del punto de *grid* más próximo al átomo del ligando tratado. Con ello se consigue ahorrar una gran cantidad de tiempo al no tener que probar átomos del receptor que se sabe que están alejados y no van a producir solapamiento.

Una vez que *CDOCK* ha completado sus tareas para la pose de un ligando, es capaz de obtener las desolvataciones llamando directamente a las funciones específicas del cálculo de desolvatación en la implementación del modelo *ISM*.

3.2.3.2.3 Enlaces por Puente de Hidrógeno

Los enlaces por puente de hidrógeno se calculan del mismo modo que se hace en el modelo de *ISM* para realizar la corrección (ver apartado 3.2.2.2 de la página 63). Para ello en *CGRID* se clasifican los átomos del receptor en donadores, aceptores o ambos. En *CDOCK* se hace uso de la lista de vecinos (que también se calculó en *CGRID*) para las desolvataciones. Para cada átomo del ligando (que sea aceptor, donador o ambos) se mira el punto de *grid* más cercano, se obtienen sus vecinos, y para cada uno de ellos se comprueba si es posible formar el punte de hidrógeno. Se considera que cada puente de hidrógeno contribuye con -1 *kcal/mol* a la energía total de interacción. Además, utilizando las funciones implementadas en *ISM* también se realiza la clasificación de cada puente de hidrógeno dependiendo del tipo de átomos involucrados en el enlace (carga-carga, carga-neutro o neutro-neutro) por lo que se puede obtener el término de corrección.

3.2.3.2.4 Componente No Polar

Para cuantificar el efecto hidrofóbico se asume una relación lineal con la superficie accesible al solvente de la molécula como se muestra en la ecuación:

$$\Delta G_{hidrofóbico} = a + b \cdot SASA \qquad \textbf{[3-16]}$$

dónde $a = 0.092$ *kcal/mol* y $b = 0.00542$ *kcal/mol*Å^2. Cuando se quiere calcular la componente no polar para interacciones proteína-ligando lo que se hace es calcular la *SASA* tanto del complejo como de la proteína y el ligando por separado, tomando la diferencia:

$$\Delta G_{nopolar} = \Delta G_{hid-complejo} - \Delta G_{hid-proteína} - \Delta G_{hid-ligando} \qquad \textbf{[3-17]}$$

3.2.4. Interfaz Gráfico de Usuario para COMBINE

gCOMBINE es el nombre del interfaz gráfico de usuario (*GUI*) desarrollado para facilitar el uso del programa *COMBINE* original, el cual funcionaba a través de la línea de comandos. Además de mejorar su uso, también facilita el análisis y la rápida interpretación de los resultados mediante representaciones gráficas. Esta *GUI* ha sido escrita en lenguaje *Java* (v. 1.6.0_10), lo que asegura la portabilidad entre diferentes plataformas. La funcionalidad gráfica y la interactividad se han añadido haciendo uso de las *Java Foundation Classes* (*JFC*) y los componentes *Swing* respectivamente (ver apartado 3.1.2.2.3 de la página 43). Para el desarrollo de la *GUI* se ha utilizado el

entorno integrado de desarrollo (*IDE*) *NetBeans 6.1* (http://netbeans.org/) incluyendo además el *Swing Application Framework* (ver apartado 3.1.2.2.3 de la página 43). La generación de las diferentes gráficas se apoya en las librerías *JFreeChart 1.0.11* y *JCommon 1.0.14* (ver apartado 3.1.2.2.4 de la página 43). Ambas se distribuyen bajo licencia *GNU Lesser General Public License*. Estas librerías permiten a *gCOMBINE* generar gráficas interactivas con los resultados más relevantes facilitando así su manipulación y análisis. Y ya que la *GUI* es independiente de plataforma y el programa *COMBINE* está escrito en *GNU Fortran* estándar, la aplicación completa (*COMBINE + gCOMBINE*) puede ser utilizada en sistemas operativos *Linux*, *Windows* o *Mac* que tengan un compilador *gcc* (versión 3.4.6 o superior, o bien que *COMBINE* haya sido ya compilado en ese sistema) y la maquina virtual *Java* (v. 1.6.0_10 o superior). La estructura interna de *gCOMBINE* tiene un diseño orientado a objetos basado en el patrón Modelo-Vista-Controlador (*MVC*) (más detalles en la página *web*: http://java.sun.com/blueprints/patterns/MVC.html). La clase principal del Modelo es *CombineModel*. Una instancia de esta clase almacena la información de un modelo específico (o una configuración de modelo) generado desde una ejecución de *COMBINE*: nombre del modelo, un comentario descriptivo, la carpeta de trabajo, los parámetros de configuración, los ficheros de salida, las tablas y las gráficas. Los parámetros se almacenan en una instancia de la clase *Parameters*, que usa objetos del tipo *ComplexesListItem* para guardar los diferentes complejos ligando-receptor relacionados con el modelo *COMBINE* bajo estudio. Las tablas y gráficas son paneles generados a través de métodos estáticos de las clases *CombineTables* y *CombineGraphs* respectivamente, tomando una instancia de *CombineModel* como entrada. La Vista es arrancada por la clase *CombineGUIApp* que crea una instancia de la clase *CombineGUIView*. Ésta actúa como almacén para los diferentes objetos gráficos y también funciona como Controlador para las diferentes acciones (incluyendo validaciones internas) que pueden ser realizadas en la interacción con los objetos. Una clase interna (*CombineThread*) se emplea para ejecutar el programa *COMBINE* en un hilo de ejecución diferente al de la *GUI* de modo que se evita el bloqueo de la aplicación gráfica mientras el programa está corriendo. *CombineThread* usa una instancia de la clase *CombineWrapper* para preparar la ejecución de *COMBINE*: lanza el cálculo, controla el proceso (tomando los *logs* de salida con la clase *StreamGlobber*) y carga los resultados tras la finalización de la ejecución. Hay otras tres clases que se usan a lo largo del ciclo de vida de la aplicación: a) *CombineConstants* (contiene

diferentes constantes); b) *CombineException* (para propagar errores y avisos personalizados); y c) *Useful* (para almacenar algunos métodos comunes).

3.2.5. Plataforma de Cribado Virtual

VSDMIP (*Virtual Screening Data Management on an Integrated Platform*) es una plataforma flexible y automática para realizar cribados virtuales que combina todos los pasos necesarios con el fin de generar una lista reducida de candidatos a partir de una base de datos de estructuras 2D de moléculas. Su arquitectura (ver Figura 3-8) consiste en (1) una base de datos sobre un sistema gestor de base de datos relacionales (*MySQL*) multi-hilo y multi-usuario que permite la realización de consultas estructuradas mediante el lenguaje *SQL* (*Structured Query Language*), (2) una librería de interfaces de servicio y *plugins*, y (3) un conjunto de flujos de trabajo y los comandos que los implementan. Todos los datos referentes a las moléculas pequeñas y a los resultados de los cribados se almacenan en la base de datos de *VSDMIP*, llamada *VSDB* (*Virtual Screening Data Base*). El usuario controla la plataforma a través de diferentes utilidades en la línea de comandos, y la configura haciendo uso de ficheros *XML*. *VSDMIP* puede ejecutarse en plataformas *Linux/x86*.

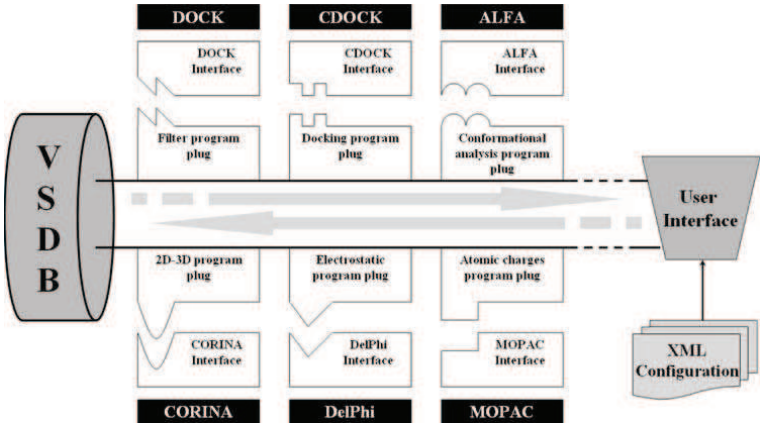


Figura 3-8. Representación de la arquitectura de *VSDMIP*. *VSDB* se refiere a la base de datos relacional (*Virtual Screening Data Base*) donde se almacenan los datos.

3.2.5.1. Base de Datos Relacional

La base de datos (ver Figura 3-9) contiene una serie de tablas para las librerías de compuestos (*ORIGINDB*, *MOLECULE*, *CONFIGURATION* y *CONFORMER*), resultados para filtros de cribado (*DOCK_FILTER* y *RESULT_DF*), y resultados para experimentos de cribado (*VIRTUAL_SCREENING* y *RESULT_VS*). El origen de cada

compuesto se almacena como una entrada en la tabla *ORIGINDB*. Cada molécula tiene una entrada en la tabla *MOLECULE* junto con su carga global y una representación de su topología en formato *SMILES*. Los diferentes estereoisómeros de cada molécula en combinación con las diferentes conformaciones de anillo, junto con la estructura 3D, se almacenan en la tabla *CONFIGURATION*; los cambios discretos generados en las diferentes estructuras 3D a través de la rotación de enlaces sencillos (análisis conformacional) se almacenan en la tabla *CONFORMER*, junto con su energía interna (*VDW_ENERGY*). Los experimentos de filtrado se agrupan mediante una entrada en la tabla *DOCK_FILTER*. La mejor pose obtenida para cada uno de los confórmeros incluidos en el experimento se almacena en la tabla *RESULT_DF*, junto con su puntuación. Estas poses se almacenan como una translación y rotación (mediante ángulos de *Euler*) del confórmero con respecto a las coordenadas originales almacenadas en la base de datos. Los experimentos de *docking* más precisos se agrupan en la tabla *VIRTUAL_SCREENING*, y cada solución de un confórmero individual tiene su entrada en la tabla *RESULT_VS*, con la información sobre translación y rotación al igual que en la tabla *RESULT_DF*. Además, la puntuación del *docking* en el caso de los experimentos más precisos se almacena detallada a través de sus términos de interacción energéticos (*van der Waals*, coulómbico, desolvataciones, componente no polar de la desolvatación) y los enlaces por puente de hidrógeno así como su tipología.

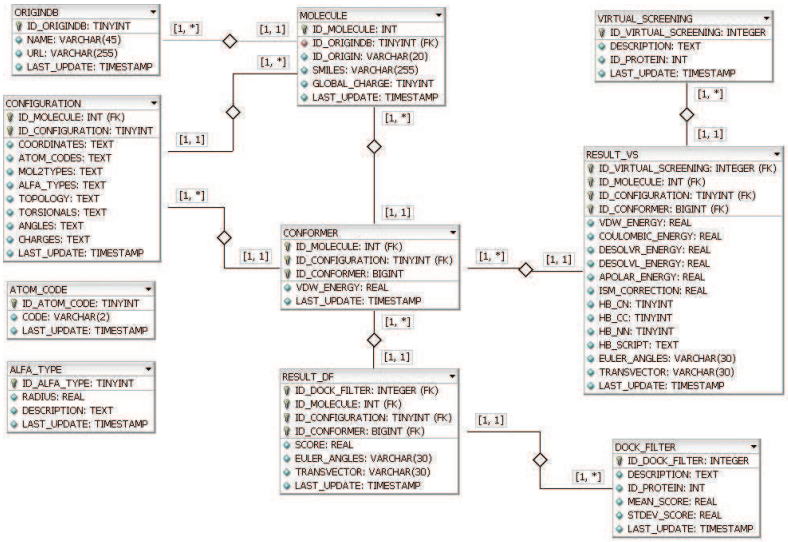


Figura 3-9. Esquema de la base de datos relacional utilizada en *VSDMIP*. Generado con *DBDesigner 4* (http://fabforce.net/dbdesigner4/).

3.2.5.2. Librería de Software de VSDMIP

Se trata de una librería escrita en lenguaje *C/C++* y que se encarga de interaccionar con la base de datos (mediante la librería *MySQL++*, ver apartado 3.1.2.2.6 de la página 44), interconectar las diferentes aplicaciones usadas, y ofrecer utilidades bioquímicas y geométricas. Para cada aplicación externa es posible añadir funcionalidades a la plataforma a través de la creación de una clase interfaz y una clase de almacenamiento de resultados, todo bajo una estructura común. La clase interfaz proporciona métodos para el manejo de los atributos de configuración de las aplicaciones, prepara la ejecución de la aplicación, lleva a cabo la ejecución, y gestiona los errores y el almacenamiento de los resultados en la base de datos. En la versión actual existen seis servicios: generación de estructuras 3D, análisis conformacional, cálculo de cargas atómicas, filtrado, *virtual screening*, y cálculos electrostáticos. Se han desarrollado algunos *plugins* para servir de interfaz con *CORINA 3.0.5* (generación 3D), *ALFA* (análisis conformacional), *MOPAC 7* (cargas atómicas), *DOCK 3.5* y *Fred 2.2* (filtros), *CDOCK* y *Autodock 3.0.5* (*virtual screening*), y *DelPhi 4* e *ISM* (cálculos electrostáticos). En el apartado 3.1.2.1 de la página 34 puede encontrarse más información sobre estos programas.

3.2.5.3. Flujos de Trabajo

VSDMIP contiene comandos para la inserción de moléculas en la base de datos, realizar el *docking* de un conjunto de moléculas en el centro activo de una proteína (ya sea como filtro o como *virtual screening*), *rescoring* de resultados de un experimento de *docking*, y obtención de los resultados de un filtro o un *virtual screening*. Cada comando utiliza uno o varios de los servicios incluidos en *VSDMIP*, y la ejecución se puede realizar en un *cluster*.

3.2.5.4. Paralelización

Se han desarrollado tres aproximaciones al problema de paralelizar las acciones más pesadas computacionalmente llevadas a cabo con *VSDMIP*. En los tres casos, se toma como unidad de cálculo la molécula y las acciones que se pueden realizar con ella. Así, la paralelización consiste básicamente en la división de las unidades de cálculo (moléculas) entre los recursos disponibles (procesadores). A continuación se describen las tres aproximaciones.

3.2.5.4.1 PBS

Para implementar la paralelización mediante *PBS* (ver apartado 3.1.2.2.5 de la página 44), en primer lugar se crean una serie de modos de ejecución en los programas que implementan los flujos de trabajo. Dichos modos de ejecución se describen en la Tabla 3-IX.

Modo de Ejecución	Descripción
0	Ejecución del flujo de trabajo en el procesador actual
1	Preparación y reparto de los trabajos para ejecutarse en procesadores remotos
2	Gestión de la ejecución remota y devolución de los resultados

Tabla 3-IX. Modos de ejecución en los flujos de trabajo de *VSDMIP*.

Así pues, el camino seguido inicialmente por el flujo de trabajo depende del valor del parámetro que identifica el modo de ejecución. Un resumen de las acciones realizadas en cada caso se presenta en la Figura 3-10.

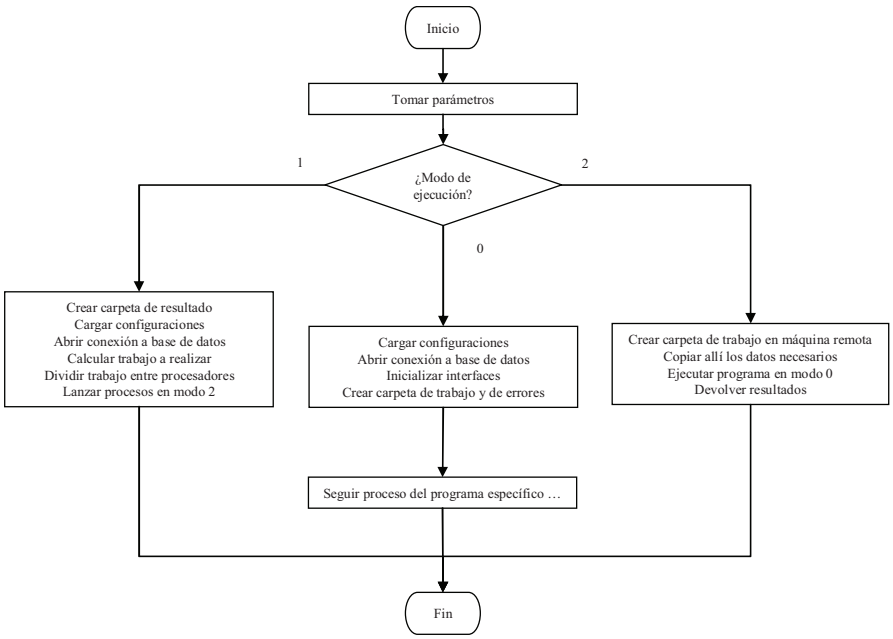


Figura 3-10. Esquema general del comienzo de los flujos de trabajo en *VSDMIP*.

Para lanzar un proceso en modo 2 estando en modo 1 se genera un fichero de tipo *PBS* el cual será enviado al gestor de colas. Dicho fichero contiene el comando de ejecución del programa en modo 0 (ejecución en el procesador actual) para que se ejecute en el procesador remoto.

3.2.5.4.2 GridSuperscalar

Para adaptar *VSDMIP* a *GridSuperscalar* de *MareNostrum* (ver apartados 3.1.2.2.2 y 3.1.3.2 de las página 43 y 46 respectivamente), lo que se hace es aprovechar la división en paquetes de trabajo que ya de por sí hace *VSDMIP* cuando se utiliza *PBS*. Cada uno de estos paquetes será tratado por un *worker* del *GSS*, el cual ejecutará el comando extraído del fichero *.pbs* que se generó. Además, también hay que tener en cuenta que en *MareNostrum* no se tiene una máquina que esté continuamente dedicada a ser servidor de base de datos, por lo que se necesita que uno de los procesadores haga esta función. Por ello, se hace que el nodo *master* arranque una instancia de la base de datos. Todas estas adaptaciones se hacen a través de un programa en lenguaje *Java* que se encarga de preparar y lanzar los *workers* (cada uno con su paquete de trabajo). Este programa, llamado *RunVSDBonGSS*, se ejecuta desde un *script* para el sistema de colas de *MareNostrum*. A continuación se muestra el algoritmo de las principales tareas de este programa:

```
arrancar base de datos
arrancar hilo de parada de base de datos
componer nombre de fichero .pbs y fichero de configuración .xml
establecer nodo master como base de datos en el .xml
ejecutar comando que genera las tareas (modo 1 en PBS)
para cada tarea
     tomar comando desde su fichero .pbs
     crear script que lanza el comando
     ejecutar script como un worker de GSS
fin para
parar base de datos (si no ha sido parada por el hilo)
```

Algoritmo 3-5. Ejecución de *VSDMIP* sobre *GridSuperscalar*.

Como puede verse en la segunda línea, también se arranca un nuevo hilo de ejecución que se encarga de parar la base de datos. Este hilo lleva un contador de tiempo, y permanece suspendido hasta que este contador expira, momento en el que se activa y para la base de datos. El motivo de hacer ésto es que en el sistema de colas de *MareNostrum* el tiempo que puede estar corriendo un trabajo es limitado, y al finalizar este tiempo el trabajo es detenido abruptamente. Esto podría hacer que la base de datos no se cerrarse correctamente y que se provocara una corrupción en los datos.

3.2.5.4.3 Ibercivis

El proyecto Ibercivis es una arquitectura de cálculo voluntario sobre *BOINC* que comparte diversas aplicaciones (ver apartados 3.1.2.2.1 y 3.1.3.3 de las páginas 43 y 46 respectivamente). En este caso, sólo se ha adaptado la parte del *docking* ya que es la que más tiempo consume. Para hacerlo, se ha generado un conjunto de ficheros (uno para

cada molécula en *VSDMIP*) y dicho conjunto se ha colocado en los servidores de Ibercivis. El programa de *docking*, *CDOCK*, se ha compilado en diferentes arquitecturas (*Windows*, *Linux* y *Mac*) para que la plataforma de Ibercivis pueda enviar a cada voluntario su ejecutable apropiado. El resumen de cada resultado es almacenado en un fichero en formato tabulado en el servidor de Ibercivis. Dicho fichero es descargado periódicamente en la Unidad de Bionformática y a través de un script en *Perl* se inserta en la base de datos de *VSDMIP*.

3.2.6. Protocolo de Cribado Virtual

Un protocolo de cribado virtual se compone de diferentes etapas sucesivas y de un modo de preparación de los elementos que intervienen en ellas (receptor y ligandos). Normalmente se intenta que los datos que se preparen puedan ser utilizados en el mayor número de circunstancias (por ejemplo diferentes *dockings* o diferentes cribados). Tal es el caso de los ligandos, que una vez preparados pueden utilizarse en diferentes cribados, o de la proteína, que si se prepara con independencia del ligando puede reutilizarse en diferentes *dockings*. En cambio, con la idea de la división en etapas lo que se pretende básicamente es encontrar el equilibrio entre eficacia necesaria y los recursos disponibles, de tal modo que se maximice la eficacia del cribado dentro de los límites impuestos por los recursos (tiempo, procesadores...). En base a esto, se ha desarrollado un protocolo de cribado virtual cuyo esquema general se muestra en la Figura 3-11.

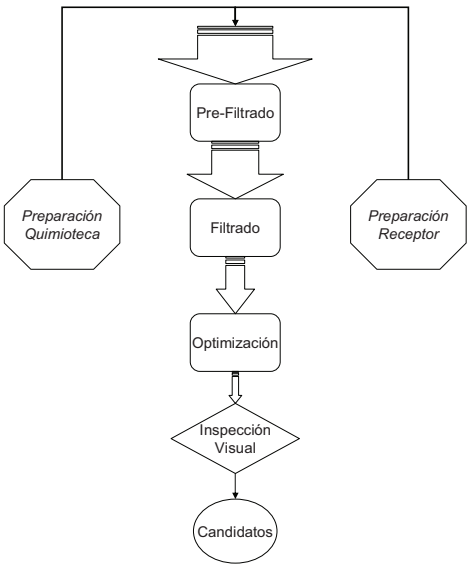


Figura 3-11. Esquema general del protocolo de cribado virtual.

A continuación se describe la implementación habitual que se hace de este protocolo, aunque ésta puede modificarse según las necesidades.

3.2.6.1. Preparación de la Quimioteca

El objetivo es obtener las diferentes estructuras tridimensionales que puede adoptar cada ligando, los parámetros necesarios de sus átomos (cargas, radios...) y cualquier otra información relevante para etapas sucesivas. Dicha información se almacenará en una base de datos relacional con la idea de poder utilizarla en diferentes cribados. En la implementación realizada en este trabajo se parte de ligandos en formato *SMILES*, por lo que primero es necesario obtener su estructura tridimensional. Para realizar esta conversión se emplea el programa *CORINA* de modo que considere hasta 6 centros quirales, genere diferentes conformaciones de anillos y añada los átomos de hidrógeno. Para cada estructura generada por *CORINA* se calculan las cargas atómicas de tipo *ESP* (Bayly et al., 1993) con el programa *MOPAC* usando el método *MNDO* (*Modified Neglect of Differential Overlap*) (Dewar & Thiel, 1977). Por último, con el programa *ALFA* se realiza el análisis conformacional, permitiendo la asignación automática de tipos de átomos y radios (basados en *AMBER*), detección de enlaces rotables, asignación de posibles estados rotaméricos, y la generación de confórmeros junto con el cálculo de su energía interna. De todos los confórmeros generados, suelen seleccionarse los 200 mejores (en función de su energía interna) para las etapas sucesivas del cribado.

3.2.6.2. Preparación del Receptor

Se parte de la estructura tridimensional de una proteína, ya sea obtenida a través de rayos-X, RMN o mediante modelado por homología. En primer lugar, será necesario seleccionar sólo aquella parte de la proteína que interesa y puede ejercer influencia sobre el centro activo tratado. Tras ello se añaden los átomos de hidrógeno, ya que las estructuras de rayos-X (las más habituales y de mejor resolución) solo tienen los átomos pesados. Para añadirlos se utiliza un método sencillo que contempla únicamente estados de protonación estándar de los grupos titrables, como es el programa *protonate* de *AMBER*, o bien uno más refinado como el servidor de *H++* que también utiliza módulos de *AMBER* pero tiene en cuenta los diferentes estados de protonación de los grupos titrables. A continuación se añaden los parámetros de cargas y radios atómicos, también basándose en el campo de fuerzas de *AMBER*. Como último paso se caracteriza

energéticamente el sitio activo mediante los programas *CGRID*, *CDOCK* y *GAGA*. Con el primero se obtienen las *grids* de interacción con un espaciado de 0.5 ó 0.375 Å según el grado de precisión requerido. Estas *grids* las utilizará *CDOCK* para realizar los *dockings*. En este paso de preparación se utiliza el programa *CDOCK* para generar mapas de potenciales de interacción molecular para tres sondas: 1) benceno (localización de áreas hidrofóbicas); 2) metanol (áreas susceptibles de formar enlaces por puente de hidrógeno); y 3) agua (localización de zonas hidrofílicas). El programa *GAGA* comprime esta información y la transforma en funciones gaussianas con el objetivo de capturar las zonas de mayor interés para cada tipo de interacción. El resultado de este cálculo es lo que se conoce como una imagen negativa del centro activo.

3.2.6.3. Pre-Filtrado

En esta etapa es posible realizar un filtro o una combinación de éstos. Los filtros pueden ser de muchos tipos: propiedades físico-químicas, *fingerprints* de interacción, similitud con moléculas activas, complementariedad de forma... En la implementación realizada aquí el filtro está basado en complementariedad de forma. Esto se realiza mediante el programa *DOCK* y haciendo uso de las funciones gaussianas generadas por *GAGA*. Cada una de estas funciones puede representarse como una esfera en el sitio activo, y el programa *DOCK* se encarga de encontrar, para una estructura 3D de un ligando dado, la posición que mejor ajusta sus centros atómicos a los centros de las esferas. En base a este ajuste se asigna una puntuación ($score_i$) mediante la función de contactos de *DOCK*. Estos valores son convertidos a *ZScore* usando la media ($\overline{score}$) y la desviación estándar (σ):

$$ZScore = (score_i - \overline{score}) / \sigma \qquad \textbf{[3-18]}$$

Habitualmente, sólo aquellos ligandos cuyo *ZScore* sea superior a un valor de corte (normalmente 5.0) son seleccionados para pasar a la siguiente etapa.

3.2.6.4. Filtrado

En este paso se suele utilizar un algoritmo de *docking* más preciso. En el caso de esta implementación se utiliza el programa *CDOCK* que hace uso de las *grids* calculadas con *CGRID* en la caracterización del sitio activo. Tras esta etapa ya se obtienen las energías de interacción y unas poses más refinadas.

3.2.6.5. Optimización

Aquí se consideran tanto los procesos para dotar de flexibilidad completa al sistema (dinámica molecular), como los procesos para el análisis de las relaciones cuantitativas estructura-actividad tridimensionales (*QSAR-3D*, con *COMBINE* por ejemplo). En la implementación de este protocolo se realizan dinámicas moleculares con *AMBER* de los mejores resultados de la etapa anterior. Usualmente el tiempo de la dinámica es de 1 ns. Todas las simulaciones se realizan a presión y temperatura constante (1 atm y 300 K) con un paso de integración de 2 fs. Se utiliza el algoritmo *SHAKE* (Ryckaert et al., 1977) para fijar en su distancia de equilibrio a todos los enlaces que impliquen átomos de hidrógeno. Los métodos de condiciones periódicas de contorno (*Periodic Boundary Condition*) y *Particle Mesh Ewald* se utilizan para dotar de periodicidad al sistema y tratar los efectos electrostáticos de largo alcance (Darden et al., 1993). Los campos de fuerza que habitualmente se utilizan son *AMBER-99* (Cornell et al., 1995) y *TIP3P* (Jorgensen et al., 1983). Los complejos son inicialmente 1) hidratados usando cajas con moléculas de agua explícitas *TIP3P*; 2) optimizados; 3) calentados (20 ps); y 4) equilibrados (100 ps). Tras el equilibrado, las trayectorias de dinámica molecular se continúan durante 1 ns. Utilizando la aproximación *MM-GBSA* (Massova & Kollman, 2000) se estiman las energías libres de unión para cada ligando. El método *MM-GBSA* aproxima la energía libre de unión mediante la suma de un término de interacción de mecánica molecular (*MM*), la contribución a la solvatación a través del modelo generalizado de *Born* (*GB*) (Still et al., 1990), y la contribución del área de superficie (*SA*) para tener en cuenta la parte no polar de la desolvatación. Estos cálculos se realizan para una serie de estructuras moleculares a lo largo de la dinámica (*snapshots*) usando el módulo apropiado de *AMBER* y calculando su media. Finalmente, de todo este proceso se obtendrá la lista priorizada de compuestos para ser examinada en la siguiente etapa.

3.2.6.6. Inspección Visual

Para seleccionar los ligandos candidatos a comprar y probar experimentalmente, se realiza una inspección visual de los mejores resultados tras la etapa de optimización. Esta etapa suele consistir en la visualización de la estructura promedio de la dinámica o de la trayectoria completa (con los programas *PyMOL* y *VMD* respectivamente), y el

estudio de energía total de interacción calculada con *MM-GBSA* así como el de las diferentes contribuciones energéticas por residuo.

3.2.7. Herramientas de Validación

Normalmente, cuando se realizan pruebas de software lo que se hace es buscar errores en su ejecución, es decir, que no se obtengan los resultados esperados. Pero el caso del software científico tiene la complicación adicional de que en la mayoría de las ocasiones se desconoce cuáles deberían de ser esos resultados. Pero al fin y al cabo, lo que se trata es de simular procesos biológicos, por lo que bastaría con encontrar un conjunto de pruebas para las que se conozca este resultado experimental. Este es el modo más habitual de probar las técnicas involucradas en *docking* y cribado virtual, realizando pruebas retrospectivas. Así que, para demostrar la validez de *ALFA* y *CDOCK*, se comparan las estructuras de ligando y complejos generados con estructuras cristalográficas conocidas; para el cribado virtual se toman conjuntos de ligandos para los que se sabe cuales de ellos son activos. Pero en otro tipo de pruebas, como en los casos de *ISM* y *gCOMBINE*, lo que se busca es compararlo con otro método, una referencia, que en este caso son *DelPhi* y *COMBINE* respectivamente; se desea determinar si son válidos como una mejor alternativa, es decir, que además de obtener datos similares, también aportasen una mayor velocidad de cálculo o bien facilitasen su uso. Finalmente, también se realizan aplicaciones en casos reales: se hacen cribados en los que las moléculas finalmente seleccionadas serán compradas y probadas experimentalmente. En los siguiente apartados se describen algunas de las medidas de validación empleadas en las pruebas retrospectivas y comparativas, así como una breve descripción de los tests que se van a realizar.

3.2.7.1. Medidas de Validación

3.2.7.1.1 RMSD (Root Mean Square Deviation)

Es un parámetro que sirve para evaluar cómo de parecido es un resultado de una solución de *docking* (o de un análisis conformacional) con la que se conoce experimentalmente para un ligando dado. Dicha comparación se hace en términos geométricos. Consiste en calcular la desviación cuadrática media entre las coordenadas de cada par de átomos pesados homónimos para la pose experimental y la teórica. En la Ecuación [3-19] se muestra cómo calcularlo, siendo N el número de átomos pesados y δ la desviación entre ellos.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\delta_i^2} \qquad \textbf{[3-19]}$$

Normalmente se considera que un resultado de un *RMSD* menor de 2 Å es bueno para *docking* ya que reproduce bastante bien la estructura experimental. En el caso de un análisis conformacional, se consideran buenos los valores por debajo de 1 Å. En este último caso es necesario superponer antes las estructuras (McLachlan, 1979).

3.2.7.1.2 Factores de Enriquecimiento

Se utilizan para evaluar cómo de bueno es un cribado a la hora de priorizar los ligandos activos sobre los inactivos. Una vez se tienen los resultados para todos los ligandos ordenados por su afinidad predicha, de lo que se trata es de calcular para un porcentaje de la lista dado (*subset*) cuántas veces la fracción de activos del *subset* está por encima de los esperados al azar. El modo de calcularlo es:

$$EF_{subset} = \frac{\{NAc_{subset} / NT_{subset}\}}{\{NAc_{total} / NT\}} \qquad \textbf{[3-20]}$$

dónde *NAc* es el número de activos y *NT* es el número total de ligandos. La variación de los factores de enriquecimiento suele representarse en una gráfica en la que en el eje *X* se muestra el porcentaje de lista ordenada final escaneada, y en el eje *Y* el factor de enriquecimiento encontrado hasta ese momento. En la Figura 3-12 se muestra un ejemplo de este tipo de gráficas.

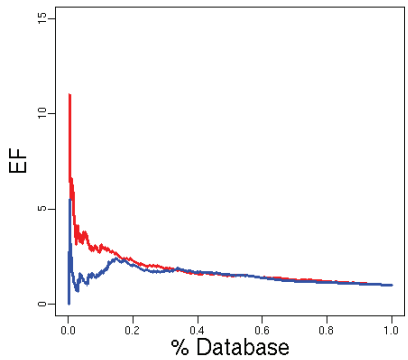


Figura 3-12. Ejemplo de gráfica de factores de enriquecimiento.

3.2.7.1.3 Curvas ROC (Receiver Operating Characteristic)

Se trata de una técnica (Triballeau et al., 2005) basada en la teoría de decisión de *Neyman* y *Pearson* (Neyman & Pearson, 1933a; Neyman & Pearson, 1933b). Sirve para

evaluar de un modo objetivo la habilidad de cierto test a la hora de distinguir entre dos poblaciones (compuestos activos e inactivos en este caso). En la Figura 3-13 puede verse cuál sería el proceso a la hora de realizar una curva *ROC*. En el apartado a) se muestra una representación del conjunto de moléculas activas y el de las inactivas suponiendo para ambos una distribución normal. Se puede apreciar que los compuestos con mayor puntuación (*Score*) tienden a ser los activos, y los de menor puntuación los inactivos. Es necesario fijar un límite de puntuación (*Threshold*) a partir del cual se considerará que los compuestos con *Score* mayor a ese límite son activos. Pero resulta inevitable tener compuestos inactivos que superen ese límite y a la vez compuestos activos que no lo alcanzan; son los llamados falsos positivos y falsos negativos respectivamente. Estos pueden representarse como se muestra en el apartado b), en una matriz, llamada de confusión, donde también se muestran los verdaderos positivos y los verdaderos negativos. A partir de estos datos se puede calcular la sensibilidad a la hora de detectar activos (apartado d). Del mismo modo, también puede calcularse la especificidad (apartado d), que indica cómo de bueno es el método a la hora de descartar los compuestos inactivos. Normalmente, a mayor sensibilidad menor especificidad y viceversa, por lo que habrá que tenerlo en cuenta a la hora de seleccionar cuál es la puntuación mínima para los activos según las necesidades que se tengan (probar pocos compuestos pero buenos, encontrar el mayor número de activos, recursos económicos limitados...).

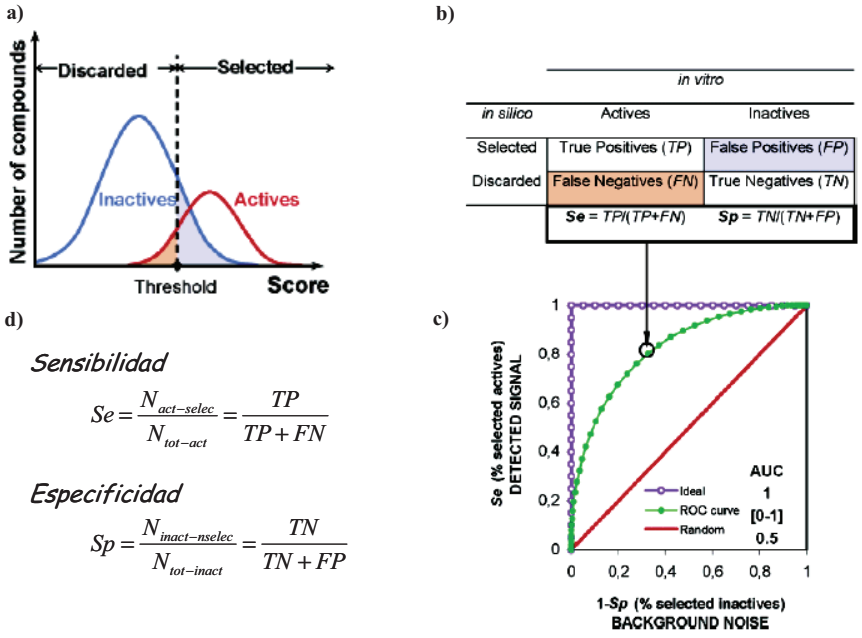


Figura 3-13. Esquema general del proceso de generación de una curva *ROC*. Tomado de (Triballeau et al., 2005)

Finalmente, en el apartado c) de la Figura 3-13 se muestra la curva *ROC* generada a partir de los datos calculados anteriormente. En verde se muestra la curva obtenida, en morado se muestra la curva ideal, es decir, aquella en la que los compuestos activos están los primeros en la clasificación; en rojo se muestra la curva teórica que se obtendría si los compuestos se ordenaran aleatoriamente. También se muestra el área bajo la curva (*AUC – Area Under Curve*), que sirve para ver numéricamente los resultados. Este valor sería 1.0 para el caso ideal y 0.5 para el aleatorio. De este modo se tiene la ventaja de poder utilizar esta representación para comparar cómo de bueno es el método a la hora de seleccionar activos y descartar inactivos en relación con otros métodos. También se tiene la ventaja de que ayuda en la selección de cuál será el límite de puntuación para considerar compuestos activos en función de las pruebas realizadas y las necesidades que se tengan, de modo que será este valor el que se utilizará en los cribados reales a la hora de seleccionar compuestos que se suponen activos para probarlos experimentalmente.

3.2.7.2. Pruebas de Generación de Conformación Bioactiva

En un primer caso se utilizan 36 ligandos seleccionados de la publicación de Boström et al. (ver apartado 3.1.1.1.1 de la página 27). No se usa el algoritmo *MCSA* para el muestreo del espacio conformacional de los ligandos con muchos enlaces rotables de modo que pueda apreciarse el efecto que sobre el tiempo de cálculo tienen el número de torsionales. Los 36 ligandos cumplen las siguientes propiedades: a) alta resolución de las estructuras de rayos-X (< 2 Å); b) bajos factores de temperatura (*B-factor* < 30 $Å^2$) para sus átomos; c) sin enlaces rotables no detectables por cristalografía; d) razonablemente pequeños, flexibles y con número de enlaces rotables entre 1 y 11.

Para prepararlos se obtienen del *Protein Data Bank*, se extraen de la proteína con la que se encontraban cristalizados y se les añaden los átomos de hidrógeno. En cuanto a la evaluación de la aplicación de *ALFA*, se calculan los siguientes datos: *RMSD*, porcentaje de confórmeros en conformación bioactiva (aquellos que tienen un *RMSD* por debajo de 1 Å con respecto a la estructura cristalográfica) y tiempos de cálculo (*CPU* que se consume en el proceso completo).

En un segundo caso se emplea un conjunto de 30 ligandos seleccionados de la publicación de Good et al. (ver apartado 3.1.1.1.2 de la página 27). Dichos ligandos presentan diferente grado de flexibilidad (baja: 3-5 enlaces rotables; media: 6-8; alta: 9-14) y en este caso sí que se emplea el algoritmo *MCSA* para los ligandos de mayor

flexibilidad. Además se compararán los resultados obtenidos por *ALFA* con los obtenidos mediante otra serie de programas con los que puede hacerse el análisis conformacional: *DOCK*, *CONFIRM*, *CONFORT* y *OMEGA*. También se comparará con el resultado de hacer una simple minimización de la energía sobre la conformación de rayos-X. En cada programa se selecciona como resultado un determinado número de confórmeros. En *ALFA* son 200; en *OMEGA* son 100; en *CONFORT* son 500, 10 y 5; para *DOCK* 10, 5 y 3; para *CONFIRM* se tienen los modos de configuración *BEST* y *FAST* seleccionando en cada uno 100, 10 y 5 confórmeros; con *CONCORD* se obtiene una única conformación 3D; y finalmente la minimización consiste en tomar la estructura del ligando en rayos-X y minimizarla con el campo de fuerza de *Tripos*. El método de comparación vuelve a ser el mismo que en el primer caso: medir la habilidad para reproducir la conformación bioactiva de un ligando que se tiene en un fichero *PDB* determinado experimentalmente, siendo la medida de similitud de nuevo el *RMSD*.

Por último, se realiza una prueba más completa con el conjunto de 85 ligandos de *Astex Therapeutics* (ver apartado 3.1.1.1.3 de la página 27). En este caso cada uno de los 85 ligandos es convertido a formato *SMILES*. Partiendo de esta representación, se generan diferentes estructuras 3D con el programa *CORINA*; el motivo de que pueda generar varias estructuras es porque tiene en cuenta los centros quirales y las posibles configuraciones de anillos. Partiendo de moléculas *SMILES* se evita la influencia de la estructura original en los resultados, ya que ésta no se utiliza en ninguna etapa de la generación de confórmeros. Así pues, para cada una de las estructuras generadas con *CORINA* se utiliza el programa *ALFA* para generar los confórmeros, y finalmente, del total de confórmeros generados, se obtienen los 200 mejores en energía. En teoría estos serían los que se pasarían al programa de *docking* (*CDOCK* en el protocolo habitual), así que el objetivo es que al menos uno de ellos tenga una estructura similar al confórmero cristalográfico (*RMSD* por debajo de 1 Å).

3.2.7.3. Comparación del Nuevo Modelo de Solvente Implícito

Ya que la utilidad de la ecuación de *Poisson* a la hora de calcular los efectos electrostáticos en interacciones intermoleculares está bien contrastada, las pruebas se llevarán a cabo comparando cómo el modelo *ISM* es capaz de aproximar la energía libre de unión electrostática a la calculada haciendo uso de la ecuación de *Poisson*. Para ello todos los cálculos se realizan mediante la resolución numérica de dicha ecuación usando el método de las diferencias finitas implementado en *DelPhi*. Para la preparación de las

proteínas se asignan radios atómicos y cargas basándose en el campo de fuerzas de *AMBER*. Para los ligandos las cargas se calculan con *MOPAC*. Finalmente, se comparan tanto la energía libre total de unión electrostática como cada una de sus componentes obtenidas con *ISM* y con *DelPhi* para un conjunto de 826 complejos proteína ligando (ver apartado 3.1.1.3 de la página 28). Además, para probar la robustez del modelo, se utiliza el procedimiento de validación cruzada *Leave One Out* (*LOO*). En dicho procedimiento, uno a uno se eliminan del conjunto de entrenamiento todas las poses para un receptor dado; el modelo se construye en base al conjunto restante y los complejos eliminados son evaluados con el nuevo modelo generado. En cada paso se calcula el *RMSD* y los coeficientes de regresión entre los valores obtenidos por *ISM* para este conjunto y los valores obtenidos mediante *Poisson*.

3.2.7.4. Reproducción de Poses Cristalográficas en Docking

Se realizan varias pruebas empleando el conjunto de 85 ligandos de *Astex Therapeutics* (ver apartado 3.1.1.1.3 de la página 27): *docking* rígido, *docking* flexible y convergencia del algoritmo *MCSA* en el *docking*. El *docking* rígido consiste en comprobar si el programa de *docking* es capaz de reproducir la pose cristalográfica si únicamente dispone del confórmero correcto del ligando (en este caso el cristalográfico), y el *docking* flexible consiste en hacer esta comprobación pero utilizando para cada ligando el conjunto de sus confórmeros generados por *ALFA*. En ambos casos se calcula el valor del *RMSD* entre el resultado del *docking* y la pose del ligando en el complejo cristalográfico. En cada prueba se realizan 20 ejecuciones y se toma como resultado el promedio. Esto se hace para evitar la posible influencia del uso de números aleatorios en el algoritmo de *docking*. En cuanto a la prueba de convergencia del algoritmo *MCSA*, se trata de verificar que es capaz de alcanzar el mismo mínimo energético que el método exhaustivo.

3.2.7.5. Comparación de COMBINE frente a gCOMBINE

El principal objetivo de estos tests es validar que los resultados obtenidos por *COMBINE* (ver apartado 3.1.2.1.8 de la página 38) a través de su implementación gráfica son los mismos que se obtenían en las publicaciones donde se han realizado con éxito análisis *COMBINE*. En concreto se han seleccionado dos de ellas (ver Materiales y Métodos, apartado 3.1.1.4 de la página 29).

3.2.7.6. Discriminación entre Compuestos Activos e Inactivos en Cribados Virtuales

El método de evaluación en este caso consiste en determinar la capacidad de la plataforma de cribado virtual a la hora de distinguir ligandos activos e inactivos para un receptor dado. Para ello, lo primero que se necesita es completar los conjuntos de ligandos activos con un conjunto de inactivos de características similares: peso molecular, número de aceptores de enlaces por puente de hidrógeno, número de donadores de enlaces por puente de hidrógeno, número de enlaces rotables y *logP* (coeficiente de partición octanol/agua). La selección de ligandos inactivos para cada conjunto fue realizada por Jorge Estrada, del *Protein Folding and Stability and Molecular Design Group* dirigido por el Dr. Javier Sancho en la Universidad de Zaragoza. Para ello se desarrolló un programa que, utilizando un algoritmo basado en *MCSA*, realiza la selección de un conjunto de ligandos inactivos con propiedades físico-químicas (calculadas con la herramienta *Filter*) similares a las del conjunto de los activos. Con estos conjuntos se realizan 11 protocolos de cribados diferentes para cada uno de ellos. Estos protocolos se diferencian en la configuración usada o el modo de abordarlos. La Tabla 3-X resume las características de los 11 tipos de cribados. Los valores de *ZScore* usados son muy bajos (3.0 y 1.5) ya que al tener pocos datos realmente ninguno puede diferenciarse mucho del resto.

ID	DESCRIPCIÓN
1	Filtro con *DOCK* (función de contactos) de los 100 mejores confórmeros de cada una de las moléculas, tomando finalmente sólo el mejor resultado para cada una
2	Igual que el caso 1 pero utilizando la función de campo de fuerza de *DOCK*
3	Igual que el caso 1 pero reordenando el resultado utilizando la función de tanteo *X-Score* (Wang et al., 2002)
4	*Docking* con *CDOCK* de los 50 mejores confórmeros de cada molécula ordenando los resultados en función de la suma de las energías de *van der Waals* y coulómbica
5	Igual que el caso 4 pero utilizando sólo la energía de *van der Waals* para la reordenación
6	Igual que el caso 4, pero tras ello, para los 500 mejores resultados se calculan las energías coulómbicas, de desolvatación del receptor y de desolvatación del ligando con *DelPhi*, y la componente no polar con *Apolar.* La reordenación final se hace a partir de la suma de las energías de *van der Waals*, coulómbica (de *DelPhi*), desolvataciones y componente no polar
7	Filtro con *DOCK* (función de contactos) de los 100 mejores confórmeros de cada una de las moléculas, tomando finalmente sólo el mejor resultado para cada una. De estos resultados se seleccionan sólo aquellos que tengan un *ZScore* de puntuación mayor o igual a 3.0 para realizar *docking* con *CDOCK*. La ordenación final de estos resultados se hace por la suma de las energías de *van der Waals* y coulómbica
8	Filtro con *DOCK* (función de contactos) de los 100 mejores confórmeros de cada una de las moléculas, tomando finalmente los 10 mejores resultados para cada una. De estos resultados se seleccionan sólo aquellos que tengan un *ZScore* de puntuación mayor o igual a 3.0 para realizar *docking* con *CDOCK*. La ordenación final de estos resultados se hace por la suma de las energías de *van der Waals* y coulómbica
9	Igual que el caso 7 pero usando un ZScore de 1.5
10	Igual que el caso 8 pero usando un ZScore de 1.5
11	Igual que el caso 7 pero tras ello para los 500 mejores resultados se calculan las energías coulómbicas, de desolvatación del receptor y de desolvatación del ligando con *DelPhi*, y la componente no polar con *Apolar.* La reordenación final se hace a partir de la suma de las energías de *van der Waals*, coulómbica (de *DelPhi*), desolvataciones y componente no polar

Tabla 3-X. Diferentes configuraciones de cribado virtual para las pruebas de la plataforma.

Con el objetivo de evaluar y comparar los resultados de las pruebas, para cada caso se genera una curva *ROC* y se calcula el área bajo esta curva (*AUC*). Además, también se calculan los tiempos empleados y los factores de enriquecimiento para cada caso. Todo ello mediante programas que utilizan el paquete estadístico *R* (http://www.r-project.org/) y que también fueron desarrollados por Jorge Estrada.

3.2.7.7. Aplicación en Casos Reales

A continuación se describen los métodos empleados en cada uno de los casos de estudio desarrollados. Todos ellos hacen uso de la plataforma *VSDMIP* y sus utilidades asociadas.

3.2.7.7.1 MGMT

Se ha utilizado como receptor la cadena A de la estructura cristalográfica con código *PDB 1t39*, y se ha preparado según el protocolo estándar. El sitio activo se ha definido alrededor del ligando cocristalizado (*E1X*), añadiendo 5 Å a sus dimensiones máximas. El espaciado de la *grid* construida es de 0.5 Å. Se ha seleccionado esta zona ya que es donde se realiza el reconocimiento y reparación de las guaninas metiladas (el residuo *Cys145* es el que recibe la lesión alquílica). Como ejemplo, en la Figura 3-14 se muestra una imagen con la proteína, la caja que delimita el sitio activo, el residuo *Cys145* en esferas, y en azul el isocontorno a -1.5 *kcal/mol* de la *grid* para los átomos de carbono.

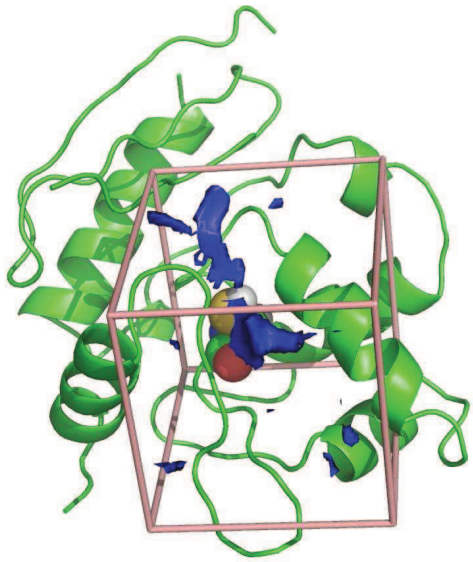


Figura 3-14. Delimitación del sitio activo de la proteína *MGMT* y la *grid* de carbono con un isocontorno de -1.5 *kcal/mol*. En esferas se representa el residuo *Cys145*.

En la Figura 3-15 se muestra una representación de la superficie de la proteína junto con las esferas de las funciones gaussianas generadas con *GAGA*.

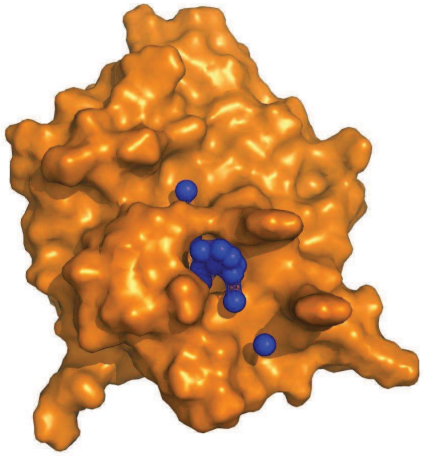


Figura 3-15. Esferas seleccionadas de *GAGA* para el pre-filtrado en el cribado de *MGMT*.

Una vez se tiene preparado el sitio activo se continua con la etapa de pre-filtrado del protocolo estándar de cribado virtual. Las moléculas que superan esta etapa son filtradas con *CDOCK* utilizando sus funciones energéticas de *van der Waals* y coulómbico, para posteriormente realizar un corrección de los términos energéticos electrostáticos (añadiendo desolvataciones) usando *DelPhi*, y de la parte no electrostática de la solvatación utilizando el programa *Apolar*. En este caso, la selección de los ligandos mediante inspección visual se hace tras esta etapa, llevando a cabo la etapa de optimización con dinámica molecular sólo para los ligandos que finalmente resultan activos experimentalmente con el objetivo de estudiar sus modos de unión de una manera más exacta. Los ensayos experimentales consistirán en test tanto *in vivo* como *in vitro*, y serán llevados a cabo por los colaboradores en este proyecto.

3.2.7.7.2 Ape1

Se eliminan de la estructura cristalográfica todos los átomos no pertenecientes a la proteína, excepto el metal divalente (plomo) en el sitio activo. Para simular las condiciones in vivo, dicho metal es sustituido por magnesio, que es el cofactor metálico preferido de *Ape1* (Barzilay et al., 1995). La proteína se ha preparado en el modo estándar, empleando el servidor *H++* (en sus valores por defecto) para añadir los átomos de hidrógeno. Para delimitar el centro activo se utiliza un fragmento de ADN del complejo *Ape1*-ADN (código *PDB 1dew*) como referencia para generar una *grid* de

5 Å alrededor de dicho fragmento. El espaciado de *grid* utilizado es de 0.375 Å. La Figura 3-16 muestra las esferas seleccionadas para realizar el filtro pre-filtrado.

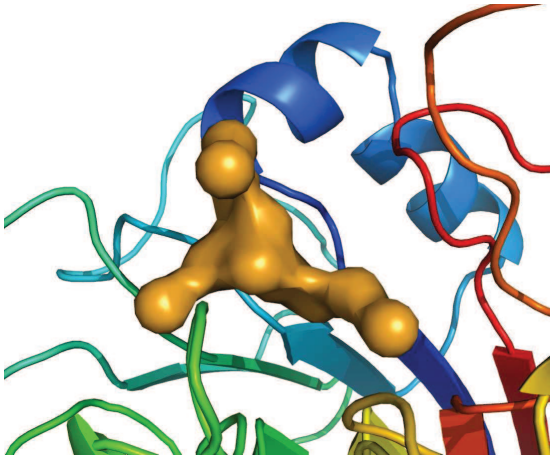


Figura 3-16. Esferas seleccionadas de *GAGA* para el pre-filtrado en el cribado de *Ape1*.

En este cribado también se realiza un protocolo alternativo, en el que se utiliza el programa *FRED* como pre-filtro de complementariedad de forma. La preparación del centro activo para ser usada con *FRED* se hace mediante la interfaz gráfica de *fred_receptor*, incluido también en el paquete de *FRED*. Utilizando el fragmento de ADN para delimitar el centro activo, se obtiene una caja con un volumen de 5457 $Å^3$, siendo el contorno externo de 1974 $Å^3$ y el contorno interno de 41 $Å^3$. No se configura ninguna restricción. El pre-filtrado se realiza utilizando la función de puntuación *PLP* (*Piecewise Linear Potential*) sin optimización posterior.

Finalmente, se continúa en cada caso con el filtrado y el resto de etapas del protocolo estándar.

3.2.7.7.3 HDC

A partir de la estructura modelada y refinada, se sitúa el centro activo colocando una caja tridimensional de 5 Å alrededor de la histidina en su conformación de aldimina externa. La Figura 3-17 muestra una imagen de la estructura dimérica de *HDC*. En ella se aprecia la localización del centro activo, el cual está cerca de la aldimina interna, representada en esferas. Puede apreciarse que se encuentra muy enterrado.

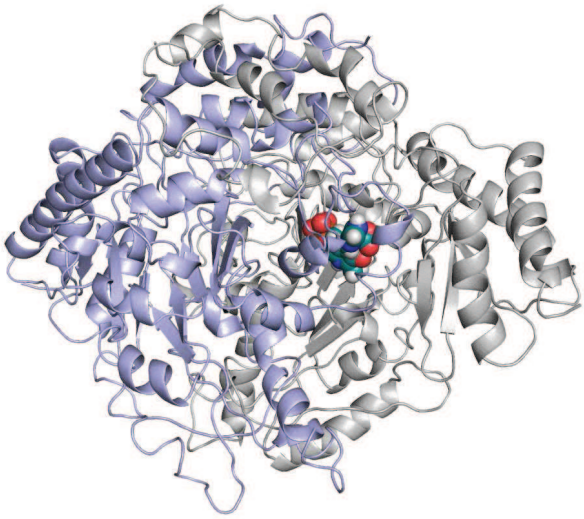


Figura 3-17. Representación de la estructura 3D de la forma dimérica de *HDC*. Cada monómero está representado en un color diferente. La aldimina interna se muestra en esferas.

El espaciado de *grid* es de 0.5 Å. Los átomos de hidrógeno se añaden empleando el servidor *H++* en su configuración por defecto. En este caso se aplica el protocolo estándar de cribado pero suprimiendo la etapa de pre-filtrado. Para que resulte factible eliminar esta etapa, se hace uso de la gran capacidad de cálculo del supercomputador *MareNostrum* y se realiza directamente la etapa de filtrado con la quimioteca completa.

3.2.7.7.4 PCNA

De la estructura del *PDB* (ver apartado 3.1.1.6.4 de la página 32) se eliminan todos aquellos átomos que no pertenezcan a la proteína, y ésta se prepara según el protocolo estándar para añadir los átomos de hidrógeno y los parámetros de cargas y radios atómicos. El centro activo se delimita alrededor de la zona donde se encontraba el péptido *FEN1*, y el espaciado de *grid* usado es de 0.5 Å. El protocolo de cribado se aplica también en su modalidad estándar, pero tras seleccionar las moléculas, comprarlas y probarlas, se realiza una búsqueda de análogos tanto en *ZINC* con la función específica para ello como con el programa *ROCS* de *Openeye* (superposición estructural, seleccionando aquellas cuyo índice de *Tanimoto* es ≥ 0.9). Las moléculas análogas encontradas también siguen el protocolo estándar a partir de la etapa de filtrado con *CDOCK*.

3.2.7.7.5 FtsZ

Para preparar la estructura 3D, se eliminan del fichero *PDB* (ver apartado 3.1.1.6.5 de la página 33) todos los átomos que no pertenezcan a la cadena *A* de la proteína *FtsZ*, y se añaden hidrógenos, cargas y radios atómicos del modo estándar. El centro activo se delimita mediante una caja de 7.5 Å alrededor de la posición donde se encontraba la molécula de *GDP*. El espaciado de *grid* usado es el habitual de 0.5 Å. El protocolo de cribado se aplica del modo estándar, pero como en el caso del *PCNA*, se realizan posteriormente búsquedas en *ZINC* de compuestos análogos a los originalmente seleccionados. Dichos análogos también siguen el protocolo estándar a partir de la etapa de filtrado con *CDOCK*.

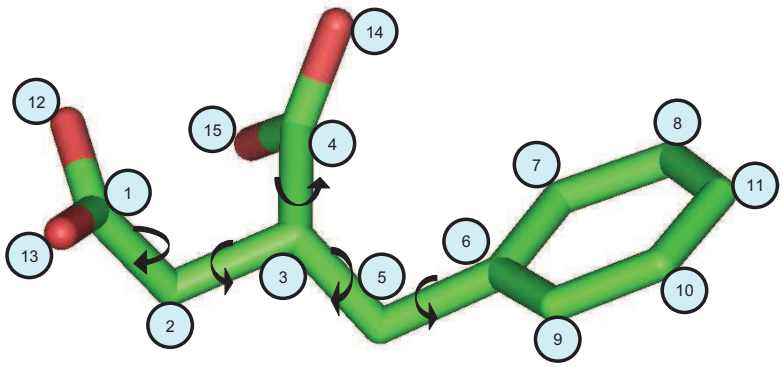
RESULTADOS

4. Resultados

A continuación se exponen los resultados obtenidos en las pruebas realizadas con la plataforma *VSDMIP*, tanto retrospectivas como prospectivas. En primer lugar se describen los resultados para tres de sus componentes: análisis conformacional, modelo de solvente implícito y algoritmo de *docking*. Se termina con las pruebas de integración de la plataforma y su aplicación en casos reales de cribado virtual.

4.1. ALFA para el Análisis Conformacional de Ligandos

El nuevo *ALFA* presenta como principales ventajas el hecho de poder detectar y clasificar automáticamente los enlaces rotables, asignando a cada uno de ellos sus posibles estados rotaméricos y generar los diferentes confórmeros a partir de sus combinaciones. En primer lugar se exponen dos ejemplos de generación automática de confórmeros para el caso de ligandos sencillos, para posteriormente mostrar una serie de resultados de varios estudios con diferentes conjuntos de ligandos (ver en Materiales y Métodos el apartado 3.1.1.1 de la página 27). La primera molécula que se utiliza en el ejemplo es el *L-benzilsuccinato*, que puede verse en la Figura 4-1 en la cuál se han omitido los átomos de hidrógeno para simplificar. Cada átomo se ha etiquetado con un número y se han indicado con una flecha negra los enlaces que pueden rotar. También se muestra una tabla con los ángulos que puede adoptar cada uno de los torsionales de esta molécula.



Dihedro	Números de Átomo	Ángulos asignados
1	14 4 3 2	0. 60. -60. 120. -120. 180.
2	7 6 5 3	0. 60. -60. 120. -120. 180.
3	6 5 3 2	60. -60. 180.
4	4 3 2 1	60. -60. 180.
5	3 2 1 12	0. 60. -60. 120. -120. 180

Figura 4-1. Ejemplo de enlaces rotables para un ligando y sus posibles estados rotaméricos.

Como puede apreciarse en la Figura 4-1, el programa detecta correctamente los 5 enlaces rotables de la molécula. Los ángulos asignados a cada torsional son los especificados automáticamente por el programa en función de los grupos químicos que se encuentran a ambos lados del enlace a rotar. Del mismo modo, puede observarse que se encuentran correctamente asignados. Se generan 1944 (6x6x3x3x6) confórmeros, de entre los cuales el de mejor *RMSD* encontrado con respecto a la estructura cristalográfica es 0.44 Å. El tiempo de cálculo es ~230 ms. En la Figura 4-2 se puede ver una superposición entre la conformación bioactiva y la conformación generada más parecida.

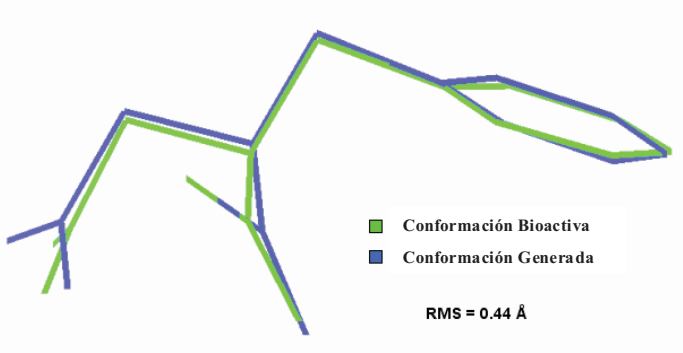
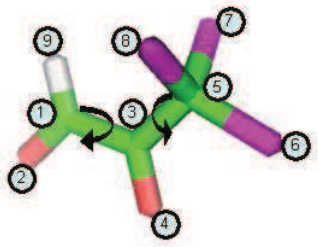


Figura 4-2. Superposición de conformaciones bioactiva y generada.

En la Figura 4-3 se muestra el otro ejemplo que tiene como objetivo comprobar la aplicación de las reglas de *ALFA* para detectar un torsional de grupo carbonilo y otro de grupo triclorometilo. Se trata de un 3,3,3-tricloro-2,2-oxopropanal. Como puede verse, se asignan los ángulos correctos del carbonilo, dejando el triclorometilo sin rotación.



Dihedro	Números de Átomo	Ángulos asignados
1	6 5 3 1	-
2	4 3 1 2	-120. -90. 0. 90. 120. 180.

Figura 4-3. Ejemplo de detección de grupos funcionales en *ALFA*.

Tras la comprobación de la funcionalidad básica de *ALFA*, se procede a realizar tests más completos con conjuntos de ligandos para los que se tiene información cristalográfica, y por lo tanto puede comprobarse si alguna de las conformaciones generadas por *ALFA* está próxima a la bioactiva. El primero de estos tests es el realizado

con el conjunto de Boström et al. (ver apartado 3.1.1.1.1 de la página 27), donde no se hace uso de la búsqueda *MCSA*. En la Tabla 4-I se muestran algunos datos de interés como pueden ser el número de confórmeros generados y seleccionados por ligando, las mejores energías, los mejores *RMSDs*, o el tiempo de computación. Dicha tabla contiene el identificador del ligando (*LIG*), el número de átomos (*N.AT*), el número de torsionales (*N.TOR*), el *RMSD* mínimo encontrado (*MIN.GLOBAL RMS*), el *RMSD* mínimo seleccionado (*MIN.SEL.RMS*), el porcentaje de confórmeros con un *RMSD* menor a 1 Å (*% RMS<1 Å SEL.*), el tiempo de cálculo en milisegundos (*TIEMPO CALC.*), el número de confórmeros seleccionados (*N.CONF.SEL.*) y el número de confórmeros generados (*N.CONF.GEN.*). Además, se presenta una serie de gráficos con algunas relaciones interesantes (número de enlaces rotables frente a tiempo de computación, frente al porcentaje de conformaciones bioactivas seleccionadas y tiempo que tarda en generarse cada confórmero).

Tabla 4-I. Resumen de resultados de *ALFA* para el conjunto de Böstrom et al.

LIG.	*N. AT*	*N.TOR*	*MIN.GLOBAL RMS*	*MIN. SEL.RMS*	*% RMS<1 Å SEL.*	*TIEMPO CALC. (ms)*	*N.CONF. SEL.*	*N.CONF. GEN.*
lig_1a28	53	1	0.052	0.052	100.00 %	0	6	6
lig_1tng	24	1	0.110	0.110	100.00 %	0	3	3
lig_1tnh	17	1	0.053	0.053	100.00 %	0	6	6
lig_1qft	19	2	0.097	0.097	88.89 %	0	18	18
lig_1ftm	22	3	0.315	0.315	16.67 %	10	108	108
lig_1phg	31	3	0.626	0.626	8.33 %	10	72	72
lig_3bto	27	3	0.130	0.130	44.44 %	0	27	27
lig_1ia3	43	3	0.620	0.620	8.33 %	70	216	216
lig_1fcy	54	3	0.670	0.670	25.00 %	0	8	8
lig_1dg3	42	3	0.521	0.521	50.00 %	0	8	8
lig_1c83	24	4	0.309	0.309	18.75 %	0	48	48
lig_1ecv	19	4	0.199	0.199	18.75 %	0	48	48
lig_1fcz	52	5	0.750	0.750	9.38 %	10	32	32
lig_1gr2	29	4	0.222	0.223	15.74 %	110	648	648
lig_1ian	43	4	0.550	0.550	20.83 %	0	48	48
lig_1frb	40	5	0.467	0.467	4.50 %	420	1000	1944
lig_1bju	34	6	0.723	0.723	3.13 %	10	64	64
lig_1dyr	39	5	0.159	0.159	19.50 %	1220	1000	7776
lig_2izg	31	5	0.249	0.249	7.61 %	110	486	486
lig_1cbx	25	5	0.442	0.442	7.20 %	230	1000	1944
lig_5std	47	5	0.938	0.938	0.20 %	420	1000	1296
lig_6std	38	5	0.323	0.323	2.31 %	120	432	432
lig_7std	38	5	0.378	0.378	4.63 %	110	432	432
lig_1cbs	49	9	0.438	0.438	1.95 %	200	512	512
lig_1dam	32	6	0.210	0.210	6.20 %	270	1000	1458
lig_3std	48	7	0.901	0.901	0.40 %	1860	1000	7776
lig_1ejn	53	7	0.414	0.414	4.30 %	670	1000	1728
lig_1if8	47	7	0.771	0.771	0.80 %	700	1000	2592
lig_1caq	64	8	0.543	0.543	0.80 %	89840	1000	248832
lig_1mtv	65	8	0.649	0.649	2.00 %	17680	1000	46656
lig_1mtw	62	9	0.420	0.420	2.10 %	31840	1000	93312

LIG.	N. AT	N.TOR	MIN.GLOBAL RMS	MIN. SEL.RMS	% RMS<1 Å SEL.	TIEMPO CALC. (ms)	N.CONF. SEL.	N.CONF. GEN.
lig_1pph	59	8	0.875	0.875	0.10 %	25180	1000	82944
lig_1f0u	66	11	0.972	0.972	0.60 %	191040	1000	497664
lig_1fkg	68	11	0.626	1.499	0.00 %	11020400	1000	26873856
lig_1fkh	74	11	0.597	1.377	0.00 %	6436150	1000	13436928
lig_1ppc	69	11	0.418	0.812	0.30 %	835760	1000	1990656

En la Figura 4-4 se representa el número de enlaces rotables frente al porcentaje de conformaciones seleccionado con un *RMSD* por debajo de 1 Å con respecto a la conformación cristalográfica. La fracción de confórmeros por debajo de 1 Å es un indicador de la probabilidad de éxito del proceso de *docking* posterior. Como se verá más adelante, la población de confórmeros por debajo de 1 Å debe estar por encima de ~1% para garantizar el éxito en el docking. Se observa que cuánto mayor es el número de enlaces rotables más difícil es seleccionar la conformación bioactiva. Por encima de ~7 enlaces rotables la fracción de confórmeros correctos (bioactivos) cae excesivamente por debajo del ~1% (ver también Tabla 4-I). Afortunadamente, la mayor parte de los compuestos de las librerías virtuales obedecen las reglas de Lipinski (Lipinski et al., 2001), lo que implica que en su gran mayoría presentan menos de 8 enlaces rotables. Por tanto, estos resultados parecen indicar que el método de generación de confórmeros resulta adecuado para la creación de flexibases a partir de catálogos comerciales.

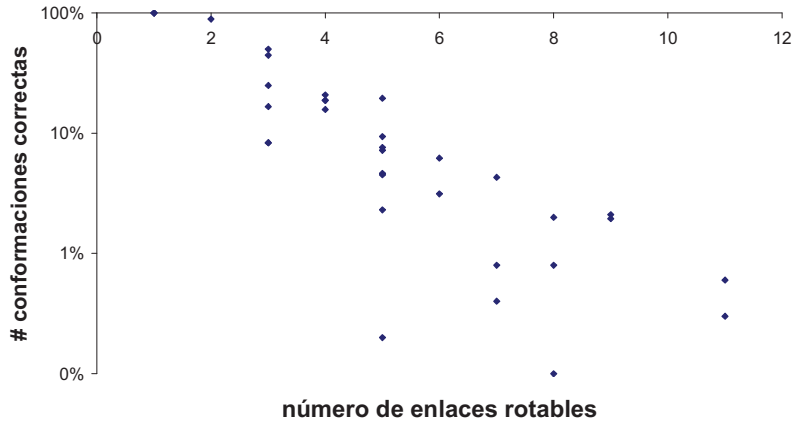


Figura 4-4. Enlaces rotables frente a conformaciones correctas.

El segundo aspecto clave en la generación de confórmeros en protocolos de cribado virtual es el tiempo de cálculo y la dependencia de éste con el número de torsionales. En la Figura 4-5 se comparan el número de enlaces rotables frente al tiempo de computación, en escala semilogarítmica. Puede observarse una progresión exponencial. Por encima de ~7 enlaces rotables se disparan los tiempos (>10 s en la generación de los confórmeros, ver también Tabla 4-I), lo que, junto al resultado de la

Figura 4-4, indica que es necesario tratar de un modo diferente las moléculas grandes, por lo que se introdujo el algoritmo *MCSA* en *ALFA* para estos casos.

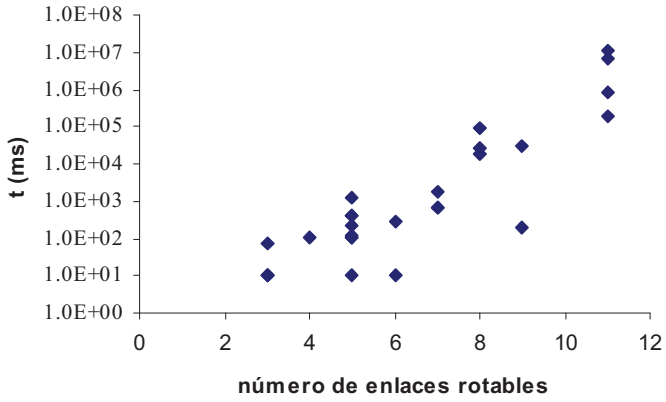


Figura 4-5. Enlaces rotables frente a tiempo de cálculo.

Finalmente, se estudia el tiempo de generación por confórmero. Desviaciones de la linealidad en gráficos del tiempo de generación frente al número de confórmeros son indicativos de un cambio de régimen en el algoritmo, y por tanto de posibles problemas en el código (cuellos de botella en el cálculo de las energías, deficiencias en los accesos a memoria, etc). En la Figura 4-6 se muestra una gráfica dónde se observa que existe una relación lineal casi perfecta entre el número de confórmeros generados y el tiempo invertido, de forma que se manejan con igual eficiencia moléculas pequeñas y moléculas grandes. El programa emplea ~0.43 ms en promedio para la generación de cada confórmero. En conjunto, los tiempos por confórmero y la linealidad del gráfico indican que la implementación es correcta.

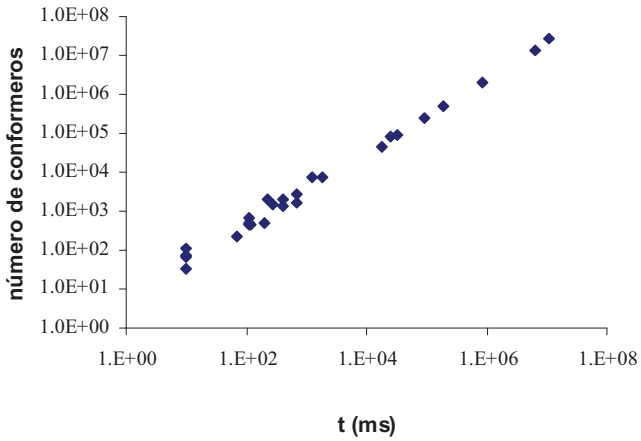


Figura 4-6. Confórmeros generados frente a tiempo de cálculo.

En el segundo de los tests, basado en los 30 ligandos seleccionados en la publicación de Good et al. (ver apartado 3.1.1.1.2 de la página 27), se compara *ALFA* frente a diferentes programas para el análisis conformacional de ligandos permitiendo utilizar el algoritmo *MCSA* en este caso. Los 30 ligandos se pueden clasificar en 3 grupos según su flexibilidad, como se muestra en la Tabla 4-II dónde además se indica el *RMSD* medio obtenido por *ALFA*.

