Universitat Ramon Llull

TESIS DOCTORAL

Título	**Herramientas de cribado virtual aplicadas a inhibidores de entrada del VIH. Diseño de nuevos compuestos anti-VIH.**
Realizada por	Violeta Isabel Pérez Nueno
en el Centro	Escola Tècnica Superior IQS
y en el Departamento	Química Orgánica
Dirigida por	Dr. Jordi Teixidó i Closa, Dr. Xavier Batllori Aguilà

C.Claravall, 1-3
08022 Barcelona
Tel. 936 022 200
Fax. 936 022 249
E-mail: urlsc@sec.url.es
www.url.es

A mis padres, Vicente y Guadalupe
A mi hermana, Eva
A mi marido, Sergi
A mi hijito, Aleix

Far away there in the sunshine are my aspirations
I may not reach them, but I can look up and see their beauty,
believe in them and try to follow where they may lead

Louisa May Alcott

Agradecimientos

El presente trabajo se ha realizado en el Departamento de Química Orgánica del Instituto Químico de Sarriá bajo la dirección del Dr. Jordi Teixidó, a quien quisiera agradecer su constante apoyo tanto en el ámbito académico como en el personal, así como su ilusión, dedicación y amistad, y bajo la dirección del Dr. Xavier Batllori a quien quisiera agradecer su apoyo y colaboración siempre prestada a pesar de trabajar en campos distintos.

También quisiera expresar mi gratitud:

Al Dr. José Ignacio Borrell por su apoyo, ayuda, colaboración y amistad así como a todo el GEM, Grup d' Enginyeria Molecular, en particular al Dr. Santiago Nonell por su atención siempre conmigo.

Al Dr. Jordi Cuadros por su apoyo, amistad y ayuda, especialmente al inicio de la tesis, y a Sergi por su amabilidad y ayuda con los problemas informáticos.

Al Dr. Alberto Barrera por su apoyo, ayuda y amistad, tanto en el ámbito académico como personal a lo largo de toda la carrera y el doctorado.

Al Dr. Dave Ritchie por sus facilidades y atenciones durante mi estancia en el Department of Computing Science, King's College, University of Aberdeen. Asimismo agradezco toda su paciencia, su constante apoyo, dedicación y ayuda en la promoción de mi carrera científica. Thank you very much for your constant support, dedication and help in the promotion of my research career. Igualmente a todos los miembros de su equipo, en especial a Lazaros Mavridis por acogerme tan amablemente.

Al Dr. Antonio Carrieri por su colaboración y ayuda en el desarrollo final de esta tesis. Thank you very much for your collaboration and help sharing with me your broad experience in CCR5 co-receptor modelling.

A los compañeros que han pasado por el departamento de orgánica del IQS a lo largo de estos años, especialmente a Obdulia Rabal por su amistad, ayudarme y aconsejarme siempre, así como introducirme en este mundo (¡como costó el software inicial!;)), y tenías razón, el hecho de espabilarse solo es un valor añadido que al final se nota. Oscar Rey, y Rosalía Pascual, compañeros de viaje en diversos congresos, por su colaboración, ayuda y amistad. Gracias por todos tus consejos, Oscar, siempre tan oportunos. Gracias por tus sugerencias, Rosalía, tanto en el trabajo final de carrera como en las ocasiones en que hemos coincidido después.

Asimismo, gracias a todos los compañeros de los laboratorios de síntesis y fotoquímica por animar los momentos en el IQS. Gracias a Roger y Laia, integrantes de "els quatre gats" ;) de diseño molecular con los que he pasado todos estos años, por su amistad y apoyo, así como por la multitud de favores que siempre han estado dispuestos ha hacerme. Gracias a Sofia por su colaboración, amistad y ayuda tanto en los aspectos más computacionales como en los químico/experimentales. Gracias también por escucharme y aconsejarme siempre. Gracias a Gemma y Maia con las que he compartido muchas comidas el último año, por su amistad y consejos maternales.

A todos mis amigos, a los que he hablado tanto de esta tesis y me han dado siempre ánimos, en especial a Marta y Pablo, Lali y Noe.

A toda mi familia, en especial a mis abuelos que aunque hablo poco con ellos se que siempre están pensando en mi.

A mi hermana Eva, por haberme ayudado tanto en los inicios a la programación en C. Tal como le dije, todas las palabras son pocas para agradecerle lo mucho que me ha ayudado en esta tesis, tanto en el aspecto más material como en el personal. Gracias por las comidas juntas en el IQS, gracias por aquellos ratos en que dimos clase juntas a los de primero de licenciatura/ingeniería química en el IQS, gracias por las discusiones de problemas matemáticos mediante chat cuando estaba en Escocia, gracias por las tardes/noches en tu universidad de matemáticas compilando códigos, gracias por esos abrazos y besos que me daban tantos ánimos… Gracias por escucharme siempre, darme ánimos y estar siempre ahí.

A mis padres, Vicente y Guadalupe, por todo el esfuerzo que siempre habéis hecho por mí para ayudarme a conseguir mis metas y objetivos. Gracias por ayudarme siempre tanto, aconsejarme y estar a mi lado en todo momento. Gracias por ponerme desde pequeñita a estudiar, por esa constancia que me habéis inculcado, gracias por esos veranos en los que estudiábamos las asignaturas del curso juntos en algunos ratos por la mañana para que al curso siguiente me "sonara" todo desde el principio ;), gracias por la carrera de piano que aunque costó me ayudó a desarrollar más ciertos sentidos y trabajar con constancia hasta que las cosas sonaran "de maravilla", gracias por la calidad humana con que me habéis educado, gracias por todos los valores que me habéis dado, porque por todo eso he llegado a ser como soy, conseguir lo que me propongo y valorar lo que tengo. Gracias por vuestro amor y cariño.

Finalmente, gracias a mi marido, Sergi. No hay palabras para describir lo importante que has sido para mí en todo este tiempo, amor. Me has ayudado, me has dado soporte, me has aconsejado, has estado siempre a mi lado en todo momento. Gracias por ayudarme tanto mientras daba clases a los de primero, gracias por tus idas y venidas a Escocia, gracias por escuchar mis problemas, que aunque no seas químico computacional ya entiendes de la mayoría de ellos ;), gracias por tu comprensión, por tu cariño, por tus bromas para que sonriera cuando estaba estresada pensando en mis cosas ;), en definitiva gracias por quererme tanto. Cariño, ha pasado mucho tiempo desde que aprendiste lo que era un benceno, me acuerdo de aquel día estudiando orgánica "juntos" como si fuera hoy, y ahora ya nos ves, padres de Aleix, un hijito maravilloso cuyas risas nos hacen sonreír cada día y nos dan fuerzas y ánimos para superarnos día a día y conseguir nuevos retos. Y por que no, gracias a Aleix, que a pesar de que tiene 5 meses y medio, algún día leerá esto y quizás se acuerde de los muchos ratos que pasó en mi regazo mientras escribía esta tesis en sus primeros meses de vida, de cuando aprendió a tocar las teclas del ordenador y me hacía saltos de línea inesperados o me escribía letras sin querer. Gracias a ti también, cariño, por estar desde el primer día a mi lado transmitiéndome esa felicidad y alegría que te caracteriza ya desde tan pequeño. Algún día sabrás lo que eso significa para mí.

A la Generalitat de Catalunya, DURSI por una beca en el plan de formación de personal investigador (2008FI), y al Instituto Químico de Sarriá por una beca IQS (2006) y los medios que ha puesto a mi disposición.

Abstract

HIV entry inhibitors have emerged as a new generation of antiretroviral drugs that block viral fusion with the CXCR4 and CCR5 membrane co-receptors. Several small molecule antagonists for these co-receptors have been developed, some of which are currently in clinical trials. However, because no crystal structures for the co-receptor proteins are available, the binding modes of the known inhibitors within the co-receptor extracellular pockets need to be analyzed by means of site-directed mutagenesis and computational experiments. Generally, the objective of these computational approaches is to screen large numbers of candidate drug compounds rapidly. Virtual screening has recently become a useful complement to laboratory-based high-throughput screening methods for large libraries of compounds. Hence, in this thesis, a virtual screening protocol, using several receptor-based and ligand-based approaches, has been performed to find CXCR4 and CCR5 antagonists that could potentially serve as HIV entry inhibitors.

For receptor-based virtual screening, homology models of CXCR4 and CCR5 co-receptors built in our research group have been improved, and preliminary binding mode analyses using these models and high affinity known ligands have been carried out. Also, the performance in virtual screening and docking post-processing of different interaction fingerprints, compared to the results obtained with a new interaction fingerprint (APIF) developed in our research group, has been analysed.

For ligand-based virtual screening, pharmacophore modelling and several shape-based and property-based molecular comparison approaches have been compared, using high-affinity ligands as query molecules. Also, a novel consensus shape-based virtual screening approach has been developed and used to investigate and add further evidence for multiple binding sites within the CCR5 extracellular pocket hypothesis.

All the receptor-based and ligand-based methods have been firstly applied in a retrospective virtual screening, using a large database of known CXCR4/CCR5 inhibitors and similar presumed inactive molecules assembled in this thesis. For each receptor, the library has been queried using known binders, and the enrichment factors and diversity of the resulting virtual hit lists have been analyzed. Moreover, receiver-operator-characteristic analyses for both CXCR4 and CCR5 inhibitors have been carried out in order to compare the performance of the new consensus shape matching algorithm with the other screening approaches used.

Once the different virtual screening approaches have been validated and the best parameters for each one have been selected, prospective virtual screening of a combinatorial library designed by our research group and *de novo design* methods have been applied to identify new HIV entry blockers.

Sumario

Los inhibidores de entrada del VIH han surgido recientemente como una nueva generación de fármacos antiretrovirales, los cuales bloquean la unión del virus con los coreceptores de membrana CXCR4 y CCR5. Se han desarrollado diversas moléculas pequeñas antagonistas de estos coreceptores, algunas de las cuales están actualmente en fase de ensayo clínico. Sin embargo, dado que no existen estructuras cristalográficas para estos coreceptores proteicos, es necesario analizar los modos de unión de inhibidores conocidos a la cavidad de unión extracelular de los coreceptores mediante experimentos de mutagénesis dirigida y estudios computacionales. En general, el objetivo de estas aproximaciones computacionales es cribar un gran número de compuestos candidatos a fármacos rápidamente. El cribado virtual se ha convertido recientemente en un complemento útil de los métodos de cribado experimentales *high-throughput screening* para grandes librerías de compuestos. Por lo tanto, en esta tesis se ha llevado a cabo un protocolo de cribado virtual, mediante aproximaciones basadas en el receptor y en ligandos activos conocidos, con el fin de encontrar antagonistas de CXCR4 y CCR5 que puedan servir como potenciales inhibidores de entrada del VIH.

Para el cribado virtual basado en el receptor, se han mejorado los modelos de los coreceptores CXCR4 y CCR5 desarrollados en el laboratotio de diseño molecular del IQS, y se han llevado a cabo ensayos preliminares de modo de unión utilizando estos modelos y ligandos conocidos de elevada afinidad. Asimismo, se ha analizado el comportamiento en el cribado virtual y en el post-procesado de resultados de *docking* de diferentes *fingerprints* de interacción en comparación con los resultados obtenidos por un nuevo *fingerprint* de interacción (APIF) desarrollado en el laboratotio de diseño molecular del IQS.

Para el cribado virtual basado en ligandos, se han comparado modelos farmacofóricos y diversas aproximaciones basadas en la forma y propiedades moleculares utilizando ligandos de elevada afinidad como moléculas de referencia. Además, se ha desarrollado una nueva aproximación basada en la forma molecular, la cual se ha utilizado para estudiar en profundidad la hipótesis de la multi-región de unión de la cavidad de unión extracelular del coreceptor CCR5.

Todos los métodos, ya sean basados en el receptor o en ligandos conocidos, se han aplicado en primer lugar de manera retrospectiva utilizando una extensa base de datos de inhibidores de CXCR4/CCR5 y de moléculas supuestamente inactivas, similares en propiedades a los activos, recopilada en esta tesis. Para cada receptor, la quimioteca ha sido cribada utilizando inhibidores conocidos, Se han analizado los factores de enriquecimiento y la diversidad en las listas finales de *hits*. Además, se han llevado a cabo análisis ROC (*receiver-operator-characteristic*) para ambos inhibidores de CXCR4 y CCR5 con el fin de comparar la habilidad del nuevo algoritmo basado en la igualdad de formas de ligandos con el resto de aproximaciones de cribado utilizadas.

Una vez validadas las diferentes aproximaciones de cribado y seleccionados los mejores parámetros para cada una de ellas, se han aplicado las herramientas de cribado virtual de manera prospectiva sobre una quimioteca combinatoria diseñada en el laboratorio de diseño molecular del IQS, así como técnicas de diseño *de novo* de ligandos para identificar nuevos bloqueadores de la entrada del VIH en las células.

Índice

Introducción

I.1. Diseño de fármacos asistido por ordenador

El uso de métodos computacionales en el diseño de nuevos ligandos para una determinada diana terapéutica se denomina comúnmente diseño de fármacos asistido por ordenador (*Computer Assisted Molecular Design*). La genómica, proteómica y bioinformática están descubriendo nuevas dianas terapéuticas y contribuyendo cada vez más al descubrimiento de fármacos [1]. Muchas estructuras 3D de dianas terapéuticas han sido resueltas mediante diversas técnicas, como cristalografía de rayos X o resonancia magnética nuclear (RMN). A pesar de que estas estructuras sean conocidas, el diseño de fármacos candidatos que puedan interaccionar con estas dianas es una tarea difícil [2]. Por ello, el cribado de quimiotecas virtuales (*virtual screening*) es ahora un método bien establecido para la selección/identificación de nuevos *leads* o cabezas de serie.

Asimismo, el desarrollo y lanzamiento de un nuevo fármaco al mercado requiere a la industria farmacéutica una media de 12 a 20 años y unos costes de aproximadamente 850 millones de euros. Las técnicas de cribado virtual resultan baratas (ahorran en compra de reactivos y robotización), rápidas, y permiten tener en cuenta un gran número de compuestos *in silico* del orden de billones, cifra impensable experimentalmente. Típicamente, en una cascada de cribado, una quimioteca virtual que contiene unas 10^6-10^{12} estructuras es sucesivamente filtrada y reducida a una colección de unos 100-1000 candidatos [3]. Por ello, se puede considerar un buen complemento para las técnicas de *High-throughput screening* (HTS) como fuente de obtención de nuevos *leads*. Ahora bien, si no se aplican restricciones, el cribado virtual puede sugerir potenciales *hits* no accesibles sintéticamente. Sin embargo, proporciona información acerca del modo de interacción fármaco-diana y es un buen criterio para la priorización de moléculas a sintetizar.

Existen dos aproximaciones al cribado de compuestos. Cuando se dispone de la estructura tridimensional de la diana terapéutica, bien obtenida por métodos experimentales (cristalografía de rayos X o RMN) o bien a través de la construcción de modelos moleculares, se sigue la metodología de diseño de ligandos basado en su estructura o *structure-based*. Se incluyen aquí las técnicas de *docking* (intento de encontrar el "mejor" acoplamiento entre dos moléculas: un receptor y un ligando). En caso contrario o en caso de que se prescinda, el cribado virtual de ligandos se puede realizar mediante búsquedas *ligand-based*, basadas en el análisis y comparación de propiedades moleculares y datos de afinidad por el receptor para ligandos conocidos, sin tener en cuenta la estructura de dicho receptor. Se incluyen aquí las técnicas de búsqueda de similitud mediante descriptores 2D/3D, QSAR (*Quantitative Structure-Activity Relationship* o relación cuantitativa de los cambios estructurales de un conjunto de compuestos con los cambios en actividad), desarrollo de modelos farmacofóricos (identificación del conjunto de características estructurales de un ligando directamente relacionadas con los sitios claves de interacción de un receptor), y técnicas de *shape matching* (superposición de la forma de dos ligandos). Por lo tanto, las herramientas de cribado virtual requieren inevitablemente conocer la actividad de algunos compuestos o la estructura de la diana biológica.

Dichas aproximaciones se pueden llevar a cabo mediante un cribado virtual retrospectivo, en el que se utilizan compuestos activos conocidos, con el fin de validar las herramientas de cribado sobre la química concreta de la diana farmacológica de estudio, y seleccionar los mejores parámetros con los que se deben utilizar dichas herramientas. Una vez se dispone de un buen protocolo de cribado,

se puede realizar un cribado virtual prospectivo, en el que las técnicas de cribado se aplican a un caso real con el fin de identificar nuevos compuestos activos.

Una cascada típica de cribado virtual contiene diferentes pasos de filtrado que conllevan a una reducción del número de compuestos candidatos a ser examinados experimentalmente, como mínimo en nueve órdenes de magnitud, acabando con unos 1000 compuestos para su ensayo clínico. Dichos filtros se aplican de forma secuencial de acuerdo con el nivel de requerimientos computacionales que utiliza cada una de las técnicas (de menor a mayor coste de cálculo) y la complejidad de la información necesaria para cada una de ellas. Para la identificación de *hits* es necesario partir del análisis conformacional de las estructuras de la quimioteca generada. A partir de ellas se realiza un pre-cribado en el que se seleccionan las moléculas que poseen estructura y propiedades de fármaco (*drug likeness*), así como propiedades ADMET (Absorción, Distribución, Metabolismo, Eliminación y Toxicidad) de interés con el fin de optimizar simultáneamente la potencia y la farmacocinética. Después se aplican filtros de similitud 2D/3D para seleccionar compuestos focalizados hacia un tipo de estructura concreta necesaria para la interacción con la diana de estudio. El siguiente paso acostumbra a ser la búsqueda de modelos farmacofóricos con los que seleccionar compuestos de manera aún más focalizada según unas características estructurales y tridimensionales concretas. Una vez se ha focalizado suficiente la quimioteca se procede con la fase de descubrimiento y optimización de *leads*. Para ello se aplican técnicas de *docking*. Se utiliza información de la estructura de la proteína diana para muestrear el complejo ligando-receptor y una función de *scoring* para evaluar la interacción ligando-macromolécula [4]. Así mismo, las mejores configuraciones obtenidas por *docking* de todas las moléculas de la base de datos en la estructura de un receptor se pueden trasladar a una cadena de bits derivada del número de residuos/átomos en la cavidad de unión de la proteína, lo que se denomina *fingerprint* de interacción (*interaction fingerprint*), con el fin de analizar las interacciones proteína-ligando. Finalmente se post-procesa la lista de *hits* obtenidos, se aplican técnicas de *clustering* (o agrupación de moléculas según su similitud) y se evalúan los *leads* hallados [5].

En esta tesis se ha utilizado un protocolo de cribado el cual comprende las técnicas básicas presentes en una cascada típica de cribado virtual (véase Figura I.1). Primero se compila una base de datos de compuestos activos conocidos, compuestos tipo fármaco para ser cribados y supuestos compuestos inactivos con propiedades 1D similares a las de los activos para evitar tendencias en los resultados debidas a grandes diferencias en propiedades básicas (masa molecular, número de átomos dadores o aceptores, número de enlaces simples rotables, número de átomos hidrofóbicos, y coeficiente de partición octanol-agua). Se aplican filtros *drug likeness* a los compuestos de la base de datos. Después se aplican filtros *docking-based* e *interaction fingerprints* previo modelado del receptor, análisis del sitio activo de la proteína, y predicción de la unión ligando-receptor, y filtros *ligand-based* (*shape matching*, búsquedas de similitud, propiedades ADME, QSAR, y modelos farmacofóricos). Finalmente todos los compuestos se ordenan en listas según la puntuación (*score*) obtenida en cada aproximación. Asimismo se aplican técnicas de consenso para seleccionar las moléculas a ser sintetizadas y testadas. Por otra parte se aplican técnicas de diseño *de novo* (*de novo design*) para construir nuevas moléculas inhibidoras de las dianas de estudio. En las secciones siguientes se detallan brevemente cada uno de estos pasos.

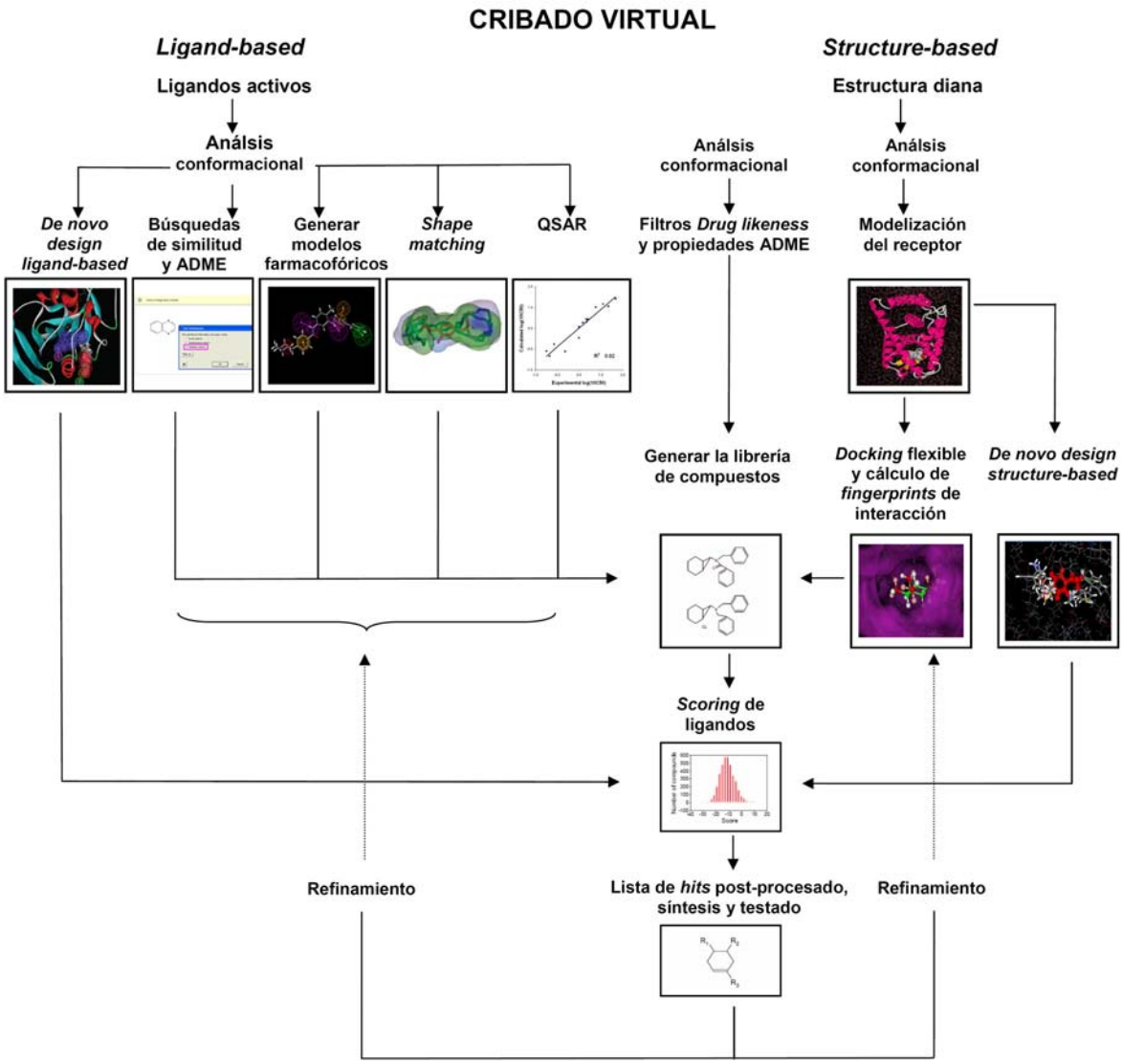


Figura I.1 Protocolo de cribado virtual utilizado.

I.1.1. Pre-filtrado: filtros *Drug Likeness* y propiedades ADMET

Como primer paso en un cribado virtual se utilizan filtros que eliminen aquellas estructuras que posean propiedades de no fármaco en cuanto a los grupos funcionales que presentan y a sus propiedades físicas (*drug likeness*) [6, 7]. Ahora bien, no todos los fármacos actuales satisfacen estos criterios. Los filtros utilizados para considerar un compuesto tipo fármaco (*drug-like*) son los siguientes:

a) La "regla de los cinco" de Lipinski [8] filtra las moléculas en función de su peso molecular (≤500g/mol), su lipofilia, medida en función del coeficiente de partición octanol-agua (LogP) (≤5) y el número de dadores (≤5) y aceptores (≤10) de puente de hidrógeno. Se considera que un compuesto que no satisfaga dos o más de estos criterios, tiene una baja probabilidad de convertirse en un buen fármaco. Dichos márgenes parecen demasiado

estrictos, por lo que se han realizado estudios que establecen márgenes de variabilidad para estos descriptores [9].

b) Presencia de grupos funcionales de fármacos establecidos [10].
c) Eliminación de grupos funcionales tóxicos o muy inestables [11].
d) Uso de árboles de decisión [12], redes neuronales [13] y algoritmos genéticos [14] para clasificar los compuestos de bases de datos como *drug-like* o *non drug-like*.
e) Cálculo de propiedades ADMET (Absorción, Distribución, Metabolismo, Eliminación y Toxicidad), como son la capacidad de atravesar la barrera hematoencefálica (BBB), predicción del metabolismo mediado por el citocromo P450, unión a la albúmina, solubilidad en agua y en DMSO... [15]

I.1.2. Cribado virtual basado en ligandos (*Ligand-Based Virtual Screening*)

El cribado virtual *ligand-based* se basa en que moléculas estructuralmente relacionadas deberían mostrar actividades biológicas similares [16]. Ahora bien, existen puntos críticos como el hecho de que en ocasiones pequeños cambios estructurales conducen a un gran cambio en la actividad del compuesto o que moléculas similares a veces muestren modos de unión diferentes [17, 18]. A pesar de ello, estas técnicas se han mostrado de gran utilidad dado su bajo coste computacional cuando no se dispone, o se prescinde, de la información contenida en la estructura del receptor. A continuación se comentan brevemente dichos métodos.

Las **búsquedas de similitud** parten de una o varias estructuras diana y su descripción por uno o más descriptores estructurales, junto con la de los compuestos candidatos contenidos en la quimioteca virtual [19, 20]. Los descriptores utilizados para caracterizar dichas quimiotecas virtuales están clasificados como 1D, los cuales únicamente especifican el tipo atómico; 2D, que incluyen información topológica, es decir, la conectividad de la molécula, y 3D, cuando contemplan la estructura tridimensional de la molécula [21] (véase Sección 1.5). En lo que se refiere a la medida de similitud, existen diferentes coeficientes de similitud y distancia (véase Artículo IV) [19]. Dichas métricas se comportan mejor o peor en función del conjunto de descriptores utilizado y de las moléculas a comparar. No existe un consenso en el uso tanto de descriptores como coeficientes de similitud por lo que se han llevado a cabo diferentes estudios dirigidos a establecer una combinación de descriptores/coeficientes óptima para la búsqueda de similitud [22] o criterios para la validación de dichos descriptores [23].

El **diseño de modelos farmacofóricos** es una técnica muy útil cuando se dispone de una serie de compuestos activos. Se basa en la identificación del ordenamiento tridimensional común de los sitios de interacción claves con un receptor a partir de un conjunto accesible de conformaciones de un grupo de ligandos activos [24]. Un farmacóforo se puede definir pues, como un conjunto de características estructurales en un ligando que están directamente relacionadas con el reconocimiento del ligando en el sitio de unión del receptor y, por lo tanto, con su actividad biológica. A un ligando activo se le asigna un conjunto de puntos en el espacio que reflejan la presencia o ausencia de características farmacofóricas (componentes esenciales para el reconocimiento molecular, es decir, los componentes de una molécula que la hacen activa). Ello da lugar a lo que se denomina *query* o plantilla con la que se "interrogará" la quimioteca (Figura I.2).

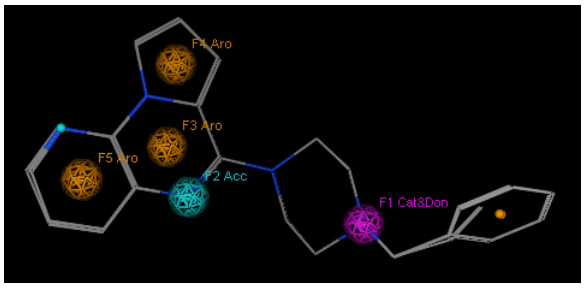


Figura I.2 *Query* farmacofórica, la cual contiene tres centros aromáticos, un aceptor de puente de hidrógeno, y un átomo catiónico o dador de puente de hidrógeno como características farmacofóricas (*features*). Se muestra también el alineamiento de una conformación *hit* representada en *lines*.

La obtención de farmacóforos parte del alineamiento de moléculas activas para superponer e identificar todos los grupos farmacofóricos conservados entre ellas y así obtener la configuración espacial de las características químicas clave, responsables de la interacción con el receptor. Los grupos farmacofóricos comúnmente utilizados son átomos con cargas positiva y negativa, dadores y aceptores de puente de hidrógeno y átomos con carácter hidrofóbico. Una vez se obtiene el modelo, se puede utilizar para buscar en bases de datos de moléculas previamente alineadas frente a la hipótesis farmacofórica seleccionada, compuestos que contengan el mismo farmacóforo, para explicar relaciones de estructura-actividad o como punto de partida para el diseño de nuevas moléculas potencialmente activas. A la hora de diseñar el modelo farmacofórico cabe tener en cuenta la flexibilidad molecular de los activos de partida (plantillas) y su superposición, así como la conservación y tolerancia de los modelos farmacofóricos, es decir, el número de características farmacofóricas que ha de estar presente en todas o en parte de las moléculas alineadas. En este sentido, aparte de la generación manual de hipótesis, se han desarrollado programas para derivar automáticamente hipótesis, basados en superposiciones y alineamientos múltiples, aunque sigue siendo necesaria la intervención del usuario para seleccionar la mejor propuesta (véase Sección 1.6).

Los **métodos QSAR (*Quantitative Structure-Activity Relationship*)** permiten relacionar cuantitativamente los cambios estructurales de una serie de compuestos con los cambios en la actividad. Actualmente, se utilizan múltiples descriptores de la estructura química combinados con la aplicación de técnicas de optimización lineales y no lineales (algoritmos genéticos, redes neuronales…) para derivar modelos. El 3D-QSAR utiliza descriptores espaciales y técnicas de análisis multivariante PLS (*partial least squares*). Se utilizan los descriptores de campo molecular, basados en describir las interacciones receptor-ligando a través de potenciales de interacción molecular (*Molecular Interaction Potential*, MIP). Los MIP se calculan a partir de una malla o *grid* que engloba todos los compuestos alineados sobre un mismo marco de referencia, y donde en cada punto se sitúan distintos grupos químicos o sondas.

Las **aproximaciones *shape matching*** se basan en la comparación/superposición de la forma tridimensional de un conjunto de moléculas frente a una molécula activa conocida (Figura I.3). La forma tridimensional de una molécula activa frente a una diana específica es la adecuada/complementaria para la interacción con el sitio activo de dicha diana, por lo tanto los compuestos que posean una forma tridimensional similar a la del activo conocido (molécula plantilla, *shape-matching query*) tendrán mayor probabilidad de encajar en el receptor biológico y consecuentemente mayor actividad (véase Sección 1.7). El mayor problema asociado a estas técnicas es la selección de la conformación de la *query*, dado que si se utiliza una molécula activa u otra, o una conformación de dicha molécula activa u otra, la forma tridimensional utilizada como *query* sobre la que han de solaparse los compuestos de la base de datos a cribar puede ser bastante diferente. Normalmente se utiliza como *query* la conformación cristalográfica del ligando

complejado; ahora bien, si se trabaja con una diana de la que no se dispone información cristalográfica, se deben recurrir a métodos computacionales. Se suele utilizar en estos casos la conformación calculada de menor energía, o bien calcular diferentes conformaciones como *queries* y utilizar las más similares en cada caso a los compuestos superpuestos. Asimismo, si una diana posee un sitio activo en el que pueden encajar ligandos con diferentes modos de unión, la *query* seleccionada será representante de un solo modo de unión, y por lo tanto, solo el conjunto de compuestos de la base de datos a cribar con dicho modo de unión se superpondrán correctamente con la *query*. De esta manera, compuestos con diferente modo de unión que podrían ser activos no serán seleccionados por no tener elevada similitud con la conformación de la *query* seleccionada. En esta tesis se ha desarrollado una nueva aproximación *shape matching* para mejorar el tratamiento de estos problemas.

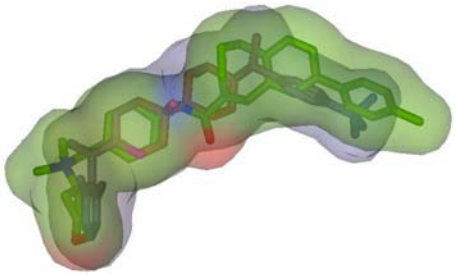


Figura I.3 *Shape matching query* representada en *lines* del color de los elementos (azul, nitrógeno; verde, carbono; rojo, oxígeno). Superposición sobre la *query* de una molécula representada en *lines* de color rojo/azul. Se observa el solapamiento de la forma tridimensional de la *query* y la molécula superpuesta.

I.1.3. Cribado virtual basado en el receptor (*Structure-Based virtual screening*)

El cribado virtual *structure-based* utiliza la estructura del receptor para explorar el espacio químico identificando ligandos de bases de datos de compuestos orgánicos mediante *docking*, o bien para diseñar *de novo* compuestos a partir de la complementariedad con el sitio de unión de dicho receptor [7, 25].

Las **técnicas de *docking*** evalúan la actividad potencial de una quimioteca de compuestos, posicionada en el sitio de unión del receptor, a partir de la interacción proteína-ligando (Figura I.4). Asimismo, dichas técnicas permiten la identificación del modo de unión, es decir, la orientación y conformación que el ligando adopta en la cavidad de la proteína, y, menos frecuentemente, se utilizan para identificar el sitio de unión (*blind docking*) [26]

Un protocolo de *docking* se caracteriza tradicionalmente por dos aspectos: el *docking* o método seguido para muestrear el espacio conformacional del complejo ligando-receptor, y la función de *scoring* utilizada para evaluar la afinidad de la interacción ligando-macromolécula [4]. La parte más conflictiva es la función de *scoring* para predecir la afinidad de la unión macromolécula–ligando. (véase Sección 1.4).

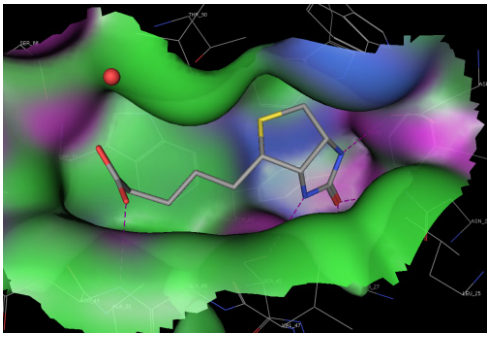


Figura I.4 *Docking* de un ligando, representado en *lines*, en el sitio activo de la proteína, mostrado mediante una superficie de interacción de van der Waals coloreada por regiones: unión por puente de hidrógeno (lila), hidrofóbica (verde), y polar (azul).

I.1.4. Combinación de métodos basados en el receptor y en ligandos

Existen métodos que combinan la información procedente del receptor y de los ligandos. Una aproximación es la introducción en el *docking* del modo de unión al receptor, extraída de complejos co-cristalizados con otros ligandos, de átomos prueba o de grupos funcionales. Ello se conoce como *docking* dirigido directamente (*direct guided-docking*) [27]. Cabe tener en cuenta que normalmente se acepta que el modo de unión se conserva entre distintos ligandos, aunque no siempre esta afirmación se cumple. Otra aproximación es la construcción de modelos farmacofóricos teniendo en cuenta tanto las características de ligandos activos como la estructura del sitio activo del receptor, especialmente si se dispone de complejos ligando-proteína co-cristalizados. Se combinan las características farmacofóricas extraídas del alineamiento de ligandos activos conocidos con las extraídas de la generación de un mapa de interacciones del sitio activo del receptor (dadores de puente de hidrógeno, aceptores de puente de hidrógeno y regiones hidrofóbicas). Otro tipo de aproximaciones se basan en codificar los contactos 3D proteína-ligando en cadenas de bits derivadas del número de residuos/átomos en la cavidad de unión de la proteína. Cada bit denota la presencia (1) o ausencia (0) de una interacción particular: puente de hidrógeno, contacto hidrofóbico o de van der Waals. Reciben el nombre de *fingerprints* de interacción (*Interaction Fingerprints*). La implementación de estos *fingerprints* puede variar dependiendo de la definición de la cadena de bits y el tipo de interacciones consideradas. Las interacciones de un ligando co-cristalizado con un receptor diana se trasladan a un *fingerprint* de interacción, el cual se utiliza como referencia para la comparación con los *fingerprints* de interacción extraídos de las mejores configuraciones obtenidas por *docking* de todas las moléculas de la base de datos a cribar (véase Artículo IV).