Flexibilidad	# Enlaces rotables	Número de ligandos	*RMSD* Medio
Baja	3-5	15	0.59
Media	6-8	7	0.64
Alta	9-14	8	1.67

Tabla 4-II. Número de ligandos según su flexibilidad y RMSD medio obtenido.

En la Figura 4-7 se muestra el resumen de los resultados obtenidos. En el eje *Y* se representa la media de los mejores *RMSDs* obtenidos con cada ligando. En el eje *X* están los diferentes métodos utilizados en la generación de la conformación bioactiva de cada ligando. Resulta curioso que una minimización de una estructura original de rayos-X dé como resultado desviaciones tan altas de *RMSD*. Puede apreciarse que con *ALFA* se consigue un nivel muy similar al de los programas comerciales, e incluso mejor en algún caso.

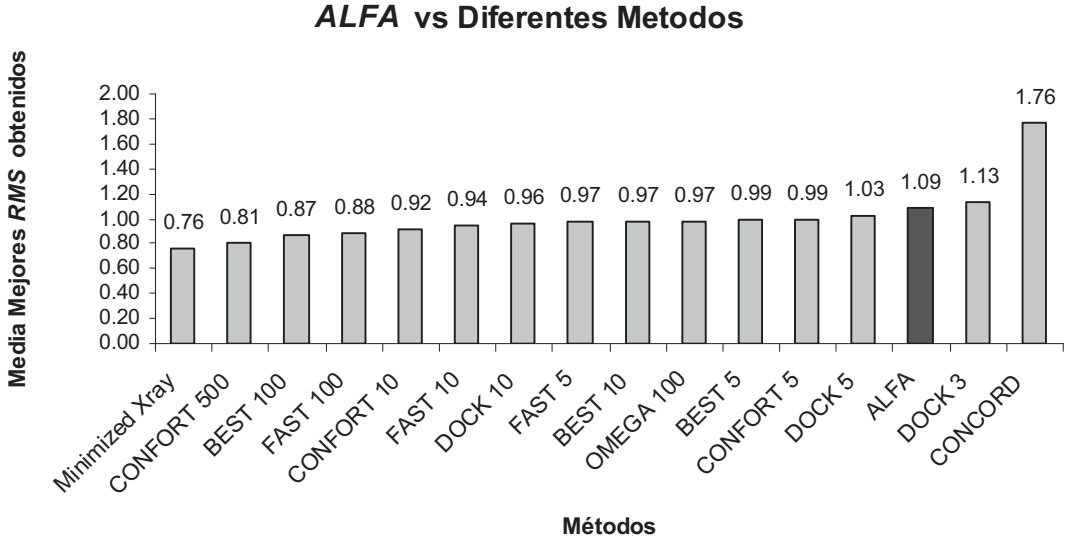


Figura 4-7. Media de *RMSD* obtenida con diferentes métodos en la publicación de Good et al. comparada con *ALFA*.

Un análisis posterior de los resultados obtenidos con *ALFA* indica que podrían mejorarse notablemente si la función de energía que evalúa las conformaciones para hacer su clasificación se complementase con otros términos energéticos (solvatación, entropía...). Lo demuestra el hecho de que el promedio de los mejores *RMSDs*

generados es 0.83 mientras que el promedio de los mejores *RMSDs* seleccionados es 1.09 como se mostró en la Figura 4-7. Por lo tanto, este será un aspecto importante a mejorar. En la Figura 4-8 se muestra gráficamente la relación entre el mejor *RMSD* generado y el seleccionado para cada ligando. Puede apreciarse que, salvo algunas excepciones, la mayoría coinciden con la bisectriz de la gráfica o cerca de ella, por lo que los valores se corresponden. La línea de regresión asociada a los puntos tiene una pendiente de 1.6 y una ordenada en el origen de 0.2, indicando como ya se ha comentado, que para algunos valores de *RMSD* generado, el seleccionado es más alto.

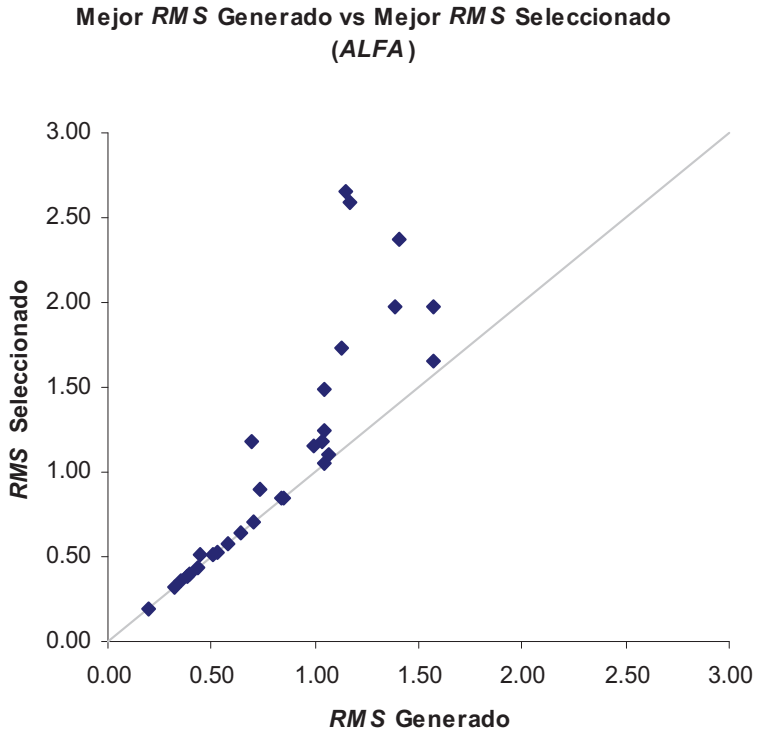


Figura 4-8. Relación entre el *RMSD* generado y el seleccionado.

Otro aspecto a mejorar en *ALFA* es el muestreo conformacional para ligandos grandes, es decir, para aquellos que tienen muchos enlaces rotables y por lo tanto es necesario ejecutar el algoritmo de *MCSA*. En la Figura 4-9 se muestra la relación entre el mejor *RMSD* seleccionado para estos ligandos con el método *MCSA* y el método exhaustivo (sin poner límite al número de confórmeros generados). Puede apreciarse que la mayoría de los puntos caen sobre la bisectriz o cerca de ella, por lo que existe correspondencia entre lo obtenido con el método exhaustivo y utilizando *MCSA*. La pendiente de la línea de regresión asociada a los puntos es 1 y la ordenada en el origen 0.1, por lo que salvo pocas excepciones todos los puntos ajustan bien. Hay casos que podrían parecer extraños, como aquellos en los que se obtiene mejor resultado con

MCSA. Esto es debido a que en el método exhaustivo, al generarse más confórmeros, se llega a obtener alguno con mejor energía pero peor *RMSD* que en el método *MCSA*. Como la selección se hace en función de la energía, se toman esos confórmeros a pesar de tener un peor *RMSD*.

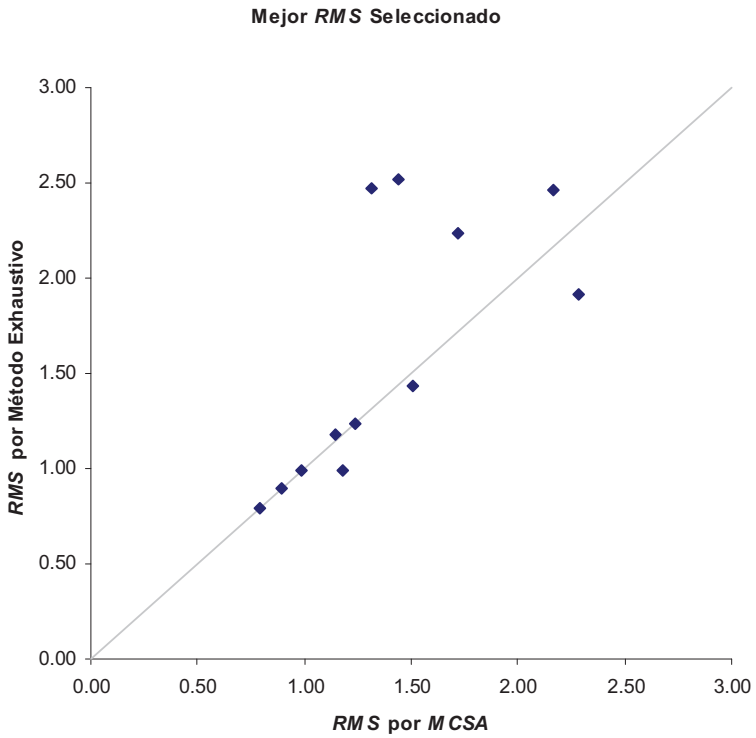


Figura 4-9. Relación entre el *RMSD* seleccionado con el método *MCSA* y el método exhaustivo.

Para demostrar la convergencia del método *MCSA* se presenta en la Figura 4-10 una comparación de la mejor energía obtenida para un confórmero en el método *MCSA* y en el exhaustivo en cada uno de los 12 ligandos para los cuales es necesario aplicar el método *MCSA*. Se observa que con ambos métodos se llega al mismo mínimo energético en todos los ligandos.

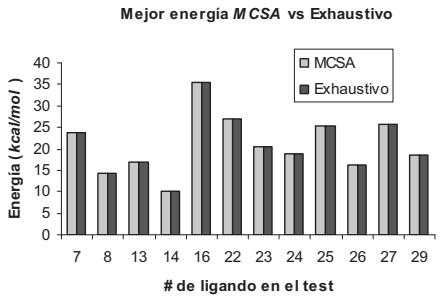


Figura 4-10. Mejor energía en el método *MCSA* frente al exhaustivo.

Por último, para realizar una prueba más completa, se utilizan los ligandos del conjunto de 85 complejos proteína-ligando publicado por *Astex Therapeutics* (ver Materiales y Métodos, apartado 3.1.1.1.3 de la página 27). Dicho conjunto es representativo de un gran número de tipos de proteínas y fue concebido a modo de referente para la realización de pruebas de algoritmos de *docking*. En la Tabla 4-III se presentan los resultados obtenidos al realizar el análisis conformacional con ALFA para los 85 ligandos del conjunto. Dicha tabla muestra para cada complejo: 1) el número de estructuras de partida para el análisis conformacional; 2) el número de enlaces rotables del ligando; 3) el número de confórmeros posibles; 4) el número de confórmeros seleccionados; y 5) el mejor *RMSD* encontrado de entre las estructuras seleccionadas, es decir, a qué distancia de la estructura cristalográfica se encuentra la más cercana de las generadas.

Tabla 4-III. Resumen de resultados de *ALFA* para el conjunto de *ASTEX*.

Complejo	# Estructuras Iniciales	# Torsionales	# Confórmeros Posibles	# Confórmeros Seleccionados	Mejor *RMSD* (Å)
1g9v	1	7	79380	200	1.04
1gkc	1	11	839808	200	0.68
1gm8	5	5	8640	200	0.81
1gpk	1	0	1	1	0.32
1hnn	4	1	28	32	0.49
1hp0	6	2	90	60	0.79
1hq2	3	1	18	21	0.13
1hvy	1	10	1306368	200	1.23
1hwi	1	8	139968	200	0.53
1hww	8	0	8	8	0.48
1ia1	1	2	24	12	0.34
1ig3	1	4	756	200	0.30
1j3j	1	2	18	5	0.27
1jd0	1	3	84	64	0.24
1jje	2	7	72576	200	0.56
1jla	1	7	46656	200	0.56
1k3u	1	7	72576	200	0.76
1ke5	1	4	336	200	0.39
1kzk	8	12	35831808	200	1.08
1l2s	1	4	504	200	0.87
1l7f	8	9	326592	162	0.78
1lpz	1	7	3456	200	1.03
1lrh	1	2	36	26	0.29
1m2z	2	2	84	39	0.26
1meh	1	6	27216	200	0.72
1mmv	1	9	81648	200	0.54
1mzc	10	6	226800	200	0.77
1n1m	5	3	720	200	0.34
1n2j	1	3	81	55	0.08
1n2v	1	3	54	43	0.20
1n46	1	4	1344	57	0.64
1nav	1	5	18816	200	0.59
1of1	2	2	24	22	0.54
1of6	1	3	72	53	0.16
1opk	1	4	1029	200	0.63
1oq5	1	3	105	15	0.92
1owe	1	4	40	13	0.50
1oyt	10	4	630	200	0.90
1p2y	7	1	28	32	0.19
1p62	7	2	84	36	0.18
1pmn	15	7	311040	200	2.25
1q1g	8	3	360	112	0.87
1q41	1	0	1	1	0.38

Complejo	# Estructuras Iniciales	# Torsionales	# Confórmeros Posibles	# Confórmeros Seleccionados	Mejor *RMSD* (Å)
1q4g	1	3	108	91	0.32
1r1h	1	10	559872	200	0.82
1r55	1	11	279936	200	0.32
1r58	1	10	23328	200	1.03
1r9o	1	3	180	127	0.25
1s19	10	5	32400	200	1.20
1s3v	2	6	36288	200	0.81
1sg0	1	2	32	33	0.12
1sj0	10	6	19440	200	1.30
1sq5	1	7	11664	200	0.61
1sqn	4	0	4	4	0.25
1t40	1	7	18144	200	1.30
1t46	17	8	1028160	200	1.66
1t9b	1	5	1008	200	0.89
1tow	1	4	504	200	0.29
1tt1	7	4	4536	200	0.25
1tz8	1	4	900	50	0.26
1u1c	1	6	3888	200	0.39
1u4d	5	0	5	5	0.16
1uml	1	11	2239488	200	0.99
1unl	1	8	69984	200	1.07
1uou	4	2	112	83	0.16
1v0p	1	8	381024	183	0.80
1v48	1	6	972	160	0.45
1v4s	1	5	864	200	0.79
1vcj	2	8	69984	200	0.61
1w1p	6	0	6	6	0.07
1w2g	7	2	84	47	0.25
1x8x	1	3	72	53	0.14
1xm6	3	5	3888	200	0.41
1xoq	1	8	19440	192	0.48
1xoz	6	1	30	26	0.54
1y6b	1	9	367416	170	1.72
1ygc	1	12	17635968	200	0.93
1yqy	9	6	27216	200	1.14
1yv3	3	1	9	12	0.26
1yvf	1	7	21168	200	1.01
1ywr	10	6	72000	200	1.41
1z95	1	6	6048	200	1.11
2bm2	9	7	296352	200	1.47
2br1	1	7	48600	200	0.79
2bsm	1	6	3600	133	0.55

Como puede verse en la tabla, para la mayoría de los ligandos se consigue obtener que al menos una estructura de las seleccionadas esté muy próxima a la cristalográfica ($RMSD \leq 1$ Å), aún teniendo en cuenta que en muchos casos se están seleccionando muy pocas en relación a todas las posibles. Pero esta selección, a pesar de ser un conjunto pequeño, en teoría resultaría apropiado para ser usado en un algoritmo de docking como aproximación a la flexibilidad del ligando, ya que cuenta con conformaciones similares a la cristalográfica.

Aunque el tiempo medio que tarda el análisis conformacional para cada ligando es de ~89 segundos (teniendo en cuenta que puede tener varias estructuras *3D* de partida), se tiene la ventaja de que puede utilizarse en cada *docking* que se realice (independientemente de la proteína) sin tener que realizar de nuevo el análisis.

En la Figura 4-11 se agrupan los resultados en tres categorías: buenos (*RMSD* ≤ 1.0 Å), regulares (1.0 < *RMSD* ≤ 1.5 Å) y malos (*RMSD* > 1.5 Å). Como puede verse se alcanza un alto porcentaje de análisis conformacionales buenos.

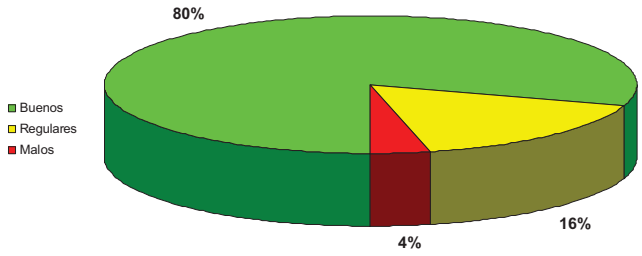


Figura 4-11. Distribución de resultados de *ALFA* para el test de *ASTEX.*

Analizando los resultados, se ha visto que existe una ligera influencia del número de estructuras iniciales sobre el *RMSD* obtenido, como puede apreciarse en la Figura 4-12. Esto puede ser debido a que dichas estructuras iniciales se diferencian fundamentalmente en la configuración de los anillos. Cambios de esta configuración no producen grandes cambios estructurales en la molécula, y debido a esto, grupos de estructuras similares tienen un mayor número de elementos lo que conduce a que haya una menor variabilidad en la selección final. El conjunto de soluciones buenas tiene un 7% de ligandos con 8 o más estructuras iniciales, el de regulares tiene un 43%, y el de malos un 67%. Una posible solución sería quizá realizar grupos de estructuras y como solución final tomar los mejores representantes de cada grupo.

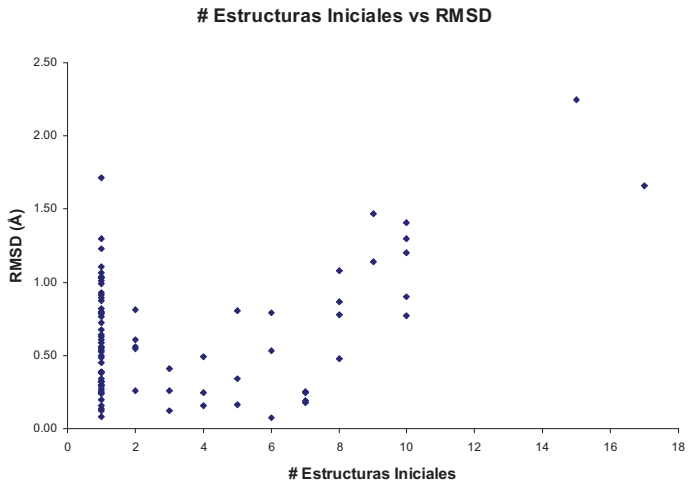


Figura 4-12. Comparación entre el valor de *RMSD* obtenido y el número de estructuras iniciales por ligando para *ALFA.*

Y como se comentó con anterioridad, también el número de torsionales influye en la dificultad para encontrar una estructura similar a la cristalográfica, como puede verse

en la Figura 4-13, donde se aprecia una tendencia a incrementar el *RMSD* según aumenta el número de torsionales de la molécula. De hecho, dentro del conjunto de soluciones malas, el 67% de los ligandos tienen 8 o más torsionales, en el conjunto regular son el 29%, y en el bueno son sólo el 16%.

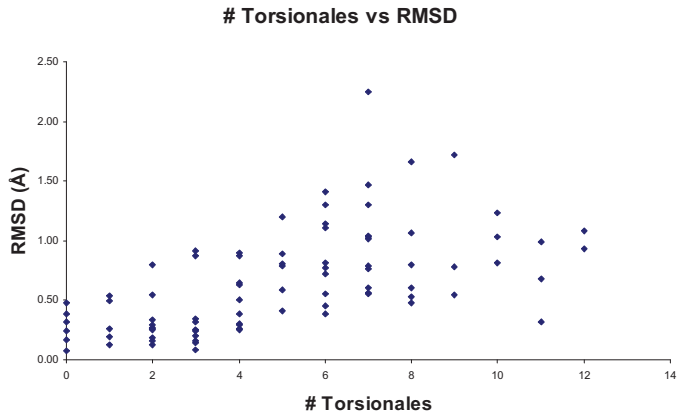


Figura 4-13. Número de torsionales frente al valor de *RMSD* en las pruebas de *ALFA* con el conjunto de *ASTEX*.

También se han realizado estudios (empleando el conjunto de *ASTEX*) en los que se sustituía la función de energía en *ALFA* (*van der Waals 1-4*) por: 1) el área de superficie accesible al solvente (*SASA*), ó 2) la energía calculada con el campo de fuerzas de *AMBER* (Case et al., 2005). En ninguno de los dos casos se observó una mejora significativa (datos no mostrados). Esto apoya la idea de que quizá una opción para mejorar los resultados sea, como se comentó anteriormente, el uso de grupos de soluciones para tomar los mejores elementos de cada uno de modo que se tenga una representación homogénea del espacio de soluciones.

4.2. ISM como Modelo de Solvente Implícito Rápido y Eficaz

La solvatación juega un papel importante en la caracterización energética de la unión proteína-ligando. Debido a que la incorporación del tratamiento explícito del solvente durante el docking resulta algo impracticable, se ha desarrollado un método que lo trata de forma implícita mediante aproximaciones que necesitan ser ajustadas. Un aspecto clave en el desarrollo de cualquier método que involucre ajuste de parámetros es el conjunto de datos que se emplea. En este caso se han utilizado 826 poses proteína-ligando para un conjunto de 23 proteínas (ver el apartado 3.1.1.3 de la página 28). Dichas poses cubren un amplio rango (alrededor de 40 *kcal/mol*) de energías libres de unión electrostáticas. También hay una amplia variedad estructural, tanto en la

arquitectura de las proteínas como en los grupos funcionales de los ligandos. Además los ligandos presentan distintas distribuciones de carga. Por último, hay también un número considerable de orientaciones representativas para cada complejo (20 en promedio), cubriendo de esta manera un amplio espectro de *RMSDs* (desde estructuras similares a la nativa hasta estructuras que se alejan ~10 Å). Así pues, el conjunto de datos utilizado presenta la variabilidad suficiente como para garantizar la generalidad de los resultados.

Otro aspecto a tener en cuenta es el número de parámetros a ajustar, que en el modelo *ISM* es relativamente pequeño (ver Tabla 3-VII en la página 65). Dejando a un lado los parámetros referentes a cargas, radios y los involucrados en cálculo de superficie, *ISM* tiene un total de 17 ó 19 parámetros, dependiendo del tipo de modelo que se elija. El básico, correspondiente a la Ecuación [3-10] de la página 62, contiene ocho parámetros: h_+, h_-, g, λ_+, λ_-, ε, r_{probe}, y el factor de escala para los radios atómicos. Seis de estos ocho parámetros se utilizan para la optimización, mientras que ε y r_{probe} se mantienen fijos. Para lograr reproducir los resultados obtenidos mediante la resolución numérica de la ecuación de *Poisson*, es necesario introducir una corrección basada en los enlaces por puente de hidrógeno. Estos enlaces añaden dos nuevos parámetros al modelo: la distancia donador-aceptor, y el ángulo formado entre el donador, el aceptor y su antecedente. Además, para obtener un ajuste razonable, se utilizan otros cinco parámetros que tienen en cuenta la naturaleza electrostática de los átomos que forman el enlace por puente de hidrógeno (neutros o cargados).

La comparación de las energías libres de unión electrostáticas usando la ecuación de *Poisson* y el modelo *ISM* (usando el conjunto de parámetros optimizado) muestra una relación de tipo exponencial entre ambas, como puede verse en el apartado a) de la Figura 4-14. El motivo de que esta relación sea exponencial no está claro y podría ser materia de estudio en el futuro. Teniendo en cuenta esta dependencia, se trata de realizar el ajuste de dos maneras: 1) directamente, llamado modelo 1, mediante la ecuación $Poisson = A\ exp(B \times SCP\text{-}ISM)$; y 2) un ajuste lineal del modelo exponencial, llamado modelo 2, mediante la ecuación $Poisson = C + D(A\ exp(B \times SCP\text{-}ISM))$, para tener en cuenta las desviaciones sistemáticas respecto a la línea exponencial. Con estos nuevos parámetros (dos o cuatro dependiendo del modelo seleccionado), se completa el conjunto final de 17 ó 19 parámetros para *ISM*.

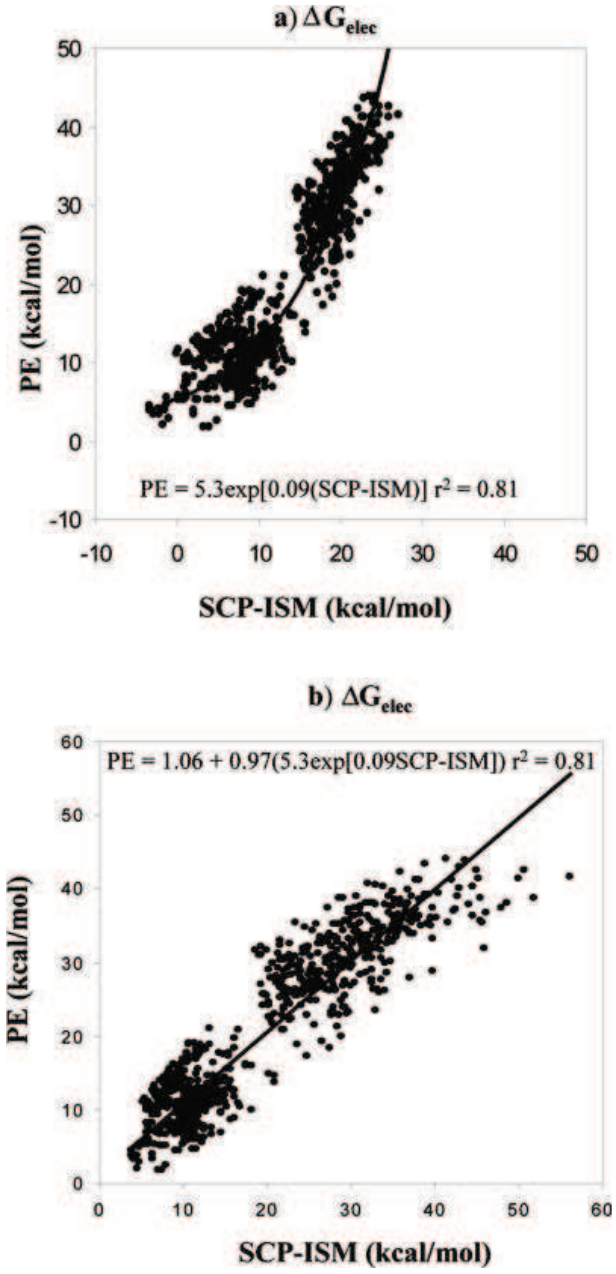


Figura 4-14. Correlaciones entre la energía libre de unión electrostática total (en *kcal/mol*) obtenidas mediante la resolución de la ecuación de *Poisson* (*PE*) y el método *SCP-ISM*. Siendo a) la correlación directa, y b) la correlación tras una corrección logarítmica.

El resumen de los resultados de los dos ajustes (modelo 1 y 2), así como los tests *LOO*, pueden verse en la Tabla 4-IV. La fila etiquetada como *Todos* se corresponde con el caso de validación cruzada estándar: cada conjunto de poses para una proteína dada se elimina, se genera un modelo con los restantes conjuntos, y con dicho modelo se predice el conjunto que fue eliminado. Para este caso, como es de esperar, el *RMSD* obtenido en la validación cruzada es ligeramente superior que el ajustado específico. El resto de filas de la Tabla 4-IV corresponden a los resultados parciales del caso *Todos*.

Por ejemplo, la primera columna muestra el resultado del ajuste para el modelo obtenido tras eliminar el conjunto de poses para *1HVI*, utilizando el resto como conjunto de entrenamiento, y las columnas de validación cruzada *LOO* presentan el resultado de este modelo aplicado a las poses de *1HVI*. El modelo 2 parece obtener resultados de *RMSD* ligeramente mejores que con el modelo 1: 4.16 *kcal/mol* frente a 4.20 respectivamente. El mejor modelo presenta un r^2 (o q^2) de validación cruzada de 0.81, una pendiente de 0.97, y una ordenada en el origen en 1.06 *kcal/mol*. En el apartado b) de la Figura 4-14 se representa una comparación con los datos obtenidos de *Poisson*. El hecho de que los valores de *RMSD* entre la energías libres de unión ajustadas y las de validación cruzada en los tests *LOO* (4.20 contra 4.33) sean relativamente similares indica que no hay evidencia de que se esté produciendo un sobreajuste. Así, se supone que los resultados presentados aquí se mantendrían aunque se utilizasen conjuntos diferentes de complejos.

	Ajuste								**Validación Cruzada *LOO***		
	PE = A exp(B × ISM)				***PE = C + D(A exp(B × ISM))***						
Compl.	**A**	**B**	r^2	***RMSD***	**C**	**D**	r^2	***RMSD***	***RMSD* Mod. 1**	***RMSD* Mod. 2**	q^2
1HVI	5.33	0.09	0.80	4.11	0.79	0.99	0.86	4.07	5.88	5.70	0.70
1HVJ	5.34	0.09	0.80	4.04	0.85	0.98	0.87	4.00	6.14	5.96	0.40
1HVK	5.31	0.09	0.80	4.17	0.85	0.98	0.86	4.14	4.76	4.61	0.81
1HIH	5.32	0.09	0.80	4.15	0.82	0.98	0.87	4.12	4.96	4.69	0.76
1HPX	5.30	0.09	0.80	4.14	0.84	0.98	0.87	4.10	4.69	4.53	0.50
1MCJ	5.47	0.09	0.83	4.13	1.29	0.96	0.87	4.08	5.21	5.76	0.71
1RBP	5.52	0.09	0.81	4.22	1.10	0.97	0.86	4.18	3.37	4.12	0.44
2UPJ	5.25	0.09	0.81	4.16	0.92	0.98	0.87	4.12	5.04	5.00	0.36
1ABE	4.69	0.09	0.85	4.29	1.54	0.94	0.87	4.22	5.19	4.19	0.93
1AJX	5.46	0.09	0.81	4.28	1.31	0.96	0.86	4.23	2.46	3.18	0.73
5ABP	4.86	0.09	0.84	4.23	1.46	0.95	0.87	4.16	5.96	4.97	0.86
1DBB	5.30	0.09	0.81	4.21	1.06	0.97	0.86	4.17	1.55	2.28	0.98
1FKG	5.32	0.09	0.81	4.25	1.12	0.97	0.86	4.20	1.48	2.16	0.66
1FKH	5.32	0.09	0.81	4.25	1.11	0.97	0.86	4.20	1.26	1.98	0.95
1MRK	5.28	0.09	0.81	4.20	1.06	0.97	0.87	4.16	2.44	2.17	0.62
1STP	5.57	0.09	0.82	4.25	1.06	0.97	0.86	4.21	2.59	3.22	0.80
1B9V	5.29	0.09	0.81	4.24	1.00	0.98	0.87	4.19	4.38	3.92	0.92
1DBM	5.24	0.09	0.82	4.23	1.05	0.97	0.87	4.18	3.67	3.08	0.68
1TNG	5.32	0.09	0.81	4.20	1.03	0.97	0.86	4.16	2.59	3.41	0.87
1TNI	5.41	0.09	0.80	4.23	0.91	0.98	0.86	4.19	1.17	1.77	0.68
1TNK	5.37	0.09	0.81	4.22	1.07	0.97	0.86	4.18	2.49	3.08	0.69
1TNL	5.30	0.09	0.81	4.21	1.07	0.97	0.86	4.17	1.36	1.80	0.73
1BMA	5.27	0.09	0.81	4.24	1.12	0.97	0.86	4.19	1.85	1.53	0.19
Todos	5.30	0.09	0.81	**4.20**	1.06	0.97	0.87	**4.16**	**4.40**	**4.33**	**0.81**

Tabla 4-IV. Resultados de la evaluación del modelo *ISM*.

También se obtienen resultados satisfactorios para las distintas componentes de la energía libre de unión electrostática. Los coeficientes de correlación cuadrática oscilan entre 0.79 y 0.88 (ver Figura 4-15). Las pendientes están próximas a uno (1.13 para el

término coulómbico y 1.01 para la desolvatación del ligando) excepto para la desolvatación del receptor (1.72). La ordenada en el origen está próxima a cero en todos los casos. Así, el modelo *ISM* reproduce, no sólo la energía total, sino también sus contribuciones individuales.

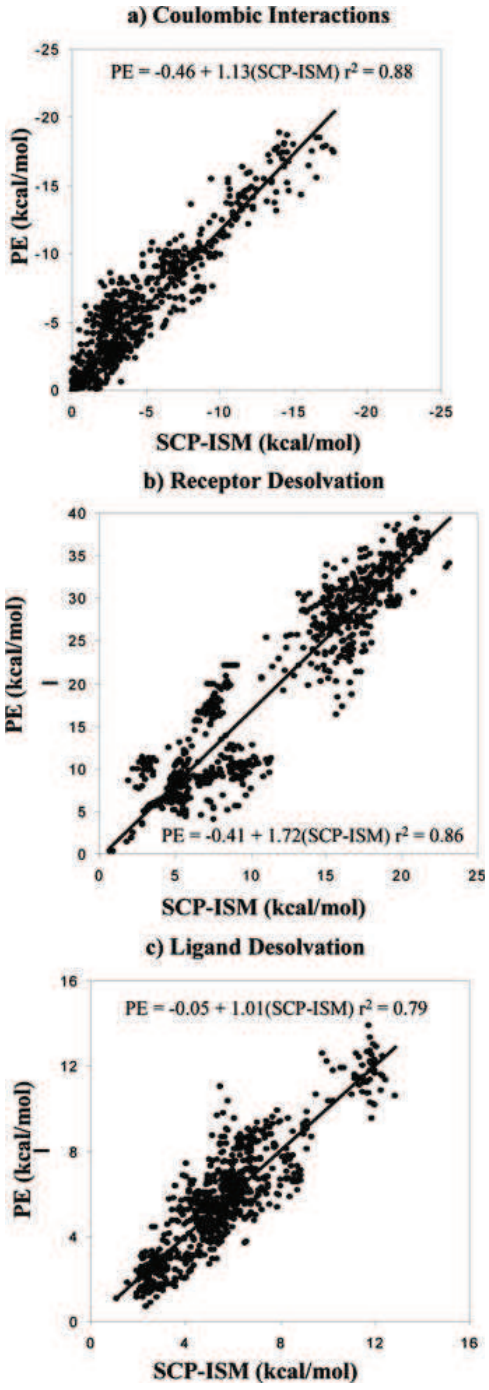
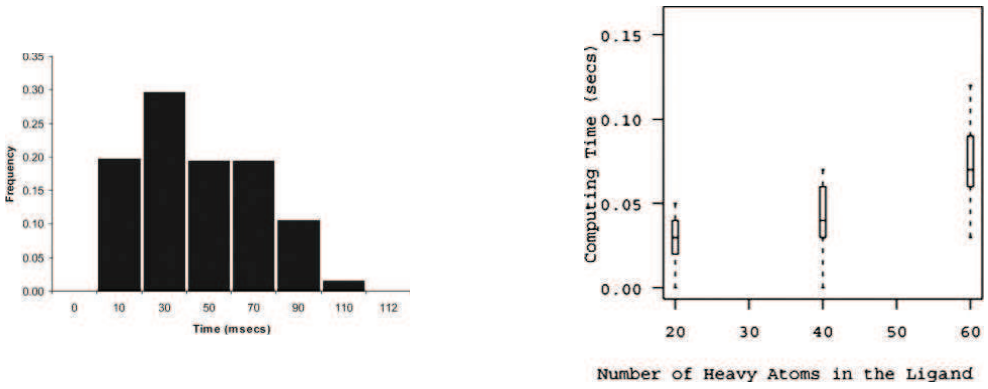


Figura 4-15. Comparación de las diferentes contribuciones a la energía libre de unión electrostática resolviendo la ecuación de Poisson y utilizando el modelo *ISM*. (a) Contribución coulómbica; (b) desolvatación del receptor; y (c) desolvatación del ligando.

Respecto a los tiempos de computación, el apartado a) de la Figura 4-16 muestra un histograma de éstos para los cálculos realizados sobre las 826 poses. El tiempo medio de computación es de 40 ms, teniendo la moda en 30 ms. Estos resultados mejoran la mayoría de las aproximaciones basadas en *GB* (Bashford & Case, 2000; Still et al., 1990). La dependencia del tiempo de computación con el número de átomos pesados se muestra en el apartado b) de la Figura 4-16. Para cada caja, los datos se dividen en cuatro intervalos: un cuarto de los datos (percentil 25%) está entre el extremo inferior de la línea punteada y la base de la caja, otro cuarto está entre la base de la caja y la línea de la mediana, otro entre esta línea y la parte superior de la caja, y el último va desde la parte superior de la caja hasta la parte superior de la línea punteada. Se puede ver que se tiene aproximadamente una dependencia lineal entre el número de átomos pesados y el tiempo de computación.



a) b)

Figura 4-16. a) Distribución de frecuencia para el tiempo de computación requerido por pose en el método *ISM*. b) Relación entre el tamaño del ligando y el tiempo de computación.

4.3. Docking Proteína-Ligando con CDOCK

En primer lugar se evaluó la capacidad de *CDOCK* para la correcta realización del *docking* rígido, es decir, aquel en el que la estructura interna del ligando permanece fija en su conformación cristalográfica. En la Figura 4-17 se muestra la clasificación de los resultados obtenidos tras el *docking* rígido con el conjunto de ligandos de *Astex Therapeutics*.

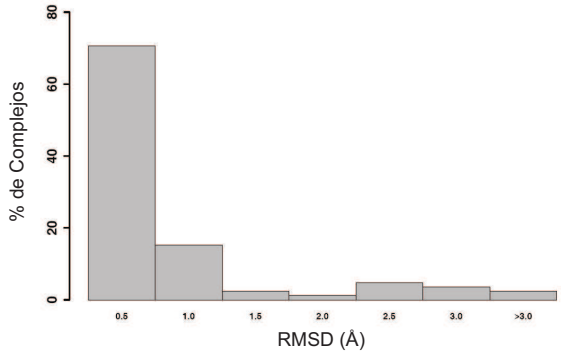


Figura 4-17. Histograma de resultados de *docking* rígido para el conjunto de datos de *Astex Therapeutics*. Los resultados se agrupan según el rango de *RMSD* en el que se encuentra la mejor solución.

En la gráfica se aprecia la habilidad de *CDOCK* a la hora de situar correctamente en el centro activo la conformación cristalográfica del ligando. Los resultados son realmente buenos, aproximadamente el 90% de los ligandos son colocan con un *RMSD* ≤ 2 Å, tomando en promedio sólo ~12 s por *docking*. Resulta también notable que algo más del 70% presenta soluciones con *RMSD* ≤ 0.5 Å. La función de puntuación es lo suficientemente buena como para distinguir la pose cristalográfica de entre todas las posibles.

La segunda prueba consistía en evaluar el *docking* flexible, que utiliza el algoritmo de *MCSA* utilizando las estructuras generadas por *ALFA*. Como se puede apreciar en la Figura 4-18 los resultados no son excesivamente buenos, hay demasiadas estructuras con un *RMSD* > 2 Å. Aunque se puede obtener una conclusión importante de esta gráfica: no se obtienen dockings buenos a partir de un *RMSD* de 1 Å en *ALFA*.

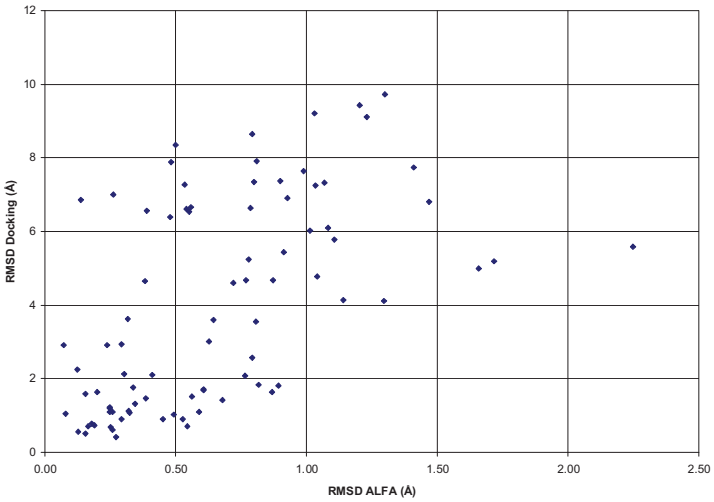


Figura 4-18. Valores de *RMSD* obtenidos en *ALFA* frente a los valores de *RMSD* obtenidos por *CDOCK* en el *docking* flexible.

Con el objetivo de evaluar sólo el algoritmo del *docking* flexible, se han eliminado del conjunto de datos de *Astex Therapeutics* aquellos ligandos para los que no se consigue en el *docking* rígido una solución con *RMSD* ≤ 2 Å, y también aquellos cuyo análisis conformacional con *ALFA* no pueda generar algún confórmeros con *RMSD* ≤ 1 Å. Los resultados se muestran en la Figura 4-19.

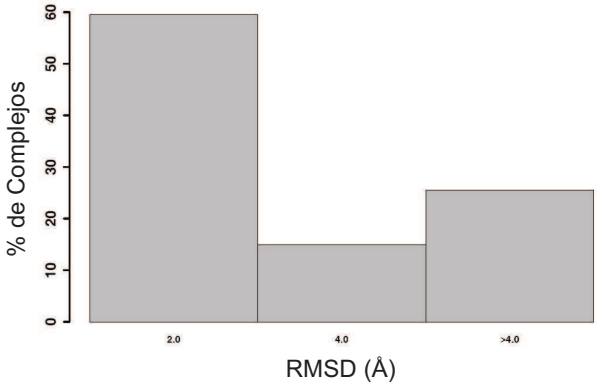


Figura 4-19. Histograma de resultados de *docking* flexible para el conjunto seleccionado de datos de *Astex Therapeutics*. Los resultados se agrupan según el rango de *RMSD* en el que se encuentra la mejor solución.

Como muestra la gráfica, el 60% de los resultados de *docking* tienen un *RMSD* ≤ 2 Å, aunque aún se tiene un número relativamente alto de casos malos. El tiempo medio por *docking* flexible es de ~298 s. En la Figura 4-20 se muestran un par de ejemplos del resultado del *docking* flexible para dos casos particularmente difíciles.

1xm6

RMSD: 0.67 Å

1kzk

RMSD: 1.32 Å

Figura 4-20. Ejemplo de resultados de *docking* para un par de casos particularmente difíciles: *1xm6* y *1kzk*. Con los carbonos en verde se muestra la solución de *docking*, y en gris la estructura cristalográfica. Se han omitido los átomos de hidrógeno por claridad.

Viendo los ejemplos de la Figura 4-20 resulta llamativo que no se consiga un mayor número de soluciones con un *RMSD* menor de 2 Å. Dado que la función de puntuación parece ser bastante efectiva como se vio en el *docking rígido*, se podría

pensar que quizá el problema esté en el método *MCSA* implementado en *CDOCK* para tratar la flexibilidad del ligando (a partir de sus confórmeros generados con *ALFA*). Por ello también se estudia si este algoritmo, como alternativa a la exploración exhaustiva, 1) está implementado correctamente, es decir, converge al mismo mínimo energético que en el caso exhaustivo, y 2) ahorra tiempo de computación en la exploración. La Figura 4-21 muestra el resultado de la comparación de los mínimos energéticos alcanzados por ambos modos de exploración.

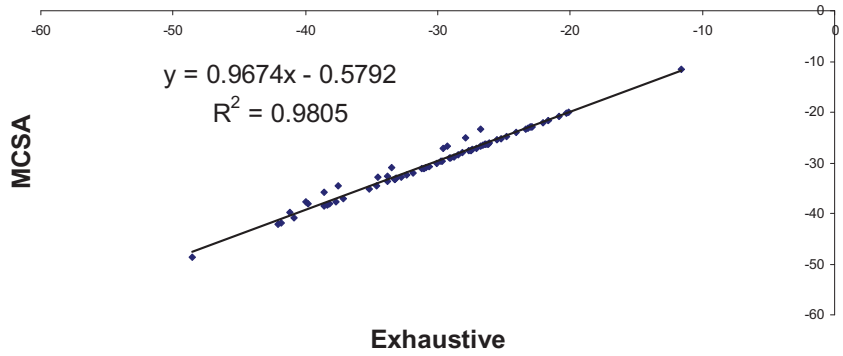


Figura 4-21. Comparación del mínimo energético alcanzado en *CDOCK* para el método de exploración exhaustiva del sitio activo (*Exhaustive*) y el método de exploración *MCSA*.

Como se aprecia en la gráfica, la correlación es casi perfecta, con un valor de R^2 de 0.98, una pendiente de aproximadamente 1 (~0.97) y una ordenada en el origen casi de cero. Además, y como era uno de los objetivos, los tiempos de computación con uno y otro método son muy diferentes para la etapa de exploración del centro activo: en promedio, ~137 s para el método exhaustivo y ~45 s para el método *MCSA* (tres veces más rápido).

Viendo que el problema no está en el método de exploración, se decide aislar un conjunto de posibles errores y se trata de encontrar cuál es la influencia de cada uno de ellos con el objetivo de poder solucionarlos por separado en un futuro. De este modo, dejando a un lado los posibles fallos de preparación de las proteínas o de los ligandos, se han identificado cuatro tipos de problemas: 1) problema en la generación de confórmeros bioactivos (*CORINA* + *ALFA*); 2) problemas en el *docking* rígido; 3) problema en el minimizador de poses en *CDOCK* por el método *SIMPLEX*; y 4) problema en el *docking* flexible. La Figura 4-22 muestra el resumen del análisis por tipología de errores.

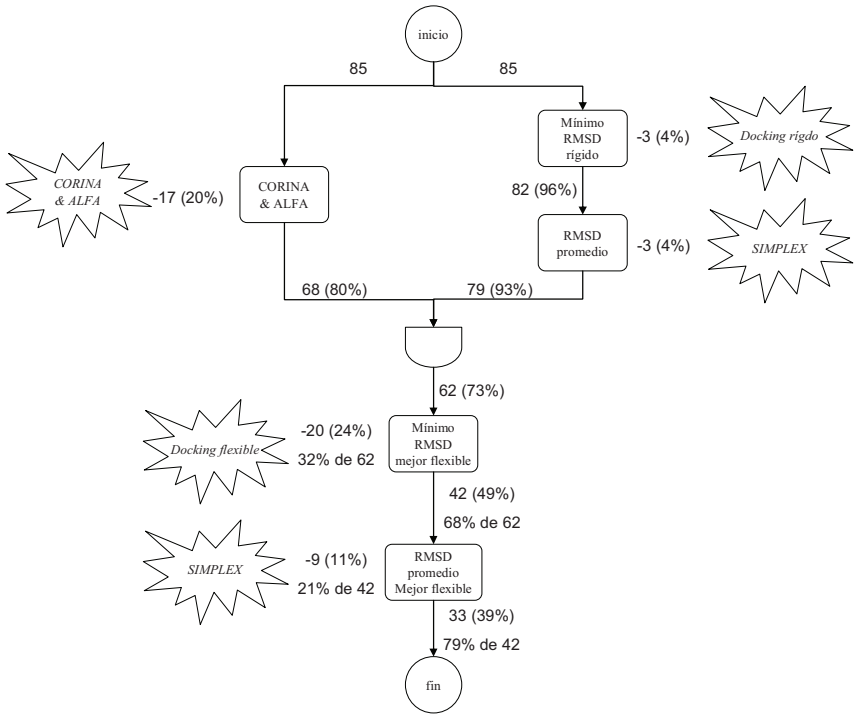


Figura 4-22. Análisis de la tipología de errores en el docking con *CDOCK* para el conjunto de datos de *Astex Therapeutics*.

Para realizar la tipología de errores de la Figura 4-22, en primer lugar se estableció como requisito que los ligandos que pasen a la etapa de *docking* flexible hayan realizado con éxito estos dos caminos: a) generación de conformación bioactiva, y b) *docking* rígido. En el primero se requiere, para cada ligando, que de los confórmeros seleccionados (tras la transformación a 3D con *CORINA* y el análisis conformacional con *ALFA*) al menos uno de ellos tenga un *RMSD* ≤ 1 Å. De aquí, 68 de 85 ligandos pasan esta prueba (el 80%). En el segundo caso se requiere que tengan éxito en el docking rígido (*RMSD* ≤ 2 Å). Aquí los errores pueden ser: 1) que de las 20 ejecuciones ninguna de ellas obtenga un resultado válido (este sería un problema en el propio *docking* rígido); 2) que el *RMSD* promedio tras las 20 ejecuciones sea mayor de 2 Å aun existiendo algunas soluciones individuales por debajo de esa cifra (esto significaría que el minimizador *SIMPLEX* no es capaz de alcanzar el mínimo en todas las ejecuciones).

A la etapa de *docking* flexible solo pasan 62 ligandos, un 73% (aquellos que logran al mismo tiempo conformaciones bioactivas y además buenos resultados de *docking* rígido). En esta etapa de *docking* flexible, con el objetivo de evitar posibles artefactos, se utilizará para cada ligando sólo el mejor confórmero generado (en base a su *RMSD*). De esta manera se puede evaluar si la función de puntuación es capaz de posicionar correctamente un ligando que no está exactamente en su conformación

cristalográfica (además ya se comprobó previamente que el algoritmo de muestreo del centro activo con diferentes confórmeros funciona correctamente). En esta etapa se comprueba que el 32% de ligandos (20 de 62) no logra obtener al menos un resultado con *RMSD* ≤ 2 Å en de las 20 ejecuciones. De los 42 restantes, el 21% fallan en el proceso de *SIMPLEX* (de modo análogo a como sucedía en el *docking* rígido). Así pues, si se consideran porcentajes totales, el ~20% de los ligandos tendrían un fallo en el análisis conformacional, el ~5% fallarían en el docking rígido, el ~15% (4% + 11%) en la minimización por *SIMPLEX*, y el ~24% en el docking flexible. Tan solo 33 de los 85 ligandos (~39%) lograrían un *RMSD* ≤ 2 Å. Es una cifra muy baja si se tiene en cuenta en su conjunto, aunque el objetivo de este último estudio no era otro que determinar los puntos donde el proceso es mejorable. De todos modos, si se tienen en cuenta sólo los ligandos válidos para el *docking* flexible (aquellos cuyo análisis conformación y *docking* rígido son correctos), se alcanza una respetable cifra del ~60% de aciertos (ver Figura 4-19).

Resulta curioso ver como al método *SIMPLEX* le cuesta más encontrar siempre (o al menos la mayoría de las veces) la mejor solución cuando se está utilizando una estructura generada (21% de fallos) que cuando se está utilizando la estructura cristalográfica (4% de fallos). Esto puede deberse a que el encaje de una estructura generada no sería tan bueno como en la estructura cristalográfica por lo que podría tener varios mínimos energéticos diferentes en cuanto a poses se refiere. Para cuantificar de una manera aproximada cuanto afecta cada problema al resultado final, se ha hecho un estudio utilizando las probabilidades mostradas en el gráfico. En este estudio se considera, para cada tipo de problema, cual sería el resultado final (soluciones válidas, RMSD ≤ 2 Å) que se obtendría si se solucionase al 100% el problema en cuestión. Así se obtiene que solucionando el problema en *CORINA+ALFA* se alcanzaría un 51% de soluciones buenas (frente al 39% original). Esto en principio resultaría factible introduciendo nuevas reglas de rotación y quizá un algoritmo de agrupación para la selección de resultados en cuanto a su estructura, en lugar de basarse sólo en la energía interna. Reparar el *docking* rígido tan solo conduciría a alcanzar un 41% de soluciones buenas (frente al 39% original). Aquí habría que estudiar los complejos cristalográficos, pues en algunos casos el problema es que ya de por sí la estructura inicial está muy forzada y contiene choques (como es el caso de *1r58*). En cambio, reparando los problemas en el minimizador de *SIMPLEX* se alcanzaría un 52% (frente al 39% original); en este caso habría que considerar el uso de otro tipo de minimizador, con

mejor convergencia y previsiblemente más rápido (ya que el empleo de este método es una de las partes que más tiempo de computación consume en *CDOCK*). Por último, solucionar los problemas en el *docking* flexible parece que sería lo más rentable, pues se lograría alcanzar un 58% de resultados buenos (frente al 39% original). Para resolverlo seguramente habría que tratar con dos alternativas: 1) tener una función de puntuación más permisiva, en cuanto a que sea capaz de reconocer soluciones buenas que estén ligeramente alejadas de la mejor; ó 2) permitir ligeros movimientos de los torsionales del ligando durante el *docking*. Esta última solución parece más aconsejable, por varios motivos: a) se obtendrían soluciones aún más parecidas a la cristalográfica; b) no habría que modificar la función de puntuación, pues parece funcionar muy bien en el *docking* rígido; y c) llevar al ligando a una conformación más próxima a la cristalográfica, mejorando *ALFA*, mejoraría también los resultados en el *SIMPLEX* (pues como se ha visto, estos empeoran con ligandos más alejados de la conformación cristalográfica). En la Tabla 4-V se muestra un resumen de los diferentes problemas y los complejos del conjunto de datos de *Astex Therapeutics* a los que afectan.

Problema	Complejos Afectados
Generación de conformaciones bioactivas (*CORINA+ALFA*)	1g9v 1hvy 1kzk 1lpz 1pmn 1r58 1s19 1sj0 1t40 1t46 1unl 1y6b 1yqy 1yvf 1ywr 1z95 2bm2
Docking rígido	1jd0 1r58 1tz8
Minimizador *SIMPLEX*	1ke5 1sg0 1x8x 1gm8 1ia1 1k3u 1owe 1q41 1tow 1u4d 1uml 1v4s
Docking flexible	1gpk 1hnn 1hp0 1hww 1ig3 1l7f 1mmv 1mzc 1oq5 1oyt 1q1g 1r1h 1s3v 1t9b 1v0p 1xoq 1xoz 1ygc 2br1 2bsm

Tabla 4-V. Resumen de problemas para el conjunto de *Astex Therapeutics* y complejos que se ven afectados.

4.4. gCOMBINE como Herramienta para QSAR-3D

A continuación se describe la *GUI* para *COMBINE* y las razones por las que es necesaria, y posteriormente se analizan los resultados de la comparación de las aplicaciones de *gCOMBINE* para generar los modelos publicados en dos destacados artículos de *COMBINE*.

4.4.1. GUI para COMBINE

COMBINE es una herramienta quimiométrica que ha dado excelentes resultados en un gran número de estudios. Pero a pesar de su probada utilidad, su uso ha estado casi exclusivamente limitado a las personas que intervinieron en su desarrollo debido fundamentalmente a dos razones: 1) su preparación y ejecución no resultan sencillos; y 2) los resultados obtenidos no están en un formato cómodo ni son directamente interpretables. A continuación se muestra un ejemplo del fichero de configuración que se necesita preparar para *COMBINE*:

```
    0    0    0    #Scramble; Scaling; InputEnergy???
   10    1  4.0    #NumLatVar; DielecModel; DielecCnst
    1    2         #AddVar; NumAddVar
   32   16         #TrainVar; TestVar
    1  0.1         #Pretreat; CutPretreat
    1    3         #RandCrossVal; NumElement
hiv_M01    M01    9.60210  16.96000  16.96000
hiv_M03    M03    8.11350  17.32800  17.32800
hiv_M04    M04    9.72120  18.60400  18.60400
                .
                .
                .
hiv_M48    M48    6.64020  17.04500  17.04500
hiv_M49    M49    5.32790  16.89500  16.89500
hiv_M50    M50    5.86170  16.37900  16.37900
```

No resulta fácil preparar un fichero de este tipo: hay que conocer el significado de cada posición, tratar con el formato rígido de columnas en *Fortran*, conocer los posibles valores que puede tomar cada parámetros y su significado, etc. Además no es complicado incurrir en errores como por ejemplo la asignación de valores incompatibles entre parámetros. El hecho de trabajar con este formato de entrada también dificulta la generación de modelos diferentes haciendo cambios en la configuración: añadir/eliminar complejos, intercambiar sus tipos (entrenamiento o test), etc. Cuando el fichero de configuración está listo, se ejecuta *COMBINE* desde la línea de comandos y a la finalización del cálculo del modelo se obtienen numerosos ficheros, algunos de los cuales se muestran en la Figura 4-23 a modo de ejemplo.

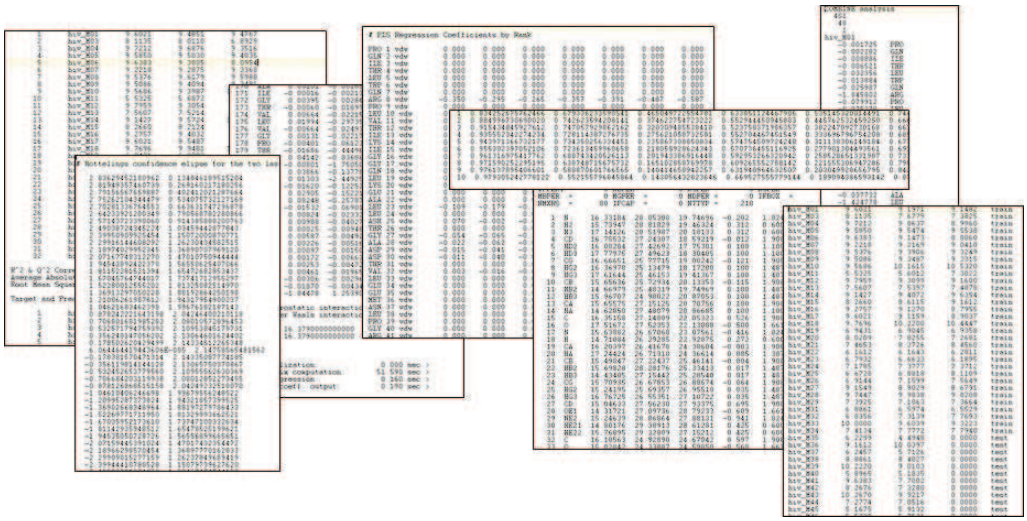


Figura 4-23. Ejemplo de algunos de los ficheros generados por *COMBINE*.

Incluso conociendo el significado de cada uno de los ficheros obtenidos, su interpretación no es sencilla. Se tienen que realizar a mano gráficas y resúmenes para cada modelo generado.

gCOMBINE ha sido diseñado teniendo en mente la necesidad de una herramienta quimiométrica simple y fácil de usar. Una vez que se arranca el programa, éste muestra una barra de menú con dos submenús (*File* y *Help*) y un panel con dos pestañas (ver Figura 4-24): una para la configuración del modelo (*COMBINE Model*) y otra para manejar los resultados (*Results*).

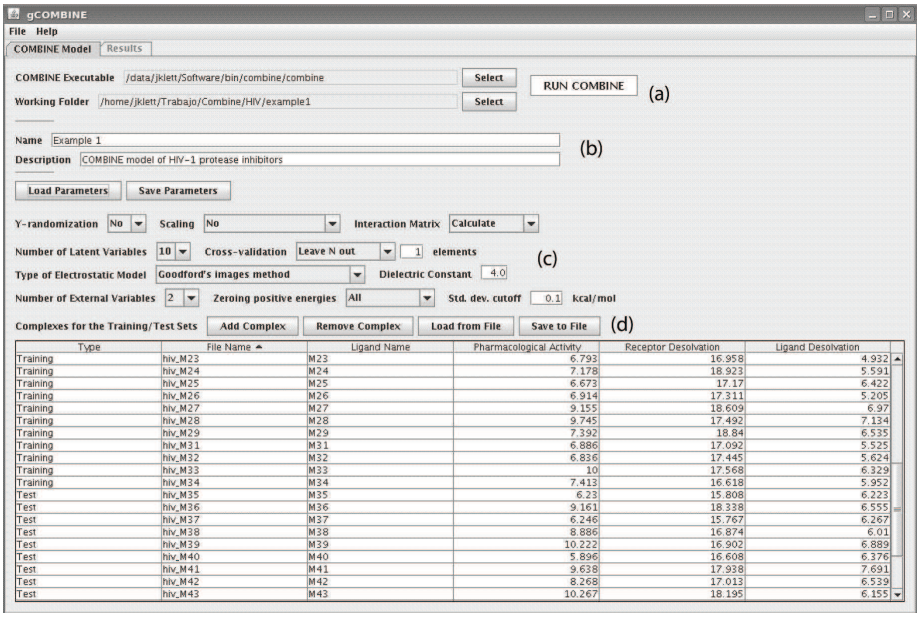
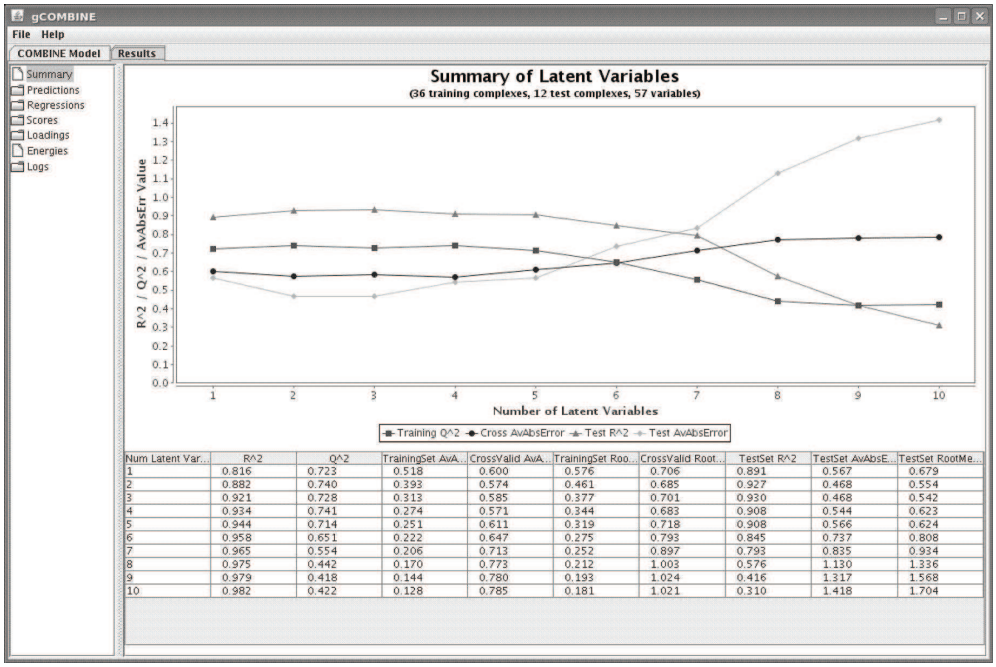


Figura 4-24. Ventana principal de *gCOMBINE*. Las letras a-d se refieren a los cuatro bloques principales de datos.

El submenú *File* contiene cuatro opciones: (a) *New Model*, para limpiar los datos de las pestañas, (b) *Load Model*, para cargar en las pestañas un modelo salvado previamente, (c) *Save Model*, para salvar el modelo actual en un fichero específico, y (d) *Exit*, que cierra la aplicación. En el submenú *Help* se ofrece información acerca de la publicación original de *COMBINE*, el autor principal, y la información de contacto. En la pestaña *COMBINE Model* se pueden apreciar cuatro áreas principales (a-d en la Figura 4-24): (a) parte superior donde el usuario puede seleccionar la ruta del ejecutable de *COMBINE* y la carpeta de trabajo donde los complejos receptor-ligando están almacenados. Pinchando en el botón *RUN COMBINE* comienza el cálculo del modelo; (b) sección donde el usuario puede introducir comentarios relacionados con el modelo, como son un nombre y una pequeña descripción; (c) esta sección permite al usuario cargar los parámetros desde un cálculo previo o salvar los parámetros actuales a través de los botones *Load*/*Save Parameters*. Todos los parámetros para configurar el modelo se introducen aquí: asignación aleatoria de actividades farmacológicas (No por defecto), escalado (No por defecto), matriz de interacción (puede ser calculada y escrita por *gCOMBINE* o bien puede ser leída desde un fichero externo), número de variables latentes (son extraídas 5 por defecto), método de validación cruzada (dejar *N* fuera o generar grupos aleatorios), tipo de modelo dieléctrico (constante dieléctrica uniforme, implementación de *Goodford* del método de las imágenes (Goodford, 1985), constante dieléctrica dependiente de la distancia, energías de interacción electrostáticas de *Poisson–Boltzmann* leídas de un fichero externo, y un modelo sigmoidal (Mehler & Solmajer, 1991)), valor de constante dieléctrica (4 por defecto), número de variables externas, pretratamiento de los datos y valor de corte para el pretratamiento; y (d) el área de los complejos receptor-ligando, donde el usuario puede añadir/borrar/cargar/salvar complejos desde/a un fichero (botones *Add*/*Remove Complex* y *Load*/*Save from*/*to File*). Bajo estos botones se encuentra una tabla con la siguiente información: el tipo asociado a cada complejo (de entrenamiento, de test o no usado, según defina el usuario) que puede modificarse en cualquier momento para probar modelos alternativos, el nombre del fichero del complejo, el nombre del ligando dentro de este fichero, su actividad farmacológica, y una columna para cada variable externa considerada. Una vez los parámetros y los complejos han sido proporcionados y se ha pulsado el botón *RUN COMBINE*, lo primero que hace la aplicación es comprobar si todos los parámetros necesarios han sido proporcionados, tienen valores válidos, y si existen los ficheros necesarios.



Num Latent Var...	R^2	Q^2	TrainingSet AvA...	CrossValid AvA...	TrainingSet Roo...	CrossValid Root...	TestSet R^2	TestSet AvAbsE...	TestSet RootMe...
1	0.816	0.723	0.518	0.600	0.576	0.706	0.891	0.567	0.679
2	0.882	0.740	0.393	0.574	0.461	0.685	0.927	0.468	0.554
3	0.921	0.728	0.313	0.585	0.377	0.701	0.930	0.468	0.542
4	0.934	0.741	0.274	0.571	0.344	0.683	0.908	0.544	0.623
5	0.944	0.714	0.251	0.611	0.319	0.718	0.908	0.566	0.624
6	0.958	0.651	0.222	0.647	0.275	0.793	0.845	0.737	0.808
7	0.965	0.554	0.206	0.713	0.252	0.897	0.793	0.835	0.934
8	0.975	0.442	0.170	0.773	0.212	1.003	0.576	1.130	1.336
9	0.979	0.418	0.144	0.780	0.193	1.024	0.416	1.317	1.568
10	0.982	0.422	0.128	0.785	0.181	1.021	0.310	1.418	1.704

Figura 4-25. Pestaña *Results* de *gCOMBINE* mostrando la evolución de los índices quimiométricos en modo gráfico y tabular.

El usuario puede detener la ejecución o esperar hasta que se haya completado. En este último caso, si no se han producido errores, la aplicación se coloca en la pestaña de resultados (ver Figura 4-25) donde el usuario tiene acceso a las diferentes tablas y gráficas agrupadas en nodos y subnodos: un resumen de cada modelo (un gráfico que muestra la evolución de los índices quimiométricos y una tabla conteniendo las valores numéricos de esos índices), una gráfica para los coeficientes *PLS* (ver Figura 4-26), predicciones (para cada modelo se muestra un gráfico de actividad experimental frente a actividad predicha, como puede apreciarse en el panel a de la Figura 4-27), gráficas de contribuciones (*loadings*) y puntuaciones (*scores*) (Figura 4-27, paneles b y c respectivamente), y una gráfica con las variables de energía de interacción de entrada al *PLS* descompuestas por residuo (muy útil para detectar posibles anomalías en los valores energéticos, ver panel d en la Figura 4-27). El usuario además puede interaccionar con las gráficas en diferentes modos: zoom para acercarse/alejarse, ver descripciones emergentes de los datos, establecer etiquetas, cambiar la apariencia, salvarla como imagen, imprimirla, etc. Por último, hay un nodo de *logs* para mantener los mensajes de salida de la ejecución de *COMBINE*.

La distribución de *gCOMBINE* y su manual pueden obtenerse a través de su página *web* (http://ub.cbm.uam.es/gCOMBINE).

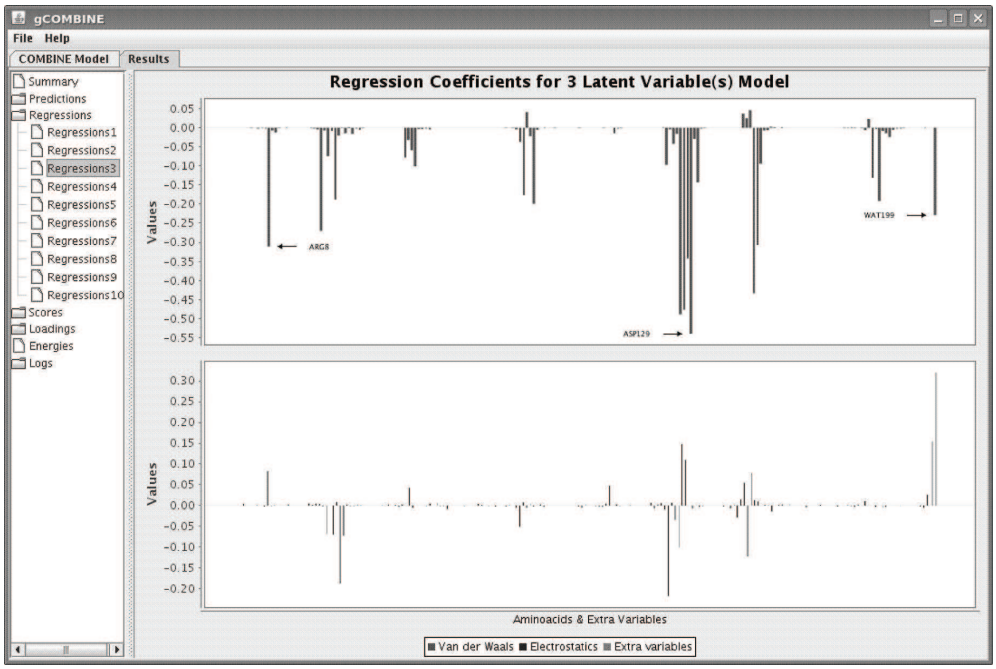


Figura 4-26. Gráficas mostrando los pesos asignados a los valores de energías de interacción de *van der Waals* y electrostática por residuo en un modelo *COMBINE* realizado con cuatro componentes principales para tener en cuenta las diferencias en actividad para la serie de inhibidores de la proteasa *HIV-1*.

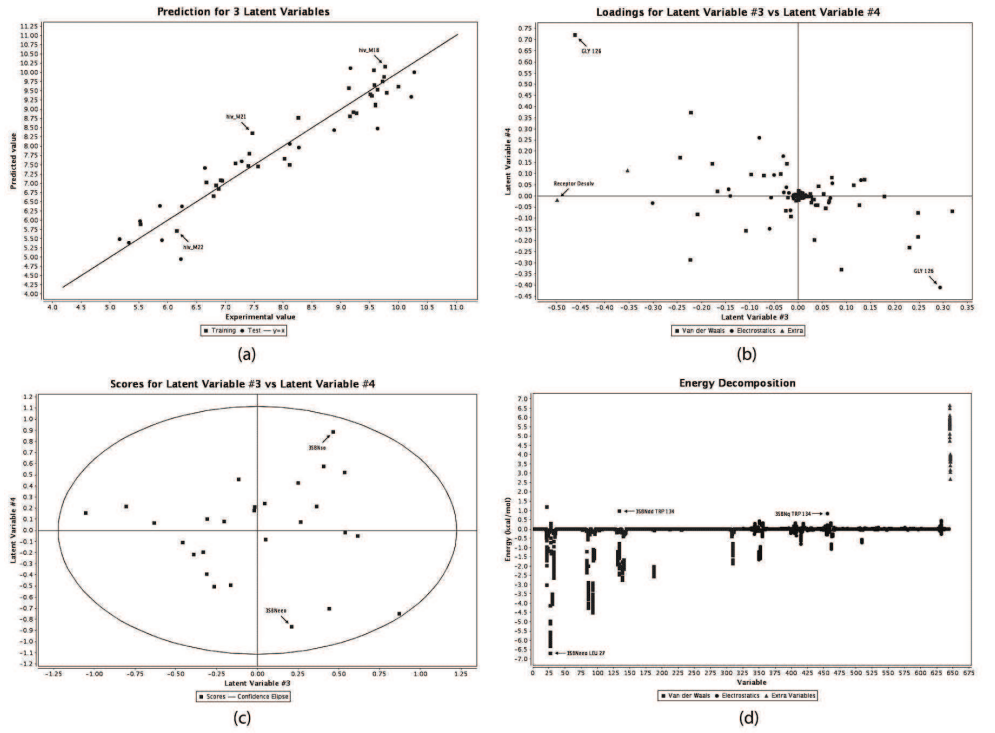


Figura 4-27. (a) Gráfica de la actividad experimental frente a la predicha, (b) gráfica mostrando las contribuciones (*loadings*) a los componentes principales de las variables originales, (c) gráfica de puntuaciones (el dominio de aplicación está encerrado en una elipse de confianza (Rocchia et al., 2002)), (d) gráfica de la descomposición de la energía de interacción receptor-ligando por residuo para las variables originales de entrada al análisis *PLS*.

4.4.2. COMBINE versus gCOMBINE

Para la comparación de los resultados obtenidos con *COMBINE* y *gCOMBINE* se han reproducido dos de los estudios en los que se obtuvieron modelos *COMBINE* exitosos (ver Materiales y Métodos, apartado 3.1.1.4 de la página 29).

El primero de estos estudios estuvo motivado principalmente por dos aspectos: (a) las energías de interacción ligando-receptor (tal y como fueron calculadas por los investigadores de *Merck* usando el campo de fuerzas *MM2X*) ya de por sí correlacionaban bastante bien (r^2 = 0.74 en el análisis de regresión) con los datos de inhibición determinados experimentalmente (valores de IC_{50}) (Holloway et al., 1995). Además, el poder predictivo de dicho modelo con un conjunto de prueba de 16 compuestos (no usados en la generación del modelo) fue también notable (q^2 = 0.75), con un error medio absoluto de 1 a través de un rango de 5 unidades logarítmicas, y (b) el hecho de que no hubiera mejora tras la incorporación de los efectos del solvente (mediante una descripción continua) o usando otro campo de fuerzas (*CHARMM* en su caso). Los principales objetivos de los análisis *COMBINE* realizados fueron: (i) buscar posibles dependencias de la correlación obtenida con el campo de fuerzas usado, y (ii) probar y desarrollar modelos *QSAR* más precisos. En este ejercicio, se demostró que se obtenían resultados similares usando el campo de fuerzas de *AMBER*, por lo que no parecía haber dependencias con el campo de fuerza (ver los índices quimiométricos para L_{MM2X} y L_{AMBER} en la Tabla 4-VI). Por otro lado, se lograron mejoras considerables mediante el uso de modelos *COMBINE*, especialmente cuando los efectos parciales de desolvatación para el ligando y el receptor tras la formación del complejo fueron incluidos usando una descripción continua (resolviendo la ecuación de *Poisson-Boltzmann*) y sustituyendo el término electrostático estándar basado en el coulómbico dependiente de distancia por valores de solvatación corregidos calculados para cada residuo. La principal conclusión de este trabajo fue que simplemente reemplazando el término coulómbico con la descripción continua del electrostático e incluyendo los efectos de desolvatación no producía una notable mejor cuando se usaba *MLR* (*Multiple Linear Regression*), pero en cambio sí que se producía cuando se hacía uso del correspondiente modelo *COMBINE* empleando una descripción continua del modelo de solvente. Aunque en el artículo original se incluyeron diferentes mejoras sobre el término coulómbico estándar, en las pruebas para *gCOMBINE* se utiliza solo el caso más simple (y a la vez más usado) con el objetivo de reproducir los datos previamente

publicados. En particular se demuestra la reproducibilidad del modelo llamado C_{AMBER}, donde las contribuciones de *van der Waals* y electrostáticas (termino Coulómbico usando una constante dieléctrica de 4) fueron tomadas directamente del campo de fuerzas de *AMBER* usando el módulo *ANAL*. Los resultados no pueden ser exactamente los mismos ya que la técnica de validación cruzada usa componentes aleatorios: los compuestos se asignan aleatoriamente a cinco grupos de aproximadamente el mismo tamaño, cada grupo es excluido del análisis por turnos, y el proceso total es repetido 20 veces. De todos modos, los resultados son claramente similares (ver los índices quimiométricos para C_{AMBER} y gC_{AMBER} en la Tabla 4-VI).