I.1.5. Técnicas de diseño de novo (*De novo design*)

Las técnicas de diseño *de novo* [28] permiten diseñar nuevas moléculas a partir de la zona activa del receptor biológico de interés o un farmacóforo tridimensional. Se distinguen principalmente los métodos basados en energía y los basados en bases de datos de conocimiento o *knowledge-based*. Los primeros se basan en colocar fragmentos en la zona activa y permitir una exploración de ésta por minimización de energía o técnicas de simulación. O bien colocar los fragmentos sucesivamente en los vértices de una malla de puntos alrededor de la zona activa y calcular la energía de interacción del fragmento prueba con el receptor. Los *knowledge-based methods* emplean bases de datos con información de geometrías e interacciones preferentes obtenidas mediante el análisis de bases de datos cristalográficas. Identifican las regiones de la zona activa donde es favorable la unión de fragmentos mediante interacciones como la formación de puentes de hidrógeno o interacciones hidrofóbicas, y posteriormente posicionan los fragmentos con la disposición geométrica adecuada. Estos métodos permiten también la generación de compuestos a partir de farmacóforos (véase Sección 1.8). Tras la colocación de los fragmentos en el sitio de unión por cualquier tipo de método, es necesaria su interconexión. Para ello se suelen emplear bases de datos de conectores predefinidos (*building blocks*) y algoritmos de tipo heurístico. Dicha etapa es la más problemática, dada la cantidad de posibilidades de interconexión y la dificultad en evaluar la facilidad sintética. A pesar de que no existe una buena valoración de facilidad sintética acoplada a los programas de diseño *de novo*, sí se han desarrollado programas de análisis retrosintético, los cuales a partir de una molécula objetivo proponen distintas vías sintéticas para su obtención [29], con lo que pueden ayudar a filtrar los resultados obtenidos por técnicas *de novo design*.

I.2. Aplicación de las técnicas de cribado virtual a inhibidores de entrada del VIH

El presente trabajo se desarrolla en el Grupo de Ingeniería Molecular, GEM, en el Instituto Químico de Sarriá y dentro del marco de la línea de investigación dirigida hacia el diseño, síntesis y evaluación de la actividad antiviral de nuevos compuestos anti-VIH que lleva a cabo el departamento de Química Orgánica (laboratorios de diseño molecular y síntesis) del Instituto Químico de Sarriá en colaboración con el laboratorio de retrovirología de la Fundación IrsiCaixa en el Hospital Germans Trias i Pujol en Badalona. Dicha línea de investigación se inició en el departamento hace ocho años con el diseño y selección de quimiotecas combinatorias de inhibidores de la transcriptasa inversa del tipo HEPT (derivados de 1-[(2-hidroxietoxi)metil]-6-(feniltio)timina) [30]. Con ello se intentaba inhibir el ciclo de reproducción viral del Virus de Inmunodeficiencia Humana (VIH) una vez el virus ya ha infectado la célula, en la etapa de transcripción inversa, de tal manera que los compuestos diseñados inhibieran la transcriptasa inversa y ésta no pudiera realizar sus funciones de actividad DNApolimerasa (síntesis de una cadena sencilla de DNA complementaria al RNA viral, que dirige la producción de su complementaria) y actividad ribonucleasa (degradación de la secuencia de RNA original del virus). Este trabajo permitió el desarrollo del programa Pralins (*Program for Rational Analysis of Libraries in silico*) [31], como herramienta computacional para la selección de quimiotecas combinatorias y la colaboración con el Prof. Dr. Johann Gasteiger (Computer-Chemie-Centrum and Institute for Organic Chemistry, University of Erlangen-Nürnberg) [32]. Poco después el GEM se centró en el ataque directo de la entrada del virus a las células, una diana farmacológica poco estudiada hasta ese momento. De esta manera no se trataba la infección una vez el virus había entrado en las células sino que no se permitía que las células fueran infectadas. Para ello se diseñaron y sintetizaron potenciales moléculas inhibidoras de los coreceptores CXCR4 y CCR5, receptores celulares transmembrana que forman un complejo con la proteína CD4 celular al cual se une la unidad glicoproteica gp120 del virus. Ello dio lugar a la colaboración con el Laboratorio de Retrovirología de la Fundación IrsiCaixa en el Hospital Germans Trias i Pujol (el cual participó en el descubrimiento y caracterización del primer agente antagonista de CXCR4 [33]) mediante el proyecto "Disseny, síntesi i avaluació de l'activitat antiviral de nous compostos anti-VIH" financiado (2003-5) por la edición SIDA-2001 de la Fundació La Marató de TV3 (Corporació Catalana de Ràdio i Televisió, CCRTV). Con los resultados de este proyecto se publicó una patente referente a compuestos inhibidores del coreceptor CXCR4 [34]. En lo que respecta al diseño molecular, se construyó el modelo inicial para los coreceptores CXCR4 y CCR5 [30, 35] y se empezaron a aplicar técnicas de cribado virtual a otros casos de estudio [35]. Ello dio lugar a la colaboración con el Prof. Dr. Gisbert Schneider (Institut für Organische Chemie & Chemische Biologie, Johann Wolfgang Goethe-Universität) [36] y la Prof. Dra. Valery Gillet (Department of Information Studies, University of Sheffield) [37]. El presente trabajo se inició en el año 2005, formando parte del último año del proyecto de la Fundació La Marató de TV3, siguiendo posteriormente como parte de un nuevo proyecto dentro de la misma línea de investigación, "Diseño y síntesis de nuevos inhibidores de entrada del VIH" financiado (2007-10) por el Programa Nacional de Biomedicina (Ministerio de Educación y Ciencia, SAF2007-63622-C02-01), en colaboración también con el Laboratorio de Retrovirología de la Fundación IrsiCaixa. Dicho trabajo ha llevado a la colaboración con el Prof. Dr. Dave Ritchie, inicialmente en Department of Computing Science, University of Aberdeen, y después en Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA), Vandoeuvre-les-Nancy, así como la colaboración con el Prof. Dr. Antonio Carrieri (Dipartimento Farmaco-Chimico, Università degli Studi di Bari). En el marco de esta línea y en el contexto del diseño molecular, esta tesis intenta dar respuesta a

diferentes aspectos metodológicos planteados en relación con la caracterización estructural de los coreceptores CXCR4 y CCR5, las interacciones con sus ligandos y el diseño de éstos por medio de métodos computacionales. Para ello, se lleva a cabo un protocolo de cribado virtual retrospectivo, con el fin de establecer y validar los parámetros que mejor se ajustan, para la posterior identificación de compuestos con potencial actividad inhibidora de los coreceptores CXCR4 y CCR5 en un análisis prospectivo, el cual permita priorizar los candidatos a ser sintetizados y posteriormente testados en el laboratorio de retrovirología de la Fundación Irsi-Caixa. Dicho protocolo incluye varias de las estrategias expuestas previamente.

Teniendo el cuenta estos antecedentes, el interés farmacológico de esta tesis, en concordancia con el proyecto de investigación en que está enmarcada, es el bloqueo de los coreceptores CXCR4 y CCR5, los cuales pertenecen al grupo de receptores acoplados a proteínas G (GPCRs, *G Protein-Coupled Receptors*), para evitar la entrada del Virus de Inmunodeficiencia Humana (VIH) a las células. De esta forma, el virus no infectaría nuevas células y el propio sistema inmunológico eliminaría la población celular previamente infectada. Las siguientes secciones describen el Síndrome de Inmunodeficiencia Adquirida, la estructura del VIH que lo provoca, su ciclo vital, el papel de los coreceptores en la entrada del VIH, su interés terapéutico, su caracterización estructural, y los diferentes mecanismos de inhibición desarrollados.

I.2.1 Síndrome de Inmunodeficiencia Adquirida (SIDA)

El SIDA (Síndrome de Inmunodeficiencia Adquirida) es una enfermedad del sistema inmunitario producida por el Virus de Inmunodeficiencia Humana (VIH), el cuál causa una grave inmunodepresión provocando que el organismo no sea capaz de ofrecer una respuesta inmune adecuada contra las infecciones.

El Síndrome de Inmunodeficiencia Adquirida constituye sin duda la primera pandemia de la segunda mitad del siglo XX. Detectado en 1981, sus orígenes hay que buscarlos en África central donde probablemente se produjo la primera infección de un ser humano, posiblemente ya en la década de 1950. Desde esa zona se propagó al Caribe y posteriormente a Estados Unidos y Europa. Hoy en día, de acuerdo con el Programa Conjunto de las Naciones Unidas sobre el VIH/SIDA (ONUSIDA) y la Organización Mundial de la Salud (OMS), existe un total de 33,2 millones [30,6–36,1 millones] de personas infectadas por VIH. Un total de 30,8 millones [28,2–33,6 millones] de adultos infectados, 15,4 millones [13,9–16,6 millones] de mujeres infectadas, y 2,1 millones [1,9-2,4 millones] de niños menores de 15 años infectados (datos de 2007) [38]. En 2007 se registraron 2,5 millones [1,8–4,1 millones] de casos nuevos y 2,1 millones [1,9–2,4 millones] de muertes.

El sistema inmunitario está constituido por dos grandes grupos de células: los linfocitos (células B, células T y células asesinas) y los fagotitos (macrófagos, monocitos y neutrófilos). El colapso total de las defensas inmunitarias de las víctimas del SIDA se debe en gran parte a la reducción del número de linfocitos T4 (también denominados CD4), uno de los muchos tipos celulares que constituye el sistema inmunitario. Una vez el virus se encuentra en el interior del organismo, su objetivo son las células que expresan en su superficie la proteína CD4, como los linfocitos T4, monocitos y macrófagos. El descenso de la población de células T4 se debe, obviamente, a la muerte de las células infectadas, que a su vez interrumpe la normal proliferación de los linfocitos, inherente a su función inmunológica. En condiciones normales, cuando la célula T4 interactúa con un macrófago, ésta se activa. La célula T4 produce sustancias que estimulan la maduración de otro tipo de linfocitos, las células B. Cuando madura la célula B se transforma en célula plasmática

especializada en excretar anticuerpos, tarea para la que puede recibir ayuda de las T4. Otras señales emitidas por la célula T4 desencadenan un proceso de maduración de un segundo subtipo de células T, las T8, que atacan y matan a las células infectadas por patógenos. Cuando se ha controlado la infección, las células T4 se encargan también de suprimir la maduración de más células B y T8. Finalmente, y para conferir una última protección, a partir de las células T4 se forma un clon de "memoria" inmunológica (compuesto por unos mil descendientes), que circulan por la sangre dispuestos a reconocer un patógeno específico y reaccionar ante su presencia.

La Figura I.5 esquematiza el proceso de activación de las células T4, las cuales estimulan a las células B para que secreten anticuerpos contra un antígeno vírico. Una célula T4 se activa por medio de la interleuquina (IL-1) tras reconocer el complejo que forman el antígeno y una proteína del complejo principal de histocompatibilidad de la clase II (MHC II), presentada por un macrófago o otra célula de presentación del antígeno. La célula T4 se une entonces a una célula B que haya reconocido también al antígeno sobre una célula que lo exhiba. El contacto con la célula T4 estimula la maduración, multiplicación y diferenciación de la célula B en un clon de células de memoria y un clon de células plasmáticas que secretan anticuerpos; éstos se unen al virus rodeándolo e inactivándolo. Las linfoquinas secretadas por la célula T4 colaboran en la maduración. Las células B pueden reconocer también antígeno libre en solución en la sangre o en la linfa (Figura I.5 arriba, a la izquierda) [39].

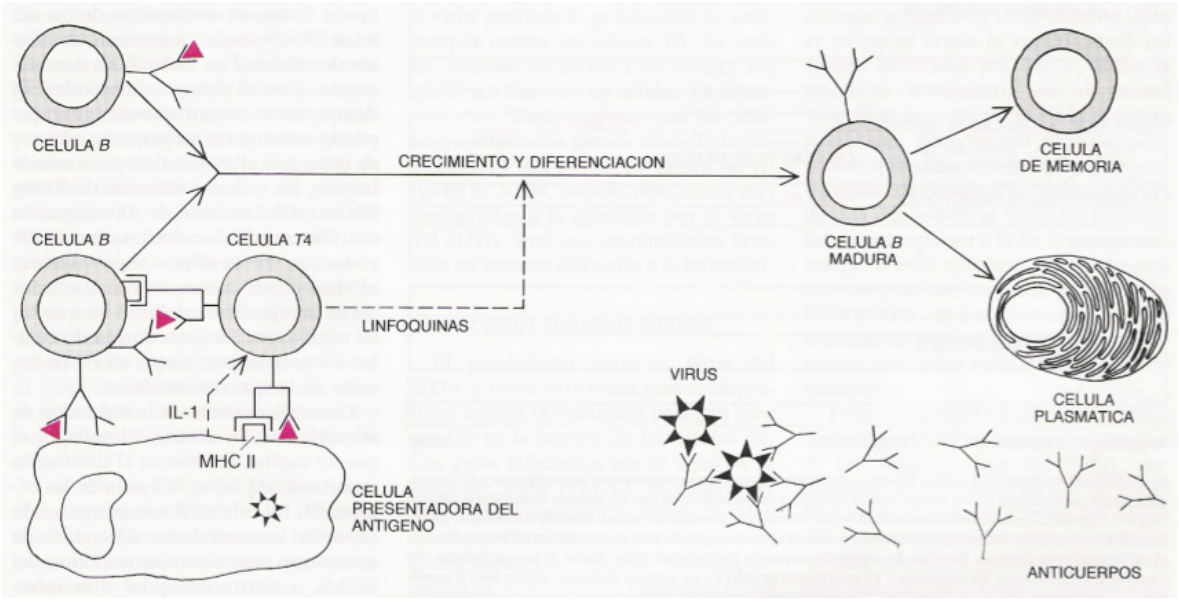


Figura I.5 Activación de las células T4, las cuales estimulan a las células B para que secreten anticuerpos contra un antígeno vírico.

Cuando la célula T4 que se activa está infectada por el virus de SIDA, el resultado es bien distinto. En vez de producir un clon de un millar de descendientes, la célula T infectada forma un clon de una decena escasa de células. Cuando éstas llegan a la sangre y las estimula un antígeno, empiezan a producir virus y mueren. Así pues, la presencia del virus de SIDA provoca la muerte de los linfocitos infectados y el frenado del clon que confiere memoria inmunológica [40].

La Figura I.6 muestra la formación de un clon de un millar de células T4 en comparación con un clon formado por una decena escasa de células cuando la célula T4 que se activa está infectada. Los

cinco dibujos de la fila superior recogen la formación de un clon normal de células T4 de memoria inmunológica. Antes de la infección, la célula T se encuentra en reposo (*1*). Durante la infección, los macrófagos excretan la proteína IL-1 y presentan un antígeno (proteína del organismo invasor) a la célula T4. Se activa entonces la célula T4 y se expresan algunos de sus genes, entre ellos los que determinan el factor IL-2 y su receptor (*2*). La célula activada excreta IL-2 y en su superficie aparecen receptores para aquella proteína (*3*). La unión IL-2-receptor (*4*) desencadena un proceso de proliferación que culmina en la formación de un clon integrado por unas mil células, que constituyen la memoria inmunológica (*5*). Todas las células del clon están preparadas para reaccionar contra el antígeno que inició el proceso. Los dibujos de la fila inferior representan lo que ocurre en la activación de una célula infectada. El ADN vírico integrado ya en los cromosomas de la célula se conoce como provirus (*1*). La interacción con el macrófago (*2*) provoca la activación de los genes celulares y del provirus. Se sintetizan ARN y proteínas víricas (*3*), que se auto ensamblan y forman partículas que abandonan posteriormente la célula, a menudo matándola (*4*). En consecuencia, la memoria inmunológica se reduce a un clon de unas diez células (*5*). El modelo explica la pérdida de la inmunidad gobernada por células T que padecen los individuos infectados por el virus. El proceso de destrucción total de todas las células T4 sin aplicación de ningún tratamiento es de unos diez años aproximadamente [41].

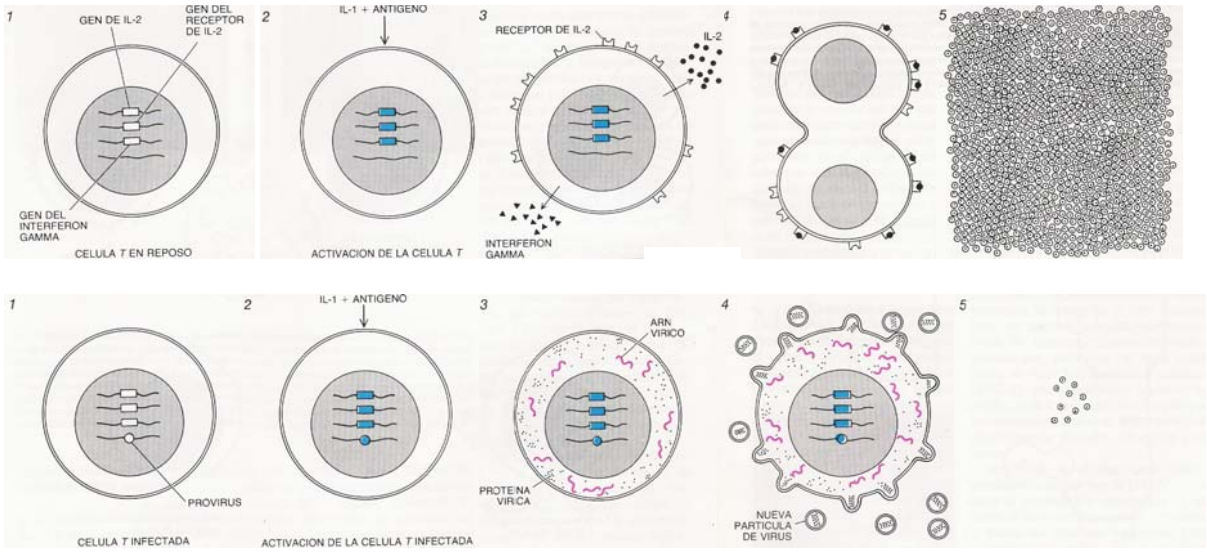


Figura I.6 Formación de un clon de un millar de células T4 de memoria inmunológica (fila superior) en comparación con la formación de un clon de una decena escasa de células en el caso de que la célula T4 que se activa esté infectada (fila inferior).

I.2.2 El virus de inmunodeficiencia humana

El virus del SIDA es un retrovirus, perteneciente a una gran familia de lentivirus de ácido ribonucleico (ARN) caracterizados por su asociación con enfermedades inmunosupresivas o del sistema nervioso central y por sus largos tiempos de incubación desde la infección hasta que se manifiesta la enfermedad.

Se conocen dos tipos de VIH: VIH-1 y VIH-2, de los cuales el primero es el predominante y el que recibe el nombre genérico de VIH. El VIH-2 es el causante de infecciones principalmente en el África Occidental y se diferencia del VIH-1 por la secuencia genómica (sólo tienen un 40% de homología de secuencia), las propiedades antigénicas, en el tamaño de sus proteínas y en que se

transmite peor. Además, dentro del tipo VIH-1 existen tres grupos: M, N y O, y a su vez, el M contiene 10 subtipos que van de la A a la J.

En unas condiciones idóneas, se considera que el VIH es una partícula esférica con un diámetro entre 80 y 110 nanómetros. Esta partícula presenta tres capas concéntricas:

- Capa interna. Contiene el "núcleo" del virión, formado por el ARN del virus, las enzimas retrotranscriptasa o transcriptasa inversa (que cataliza la síntesis de ADN vírico), integrasa y proteasa y la nucleoproteína p24. El genoma del VIH incluye tres genes principales (*env*, *gag* y *pol*) que codifican los principales elementos estructurales y funcionales del mismo.

- Capa intermedia. Es una nucleocápside icosahédrica que contiene la proteína p18.

- Capa externa o envoltura. Es una bicapa lipídica que procede de la membrana externa de la célula huésped; está constituida por la inserción de glicoproteínas del virus, cada una de las cuales posee dos componentes, gp41, que atraviesa la membrana de un lado a otro, y gp120, que sobresale de ella, (gp, abreviatura de glicoproteína) y por una alta concentración de proteínas celulares entre las que destacan antígenos de histocompatibilidad de clases I y II (HLA I y II). La cubierta glicoproteica de la capa externa desempeña un papel destacado tanto en la entrada del virus en su célula huésped como en la muerte de ésta. Algunas regiones de la proteína son comunes a todas las estirpes víricas o presentan un grado intermedio de variabilidad (color oscuro en la Figura I.7), otras son muy variables (color claro en la figura). La entrada en la célula depende de la interacción entre una o más de las regiones constantes y algunas moléculas de la membrana celular. La proteína de la cubierta participa también en la salida de las nuevas partículas víricas, durante la cual se abren huecos en la superficie celular.

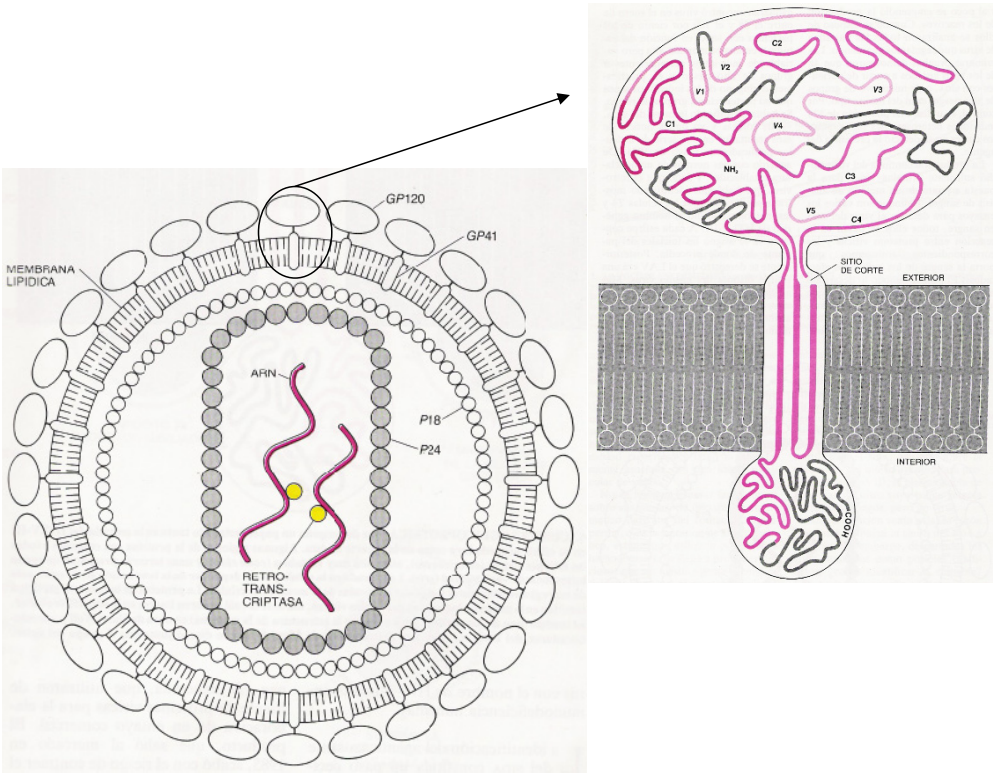


Figura I.7 Estructura del VIH y detalle de la cubierta glicoproteica [42].

Las diferentes etapas del ciclo vital del VIH[43], representadas en la Figura I.8, son:

1) Acoplamiento y fusión de la membrana

Para que el VIH penetre en la célula se debe producir la fusión de las membranas viral y celular. La entrada del VIH-1 en la célula se produce por la interacción del virus con al menos dos tipos de receptores. Primeramente, el virus interacciona con la proteína CD4 de la superficie celular. Se cree que esta proteína CD4 es específica, y que la afinidad de la gp120 viral por la CD4 es mayor que la afinidad de ésta por su ligando natural, una molécula del complejo mayor de histocompatibilidad de clase II. A continuación, el virus interacciona con los coreceptores celulares del tipo CC o CXC, según unan quimiocinas (proteínas pequeñas solubles que inducen inflamación al dirigir la migración de glóbulos blancos a sitios de infección) con los residuos de cisteína (C) cercanos al extremo *N*-terminal en posición adyacente o con las dos primeras cisteínas separadas por un único residuo (X), respectivamente. El coreceptor CCR5 es fundamentalmente utilizado por las cepas del VIH con tropismo por los monocitos, es decir, que muestra preferencia por infectar las células del sistema monocito-macrófago. Estas cepas son denominadas monocitotrópicas, M-trópicas o cepas R5. Por otra parte, el CXCR4 es utilizado por las cepas que presentan linfocitotropismo. Estas cepas son denominadas T-trópicas o X4. Se cree que las cepas inductoras de sincitios (agrupación de células cuyas membranas se han unido) pueden utilizar ambos coreceptores, aunque fundamentalmente utilicen los CXCR4, estas cepas son denominadas duales o R5X4.

Una vez tiene lugar la interacción entre la gp120 viral y los dos receptores, se produce la fusión entre las membranas de la célula y del virus. El principal responsable es la gp41, la cual se inserta en la membrana celular permitiendo la internalización de la nucleocápside del virus y la desencapsidación de su genoma. Esta tesis se ha centrado en esta etapa del ciclo vital del VIH, por lo que se detalla con más profundidad más adelante.

2) Transcripción inversa

Tras la entrada se inicia la reproducción del virus (replicación) por transcripción inversa o retrotranscripción mediada por la transcriptasa inversa viral, lo cual conduce a la formación de la primera cadena de ADN a partir del ARN viral. La segunda cadena de ADN requiere la acción de la ribonucleasa H.

3) Integración

La doble cadena así generada es integrada por medio de la integrasa viral en el ADN de la célula, aunque parte del ADN formado puede persistir en el citoplasma de la célula sin integrarse dentro del genoma celular. La copia del material genético del VIH como ADN se almacena en el citoplasma de la célula (latencia preintegración) y se va integrando en los cromosomas de la célula a medida que pasa el tiempo, como consecuencia de estímulos sobre la célula. Una vez integrado en el material genético de la célula, el provirus puede permanecer latente o empezar a multiplicarse de una forma controlada o de una forma masiva, en cuyo caso ocasionará efectos citopáticos sobre la célula, mientras que en la fase de latencia no se producen alteraciones patológicas.

4) Transcripción

La activación celular por diferentes factores como antígenos, mitógenos, citoquinas o virus heterólogos puede producir una cascada de acontecimientos que lleven a la expresión del genoma

viral. Estos factores, entre los que destaca el NF-bK, conducen a una nueva transcripción que supone la síntesis de ARN del virus a partir del ADN proviral integrado en la célula. Este ARN se sintetiza como un único transcrito que debe volver al citoplasma de la célula para procesarse en transcritos de diferente tamaño, por lo que son fundamentales las proteínas Tat y Rev.

5) Traducción

A continuación se traduce el ARN viral a las proteínas precursoras del virus. Las proteínas codificadas por el gen *gag* se traducen como una única poliproteína (gag). Lo mismo ocurre con las del gen pol, que forman la poliproteína gag-pol. Finalmente se glicosidan las proteínas sintetizadas.

6) Ensamblaje

El ensamblaje del núcleo ocurre en la membrana celular y parece comenzar con la asociación de la proteína p17 de la matriz con el dominio citoplasmático de la proteína gp41. La síntesis de las proteínas de la envoltura viral se producen en el retículo endoplasmático de la célula huésped a partir de la gp160; dicha glicoproteína es escindida en el aparato de Golgi por una proteasa para producir gp120 y gp41. Estas proteínas, junto con las poliproteínas gag y gag-pol sintetizadas y dos cadenas de ARN se agregan formándose la nucleocápside. Otras proteínas no son empaquetadas en los viriones y sólo actúan en los pasos que preceden a la liberación de los virus.

7) Salida de viriones

El virión inmaduro se desprende de la célula. Tras madurar el virión, el virus resulta infeccioso. Se piensa que la vida libre de los viriones es muy corta, aproximadamente de 0,3-0,5 días (lo que corresponde a 8-12 horas), y que en 2,6 días se realiza un ciclo viral completo con infección productiva, salida desde la célula infectada, vida libre, infección de otro linfocito, replicación intracelular y salida de nuevos viriones. De este modo se producirían unos 140 ciclos de replicación al año.

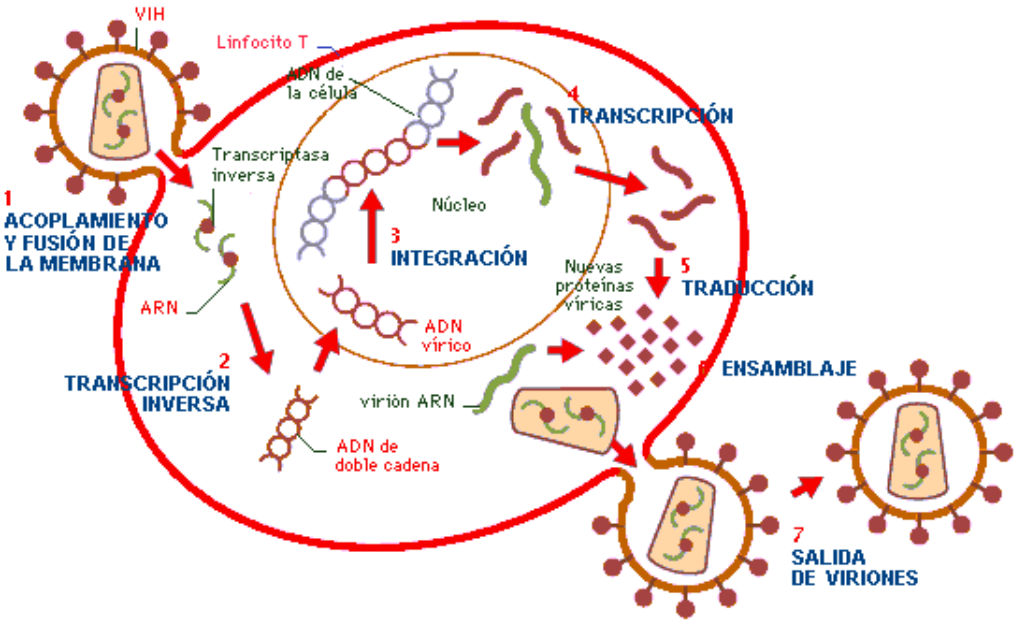


Figura I.8 Ciclo vital del VIH-1 [44].

En la Figura I.9 se esquematiza el proceso concreto de entrada/fusión del virus a las células diana, punto de ataque en el que se centra el presente trabajo:

A) La infección del virus se inicia debido a la interacción entre la envoltura glicoproteica del virus (Env), compuesta por la subunidad gp120 y la subunidad transmembrana gp41 unidas no covalentemente, y el complejo comprendido por la CD4 y un receptor de quimiocinas (CXCR4 o CCR5).

B) La unión de la subunidad gp120, concretamente el *loop* V3, con la CD4 provoca un cambio conformacional que expone o crea el sitio de unión de la gp120 a uno de los dos coreceptores citados.

C) La gp120 se une al receptor de quimiocinas, lo cual induce cambios estructurales en la gp41.

D) Cada gp41 consta de dos hélices α que se reorganizan para formar un 'intermedio pre-*hairpin*', insertando un péptido de fusión hidrofóbico en la célula a infectar y dejan al virus y a la célula en una posición óptima para iniciar la fusión.

E) Las hélices de la gp41 se pliegan entonces en un manojo de seis hélices, uniendo los extremos *N*-terminal y *C*-terminal, y con ello también las membranas vírica y celular.

F, G) El contacto entre las membranas crea un poro de fusión con lo que la información del virus puede acceder al citoplasma.

Mientras que todas las cepas preferentemente de VIH-1 requieren la inmunoglobulina CD4 para entrar e infectar células, algunas utilizan el receptor de quimiocinas CCR5 (cepas R5), otras usan CXCR4 (cepas X4), y algunas pueden usar ambos coreceptores (cepas R5X4). Las cepas de R5 son las necesarias para la transmisión del virus de persona a persona, así como para el establecimiento y mantenimiento de la infección. Las cepas X4 y R5X4 aparecen más tarde en el 30-50% de los individuos infectados. Aunque los individuos pueden progresar en la enfermedad del SIDA en ausencia de estas dos últimas variantes, su aparición está fuertemente asociada a un proceso de aceleración de la enfermedad. No se sabe exactamente por qué la aparición de variantes que usan el coreceptor CXCR4 es una causa o una consecuencia de la destrucción inmunológica. No obstante, esta asociación indica un papel importante del coreceptor CXCR4 en el progreso de la enfermedad. Incluso cuando las variantes X4 y R5X4 aparecen, como síntoma del progreso de la enfermedad, las variantes R5 siguen persistiendo, con lo que se demuestra que el coreceptor CCR5 es crítico a lo largo de la enfermedad. Otros GPCRs también son utilizados por el VIH-1 para entrar en las células diana en modelos *in vitro*. Sin embargo, no hay evidencia de que ningún coreceptor diferente a CCR5 y CXCR4 sea utilizado *in vivo* o juegue un papel en la patogénesis [45]. La selectividad del coreceptor es determinada por secuencias genéticas que contiene la gp120, particularmente una región muy variable y estructuralmente flexible (V3).

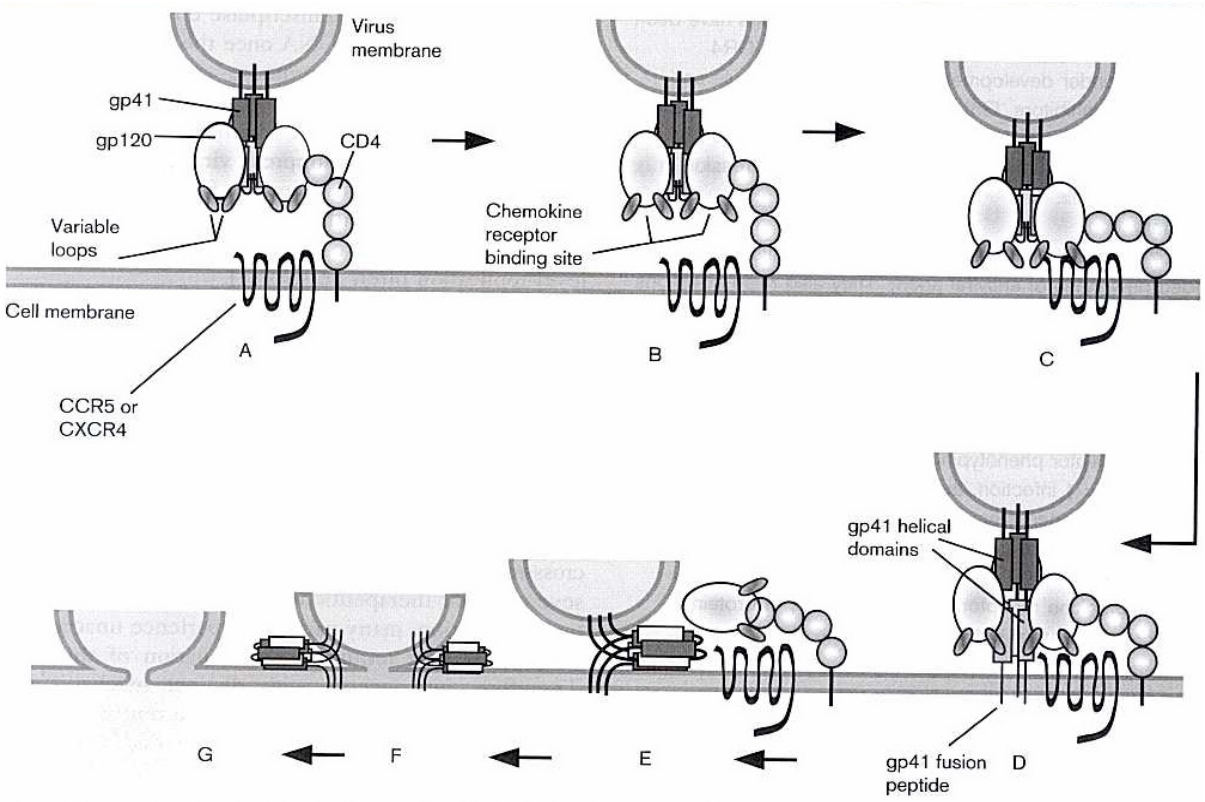


Figura I.9 Entrada del VIH-1 a la célula huésped de acuerdo con los modelos actuales. La gp120 se omite en los dibujos F y G con el fin de ofrecer mayor claridad [45].

Cada paso en la entrada del virus puede considerarse una diana farmacológica:

A) *Unión gp120 - CD4*. El mecanismo consiste en intentar bloquear la inmunoglobulina CD4 de la célula diana para evitar la unión gp120-CD4. Hace una década se apuntó hacia la unión gp120-CD4 como diana farmacológica, lo cual condujo a no muy buenos resultados [46]. La aproximación falló probablemente porque las interacciones gp120-CD4 de VIH-1 en su estado natural tienen relativamente poca afinidad (a diferencia de las cepas con las que se trabajó en los estudios preclínicos en el laboratorio), ya que son estabilizadas por interacciones multivalentes entre las diferentes glicoproteínas del Env del virus y las múltiples moléculas CD4 de las células. Sin embargo, la interacción gp120-CD4 ha vuelto a ser estudiada utilizando nuevas aproximaciones [47].

B) *Cambio conformacional que expone el sitio de unión de la gp120*. El mecanismo consiste en intentar bloquear la gp120 del virus con el fin de evitar los cambios conformacionales necesarios para poder interactuar con el receptor de quimiocinas.

C) *Unión gp120-receptor de quimiocinas*. El mecanismo consiste en intentar bloquear el sitio de unión del receptor de quimiocinas (tanto para el CXCR4 como CCR5) y así evitar la unión del receptor de la célula diana a la gp120 del virus. Esta es la diana farmacológica abordada en esta tesis.