Modelo	Objetos	Variables	*LV*	r^2	q^2	$SDEP_{CV}$	$SDEP_{ex}$
L_{MM2X}	32	1	1	0.74	0.75	-	1.00
L_{AMBER}	32	1	1	0.81	0.79	0.61	1.08
C_{AMBER}	32	48	2	0.89	0.70	0.72	0.83
gC_{AMBER} [a]	32	48	2	0.89	0.70	0.72	0.80

Tabla 4-VI. Índices quimiométricos para los diferentes modelos en el estudio de inhibidores de la proteasa del *HIV-1*. [a] Calculado con *gCOMBINE*

La segunda publicación reproducida trató el estudio de 27 inhibidores de *HIV-1 RT* no nucleosídicos (*NNRTI*). En este caso se disponía de una gran cantidad de datos sobre la actividad en enzimas *RT* con diferentes mutaciones en el sitio de unión de los *NNRTI*. Se generaron modelos *COMBINE* para cuantificar la relación estructura-actividad y posiblemente también los efectos de las mutaciones. Las energías de interacción de *van der Waals* entre el ligando y los residuos del sitio activo se calcularon usando los parámetros de *AMBER* (*parm99*), mientras que la parte electrostática para la corrección de la solvatación se obtuvo resolviendo la ecuación de *Poisson-Boltzmann* implementada en *DelPhi*. Se incluyeron como variables externas los cambios en la desolvatación para receptor y ligando tras la formación del complejo. Para llevar a cabo la prueba de *gCOMBINE*, estos cálculos electrostáticos se han realizado ejecutando *DelPhi* por separado y después cargando los resultados en la aplicación desde el menú *Type of Dielectric Model* (ver el bloque c de la figura Figura 4-24 de la página 117) y seleccionando la opción *Poisson-Boltzmann from .dph files*. De esta manera, los resultados previamente publicados son reproducidos con precisión por *gCOMBINE* obteniéndose sólo pequeñas variaciones, como es de esperar cuando se emplea la validación cruzada por formación de grupos aleatorios (ver la Tabla 4-VII). Algunos de los resultados más relevantes se muestran en la Figura 4-27 de la página 120: en el panel (a) se tiene la correlación entre la actividad experimental y la predicha

para los 27 inhibidores *NNRTI*, las contribuciones y las puntuaciones en los paneles (b) y (c) respectivamente, y las variables de energías de interacción que sirvieron de entrada para el análisis *PLS* en el panel (d). En la Figura 4-26 de la página 120 se muestran los coeficientes de regresión. Finalmente, la evolución de r^2, q^2 y *SDEP* (ver Tabla 4-VII) para el conjunto de los 27 inhibidores *NNRTI* según se incrementa el número de componentes principales extraídas se muestra en la Figura 4-25 de la página 119.

LV	r^2	q^2	$SDEP_{CV}$
1	0.84	0.71	0.62
2	0.88	0.78	0.55
3	0.93	0.82	0.49
4	0.94	0.86	0.44
5	0.95	0.86	0.44

Tabla 4-VII. Índices quimiométricos calculados por *gCOMBINE* para el conjunto completo de 27 *NNRTI*.

Viendo los resultados obtenidos para los dos estudios reproducidos, esta claro que *gCOMBINE* es capaz de reproducir fielmente los resultados de un modo más simple y cómodo.

4.5. Plataforma de Cribado Virtual con Diferentes Protocolos

Los conjuntos de datos utilizados en las pruebas de la plataforma con las proteínas *fXa*, *AChE*, *ERa*, *CDK2*, neuraminidasa y *p38MAP* han sido descritos en el apartado 3.1.1.5 de la página 29. Los diferentes protocolos utilizados se resumen en la Tabla 3-X de la página 85. A continuación se muestran las tablas con los resultados obtenidos (Tabla 4-VIII, Tabla 4-IX y Tabla 4-X), en las que los datos más relevantes son el área bajo la curva (*AUC*) obtenida en las diferentes curvas *ROC*, los factores de enriquecimiento (tanto el mejor, EF_{best}, como el máximo que podría obtenerse teóricamente, EF_{max}), y el tiempo empleado por molécula (segundos en escala logarítmica, *log t*). Cada tipo de protocolo usado se identifica con un número del 1 al 11 (*ID VSP*, *Virtual Screening Protocol*). Los valores entre paréntesis en la columna de factores de enriquecimiento se refieren al porcentaje (en forma decimal) de la base de datos analizados para alcanzar el mejor factor de enriquecimiento (EF_{best}).

fXa									
	Conjunto *Fontaine*			Conjunto *Jacobsson*			Conjunto *Jorissen-Gilson*		
ID VSP	*AUC*	EF_{best} (EF_{max} = 2.16)	*log t*	*AUC*	EF_{best} (EF_{max} = 4.94)	*log t*	*AUC*	EF_{best} (EF_{max} = 11)	*log t*
1	0.39	1.00 (0.99)	1.30	0.41	1.00 (0.99)	0.99	0.36	1.00 (1.00)	0.93
2	0.37	1.00 (1.00)	1.32	0.41	1.01 (0.96)	0.91	0.39	1.02 (0.98)	0.93
3	0.33	1.00 (1.00)	1.34	0.33	0.98 (0.99)	1.05	0.31	1.00 (1.00)	0.99
4	0.67	2.16 (0.01)	3.62	0.61	1.65 (0.05)	3.42	0.68	3.30 (0.04)	3.38
5	0.70	1.94 (0.01)	3.62	0.70	1.97 (0.05)	3.42	0.68	3.30 (0.02)	3.38
6	0.57	1.51 (0.14)	3.72	0.52	1.23 (0.22)	3.55	0.60	1.80 (0.20)	3.51
7	0.54	1.51 (0.01)	1.43	0.55	1.73 (0.06)	1.28	0.58	3.30 (0.02)	1.26
8	0.59	1.44 (0.03)	2.19	0.55	1.32 (0.05)	-	0.65	2.57 (0.05)	1.99
9	0.58	1.58 (0.03)	1.76	0.56	1.38 (0.32)	1.53	0.61	2.93 (0.05)	1.46
10	0.64	1.43 (0.09)	2.58	0.57	1.29 (0.33)	2.37	0.65	2.48 (0.07)	2.31
11	0.54	1.25 (0.16)	2.51	0.54	1.48 (0.02)	2.29	0.58	1.83 (0.05)	2.29

Tabla 4-VIII. Resultados de los cribados para *fXa*.

	AChE				*ERa*				
	Conjunto de *Jacobsson*			Conjunto de *Jacobsson*			Conjunto de *Stahl*		
ID VSP	*AUC*	EF_{best} (EF_{max} = 19.52)	*log t*	*AUC*	EF_{best} (EF_{max} = 8.04)	*log t*	*AUC*	EF_{best} (EF_{max} = 21)	*log t*
1	0.30	1.08 (0.86)	0.84	0.56	2.01 (0.01)	0.90	0.28	1.00 (1.00)	1.24
2	0.32	1.02 (0.98)	0.85	0.60	3.35 (0.04)	0.91	0.57	1.73 (0.22)	1.23
3	0.26	1.02 (0.98)	1.17	0.36	1.03 (0.97)	0.99	0.17	1.00 (1.00)	1.29
4	0.97	15.97 (0.01)	3.09	0.35	1.00 (1.00)	3.02	0.55	11.45 (0.01)	3.13
5	0.94	9.58 (0.05)	3.09	0.33	0.99 (1.00)	3.02	0.52	7.64 (0.01)	3.13
6	0.93	15.97 (0.01)	3.46	0.47	0.40 (0.05)	-	0.64	5.73 (0.01)	3.32
7	0.49	5.32 (0.01)	1.08	0.43	0.31 (0.14)	1.15	0.63	13.36 (0.01)	1.35
8	0.69	12.42 (0.01)	-	0.42	0.58 (0.23)	1.74	0.66	17.18 (0.01)	1.80
9	0.71	14.20 (0.01)	1.25	0.43	0.78 (0.25)	-	0.65	7.64 (0.01)	1.46
10	0.93	15.97 (0.01)	-	0.45	0.90 (0.45)	-	0.71	10.5 (0.02)	2.09
11	0.49	1.77 (0.01)	2.75	0.43	0.30 (0.12)	2.44	0.63	11.45 (0.01)	2.47

Tabla 4-IX. Resultados de los cribados para *AChE* y *ERa*.

	CDK2			Neuraminidasa			*p38MAP*		
	Conjunto de *Jorissen-Gilson*			Conjunto de *Stahl*			Conjunto de *Stahl*		
ID VSP	*AUC*	EF_{best} (EF_{max} = 21)	*log t*	*AUC*	EF_{best} (EF_{max} = 59.82)	*log t*	*AUC*	EF_{best} (EF_{max} = 46.45)	*log t*
1	0.36	0.97 (0.98)	1.20	0.35	1.09 (0.87)	1.26	0.23	0.96 (0.99)	0.95
2	0.41	1.91 (0.01)	1.18	0.56	1.36 (0.39)	0.37	0.42	1.17 (0.31)	1.00
3	0.41	1.02 (0.98)	1.27	0.20	1.00 (1.00)	1.35	0.26	1.03 (0.97)	1.10
4	0.67	2.45 (0.07)	3.02	0.89	10.88 (0.01)	3.25	0.64	4.22 (0.02)	-
5	0.63	2.12 (0.09)	3.02	0.42	1.06 (0.78)	3.25	0.60	4.22 (0.01)	-
6	0.57	2.06 (0.14)	3.33	0.72	6.22 (0.08)	3.24	0.55	1.92 (0.12)	-
7	0.55	1.91 (0.13)	1.28	0.55	10.88 (0.01)	1.40	0.65	4.75 (0.09)	1.36
8	0.63	3.82 (0.01)	1.67	0.63	5.44 (0.01)	1.96	0.73	4.22 (0.03)	2.02
9	0.59	2.67 (0.05)	1.40	0.69	10.88 (0.01)	1.56	0.72	4.22 (0.01)	1.57
10	0.61	2.10 (0.10)	1.95	0.82	13.6 (0.02)	-	0.75	4.22 (0.01)	-
11	0.55	1.91 (0.02)	2.54	0.53	1.81 (0.03)	2.56	0.64	3.75 (0.10)	2.60

Tabla 4-X. Resultados de los cribados para *CDK2*, neuraminidasa, *p38MAP*.

En los siguientes apartados se comentan los resultados obtenidos en relación a diferentes aspectos.

4.5.1. Rendimiento de las Herramientas de Docking por Separado

El rendimiento de *DOCK* es realmente pobre en todos los casos excepto uno (el conjunto de *Jacobsson* para *ERa*), donde es ligeramente mejor que la selección aleatoria. Los resultados obtenidos con la función de puntuación de campo de fuerzas (*VSP2*) son siempre mejores que aquellos obtenidos con la función de puntuación basada en contactos (*VSP1*), pero la diferencia no es significativa, teniendo el mejor un área bajo la curva de 0.3. Este es también el rango de variación en valores de *AUC* para

diferentes proteínas con ambos métodos de puntuación. Los factores de enriquecimiento son muy bajos y además están muy alejados del factor de enriquecimiento máximo en todos los casos, no apreciándose diferencias entre ambas funciones. Los mejores factores de enriquecimiento se obtienen en el conjunto de *Jacobsson* para *ERa* (EF_{best} = 3.35). En la mayoría de los casos, el *AUC* obtenido con *CDOCK* está por encima de 0.6, con la excepción de *ERa*, para el cuál los resultados son algo peores que en el caso aleatorio (conjunto de *Jacobsson*) o ligeramente mejores (conjunto de *Stahl*). En todos los casos, excepto aquellos para los conjuntos de *fXa*, la selección de compuestos basada sólo en las interacciones de *van der Waals* (*VSP5*) conduce a valores de *AUC* más pequeños, aunque parece que no es un efecto excesivamente importante. Una excepción notable es el ejemplo de neuraminidasa, donde el término de energía electrostática mejora en 0.47 el *AUC*. Son significativas las variaciones en los valores de *AUC* para diferentes proteínas (alrededor de 0.6 unidades) pero parecen independientes de la función de puntuación utilizada (*van der Waals* con electrostático o *van der Waals* solo). Para *CDOCK* utilizando conjuntamente *van der Waals* más electrostático como función de puntuación (*VSP4*), los factores de enriquecimiento son siempre mejores que usando *DOCK* (con cualquiera de sus funciones) y por encima del 10% del EF_{max}, alcanzando este valor en un caso (conjunto de *Fontaine* para *fXa*), 80% (*AChE*), y 55% (conjunto de *Stahl* para *ERa*). Si sólo se utiliza el término de *van der Waals* (*VSP5*), el factor de enriquecimiento se reduce en gran medida, siendo este decrecimiento incluso mayor en *AChE* (30%), *ERa* (conjunto de *Stahl*, 20%), y neuraminidasa (16%).

4.5.2. Rendimiento con Herramientas de Docking Conjuntas

4.5.2.1. Considerando Diferentes ZScores

Se realiza un filtro mediante *DOCK*, seleccionando aquellas moléculas con un *ZScore* superior a 3.0, para después emplear *CDOCK* (*VSDP7*). De este modo se obtienen valores que son muy próximos a los que se obtendrían en una selección aleatoria. No se aprecian diferencias significativas cuando se emplea un *ZScore* más restrictivo (1.5 en lugar de 3.0, como en el *VSP3*). Las dos excepciones son los casos de *AChE* y neuraminidasa. En el primero, se pasa de un valor de *AUC* casi aleatorio (0.49) a un valor bastante bueno de 0.71, es decir, una diferencia de 0.22 unidades. En el segundo el incremento es de 0.14 unidades, alcanzando un valor de 0.70 en *AUC*. Las

diferencias entre proteínas son del mismo orden (0.2 para *VSP7* y 0.3 para *VSP9*). Sólo en dos casos el factor de enriquecimiento está por encima del 50% del EF_{max} (64% en el conjunto de *Stahl* para *ERa*, y 70% en el de *Fontaine* para *fXa*), mientras que en resto el rango varía de 4% a 35%. Se observan pequeñas variaciones en los factores de enriquecimiento para la mayoría de los conjuntos cuando se usa el *ZScore* más pequeño. Estas variaciones son más importantes en *AChE* (incremento del 18%) y *ERa* (en el conjunto de *Stahl*, 28% de decremento).

4.5.2.2. Seleccionando 10 (en lugar de 1) conformaciones para cada ligando de DOCK

Cuando el valor del *ZScore* es 3.0 (*VSP8*) los valores de *AUC* están siempre por encima de 0.5, siendo el ejemplo de *ERa* (conjunto de *Jacobsson*) la excepción. Las variaciones a través de diferentes proteínas son del orden de 0.3 unidades de *AUC*. Para un valor *ZScore* de 1.5 (*VSP10*) se obtienen los mejores resultados. En general, son comunes los valores de *AUC* por encima de 0.6, de nuevo excepto para el caso de *ERa* (conjunto de *Jacobsson*). Aquí, las variaciones entre diferentes proteínas llegan a ser incluso de 0.5 unidades de *AUC*. Cuando el *ZScore* es 3.0, tres casos presentan factores de enriquecimiento por encima del 50% del enriquecimiento máximo, mientras que los otros varían entre 7% y 27%. Con el valor de *ZScore* más pequeño (1.5), las mayores diferencias se observan en *AChE* (incremento del 18%), neuraminidasa (incremento del 14%), y *ERa* (en el conjunto de *Stahl*, un decremento del 32%).

4.5.3. Inclusión del Efecto del Solvente: PBSA como Término de Corrección

Las energías de interacción basadas en mecánica molecular (*van der Waals* y coulómbico de *CDOCK*) son corregidas para los efectos de la desolvatación utilizando el método *PBSA* (Kollman et al., 2000) directamente sobre los resultados obtenidos de *CDOCK* (*VSP6*) o bien tras un protocolo combinado que incluya *DOCK* y *CDOCK* (*VSP11*). En el primer caso, los *AUCs* obtenidos están por encima de la curva aleatoria en la mayoría de los casos. Se obtienen los mejores *AUCs* para las proteínas *AChE* y neuraminidasa, mientras que *ERa* (conjunto de *Jacobsson*) es el único caso con valor de *AUC* peor que el aleatorio. En este caso, la introducción de los efectos del solvente a través de *PBSA* no conduce a una mejora sobre los resultados de *CDOCK* con *van der Waals* y coulómbico. Cinco de los conjuntos muestran factores de enriquecimiento

superiores al 30%, con *fXa* (en el conjunto de *Fontaine*) logrando el EF_{max}. Excepto para *AChE*, la inclusión de la desolvatación siempre reduce los factores de enriquecimiento. El rango de la reducción varía entre prácticamente inapreciable (2% en *ERa* para el conjunto de *Jacobsson*, y 5% en p38) a notorio (30% en *fXa* para el conjunto de *Fontaine*, 28% en *ERa* para el conjunto de *Stahl*). En el segundo caso, los valores de *AUC* permanecen casi sin alteración comparados con la situación en la cual no se introducen los efectos del solvente. De nuevo, el rendimiento en *ERa* (conjunto de *Jacobsson*) es peor que aleatorio. En términos de variación a través de las diferentes proteínas, se obtiene un valor de 0.46 unidades de *AUC* cuando se aplica el método *PBSA* a *CDOCK* sin el filtro previo de *DOCK*, y 0.21 con éste. Los factores de enriquecimiento obtenidos son similares a los comentados anteriormente, pero solo cuatro conjuntos (en lugar de cinco) alcanzan una cifra mayor al 30%, y ninguno alcanza el enriquecimiento máximo. Igualmente, se observa una reducción general en estos valores cuando la desolvatación se tiene en cuenta, con variaciones que van desde el 2% (*p38*) hasta el 18% (*AChE*).

Finalmente, para alcanzar el máximo factor de enriquecimiento es necesario explorar casi toda la base de datos cuando se utiliza DOCK. Hay, sin embargo, algunas excepciones (6 de 27): *VSP2* para *ERa* (conjuntos de *Jacobsson* y *Stahl*), *CDK2*, neuraminidasa y *p38MAP*; y *VSP1* para *ERa* (conjunto de *Jacobsson*). Para el resto de los protocolos, el mejor factor de enriquecimiento se alcanza muy pronto en la mayoría de los casos, excepto en aquellos en los cuales se tiene que explorar el 25% de la base de datos para lograrlo: *ERa* (conjunto de *Jacobsson*, *VSP9* y *VSP10*), neuraminidasa (*VSP5*), y *fXa* (conjunto de *Jacobsson*, *VSP9* y *VSP10*).

4.5.4. Rendimiento General frente a Tiempo de Cálculo

En la Figura 4-28 se muestra, a modo de referencia, los *AUCs* obtenidos para todos los protocolos aplicados a los conjuntos que utilizan *fXa* y neuraminidasa. En la Figura 4-29 se muestran los tiempos empleados por molécula (segundos en escala logarítmica).

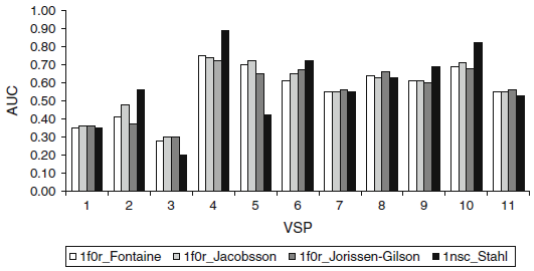


Figura 4-28. Área bajo la curva (*AUC*) para cada uno de los protocolos de cribado empleados (*VSP*) en los tres conjuntos de datos de *fXa* y neuraminidasa.

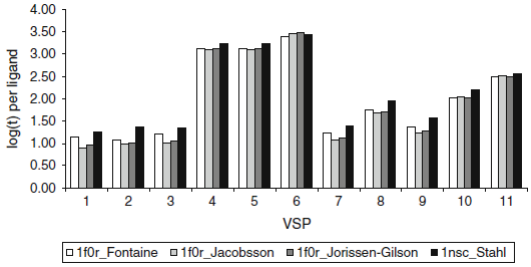


Figura 4-29. Tiempo de *CPU* (segundos, en escala logarítmica) por ligando (*log(t) per ligand*) requerido para cada protocolo de cribado virtual (*VSP*) en los tres conjuntos de *fXa* y neuraminidasa.

Como es de esperar, se obtienen mejores resultados cuando el cribado se realiza directamente con *CDOCK* (ver comparación de *VSP4-VSP6* con *VSP1-VSP3*) debido a su mejor función de puntuación y a la búsqueda exhaustiva que se realiza en el centro activo. Pero sin embargo, resulta más interesante comprobar que cuando *DOCK* se utiliza como filtro previo a *CDOCK* (*VSP7-VSP11*) se obtienen resultados bastante razonables, lo que indica que el filtro resulta aconsejable a la hora de eliminar estructuras no deseadas de ligandos. Por otro lado, como puede apreciarse en la Figura 4-29, el protocolo *VSP4* es uno de los que más tiempo consume, aunque es el que mejores resultados obtiene (ver Figura 4-28). Comparando *VSP10* con *VSP4* se aprecia que los resultados no difieren mucho en cuanto a sus *AUCs*, incrementándose el tiempo de computación en 1 ó 2 órdenes de magnitud en el segundo respecto al primero. Por lo tanto, el uso de un filtro ahorra tiempo de cálculo manteniendo aproximadamente los mismos valores de *AUC*. A modo de ejemplo, en la Figura 4-30 se muestra una curva *ROC* que ilustra este efecto.

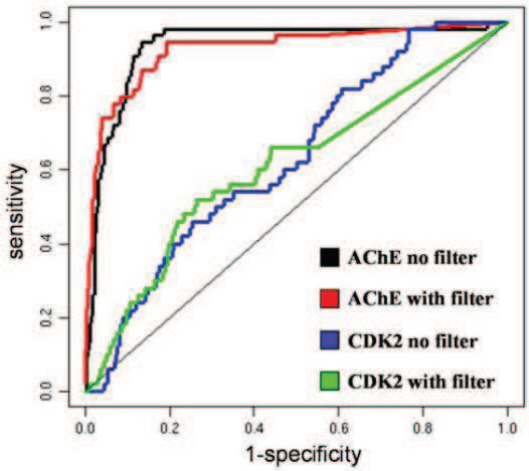


Figura 4-30. Curva *ROC* para los conjuntos de *AChE* y *CDK2* usando los protocolos *VSP4* (sin filtro de *DOCK*) y *VSP10* (con filtro de *DOCK*).

4.6. Casos de Estudio

A continuación se describen los resultados obtenidos en los estudios reales de cribado virtual realizados.

4.6.1. MGMT

Como resultado de la preparación de proteína *MGMT* según el protocolo estándar desarrollado en esta Tesis, se obtuvo una caracterización de la forma del centro activo mediante una imagen negativa de éste con el programa *GAGA* (ver Figura 3-15 en la página 87). Puede verse cómo cubre el bolsillo y además se extiende hasta la vecindad del residuo *Tyr114*, que juega un papel clave en la rotación del nucleótido dañado que va a ser reparado. Además, la forma de esta imagen negativa imita a la del nucleótido cuando se une a *MGMT*. Usando las esferas que componen esta imagen se realiza el pre-filtrado con el programa *DOCK*, asignando así una puntuación a cada molécula en función al grado de complementariedad que tiene con la forma del sitio activo. Del conjunto inicial de 2.3 millones de moléculas, 1664 pasaron el valor de corte (un *ZScore* de 5.0). En la Figura 4-31 se muestra un histograma con la distribución de puntuaciones (*DOCK Scores*).

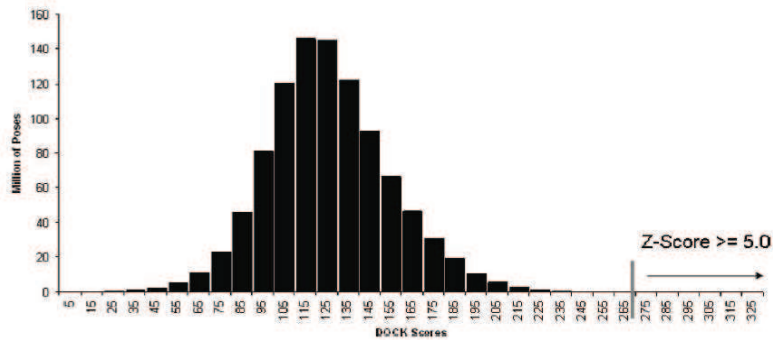


Figura 4-31. Histograma de la distribución de puntuaciones (*DOCK Scores*) obtenidas en el pre-filtrado para el cribado sobre la proteína *MGMT*. El valor de corte utilizado para la selección (*Z-Score*) es de 5.0.

Tras la etapa de filtrado (*CDOCK*, *DelPhi* y *Apolar*, ver apartado 3.2.7.7.1 de la página 86) e inspección visual, se seleccionaron 17 compuestos que fueron comprados y probados experimentalmente. De este conjunto, 4 resultaron activos para *MGMT* (ver Figura 4-32) en el rango micromolar y fueron estudiados mediante dinámica molecular (ver etapa de Optimización del protocolo estándar de cribado) con el fin de analizar sus modos de unión. Estos cuatro compuestos se clasifican en dos familias según su estructura química: derivados de quinolina (1 y 2) y de triazolpirimidina (3 y 4).

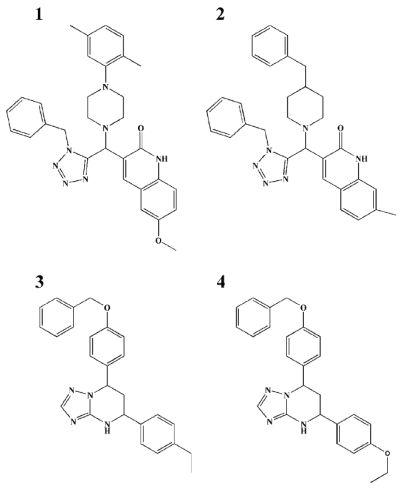


Figura 4-32. Estructura química de los cuatro compuestos que muestran inhibición de *MGMT* en el rango micromolar.

En la Tabla 4-XI se muestran los 17 compuestos, con las energías calculadas tras la etapa de filtrado, los resultados de las energías libres de unión calculadas tras la optimización de los 4 compuestos activos, junto con algunas propiedades almacenadas en la base de datos de *ZINC* para cada ligando. Además, también se realizó un análisis de la energía de interacción entre el ligando y los residuos más relevantes en el sitio de unión (ver Tabla 4-XII).

Compuesto (Código ZINC)	logP	Don.	Acc.	Carga	Peso Mol.	Energía Filtrado	Energía Optimización	IC_{50} *in vitro* (µM)	IC_{50} *in vivo* (µM)
1 (ZINC00910802)	3.52	2	9	1	536	-34.57	-32.26 (2.59)	54	10
2 (ZINC00889422)	4.24	2	7	1	505	-31.91	-43.54 (3.35)	34	50
3 (ZINC03642335)	6.18	1	5	0	410	-32.42	-46.52 (3.26)	24	10
4 (ZINC02487935)	5.61	1	6	0	426	-32.37	-56.90 (3.24)	22	10
5 (ZINC01327643)	4.23	2	6	0	437	-31.24	-	> 100	-
6 (ZINC00714917)	6.41	1	6	0	503	-31.84	-	> 100	-
7 (ZINC01360953)	5.01	0	7	0	563	-32.06	-	> 100	-
8 (ZINC01360953)	4.51	2	6	0	433	-31.73	-	> 100	-
9 (ZINC02809317)	1.34	0	11	0	463	-32.95	-	> 100	-
10 (ZINC03404767)	4.88	1	8	0	481	-34.08	-	> 100	-
11 (ZINC01437200)	2.81	2	8	2	479	-32.71	-	> 100	-
12 (ZINC03052303)	3.50	2	7	0	516	-33.90	-	> 100	-
13 (ZINC00784955)	0.90	3	9	2	452	-32.61	-	> 100	-
14 (ZINC00892609)	4.45	2	7	1	505	-31.56	-	> 100	-
15 (ZINC01352201)	3.06	1	9	0	472	-33.19	-	> 100	-
16 (ZINC02835223)	4.34	1	6	0	433	-28.35	-	> 100	-
17 (ZINC00738815)	4.48	2	6	0	460	-32.52	-	> 100	-

Tabla 4-XI. Lista de los 17 compuestos obtenidos en el cribado virtual para *MGMT*. Se muestran diversas propiedades químicas obtenidas de la base de datos de *ZINC*: coeficiente de partición octanol/agua (*logP*), átomos donadores enlaces por puente de hidrógeno (*Don.*), átomos aceptores de enlace por puente de hidrógeno (*Acc.*), carga global y peso molecular. También se muestran en *kcal/mol* las energías calculadas en el filtro y en la optimización (sólo realizada para los compuestos activos). Esta última con la desviación estándar entre paréntesis ya que se trata de un promedio durante la simulación por dinámica molecular. Las dos últimas columnas son las actividades (IC_{50}) *in vitro* y las actividades *in vivo* (sólo para aquellos compuestos activos *in vitro*).

Residuo	Compuesto			
	1	**2**	**3**	**4**
ARG128	-5.99 (0.65)	-4.09 (1.51)	-6.01 (0.99)	-6.71 (0.64)
TYR114	-1.92 (0.31)	-4.33 (0.47)	-5.46 (0.65)	-4.87 (0.46)
ARG135	-5.52 (0.77)	-1.39 (0.56)	-3.51 (0.84)	-4.75 (0.95)
TYR158	-1.53 (0.44)	-3.17 (0.41)	-1.39 (0.24)	-4.03 (0.55)
GLY131	-	-2.31 (0.44)	-2.83 (0.36)	-3.17 (0.38)
ASN157	-2.69 (0.65)	-3.34 (0.42)	-1.38 (0.27)	-2.93 (0.52)
MET134	-	-3.11 (0.51)	-2.25 (0.34)	-2.58 (0.41)
ALA127	-1.11 (0.24)	-	-	-1.68 (0.29)
SER159	-1.30 (0.29)	-1.65 (0.35)	-	-1.55 (0.29)
GLN115	-	-	-2.36 (0.62)	-1.37 (0.33)
CYS150	-	-	-	-1.18 (0.46)
CYS145	-	-1.21 (0.32)	-	-1.01 (0.42)
Total	**-20.06 (0.48)**	**-24.60 (0.54)**	**-25.19 (0.51)**	**-35.83 (0.48)**

Tabla 4-XII. Análisis de las energías de interacción (en *kcal/mol*) por residuo para los inhibidores de *MGMT*. Son promedios calculados con el método *MM-GBSA* a partir de las simulaciones de dinámica molecular. Entre paréntesis se muestra la desviación estándar.

En los siguientes apartados se describen los resultados obtenidos en los ensayos experimentales, así como el análisis de los modos de unión obtenidos tras la etapa de optimización.

4.6.1.1. Ensayos Experimentales

Estos ensayos fueron realizados en el CNIO (Centro Nacional de Investigaciones Oncológicas) por la Dra. Carme Fábrega y el Dr. Federico M. Ruiz del antiguo Grupo de Transducción de Señales que estaba dirigido por el Dr. Jerónimo Bravo. Se hicieron tanto ensayos *in vitro* como ensayos *in vivo*. Los detalles de su realización pueden

consultarse en el artículo original (Ruiz et al., 2008). A continuación se muestran los resultados más relevantes de estas pruebas. Tras los ensayos *in vitro*, se comprobó que 4 de los 17 compuestos mostraban una actividad inhibitoria significativa (ver valores de IC_{50} en Tabla 4-XI), en el rango micromolar. La Figura 4-33 presenta una gráfica con las diferentes curvas obtenidas para el grado de inhibición de *MGMT* con cada uno de estos cuatro compuestos a diferentes concentraciones.

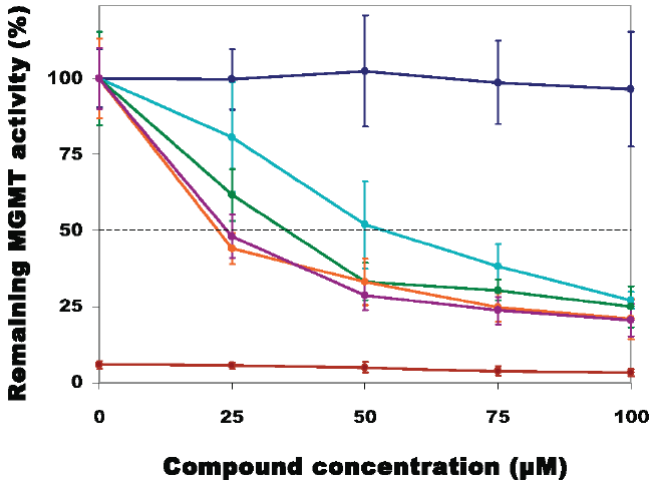


Figura 4-33. Curva de concentración mostrando el grado de inactivación de la proteína *MGMT* para los compuestos 1 (cian), 2 (verde), 3 (naranja) y 4 (violeta). En marrón se muestra el control negativo, y en azul el efecto del solvente (*DMSO*) sobre la actividad de *MGMT*. La línea punteada marca el nivel de 50% de actividad de *MGMT*.

Como puede apreciarse, los cuatro compuestos son capaces de inactivar *MGMT* en el rango bajo-medio micromolar. También se ve como el solvente no afecta significativamente en la actividad. Tampoco se aprecia modificación de la actividad al utilizar un mutante de *MGMT* inactivo (control negativo). Las dos familias de compuestos activos también muestran una ligera diferencia en su actividad (ver Tabla 4-XI): 54 y 34 µM para los derivados de quinolina, y 24 y 22 µM para los derivados de triazolpirimidina.

Los ensayos *in vivo* se realizaron sólo con estos cuatro compuestos. Los resultados se muestran en la Figura 4-34.

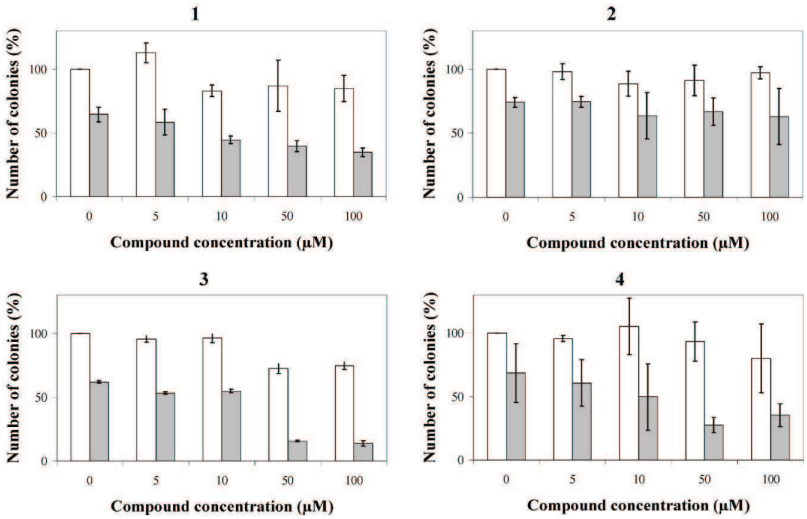


Figura 4-34. Efecto de los diferentes compuestos y a diferentes concentraciones sobre la supervivencia de células solas (barras blancas) o tratadas con el agente quimioterapéutico *BCNU* (barras grises).

Se aprecia como los cuatro compuestos son capaces de mejorar la citotoxicidad del agente quimioterapéutico *BCNU*. Dicho agente por sí solo es capaz de reducir el número de colonias en un 30%, pero aumenta hasta el 50% al añadir una concentración de 10 μM de los compuestos 1, 3 ó 4. En cambio el compuesto 2 necesita una concentración de 50 μM para obtener el mismo efecto, a pesar de ser uno de los mejores compuestos *in vitro*. Esto quizá es debido a una pobre penetración celular para el compuesto 2. Y como era de esperar tras las pruebas *in vitro*, los compuestos 3 y 4 siguen siendo mejores que el compuesto 1 en las pruebas in vivo. Aquí también influye que los dos compuestos derivados de triazolpirimidina tienen unos valores de coeficiente de partición octanol/agua (*logP*) mayores. Observando la Figura 4-34, se aprecia también que en general los cuatro compuestos no resultan tóxicos por sí solos, por lo que se considera que la muerte celular producida es debida a la acción conjunta de *BCNU* y los compuestos estudiados.

4.6.1.2. Descripción de los Modos de Unión Teóricos

La predicción de las interacciones y de los modos de unión para los cuatro compuestos activos se muestra en la Figura 4-35. Se trata de las estructuras promedio minimizadas de las simulaciones de dinámica molecular realizadas tras el *docking*.

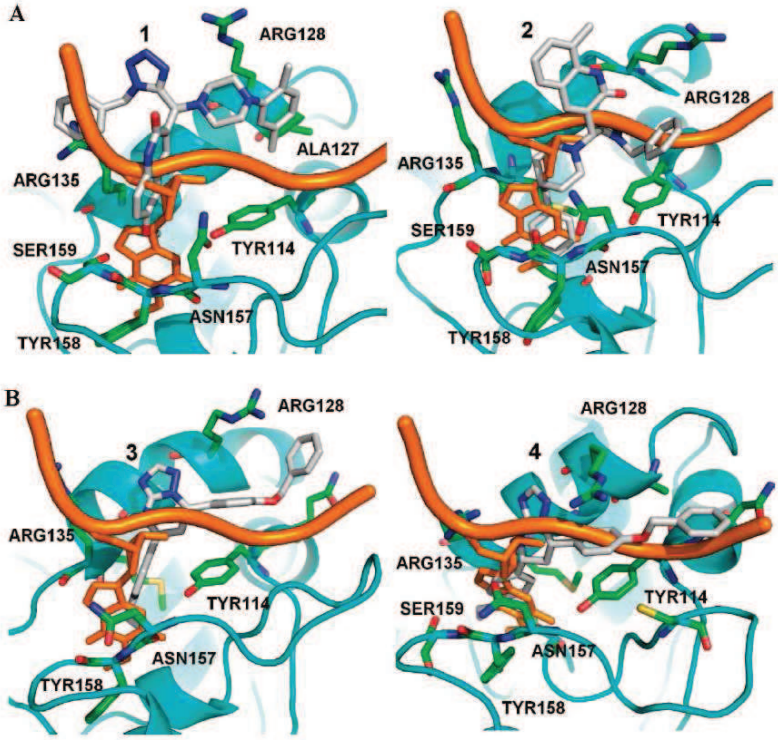


Figura 4-35. Estructuras promedio minimizadas de los complejos compuesto-*MGMT* tras las simulaciones de dinámica molecular. En naranja se muestra superpuesta la estructura del nucleótido girado. La proteína está representada en cintas 3D color cian; las cadenas laterales de los principales residuos en la interacción están representados en varillas y coloreados por tipos de átomos: carbono en verde, nitrógeno en azul, oxígeno en rojo, y azufre en amarillo. Los compuestos 1-4 están con los carbonos en gris, y se han omitido los átomos hidrógenos por claridad. Las letras A y B corresponden a las dos familias de compuestos.

Los modos de unión predichos sugieren que la conformación unida del inhibidor imita la observada del nucleótido al ser reparado por *MGMT*. Así pues, tanto el fragmento de quinolina en la familia 1 (Figura 4-35A) como el fragmento de triazolpirimidina de la familia 2 (Figura 4-35B) ocupan el surco catalítico de *MGMT*, haciendo el rol de análogos del motivo O^6-guanina del sustrato natural, que profundiza en el hueco y reacciona con el residuo catalítico *Cys145*. En ambos casos, se predice que la región restante de ambos inhibidores sale del hueco catalítico y ocupa la posición cercana a los residuos *Arg135* y *Tyr114*. En la Tabla 4-XII se muestra un resumen de las interacciones más importantes para cada uno de los cuatro inhibidores, tal como se calculó con el método *MM-GBSA* en base a las simulaciones de dinámica molecular. Por lo tanto, el grupo tetrazol en la familia 1 y el grupo triazol en la familia 2 ocuparían la zona junto al hueco catalítico y actuarían como un grupo isostérico del fosfato 5´ de la base dañada, permitiendo así a los grupos previamente mencionados interaccionar con el residuo *Arg135*. Esto es consistente con el hecho de que los grupos tetrazol son

isósteros de grupos aniónicos, como los ácidos carboxílicos o los fosfonatos. Del mismo modo, los grupos bencilpiperazinilo y bencilpiperadinilo de los inhibidores en la familia 1, y el radical benciloxibencilo en la familia 2, harían un apilamiento contra el anillo aromático de la *Tyr114*. El orden de las energías de interacción promedio calculadas con *MM-GBSA* correlaciona bien con las diferencias observadas en afinidad (ver Tabla 4-XI), proporcionando soporte a los modos de unión predichos. Por último, como puede verse en la Tabla 4-XII, las interacciones más relevantes tienen lugar con los residuos *Arg128*, *Arg135*, *Tyr114* y *Tyr158*. Esto es consistente con la importancia que dichos residuos tienen en estudios experimentales de mutagénesis.

4.6.2. Ape1

Hasta el momento sólo se han analizado los resultados del pre-filtro realizado con *DOCK* (falta el pre-filtrado con *FRED*). Tras aplicar el corte, se obtienen 2288 moléculas que pasan a la etapa de filtrado con *CDOCK*. Al concluir esta etapa, se realizaron las dinámicas moleculares de las 100 mejores, y su inspección visual permitió seleccionar 16 moléculas para comprar y probar experimentalmente. Dichas moléculas proceden de casas comerciales como *Specs*, *ChemDiv*, *IBScreen*, *ChemBridge*, etc. Los ensayos experimentales fueron realizados inicialmente en el CNIO (Centro Nacional de Investigaciones Oncológicas) por la Dra. Carme Fábrega, Sandrea M. Francis, y el Dr. Federico M. Ruiz del antiguo Grupo de Transducción de Señales que estaba dirigido por el Dr. Jerónimo Bravo. En la actualidad la Dra. Carme Fábrega es la encargada de continuar con el proyecto en el *IRB* (*Institute for Research in Biomedicine*) de Barcelona.

Por ahora sólo se han llevado a cabo los ensayos *in vitro*. De las 16 moléculas, 7 mostraron algo de actividad (en valores de IC_{50}): dos de ellas en el rango 100-150 μM, por lo que sólo se tendrán en cuenta para realizar futuras modificaciones y tratar de buscar mejores hits a partir de ellas; otras dos estaban en el rango 50-125 μM, pero es demasiado amplio ya que los resultados no fueron excesivamente consistentes por lo que se necesitarán más pruebas; un problema parecido lo tenía otra de las moléculas, con un rango de actividad de 5-25 μM, pero tanto la inconsistencia de los resultados como la amplitud del rango en esos valores igualmente invitan a la realización de más pruebas; las dos últimas moléculas activas mostraron consistentemente actividades de aproximadamente 5 μM y 7.5 μM respectivamente. La Figura 4-36 muestra la gráfica con los resultados preliminares de actividad para las 4 mejores moléculas (cada una

representada en un color). Las líneas verticales en cada punto indican la desviación estándar.

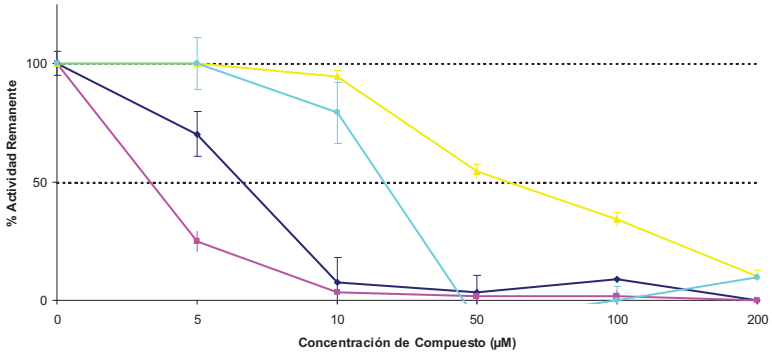


Figura 4-36. Resultados preliminares del nivel de actividad de *Ape1* para diferentes concentraciones de los cuatro mejores compuestos.

La Dra. Carme Fábrega continuará refinando los ensayos *in vitro* de las moléculas que mostraron las actividades más prometedoras, y además realizará posteriormente los ensayos *in vivo*. En base a los resultados obtenidos se podrá abordar la etapa de optimización para obtener compuestos más activos.

4.6.3. HDC

En base a los resultados obtenidos se comprueba que los métodos empleados tienden a fallar a la hora de intentar ajustar en el centro activo compuestos con un gran número de átomos. Esto hace que se reduzca el número de posibles candidatos. Estos resultados son acordes con las observaciones realizadas por Wu et al. (Wu et al., 2008). Por otro lado, los compuestos identificados tras el cribado parecen adoptar conformaciones en el centro activo que permiten realizar interacciones con algunos residuos clave involucrados en la estabilización del substrato (Moya-Garcia et al., 2008), como son *Y83* e *Y337* para favorecer la recepción del ligando, y *H351* y *H197*. Tras la selección de moléculas para comprar, los ensayos experimentales serán realizados por Almudena Pino y el Dr. Aurelio Moya-García del grupo dirigido por la Dra. Francisca Sánchez-Jiménez en el Departamento de Biología Molecular y Bioquímica de la Universidad de Málaga. Dichos colaboradores pertenecen también al CIBER (Centro de Investigaciones Biomédicas en Red) de Enfermedades Raras (CIBERER).

4.6.4. PCNA

Tras la etapa de pre-filtrado 29475 moléculas, aquellas con un *ZScore* ≥ 3.4, pasaron a filtrarse con el programa *CDOCK*. Las 100 mejores tras este filtro fueron refinadas mediante simulaciones de dinámica molecular según el protocolo estándar. Tras la inspección visual, se seleccionaron 13 moléculas para ser compradas y probadas experimentalmente. La verificación de la unión entre los ligandos y *PCNA* se realizó mediante RMN (Tugarinov et al., 2004) y ensayos de fluorescencia. Dichas pruebas fueron llevadas a cabo por el Dr. David Pantoja y el Dr. Ramón Campos-Olivas en el antiguo Grupo de RMN del CNIO. Las moléculas resultaron insolubles unas y muy poco solubles otras. Por ello se realizó una búsqueda de análogos de estas 13 moléculas, pero esta vez teniendo en cuenta sus valores de lipofilicidad (*logP*) y solubilidad en agua (*logS*) a la hora de realizar la selección final. De esta búsqueda se eligieron 8 nuevas moléculas de las cuales sólo 7 pudieron comprarse. En los nuevos ensayos 2 de las moléculas mostraron algo de actividad, aunque al parecer para una de ellas la constante de disociación era superior a 100 μM.

4.6.5. FtsZ

A la etapa de filtrado con *CDOCK* pasan tan solo 2055 moléculas, que son aquellas cuyo valor de *ZScore* obtenido a partir del pre-filtrado con *DOCK* es mayor o igual que 4.5. Tras la etapa de inspección visual, se seleccionaron 17 compuestos para ser comprados y probados experimentalmente. Las pruebas fueron realizadas por Claudia Schaffner-Barbero y el Dr. Antonio J. Martín-Galiano del laboratorio "Microtubulos y FtsZ: modulación del ensamblaje de proteínas" dirigido por el Profesor José Manuel Andreu, en el CIB (Centro de Investigaciones Biológicas) de Madrid. Pero al igual que sucedió con las primeras moléculas de *PCNA*, los valores de solubilidad no permitieron llevar a cabo los experimentos. Por ello se buscaron moléculas análogas a las seleccionadas pero teniendo en cuenta que su valor de *logS* (calculado con *QikProp* de la compañía *Schrödinger*) estuviera entre -5 y -1. Tras continuar el protocolo (desde la etapa de filtrado) con los 2337 análogos que cumplían las condiciones de *logS*, se seleccionaron 8 nuevas moléculas para probar. Por desgracia ninguna de estas moléculas resulto activa. Por ello en estos momentos se están planteando otras alternativas, como la utilización del sitio análogo al de unión de los taxanos (Desai & Mitchison, 1998) en β-tubulina (la proteína homóloga eucariota de *FtsZ*) como centro activo, o la selección del centro activo descrito por Haydon et al. (Haydon et al., 2008).

4.6.6. Otros Proyectos de Cribado Virtual

Actualmente se están llevando a cabo otros proyectos de cribado virtual, alguno de los cuales se encuentra en un estado muy avanzado. Pero debido al alto nivel de confidencialidad que requieren no es posible proporcionar excesivos detalles sobre ellos.

Uno de los proyectos consiste en la identificación de fármacos basándose en un nuevo mecanismo de inactivación descubierto para la proteína *p38 MAPK* (*Mitogen-Activated Protein Kinases*) (Chang et al., 2002; Peregrin et al., 2006), que regula importantes procesos celulares, como son las respuestas al estrés, diferenciación, y control del ciclo celular. Dicho proyecto está dirigido por el Profesor Federico Mayor Menéndez y la Dra. Cristina Murga Montesinos, del Departamento de Biología Molecular (UAM) y el Centro de Biología Molecular Severo Ochoa (CSIC) de la Universidad Autónoma de Madrid. Los ensayos experimentales son llevados a cabo por Pedro Campos, del mismo grupo. A lo largo del proyecto se han obtenido varias moléculas activas (en el rango μM), tanto in vitro como en células. A partir de estas moléculas se han buscado análogos, alguno de los cuales también han resultado activos. Lo mismo sucede con las moléculas que han sido sintetizadas a partir de pequeñas modificaciones sobre las originalmente activas. Esta síntesis ha sido realizada gracias a la colaboración con el Dr. Florenci V. González, del Departament de Química Inorgànica i Orgànica de la Universitat Jaume I en Castellón. En la actualidad se están terminando los últimos ensayos con el fin de seleccionar entre 1 y 3 compuestos para la realización de ensayos en animales. Además, con la información obtenida sobre los datos experimentales de inhibición se realizarán análisis *COMBINE* para poder sugerir modificaciones que aumenten la potencia de los inhibidores. Este proyecto está financiado por la Fundación Genoma España y además se prevé que en breve se pueda depositar una patente.

Otro de los proyectos trata de encontrar inhibidores para *TC21* (*RRas2*), que es un oncogén de la subfamilia *RRas* de las *Ras GTPasas*. Existen evidencias (Delgado et al., 2009) de que *TC21* es fundamental para la supervivencia y la proliferación homeostática de células T y B. Por lo tanto los inhibidores de *TC21* podrían convertirse en futuros fármacos antitumorales. Este proyecto está coordinado por el Profesor Balbino Alarcón, del Centro de Biología Molecular Severo Ochoa. El cribado virtual ha sido realizado en el supercomputador *MareNostrum*. De las 16 moléculas finalmente adquiridas, la mitad

muestran buenos valores de actividad *in vivo* (algunos por debajo de 1 μM). En la actualidad se están realizando pruebas para ver si también resultan activas *in vitro*, de modo que se puedan seleccionar una serie de compuestos a partir de los cuales generar derivados más potentes y patentables.

Recientemente en otro proyecto de cribado se han seleccionado 20 moléculas para probar experimentalmente su grado de inhibición sobre la proteína *Vav1* (Chrencik et al., 2008). Dicha proteína es un factor de intercambio de guanina (*GEF – Guanine-nucleotide exhange factor*) que juega un importante papel en la activación de células T y generación de tumores. Este otro proyecto está coordinado por el Profesor Xosé Bustelo, del CIC (Centro de Investigación del Cáncer) de Salamanca. El cribado también ha sido realizado en el supercomputador *MareNostrum*. Se está a la espera de los primeros resultados experimentales.

DISCUSIÓN

5. Discusión

La presente Tesis se ha desarrollado completamente en el campo de la búsqueda de nuevos fármacos desde una aproximación computacional. Por lo tanto, tiene un alto componente informático, aunque apoyado sobre conocimientos químicos, físicos, biológicos y matemáticos. También se le ha dado una gran importancia a la validación de los resultados obtenidos, no sólo a nivel teórico, sino también a nivel experimental, ya que es aquí donde en última instancia un trabajo de este tipo debe demostrar su verdadera utilidad. Esta validación ha sido posible gracias al trabajo de los colaboradores de laboratorios experimentales interesados en este tipo de aproximaciones.

La búsqueda de fármacos desde un punto de vista computacional no consiste en una única técnica que se aplica en un proceso simple; más bien se trata de un conjunto de técnicas que deben aplicarse conjuntamente en diferentes procesos para lograr el resultado deseado. Hoy en día están disponibles una inmensa variedad de métodos para llegar a proponer moléculas candidatas a fármacos partiendo de librerías químicas formadas por representaciones 2D de compuestos. Entonces debería ser posible integrar todos los elementos de *software* necesarios para crear un flujo de trabajo personalizado para un proyecto en cuestión. Pero el flujo de los datos entre los diferentes pasos (las conexiones de entrada/salida) resultan ser un problema importante principalmente debido a la gran variedad de formatos que pueden ser utilizados a la hora de describir las estructuras moleculares. Aunque se han hecho algunos avances al respecto (por ejemplo *SMILES* e *InChI* [http://www.iupac.org/inchi/]), aún se está lejos de adoptar un formato consenso. Por otro lado, en las ciencias de la vida en general, y en el diseño computacional de fármacos en particular, se debe tratar con una gran cantidad de datos provenientes tanto por parte del receptor (por ejemplo las estructuras 3D de proteínas disponibles de los proyectos genómicos), como por parte del ligando (enormes quimiotecas con millones de moléculas para ser cribadas). La cantidad de datos para ser procesados, almacenados y gestionados requiere el uso de potentes motores de bases de datos. Finalmente, los métodos utilizados en este tipo de proyectos cada vez son más precisos pero al mismo tiempo también van requiriendo de más recursos computacionales. Por ello, hoy en día resulta indispensable el uso de la computación paralela en diferentes procesadores y diferentes sistemas. Estos cuatro aspectos mencionados son los que han motivado el desarrollo de *VSDMIP* como una plataforma

integrativa para el manejo de estos datos, libre, de código abierto, y disponible para toda la comunidad científica. Sus principales ventajas son: 1) la posibilidad de realizar automáticamente experimentos de cribado virtual; 2) facilitar la comparación de diferentes protocolos; 3) total flexibilidad a la hora de diseñar protocolos de cribado; 4) implementación de un mecanismo basado en *XML* para añadir nuevos componentes software, y configurar protocolos a voluntad; 5) la generación de una base de datos relacional acoplada para mantener todos los datos organizados y listos para usarse; y 6) mecanismos para la ejecución de tareas en paralelo en diferentes procesadores. *VSDMIP* carece de un interfaz gráfico de usuario, pero su arquitectura permite que pueda añadirse en el futuro. Cuando se implemente, se podrá además proporcionar información relacionada con el receptor y las interacciones del ligando en el sitio activo. Con pequeñas modificaciones en el esquema de la base de datos sería posible almacenar resultados de programas de *docking* que generasen como solución nuevas conformaciones de un ligando, extendiendo así la capacidad actual en la que básicamente se trabaja con programas de *docking* que utilizan conjuntos de confórmeros precalculados (o que se han configurado de esta manera).

Recientemente se han publicado algunas aproximaciones similares, a continuación se comentan brevemente y se comparan con *VSDMIP*. Probablemente la aproximación más parecida es la reportada por *SciTegic, Inc.* (Hassan et al., 2006; SciTegic, Inc. 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA) llamada *Pipeline Pilot*. Es un software comercial que utiliza una tecnología conocida como *tubería de datos* para construir y ejecutar flujos de trabajo configurables usando componentes que encapsulan principalmente algoritmos quimioinformáticos (aunque el *docking* pude ser realizado también utilizando los programas *GOLD* o *FLEXX*). Hassan et al. han demostrado la utilidad de esta plataforma en una revisión reciente (Hassan et al., 2006) donde presentan resultados de cribados virtuales utilizando técnicas Bayesianas de aprendizaje sobre varias dianas proteicas.

En esta misma línea, *Astex Therapeutics* (Watson et al., 2003) posee una plataforma *web* propietaria que integra una base de datos relacional *ORACLE* para almacenar moléculas (procedentes de su base de datos *ATLAS* [*Astex Technology Library of Available Substances*]), propiedades, datos de la diana, interacciones de unión, y resultados. Los experimentos de cribado se pueden realizar usando compuestos obtenidos directamente de la base de datos o creándolos virtualmente a partir de ciertas reacciones químicas. Las moléculas 2D en formato *SMILES* son convertidas a

estructuras 3D usando *CORINA* para después hacer *docking* con el programa *GOLD*. También contiene una interfaz gráfica para configurar los experimentos y visualizar los resultados. *VSDMIP* tiene una serie de ventajas sobre este sistema, como son su modularidad, el ofrecer a los usuarios la posibilidad de crear sus propias quimiotecas, y el hecho de que esté disponible libremente para la comunidad científica. Por contra, a día de hoy *VSDMIP* trabaja a través del interfaz por línea de comandos.

La idea de emplear ficheros *XML* configurables por el usuario para realizar flujos de trabajo personalizados en el campo del diseño de fármacos (tal como se hace en *VSDMIP*) ha sido implementada también por Lehtovuori y Nyrönen (Lehtovuori & Nyronen, 2006). En su aproximación *SOMA*, el usuario decide qué protocolo usar seleccionando, a través de un interfaz *web*, los pasos y los parámetros adecuados (a pesar de que en la implementación publicada sólo podían realizarse cálculos de propiedades moleculares estructurales y/o experimentos de docking con *GOLD*). El programa *Grape* gestiona el flujo de trabajo encadenando las aplicaciones necesarias en el orden establecido por el usuario a través de los nodos de ejecución (usados para ejecutar y manejar las aplicaciones). Finalmente, el mismo interfaz *web* se usa para recuperar, visualizar y analizar los resultados. La salida producida por cada nodo ejecutable es codificada, etiquetada (para un seguimiento del proceso en cualquier momento), y actualizada con la información de salida generada por los pasos siguientes (dependiendo del protocolo seleccionado). Otro componente importante es el repositorio de herramientas, que contiene una serie de programas útiles para realizar tareas intermedias (conversiones de formatos de ficheros y generación de ficheros ejecutables entre otras). El protocolo completo de *SOMA* se codifica en un único fichero *XML*, lo que significa que, una vez se han proporcionado los parámetros de entrada, el usuario no necesita interaccionar más con el sistema hasta que el trabajo completo ha finalizado. Esto resulta muy conveniente para protocolos ya establecidos. Por contra, *VSDMIP* está más orientado a la toma de decisiones tras cada paso. Esto es así porque si se va a utilizar más de un programa de *docking*, o hay disponibles diferentes filtros, es necesario parar tras cada paso para comprobar los resultados antes de pasar al siguiente. A este nivel, realizar un ajuste fino de un protocolo automático resulta difícilmente tratable o factible. Otro aspecto que merece comentarse en comparación con *VSDMIP* es la carencia de una base de datos para gestionar los resultados. *SOMA* almacena los resultados de todo el flujo de trabajo en un gran fichero que es mostrado como una tabla en el interfaz. No está claro si las moléculas y los resultados ya obtenidos pueden ser

reutilizados en otro conjunto de experimentos. *VSDMIP* es totalmente flexible en ese aspecto (una vez una molécula ha sido insertada en *VSDMIP*, esta puede ser reutilizada tantas veces como se quiera). La disponibilidad del esquema *XML* de *SOMA* y su interfaz gráfica en web hacen que esta aplicación resulte atractiva para los diseñadores de fármacos con pocos conocimientos de programación y que quieran centrarse más en los aspectos químicos/biomédicos del problema a tratar en lugar de en los detalles técnicos. Finalmente, un aspecto en común con *VSDMIP* es la posibilidad de incorporar nuevas aplicaciones de una manera relativamente simple empleando scripts *XML*.

Recientemente se ha publicado una nueva plataforma para la realización automática de experimentos de docking que tiene como característica más significativa la posibilidad de evaluar en cierta medida si un cribado prospectivo podría ser exitoso. Se trata de *DOCK Blaster*, de Irwin et al. (Irwin et al., 2009), que incluye un sistema experto capaz de preparar la proteína para el *docking* con el programa *DOCK*, y permite proporcionar uno o varios ligandos activos (y también inactivos) con los que realiza automáticamente un cribado retrospectivo antes de realizar el cribado prospectivo. El cribado retrospectivo se puede realizar con diferentes configuraciones. Para decidir cuál de ellas resulta mejor se usan dos parámetros: 1) fidelidad de la pose, que consiste en evaluar el *RMSD* de la pose obtenida y la proporcionada; y 2) el factor de enriquecimiento, entendiéndose en este caso como la habilidad para obtener activos sobre los inactivos. En base a estos datos es capaz de seleccionar la mejor estrategia, o bien desestimar el cribado al considerarlo no factible. También tiene las ventajas de poseer un interfaz *web* sencillo, funciona automáticamente en paralelo sobre un *cluster* con cientos de procesadores, y además está integrado con la base de datos de *ZINC* y con el *DUD Decoy Maker* (Huang et al., 2006) (para la obtención de ligandos inactivos con propiedades similares a los activos). Pero aunque se trata de una aproximación novedosa en lo que se refiere a la selección automática de la mejor configuración para el cribado, presenta algunas desventajas en comparación con *VSDMIP*. Como por ejemplo la rigidez, ya que en principio no permite utilizar otros programas que no sean *DOCK*, ni configurar diferentes protocolos. Además, los datos generados no se almacenan en una base de datos (al menos que esté accesible para el usuario desde la interfaz) ni tampoco parece que puedan ser reutilizados en sucesivas etapas o nuevos cribados. Otro factor negativo es que por ahora no permite incluir quimiotecas propias, ya que utiliza directamente la base de datos *ZINC*.

La importancia que está adquiriendo en cualquier proyecto de búsqueda de nuevos fármacos la gestión de los datos, flujos de trabajo y entornos de cálculo se hace patente cada día como demuestra la aparición reciente de estas plataformas computacionales. Otro ejemplo más es *DVSDMS* (*Data Management System for Distributed Virtual Screening*) de Zhou y Caflisch (Zhou & Caflisch, 2009), muy similar a *VSDMIP* pero mucho más orientado al control continuo de los procesos que se están ejecutando. Y aunque no son estrictamente comparables con *VSDMIP*, existen otras herramientas relacionadas que merecen ser comentadas. Por ejemplo, el paquete de *AutoGrid*/*AutoDock*, que ha atraído gran interés en los últimos años y probablemente esté llamado a convertirse en uno de los programas más usados para tareas de *docking*. Por encima de otras herramientas automáticas disponibles en la página *web* de desarrolladores, hay dos aplicaciones para realizar experimentos de cribado virtual interesantes: *BDT* (Vaque et al., 2006) y *DOVIS* (Zhang et al., 2008). El primero permite al usuario interactuar con el código del programa a través de una aplicación gráfica que automáticamente realiza la preparación de las *grids* y su combinación (para permitir flexibilidad del receptor), cálculo del *docking* y análisis de los resultados. El segundo también incorpora un paso adicional para la preparación de los ligandos. La principal ventaja de *DOVIS* sobre *BDT* es que *DOVIS* permite que el *docking* se realice en paralelo usando un *cluster* de *Linux* (con o sin sistema gestor de colas). Sin embargo, en ambos casos el usuario está restringido al uso de un único programa de *docking* y no existe una base de datos para manejar los diferentes proyectos. De todas formas, la distribución libre de los programas y los interfaces de usuario fáciles de utilizar hacen ideales estas herramientas para investigadores que están más interesados en obtener respuestas a sus problemas particulares que en el proceso de *docking* en sí mismo.

Una vez ya se cuenta con un marco de trabajo sobre el que implementar diferentes protocolos de cribado, se hace necesario encontrar cuál sería el más apropiado en cada caso, o al menos encontrar uno que en términos generales pueda funcionar bien y que se ajuste a los requisitos de precisión, tiempo y capacidad de computación. Dada la cantidad de moléculas presentes en las bases de datos estándar, hoy en día no es computacionalmente factible realizar experimentos de cribado virtual usando únicamente programas de *docking* de alta precisión. En su lugar, una alternativa común es usar filtros concatenados para reducir el número de moléculas con las que se hará *docking*. En otros estudios, tras la etapa inicial de filtrado y *docking* fino, se utilizan diferentes funciones de puntuación y los candidatos son seleccionados en base a un

criterio de consenso (Yang et al., 2005). Pero otras alternativas más prometedoras utilizan más de un programa de docking en orden creciente de eficacia (Maiorov & Sheridan, 2005; Miteva et al., 2005). En esta Tesis se ha elegido *DOCK* como programa inicial de *docking* (etapa de Pre-Filtrado) porque es lo suficientemente rápido como para cribar una molécula en pocos segundos (el tiempo total depende del número de confórmeros y del número de esferas usadas para describir el sitio de unión). En las diferentes pruebas realizadas en este trabajo se ha visto que *CDOCK* claramente mejora los resultados obtenidos con *DOCK*, pero a costa de requerir más tiempo de computación. Por ello se han probado diferentes protocolos haciendo uso primero de *DOCK* y después de *CDOCK*. Las variaciones entre los distintos protocolos consistían en cambios en los parámetros que controlan el número de compuestos y conformaciones pasados de *DOCK* a *CDOCK*. Se ha observado que aunque la mayoría de las moléculas que pasan el filtro inicial con *DOCK* son en realidad inactivas, pueden ser reconocidas por *CDOCK* como tal y ser descartadas. En general, la combinación de estos dos programas de *docking* muestra un grado de rendimiento intermedio en comparación con los rendimientos de ambos por separado. Además, también se ha comprobado que los resultados tras *CDOCK* son mejores cuanto mayor es el número de conformaciones de una misma molécula que se le pasan tras *DOCK*. Esto es un efecto en general observado (siempre que el muestreo conformacional se haya realizado adecuadamente), ya que cuantas más conformaciones se tengan por molécula mayor será la probabilidad de que el programa de *docking* detecte la pose de unión correcta (Knox et al., 2005). Así pues, en función de la experiencia acumulada en el desarrollo de esta Tesis y las pruebas realizadas, se ha elaborado un protocolo de cribado virtual que pretende servir como referencia para los nuevos proyectos de búsqueda de fármacos. Dicho protocolo incluye las fases previas de preparación de la proteína y la quimioteca, y además también tiene en cuenta la realización de optimizaciones de los resultados utilizando dinámica molecular con el objetivo de dotar de flexibilidad a los complejos proteína-ligando generados y realizar cálculos energéticos más precisos con *MM-GBSA*.

Pero no nos hemos limitado a la creación de un marco de trabajo para la búsqueda de nuevos fármacos y a la determinación de un protocolo estándar de cribado, sino que hemos profundizado en algunos de los componentes más importantes en proyectos de este tipo, como son: 1) las funciones de puntuación basadas en términos energéticos para complejos proteína-ligando, que son las que permiten reconocer la pose correcta de un ligando y discriminar entre activos e inactivos; 2) la flexibilidad en los ligandos, lo

que permite su correcta adaptación al centro activo de la proteína; y 3) el propio método de *docking* proteína-ligando.

En cuanto a las funciones de puntuación, se ha presentado *ISM*, un nuevo modelo para el cálculo rápido de las energías electrostáticas libres de unión en complejos proteína-ligando. La formulación es una modificación del modelo original propuesto por Hassan et al. (Hassan et al., 2000a; Hassan et al., 2000b) para tratar los efectos electrostáticos en las proteínas. Como en su caso, no se requiere una superficie de separación entre la parte de alta constante dieléctrica (solvente) y la de baja (receptor o ligando). Esto se logra definiendo la función dieléctrica de modo sigmoidal dependiente de la distancia. De un modo similar, los radios efectivos de *Born* son calculados solo a partir del área de la superficie accesible al solvente para el átomo de interés. Para reproducir correctamente los resultados obtenidos resolviendo la ecuación de *Poisson* (*PE*, tomados como referencia), se hace necesaria la introducción de un término corrector debido a la formación de enlaces por puente de hidrógeno. Aunque esta corrección pueda parecer extraña, las uniones por puentes de hidrógeno tienen una componente electrostática muy importante que debe ser capturada por el modelo. Sin embargo, en estudios de Swanson et al. (Swanson et al., 2005) se establece claramente que el uso de superficies centradas en los átomos, tal como se hace en el método *LCPO*, o la presencia de transiciones suaves entre las regiones de constantes dieléctricas altas y bajas, como en el caso de la función dieléctrica sigmoidal, incrementa la fracción de regiones intersticiales de alto dieléctrico en el interior de la proteína. Se ha demostrado que la presencia de estas regiones suprime las barreras de energía libre electrostática características en la formación de puentes de hidrógeno cuando se comparan con simulaciones de potencial de fuerza media, produciendo una sobrestimación de las energías de solvatación, en concreto para los grupos envueltos en interacciones por puentes de hidrógeno (Swanson et al., 2005). El término empírico de puente de hidrógeno utilizado en *ISM* actúa como una corrección de este efecto, aunque deberían realizarse estudios futuros sobre este tema.

En base a las pruebas realizadas con el conjunto de 826 poses (las cuales cubren substancialmente una gran variedad estructural tanto en proteínas como en ligandos), el modelo *ISM* es capaz de obtener resultados satisfactorios. El coeficiente cuadrático de validación cruzada con las energías electrostáticas libres de unión obtenidas mediante *PE* es de 0.81, con una pendiente de 0.97 y una ordenada en el origen de 1.06 *kcal/mol*, presentando un valor de *RMSD* de ~4.33 *kcal/mol* (ver apartado b de la Figura 4-14 y la

Tabla 4-IV de las páginas 106 y 107 respectivamente). Las diferentes contribuciones para las energías electrostáticas libres de unión también se reproducen con una precisión similar (ver Figura 4-15 de la página 108). Estos resultados son coherentes con los obtenidos por Liu y Zou (Liu & Zou, 2006), quienes estudiaron la habilidad de *GB* para reproducir las energías electrostáticas libres de unión calculadas con *PE*. En su estudio usaron estructuras cristalinas de 15 complejos para el ajuste y otras 15 para la validación cruzada, Liu y Zou obtuvieron un coeficiente de correlación cuadrático de 0.81 y un *RMSD* de 4.05 *kcal/mol* en la fase de ajuste, mientras que en la de validación cruzada los valores obtenidos fueron 0.81 y 5.14 respectivamente. Esta comparación sugiere que los métodos *GB* e *ISM* logran rendimientos similares a la hora de modelar las energías de unión de complejos proteína-ligando. Pero el modelo *ISM* necesita una transformación logarítmica para ajustar la energía total electrostática libre de unión a los resultados obtenidos con la *PE*. La razón de este efecto no lineal es la relativa sobreestimación de la energía libre de desolvatación para proteínas que contienen canales hidrofóbicos largos y angostos en el sitio de unión, como por ejemplo la proteína de unión a retinol o la de unión a biotina (ver Tabla 3-I en la página 28 para una descripción de los complejos). El motivo de esta sobreestimación no está claro, y es algo que se continuará investigando. Por otro lado, hay que hacer notar que los cálculos *PE* se utilizan aquí más como una guía que como una referencia o estándar. El propio método *PE* depende de una serie de parámetros empíricos, como las contantes dieléctricas externas e internas y la definición de límites entre otros. Estos parámetros no han sido determinados únicamente y están sujetos a debate. Por ello no se ha intentado mejorar el ajuste del modelo *ISM* a *PE* introduciendo nuevos conjuntos de parámetros o más empirismo al modelo; sino que se ha preferido mantener sólo aquellos parámetros que corresponden a cantidades físicas significativas.

El tiempo medio de cálculo por pose para el modelo *ISM* es de 30-40 ms (ver apartado a de la Figura 4-16 en la página 109), y se observa una relación aproximadamente lineal entre el tamaño del ligando y el tiempo de computación (ver apartado b de la Figura 4-16 en la página 109). Se espera que este tiempo pueda reducirse aún más mediante el empleo de una tabla de átomos vecinos de cada punto de *grid* (tal como se hace en la implementación de *ISM* dentro de *CGRID-CDOCK* para la obtención de los puentes de hidrógeno), acelerando de este modo el cálculo de los radios efectivos de *Born*. Así pues, el método *ISM* muestra, tanto en términos de tiempo como de precisión, que es lo suficientemente bueno como para ser implementado

directamente dentro de un algoritmo de *docking*, y que es comparable a otras aproximaciones. Por ejemplo, el método *GB* implementado originalmente por Zou et al. (Zou et al., 1999) en el programa *DOCK* requería ~10 s por complejo en una estación de trabajo *SGI Octane*. El mismo grupo propuso posteriormente una aproximación para calcular los radios de *Born* basada en interacciones por pares de átomos, reduciendo así el tiempo de computación a 0.5 s por complejo (Liu et al., 2004). Estos tiempos hacen que sea desaconsejable el empleo de este método en la etapa de *docking*, siendo relegado su uso únicamente al postprocesado. Por otro lado, Majeux et al. (Majeux et al., 2001) propusieron un método continuo simplificado basado en la asunción de que la desolvatación electrostática puede aproximarse a través de la eliminación de la primera capa de moléculas de agua en la superficie de unión, y que la contribución coulómbica se puede calcular mediante un modelo dieléctrico dependiente de distancia. Precalcular las contribuciones energéticas para un conjunto de *grids* permitió a los autores estimar la energía electrostática libre de unión en solución empleando aproximadamente 3-4 ms por cada fragmento de 5-10 átomos pesados en una máquina *Pentium* III a 550 *MHz*. Sin embargo, su método estaba restringido al *docking* de moléculas rígidas, ya que tanto el ligando como el receptor necesitan que las *grids* sean preprocesadas, mientras que el método *ISM* puede ser empleado tanto en casos rígidos como en flexibles. Esto limita la aplicación del método de Majeux et al. principalmente al *docking* de fragmentos rígidos pequeños en sitios activos también rígidos. La precisión para la energía total electrostática libre de unión obtenida con el método de Majeux et al. es ligeramente inferior a la obtenida con el método *ISM*, en vista de los coeficientes de correlación cuadráticos cuando se compara con resultados de *PE* (~0.75 frente a ~0.81, respectivamente). Las contribuciones a la energía libre de unión electrostática son reproducidas similarmente por ambos métodos (con un r^2 de ~0.81 en ambos casos), pero el modelo *ISM* proporciona una pendiente próxima a 1.0 (ver Figura 4-15), mientras que en el método de Majeux et al. las pendientes son mayores y muestran también mayor dispersión (de 1.49 a 2.95) de los datos.

En lo que respecta la simulación de la flexibilidad, gracias a *ALFA* es posible dotar de flexibilidad a las estructuras rígidas de ligandos generadas con un conversor de 2D a 3D como puede ser *CORINA*. La generación de confórmeros independientemente de la estructura de un centro activo de una proteína tiene como ventaja principal el poder reutilizar los resultados con múltiples receptores; pero como se ha podido comprobar, no siempre es efectiva. En base a los resultados obtenidos se aprecia que el

problema radica en gran medida en la selección de los confórmeros de entre todos los generados. Las funciones de energía empleadas (en este caso es un *van der Waals* 1-4) evitan que se seleccionen estructuras con choques estéricos pero no son capaces de obtener un conjunto de ligandos representativos de todo el espacio conformacional. Cuantos más representantes se generan para los diferentes mínimos energéticos, más difícil es seleccionar soluciones diferentes a las del mínimo global. Para obtener una representación homogénea sería necesario generar grupos de soluciones factibles (aquellas que no presentasen choques estéricos) y tomar al menos un representante de cada uno. Aunque en otras ocasiones el problema se encuentra en la asignación de estados rotaméricos. En este sentido sería importante disponer de un método automático para actualizar los patrones de los enlaces rotables y sus posibles ángulos en función de información experimental, como hacen Sadowski y Boström (Sadowski & Bostrom, 2006) utilizando los principios del programa *MIMUMBA* de Klebe y Mietzner (Klebe & Mietzner, 1994). En su trabajo, Sadowski y Boström generan patrones *SMARTS* para todas las combinaciones de 4 átomos que pueden obtenerse a partir de los tipos de átomos del campo de fuerzas de *Tripos*, y hacen un análisis estadístico para cada patrón de torsional sobre la librería *Cambridge Crystallographic Database* (*CSD*) (Allen et al., 1991) con el objetivo de encontrar los ángulos representativos. Las reglas derivadas se utilizan en el programa *OMEGA* (de la compañía *OpenEye*), mejorando sus resultados. *OMEGA* es un programa de análisis conformacional basado en reglas al igual que *ALFA*, pero en lugar de utilizar un algoritmo de *MCSA* para generar los confórmeros, lo que hace es dividir la molécula en fragmentos de hasta cinco enlaces rotables contiguos, y los reensambla basándose en el orden obtenido según las energías de cada fragmento realizando una búsqueda en profundidad. Además también tiene en cuenta las conformaciones de anillos mediante el uso de librerías de estructuras pregeneradas. En *ALFA* no se tienen en cuenta las modificaciones en los anillos ya que deja esta tarea al programa *CORINA* cuando genera las estructuras 3D. *OMEGA* considera que dos estructuras generadas son diferentes si tras superponerlas al menos una pareja de átomos equivalentes están a una distancia superior a cierto valor de corte. En *ALFA* se hicieron pruebas similares empleando el *RMSD*, pero la gran cantidad de tiempo empleada en las superposiciones desaconsejaba el uso de este método (datos no mostrados).

Para evitar el problema de la explosión combinatoria al realizar las combinaciones de los ángulos de los torsionales se suelen utilizar algoritmos estocásticos. Además de los basados en técnicas de *MCSA*, como *ALFA*, otros hacen uso de algoritmos genéticos.

Es el caso del programa *Balloon*, de Vainio y Johnson (Vainio & Johnson, 2007), que se distribuye libremente al igual que *ALFA*. En *Balloon* se utiliza un algoritmo genético multiobjetivo, es decir, se trata de optimizar tanto la energía de los torsionales como la energía de *van der Waals* general, basándose ambas en un campo de fuerzas de tipo *MMFF94* (Halgren, 1996). Los cromosomas del algoritmo genético se codifican de modo que almacenen información sobre los ángulos en los enlaces rotables, la posición de los centros quirales, la estereoquímica de los dobles enlaces y las conformaciones de anillos. Para lograr diversidad en los resultados hace uso de comparaciones basadas en el *RMSD*. Dos aspectos importantes son su capacidad de obtener una estructura de partida a partir de una cadena de *SMILES* (utilizando algoritmos de geometrías de distancia) y además el hecho de que realiza un postprocesado para optimizar las estructuras obtenidas tras la ejecución del algoritmo genético. En base a los resultados publicados parece que su rendimiento es ligeramente inferior al de a *ALFA* (~60% frente a ~80% de ligandos que obtienen conformaciones por debajo de 1 Å en *RMSD* con respecto a la nativa), y el tiempo medio empleado es superior (~200 s frente a ~90 s), siendo el paso de optimización de estructuras una de las partes que más tiempo consume. Aunque cabe destacar el hecho de que como resultado final seleccionan un conjunto muy reducido de confórmeros (una media de 14 por ligando). Esto resulta contradictorio con el estudio realizado por Borodina et al. (Borodina et al., 2007), en el que demuestran la dependencia existente entre el tamaño del conjunto seleccionado, la flexibilidad de los ligandos, y el nivel de cobertura del espacio conformacional alcanzado. Según Borondina et al. con un conjunto de 14 confórmeros seleccionados la predicción del *RMSD* obtenido (con respecto a la conformación nativa) sería menor de 1 Å sólo para ligandos con 6 o menos enlaces rotables. El número de confórmeros seleccionados debería incrementarse en cada caso según el número de enlaces rotables de la molécula, y también en base a la resolución deseada. En el caso de las pruebas realizadas con *ALFA* se decidió tomar un máximo de 200 confórmeros para cada ligando pues ese era aproximadamente el número necesario para lograr valores de RMSD por debajo de 1 Å con moléculas de hasta diez enlaces rotables, que es el caso correspondiente al 80% de los ligandos utilizados en el estudio de Borodina et al.

Otro método que utiliza algoritmos estocásticos es *SPE* (*Stochastic Proximity Embedding*), de Agrafiotis *y* Xu (Agrafiotis, 2003; Agrafiotis & Xu, 2002; Xu et al., 2003). Se basa en la generación de estructuras que satisfacen una serie de restricciones referentes a distancias entre átomos y volúmenes de cuartetos de átomos. Partiendo de

una semilla para su generador de números aleatorios, sigue un proceso iterativo que conduce a la obtención de un confórmero. Cada nuevo confórmero se obtiene utilizando nuevas semillas. Mediante una simple heurística, basada en un algoritmo de impulso (*boosting*) (Agrafiotis et al., 2007), se consiguen conformaciones más extendidas tras cada iteración. Este sesgo se introduce en base al hecho de que las conformaciones bioactivas tienden a ser más extendidas que las generadas aleatoriamente, como demostraron Diller y Merz (Diller & Merz, 2002). El programa *Rubicon*, de la compañía *Daylight*, usa un método de geometría de distancia muy similar a *SPE*, aunque realiza una minimización por gradientes conjugados que lo hace más lento.

Recientemente Sperandio et al. (Sperandio et al., 2009) publicaron otro generador de confórmeros, *MED-3DMC*, muy similar en concepto a *ALFA*. También utilizan un algoritmo basado en *Monte Carlo* y Metropolis, pero en su caso no hacen uso de un conjunto de reglas predefinidas, sino que realizan variaciones del ángulo pudiendo tomar valores entre 30º y 180º a intervalos de 30º. Utilizan patrones *SMARTS* para detectar los enlaces rotables. De un modo similar a *ALFA*, tienen una ventana de energía para eliminar de la lista final aquellos confórmeros cuya energía es superior en cierta cantidad a la del mínimo global; Sperandio et al. usan también la energía de *van der Waals* pero además tienen en cuenta la energía por torsional. Finalmente aplica un método para conseguir diversidad de estructura consistente en dos pasos. En el primero se calcula el *RMSD* de pares de ángulos diedros equivalentes para el conjunto de todos los ligandos sobre unos diedros de referencia; esto permite descartar rápidamente conformaciones muy parecidas sin tener que superponerlas. En el segundo paso se calcula el *RMSD* entre pares de ligandos, pero sólo para aquellos que pasaron el primer filtro. Los resultados obtenidos con *MED-3DMC* parecen ligeramente mejores que los de ALFA, a pesar de que sólo seleccionan 50 estructuras por ligandos. Aunque se trata de un programa comercial.

Liu et al. (Liu et al., 2009) también han publicado recientemente un generador de confórmeros, *Cyndi*, basado en un algoritmo evolutivo multiobjetivo de un modo similar a como se hace en el programa *Balloon*, pero en este caso considerando además la diversidad geométrica como tercer objetivo independiente. Utilizan la implementación *o-MOEA* desarrollada por Deb et al. (Deb et al., 2003; Laumanns et al., 2002) para obtener soluciones óptimas de *Pareto*, es decir, soluciones que no son dominadas por ninguna otra (respecto a los tres objetivos) del espacio de soluciones. Al igual que *ALFA* solo mueve enlaces rotables (no anillos), y consigue una eficacia

ligeramente mejor (el *RMSD* promedio es de 0.864 Å para los tests realizados) empleando un tiempo similar: aunque Liu et al. afirman que *Cyndi* tarda sólo una media de 0.5 s por ligando, lo cierto es que sumando la etapa de minimización el tiempo necesario alcanza los 24 s de media por estructura. ALFA requiere aproximadamente en promedio 86 s por ligando, pero en este caso el ligando se compone de varias estructuras de partida (*CORINA* generó aproximadamente 4 estructuras de media en los conjuntos probados; datos no mostrados).

Los métodos para generar confórmeros de ligandos se han venido evaluando en función de si encuentran alguna estructura lo suficientemente parecida, en términos de valores *RMSD*, a una cristalográfica conocida, la llamada nativa o bioactiva. Resulta complicado evaluar si un método es capaz de generar todas las estructuras bioactivas posibles ya que normalmente para cada ligando sólo se dispone como mucho de un par de estructuras cristalográficas con diferentes proteínas. Por ello Takagi et al. (Takagi et al., 2009) han publicado recientemente un trabajo en el que ponen de manifiesto este problema y proponen que la evaluación se haga en función de la tasa de ocupación del espacio conformacional bioactivo logrado por el método a evaluar. Este espacio conformacional bioactivo lo generan mediante el programa *MOE* (de la compañía Chemical Computing Group - http://www.chemcomp.com/) y basándose en las restricciones para la energía de los torsionales obtenidas por Perola y Charifson (Perola & Charifson, 2004). Aunque constituye una alternativa interesante a la hora de evaluar los generadores de confórmeros, tiene el problema de que resulta difícil obtener el espacio conformacional bioactivo; de hecho Takagi et al. lo obtienen sólo para moléculas con seis o menos enlaces rotables.

Los métodos de *docking* son los que finalmente determinan la pose correcta para un complejo proteína-ligando, asignando además una puntuación que permitirá discriminar entre diferentes ligandos, que es el objetivo del cribado virtual. En este caso el estudio se hace sobre el programa *CDOCK*, que como una gran parte de los algoritmos de inteligencia artificial (en este caso de simulación de un proceso biológico) se basa en operaciones de movimiento (rotación, translación y flexibilidad del ligando) y evaluación (mediante la función de puntuación) para lograr el objetivo, la pose de mejor puntuación. Así pues, las tareas principales en las que interviene el *docking* se pueden resumir en: 1) predicción del modo de unión; 2) cribado virtual para la identificación de *hits*; y 3) ordenación por afinidad para la optimización de *hits*. Estas tres tareas son las evaluadas en el exhaustivo estudio realizado por Warren et al.

(Warren et al., 2006), que a pesar de haber sido publicado en el año 2006 sigue siendo un referente a la hora de conocer el estado del arte en lo que se refiere a programas de *docking* y funciones de puntuación. En dicho estudio evalúan 10 programas de *docking* con 37 funciones de puntuación empleando para ello ocho proteínas pertenecientes a siete clases diferentes. Determinan que en general los programas de *docking* son capaces de muestrear el espacio de poses de modo que se obtengan estructuras similares a las nativas, pero las funciones de puntuación tienen menos éxito a la hora de seleccionar la pose más próxima a la conformación del cristal. Esto va en concordancia con los resultados obtenidos en esta Tesis, aunque bien es cierto que en el caso de la función de puntuación de *CDOCK* el problema parece estar principalmente en que las pequeñas modificaciones en la conformación bioactiva del ligando en ocasiones impiden que alguna de las posibles poses alcance el mínimo energético global, ya que en las pruebas de *docking* rígido (tomando la estructura cristalográfica del ligando) se comprobó que la tasa de éxito era muy elevada (~90% con un *RMSD* por debajo de 2 Å, y ~70% por debajo de 0.5 Å). En base a esto parece que el empleo de conjuntos de confórmeros para simular la flexibilidad del ligando en el *docking* es una técnica limitada, por lo que debería emplearse en combinación con otras capaces de explorar el espacio conformacional del ligando en el contexto del centro activo de la proteína. De este modo el *docking* se beneficiaría de una mayor precisión al generarse el confórmero en función de los requisitos del receptor, y de una mayor velocidad al limitarse las estructuras de partida mediante el uso de conjuntos de confórmeros. Por contra habría que hacer los análisis conformacionales cada vez que se hace *docking*.

Otra de las conclusiones del estudio de Warren et al. es que los programas de *docking* son capaces de identificar moléculas activas de entre un conjunto de inactivas, aunque éstas tengan unas propiedades similares a las activas. *CDOCK* también parece ser capaz de realizar esta distinción, como se ha visto en la evaluación de diferentes protocolos con la plataforma *VSDMIP*. Y al igual que ocurre con *CDOCK*, Warren et al. determinaron que el rendimiento del cribado virtual puede variar mucho en función del tipo de proteína. Este es el motivo por el que puede resultar útil una plataforma como la mencionada en la discusión sobre *VSDMIP*: *DOCK Blaster* de Irwin et al. Gracias a la predicción del grado de éxito del cribado, se podría determinar cual sería el método de *docking* que obtendría el mayor rendimiento. Aunque para ello sería necesario que *DOCK Blaster* incluyera diferentes métodos de *docking*. De todos modos esta aproximación sólo sería factible en caso de disponer de moléculas activas ya conocidas.

En el trabajo de Warren et al. también se estudiaron otros asuntos importantes que no se han abordado en esta Tesis: 1) *docking* cruzado, es decir, ver como funciona el algoritmo de *docking* utilizando diferentes estructuras de la misma proteína. Aunque una prueba de este tipo tiene más sentido a la hora de evaluar un programa de *docking* que trate la flexibilidad de la proteína; 2) identificación de los diferentes quimiotipos activos en los compuestos obtenidos en el cribado virtual. Esto es algo que podría incluirse en el futuro mediante el uso de un método para agrupar las soluciones del cribado en función de la estructura química de los ligandos o su modo de unión; y 3) correlación entre el valor experimental de la afinidad y el valor obtenido con las funciones de puntuación, algo que en la actualidad aún no está dando buenos resultados.

En un trabajo posterior, Hartshorn et al. (Hartshorn et al., 2007) evalúan una versión propia del programa *GOLD* (Jones et al., 1997) empleando el conjunto diverso de *Astex Therapeutics* obtenido también en ese trabajo (ver apartado 3.1.1.1.3 de la página 27). El método de muestreo de *GOLD* está basado en un algoritmo genético, y utiliza *Goldscore* (Jones et al., 1997) como función de puntuación. Dicha función tiene en cuenta las interacciones de *van der Waals*, la energía interna del ligando, y la formación de enlaces por puentes de hidrógeno. En este estudio logran muy buenos resultados, con aproximadamente el 80% de los complejos proteína-ligando con un *RMSD* por debajo de 2 Å con respecto a la estructura cristalográfica. El tiempo medio empleado oscila entre 1 y 5 min, sin tener en cuenta la inicialización de la proteína, por lo que se puede considerar que se mueve en un intervalo de tiempo similar al requerido por *CDOCK*. Un dato curioso es que cuando para el *docking* utilizan estructuras generadas por *CORINA*, como se hace en esta Tesis, la tasa de éxito decae aproximadamente en un 5%. Además, la tasa de éxito también depende en gran medida del protocolo utilizado, como ha podido comprobarse igualmente en las pruebas realizadas con *VSDMIP*.

Por último, se ha realizado una aproximación inicial a la etapa de optimización de *hits* mediante el desarrollo de una aplicación gráfica (*gCOMBINE*) para el uso de una herramienta *QSAR-3D* (*COMBINE*) y la evaluación e interpretación de los resultados obtenidos. Desde la aparición del método *COMBINE* en el año 1995 para el estudio de las diferencias en actividad en una serie de inhibidores de la fosfolipasa humana A_2 del fluido sinovial (Ortiz et al., 1995), dicho método ha sido aplicado en varios estudios relacionados con la unión de pequeñas moléculas a diferentes dianas proteicas (proteasa *HIV-1* (Perez et al., 1998), citocromo humano *P450 1A2* (Lozano et al., 2000), elastasa

neutrófila humana (Cuevas et al., 2001), transcriptasa inversa de *HIV-1* (Rodriguez-Barrios & Gago, 2004), acetilcolinesterasa (Guo et al., 2004; Lushington et al., 2005; Martin-Santamaria et al., 2004), haloalcano dehalogenasa *DhlA* (Kmunicek et al., 2001) y *LinB* (Damborsky et al., 2004)) y también para interacciones péptido-proteína (Schleinkofer et al., 2004; Wang & Wade, 2002), proteína-proteína (Tomic et al., 2007) y proteína-ADN (Tomic et al., 2000). El método ha adquirido cierta relevancia en el campo de las herramientas *QSAR-3D* y ha sido revisado en detalle por Wade et al. (Wade et al., 2004; Wade et al., 1998), Damborsky et al. (Damborsky et al., 2004), y más recientemente por Lushington et al. (Lushington et al., 2007), quienes incluso propusieron algunas ideas para mejorar las funcionalidades de *COMBINE* en el futuro. Un hito importante en su desarrollo fue la incorporación de múltiples estructuras en el análisis, lo cual permite el tener en cuenta, al menos en parte, la flexibilidad del receptor (Mou et al., 2006; Murcia & Ortiz, 2004; Pastor et al., 1997; Tomic et al., 2000). Todos estos estudios demuestran que es posible obtener modelos *COMBINE* de calidad empleando múltiples representaciones de la estructura del receptor, aunque se debe tener especial cuidado al tratar de realizar análisis cuantitativos debido a la dependencia conformacional de los modelos. Otro tema interesante relacionado con la variación estructural es el estudio conjunto de afinidad y selectividad por el uso de diferentes estructuras del receptor pertenecientes a la misma familia, lo cual puede proporcionar una importante guía para el diseño de fármacos. Son claros ejemplos de esto los estudios de Wang y Wade (Wang & Wade, 2001) sobre los análogos de ácidos siálicos y benzoicos que unen a los subtipos *N2* y *N9* de neuraminidasa, y el estudio de unión de ligandos a tres serín proteasas (tripsina, trombina y el factor de coagulación *Xa*) realizado por Murcia et al. (Murcia et al., 2006). Esta aproximación puede, en principio, ser extendida a un número arbitrario de receptores de la misma familia de proteínas.

El análisis *COMBINE* también puede relacionarse con un algoritmo de docking, como demostraron Murcia y Ortiz (Murcia & Ortiz, 2004), para mejorar las conformaciones de unión de los ligandos putativos al realizar cribados virtuales, para mejorar la capacidad predictiva de los modelos de regresión, y para incrementar los factores de enriquecimiento.

En base a todos estos trabajos, está claro que el análisis COMBINE ocupa una posición privilegiada como herramienta para guiar el diseño y la optimización de las moléculas candidatas a fármacos. Pero a pesar de ello, y como señalan Lushington et al. (Lushington et al., 2007), el número de grupos de investigación que utilizan *COMBINE*

es relativamente pequeño. De ahí la importancia de desarrollar *gCOMBINE*, un interfaz gráfico de usuario, sencillo y amigable, que permite manipular fácilmente los datos y los ficheros de entrada/salida. El hecho de poner esta herramienta disponible libremente para la comunidad científica, tal como se hizo en su día con *COMBINE*, potenciará el uso de este método cuya utilidad ha sido ampliamente demostrada en el campo del diseño de fármacos *in silico*.

Esta Tesis ha abarcado el proceso de la búsqueda de nuevos fármacos, desde que se cuenta con sólo una diana de interés terapéutico y una quimioteca, hasta que finalmente se seleccionan moléculas para ser probadas experimentalmente. Además, también se ha proporcionado una importante herramienta para la etapa de optimización de *hits*. La infraestructura ha comenzado a dar sus frutos, ya que se han obtenido moléculas que muestran un nivel de actividad interesante para su diana particular, habiéndose patentado algunas de ellas. Aun así, la búsqueda de nuevos fármacos sigue siendo un proceso largo, incluso empleando técnicas computacionales. Esto, junto con el hecho de que suelen ser proyectos altamente confidenciales debido a la inversión económica que requieren, hace que la presentación exacta de resultados se retrase hasta que estos están protegidos por una patente. Por este motivo, en la presente Tesis sólo se han podido presentar resultados completos para el cribado virtual con *MGMT*, a pesar de que la metodología desarrollada se continúa aplicando con otras dianas para las que se están obteniendo excelentes resultados.

Aun así los resultados podrían ser mejores, tanto en eficacia a la hora de distinguir la actividad de los compuestos como en el tiempo y los recursos empleados. Se ha observado que para unos tipos de receptores se obtiene mejores resultados que para otros, lo que indica que deberían refinarse los modelos para esos casos. También se ha observado que hay ciertas etapas del cribado que podrían automatizarse más, como puede ser por ejemplo la preparación del receptor. Otro de los aspectos a mejorar sería el tema de la flexibilidad del receptor, problema complejo de resolver, no sólo por el hecho de obtener un conjunto con las estructuras más representativas que puede tener un receptor, o reproducir los cambios que sufre al unirse con un ligando, ya que incluso resolviendo estos temas queda el problema de poder abordar computacionalmente el cribado debido al aumento en el número de combinaciones: a las que ya se tenían para un cribado normal, habría que multiplicarlas por las *N* estructuras del receptor. En la actualidad se pueden obtener diferentes estructuras a partir de datos de dinámica molecular o bien utilizar la técnica de análisis de modos normales (*NMA – Normal*

Mode Analysis (Cavasotto et al., 2005)) para predecir el movimiento de la proteína. También resulta muy importante la predicción de propiedades físico-químicas para usar estos datos tanto como filtro previo de la quimioteca como para la etapa de optimización de *hits*. Conocer como afectan las modificaciones en el ligando a sus propiedades *ADME* y a su toxicidad ayudará a obtener inhibidores más eficientes y seguros. El disponer de las propiedades de los ligandos, tanto físico-químicos como estructurales, permitirá a su vez el poder agruparlos y clasificarlos, de modo que se facilite la búsqueda de compuestos análogos a otros ya conocidos (usando *fingerprints* por ejemplo), o bien realizar aproximaciones no secuenciales del cribado virtual. Y para explorar en mayor profundidad el espacio químico de moléculas pequeñas, el empleo de técnicas basadas en fragmentos se convierte en algo fundamental. Gracias a estas técnicas se pueden llegar a obtener moléculas más potentes y más optimizadas. Por último, y como sucedía en el caso del programa *COMBINE*, será importante de cara al futuro el desarrollo de una interfaz gráfica de usuario desde la que se puedan utilizar fácilmente todas las funcionalidades ofrecidas por *VSDMIP*, de modo que el mayor número de usuarios pueda beneficiarse de esta plataforma y colaborar directa o indirectamente en su desarrollo y mejora.

CONCLUSIONES

6. Conclusiones

Las principales conclusiones que se pueden derivar del trabajo presentado en esta Tesis se pueden resumir en los siguientes puntos:

1. Se ha desarrollado una plataforma computacional donde implementar e integrar diferentes protocolos y herramientas relacionados con el proceso de obtención de nuevos fármacos. Esto permite automatizar fácilmente las tareas implicadas en el cribado virtual de quimiotecas y configurar un protocolo que se adapte a los requerimientos de precisión y tiempo en función de los recursos disponibles. Además, la información generada se mantiene de una manera organizada, facilitando tanto su obtención y análisis como la comparación entre diferentes protocolos y herramientas. Por último, se ha demostrado su utilidad en diversas aplicaciones reales.

2. El método *ISM* proporciona un nuevo paso hacia la incorporación realista de un modelo de solvente en *dockings* a gran escala.

3. Cuanto más precisa es una función de puntuación para el *docking* proteína-ligando resulta también más sensible a pequeñas variaciones en la conformación del ligando, por lo que en ocasiones es difícil alcanzar una solución de *docking* similar a la nativa cuando se utilizan conjuntos de conformaciones pre-calculadas del ligando para simular su flexibilidad. Esto implica que en el proceso de *docking* se debe permitir la realización de pequeñas modificaciones en la conformación del ligando para ajustarse mejor a la forma del centro activo.

4. El uso de interfaces gráficos de usuario para el empleo de métodos *QSAR-3D*, como es el programa *COMBINE*, facilita la preparación, generación, interpretación y reutilización de modelos que ayuden en el desarrollo de *leads* a partir de *hits*.

BIBLIOGRAFÍA

7. Bibliografía

Agrafiotis, DK (2003) Stochastic proximity embedding. *J Comput Chem* 24(10)**:** 1215-1221.

Agrafiotis, DK, Gibbs, AC, Zhu, F, Izrailev, S, Martin, E (2007) Conformational sampling of bioactive molecules: a comparative study. *J Chem Inf Model* 47(3)**:** 1067-1086.

Agrafiotis, DK, Xu, H (2002) A self-organizing principle for learning nonlinear manifolds. *Proc Natl Acad Sci U S A* 99(25)**:** 15869-15872.

Alvarez, JC (2004) High-throughput docking as a source of novel drug leads. *Curr Opin Chem Biol* 8(4)**:** 365-370.

Allen, FH, Davies, JE, Galloy, JJ, Johnson, O, Kennard, O, Macrae, CF, Mitchell, JF, Smith , JM, Watson, DG (1991) The Developtment of Version 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* 31**:** 187-204.

Antolí, B, Benito, D, Cárdenas-Montes, M, Castejón, F, Castellanos, C, Castelo, V, de Alfonso, C, de Miguel, T, Gavela, R, Hernández, V, Herraiz, L, Ibar, J, Jiménez, L, Ramos, R, Rivero, A, Sáenz, JF, Serrano, F, Rubio, M, Tarancón, A, Vuillemin, P (2008) Citizen Volunteer Infrastructure for Computing. *Proceedings of 2nd Ibergrid Conference, Porto, Portugal*.

Arris, CE, Boyle, FT, Calvert, AH, Curtin, NJ, Endicott, JA, Garman, EF, Gibson, AE, Golding, BT, Grant, S, Griffin, RJ, Jewsbury, P, Johnson, LN, Lawrie, AM, Newell, DR, Noble, ME, Sausville, EA, Schultz, R, Yu, W (2000) Identification of novel purine and pyrimidine cyclin-dependent kinase inhibitors with distinct molecular interactions and tumor cell growth inhibition profiles. *J Med Chem* 43(15)**:** 2797-2804.

Badia, RM, Labarta, J, Sirvent, R, M., PJ, Cela, JM, Grima, R (2003) Programming Grid Applications with GRID superscalar. *Journal of GRID Computing* 1(2)**:** 151-170.

Baker, D, Sali, A (2001) Protein structure prediction and structural genomics. *Science* 294(5540)**:** 93-96.

Barzilay, G, Mol, CD, Robson, CN, Walker, LJ, Cunningham, RP, Tainer, JA, Hickson, ID (1995) Identification of critical active-site residues in the multifunctional human DNA repair enzyme HAP1. *Nat Struct Biol* 2(7)**:** 561-568.

Bashford, D, Case, DA (2000) Generalized born models of macromolecular solvation effects. *Annu Rev Phys Chem* 51**:** 129-152.

Bayly, CE, Cieplak, P, Cornell, WD, Kollman, PA (1993) A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J Phys Chem* 97(40)**:** 10269-10280.

Beernink, PT, Segelke, BW, Hadi, MZ, Erzberger, JP, Wilson, DM, 3rd, Rupp, B (2001) Two divalent metal ions in the active site of a new crystal form of human apurinic/apyrimidinic endonuclease, Ape1: implications for the catalytic mechanism. *J Mol Biol* 307(4)**:** 1023-1034.

Bissantz, C, Folkers, G, Rognan, D (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. *J Med Chem* 43(25)**:** 4759-4767.

Bliznyuk, AA, E.Gready, J (1999) Simple method for locating possible ligand binding sites on protein surfaces. *J. Comput. Chem.* 20**:** 983-988.

Bohm, HJ (1998) Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J Comput Aided Mol Des* 12(4)**:** 309-323.

Borodina, YV, Bolton, E, Fontaine, F, Bryant, SH (2007) Assessment of conformational ensemble sizes necessary for specific resolutions of coverage of conformational space. *J Chem Inf Model* 47(4)**:** 1428-1437.

Boström, J, Greenwood, JR, Gottfries, J (2003) Assessing the performance of OMEGA with respect to retrieving bioactive conformations. *J Mol Graph Model* 21(5)**:** 449-462.

Bower, MJ, Cohen, FE, Dunbrack, RL, Jr. (1997) Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. *J Mol Biol* 267(5)**:** 1268-1282.

Brent, TP, Houghton, PJ, Houghton, JA (1985) O6-Alkylguanine-DNA alkyltransferase activity correlates with the therapeutic response of human rhabdomyosarcoma xenografts to 1-(2-chloroethyl)-3-(trans-4-methylcyclohexyl)-1-nitrosourea. *Proc Natl Acad Sci U S A* 82(9)**:** 2985-2989.

Brooks, BR, Bruccoleri, RE, Olafson, BD, States, DJ, Swaminathan, S, Karplus, MJ (1983) A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* 4**:** 187-217.

Bruning, JB, Shamoo, Y (2004) Structural and thermodynamic analysis of human PCNA with peptides derived from DNA polymerase-delta p66 subunit and flap endonuclease-1. *Structure* 12(12)**:** 2209-2219.

Burmeister, WP, Henrissat, B, Bosso, C, Cusack, S, Ruigrok, RW (1993) Influenza B virus neuraminidase can synthesize its own inhibitor. *Structure* 1(1)**:** 19-26.

Canutescu, AA, Shelenkov, AA, Dunbrack, RL, Jr. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci* 12(9)**:** 2001-2014.

Case, DA, Cheatham, TE, 3rd, Darden, T, Gohlke, H, Luo, R, Merz, KM, Jr., Onufriev, A, Simmerling, C, Wang, B, Woods, RJ (2005) The Amber biomolecular simulation programs. *J Comput Chem* 26(16)**:** 1668-1688.

Cavasotto, CN, Abagyan, RA (2004) Protein flexibility in ligand docking and virtual screening to protein kinases. *J Mol Biol* 337(1)**:** 209-225.

Cavasotto, CN, Kovacs, JA, Abagyan, RA (2005) Representing receptor flexibility in ligand docking through relevant normal modes. *J Am Chem Soc* 127(26)**:** 9632-9640.

Claussen, H, Buning, C, Rarey, M, Lengauer, T (2001) FlexE: efficient molecular docking considering protein structure variations. *J Mol Biol* 308(2)**:** 377-395.

Congreve, M, Chessari, G, Tisi, D, Woodhead, AJ (2008) Recent developments in fragment-based drug discovery. *J Med Chem* 51(13)**:** 3661-3680.

Cordell, SC, Robinson, EJ, Lowe, J (2003) Crystal structure of the SOS cell division inhibitor SulA and in complex with FtsZ. *Proc Natl Acad Sci U S A* 100(13)**:** 7889-7894.

Cornell, WD, Cieplak, P, Bayly, CI, Gould, IR, Merz, KM, Ferguson, DM, Spellmeyer, DC, Fox, T, Caldwell, JW, Kollman, PA (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* 117**:** 5179-5197.

Cramer, RD, Patterson, DE, Bunce, JD (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *JACS* 110**:** 5959-5967.

Cuevas, C, Pastor, M, Perez, C, Gago, F (2001) Comparative binding energy (COMBINE) analysis of human neutrophile elastase inhibition by Pyridone-containing Trifluoromethylketones. *Comb Chem High Throughput Screen* 4**:** 627-642.

Chang, CI, Xu, BE, Akella, R, Cobb, MH, Goldsmith, EJ (2002) Crystal structures of MAP kinase p38 complexed to the docking sites on its nuclear substrate MEF2A and activator MKK3b. *Mol Cell* 9(6)**:** 1241-1249.

Charifson, PS, Corkery, JJ, Murcko, MA, Walters, WP (1999) Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J Med Chem* 42(25)**:** 5100-5109.

Chrencik, JE, Brooun, A, Zhang, H, Mathews, II, Hura, GL, Foster, SA, Perry, JJ, Streiff, M, Ramage, P, Widmer, H, Bokoch, GM, Tainer, JA, Weckbecker, G, Kuhn, P (2008) Structural basis of guanine nucleotide exchange mediated by the T-cell essential Vav1. *J Mol Biol* 380(5)**:** 828-843.

Damborsky, J, Kmunicek, J, Jedlicka, T, Luengo, S, Gago, F, Ortiz, AR, Wade, RC (2004) Rational Redesign of Haloalkane Dehalogenases Guided by Comparative Binding Energy Analysis. In: *Enzyme Functionality: Design, Engineering and Screening*, Svendsen, A, Dekker, M (eds), pp 79-96. New York.

Daniels, DS, Woo, TT, Luu, KX, Noll, DM, Clarke, ND, Pegg, AE, Tainer, JA (2004) DNA binding and nucleotide flipping by the human DNA repair protein AGT. *Nat Struct Mol Biol* 11(8)**:** 714-720.

Darden, T, York, D, Pedersen, L (1993) Particle mesh Ewald: An N*log(N) method for Ewald sums in large systems. *J. Chem. Phys.* 98**:** 10089-10092.

Davis, AM, Teague, SJ, Kleywegt, GJ (2003) Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew Chem Int Ed Engl* 42(24)**:** 2718-2736.

Deb, K, Mohan, M, Mishra, S (2003) *A fast multi-objective evolutionary algorithm for finding well-spread Pareto-optimal solutions*. Kangal Rep No.2003002, Indian Institute of Technology: Kanpur, India.

Delgado, P, Cubelos, B, Calleja, E, Martinez-Martin, N, Cipres, A, Merida, I, Bellas, C, Bustelo, XR, Alarcon, B (2009) Essential function for the GTPase TC21 in homeostatic antigen receptor signaling. *Nat Immunol* 10(8)**:** 880-888.

Desai, A, Mitchison, TJ (1998) Tubulin and FtsZ structures: functional and therapeutic implications. *Bioessays* 20(7)**:** 523-527.

Dewar, MJS, Thiel, W (1977) MIND0/3 Study of the Addition of Singlet Oxygen (1deltagO2) to 1,3-Butadiene. *J. Am. Chem. Soc.* 99**:** 2338-2339.

Diller, DJ, Merz, KM, Jr. (2002) Can we separate active from inactive conformations? *J Comput Aided Mol Des* 16(2)**:** 105-112.

Dobson, CM (2004) Chemical space and biology. *Nature* 432(7019)**:** 824-828.

Dunbrack, RL, Jr. (1999) Comparative modeling of CASP3 targets using PSI-BLAST and SCWRL. *Proteins* Suppl 3**:** 81-87.

Eldridge, MD, Murray, CW, Auton, TR, Paolini, GV, Mee, RP (1997) Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J Comput Aided Mol Des* 11(5)**:** 425-445.

Erlanson, DA (2006) Fragment-based lead discovery: a chemical update. *Curr Opin Biotechnol* 17(6)**:** 643-652.

Esteller, M, Garcia-Foncillas, J, Andion, E, Goodman, SN, Hidalgo, OF, Vanaclocha, V, Baylin, SB, Herman, JG (2000) Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *N Engl J Med* 343(19)**:** 1350-1354.

Gerson, SL (2002) Clinical relevance of MGMT in the treatment of cancer. *J Clin Oncol* 20(9)**:** 2388-2399.

Gerson, SL (2004) MGMT: its role in cancer aetiology and cancer therapeutics. *Nat Rev Cancer* 4(4)**:** 296-307.

Gilson, MK, Given, JA, Bush, BL, McCammon, JA (1997) The statistical-thermodynamic basis for computation of binding affinities: a critical review. *Biophys J* 72(3)**:** 1047-1069.

Gohlke, H, Hendlich, M, Klebe, G (2000) Knowledge-based scoring function to predict protein-ligand interactions. *J Mol Biol* 295(2)**:** 337-356.

Good, AC, Cheney, DL (2003) Analysis and optimization of structure-based virtual screening protocols (1): exploration of ligand conformational sampling techniques. *J Mol Graph Model* 22(1)**:** 23-30.

Goodford, PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 28(7)**:** 849-857.

Grosdidier, A, Zoete, V, Michielin, O (2007) EADock: docking of small molecules into protein active sites with a multiobjective evolutionary optimization. *Proteins* 67(4)**:** 1010-1025.

Guo, J, Hurley, MM, Wright, JB, Lushington, GH (2004) A docking score function for estimating ligand-protein interactions: application to acetylcholinesterase inhibition. *J Med Chem* 47(22)**:** 5492-5500.

Halgren, TA (1996) Merck Molecular Force Field. 1. Basis, Form, Scope, Parametrization, and Performance of MMFF94. *J. Comput. Chem.* 17**:** 490-519.

Hann, MM, Leach, AR, Harper, G (2001) Molecular complexity and its impact on the probability of finding leads for drug discovery. *J Chem Inf Comput Sci* 41(3)**:** 856-864.

Hansch, C, Hoekman, D, Gao, H (1996) Comparative QSAR: Toward a Deeper Understanding of Chemicobiological Interactions. *Chem Rev* 96(3)**:** 1045-1076.

Hansch, C, Hoekman, D, Leo, A, Weininger, D, Selassie, CD (2002) Chem-bioinformatics: comparative QSAR at the interface between chemistry and biology. *Chem Rev* 102(3)**:** 783-812.

Hartshorn, MJ, Verdonk, ML, Chessari, G, Brewerton, SC, Mooij, WT, Mortenson, PN, Murray, CW (2007) Diverse, high-quality test set for the validation of protein-ligand docking performance. *J Med Chem* 50(4)**:** 726-741.

Hassan, M, Brown, RD, Varma-O'brien, S, Rogers, D (2006) Cheminformatics analysis and learning in a data pipelining environment. *Mol Divers* 10(3)**:** 283-299.

Hassan, SA, Guarnieri, F, Mehler, EL (2000a) Characterization of hydrogen bonding in a continuum solvent model. *J Phys Chem B* 104**:** 6490-6498.

Hassan, SA, Guarnieri, F, Mehler, EL (2000b) General treatment of solvent effects based on screened Coulomb potentials. *J Phys Bhem B* 104**:** 6478-6489.

Hassan, SA, Mehler, EL (2002) A critical analysis of continuum electrostatics: the screened Coulomb potential--implicit solvent model and the study of the alanine dipeptide and discrimination of misfolded structures of proteins. *Proteins* 47(1)**:** 45-61.

Haydon, DJ, Stokes, NR, Ure, R, Galbraith, G, Bennett, JM, Brown, DR, Baker, PJ, Barynin, VV, Rice, DW, Sedelnikova, SE, Heal, JR, Sheridan, JM, Aiwale, ST, Chauhan, PK, Srivastava, A, Taneja, A, Collins, I, Errington, J, Czaplewski, LG (2008) An inhibitor of FtsZ with potent and selective anti-staphylococcal activity. *Science* 321(5896)**:** 1673-1675.

Hegi, ME, Diserens, AC, Gorlia, T, Hamou, MF, de Tribolet, N, Weller, M, Kros, JM, Hainfellner, JA, Mason, W, Mariani, L, Bromberg, JE, Hau, P, Mirimanoff, RO, Cairncross, JG, Janzer, RC, Stupp, R (2005) MGMT gene silencing and benefit from temozolomide in glioblastoma. *N Engl J Med* 352(10)**:** 997-1003.

Ho, CM, Marshall, GR (1990) Cavity search: an algorithm for the isolation and display of cavity-like binding regions. *J Comput Aided Mol Des* 4(4)**:** 337-354.

Holloway, MK, Wai, JM, Halgren, TA, Fitzgerald, PM, Vacca, JP, Dorsey, BD, Levin, RB, Thompson, WJ, Chen, LJ, deSolms, SJ (1995) A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site. *J Med Chem* 38(2)**:** 305-317.

Honig, B, Nicholls, A (1995) Classical electrostatics in biology and chemistry. *Science* 268(5214)**:** 1144-1149.

Hopkins, AL, Groom, CR, Alex, A (2004) Ligand efficiency: a useful metric for lead selection. *Drug Discov Today* 9(10)**:** 430-431.

Huang, N, Shoichet, BK, Irwin, JJ (2006) Benchmarking sets for molecular docking. *J Med Chem* 49(23)**:** 6789-6801.

Huey, R, Morris, GM, Olson, AJ, Goodsell, DS (2007) A semiempirical free energy force field with charge-based desolvation. *J Comput Chem* 28(6)**:** 1145-1152.

Irwin, JJ, Shoichet, BK (2005) ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model* 45(1)**:** 177-182.

Irwin, JJ, Shoichet, BK, Mysinger, MM, Huang, N, Colizzi, F, Wassam, P, Cao, Y (2009) Automated docking screens: a feasibility study. *J Med Chem* 52(18)**:** 5712-5720.

Jacobsson, M, Liden, P, Stjernschantz, E, Bostrom, H, Norinder, U (2003) Improving structure-based virtual screening by multivariate analysis of scoring data. *J Med Chem* 46(26)**:** 5781-5789.

Johnson, M, Maggiora, GM (1990) *Concepts and Applications of Molecular Similarity*. Wiley: New York.

Jones, G, Willett, P, Glen, RC, Leach, AR, Taylor, R (1997) Development and validation of a genetic algorithm for flexible docking. *J Mol Biol* 267(3)**:** 727-748.

Jorgensen, W, Chandrasekhar, J, Madura, J, Impey, R, Klein, M (1983) Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79**:** 926-935.

Jorgensen, WL (2004) The many roles of computation in drug discovery. *Science* 303(5665)**:** 1813-1818.

Kasimova, MR, Kristensen, SM, Howe, PW, Christensen, T, Matthiesen, F, Petersen, J, Sorensen, HH, Led, JJ (2002) NMR studies of the backbone flexibility and structure of human growth hormone: a comparison of high and low pH conformations. *J Mol Biol* 318(3)**:** 679-695.

Klebe, G (2006) Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov Today* 11(13-14)**:** 580-594.

Klebe, G, Abraham, U, Mietzner, T (1994) Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J Med Chem* 37(24)**:** 4130-4146.

Klebe, G, Mietzner, T (1994) A fast and efficient method to generate biologically relevant conformations. *J Comput Aided Mol Des* 8(5)**:** 583-606.

Klon, AE, Glick, M, Thoma, M, Acklin, P, Davies, JW (2004) Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results. *J Med Chem* 47(11)**:** 2743-2749.

Kmunicek, J, Luengo, S, Gago, F, Ortiz, AR, Wade, RC, Damborsky, J (2001) Comparative binding energy analysis of the substrate specificity of haloalkane dehalogenase from Xanthobacter autotrophicus GJ10. *Biochemistry* 40(30)**:** 8905-8917.

Knox, AJ, Meegan, MJ, Carta, G, Lloyd, DG (2005) Considerations in compound database preparation--"hidden" impact on virtual screening results. *J Chem Inf Model* 45(6)**:** 1908-1919.

Kollman, PA, Massova, I, Reyes, C, Kuhn, B, Huo, S, Chong, L, Lee, M, Lee, T, Duan, Y, Wang, W, Donini, O, Cieplak, P, Srinivasan, J, Case, DA, Cheatham, TE, 3rd (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Accounts of chemical research* 33(12)**:** 889-897.

Kontoyianni, M, McClellan, LM, Sokol, GS (2004) Evaluation of docking performance: comparative data on docking algorithms. *J Med Chem* 47(3)**:** 558-565.

Kryger, G, Silman, I, Sussman, JL (1999) Structure of acetylcholinesterase complexed with E2020 (Aricept): implications for the design of new anti-Alzheimer drugs. *Structure* 7(3)**:** 297-307.

Kuntz, ID, Blaney, JM, Oatley, SJ, Langridge, R, Ferrin, TE (1982) A geometric approach to macromolecule-ligand interactions. *J Mol Biol* 161(2)**:** 269-288.

Lahana, R (1999) How many leads from HTS? *Drug Discov Today* 4(10)**:** 447-448.

Lattman, EE (1972) Optimal sampling of the rotation function. The molecular replacement method. *Gordon and Breach, Science Publishers Inc.***:** 179-185.

Laumanns, M, Thiele, L, Deb, K, Zitzler, E (2002) Combining convergence and diversity in evolutionary multiobjective optimization. *Evol Comput* 10(3)**:** 263-282.

Leach, AR, Shoichet, BK, Peishoff, CE (2006) Prediction of protein-ligand interactions. Docking and scoring: successes and gaps. *J Med Chem* 49(20)**:** 5851-5855.

Lehtovuori, PT, Nyronen, TH (2006) SOMA--workflow for small molecule property calculations on a multiplatform computing grid. *J Chem Inf Model* 46(2)**:** 620-625.

Lipinski, CA, Lombardo, F, Dominy, BW, Feeney, PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* 46(1-3)**:** 3-26.

Liu, H-Y, Zou, X, Kuntz, ID (2004) Pairwise GB/SA Scoring Functions for Structure-based Drug Design. *J. Phys. Chem. B.* 108**:** 5453-5462.

Liu, HY, Zou, X (2006) Electrostatics of ligand binding: parametrization of the generalized Born model and comparison with the Poisson-Boltzmann approach. *J Phys Chem B* 110(18)**:** 9304-9313.

Liu, X, Bai, F, Ouyang, S, Wang, X, Li, H, Jiang, H (2009) Cyndi: a multi-objective evolution algorithm based method for bioactive molecular conformational generation. *BMC Bioinformatics* 10**:** 101.

López-Romero, PG, M. J.; Gómez-Puertas, P.; Valencia, A (2004) Prediction of Functional Sites in Proteins by Evolutionary Methods. In: *Principles and Practice. Methods in Proteome and Protein Analysis*, R.M. Kamp, JJC, T.Choli-Papadopoulou (Eds.) Springer-Verlag Berlin Heidelberg (ed) Vol. 22, pp 320-340.

Lorber, DM, Shoichet, BK (1998) Flexible ligand docking using conformational ensembles. *Protein Sci* 7(4)**:** 938-950.

Lozano, JJ, Pastor, M, Cruciani, G, Gaedt, K, Centeno, NB, Gago, F, Sanz, F (2000) 3D-QSAR methods on the basis of ligand-receptor complexes. Application of COMBINE and GRID/GOLPE methodologies to a series of CYP1A2 ligands. *J Comput Aided Mol Des* 14(4)**:** 341-353.

Lushington, GH, Guo, JX, Wang, JL (2007) Whither combine? New opportunities for receptor-based QSAR. *Curr Med Chem* 14(17)**:** 1863-1877.

Lushington, GH, Wallace, NM, Guo, JX (2005) Reliable Prescreening of Candidate NerveAgent Prophylaxes via 3D QSAR. *DTIC Monitor Series***:** 1-28.

Madhusudan, S, Hickson, ID (2005) DNA repair inhibition: a selective tumour targeting strategy. *Trends Mol Med* 11(11)**:** 503-511.

Maga, G, Hubscher, U (2003) Proliferating cell nuclear antigen (PCNA): a dancer with many partners. *J Cell Sci* 116(Pt 15)**:** 3051-3060.

Maignan, S, Guilloteau, JP, Pouzieux, S, Choi-Sledeski, YM, Becker, MR, Klein, SI, Ewing, WR, Pauls, HW, Spada, AP, Mikol, V (2000) Crystal structures of human factor Xa complexed with potent inhibitors. *J Med Chem* 43(17)**:** 3226-3232.

Maiorov, V, Sheridan, RP (2005) Enhanced virtual screening by combined use of two docking methods: getting the most on a limited budget. *J Chem Inf Model* 45(4)**:** 1017-1023.

Majeux, N, Scarsi, M, Caflisch, A (2001) Efficient electrostatic solvation model for protein-fragment docking. *Proteins* 42(2)**:** 256-268.

Margison, GP, Povey, AC, Kaina, B, Santibanez Koref, MF (2003) Variability and regulation of O6-alkylguanine-DNA alkyltransferase. *Carcinogenesis* 24(4)**:** 625-635.

Martin-Santamaria, S, Munoz-Muriedas, J, Luque, FJ, Gago, F (2004) Modulation of binding strength in several classes of active site inhibitors of acetylcholinesterase studied by comparative binding energy analysis. *J Med Chem* 47(18)**:** 4471-4482.

Massova, I, Kollman, PA (2000) Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspect Drug Discov Des* 18**:** 113-135.

Mattern, J, Eichhorn, U, Kaina, B, Volm, M (1998) O6-methylguanine-DNA methyltransferase activity and sensitivity to cyclophosphamide and cisplatin in human lung tumor xenografts. *Int J Cancer* 77(6)**:** 919-922.

McDonald, IK, Thornton, JM (1994) Satisfying hydrogen bonding potential in proteins. *J Mol Biol* 238(5)**:** 777-793.

McGovern, SL, Shoichet, BK (2003) Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. *J Med Chem* 46(14)**:** 2895-2907.

McInnes, C (2007) Virtual screening strategies in drug discovery. *Curr Opin Chem Biol* 11(5)**:** 494-502.

McLachlan, AD (1979) Gene duplications in the structural evolution of chymotrypsin. *J Mol Biol* 128(1)**:** 49-79.

Mehler, EL, Solmajer, T (1991) Electrostatic effects in proteins: comparison of dielectric and charge models. *Protein Eng* 4(8)**:** 903-910.

Metropolis, N, Rosenbluth, AW, Rosenbluth, AH, Teller, AH, Teller, E (1953) Fast Computing Machines. *J Chem Phys* 21**:** 1087-1092.

Miteva, MA, Lee, WH, Montes, MO, Villoutreix, BO (2005) Fast structure-based virtual ligand screening combining FRED, DOCK, and Surflex. *J Med Chem* 48(19)**:** 6012-6022.

Morreale, A, Gil-Redondo, R, Ortiz, AR (2007) A new implicit solvent model for protein-ligand docking. *Proteins* 67(3)**:** 606-616.

Morris, GM, Goodsell, DS, Halliday, RS, Huey, R, Hart, WE, Belew, RK, Olson, AJ (1998) Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J Comput Chem*.

Mou, TC, Gille, A, Suryanarayana, S, Richter, M, Seifert, R, Sprang, SR (2006) Broad specificity of mammalian adenylyl cyclase for interaction with 2',3'-substituted purine- and pyrimidine nucleotide inhibitors. *Mol Pharmacol* 70(3)**:** 878-886.

Moya-Garcia, AA, Pino-Angeles, A, Gil-Redondo, R, Morreale, A, Sanchez-Jimenez, F (2009) Structural features of mammalian histidine decarboxylase reveal the basis for specific inhibition. *Br J Pharmacol* 157(1)**:** 4-13.

Moya-Garcia, AA, Pino-Angeles, A, Sanchez-Jimenez, F (2006) New structural insights to help in the search for selective inhibitors of mammalian pyridoxal 5'-phosphate-dependent histidine decarboxylase . 4. Synthesis, metabolism and release of histamine. *Inflamm Res* 55 Suppl 1**:** S55-56.

Moya-Garcia, AA, Ruiz-Pernia, J, Marti, S, Sanchez-Jimenez, F, Tunon, I (2008) Analysis of the decarboxylation step in mammalian histidine decarboxylase. A computational study. *J Biol Chem* 283(18)**:** 12393-12401.

Muegge, I, Martin, YC (1999) A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J Med Chem* 42(5)**:** 791-804.

Munos, B (2009) Lessons from 60 years of pharmaceutical innovation. *Nat Rev Drug Discov* 8(12)**:** 959-968.

Murcia, M, Morreale, A, Ortiz, AR (2006) Comparative binding energy analysis considering multiple receptors: a step toward 3D-QSAR models for multiple targets. *J Med Chem* 49(21)**:** 6241-6253.

Murcia, M, Ortiz, AR (2004) Virtual screening with flexible docking and COMBINE-based models. Application to a series of factor Xa inhibitors. *J Med Chem* 47(4)**:** 805-820.

Murray, CW, Baxter, CA, Frenkel, AD (1999) The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. *J Comput Aided Mol Des* 13(6)**:** 547-562.

Nelder, JAM, R. (1965) A simplex method for function minimization. *Computer J.* 7**:** 308-313.

Neyman, J, Pearson, ES (1933a) On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. R. Soc, London, Ser. A* 231**:** 289-337.

Neyman, J, Pearson, ES (1933b) The testing of statistical hypotheses in relation to probabilities a priori. *Proc. Cambridge Philos. Soc.* 20**:** 492-510.

Oliva, MA, Cordell, SC, Lowe, J (2004) Structural insights into FtsZ protofilament formation. *Nat Struct Mol Biol* 11(12)**:** 1243-1250.

Ortiz, AR, Pisabarro, MT, Gago, F, Wade, RC (1995) Prediction of drug binding affinities by comparative binding energy analysis. *J Med Chem* 38(14)**:** 2681-2691.

Oshiro, C, Bradley, EK, Eksterowicz, J, Evensen, E, Lamb, ML, Lanctot, JK, Putta, S, Stanton, R, Grootenhuis, PD (2004) Performance of 3D-database molecular docking studies into homology models. *J Med Chem* 47(3)**:** 764-767.

Pastor, M, Perez, C, Gago, F (1997) Simulation of alternative binding modes in a structure-based QSAR study of HIV-1 protease inhibitors. *J Mol Graph Model* 15(6)**:** 364-371, 389.

Paunesku, T, Mittal, S, Protic, M, Oryhon, J, Korolev, SV, Joachimiak, A, Woloschak, GE (2001) Proliferating cell nuclear antigen (PCNA): ringmaster of the genome. *Int J Radiat Biol* 77(10)**:** 1007-1021.

Pepponi, R, Marra, G, Fuggetta, MP, Falcinelli, S, Pagani, E, Bonmassar, E, Jiricny, J, D'Atri, S (2003) The effect of O6-alkylguanine-DNA alkyltransferase and mismatch repair activities on the sensitivity of human melanoma cells to temozolomide, 1,3-bis(2-chloroethyl)1-nitrosourea, and cisplatin. *J Pharmacol Exp Ther* 304(2)**:** 661-668.

Peregrin, S, Jurado-Pueyo, M, Campos, PM, Sanz-Moreno, V, Ruiz-Gomez, A, Crespo, P, Mayor, F, Jr., Murga, C (2006) Phosphorylation of p38 by GRK2 at the docking groove unveils a novel mechanism for inactivating p38MAPK. *Curr Biol* 16(20)**:** 2042-2047.

Perez, C, Ortiz, AR (2001) Evaluation of docking functions for protein-ligand docking. *J Med Chem* 44(23)**:** 3768-3785.

Perez, C, Pastor, M, Ortiz, AR, Gago, F (1998) Comparative binding energy analysis of HIV-1 protease inhibitors: incorporation of solvent effects and validation as a powerful tool in receptor-based drug design. *J Med Chem* 41(6)**:** 836-852.

Perola, E, Charifson, PS (2004) Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J Med Chem* 47(10)**:** 2499-2510.

Ponder, JW, Richards, FM (1987) Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol* 193(4)**:** 775-791.

Prasanth, SG, Prasanth, KV, Siddiqui, K, Spector, DL, Stillman, B (2004) Human Orc2 localizes to centrosomes, centromeres and heterochromatin during chromosome inheritance. *Embo J* 23(13)**:** 2651-2663.

Ramesha, CS (2000) How many leads from HTS? - Comment. *Drug Discov Today* 5(2)**:** 43-44.

Rarey, M, Kramer, B, Lengauer, T, Klebe, G (1996) A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* 261(3)**:** 470-489.

Rocchia, W, Sridharan, S, Nicholls, A, Alexov, E, Chiabrera, A, Honig, B (2002) Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J Comput Chem* 23(1)**:** 128-137.

Roche, O, Kiyama, R, Brooks, CL, 3rd (2001) Ligand-protein database: linking protein-ligand complex structures to binding data. *J Med Chem* 44(22)**:** 3592-3598.

Rodriguez-Barrios, F, Gago, F (2004) Chemometrical identification of mutations in HIV-1 reverse transcriptase conferring resistance or enhanced sensitivity to arylsulfonylbenzonitriles. *J Am Chem Soc* 126(9)**:** 2718-2719.

Ruiz, FM, Gil-Redondo, R, Morreale, A, Ortiz, AR, Fabrega, C, Bravo, J (2008) Structure-based discovery of novel non-nucleosidic DNA alkyltransferase inhibitors: virtual screening and in vitro and in vivo activities. *J Chem Inf Model* 48(4)**:** 844-854.

Ryckaert, J, Ciccotti, G, Berendsen, H (1977) Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comp. Phys.* 23**:** 327-341.

Sadowski, J, Bostrom, J (2006) MIMUMBA revisited: torsion angle rules for conformer generation derived from X-ray structures. *J Chem Inf Model* 46(6)**:** 2305-2309.

Sakakura, C, Hagiwara, A, Tsujimoto, H, Ozaki, K, Sakakibara, T, Oyama, T, Ogaki, M, Takahashi, T (1994) Inhibition of gastric cancer cell proliferation by antisense oligonucleotides targeting the messenger RNA encoding proliferating cell nuclear antigen. *Br J Cancer* 70(6)**:** 1060-1066.

SciTegic, Inc. 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA.

Schleinkofer, K, Wiedemann, U, Otte, L, Wang, T, Krause, G, Oschkinat, H, Wade, RC (2004) Comparative structural and energetic analysis of WW domain-peptide interactions. *J Mol Biol* 344(3)**:** 865-881.

Schneider, G, Fechner, U (2005) Computer-based de novo design of drug-like molecules. *Nat Rev Drug Discov* 4(8)**:** 649-663.

Shiau, AK, Barstad, D, Loria, PM, Cheng, L, Kushner, PJ, Agard, DA, Greene, GL (1998) The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* 95(7)**:** 927-937.

Shoichet, BK (2004) Virtual Screening of chemical libraries. *Nature* 432**:** 862-865.

Smith, A (2002) Screening for drug discovery: the leading question. *Nature* 418(6896)**:** 453-459.

Sneader, W (2005) *Drug Discovery*. John Wiley & Sons: Chichester.

Sousa, SF, Fernandes, PA, Ramos, MJ (2006) Protein-ligand docking: current status and future challenges. *Proteins* 65(1)**:** 15-26.

Sperandio, O, Souaille, M, Delfaud, F, Miteva, MA, Villoutreix, BO (2009) MED-3DMC: a new tool to generate 3D conformation ensembles of small molecules with a Monte Carlo sampling of the conformational space. *Eur J Med Chem* 44(4)**:** 1405-1409.

Stahl, M, Rarey, M (2001) Detailed analysis of scoring functions for virtual screening. *J Med Chem* 44(7)**:** 1035-1042.

Stewart, JJ (1990) MOPAC: a semiempirical molecular orbital program. *J Comput Aided Mol Des* 4(1)**:** 1-105.

Still, W, Tempczyk, A, Hawley, R, Hendrickson, T (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. *J Am Chem Soc* 112**:** 6127-61129.

Swanson, JM, Mongan, J, McCammon, JA (2005) Limitations of atom-centered dielectric functions in implicit solvent models. *J Phys Chem B* 109(31)**:** 14769-14772.

Tagliabue, G, Citti, L, Massazza, G, Damia, G, Giavazzi, R, D'Incalci, M (1992) Tumour levels of O6-alkylguanine-DNA-alkyltransferase and sensitivity to BCNU of human xenografts. *Anticancer Res* 12(6B)**:** 2123-2125.

Takagi, T, Amano, M, Tomimoto, M (2009) Novel method for the evaluation of 3D conformation generators. *J Chem Inf Model* 49(6)**:** 1377-1388.

Teague, SJ (2003) Implications of protein flexibility for drug discovery. *Nat Rev Drug Discov* 2(7)**:** 527-541.

Thomas, MP, McInnes, C, Fischer, PM (2006) Protein structures in virtual screening: a case study with CDK2. *J Med Chem* 49(1)**:** 92-104.

Tomic, S, Bertosa, B, Wang, T, Wade, RC (2007) COMBINE analysis of the specificity of binding of Ras proteins to their effectors. *Proteins* 67(2)**:** 435-447.

Tomic, S, Nilsson, L, Wade, RC (2000) Nuclear receptor-DNA binding specificity: A COMBINE and Free-Wilson QSAR analysis. *J Med Chem* 43(9)**:** 1780-1792.

Triballeau, N, Acher, F, Brabet, I, Pin, JP, Bertrand, HO (2005) Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *J Med Chem* 48(7)**:** 2534-2547.

Tugarinov, V, Hwang, PM, Kay, LE (2004) Nuclear magnetic resonance spectroscopy of high-molecular-weight proteins. *Annu Rev Biochem* 73**:** 107-146.

Vainio, MJ, Johnson, MS (2007) Generating conformer ensembles using a multiobjective genetic algorithm. *J Chem Inf Model* 47(6)**:** 2462-2474.

Vaque, M, Arola, A, Aliagas, C, Pujadas, G (2006) BDT: an easy-to-use front-end application for automation of massive docking tasks and complex docking strategies with AutoDock. *Bioinformatics* 22(14)**:** 1803-1804.

Verdonk, ML, Cole, JC, Hartshorn, MJ, Murray, CW, Taylor, RD (2003) Improved protein-ligand docking using GOLD. *Proteins* 52(4)**:** 609-623.

Wade, RC, Henrich, S, Wang, T (2004) Using 3D protein structures to derive 3D-QSARs. *Drug Discovery Today: Technologies* 1**:** 241-246.

Wade, RC, Ortiz, AR, Gago, F (1998) Comparative Binding Energy Analysis. In: *3D-QSAR in Drug Design, Vol. 2: Ligand-Protein Interactions and Molecular Similarity*, Kubinyi, H, Folkers, G, Martin, Y (eds), pp 19-34. Kluwer-ESCOM, Dordrecht (Netherland).

Wang, K, Murcia, M, Constans, P, Perez, C, Ortiz, AR (2004) Gaussian mapping of chemical fragments in ligand binding sites. *J Comput Aided Mol Des* 18(2)**:** 101-118.

Wang, R, Lai, L, Wang, S (2002) Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J Comput Aided Mol Des* 16(1)**:** 11-26.

Wang, R, Lu, Y, Wang, S (2003) Comparative evaluation of 11 scoring functions for molecular docking. *J Med Chem* 46(12)**:** 2287-2303.

Wang, T, Wade, RC (2001) Comparative binding energy (COMBINE) analysis of influenza neuraminidase-inhibitor complexes. *J Med Chem* 44(6)**:** 961-971.

Wang, T, Wade, RC (2002) Comparative binding energy (COMBINE) analysis of OppA-peptide complexes to relate structure to binding thermodynamics. *J Med Chem* 45(22)**:** 4828-4837.

Wang, Z, Harkins, PC, Ulevitch, RJ, Han, J, Cobb, MH, Goldsmith, EJ (1997) The structure of mitogen-activated protein kinase p38 at 2.1-A resolution. *Proc Natl Acad Sci U S A* 94(6)**:** 2327-2332.

Warren, GL, Andrews, CW, Capelli, AM, Clarke, B, LaLonde, J, Lambert, MH, Lindvall, M, Nevins, N, Semus, SF, Senger, S, Tedesco, G, Wall, ID, Woolven, JM, Peishoff, CE, Head, MS (2006) A critical assessment of docking programs and scoring functions. *J Med Chem* 49(20)**:** 5912-5931.

Watson, P, Verdonk, M, Hartshorn, MJ (2003) A web-based platform for virtual screening. *J Mol Graph Model* 22(1)**:** 71-82.

Weiner, SJ, Kollman, PA, Case, DA, Singh, UC, Ghio, C, Alagona, G, Profeta, S, Weiner, P (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. *J Am Chem Soc* 106**:** 765-784.

Weiser, J, Shenkin, PS, Still, WC (1999) Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). *J Comput Chem* 20(2)**:** 217-230.

Wilson, SR, Cui, WL (1990) Applications of simulated annealing to peptides. *Biopolymers* 29(1)**:** 225-235.

Wu, F, Yu, J, Gehring, H (2008) Inhibitory and structural studies of novel coenzyme-substrate analogs of human histidine decarboxylase. *FASEB J* 22(3)**:** 890-897.

Xu, H, Izrailev, S, Agrafiotis, DK (2003) Conformational sampling by self-organization. *J Chem Inf Comput Sci* 43(4)**:** 1186-1191.

Yang, JM, Chen, YF, Shen, TW, Kristal, BS, Hsu, DF (2005) Consensus scoring criteria for improving enrichment in virtual screening. *J Chem Inf Model* 45(4)**:** 1134-1146.

Zhang, S, Kumar, K, Jiang, X, Wallqvist, A, Reifman, J (2008) DOVIS: an implementation for high-throughput virtual screening using AutoDock. *BMC Bioinformatics* 9**:** 126.

Zhou, HX, Gilson, MK (2009) Theory of free energy and entropy in noncovalent binding. *Chem Rev* 109(9)**:** 4092-4107.

Zhou, T, Caflisch, A (2009) Data management system for distributed virtual screening. *J Chem Inf Model* 49(1)**:** 145-152.

Zou, X, Sun, Y, Kuntz, ID (1999) Inclusion of Solvation in Ligand Binding Free Energy Calculations Using the Generalized-Born Model. *J. Am. Chem. Soc.* 121(35)**:** 8033-8043.

ARTÍCULOS Y PATENTE

8. Artículos y Patente

El trabajo realizado en esta Tesis dio lugar a las siguientes publicaciones:

1. Morreale A, Gil-Redondo R, Ortiz AR. *A new implicit solvent model for protein-ligand docking*. Proteins (2007), 67(3):606-16
2. Ruiz FM, Gil-Redondo R, Morreale A, Ortiz AR, Fábrega C, Bravo J. *Structure-based discovery of novel non-nucleosidic DNA alkyltransferase inhibitors: virtual screening and in vitro and in vivo activities*. J Chem Inf Model (2008), 48(4):844-54
3. Gil-Redondo R, Estrada J, Morreale A, Herranz F, Sancho J, Ortiz AR. *VSDMIP: virtual screening data management on an integrated platform*. J Comput Aided Mol Des (2009), 23(3):171-84
4. Moya-García AA, Pino-Angeles A, Gil-Redondo R, Morreale A, Sánchez-Jiménez F. *Structural features of mammalian histidine decarboxylase reveal the basis for specific inhibition*. Br J Pharmacol (2009), 157(1):4-13
5. Gil-Redondo R, Klett J, Gago F, Morreale A. *gCOMBINE: a graphical user interface to perform structure-based Comparative Binding Energy (COMBINE) analysis on a set of ligand-receptor complexes*. Proteins (2010), 78(1):162-172

Y también se depositó la patente:

WO2009060114-A1. *Use of a compound derived from piperadinyl-methyl-tetrazol-quinolinone and diphenyl-triazolo-pyrimidine for the development of a medicine or composition for the treatment of cancer*. Asignees: Consejo Superior de Investigaciones Científicas (CSIC) y Centro Nacional de Investigaciones Oncológicas (CNIO). Inventors: Bravo Sicilia J, Fabregas Claveria M C, Gil Redondo R, Morreale de León A J, Ortiz M L, Ruiz F.

Una copia de cada una de las publicaciones y de la primera página de la patente se adjunta en las siguientes páginas.

PROTEINS: Structure, Function, and Bioinformatics 67:606–616 (2007)

A New Implicit Solvent Model for Protein–Ligand Docking

Antonio Morreale, Rubén Gil-Redondo, and Ángel R. Ortiz*
Bioinformatics Unit, Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Universidad Autónoma de Madrid, Cantoblanco, Madrid 28049, Spain

ABSTRACT **A new implicit solvent model for computing the electrostatics binding free energy in protein–ligand docking is proposed. The new method is based on an adaptation of the screening coulombic potentials proposed originally by Hassan et al. (J Phys Chem B 2000;104:6490–6498). In essence, it relies on two basic assumptions; (i) solvent screening can be accounted for by means of radially dependent sigmoidal dielectric functions and; (ii) the effective atom Born radii can be expressed only as a function of the exposed atom surface. Parameters of the model other than radii and charges are generic. These were optimized for a dataset of 826 protein–ligand complexes, comprising both X-ray complexes for 23 receptors as well as decoys generated by docking computations. We show that the new model provides satisfactory results when benchmarked against reference values based on the numerical solution of the Poisson equation, with a root mean square error of 4.2 kcal/mol over a range of ~40 kcal/mol in electrostatics binding free energies, a cross-validated r^2 of 0.81, a slope of 0.97, and an intercept of 1.06 kcal/mol. We show that the model is appropriate for ligands of different sizes, polarities, overall charge, and chemical composition. Furthermore, not only the total value of the electrostatic contribution to the binding free energy, but also its components (coulombic term, receptor desolvation, and ligand desolvation) are reasonably well reproduced. Computation times of ~0.030 s per pose are obtained on a single processor desktop workstation. Proteins 2007;67:606–616.**

Key words: electrostatics; force fields; solvation; binding free energies; virtual screening; docking

INTRODUCTION

An adequate treatment of solvation is yet an unsolved problem in protein–ligand docking.[1] Solvation (or hydration, for simplicity throughout this manuscript we will interchange both terms) plays an important role in the energetics of ligand–protein association,[2–5] and when using molecular mechanics energy functions, its physical model influences ranking in virtual screening,[6–10] and to a more limited extent, docking geometry.[11,12] Introduction of explicit solvent would possibly be the most rigorous means of incorporating the solvent effect, but this is impractical in docking computations. On the other hand, there is ample consensus nowadays that implicit solvation methods, while introducing various approximations to hydration effects,[13] provide an adequate balance between computational efficiency and physical soundness.

The implicit hydration free energy is usually divided into several components: cavity formation, short-range solvent–solute interactions, and electrostatic solvation. In this paper we will only consider the later. Implicit electrostatic solvation is achieved by presuming that the solvent is a continuum high-dielectric-constant medium that responds to the partial charges of a low-dielectric-constant solute. Two major continuum models have been applied in protein–ligand docking, one that uses the numerical solutions of the Poisson equation (PE),[14] and another that applies the generalized-Born (GB) approximation.[15,16] While the use of the PE method for docking has been explored, it is the GB formulation, due to its improved computational efficiency, the method most widely investigated to account for electrostatic hydration effects in protein–ligand docking.[17–20] Nevertheless, the computational cost of an straightforward implementation of these models is still significant, and in practice most docking protocols employ PE or GB models only as a second rescoring step, while the docking scoring functions employ a crude treatment of electrostatics and solvation.[6,10,19,21]

It would be beneficial to develop new or improved approaches to the calculation of the electrostatics binding free energy with a more appropriate balance of accuracy and speed for protein–ligand docking. Adaptations of

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: MEC; Grant numbers: BIO2001-3745, BIO2005-0576 and GEN2003-206420-C09-08; Grant sponsor: Comunidad de Madrid; Grant numbers: GR/SAL/0306/2004 and 200520M157; Grant sponsor: CSIC, intramural program (PIF 2005, project CAR); Grant number: PIF2005; Grant sponsor: Fundación Ramón Areces.

*Correspondence to: Ángel R. Ortiz Bioinformatics Unit, Centro de Biología Molecular "Severo Ochoa" (CSIC-UAM), Universidad Autónoma de Madrid, Cantoblanco, 28049 Madrid, Spain.
E-mail: aro@cbm.uam.es

Received 7 March 2006; Revised 10 August 2006; Accepted 15 September 2006

Published online 28 February 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21269

PE and GB methods have been proposed earlier to make them more efficient in docking. These are based on the realization that when two molecules are brought together, the main contribution to electrostatic desolvation originates from the displacement of the first shell of water molecules occupying the interacting surface, the so-called first shell approximation. Based on this idea, Arora and Bashford[22] presented the solvation energy density occlusion (SEDO) method as an approximation PE electrostatic desolvation. With SEDO, a solvation energy density is stored on a grid for each molecule in isolation. Desolvation is then calculated by integrating the solvation energy density of one molecule that is occluded by the other when the pair associates. Similarly, Caflisch and coworkers have used the first shell approximation to modify the GB model by using the Coulomb approximation of the electric displacement via volume-integration.[23] The modification allows estimating the energy in solution of ~300 protein-fragment binding modes per second on a 550 MHz Pentium III. For two different proteins binding to a set of small molecule fragments a r^2 of ~0.72 with respect to PE energies and slopes close to the unit were obtained in each case. However, no data are available on the overall performance with a large dataset of structurally different proteins. A limitation in both approaches is the requirement of a rigid protein structure, in order to efficiently use the grid technology.

Herein, we propose a new implicit solvent model (ISM) for computing the electrostatics binding free energy in protein–ligand docking, based on an adaptation to protein–ligand binding of the screening coulombic potentials (SCP) proposed originally by Hassan et al. to treat electrostatics interactions in proteins.[24–26] The model (SCP-ISM) shares similarities with GB approaches, but also differs from them in a number of important aspects. With SCP-ISM the system is described as immersed in a continuum that permeates all space and is completely characterized by the screening function. In this way the model departs from GB approaches, eliminating the need to define an internal dielectric constant and a (discontinuous) boundary between protein and solvent. This is advantageous in the computation of protein–ligand docking interactions, which mostly take place in the protein–solvent interface, where the precise location of the boundary is not well defined. A second feature of the model is the use of the first shell approximation, that is, the effective Born radii for each atom is expressed only as a function of the atom exposed surface accessible area. This allows an improved trade off between speed and accuracy. A third advantage is that the effective Born radii only appear in the self-energy terms, playing no role in the calculation of the interaction contributions, allowing storing the protein electrostatic potential in a grid for a fast computation of the pairwise electrostatic contribution to the electrostatic binding free energy, the so-called test charge approximation. Finally, the model parameters, other than radii and charges, are generic, allowing the model to be easily parametrized and therefore it is adaptable to large scale docking computations.