D), E) *Cambios estructurales en la gp41*. El mecanismo consiste en intentar bloquear la gp41 del virus con el fin de evitar los cambios estructurales necesarios para la fusión.

F),G) *Contacto entre las membranas y fusión del virus*: El mecanismo consiste en intentar bloquear la desestabilización de la bicapa lipídica y fusión de la membrana de virus con la membrana celular.

I.2.3 Receptores acoplados a proteínas G (GPCRs). Coreceptores CXCR4 y CCR5

Los receptores acoplados a proteínas G (GPCRs, *G Protein-Coupled Receptors*), también denominados 7TM, son proteínas que comparten una topología común compuesta por siete hélices α transmembrana hidrofóbicas. El extremo *N*-terminal se localiza en la cara extracelular de la membrana y está a menudo glicosidado, mientras que el extremo *C*-terminal se encuentra en la zona citoplasmática y generalmente está fosforilado. Tres *loops* extracelulares se alternan con tres *loops* intracelulares para unir las siete regiones transmembrana. Las partes más conservadas de estas proteínas son las regiones transmembrana y los dos primeros *loops* citoplasmáticos. Suele conservarse el triplete ácido-Arg-aromático en el extremo *N*-terminal del segundo *loop* citoplasmático [48], el cual puede estar implicado en la interacción con las proteínas G.

La superficie extracelular de los receptores GPCR (la cual incluye el extremo *N*-terminal, los *loops* extracelulares y las zonas exofaciales de diversos dominios transmembrana) está involucrada en la unión de ligandos, mientras que la superficie intracelular de dichos receptores (la cual comprende el extremo *C*-terminal, los *loops* intracelulares y las regiones citoplasmáticas de diversos dominios transmembrana) es importante para el acoplamiento de las proteínas G.

La Figura I.10 se muestra una representación esquemática del esqueleto común de todos los GPCRs.

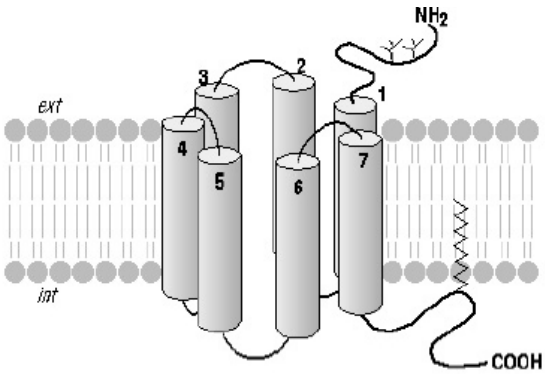


Figura I.10 Esqueleto común de todos los GPCRs consistente en siete hélices transmembrana unidas por *loops* intracelulares y extracelulares alternados, con un extremo *N*-terminal extracelular y un extremo *C*-terminal intracelular [49].

Los GPCRs son los responsables de la traducción de señales endógenas a una respuesta celular. La unión de sus correspondientes ligandos (como aminas, lípidos, proteínas, nucleósidos, nucleótidos, y aminoácidos) a las regiones extracelulares o transmembrana provoca la activación del receptor. Dicho receptor se encuentra inicialmente en un estado latente, unido por su cara intracelular a la proteína G (proteína reguladora de la unión a un nucleótido de guanina), la cual es un trímero compuesto por las subunidades α, β, y γ. En este estado la proteína G une al nucleótido GDP. Al activarse el receptor, se produce un cambio conformacional en la subunidad α de la proteína G que favorece el intercambio de GDP (forma inactiva) por GTP (forma activa), lo cual lleva a la disociación de sus subunidades α y βγ. Esta disociación desencadena a su vez diversos mecanismos

de señalización secundaria inhibiendo o estimulando así la producción de mensajeros intracelulares secundarios. Existen distintas proteínas G dependiendo de la naturaleza de la subunidad α, y cada una de ellas promueve una acción intracelular distinta, como por ejemplo, α_i, la cual inhibe la producción de adenilato ciclasa, cAMP, y estimula la producción de iones K^+ o enzima fosfolipasa; α_q, la cual estimula la producción de PLC_β, diacilglicerol (DAG), PKC, e iones Ca^{2+}; α_s, la cual estimula la producción de adenilato ciclasa e iones Ca^{2+}. También la subunidad βγ puede promover la activación de alguna ruta bioquímica, como la regulación de los canales de K^+, o de las encimas adenil ciclasa y fosfolipasa C. De esta manera, se desencadena una cascada de señales internas que culminan en un cambio en el comportamiento de la célula [48].

La Figura I.11 muestra de forma más detallada el proceso de activación de un GPCR y el camino que toma la señal que llega a éste:

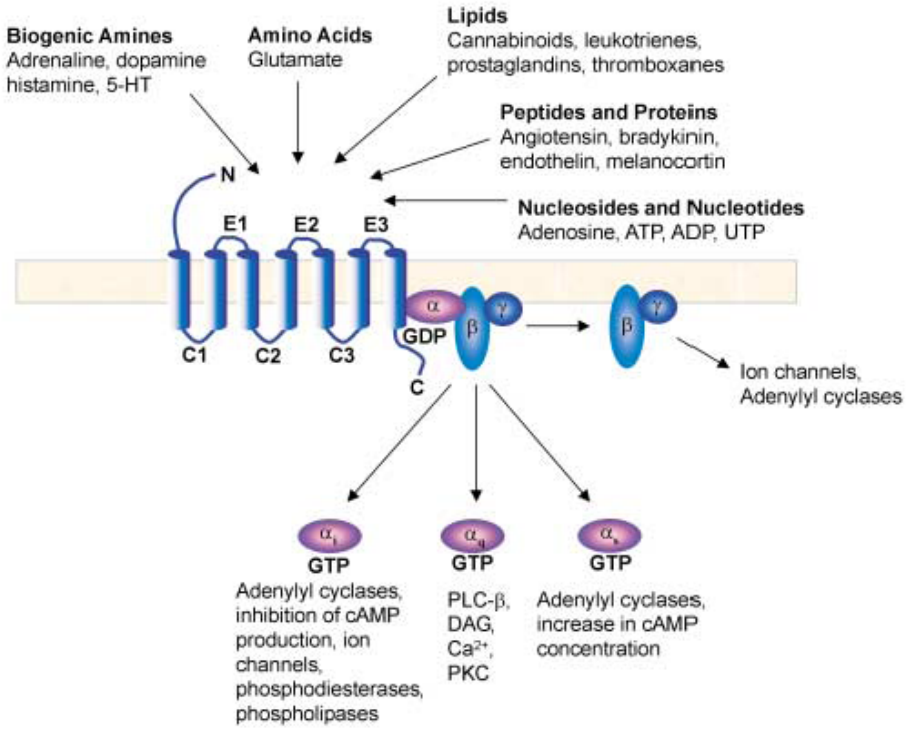


Figura I.11 Diagrama "*snake*" de un GPCR y camino que toma la señal que llega a éste [50].

Cerca de la mitad de los fármacos del mercado actúan sobre los GPCRs [51], [52] con unas ventas anuales superiores a los 30 billones de euros [53]. De entre los 100 fármacos más vendidos actualmente, el 25% van dirigidos a miembros de esta familia de proteínas. Muchos fármacos del mercado se dirigen hacia el centro activo de un GPCR específico. En el curso de los últimos años, se ha descubierto que la actividad de un GPCR se puede modificar también mediante la unión a otras regiones del mismo. De esta forma, moléculas de tamaño reducido dirigidas hacia esos puntos de ataque pueden activar o inhibir los GPCRs implicados en varias enfermedades. Ello abre nuevas vías terapéuticas.

Desafortunadamente, a pesar de la importancia farmacológica de los GPCRs, no hay suficiente información estructural. La falta de estructuras tridimensionales para los GPCRs es debida a la dificultad en la obtención de cristales de proteínas transmembrana con una calidad suficiente como para obtener buena resolución en la técnica de difracción de rayos-X y la dificultad de uso de RMN para determinar la estructura en estos sistemas transmembrana [54]. A nivel atómico solamente se han resuelto las estructuras de los GPCRs rodopsina bovina [55, 56] y β2-adrenérgico [57, 58],, las cuales presentan una baja homología (inferior al 35%) respecto a la mayoría de GPCRs de interés farmacológico. Por ello, es importante desarrollar métodos teóricos para poder predecir la estructura y función de los GPCRs [59, 60].

Los GPCR más interesantes se pueden clasificar en tres familias o clases, dependiendo de su secuencia de aminoácidos [61]:

- La clase I de GPCRs pertenece a la familia de receptores semejantes a la rodopsina o receptores adrenérgicos cuyos ligandos son aminas (dopamina, serotonina, histamina, etc.) y pequeños péptidos (quimiocinas y neuropéptidos). Se caracterizan por un pequeño dominio *N*-terminal extracelular, un dominio siete-transmembrana (7TM), un largo dominio *C*-terminal intracelular, y una elevada conservación de residuos aminoácidos en cada hélice transmembrana. En muchos casos, la cavidad de unión del ligando está delimitada por el dominio 7TM, aunque receptores peptídicos específicos pueden usar dos de los tres *loops* extracelulares para rodear el lugar de unión del péptido. Esta clase es la que contiene un mayor número de GPCRs (más de 240) [62].

- La clase II de GPCRs pertenece a la familia de los receptores semejantes a la secretina o semejantes al glucagón cuyos ligandos son hormonas proteicas (péptido vasointestinal, glucagón, etc.). Difieren de la clase anterior por tener un dominio *N*-terminal más largo (con un grupo de seis residuos cisteína conservado en el grupo) que delimita el sitio de unión de la hormona y un dominio *C*-terminal también más largo. Pertenecen a esta clase alrededor de 60-65 GPCRs [63].

- La clase III pertenece a la familia de los receptores semejantes al receptor metabotrópico de glutamato, los cuales reconocen ligandos con carga de bajo peso molecular (glutamato, calcio, ácido γ-aminobutírico, etc.) a través de un dominio *N*-terminal muy largo (entre 500 y 600 residuos) compuesto por dos lóbulos simétricos. Se caracterizan por tener los *loops* intracelulares más cortos y un dominio *C*-terminal largo. Solo pertenecen a esta clase unos 15 GPCRs [62].

Esta tesis trata con la clase I de GPCRs, en concreto los coreceptores CXCR4 y CCR5 pertenecientes a dicha clase.

En 1996 se identificó el CXCR4 [64] (número de acceso en la base de datos SWISS-PROT P61073 y nombre de entrada CXCR4_HUMAN) [65]como coreceptor del VIH-1 junto con la CD4. También se denomina LESTR o fusina, ya que interviene en el proceso de fusión del VIH-1, aunque también actúa como receptor primario en algunas cepas de VIH-2 [66]. A diferencia de muchos otros receptores de quimiocinas, los cuales tienden a tener diferentes ligandos, el coreceptor CXCR4 solo tiene un ligando natural, una quimiocina del tipo CXC denominado *stromal cell-derived factor* (SDF-1) [64], la cual presenta dos isoformas, α y β. El SDF-1 es un potente inhibidor de las cepas T-trópicas porque se une a su receptor natural CXCR4 y bloquea la entrada del VIH en los linfocitos.

En 1996 se identificó también el CCR5 (número de acceso en la base de datos SWISS-PROT P51681 y nombre de entrada CCR5_HUMAN) [67] como coreceptor del VIH-1 junto con la CD4.

También interviene en el proceso de fusión del VIH [68]. Los ligandos naturales de la proteína CCR5 son las quimiocinas RANTES (*Regulated Upon Activation, normal T-cell expressed and secreted*), MIP-1α y MIP-1β (*Macrophage Inflamatory Protein*) secretados por las células T CD8+ (T citotóxicas) [69, 70]. El receptor CCR5, en respuesta a dichas quimiocinas CC, traduce señales que hacen que el monocito muestre una quimiotaxis hacia las áreas de inflamación. Estas quimiocinas pues, inhiben la infección de las cepas M-trópicas primarias del VIH-1 y células facilitadoras T CD4+. Sin embargo las cepas T-trópicas no son inhibidas por las quimiocinas CC.

Actualmente no existe ninguna estructura resuelta de las proteínas CXCR4 y CCR5 depositadas en el *Protein Data Bank* [203], pero se dispone de bastante información acerca de los sitios de unión de los diferentes ligandos de ambas proteínas (sus ligandos naturales, agentes derivados de quimiocinas, su unión con la proteína gp120 del VIH, anticuerpos monoclonales y unión con diferentes moléculas que actúan como inhibidores de entrada del VIH), obtenida básicamente a partir de estudios de mutagénesis dirigida y de experimentos con quimeras [66, 70, 71].

La Figura I.12 presenta la estructura de los coreceptores CXCR4 y CCR5. Se observan las siete hélices α transmembrana características de los GPCRs conectadas por tres *loops* extracelulares (E1, E2, E3) y tres *loops* intracelulares (I1, I2, I3). Se observa que ambas proteínas contienen cuatro residuos de cisteína (dos puentes disulfuro, marcados con * en la estructura de CXCR4 y con − en la estructura de CCR5), dos cisteínas de las cuales se cree que contribuyen a la estabilización de la conformación del receptor formando un puente disulfuro entre los *loops* extracelulares E1 y E2.

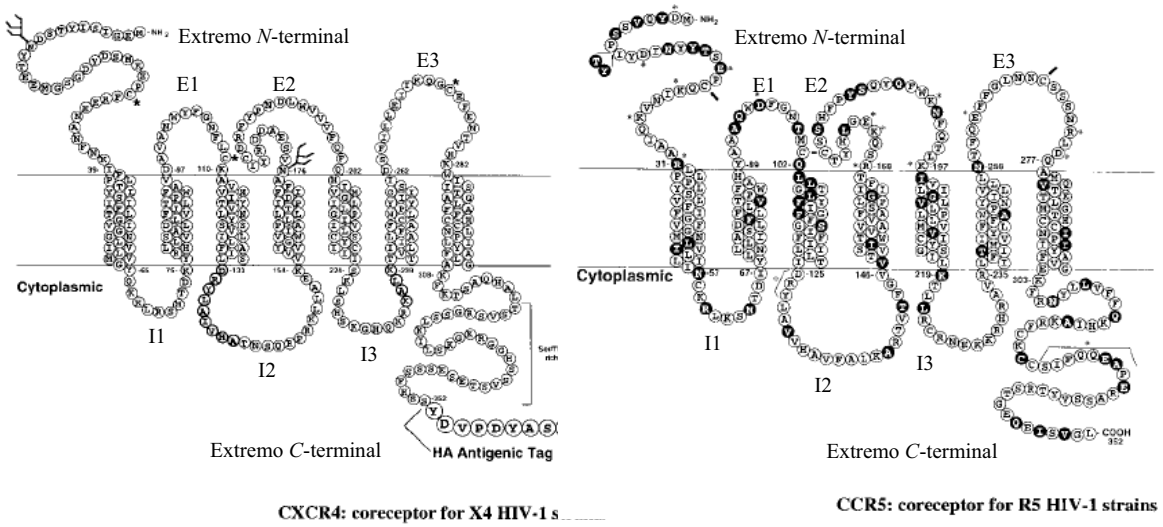


Figura I.12 CXCR4 [64]: receptor de la quimiocina CXC SDF-1 y coreceptor de las cepas T-trópicas (X4) de VIH-1. CCR5 [72]: receptor de las quimiocinas CC MIP-1a, MIP-1b, RANTES y coreceptor de las cepas M-trópicas (R5) de VIH-1 [73].

I.2.4 Fármacos anti-VIH

Tal y como se muestra en la Figura I.13, el ciclo de vida del VIH ofrece abundantes dianas terapéuticas. A continuación se muestran tales dianas junto con algunos compuestos inhibidores representantes de cada una de ellas [74, 75, 76]:

- Unión del virus a los linfocitos CD4, mediada por la interacción entre la gp120 del virus y el receptor celular CD4.

Compuestos inhibidores de la unión virus-CD4: Ciclotriazadisulfonamida (CADA), como modulador del receptor CD4. Lectinas de plantas derivadas de GNA (*Snowdrop*) y HHA (*Amaryllis*), como inhibidoras de la unión gp120-CD4.

- Interacción de la gp120 con los coreceptores CXCR4 y/ó CCR5, inserción de la gp41 en la membrana celular y fusión de la envoltura viral con la membrana celular.

Compuestos inhibidores de los coreceptores: AMD3100, AMD070, CS-3955 para CXCR4, TAK779, TAK220, SCH-D (Vicriviroc), UK-427857 (Maraviroc), GW873140 (Aplaviroc) para CCR5.

Compuestos inhibidores de la fusión (FI): Enfuvirtide (T20), T-1249, RPR 103611.

- Transcripción inversa (mediante la transcriptasa reversa) del ARN viral de una sola cadena en ADN proviral de dos cadenas, el cual es transportado al núcleo de la célula huésped.

Compuestos inhibidores de la transcriptasa reversa (RT) cuya diana es el sitio de unión del substrato. Son los llamados inhibidores nucleósidos o Nucleoside Reverse Transcriptase Inhibitors (NRTIs, ddNs o Nukes) y inhibidores nucleotídicos o Nucleotide Reverse Transcriptase Inhibitors (NtRTI): adefovir dipivoxil, tenofovir disoproxil, emtricitabine [(-)FTC], amdoxovir, dOTC, FdOTC, reverset, elvucitabine, 4'-E-dC, 4'-E-dCA, 4'-E-dDAP, alovudine, ariloxifosfodamidato de d4T, ciclosaligenil d4TMP.

Compuestos inhibidores de la transcriptasa reversa (RT) cuya diana no es el sitio de unión del substrato, son moduladores alostéricos. Son los llamados inhibidores no nucleósidos o Non-Nucleoside Reverse Transcriptase Inhibitors (NNRTIs): nevirapina, delavirdina, efavirenz, UC-781, DPC 083, TMC125 (R165335), SJ-3366, Capravina (S-1153, AG1549), PNU-142721, (+)-Calanolide A.

- Integración del ADN proviral en ADN cromosómico del núcleo de la célula hospedadora mediante la integrasa del virus.

Compuestos inhibidores de la integrasa: S-1360, L-870,812, PDPV-165, L-870,810.

- Transcripción de ADN proviral a ARN viral.

Compuestos inhibidores de la transcripción: fluoroquinolinas (K12, K37), temacrazina, CGP64222, flavopiridol, EM2487, PKF 050-638.

- Síntesis de precursores de proteínas a partir del ARN viral. Proceso proteolítico (mediante la proteasa viral) de los precursores proteicos para dar lugar a la estructura proteica madura (cápside) y a proteínas funcionales (proteasa, transcriptasa reversa, integrasa) del virus.

Compuestos inhibidores de la proteasa (PI): tipranavir, TMC-114, saquinavir, indinavir, ritonavir, nelfinavir, lopinavir, atazanavir.

- Ensamblaje del ARN vírico con las proteínas víricas en la membrana celular y posterior salida de los viriones de la membrana celular.

Compuestos inhibidores del ensamblaje o maduración: PA-457, α-HGA.

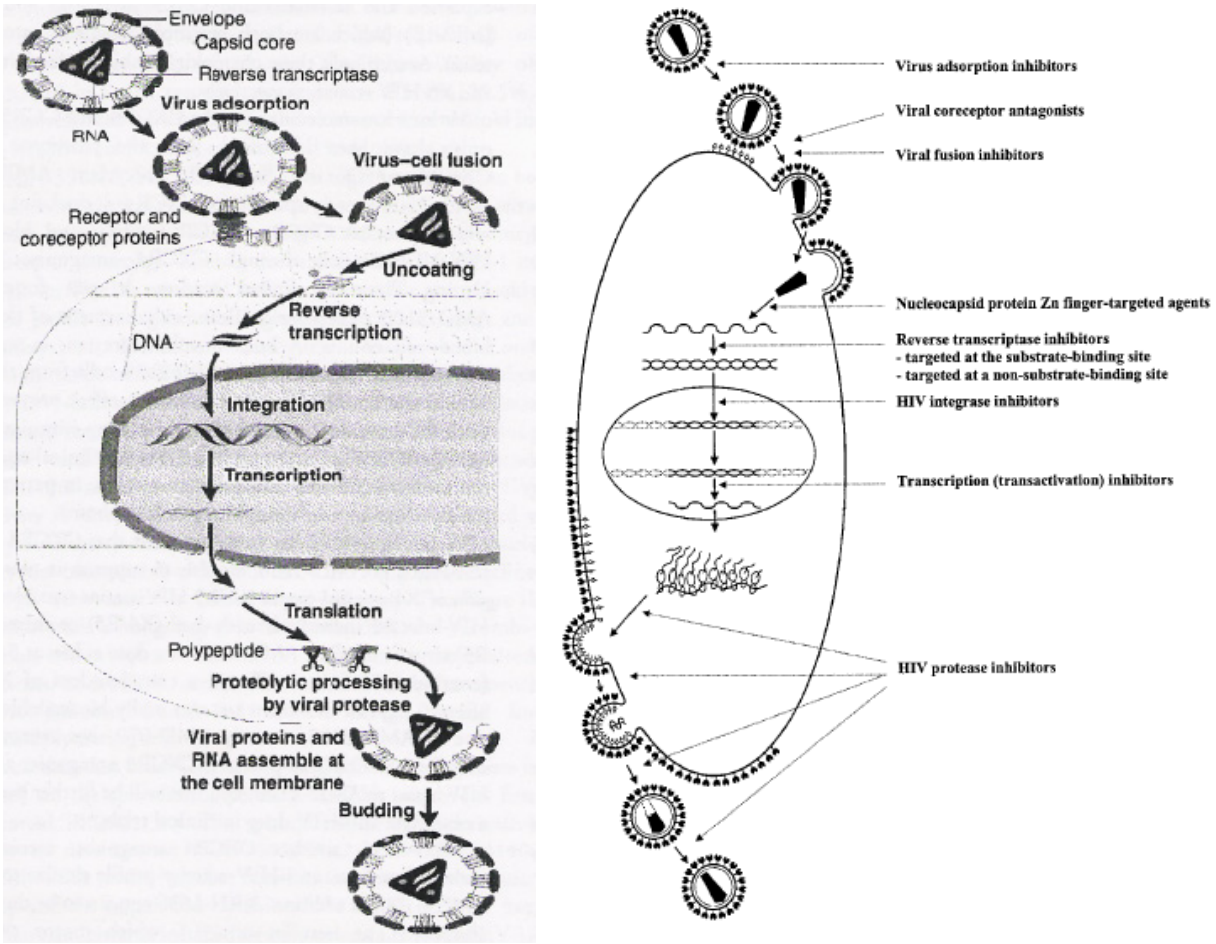


Figura I.13 Ciclo de vida del VIH (izquierda) junto a las dianas terapéuticas con las cuales interaccionan los agentes antivirales actuales (derecha) [74, 75].

Actualmente existen más de 40 compuestos que se encuentran en desarrollo (pre)clínico y 25 fármacos antivirales formalmente aprobados por la FDA (*Food and Drug Administration*) para el tratamiento del SIDA [77, 78]: 7 NRTIs (zidovudina, didanosina, zalcitabina, estavudina, lamivudina, abacavir, emtricitabina), 1 NtRTI (tenofovir disoproxil fumarato), 4 NNRTIs (nevirapina, delavirdina, efavirenz, y etravirina), 10 PIs (saquinavir mesilato, ritonavir, indinavir, nelfinavir mesilato, amprenavir, lopinavir, atazanavir sulfato, fosamprenavir, tipranavir, y darunavir), 1 FI (enfuvirtide), 1 inhibidor de la integrasa (raltegravir), 1 inhibidor de CCR5 (maraviroc). Ninguna

de estas clases es tan eficiente como para acabar con el VIH, pero cada una de ellas ralentiza el proceso de replicación del VIH de una manera particular. Dichos fármacos antivirales aprobados por la FDA se muestran en la Tabla 2.1.

Clase, nombre del fármaco	Abreviatura	Nombre comercial	Compañía
Inhibidores de la transcriptasa reversa nucleosídicos y nucleotídicos			
Zidovudina	AZT	Retrovir®	GlaxoSmithKline
Didanosina	ddI	Videx®,Videx®EC	Bristol-Myers Squibb
Zalcitabina	ddC	Hivid®	Hoffmann - La Roche
Lamivudina	3TC	Epivir®	GlaxoSmithKline
Estavudina	d4T	Zerit®	Bristol-Myers Squibb
Abacavir	ABC	Ziagen®	GlaxoSmithKline
3TC más AZT		Conbivir®	GlaxoSmithKline
ABC más 3TC más AZT		Trizivir®	GlaxoSmithKline
ABC más 3TC		Epzicom®	GlaxoSmithKline
Emtricitabina	FTC	Emtriva®	Gilead Sciences
Tenofovir	TDF	Viread®	Gilead Sciences
TDF más FTC		Truvada®	Gilead Sciences
Inhibidores de la transcriptasa reversa no nucleosídicos			
Nevirapina	NVP	Viramune®	Boehringer Ingelheim
Delavirdina	DLV	Rescriptor®	Pfizer
Efavirenz	EFV	Sustiva®	Bristol-Myers Squibb
Etravirina		Intelence®	Tibotec Therapeutics
Inhibidores de proteasa			
Saquinavir	SQV	Fortovase®,Invirase®	Hoffmann - La Roche
Indinavir	IDV	Crixivan®	Merck
Ritonavir	RTV	Norvir®	Abbot Laboratoires
Nelfinavir	NFV	Viracept®	Agouron Pharmaceuticals
Amprenavir	AMP	Agenerase®	GlaxoSmithKline
Fosamprenavir	FOS-APV	Lexiva®	GlaxoSmithKline
Lopinavir	LPV	Aluviran®	Abbot Laboratoires
LPV más RTV (4:1)		Kaletra®	Abbot Laboratoires
Atazanavir	ATV	Reyataz®	Bristol-Myers Squibb
Tipranavir	TPV	Aptivus®	Boehringer Ingelheim
Darunavir		Prezista®	Tibotec, Inc.
Inhibidores de fusión			
Enfuvirtide	T-20	Fuzeon®	Hoffmann - La Roche / Trimeris
Inhibidores de entrada			
Maraviroc		Selzentry®	Pfizer
Inhibidores de integrasa			
Raltegravir		Isentress®	Merck & Co., Inc.

Tabla 2. 1 Clases de fármacos aprobados en el 2008 por la US Food and Drug Administration (FDA) para el tratamiento de VIH/SIDA [77].

Las terapias antiretrovirales actuales contra el VIH están basadas en la combinación de diferentes fármacos (*Highly Active Antiretroviral Therapy, HAART*). Generalmente se componen de uno o más inhibidores de la transcriptasa reversa nucleosídicos (NRTIs) o uno o más inhibidores de la transcriptasa reversa nucleotídicos (NtRTIs), junto con un inhibidor de la transcriptasa reversa no nucleosídico (NNRTI) o bien un inhibidor de proteasa (PI), a veces complementados o no por un inhibidor de fusión (FI) [76, 77, 78].

A pesar del gran número de fármacos disponible, existen diversas preocupaciones acerca de los regímenes antiretrovirales. El problema de estas terapias es la complicidad de algunos regímenes obligando a los pacientes a tomar diversas píldoras en diferentes momentos del día, la resistencia del virus que emerge debido a mutaciones de éste (la cual provoca que los pacientes no alcancen o

mantengan significativamente la supresión viral), y los efectos secundarios que producen [74, 79]. Por ello se buscan nuevos fármacos alternativos, con otros modos de acción, que consigan:

- Una mayor potencia.
- Menores efectos secundarios.
- Menor riesgo en el desarrollo de la resistencia del virus.

Además, es importante trabajar en la búsqueda de nuevos fármacos contra el VIH accesibles a todo el mundo, para poder reducir los costes sustancialmente y así poder asegurar que llegaran a las poblaciones que no pueden abordar precios altos en medicamentos, lugares donde precisamente el SIDA prevalece mayoritariamente.

Así pues, en los años recientes, se ha progresado significantemente en lo que respecta a la quimioterapia del VIH. El progreso se sitúa a tres niveles diferentes [80]:

- Nuevos fármacos anti-VIH han sido aprobados para el uso clínico y han entrado al mercado: el FI enfuvirtide (Fuzeon®, 2003), el NRTI emtricitabina (Emtriva®, 2003), el NtRTI tenofovir disoproxil fumarato (Viread®, 2001), el NNRTI etravirina (Intelence, 2008), los PIs atazanavir (Reyataz®, 2003), fosamprenavir (Lexiva®, 2003), tripanavir (Aptivus®, 2005), y darunavir (Prezista®, 2006), el inhibor de entrada maraviroc (Selzentry®, 2007) y el inhibidor de integrasa raltegravir (Isentress®, 2007), véase Tabla 2.1.

- Otros compuestos han pasado a desarrollo clínico o preclínico: antagonistas del CXCR4 como AMD070 o KRH-1636, antagonistas del CCR5 como SCH-C o TAK220, NRTIs como amdoxovir o Reverset™, NNRTIs como capravirine o dapivirine (Tabla 2.2).

- Además se han identificado nuevos compuestos, que actúan por diferentes mecanismos, como agentes antivirales que están en desarrollo (pre)clínico un poco más lejano: moduladores del receptor celular CD4 (ciclotriazadisulfonamidas), agentes bloqueantes de la unión gp120-CD4 como las lectinas de las plantas y antibióticos glicopeptídicos, inhibidores de la integrasa del virus como la piranodipirimidina V-165, y dos nuevas clases de compuestos (*N*-aminoimidazoles y derivados de óxido de piridina), los cuales parecen interferir en los procesos de post-integración, transcripción, y síntesis de precursores de proteínas a partir del ARN viral (véase Tabla 2.2).

Las siete clases de fármacos formalmente aprobados por la FDA mencionados anteriormente (NRTI, NtRTI, NNRTI, PI, inhibidores de integrasa, FI, e inhibidores de entrada), a excepción de los dos últimos, operan dentro de la célula infectada. El propósito principal de los nuevos fármacos antiretrovirales es que operen fuera de la célula a infectar. Estos nuevos fármacos actúan inhibiendo la entrada del virus en la célula diana, deteniendo así el primer paso de la replicación. Además de su nuevo mecanismo de acción, estos fármacos tienen también una acción potencial contra la resistencia de las cadenas de VIH, causan menos efectos secundarios, y se pueden administrar en un régimen de dosis más simple [80, 82, 83].

Asimismo, un número creciente de pacientes infectados por VIH no puede usar los fármacos anti-VIH actualmente aprobados debido a sus efectos secundarios y resistencia que ejerce el virus. Los productos naturales (sean de origen natural o sintético) son fuentes importantes para nuevos fármacos. Los siguientes productos naturales pueden citarse como promesas de agentes anti-VIH provenientes de las plantas: baicalina (un flavonoide), calanolidas (cumarinas), ácido botulínico (un

triterpeno), policetona A (un alcaloide), ácido litospérmico (un polifenol), polisacáridos sulfatados, cianovirina-N y alfa-tricobitacina (proteínas) [84].

Clase, subclase, nombre del fármaco	Estado clínico	Compañía
Inhibidores de entrada		
Inhibidores de CD4 y gp120		
Ciclotriazadisulfonamida	Preclínica	Rega Institute K.U.Leuven/Reno University
BMS-378806	Fase II	Bristol-Myers Squibb
BMS-378806	Fase II	Bristol-Myers Squibb
Cyanovirin-N	Preclínica	Biosyn
Inhibidores de CXCR4		
AMD3100	Fase II	AnorMED
AMD070	Fase I/II	AnorMED
CS-3955, KRH-2731	Preclínica	Sankyo/Kureha Chemical Industry
Inhibidores duales CXCR4/ CCR5		
AMD3451	Preclínica	AnorMED
Inhibidores de CCR5		
SCH-C (SCH-351125)	Fase III	Schering-Plough
SCH-D (SCH-417690)	Fase III	Schering-Plough
GW-873140,ONO-4128,AK-602	Fase III	Ono Pharmaceutical GlaxoSmithKline
TAK 220, TAK 652	Fase II	Takeda Pharmaceutical
Inhibidores de de la transcriptasa reversa nucleosídicos		
Racivir®	Fase II	Emory University
SPD-745	Fase II	Shire Pharmaceuticals (BioChem Pharma)
Reverset^{TM} (DPC-817)	Fase II	Emory University, Pharmasset, Incyte
Elvucitabine (ACH-126443)	Fase II	Yale University, Achillion
Alovudine (MIV-310)	Fase II	Medivir, Boehringer Ingelheim
MIV-210	Fase I	Medivir
Amdoxovir	Fase II	Emory University
1-(*β-D* -Dioxolano)timina (DOT)	Fase II	University of Georgia
Inhibidores de la transcriptasa reversa no nucleosídicos		
Tiocarboxanilida (UC-781)	Fase I/II	Uniroyal Chemical, Biosyn
Capravirine (S-1153, AG1549)	Fase III	Shionogi/Pfizer
Dapivirine (TMC-120,R-147681)	Fase I/II	Tibotec (Johnson & Johnson)
Rilpivirine (R-278474)	Fase II	Janssen Pharmaceutica, Tibotec (J & J)
Inhibidores de integrasa		
S-1360, GW-810781	Fase II	GlaxoSmithKline and Shionogi
L-870810	Fase I/II	Merck
L-870812	Fase I	Merck
Piranodipirimidina (V165)	Fase I/II	K.U. Leuven Research & Development, Ampharm
GS-9160	Fase I	Gilead Sciences
Inhibidores de la maduración del VIH		
PA-457	Fase II	University of North Carolina, Panacos
α-Hidroxi-glicinamida (α-HCG)	Fase II	Tripep

Tabla 2. 2 Fármacos bajo desarrollo o en proceso de testado para el tratamiento de VIH/SIDA en 2008. Algunos de ellos han parado su desarrollo clínico como es el caso del AMD3100 [74, 81, 85, 86].

I.2.5 Inhibidores de entrada del VIH

Tal y como se ha comentado anteriormente, los inhibidores de los receptores de quimiocinas son el punto de ataque de muchos de los nuevos fármacos anti-VIH, [87, 88, 89, 90], ya que intentan detener al virus antes de que éste entre a la célula. Estos agentes tienen además aplicaciones importantes hacia otras situaciones, como cáncer [91] o enfermedades inflamatorias [92, 93, 94, 95].

A continuación se presentan las clases de inhibidores del CXCR4 más importantes divididos en pequeñas moléculas, antagonistas peptídicos y agentes basados en quimiocinas [45, 82].

Pequeñas moléculas

1.- Derivados **biciclamos** desarrollados por AnorMED, los cuales se encuentran en ensayos clínicos de fase II / III [82,96, 33, 97]. Los biciclamos contienen dos anillos macrocíclicos (1,4,8,11-tetraazaciclotetradecano) conectados por un fragmento alifático (como es el caso del AMD2763) o aromático (como ocurre en el AMD3100). Entre ellos destaca el AMD3100, el cual muestra una potente actividad frente a diversas cepas de VIH, pero tiene la desventaja de requerir administración intravenosa o subcutánea.

Los inhibidores de estructura biciclamo se consideran de los antagonistas más potentes del CXCR4 hasta hoy día conocidos. Sin embargo, dichos biciclamos carecen de biodisponibilidad debido a su carga total positiva. Actualmente se está trabajando en este punto, por lo que AnorMED ha desarrollado una nueva generación de biciclamos, protagonizada por un derivado oral biodisponible del AMD3100, AMD070 (de estructura no revelada), el cual está en fase clínica I/II.

Asimismo se han realizado estudios con derivados monociclamos como AMD3465 [98] o AMD3451 [99]. El N-piridinilmetilen ciclamo (AMD3465), con solo un anillo de ciclamo (1,4,8,11-tetraazaciclotetradecano), es 10 veces más efectivo como antagonista del CXCR4, no mostrando interacción con el CCR5. Sin embargo, el AMD3451, es antagonista tanto para CXCR4 como para CCR5. La biodisponibilidad de dichos monociclamos no se ha logrado todavía. En comparación con los biciclamos, gracias a su menor carga molecular, son un importante paso hacia el diseño de antagonistas orales del CXCR4.

La Figura I.14 muestra las estructuras de algunos de los derivados biciclamos (izquierda y centro) y monociclamos (derecha) en estudio clínico.