METHODS

Implicit Solvent Model Based on Screened Coulomb Potential

We review here for completeness of the most salient features of the SCP-ISM theory developed by Hassan et al. More details can be obtained from the original publications.[25,26] The model starts from the Lorentz–Debye–Sack theory of polar liquids, which establishes that the screening effect due to the solvent shows a sigmoidal-distance dependent dielectric function of the form:

$$D(r) = \frac{\varepsilon + 1}{1 + k\ \exp[-\lambda(\varepsilon + 1)r]} - 1 \tag{1}$$

where ε is the solvent dielectric constant, $k = (\varepsilon - 1)/2$ and λ is a parameter controlling the rate of change of $D(r)$. A similar screening function has also been previously introduced in the docking program Autodock.[27] A second key aspect of the model is the assumption that the main contribution to electrostatic desolvation of an atom originates from the displacement of the first shell of water molecules surrounding the atom and occupying the atomic surface. The model defines parameters $R_{i,\mathrm{B_S}}$ and $R_{i,\mathrm{B_V}}$ as the effective Born radii for the processes of transferring an atom from the vacuum into a protein interior, surrounded by solvent or vacuum, respectively. According to the first shell approximation, these radii are calculated using linear relationships of the form:

$$R_{i,\mathrm{B_S}} = R_{i,w}\xi_i + R_{i,p}(1 - \xi_i) \tag{2}$$

and,

$$R_{i,\mathrm{B_V}} = R_{i,\mathrm{V}}\xi_i + R_{i,p}(1 - \xi_i) \tag{3}$$

where ξ_i is the fraction of SASA (A_i) of the i atom: $\xi_i = A_i/4\pi(r_{vdw,i} + r_{\mathrm{probe}})^2$. With $R_{i,W} = R_{i,\mathrm{COV}} + h_{(+,-)}$, $R_{i,P} = R_{i,\mathrm{COV}} + g$, and $R_{i,\mathrm{V}} = R_{\mathrm{COV}}$. R_{COV} is the covalent radius, and $h_{(+,-)}$ and g are positive quantities that account for the enlargement of the cavity due to charge effects. In particular, $h_{(+,-)}$ depends on the atomic charge (see below). Applying this function to the solvation process (a detailed derivation of the model can be found in the original papers by Hassan et al.[25,26]) the following equation is obtained:

$$\Delta G_{\mathrm{elec}} = \sum_{i<j} \frac{q_i q_j}{r_{ij}} \left[\frac{1}{D_{\mathrm{S}}(r_{ij})} - \frac{1}{D_{\mathrm{V}}(r_{ij})}\right] + \frac{1}{2}\sum_{i=1}^{N} q_i^2 \left\{\frac{1}{R_{i,\mathrm{B_S}}}\left[\frac{1}{D_{\mathrm{S}}(R_{i,\mathrm{B_S}})} - 1\right] - \frac{1}{R_{i,\mathrm{B_V}}}\left[\frac{1}{D_{\mathrm{V}}(R_{i,\mathrm{B_V}})} - 1\right]\right\} \tag{4}$$

where the first term represents Coulombic interactions between charged particles screened by dielectric sigmoidal function depicted in Eq. (1), s stands for the solvent and v for vacuum.

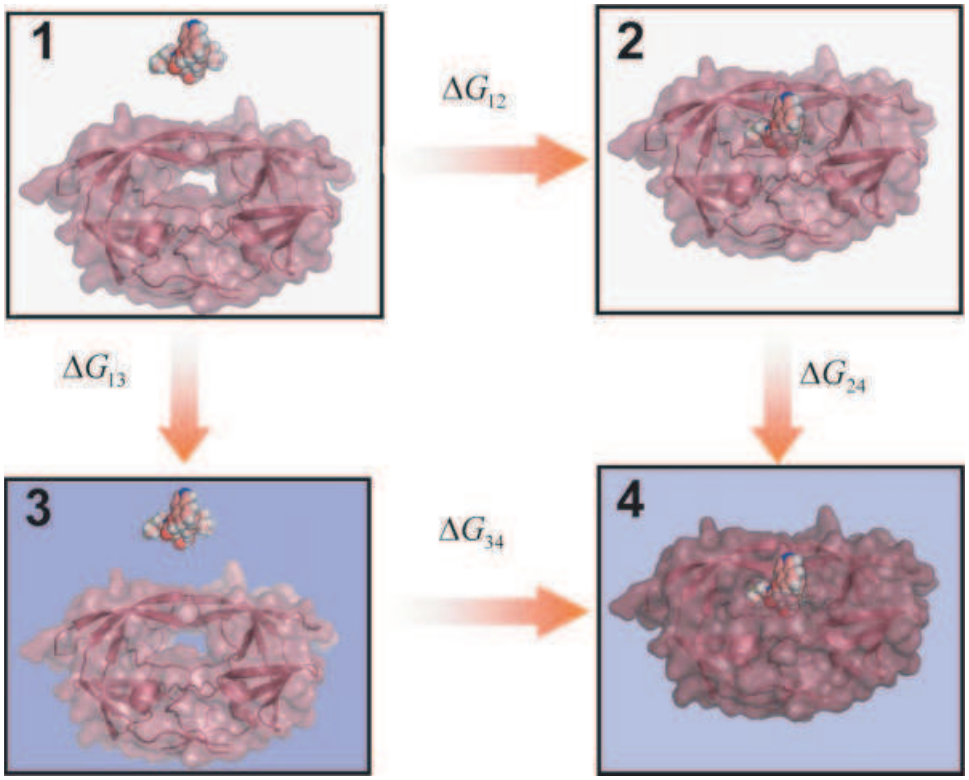


Fig. 1. Thermodynamic cycle employed for the calculation of the electrostatics binding free energy. The process of binding in solution (ΔG_{34}), can be alternatively computed as the process of desolvating the unbound species ($-\Delta G_{13}$), letting them interact in vacuum (ΔG_{12}), and solvating the resulting complex (ΔG_{24}). See text for additional details.

Extending SCP–ISM to the Ligand-Receptor Docking Problem

To extend Eq. (4) to the problem of ligand–receptor interactions, the thermodynamic cycle depicted in Figure 1 was built. Boxes represent vacuum (1, 2) and solvent (3, 4) states, in both unbound (1, 3) and bound (2, 4) forms. ΔG_{elec} is then calculated from Eq. (5):

$$\Delta G_{\text{elec}} = \Delta G_{34} = \Delta G_{12} + (\Delta G_{24} - \Delta G_{13}) \tag{5}$$

ΔG_{13}, ΔG_{24}, and ΔG_{12} that can be obtained directly from Eq. (4), and assuming a rigid ligand–protein binding, after reorganizing terms, Eq. (6) is obtained:

$$\Delta G_{\text{elec}} = \sum_{i=1}^{N_L}\sum_{j=1}^{N_R}\frac{q_i q_j}{D_s(r_{ij})r_{ij}} + \frac{1}{2}\sum_{i=1}^{N_L+N_R} q_i^2 \times \left[\left(\frac{1}{D_S(R_S^C)R_S^C} - \frac{1}{D_S(R_S^U)R_S^U}\right) + \left(\frac{1}{R_S^U} - \frac{1}{R_S^C}\right)\right] \tag{6}$$

N_L and N_R are the number of atoms in ligand and receptor, respectively, R_S^C and R_S^U are the effective Born radii for complexed and uncomplexed forms of both, ligand and receptor. In analogy with Eq. (4), the first term represents charge–charge interactions between the ligand and the receptor, and the second contains desolvation penalties to be paid for removing atoms from the solvent to form the complex. Eq. (6) is the main result of applying SCP–ISM theory to the protein–ligand binding problem.

Surface Model

SASA values required in Eqs. (2) and (3) were obtained with the LCPO approximation[28] with a solvent-probe radius (r_{probe}) of 1.4 Å. SASAs were computed using Eq. (7):

$$A_i = P_1 S_1 + P_2 \sum_{j\in N(i)} A_{ij} + P_3 \sum_{\substack{j,k\in N(i)\\ k\in N(j)\\ k\neq j}} A_{jk} + P_4 \sum_{j\in N(i)} A_{ij} \left(\sum_{\substack{j,k\in N(i)\\ k\in N(j)\\ k\neq j}} A_{jk} \right) \tag{7}$$

In this equation, S_1 in the first term is the surface area of the isolated sphere corresponding to atom i. A_{ij} is the area of sphere i buried inside sphere j, while $N(i)$

stands for the neighbor list of i, that is, the list of spheres that overlap with sphere i. Thus, the second term involves the sum of pairwise overlaps of sphere i with its neighbors. The third term is the sum of overlaps of neighbors of i with each other. The fourth and last term is a further correction for multiple overlaps. See Weiser et al.[28] for additional details. New parameters P_1 to P_4 were derived using our own training set of proteins. See the supporting information.

Hydrogen Bond Correction

A correction term was introduced into Eq. (6) to account for the effect of hydrogen bond interactions on the ligand desolvation energies. The origin for this correction term is analyzed in the discussion section. A regression analysis was performed between the difference in desolvation energies calculated with Poisson and SCP–ISM methods and the number and type of hydrogen bonds. The number of hydrogen bonds for each decoy was deduced using a donor–acceptor cut off distance of 3.4 Å and an angle of 120° between donor–acceptor–acceptor antecedents. Definitions for donor and acceptor atoms types were taken from HBPLUS[29] in the case of protein atoms. For the ligands, all the oxygen and nitrogen atoms were visually inspected and assigned. Hydrogen bonds were then classified by the type of interactions in charged–charged (cc), neutral–charged and charged–neutral (nc), and neutral–neutral (nn). We observed that for those cases involving protonated amino groups, an additional binary variable describing the presence or absence of this group was required (npn). This is easily understood bearing in mind the way in which LCPO surface parameters were obtained, by linear fitting over the surface accessible area of the residues in the native structures (see supplementary material). Therefore, parameters for protonated nitrogen atoms where derived from lysine, whose distribution of exposed surface areas is skewed towards high values, and consequently resulting in underrepresented counts for buried protonated amino groups. All in all, the correction term is then of the form:

$$\Delta G_{\mathrm{corr}} = a + b \cdot hb_{\cdot\,\mathrm{cc}} + chb_{\cdot\,\mathrm{nc}} + dhb_{\cdot\,\mathrm{nn}} + enpn \quad (8)$$

where hb_{xx} is the number of hydrogen bonds of type xx. The final electrostatics binding free energy computed with the SCP–ISM approximation is therefore:

$$\Delta G_{\mathrm{ISM}} = \Delta G_{\mathrm{elec}} + \Delta G_{\mathrm{corr}} \quad (9)$$

Comparison With Electrostatics Binding Free Energies Obtained With Numerical Solutions of the Poisson Equation

Because the usefulness of the PE for calculating electrostatics effects in intermolecular interactions is well established,[14] the tests of the SCP–ISM approximation presented in this article are mainly intended to see how well it approximates the electrostatics binding free energy calculated using the full PE approach. The total electrostatics binding free energy (ΔG_{ele}) was calculated from the total electrostatic energy of the system obtained when solving the PE by running three consecutive calculations on the same grid: one for all the atoms in the complex ($G^{\mathrm{LR}}_{\mathrm{ele}}$), one for the ligand atoms alone ($G^{\mathrm{L}}_{\mathrm{ele}}$), and a third one for the receptor atoms alone ($G^{\mathrm{R}}_{\mathrm{ele}}$). Since the grid definition is the same in the three calculations, the artifactual grid energy cancels out when the electrostatic contribution to the binding free energy is expressed as the difference in energy between the product and the reactants:

$$\Delta G_{\mathrm{ele}} = G^{\mathrm{LR}}_{\mathrm{ele}} - (G^{\mathrm{L}}_{\mathrm{ele}} + G^{\mathrm{R}}_{\mathrm{ele}}) \quad (10)$$

Alternatively, we also considered a different description of the binding process, consisting of first desolvating the opposing surfaces of both ligand and receptor and then letting the charges of the two molecules interact, in order to isolate the three different contributions to the electrostatics binding free energy: the ligand–receptor interaction energy in the presence of the surrounding solvent ($E^{\mathrm{LR}}_{\mathrm{ele}}$), the change in solvation energy of the ligand upon binding ($\Delta G^{\mathrm{L}}_{\mathrm{desolv}}$), and the change in solvation energy of the receptor upon binding ($\Delta G^{\mathrm{R}}_{\mathrm{desolv}}$):

$$\Delta G_{\mathrm{ele}} = E^{\mathrm{LR}}_{\mathrm{ele}} + (\Delta G^{\mathrm{L}}_{\mathrm{desolv}}) + \Delta G^{\mathrm{R}}_{\mathrm{desolv}}) \quad (11)$$

The first term, the Coulombic contribution to ΔG_{elec}, was obtained by computing the product of ligand charges and the electrostatic potential generated by the protein on the ligand charge centres. On the other hand, receptor and ligand electrostatic desolvation energies were calculated in two successive steps: a first one, where a calculation is performed for receptor and ligand alone; and the second one, for the ligand, with uncharged receptor, and the receptor, with uncharged ligand.

All these calculations were performed by numerically solving the linear PE using the finite difference method as implemented in DelPhi.[30] For all calculations with DelPhi PARSE atomic radii[31] were used while charges were assigned either based on the AMBER force field[32] (protein case), or computed with MOPAC[33] (ligand case). Each complex was immersed in a cubic box occupying 65% of the total volume, with a grid spacing of 0.5 Å. Solute dielectric constant was set to 4, while the solvent and the dielectric medium was set to 80. We note that although a bulk dielectric constant of 1 or 2 is commonly employed to model the solute interior in continuum calculations, the microscopic value is environment dependent.[34] Higher values can be employed when energetics of processes involving polar sites are of interest, such as in pKa calculations or when modelling ligand–DNA interactions.[35,36] Although a matter of debate, and since ligand binding sites are moderately polar, we[5] and

TABLE I. Summary of the Decoy Dataset Used in this Paper

PDB ID	Description	No.of atoms (heavy)	No. of decoys	RMSD min–max (Å)	Total charge	No. of HBD	No. of HBA
1HVI	HIV-1 protease	116 (58)	43	0.3–15.1	0	6	8
1HVJ	HIV-1 protease	115 (57)	65	0.2–14.1	0	5	7
1HVK	HIV-1 protease	116 (58)	50	0.3–15.8	0	6	8
1HIH	HIV-1 protease	91 (41)	57	0.4–15.8	0	5	5
1HPX	HIV-1 protease	87 (46)	85	0.1–9.6	0	4	6
1MCJ	Immunoglobulin	60 (32)	44	0.3–11.3	0	4	6
1RBP	Retinol binding protein	51 (21)	40	0.2–3.4	0	1	1
2UPJ	HIV-1 protease	81 (41)	48	0.2–10.1	0	3	4
1ABE	L-arabinose binding protein	20 (10)	60	0.1–3.8	0	4	4
1AJX	HIV-1 protease	74 (40)	64	0.1–3.7	0	3	3
5ABP	L-arabinose binding protein	22 (12)	34	0.1–2.5	0	5	5
1DBB	Immunoglobulin	53 (23)	3	1.4–6.9	0	0	2
1FKG	FK506 binding protein	68 (33)	20	0.4–8.5	0	0	3
1FKH	FK506 binding protein	74 (33)	20	0.4–8.0	0	0	3
1MRK	α-trichosanthin	32 (19)	9	0.5–4.5	0	6	8
1STP	Biotin binding protein	31 (16)	78	0.1–8.5	−1	3	7
1B9V	Influenza virus neuraminidase	50 (25)	20	0.4–5.0	−1	0	6
1DBM	Immunoglobulin	64 (31)	20	0.5–3.3	−1	0	5
1TNG	Trypsin	24 (8)	3	0.7–1.8	+1	3	0
1TNI	Trypsin	27 (11)	20	0.6–3.2	+1	3	0
1TNK	Trypsin	24 (10)	20	0.6–2.9	+1	3	0
1TNL	Trypsin	22 (10)	3	1.6–2.1	+1	3	0
1BMA	Trypsin	73 (37)	20	0.6–4.2	+1	0	3

See Methods for additional details.

others[2,37,38] consider more appropriate a value of four. The dielectric boundary was calculated using a solvent probe radius of 1.4 Å. A minimum separation of 11 Å was allowed between any solute atom and the box walls. The potentials at the grid points delimiting the box were calculated analytically by treating each charge atom as a Debye–Hückel sphere.

Training Set of Complexes Used in the Development of the SCP–ISM Model

A training set was formed with 23 different proteins (see Table I) summing up a total of 826 decoys. Some of the decoys (those for 1HVI, 1HVJ, 1HVK, 1HIH, 1HPX, 1MCJ, 1RBP, 1UPJ, 1ABE, 1AJX, 5ABP, and 1STP) were taken directly from LPDB database.[39] Atom types and hydrogen atoms definitions where translated from the CHARMm force field into their equivalents in AMBER. Hydrogen atoms were removed, and added back with protonate program from the AMBER 8.0 package, were also charge and radii for all the atoms in the proteins assigned. For the ligands, AMBER radii and semiempirical charges fitted to electrostatic potentials with MOPAC[33] were used (keywords 1SCF, MNDO, ESP, and DIPOLE). The rest of the decoys (those for 1DBB, 1FKG, 1FKH, 1B9V, 1DBM, 1TNG, 1TNI, 1TNK, 1TNL, and 1BMA) were generated by us using our in-house docking program.[11,40] They were prepared in exactly the same way as stated before for LPDB data set.

Parametrization and Validation of the SCP–ISM Model

Description of the SCP–ISM parameters and their optimized values are listed in Table II, and in Table I of the supplementary material. We arrived at these parameters by performing exhaustive searches in a subset of the parameter space, using the quadratic error between the SCP–ISM and PE values as fitness function. The following parameters were systematically modified: scale factor for atomic radii (from 0.3 to 1.3 Å in 0.1 Å intervals; enlargement factor $h_{(+,-)}$, from 0.35 to 0.85 Å in 0.1 Å intervals; enlargement factor g, from 0.0 to 1.0 Å in 0.1 Å intervals; λ, from 0.001 to 0.020 in 0.001 intervals; fixed values were used for ε (78.39) and the solvent-probe radius (1.4 Å). To test for the robustness of the fitted parameters, a Leave One Out (LOO) procedure was applied. One by one, all decoys from a single target were removed from the complete set, the model was rebuilt (internal test) and the excluded decoys blindly predicted (external validation). At each step the RMSD and regression coefficients between the values for the external set and their Poisson counterparts were compared. Calculations were carried out with the R package (http://www.r-project.org/).

RESULTS

A key aspect of any method development involving parameter fitting is the training set of examples employed in the fitting process. Here, 826 different decoys have

TABLE II. Relevant Parameters Used in our SCP-ISM Model

Parameter	Value
Atomic Radii[a]	
Scale	0.6
$h_{(+,-)}$	0.85 (0.35)
g	0.5
Solvent[b]	
$\lambda_{(+,-)}$	0.013 (0.007)
ε	78.39
r_{probe}	1.4
Hydrogen bond[c]	
r	3.4
α	120
a	−0.29
b	1.02
c	−0.25
d	0.02
e	10.52
PE/SCP-ISM[d]	
A	5.3
B	0.09
C	1.06
D	0.97

[a]Initial AMBER radii for all the atoms are scaled down by the scale parameter. $h_{(+,-)}$ and g account for the enlargement of the radii when immersed in solvent. $h_{(+,-)}$ depends on the type of charge (0.85 for positive and 0.35 for negative), g is independent of the charge and it is always equals to 0.5, both in Å.

[b]Solvent related parameters are the slope of sigmoidal dielectric function ($\lambda_{(+,-)}$) with two values: 0.013 for all of the atoms except for those with a formal positive charge, and 0.007 for these last ones. ε is the dielectric constant of the bulk solvent and r_{probe} is the radius, in Å, of the water probe molecule employed to calculate the solvent accessible surface.

[c]Hydrogen bond parameters: r and α are the minimum radii (in Å) and angle (in degrees) between donor and acceptor, and donor–acceptor–acceptor antecedents, respectively. a, b, c, and d corresponds to the fitted parameters to account for the hydrogen bond correction (see Methods) according to the equation: $\Delta G_{\text{corr}} = a + b \cdot hb_{\text{cc}} + c \cdot hb_{\text{nc}} + d \cdot hb_{\text{nn}} + e \cdot npn$.

[d]Final parameters obtained from the comparison between PE and SCP-ISM according to the equation: $PB = C + D(A \exp(B(ISM))$, see Table III and main text.

been employed to test the new solvation model described in this paper (Table I). Decoys cover a wide range (around 40 kcal/mol) of electrostatics binding free energies. There is also ample structural variety in the complexes employed, both in protein architectures as well as in the ligand functional groups. The set includes neutral, zwitterionic, as well as formally charged (both positively and negatively) ligands. Finally, there is also a considerable number of representative orientations of each complex within each binding site, about 20 on average, covering a large spectrum of RMSD values, ranging from close natives to more than 10 Å RMSD. Thus, we feel confident that our dataset, while not perfect, has enough variety to warrant the generality of our results. The LOO tests (see below) seem to confirm this.

A second important aspect is the number of parameters to fit. Our SCP–ISM model has a relatively small number of generic parameters (Table II). Leaving aside the parameters related to charges, radii, and those involved in the surface calculation, the model has a total of 17 or 19 parameters (Table II), depending on the exact choice of the model (see below). The basic model, corresponding to Eq. (6), contains eight parameters: h_+, h_-, g, λ_+, λ_-, ε, r_{probe}, and the scale factor of the atomic radii (see Methods and Table III for definitions and values, respectively). Six out of these eight were considered for optimization, while ε and r_{probe} were kept fixed. In order to properly reproduce PE results with the version of the SCP–ISM model developed here, a hydrogen bond correction was deemed necessary. The definition of the hydrogen bond itself added two parameters to the model. In order to obtain a reasonable fit, an additional set of five parameters, accounting for the nature of the hydrogen bond, were required (Table II). A comparison of the SCP–ISM and PE electrostatics binding free energies with the optimized set of parameters is shown in Figure 2(a). The direct comparison suggests that an exponential-type relationship exists between them. The reason for this dependence is unclear to us, and its investigation will be left for future work. Noting this dependence, two different fittings between the two sets of data were attempted: an exponential one (model1, Poisson $= A \exp(B \times$ SCP–ISM)); and a second linear fitting of the exponential model (i.e., model1) was also investigated (model2, Poisson $= C + D(A \exp(B \times$ SCP–ISM)), to account for systematic deviations from the exponential behavior. Adding the fitting parameters (two or four, depending on which model is used) yields the final set of 17 or 19 parameters comprising our complete SCP–ISM model.

Results for the two fittings (model1 and model2), including LOO tests, can be found in Table III. The results for the ALL row correspond to the standard crossvalidation case, where each set of decoys for a given protein were removed, a model derived, and based on the model the removed complexes were predicted. In this case, as expected, the crossvalidated RMSD is slightly larger than fitted one. The rest of the rows in Table III correspond to the partial results of the ALL case. For example, the first row shows the in-fitting columns, the model obtained after removing decoys for 1HVI, with all other decoys as training set. The LOO crossvalidation columns show the result of this model as applied to the 1HVI decoys. A slightly better RMSD value is obtained with model2 as compared to model1 with 4.16 versus 4.20 kcal/mol, respectively. A crossvalidated r^2, or q^2 of 0.81, a slope of 0.97, and an intercept of 1.06 kcal/mol was found for the best model. A comparison with PE data is shown in Figure 2(b). The similarity for the RMSD values between the fitted and crossvalidated electrostatics binding free energies shown in the LOO tests (4.20 versus 4.33, see Table III) indicates that there is no evidence of overfitting. Thus, our results are likely to hold using different sets of complexes. Satisfactory results are also found for the different components of the electrostatics binding free energy. The squared corre-

TABLE III. Evaluation of the SCP-ISM Model

	Fitting								LOO crossvalidation		
	PE = A exp(B × ISM)				PE = C + D(A exp(B × ISM))						
Complex	A	B	r^2	RMSD[a]	C	D	r^2	RMSD[a]	Model1 RMSD[a]	Model2 RMSD[a]	q^2
1HVI	5.33	0.09	0.80	4.11	0.79	0.99	0.86	4.07	5.88	5.70	0.70
1HVJ	5.34	0.09	0.80	4.04	0.85	0.98	0.87	4.00	6.14	5.96	0.40
1HVK	5.31	0.09	0.80	4.17	0.85	0.98	0.86	4.14	4.76	4.61	0.81
1HIH	5.32	0.09	0.80	4.15	0.82	0.98	0.87	4.12	4.96	4.69	0.76
1HPX	5.30	0.09	0.80	4.14	0.84	0.98	0.87	4.10	4.69	4.53	0.50
1MCJ	5.47	0.09	0.83	4.13	1.29	0.96	0.87	4.08	5.21	5.76	0.71
1RBP	5.52	0.09	0.81	4.22	1.10	0.97	0.86	4.18	3.37	4.12	0.44
2UPJ	5.25	0.09	0.81	4.16	0.92	0.98	0.87	4.12	5.04	5.00	0.36
1ABE	4.69	0.09	0.85	4.29	1.54	0.94	0.87	4.22	5.19	4.19	0.93
1AJX	5.46	0.09	0.81	4.28	1.31	0.96	0.86	4.23	2.46	3.18	0.73
5ABP	4.86	0.09	0.84	4.23	1.46	0.95	0.87	4.16	5.96	4.97	0.86
1DBB	5.30	0.09	0.81	4.21	1.06	0.97	0.86	4.17	1.55	2.28	0.98
1FKG	5.32	0.09	0.81	4.25	1.12	0.97	0.86	4.20	1.48	2.16	0.66
1FKH	5.32	0.09	0.81	4.25	1.11	0.97	0.86	4.20	1.26	1.98	0.95
1MRK	5.28	0.09	0.81	4.20	1.06	0.97	0.87	4.16	2.44	2.17	0.62
1STP	5.57	0.09	0.82	4.25	1.06	0.97	0.86	4.21	2.59	3.22	0.80
1B9V	5.29	0.09	0.81	4.24	1.00	0.98	0.87	4.19	4.38	3.92	0.92
1DBM	5.24	0.09	0.82	4.23	1.05	0.97	0.87	4.18	3.67	3.08	0.68
1TNG	5.32	0.09	0.81	4.20	1.03	0.97	0.86	4.16	2.59	3.41	0.87
1TNI	5.41	0.09	0.80	4.23	0.91	0.98	0.86	4.19	1.17	1.77	0.68
1TNK	5.37	0.09	0.81	4.22	1.07	0.97	0.86	4.18	2.49	3.08	0.69
1TNL	5.30	0.09	0.81	4.21	1.07	0.97	0.86	4.17	1.36	1.80	0.73
1BMA	5.27	0.09	0.81	4.24	1.12	0.97	0.86	4.19	1.85	1.53	0.19
ALL	5.30	0.09	0.81	***4.20***	1.06	0.97	0.87	***4.16***	***4.40***	***4.33***	***0.81***

Fitted (for model1, PE = A exp(B × ISM); and model2, PE = C + D(A exp(B × ISM)), see main text for more details) as well as crossvalidated results are shown. The results for the ALL row correspond to the standard crossvalidation case, where each set of decoys for a given protein were removed, a model derived, and based on the model the removed complexes were predicted. Partial results obtained excluding decoys of specific system shown in the corresponding row during the fitting phase are also presented. In these cases the fitting values correspond to those obtained with the model generated using the rest of the decoys. Reported q^2 values correspond to model2.
[a]Root mean square deviation (in kcal/mol) between the PE and SCP-ISM results.

lation coefficients oscillate between 0.79 and 0.88 (see Fig. 3). Slopes are also close to unity (1.13 for the coulombic term and 1.01 for the ligand desolvation term), except for receptor desolvation term (1.72) (see Fig. 3). Intercepts are close to zero in all cases (see Fig. 3). Thus, not only the total energy, but also its contributions are well reproduced by the SCP–ISM model.

As to the computation times, Figure 4 shows a histogram of the computing times for the 826 decoys. The average computing time is 40 ms, with a mode at 30 ms. These times are well below most GB approaches. The dependency of the computing times with the number of heavy atoms in the ligand can be found in Figure 5, which shows a "box and whisker plot". For each bin, the data is divided into four intervals: a quarter of the data (25% percentile) is between the lower-lying whisker and the baseline of the box, another quarter is between this line and the median line, other quarter is between the median line and the top line of the box, and finally, the last quarter is between this last line and the end of the higher-lying whisker. An approximately linear dependency between number of heavy atoms and computing time is observed.

DISCUSSION

Herein we present a new model for the fast calculation of electrostatics binding free energies in protein–ligand binding problems. The formulation is a modification of the original model proposed by Hassan et al. to treat electrostatics effects in proteins.[25,26] As in their case, no boundary surface between the high (solvent) and low (receptor and/or ligand) dielectric media is required. This is achieved by defining the dielectric function in a sigmoidal distance-dependent manner. Similarly, the effective Born radii are readily computed only from the exposed surface accessible area of the atom of interest. In order to properly account for the PE results, a hydrogen bond correction term in the SCP–ISM model was necessary. At face value, this requirement may seem odd, since both models (PE and SCP–ISM) attempt describe the same process, the electrostatics binding free energy, and hydrogen bonding has a strong electrostatics component which should be captured by the model. However, recent studies by McCammon and coworkers[41] have clearly established that the use of atom-centered surfaces, such as the ones employed here the LCPO method, or the presence of smooth transitions between

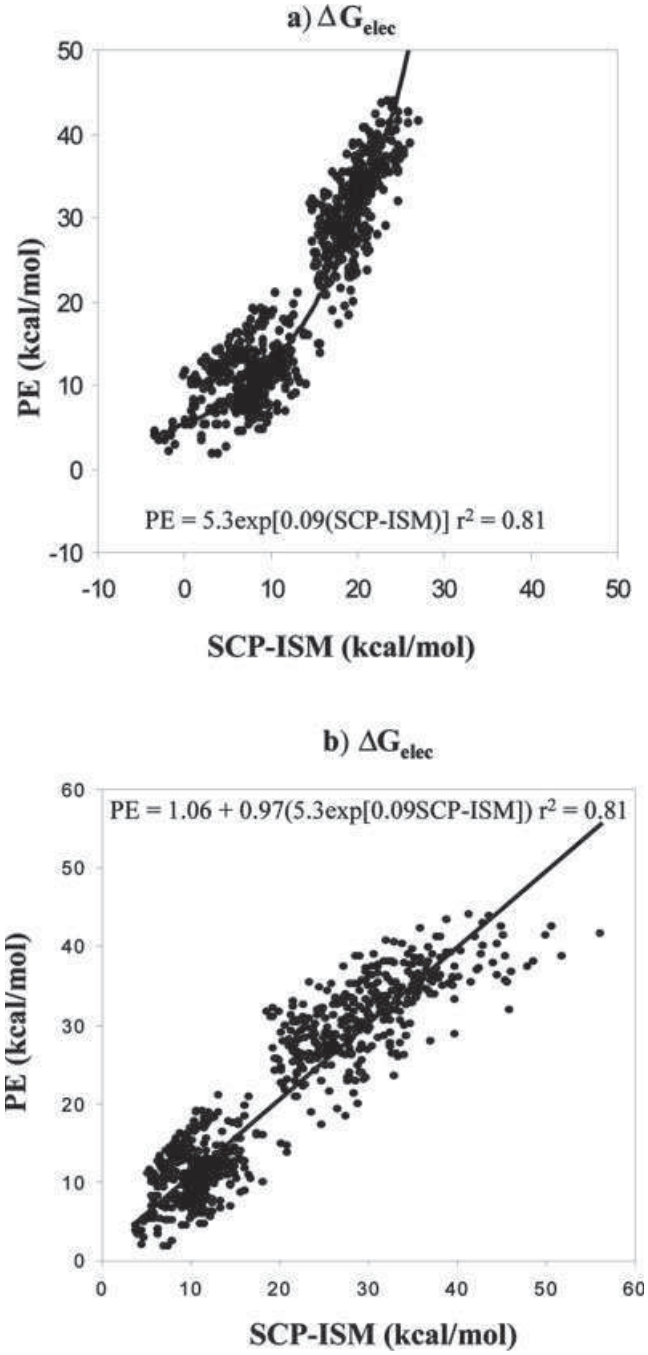


Fig. 2. Correlation between the total electrostatics binding free energy (in kcal/mol), as obtained by numerically solving the PE (*y*-axis), and by the SCP–ISM method (*x*-axis). Each point represents the corresponding energy pair for a different decoy. About 826 different decoys (summarized in Table I) have been employed. (**a**) Direct correlation. (**b**) Correlation after logarithmic correction. See text for details.

low and high dielectric regions, as in our sigmoidal dielectric function, increase the fraction of interstitial high dielectric regions in the protein interior. The presence of these regions has been shown to suppress the electrostatic free energy barriers characteristic of hydro-

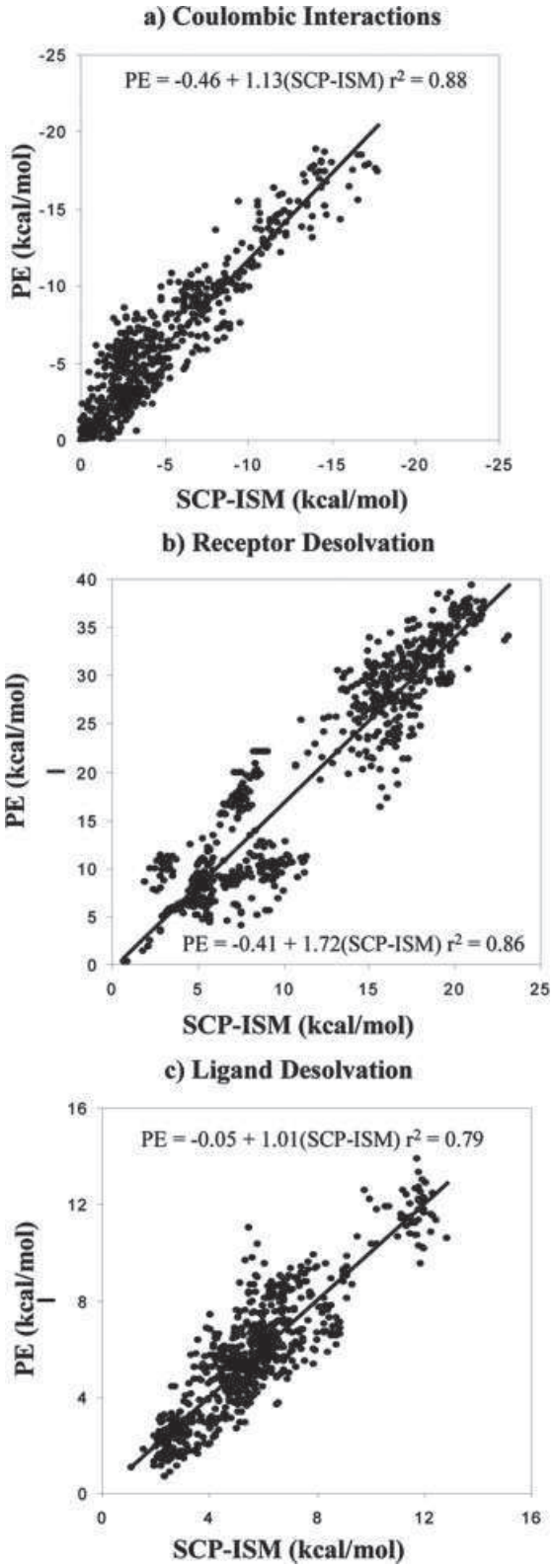


Fig. 3. Comparison of the different energy contributions to the electrostatics binding free energy, as obtained by solving the PE and by using SCP–ISM. (**a**) Coulombic contribution; (**b**) receptor desolvation, and (**c**) ligand desolvation. Each point represents the corresponding energy pair for a different decoy. About 826 different decoys were used.

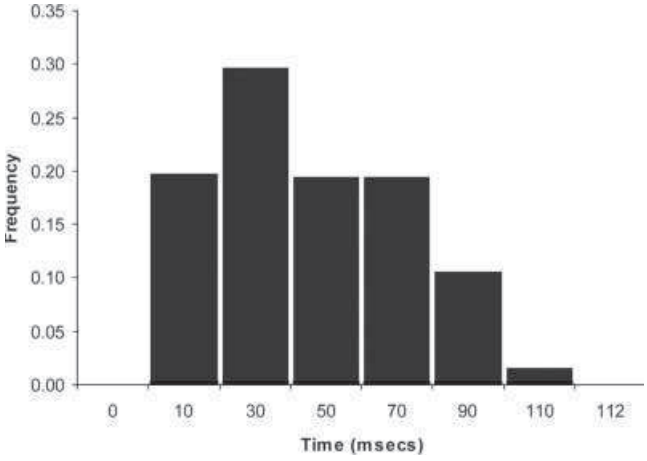


Fig. 4. Frequency distribution of the SCP–ISM computing time required to obtain the electrostatics binding free energy per decoy. The *x*-axis shows the range of computing times (in ms), while the *y*-axis shows the corresponding frequency in the set. The set of 826 decoys summarized in Table I has been employed. Calculations were performed in a 3.0 GHz Pentium IV computer.

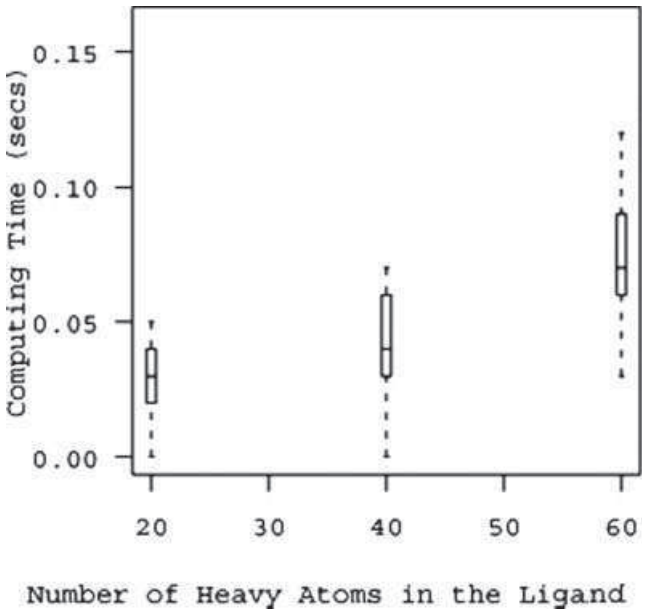


Fig. 5. Box and whisker plot showing the relationship between ligand size and computing time. The number of heavy atoms in the ligand is plotted versus the corresponding computing time. The set of 826 decoys summarized in Table I has been employed. Calculations were performed in a 3.0 GHz Pentium IV computer. See text for details.

gen bond formation when compared to atomistic potential of mean force simulations, leading to overestimation of solvation energies, particularly for groups involved in hydrogen bond interactions.[41] We speculate that our empirical hydrogen bond term might act as an "ad-hoc" correction to account for this effect. Nevertheless, future studies will address this matter in detail.

On the basis of studies with a large set of 826 decoys, covering substantial structural variety both in targets and ligands, satisfactory results have been obtained with the SCP–ISM model. The new method has a squared crossvalidated correlation coefficient with the electrostatics binding free energies obtained with the PE of 0.81, a slope of 0.97, an intercept of 1.06 kcal/mol, and a RMSD of about 4.33 kcal/mol [Fig. 2(b) and Table III]. The different contributions to the electrostatics binding free energies are also reproduced with similar accuracy (see Fig. 3). These results compare well with those recently obtained by Liu and Zou,[37] who studied the ability of GB to reproduce electrostatics binding free energies computed with PE. In their study, using crystal structures for 15 complexes in fitting and another 15 in crossvalidation, Liu and Zou obtained a squared correlation coefficient of 0.81 and a RMSD of 4.05 kcal/mol in fitting phase, while in crossvalidation the values obtained were 0.81 and 5.14, respectively. The comparison suggests that GB and SCP–ISM achieve similar performances in modelling protein–ligand binding energetics. Nevertheless, a log transformation in the SCP–ISM model was required to linearly fit the total electrostatics binding free energy to the PE results. We have found that the reason for this non linear effect rests on a relative (i.e., with respect the reference value computed with PE) overestimation of the desolvation free energy for proteins hosting long, narrow hydrophobic channels in the ligand binding site, such as the retinol binding protein or the biotin binding protein (see Table I for a description of the complexes). The reasons for this overestimation are unclear and are being investigated at present. On the other hand, we wish to emphasize that PE calculations are employed here more as a guideline, upon which our method should qualitatively conform, than as a reference quantitative golden standard. The PE method itself depends on a number of empirical parameters such as internal and external dielectric constants, boundary definitions, and so forth, which are not uniquely determined and are subjected to debate. For this reason we have not attempted to further "improve" fitting by introducing new sets of parameters or more empiricism into our model, and we have restricted our parameter search to physically meaningful quantities. The fact that in these conditions we obtain reasonable fits attests to the physical soundness of the SCP–ISM model.

The mean pose calculation time for the SCP–ISM model is about 30–40 ms (see Fig. 4), and an approximately linear relationship between ligand size and computing time is observed (see Fig. 5). This time is expected to be reduced further by employing a look-up table containing neighbouring atoms at each grid point, accelerating the calculation of the effective Born radii. Thus, the new method is shown, both in terms of timing and accuracy, to be good enough to be implemented directly into a docking algorithm, and compares favorably with other approaches. For example, the GB method implemented originally by Kuntz and coworkers in DOCK required ~10 s per complex on a SGI Octane workstation.[17] The same group later proposed a pair-

wise approach to compute the Born radii, reducing the computational time to 0.5 s per complex.[20] These timings prevent its direct use in the docking step, remaining only as a post-DOCK filter. On the other hand, Caflisch and coworkers proposed a simplified continuum method based on the assumption that electrostatic desolvation can be approximated by the removal of the first layer of water molecules at the binding interface, and the coulombic contribution can be approached by a distance dependent dielectric model.[23] Precomputation of the energy contributions on a set of grids allowed the authors to estimate the electrostatics binding free energy in solution in about 3–4 ms for fragments of 5–10 heavy atoms on a 550 MHz Pentium III. However, their method is restricted to docking of rigid molecules, since both ligand and receptor need to be grid-preprocessed, while our SCP–ISM can be employed for both rigid and flexible docking cases. This limits the applicability of the method of Caflisch and coworkers mainly to the docking of small rigid fragments in rigid binding sites. The accuracy of the total fitted electrostatics binding free energy obtained with the method of Caflisch and coworkers is also slightly worst than the one obtained with SCP–ISM, as judged by the squared correlation coefficients when compared with PE results ($\sim$0.75 vs. $\sim$0.81, respectively). Contributions to the electrostatics binding free energy are similarly reproduced by both methods (r^2 of $\sim$0.81 in both cases), but the SCP–ISM provides slopes close to 1.0 (see Fig. 3), while in the method of Caflisch and coworkers, the slopes are larger and show more dispersion (from 1.49 to 2.95).

In summary, although some descriptions are available to consider solvent effects in protein–ligand binding, they are either time consuming, inaccurate, or only applicable in very restricted conditions. This limits their usefulness in virtual screening projects, where millions of molecules with different conformers, tautomeric, and protonation states, need to be considered. The method presented here is a step in the direction of incorporating realistic, but fast, solvent models in large scale docking. We are currently incorporating the ISM method in our in-house docking program. Impact of the new electrostatics model in docking and virtual screening is being evaluated and will be presented in due time.

ACKNOWLEDGMENTS

Generous allocation of computer time at the Barcelona Supercomputer Center is gratefully acknowledged.

REFERENCES

1. Mohan V, Gibbs AC, Cummings MD, Jaeger EP, DesJarlais RL. Docking: successes and challenges. Curr Pharm Des 2005;11: 323–333.
2. Schwarzl SM, Tschopp TB, Smith JC, Fischer S. Can the calculation of ligand binding free energies be improved with continuum solvent electrostatics and an ideal-gas entropy correction? J Comput Chem 2002;23:1143–1149.
3. Sims PA, McCammon JA, Wong CF. A computational model of binding thermodynamics: the design of cyclin-dependent kinase 2 inhibitors. J Med Chem 2003;46:3314–3325.
4. Wang W, Donini O, Reyes CM, Kollman PA. Biomolecular simulations: recent developments in force fields, simulations of enzyme catalysis, protein–ligand, protein–protein, and protein–nucleic acid noncovalent interactions. Annu Rev Biophys Biomol Struct 2001;30:211–243.
5. Checa A, Ortiz AR, de Pascual-Teresa B, Gago F. Assessment of solvation effects on calculated binding affinity differences: trypsin inhibition by flavonoids as a model system for congeneric series. J Med Chem 1997;40:4136–4145.
6. Bernacki K, Kalyanaraman C, Jacobson MP. Virtual ligand screening against *Escherichia coli* dihydrofolate reductase: improving docking enrichment using physics-based methods. J Biomol Screen 2005;10:675–681.
7. Kalyanaraman C, Bernacki K, Jacobson MP. Virtual screening against highly charged active sites: identifying substrates of α–β barrel enzymes. Biochemistry 2005;44:2059–2071.
8. Shoichet BK, Leach AR, Kuntz ID. Ligand solvation in molecular docking. Proteins Struct Funct Genet 1999;34:4–16.
9. Huang D, Caflisch A. Efficient evaluation of binding free energy using continuum electrostatics solvation. J Med Chem 2004;47: 5791–5797.
10. Kuhn B, Gerber P, Schulz-Gasch T, Stahl M. Validation and use of the MM-PBSA approach for drug discovery. J Med Chem 2005;48:4040–4048.
11. Perez C, Ortiz AR. Evaluation of docking functions for protein–ligand docking. J Med Chem 2001;44:3768–3785.
12. Ferrara P, Gohlke H, Price DJ, Klebe G, Brooks CL, III. Assessing scoring functions for protein–ligand interactions. J Med Chem 2004;47:3032–3047.
13. Orozco M, Luque FJ. Theoretical methods for the description of the solvent effect in biomolecular systems. Chem Rev 2000;100: 4187–4225.
14. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144–1149.
15. Bashford D, Case DA. Generalized born models of macromolecular solvation effects. Annu Rev Phys Chem 2000;51:129–152.
16. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 1990;112:6127–6129.
17. Zou X, Sun Y, Kuntz ID. Inclusion of solvation in ligand binding free energy calculations using the generalized-born model. J Am Chem Soc 1999;121:8033–8043.
18. Majeux N, Scarsi M, Apostolakis J, Ehrhardt C, Caflisch A. Exhaustive docking of molecular fragments with electrostatic solvation. Proteins 1999;37:88–105.
19. Taylor RD, Essex JW, Jewsbury PJ. FDS: flexible ligand and receptor docking with a continuum solvent model and soft-core energy function. J Comput Chem 2003;24:1637–1656.
20. Liu H-Y, Zou X, Kuntz ID. Pairwise GB/SA scoring function for structure-based drug design. J Phys Chem B 2004;108:5453–5462.
21. Wang J, Kang X, Kuntz ID, Kollman PA. Hierarchical database screenings for HIV-1 reverse transcriptase using a pharmacophore model, rigid docking, solvation docking, and MM-PB/SA. J Med Chem 2005;48:2432–2444.
22. Arora N, Bashford D. Solvation energy density occlusion approximation for evaluation of desolvation penalties in biomolecular interactions. Proteins 2001;43:12–27.
23. Majeux N, Scarsi M, Caflisch A. Efficient electrostatic solvation model for protein-fragment docking. Proteins 2001;42:256–268.
24. Hassan SA, Mehler EL. A critical analysis of continuum electrostatics: the screened Coulomb potential–implicit solvent model and the study of the alanine dipeptide and discrimination of misfolded structures of proteins. Proteins 2002;47:45–61.
25. Hassan SA, Guarnieri F, Mehler EL. Characterization of hydrogen bonding in a continuum solvent model. J Phys Chem B 2000;104:6490–6498.
26. Hassan SA, Guarnieri F, Mehler EL. General treatment of solvent effects based on screened Coulomb potentials. J Phys Chem B 2000;104:6478–6489.
27. Morris GM, Goodsell DS, Huey R, Olson AJ. Distributed automated docking of flexible ligands to proteins: parallel applica-

tions of AutoDock 2.4. J Comput Aided Mol Des 1996;10:293–304.
28. Weiser J, Shenkin PS, Still WC. Approximate atomic surfaces from linear combinations of pairwise overlaps (LCPO). J Comput Chem 1999;20:217–230.
29. McDonald IK, Thornton JM. Satisfying hydrogen bonding potential in proteins. J Mol Biol 1994;238:777–793.
30. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. J Comput Chem 2002;23:128–137.
31. Sitkoff DS, Sharp KA, Honig B. Accurate calculation of hydration free energies using macroscopic solvent models. J Phys Chem 1994;98:1978–1988.
32. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1995;117:5179–5197.
33. Stewart JJ. MOPAC: a semiempirical molecular orbital program. J Comput Aided Mol Des 1990;4:1–105.
34. Schutz CN, Warshel A. What are the dielectric "constants" of proteins and how to validate electrostatic models? Proteins 2001;44:400–417.
35. Misra VK, Sharp KA, Friedman RA, Honig B. Salt effects on ligand–DNA binding. Minor groove binding antibiotics. J Mol Biol 1994;238:245–263.
36. Misra VK, Honig B. On the magnitude of the electrostatic contribution to ligand–DNA interactions. Proc Natl Acad Sci USA 1995;92:4691–4695.
37. Liu HY, Zou X. Electrostatics of ligand binding: parametrization of the generalized Born model and comparison with the Poisson–Boltzmann approach. J Phys Chem B Condens Matter Mater Surf Interfaces Biophys 2006;110:9304–9313.
38. Grater F, Schwarzl SM, Dejaegere A, Fischer S, Smith JC. Protein/ligand binding free energies calculated with quantum mechanics/molecular mechanics. J Phys Chem B Condens Matter Mater Surf Interfaces Biophys 2005;109:10474–10483.
39. Roche O, Kiyama R, Brooks CL, III. Ligand–protein database: linking protein–ligand complex structures to binding data. J Med Chem 2001;44:3592–3598.
40. Murcia M, Ortiz AR. Virtual screening with flexible docking and COMBINE-based models. Application to a series of factor Xa inhibitors. J Med Chem 2004;47:805–820.
41. Swanson JM, Mongan J, McCammon JA. Limitations of atom-centered dielectric functions in implicit solvent models. J Phys Chem B Condens Matter Mater Surf Interfaces Biophys 2005;109:14769–14772.

Structure-Based Discovery of Novel Non-nucleosidic DNA Alkyltransferase Inhibitors: Virtual Screening and in Vitro and in Vivo Activities

Federico M. Ruiz,[†] Rubén Gil-Redondo,[‡] Antonio Morreale,[‡] Ángel R. Ortiz,*[,‡] Carmen Fábrega,*[,†] and Jerónimo Bravo*[,†]

Signal Transduction Group, Structural Biology and Biocomputing Programme, Centro Nacional de Investigaciones Oncológicas (CNIO), Melchor Fernández Almagro 3, E-28029 Madrid, Spain, and Bioinformatics Unit, Centro de Biología Molecular Severo Ochoa (CSIC-UAM), Universidad Autónma de Madrid, Nicolás Cabrera, 1. Cantoblanco, 28049 Madrid, Spain

Received November 30, 2007

The human DNA-repair O^6-alkylguanine DNA alkyltransferase (MGMT or hAGT) protein protects DNA from environmental alkylating agents and also plays an important role in tumor resistance to chemotherapy treatment. Available inhibitors, based on pseudosubstrate analogs, have been shown to induce substantial bone marrow toxicity in vivo. These deficiencies and the important role of MGMT as a resistance mechanism in the treatment of some tumors with dismal prognosis like glioblastoma multiforme, the most common and lethal primary malignant brain tumor, are increasing the attention toward the development of improved MGMT inhibitors. Here, we report the identification for the first time of novel non-nucleosidic MGMT inhibitors by using docking and virtual screening techniques. The discovered compounds are shown to be active in both in vitro and in vivo cellular assays, with activities in the low to medium micromolar range. The chemical structures of these new compounds can be classified into two families according to their chemical architecture. The first family corresponds to quinolinone derivatives, while the second is formed by alkylphenyl-triazolo-pyrimidine derivatives. The predicted inhibitor protein interactions suggest that the inhibitor binding mode mimics the complex between the excised, flipped out damaged base and MGMT. This study opens the door to the development of a new generation of MGMT inhibitors.

In spite of the considerable progress in cancer cell biology, most cancer treatments are still multimodal, involving extensive surgery, radiotherapy, and chemotherapy treatment. Chemotherapy remains the most important pharmacological approximation to cancer therapy. Cytotoxic alkylating agents (e.g., streptozotocin, procarbazine, or dacarbazine) are the oldest family of anticancer drugs.[1] The sites of reaction of alkylating agents in guanine include N^1, N^3, N^7, and O^6. The N^7 position is the most reactive site,[2–4] however the DNA functions are most strongly affected by alkylation in the O^6 position.[5] 1,3-Bis-(2-chloroethyl)-1-nitrosourea (BCNU) attacks initially at the O^6 guanine position followed by formation of a cyclic intermediate with attack at the N^1 position of guanine, giving rise to N^1O^6 ethanoguanine. Finally the structure rearranges from the O to form a cross-link with the opposite cytosine.[6] Eventually, DNA replication is blocked, producing G2/M rest.[7] In addition to the well-known side effects and limitations of chemotherapeutic agents, they also present problems of acquired tumor resistance. Particularly, the human DNA-repair O^6-alkylguanine DNA alkyltransferase (MGMT or hAGT), an important protein that protects DNA from environmental alkylating agents, also plays an important role as a resistance mechanism. It is well established that resistance to both chloroethylating and methylating chemotherapeutic agents and cyclophosphamide (a nitrogen mustard alkylating agent) is mediated by MGMT activity,[8–11] as tumor cells frequently express high levels of this enzyme. This effect has been observed in a number of cancers, ranging from colon cancer, lung tumors, breast cancer, pancreatic tumors and non-Hodgkin's lymphomas to myelomas and gliomas, among others.[12–14] It is significant that MGMT promoter methylation, and consequently complete MGMT depletion, has been statistically associated with longer survival in patients with high grade gliomas under radiation-chemotherapy combined treatment.[15,16] Pharmacological inhibition of MGMT, therefore, has the potential to enhance the cytotoxic effect of a diverse range of anticancer agents, particularly in colon and brain tumors.

Despite pharmacological interest and the efforts in the structural biology of DNA repair, the first two structures of MGMT-damaged DNA have only been published recently.[17] These structures, an inactive C145S mutant bound to an O^6-methylguanine-containing oligonucleotide and an active MGMT covalently cross-linked to an oligonucleotide containing N^1,O^6-ethonaxanthosine, provided both, novel protein-DNA architecture and the structural basis for the reaction mechanism. The DNA-binding helix-turn-helix (HTH) motif is placed in the MGMT C-terminal domain of this 22 kDa protein (207 AA) with a two-domain α/β fold.[18] The second helix of this motif binds deep within the DNA minor groove. MGMT binds without changes but widening the minor groove and therefore bending the DNA. Arg128 is positioned

* To whom correspondence should be addressed. Tel.: 34-912246900 (C.F. and J.B.); 34-911964633 (A.R.O.). Fax: 34-912246976 (C.F. and J.B.); 34-911964420 (A.R.O). E-mail: cfabrega@cnio.es (C.F.); jbravo@cnio.es (J.B.); aro@cbm.uam.es (A.R.O.).
[†] Centro Nacional de Investigaciones Oncológicas (CNIO).
[‡] Centro de Biologia Molecular Severo Ochoa (CSIC-UAM).

10.1021/ci700447r CCC: $40.75
Published on Web 03/20/2008

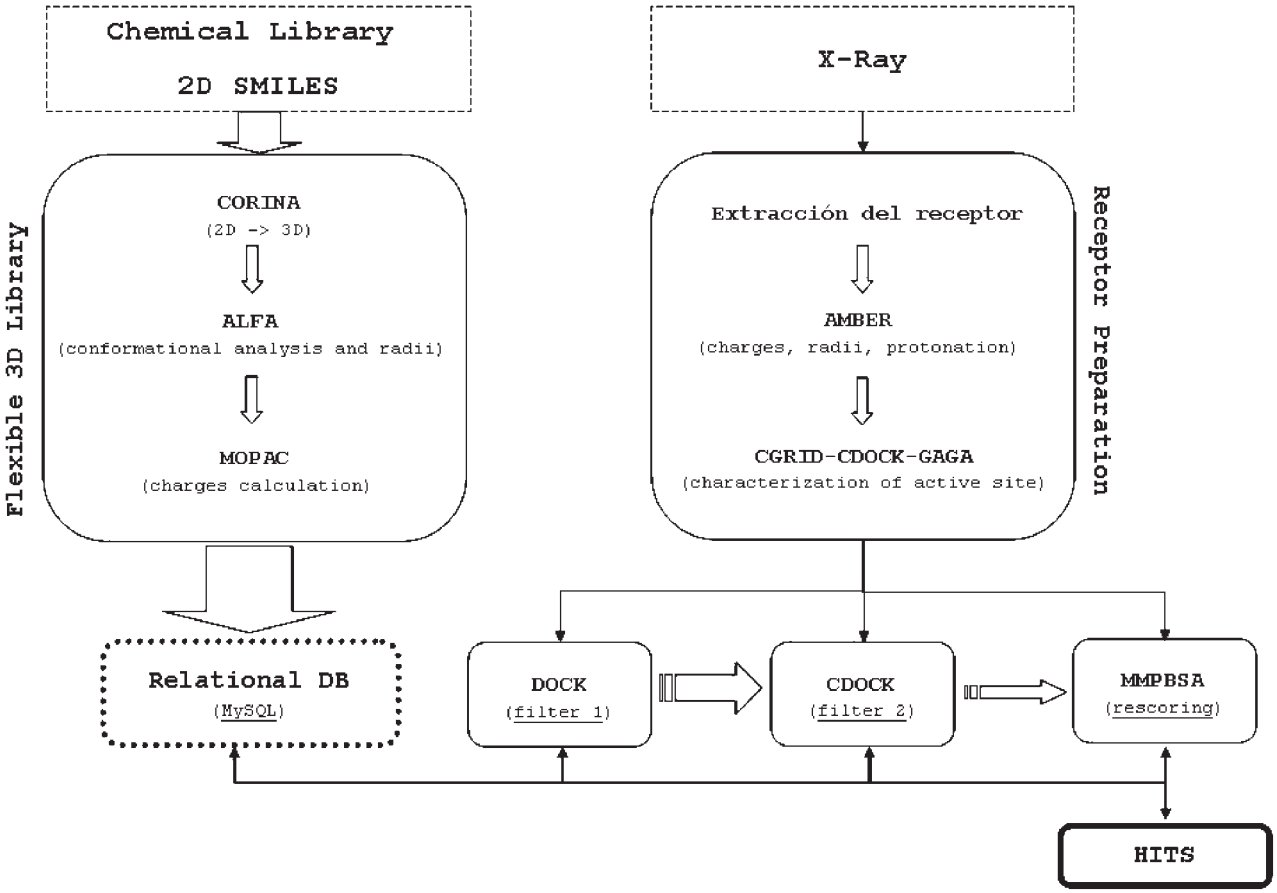


Figure 1. Flowchart of the virtual screening procedure applied in this work. See the main text for details.

inside the DNA duplex, stabilizing the structure by hydrogen bonds with the orphaned cytosine. This 'arginine finger,' also seems to promote the flipping of nucleotides out from the base stack, and into the MGMT hydrophobic active site. The observation that MGMT activity is decreased more than 1000 times when Arg128 is replaced by alanine confirms the essential role played by this residue.[19]

The nucleophilic residue, Cys145, is buried deeply into the active site, near the bottom of the hydrophobic active site pocket. This residue participates in a hydrogen bond network involving His146, Glu172, and a water molecule, which may act as a relay able to increase cysteine reactivity upon substrate binding. In the proposed enzymatic mechanism, His146 acts as a water-mediated base to deprotonate Cys145, which receives the alkyl lesion in a SN2 manner.[20] The DNA is restored but the protein, acting as a suicidal enzyme, inactivates itself in the process. Finally, the alkylated protein is destabilized,[21] recognized by an E3 ubiquitin ligase with its associated E2 ubiquitin-conjugating enzyme and degraded by an ATP-dependent pathway.[22,23]

Several small molecules acting as MGMT inhibitors have been developed during the past few years, all of them pseudosubstrates. The guanine analogue O^6-methylguanine (O^6-MG) was the first inhibitor discovered. It reduces MGMT activity in cells; however, it is not effective enough to be used in animal or clinical assays. A more rapid and effective depletion of MGMT was obtained with O^6-benzylguanine (O^6-BG).[24] O^6-BG acts as an alternative substrate for MGMT, transferring the benzyl group to the Cys145 and irreversibly inactivating the protein.[25] It has an IC_{50} value of 0.2 μM against MGMT,[26] significantly enhancing the cytotoxic effect of BCNU in prostate, breast, colon and lung tumor cells[27] and in tumor xenograft studies.[28] Several phase I and II clinical studies of the combination of O^6-BG and BCNU have been completed.[29] However, significant therapeutic limitations have been observed: O^6-BG has low bioavailability, poor water solubility and rapid plasma clearance.[30] There is no evidence of toxicity associated to O^6-BG alone, however, when combined with BCNU, it increases myelo-suppression.[31,32] In addition, repeated administration of O^6-BG with BCNU raises the possibility of developing O^6-BG-resistantance. A point mutation in Lys165 has been associated with acquired O^6-BG and BCNU resistance in MMR-deficient medulloblastoma cell lines.[33] Direct evidence of the relation between Lys165 mutation and BCNU activity has been shown in MMR-deficient colon cancer cells.[34] In contrast MGMT independent resistance to O^6-BG has been found in breast cancer cells after treatment with this MGMT inhibitor plus BCNU.[35] Recently different guanine derivatives have been used in MGMT inhibition studies. Among them, lomeguatrib [6-(4-bromo-2-thienyl) methoxy]purin-2-amine] is more active in vitro than O^6-BG, having an IC_{50} value of 0.018 μM.[34] This O^6-thenyl analogue of O^6-BG has been selected for clinical experiments and successfully used in combination with Temozolomide in a first phase I trial.[36]

Searching for new molecules has become in one of the most actives areas in computational chemistry and biology. Virtual screening protocols are being used routinely in this regard and there are many successful cases where

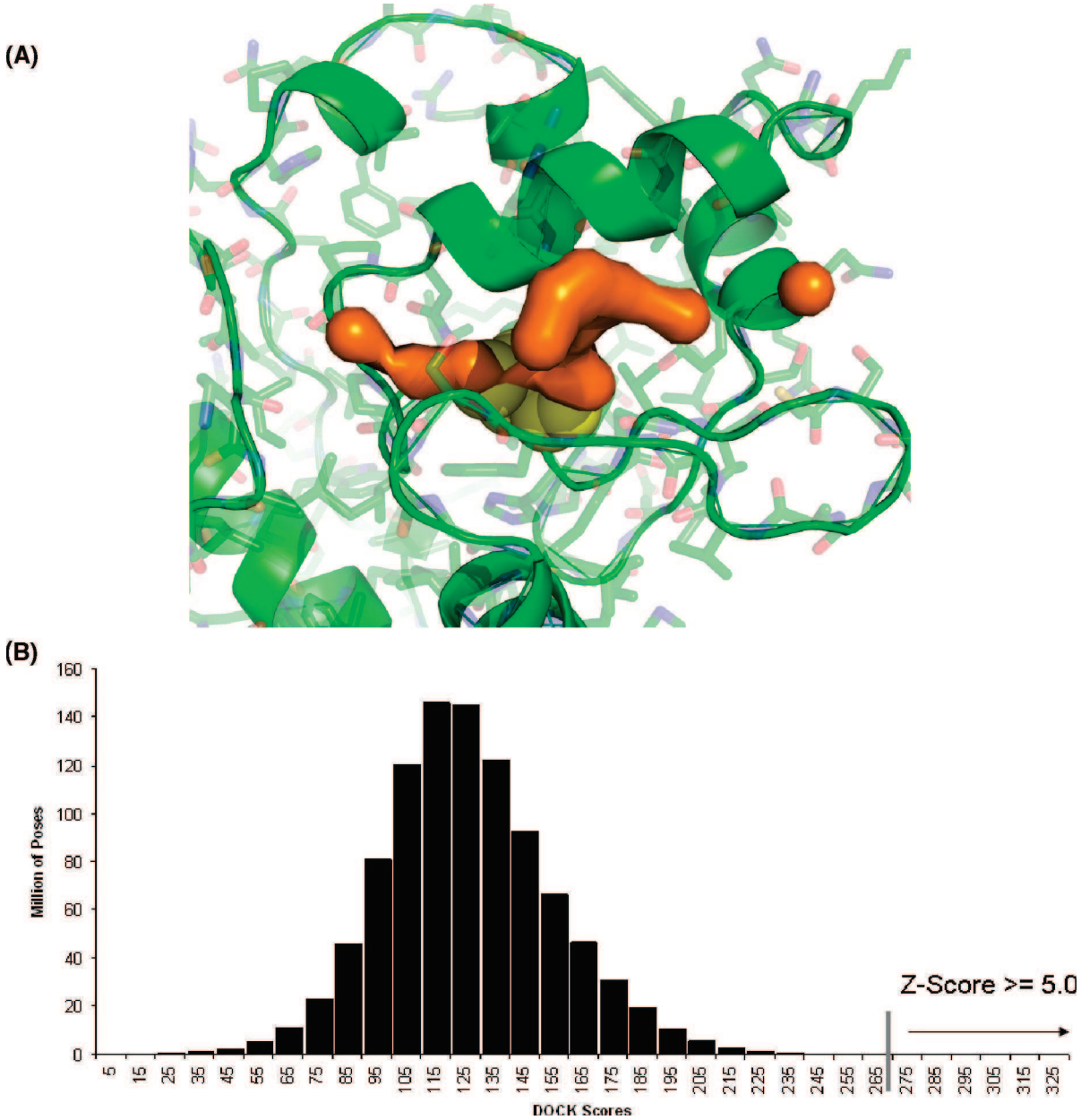


Figure 2. (A) Negative spheres image of the MGMT active site computed with GAGA. (B) DOCK score distribution showing the Zscore cutoff value applied.

novel inhibitors have been found.[37,38] The underlying engine that moves virtual screening consists of two pieces: a docking algorithm to sample the binding site of a receptor target, and a mathematical scoring function to assign a score to each binding site pose.[39] Usually, only the best solution for each molecule, the lowest in energy, is considered. With the advent of supercomputer virtual screening of chemical libraries is becoming a feasible issue, and it is customary to screen thousands or even millions of molecules in some virtual experiments. Screening such a large number of molecules comes with the problem that extremely large amount of time is required, and then accuracy is reduced to maintain time in a reasonable range. Accordingly, the receptor is kept rigid along the experiment, environmental effects (mainly due to solvent) are complete ignored or treated at a very low theoretical level, entropic effect are rarely taken into account. These shortcomings are often alleviated by post processing the highest ranked candidates (between 50 and 100) to more elaborated protocols as molecular dynamics simulations.[40,41] Then, selected snapshots from these trajectories are treated with approximations as MMPBSA,[42] MMGBSA,[43] or LIE[44] among others, to obtain an estimation of the free energy of binding, a measure comparable with experimental inhibition constants.

Motivated by the deficiencies of known inhibitors and the important role of MGMT in difficult tumors like gliomastoma multiforme (the most common and lethal primary malignant brain tumor), we report here the identification, for the first time, of novel non-nucleosidic MGMT inhibitors. Docking and virtual screening techniques, followed by molecular mynamics simulations and free energy of binding calculation with MMGBSA method yielded four promising compounds with experimental probed activities both in vitro and in vivo.

Table 1. List of the 17 Top-Ranked Compounds Obtained in the Virtual Screening Computation[a]

	compound (ZINC code)	log P	H-bond donors	H-bond acceptors	charge	MW[b]	CDOCK energy	MMGBSA energy[c]	IC_{50} (μM)[d]	IC_{50} (μM)[e]
1	(ZINC00910802)	3.52	2	9	1	536	−34.57	−32.26 (2.59)	54	10
2	(ZINC00889422)	4.24	2	7	1	505	−31.91	−43.54 (3.35)	34	50
3	(ZINC03642335)	6.18	1	5	0	410	−32.42	−46.52 (3.26)	24	10
4	(ZINC02487935)	5.61	1	6	0	426	−32.37	−56.90 (3.24)	22	10
5	(ZINC01327643)	4.23	2	6	0	437	−31.24	ND[f]	>100	ND
6	(ZINC00714917)	6.41	1	6	0	503	−31.84	ND	>100	ND
7	(ZINC01360953)	5.01	0	7	0	563	−32.06	ND	>100	ND
8	(ZINC01360953)	4.51	2	6	0	433	−31.73	ND	>100	ND
9	(ZINC02809317)	1.34	0	11	0	463	−32.95	ND	>100	ND
10	(ZINC03404767)	4.88	1	8	0	481	−34.08	ND	>100	ND
11	(ZINC01437200)	2.81	2	8	2	479	−32.71	ND	>100	ND
12	(ZINC03052303)	3.50	2	7	0	516	−33.90	ND	>100	ND
13	(ZINC00784955)	0.90	3	9	2	452	−32.61	ND	>100	ND
14	(ZINC00892609)	4.45	2	7	1	505	−31.56	ND	>100	ND
15	(ZINC01352201)	3.06	1	9	0	472	−33.19	ND	>100	ND
16	(ZINC02835223)	4.34	1	6	0	433	−28.35	ND	>100	ND
17	(ZINC00738815)	4.48	2	6	0	460	−32.52	ND	>100	ND

[a] The computed chemical properties (as found in the ZINC database), the computed binding energies (computed both with CDOCK and the MMGBSA method, see Materials and Methods for details), and the in vitro and in vivo activities of the active compounds are shown. [b] MW molecular weight. [c] Average interaction energy during the MD simulation in kilocalories per mole; standard deviation is shown in parenthesis. [d] IC_{50} in vitro value (concentration of the compounds required to produce 50% reduction in the MGMT activity). [e] IC_{50} in vivo value (concentration of the compounds required to produce 50% cell killing in the presence of 80 μM sBCNU). [f] ND not determined.

Table 2. Interaction Energy Analysis (Standard Deviations in Parentheses), As Computed from the Molecular Dynamics Simulations by the MMGBSA Approach, for the Four Active Molecules Found in This Work[a]

	compound			
res no.	**1**	**2**	**3**	**4**
ARG128	−5.99 (0.65)	−4.09 (1.51)	−6.01 (0.99)	−6.71 (0.64)
TYR114	−1.92 (0.31)	−4.33 (0.47)	−5.46 (0.65)	−4.87 (0.46)
ARG135	−5.52 (0.77)	−1.39 (0.56)	−3.51 (0.84)	−4.75 (0.95)
TYR158	−1.53 (0.44)	−3.17 (0.41)	−1.39 (0.24)	−4.03 (0.55)
GLY131		−2.31 (0.44)	−2.83 (0.36)	−3.17 (0.38)
ASN157	−2.69 (0.65)	−3.34 (0.42)	−1.38 (0.27)	−2.93 (0.52)
MET134		−3.11 (0.51)	−2.25 (0.34)	−2.58 (0.41)
ALA127	−1.11 (0.24)			−1.68 (0.29)
SER159	−1.30 (0.29)	−1.65 (0.35)		−1.55 (0.29)
GLN115			−2.36 (0.62)	−1.37 (0.33)
CYS150				−1.18 (0.46)
CYS145		−1.21 (0.32)		−1.01 (0.42)
total	**−20.06 (0.48)**	**−24.60 (0.54)**	**−25.19 (0.51)**	**−35.83 (0.48)**

[a] All values are in kilocalories per mole.

RESULTS AND DISCUSSION

Virtual Screening (VS). The virtual screening protocol employed is summarized in Figure 1. and briefly described in the Materials and Methods section. An essential part in the procedure is to characterize the shape of the active site. For this we use our algorithm GAGA (see Materials and Methods) to obtain a negative image of the binding site (see Figure 2A). It can be seen that it covers the active site pocket and protrudes toward the neighborhood of Tyr114. Overall, the shape of the negative image resembles the shape of the target nucleotide bound to MGMT. Upon charactering the binding site, a library of 2.3 million compounds was screened. Compounds were first filtered with DOCK[45] using the negative image of the binding site as computed with GAGA. We chose to employ a Zscore (see Materials and Methods) cutoff value of 5 in the filtering step. From the initial set of 2.3 million of molecules, 1664 passed the ZScore cut off (Figure 2B). These molecules were then further screened with the CDOCK program, our in-house docking program. The CDOCK energies, computed with the CGRID molecular mechanics energy function, were solvent corrected in order to obtain the final scoring (see Materials and Methods). From the highest scoring compounds, and upon visual examination, 17 compounds were finally selected, purchased, and tested experimentally. Four out of the 17 showed activity against MGMT, and were further analyzed by means of molecular dynamic simulations in explicit solvent (see Materials and Methods). For these four compounds, the commonly employed MMGBSA method to estimate free energy of binding from molecular dynamic trajectories was used. CDOCK binding energies for all the 17 compounds, MMGBSA binding energies for the four more active compounds, together with other chemical and physicochemical properties of the molecules, as stored in the ZINC database,[46] are shown in Table 1. In addition, interaction energy analysis between ligand and the more relevant residues in the binding site were computed (with MMGBSA) and are contained in Table 2.