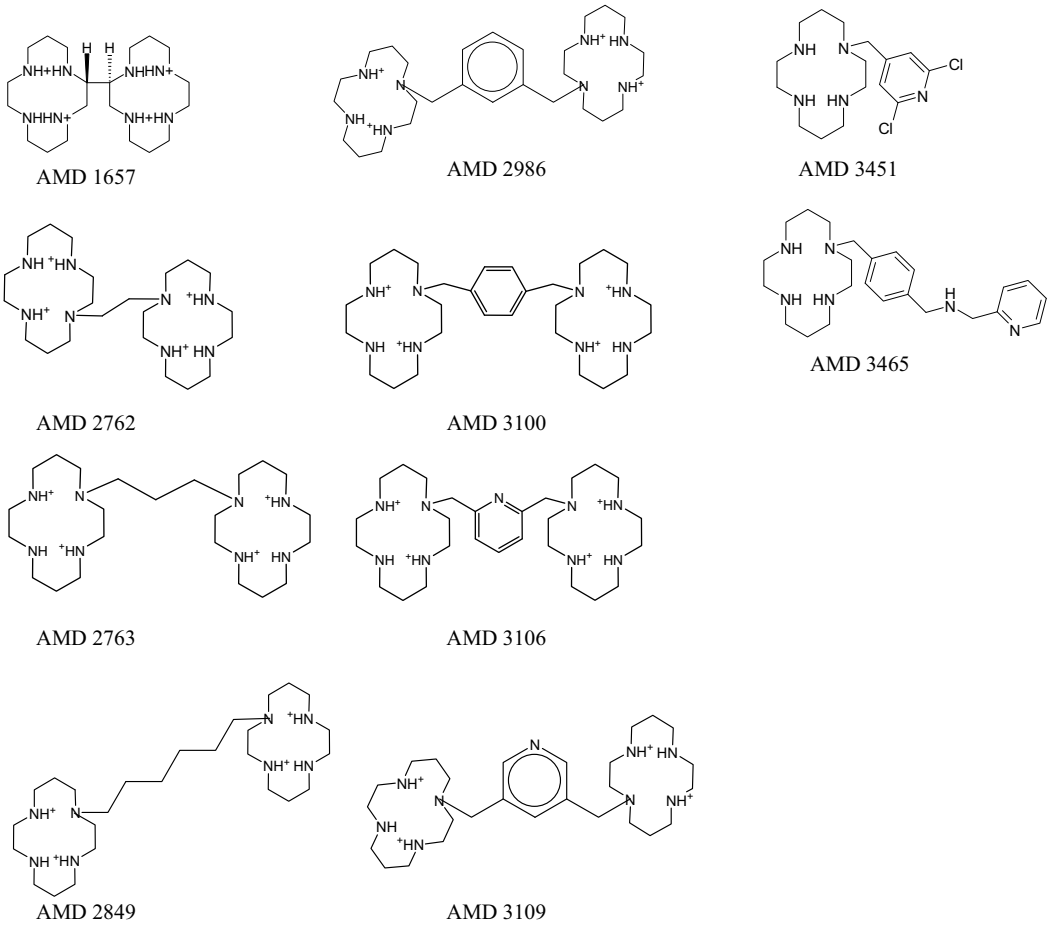


Figura I.14 Derivados biciclamos (izquierda y centro) y derivados monociclamos (derecha) [96, 98, 99].

2.- Derivados de **azamacrociclos** [100] como los derivados de 1,10-fenantrolina, 2,2'bipiridilo o derivados de bis-piridil macrociclos cuyo tamaño de anillo varía entre 12 y 16 miembros por anillo (Figura I.15).

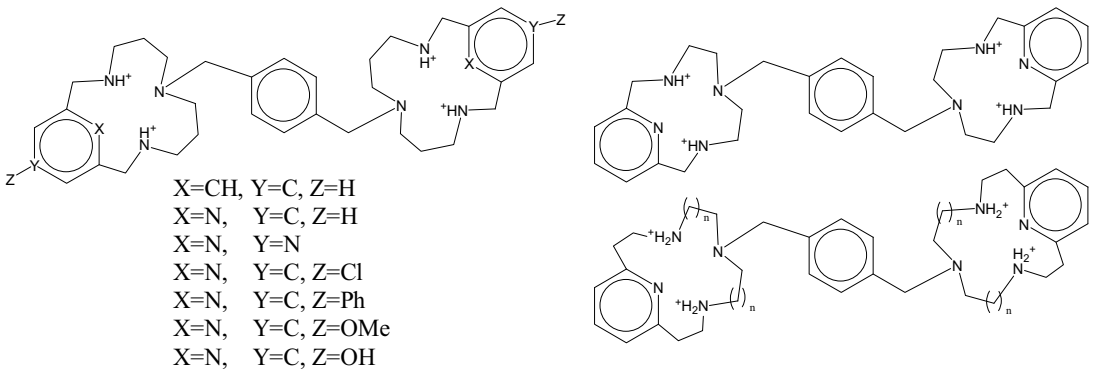


Figura I.15 Derivados azamacrocíclicos [100].

3.- Derivados de **Tetrahidroquinolinamina** [82, 101, 102, 103, 104, 105] desarrollados por AnorMED, en los cuales se remplaza la estructura de ciclamo por *N*-(1H-benzimidazol-2-ilmetil)- 5,6,7,8-tetrahidro-8-quinolinamina (Figura I.16).

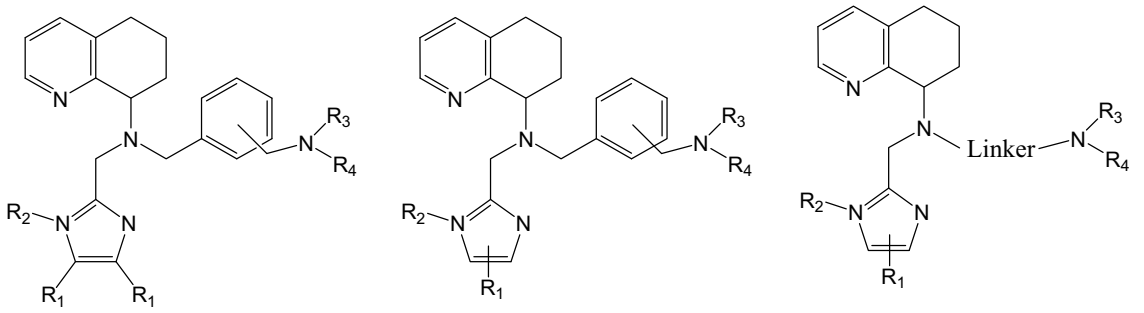


Figura I.16 Derivados Tetrahidroquinolinamina [82].

4.- Derivados de **KRH1636** [82, 106, 107, 108, 109], antagonista potente del CXCR4, el cual muestra una potencia similar al AMD3100. Desarrollados por Kuhrea Chemical Insdustry Co Ltd (Figura I.17).

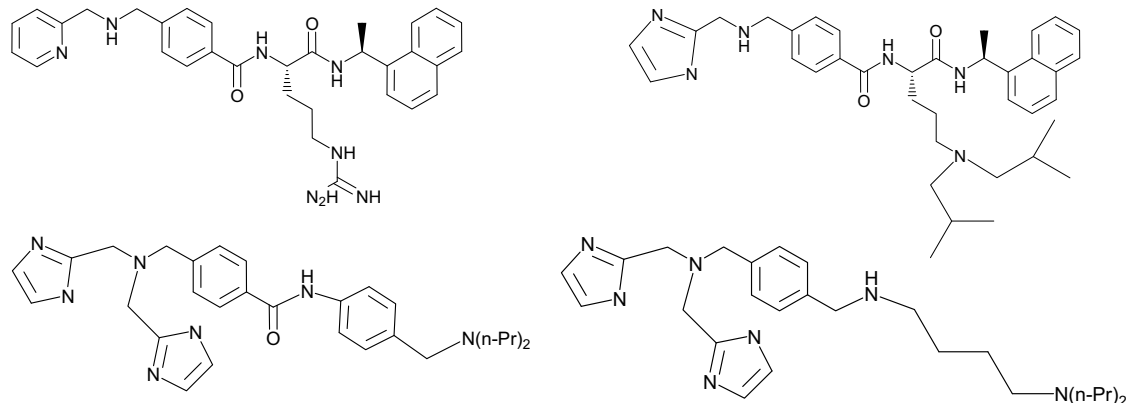


Figura I.17 Derivados de KRH1636 [82].

5.- Complejos de **dipicolil amina zinc(II)** [110], antagonistas potentes de CXCR4 con motivos aromáticos semejantes a los de los inhibidores T140, FC131, AMD3100 and KRH-1636, lo cual sugiere que estos fragmentos pueden ser importantes para la interacción con el CXCR4.

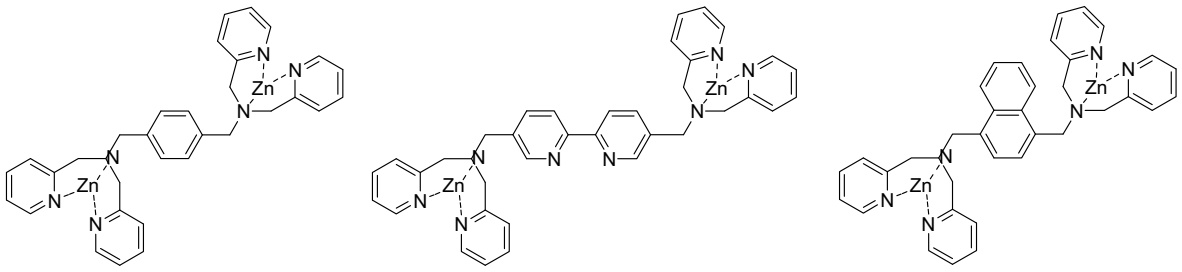


Figura I.18 Derivados de dipicolil amina zinc(II) [110].

Antagonistas peptídicos y proteínas

1.- T22 [111, 112], **T134** [113] y **14-mer T140** [114]. Péptidos bloqueantes de la función del coreceptor CXCR4 y la entrada de VIH X4, desarrollados basándose en un péptido antimicrobial natural derivado del cangrejo (polifemusinas). La molécula *lead*, T22 (Figura I.19), es un péptido que consta de 18 aminoácidos, el cual bloquea la infección para diferentes subtipos de VIH. A partir de ésta, se desarrollan análogos de menor peso molar como T134 o 140. El 14-mer T140, o simplemente T140 (Figura I.20), es un agonista inverso, el cual presenta elevada actividad; no tiene una actividad agonista intrínseca, bloquea la activación que causa el ligando natural y por lo tanto inhibe la actividad del receptor.

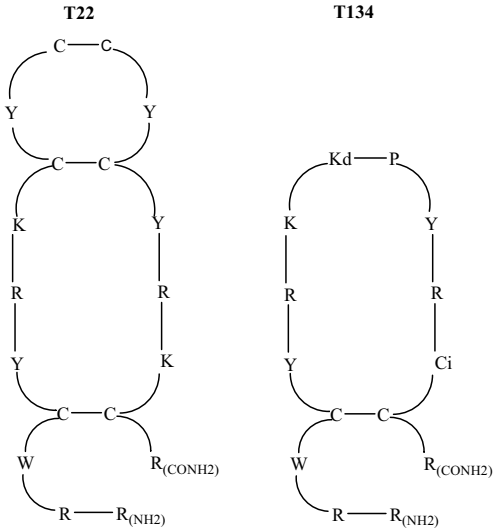


Figura I.19 Estructura del T22 (izquierda) y T134 (derecha) [68].

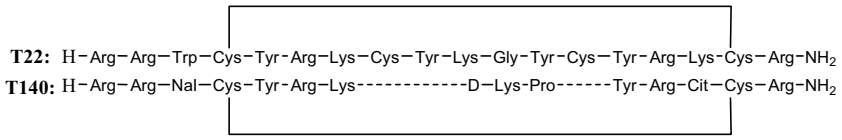
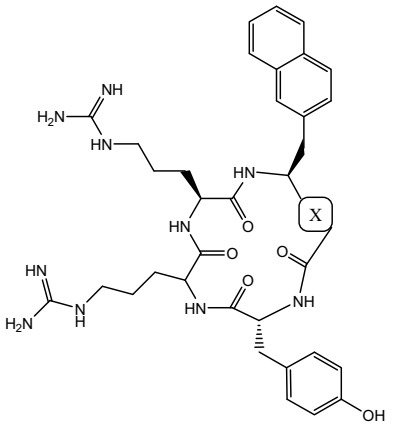


Figura I.20 T22 y T140: antagonistas peptídicos del CXCR4 [82].

2.- Derivados de **péptidos cíclicos** [115].



Sequence (*cyclo*)	X
(-γ-Na1[1]-D-Tyr[2]-D-Arg[3]-Arg[4]-)	$-CH_2-CH_2-$
(-γ-Na1[1]-D-Tyr[2]-Arg[3]-Arg[4]-)	$-CH_2-CH_2-$
(-γ-(*E*)-Na1[1]-D-Tyr[2]-D-Arg[3]-Arg[4]-)	-CH=CH-
(-γ-(*E*)-Na1[1]-D-Tyr[2]-Arg[3]-Arg[4]-)	-CH=CH-

Figura I.21 Péptidos cíclicos, antagonistas del CXCR4 [115].

3.- ALX40-4C [116] y **CGP 64222** [117]. El *N*-α-acetil-nona-D-arginina amida acetato, ALX40-4C, (Figura I.22, arriba) ha sido desarrollado como inhibidor competitivo de la unión de la proteína Tat del VIH a su diana, TAR, en la región situada en el extremo 5' del ARN viral.

CGP 64222 (Figura I.22, abajo) es un análogo peptídico de la proteína Tat, el cual bloquea el proceso de transactivación del Tat.

Asimismo, ambos compuestos se han mostrado activos como inhibidores de entrada del VIH bloqueando el coreceptor CXCR4 [118, 119, 120].

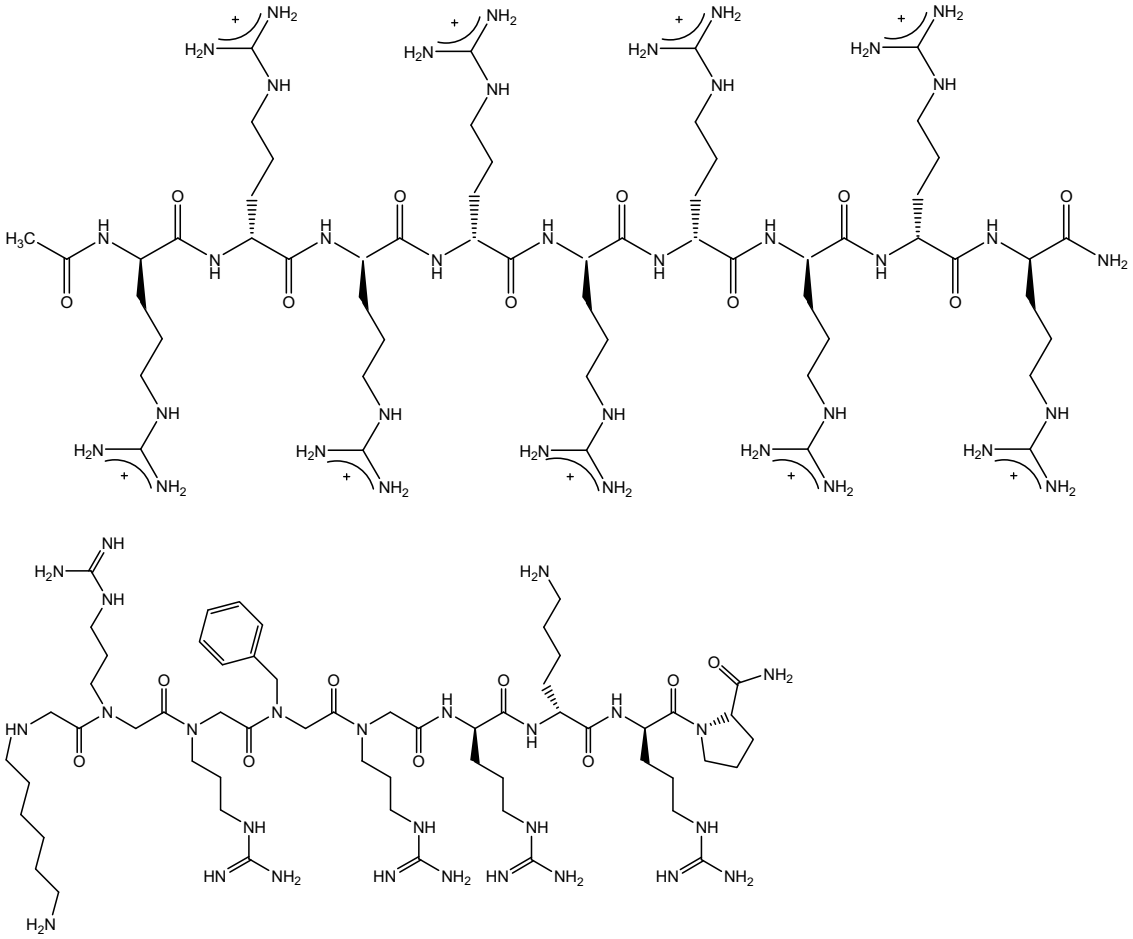


Figura I.22 ALX40-4C (arriba) y CGP 64222 (debajo) [68].

Agentes basados en quimiocinas

1.- Quimiocinas naturales o modificadas como la "***human herpesvirus 8-derived viral macrophage inflammatory protein II***", la cual bloquea CXCR4, así como CCR5 y CCR3 [121].

2.- Derivados de **SDF1-α modificado** [122].

A continuación se presentan las clases de inhibidores del CCR5 más importantes divididos en pequeñas moléculas, anticuerpos monoclonales, quimiocinas modificadas y otros compuestos [45, 82].

Pequeñas moléculas

1.- Derivados de **fenilciclohexilamina** [123, 124, 125, 126, 127, 128].

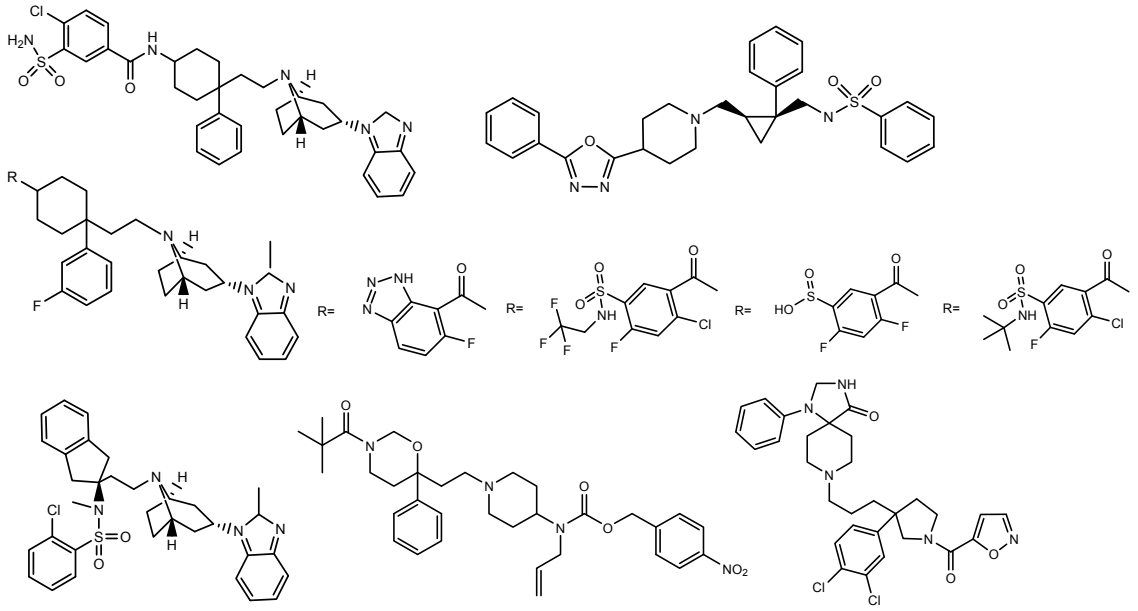


Figura I.23 Inhibidores de CCR5 de GlaxoSmithKline.Derivados de fenilciclohexilamina [82].

2.- Derivados de **dicetopiperazina**. ONO-4128/GW873140. E913 [129, 130, 131, 132, 133].

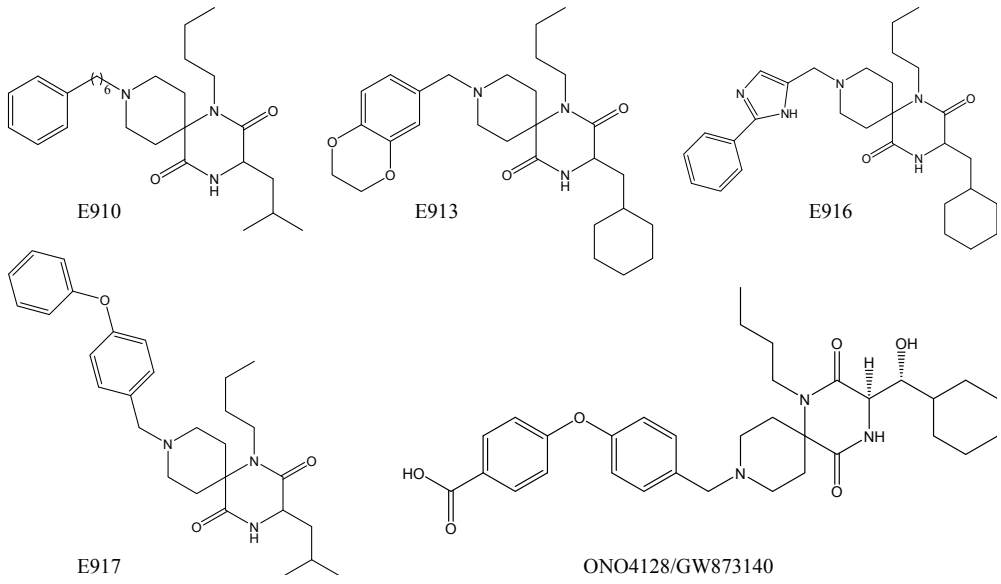


Figura I.24 Inhibidores de CCR5 de ONO Pharmaceuticals. Derivados de dicetopiperazina [82].

3.- Derivados de **SCH-C**. SCH-D [134, 135, 136].

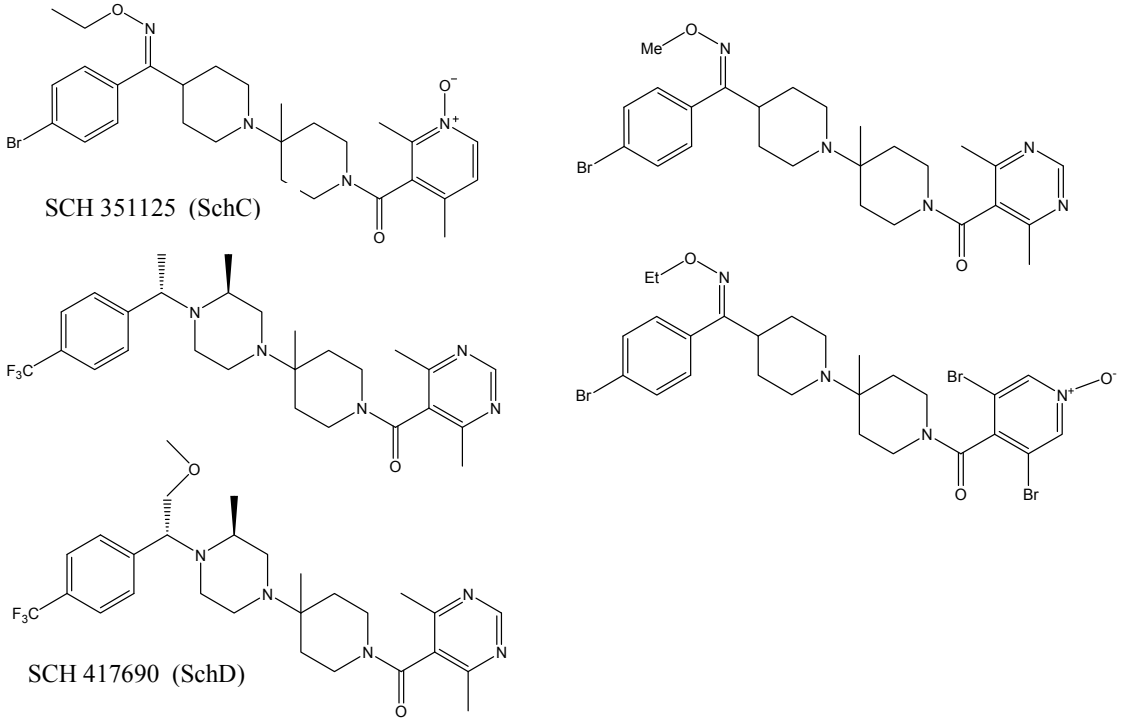


Figura I.25 Inhibidores de CCR5 piperidino-piperidina Schering-Plough [82].

4.- Derivados de **TAK779**. TAK770, TAK220 [137,138].

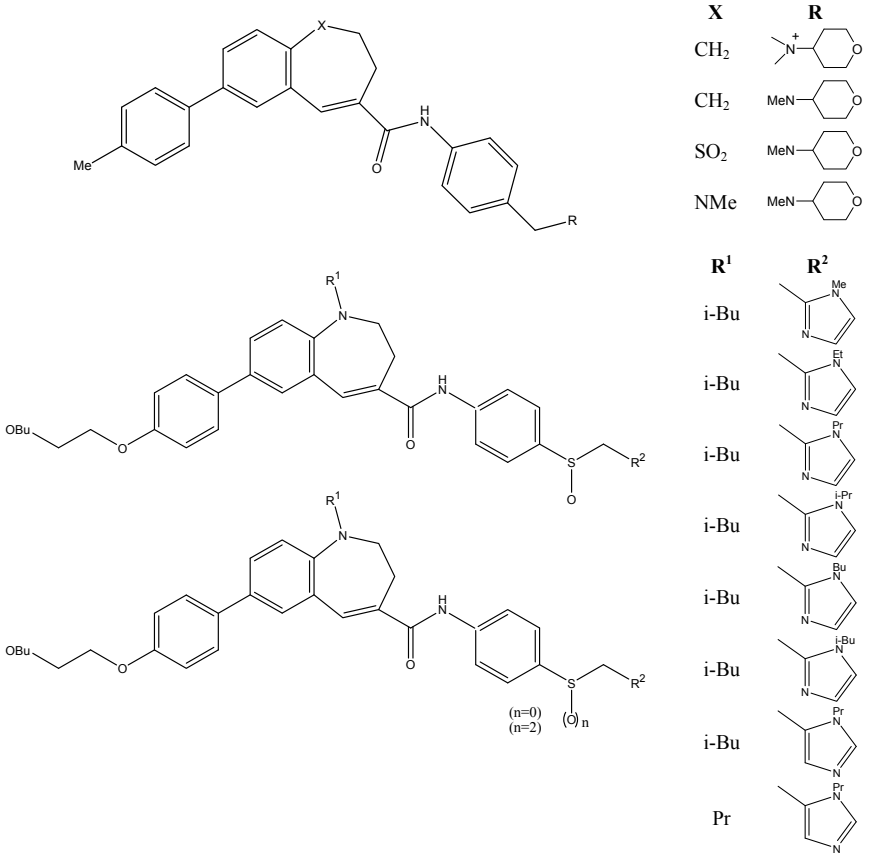
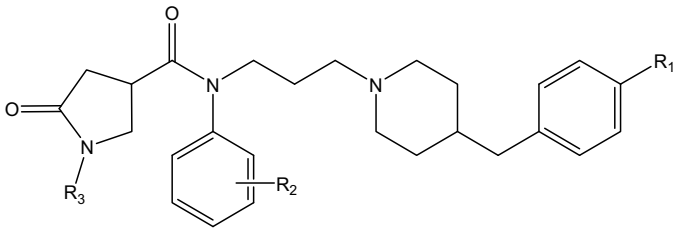


Figura I.26 Inhibidores de CCR5 de Takeda. TAK779 (arriba izquierda) y análogos [82].

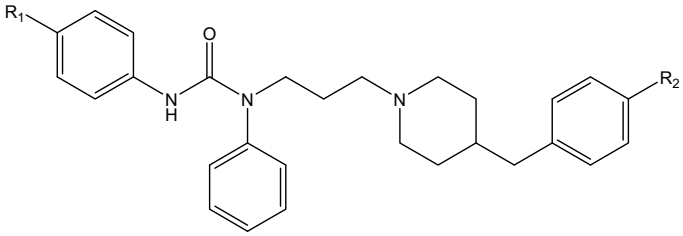
5.- Derivados de **5-oxopirrolidina-3-carboxamida** [139].



R_1	R_2	R_3
F	H	Me
H	3,4-diCl	Me
F	3,4-diCl	Me
H	H	$PhCH_2$
H	3,4-diCl	$PhCH_2$

Figura I.27 Inhibidores de CCR5 de Takeda derivados de 5-oxopirrolidina-3-carboxamida [82].

6.- Derivados de **urea** [140].



R_1	R_2
H	H
Cl	H
Cl	SO_2Me
Cl	SO_2(morfolino)

Figura I.28 Inhibidores de CCR5 de Takeda derivados de *N,N'*-difenilurea [82].

7.- Derivados de **1,3,4-pirrolidina-piperidina trisustituida** [141].

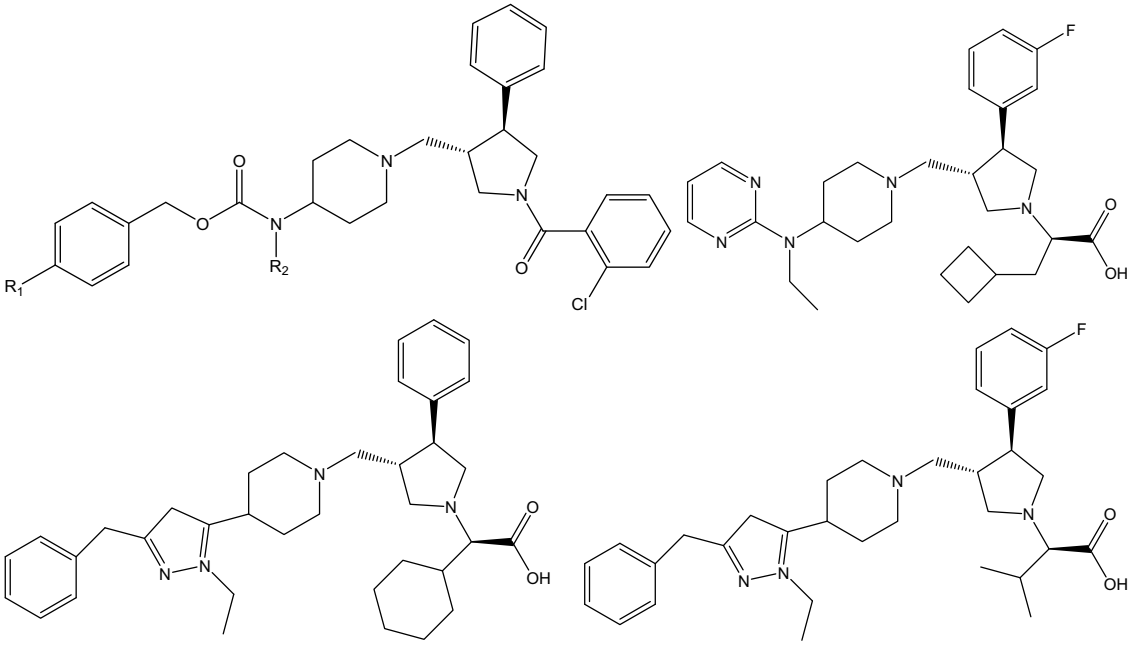
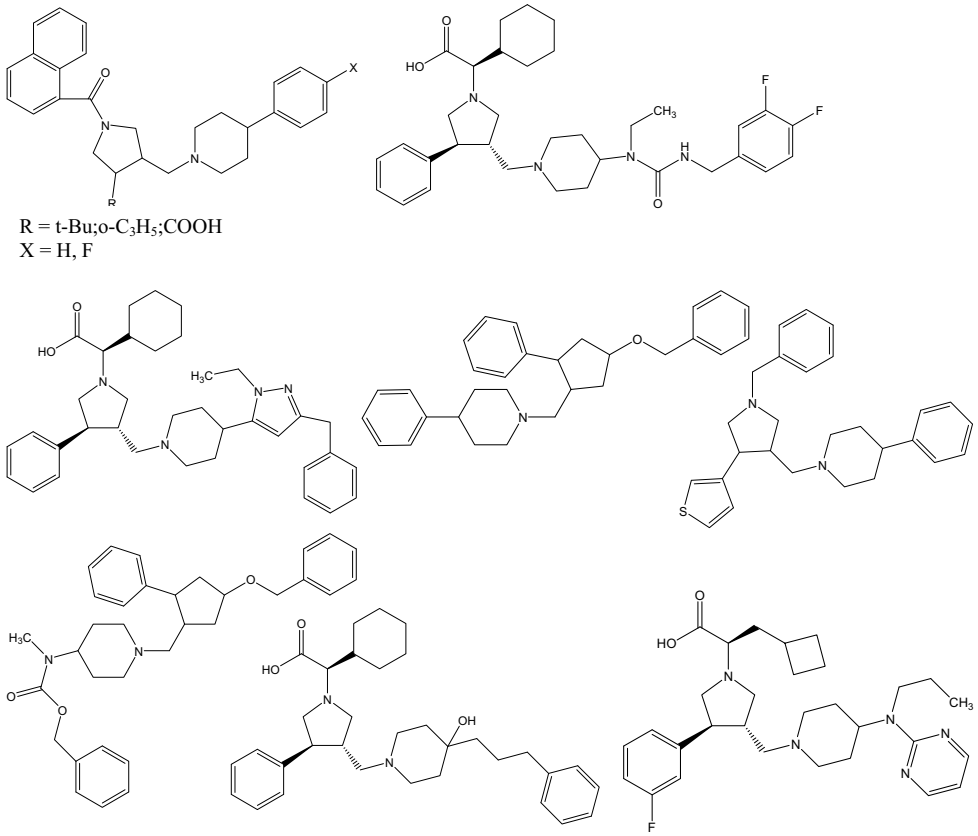


Figura I.29 Inhibidores de CCR5 de Merck Research Laboratorios derivados de 1,3,4-pirrolidina-piperidina trisustituida [82].

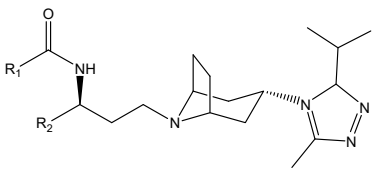
8.- Derivados de **anillos pentacíclicos 1,3,5 trisustituidos** [142].



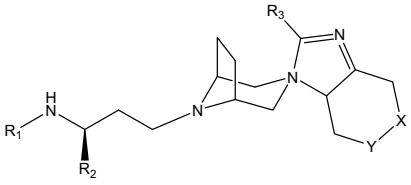
R = t-Bu;o-C_3H_5;COOH
X = H, F

Figura I.30 Inhibidores de CCR5 de Merck Research Laboratorios derivados de anillos pentacíclicos 1,3,5 trisustituidos [142].

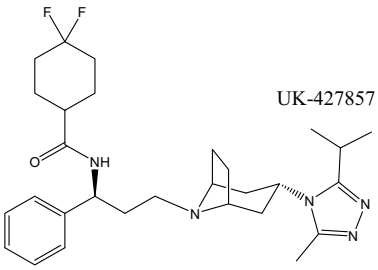
9.- Derivados de **1-fenil-1,3-propanodiamina**. Un representativo de esta clase es UK-427.857 o Maraviroc (Figura I.31).[143, 144, 145].



R_1 = cicloalquil, alquil, o cicloalquilmetil (todos ellos sustituidos o no por uno o más átomos de F)
R_2 = fenil sustituido opcionalmente por uno o más átomos de F



X, Y = CH_3 o NR
R_1 = COR, CO_2R, CONR, SO_2R
R_2 = fenil sustituido
R_3 = alquil sustituido



UK-427857

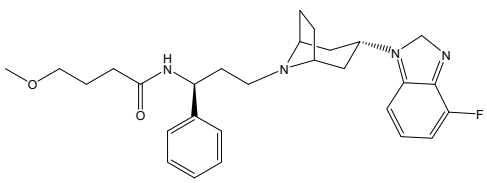


Figura I.31 Inhibidores de CCR5 de Pfizer derivados de 1-fenil-1,3-propanodiamina [82].

10.- Derivados de **4-amino-piperidina o tropano** [146, 147, 148].
11.- Derivados de **4-piperidina** [148, 149].
12.- Derivados de **anilida *N*-óxido de piperidina** [150].

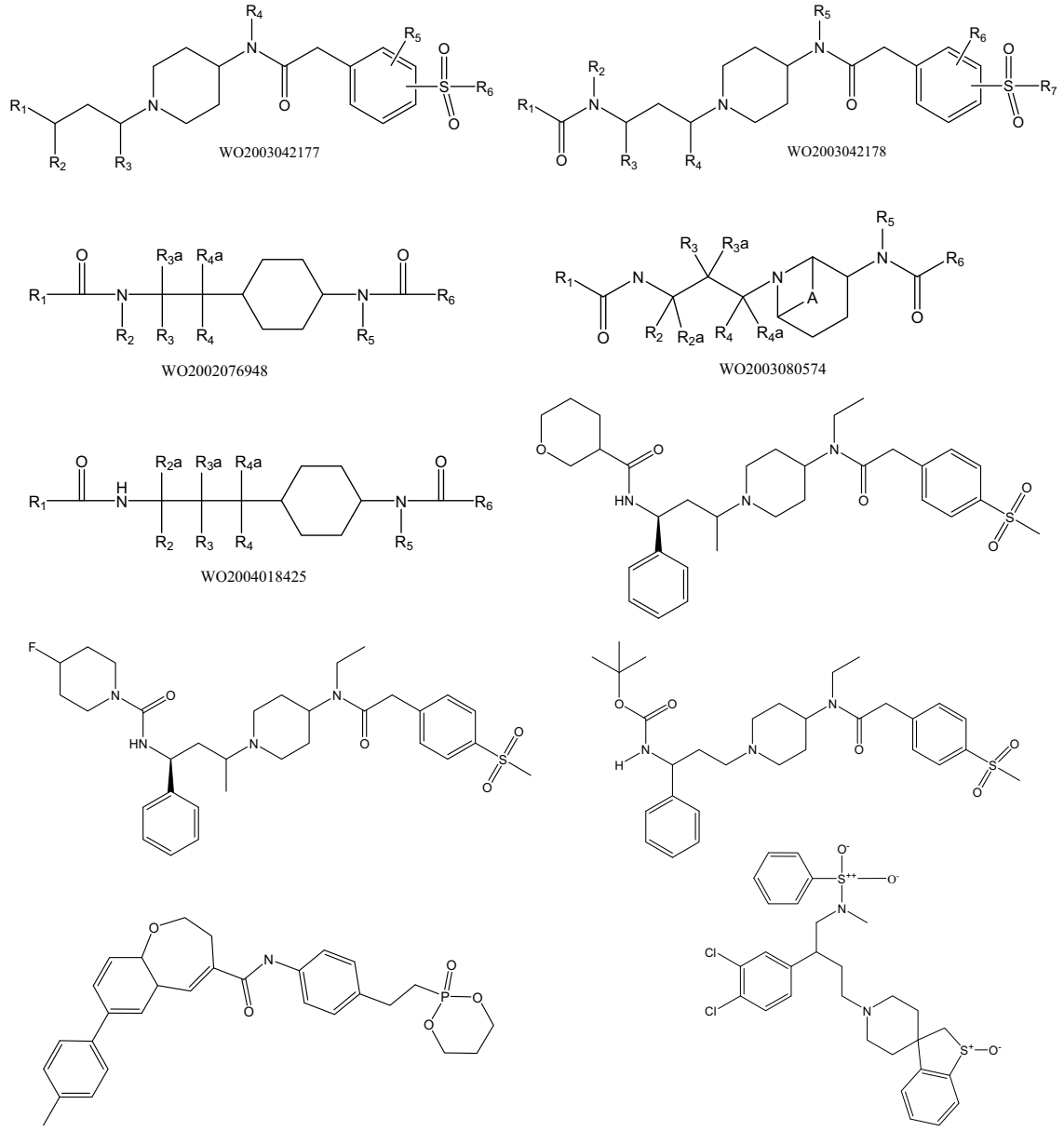


Figura I.32 Inhibidores de CCR5 de Astra Zeneca derivados de 4-amino piperidina o tropano. Inhibidores derivados de anilida *N*-óxido de piperidina y de 4-piperidina [82].

12.- Derivados de **guanilhidrazona** [151].

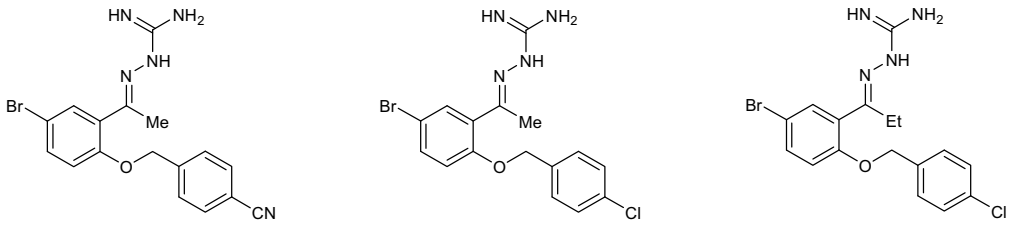


Figura I.33 Inhibidores de CCR5 derivados de guanilhidrazona [151].

13.- Derivados de **4-hidroxipiperidina** [152].

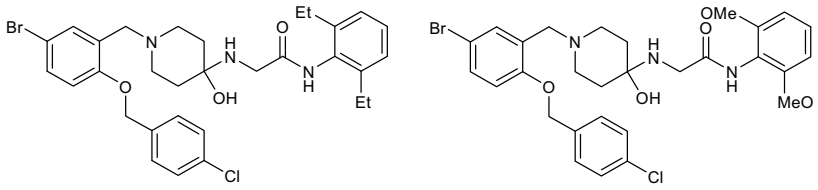


Figura I.34 Inhibidores de CCR5 derivados de 4-hidroxipiperidina [152].

Anticuerpos monoclonales

1.- PRO 140 (Progenies Pharmaceuticals Inc. Tarrytown, NY, USA), el cual está en ensayos clínicos de fase II [45, 153].

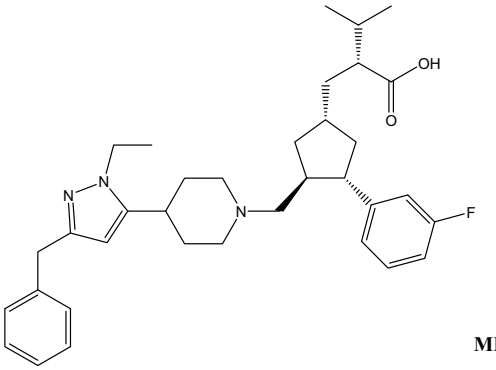
Quimiocinas modificadas

1.- Derivados de la quimiocina natural RANTES: aminooxipentano-RANTES, *n*-nonanoil-RANTES y PSC-RANTES [45, 154].

2.- Análogos de RANTES sintéticos llamados **nonaquinas** [45, 155].

Otros compuestos

1.- MRK-1 [156], Merck1, Merck2, Merck3, Merck compuesto 167 (CMPD 167) [157, 158, 159] desarrollados por Merck Research Laboratorios.



MRK-1

Figura I.35 Inhibidor de CCR5 de Merch Research Laboratorios [68].

I.2.6 Interacciones inhibidor-CXCR4 descritas. Predicción de la unión AMD3100-CXCR4

El AMD3100 [160, 161] (véase sección I.2.5) es un bicíclamo simétrico compuesto de dos 1,4,8,11-tetraazaciclotetradecano (cíclamo) conectados por un *linker* aromático que provoca restricción conformacional. Se conoce como el prototipo de los antagonistas no peptídicos del coreceptor CXCR4. El compuesto fue descubierto como un agente anti-VIH mucho antes del descubrimiento de su función de bloqueo del coreceptor CXCR4. Dicho compuesto es un antagonista específico del CXCR4 que inhibe la unión y función de su ligando natural SDF-1α con elevada afinidad y potencia. Sustituciones mutacionales en 16 aminoácidos situados en las hélices transmembranas TM3, TM4, TM5, TM6 y TM7 han identificado tres residuos ácidos: Asp171, Asp262 y Glu288 como puntos principales de interacción para el AMD3100; dos de los cuales están en un extremo (Asp262 en la TM6 y Glu288 en la TM7) y el tercero está en el extremo opuesto (Asp171 en la TM4) de la cavidad de unión identificada. Además, Asp171 y Asp262 también han sido identificados como esenciales para la función del receptor CXCR4 como coreceptor del VIH [33, 162-170].

Mediante los experimentos de mutagénesis dirigida se observa si la sustitución de un residuo por otro provoca un aumento o disminución de actividad, lo cual ayuda a deducir qué residuos son los que intervienen en la unión del inhibidor al coreceptor. Los residuos Asp e His del coreceptor CXCR4 son las dianas de la mutagénesis, de acuerdo con la suposición de que cada anillo de ciclamo del AMD3100 y análogos podría estar doblemente protonado a pH fisiológico, o podría complejar a metales de transición como zinc *in vitro* para dar un complejo metal con una carga total de +2. Sea cual sea el análogo biciclamo, los residuos Asp e His podrían intervenir en el complejo. Por ello todos los residuos His situados en los *loops* extracelulares o en los dominios transmembrana se mutan por residuos Ala y los residuos Asp situados en las transmembranas IV y VI y en el *loop* extracelular II son mutados por residuos Asn [171]. En la Figura I.36 se muestran los residuos identificados por mutagénesis dirigida como importantes para la unión del AMD3100 y AMD3100 complejado con Zn^{2+} al coreceptor CXCR4. En algunos casos, se ha llevado a cabo una mutagénesis con impedimento estérico introduciendo grandes cadenas laterales como Phe o Trp por Ala, Gly o Ile. De todas las sustituciones realizadas, solo cinco de ellas han ocasionado problemas en la unión del AMD3100, decreciendo su afinidad por el AMD3100 más de 10 veces [171].

La unión de metales a los anillos de ciclamo del AMD3100 parece que hace aumentar su dependencia del Asp262 y proporciona una mayor afinidad para el coreceptor CXCR4. Parece ser que el efecto del metal es mediado aparentemente únicamente a través de la interacción con el Asp262. Mientras que el Asp171 y Glu288 son ambos muy importantes para la unión del compuesto, el efecto del metal se confina al Asp262 [167]. Los iones metales como Zn^{2+}, y Cu^{2+} se unen al anillo de ciclamo fuertemente (valores de $\log K_i \approx 15$ y 27 respectivamente) y relativamente rápido. Para dichos complejos metal-biciclamo, existe una correlación entre actividad antiviral y unión al coreceptor CXCR4 como se monitoriza por la inhibición de la unión del anticuerpo monoclonal 12G5 y la señal intracelular de Ca^{2+} inducida por la quimicocina SDF1α, siendo el orden de actividad decreciente: $Zn^{2+} \approx Ni^{2+} > Cu^{2+} >> Co^{3+} >> Pd^{2+}$. La afinidad del AMD3100 por el coreceptor CXCR4 se incrementa en factores de 7, 36 y 50 incorporando Cu^{2+}, Zn^{2+}, o Ni^{2+}, respectivamente, en los anillos de ciclamo [169].

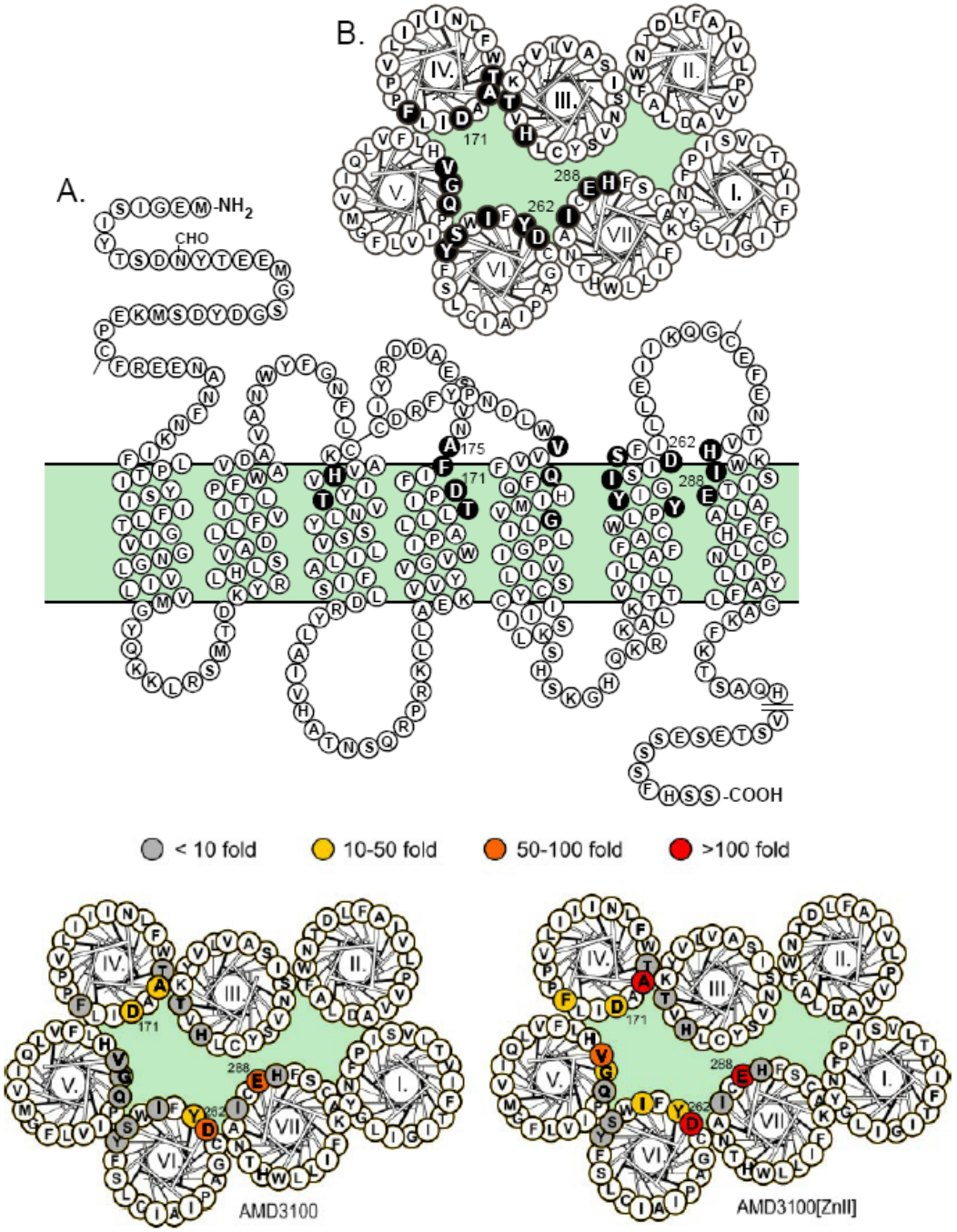


Figura I.36 Residuos identificados por mutagénesis dirigida importantes en la unión del AMD3100 y AMD3100 complejado al coreceptor CXCR4. En la parte superior de la figura, diagrama de serpentín (A) y helicoidal (B) del coreceptor CXCR4. Las letras blancas en círculos negros indican residuos mutados, sustituidos por otros aminoácidos para reducir el tamaño de la cadena lateral, neutralizar la carga (Ala o Asn) o incrementar el tamaño de la cadena lateral (Phe o Trp) como aproximación estérica. En la parte inferior de la figura, se presentan en un diagrama helicoidal del coreceptor CXCR4, los residuos identificados por mutagénesis importantes en la unión del AMD3100 y el AMD3100 complejado con Zn. El color de fondo indica la magnitud del efecto de la mutación en la unión de los ligandos. El fondo gris indica una reducción de la afinidad unas 10 veces menor; amarillo indica una reducción de la afinidad de 10 a 50 veces menor; naranja indica una reducción de la afinidad de 50 a 100 veces menor; y rojo implica una reducción de la afinidad más de 100 veces menor [171].

Según los trabajos computacionales descritos hasta la fecha, parece ser que, en principio, los anillos de ciclamo de un compuesto biciclamo tienen preferencia por unirse a grupos ácido carboxílico [167-171]; ahora bien, lo pueden hacer de diversas maneras:

· Dos nitrógenos de un anillo de ciclamo interaccionando con los dos oxígenos del grupo ácido del Asp171 y dos nitrógenos del otro anillo de ciclamo interaccionando con los dos oxígenos del grupo ácido del Asp262. Este modelo es el propuesto por el grupo de Schwartz [168] para la unión AMD3100-CXCR4 (véase Figura I.37.d).

· Dos nitrógenos de un anillo de ciclamo interaccionando con los dos oxígenos del grupo ácido del Asp262 y dos nitrógenos del otro anillo de ciclamo interaccionando con los dos oxígenos del grupo ácido del Glu288. Este modelo es el propuesto por el grupo de Trent [170] para la unión AMD3100-CXCR4 (véase Figura I.37.c).

· Los nitrógenos de los dos anillos de ciclamo pueden interaccionar con los tres carboxilatos de los residuos ácidos Asp262, Asp171 y Glu288 a la vez. Estudios computacionales describen que con un modelo de CXCR4 obtenido a partir de la forma inactiva de la rodopsina sin mover el esqueleto, básicamente solo ajustando las conformaciones de las cadenas laterales, se ha encontrado que dos nitrógenos de un anillo de ciclamo interaccionan con los dos oxígenos del ácido carboxílico del Asp171 de la TM4, mientras que el otro anillo de ciclamo interacciona por una cara con los dos oxígenos del ácido carboxílico del Asp262 de la TM6 y por la cara opuesta con los dos oxígenos del ácido carboxílico del Glu288 de la TM7, formando un "sandwich". Lo mismo ocurre si el anillo de ciclamo está complejado con un metal. Una posible configuración de dicho anillo unido a dos residuos ácidos es la conformación cis-V propuesta por Sadler [169], en la que el Asp262 coordina al metal y el Asp171 se une a través de dos puentes de hidrógeno al otro anillo de ciclamo en configuración trans-I. Este modelo es el propuesto por el grupo de Schwartz [171] y el grupo de Sadlder [169] para la unión AMD3100(Zn_2)-CXCR4 (Figuras I.37.a y .b).

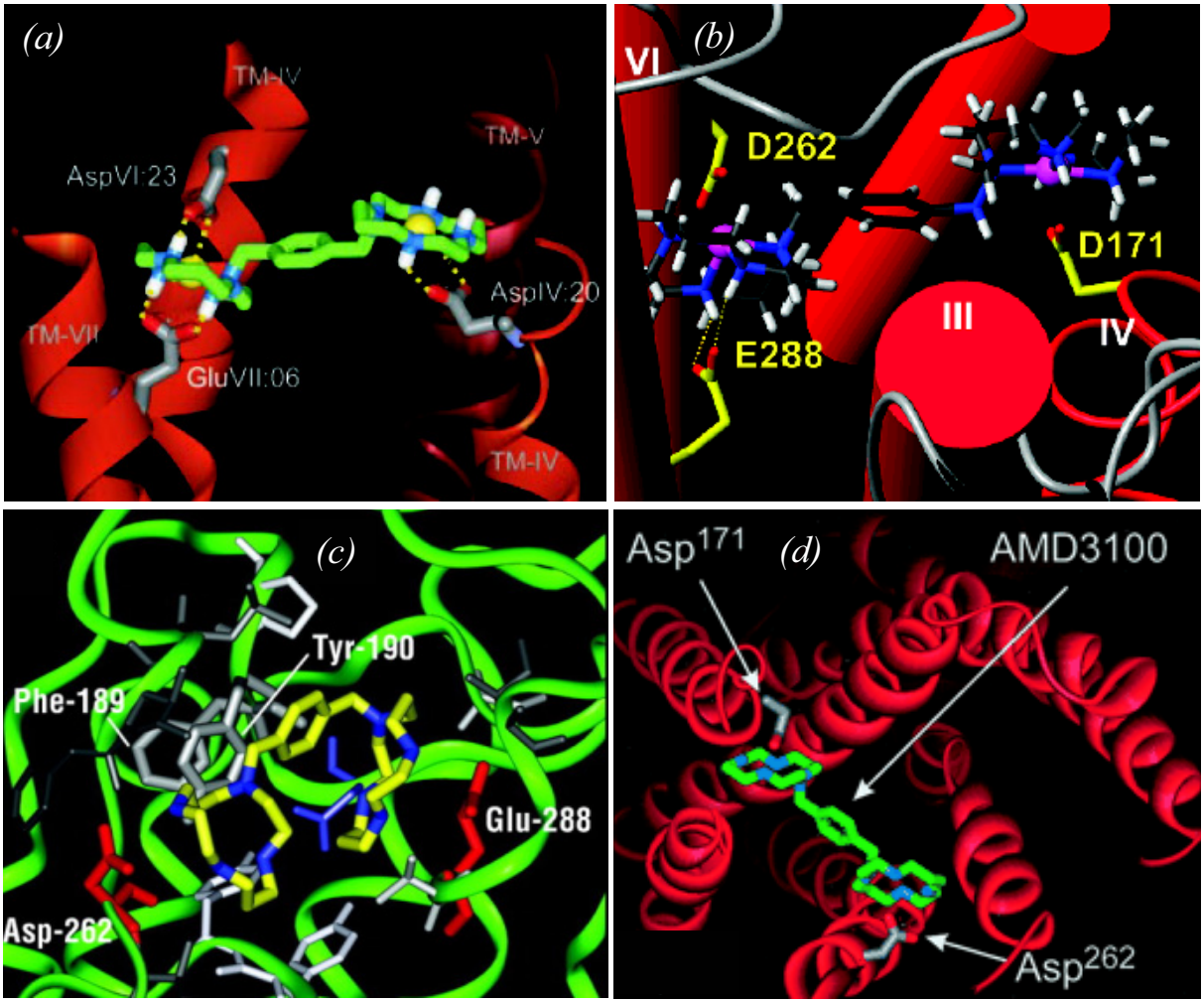


Figura I.37 Modelo molecular del modo de unión supuesto entre el AMD3100(Zn2) y el coreceptor CXCR4 (a,b) y el modo de unión supuesto entre el AMD3100 y el coreceptor CXCR4 (c,d).(a) Modelo propuesto por Schwartz y su equipo para el AMD3100(Zn2). El modelo del receptor está construido tomando como plantilla la estructura cristalográfica de la rodopsina y muestra la interacción de uno de los anillos de ciclamo del AMD3100(Zn2) con el Asp171, mientras que el otro anillo de ciclamo forma un "sándwich" entre el Asp262 y el Glu288 [171]. (b) Modelo propuesto por Sadler y su equipo del *docking* entre el cis-V/trans-I AMD3100(Zn2) y el coreceptor CXCR4 modelado por homología a partir de la rodopsina bovina. Las hélices transmembrana se muestran como cilindros rojos, excepto la hélice IV, para mayor claridad. El anillo de ciclamo cis-V se coordina con el Asp262 mediante enlaces Zn(II)-carboxilato de longitud similar (2.27 Å) y vía doble puente de hidrógeno por la cara opuesta entre los NH del ciclamo y los oxígenos del Glu288 (COO·····H-N, 2.01 Å). El anillo de ciclamo trans-I se une a los oxígenos del Asp171 (Zn-O, 2.28 Å) [169]. (c) Modelo propuesto por Trent y su equipo para el AMD3100. El modelo es obtenido mediante dinámica molecular con el force field Amber-96. En él se muestran los residuos críticos para la unión del AMD3100 al CXCR4. El CXCR4 se presenta mediante hélices verdes con los residuos críticos para la unión del ligando encontrados experimentalmente según un ensayo de fusión, en representación de *stick*, etiquetados de color blanco: Phe189, Tyr190, Asp262, Glu288. Se muestra cómo un anillo de ciclamo interacciona con el Asp262, mientras que el otro interacciona con el Glu288 [170]. (d) Modelo propuesto por Schwartz y su equipo para el AMD3100. El modelo del receptor es el construído a partir de la rodopsina por homología. En él se muestra cómo un anillo de ciclamo interacciona con el Asp171, mientras que el otro interacciona con el Asp262 [168].

I.2.7 Interacciones inhibidor-CCR5 descritas. Predicción de la unión TAK779-CCR5

Estudios inmunológicos y de mutagénesis dirigida sobre el coreceptor CCR5 han determinado que la mayoría de inhibidores interaccionan principalmente con una cavidad de unión común en el dominio transmembrana del coreceptor CCR5; ahora bien, algunos compuestos dan lugar a contactos adicionales [135, 172, 173].

Los estudios de mutagénesis dirigida para la localización del sitio de unión del TAK779, AD101 y SCH-C (véase sección I.2.5) indican que las cadenas laterales de los residuos Glu283, Trp86, Tyr37, Tyr108, Leu33, Val83, Ala90 y Gly286 en las transmembranas TM1, TM2, TM3 y TM7, forman el sitio de unión del CCR5 para estas moléculas [173] (Figura I.38). Además, los resultados sugieren que el Glu283 hace la función de contraión para el átomo de nitrógeno cargado positivo común a TAK779, AD101 y SCH-C. Este Glu283 no es accesible a la superficie, con lo que no tiene un contraión aparente en el coreceptor CCR5. La ausencia de contraiones cargados positivos en las cercanías del Glu283 facilita una interacción iónica con los inhibidores cargados positivos. Otro elemento clave del supuesto sitio de unión es el conjunto de residuos aromáticos. En particular, el Trp86 es indispensable para la interacción del CCR5 con TAK779, AD101 y SCH-C. Ello es debido posiblemente a la interacción del anillo indol del Trp86 con las regiones hidrofóbicas de los antagonistas. Sin embargo, debido al momento cuadrupolar del anillo de indol, también son posibles interacciones del Trp86 con grupos polares de los compuestos. La Tyr37 también es importante en la interacción de los tres compuestos con el CCR5, aunque, a diferencia del Trp86, la interacción parece ser diferente para AD101 y SCH-C frente al TAK779 [173].En cuanto a los residuos Val83, Ala90 y Gly286, el primero no muestra una reducción significante de la inhibición frente a la sustitución con cadenas laterales polares o alifáticas de longitud similar, pero sí una moderada reducción de la inhibición al sustituirlo por alanina. Estas observaciones sugieren que la Val83 puede estar situada en la periferia de la cavidad de unión. Los residuos Ala90 y Gly286, los cuales se piensa que definen la entrada y la parte inferior de la cavidad de unión respectivamente, según el grupo de Seibert [173], no contribuyen mucho a la energía de unión de los inhibidores debido a sus pequeñas cadenas laterales. Sin embargo, estos residuos pueden ser importantes para conformar la cavidad de unión y crear suficiente espacio para que las moléculas antagonistas puedan entrar en su interior. Estos residuos, al ser sustituidos por residuos con largas cadenas alifáticas, bloquean la interacción del CCR5 con TAK779, AD101 y SCH-C.

Estudios computacionales sobre un modelo de CCR5, realizados por el grupo de Seibert [173], predicen que existen otros residuos, un poco más alejados del sitio de unión propuesto, que también afectan a la actividad, como son las cadenas laterales de los residuos Arg31, Phe113 o Ile198. La mutación de estos residuos puede afectar, de manera indirecta, a la conformación del sitio de unión de TAK779, AD101 y SCH-C. O bien, es posible que el efecto inhibidor de estas moléculas requiera cambios conformacionales en el CCR5, los cuales no pueden ocurrir con los residuos mutados del CCR5. Por otra parte, la sustitución de los residuos Asp76, Phe79 y Thr82, los cuales se sitúan cerca del sitio de unión, podrían actuar indirectamente de una de estas formas. Sin embargo, es también posible que estos residuos puedan interaccionar directamente con las regiones periféricas de los inhibidores.

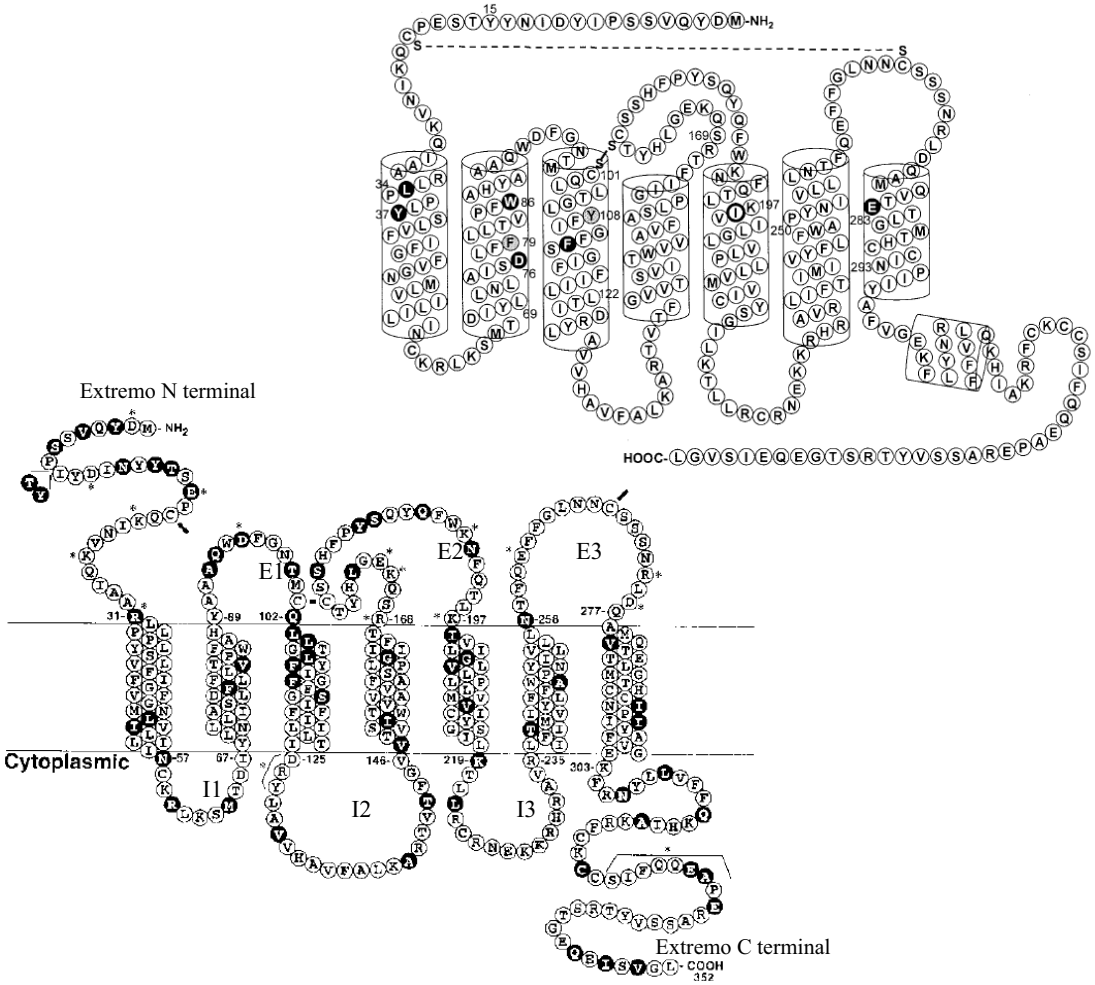


Figura I.38 Coreceptor CCR5 (arriba) [135]. Los residuos importantes para la unión de las moléculas activas AD101 y SCH-C están marcados en negro en la zona transmembrana. Modelo del coreceptor CCR5 (abajo) [72] en el que se muestran los *loops* extracelulares, las siete hélices transmembrana, y los *loops* intracelulares. Los residuos estudiados por mutagénesis dirigida están marcados en negro.

A continuación se analizan en detalle los estudios de mutagénesis dirigida que han caracterizado el sitio de unión del TAK779. Las posiciones de los residuos transmembrana implicados en la unión del TAK779 se muestran en el modelo de CCR5 de Dragic *et al.* [172] representado en las Figuras I.39 y I.40. La mutagénesis dirigida, así como el modelo de Dragic *et al.*, sugieren que el sitio de unión del TAK779 es una cavidad rodeada por las hélices transmembrana 1, 2, 3 y 7. Los residuos prolina en las posiciones 34, 35 y 84 en las hélices transmembrana 1 y 2 pueden facilitar la apertura de la cavidad de unión para el TAK779. La profundidad de la cavidad es aproximadamente igual a una tercera parte de la membrana y es aproximadamente de la misma longitud que el grupo metilfenil-benzocicloheptenil del TAK779. Este grupo hidrofóbico conjugado es planar y está conectado al bencil-piran-amonio, positivamente cargado, por un enlace amida. El enlace covalente amida está en "meta" con el grupo ciclobencil en el anillo heptenil, dando lugar a una torsión de aproximadamente 90º en la mitad de la estructura del TAK779 minimizada (Figura I.40 A). Dado que esta estructura posee extremos hidrofóbicos y polares distinguibles, Dragic *et al.* sugieren que el grupo metilfenilbenzocicloheptenil esté insertado en la transmembrana hidrofóbica de la cavidad, permitiendo a la parte cargada del TAK779 sobresalir e interaccionar con el dominio extracelular polar.

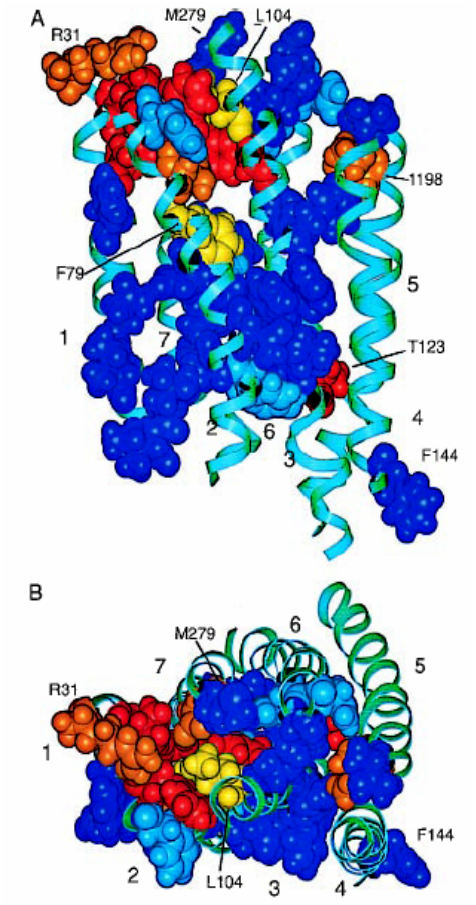


Figura I.39 Modelo estructural del dominio transmembrana del coreceptor CCR5. Los segmentos helicoidales transmembrana, marcados del 1 al 7, se muestran como cintas de color cian. Los residuos aminoácidos sustituidos por alanina se muestran con átomos representados como bolas coloreados según el siguiente código: las sustituciones por alanina de los residuos coloreados en rojo tienen un fuerte efecto inhibidor de la actividad antiviral del TAK779 (Leu33, Tyr37, Trp86, Tyr108, Thr123); las sustituciones por alanina de los residuos coloreados en naranja tienen un efecto inhibidor intermedio (Arg31, Thr82, Ile198, Glu283); las sustituciones por alanina de los residuos coloreados en amarillo tienen un efecto inhibidor dudoso (Phe79, Leu104); las sustituciones por alanina de los residuos coloreados en azul oscuro no tienen efecto inhibidor (Phe41, Asn48, Ile52, Leu55, Ile56, Leu69, Asn71, Asp76, Thr105, Phe112, Phe113, Phe117, Phe118, Leu121, Leu122, Phe144, Thr195, Leu255, Asn258, Thr259, Met279, His289, Tyr297). Los residuos coloreados de azul claro indican mutaciones que causan una mala expresión del CCR5 (Tyr68, Phe85, Tyr251, Asn252, Asn293). Estos residuos no podrían ser evaluados para la entrada del VIH-1. (A) Vista del CCR5 desde el plano de la membrana. La superficie extracelular se encuentra en la parte superior de la figura, la superficie citoplasmática en la parte inferior. Para mayor orientación, Arg31 se encuentra en la parte superior izquierda en naranja, y Phe144 se encuentra en la parte inferior derecha en azul. (B) Vista del CCR5 desde la superficie extracelular. El modelo está rotado unos 90º respecto la orientación anterior [172].

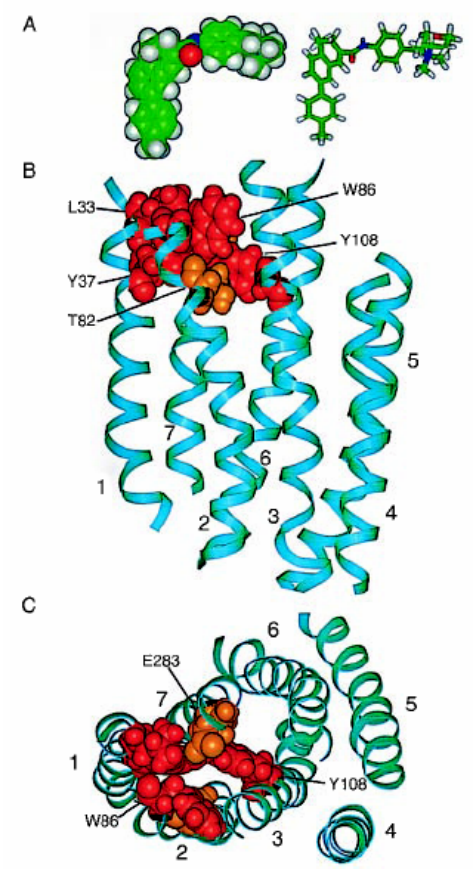


Figura I.40 Modelos estructurales del inhibidor TAK779 y del coreceptor CCR5. (A) Representaciones en bolas y palos de la estructura del TAK779 minimizado. Los átomos están coloreados según el código: carbono, verde; oxígeno, rojo; nitrógeno, azul; hidrógeno, gris. El TAK779 tiene dos segmentos aproximadamente planares conectados por un enlace amida. Se cree que su anillo 4-metilfenil hidrofóbico interacciona con los residuos críticos del CCR5, mientras que el extremo del TAK779 amonio tetrahidro-2H-pirano, cargado positivo, se orienta a lo largo de la superficie extracelular, donde puede bloquear la unión de los ligandos quimiocinas y el complejo gp120-CD4. (B) Modelo estructural del CCR5 visto desde el plano de la membrana. El código de colores es el mismo que en la Figura I.39. Los residuos aminoácidos más cercanos a la superficie extracelular son Leu33, Trp86 y Glu283. Los residuos Tyr37, Thr82 y Tyr108 se encuentran más profundos en el receptor. El grupo de residuos aminoácidos mencionados para la unión del TAK779 incluye algunos residuos aromáticos (Tyr37, Trp86 y Tyr108) que pueden formar interacciones favorables con los anillos aromáticos del TAK779. (C) Vista del CCR5 desde la cara extracelular del coreceptor para mostrar la orientación de la unión del TAK779. Los colores son los mismos que en B, pero el modelo está rotado 90º. La escala es la misma en A, B y C [172].