MGMT in Vitro Assays. The 17 top-ranked compounds selected from the VS computations where purchased and

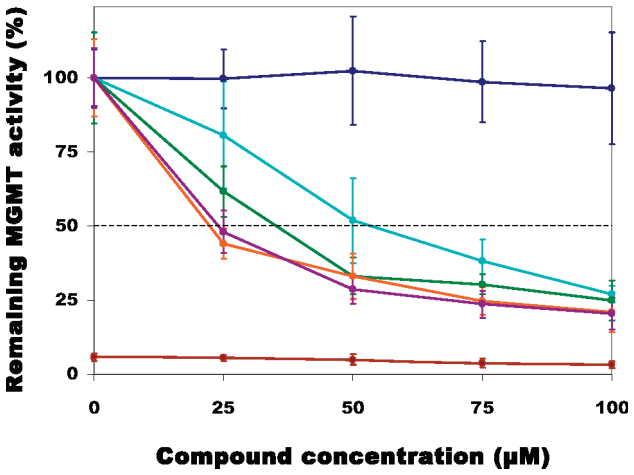


Figure 3. Concentration curve showing the inactivation of human alkyltransferase by compounds **1**, **2**, **3**, and **4**. Remaining MGMT activity vs compound concentration is shown relative to untreated control samples. Compounds **1** (cyan), **2** (green), **3** (orange), and **4** (violet) show inhibitory effect on MGMT activity in the micromolar range. The negative control (inactive C145S MGMT protein) is shown in brown, and the effect of the compound solvent (DMSO) on MGMT activity is shown in blue. The dotted line marks the 50% remaining MGMT activity.

dissolved in DMSO. The ability of those 17 candidates to inhibit the recombinant MGMT activity in vitro was determined by measuring the radioactivity transfer from ^{3}H-methyled DNA to the active site residue Cys145, as described in the Experimental Section. The resultant IC_{50} values obtained in these experiments for all compounds are shown in Table 1 and in Figure 3. It was found that compounds 5 to 17 did not exhibit significant inhibitor activity in the concentration range used in this assay (Table 1) and therefore were discarded for the subsequent in vivo assays (see below). On the other hand, compounds 1, 2, 3, and 4 inactivated MGMT in the low to medium micromolar range (Figure 3 and Table 1). We checked that MGMT activity was unaffected by any of the solvent (DMSO) volumes used to achieve the desired compound concentration. We did not detect changes, over all the tested compound concentration range, in the remaining radioactivity when an inactive C145S MGMT mutant was used as negative control (Figure 3). The inhibition at 50 μM compound concentration was not attenuated by addition of 0.001% Triton X-100 or BSA 2 mg/mL (data not shown), confirming that the four compounds were not promiscuous inhibitors.[47,48]

The chemical structures of these compounds are shown in Figure 4. They can be classified into two families according to their chemical architecture. The first family corresponds to quinolinone derivatives 3-[(4-(2,5-dimethylphenyl)-1-piperazinyl)(1-(phenylmethyl)-1*H*-tetrazol-5-il)-methyl]-6-methoxy-2(1*H*)-quinolinone (**1**) and 3-[(4-benzyl-1-piperadinyl)(1-benzyl-1*H*-5-tetrazolyl)]methyl-7-methyl-2(1*H*)-quinolinone (**2**). According to the in vitro activity assays, their affinity values are 54 and 34 μM, respectively, slightly larger than the affinities obtained for the second family. The second family is composed by triazolo-pyrimidine derivatives 7-(4-(benzyloxy)phenyl)-5-(4-ethylphenyl)-4,5,6,7-tetrahydro-(1,2,4)triazol(1,5-a)pyrimidine (**3**) and 7-(4-(benzyloxy)phenyl)-5-(4-ethoxyphenyl)-4,5,6,7 -tetrahydro-(1,2,4)triazol(1,5-a)pyrimidine (**4**). The corresponding affinities for these compounds are 24 and 22 μM, respectively.

MGMT in Vivo Assays. On the basis of IC_{50} values obtained for the inactivation of pure recombinant MGMT in the in vitro assay, compounds **1**−**4** were further analyzed and validated with in vivo assays. The colony forming assay[49,50] was used in order to study the capability of compounds **1**−**4** to enhance BCNU cytotoxicity using HTB-38 cells. Cells were incubated with these compounds before, during and after BCNU treatment to ensure that inhibitors were present during the entire period of time needed for DNA adducts to be formed. As shown in Figure 5, all four compounds enhance BCNU citotoxicity. BCNU alone reduced the number of colonies by 30%. The cell sensitivity to this chemotherapeutic agent was increased to the 50% when compounds **1**, **3**, and **4** were added at 10 μM. Only compound **2** needed to be added up to a concentration of 50 μM to get the same effect. Compound **2** presented little if any sensitizing effect (Figure 5), despite its ability to inhibit MGMT activity in vitro. The ineffectiveness of compound **2** is probably related to a reduced cell penetration. Compound **1** is less effective sensitizing HTB 38 cell to BCNU than compounds **3** and **4**. This is consistent both with their slightly larger in vitro affinity and with their larger log P values. Finally, colony forming experiments have also been carried out in the absence of BCNU, showing an average decrease of 15% using a compound concentration of 50 μM. These results confirm that cell killing is the final outcome of the joint action between BCNU and the studied compounds.

Description of the Docking Modes. The predicted interactions and docking modes of these 4 active compounds, as obtained after docking and molecular dynamics simulations, can be seen in Figure 6. The predicted binding modes suggest that the bound conformation of the inhibitors strongly mimics the observed conformation of the excised, flipped-out nucleotide bearing the damaged base in the complex with MGMT (Figure 6). Thus, the quinolinone fragment of the inhibitors in family 1 (Figure 6A), as well as the alkylphenyl radical of the triazolopyrimidine core in family 2 (Figure 6B), are predicted to occupy the MGMT catalytic cleft, playing the role of analogs of the O^6-guanine moiety in the natural substrate, burying deep within the binding groove and reaching the catalytic residue Cys145. In both cases, the remaining portion of the inhibitors is predicted to protrude outside the catalytic pocket and occupy the neighborhood of both the Arg135 and Tyr114. A summary of the most important interactions for each one of the four inhibitors, as computed with the MMGBSA method on the basis of the molecular dynamics simulations, can be found in Table 2. Thus, the tetrazol moiety in family 1 and the triazol group in family 2 are predicted to occupy the site and act as an isosteric group of the 5′ phosphate binding site of the damaged base, adjacent to the active site pocket, and allowing the group to interact with Arg135. This is consistent with the fact that tetrazol groups are well-known isosters of anionic groups, such as carboxylic acids or phosphonates. Similarly, the benzyl-piperazinyl and benzyl-piperadinyl moieties of the inhibitors in family 1, and the benzyloxi-benzyl radical in family 2, are predicted to stack against the aromatic ring of Tyr114. The ranking of the MMGBSA computed average interaction energies correlate well with the observed affinity differences (Table 1), providing some support to the predicted docking modes. Finally, for each complex we computed the most important interactions

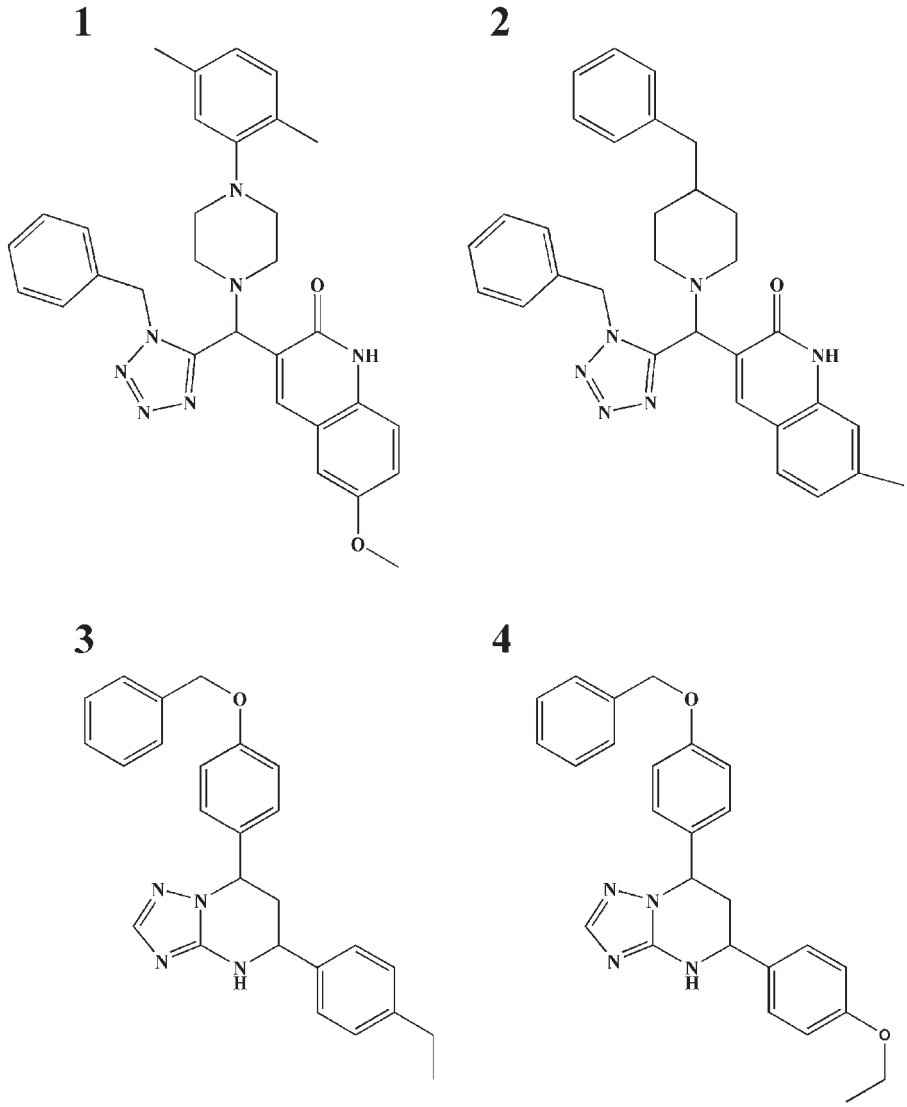


Figure 4. Chemical structures of the four compounds that have shown MGMT inhibition in the micro molar range: **1** 3-[(4-(2,5-dimethylphenyl)-1-piperazinyl)(1-(phenylmethyl)-1*H*-tetrazol-5-il)methyl]-6-methoxy-2(1*H*)-quinolinone; **2** 3-[(4-benzyl-1-piperadinyl)(1-benzyl-1*H*-5-tetrazolyl)]methyl-7-methyl-2(1*H*)-quinolinone; **3** 7-(4-(benzyloxy)phenyl)-5-(4-ethylphenyl)-4,5,6,7-tetrahydro-(1,2,4)triazol(1,5-a)pyrimidine; **4** 7-(4-(benzyloxy)phenyl)-5-(4-ethoxyphenyl)-4,5,6,7 -tetrahydro-(1,2,4)triazol(1,5-a)pyrimidine.

between ligand and protein, as obtained from the MMGBSA analysis of the molecular dynamics simulations (Table 2). For all inhibitors the most relevant interactions take place with Arg128 (the arginine finger) and Arg135, as well as Tyr114 and Tyr158. This is consistent with the importance of the residues surrounding the catalytic site as observed in site mutagenesis studies.

CONCLUSION

We have applied docking and virtual screening techniques to identify novel MGMT inhibitors. Out of 17 inhibitors selected from the ZINC database, four were found to be active. Thus, a success rate of about ~20% for our screening procedure can be estimated. This success rate is consistent with results reported by other groups and highlights the increasingly important role of virtual screening techniques in the search for bioactive molecules. The new compounds belong to two new families of non-nucleosidic inhibitors which, with further optimization, could help to overcome some of the side effects of the existing MGMT inhibitors when combined with alkylating agents. Both families are active in vitro and in vivo in the low to medium micromolar range. The predicted binding modes suggest that the bound conformation of the inhibitors mimics the observed conformation of the flipped-out nucleotide in complex with MGMT. In summary, these novel compounds may form the basis for the development of a new generation of non-nucleosidic MGMT inhibitors with improved pharmacological properties as coadjuvants in cancer chemotherapy.

EXPERIMENTAL SECTION

Materials and Methods. Human colorectal adenocarcinoma cells (ATCC Number HTB-38) were cultured in RPMI 1640 medium (Genycell) supplemented with 10% fetal bovine serum (Biowhittaker). BCNU (1,3-bis-(2-chloroethyl)-1-nitrosourea) was obtained from Sigma and dissolved in 50% phosphate-buffered saline buffer (PBS)−50% ethanol at 4 mM stock solution. *N*-[^{3}H]Methyl-*N*-nitrosourea (MNU) (18.5 MBq/mL, 5 mCi/mL) was purchased from Amersham Biosciences. Candidate compounds were purchased from different companies, in particular compounds **1** and **2** were obtained from Asinex; compounds **3** and **4** were obtained

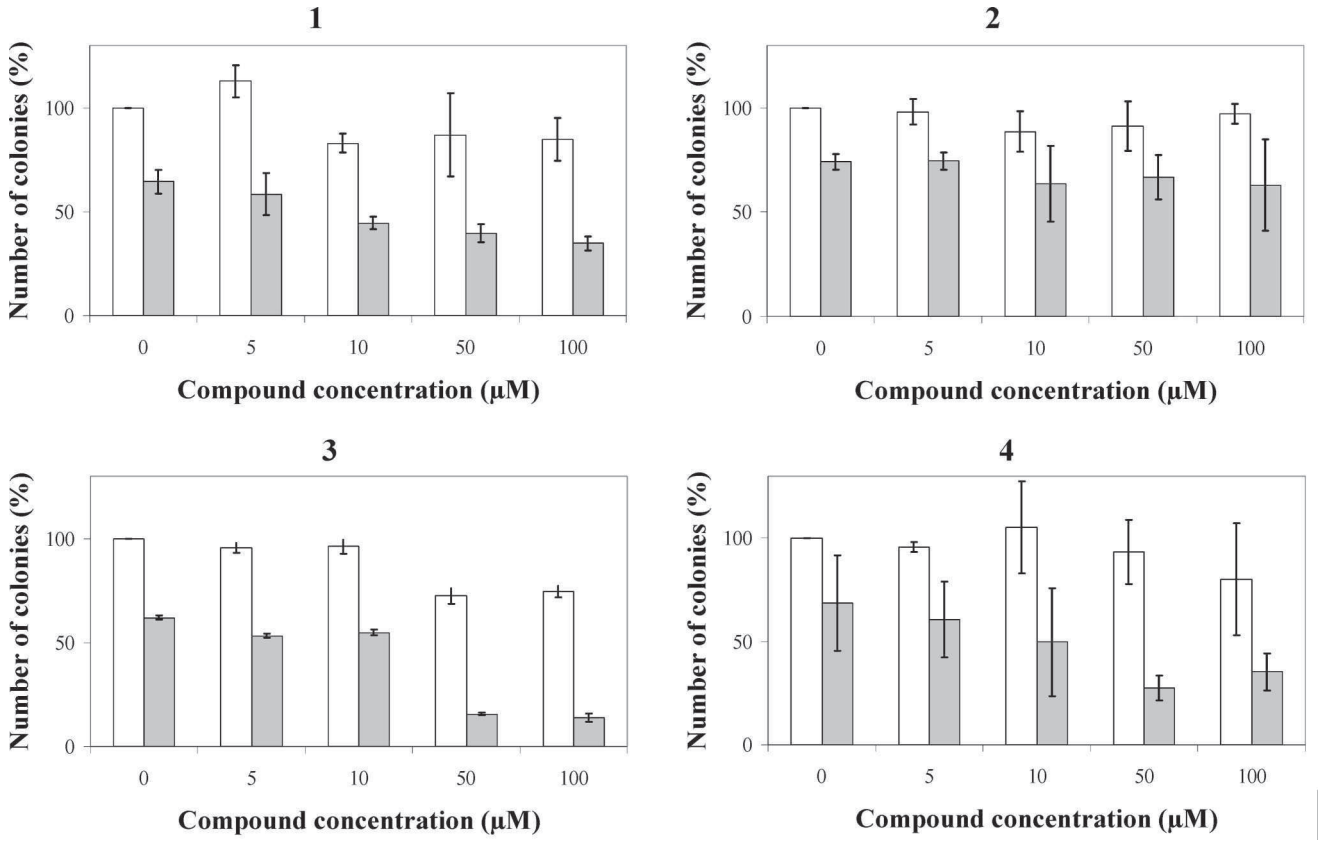


Figure 5. Effect of compounds in HTB-38 cells survival, relative to untreated cells. White bars show samples with no BCNU added in the presence of different concentrations of each of the four compounds, and gray bars show the same experiment in the presence of BCNU at 80 μM. BCNU alone reduce the cell number to an average of 68%; 10, 50, 10, and 100 μM are the concentrations required to produce 50% cell killing of compounds **1**, **2**, **3**, or **4**, respectively.

from ChemDiv. All of them were dissolved in DMSO (Sigma) at a final concentration of 1 mM and kept at −20 °C until used.

Virtual Screening (VS). All VS calculations have been performed within the VSDB platform (to be published, see Figure 1). For clarity, we briefly describe here the main steps comprising the protocol.

Protein Preparation. Since no substantial difference appear in the active site among available MGMT structures, we selected the A chain of 1t39[18] (PDB ID code) as receptor. The AMBER8 ff99 force field[51] was then used to assign atom types and charges for each atom in the protein. Hydrogen atoms were added assuming standard protonation states of titratable groups.

Binding Site Definition and Characterization. The binding site was built around the cocrystallized ligand (E1X) adding a 5.0 Å *cushion* to the maximum dimensions of the ligand. An equally spaced grid of 0.5 Å was built. In each grid point, the interaction energy of typical atom types (C, N, O, S, P, F, Cl, Br, I, and H) and e^- with all the atoms in the protein were calculated using a combination of 12–6 Lennard-Jones potential and a sigmoidal screening function for van der Waals and electrostatic interactions respectively, with CGRID program.[52] Employing our CDOCK docking program[52] (see below) we docked benzene, water, and methanol molecules generating intermolecular interaction energy maps. Benzene molecule was used here to locate favorable hydrophobic areas, water for hydrophilic sites, and methanol for hydrogen bonds; in a last step the generated energy maps were compressed using GAGA algorithm[42] in form of Gaussian functions trying to capture the most likely areas of interest for each kind of interaction (hydrophobic, hydrophilic, and hydrogen bond). The result of this calculation is the characterization of a sort of negative image of the interaction site. The putative active ligands in the library must conform to this approximate shape.

Chemical Library Preparation. Ligands for VS were obtained from the publicly available ZINC database in SMILES format.[53] Multiple protonation states and tautomeric forms are considered as implemented by default in ZINC database. The database was then processed within VSDB as follows: 2D to 3D conversion was carried out with CORINA,[54] up to 6 stereochemical centers were considered, ring conformations generated, hydrogen atoms added, and salt ions removed. Charges were assigned to ligand atoms with MOPAC (MOPAC[55] ESP with MINDO method[56] and radii assignment (AMBER-type[51])). Conformational analysis was carried out with ALFA.[57] ALFA allows the automatic assignment of atom types, detection of rotatable bonds, assignment of possible rotameric states, and generation of conformers. From ZINC we selected around 2.3 million fulfilling the Lipinsky rule of five with up to 7 rotatable bonds.

Filter 1. An initial filter was performed with the DOCK program[45] to discard those molecules that do not geometrically fit within the binding site. The spheres needed by DOCK were generated with our algorithm GAGA (see above). 3D molecules were scored with DOCK's contact scoring function. Finally, score values ($score_i$) are converted into Zscore using mean ($\overline{score}$) and standard deviation (σ)

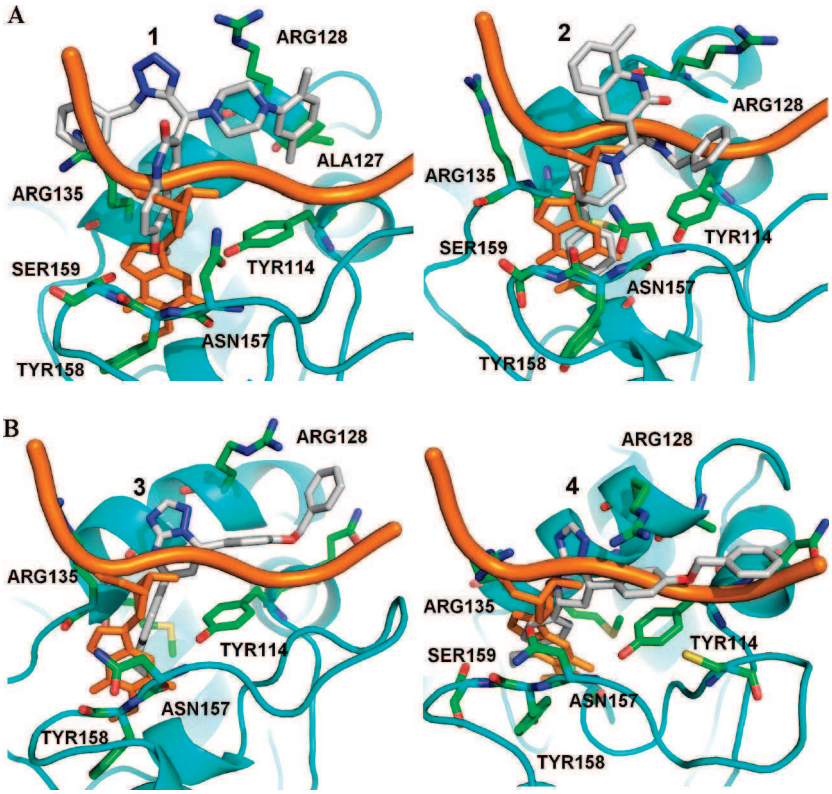


Figure 6. Average minimized structure of the compound-MGMT complexes after the molecular dynamics simulations. The structure of the flipped out nucleotide is also shown to highlight the structural similarity between the predicted complexes and the experimental structure. The color code is as follows: the protein is represented as cartoons in cyan; the flipped out nucleotide and part of the DNA backbone is colored in orange; side chains of main interacting residues are colored by atom type: C in green, N in blue, O in red, and S in yellow. The C atoms of compounds **1**−**4** are in gray, and hydrogen atoms are omitted for clarity. A and B correspond to families 1 and 2, respectively, as defined in the Results and Discussion section.

values ($Zscore_i = (score_i - \overline{score})/\sigma$). Only molecules with a Zscore beyond the cutoff value of 5 were used to select initial hits. From the initial set of 2.3 million molecules, only 1664 passed the Zscore filter.

Filter 2. This last number is an affordable amount of molecules to be studied with a more accurate docking algorithm as CDOCK. CDOCK exhaustively docks each molecule within the binding site using the interaction energy grids calculated with CGRID (see above). The centers of mass of the molecules are positioned on grid points equally spaced 1 Å where discrete rotations of 27° arc on each axel are performed. The "docking energy" for each pose (van der Waals and electrostatic) was then calculated using a trilinear interpolation method. CDOCK program has been proved to be accurate in reproducing native-like conformation starting both from X-ray structure[52] or building the structures from scratch.[58]

Rescoring. After docking with CDOCK, electrostatic interaction was corrected for desolvation of ligand and receptor by numerically solving Poisson equation using DelPhi.[59] Details of the calculations can be found elsewhere.[60] All molecules were finally ranked according to their corrected interaction energies, namely, van der Waals plus Coulombic term and desolvation values for receptor and ligand. Finally, the non electrostatic part of solvation was calculated assuming a linear relationship with the solvent accessible surface area. No correction was applied to account for conformational entropy.

Selection of Candidates. The best 17 molecules were selected upon analyzing the binding energy, the physicochemical properties of the molecule and visualizing the complexes with Pymol[61] purchased and tested experimentally.

Molecular Dynamics Simulations. With the 4 active molecules we carried out a 1 ns molecular dynamics simulation. All simulations were performed at a constant pressure and temperature (1 atm and 300 K) with an integration time step of 2 fs. SHAKE[62] was used to constrain all the bonds involving H atoms at their equilibrium distances. Periodic boundary conditions and the Particle Mesh Ewald methods were used to treat long-range electrostatic effects.[63] AMBER-99[51] and TIP3P[64] force-fields were used in all cases. All the trajectories were performed using the AMBER 8 computer program and associated modules.[65] The four starting models corresponded to the CDOCK predicted complexes. The 4 complexes were hydrated by using boxes containing explicit water molecules, optimized, heated (20 ps), and equilibrated (100 ps). After equilibration, MD trajectories were continued for 1 ns.

Effective binding free energies were qualitatively estimated using the MM-GBSA approach.[66] MM-GBSA method approaches free energy of binding as a sum of a molecular mechanics (MM) interaction term, a solvation contribution thorough a generalized Born (GB) model, and a surface area (SA) contribution to account for the non polar part of desolvation. These calculations were performed for each snapshot from the simulations using the appropriate module within AMBER 8 and averaged out.

Inhibition of MGMT Activity. *Protein Production and Purification.* In vitro assays were carried out using recombinant MGMT cloned in the pet-21a(+) (Novagen)

vector. The protein was expressed in the *E. coli* strain Rosetta and once the culture reached an OD_{600} value of 0.8 it was induced by adding 1 mM IPTG during 4 h at 30 °C. The pellet from a 3 L culture was disrupted by sonication and centrifuged. The supernatant was filtered, loaded into a HiTrap FF column (GE Healthcare) and eluted with an Imidazole (Fluka) gradient. Finally the protein was loaded into a Superdex 75 16/60 column (GE Healthcare) being the buffer 150 mM NaCl, 10 mM DTT (Sigma) and 0,1 mM EDTA. The protein was concentrated in this buffer and kept at −20 °C in presence of 40% glycerol. The same protocol has been used for the purification of the inactive mutant MGMT-C145S, cloned in the pet-28a(+) vector (Novagen) and expressed in the *E. coli* strain BL21.

Substrate Preparation. The ^{3}H-methylation of DNA has been described by Bodgen et al.[67] Briefly, calf thymus DNA (Sigma) was dissolved in 10 mM Sodium Cacodylate pH 7 buffer at 1 mg/mL stock solution. A 7 μL portion of MNU was added to 1 mL of the DNA stock solution and was then incubated at 37 °C during 2 h. The DNA was precipitated by adding sodium acetate, to a final concentration of 25 mM, and two volumes of cold ethanol 96%. After centrifugation, the DNA pellet was washed twice with cold ethanol, dried, and redissolved overnight at 4 °C in 0.15 N sodium chloride−0.015 N sodium citrate, pH 7.0. The DNA was reprecipitated, washed and dried as described above, redissolved in the reaction buffer (50 mM Tris pH 7, 8; 1 mM DTT; 5 mM EDTA) and stored at −20 °C until use.

AGT Activity Assay. The in vitro alkyltransferase activity assay has been previously described.[11,68] Purified protein was incubated with a defined concentration of compound in the reaction buffer at 37 °C. After 30 min [^{3}H]-methylated DNA was added, and the incubation was continued for an additional 90 min. The final volume was 200 μL, being the DNA concentration 100 times higher than the protein. The reaction was stopped by the addition of 400 μl of 13% trichloroacetic acid (TCA). DNA was then hydrolyzed by heating the sample at 95 °C for 30 min. The precipitated protein was washed twice with TCA 4% and redissolved in 0.2 M Tris pH 8. The activity corresponding to the [^{3}H]M-ethyl group transferred to the protein was analyzed by liquid scintillation counting using the Optiphase HiSafe 3 cocktail (Perkin-Elmer) and a Wallac 1414 liquid scintillation analyzer (GMI Inc.). Each compound concentration was assayed in quadruplicate and experiments repeated two times. Percent inhibition was calculated relative to untreated control samples. The IC_{50} values were determined graphically from plots of percent inhibition vs inhibitor concentration.

Cell Culture Cytotoxic Assay. The effect of compounds **1**−**4** on the sensitivity of HTB-38 cells to BCNU was determined using colony forming assays as has been described previously in MGMT inhibition studies.[49,50] HTB-38 cells were seeded at 15×10^3 cells per well density in 6-well, flat-bottomed plates (Falcon) and incubated in a humidified, 5% CO_2 incubator at 37 °C for 48 h. Compound solutions were diluted in the culture medium at final concentrations of 100, 50, 10 and 5 μM, and were immediately used to treat the cells. Cells were incubated with these compounds solutions for 6 h and then BCNU (or the equivalent volume of the vehicle) was added to a final concentration of 80 μM. After 2 h incubation the medium was replaced with fresh medium containing same compound concentration, and cells were left to grow for an additional 16 h. The cells were then replated at densities of 2000 cells per well in 6-well plates and grown for 12 days until discrete colonies were formed. Colonies were washed twice with PBS and stained with a 0.5% crystal violet−20% ethanol solution. Cells were rinsed with deionized water and air-dried. Finally crystal violet was solubilized with 10% acetic acid solution and the absorbance was measured in a Benchmark Microplate Reader (Bio-Rad). Samples were assayed in duplicate and experiments repeated three times. The percent of remaining cells was calculated relative to untreated control samples.

ACKNOWLEDGMENT

Work was partially supported by a grant from the "Comunidad de Madrid" (SBIO-0214-2006). Work at CNIO was supported by grants from "Fondo de Investigaciones Sanitarias" FIS (PI030989) and the Education and Science Ministry of Spain (GEN2003-20642-C09-02/NAC). Work at CBMSO was supported by grants from the Education and Science Ministry of Spain (BIO2005-0576 and GEN2003-206420-C09-08), Comunidad de Madrid (200520M157), and by an institutional grant from the Ramón Areces Foundation. Generous allocation of computer time at the Barcelona Supercomputing Center is gratefully acknowledged. C.F. was supported by Fondo de Investigaciones Sanitarias, Ministerio de Sanidad y Consumo (Spain). We thank Dr. Susana Gonzalez for providing HTB-38 cells.

REFERENCES AND NOTES

(1) Middleton, M. R.; Margison, G. P. Improvement of chemotherapy efficacy by inactivation of a DNA-repair pathway. *Lancet Oncol.* **2003**, *4*, 37–44.

(2) Friedberg, E. C.; Walker, G. C. *DNA Repair and Mutagenesis*, 1st ed.; American Society for Microbiology: Washington, DC, 1995; p 32–33.

(3) Bren, U.; Zupan, M.; Guengerich, F. P.; Mavri, J. Chemical reactivity as a tool to study carcinogenicity: reaction between chloroethylene oxide and guanine. *J. Org. Chem.* **2006**, *71*, 4078–4084.

(4) Bren, U.; Guengerich, F. P.; Mavri, J. Guanine alkylation by the potent carcinogen aflatoxin B1: quantum chemical calculations. *Chem. Res. Toxicol.* **2007**, *20*, 1134–1140.

(5) McMurry, T. B. MGMT inhibitors—The Trinity College-Paterson Institute experience, a chemist's perception. *DNA Repair (Amst)* **2007**, *6*, 1161–1169.

(6) Tong, W. P.; Kirk, M. C.; Ludlum, D. B. Formation of the cross-link 1-[N3-deoxycytidyl),2-[N1-deoxyguanosinyl]ethane in DNA treated with N,N′-bis(2-chloroethyl)-N-nitrosourea. *Cancer Res.* **1982**, *42*, 3102–3105.

(7) Yan, L.; Donze, J. R.; Liu, L. Inactivated MGMT by O6-benzylguanine is associated with prolonged G2/M arrest in cancer cells treated with BCNU. *Oncogene* **2005**, *24*, 2175–2183.

(8) Brent, T. P.; Houghton, P. J.; Houghton, J. A. O6-Alkylguanine-DNA alkyltransferase activity correlates with the therapeutic response of human rhabdomyosarcoma xenografts to 1-(2-chloroethyl)-3-(trans-4-methylcyclohexyl)-1-nitrosourea. *Proc. Natl. Acad. Sci.* **1985**, *82*, 2985–2989.

(9) Tagliabue, G.; Citti, L.; Massazza, G.; Damia, G.; Giavazzi, R.; D'Incalci, M. Tumour levels of O6-alkylguanine-DNA-alkyltransferase and sensitivity to BCNU of human xenografts. *Anticancer Res.* **1992**, *12*, 2123–2125.

(10) Pepponi, R.; Marra, G.; Fuggetta, M. P.; Falcinelli, S.; Pagani, E.; Bonmassar, E.; Jiricny, J.; D'Atri, S. The effect of O6-alkylguanine-DNA alkyltransferase and mismatch repair activities on the sensitivity of human melanoma cells to Temozolomide, 1,3-bis(2-chloroethyl)1-nitrosourea, and cisplatin. *J. Pharmacol. Exp. Ther.* **2003**, *304*, 661–668.

(11) Mattern, J.; Eichhorn, U.; Kaina, B.; Volm, M. O6-methylguanine-DNA methyltransferase activity and sensitivity to cyclophosphamide and cisplatin in human lung tumor xenografts. *Int. J. Cancer* **1998**, *77*, 919–922.

(12) Margison, G. P.; Povey, A. C.; Kaina, B.; Santibanez Koref, M. F. Variability and regulation of O6-alkylguanine-DNA alkyltransferase. *Carcinogenesis* **2003**, *24*, 625–635.
(13) Gerson, S. L. MGMT: its role in cancer aetiology and cancer therapeutics. *Nat. Rev. Cancer* **2004**, *4*, 296–307.
(14) Gerson, S. L. Clinical relevance of MGMT in the treatment of cancer. *J. Clin. Oncol* **2002**, *20*, 2388–2399.
(15) Hegi, M. E.; Diserens, A. C.; Gorlia, T.; Hamou, M. F.; de Tribolet, N.; Weller, M.; Kros, J. M.; Hainfellner, J. A.; Mason, W.; Mariani, L.; Bromberg, J. E.; Hau, P.; Mirimanoff, R. O.; Cairncross, J. G.; Janzer, R. C.; Stupp, R. MGMT gene silencing and benefit from Temozolomide in glioblastoma. *N. Engl. J. Med.* **2005**, *352*, 997–1003.
(16) Esteller, M.; Garcia-Foncillas, J.; Andion, E.; Goodman, S. N.; Hidalgo, O. F.; Vanaclocha, V.; Baylin, S. B.; Herman, J. G. Inactivation of the DNA-repair gene MGMT and the clinical response of gliomas to alkylating agents. *N. Engl. J. Med.* **2000**, *343*, 1350–1354.
(17) Daniels, D. S.; Woo, T. T.; Luu, K. X.; Noll, D. M.; Clarke, N. D.; Pegg, A. E.; Tainer, J. A. DNA binding and nucleotide flipping by the human DNA repair protein AGT. *Nat. Struct. Mol. Biol.* **2004**, *11*, 714–720.
(18) Daniels, D. S.; Mol, C. D.; Arvai, A. S.; Kanugula, S.; Pegg, A. E.; Tainer, J. A. Active and alkylated human AGT structures: a novel zinc site, inhibitor and extrahelical base binding. *EMBO J.* **2000**, *19*, 1719–1730.
(19) Kanugula, S.; Goodtzova, K.; Edara, S.; Pegg, A. E. Alteration of arginine-128 to alanine abolishes the ability of human O6-alkylguanine-DNA alkyltransferase to repair methylated DNA but has no effect on its reaction with O6-benzylguanine. *Biochemistry* **1995**, *34*, 7113–7119.
(20) Mishina, Y.; Duguid, E. M.; He, C. Direct reversal of DNA alkylation damage. *Chem. Rev.* **2006**, *106*, 215–232.
(21) Rasimas, J. J.; Dalessio, P. A.; Ropson, I. J.; Pegg, A. E.; Fried, M. G. Active-site alkylation destabilizes human O6-alkylguanine DNA alkyltransferase. *Protein Sci.* **2004**, *13*, 301–305.
(22) Srivenugopal, K. S.; Yuan, X. H.; Friedman, H. S.; Ali-Osman F. Ubiquitination-dependent proteolysis of O6-methylguanine-DNA methyltransferase in human and murine tumor cells following inactivation with O6-benzylguanine or 1,3-bis(2-chloroethyl)-1-nitrosourea. *Biochemistry* **1996**, *35*, 1328–1334.
(23) Xu-Welliver, M.; Pegg, A. E. Degradation of the alkylated form of the DNA repair protein, O(6)-alkylguanine-DNA alkyltransferase. *Carcinogenesis* **2002**, *23*, 823–830.
(24) Dolan, M. E.; Moschel, R. C.; Pegg, A. E. Depletion of mammalian O6-alkylguanine-DNA alkyltransferase activity by O6-benzylguanine provides a means to evaluate the role of this protein in protection against carcinogenic and therapeutic alkylating agents. *Proc. Natl. Acad. Sci.* **1990**, *87*, 5368–5372.
(25) Pegg, A. E.; Boosalis, M.; Samson, L.; Moschel, R. C.; Byers, T. L.; Swenn, K.; Dolan, M. E. Mechanism of inactivation of human O6-alkylguanine-DNA alkyltransferase by O6-benzylguanine. *Biochemistry* **1993**, *32*, 11998–12006.
(26) Xu-Welliver, M.; Kanugula, S.; Pegg, A. E. Isolation of human O6-alkylguanine-DNA alkyltransferase mutants highly resistant to inactivation by O6-benzylguanine. *Cancer Res.* **1998**, *58*, 1936–1945.
(27) Pegg, A. E.; Swenn, K.; Chae, M. Y.; Dolan, M. E.; Moschel, R. C. Increased killing of prostate, breast, colon, and lung tumor cells by the combination of inactivators of O6-alkylguanine-DNA alkyltransferase and N,N′-bis(2-chloroethyl)-N-nitrosourea. *Biochem. Pharmacol.* **1995**, *50*, 1141–1148.
(28) Kreklau, E. L.; Kurpad, C.; Williams, D. A.; Erickson, L. C. Prolonged inhibition of O(6)-methylguanine DNA methyltransferase in human tumor cells by O(6)-benzylguanine in vitro and in vivo. *J. Pharmacol. Exp. Ther.* **1999**, *291*, 1269–1275.
(29) Rabik, C. A.; Njoku, M. C.; Dolan, M. E. Inactivation of O6-alkylguanine DNA alkyltransferase as a means to enhance chemotherapy. *Cancer Treat. Rev.* **2006**, *32*, 261–276.
(30) Dolan, M. E.; Chae, M. Y.; Pegg, A. E.; Mullen, J. H.; Friedman, H. S.; Moschel, R. C. Metabolism of O6-benzylguanine, an inactivator of O6-alkylguanine-DNA alkyltransferase. *Cancer Res.* **1994**, *54*, 5123–5130.
(31) Schilsky, R. L.; Dolan, M. E.; Bertucci, D.; Ewesuedo, R. B.; Vogelzang, N. J.; Mani, S.; Wilson, L. R.; Ratain, M. J. Phase I clinical and pharmacological study of O6-benzylguanine followed by carmustine in patients with advanced cancer. *Clin. Cancer Res.* **2000**, *6*, 3025–3031.
(32) Gajewski, T. F.; Sosman, J.; Gerson, S. L.; Liu, L.; Dolan, E.; Lin, S.; Vokes, E. E. Phase II trial of the O6-alkylguanine DNA alkyltransferase inhibitor O6-benzylguanine and 1,3-bis(2-chloroethyl)-1-nitrosourea in advanced melanoma. *Clin. Cancer Res.* **2005**, *11*, 7861–7865.
(33) Bacolod, M. D.; Johnson, S. P.; Pegg, A. E.; Dolan, M. E.; Moschel, R. C.; Bullock, N. S.; Fang, Q.; Colvin, O. M.; Modrich, P.; Bigner, D. D.; Friedman, H. S. Brain tumor cell lines resistant to O6-benzylguanine/1,3-bis(2-chloroethyl)-1-nitrosourea chemotherapy have O6-alkylguanine-DNA alkyltransferase mutations. *Mol. Cancer Ther.* **2004**, *3*, 1127–1135.
(34) Liu, L.; Schwartz, S.; Davis, B. M.; Gerson, S. L. Chemotherapy-induced O(6)-benzylguanine-resistant alkyltransferase mutations in mismatch-deficient colon cancer. *Cancer Res.* **2002**, *62*, 3070–3076.
(35) Phillips, W. P., Jr.; Gerson, S. L. Acquired resistance to O6-benzylguanine plus chloroethylnitrosoureas in human breast cancer. *Cancer Chemother. Pharmacol.* **1999**, *44*, 319–326.
(36) Ranson, M.; Middleton, M. R.; Bridgewater, J.; Lee, S. M.; Dawson, M.; Jowle, D.; Halbert, G.; Waller, S.; McGrath, H.; Gumbrell, L.; McElhinney, R. S.; Donnelly, D.; McMurry, T. B.; Margison, G. P. Lomeguatrib, a potent inhibitor of O6-alkylguanine-DNA-alkyltransferase: phase I safety, pharmacodynamic, and pharmacokinetic trial and evaluation in combination with Temozolomide in patients with advanced solid tumors. *Clin. Cancer. Res.* **2006**, *12*, 1577–1584.
(37) Stoermer, M. J. Current status of virtual screening as analysed by target class. *Med. Chem.* **2006**, *2*, 89–112.
(38) Klebe, G. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discov. Today* **2006**, *11*, 580–594.
(39) Warren, G. L.; Andrews, C. W.; Capelli, A. M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A critical assessment of docking programs and scoring functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.
(40) Michel, J.; Verdonk, M. L.; Essex, J. W. Protein-ligand binding affinity predictions by implicit solvent simulations: a tool for lead optimization. *J. Med. Chem.* **2006**, *49*, 7427–7439.
(41) Weis, A.; Katebzadeh, K.; Soderhjelm, P.; Nilsson, I.; Ryde, U. Ligand affinities predicted with the MM/PBSA method: dependence on the simulation method and the force field. *J. Med. Chem.* **2006**, *49*, 6596–6606.
(42) Kollman, P. A.; Massova, I.; Reyes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Donini, O.; Cieplak, P.; Srinivasan, J.; Case, D. A.; Cheatham, T. E., 3rd. Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.* **2000**, *33*, 889–897.
(43) Massova, I.; Kollman, P. A. Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspect. Drug Discov. Des.* **2000**, *200*, 113–135.
(44) Aqvist, J.; Medina, C.; Samuelsson, J. E. A new method for predicting binding affinity in computer-aided drug design. *Protein Eng.* **1994**, *7*, 385–391.
(45) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **1982**, *161*, 269–288.
(46) Irwin, J. J.; Shoichet, B. K. ZINC—a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(47) McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* **2002**, *45*, 1712–1722.
(48) McGovern, S. L.; Helfand, B. T.; Feng, B.; Shoichet, B. K. A specific mechanism of nonspecific inhibition. *J. Med. Chem.* **2003**, *46*, 4265–4272.
(49) Pegg, A. E.; Goodtzova, K.; Loktionova, N. A.; Kanugula, S.; Pauly, G. T.; Moschel, R. C. Inactivation of human O(6)-alkylguanine-DNA alkyltransferase by modified oligodeoxyribonucleotides containing O(6)-benzylguanine. *J. Pharmacol. Exp. Ther.* **2001**, *296*, 958–965.
(50) Nelson, M. E.; Loktionova, N. A.; Pegg, A. E.; Moschel, R. C. 2-amino-O4-benzylpteridine derivatives: potent inactivators of O6-alkylguanine-DNA alkyltransferase. *J. Med. Chem.* **2004**, *47*, 3887–3891.
(51) Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179–5197.
(52) Perez, C.; Ortiz, A. R. Evaluation of docking functions for protein-ligand docking. *J. Med. Chem.* **2001**, *44*, 3768–3785.
(53) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
(54) *Corina Molecular Networks*; GmbH Computerchemie: Germany, 2000.
(55) Stewart, J. J. MOPAC: a semiempirical molecular orbital program. *J. Comput. Aided Mol. Des.* **1990**, *4*, 1–105.
(56) Dewar, M. J. S.; Thiel, W. MIND0/3 Study of the Addition of Singlet Oxygen (1deltagO2) to 1,3-Butadiene. *J. Am. Chem. Soc.* **1977**, *99*, 2338–2339.
(57) Gil-Redondo, R. Implementación de una plataforma para el cribado virtual de quimiotecas. Master Thesis; Universidad Nacional de Educación a Distancia, Madrid, Spain, 2006.

(58) Murcia, M.; Morreale, A.; Ortiz, A. R. Comparative binding energy analysis considering multiple receptors: a step toward 3D-QSAR models for multiple targets. *J. Med. Chem.* **2006**, *49*, 6241–6253.
(59) Rocchia, W.; Sridharan, S.; Nicholls, A.; Alexov, E.; Chiabrera, A.; Honig, B. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J. Comput. Chem.* **2002**, *23*, 128–137.
(60) Morreale, A.; Gil-Redondo, R.; Ortiz, A. R. A new implicit solvent model for protein-ligand docking. *Proteins* **2007**, *67*, 606–616.
(61) DeLano, W. L. *The PyMOL Molecular Graphics System* DeLano Scientific: Palo Alto, CA, 2002.
(62) Ryckaert, J.; Ciccotti, G.; Berendsen, H. Numerical Integration of the Cartesian Equations of Motion of a System with Constraints: Molecular Dynamics of n-Alkanes. *J. Comp. Phys.* **1977**, *23*, 327–341.
(63) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.
(64) Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935.
(65) McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* **2002**, *45*, 1712–1722.
(66) Still, W.; Tempczyk, A.; Hawley, R.; Hendrickson, T. Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **1990**, *112*, 6127–6129.
(67) Bogden, J. M.; Eastman, A.; Bresnick, E. A system in mouse liver for the repair of O6-methylguanine lesions in methylated DNA. *Nucleic Acids Res.* **1981**, *9*, 3089–3103.
(68) Reinhard, J.; Hull, W. E.; von der Lieth, C. W.; Eichhorn, U.; Kliem, H. C.; Kaina, B.; Wiessler, M. Monosaccharide-linked inhibitors of O(6)-methylguanine-DNA methyltransferase (MGMT): synthesis, molecular modeling, and structure-activity relationships. *J. Med. Chem.* **2001**, *44*, 4050–4061.

CI700447R

J Comput Aided Mol Des (2009) 23:171–184
DOI 10.1007/s10822-008-9249-9

VSDMIP: virtual screening data management on an integrated platform

Rubén Gil-Redondo · Jorge Estrada · Antonio Morreale · Fernando Herranz · Javier Sancho · Ángel R. Ortiz

Received: 29 May 2008 / Accepted: 28 September 2008 / Published online: 22 October 2008

Abstract A novel software (VSDMIP) for the virtual screening (VS) of chemical libraries integrated within a MySQL relational database is presented. Two main features make VSDMIP clearly distinguishable from other existing computational tools: (i) its database, which stores not only ligand information but also the results from every step in the VS process, and (ii) its modular and pluggable architecture, which allows customization of the VS stages (such as the programs used for conformer generation or docking), through the definition of a detailed workflow employing user-configurable XML files. VSDMIP, therefore, facilitates the storage and retrieval of VS results, easily adapts to the specific requirements of each method and tool used in the experiments, and allows the comparison of different VS methodologies. To validate the usefulness of VSDMIP as an automated tool for carrying out VS several experiments were run on six protein targets (acetylcholinesterase, cyclin-dependent kinase 2, coagulation factor Xa, estrogen receptor alpha, p38 MAP kinase, and neuraminidase) using nine binary (actives/inactive) test sets. The performance of several VS configurations was evaluated by means of enrichment factors and receiver operating characteristic plots.

Ángel R. Ortiz deceased on May 5, 2008.

R. Gil-Redondo · A. Morreale (✉) · F. Herranz · Á. R. Ortiz
Unidad De Bioinformática, Centro De Biología Molecular Severo Ochoa (CSIC-UAM), C/Nicolás Cabrera 1, Campus De Cantoblanco, Madrid 28049, Spain
e-mail: amorreale@cbm.uam.es

J. Estrada · J. Sancho
Departamento de Bioquímica y Biología Molecular y Celular, Facultad de Ciencias and BIFI –Instituto de Biocomputación y Física de Sistemas Complejos, c/Pedro Cerbuna 12, Universidad de Zaragoza, Zaragoza 50009, Spain

Keywords Docking · Virtual screening · Drug design · Database · Platform

Abbreviations

AChE	Acetylcholinesterase
fXa	Coagulation factor Xa
CDK2	Cyclic dependant kinase 2
Era	Estrogen receptor a
p38MAP	MAP Kinase P38
VS	Virtual Screening
EF	Enrichment Factor
ROC	Receiver Operating Characteristic

Introduction

Launching a new molecule to the market requires tremendous effort in research, development and money investment. Recent studies estimate in 15 years and around 800 million dollars the average time and cost per approved molecule [1]. After more than 30 years of using combinatorial chemistry and high-throughput screening (the two techniques that researchers thought would be the solution to the drug discovery bottleneck), the ratio between the number of new drugs obtained and the funds invested in their generation is well below the initial expectations [2, 3]. In some sense this has fuelled the development of theoretical techniques that attempt to accelerate the initial steps in the drug design cycle. Theoretical methods, if correctly derived, allow pinpointing the more promising candidates (hits) out of pools of thousands or even millions of molecules (chemical libraries). This reduced set can then be subjected to experimental analysis, and any promising compound can be subsequently optimized to attain the desired pharmacological profile and become a lead.

From a theoretical perspective different scenarios can be envisaged depending on the structural data available [4, 5]. The most favourable one is when the structures of both the target and the ligand(s) are known. In this case docking and VS techniques are the methods of choice [6]. More elaborate approaches based on molecular dynamics [7] or free energy perturbation coupled to thermodynamic integration can also be used [8]. The problem here is that the amount of time required to perform the calculations becomes prohibitive when a large collection of molecules is used.

The goal of docking is to identify, among the large number of possible orientations of a ligand within the binding site of the target, the one closest to the experimental structure of the complex [4]. This is done by using a mathematical function that accounts for the goodness of the coupling between ligand and target. Consequently, two key elements of the docking problem are: (a) a good sampling method, and (b) an accurate scoring function. VS is the extrapolation of docking to the case in which a large database of molecules is going to be processed [9]. Here the primary goal is somewhat different from that in docking because promising hit candidates are picked out from a database of mostly non-binders and no attempt is made to correctly classify all the molecules in the library or to identify all the actives.

Generating the experimental pose for a ligand is feasible with today's sampling techniques. However, positioning this pose at the top of a prioritized list of candidates is more problematic [10]. The main reason for this limitation is that the physical effects that describe the binding process are incompletely represented in the scoring function despite the fact that the underlying principles are reasonably well understood. In particular, the influence of the solvent, entropic effects and target flexibility are difficult to implement without compromising speed. Although many advances are being made to overcome these deficiencies, we are still at an early developmental stage.

Besides the problems outlined above, filtering millions of molecules using molecular descriptors and properties, docking them with one or more programs endowed with different accuracies, dealing with complex effects such as desolvation and/or protein flexibility at post-docking stages, etc. generates huge amounts of data that cannot be easily stored or utilized. It would then seem that the introduction of relational databases and potent database managing tools could be of aid in this regard.

Here we propose a flexible, fully automated computational platform to perform VS experiments that combines all the necessary steps to generate a short list of candidates starting from a database of 2D molecular structures. VSDMIP is intended to fill an existing gap in the docking and VS fields in relation to the storage and handling of the data. We are committed to making this platform available to interested parties so that the scientific community at large can benefit from it.

Methods

VSDMIP software and architecture

The VSDMIP architecture (Fig. 1) consists of (1) a database (in a multithreaded multi-user Structured Query Language [MySQL] DataBase Manager System), (2) a library of service interfaces and plugins, and (3) a set of workflows and its implementing commands. All small molecule data and VS results are stored in the VSDMIP database. The user controls the platform through different command line utilities and configures it using XML files. VSDMIP currently runs on Linux/x86 platforms (in our

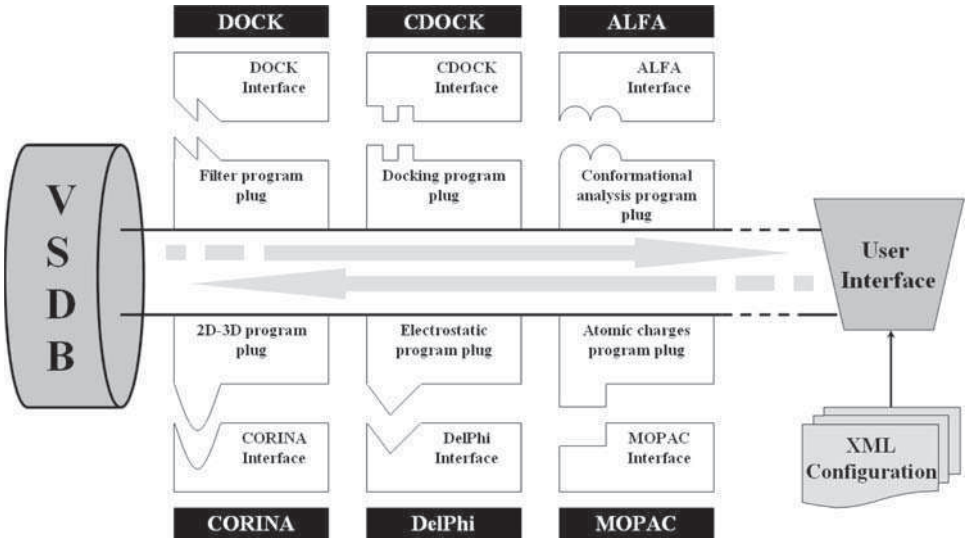


Fig. 1 Pictorial representation of the VSDMIP's architecture. VSDB refers to the Virtual Screening Data Base where data is stored

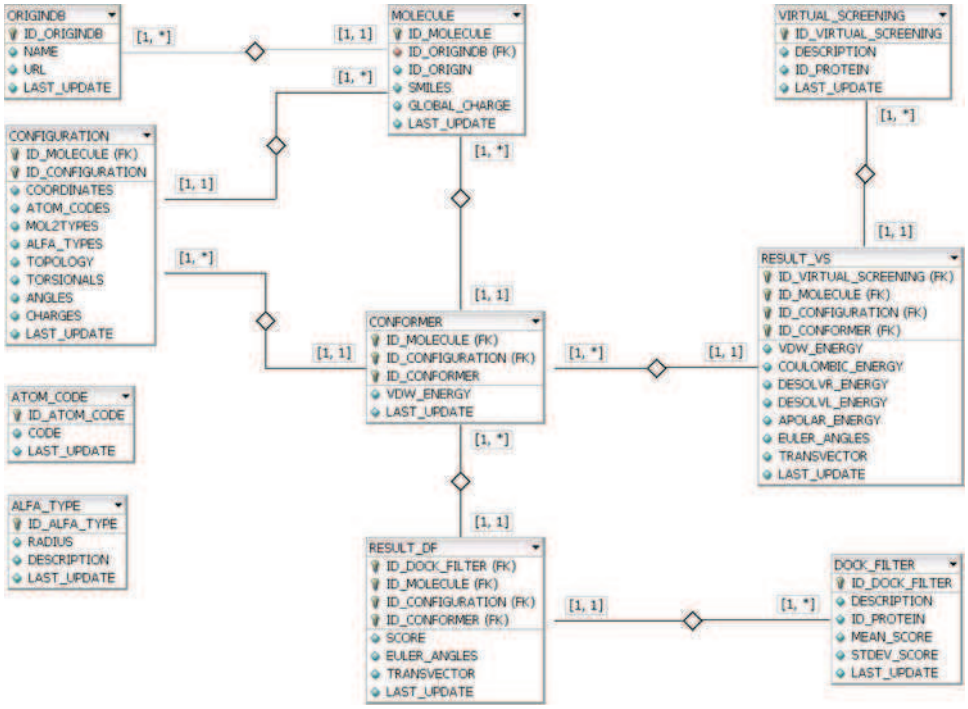


Fig. 2 Entity-Relationship schema of the database used by VSDMIP

case, a cluster of ten 2.4 and thirty-two 3.0 GHz Xeon Biprocessor CPUs with 2 GB RAM each).

Relational database

The database (Fig. 2) contains tables for the compound libraries (ORIGINDB, MOLECULE, CONFIGURATION, CONFORMER), results from filtering experiments (DOCK_FILTER, RESULT_DF), and results from VS experiments (VIRTUAL_SCREENING, RESULT_VS). Each compound library has its origin entered in an ORIGINDB entry. Each molecule has a MOLECULE entry with its global charge and a SMILES string representing its topology. The different stereoisomers of each molecule in combination with the different ring conformations are stored, together with a 3D structure, in a CONFIGURATION entry; discrete conformational changes on each of these entries are stored in the CONFORMER table, with a score resembling its internal energy (VDW_ENERGY). Filtering experiments are identified by a DOCK_FILTER entry. The best docking pose for each of the selected conformers included in the experiment is stored in a RESULT_DF entry, together with its score. These poses are stored as a translation and rotation of the conformer with respect to the original ligand coordinates stored in the database. More accurate docking experiments are identified by a VIRTUAL_SCREENING entry, and each individual conformer solution has an entry in RESULT_VS, with the same translation and rotation information as explained for RESULT_DF. Besides, the docking score can be stored with the details of its interaction energy terms.

VSDMIP software library

This C/C++ library manages the database, interconnects the different applications used, and offers biochemical and geometrical utilities. For each external application it is possible to add functionalities to the platform by creating an interface class and a storage class, under a common framework. The interface class provides methods for managing the application configuration attributes, preparing the execution of the application, performing the execution, managing errors and storing the results in the database. Six service interfaces currently exist: 3D structure generation, conformational analysis, atomic charge calculation, filtering, VS, and electrostatic calculations. Several plugins have been developed to interface with CORINA 3.0.5 [11] (3D generation), ALFA [12] (conformational analysis), MOPAC 7 [13] (atomic charges), DOCK 3.5 [14] and FRED 2.2 [15] (filter interface), CDOCK [16] and Autodock 3.0.5 [17] (VS interface), and for DelPhi 4 [18] and ISM [19] (electrostatic calculations).

Process workflows

The currently included commands allow inserting molecules into the database, docking a set of molecules

within a protein binding site (for filtering or VS), rescoring the results of a docking experiment, and retrieving the results from a filtering or docking step. Each command uses one or more of the services plugged into VSDMIP and task execution is allowed to take place in a computer cluster.

Validation tests

Eleven different VS protocols (VSP) were designed and applied to nine different ligand datasets involving six protein targets. The protocols differ in several aspects: (i) the use of a filter and the number of molecules and conformers allowed to pass the filter, (ii) the docking engine used, and (iii) the scoring function(s) employed for ranking (Table 1). The active compounds for the validation tests (Table 2) were obtained from the binary (active/inactive) datasets available from CHEMINFORMATICS.ORG. Inactive compounds were randomly selected from a 9862 subset of the Maybridge Hit-Finder collection. Most of these molecules follow Lipinski's rule of 5 [20]. Table 3 shows the general properties of the database. The results were evaluated using receiver operating characteristic (ROC) plots [21], which represent the sensitivity (y-axis, true positives rate, see Eq. 1) versus (1—specificity) (x-axis, false positives rate, see Eq. 2), as well as areas under the ROC curves (AUC), enrichment factors (EF), and computing time.

Table 2 Relevant information for each of the datasets used in the virtual screening experiments

VS experiment	# of actives	# of inactives	Total
fXa (Fontaine)	432	500	932
fXa (Jacobsson)	127	500	627
fXa (Jorissen-Gilson)	50	500	550
AChE (Jacobsson)	54	1000	1054
CDK2 (Jorissen-Gilson)	50	1000	1050
ERa (Jacobsson)	142	1000	1142
ERa (Stahl)	50	1000	1050
Neuraminidase (Stahl)	17	1000	1017
p38MAP (Stahl)	22	1000	1022

$$\text{Sensitivity} = \frac{\text{TruePositives}}{\text{True Positives} + \text{False Negatives}} \quad (1)$$

$$\text{Specificity} = \frac{\text{TrueNegatives}}{\text{True Negatives} + \text{False Positives}} \quad (2)$$

EF represents the ratio of active compounds detected in a fixed percentage of the scanned database to the total number of actives in the whole database (Eq. 3),

Table 1 Virtual screening protocols used in the validation of VSDMIP

Virtual screening tests		
VSP id.	Description	Description
1	DOCK_100_1_CS	DOCK is performed on the 100 lower-energy conformers per molecule. The best conformer is selected using contact scoring.
2	DOCK_100_1_FF	DOCK is performed on the 100 lower-energy conformers per molecule. The best conformer per molecule is selected using force-field scoring.
3	DOCK_100_1_CS-XS	The same as in #1 but a final re-scoring is performed with XSCORE.
4	CDOCK_50_VDW_COUL	CDOCK is performed on the 50 lower-energy conformers per molecule. The best conformer per molecule is selected using the sum of van der Waals and coulombic interaction energies.
5	CDOCK_50_VDW	The same as in #1 but the final rank is done only with van der Waals interaction energies
6	CDOCK_50_ALL500	#4 is followed, and then for the 500 top-ranking molecules the solvation correction term is calculated using DelPhi. The final score is obtained by summing up the van der Waals and desolvation energies.
7	DOCK_100_1_CS + CDOCK_ZS3.0	#1 is followed, and then those compounds with ZScore $\geq$ 3.0 are submitted to CDOCK.
8	DOCK_100_10_CS + CDOCK_ZS3.0	The same as in #7 but now the best 10 conformer for each molecule are passed on to CDOCK.
9	DOCK_100_1_CS + CDOCK_ZS1.5	The same as in #7 but with ZScore $\geq$ 1.5.
10	DOCK_100_10_CS + CDOCK_ZS1.5	The same as in #8 but with ZScore $\geq$ 1.5.
11	DOCK_100_1_CS + CDOCK_ZS3.0_ALL500	#7 is followed, and then solvation correction using DelPhi is calculated for the first 500 molecules. The final score is obtained by summing up the van der Waals and desolvation energies.

Table 3 Mean and standard deviation of Lipinsky's properties for the Maybridge database ligands extended with the number of rotatable bonds

	MW[a]	HBA[b]	HBD[c]	RB[d]	logP[e]
Mean	311	4.82	1.12	4.58	2.65
Standard deviation	75.7	1.87	1.06	2.39	1.65

As calculated with OpenEye's Filter 2.0. The rules are MOL_WT [0, 500], ROT_BONDS [0, 20], LIPINSKI_DONORS [0, 5], LIPINSKI_ACCEPTORS [0, 10] and XLOGP [−5.0, 5.0]. 743 molecules failed Lipinski's test

[a] Molecular weight (in Da); [b] Number of hydrogen bond acceptors; [c] Number of hydrogen bond donors; [d] Number of rotatable bonds; [e] log of octanol/water partition coefficient

$$EF = \frac{\{NAc_{subset}/NT_{subset}\}}{\{NAc_{total}/NT\}} \quad (3)$$

where subset is the fixed percentage of the database, NAc_{subset} is the number of active molecules found in the subset, NT_{subset} is the total number of molecules in the subset, NAc_{total} is the total number of active molecules in the entire database, and NT is the total number of molecules in the database. For comparative purposes we have computed the best EF found (EF_{best}), the maximum EF (EF_{max}) possible for each experiment, and the percentage of the database at which EF_{best} is obtained. To calculate EF_{best}, only subsets with an NT_{subset} multiple of b are considered, where b is max{10, 0.01 × NT}. This avoids artefacts in EF_{best} at the start of database scanning.

Setup of small molecule databases

All molecules were first converted to isomeric Simplified Molecular Input Line Entry Specification (SMILES) [22] format, using the OpenEye OEChem 1.4 library. A 10,000 subset of the Maybridge HitFinder database was obtained in 2D Structure Data Format (SDF) [23]. The stereo information was not considered, and only those molecules bearing 3 or less stereogenic centres were selected. An additional set of 12 molecules had improbable connectivities and were also discarded by the VSDMIP insertion application, leaving a total of 9,862 molecules. The active datasets were prepared from SDF or SMILES input files. Duplicates (most likely arising from different stereoisomers devoid of the stereo information) were removed. When known, stereo and protonation information was preserved. Otherwise, protonation was assigned following OpenEye's Filter 2.0 rules for pH 7.4. Changes in bond order and number of hydrogens were done in some molecules to obtain standard valences. Using the final SMILES strings, the totally automatic process, ***insertVSDB***, inserted the molecules into the database after carrying out the following steps: (i) conversion from SMILES to 3D MOL2 using CORINA: up to 6 stereogenic centres were considered, ring conformations were generated, hydrogen atoms were added, and salt ions were removed; (ii) atomic charge calculations with MOPAC: single point calculations were performed with the MNDO semiempirical method [24] obtaining atom centred charges via electrostatic potential fitting techniques on each single structure provided from CORINA (one or more for each SMILE string depending on the number of stereogenic centres and ring conformations); (iii) atom type assignment and conformational analysis using ALFA.

Protein set up

All protein structures were obtained from the PDB and correspond to X-ray experiments with resolutions of 2.6 Å or better. The structures were chosen based on previous published VS and docking experiments: 1f0r (chain A) for coagulation factor Xa (fXa) [25, 26]; 1eve for acetylcholinesterase (AChE) [27, 28], 1e1x for cyclin-dependent kinase 2 (CDK2) [29, 30], 3ert for estrogen receptor alpha (Era) [31, 32], 1nsc (chain B) for neuraminidase [33, 34], and 1p38 for p38 MAP kinase [35, 36]. All HETATM records, including water molecules and ions, were removed from the PDB files. Side chains with missing atoms were rebuilt using SCWRL3 software[37]. In 1e1x, MODELLER v6.2 [38] was employed to reconstruct a missing loop. Then, the AMBER 8 [39] ff99 force field [40] was used to assign atom types and partial charges to each atom in the proteins, and hydrogen atoms were added assuming standard protonation states of titratable groups.

The H++ web server [41] was employed to study and assign protonation states for key interacting residues in the binding site (see below). The Poisson-Boltzmann (PB) method [42, 43] was used, at pH 7.4, 0.15 M salt concentration, and using internal and external dielectric constants of 4 and 80, respectively. Based on the information provided by H++, the following residues were protonated: HIS57 and ASP189 in 1f0r; HIS440, GLU278, and GLU443 in 1eve; HIS125 in CDK2; and ASH323 in 1nsc. The modified proteins were subjected to 10,000 steps of energy minimization (500 initial steps of steepest descent, followed by conjugate gradient), in vacuum, and only the hydrogen atoms were allowed to move. A further 10,000 steps of energy minimization were done with a Generalized-Born (GB) implicit solvent model [44–46] during which all atoms were allowed to move but heavy atoms were positionally restrained with a harmonic potential (100 kcal/mol$Å^2$). To check for consistency, the optimized structures were submitted again to the H++ web server. No significant changes were observed. The final

minimized structures differed less than 0.05 Å (for Cα atoms) compared to the initial X-ray structure.

Binding site definition and characterization

For each protein structure, the initial binding site was defined as the space delimited by the axis-parallel box containing the co-crystallized ligand, augmented by 5 Å in each axis direction. Structure 1p38 was structurally aligned to PDB 1ywr [47], and the co-crystallized ligand found in 1ywr was used to define the binding site in the p38MAP studies. Protein interaction grids covering the binding site (1.0 Å spacing in all directions) were calculated for atom probes C, N, O, S, P, H, F, Cl, Br, and I. Each grid point represents the interaction between the protein and the probe atom as the sum of a van der Waals Lennard-Jones 12–6 potential and an electrostatic term modelled with a sigmoidal dielectric screening function [48]. For each binding site, a set of about 20 interaction points were defined to guide docking studies. These points were selected from the GAGA [49] centers of the gaussian sphere functions that best captured the interaction maps between the protein and benzene, water, and methanol molecules, as calculated by docking experiments using CDOCK. The set of points thus selected represent hydrophobic, hydrophilic and hydrogen bonding interactions. For docking using DOCK, the binding site was defined as the space delimited by the axis-parallel box containing the selected interaction points, augmented by 7.5 Å in each axis direction. The interaction grids used by DOCK and covering the binding site had a spacing of 0.3 Å. DOCK grids (see DOCK documentation) represent electrostatic, van der Waals and contact scores. For docking studies using CDOCK, the binding site was the same as that defined for DOCK, adding an extra 2.5 Å (for a total of 10 Å) in each axis direction. CDOCK grids had a spacing of 0.5 Å, and considered both protein interactions with different atom probes and electrostatic interactions, as explained above.

Filter plugin

For the tests shown in this article, DOCK 3.5 was used as a fast initial filter of the chemical libraries to be screened. DOCK uses a sphere-matching algorithm to fit ligand atoms to spheres in the binding pocket. We defined such points using GAGA as explained before. We chose to evaluate docking poses using the DOCK contact score in most cases and the DOCK force-field score in one case (VSP 2). For each conformer the single best DOCK solution obtained was retained. For VSP 1, 2, 3, 7, 9, and 11 DOCK scores for the best conformer of each molecule were stored; for VSP 8 and 10 the scores for the 10 best conformers of each molecule were kept. For VSP 3 the best solution per molecule was rescored using XSCORE [50] v. 1.2.1. but no improved performance was found and therefore, in the following, no further reference to this program will be made. Stored scores were normalized to ZScore (Eq. 4),

$$\mathrm{ZScore}_i = \frac{\mathrm{score}_i - \overline{\mathrm{score}}}{\sigma} \tag{4}$$

where $\overline{\mathrm{score}}$ is the mean and σ is the standard deviation values for the scores.

The VSDMIP application ***runDOCKFilter*** extracts information for all selected molecules from the database, performs the docking and stores the results in the database. Then ***getResultsFromDOCK*** extracts the docking results in a single coordinate file in MOL2 format for all the conformers with a ZScore higher than that provided by the user.

Docking plugin

We have used our in-house program CDOCK for the detailed docking phase of our protocols. Using the interaction energy grids calculated with CGRID, CDOCK exhaustively docks each molecule within the binding site. The centres of mass of the molecules are positioned on grid points equally spaced 1 Å, and discrete rotations of 27° on each axis are performed. A molecular mechanics force-field scoring function is used to score each pose (Eq. 5),

$$E_{\mathrm{MM}} = \sum_{i}^{\mathrm{prot}} \sum_{j}^{\mathrm{lig}} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{D(r_{ij}) r_{ij}} \right] \tag{5}$$

where A_{ij} and B_{ij} are the van der Waals parameters of the atom types to which atom i and j belong, r_{ij} is the distance between the ith atom in the protein and the jth from the ligand, q_i and q_j are the partial charges of atom i and j, respectively. $D(r_{ij})$ is a sigmoidal dielectric function that accounts for solvent screening (Eq. 6),

$$D(r_{ij}) = \frac{\varepsilon + 1}{1 + \mathrm{k}e^{-\lambda(\varepsilon+1)r_{ij}}} \tag{6}$$

where ε is the dielectric constant for water (78.39), k is $(\varepsilon - 1)/2$, λ is $\alpha/(\varepsilon + 1)$ and α is 1.0367. The docking score for each pose (van der Waals plus electrostatic) is then calculated using a trilinear interpolation method.

The VSDMIP application ***runCDOCK*** extracts information for all selected molecules from the database, performs the docking and stores the results in the database. Then ***getResultsFromVS*** extracts the docking results in a single coordinate file in MOL2 format [51] containing the number of molecules defined by the user, and creates a text file with the score for each result.

Rescoring plugin

CDOCK scores were corrected with solvation energies (Eq. 7) obtained by solving the Poisson equation (electrostatic part, hereafter PB).

$$\Delta G_{\text{desolv}} = \Delta G_{\text{ele}} + \Delta G_{\text{np}} \quad (7)$$

The electrostatic part of the energy (Eq. 8) is the sum of the electrostatic interactions between the ligand and protein in the complex (E^{LR}_{ele}), the change in solvation energy of the ligand upon binding ($\Delta G^{L}_{\text{desolv}}$) and the change in solvation energy of the receptor upon binding ($\Delta G^{R}_{\text{desolv}}$).

$$\Delta G_{\text{ele}} = E^{LR}_{\text{ele}} + (\Delta G^{L}_{\text{desolv}} + \Delta G^{R}_{\text{desolv}}) \quad (8)$$

The first term in Eq. 8 was computed as the product of ligand charges times the electrostatic potential created by the protein at each charge. The ligand desolvation energy was computed as the difference in energy between the solvated ligand and the energy of the ligand complexed to the uncharged receptor. The receptor desolvation term was computed analogously. All calculations were performed by numerically solving the linear Poisson equation using the finite difference method as implemented in DelPhi, PARSE atomic radii [52], AMBER ff99 partial charges for protein atoms, and MOPAC-calculated charges for ligand atoms, as described above. Each complex was immersed in a cubic box occupying 65% of the total volume with a grid spacing of 0.5 Å. The solute dielectric constant was set to 4 while that of the solvent was set to 80. The dielectric boundary was calculated using a solvent probe radius of 1.4 Å and a minimum separation of 11 Å was allowed between any solute atom and the box walls. The potentials at the grid points delimiting the box were calculated analytically by treating each charge atom as a Debye-Hückel sphere. The non-polar part of the desolvation (Eq. 9) was modelled as a linear relationship to the change of solvent-accessible surface area (SASA),

$$\Delta G_{\text{np}} = a + b\Delta\text{SASA} \quad (9)$$

where a is 0.092 kcal/mol, b is 0.00542 kcal/molÅ^2, and the change in SASA refers to the complex SASA minus the sum of that of the protein and the ligand alone. SASAs were calculated using the analytical method implemented in TINKER [53].

The VSDMIP application ***runDelPhiAndApolar*** extracts information for all selected molecules from the database, performs the calculation and stores the results as new entries into the database.

Visualization

Visualization of results is accomplished using the well-documented, free, open-source program Pymol [54].

Results

VSDMIP architecture and relational database

The architecture of VSDMIP is depicted in Fig. 1, and has been commented on, to some extent, in the Methods section. The main role played by the VSDMIP database is to act as a common origin and destination for all stages of the VS process. The XML configuration files and the different plug-in interfaces allow the user to configure different custom-made VS protocols, as exemplified below.

The Entity-Relationship schema of the database used by VSDMIP is shown in Fig. 2. The table primary keys are depicted with a key symbol, and foreign keys are marked with FK in brackets. The schema shows a compact way of storing and organizing chemical libraries and VS results through the MOLECULE, CONFIGURATION, and CONFORMER tables, and their relationships with RESULT_DF and RESULT_VS. Many different VS protocols can be composed by using as many DOCK_FILTERs (and their related RESULT_DF) and VIRTUAL_SCREENINGs (and their related RESULT_VS) as desired, and allow the results to be easily reused. Only the results from the last step are needed to run the following step in the workflow.

Performance of the individual docking tools (Tables 4–6)

DOCK performance is really poor for all but one system (ERa Jacobsson set), where it is marginally better than random selection. The results obtained using the force-field scoring function (VSP 2) are always better than those obtained with the contact scoring function (VSP 1), but the difference is not significant, the greater being around 0.3 U of AUC. This is also the range of variation in AUC values for different proteins with both scoring schemes. The EFs are very low and far from EF_{max} in all the cases and no significant differences are found when the force-field scoring function is used instead of the contact one. The best EFs are obtained for the ERa Jacobsson set ($EF_{\text{best}} = 3.35$). In most cases, the CDOCK AUCs are above 0.6, with the exception of the ERa receptor, for which results are below random (Jacobsson set) and slightly above random (Stahl set). In all the cases, except for the three fXa test sets, compound selection based only on van der Waals interaction energies (VSP 5) leads to smaller AUC values but this effect is not important. A noteworthy exception is the neuraminidase example where the electrostatic energy term improves the AUC by 0.47. Variations in AUC values across protein systems are significant (around 0.6 U) but independent from the scoring function (van der Waals plus electrostatic or van der Waals alone). For complete

CDOCK scoring (van der Waals plus electrostatic, VSP 4) EF are always better than when using DOCK (with both scoring functions, VSP 1 and 2) and above 10% of EF_{max}, reaching this value in one case (fXa Fontaine set), 80% (AChE), and 55% (ERa, Sthal set). If only the van der Waals term is used (VSP 5), the EF is greatly reduced, and this decrease is most dramatic in the AChE (30%), ERa (Sthal set, 20%), and neuraminidase (16%) examples.