Otros grupos han realizado estudios computacionales de *docking* TAK779 – CCR5 [173-177], mostrando interacciones del conjunto bencilo-pirano-amonio con las hélices TM1, TM2 y TM7, y contacto cercano del nitrógeno amónico del TAK779 con el residuo Glu283 del CCR5. Concretamente, la orientación del *docking* se caracteriza por una interacción electrostática entre el grupo amonio del TAK779 y el residuo Glu283. Este glutamato es accesible por el solvente y es el único residuo ácido del extremo extracelular de la hélice TM7. Los estudios de mutagénesis muestran la importancia de este glutamato, no solo para la unión del TAK779 al CCR5, sino también para antagonistas de otros receptores de quimiocinas [178-181]. Los residuos de las TM3, TM5 y TM6 interaccionan con el grupo aromático metilfenil-benzocicloheptenil.

Los residuos que han sido encontrados como importantes en la unión del TAK779 mediante *docking* son el Glu283 (en la TM7) y residuos aromáticos cercanos como Tyr37 (en la TM1), Trp86 (en la TM2) y Tyr108 (en la TM3). Las interacciones del TAK779 con las TM5 y TM6 incluyen a Thr195 y Ile198 en la TM5 y Tyr251, Asn252 y Leu255 en la TM6. Los datos de mutagénesis dirigida que dan lugar a un cambio del 20% o más en la eficacia del TAK779 para bloquear la entrada del VIH-1, también sugieren la interacción con Ile198, pero una contribución negligible de Thr195 y Leu255. No se tienen datos de mutagénesis dirigida para Tyr251 y Asn252, ya que la mutación de estos dos aminoácidos a Ala da lugar a una mala expresión del receptor [175].

En las figuras siguientes se muestra el modelo estructural de la transmembrana y regiones intracelulares del coreceptor CCR5 (Figura I.41) y la orientación del TAK779 en la región transmembrana del CCR5 (Figura I.42) según Paterlini [174].

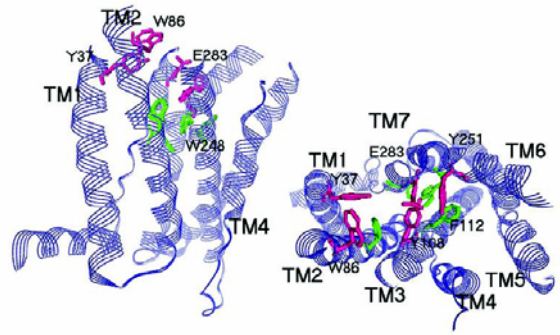


Figura I.41 Modelos estructurales de la transmembrana y regiones intracelulares de coreceptor CCR5. En rojo se muestran los residuos aromáticos y ácidos conservados en los receptores CCR1 y CCR5 (Tyr37, Trp86, Tyr108, Tyr251 y Glu283) y en verde se muestra un conjunto de residuos aromáticos conservados (Phe79, Trp248) y específicos de CCR5 (His289 y Phe112). Los residuos Tyr37, Trp86, Tyr108 y Glu283 forman parte de la cavidad de unión del TAK779 [174].

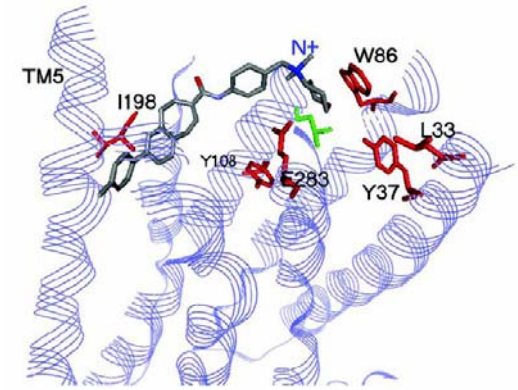


Figura I.42 Orientación del TAK779 en la región transmembrana del CCR5 después de 1 ns de simulación por dinámica molecular según estudios de Paterlini [174]. Las cadenas laterales a distancia de 5 Å o menos del TAK779 se muestran en representación de *stick*. Los residuos marcados en rojo están implicados en la unión del TAK779. Los residuos en verde son cadenas laterales cuya sustitución por Ala no interfiere en la actividad antiviral del TAK779 [174].

Los sitios de interacción identificados hasta la fecha, expuestos en este apartado, son los sitios de mayor interacción, pero pueden existir residuos adicionales, los cuales constituyen sitios de menor interacción. Estas interacciones débiles se pueden escapar de la detección experimental, debido al límite de sensibilidad del ensayo de entrada del VIH-1 o por limitaciones en la mutagénesis de los residuos por alanina. Además, es posible que existan interacciones por puente de hidrógeno entre la cadena principal peptídica y las moléculas inhibidoras. Estas interacciones no se detectarían mediante un estudio de mutagénesis convencional, el cual se centra en las cadenas laterales. Por ello, a pesar de que no se tengan indicaciones de alguna interacción entre los residuos del dominio extracelular y TAK779, AD101 y SCH-C, no se puede asegurar la posibilidad de que exista alguna.

Cabe mencionar que en lo que se refiere a la interacción de otros inhibidores de CCR5, diversos estudios corroboran que el coreceptor CCR5 presenta diferentes subsitios en la cavidad de unión. Es decir, que diferentes inhibidores se unen al coreceptor con diferentes modos de unión [135, 172, 173, 174, 177, 182-186]. Se puede plantear, pues, la hipótesis de que los inhibidores de CCR5 se pueden agrupar en dos o más grupos, cuyos miembros se unen a una región de la cavidad de unión del CCR5 similar. Esta hipótesis ha sido estudiada en esta tesis mediante una nueva aproximación de *shape matching* desarrollada (véase Artículo II). Esta hipótesis de la multi región de unión está también soportada por Castonguay *et al.* [184], los cuales han determinado que el sitio de unión para los inhibidores de 2-aryl-4-(piperidin-1-il)butanamina y 1,3,4-pirrolidina trisustituida está situado en una región similar a la propuesta para la unión de pequeñas moléculas a otras GPCRs, la cual solapa parcialmente la región propuesta para la unión del TAK779. Asimismo Kellenberger *et al.* [182] han citado también evidencia experimental sobre un sitio de unión deslocalizado para CCR5 (véase Figura I.43). El grupo de Maeda [185] ha discutido también las diferencias en los modos de unión al CCR5 de la espirodicetopiperazina Aplaviroc, el SCH-C y el TAK779. El grupo de Tsamis [135] ha recalcado que no todos los residuos que son importantes para la unión de AD101 y SCH-C, intervienen en la unión del TAK779. El grupo de Fujisawa [186] recalca lo mismo mostrando que el TAK779, el TAK220 y el AD101 se unen a una región de la cavidad de unión la cual se solapa pero es distinta. Lo mismo destaca el grupo de Seibert [173] en su estudio con TAK779, AD101 y SCH-C.

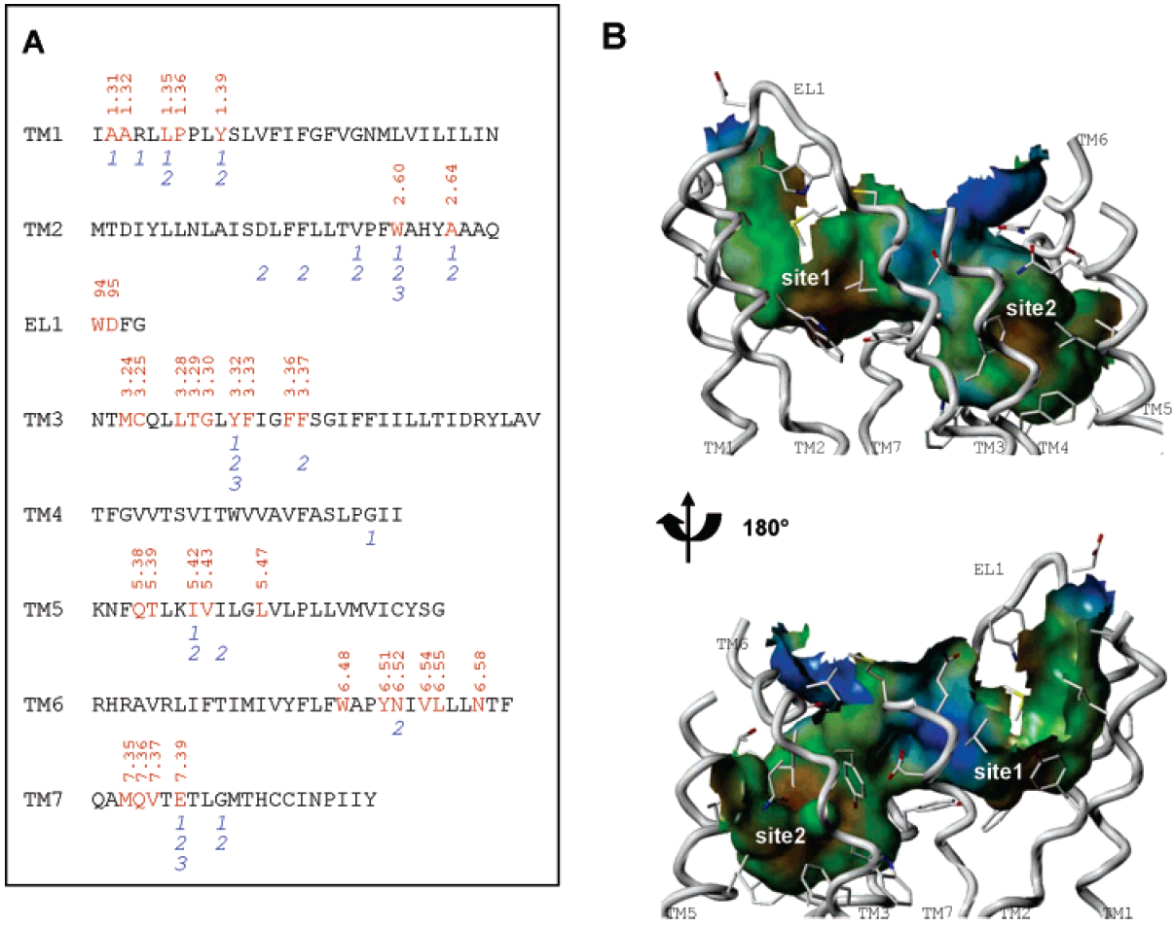


Figura I.43 Cavidad de unión del coreceptor CCR5. A) Secuencia de aminoácidos. Los residuos con las cadenas laterales hacia la cavidad se muestran en rojo y marcados según la numeración de Ballesteros [187], excepto en el dominio extracelular (EL1). Los números mostrados bajo la secuencia en azul resumen el mapeado experimental del sitio de interacción del receptor para los antagonistas TAK779, SCH-C, AD101, 1-amino-2-fenil-4-(piperidin-1-il)butano, y 1,3,4-pirrolidina trisustituida. *1*: residuos importantes para la eficiencia de TAK779. *2*: residuos importantes para la eficiencia de SCH-C y AD101. *3*: residuos importantes para la unión de 1-amino-2-fenil-4-(piperidin-1-il)butano y 1,3,4-pirrolidina trisustituida. B) La superficie de Connolly de la cavidad del coreceptor CCR5 (coloreada de acuerdo al potencial lipofílico) se muestra junto con el diagrama en forma de cintas de las siete hélices transmembrana del coreceptor. Las cadenas laterales de los residuos marcados en la secuencia están representadas utilizando líneas. La vista de la parte inferior está rotada 180° en el eje vertical respecto a la vista de la parte superior. [182].

Objetivos

Teniendo en cuenta todo lo expuesto en la introducción, los objetivos del presente trabajo son:

- Diseño de una base de datos de compuestos activos conocidos inhibidores de los coreceptores CXCR4 y CCR5, y supuestos inactivos tipo fármaco para realizar un cribado virtual retrospectivo. Asimismo, diseño de una base de datos de compuestos tipo fármaco para su cribado virtual prospectivo.

- Aplicación de filtros *structure-based* basados en *docking* a las quimiotecas compiladas. Para ello, primero se deben refinar los modelos de los coreceptores CXCR4 y CCR5, previamente desarrollados en el laboratotio de diseño molecular del Instituto Químico de Sarriá, así como realizar análisis preliminares del modo de unión de ligandos conocidos.

- Aplicación de filtros *ligand-based* a las quimiotecas compiladas: búsquedas de similitud, modelos farmacofóricos y *shape matching*.

- Implementación de una nueva técnica de *shape matching* y aplicación de ésta al estudio de la hipótesis de la multi-región de unión en la cavidad del coreceptor CCR5 para las diversas familias de inhibidores conocidas.

- Aplicación del nuevo *fingerprint* de interacción estructural APIF, desarrollado en el laboratotio de diseño molecular del Instituto Químico de Sarriá, como herramienta en el post-procesado de *docking* y filtro en un cribado virtual.

- Desarrollo de nuevos compuestos inhibidores de los coreceptores CXCR4 y CCR5 mediante técnicas *de novo design*.

1. Fundamentos Teóricos

1.1. Modelización Molecular

Existen dos grandes áreas en química computacional para el desarrollo de modelos moleculares: las técnicas de Mecánica Molecular (modelos clásicos) y las de Mecánica Cuántica (modelos cuánticos). Ambos métodos pretenden:

- Calcular la energía asociada a una estructura molecular determinada y así poder derivar propiedades asociadas a ésta.

- Encontrar la estructura molecular con menor energía (optimización de geometría o minimización de energía).

- Simular la variación a lo largo del tiempo de una determinada estructura molecular (dinámica molecular). Aunque tradicionalmente este estudio era exclusivo de la mecánica molecular por su coste y recursos computacionales necesarios, actualmente se empieza a utilizar en sistemas pequeños a través de modelos cuánticos.

Ahora bien, estos problemas se abordan de manera diferente. La Mecánica Molecular (MM) estudia el comportamiento dinámico de los átomos, mientras que la Mecánica Cuántica (QM) estudia la dinámica electrónica. Así pues, la Mecánica Molecular trata la geometría en el equilibrio o análisis conformacional entre otras cosas, y es aplicable a entidades moleculares grandes, de entre cien y cien mil átomos. La Mecánica Cuántica trata la rotura y formación de enlaces, espectroscopia UV-vis, reacciones químicas, estudio de orbitales, densidades de carga, órdenes de enlace, entre otras cosas, y es apropiada para entidades moleculares de menos de cien átomos. Se destacan los métodos *ab initio* y los métodos semiempíricos. En los primeros la distribución electrónica se incluye explícitamente mediante su codificación en la función de onda (Ψ), relacionada con la energía (E) a través de la ecuación de Schrödinger independiente del tiempo:

$$H\Psi(r) = E\Psi(r) \qquad \textbf{[1. 1]}$$

siendo H una función diferencial que incluye la energía cinética y potencial de núcleos y electrones, denominada operador hamiltoniano. La resolución de esta ecuación es compleja y requiere la introducción de diversas aproximaciones (aproximación de Born-Oppenheimer, combinación linear de orbitales atómicos…). Estos métodos no contemplan ningún tipo de parametrización empírica "externa". En los métodos semiempíricos sí que hay una parametrización empírica para la descripción de los electrones internos (*core*) mientras que los electrones externos se caracterizan mediante funciones de onda cuánticas.

También se han desarrollado métodos mixtos (QM/MM) que tratan el sistema parcialmente de forma cuántica y de forma clásica. La mecánica cuántica se utiliza en la zona reactiva para estudiar la ruptura y formación de enlaces implicados en la reacción, mientras que la mecánica molecular se utiliza en el entorno en el que se da la reacción.

1.1.1. Mecánica molecular

La mecánica molecular considera los átomos como bolas unidas por muelles (representando los enlaces entre átomos) según los criterios de la física clásica. Por lo tanto, en dicha simplificación, se consideran solo los núcleos de los átomos, no teniendo en cuenta explícitamente los electrones.

En el presente trabajo se aplican técnicas de Mecánica Molecular, dado que:

- El sistema de estudio es un sistema grande (macromoléculas).
- Interesa conocer la estructura de una proteína, obtener información de su geometría.
- Se desea realizar análisis conformacionales.
- Se pretende calcular la energía de unión entre proteína y una serie de compuestos, así como la minimización de energía tanto de dichos compuestos como de la proteína problema.

Todos ellos son objetivos alcanzables con la mecánica molecular.

Para definir el sistema en Mecánica Molecular, se utilizan los denominados campos de fuerza o *force fields* [188]. Se denomina *force field* al conjunto que forman las ecuaciones para calcular las contribuciones a la energía y los valores de los parámetros ajustables al equilibrio. Cada *force field* define una ecuación de energía potencial, de manera que la energía de una molécula en una conformación determinada se calcula a partir de la que tendrían idealmente las partes que la constituyen. La energía es relativa a un estado de referencia y se calcula como la suma de los diferentes términos que indican la penalización por el alejamiento de la idealidad de las distancias de enlace, ángulos, ángulos diedros, torsiones…

Para efectuar un cálculo en Mecánica Molecular se requieren tres elementos:

- Las coordenadas atómicas del sistema y las conectividades.

- El *force field* que se quiera utilizar, el cual dictará los *atom types* a usar, así como el método de cálculo de cargas parciales para la molécula. El *force field* está compuesto de:

  - *Atom types* para cada uno de los átomos del sistema según el *force field* empleado. Éstos permiten asignar a cada átomo sus características según su hibridación, carga y tipo de átomos a los que está unido.

  - Ecuación de energía potencial, la cual calcula la energía de la molécula en una configuración determinada. Dicha energía se calcula como la suma de diferentes términos que indican la penalización por el alejamiento de un estado ideal de las distancias de enlace, ángulos de enlace, torsiones…

  - Un conjunto de parámetros, los cuales permiten ajustar las ecuaciones a los diferentes tipos de átomos. Se suelen obtener a partir de datos experimentales o de resultados de cálculos de mecánica cuántica.

Existe una gran diversidad de *force fields* según se apliquen a diferentes aspectos de la química bioorgánica. Así, se pueden diferenciar *force fields* dirigidos a moléculas pequeñas y medianas (MM2 [189], MM3 [190], MM4 [191], TRIPOS [192], MMFF94 [193]) y a macromoléculas (AMBER [194], CHARMM [195], GROMOS [196], OPLS [197]).

De manera general, la función de potencial incluye los siguientes términos:

- Interacciones enlazantes: intervienen átomos unidos por enlaces químicos. Son considerados interacciones enlazantes los términos de enlace, ángulos, ángulos diedros o torsiones y ángulos fuera del plano.

- Interacciones no enlazantes: intervienen átomos no unidos de manera directa por enlaces. Constan de términos electrostáticos e interacciones de Van der Waals.

- Interacciones cruzadas: modelan el acoplamiento entre los términos anteriores. Se encuentran aquí los términos de ángulo-enlace, ángulo-ángulo …

Estos tres términos se muestran en la Figura 1.1.

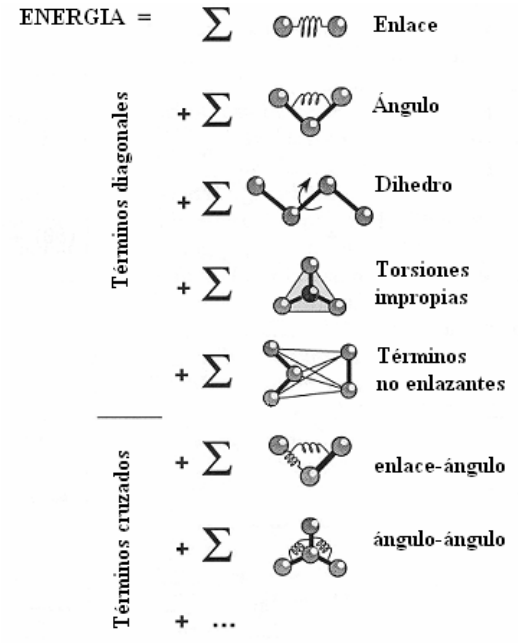


Figura 1.1 Términos incluidos en la ecuación de energía potencial de un *force field*. La energía se calcula como el sumatorio de diferentes términos. Términos enlazantes: enlace o extensión, ángulo, ángulos diedros o torsiones, ángulos fuera del plano (torsiones impropias). Términos no enlazantes. Términos cruzados: enlace-enlace, ángulo-ángulo, enlace-ángulo, enlace-diedro, ángulo fuera del plano-diedro, ángulo-diedro, ángulo-ángulo fuera del plano- diedro… [30]

Los *force field* destinados a macromoléculas, normalmente no tienen en cuenta términos cruzados y los términos enlazantes tienen ecuaciones sencillas, ya que el requerimiento computacional asciende.

En este proyecto se ha trabajado con el *force field* CHARMM, implementado en el análisis conformacional con Congen, el *force field* AMBER para la optimización de la geometría de las proteínas de estudio y dinámica molecular, y el *force field* MMFF94 para la optimización de los ligandos de las proteínas estudiadas.

- *CHARMM (Mackerell & Karplus, et al., 1995)* [195]

CHARMM (Chemistry at HARvard Macromolecular Mechanics *force field*) se refiere tanto a un programa de dinámica macromolecular y mecánica como a la función de energía potencial desarrollada para su uso en el programa. CHARMM es un *all-atom force field*, lo cual significa que considera todos los átomos, en contraposición a los denominados *united-atoms*, los cuales no incluyen explícitamente los hidrógenos no polares. En la versión más reciente (CHARMM22 y

CHARMM27), los parámetros fueron creados usando datos experimentales y suplementados con resultados *ab initio*. Puede trabajar con un amplio margen de entidades moleculares, desde pequeñas moléculas a complejos solvatados de grandes macromoléculas biológicas, aunque se utiliza principalmente para el estudio de macromoléculas.

La ecuación del force field CHARMM presenta la forma siguiente [195]:

$$U(\vec{R}) = \sum_{enlaces} k_b (b-b_0)^2 + \sum_{ángulos} k_\theta (\theta-\theta_0)^2 + \sum_{diedros} k_\chi \left[1+\cos(n\chi-\delta\right]+ \sum_{impropias} k_{imp} (\phi-\phi_0)^2 + \sum_{Urey-Bradley} k_u (u-u_0)^2 + \sum_{no\ enlazantes} \varepsilon\left[\left(\frac{R_{min\ ij}}{r}\right)^{12} - \left(\frac{R_{min\ ij}}{r}\right)^{6}\right] + \sum_{\substack{no\ enlazantes \\ electrostáticas}} \frac{q_i q_j}{\varepsilon_{ij} r_{ij}} \qquad \textbf{[1. 2]}$$

donde los dos primeros términos de la ecuación penalizan el alejamiento de los enlaces y ángulos de su valor de equilibrio, b_0 y θ_0 respectivamente, mediante un potencial harmónico con constantes de fuerza k_b y k_θ. El tercer término representa el potencial de torsión, donde k_χ es la constante de fuerza del diedro, n es la multiplicidad de la función, χ es el valor del ángulo diedro y δ es la fase. El cuarto término tiene en cuenta las torsiones impropias, fuera del plano, donde ϕ es el valor del ángulo impropio y ϕ_0 es el valor ideal del ángulo impropio. El quinto término lo constituye la componente de Urey-Bradley (término cruzado que tiene en cuenta la flexión del ángulo mediante interacciones 1-3 no enlazantes), en el que k_u es la respectiva constante de fuerza y u es la distancia entre los átomos 1,3 en el potencial harmónico. Las interacciones no-enlazantes entre dos pares de átomos (i,j) están representadas por los últimos dos términos. Por definición, las fuerzas no enlazantes se aplican a pares de átomos separados por lo menos tres enlaces. La energía de Van der Waals se calcula con un potencial estándar 12-6 de Lennard-Jones y la energía electrostática mediante el potencial de Coulomb. En el potencial de Lennard-Jones, ε es el parámetro de profundidad de Lennard-Jones, el término $R_{min\ ij}$ se refiere a la distancia en el mínimo de Lennard-Jones; por lo que respecta a la energía electrostática, q_i y q_j son la carga de los átomos, ε_{ij} es la constante dieléctrica efectiva y r_{ij} es la distancia entre átomos.
La energía del enlace de hidrógeno no se incluye como un término separado, sino que está implícita en la combinación de los términos de van der Waals y electrostático.

- *AMBER (Cornell, et al., 1995)* [194]

AMBER fue originalmente parametrizado para un número limitado de sistemas orgánicos, como pequeñas moléculas o polímeros, y ha sido utilizado ampliamente para proteínas y ácidos nucleicos. Como otros *force fields* desarrollados para este fin, utiliza *atom types* específicos [198]. Ofrece buenos resultados para geometrías de modelos en fase gas, energía libre de solvatación, frecuencias de vibración y energías conformacionales.

AMBER utiliza una representación *united atom*, en contraposición al *force field* CHARMM (que utiliza *all-atom representation*), lo cual implica que los hidrógenos no polares no se representan explícitamente, pero se tienen en cuenta en la descripción de los átomos pesados a los que están unidos. Ello resulta en una rapidez adicional significativa en cálculos basados en AMBER en comparación a otros *force fiel*ds.

Otra distinción del *force fie*ld AMBER se encuentra en que utiliza términos de torsión generales. Es un intermedio entre los *force fields* que utilizan diversos términos para las torsiones y los que sólo utilizan un término. Los *force fiels united atom*, como AMBER, utilizan términos de torsiones impropias para mantener la estereoquímica y los centros quirales.

La ecuación del *force field* AMBER presenta la forma siguiente [194]:

$$E_{pot} = \sum_{enlaces} k_r (r - r_0)^2 + \sum_{ángulos} k_\theta (\theta - \theta_0)^2 + \sum_{diedros} \frac{V_n}{2}[1 + \cos(n\phi - \phi_0)] + \sum_{\substack{no\ enlazantes \\ van\ der\ Waals}} \varepsilon\left[\left(\frac{r^*}{r}\right)^{12} - 2\left(\frac{r^*}{r}\right)^{6}\right] + \sum_{\substack{no\ enlazantes \\ electrostáticas}} \frac{q_i q_j}{\varepsilon_{ij} r_{ij}} \qquad \textbf{[1. 3]}$$

donde los dos primeros términos penalizan el alejamiento de los enlaces y ángulo de su valor de equilibrio, r_0 y θ_0 respectivamente, mediante un potencial harmónico simple, con constantes de fuerza k_r y k_θ. El potencial de torsión se representa mediante una serie de Fourier truncada donde V_n es el potencial máximo, n es la periodicidad y δ es la fase. Las interacciones de Van der Waals vienen dadas por el potencial de Lennard-Jones y las interacciones electrostáticas se calculan según la ley de Coulomb.

- *MMFF94 (Halgren, 1996)* [193]

MMFF94 (Halgren, 1996) fue desarrollado a través de métodos *ab initio* y verificado por datos experimentales. Se desarrolló para la aplicación en el estudio de las interacciones receptor-ligando, con proteínas y ácidos nucleicos como receptores y un gran número de biomoléculas y estructuras orgánicas como ligandos. En concreto, el *force field* debe poder describir cuantitativamente el ligando y la diana individualmente, así como su enlace [193]. Ello requiere que el *force field* tenga una buena predicción de las energías conformacionales y de las geometrías moleculares para poder así estimar la conformación requerida para la unión ligando-receptor.

Halgren, en el desarrollo del *force field* MMFF, utilizó una nueva manera de modelar con más precisión las interacciones de Van der Waals. Presenta un potencial fácil de utilizar en los cálculos de mecánica molecular y que mejora la reproducción de los datos experimentales. La energía de Van der Waals se calcula con un potencial 14-7 de Lennard-Jones, en lugar del potencial estándar 12-6 de Lennard-Jones.

La ecuación del *force field* MMFF94 está compuesta por términos enlazantes (enlaces, ángulos, extensión-flexión o *strech-bend*, flexión fuera del plano o *out-of-plane bending* y diedros) y por términos no enlazantes (van der Waals y electrostáticos). La expresión de energía se compone de siete términos, (Ecuación 1.4): el término de energía de enlace corresponde a una expansión de cuarto orden de la función de Morse, el término de ángulo de enlace se expresa mediante un desarrollo cúbico, o un desarrollo de Fourier truncado para enlaces lineales o casi lineales. Comparativamente a los dos *force fields* anteriores destaca el término de interacción enlace-ángulo, donde θ es el ángulo formado por r y r', seguido del término de torsiones impropias para centros tricoordenados, y finalmente las interacciones de torsión, van der Waals y electrostáticas, donde δ es la constante de *buffering* electrostático y n suele tomar el valor de 1 o 2.

$$E = \sum_{enlaces} 143.9325 \cdot \frac{k_r}{2} \cdot (r - r_o)^2 \cdot \left[1 - 2 \cdot (r - r_{\theta 0}) + \frac{7}{3} \cdot (r - r_0)^2 \right] +$$
$$\sum_{ángulos} \left| \begin{array}{c} \sum_{angulos} 0.043844 \cdot \frac{k_\theta}{2} \cdot (\theta - \theta_o)^2 \cdot \left[1 - 0.4 \cdot (\theta - \theta_0) \right] \\ 143.9325 \cdot k_\theta \cdot \left[1 + \cos(\theta - \theta_o) \right] \end{array} \right| +$$
$$\sum_{\substack{ángulos-enlaces \\ no\ lineales}} 2.5121 \cdot \left[k_r \cdot (r - r_0) + k_{r'} \cdot (r' - r_0') \right] \cdot (\theta - \theta_0) + \sum_{\substack{centros \\ tricoordenados}} 0.043844 \cdot \frac{k_i}{2} \cdot \omega_i^2 + \quad \textbf{[1. 4]}$$
$$\sum_{diedros} 0.5 \cdot \left[\left[V_1 \cdot (1 + \cos\phi) + V_2 \cdot (1 - \cos 2\phi) + V_3 (1 + \cos 3\phi) \right] \right] +$$
$$\sum_{\substack{no\ enlazantes \\ van\ der\ Waals}} \varepsilon_{ij} \cdot \left[\left(\frac{1.07 \cdot r^*}{r + 1 - 07 \cdot r^*} \right)^7 \cdot \left(\frac{1.12 \cdot r^{*7}}{r^7 + 0.12 \cdot r^{*7}} - 2 \right) \right] + \sum_{\substack{no\ enlazantes \\ electrostáticas}} \frac{332.071 \cdot q_i \cdot q_j}{\varepsilon_r \cdot (r_{ij} + \delta)^n}$$

1.1.2. Métodos de minimización

La optimización de la geometría de un sistema molecular consiste en localizar la estructura molecular con menor energía; por ello se utilizan de manera indistinta los términos de minimización de energía y optimización de geometría.

Los métodos habituales de optimización de geometría se basan en el cálculo de las derivadas de la energía con respecto a los grados de libertad geométricos. Los métodos derivativos se diferencian entre los de orden uno y orden dos. Los métodos derivativos de orden uno son los basados en el gradiente de energía, los cuales buscan desplazar el sistema en una dirección que conduzca a un valor menor de energía. Estos métodos presentan el inconveniente de tender a conducir el sistema hacia mínimos de energía próximos a la posición de partida (mínimos locales), los cuales no tienen por qué coincidir con el mínimo global correspondiente a la geometría óptima que se pretende hallar.

El *vector gradiente* $g = \nabla f = \{\delta f / \delta x_i\}$ o primera derivada de la función en un punto indica la dirección de la pendiente.

La *matriz hessiana* $H = \nabla^2 f = \nabla \otimes \nabla f = \{\delta^2 f / \delta x_i \, \delta x_j\}$ o segunda derivada de la función indica las direcciones en que la función es cóncava (mínimos, valores propios positivos) o convexa (máximos, valores propios negativos).

El gradiente es cero en los puntos estacionarios (máximos, mínimos, puntos de ensilladura o *saddle points*). Para distinguirlos, hace falta evaluar la hessiana; si ésta es convexa en todas direcciones se tratará de un máximo, si es cóncava en todas direcciones (todos los valores propios positivos) será un mínimo, si es cóncava en todas las direcciones menos una, se tratará de un *saddle point* de primer orden, si es cóncava en todas las direcciones menos dos, un *saddle point* de segundo orden, etc.

Algunos métodos habituales basados en el gradiente son *Steepest descent (Line Search)* o *Conjugate Gradient* (con variantes como *Fletcher-Reeves* o *Polack-Ribiere*). Dentro de los métodos derivativos de orden dos se encuentran *Newton-Raphson* y variantes. Estos métodos son

muy costosos computacionalmente ya que hace falta calcular la inversa de la hessiana. Para facilitar el cálculo de la hessiana, otros métodos de orden dos la construyen iterativamente como Quasi Newton o *variable metric methods*: *Davidon-Fletcher-Powell (DFP)* o *Broydon-Fletcher-Goldfarb-Shanno (BFGS)*. Dichos métodos se diferencian en la estimación inicial de la hessiana y en el proceso iterativo de construcción gradual de una mejor inversa de dicha hessiana.

A continuación se describen brevemente los tres métodos de minimización más comunes:

- *Steepest descent*: resulta útil como primera aproximación en sistemas que están alejados del mínimo local más cercano, ya que desciende rápidamente hacia una zona próxima al mínimo pero no suele converger. Sigue la estrategia de minimización en una línea (*line search*). Mediante cálculos del gradiente, se evalúan los cambios en la energía del sistema asociados a perturbaciones sobre los grados de libertad geométricos, y se modifica el sistema en la dirección indicada por el gradiente. El proceso se repite hasta que el cambio energético generado por la perturbación de cualquier grado de libertad geométrico es menor que un umbral predefinido. Cada búsqueda es perpendicular a la anterior. Repitiendo el proceso se encuentra el mínimo. Si el mínimo es muy estrecho, el número de evaluaciones de la función puede ser muy elevado.

- *Conjugate Gradient*: a diferencia del método anterior, no solo utiliza el gradiente del paso actual, sino que tras cada evaluación de gradiente se tiene en cuenta la información obtenida de evaluar el gradiente del paso anterior para el paso actual. Ello supone un mayor coste de cálculo, por lo que este método se emplea para explorar geometrías que estén cerca de un mínimo de energía. Según la manera de implementar este método existen diferentes variantes. *Fletcher-Reeves* propone que una vez encontrado el gradiente y el primer mínimo como en un *steepest descent*, el siguiente punto se encuentre en una dirección no exactamente perpendicular, sino en la dirección conjugada que garantiza encontrar el mínimo de la función si fuera cuadrática. *Polack-Ribiere* propone otra dirección, ya que según ellos normalmente las funciones de trabajo no son exactamente cuadráticas.

- *Newton-Raphson*: método más costoso computacionalmente, ya que necesita calcular la hessiana e invertirla para encontrar un nuevo mínimo. Se obtiene así información sobre la velocidad de cambio en la función gradiente y se puede modular adecuadamente la evolución del sistema hacia la geometría óptima.

Existen también métodos que exploran la superficie de energía sin información del gradiente como Simplex y Univariable Secuencial.

Todos los métodos de minimización tenderán hacia el mínimo local más cercano. El problema reside en hallar el mínimo global. Para comunicar al sistema cuando se ha iterado suficiente como para haber llegado al posible mínimo global se atiende a:

- Diferencias de energía en iteraciones sucesivas por debajo de un valor establecido.
- Diferencias en las coordenadas. Configuraciones sucesivas prácticamente iguales.
- *Rms* de los gradientes o fuerzas y valor absoluto de ninguno de ellos por debajo de unos valores establecidos.

Para asegurar que se ha hallado un mínimo (segunda derivada positiva), se puede realizar el cálculo de la hessiana al final de la optimización. Ello es costoso, tal y como se ha mencionado

anteriormente, pero puede ofrecer información adicional como el conocimiento de las vibraciones (análisis de los modos normales de vibración con los que se puede conocer el espectro IR o Raman) o la energía en el punto cero (ZPE), y así poder estimar propiedades termodinámicas.

1.2. Modelización de proteínas

La forma ideal de obtener información estructural para una proteína dada es determinar la estructura por cristalografía de rayos X o espectroscopia de resonancia magnética nuclear (RMN). Sin embargo, existen una serie de problemas:

- Algunas proteínas no pueden ser cristalizadas fácilmente. Se encuentran a menudo dificultades en la purificación de la cantidad suficiente de proteína, obtención de cristales para la difracción de rayos X y otros aspectos técnicos.

- La cristalografía puede tardar desde varios meses a varios años para resolver/analizar la estructura de una proteína simple.

- La resonancia magnética nuclear es, en promedio, más rápida que la cristalografía pero no puede aplicarse hoy en día a proteínas mayores de 100 residuos.

- Algunas proteínas que se obtienen son demasiado grandes para el análisis RMN y no pueden ser cristalizadas para aplicar la difracción de rayos X [199].

Dado que determinados residuos presentan una mayor frecuencia relativa de aparición en las estructuras secundarias y que la información necesaria para el correcto plegado de una proteína parece estar contenida en su estructura primaria, uno de los grandes retos de la bioinformática es la predicción de la estructura terciaria de las proteínas.

1.2.1. Métodos de modelización de proteínas

La predicción de la estructura a partir de la secuencia de aminoácidos resulta difícil, fundamentalmente debido a las interacciones de largo alcance que estabilizan las estructuras secundaria y terciaria. Para intentar resolver este problema se han diseñado tres aproximaciones: modelización por homología, *threading* o reconocimiento de plegamiento y métodos *ab initio* (no confundir con los métodos de modelización molecular cuánticos que reciben el mismo nombre).

En la modelización por homología, la construcción del modelo tridimensional de la proteína de estructura desconocida se basa en una o más proteínas relacionadas de estructura conocida (plantilla). Los métodos de reconocimiento de plegamiento o *threading* se basan en el hecho de que las proteínas generalmente adoptan plegamientos similares a pesar de que no haya una similitud significativa de secuencia o funcional. Los métodos *ab initio* [200] predicen la estructura 3D a partir únicamente de su secuencia, lo que equivale a conocer el mecanismo de plegamiento de las proteínas. Parten de la conformación extendida de un péptido, reducen los grados de libertad de la proteína, mediante los denominados "modelos de complejidad reducida" (*reduced complexity models*) y utilizan funciones energéticas, normalmente derivadas a partir de bases de datos (*knowledge-based*) para evaluar cada una de las geometrías obtenidas.

En el presente trabajo se han utilizado los métodos *ab initio* para el refinado de los *loops* de las proteínas de estudio, y la modelización por homología para la predicción de la estructura tridimensional de dichas proteínas. Esta última aproximación se detalla en el apartado siguiente.

1.2.2. Modelización por homología

La modelización por homología usa estructuras determinadas experimentalmente (plantillas) para predecir la conformación de otra proteína (diana) que tiene una secuencia de aminoácidos similar a la plantilla. Esta aproximación a la modelización de proteínas es posible ya que un pequeño cambio en la secuencia de la proteína usualmente resulta en un pequeño cambio en su estructura 3D. Se asume que la estructura está más conservada que la secuencia, de modo que si la proteína que se quiere modelar presenta más de un 30% de identidad con una proteína de estructura conocida, ambas proteínas pueden considerarse estructuralmente semejantes. Las condiciones necesarias son pues que la similitud entre la secuencia diana y la/las plantilla/s sea detectable y que se pueda construir un correcto alineamiento entre ellas.

La modelización por homología es el único método que actualmente puede proporcionar modelos con un error en la desviación cuadrática media (*rms*) respecto al experimental inferior a 2 Å [201].

La modelización por homología se basa en dos grandes observaciones:

- La estructura de una proteína está únicamente determinada por su secuencia de aminoácidos. Conocer la secuencia debería ser suficiente, al menos en teoría, para obtener la estructura.

- Durante la evolución, la estructura es más estable y cambia más lentamente que la secuencia asociada, con lo que secuencias similares adoptan prácticamente idénticas estructuras, y secuencias relacionadas distantes en el tiempo todavía se pliegan en estructuras similares. Esta relación fue identificada primeramente por Chothia and Lesk (1986) y después cuantificada por Sander and Schneider (1991) [199].

La modelización por homología usualmente consiste en los siguientes cinco pasos: búsqueda de plantillas, selección de una o más plantillas, alineamiento plantilla-diana, construcción del modelo y evaluación del modelo (véase Figura 1.2). Si el modelo no es satisfactorio, todos o alguno de estos pasos se pueden repetir. Cada uno de estos pasos se describe a continuación.

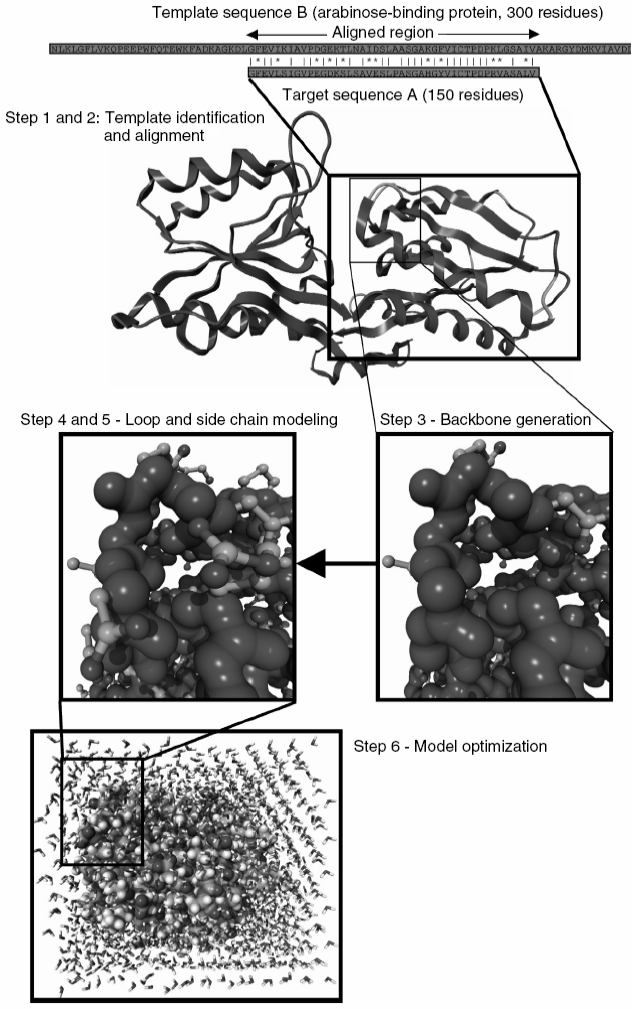


Figura 1.2 Pasos en la modelización por homología [202].

Búsqueda de estructuras y secuencias relacionadas con la secuencia diana

La modelización por homología empieza con la búsqueda de estructuras de proteínas conocidas en una base de datos, típicamente el *Protein Data bank*, PDB [203], usando la secuencia diana como referencia. Se compara la secuencia diana con cada una de las secuencias de las estructuras encontradas en la base de datos. Existen diferentes métodos de comparación secuencia-secuencia [204] así como muchos servidores de buscadores en bases de datos [202] como BLASTT, FASTA, 123D, PHDTHREADER, UCLA-DOE FRSVR o PROFIT entre otros.

Una vez se tiene la lista de plantillas potenciales, se seleccionan las apropiadas para la modelización a realizar. Éstas son usualmente las que tengan mayor similitud con la secuencia diana, es decir, mayor porcentaje de residuos idénticos y menor número de huecos en el alineamiento, así como menor longitud de éstos. Sin embargo, se han de tener en cuenta también otros factores como la construcción de árboles filogenéticos para seleccionar la plantilla más cercana a la subfamilia de la secuencia diana, la similitud del entorno de la plantilla y el entorno en el cual la diana necesita ser modelada (entendiendo como entorno todo lo que rodee a la proteína que no forme parte de ésta: solvente, pH, ligandos, interacciones cuaternarias, etc.) o la calidad de la estructura experimental (la resolución, y el factor R de una estructura cristalográfica, así como el

número de restricciones por residuo para una resonancia magnética nuclear de una estructura, son indicativos de la exactitud con que está resuelta dicha estructura).

El primer paso consiste pues en decidir qué partes de qué plantillas son lo bastante similares en cuanto a conformación de la proteína diana como para utilizarlas en el modelo inicial. Los modelos se pueden construir a partir de una única estructura de plantilla, pero se aprecia una mejoría significativa si se utilizan diversas estructuras plantilla. La dificultad está en la elección de qué plantilla usar para cada una de las partes de la proteína diana. Generalmente, la similitud de secuencia es bastante obvia para regiones de estructura secundaria, pero resulta poco fiable en las regiones de *loops*, dificultando así su construcción.

Alineamiento de la secuencia diana con la plantilla

Este paso en el que se establecen equivalencias estructurales y/o secuenciales entre la secuencia diana y las plantillas es el más importante. El alineamiento es relativamente simple de obtener cuando la identidad de la secuencia diana-plantilla está por encima del 40% con métodos de alineamiento secuencia-secuencia automáticos estándares. Altas identidades de secuencia, del orden del 70%, indican que los modelos obtenidos por homología serán de alta precisión, con desviaciones *rms* de la estructura nativa del orden de 1 o 2 Å. Sin embargo, si la identidad de secuencia diana-plantilla es menor que el 40%, el alineamiento generalmente tiene huecos y necesita intervención manual para minimizar el número de residuos no alineados. En estos casos de baja identidad, la precisión del alineamiento es el factor más importante que afecta a la calidad del modelo final, obteniéndose así modelos pobres con desviaciones *rms* mayores a 4 Å. Identidades de secuencias intermedias (entre 40% y 70%) tendrán errores dentro de estos valores.

Construcción del modelo

Una vez construido el alineamiento diana-plantilla, se pueden utilizar diferentes métodos para la construcción del modelo 3D de la proteína diana [202]:

- El método original y aún muy utilizado es el montaje de cuerpo rígido o *rigid-body assembly* [205]. Este método construye el modelo a partir de pequeñas regiones del *core*, de *loops* y de cadenas laterales, obtenidas de desagrupar y observar estructuras relacionadas.

- Otra familia de métodos se basa en las posiciones aproximadas de átomos conservados de las plantillas para calcular las coordenadas de otros átomos [206].

- El tercer grupo de métodos, el cual modela por satisfacción de restricciones espaciales, usa tanto distancias como técnicas de optimización para satisfacer restricciones espaciales obtenidas del alineamiento de la secuencia diana con las estructuras plantilla. Estos métodos son implementados por programas como Modeller [207].

Generación del esqueleto y construcción de las cadenas

Se han descrito diversos algoritmos los cuales construyen con elevada precisión las cadenas laterales del esqueleto de la proteína a partir de una cadena principal. Sin embargo, aplicando estos algoritmos a problemas reales, se encuentra que las cadenas laterales están construidas con una exactitud menor a la esperada. La explicación a esto parece ser que la conformación de la cadena

principal no está generalmente modelada con la bastante exactitud para poder predecir la conformación de las cadenas laterales de manera precisa [202].

Construcción de los loops

Los *loops* son regiones de la estructura de las proteínas, muy flexibles y sin periodicidad, que conectan elementos de estructura secundaria como hélices α u hojas β. Originariamente, los *loops* fueron considerados como "conformaciones irregulares" y a veces denominados como "*random coils* (espirales al azar)", pero hoy en día están reconocidos como una clase estructural adicional. Se pueden distinguir diferentes tipos de *loops* de acuerdo con la estructura secundaria que conectan: α-α, α-β, β-α, β-β hairpin, β-β link. Se encuentran principalmente en la superficie de proteínas globulares y representan una parte importante de la estructura proteica dando estabilidad en el proceso de plegamiento de las proteínas. Debido a su elevada flexibilidad son difíciles de describir mediante cristalografía de rayos X o espectroscopia RMN.

En general, los métodos de modelización por homología tienden a predecir correctamente el *core* o esqueleto básico de la proteína cuando se puede disponer de la estructura de una proteína homóloga cercana a la proteína diana, pero no las regiones de los *loops*. Los errores en los *loops* son el mayor problema en modelización por homología por encima del 35% en identidad de secuencia. En este rango de identidad de secuencia, los *loops* entre los homólogos varían mientras las regiones del *core* se conservan aún relativamente y se alinean correctamente. La modelización de los *loops* es por lo tanto aún una parte difícil en la predicción de la estructura de las proteínas, especialmente en lo que respecta a *loops* largos. De manera similar a la predicción de toda la estructura de la proteína, existen dos aproximaciones básicas en la predicción de estructuras de *loops*: métodos que utilizan bases de datos *database-search* (también llamados métodos *knowledge-based*) [208, 209] y métodos *ab initio* (también llamados métodos *conformational search*) [205, 210, 211].

Dado el importante papel de los *loops* en las funciones biológicas de las proteínas, se prestará especial atención a la modelización de los *loops* de los coreceptores proteicos de interés de este proyecto. Se utiliza el programa Congen (CONformation GENerador) [212] para modelar *loops ab initio*, método por el cual se realizan búsquedas conformacionales con el fin de explorar minuciosamente las posibles conformaciones de un *loop*. Utilizando una función energética de CHARMM se ranquean estas conformaciones para obtener la predicción final del *loop* (Artículo I).

Validación del modelo

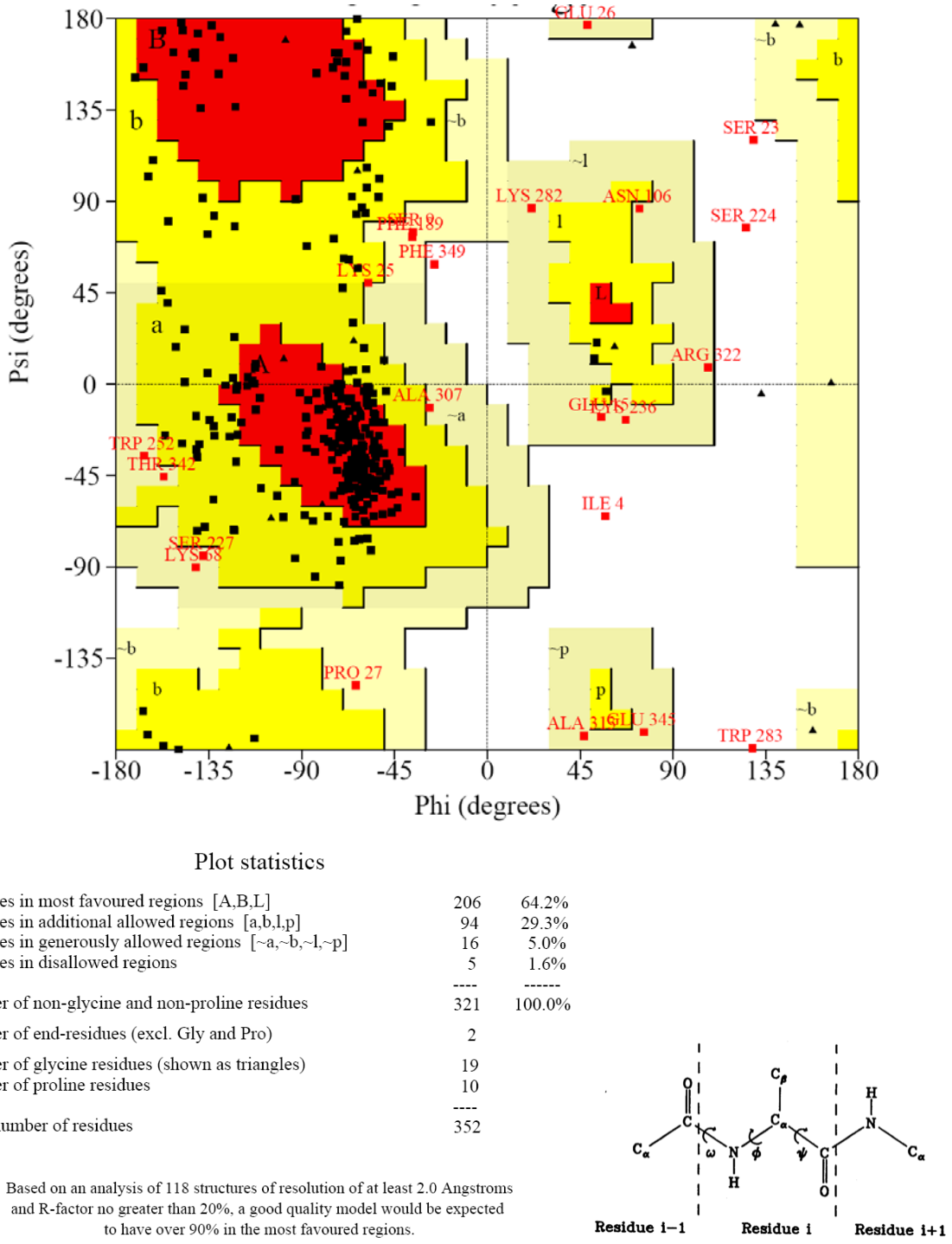
Después de la construcción del modelo, es importante comprobar la presencia de errores. Se pueden llevar a cabo dos tipos de validaciones [213]:

- Evaluación interna de la consistencia del modelo, la cual comprueba si el modelo satisface las restricciones usadas para calcularlo.
- Evaluación externa, la cual se refiere a información que no ha sido usada en el cálculo del modelo. Existen dos tipos de evaluación externa:
    - Testar si se ha utilizado una plantilla correcta. Se realiza este tipo de evaluación cuando el modelo está basado en menos de un 30% de identidad de secuencia respecto la plantilla. Una manera de predecir si una plantilla es mejor o peor es la comparación del Prosa II Z-*score* [214] de las estructuras del modelo y la plantilla. El

Z-*score* de un modelo es una medida de compatibilidad entre su secuencia y su estructura. El Z-*score* del modelo debería ser comparable con el Z-*score* de la plantilla. Ahora bien, cabe decir que esta evaluación no siempre funciona debido a que la función de potencial utilizada en Prosa II no es apropiada para ciertos tipos de plegamientos.

- Reconocer regiones poco fiables en el modelo. Una vía de aproximación es calculando un perfil de energías con programas como ProsaII [214]. Dicho perfil reporta la energía de cada posición en el modelo. Es posible detectar errores en el modelo porque aparecen como picos de energía positiva en el perfil de energías. Estas regiones del modelo deben ser inspeccionadas cuidadosamente. Otra vía de encontrar regiones poco fiables es evaluar la estereoquímica del modelo (longitud de enlaces, ángulos de enlace, ángulos diedros, solapamiento entre átomos) con programas como Procheck [215] y Whatcheck [216]. Finalmente, una herramienta de evaluación importante es el conocimiento experimental acerca de la estructura de la proteína y su función. Un modelo debería ser consistente con las observaciones experimentales como los datos obtenidos por mutagénesis dirigida, datos de *crosslinking*, unión de ligandos activos, etc.

En este proyecto se ha utilizado el programa de validación Procheck (Artículo I), el cual comprende a su vez una serie de subprogramas los cuales proporcionan una comprobación detallada de la estereoquímica de la estructura de una proteína. Los parámetros utilizados para evaluar dicha estructura son los encontrados por Morris, MacArthur, Hutchinson y Thornton (1992), obtenidos de estructuras de alta resolución, contra las cuales se compara la estructura problema en una base de datos de residuos (Cambridge Structural Database) [217]. La entrada principal del programa es un archivo en formato Brookhaven [218] que contiene las coordenadas de la estructura. La salida del programa comprende un número de gráficos en formato PostScript y una lista de residuos. Ello da una idea de la calidad total de la estructura comparada con estructuras bien refinadas de la misma resolución y resalta regiones que pueden necesitar mayor investigación. Procheck ofrece diez gráficos diferentes (mapa de Ramachandran, mapa de Ramachandran para Gly y Pro, ángulos de torsión $\chi 1$ versus $\chi 2$ para cada residuo donde sea aplicable, propiedades de la cadena principal, propiedades de las cadenas laterales, propiedades de los residuos, distribuciones de la longitud de enlace de los enlaces de la cadena principal, distribuciones de los ángulos de enlace de los enlaces de la cadena principal, distancias *rms* respecto a la planaridad de los diferentes grupos planares, geometrías distorsionadas de las longitudes de enlace, ángulos de enlace de la cadena principal y grupos planares). Entre ellos destaca el mapa de Ramachandran (Figura 1.3) por la información global que aporta. Dicho mapa permite visualizar ángulos diedros φ contra ψ de residuos aminoácidos en una estructura proteica. Muestra las posibles conformaciones de los ángulos φ y ψ para un polipéptido. Asimismo, también se obtienen mapas de Ramachandran para visualizar los ángulos diedros φ contra ψ para los residuos Gly y Pro especialmente, ya que estos residuos poseen regiones favorables/desfavorables diferentes del resto de residuos.



Plot statistics

Residues in most favoured regions [A,B,L]	206	64.2%
Residues in additional allowed regions [a,b,l,p]	94	29.3%
Residues in generously allowed regions [~a,~b,~l,~p]	16	5.0%
Residues in disallowed regions	5	1.6%
	----	------
Number of non-glycine and non-proline residues	321	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	19	
Number of proline residues	10	

Total number of residues	352	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Figura 1.3 Ejemplo de mapa de Ramachandran obtenido con Prochek para e1 coreceptor CXCR4 modelado por homología. Los residuos que se encuentran en las zonas rojas (A, B, L) son los que se hallan en las regiones más favorables. Los residuos que se encuentran en las zonas naranjas (a, b, l, p) son los que se hallan en las regiones permitidas. Los residuos que se encuentran en las zonas beix (∽a, ∽b, ∽l, ∽p) son los que se hallan en regiones menos permitidas y los residuos que se encuentran en las zonas blancas son los que se hallan en regiones no permitidas.

1.3. Dinámica Molecular

La dinámica molecular es una disciplina particular del modelado molecular que explora el movimiento de las moléculas. Utiliza soluciones numéricas de la ecuación de movimiento de Newton, sobre un modelo que representa un sistema molecular, para simular el movimiento atómico y así obtener información a cerca de las propiedades dependientes del tiempo de dicho sistema. Son precisamente estos movimientos y fluctuaciones de las moléculas los que permiten la interacción fármaco-receptor, y por tanto dicho estudio dinámico de las moléculas es importante en el diseño de fármacos por ordenador.

Para definir, pues, la superficie de energía potencial dejando que el sistema evolucione a lo largo del tiempo, se empieza con una conformación de partida químicamente plausible y se calculan las cargas parciales. Se calculan las fuerzas que actúan mediante la integración de las ecuaciones de Newton, se deja que el sistema evolucione a lo largo del tiempo, con lo que cambian las posiciones de los átomos, se calcula la energía de la nueva conformación y se vuelven a calcular las fuerzas, volviendo a integrar las ecuaciones de Newton, repitiendo así el ciclo a lo largo del tiempo. Contra más tiempo pase más cambios se podrán calcular, lo cual implica mayor número de conformaciones exploradas y mejor descripción del sistema.

Los estudios de dinámica molecular ofrecen información de diversas propiedades de un sistema:

- Estabilidad del sistema. Desviación con respecto a la geometría inicial.
- Exploración conformacional. Se exploran las geometrías posibles del sistema de una manera más exhaustiva que mediante los métodos de optimización de geometría basados en el gradiente. El problema reside en que los cambios conformacionales implicados en fenómenos biológicos, como la activación de un receptor al unirse un ligando, ocurren a escalas de tiempo mucho mayores que las que se pueden simular. Por ello, se han desarrollado campos de fuerza simplificados [219, 220], para poder simular períodos de tiempo mayores.
- Cálculo promedio de propiedades moleculares mediante un muestreo sistemático de valores de la propiedad que se quiera medir.

Las propiedades que ha de cumplir una dinámica molecular para poderla llevar a cabo son las siguientes:

- El movimiento ha de ser debido a fuerzas internas, a la energía potencial de las otras partículas (el campo externo ha de ser cero).
- Determinista. En cada paso de la dinámica siempre se obtendrá lo mismo en las mismas condiciones.
- Se puede definir la trayectoria del sistema ya que se conoce la posición y velocidad de los átomos gracias a la integración de las ecuaciones de Newton por diferencias finitas Δt.
- Conociendo el potencial, se conocen las fuerzas, y con éstas las aceleraciones, que junto con las velocidades y las posiciones a tiempo t, se calculan para tiempo $t + \Delta t$.
- Las fuerzas se consideran constantes durante el Δt y se recalculan a $t + \Delta t$.
- Las velocidades iniciales para que empiece a moverse el sistema son aleatorias, dependientes de la temperatura. El sistema se irá acercando hacia puntos más estables.
- Las posiciones iniciales se obtienen a partir de difracción de rayos X.

La velocidad y aceleración de los átomos viene dictada por las fuerzas que ejercen éstos entre ellos, según describe la segunda ley de Newton. Así, la fuerza que actúa sobre una partícula *i* puede expresarse en función de su masa *m* y aceleración *a*:

$$F(t_i) = ma_i = m\frac{\delta^2 r_i}{\delta t_i^{\,2}} \qquad \textbf{[1. 5]}$$

La fuerza en un átomo se obtiene como la derivada de la energía respecto al cambio en la posición del átomo.

$$F(t_i) = -\frac{\delta \mathrm{V(r_i)}}{\delta \mathrm{r_i}} \qquad \textbf{[1. 6]}$$

Si se igualan las Ecuaciones 1.5 y 1.6 se tiene la ecuación del movimiento:

$$-\frac{\delta \mathrm{V(r_i)}}{\delta \mathrm{r_i}} = m\frac{\delta^2 r_i}{\delta t_i^{\,2}} \qquad \textbf{1. 7]}$$

La integración de la ecuación del movimiento resuelve la trayectoria:

$$r_i(t_2) = r_i(t_1) + \int_{t_1}^{t_2} \frac{p(t)}{m} dt$$
$$p(t_2) = p(t_1) + m\int_{t_1}^{t_2} a(t)dt \qquad \textbf{[1. 8]}$$

En pocos casos se dispone de ecuaciones analíticas para la posición y el momento en función del tiempo, con lo que las Ecuaciones 1.7 se aproximan numéricamente discretizando el tiempo en fracciones de tiempo, *time step* (δt), el cual interesa que sea pequeño (del orden de femtosegundos), ya que cuanto más pequeño, más se parece a la trayectoria de simulación, y menor es el error de cálculo que se propaga a lo largo de ésta, aunque supone aumentar el coste computacional. Para reducir este coste, se suele trabajar con un *time step* de 1fs, una décima parte del tiempo del movimiento más rápido del sistema de estudio: la vibración de un enlace de un átomo de hidrógeno, la cual es del orden de 10 fs, con lo que se evita que los átomos colapsen en el espacio. Esto hace que las dinámicas de un sistema biológico, constituido por miles de átomos, no superen el orden de unos pocos nanosegundos, ya que dado el coste computacional, las simulaciones de mucho más de pico o nanosegundos no son razonables. Ello impide simulaciones que afecten a grandes cambios conformacionales.

Existen diferentes algoritmos para integrar numéricamente las ecuaciones del movimiento. El método de integración más común es el algoritmo de Verlet [221]. Si se discretiza el tiempo en fracciones de tiempo, *time step* (δt), y se considera que los átomos tienen velocidades iniciales (v_0), se pueden calcular las nuevas posiciones atómicas (r_i) y velocidades (v_i) para el momento ($t + \delta t$) después de haber aplicado una fuerza. Para las nuevas posiciones a $t + \delta t$:

$$\left.\begin{aligned} r(t+\delta t) &= r(t) + \delta t v(t) + \frac{1}{2}\delta t^2 a(t) + \ldots \\ r(t-\delta t) &= r(t) - \delta t v(t) + \frac{1}{2}\delta t^2 a(t) - \ldots \end{aligned}\right\} \; r(t+\delta t) = 2r(t) - r(t-\delta t) + a(t)\delta t^2 \qquad \textbf{[1. 9]}$$

Las velocidades no aparecen explícitamente y hay que calcularlas de otra manera:

$$v(t) = \frac{\left[r\left(t+\delta t\right) - r(t-\delta t)\right]}{2\delta t} \qquad \textbf{[1. 10]}$$

O bien estimándolas a media etapa ($t + 1/2\delta t$):

$$v(t+\frac{1}{2}\delta t) = \frac{\left[r\left(t+\delta t\right) - r(t)\right]}{\delta t} \qquad \textbf{[1. 11]}$$

Las velocidades iniciales se asignan aleatoriamente mediante la distribución de probabilidad de Maxwell-Boltzman.

Existen modificaciones de este método como Verlet Leap-Frog [222] (implementado en el módulo Sander del programa Amber [223] utilizado en las dinámicas moleculares realizadas en este trabajo), el cual calcula explícitamente posiciones y velocidades, pero desfasados de tal manera que no se pueden definir posiciones al mismo tiempo exacto que la energía cinética, Verlet Velocity, que calcula explícitamente posiciones y velocidades sin estar desfasadas, conociendo la aceleración (fuerzas) a tiempo t y $t + \delta t$, o Beeman y métodos predictores-correctores, los cuales utilizan algoritmos todavía más precisos.

Las dinámicas moleculares se pueden llevar a cabo, bien en condiciones de número de átomos, volumen y energía constantes, lo que se denomina *NVE ensemble* o microcanónico, que es la opción clásica, bien en condiciones de número de átomos, volumen y temperatura constante, esto es *NVT ensemble* o canónico, o bien con número de átomos, temperatura y presión constante, denominado *NPT ensemble* o isotérmico-isobárico. Para conseguir estas dos últimas es necesario poder ajustar, bien la temperatura para el *ensemble* NVT o bien la presión para el *ensemble* NPT. En el presente trabajo se ha utilizado el protocolo publicado por Orozco *et al.*[224] en que se realiza un calentamiento inicial del sistema hasta adquirir la temperatura deseada y después se utilizan condiciones de *ensemble* canónico, por lo que se describirán brevemente los métodos de termostatación.

En una dinámica es importante el control de la temperatura, para asegurarse de cumplir el principio de conservación de la energía (Ecuación 1.12). La manera más obvia de dicho control sería reescalando las velocidades multiplicándolas por una constante λ, pero este simple reescalado puede hacer que las trayectorias dejen de ser newtonianas conduciendo a propiedades menos precisas.

$$\sum_i \frac{1}{2} m_i \left(v_i \lambda\right)^2 = \sum_i \frac{1}{2} m_i \left(v_i'\right)^2$$
$$\lambda = \sqrt{\frac{T_{\text{deseada}}}{T_{actual}}} \qquad \textbf{[1. 12]}$$

Una alternativa es acoplar el sistema a un baño externo a la temperatura deseada, el llamado baño de Berendsen [225]. El baño aporta o extrae calor del sistema escalando las velocidades proporcionalmente a la diferencia de temperaturas:

$$\lambda = \sqrt{1 + \frac{\Delta t}{\tau}\left(\frac{T_{\text{b año}}}{T_{actual}} - 1\right)} \qquad \textbf{[1. 13]}$$

donde τ es un parámetro de acoplamiento que determina cuánto de acoplados están el sistema y el baño. De esta manera, cuando τ es grande, el acoplamiento es débil y para τ pequeñas, el acoplamiento es fuerte; τ suele tener un valor de $400\delta t$. Los acoplamientos débiles tienen la ventaja de dejar fluctuar la temperatura del sistema alrededor del valor deseado y son los que se emplean el programa Sander bajo condiciones de control de temperatura.

Los dos métodos expuestos aunque útiles y muy empleados no generan condiciones rigurosamente canónicas, ya que el escalado de velocidades prolonga artificialmente diferencias de temperatura entre componentes del sistema. Por ello se han desarrollado otros métodos alternativos, como el método *stochastic collisions* [226] y el método *extended system* que implementados adecuadamente, sí generan *ensembles* estrictamente canónicos [227, 228].

Otro aspecto importante a tener en cuenta en la simulación de un sistema químico es el tratamiento del solvente. En algunos casos las moléculas de disolvente están directamente implicadas en el proceso a simular, como en el caso de una reacción de hidrólisis. En otros casos el solvente no interactúa directamente pero proporciona un entorno que afecta fuertemente el comportamiento del soluto. En ocasiones, sin embargo, el solvente actúa meramente de medio, afectando al soluto sólo en cuanto a propiedades dieléctricas. De este modo se han desarrollado diferentes alternativas para simular sistemas en solución:

- Métodos de solvente explícito: incluyen explícitamente una cantidad determinada de moléculas de solvente. Concretamente en el programa Amber se pueden añadir un *cap* (una porción esférica de moléculas de solvente centradas en un punto del soluto), una *box*, (una caja de solvente de las dimensiones especificadas), un octaedro truncado (una caja con los vértices cortados, consiguiendo una distribución de solvente más uniforme alrededor del soluto), o una *shell* (añade una capa de solvente irregular siguiendo la superficie del soluto).
- Métodos aproximados: grupo de técnicas que incorporan el efecto del solvente sin que las moléculas de solvente se encuentren de forma explícita en el sistema.

Los métodos de coste computacional más bajo se basan en modificar la permitividad. Entre ellos, la alternativa más simple consiste en multiplicar la permitividad del espacio libre o vacío ε_0 por la constante dieléctrica del disolvente ε_r. Dicha alternativa puede aplicarse en solventes homogéneos y soluciones muy diluidas, sin embargo en sistemas biológicos la constante dieléctrica depende de la distancia entre grupos cargados. La implementación más sencilla consiste en una dependencia lineal de la constante dieléctrica con la distancia. Asimismo se han desarrollado otros modelos más complejos para la dependencia de la constante dieléctrica, muchos de ellos con una forma aproximadamente sigmoidea. Otras técnicas más complejas para incorporar el efecto del solvente específicamente en dinámica molecular son principalmente *Potentials of Mean Force* y *Stochastic Dynamics* [229], así como otras técnicas denominadas '*continuum' solvent models*, entre los que destacan los basados en las ecuaciones de Poisson-Boltzmann [230, 231] y el modelo generalizado de Born [232].

1.4. Evaluación de la interacción Proteína-Ligando

En función de la información experimental de la que se disponga, la predicción de la unión proteína-ligando puede ser más o menos costosa. Si se dispone de la proteína (receptor) cristalizada, y además ésta tiene un elevado interés farmacológico, acostumbran a existir complejos ligando-receptor cristalizados, los cuales permiten estudiar la posición de distintos ligandos en el sitio activo de la proteína. Con esto, se puede conocer el sitio de unión de posibles nuevos fármacos. Por otra parte, si se disponen de datos biológicos que informen sobre las interacciones entre fármaco-proteína (como experimentos de mutagénesis dirigida), se puede deducir también el modo de unión del ligando al receptor. Con estas informaciones, se puede realizar un acoplamiento manual entre ligando y receptor mediante programas de modelización molecular para obtener así modelos que expliquen en lo posible los datos experimentales.

Sin embargo, si no se disponen de datos experimentales, existen métodos automáticos para explorar las posibles uniones entre ligando y receptor. Son los denominados programas de *docking* [233], los cuales realizan una exploración de todas las posibles posiciones relativas ligando-receptor, evaluando la interacción intermolecular entre ambos. Como resultado de esta exploración, se obtienen una serie de posibles conformaciones para la unión ligando-receptor. Cada una de estas soluciones se evalúa mediante una función de *scoring*.

Los sitios de unión son áreas de la proteína, las cuales se conoce que son activas al formarse el complejo ligando-receptor. Los programas de *docking* identifican posibles conformaciones de unión entre ligando y receptor, cada una de las cuales está unida a la cavidad de unión según un modo de unión. En la forma más general del *docking* no se disponen de datos bioquímicos adicionales, pero disponer de información complementaria facilita el problema del *docking* considerablemente. Sin embargo, hay que tener en cuenta que existen potenciales sitios de unión adicionales en la superficie de las proteínas, por lo que se asume que el sitio conocido es el que participa en la conformación de unión ligando-receptor.

Existen tres ingredientes claves en el *docking*: representación del sistema, búsqueda conformacional y *ranking* de las soluciones potenciales. Dichos aspectos están mutuamente interrelacionados: la elección de la representación del sistema (superficie) decide los tipos de algoritmos para la búsqueda conformacional y la manera de ranquear las potenciales soluciones.

El *docking* estimula esencialmente la interacción con la superficie del receptor. Dicha superficie se puede describir mediante modelos matemáticos, como descriptores, o una *grid.* Además, dicho receptor puede tratarse de manera estática o dinámica (proteína flexible o rígida). La mayoría de las aproximaciones consideran la proteína casi totalmente rígida o con flexibilidad parcial en las cadenas laterales y permiten flexibilidad al ligando.

Básicamente son esenciales tres pasos para una buena predicción del complejo receptor/ligando: definición de la estructura de la molécula diana, localización del sitio de unión y determinación del modo de unión. Idealmente la estructura de la molécula diana debería estar determinada experimentalmente, aunque algunas aplicaciones de *docking* utilizan dianas modeladas si ésta no se puede obtener de manera experimental. El segundo paso, la localización del sitio de unión, se puede llevar a cabo en la mayoría de veces de manera satisfactoria. El tercer paso, es la "típica" aplicación de los algoritmos de *docking*: dado el sitio de unión de una molécula, determinar el modo de unión a un ligando [26].

Un algoritmo de búsqueda riguroso podría muestrear todos los modos de unión entre dos moléculas. Sin embargo, esto es impráctico debido al gran tamaño del espacio conformacional. A modo de ejemplo se considera un sistema simple formado por un ligando con cuatro enlaces rotables, seis parámetros de alineamiento de cuerpo rígido y un centro activo cúbico que mide 10^3 $Å^3$. Se tienen pues, tres grados de libertad translacionales y tres rotacionales. Si los ángulos se consideran en incrementos de 10 grados y los parámetros translacionales en una *grid* de 0.5 Å, existen aproximadamente 4 x 108 parámetros de cuerpo rígido a muestrear, correspondientes a 6 x 1014 configuraciones a ser muestreadas. Este trabajo requeriría aproximadamente 2000000 de años de tiempo computacional, a una velocidad de 10 configuraciones por segundo. Consecuentemente, solo una pequeña parte del espacio conformacional puede ser muestreado, por lo que se debe poner en balanza el coste computacional y la cantidad de espacio conformacional examinado [234]. Además, el problema computacional se incrementa si se considera la flexibilidad de la proteína (receptor) y la demanda creciente de cribar grandes bases de datos (de potenciales fármacos o de estructuras de proteínas).