Performance of mixed protocols: DOCK as a filter for CDOCK (Tables 4–6)

(a) Considering different ZScores. Filtering using DOCK, selecting molecules with a ZScore above 3.0, and then employing CDOCK (VSP 7) yields AUC values that are very close to those obtained from a random screen. No appreciable differences are observed when a more restrictive ZScore is used (1.5 vs. 3.0, VSP 9). The two exceptions in this case are AChE and neuraminidase. In the former, from an almost random performance (0.49) to a respectable value of 0.71, a difference of 0.22 units of AUC, while in the latter the difference is 0.14, thus reaching an AUC value close to 0.70. The differences across target proteins are of the same order (0.2 for VSP 7 and 0.3 for VSP 9). In only two cases the EF is above 50% of EF_{max} (ERa Sthal set 64% and fXa Fontaine set 70%) while the rest range between 4% and 35%. Small variations in EFs are observed for most of the sets when the lower ZScore value is used. These variations are more important in AChE (18% increase) and ERa (Sthal set, 28% decrease).

(b) Selecting 10 (instead of 1) conformations for each ligand from DOCK. When the ZScore is 3.0 (VSP 8) the AUC values are always above 0.5, the exception being the ERa example (Jacobsson set). Variations across targets are of the order of 0.3 AUC units. For a ZScore value of 1.5 (ID VSP 10) better results were obtained. In general, values for AUC above 0.6 are common, except for ERa (Jacobsson set) once again. Here, variations across the targets are as high as 0.5 AUC units. Three cases present EFs values above 50% of EF_{max} while the others range between 7% and 27%, when ZScore is 3.0. With the lower value, i.e. 1.5, the major differences are observed in AChE (18% increase), neuraminidase (14% increase), and ERa (Sthal set, 32% decrease).

Including solvent effect: PBSA as a correction term (Tables 4–6)

Molecular mechanics interaction energies (CDOCK scoring function) are corrected for desolvation effects using the PBSA method [55] on results directly obtained from CDOCK (VSP 6) or after a combined protocol encompassing DOCK and CDOCK (VSP 11). In the first case, the AUCs obtained are somewhat above random in many of the tests. The best AUC are for AChE and neuraminidase targets, while ERa (Jacobsson test) is the only case with AUC below random. In this case, the introduction of solvent effects via PBSA does not lead to an improvement over plain CDOCK results. Five sets show EFs above 30%, with fXa (Fontaine set) achieving EF_{max}. Except for AChE, the inclusion of desolvation always reduces the EF values. The reduction range goes from negligible (ERa, Jacobsson set, 2% or p38, 5%) to notorious (fXa Fontaine set, 30% or ERa Sthal set, 28%). In the second case, the AUC values

Table 4 Area Under the Curve (AUC), EF_{best}, EF_{max} for the three fXa sets obtained for eachVSP used

VSP id.	fXa								
	Fontaine set			Jacobsson set			Jorissen-Gilson set		
	AUC	EF_{best} (EF_{max} = 2.16)	log *t*	AUC	EF_{best} (EF_{max} = 4.94)	log *t*	AUC	EF_{best} (EF_{max} = 11)	log *t*
1	0.39	1.00 (0.99)	1.30	0.41	1.00 (0.99)	0.99	0.36	1.00 (1.00)	0.93
2	0.37	1.00 (1.00)	1.32	0.41	1.01 (0.96)	0.91	0.39	1.02 (0.98)	0.93
3	0.33	1.00 (1.00)	1.34	0.33	1.00 (1.00)	1.05	0.31	1.00 (1.00)	0.99
4	0.67	2.16 (0.01)	3.62	0.61	1.65 (0.05)	3.42	0.68	3.30 (0.04)	3.38
5	0.70	1.94 (0.01)	3.62	0.70	1.97 (0.05)	3.42	0.68	3.30 (0.02)	3.38
6	0.57	1.51 (0.14)	3.72	0.52	1.23 (0.22)	3.55	0.60	1.80 (0.20)	3.51
7	0.54	1.51 (0.01)	1.43	0.55	1.73 (0.06)	1.28	0.58	3.30 (0.02)	1.26
8	0.59	1.44 (0.03)	2.19	0.55	1.32 (0.05)	NA	0.65	2.57 (0.05)	1.99
9	0.58	1.58 (0.03)	1.76	0.56	1.38 (0.32)	1.53	0.61	2.93 (0.05)	1.46
10	0.64	1.43 (0.09)	2.58	0.57	1.29 (0.33)	2.37	0.65	2.48 (0.07)	2.31
11	0.54	1.25 (0.16)	2.51	0.54	1.48 (0.02)	2.29	0.58	1.83 (0.05)	2.29

Values in brackets refer to the percent (in decimal form) of the database scanned at which EF_{best} is found

Table 5 Area Under the Curve (AUC), EF_{best}, EF_{max} for the AChE and ERa sets obtained for each VSP used

VSP id.	AChE			ERa					
	Jacobsson set			Jacobsson set			Stahl set		
	AUC	EF_{best} ($EF_{max} = 19.52$)	log *t*	AUC	EF_{best} ($EF_{max} = 8.04$)	log *t*	AUC	EF_{best} ($EF_{max} = 21$)	log *t*
1	0.30	1.08 (0.86)	0.84	0.56	2.01 (0.01)	0.90	0.28	1.00 (1.00)	1.24
2	0.32	1.02 (0.98)	0.85	0.60	3.35 (0.04)	0.91	0.57	1.73 (0.22)	1.23
3	0.26	1.02 (0.98)	1.17	0.36	1.03 (0.97)	0.99	0.17	1.00 (1.00)	1.29
4	0.97	15.97 (0.01)	3.09	0.35	1.00 (1.00)	3.02	0.55	11.45 (0.01)	3.13
5	0.94	9.58 (0.05)	3.09	0.33	1.00 (1.00)	3.02	0.52	7.64 (0.01)	3.13
6	0.93	15.97 (0.01)	3.46	0.47	1.00 (1.00)	NA	0.64	5.73 (0.01)	3.32
7	0.49	5.32 (0.01)	1.08	0.43	1.00 (1.00)	1.15	0.63	13.36 (0.01)	1.35
8	0.69	12.42 (0.01)	NA	0.42	1.00 (1.00)	1.74	0.66	17.18 (0.01)	1.80
9	0.71	14.20 (0.01)	1.25	0.43	1.00 (1.00)	NA	0.65	7.64 (0.01)	1.46
10	0.93	15.97 (0.01)	NA	0.45	1.00 (1.00)	NA	0.71	10.5 (0.02)	2.09
11	0.49	1.77 (0.01)	2.75	0.43	1.00 (1.00)	2.44	0.63	11.45 (0.01)	2.47

Values in brackets refer to the percent (in decimal form) of the database scanned at which EF_{best} is found

Table 6 Area Under the Curve (AUC), EF_{best}, EF_{max} for the CDK2, neuraminidase and p38MAP sets obtained for each VSP used

VSP id.	CDK2			Neuraminidase			p38MAP		
	Jorissen-Gilson set			Stahl set			Stahl set		
	AUC	EF_{best} ($EF_{max} = 21$)	log *t*	AUC	EF_{best} ($EF_{max} = 59.82$)	log *t*	AUC	EF_{best} ($EF_{max} = 46.45$)	log *t*
1	0.36	1.00 (1.00)	1.20	0.35	1.09 (0.87)	1.26	0.23	1.00 (1.00)	0.95
2	0.41	1.91 (0.01)	1.18	0.56	1.36 (0.39)	0.37	0.42	1.17 (0.31)	1.00
3	0.41	1.02 (0.98)	1.27	0.20	1.00 (1.00)	1.35	0.26	1.03 (0.97)	1.10
4	0.67	2.45 (0.07)	3.02	0.89	10.88 (0.01)	3.25	0.64	4.22 (0.02)	NA
5	0.63	2.12 (0.09)	3.02	0.42	1.06 (0.78)	3.25	0.60	4.22 (0.01)	NA
6	0.57	2.06 (0.14)	3.33	0.72	6.22 (0.08)	3.24	0.55	1.92 (0.12)	NA
7	0.55	1.91 (0.13)	1.28	0.55	10.88 (0.01)	1.40	0.65	4.75 (0.09)	1.36
8	0.63	3.82 (0.01)	1.67	0.63	5.44 (0.01)	1.96	0.73	4.22 (0.03)	2.02
9	0.59	2.67 (0.05)	1.40	0.69	10.88 (0.01)	1.56	0.72	4.22 (0.01)	1.57
10	0.61	2.10 (0.10)	1.95	0.82	13.6 (0.02)	NA	0.75	4.22 (0.01)	NA
11	0.55	1.91 (0.02)	2.54	0.53	1.81 (0.03)	2.56	0.64	3.75 (0.10)	2.60

Values in brackets refer to the percent (in decimal form) of the database scanned at which EF_{best} is found

are almost unaltered compared with the situation in which no solvent effects are introduced. Again, performance on the ERa target (Jacobsson test) was worse than random. In terms of variations across the targets, 0.46 AUC units are found when PBSA is applied to CDOCK without previous DOCK filter and 0.21 with the DOCK filter. The EFs here are similar to those commented on before, but only four sets are above 30% (instead of five) and none reaches EF_{max}. Again, a general reduction in these values is observed when desolvation is taken into account, the range being between 2% (p38) and 18% (AChE).

Finally, and related to the percentage of the database at which the best EF is found, almost the entire database has to be explored when DOCK is used. There are, however, some exceptions (6 out of 99): VSP 2 for ERa (Jacobsson and Sthal sets), CDK2, neuraminidase and p38MAP, and VSP 1 for ERa (Jacobsson test). For the rest of the protocols, the best EF is achieved very early for most of the cases, except in those for which over 25% of the database has to be screened: ERa (Jacobsson set, VSP 9 and 10), neuraminidase (VSP 5), and fXa (Jacobsson set, VSP 9 and 10).

Discussion

A plethora of methods are available to propose new drug candidates starting simply from 2D sketches of virtual chemical libraries. It should then be possible to integrate all

the needed software elements to create customized workflows for any desired project. But the data flow between the different steps (input/output connections) seems to be an important problem mainly due to the great variety of formats that can be used to describe molecular structures. Although some advances have been done (e.g., SMILES and InChI [IUPAC International Chemical Identifier]), a consensus format is far from having been adopted. On the other hand, life sciences, in general, and computer-aided drug design, in particular, witness a data deluge coming from the target side, i.e., 3D protein structures available from structural genomics projects, as well as from the ligand side, i.e., huge chemical libraries with millions of molecules to be screened. Finally, the amount of data to be processed, stored and managed requires potent database engines. These three aspects have motivated us to develop VSDMIP as an integrative platform for handling these data. The main advantages of VSDMIP are: (1) the possibility to perform automated VS experiments; (2) easy comparison of different protocols; (3) total flexibility to design VS protocols; (4) the implementation of an XML mechanism to plug in new software pieces to customize protocols at will; and (5) the generation of a coupled relational database to have all the data organized and ready to use. VSDMIP presently lacks a graphical user interface (GUI), but it may be added in the future. When this is done, it will be able to additionally provide information regarding both the receptor and the ligand-binding site interactions. Small changes in the database schema will make it possible to store the results from docking engines that generate new ligand conformations as solutions, thus extending the current capability that basically works with docking engines that rely on pre-computed sets of conformers.

As similar approaches have been published recently, a brief discussion of some of them in comparison with VSDMIP is in order. Probably the most similar approach is that reported by SciTegic, Inc. named Pipeline Pilot [56, 57]. This commercial software uses the technology known as data pipeline to construct and execute customizable workflows using components that encapsulate mainly cheminformatics-based algorithms (although docking can also be performed using programs GOLD and FLEXX). Hassan et al. have shown the usefulness of this platform in a recent review [57], where they also show results from virtual screening experiments using Bayesian learning technique on several targets. It has an underlying database in common with VSDMIP, while performing many different types of calculations and methods.

Along the the same line, Astex Therapeutics [58] reported a proprietary web-based platform that integrates an ORACLE relational database to store molecules (from the Astex Technology Library of Available Substances [ATLAS] database), properties, target data, binding interactions, and results. VS experiments can be performed using compounds obtained directly from the database or created virtually thorough stored chemical reactions. Molecules in SMILES 2D strings are converted into 3D structures using CORINA and are docked with GOLD. It also contains a GUI to set up the experiments and visualize the results. VSDMIP has some added advantages such as its modularity, the possibility for users to generate their own chemical libraries, and the fact that it will be available to academic users upon request. On the down side, to the present day VSDMIP works through a command line interface.

The idea of employing user-configurable XML files to perform customized drug design-related workflows (as we do in VSDMIP) has been implemented by Lehtovuori and Nyrönen [59]. In their SOMA approach, the user decides which protocol to use by selecting, from a web browser interface, the steps and adequate parameters (even though in the reported implementation only molecular structure-based properties calculation and/or docking experiments with GOLD can be performed). Then program *Grape* manages the workflow by joining the needed applications in the order established by the user through the execution nodes (to run and manage the applications). Finally, the same web browser interface is used to retrieve, visualize and analyze the results. The output produced by each executable node is encoded, labelled (to keep track of the process at any time), and updated with the output information generated by the subsequent steps (depending on the selected protocol). Another important component is the toolkit that contains utility programs to perform intermediate tasks (file format conversions and generation of execution files, among others). The entire SOMA protocol is encoded within a unique XML file, which means that, after the required input has been provided, the user does not need to interact with the system until the entire workflow has been completed. This is very convenient for already established protocols. On the contrary, VSDMIP is more focused on decision making after every step. This is so because if more than one docking program is going to be used, or different filters are available, it is necessary to stop at each step to check the results before carrying on. A fine tuning at this level is not easily tractable nor is it feasible to implement in an automatic protocol. Another aspect that deserves to be commented on in relation to VSDMIP is the lack of a database to manage the results. SOMA stores the results of the entire workflow in a large single file that is displayed as a table within the interface. It is unclear whether the molecules and results already obtained could be used in another set of experiments. VSDMIP is totally flexible in this respect (i.e. once a molecule has been inserted into VSDMIP it can be reutilised as many times as desired). The availability of the SOMA XML schema and

the web-implemented GUI makes this approach very attractive to drug designers with low programming skills who want to focus more on the chemical/biochemical aspects of the problem in hand rather than on the technical details. A final aspect in common with VSDMIP is the possibility to incorporate new applications in a relatively easy way using XML scripts.

Although not strictly comparable with VSDMIP, some other related tools also merit some comments. For example, the Autogrid/Autodock suite, which has attracted a lot of interest in recent years and is probably called to become one of the most widely used programs for docking purposes. Besides the automatic tools available at the developers' web site, two other applications for VS experiments have to be noted: BDT [60] and DOVIS [61]. The former allows the user to interact with the program code thorough a graphical front-end application that automatically performs grids preparation and their combination (to allow for receptor flexibility), docking computation and analysis of the results. The latter also incorporates an additional step for ligand preparation. The main advantage of DOVIS over BDT is that DOVIS allows docking to be performed in parallel using Linux clusters (with or without a queuing system). In both cases, however, the user is restricted to just one docking program and no database exists to manage different projects. Nonetheless, the free distribution of the programs and the easy-to-use graphical interfaces make them ideal tools for researchers who are more interested in getting answers to their particular problems than in the docking process itself.

To test the performance of our platform we used six different targets, and two of these with different sets of active compounds. The compilation of 11 VSP allows us to discuss some important effects in VS experiments, although more detailed studies will have to be conducted to assess the performance of other more sophisticated protocols. Three main questions are particularly addressed here: (1) the possible advantage of a combined docking protocol (using two programs, a first one as a filter, and a second more exhaustive one as a final docking tool) over a single one; (2) the effect of the number of molecules and conformers per molecule that are passed from the filter on to the final docking tool; and (3) the impact of incorporating desolvation using a continuum method as a rescoring function. For reference, the AUCs obtained for all the protocols applied to the fXa and neuraminidase test sets are depicted in Fig. 3 and the time employed per database molecule in these same cases is graphically shown in Fig. 4. As expected, better results are obtained when the VS is performed directly with CDOCK (compare VSP4–VSP6 with VSP1–VSP3) due to its superior scoring function and the exhaustive search within the binding site. More interesting, however, is the fact that when DOCK is

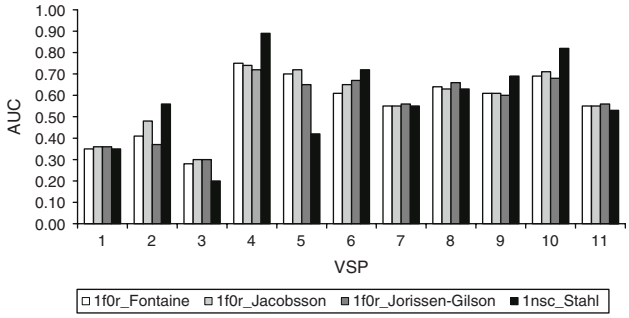


Fig. 3 Area Under the Curve (AUC) versus virtual screening protocol (VSP) identification number for the three sets of fXa and neuraminidase

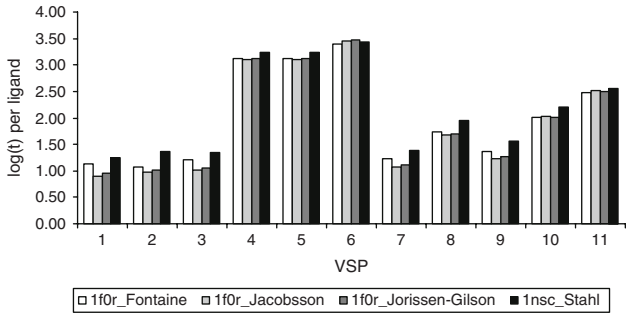


Fig. 4 CPU time (seconds, in logarithmic units) per ligand required for each virtual screening protocol (VSP) for the three sets of fXa and neuraminidase

used as a filter preceding CDOCK (VSP7–VSP11) reasonable results are obtained, which attests to its suitability to remove undesirable ligand structures. On the other hand, as clearly shown in Fig. 4, the best results are obtained for VSP4, which is also the most time-consuming protocol. On the contrary, VSP10 displays similar results compared to VSP4 but the computer time increases between 1 and 2 orders of magnitude. In other words, the use of a filter saves computational time while retaining most of the AUC values. The representative ROC plots depicted in Fig. 5 illustrate the aforementioned effect.

Given the amount of molecules present in standard databases, it is not yet computationally feasible to conduct VS experiments using extremely accurate docking programs. Instead, a common approach is to use some sort of concatenated filters to reduce the number of molecules to be docked. In other studies, after initial filtering and docking, different scoring functions are employed and then candidates are selected on the basis of a consensus criterion [62]. More promising alternatives make use of more than one docking program in increasing order of accuracy [63, 64]. We chose DOCK as an initial docking program because it is fast enough to screen a molecule in a few seconds, the total time depending on the number of conformers and the number of spheres used to describe the

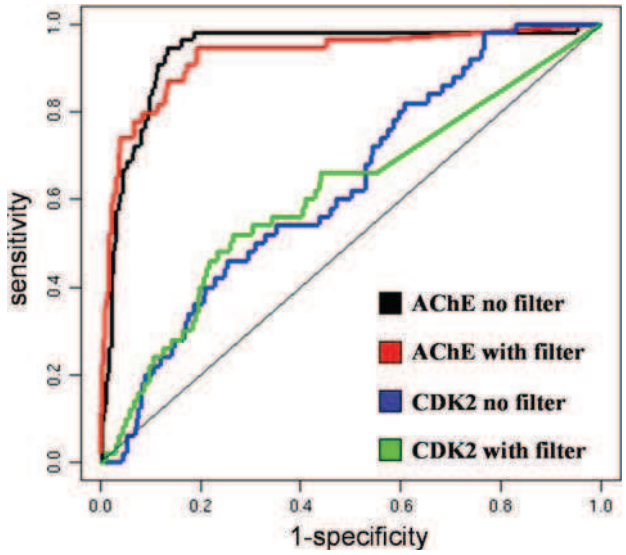


Fig. 5 ROC plots for AChE and CDK2 validation tests using VSP4 (without DOCK filter) or VSP10 (with DOCK filter)

binding site. In order to compare the relative performance of DOCK when used alone or in conjunction with another more exhaustive docking tool, an initial study was done considering DOCK alone. The performance of our configuration for DOCK is not better than random either with a contact- or a force-field-based scoring function, as can be seen from the AUC and EF_{best} values. We subjected our in-house docking program CDOCK to the same test and the results show that CDOCK clearly outperforms this particular DOCK configuration both in terms of AUC and EF_{best} values. As a trade-off for this increased accuracy, however, CDOCK employs more computer time. Another observation is that the elimination of the coulombic component does not really produce any variation in AUC or EF_{best} values, a result that is presently being scrutinized.

As stated before, a good approach consists of using a sequential combination of docking programs in increasing order of accuracy. Here we tested different protocols using first DOCK and then CDOCK. Variations among them relate to the number of compounds and conformations per compound to be passed from DOCK to CDOCK. In all cases, 100 conformers are used in DOCK. First, two cut-off values for the numbers of molecules passed to CDOCK are set using ZScore (3.0 and 1.5), DOCK contact scores and only one conformer per molecule. The AUC values do not seem to be dependent on ZScore, whereas the EF_{best} values show some degree of variation depending on the type of target. This means that although most of the molecules that survive after applying a low ZScore value are in fact inactive, they can be recognized and discarded by CDOCK. In general, the combination of both docking programs shows an intermediate degree of performance, as could be expected, between DOCK alone and CDOCK alone. Secondly, better results are obtained when instead of a single conformation 10 are passed on to CDOCK. This is a commonly observed effect, provided that conformational sampling has been performed adequately, and simply states that with more conformations per molecule the docking algorithm stands a better chance of detecting the correct pose for a binder [65]. Finally, use of the same ZScore but a different number of conformers per molecule (1 or 10) does not lead to any appreciable changes in AUC or EF_{best} values, the variation in the latter case depending on the type of target.

The third point addressed here is related to the introduction of solvent effects via PBSA as a post-scoring function. Together with flexibility issues [66], an adequate representation of the desolvation process that accompanies ligand-receptor binding is a major hurdle in VS studies. It is also a problem in traditional docking but, due to the fact that a small number of molecules are going to be studied, more elaborated solvation methods can be applied in this case, even when time consumption for such a calculation is often seen as a shortcoming. Methods based on PB or GB approximations are common but become impractical at the large scales a VS experiment requires although some promising approaches have already been published [19, 67, 68]. We have observed a small influence of solvent effects on AUC values using PB after CDOCK and no effect at all when used after the combination of DOCK and CDOCK. Again, the influence on EF_{best} is target-dependent, but in all cases a decrease in these values has been found.

VSDMIP has already proved successful in some recent scientific applications, such as one devoted to the discovery of new inhibitors of the DNA repair protein O6-alkylguanine DNA alkyltransferase [69]. Four compounds selected out of 3.5 million molecules from the ZINC database [70] showed acceptable in vivo and in vitro activities. In another example using in vivo screenings of a chemical combinatorial library and VSDMIP, we were able to develop small molecules that compete with ubiquitin E2 variant (UEV) for its interactions with ubiquitin-conjugating enzyme UBC13 and inhibit its enzymatic activity. The UEV–UBC13 complex is also implicated in mechanisms of DNA repair (unpublished results).

Conclusion

An integrated computational platform to perform VS experiments has been developed that includes an associated relational database which stores (i) molecules and molecular properties (energies, conformations, charges, etc.), (ii) results from docking filters, and (iii) final VS results. This procedure allows the inserted molecules to be reused in as many VS experiments as desired as well as the continued

incorporation of new molecules. Also, it is easy and fast to create and allows a battery of analyses to be performed in order to test a particular VS protocol. The modular idea underlying its design is one of the stronger points, as the user is able to replace existing modules with new ones to create customizable protocols. Finally, and under development, is the idea to include protein set-up in an automatic fashion within the database, allowing the storage of geometrical and energetic characteristics of the binding site, which should serve as a classification tool for binding sites. The platform has been prepared as a bundled package to be distributed to the scientific community upon request from the authors [71]. In brief, all the programs implemented in the platform (except those that need to be purchased, by a modest prize, such as CORINA or DelPhi) are either free (MOPAC, DOCK, FRED, AutoDock) or will be released under a scientific/academic non-profit and non-commercial license as is the case for ALFA, CGRID, CDOCK, and ISM. The scripts to create the database structure as well as the XML configuration files will be also provided.

Acknowledgements Work at the CBM-SO was partially supported by a grant from "Comunidad de Madrid" thorough BIPEDD project (SBIO-0214–2006) and from "Ministerio de Educación y Ciencia" (BIO2005–0576). J.S. and J.E. were funded by grants BFU2007–61476/BMC (MEC, Spain) and PM076/2006 (DGA, Spain). J.E.'s research stage at CBM "Severo Ochoa" was funded by grants DGA (CONSI + D)/CAI (Spain) and FPU (MEC, Spain). J.E. is recepient of an FPU grant (MEC, Spain). J.E. thanks Alejandra Leo-Macías for help in using the MODELLER software. A.M. and R.G.-R. thank David Abia and Rubén Muñoz for technical support. We also acknowledge the generous allocation of computer time at the Barcelona Supercomputing Center. This work would not have been possible without the encouraging help of Ángel R. Ortiz, to whose memory this article is dedicated.

References

1. Smith A (2002) Nature 418:453
2. Lahana R (1999) Drug Discov Today 4:447. doi:10.1016/S1359-6446(99)01393-8
3. Ramesha CS (2000) Drug Discov Today 5:43. doi:10.1016/S1359-6446(99)01444-0
4. Perola E, Walters WP, Charifson PS (2004) Proteins 56:235. doi:10.1002/prot.20088
5. Warren GL, Andrews CW, Capelli AM et al (2006) J Med Chem 49:5912. doi:10.1021/jm050362n
6. Kitchen DB, Decornez H, Furr JR et al (2004) Nat Rev Drug Discov 3:935. doi:10.1038/nrd1549
7. Adcock SA, McCammon JA (2006) Chem Rev 106:1589. doi:10.1021/cr040426m
8. Brandsdal BO, Osterberg F, Almlof M et al (2003) Adv Protein Chem 66:123. doi:10.1016/S0065-3233(03)66004-3
9. Shoichet BK (2004) Nature 432:862. doi:10.1038/nature03197
10. Leach AR, Shoichet BK, Peishoff CE (2006) J Med Chem 49:5851. doi:10.1021/jm060999m
11. Corina Molecular Networks (2000). GmbH Computerchemie Langemarckplatz 1, Erlangen, Germany. http://www.molecular-networks.com/software/corina/index.html. Accessed 24 Sept 2008
12. Gil-Redondo R (2006) Master Thesis: Implementación de una plataforma para el cribado virtual de quimiotecas. UNED, Madrid
13. Stewart JJ (1990) J Comput Aided Mol Des 4:1. doi:10.1007/BF00128336
14. Kuntz ID, Blaney JM, Oatley SJ et al (1982) J Mol Biol 161:269. doi:10.1016/0022-2836(82)90153-X
15. McGann MR, Almond HR, Nicholls A et al (2003) Biopolymers 68:76. doi:10.1002/bip.10207
16. Perez C, Ortiz AR (2001) J Med Chem 44:3768. doi:10.1021/jm010141r
17. Morris GM, Goodsell DS, Halliday RS et al (1998) J Comput Chem 19:1639. doi:10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B
18. Rocchia W, Sridharan S, Nicholls A et al (2002) J Comput Chem 23:128. doi:10.1002/jcc.1161
19. Morreale A, Gil-Redondo R, Ortiz AR (2007) Proteins 67:606. doi:10.1002/prot.21269
20. Lipinski CA, Lombardo F, Dominy BW et al (2001) Adv Drug Deliv Rev 46:3. doi:10.1016/S0169-409X(00)00129-0
21. Triballeau N, Acher F, Brabet I et al (2005) J Med Chem 48:2534. doi:10.1021/jm049092j
22. Weininger D (1988) J Chem Inf Comput Sci 28:31. doi:10.1021/ci00057a005
23. Ctfile Formats MDL (2007). Symyx, California. http://www.mdl.com/solutions/white_papers/ctfile_formats.jsp. Accessed 24 Sept 2008
24. Dewar MJS, Thiel W (1977) J Am Chem Soc 99:2338. doi:10.1021/ja00449a053
25. Maignan S, Guilloteau JP, Pouzieux S et al (2000) J Med Chem 43:3226. doi:10.1021/jm000940u
26. Murcia M, Ortiz AR (2004) J Med Chem 47:805. doi:10.1021/jm030137a
27. Jacobsson M, Liden P, Stjernschantz E et al (2003) J Med Chem 46:5781. doi:10.1021/jm030896t
28. Kryger G, Silman I, Sussman JL (1999) Structure 7:297. doi:10.1016/S0969-2126(99)80040-9
29. Arris CE, Boyle FT, Calvert AH et al (2000) J Med Chem 43:2797. doi:10.1021/jm990628o
30. Thomas MP, McInnes C, Fischer PM (2006) J Med Chem 49:92. doi:10.1021/jm050554i
31. Bissantz C, Folkers G, Rognan D (2000) J Med Chem 43:4759. doi:10.1021/jm0010441
32. Shiau AK, Barstad D, Loria PM et al (1998) Cell 95:927. doi:10.1016/S0092-8674(00)81717-1
33. Burmeister WP, Henrissat B, Bosso C et al (1993) Structure 1:19. doi:10.1016/0969-2126(93)90005-2
34. Murray CW, Baxter CA, Frenkel AD (1999) J Comput Aided Mol Des 13:547. doi:10.1023/A:1008015827877
35. Cavasotto CN, Abagyan RA (2004) J Mol Biol 337:209. doi:10.1016/j.jmb.2004.01.003
36. Wang Z, Harkins PC, Ulevitch RJ et al (1997) Proc Natl Acad Sci USA 94:2327. doi:10.1073/pnas.94.6.2327
37. Canutescu AA, Shelenkov AA, Dunbrack RL Jr (2003) Protein Sci 12:2001. doi:10.1110/ps.03154503
38. Fiser A, Sali A (2003) Methods Enzymol 374:461. doi:10.1016/S0076-6879(03)74020-8
39. Case DA, Darden TA, Cheatham TE et al (2004) AMBER 8. University of California, San Francisco
40. Wang J, Cieplak P, Kollman PA (2000) J Comput Chem 21:1049. doi:10.1002/1096-987X(200009)21:12<1049::AID-JCC3>3.0.CO;2-F
41. Gordon JC, Myers JB, Folta T et al (2005) Nucleic Acids Res 33:W368. doi:10.1093/nar/gki464
42. Honig B, Nicholls A (1995) Science 268:1144. doi:10.1126/science.7761829

43. Tishmack PA, Bashford D, Harms E et al (1997) Biochemistry 36:11984. doi:10.1021/bi9712448
44. Hawkins GD, Cramer CJ, Truhlar DG (1995) Chem Phys Lett 246:122. doi:10.1016/0009-2614(95)01082-K
45. Hawkins GD, Cramer CJ, Truhlar DG (1996) J Phys Chem 100:19824. doi:10.1021/jp961710n
46. Tsui V, Case DA (2000) Biopolymers 56:275. doi:10.1002/1097-0282(2000)56:4<275::AID-BIP10024>3.0.CO;2-E
47. Golebiowski A, Townes JA, Laufersweiler MJ et al (2005) Bioorg Med Chem Lett 15:2285. doi:10.1016/j.bmcl.2005.03.007
48. Mehler EL, Solmajer T (1991) Protein Eng 4:903. doi:10.1093/protein/4.8.903
49. Wang K, Murcia M, Constans P et al (2004) J Comput Aided Mol Des 18:101. doi:10.1023/B:jcam.0000030033.26053.40
50. Wang R, Lai L, Wang S (2002) J Comput Aided Mol Des 16:11. doi:10.1023/A:1016357811882
51. Tripos Mol2 File Format (2007). Tripos LP, Missouri. http://www.tripos.com/tripos_resources/fileroot/mol2_format_Dec07.pdf . Accessed 24 Sept 2008
52. Sitkoff D, Sharp KA, Honig B (1994) J Phys Chem 98:1978. doi:10.1021/j100058a043
53. Molecular Modeling Package TINKER (2004). http://dasher.wustl.edu/tinker. Accessed 24 Sept 2008
54. DeLano WL (2002). The PyMOL Molecular Graphics System DeLano Scientific, Palo Alto, CA. http://pymol.sourceforge.net. Accessed 24 Sept 2008
55. Kollman PA, Massova I, Reyes C et al (2000) Acc Chem Res 33:889. doi:10.1021/ar000033j
56. SciTegic, Inc. 10188 Telesis Court, Suite 100, San Diego, CA 92121, USA, http://accelrys.com/products/scitegic. Accesed 24 Sept 2008
57. Hassan M, Brown RD, Varma-O'brien S (2006) Mol Divers 10:283. doi:10.1007/s11030-006-9041-5
58. Watson P, Verdonk M, Hartshorn MJ (2003) J Mol Graph Model 22:71. doi:10.1016/S1093-3263(03)00137-2
59. Lehtovuori PT, Nyronen TH (2006) J Chem Inf Model 46:620. doi:10.1021/ci050388n
60. Vaque M, Arola A, Aliagas C et al (2006) Bioinformatics 22:1803. doi:10.1093/bioinformatics/btl197
61. Zhang S, Kumar K, Jiang X et al (2008) BMC Bioinformatics 9:126. doi:10.1186/1471-2105-9-126
62. Yang JM, Chen YF, Shen TW et al (2005) J Chem Inf Model 45:1134. doi:10.1021/ci050034w
63. Maiorov V, Sheridan RP (2005) J Chem Inf Model 45:1017. doi:10.1021/ci050089y
64. Miteva MA, Lee WH, Montes MO et al (2005) J Med Chem 48:6012. doi:10.1021/jm050262h
65. Knox AJ, Meegan MJ, Carta G et al (2005) J Chem Inf Model 45:1908. doi:10.1021/ci050185z
66. Teague SJ (2003) Nat Rev Drug Discov 2:527. doi:10.1038/nrd1129
67. Huang N, Kalyanaraman C, Irwin JJ (2006) J Chem Inf Model 46:243. doi:10.1021/ci0502855
68. Kuhn B, Gerber P, Schulz-Gasch T (2005) J Med Chem 48:4040. doi:10.1021/jm049081q
69. Ruiz FM, Gil-Redondo R, Morreale A (2008) J Chem Inf Model 48:844. doi:10.1021/ci700447r
70. Irwin JJ, Shoichet BK (2005) J Chem Inf Model 45:177. doi:10.1021/ci049714+
71. Gil-Redondo R, Estrada J, Morreale A, et al. (2008). VSDMIP. CBM "Severo Ochoa" (CSIC-UAM) and Universidad de Zaragoza, Spain. http://ub.cbm.uam.es/VSDMIP.htm. Accessed 24 Sept 2008

British Journal of Pharmacology (2009), 157, 4–13

0007-1188/09
www.brjpharmacol.org

REVIEW

Structural features of mammalian histidine decarboxylase reveal the basis for specific inhibition

AA Moya-García[1,2], A Pino-Ángeles[1,2], R Gil-Redondo[3], A Morreale[3] and F Sánchez-Jiménez[1,2]

[1]*Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, Campus de Teatinos, Málaga, Spain,* [2]*CIBER de Enfermedades Raras (CIBERER), Valencia, Spain, and* [3]*Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa (CSIC-UAM), C/Nicolás Cabrera 1, Campus de Cantoblanco, Madrid, Spain*

For a long time the structural and molecular features of mammalian histidine decarboxylase (EC 4.1.1.22), the enzyme that produces histamine, have evaded characterization. We overcome the experimental problems for the study of this enzyme by using a computer-based modelling and simulation approach, and have now the conditions to use histidine decarboxylase as a target in histamine pharmacology. In this review, we present the recent (last 5 years) advances in the structure–function relationship of histidine decarboxylase and the strategy for the discovery of new drugs.
British Journal of Pharmacology (2009) **157**, 4–13; doi:10.1111/j.1476-5381.2009.00219.x

Keywords: histamine; histidine decarboxylase; QM/MM; virtual screening; homology modelling; molecular dynamics

Abbreviations: α-FMH, α-fluromethylhistidine; DDC, L-amino acid decarboxylase; EGCG, epigallocatechin-3-gallate; GAD, glutamate decarboxylase; GBSA, generalized born surface area; HDC, histidine decarboxylase; HME, histidine methyl esther; MD, molecular dynamics; MM, molecular mechanics; NMR, nuclear magnetic resonance; PLP, pyridoxal-5′-phosphate; QM, quantum mechanics; VS, virtual screening

Introduction and historical background

Histamine has many different and important roles in mammalian physiopathology. Among other physiopathological processes, it is involved in allergy and other inflammatory responses, gastric acid secretion, bone loss, the control of sleep and food intake, and schizophrenia (Jorgensen *et al.*, 2007; Haas *et al.*, 2008; Ohtsu, 2008; Schubert and Peura, 2008). Currently, the effects of histamine are controlled by using modulators (mainly antagonists), tailored as specifically as possible to block the histamine binding to one of the four known membrane histamine receptors (H_1, H_2, H_3 and H_4). All of these subtypes are homologous G protein-coupled receptors specialized to elicit particular intracellular signals for different physiopathological effects (Kuramasu *et al.*, 2006; Gurevich and Gurevich, 2008; Thurmond *et al.*, 2008). The present strategy used to control unwanted effects of histamine involves finding a specific antagonist that will interfere with a given undesirable effect of the amine on a particular process, with the lowest possible side-effects on other cell types that express different histamine receptor subtypes specific for other physiological effects (see other chapters in this number). However, in practice, this is not an easy task due to the homology among the receptor proteins and the similarities in their methods of binding a common small biomolecule, namely histamine. As such, cross-reaction is one of the major problems of this strategy. In addition, the effective blockage of a given receptor by a specific antagonist may avoid one physiological effect of histamine, but does not prevent the higher synthesis and/or release of histamine that characterizes many histamine-related diseases. An increase in endogenous/newly synthesized histamine can have multiple consequences at both cellular and systemic levels; these effects need further molecular characterization to be controlled (Tanaka and Ichikawa, 2006). How can we control histamine synthesis?

Histamine is produced by α-decarboxylation of histidine and this reaction is catalysed by histidine decarboxylase (HDC). In mammals and other eukaryotic organisms, as well as in Gram-negative bacteria, HDC is a pyridoxal-5′-phosphate (PLP)-dependent enzyme (EC 4.1.1.22) expressed only in a small number of cell types, mainly mast cells, neurons located in the posterior hypothalamus and gastric enterochromaphin-like cells (Medina *et al.*, 2003; 2005). Some important aspects of HDC regulation of histamine synthesis have been elucidated in gastric cells (Chen *et al.*, 2006; Ai *et al.*, 2007), but not much is known of this process in other

Correspondence: F Sánchez-Jiménez, Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, Campus de Teatinos, Málaga 29071, Spain. E-mail: kika@uma.es
Received 15 January 2009; accepted 29 January 2009

cell types where HDC expression seems to be associated with poorly-characterized cell differentiation processes and/or subject to alterations in the control of cell proliferation (Fitz *et al.*, 2008; Tachibana *et al.*, 2008). At present, there is no clear strategy for controlling histamine production by interfering with HDC expression.

For a long time, a characterization of the activity of these HDC enzymes has been restricted by their extreme instability. There are no X-rays depicting the structure of any PLP-dependent HDC. In mammals, the problem is even more complex, as the enzyme needs to undergo post-translational maturation to reach its active form. The precursor of the active enzyme seems to be a fusion of a fragment homologous to other PLP-dependent amino acid decarboxylases, and a C-terminal portion (around 160 amino acids), with an unknown role (apart from inhibition of the enzymic activity) and no obvious homology with other functional proteins. This terminal fragment is lost during the activation of the enzyme (Engel *et al.*, 1996; Fleming *et al.*, 2004a), but its intracellular function and the proteolytic mechanisms for both its cleavage and the degradation of the active enzyme (both very rapid and complex processes) have not been completely elucidated. Different proteolytic mechanisms have been proposed: calpains, proteasome and caspase-9 (Viguera *et al.*, 1994; Olmo *et al.*, 1999; 2000; Rodriguez-Agudo *et al.*, 2000; Furuta *et al.*, 2007), but stabilization of the enzyme seems to be neither a useful nor a feasible target for a specific intervention affecting mammalian HDC levels.

In spite of its instability, even in purified preparations (Olmo *et al.*, 2000; Fleming *et al.*, 2004b), putative inhibitors able to bind directly to the protein have been tested, *in vitro*, using cell-free extracts and recombinant mammalian HDC (Olmo *et al.*, 2002; Rodriguez-Caso *et al.*, 2003a). In fact, several substrate analogues and natural products have been described as direct HDC inhibitors; these include α-fluromethylhistidine (α-FMH), histidine methyl ester (HME) (DeGraw *et al.*, 1977) and a compound derived from green tea, epigallocatechin-3-gallate (EGCG) (Rodriguez-Caso *et al.*, 2003a,b). In the case of the substrate analogues (α-FMH and HME), their mechanisms of action seem to be completely understood. Both analogues react with PLP. HME cannot be decarboxylated and blocks the enzyme in the external aldimine state; α-FMH proceeds to decarboxylation but then forms inactive derivatives of the PLP-product adduct, which are then released slowly from the catalytic site. Nevertheless, these are not specific inhibitors for the mammalian enzyme, since they also act on the homologous PLP-dependent HDC of enterobacteria (Bhattacharjee and Snell, 1990). Thus, their usefulness as therapeutic agents seems to be very limited. Of the many different natural products, EGCG is the one with the greatest inhibitory capacity against mammalian HDC (Nitta *et al.*, 2007), with promising anti-inflammatory effects when assayed in mast cells and monocytes (Melgarejo *et al.*, 2007). It binds to the enzyme and seems to change the PLP conformation inside the catalytic site, so blocking its reaction with the substrate (Rodriguez-Caso *et al.*, 2003b). Nevertheless, the nature of the binding is not yet known. In addition, EGCG is not a specific inhibitor of mammalian HDC, since it is also able to bind and effectively inactivate aromatic L-amino acid (or Dopa) decarboxylase (DDC) (Bertoldi *et al.*, 2001), another important element of our neurological and neuroendocrine system; DDC is the enzyme responsible for the synthesis of the neurotransmitters 5-hydroxytryptamine and dopamine (Haavik *et al.*, 2008). In summary, specificity is the problem in constructing a strategy against the histamine-producing enzyme (Moya-Garcia *et al.*, 2006). The ideal inhibitor should be able to inhibit the mammalian enzyme but have minimal effects on other enzymes present in the human organism and necessary for its homeostasis.

The enterobacterial HDC and both mammalian DDC and HDC are homologous enzymes; all belong to the DDC group II (Sandmeier *et al.*, 1994). Their evolutionary relationships have recently been characterized (Moya-Garcia *et al.*, 2006). The search for specific new inhibitors able to discriminate between these requires a deep structural and functional knowledge, to detect relevant differences among their structures. Then chemical structures, or chemical modifications of previously known structures, that bind preferentially to only one of them (in our case, mammalian HDC) can be designed. Only the structure for pig DDC has been elucidated from the DDC group II; however, its high sequence identity (higher than 50%) with mammalian HDC (active fragment) allowed us to obtain a 3-D model of the latter using comparative modelling techniques (Baker and Sali, 2001). This model was experimentally validated by results obtained with more than 25 direct mutants that were assayed by different biophysical techniques (Fleming *et al.*, 2004b). The initial review of the structure–function relationship of mammalian HDC integrated all the previous information about this enzyme based on its structural characteristics (Moya-Garcia *et al.*, 2005). More recently, the decarboxylation reaction (the rate-limiting step for histamine synthesis) has been analysed by applying a combined strategy of quantum mechanics (QM) and molecular mechanics (MM) simulations on the external aldimine (PLP-histidine) complex located in the catalytic site of the enzyme (Moya-Garcia *et al.*, 2008). Therefore, the exact location of all residues involved in this reaction and their behaviour along the reaction is now known, facilitating the search for new potential inhibitory compounds for this reaction. All this previous information is highly valuable for the construction of *in silicio* experiments aimed at finding new drugs.

Today, the field of *in silicio* drug development is very attractive, active and fertile, but is still very new. Genomic and proteomic studies produce vast amounts of information, facilitating the identification of new therapeutically relevant targets, which allows the generation of libraries of compounds with rational chemical combinations. The technique called virtual screening (VS) uses computers to search databases of millions of compounds (already synthesized or not) for those chemical entities able to interact with a given target, thus able to interfere with its activity (Shoichet, 2004). These chemicals can then be tested against the target in order to obtain new candidates for a specific drug. In addition to the essential role played by the advances in experimental and theoretical fields, the incredible progress in computer technology has been decisive in our understanding of biological structures and the processes in which they are involved. Modelling unknown structures from bare sequences, long simulations of enzymes and complex multimeric structures, and large-scale VS experiments are now performed routinely

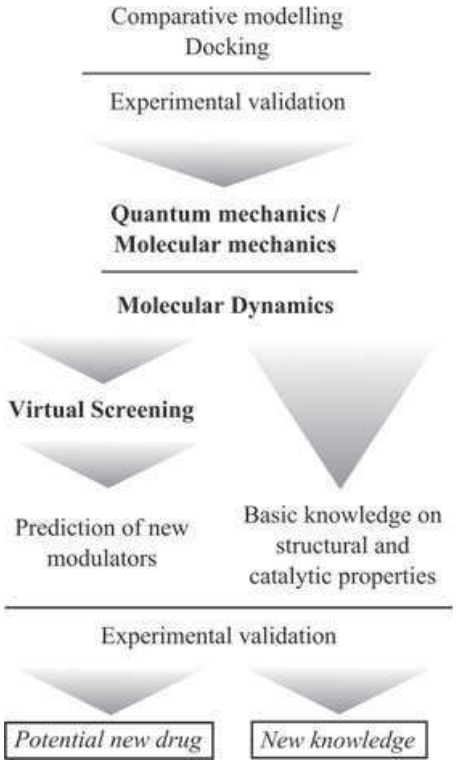


Figure 1 Sequential scheme of combined computer-based and experimental approaches followed in our studies on mammalian histidine decarboxylase over the last 5 years.

thanks to the availability of fast processors at modest prices. However, the expected revolution in rational drug discovery has not yet arrived, despite all these advances. The main limitations are the availability of reliable structural models for the target (having at hand a 3-D structure of the target in most of the cases is not enough) (Davis *et al.*, 2003) and the inclusion (at accurate levels) of some important effects such as the environment (solvent, ions, metal atoms, and so on) (Morreale *et al.*, 2007), entropic losses (of both conformation and configuration) (Carlsson and Aqvist, 2005) or the flexibility of the target (the most striking and elusive point) (Cozzini *et al.*, 2008).

In this review, we focus on computer-aided drug discovery in cases where the structure of the target has been obtained by means of comparative modelling, refined by simulation of molecular dynamics (MD), and finally, on inhibitors that have been found using rational drug discovery through VS experiments. A brief description of these methods and their application to HDC are presented in the following sections. Figure 1 shows the strategy scheme used in the case of mammalian HDC.

Comparative modelling versus experimental structure determination; the example of histidine decarboxylase

Knowing the ligand structure is a linear process, but things are much more complicated for the receptor. Protein structures are usually determined using powerful experimental techniques such as X-ray crystallography, cryoelectron microscopy and nuclear magnetic resonance (NMR). The development of these techniques, together with advances in protein expression and purification, microcrystallization (Abola *et al.*, 2000) and the use of synchrotron light (Sorensen *et al.*, 2006), has led to a separate discipline referred to as Structural Genomics (Chandonia and Brenner, 2006). Because of this development, the growth of the Protein Data Bank (PDB, i.e. the number of known protein structures) (Berman *et al.*, 2007), and the number of potential pharmacological targets, has been exponential.

Nevertheless, there is a growing gap between the number of known structures and sequences; that is, the number of newly discovered protein sequences grows faster. For example, over the last 4 years, the number of sequences in the comprehensive Swiss-Prot/TrEMBL database (Boutet *et al.*, 2007) increased by a factor of 5.44, while the number of protein structures deposited in the PDB increased by only a factor of 1.85. Therefore, the expanding field of Structural Genomics benefits from advances in computer methods for determining protein structure; comparative or homology modelling (Marti-Renom *et al.*, 2000; Xiang, 2006) attempts to bridge this sequence-structure gap.

The technique relies on the observation that, during evolution, protein structure is more stable and changes much more slowly than the underlying sequence, so similar sequences adopt practically identical structures, and distantly related sequences will fold into similar structures (Chothia and Lesk, 1986; Sander and Schneider, 1991). Thus, the unknown structure of a target protein (not to be mistaken with the drug target) can be inferred from the structure of a template protein if there is enough sequence homology between them. In order to obtain a reliable model, the threshold for sequence identity depends on the number of aligned residues, but is usually over 30%. It is important to stress that, in some cases, homology between proteins is not clear from pairwise methods such as sequence alignment. Profile-based methods can be more sensitive in homology detection; since the important factor in obtaining a reliable model is the existence of homology between target and template, the scope of comparative modelling methods can be extended, in some cases, to low sequence identities between target and template (Tramontano and Morea, 2003). Briefly, a homology modelling protocol is carried out in a few sequential steps, as described elsewhere (Marti-Renom *et al.*, 2000; Baker and Sali, 2001; Fiser and Sali, 2003; John and Sali, 2003): finding known structures related to the target sequence (templates), aligning the target sequence with the templates, building the model, and finally assessing and validating the model.

The applications of a protein structure model depend on its accuracy, which tends to decrease as the evolutionary distance between target and templates increases, so the target-template sequence identity is a good indication of the quality of a given model. Fortunately, a protein structure model does not have to be perfect to be helpful in biomedicine or biotechnology; however, the type of problem that can be tackled with a particular model does depend on its quality (Marti-Renom *et al.*, 2000), ranging from prediction of the approximate biochemical function (with models based on less than 30% sequence identity, at the low end of the accuracy

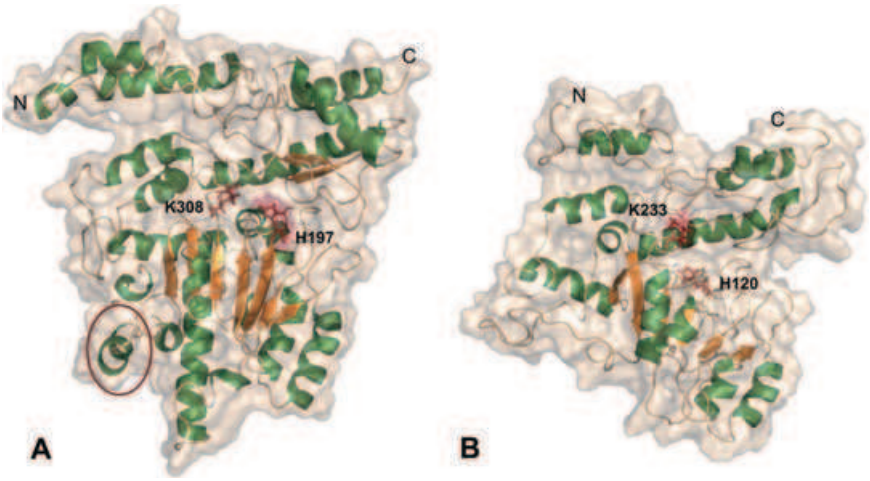


Figure 2 Structural models of both mammalian (A) and bacterial (B) PLP-dependent histidine decarboxylase (HDC) monomers. For orientation, amino(N)- and carboxy (C)-termini are shown. PLP-interacting residues K308 and H197 are depicted as dark red sticks with similar perspective. Flexible loop region of mammalian HDC, predicted to both interact and move during substrate binding, is surrounded by a dark red circle. As can be observed, the corresponding region in bacterial HDC is distorted, in agreement with the suspected differences between the enzymes in both the substrate binding surface and the quaternary structures (for further information see Moya-Garcia *et al.*, 2005; 2006). Images were made using PyMOL (DeLano, 2002).

spectrum), to predictions of important features in the target protein that do not occur in the template structure (with medium-resolution models). Moreover, the average quality of models at the highest end of the accuracy spectrum, those based on more than 50% sequence identity, is similar to that of low-resolution X-ray structures (Baker and Sali, 2001). The alignments on which these models are based contain almost no errors, so, among other applications, they can be used for structure-based drug discovery, small ligand docking and prediction of detailed ligand–protein interactions.

It is generally assumed that docking to comparative models is more challenging and less successful than docking to crystallographic structures. However, little work has been done to obtain quantitative information about the accuracy of docking to homology models, to determine in detail why the results are inferior to those obtained from experimentally determined structures. In many examples, protein homology models have supported the discovery of the optimum compounds for potency and selectivity (for a detailed review and examples see Hillisch *et al.*, 2004; Jacobson and Sali, 2004).

The PLP-dependent HDCs are good examples of proteins for which characterization by biophysical techniques has so far been impossible. The instability of the protein makes it inconceivable to reach that goal. However, comparative modelling, together with computer simulations (see below), is allowing us to gain insights into this molecular system, with a final aim of revealing new intervention strategies. In 2003, the first model of an active version of mammalian HDC was obtained (Rodriguez-Caso *et al.*, 2003a). The predicted structure was validated experimentally (Fleming *et al.*, 2004b); this allowed us to integrate the previous experimental information on the enzyme, obtaining new insights into the most promising targets at which to interfere with the protein's activity (the catalytic centre and dimerization surface). In fact, the prokaryotic and eukaryotic enzyme are homologous enzymes; however, from the comparative modelling analysis of both types of enzymes, it is clear that they do differ, mostly in residues located in the substrate binding sites and in the N-terminus of their respective monomers (Moya-Garcia *et al.*, 2006). A homology model of the Gram-negative *Klebsiella planticola* HDC was built, using as a template the structure of human PLP-dependent glutamate decarboxylase (GAD, EC 4.1.1.15), recently determined experimentally (Fenalti *et al.*, 2007). The two enzymes share a sequence identity of 24%. Figure 2 shows a comparison of the monomer of both enzymes. These deduced differences, together with other simulation-based structure-function studies, can help to identify inhibitors more selective for any of the homologous enzymes that can coexist in the human body (mammalian and enterobacterial HDCs and Dopa decarboxylases).

Molecular dynamics: a more realistic approach to protein function

In structure-based drug discovery, knowing the structure of the drug target is often not enough. Structure determines function, but it is not easy to deduce many of the activities or properties of a protein just from its structure. A protein cannot be merely reduced to its description as a rigid and static structure, as it is a dynamical reality with conformational fluctuations in time and substantial changes in the presence of ligands. Although the global structure is a key element in the function of a protein, its flexibility is an essential factor that modulates the relationship between structure and function. To go further, flexibility, understood as the capacity for conformational change in response to external stimuli, is part of the nature of all proteins and molecular systems. Thus, it is essential to understand how and why proteins change their conformations in order to be able to control and understand their biological functions. This dynamical nature needs to be considered in the study of the ligand–target interactions. Thus, for many applications,

such as VS experiments, it is advisable to represent the protein as an ensemble of different conformations that describe the inherent flexibility of the system, although there is not a clear consensus on how the ensemble should be represented. There are examples where many conformations were used (obtained from different X-ray or NMR structures or generated by computerized sampling techniques as Monte Carlo or MD) (Totrov and Abagyan, 2008), or where only one was used that comprised information of the entire ensemble (averaging energies or coordinates of many single conformations) (Osterberg *et al.*, 2002).

This blurry description is not enough to represent the protein flexibility properly, a time-dependent feature of macromolecular systems that is one of the most difficult yet essential to understand. MD is a powerful computer technique used to study flexibility (Hansson *et al.*, 2002; Norberg and Nilsson, 2003). It is valuable for understanding the dynamic behaviour of proteins at different timescales, from the fast internal fluctuations to the slower and more global movements that constitute conformational changes, or, eventually, the folding of a polypeptide into the native structure of a protein (Snow *et al.*, 2005). Furthermore, the explicit effects of solvent molecules and ions on the protein structure and stability can, and should, be taken into account to obtain accurate temporal averages of important structural and thermodynamic properties of the system under study, especially the binding energy, which are of utmost importance in the field of drug discovery.

Molecular dynamics simulates molecular time-dependent events in proteins and other biological macromolecules using the laws of classical mechanics. In particular, Newton's equations of motion are applied for an atomistic representation of a molecular system (balls for atoms and string for bonds) by employing MM force fields based on empirically deduced interaction potentials or derived from more complex quantum calculations. The energy of a molecular system within the force field approximation is a sum of different terms accounting for the distortion of the system as compared with an idealized structure where bonded (bond stretching, angle bending and torsions) and non-bonded (van der Waals and electrostatics) interaction terms are included. The main differences between the most widely used force fields [AMBER (Case *et al.*, 2004), GROMOS (van Gunsteren *et al.*, 1996), CHARMM (Brooks *et al.*, 1983), and so on] are due to parametrical issues and the functional form of the different terms entering in the force field equation. Although atomic charges are explicitly included, they remain constant over the simulation in most of the force fields (new polarized force fields are now emerging to ameliorate this drawback, see Xie *et al.*, 2008), precluding the use of MM force fields in systems undergoing chemical reactions. To model such changes adequately, as in processes such as bond-breaking/-forming, charge-transfer or electronic excitation, it is necessary to rely on the more accurate approximation obtained with QM. As noted above, it is also essential to introduce the effect of the environment. It has been demonstrated that significant changes can occur both in the biological structures and in the reactivity profile due to environmental influences. Therefore, in cases where chemical reactions need to be modelled while also taking into account environmental effects, it is essential to use a method that can account for both. Due to its high computer costs, the application of QM is still limited to relatively small systems consisting of up to tens or several hundreds of atoms, or even smaller systems when using higher levels of theory.

A solution will be a combined method able to treat the main atoms involved in the reaction with QM and the rest with MM. These methods will join the accuracy of the QM description with the low computer costs of MM; these are called hybrid methods (QM/MM) and have become very popular (Warshel, 2003). These methods are being used in the study of reactions of biological interest (Garcia-Viloca *et al.*, 2004; Marti *et al.*, 2004) and it has been demonstrated they can be used to identify key residues in catalysis (Ridder *et al.*, 2003), resolve mechanistic questions and verify the fundamental principles of catalysis (Marti *et al.*, 2004). Potential contributions, obtained from this modelling of enzyme reactions, to drug discovery have recently been reviewed by Raha *et al.* (2007) and Mulholland (2005). These include the identification of key catalytic residues and the reaction mechanism leading to the identification of transition states and other intermediates, the prediction of drug metabolism and the accurate calculation of the free energy of binding.

Our group has applied simulation techniques and MD techniques, by using the hybrid methodology QM/MM, to unravel the basis of the mammalian HDC catalytic mechanism (Moya-Garcia *et al.*, 2008). In this study, we examined the decarboxylation of the intermediate cofactor-substrate adduct (the external aldimine) in the enzymatic environment (catalysed reaction) and in an aqueous environment. In each case, the reaction environment was explicitly considered and the energy used for each process was calculated. From a comparison of the reactions in the two conditions, we obtained the differentiating elements that explain the catalysis by mammalian HDC. We consider this extensively evaluated computer model of the mammalian HDC structure, in its cofactor-substrate adduct bound state, to be the first step towards performing high-throughput screening *in silicio*.

Virtual screening techniques: searching for new molecules

The type of strategy employed in VS depends on the structural information that is readily available, and can be performed even if the structure of the target is not known [using pharmacophores (Sun, 2008), similarity techniques applied to known active ligands (Bajorath, 2001), and so on]. The most favourable case is when both of the structures (the target and the ligand) are known. Here, docking-based techniques are very promising, although far from being completely successful (Warren *et al.*, 2006). In docking, the problem is to find out of the many possible ways a ligand can be positioned within a binding site, the appropriate one that triggers/inhibits the biological activity of the target. To discern which position is the best, each of them is scored according to their reaction with the target. This is done by means of a mathematical function, the *scoring function*, built to capture the essential events that occur when a ligand binds to a target (Warren *et al.*, 2006). Much of the uncertainty (although not

all) in docking and scoring protocols has its roots in the definition of these functions. VS is an extension of the docking procedure, actually performed with a small number of molecules, to handle millions of them. In this case, the objective is somehow different than that of docking: success is achieved if we are good at separating the active compounds from the inactive ones, and if, at least, some of the active compounds are found at the top of a list based on the abovementioned scoring function. The explosion in the use of VS in the last decade can be understood when it is considered that although more money is invested in R&D projects, there are fewer newly discovered drugs reaching the market (Smith, 2002). This fact has fuelled the development of many different docking algorithms and sophisticated scoring functions (Sousa *et al.*, 2006).

A VS protocol is a sequence of filters that increase in complexity to reduce the number of molecules subjected to experimental assays to a tractable amount. It is customary to start with molecules fulfilling Lipinski's rule of five (Lipinski *et al.*, 2001) and possibly imposing some other constraints, such as adequate solubility or certain kind of chemotypes (focalized libraries). Nevertheless, brute force approximation can be also employed, based on the idea that the more molecules you can test, the higher the probability of finding promising candidates. An increasing number of databases are available to start with, for example, ZINC, with over 8 million compounds available (Irwin and Shoichet, 2005). Other databases such as DUD are useful for testing a VS protocol prior to undertaking a search for new molecules, so one can assess the performance of the protocol to decide if it is appropriate for a particular problem (Huang *et al.*, 2006).

To facilitate the choice of a particular protocol, or to compare different protocols, we have developed VSDMIP (Virtual Screening Data Management on an Integrated Platform) (Gil-Redondo *et al.*, 2009), a flexible fully automated computer platform that combines all the steps needed to generate a short list of candidates from a database of 2-D molecular structures. In brief, the VSDMIP protocol consists of (i) a database; (ii) a library of service interfaces and plugins; and (iii) a set of workflows and implementing commands. All of the data and VS results from small molecules (ligands) are stored within VSDMIP. The user controls the platform through different command line utilities and configures it using XML files. VSDMIP currently runs on Linux/x86 platforms and has been successfully implemented in the MareNostrum supercomputer at the Barcelona Supercomputing Center (BSC), making it possible to screen 4 million compounds (the actual size of our molecular database) in less than 1 month.

In general, the steps needed to initiate a VS study can be broadly divided into the preparation of the target and the ligands. (1) For the target, starting from its 3-D structure, we (i) choose only the domains surrounding the active site, (ii) add missing loops and atoms (especially if they are close to the binding site), (iii) add hydrogen atoms, (iv) assign atom types and atomic charges, and finally (v) characterize the active site. This last step entails limiting the active site by a box (where subsequent docking will be performed) and making the space covered by the box discrete with grid points spanning the three-dimensional space. Each grid point contains information on the interaction of an atomic probe atom (representing common atom types in pharmacologically relevant molecules), including the electrostatic interaction and the possibility of forming hydrogen bonds between the ligand and the residues at the binding site. (2) For the ligands, we (i) start with a 2-D topological representation of the structures (SMILES string, see Weininger, 1988) in order to avoid bias related to the conformations, (ii) transform them into 3-D (adding hydrogen atoms and generating tautomers, steroisomers, and different ring conformations if necessary), (iii) assign atomic radii and charges, and (iv) perform conformational analysis with ALFA (Gil-Redondo, 2006).

Each ligand is then docked into the target active site with CDOCK (Perez and Ortiz, 2001), our docking software, which includes a movement/evaluation/refinement strategy: (i) translate and rotate the ligand in each grid point; (ii) evaluate the energy for each configuration generated; and (iii) refine the best configurations generated using a rigid body SIMPLEX optimization program (Nelder and Mead, 1965). Finally, the best configuration of all is taken as the docking result. The scoring function implemented in CDOCK accounts for van der Waals and electrostatic forces, as well as hydrogen–bond interactions. It also includes a solvation correction term based on an implicit model (Morreale *et al.*, 2007). Within VSDMIP, the docking step can be preceded by a docking method such as DOCK (Kuntz *et al.*, 1982) or FRED (OpenEye Scientific Software, Inc.), configurated in a less accurate but faster way, or it can be replaced by Autodock (Morris *et al.*, 1998).

Representing the protein as a grid imposes the rigidity constraint into the docking calculations, which is one of the main drawbacks in computer-aided drug discovery based on structure. To overcome this drawback, the best molecules classified are submitted to a short MD simulation in an explicit solvent. MM-GBSA analysis is performed on selected snapshots to obtain an estimate of the free energy of binding (entropy not included) for each compound (Massova and Kollman, 2000). This is the value employed for the definitive classification. A visual inspection of each of the final best-ranked candidates is always mandatory.

The first attempt of drug discovery based on mammalian HDC structure

As far as we know, there has been no attempt to perform high-throughput screening on mammalian HDC in order to find new inhibitors with a potential pharmacological use, although there is interest in the characterization of this enzyme as a pharmacological target. The HDC inhibitors known to date are substrate analogues and were developed in the 1970s (DeGraw *et al.*, 1977). Recently a new strategy for developing new inhibitors based on the external aldimine, which is the common intermediate of the transformation of all amino acids catalysed by PLP-dependent enzymes, has been reported (Wu *et al.*, 2008). However, the authors do not use computer modelling to guide their inhibitor design rationale. They try to elucidate the structure–activity relationship of their synthesized compounds based on a rough computer

model of the active site of human HDC, together with the presumed intracellular form of the compound with the highest inhibition rate, namely pyridoxyl-histidine methyl ester.

Docking of known inhibitors; validation of the model

We were able to build a high-quality structural model for HDC. It shows the relevant structural features and reliably reproduces the behaviour of the enzyme. We were able to reveal key amino acids for the activity and stability of HDC (Rodriguez-Caso *et al.*, 2003a; Fleming *et al.*, 2004b) and we discern particular features of the reaction mechanism, with full atomic details (Moya-Garcia *et al.*, 2008). Nevertheless, we submitted our model to an additional validation test to check whether it can be used to discover new HDC inhibitors with potential pharmacological use. We used the above-mentioned VSDMIP with the natural substrate histidine and the two well-known inhibitors α-FMH and HME. A full standard VS protocol was followed.

Characterization of the receptor was carried out as explained above. The active site pocket was determined by a 5 Å box centred at histidine in its external aldimine conformation and using interaction grids of 0.5 Å spacing. Once the active site was demarcated, the set of test ligands was docked using CDOCK. The best results from the docking process were obtained from the database and visualized in PyMol (DeLano, 2002). The coordinates and topology files of the receptor-ligand complexes were then generated using the program LEaP of the AMBER Molecular Dynamics package (Case *et al.*, 2005). PLP and ligand parameters were obtained with Antechamber module. The systems were solvated, neutralized and submitted to a common protocol of energy minimization and MD simulation using the program sander from AMBER package. A 2 ns production stage at a temperature of 300 K was followed by a cooling process, in which the temperature was decreased gradually from 300 to 292 K.

The minimum energy configuration for each ligand with the HDC active site, according to the scoring function implemented in CDOCK, were obtained and checked visually. All of them showed similar total binding energy values, ranging from -18 kcal·mol^{-1} (for both HME and histidine) to -11 kcal·mol^{-1} for FMH. These energy values together with those obtained from the previous VS carried out using a larger number of compounds, showed that we are dealing with a closed active site, which is in agreement with the structural information derived from our HDC homology model and other PLP-dependent decarboxylases structures, whose active sites are frequently buried in the dimerization surface. Our calculations, for the pathway connecting the active site with the outside solvent, show that the substrate needs to pass over a channel of about 40 Å in length inside the enzyme to reach the active site.

The best configurations docked for histidine and the two inhibitors computed show a satisfactory fit to the active site, with the proper orientation to proceed with the transaldimination reaction and form the external aldimine with the cofactor, as can be seen in Figure 3. Both PLP and ligand were able to move freely in the active site during the MD simulation since no local restrictions were imposed, resulting in a slight separation between them. As a result of previous VS studies, where new inhibitors were found and successfully tested, the structures of receptor-inhibitor complexes were then determined by means of X-ray diffraction in order to check the similarity of the configuration predicted after docking and the validation protocol with the one obtained by crystallography (Warren *et al.*, 2006). In only a few cases, VS-derived complexes resembled those structures obtained by experimental means, indicating that the computer-derived approaches are not yet capable of giving us a perfect image of the active site of an enzyme when binding any ligand, but they are reasonably accurate for determining the accommodation and stability of those compounds attracted to the active site pocket.

Preliminary results of the docking process in our VS study over the ZINC 7 compound database (Irwin and Shoichet, 2005), which comprises 4 million molecules, showed the method tended to fail when trying to fit compounds with a large number of atoms. This would significantly reduce the number of suitable candidates as potential inhibitors of HDC activity. These results are in agreement with those observed by Wu *et al.* (2008).

On the other hand, those compounds identified with suitable configurations after docking are arranged in the active

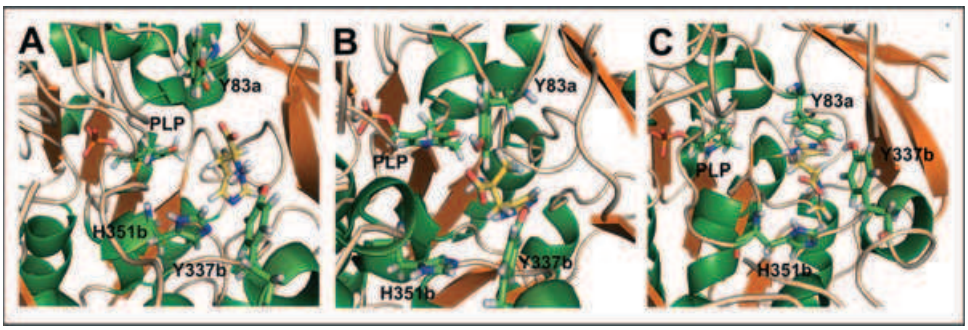


Figure 3 Final configurations of the docked histidine substrate (A), αFMH (B) and HME (C) into the HDC active site (depicted in yellow sticks) after screening with VSDMIP and molecular dynamics simulation. Residues previously described to be located in the proximity of the natural substrate and the cofactor pyridoxal-5′-phosphate (PLP) are shown as green sticks. Suffixes 'a' and 'b' indicate the residue-containing monomers, monomer 'a' being the one that binds PLP.

site as they can make interactions with key residues involved in stabilization of the substrate (Moya-Garcia *et al.*, 2008). Y83 and Y337, which have been determined as important residues in favouring the reception of the ligand into the active site (Rodriguez-Caso *et al.*, 2003a; Fleming *et al.*, 2004b), as well as H351 and H197, are located in the proximity of our best-docked compounds. Further refinement of these preliminary results by means of MD simulations will consolidate these interactions or even reveal new ones that were not obvious just after the docking process.

Discussion and conclusions

Due to the pleiotropic effects of histamine, and on the basis of results obtained with HDC knockout mice (Jorgensen *et al.*, 2007; Haas *et al.*, 2008; Ohtsu, 2008; Schubert and Peura, 2008), it is possible that selective, direct inhibition of HDC could have many different secondary effects. To be able to control the production of histamine by this method, either locally (for instance, topical use) or at the systemic level, rather than just interfering with the reception of the amine on target cells, could have important therapeutic consequences in physiopathological situations where either the local or circulating levels of histamine are excessive due to abnormal histamine production. Also, given the important roles of histamine in the central nervous system (Wijtmans *et al.*, 2008; Zhao *et al.*, 2008), special attention should be paid to the ability of any HDC inhibitor (or its derivatives) to cross the blood-brain barrier, and this should be evaluated by experimental and/or *in silicio* approaches (Kortagere *et al.*, 2008; Malakoutikhah *et al.*, 2008). In addition, therapies combining both HDC inhibitors and histamine receptor agonists/antagonists should not be ruled out.

This review presents an example of not only a potentially interesting protein for pharmacology, but also a drug target that has been very difficult to characterize by experimental approaches and, consequently, to use efficiently for drug discovery. By changing the strategy, that is by combining *in silicio* and experimental techniques, the structural and catalytic properties of HDC are now known and this knowledge can be used to discover potential, new antihistamine drugs. In addition, this strategy can be applied to many other proteins related to amine metabolism, immunology and drug discovery in general, to solve other pending problems in biomedicine, biotechnology and pharmacology. From an economical point of view, it is obvious that this strategy would also be convenient for the pharmacological industry, since the *in silicio* approach can save significant investment in experimental protein chemistry techniques and high-throughput screening protocols.

Acknowledgements

The CIBER de Enfermedades Raras is an initiative of the ISCIII. This work was supported by Grant SAF2008-02522, Ministerio de Ciencia e Innovación Work at the CBM-SO was partially supported by a grant from 'Comunidad de Madrid' thorough BIPEDD project (SBIO-0214–2006). We also acknowledge the generous allocation of computer time at the BSC.

Conflict of interest

The authors state no conflict of interest.

References

Abola E, Kuhn P, Earnest T, Stevens RC (2000). Automation of X-ray crystallography. *Nat Struct Biol* **7** (Suppl.): 973–977.

Ai W, Zheng H, Yang X, Liu Y, Wang TC (2007). Tip60 functions as a potential corepressor of KLF4 in regulation of HDC promoter activity. *Nucleic Acids Res* **35**: 6137–6149.

Bajorath J (2001). Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J Chem Inf Comput Sci* **41**: 233–245.

Baker D, Sali A (2001). Protein structure prediction and structural genomics. *Science* **294**: 93–96.

Berman H, Henrick K, Nakamura H, Markley JL (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* **35** (Database issue): D301–D303.

Bertoldi M, Gonsalvi M, Voltattorni CB (2001). Green tea polyphenols: novel irreversible inhibitors of dopa decarboxylase. *Biochem Biophys Res Commun* **284**: 90–93.

Bhattacharjee MK, Snell EE (1990). Pyridoxal 5'-phosphate-dependent histidine decarboxylase. Mechanism of inactivation by alpha-fluoromethylhistidine. *J Biol Chem* **265**: 6664–6668.

Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A (2007). UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase. *Methods Mol Biol* **406**: 89–112.

Brooks BR, Bruccoleri RE, Olafson DJ, States DJ, Swaminathan S, Karplus M (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem* **4**: 187–217.

Carlsson J, Aqvist J (2005). Absolute and relative entropies from computer simulation with applications to ligand binding. *J Phys Chem* **109**: 6448–6456.

Case DA, Darden TA, Cheatham TE, Simmerling CL, Wang J, Duke RE *et al.* (2004). *AMBER* **8**. University of California: San Francisco.

Case DA, Cheatham TE III, Darden T, Gohlke H, Luo R, Merz KM Jr *et al.* (2005). The Amber biomolecular simulation programs. *J Comput Chem* **26**: 1668–1688.

Chandonia JM, Brenner SE (2006). The impact of structural genomics: expectations and outcomes. *Science* **311**: 347–351.

Chen D, Aihara T, Zhao CM, Hakanson R, Okabe S (2006). Differentiation of the gastric mucosa. I. Role of histamine in control of function and integrity of oxyntic mucosa: understanding gastric physiology through disruption of targeted genes. *Am J Physiol Gastrointest Liver Physiol* **291**: G539–G544.

Chothia C, Lesk AM (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J* **5**: 823–826.

Cozzini P, Kellogg GE, Spyrakis F, Abraham DJ, Costantino G, Emerson A *et al.* (2008). Target flexibility: an emerging consideration in drug discovery and design. *J Med Chem* **51**: 6237–6255.

Davis AM, Teague SJ, Kleywegt GJ (2003). Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew Chem Int Ed Engl* **42**: 2718–2736.

DeGraw JI, Engstrom J, Ellis M, Johnson HL (1977). Potential histidine decarboxylase inhibitors. 1. α- and β-Substituted histidine analogs. *J Med Chem* **20**: 1671–1674.

DeLano WL (2002). *The PyMOL Molecular Graphics System*. DeLano Scientific: San Carlos, CA.

Engel N, Olmo MT, Coleman CS, Medina MA, Pegg AE, Sanchez-

Jimenez F (1996). Experimental evidence for structure-activity features in common between mammalian histidine decarboxylase and ornithine decarboxylase. *Biochem J* **320** (Pt 2): 365–368.

Fenalti G, Law RH, Buckle AM, Langendorf C, Tuck K, Rosado CJ *et al*. (2007). GABA production by glutamic acid decarboxylase is regulated by a dynamic catalytic loop. *Nat Struct Mol Biol* **14**: 280–286.

Fiser A, Sali A (2003). Modeller: generation and refinement of homology-based protein structure models. *Methods Enzymol* **374**: 461–491.

Fitz LJ, Brennan A, Wood CR, Goldman SJ, Kasaian MT (2008). Activation-induced cellular accumulation of histamine in immature but not mature murine mast cells. *Cell Mol Life Sci* **65**: 1585–1595.

Fleming JV, Fajardo I, Langlois MR, Sanchez-Jimenez F, Wang TC (2004a). The C-terminus of rat L-histidine decarboxylase specifically inhibits enzymic activity and disrupts pyridoxal phosphate-dependent interactions with L-histidine substrate analogues. *Biochem J* **381** (Pt 3): 769–778.

Fleming JV, Sanchez-Jimenez F, Moya-Garcia AA, Langlois MR, Wang TC (2004b). Mapping of catalytically important residues in the rat L-histidine decarboxylase enzyme using bioinformatic and site-directed mutagenesis approaches. *Biochem J* **379** (Pt 2): 253–261.

Furuta K, Nakayama K, Sugimoto Y, Ichikawa A, Tanaka S (2007). Activation of histidine decarboxylase through post-translational cleavage by caspase-9 in a mouse mastocytoma P-815. *J Biol Chem* **282**: 13438–13446.

Garcia-Viloca M, Gao J, Karplus M, Truhlar DG (2004). How enzymes work: analysis by modern rate theory and computer simulations. *Science* **303**: 186–195.

Gil-Redondo R (2006). *Master Thesis: Implementación de una plataforma para el cribado virtual de quimiotecas*. UNED edn: Madrid.

Gil-Redondo R, Estrada J, Morreale A, Herranz F, Sancho J, Ortiz AR (2009). VSDMIP: virtual screening data management on an integrated platform. *J Comput Aided Mol Des* **23**: 171–184.

van Gunsteren WF, Billeter SR, Eising AA, Hünenberger PH, Krüger P, Mark AE *et al*. (1996). *Biomolecular Simulation: The GROMOS96 Manual and User Guide*. Vdf Hochschulverlag: Zürich.

Gurevich VV, Gurevich EV (2008). Rich tapestry of G protein-coupled receptor signaling and regulatory mechanisms. *Mol Pharmacol* **74**: 312–316.

Haas HL, Sergeeva OA, Selbach O (2008). Histamine in the nervous system. *Physiol Rev* **88**: 1183–1241.

Haavik J, Blau N, Thony B (2008). Mutations in human monoamine-related neurotransmitter pathway genes. *Hum Mutat* **29**: 891–902.

Hansson T, Oostenbrink C, van Gunsteren W (2002). Molecular dynamics simulations. *Curr Opin Struct Biol* **12**: 190–196.

Hillisch A, Pineda LF, Hilgenfeld R (2004). Utility of homology models in the drug discovery process. *Drug Discov Today* **9**: 659–669.

Huang N, Shoichet BK, Irwin JJ (2006). Benchmarking sets for molecular docking. *J Med Chem* **49**: 6789–6801.

Irwin JJ, Shoichet BK (2005). ZINC – a free database of commercially available compounds for virtual screening. *J Chem Inf Model* **45**: 177–182.

Jacobson M, Sali A (2004). Comparative protein structure modeling and its applications to drug discovery. *Annu Rep Med Chem* **39**: 259–276.

John B, Sali A (2003). Comparative protein structure modeling by iterative alignment, model building and model assessment. *Nucleic Acids Res* **31**: 3982–3992.

Jorgensen EA, Knigge U, Warberg J, Kjaer A (2007). Histamine and the regulation of body weight. *Neuroendocrinology* **86**: 210–214.

Kortagere S, Chekmarev D, Welsh WJ, Ekins S (2008). New predictive models for blood-brain barrier permeability of drug-like molecules. *Pharm Res* **25**: 1836–1845.

Kuntz ID, Blaney JM, Oatley SJ, Landridge R, Ferrin TE (1982). A geometric approach to macromolecule – ligand interactions. *J Mol Biol* **161**: 269–288.

Kuramasu A, Sukegawa J, Yanagisawa T, Yanai K (2006). Recent advances in molecular pharmacology of the histamine systems: roles of C-terminal tails of histamine receptors. *J Pharmacol Sci* **101**: 7–11.

Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **46**: 3–26.

Malakoutikhah M, Teixido M, Giralt E (2008). Toward an optimal blood-brain barrier shuttle by synthesis and evaluation of peptide libraries. *J Med Chem* **51**: 4881–4889.

Marti S, Roca M, Andres J, Moliner V, Silla E, Tunon I *et al*. (2004). Theoretical insights in enzyme catalysis. *Chem Soc Rev* **33**: 98–107.

Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A (2000). Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* **29**: 291–325.

Massova I, Kollman PA (2000). Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspect Drug Discov Des* **200**: 113–135.

Medina MA, Urdiales JL, Rodriguez-Caso C, Ramirez FJ, Sanchez-Jimenez F (2003). Biogenic amines and polyamines: similar biochemistry for different physiological missions and biomedical applications. *Crit Rev Biochem Mol Biol* **38**: 23–59.

Medina MA, Correa-Fiz F, Rodriguez-Caso C, Sanchez-Jimenez F (2005). A comprehensive view of polyamine and histamine metabolism to the light of new technologies. *J Cell Mol Med* **9**: 854–864.

Melgarejo E, Medina MA, Sanchez-Jimenez F, Botana LM, Dominguez M, Escribano L *et al*. (2007). Epigallocatechin-3-gallate interferes with mast cell adhesiveness, migration and its potential to recruit monocytes. *Cell Mol Life Sci* **64**: 2690–2701.

Morreale A, Gil-Redondo R, Ortiz AR (2007). A new implicit solvent model for protein-ligand docking. *Proteins* **67**: 606–616.

Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK *et al*. (1998). Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J Comput Chem* **19**: 1639–1662.

Moya-Garcia AA, Medina MA, Sanchez-Jimenez F (2005). Mammalian histidine decarboxylase: from structure to function. *Bioessays* **27**: 57–63.

Moya-Garcia AA, Pino-Angeles A, Sanchez-Jimenez F (2006). New structural insights to help in the search for selective inhibitors of mammalian pyridoxal 5′-phosphate-dependent histidine decarboxylase. 4. Synthesis, metabolism and release of histamine. *Inflamm Res* **55** (Suppl. 1): S55–S56.

Moya-Garcia AA, Ruiz-Pernia J, Marti S, Sanchez-Jimenez F, Tunon I (2008). Analysis of the decarboxylation step in mammalian histidine decarboxylase. A computational study. *J Biol Chem* **283**: 12393–12401.

Mulholland AJ (2005). Modelling enzyme reaction mechanisms, specificity and catalysis. *Drug Discov Today* **10**: 1393–1402.

Nelder JA, Mead R (1965). A simplex method for function minimization. *Comput J* **7**: 308–313.

Nitta Y, Kikuzaki H, Ueno H (2007). Food components inhibiting recombinant human histidine decarboxylase activity. *J Agric Food Chem* **55**: 299–304.

Norberg J, Nilsson L (2003). Advances in biomolecular simulations: methodology and recent applications. *Q Rev Biophys* **36**: 257–306.

Ohtsu H (2008). Progress in allergy signal research on mast cells: the role of histamine in immunological and cardiovascular disease and the transporting system of histamine in the cell. *J Pharmacol Sci* **106**: 347–353.

Olmo MT, Rodriguez-Agudo D, Medina MA, Sanchez-Jimenez F (1999). The pest regions containing C-termini of mammalian ornithine decarboxylase and histidine decarboxylase play different roles in protein degradation. *Biochem Biophys Res Commun* **257**: 269–272.

Olmo MT, Urdiales JL, Pegg AE, Medina MA, Sanchez-Jimenez F

(2000). *In vitro* study of proteolytic degradation of rat histidine decarboxylase. *Eur J Biochem* **267**: 1527–1531.

Olmo MT, Sanchez-Jimenez F, Medina MA, Hayashi H (2002). Spectroscopic analysis of recombinant rat histidine decarboxylase. *J Biochem* **132**: 433–439.

Osterberg F, Morris GM, Sanner MF, Olson AJ, Goodsell DS (2002). Automated docking to multiple target structures: incorporation of protein mobility and structural water heterogeneity in AutoDock. *Proteins* **46**: 34–40.

Perez C, Ortiz AR (2001). Evaluation of docking functions for protein-ligand docking. *J Med Chem* **44**: 3768–3785.

Raha K, Peters MB, Wang B, Yu N, Wollacott AM, Westerhoff LM *et al*. (2007). The role of quantum mechanics in structure-based drug design. *Drug Discov Today* **12**: 725–731.

Ridder L, Harvey JN, Rietjens IMCM, Vervoort J, Mulholland AJ (2003). Ab initio qm/mm modeling of the hydroxylation step in p-hydroxybenzoate hydroxylase. *J Phys Chem B* **107**: 2118–2126.

Rodriguez-Agudo D, Olmo MT, Sanchez-Jimenez F, Medina MA (2000). Rat histidine decarboxylase is a substrate for m-calpain *in vitro*. *Biochem Biophys Res Commun* **271**: 777–781.

Rodriguez-Caso C, Rodriguez-Agudo D, Moya-Garcia AA, Fajardo I, Medina MA, Subramaniam V *et al*. (2003a). Local changes in the catalytic site of mammalian histidine decarboxylase can affect its global conformation and stability. *Eur J Biochem* **270**: 4376–4387.

Rodriguez-Caso C, Rodriguez-Agudo D, Sanchez-Jimenez F, Medina MA (2003b). Green tea epigallocatechin-3-gallate is an inhibitor of mammalian histidine decarboxylase. *Cell Mol Life Sci* **60**: 1760–1763.

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

Sandmeier E, Hale TI, Christen P (1994). Multiple evolutionary origin of pyridoxal-5′-phosphate-dependent amino acid decarboxylases. *Eur J Biochem* **221**: 997–1002.

Schubert ML, Peura DA (2008). Control of gastric acid secretion in health and disease. *Gastroenterology* **134**: 1842–1860.

Shoichet BK (2004). Virtual screening of chemical libraries. *Nature* **432**: 862–865.

Smith A (2002). Screening for drug discovery: the leading question. *Nature* **418**: 453–459.

Snow CD, Sorin EJ, Rhee YM, Pande VS (2005). How well can simulation predict protein folding kinetics and thermodynamics? *Annu Rev Biophys Biomol Struct* **34**: 43–69.

Sorensen TL, McAuley KE, Flaig R, Duke EM (2006). New light for science: synchrotron radiation in structural medicine. *Trends Biotechnol* **24**: 500–508.

Sousa SF, Fernandes PA, Ramos MJ (2006). Protein-ligand docking: current status and future challenges. *Proteins* **65**: 15–26.

Sun H (2008). Pharmacophore-based virtual screening. *Curr Med Chem* **15**: 1018–1024.

Tachibana M, Wada K, Katayama K, Kamisaki Y, Maeyama K, Kadowaki T *et al*. (2008). Activation of peroxisome proliferator-activated receptor gamma suppresses mast cell maturation involved in allergic diseases. *Allergy* **63**: 1136–1147.

Tanaka S, Ichikawa A (2006). Recent advances in molecular pharmacology of the histamine systems: immune regulatory roles of histamine produced by leukocytes. *J Pharmacol Sci* **101**: 19–23.

Thurmond RL, Gelfand EW, Dunford PJ (2008). The role of histamine H1 and H4 receptors in allergic inflammation: the search for new antihistamines. *Nat Rev Drug Discov* **7**: 41–53.

Totrov M, Abagyan R (2008). Flexible ligand docking to multiple receptor conformations: a practical alternative. *Curr Opin Struct Biol* **18**: 178–184.

Tramontano A, Morea V (2003). Exploiting evolutionary relationships for predicting protein structures. *Biotechnol Bioeng* **84**: 756–762.

Viguera E, Trelles O, Urdiales JL, Matés JM, Sánchez-Jiménez F (1994). Mammalian L-amino acid decarboxylases producing 1,4-diamines: analogies among differences. *Trends Biochem Sci* **19**: 318–319.

Warren GL, Andrews CW, Capelli AM, Clarke B, LaLonde J, Lambert MH *et al*. (2006). A critical assessment of docking programs and scoring functions. *J Med Chem* **49**: 5912–5931.

Warshel A (2003). Computer simulations of enzyme catalysis: methods, progress, and insights. *Annu Rev Biophys Biomol Struct* **32**: 425–443.

Weininger D (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Model* **28**: 31–36.

Wijtmans M, Celanire S, Snip E, Gillard MR, Gelens E, Collart PP *et al*. (2008). 4-benzyl-1H-imidazoles with oxazoline termini as histamine H3 receptor agonists. *J Med Chem* **51**: 2944–2953.

Wu F, Yu J, Gehring H (2008). Inhibitory and structural studies of novel coenzyme-substrate analogs of human histidine decarboxylase. *FASEB J* **22**: 890–897.

Xiang Z (2006). Advances in homology protein structure modeling. *Curr Protein Pept Sci* **7**: 217–227.

Xie W, Song L, Truhlar DG, Gao J (2008). The variational explicit polarization potential and analytical first derivative of energy: towards a next generation force field. *J Chem Phys* **128**: 234108.

Zhao C, Sun M, Bennani YL, Gopalakrishnan SM, Witte DG, Miller TR *et al*. (2008). The alkaloid conessine and analogues as potent histamine H3 receptor antagonists. *J Med Chem* **51**: 5423–5430.

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

gCOMBINE: A graphical user interface to perform structure-based comparative binding energy (COMBINE) analysis on a set of ligand-receptor complexes

Rubén Gil-Redondo,[1] Javier Klett,[1] Federico Gago,[2] and Antonio Morreale[1]*

[1] Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa, Campus UAM, c/Nicolás Cabrera 1, Madrid 28049, Spain

[2] Departamento de Farmacología, Universidad de Alcalá, Alcalá de Henares, Madrid 28871, Spain

ABSTRACT

We present gCOMBINE, a Java-written graphical user interface (GUI) for performing comparative binding energy (COMBINE) analysis (Ortiz et al. J Med Chem 1995; 38:2681–2691) on a set of ligand-receptor complexes with the aim of deriving highly informative quantitative structure-activity relationships. The essence of the method is to decompose the ligand-receptor interaction energies into a series of terms, explore the origins of the variance within the set using Principal Component Analysis, and then assign weights to selected ligand-residue interactions using partial least squares analysis to correlate with the experimental activities or binding affinities. The GUI allows plenty of interactivity and provides multiple plots representing the energy descriptors entering the analysis, scores, loadings, experimental versus predicted regression lines, and the evolution of parameters such as r^2 (correlation coefficient), q^2 (cross-validated r^2), and prediction errors as the number of extracted latent variables increases. Other representative features include the implementation of a sigmoidal dielectric function for electrostatic energy calculations, alternative cross-validation procedures (leave-N-out and random groups), drawing of confidence ellipses, and the possibility to carry out several additional tasks such as optional truncation of positive interaction energy values and generation of ready-to-use PDB files containing information related to the importance for activity of individual protein residues. This information can be displayed and color-coded using a standard molecular graphics program such as PyMOL. It is expected that this user-friendly tool will expand the applicability of the COMBINE analysis method and encourage more groups to use it in their drug design research programs.

Proteins 2010; 78:162–172.

Key words: 3D-QSAR; comparative binding energy analysis; drug design.

INTRODUCTION

Accurate prediction of biological activities for newly designed molecules is one of the greatest challenges faced in computer-aided drug research. Because the number of three-dimensional structures for pharmacologically relevant targets is continually increasing, the pioneering quantitative structure-activity relationship (QSAR) methodologies that rely solely on physico-chemical parameters of substituents in congeneric series of compounds[1] or on molecular interaction fields (MIF) calculated at discrete points in a three-dimensional (3D) lattice embedding the spatially aligned compounds, as in the popular comparative molecular field analysis (CoMFA)[2] and comparative molecular similarity indices analysis (CoMSIA),[3] have given way to other computational methods that attempt to derive as much information as possible from the structures of the ligand-receptor complexes. In this regard, continuous advances in the fields of comparative (homology) modeling of proteins of unknown experimental structure[4] and automated ligand docking,[5] as well as in computer power, have made it possible to search for new putative ligands for many different targets out of pools containing millions of candidate compounds (chemical libraries).[6] Although this "virtual screening" approach has met with some success the technique is far from being mature because of several reasons, the most important one possibly being that the scoring functions used to rank the molecules and prioritize the possible hits are not accurate enough. When the number of compounds for a given target is kept under a few hundred, however, and biological activities are known, as is usually the case in a typical medicinal chemistry project, it

The authors state no conflict of interest.
Grant sponsor: Comunidad de Madrid; Grant number: SBIO-0214-2006; Grant sponsor: Ministerio de Ciencia e Innovación (Programa Personal Técnicos de Apoyo 2008)
*Correspondence to: Antonio Morreale Unidad de Bioinformática, Centro de Biología Molecular Severo Ochoa, Campus UAM, c/ Nicolás Cabrera 1, Madrid 28049, Spain.
E-mail: amorreale@cbm.uam.es
Received 8 April 2009; Revised 19 June 2009; Accepted 10 July 2009
Published online 20 July 2009 in Wiley InterScience (www.interscience.wiley.com).
DOI: 10.1002/prot.22543

should be feasible to employ a highly precise energy function to describe the ligand-receptor interactions and then derive an accurate QSAR with predictive ability using a method such as comparative binding energy (COMBINE) analysis.[7]

COMBINE employs a number of residue-based interaction energies (both van der Waals and electrostatic) computed on a set of refined ligand-receptor complexes (rather than MIF calculated on a 3D grid for a set of unbound superimposed molecules, as is done in CoMFA) to build a data matrix that is then subjected to multivariate statistical analysis. The key idea is that partial least squares (PLS) can be used to correlate the computed energy components (plus additional optional terms such as receptor and ligand desolvation energies) with the experimental activities using an expression of the form shown in Eq. (1):

$$\Delta G = \sum_{i=1}^{2n} w_i u_i + C \qquad (1)$$

where, for n protein residues (or protein residues plus selected water molecules), there are $2n$ terms, u_i, each representing either a van de Waals or an electrostatic residue-based interaction energy with the ligand that contributes to the total binding free energy according to a weighting factor, w_i, that is determined from the PLS analysis, and C is a constant. Although other partitioning schemes are also possible,[7,8] only whole protein residues and ligands are supported in the present implementation. Obviously, if one of these contributions does not vary significantly among the complexes it cannot be used to account for activity/affinity differences within the series however important it may be for the constant term, C, and the overall free energy of binding. A pictorial representation of the COMBINE workflow is shown in Figure 1.

Since its inception in 1995, when the original approach was developed to account for the differences in activity in a series of human synovial fluid phospholipase A_2 inhibitors,[7] the COMBINE method has been applied to the study of other small molecules binding to different protein targets (HIV-1 protease,[9] human cytochrome P450 1A2,[10] human neutrophil elastase,[11] HIV-1 reverse transcriptase (RT),[12] acetylcholinesterase,[13-15] haloalkane dehalogenases DhlA[16] and LinB[17]) and also to peptide-protein,[18,19] protein-protein,[20] and protein-DNA interactions.[8]

The method has acquired some relevance within the 3D-QSAR field and has been reviewed in detail by Wade et al.,[21,22] Damborsky et al.,[17] and more recently by Lushington et al.,[23] who have also proposed some ideas to enhance COMBINE's capabilities in the future.

An important milestone in the development of the COMBINE methodology was the incorporation of multiple structures into the analysis, which allows the introduction, at least in part, of target flexibility.[8,24-26] All of

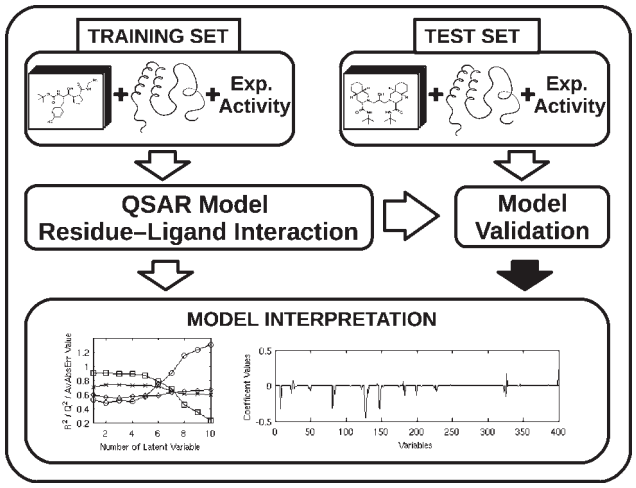


Figure 1
COMBINE workflow.

these studies demonstrate that qualitatively reliable COMBINE models can be obtained using multiple structural representations of the receptor, but care must be taken when attempting to perform a quantitative analysis, due to the conformational dependence of the models. Another interesting issue related to structural variation entering a COMBINE analysis is the joint study of affinity and selectivity by use of different protein targets belonging to the same family, which may provide important guidelines for drug design. Key examples are the study by Wang and Wade[27] of sialic and benzoic acid analogs binding to N2 and N9 subtypes of neuraminidase and the study of ligand binding to three serine proteases (trypsin, thrombin, and coagulation factor Xa) by Murcia et al.[28] This approach can, in principle, be extended to an arbitrary number of receptors from the same protein family.

COMBINE analysis can also be linked to a docking algorithm, as shown by Murcia and Ortiz,[25] when screening virtual libraries to derive more reliable bound conformations of the putative ligands, improve the predictive ability of the regression models, and increase the enrichment factors.

It can be seen therefore that COMBINE analysis occupies a privileged position as an effective tool that can aid in the design and optimization of drug candidates. However, as recently stated by Lushington et al.,[23] the small number of groups worldwide using COMBINE should join their efforts to disseminate the method and make it available to the scientific community. Although our group promoted this initiative some time ago through the free Web-based release of the COMBINE code,[29] a user-friendly graphical interface that allows easy manipulation of data and input/output files was lacking. In this contribution we fill this gap by providing the COMBINE

program with a graphical user interface (GUI) written in Java to ensure portability to different operating systems that is being released under a scientific/academic nonprofit and noncommercial license.

MATERIALS AND METHODS

Technical details of the gCOMBINE application

gCOMBINE is the GUI developed as a user-friendly wrapper to the original command-line COMBINE program. The GUI has been written in Java[30] language (v. 1.6.0_10), which ensures platform portability. Graphics functionality and interactivity have been added with Java Foundation Classes (JFC) and Swing components, respectively. For development of the GUI the NetBeans[31] IDE (Integrated Development Environment) 6.1 was used and the Swing Application Framework[32] was included. The different charts were generated using the JFreeChart[33] 1.0.11 and JCommon[34] 1.0.14 libraries. Both are distributed under a GNU Lesser General Public License.[35] These libraries allow gCOMBINE to generate interactive charts with the most relevant data for easy manipulation and analysis. Because the GUI is platform-independent and the COMBINE program is written in standard GNU Fortran, the complete application (COMBINE + gCOMBINE) can be used under Linux, Windows or Mac operating systems with the gcc compiler (v. 3.4.6 was employed in our case). gCOMBINE has an object-oriented design based on the Model-View-Controller (MVC) pattern.[36] The main class for the Model is *CombineModel*. An instance of this class stores the information about a specific model (or configuration) generated from a COMBINE run: name for the model, a description comment, the working folder, the configuration parameters, the output files, and tables and charts. The parameters are stored on an instance of the Parameters class, which uses the *ComplexesListItem* objects to keep the different ligand-receptor complexes related to the COMBINE model under study. The tables and charts are panels generated by the static methods of the classes *CombineTables* and *CombineGraphs*, respectively, taking a *CombineModel* instance as input. The View is launched by the *CombineGUIApp* class that creates an instance of the *CombineGUIView* class. This instance acts as a store for the different graphical objects and also as the Controller for the different actions (including internal validations) that can be performed when interacting with the objects. An internal class (*CombineThread*) is used to run the COMBINE program in a different execution thread to avoid the blockade of the GUI interface while COMBINE is running. *CombineThread* uses an instance of the *CombineWrapper* class to prepare the COMBINE execution: it launches the calculation, controls the process (taking the logs with the *StreamGlobber* class) and loads the results upon completion of the run. Three other classes are used through the life cycle of the application: a) *CombineConstants* (to contain different constants); b) *CombineException* (to propagate customized errors and warnings); and c) *Useful* (to store some common methods).

The statistics behind COMBINE

As in any other method focused on obtaining a quantitative view of structure-activity relationships, the core of COMBINE is a matrix containing structure-related energy descriptors (variables) to be correlated with biological activities or binding energies. The statistical method underlying COMBINE analysis is well known and widely accepted by the scientific community, and therefore it has been intensely reviewed.[37] However, a brief summary of the main steps and ideas follows.

Construction of the X matrix

The X matrix contains the entire set of variables describing the interaction energies between each ligand and every protein residue for all the complexes. Usually these are van der Waals interactions calculated using a molecular mechanics force field (typically AMBER) and electrostatic interactions calculated using point charges and either Coulomb's law (and a constant or distance-dependent dielectric definition) or the more elaborate and accurate generalized born (GB)[38] or Poisson-Boltzmann (PB)[39] methods. In addition, desolvation energy terms for both receptor and ligand can also be incorporated as "external variables". For each complex only two AMBER-type files are required, one containing the atomic coordinates (.crd extension) and the other containing the topology (i.e. atom connectivity), atom types and force-field parameters (.top extension). Alternatively, the user can generate the X matrix externally in the format described in the User Guide and load it into the program.

Pretreatment of the X matrix

To reduce the number of variables while keeping all the relevant information within the X matrix, those interaction energy values with a standard deviation below a user-defined cut-off, which can be safely assumed not to contribute to the overall variance in activity, can be removed (Pretreatment cut-off, see below). Positive energy values, which in some cases could arise from force-field inconsistencies or modeling errors, can optionally be truncated to zero (Pretreatment option, see below). Scaling of the variables can also be performed using two different approaches (Scaling option, see below): (i) standard scaling, where the mean value over the whole set of variables is subtracted from each variable and divided by the standard deviation (it is therefore similar to a Z-score), and (ii) block scaling, by means of

which the mean value of the variables is subtracted from the one being scaled and divided by the standard deviation of these variables (again, similar to a Z-score but using a group of variables).

PLS regression

This technique combines and generalizes features from PCA and Multiple Linear Regression (MLR) in the sense that not only orthogonal Principal Components (PC) are extracted, as in PCA, but also a fitting procedure is performed to describe the activities of the compounds (the dependent variable), as in MLR. There are two initial matrices in a COMBINE analysis: (i) a matrix containing the independent variables (interaction energies, and possibly additional variables such as desolvation energies), the X matrix [Eq. (2)], and (ii) a matrix (column vector) with the dependent variable (activities), the Y matrix [Eq. (3)].

$$X = \begin{pmatrix} E_1^1 & E_2^1 & \dots & E_M^1 & V_1^1 & V_2^1 & \dots & V_M^1 & A_1^1 & \dots & A_S^1 \\ E_1^2 & E_2^2 & \dots & E_M^2 & V_1^2 & V_2^2 & \dots & V_M^2 & A_1^2 & \dots & A_S^2 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ E_1^N & E_2^N & \dots & E_M^N & V_1^N & V_2^N & \dots & V_M^N & A_1^N & \dots & A_S^N \end{pmatrix} \quad (2)$$

where E_j^i, V_j^i, and A_j^i are the electrostatic, van der Waals, and additional variables, respectively. N is the number of compounds, M is the number of residues in the protein, and S is the number of additional variables.

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_N \end{pmatrix} \quad (3)$$

where y_i is the individual activity of compound i. The PLS analysis starts by decomposing the X and Y matrices into one score matrix, T, and two different loading matrices, P and Q [Eq. (4)], using the iterative NIPALS algorithm[40]:

$$X = TP^T \quad Y = TQ^T \quad (4)$$

The loading matrices P and Q contain information about the variables in the so-called LV or PC space. These are orthogonal vectors obtained as linear combinations of the original variables in the X matrix. The coefficients in a given PC provide information on the relative weight of the different terms and can be used to deduce the relevance of each individual ligand–residue interaction to explain the variance in activity/affinity. On the other hand, the score matrix T contains information about the compounds, described in terms of their projections onto the PCs. The PC space is normalized and has a mean of zero, so compounds with high scores should be checked as they could behave as outliers. In addition, clusters of compounds can be detected. A plot of the regression line between the experimentally determined and theoretically calculated activity/affinity values and calculation of the regression coefficient [r^2, Eq. (5)] allow the user to visualize the quality of the fit for the training set compounds, and also for the excluded (not used) or test compounds.

$$r^2 = \frac{\left[\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \langle\hat{y}\rangle)\right]^2}{\sum_{i-1}^{N}(y_i - \bar{y})^2 \sum_{i=1}^{N}(\hat{y}_i - \langle\hat{y}\rangle)^2} \quad (5)$$

where $\langle\hat{y}\rangle = \frac{\sum_{i=1}^{N}\hat{y}_i}{N}$.

Cross-validation

This method, which is used to check that the derived correlation is not spurious and to assess the robustness of the resulting statistical model, consists of predicting the dependent variable for some complexes that are not included in model derivation. Briefly, if C is the whole set of N compounds ($C = \{C_1,\dots,C_N\}$) with associated activities Y ($Y = \{y_1,\dots,y_N\}$), the method builds a number of subsets of s elements from C (when $s = 1$ it is the commonly employed Leave-One-Out option) and sets them apart for their activities/affinities to be predicted later, thus making up an internal test set. The compounds represented by s can be selected by following a predetermined sequential order or can be randomly assigned to a predetermined number of groups (Cross-validation Method option, see below). In any case, C can be split into k subsets S_i ($S_i = \{S_{i1},\dots,S_{is}\}$) where the subscript i represents any subset number from 1 to k. Numerically, k is the nearest integer greater than or equal to N/s (the ratio between the total number of compounds, N, and the number of those making up each subset, s). Usually, the last group would have less than s elements as it contains the remaining compounds. In the next step, k PLS regression models are built: model 1 with all N compounds except for those in S_1, model 2 with all N compounds except for those in S_2, and so on. In each case, the activities for compounds in S_1, S_2... will be estimated from their respective models ($\hat{y}_1,\dots,\hat{y}_N$) and at the end of the process a list of predicted activities for all the compounds will be obtained. The performance is then quantified by the q^2 cross-validated correlation coefficient [Eq. (6)]:

$$q^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (6)$$

where $\bar{y}$ is the average value of the activity ($\bar{y} = \sum_{i=1}^{N} y_i/N$). In simple words, this metric describes the amount of variance in the dataset that is explained by the model. Besides, a standard deviation of error of predictions [*SDEP*, Eq. (7)] and an average absolute error [*AAE*, Eq. (8)] are also calculated.

$$\mathrm{SDEP} = \sum_{i=1}^{N} \sqrt{\frac{(\hat{y}_i - y_i)^2}{N}} \quad (7)$$

$$\mathrm{AAE} = \sum_{i=1}^{N} \frac{|\hat{y}_i - y_i|}{N} \quad (8)$$

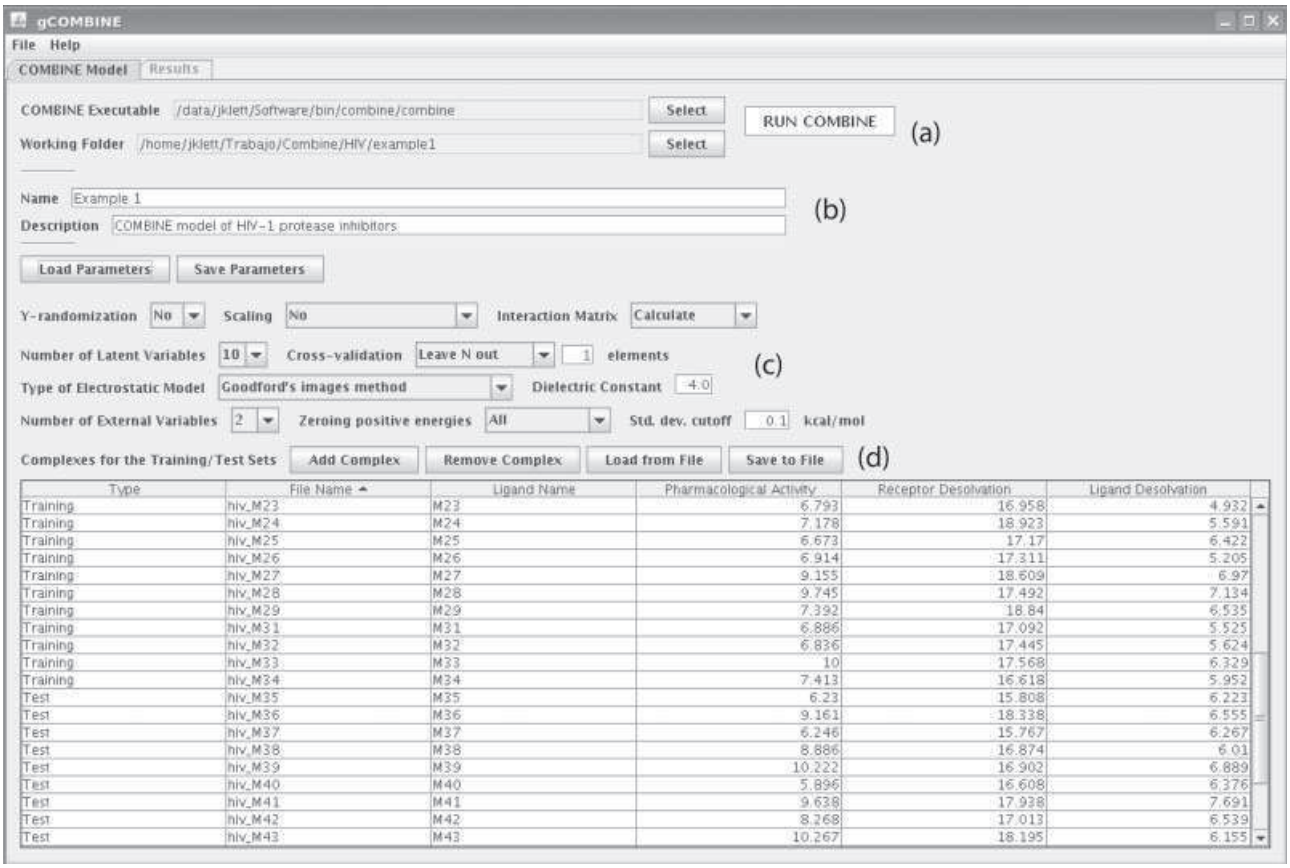
It can be argued that, despite the cross-validation procedure, the resulting model fits the data just by chance due to the selection of a fortuitous equation out of the huge amount of different PLS regression models that can be constructed with the thousands of variables contained in the *X* matrix. To check against this possibility the affinities/activities of the compounds can be randomly reassigned (permutation of activities or Y-randomization) to prove the point that in this case it is usually not possible to derive an acceptable model. gCOMBINE allows the user to carry out this task thorough the Y-randomization option in the main window (see below). This test is performed 100 times and therefore it is quite time-consuming.

Selection of the best model

As successive components are extracted from the *X* matrix, a check is made to estimate the amount of variance that is recovered (it must be borne in mind that the PLS method attempts to explain the variance not only in the *X* matrix, as does PCA, but also in the *Y* matrix). Although there is not a strict rule to select the best model resulting from a PLS analysis, the general guidance is to study the evolution of both the cross-validated correlation coefficient [q^2, Eq. (6)] and the standard deviation of the errors in prediction [SDEP, Eq. (7)]. gCOMBINE provides the user with a graphical description showing the evolution of the main chemometric indices as new components are being extracted (five by default, and up to 10, Number of Latent Variables option, see later) to facilitate the decision on the optimal dimensionality to choose for further analysis.

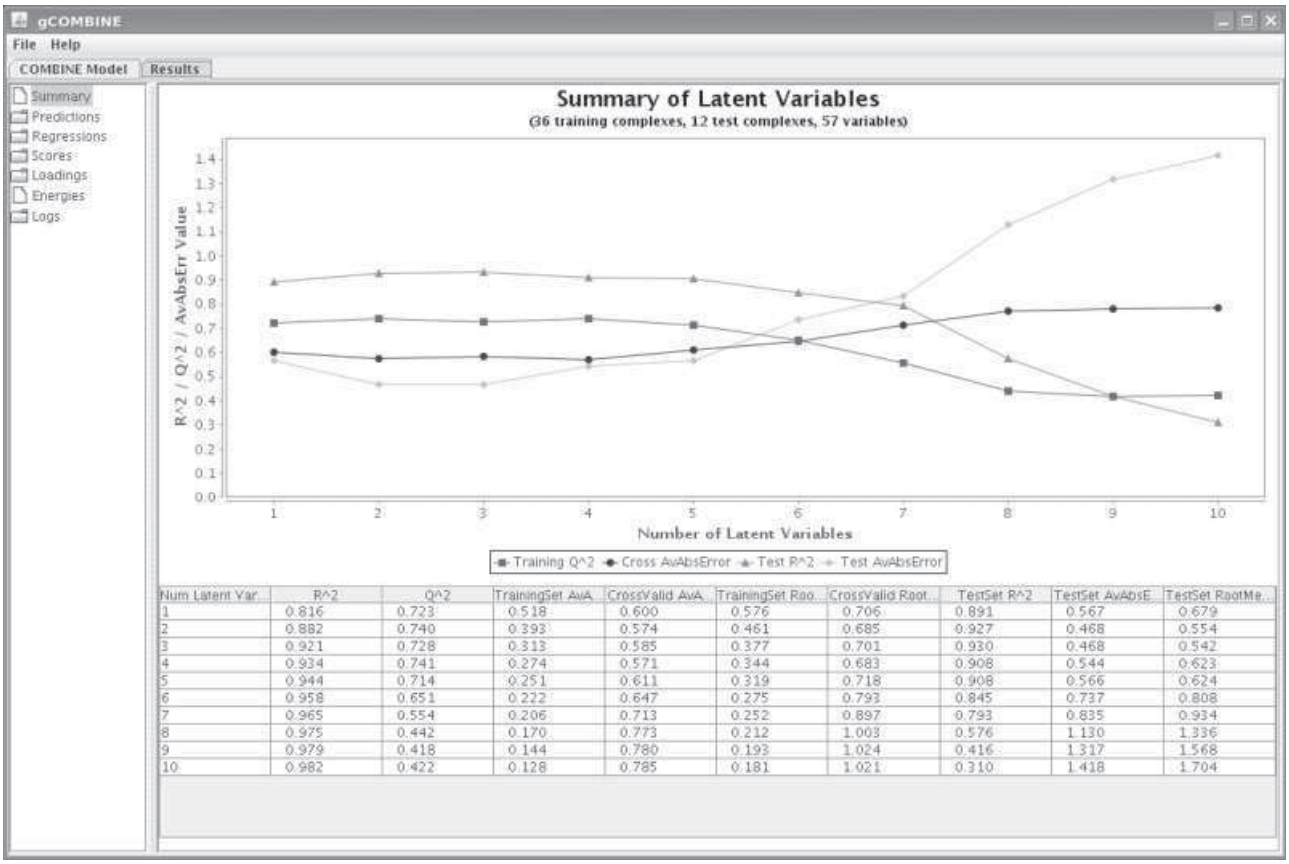
External validation

The best way to validate a PLS model is to challenge it with an external set of modeled complexes and compare the predicted affinity/activity values for the bound



Type	File Name	Ligand Name	Pharmacological Activity	Receptor Desolvation	Ligand Desolvation
Training	hiv_M23	M23	6.793	16.958	4.932
Training	hiv_M24	M24	7.178	18.923	5.591
Training	hiv_M25	M25	6.673	17.17	6.422
Training	hiv_M26	M26	6.914	17.311	5.205
Training	hiv_M27	M27	9.155	18.609	6.97
Training	hiv_M28	M28	9.745	17.492	7.134
Training	hiv_M29	M29	7.392	18.84	6.535
Training	hiv_M31	M31	6.886	17.092	5.525
Training	hiv_M32	M32	6.836	17.445	5.624
Training	hiv_M33	M33	10	17.568	6.329
Training	hiv_M34	M34	7.413	16.618	5.952
Test	hiv_M35	M35	6.23	15.808	6.223
Test	hiv_M36	M36	9.161	18.338	6.555
Test	hiv_M37	M37	6.246	15.767	6.267
Test	hiv_M38	M38	8.886	16.874	6.01
Test	hiv_M39	M39	10.222	16.902	6.889
Test	hiv_M40	M40	5.896	16.608	6.376
Test	hiv_M41	M41	9.638	17.938	7.691
Test	hiv_M42	M42	8.268	17.013	6.539
Test	hiv_M43	M43	10.267	18.195	6.155

Figure 2
gCOMBINE main window. Letters a–d refer to the four main data blocks (see text).



Num Latent Var	R^2	Q^2	TrainingSet AvA	CrossValid AvA	TrainingSet Roo	CrossValid Root	TestSet R^2	TestSet AvAbsE	TestSet RootMe
1	0.816	0.723	0.518	0.600	0.576	0.706	0.891	0.567	0.679
2	0.882	0.740	0.393	0.574	0.461	0.685	0.927	0.468	0.554
3	0.921	0.728	0.313	0.585	0.377	0.701	0.930	0.468	0.542
4	0.934	0.741	0.274	0.571	0.344	0.683	0.908	0.544	0.623
5	0.944	0.714	0.251	0.611	0.319	0.718	0.908	0.566	0.624
6	0.958	0.651	0.222	0.647	0.275	0.793	0.845	0.737	0.808
7	0.965	0.554	0.206	0.713	0.252	0.897	0.793	0.835	0.934
8	0.975	0.442	0.170	0.773	0.212	1.003	0.576	1.130	1.336
9	0.979	0.418	0.144	0.780	0.193	1.024	0.416	1.317	1.568
10	0.982	0.422	0.128	0.785	0.181	1.021	0.310	1.418	1.704

Figure 3

gCOMBINE *Results* tab showing the evolution of the chemometric indices in graphical (Top) and tabular form (Bottom).

ligands with the actual ones. This is accomplished in gCOMBINE by feeding the program with additional complexes that are marked as "Test" in lieu of "Training" in the Type column (see Main Menu in Fig. 2).

RESULTS AND DISCUSSION

The application

gCOMBINE has been designed bearing in mind the need for a simple and easy-to-use chemometric tool. Once that the program has been launched it displays a menu bar with two submenus (File and Help) and a tabbed panel with two tabs (see Fig. 2): one for the configuration of the model (*COMBINE Model*) and the other one to manage the results (*Results*). The menu bar, under File submenu, contains four options: (a) New Model, to clean the data from the tabs, (b) Load Model, to load a previously saved model into the tabs, (c) Save Model, to save the current model into a specific file, and (d) Exit, to close the application. Help submenu offers information about the original COMBINE publication, the main COMBINE author and contact information. The *COMBINE Model* tab can be divided into four main areas (a through d in Fig. 2): (a) the top part where the user can select the folder for the COMBINE executable and the working folder where the complexes are stored. Clicking on the RUN COMBINE button will start the calculation, (b) a section where the user can introduce commentaries related to the job into two boxes, Name and Description, (c) this section allows the user to load parameters from a previous calculation or to save the current parameters being used thorough the Load/Save Parameters buttons. All the parameters entering the model are configured here (the reader is referred to the "Materials and Methods" section for the definition of the different issues described here): Y-randomization (No by default), Scaling (No by default), Interaction Matrix (it can be calculated and written out by gCOMBINE or read in from an external file), Number of Latent Variables (5 are extracted by default), validation method (Leave N Out or Random Groups), Type of Electrostatic Model (uniform dielectric constant, Goodford's implementation of the images method,[41] a distance-dependent dielectric constant, PB electrostatic interaction energies read from an external file, and a sigmoidal model[42]), Dielectric

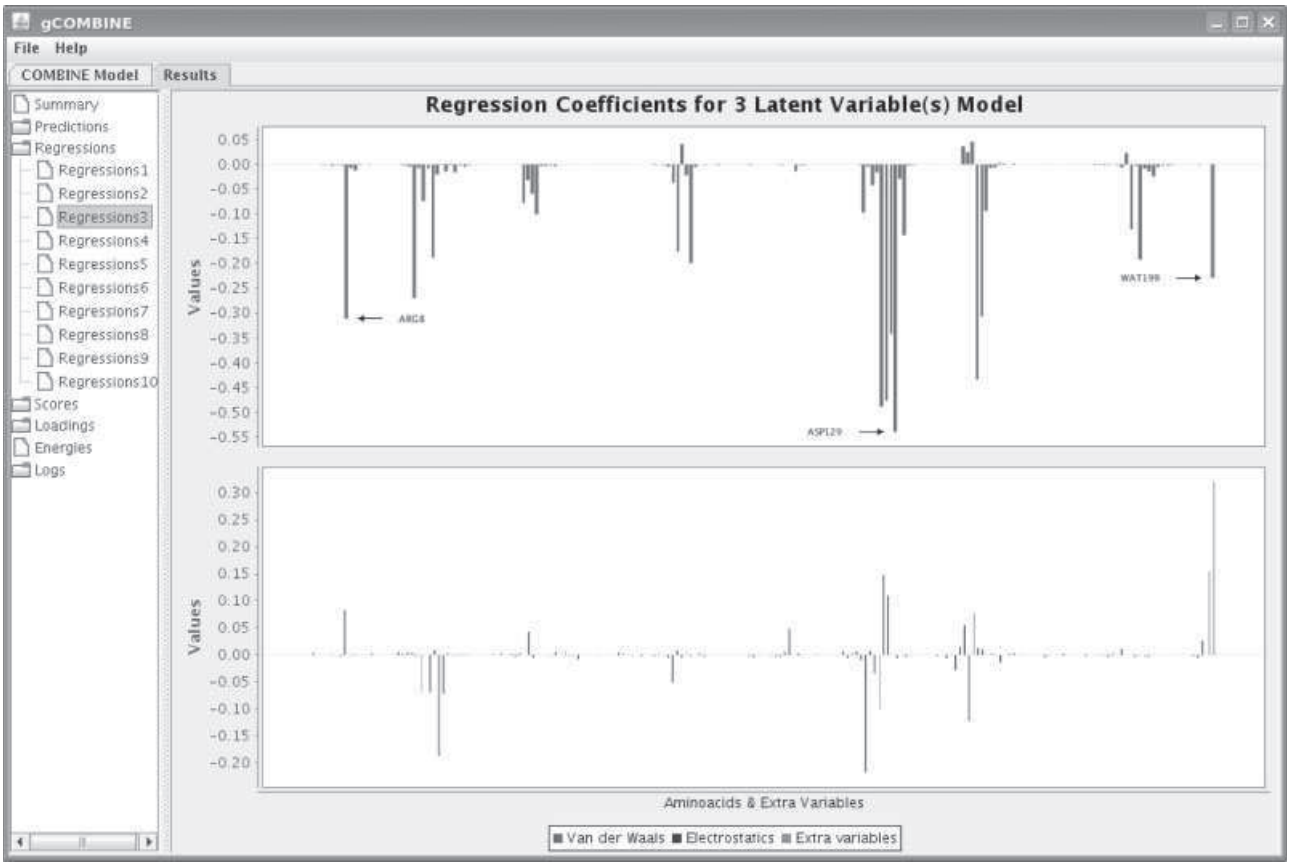


Figure 4

Plots showing the weights assigned to the residue-based van der Waals and electrostatic interaction energy values in a COMBINE model made up of four principal components to account for the differences in activity in the HIV-1 protease inhibitor series.

Constant value (four by default), Number of External Variables, Pretreatment of the data and the Pretreatment cut-off value, and (d) the complex area, where the user can add/remove/load/save complexes from/to a file (Add/Remove Complex and Load/Save from/to File buttons). Below these buttons a table is shown with the following information: the type associated to each complex (Training, Test, or Not Used, as defined by the user), which can be changed at any time to test alternative models, the File Name containing the complex, the Ligand Name, the Pharmacological Activity, and one column for each external variable to be considered. Once the parameters and the complexes have been read in and after the program begins to run, the application checks if all the parameters needed have been supplied, if they have valid values, and if the required files exist.

The user can stop the execution or wait until completion of the run. In this latter case, if no errors are found, the application focuses on the *Results* tab (see Fig. 3) where the user can have access to different tables and charts grouped by *nodes* and *sub-nodes*: a summary of each model (both a graphic showing the evolution of the chemometric indices and a table containing these indices are produced), PLS coefficients plots (Fig. 4), predictions (for each model a graphic of experimental vs. predicted activity values is shown, Fig. 5 panel a), loadings and scores plots (Fig. 5 panels b and c, respectively), and a plot of the interaction energy variables entering the PLS decomposed on a per-residue basis (a very useful plot to detect anomalous energy values, Fig. 5 panel d). The user can interact with the graphs in several ways: zoom in/out, see tooltips for specific data, set tags, change the appearance, save them as images, print, and so forth. Finally, there is a *Logs* node to keep track of the program messages.

Testing gCOMBINE

Among the many publications of successful COMBINE analysis, we have chosen two of them for testing the graphical implementation reported herein. The first one is the set accompanying the original distribution, which employed 48 (32 for the training set and 16 for the test set) inhibitors of human immunodeficiency virus Type 1

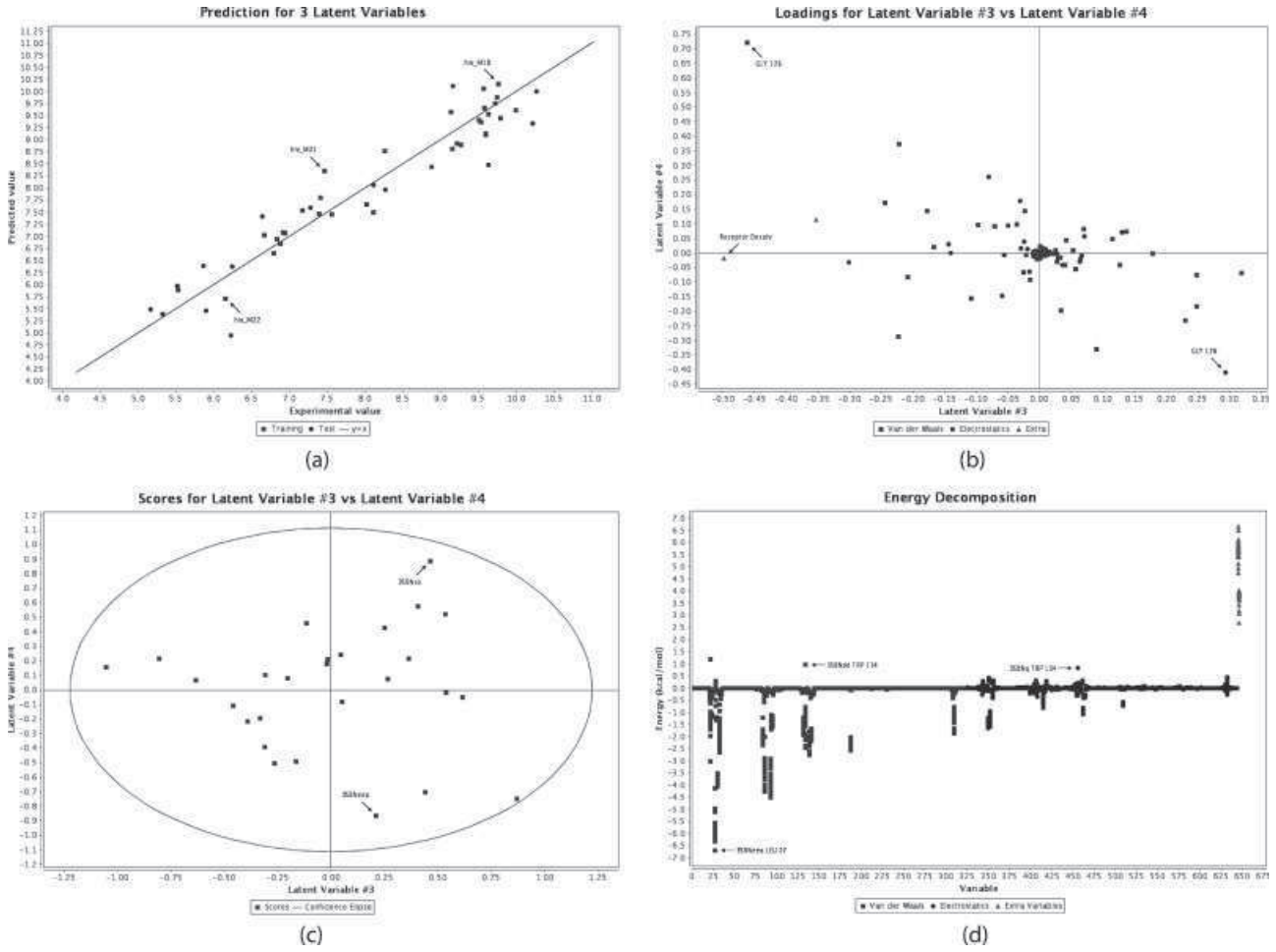


Figure 5

Selected screenshots from the gCOMBINE program displaying: a) experimental versus predicted activity plot, (b) plot showing the contributions (loadings) of the original variables to the principal components shown, (c) scores plot (the applicability domain is drawn as a confidence ellipse[43]), (d) plot of the original variables entering the PLS analysis following decomposition of the ligand-receptor interaction energy on a per-residue basis.

(HIV-1) protease.[9] In the second example, also related to another pharmacologically relevant HIV-1 target, namely RT,[12] we will show how externally generated electrostatic energy variables, as calculated with the DelPhi software,[44] can be incorporated into the model.

1. Two main issues motivated the first of these two studies: (a) bare ligand-receptor interaction energies (as computed by Merck researchers using the MM2X force field) per se correlated quite well (using linear regression analysis, $r^2 = 0.74$) with experimentally determined enzyme inhibition data (IC_{50} values, that is, compound concentrations giving rise to 50% inhibition of enzyme activity).[44] Moreover, the predictive ability of such a linear model on 16 test set compounds (not included in model derivation) was also remarkable ($q^2 = 0.75$) with an average absolute error around 1 across a range of 5 log units, and (b) the realization that no improvement upon incorporation of either solvation effects (using a continuum description) or using another force field (CHARMM in their case) was achieved. Taking these two issues into account, the two main objectives addressed by the

Table I

Chemometric Indices for the Different Models in the HIV-1 Protease Study Discussed in the Text

Model	Objects	Variables	LV	r^2	q^2	$SDEP_{CV}$	$SDEP_{ex}$
L_{MM2X}	32	1	1	0.74	0.75		1.00
L_{AMBER}	32	1	1	0.81	0.79	0.61	1.08
C_{AMBER}	32	48	2	0.89	0.70	0.72	0.83
gC_{AMBER}[a]	32	48	2	0.89	0.70	0.72	0.80

[a]Calculated with gCOMBINE.

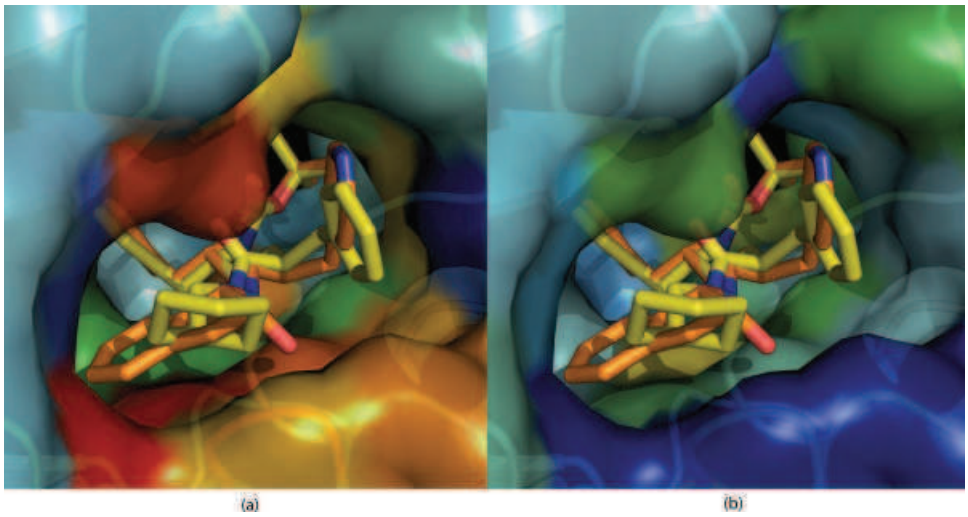


Figure 6

Visualization in PyMOL (http://pymol.sourceforge.net/) of the PLS regression coefficients plotted in Figure 4. The semitransparent surface enveloping the HIV-1 protease target has been spectrum-colored using the van der Waals (a) and electrostatic (b) PLS coefficients from the fourth column (B-factor) in the PDB file generated by gCOMBINE.

COMBINE analysis methodology were (i) to check for possible dependencies of the correlation on the force field used, and (ii) to try and develop more accurate QSAR models. In this exercise, it was shown that similar results could be obtained when comparable models were built within the framework of the AMBER force field, so no force-field dependencies were detected (see the chemometric indices for L_{MM2X} and L_{AMBER} in Table I). On the other hand, remarkable improvements were achieved through the use of COMBINE models (see the original article), especially when partial desolvation effects for ligand and receptor upon complex formation were included using a continuum description (by solving the PB equation) and the standard Coulombic distance-dependent electrostatic term was replaced with solvent-corrected values calculated for each residue. Interestingly, the main conclusion of this work was that simply replacing the Coulombic term with the continuum electrostatics description and including the desolvation effects did not lead to a significant improvement when MLR was used but the performance of the corresponding COMBINE model was dramatically enhanced when the continuum electrostatic interactions were employed. Although different improvements over the standard Coulombic term were included in the original article we will restrict ourselves here to the simplest (and more widely used) case as our intention is to reproduce the data rather than recapitulating the previously published comparison. In particular we show the reproducibility of what was called the C_{AMBER} model, where van der Waals and electrostatic contributions (the latter, a straightforward Coulombic term using a dielectric constant of four) were taken directly from the AMBER force field using the ANAL module. The results cannot be exactly the same, though, because the cross-validation technique contains a random component: the compounds are randomly assigned to one of five groups of approximately the same size, each group in turn is excluded from the analysis, and the whole procedure is repeated 20 times. Nevertheless, clearly comparable results were obtained (see chemometric indices for C_{AMBER} and gC_{AMBER} in Table I).

Finally, the relative weight assigned to individual residue-based interactions (van der Waals and electrostatic) by the COMBINE model can be color-coded and displayed on a surface representation of the protein, as shown in Figure 6.

2. The second article revisited here entailed the study of 27 6-arylsulfonyl-2-aminobenzonitrile derivatives synthesized and tested as second-generation non-nucleoside HIV-1 RT inhibitors (NNRTI).[12] In this case there was a wealth of experimental data including information about activity on RT enzymes bearing different mutations at the NNRTI binding site. COMBINE models were obtained to quantitatively characterize the observed structure-activity data and possibly to account for the effect of some of the mutations. Ligand-residue van der Waals interaction energies were calculated using AMBER parameters (parm99) while their solvent-corrected electrostatic counterparts were obtained by solving the PB equa-

Table II
Chemometric Indices Calculated by gCOMBINE for the Whole Set of 27 6-arylsulfonyl-2-aminobenzonitrile HIV-1 Reverse Transcriptase Inhibitors

LV	r^2	q^2	$SDEP_{CV}$
1	0.84	0.71	0.62
2	0.88	0.78	0.55
3	0.93	0.82	0.49
4	0.94	0.86	0.44
5	0.95	0.86	0.44

tion (as implemented in DelPhi). Also, the desolvation changes incurred by ligands and receptor upon complex formation were included as additional external variables. For testing gCOMBINE, these electrostatic energy calculations were performed using DelPhi in stand-alone mode and then loaded into the application (Type of Dielectric Model menu, Poisson-Boltzmann from.dph files option, see Figure 2 block c). By doing this, the published results were accurately reproduced with only minor variations being obtained, as expected, when random groups were employed (see Table II). Relevant results are presented in Figure 5: loading and scoring plots (panels b and c, respectively), interaction energy variables entering the PLS analysis (panel d), and the evolution of r^2, q^2, and SDEP (Table II) for the whole set of 27 NNRTI complexes as the number of PC being extracted by gCOMBINE increases. As in the previous example, no attempt will be made here to discuss or compare the results, which are already published, but it is clear that gCOMBINE faithfully reproduces the data.

CONCLUSIONS

The objective of this article has been to provide the COMBINE analysis method with an easy-to-use GUI that improves on the original command-line style implementation. The software is written in Java to allow platform portability and is made freely available to academic and/or public research institutions from a public web site (http://ub.cbm.uam.es/gCOMBINE) under an Academic License. This has been done with the idea of disseminating a user-friendly tool among the scientific community to encourage the use of a program that has proven useful in many areas related to ligand binding, structure-activity relationships and drug design.

ACKNOWLEDGMENTS

The authors thank Claire Coderch for help in testing the application. This work is dedicated to the memory of Ángel R. Ortiz, who looked forward to making this tool widely available but passed away before it was fully operative.

REFERENCES

1. Hansch C, Hoekman D, Leo A, Weininger D, Selassie CD. Chem-bioinformatics: comparative QSAR at the interface between chemistry and biology. Chem Rev 2002;102:783–812.
2. Cramer RD, Patterson DE, Bunce JD. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. JACS 1988;110:5959–5967.
3. Klebe G, Abraham U, Mietzner T. Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. J Med Chem 1994;37:4130–4146.
4. Ginalski K. Comparative modeling for protein structure prediction. Curr Opin Struct Biol 2006;16:172–177.
5. Warren GL, Andrews CW, Capelli AM, Clarke B, Lalonde J, Lambert MH, Lindvall M, Nevins N, Semus SF, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS. A critical assessment of docking programs and scoring functions. J Med Chem 2006;49:5912–5931.
6. Shoichet BK. Virtual screening of chemical libraries. Nature 2004;432:862–865.
7. Ortiz AR, Pisabarro MT, Gago F, Wade RC. Prediction of drug binding affinities by comparative binding energy analysis. J Med Chem 1995;38:2681–2691.
8. Tomic S, Nilsson L, Wade RC. Nuclear receptor-DNA binding specificity: a COMBINE and free-Wilson QSAR analysis. J Med Chem 2000;43:1780–1792.
9. Perez C, Pastor M, Ortiz AR, Gago F. Comparative binding energy analysis of HIV-1 protease inhibitors: incorporation of solvent effects and validation as a powerful tool in receptor-based drug design. J Med Chem 1998;41:836–852.
10. Lozano JJ, Pastor M, Cruciani G, Gaedt K, Centeno NB, Gago F, Sanz F. 3D-QSAR methods on the basis of ligand-receptor complexes. Application of COMBINE and GRID/GOLPE methodologies to a series of CYP1A2 ligands. J Comput Aided Mol Des 2000;14:341–353.
11. Cuevas C, Pastor M, Perez C, Gago F. Comparative binding energy (COMBINE) analysis of human neutrophile elastase inhibition by pyridone-containing trifluoromethylketones. Comb Chem High Throughput Screen 2001;4:627–642.
12. Rodríguez-Barrios F, Gago F. Chemometrical Identification of mutations in HIV-1 reverse transcriptase conferring resistance or enhanced sensitivity to arylsulfonylbenzonitriles. JACS 2004;126:2718–2719.
13. Guo J, Hurley MM, Wright JB, Lushington GH. A docking score function for estimating ligand-protein interactions: application to acetylcholinesterase inhibition. J Med Chem 2004;47:5492–5500.
14. Lushington GH, Wallace NM, Guo JX. Reliable Prescreening of Candidate NerveAgent Prophylaxes via 3D QSAR. DTIC Monitor Series, 2005. 1–28.
15. Martin-Santamaria S, Munoz-Muriedas J, Luque FJ, Gago F. Modulation of binding strength in several classes of active site inhibitors of acetylcholinesterase studied by comparative binding energy analysis. J Med Chem 2004;47:4471–4482.
16. Kmunicek J, Luengo S, Gago F, Ortiz AR, Wade RC, Damborsky J. Comparative binding energy analysis of the substrate specificity of haloalkane dehalogenase from Xanthobacter autotrophicus GJ10. Biochemistry 2001;40:8905–8917.
17. Damborsky J, Kmunicek J, Jedlicka T, Luengo S, Gago F, Ortiz AR, Wade RC. Rational redesign of haloalkane dehalogenases guided by comparative binding energy analysis. In: Svendsen A, Dekker M, editors. Enzyme functionality: design, engineering and screening. New York: Marcel Dekker; 2004. pp 79–96.

18. Schleinkofer K, Wiedemann U, Otte L, Wang T, Krause G, Oschkinat H, Wade RC. Comparative structural and energetic analysis of WW domain-peptide interactions. J Mol Biol 2004;344: 865–881.
19. Wang T, Wade RC. Comparative binding energy (COMBINE) analysis of OppA-peptide complexes to relate structure to binding thermodynamics. J Med Chem 2002;45:4828–4837.
20. Tomic S, Bertosa B, Wang T, Wade RC. COMBINE analysis of the specificity of binding of Ras proteins to their effectors. Proteins 2007;67:435–447.
21. Wade RC, Henrich S, Wang T. Using 3D protein structures to derive 3D-QSARs. Drug Discovery Today: Technol 2004;1:241–246.
22. Wade RC, Ortiz AR, Gago F. Comparative binding energy analysis. In: Kubinyi H, Folkers G, Martin Y, editors. 3D-QSAR in drug design, Vol. 2. Dordrecht (Netherland): Kluwer-ESCOM; 1998. pp 19–34.
23. Lushington GH, Guo JX, Wang JL. Whither combine? New opportunities for receptor-based QSAR. Curr Med Chem 2007;14: 1863–1877.
24. Mou TC, Gille A, Suryanarayana S, Richter M, Seifert R, Sprang SR. Broad specificity of mammalian adenylyl cyclase for interaction with 2′,3′-substituted purine- and pyrimidine nucleotide inhibitors. Mol Pharmacol 2006;70:878–886.
25. Murcia M, Ortiz AR. Virtual screening with flexible docking and COMBINE-based models. Application to a series of factor Xa inhibitors. J Med Chem 2004;47:805–820.
26. Pastor M, Perez C, Gago F. Simulation of alternative binding modes in a structure-based QSAR study of HIV-1 protease inhibitors. J Mol Graph Model 1997;15:364–371.
27. Wang T, Wade RC. Comparative binding energy (COMBINE) analysis of influenza neuraminidase-inhibitor complexes. J Med Chem 2001;44:961–971.
28. Murcia M, Morreale A, Ortiz AR. Comparative binding energy analysis considering multiple receptors: a step toward 3D-QSAR models for multiple targets. J Med Chem 2006;49:6241–6253.
29. COMBINE homepage. http://ub.cbm.uam.es/software.php. April 2009.
30. Java homepage. http://java.sun.com/. April 2009.
31. NetBeans homepage. http://www.netbeans.org. April 2009.
32. Swing Application Framework homepage. https://appframework.dev.java.net. April 2009.
33. JFreeChart homepage. http://www.jfree.org/jfreechart. April 2009.
34. JCommon homepage. http://www.jfree.org/jcommon. April 2009.
35. GNU Lesser General Public License homepage. http://www.gnu.org/licenses/lgpl.html. April 2009.
36. Model-View-Controller (MVC) pattern. http://java.sun.com/blueprints/patterns/MVC.html. April 2009.
37. Wold S, Ruhe A, Wold H, Dunn WJ. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. J Sci Stat Comp 1984;5:735–743.
38. Still WC, Tempczyk A, Hawley RC, Hendrickson T. Semianalytical treatment of solvation for molecular mechanics and dynamics. JACS 1990;112:6127–6129.
39. Honig B, Nicholls A. Classical electrostatics in biology and chemistry. Science 1995;268:1144–1149.
40. Wold H. Path models with latent variables: the NIPALS approach. In: Blalock HM, editor. Quantitative sociology: international perspectives on mathematical and statistical model building. New York: Academic Press; 1975. pp 307–357.
41. Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 1985;28:849–857.
42. Mehler EL, Solmajer T. Electrostatic effects in proteins: comparison of dielectric and charge models. Protein Eng 1991;4:903–910.
43. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B. Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. J Comput Chem 2002;23:128–137.
44. Holloway MK, Wai JM, Halgren TA, Fitzgerald PM, Vacca JP, Dorsey BD, Levin RB, Thompson WJ, Chen LJ, Desolms SJ. A priori prediction of activity for HIV-1 protease inhibitors employing energy minimization in the active site. J Med Chem 1995;38:305–317.
45. Eriksson L, Jaworska J, Worth AP, Cronin MT, Mcdowell RM, Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Perspect 2003;111:1361–1375.

(12) SOLICITUD INTERNACIONAL PUBLICADA EN VIRTUD DEL TRATADO DE COOPERACIÓN EN MATERIA DE PATENTES (PCT)

(19) Organización Mundial de la Propiedad Intelectual
Oficina internacional

(43) Fecha de publicación internacional
14 de Mayo de 2009 (14.05.2009)

PCT

(10) Número de Publicación Internacional
WO 2009/060114 A1

(51) Clasificación Internacional de Patentes:
C07D 401/06 (2006.01) *A61P 35/00* (2006.01)
A61K 31/4704 (2006.01)

(21) Número de la solicitud internacional:
PCT/ES2008/070190

(22) Fecha de presentación internacional:
20 de Octubre de 2008 (20.10.2008)

(25) Idioma de presentación: español

(26) Idioma de publicación: español

(30) Datos relativos a la prioridad:
P200702958
7 de Noviembre de 2007 (07.11.2007) ES

(71) Solicitantes *(para todos los Estados designados salvo US)*: **CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS** [ES/ES]; C/ Serrano, 117, E-28006 Madrid (ES). **CENTRO NACIONAL DE INVESTIGACIONES ONCOLÓGICAS** [ES/ES]; C/ Melchor Fernández Almagro, 3, E-28029 Madrid (ES).

(72) Inventor: **RAMÍREZ ORTIZ, Ángel** (fallecido).

(72) Inventores; e

(75) Inventores/Solicitantes *(para US solamente)*: **ORTIZ, María Luz** [ES/ES]; Pintor Rosales n°4, 8° Drcha., E-28932 Móstoles, Madrid (ES). **FÁBREGAS CLAVERÍA, María Carmen** [ES/ES]; C/ Suerte del Palomar n° 101, E-28410 Manzanares del Real, Madrid (ES). **MORREALE DE LEÓN, Antonio Jesús** [ES/ES]; Centro De Biología Molecular Severo Ochoa, C/ Nicolas Cabrera, 1, E-28049 Madrid (ES). **GIL REDONDO, Rubén** [ES/ES]; Centro De Biología Molecular Severo Ochoa, C/ Nicolas Cabrera, 1, E-28049 Madrid (ES). **RUIZ, Federico** [AR/ES]; Centro Nacional De Investigaciones Oncológicas, C/ Melchor Fernández Almagro, 3, E-28029 Madrid (ES). **BRAVO SICILIA, Jerónimo** [ES/ES]; Centro Nacional De Investigaciones Oncológicas, C/ Melchor Fernández Almagro, 3, E-28029 Madrid (ES).

(74) Mandatario: **PONS ARIÑO, Angel**; Glorieta de Rubén Darío, 4, E-28010 Madrid (ES).

(81) Estados designados *(a menos que se indique otra cosa, para toda clase de protección nacional admisible)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Estados designados *(a menos que se indique otra cosa, para toda clase de protección regional admisible)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), euroasiática (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), europea (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Publicada:
— *con informe de búsqueda internacional*
— *antes de la expiración del plazo para modificar las reivindicaciones y para ser republicada si se reciben modificaciones*

(54) Title: INHIBITORS OF ENZYME O^6-ALKYLGUANINE-DNA-METHYLTRANSFERASE FOR CANCER TREATMENT

(54) Título: INHIBIDORES DE LA ENZIMA O^6-ALQUILGUANINA-ADN-METIL- TRANSFERASA PARA EL TRATAMIENTO DEL CÁNCER

(57) Abstract: The invention relates to the use of compounds derived from piperadinyl-methyl-tetrazole-quinolinone and derived from diphenyl-triazolo- pyrimidine as inhibitors of the reparative action of the DNA produced by enzyme O^6-alkylguanine-DNA-methyltransferase, and to pharmaceutical compositions containing same, for the preparation of coadjuvant drugs for use in antitumour therapy using alkylating agents.

(57) Resumen: La presente invención se refiere al uso de compuestos derivados de piperadinil-metil-tetrazol-quinolinona y derivados de difenil-triazolo- pirimidina como inhibidores de la acción reparadora del ADN realizada por la enzima O^6 -alquilguanina-ADN-alquiltransferasa y con composiciones farmacéuticas que los contienen, para la preparación de medicamentos coadyuvantes de la terapia antitumoral basada en agentes alquilantes.

WO 2009/060114 A1