Los métodos de *docking* iniciales se basaban en el principio de la llave y cerradura [235], lo que llevó a utilizar criterios geométricos para evaluar el grado de complementariedad estérica entre ligando y sitio de unión [236]. Sin embargo, pronto la complementariedad química se empezó a tener en cuenta en las aproximaciones de *docking* para reducir el número de soluciones irreales obtenidas mediante los criterios de forma únicamente. Además, se introdujeron las funciones de *scoring*, basadas en *force fields* de mecánica molecular para estimar la buena interacción proteína-ligando. Los métodos de *docking* actuales utilizan funciones de *scoring* de dos formas diferentes. Una primera forma: se utiliza toda la función de *scoring* para ranquear una conformación proteína-ligando. El sistema es entonces modificado por el algoritmo de búsqueda, y la misma función de *scoring* es aplicada otra vez para ranquear la nueva estructura. Una segunda forma: utiliza una función de *scoring* en dos etapas. Una función reducida se utiliza para dirigir la búsqueda y una más rigurosa se utiliza para ranquear las estructuras resultantes. Algunos métodos de *docking* comunes son [234]:

- Dinámica molecular, la cual se basa en el cálculo de las ecuaciones del movimiento de Newton para hallar conformaciones de mínima energía.
- Métodos de Monte Carlo, mediante los cuales se generan movimientos aleatorios del sistema y se acepta o se rechaza el movimiento según un criterio de probabilidad de Boltzman.
- Algoritmos genéticos, los cuales parten de una población inicial aleatoria, a la que se le aplican operadores genéticos (mutaciones, *crossovers* y migraciones) para obtener una nueva población, a la que se le aplicarán funciones de cálculo, los valores de las cuales se utilizarán para ranquear las soluciones obtenidas.
- Métodos basados en el *docking* de fragmentos del ligando y posterior unión de ellos.
- Métodos que evalúan la complementariedad química entre las moléculas que interactúan.
- Métodos basados en distancias y geometrías.
- Búsquedas sistemáticas, las cuales intentan muestrear sistemáticamente todas las posibles conformaciones utilizando restricciones que permitan reducir la dimensionalidad del problema. Se asume que todas las moléculas son rígidas y la energía de interacción se evalúa gracias a un f*orce field.*

Un importante uso de los programas de *docking* proteína-ligando se centra en las aplicaciones siguientes [237]: predicción de la unión proteína-ligando, *docking* como herramienta para cribado virtual de compuestos, y funciones de *scoring* como herramientas para predecir afinidad.

En lo que respecta a la predicción de la interacción proteína-ligando, se distinguen distintas técnicas de cálculo en función de la complejidad y los modelos teóricos o empíricos utilizados, aunque principalmente se pueden clasificar en dos grandes grupos: los basados en mecánica estadística, los cuales requieren simulaciones tipo Monte Carlo y dinámica molecular, requiriendo muchos recursos computacionales, y las *scoring functions*. Estas últimas son aproximaciones muy crudas de la energía de unión, de modo que los valores de interacción corresponden a una suma de términos y se obtienen a partir de una sola estructura ligando-receptor, sin tener en cuenta promedios de *ensemble*. A su favor tienen un coste computacional muy bajo, siendo por ello adecuadas para incorporar en técnicas de *docking* así como para realizar cribados virtuales de bibliotecas de compuestos. A continuación se expone brevemente el fundamento de los métodos de predicción de unión proteína-ligando basados en mecánica estadística (Sección 1.4.1) y en la Sección 1.4.2 se detallan las funciones de *scoring* como herramientas para el cribado virtual y para predecir afinidad.

1.4.1. Métodos basados en mecánica estadística

Para evaluar la interacción proteína-ligando de una forma precisa, se debe analizar qué es lo que sucede cuando un ligando se une a su receptor (proceso esquematizado en la Figura 1.4).

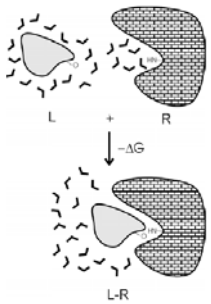


Figura 1.4 Proceso de unión entre ligando (L) y receptor (R) para formar el complejo ligando-receptor (L-R), liberando energía (-ΔG) [238].

Inicialmente, ligando y receptor están separados, solvatados por moléculas de agua. Tras la unión (*binding*), las dos moléculas se encuentran estabilizadas por interacciones intermoleculares no enlazantes, se produce el reordenamiento de las moléculas de agua, y la libertad conformacional del ligando cambia. Termodinámicamente, existe un cambio de la energía libre entre las dos situaciones:

$$\Delta G_{binding} = \Delta H - T\Delta S \qquad \textbf{[1. 14]}$$

El término entálpico *(ΔH*) está formado por las interacciones moleculares no enlazantes entre ligando y receptor, mientras que el término entrópico (*ΔS*), se refiere a la pérdida de libertad conformacional del ligando (reducción de entropía) y al reordenamiento de las moléculas de agua que pasan de estar ordenadas alrededor de cada molécula a rodear al complejo con menor superficie accesible al solvente (aumento de entropía), véase Figura 1.5.

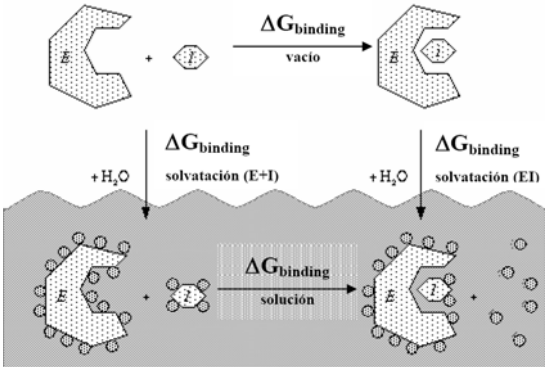


Figura 1.5 Ciclo termodinámico para la unión de un enzima, *E*, y un inhibidor, *I*, en medio solvatado y en el vacío. Las moléculas solvatadas se indican con círculos a su alrededor. Éstas tienden a estar ordenadas cuando se sitúan alrededor de *E* y de *I* por separado, ahora bien, cuando se produce la unión de *E-I*, algunas moléculas de solvente se liberan y se produce desorden. Este es un efecto entrópico y es la base del efecto hidrofóbico. El reordenamiento de solvente alrededor de *E* e *I*, cuando ambas moléculas no están unidas y si lo están, está fuertemente influenciado por los puentes de hidrógeno entre estas moléculas. Estos puentes de hidrógeno entre solvente y *E*, y solvente y *I*, contribuyen entálpicamente a la estabilización, lo cual se puede estimar en las funciones de los programas de *docking*. De acuerdo con la ley de Hess, el cambio de la energía libre de dos estados es el mismo, sea cual sea el camino, por lo que se puede calcular la energía libre de unión en solvente según la ecuación: $\Delta G_{\text{binding solución}} = \Delta G_{\text{binding vacío}} + \Delta G_{\text{solvatación (EI)}} - \Delta G_{\text{solvatación (E+I)}}$ [239].

La energía libre de Gibbs de unión ($\Delta G_{binding}$) se puede relacionar con la constante de disociación del complejo proteína-ligando (K_d) de la forma siguiente:

$$K_d = \frac{[R]\,[L]}{[RL]} = \exp\left(\frac{\Delta G_{binding}}{RT}\right) \qquad \textbf{[1. 15]}$$

Los métodos de *docking* incorporan funciones de *scoring* que permiten calcular la variación de energía de Gibbs en la unión proteína-ligando ($\Delta G_{binding}$), mediante aproximaciones de mecánica molecular o potenciales estadísticos. Ahora bien, también se puede calcular la variación de energía de Gibbs entre dos procesos como la unión de dos ligandos diferentes a una misma proteína según el siguiente ciclo termodinámico:

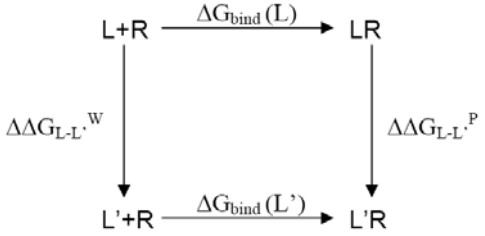


Figura 1.6 Ciclo termodinámico

De dicho ciclo termodinámico se deduce que la diferencia de energía de Gibbs entre los dos acoplamientos se puede calcular, teóricamente, como la diferencia entre cualquiera de los dos pares de ramas paralelas del ciclo:

$$\Delta\Delta G_{binding}^{(L\to L')} = \Delta G_{binding}(L) - \Delta G_{binding}(L') = \Delta\Delta G^{P}_{L\to L'} - \Delta\Delta G^{W}_{L\to L'} \qquad \textbf{[1. 16]}$$

Las ramas verticales del ciclo (transformación de un ligando *L* en otro *L'* en medios acuoso y en el entorno de la proteína) se pueden calcular por el método conocido como perturbación de energía libre (*free energy perturbation*, FEP) [240]. Existen otros métodos de cálculo de energía de *binding* como MM-PBSA[241] o el método LIE (*linear interaction energy*) [242] basados en muestrear dinámicas moleculares.

1.4.2. Funciones de *Scoring*

Las funciones de *scoring* son un componente imprescindible de los algoritmos de *docking* [243]. Tal y como se ha comentado anteriormente, se utilizan mayoritariamente para:

- Ranquear una colección virtual de compuestos de los que se quiere conocer su actividad contra una diana de estudio, distinguiendo así compuestos activos de inactivos.

- Predecir la afinidad de unión entre ligando y receptor de manera aproximada, ya que las funciones usadas para describir la química y física de la unión de un ligando a su receptor están incompletas todavía [244]. Las funciones de *scoring* pueden, en la mayoría de casos, predecir regularmente el modo de unión de ligandos unidos con afinidad nanomolar; sin embargo predicen de manera pobre el modo de unión para ligandos de menor afinidad. Ello destaca especialmente en sistemas fuera de los *sets* de prueba en el caso de algunas funciones de *scoring* empíricas.

Por lo que respeta al primer punto, ninguna función de *scoring* obtiene resultados de manera satisfactoria, lo cual conduce a un compromiso pragmático, la llamada aproximación por *consensus scoring* [245]. En dicha aproximación, se combinan diversas funciones de *scoring* y solo aquellas conformaciones en posiciones mejores del *ranking* por dos o más funciones de *scoring*, se consideran favorables. Está demostrado que este método da lugar a una gran reducción de falsos positivos cuando se aplica a la elección de ligandos con las menores energías de unión de entre un conjunto de ligandos o en la selección de las mejores conformaciones de entre diferentes configuraciones doqueadas de un ligando particular [246]. Existen diferentes estrategias para realizar un *consensus scoring* [245, 247]. *Rank-by-number* en que se realiza la media de los *scores* tras ser éstos normalizados. *Rank-by-rank* en que los compuestos son ranqueados primero para cada función de *scoring* individual, se realiza un promedio de los rangos para cada compuesto de la base de datos según las diferentes funciones de *scoring*, y finalmente los compuestos son reranqueados según su *ranking* promedio. Y *rank-by-vote* en que los compuestos deben presentarse en el x% de las listas ranquedadas para cada una de las funciones de *scoring* individuales usadas en la combinación. En esta tesis se utilizan los dos últimos métodos de consenso (Artículo I, Artículo III y Apéndice III).

Por lo que respeta a la predicción de afinidad de unión entre ligando y receptor, las funciones de *scoring* se utilizan más que para predecir los valores absolutos de la energía libre de unión, para predecir de manera correcta el orden relativo en cuanto a actividad de las moléculas doqueadas. Para cuantificar este aspecto y establecer comparaciones entre las afinidades de compuestos medidas experimentalmente y los *scores* calculados para cada ligando, se utilizan diferentes coeficientes con el fin de valorar las predicciones de afinidad para una diana dada, como son el índice predictivo de Pearlman (PI) [248, 249], el coeficiente de Spearman (Rs) [250] o el coeficiente de correlación de Pearson (r) [246].

Existen diversos requerimientos que una función de *scoring* debe satisfacer. Primero, las conformaciones deben ser ranqueadas correctamente, por ejemplo, aquellas que se parezcan más a

las estructuras experimentales han de tener mejor resultado, es decir, mejor posición en el *ranking*. Segundo, si múltiples ligandos son doqueados, sus energías libres de unión han de ser ranqueadas con precisión. En un cribado virtual de compuestos, los ligandos que se unan de manera débil al receptor han de distinguirse de los que no se unan. Tercero, una función de *scoring* debe ser lo suficientemente rápida/eficiente como para poder ser aplicada en un algoritmo de *docking*. Ello hace casi imposible usar métodos de *docking* que requieran la generación exhaustiva de conformaciones para obtener la energía de unión, aunque se han descrito cálculos de afinidades de unión basados en algoritmos que identifican las conformaciones más estables [246].

Las funciones de *scoring* se pueden agrupar en tres clases: funciones basadas en campos de fuerza o *force-field based*, funciones basadas en el conocimiento o *knowledge-based* y funciones empíricas de evaluación o *empirical scoring functions*.

- Las funciones basadas en campos de fuerza o *force-field based functions* aplican las funciones energéticas de la mecánica molecular clásica. Aproximan la energía libre de unión de los complejos proteína-ligando mediante una suma de interacciones de van der Waals y electrostáticas. La solvatación se tiene en cuenta usando una función dieléctrica dependiente de la distancia, aunque se han desarrollado también modelos de solvente basados en modelos electrostáticos continuos. Las contribuciones no polares se asume que son proporcionales a la superficie accesible por el solvente. Una desventaja de estas funciones de *scoring* es que los campos de energía asociados a los potenciales *force-field* son normalmente toscos, y por ello, se requiere de una minimización antes de cualquier evaluación energética.

- Las funciones de *scoring* empíricas o *empirical scoring functions* estiman la energía libre de unión sumando términos de interacción derivados de parámetros estructurales obtenidos de manera empírica. Se obtienen ajustando la función de *scoring* a constantes experimentales de un conjunto de prueba de complejos proteína-ligando. La función de *scoring* típica pionera de Böhm [251] consiste en cinco contribuciones, las cuales representan enlaces de hidrógeno, interacciones iónicas y lipofílicas, y la pérdida de entropía externa y configuracional en la unión. La principal desventaja de estas funciones es que no es del todo claro si son capaces de predecir la afinidad de unión de ligandos estructuralmente diferentes de los utilizados en el *set* de prueba.

En esta tesis se utilizan funciones de *scoring* de esta clase: Docked Energy del programa Autodock [252], GolScore y ChemScore del programa Gold [253], ChemScore, OechemScore, Shapegauss, Chemgauss3, Screenscore, Plp y Consensus Score del programa Fred [254], y Docked Energy del programa Hex [255]. El Artículo I muestra su uso y las características de cada una de ellas. A continuación se describen tres de las funciones de *scoring* más utilizadas en esta tesis: Autodock Docked Energy, Gold GolScore y ChemScore.

La función de energía libre que utiliza Autodock consta de cinco términos. Los primeros tres términos son los términos típicos de mecánica molecular para la dispersión-repulsión, puentes de hidrógeno y energía electrostática. ΔG_{tor} hace referencia a la restricción de las torsiones internas, la rotación y translación global del ligando; ΔG_{sol} hace referencia a la desolvatación en la unión y el efecto hidrofóbico (la entropía del solvente cambia en las superficies del soluto-solvente). La función presenta la forma siguiente:

$$\begin{aligned}\Delta G = {} & \Delta G_{vdW} \sum_{i,j} \left(\frac{A_{i,j}}{r_{ij}^{12}} - \frac{B_{i,j}}{r_{ij}^{6}} \right) \\ & + \Delta G_{hbond} \sum_{i,j} E(t) \left(\frac{C_{i,j}}{r_{ij}^{12}} - \frac{D_{i,j}}{r_{ij}^{10}} + E_{hbond} \right) \\ & + \Delta G_{elec} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} \\ & + \Delta G_{tor} N_{tor} \\ & + \Delta G_{sol} \sum_{i_c,j} S_i \cdot V_j \cdot e^{\left(\frac{-r_{ij}^2}{2 \cdot \sigma^2} \right)}\end{aligned} \qquad \textbf{[1. 17]}$$

donde los términos ΔG_{vdW}, ΔG_{hbond}, ΔG_{elec}, ΔG_{tor}, ΔG_{sol} son coeficientes empíricos determinados utilizando un análisis de regresión lineal a partir de un *set* de complejos proteína-ligando con conocidas constantes de unión. E_{hbond} es la energía promedio estimada debida a las interacciones por puente de hidrógeno del agua con un átomo polar. Los sumatorios de todos los términos de la ecuación, excepto el término de solvatación, se realizan para todos los pares de átomos del ligando, *i*, y átomos de la proteína, *j*, además de todos los pares de átomos del ligando que estén separados por tres o más enlaces. El sumatorio en el término de solvatación se realiza sobre todos los pares de átomos de carbono del ligando, i_C, y átomos de todos los tipos de la proteína, *j*. Las contribuciones en el vacío incluyen tres términos de interacción: un término de repulsión-dispersión de Lennard-Jones 12-6, un término direccional 12-10 de puente de hidrógeno, donde *E(t)* es una medida direccional basada en el ángulo, *t*, entre el átomo de prueba y el átomo diana; y un potencial electrostático de Coulomb. A la función en el vacío también se le añade una medida de la entropía no favorable de la unión del ligando debido a la restricción de grados de libertad conformacionales. Este término es proporcional al número de enlaces sp^3 en el ligando, N_{tor}. El término de solvatación corresponde a una variante del método de Souten *et al.* [256], un modelo basado en ocupaciones atómicas que contabiliza la relación entre la energía de solvatación y el volumen alrededor de un átomo que queda libre para ser ocupado por moléculas de solvente, donde V_j es el volumen fragmental de un átomo vecino *j*, r_{ij} es la distancia entre los átomos *i j*, y el término exponencial es la función que delimita el entorno de *i*, tomándose para σ el valor de 3.5 Å, que corresponde aproximadamente a la distancia de mínimo potencial de van der Waals entre dos átomos pesados.

La función de energía libre de Autodock está calibrada utilizado treinta complejos proteína-ligando con constantes de unión experimentales publicadas, escogidos del *set* de 45 moléculas utilizado por Böhm [251], omitiendo todos los complejos modelados por él (utilizando solo complejos para los que las estructuras cristalográficas son disponibles). Una limitación de estos datos de constantes de unión es que varían las condiciones bajo las cuales están determinados, lo cual limita la calidad del resultado obtenido con Autodock. Las constantes de inhibición K_i se convierten en cambio observado de la energía de unión (ΔG_{obs}) usando la siguiente ecuación:

$$\Delta G_{obs} = R \cdot T \cdot \ln K_i \qquad \textbf{[1. 18]}$$

donde *R* es la constante de los gases, 1,987 cal K^{-1} mol^{-1} y *T* es la temperatura absoluta, la cual se asume a temperatura ambiente, 298,15 K. Esta ecuación carece de signo negativo ya que la

constante de inhibición se define para la reacción de disociación, EI → E + I, mientras que ΔG_{obs} se refiere al proceso opuesto de unión, E + I → EI, donde E es el enzima e I es el inhibidor.

Cada uno de los treinta ligandos del *set* de calibración ha sido utilizado para calcular los coeficientes de energía libre empíricos para cada uno de los términos de la función de energía libre (término de contribuciones por puentes de hidrógeno, término de interacciones de van der Waals, término electrostático, término debido a las torsiones del ligando y término debido a la energía de solvatación) usando regresión lineal y estudios de *cross-validation*.

De esta manera, para efectuar el *docking* con Autodock, se realiza una interpolación trilineal sobre la función de energía libre expuesta, lo cual permite evaluar rápidamente la energía de repulsión, la energía debida a los puentes de hidrógeno, el potencial electrostático y la energía de solvatación de cada átomo del ligando, utilizando los mapas de la *grid* precalculados sobre la proteína para cada *atom type* del ligando.

Autodock devuelve dos evaluaciones: la energía de *docking* o Docked energy [252] (la cual incluye las interacciones intermoleculares ligando-proteína e intramoleculares del ligando, las cuales son utilizadas durante el *docking*) y la energía de unión o Binding energy (la cual incluye la energía intermolecular ligando-proteína y la energía libre debida a las torsiones del ligando). La energía interna o intramolecular del ligando no está incluida en el cálculo de la energía libre; sin embargo, durante el *docking*, la energía interna sí que se incluye en la energía total de *docking*, ya que cambios en la conformación del ligando pueden afectar al resultado final del *docking*.

La función de *scoring* Goldscore [253] (función de *scoring* original de Gold) es semejante a una función de mecánica molecular. Dicha función, optimizada durante el proceso de *docking*, posee cinco términos y dos opcionales:

$$GoldScore = S_{hb\_ext} + S_{vdw\_ext} + S_{hb\_int} + S_{vdw\_int} + S_{tor\_int}\left(+ S_{con} + S_{cov}\right) \qquad \textbf{[1. 19]}$$

donde S_{hb_ext} (*external H-bond*) es la puntuación (*score*) del puente de hidrógeno proteína-ligando, S_{vdw_ext} (*external vdw*) es la puntuación (*score*) de la interacción de van der Waals proteína-ligando, S_{hb_int} (*internal H-bond*) es la contribución a la función GoldScore debida a los puentes de hidrógeno intramoleculares del ligando, S_{vdw_int} (*internal vdw*) es la contribución debida a la tensión intramolecular del ligando y S_{tor_int} es la contribución debida a la tensión provocada por las torsiones del ligando. Los términos S_{con} y S_{cov} son opcionales. Si se especifica alguna restricción, se añade un término de *scoring* adicional, S_{con}, que contribuye en la modificación de la función de *scoring* final. Del mismo modo, si se realiza el *docking* de la proteína con ligandos a los cuales se les impone una unión covalente específica, se añade también a la función de *scoring* un término de covalencia S_{cov}.

El valor de *external vdw* se multiplica por un factor de 1,375 una vez se ha calculado el valor de la función de *scoring* total. Esto es una corrección empírica para tener en cuenta el peso del contacto hidrofóbico proteína-ligando en la función de *scoring*.

El valor de la función de *scoring* total se toma como menos el resultado de la suma de los términos de energía, con lo que mayores valores de la función de *scoring* son mejores.

La función GoldScore ha sido optimizada para la predicción de la posición del ligando en la unión a la proteína receptora, no para la predicción de afinidad ligando-proteína, aunque se puede encontrar alguna correlación en este último caso.

La función de *scoring* ChemScore [257] descrita por Eldridge *et al.*[258] y Baxter [259] estima la energía libre de unión de un ligando a una proteína. Está derivada empíricamente a partir de un *set* de 82 complejos proteína-ligando de los cuales se conocen los datos de afinidad de unión. Al contrario que GoldScore, la función ChemScore está construida por regresión de datos de afinidad medidos experimentalmente, aunque ello no indica con claridad que ChemScore sea superior a GoldScore en la predicción de afinidades. ChemScore estima la variación de energía libre total que ocurre al unirse el ligando a la proteína receptora de la siguiente manera:

$$\Delta G_{binding} = \Delta G_0 + \Delta G_{hbond} + \Delta G_{metal} + \Delta G_{lipo} + \Delta G_{rot} \qquad \textbf{[1. 20]}$$

Cada término de esta ecuación es el producto de un término dependiente de la magnitud de una contribución física particular a la energía libre (P), por ejemplo el puente de hidrógeno, y un factor de escalado determinado por regresión lineal a partir de un conjunto de complejos proteína-ligando de prueba (ν):

$$\begin{aligned}
\Delta G_0 &= \nu_0 \\
\Delta G_{hbond} &= \nu_1 P_{hbond} \\
\Delta G_{metal} &= \nu_2 P_{metal} \\
\Delta G_{lipo} &= \nu_3 P_{lipo} \\
\Delta G_{rot} &= \nu_4 P_{rot}
\end{aligned} \qquad \textbf{[1. 21]}$$

donde P_{hbond} , P_{metal} , P_{lipo} son *scores* para puente de hidrógeno, aceptor-metal, e interacciones lipofílicas, respectivamente. P_{rot} es un *score* representante de la pérdida de entropía conformacional del ligando al unirse a la proteína.

El valor final de la función ChemScore se obtiene añadiendo términos de enfrentamiento y torsiones internas, lo cual previene contra contactos estéricos muy cercanos en el *docking* y conformaciones internas pobres. También se pueden incluir términos covalentes y restricciones como en GoldScore. La ecuación final que presenta ChemScore queda pues de la forma siguiente:

$$ChemScore = \Delta G_{binding} + P_{clash} + c_{internal} P_{internal} + \left(c_{covalent} P_{covalent} + P_{constraint}\right) \qquad \textbf{[1. 22]}$$

donde $\Delta G_{binding}$ es la variación de energía libre expuesta en la ecuación 1.20, P_{clash} es el término referente al impedimento estérico en la unión proteína-ligando, $c_{internal}P_{internal}$ es el término que hace referencia a las torsiones internas del ligando, $c_{covalent}P_{covalent}$ y $P_{constraint}$ hacen referencia al término de enlace covalente especificado entre ligando y proteína y a restricciones impuestas, respectivamente.

- Las funciones basadas en el conocimiento o *knowledge-based functions* representan la afinidad de unión como la suma de potenciales de interacción entre los pares de átomos de ligando-proteína. Estos potenciales son derivados de los complejos proteína-ligando con estructuras conocidas, donde las distribuciones de probabilidad de distancias interatómicas son convertidas a interacciones de energía libre de unión de pares de átomos dependientes de la distancia utilizando la ley de

Boltzman "inversa". Se han propuesto diversas aproximaciones para derivar estos potenciales, las cuales difieren en la definición del estado de referencia (el cual determina los pesos entre las diversas distribuciones de probabilidad), los *atom types* de proteína y ligando, y la lista de complejos proteína-ligando a partir de la cual fueron extraídos.

1.5. Descriptores Moleculares

Un descriptor molecular es una propiedad determinada experimentalmente u obtenida mediante cálculos teóricos, o un valor que distingue y codifica cada molécula de forma particular como resultado de una transformación lógica o matemática [21]. Puede utilizarse para describir la estructura, propiedades o reactividad de una molécula, o en modelos de predicción de propiedades de otros compuestos (generación de modelos estructura-actividad, diseño de bibliotecas, manejo y gestión de información químicas y utilidades de búsqueda en bases de datos).

Tal como se ha mencionado, los descriptores pueden ser teóricos o experimentales. Excepto descriptores muy simples como los que cuentan tipos de átomos o fragmentos, o los descriptores derivados de teorías matemáticas exactas como los invariantes de grafo, tanto los experimentales como los teóricos contienen error. En el primer caso, contienen error estadístico debido a ruido experimental. En el segundo caso, tienen asociado un error inherente al método de cálculo; sin embargo la dirección, aunque no la magnitud de dicho error suele ser conocida y el error suele ser constante para una serie de compuestos relacionados. Además los descriptores teóricos poseen la ventaja de tener bajo coste, rapidez de obtención, y disponibilidad.

Los descriptores moleculares se pueden clasificar en función de: el tipo de representación química requerida (0D, 1D, 2D, 3D, 4D); el tipo de codificación matemática (valores reales, vectores reales o binarios, tensores o campos escalares discretizados según el enrejado de una malla); invariabilidad de sus propiedades (su capacidad para rendir un valor independiente de características particulares de la representación del compuesto); degeneración o capacidad de evitar asignar valores idénticos a compuestos distintos; el tipo de propiedad que describen (estéricas, electrónicas, lipofílicas, de forma, descriptores farmacofóricos…).

En el presente trabajo se utilizan diversos descriptores, principalmente para comparar quimiotecas de compuestos utilizadas en cribado virtual (Artículos I, III y IV), en búsquedas de similitud 2D/3D (Artículo V y Apéndice III), y en agrupaciones de compuestos (*clusters*) según sus propiedades (Artículo II). Para el primer caso, se calculan descriptores 1D con Moe [260]. En el segundo caso, se utilizan los 46 descriptores de propiedades físico-químicas y propiedades ADME que calcula QikProp [261] de Schrödinger. En el último caso se utilizan los descriptores topológicos, farmacofóricos basados en fragmentos 2D y *fingerprints* calculados por GenerateMD [262] de ChemAxon. Se describe a continuación brevemente el tipo de descriptores utilizados.

- Descriptores basados en índices topológicos: se basan únicamente en la estructura 2D o topología de la molécula, derivados matemáticamente del grafo estructural de la molécula. Se distinguen índices topoestructurales (que codifican sólo la información de adyacencia y distancia), índices topoquímicos (que además incluyen propiedades químicas de los átomos implicados) y los basados en teoría de la información. En general, estos índices contienen información relacionada con la forma molecular, el grado de ramificación, tamaño molecular y la flexibilidad estructural.

- Descriptores de propiedades fisico-químicas: se clasifican tradicionalmente en varias categorías, según describan propiedades hidrofóbicas, estéricas y efectos electrónicos. Además, se incluyen también propiedades mecanocuánticas (energías del HOMO y el LUMO, entalpía de formación, potencial de ionización, energía electrónica, energía de solvatación…) y propiedades estructurales (peso molecular, número de enlaces rotables, número de centros quirales…).

- Descriptores *count-based*: cuentan *building blocks* básicos de moléculas como átomos, enlaces o anillos. Son muy rápidos de calcular, pero no son muy apropiados para discriminar correctamente entre moléculas, por lo que su uso no es muy común, excepto aquellos relacionados con propiedades físico-químicas o farmacofóricas (número de enlaces rotables o de aceptores/dadores de puente de hidrógeno).

- Descriptores de forma, definen la globularidad o desviación de la molécula de la forma esférica: primera, segunda y tercera dimensión estándar (raíz cuadrada del primer, segundo y tercer valor propio mayor de la matriz de covarianza de coordenadas atómicas); radio de giro o distribución de masa atómica en la molécula; momento de inercia a lo largo de un eje principal; caracterización de la forma y el tamaño de las moléculas (índices de Jurs [263]); superficie molecular y volumen molecular.

- Descriptores farmacofóricos 2D y 3D: se basan en el concepto de pares atómicos o *atom pairs*, definidos a partir de un par de tipos atómicos y la distancia entre ellos, determinada a partir del mínimo recorrido en el grafo. El tipo atómico se puede definir a partir del elemento atómico, el número de enlaces con átomos pesados o el número de enlaces π, entre otros. Además del concepto de pares atómicos, otros fragmentos 2D típicos de subestructuras son el átomo aumentado, la secuencia atómica, la secuencia de anillo y la torsión topológica. Los descriptores farmacofóricos 3D siguen la misma idea, pero diferenciándose entre aquellos basados en distancias o en ángulos

- *Fingerprints*: consisten en una cadena de dígitos binarios que contiene información codificada, generalmente de tipo estructural. Los descriptores farmacofóricos 2D y 3D se codifican de esta forma. Los *fingerprints* se dividen en *bytes*, y las distintas características estructurales tienen asignado un intervalo de *bits* que adoptan un valor de 0 o 1, denotando la presencia (1) o ausencia (0) de determinadas características.

1.6. Obtención de modelos farmacofóricos

En esta tesis se han derivado modelos farmacofóricos utilizando los módulos *Pharmacophore Query*, *Pharmacophore Consensus* y *Pharmacophore Elucidate* de Moe [260] y los protocolos *Common Feature Pharmacophore Generation* y *Steric Refinement with Excluded Volumes* de Discovery Studio [264] (Artículo III).

Moe permite generar modelos farmacofóricos:

- Manualmente a partir de una conformación de un ligando o ajustando interactivamente las posiciones, radios, así como otras características de la *query*. Para ello se utiliza la herramienta *Pharmacophore Query Editor*, la cual permite crear manualmente una *query* consistente en un conjunto de restricciones referentes a la ubicación y tipo de características farmacofóricas de un ligando.

- Manualmente a partir de un alineamiento inicial de conformaciones de ligandos introducido previamente, el cual permanece rígido. Para ello se utiliza la herramienta *Pharmacophore Consensus*, la cual sugiere una *query* cuyas características farmacofóricas son consistentes con un conjunto especificado de conformaciones de ligandos alineadas, por ejemplo con Moe-FlexAlign [265] o a partir de los resultados de un *docking* en la proteína diana.

- Incorporando de forma automatizada la flexibilidad conformacional en la formulación de la hipótesis. Para ello se utiliza la herramienta *Pharmacophore Elucidate*, la cual genera *queries* farmacofóricas a partir de una colección de compuestos, algunos (o todos) de los cuales son activos frente a una diana biológica específica, de tal forma que todos (o la mayoría) de los compuestos activos satisfagan las *queries*.

Cada *query* puede poseer tres tipos de restricciones: *query features*, *group constraints* y *volumes* (*excluded*, *included* o *exterior*), véase Figura 1.7. Las *query features* son puntos en el espacio con un radio más una tolerancia en cuanto a proximidad en el espacio (por ejemplo existencia de un dador en la zona marcada por el radio más la tolerancia). Un *group constraint* es un grupo de *query features* el cual vincula la presencia de un *feature* (o característica) a la presencia de otro *feature*. Un volumen es una esfera o conjunto de esferas que llevan asociadas una expresión particular. Puede distinguirse entre *excluded volume*, en que el interior del volumen no debe contener ningún átomo que cumpla la expresión, *included volume* en que el interior del volumen debe contener como mínimo un átomo que cumpla la expresión y *exterior volume*, en que el exterior del volumen no debe contener ningún átomo que cumpla la expresión.

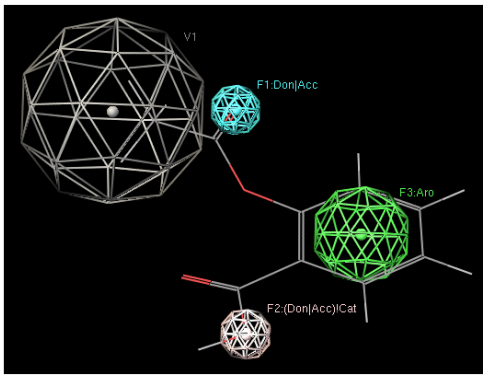


Figura 1.7 Ejemplo de restricciones de una *query* creada con Moe. Volumen de inclusión (gris), un *feature* aromático (verde), un *feature* que puede ser aceptor de puente de hidrógeno o dador de puente de hidrógeno (azul), y un *feature* que puede ser dador de puente de hidrógeno, aceptor de puente de hidrógeno o un átomo catiónico (rosa).

Tanto las moléculas sobre las que se genera la hipótesis farmacofórica como las de la base de datos de búsqueda se caracterizan según un esquema farmacofórico, que incluye el modo de anotación de los ligandos, es decir, aquellos puntos en el espacio donde se indica la ausencia/presencia de una determinada característica o *feature* farmacofórica (dadores de puente de hidrógeno, aceptores de puente de hidrógeno, cationes, aniones, centro de anillo aromático, región hidrofóbica, ligando metal y restricción de volumen). Los esquemas farmacofóricos disponibles en Moe, descritos según el motivo *Polarity-Charge-Hydrophobicity* (PCH), son:

- PCH: caracteriza puntos de ligando, átomos dadores y aceptores de puente de hidrógeno, cationes, aniones, áreas hidrófobas y centros aromáticos.

- PCH_ALL: similar a PCH, en este caso los átomos hidrofóbicos no aromáticos se caracterizan individualmente (un punto por átomo), en lugar de agruparse en un área, como en el esquema PCH.

- PCHD: incluye el esquema PCH y adicionalmente genera *site points*, que representan la posición hipotética de átomos complementarios en un receptor, determinados a partir de la posición de los átomos pesados en el ligando. Así, tiene puntos putativos proyectados a partir de dadores y aceptores de puente de hidrógeno y centros aromáticos.

- PPCH: diferencia entre aceptores/dadores de puente de hidrógeno planares (sp2) o no (sp3) y entre áreas hidrofóbicas planares o no.

- PPCH_All: de forma análoga a PCH_ALL, es un esquema derivado de PPCH en el que los átomos hidrofóbicos se anotan individualmente y no por agrupación, como en PPCH.

A partir del alineamiento de conformaciones o de una única conformación seleccionada, el usuario define las restricciones de la *query*, ajustando las posiciones, radios de tolerancia de los puntos potenciales farmacofóricos (PPPs), sus combinaciones y, adicionalmente, volúmenes.

Una vez formulado el modelo farmacofórico, la búsqueda se realiza sobre una base de datos multiconformacional previamente calculada, ya que no se generan conformaciones durante la búsqueda, sino que cada una de las entradas de la base de datos se superpone de forma rígida sobre la hipótesis (herramienta *Pharmacophore Search*). Entonces, se realiza el emparejamiento exhaustivo de todos los puntos de anotación del ligando con los PPPs del modelo. Se introduce cierta noción de conservación de éstos al permitirse, opcionalmente, que ciertas restricciones no se satisfagan por parte de la molécula en cuestión. El resultado de la búsqueda refleja la desviación cuadrática media (*rmsd*) de la superposición entre los PPPs de la hipótesis y los puntos del ligando emparejados con ellos, por lo que puede ordenarse la base de datos en función de esta *rmsd*. De esta forma, la verosimilitud del modelo farmacofórico se mide teniendo en cuenta la superposición de moléculas activas (modo de unión común), la cual se cuantifica mediante una puntuación calculada según la mejor/peor superposición de PPPs (*overlay scoring*) o el *rmsd*, y la relación con la actividad se mide mediante la exactitud en la clasificación de las moléculas según su *scoring*.

Discovery Studio, a diferencia de Moe, permite muestrear automáticamente y de manera exhaustiva todas las posibles combinaciones/alineamientos de las características farmacofóricas. El protocolo *Common Feature Pharmacophore Generation* permite alinear las conformaciones sobre las cuales se establecerán las hipótesis. Dicho protocolo utiliza los algoritmos Hypogen y HipHop para generar hipótesis y seleccionar los mejores farmacóforos producidos con características comunes a las conformaciones introducidas como input. HypoGen genera hipótesis en 3D para explicar las diferencias en la actividad a partir de la estructura de los candidatos. HipHop realiza alineaciones basándose en las características estructurales de una colección de compuestos y genera modelos de farmacóforos a partir de éstos. El resultado es una lista de hipotéticas *queries* ranqueadas según encajen mejor o peor en las conformaciones utilizadas para generar la hipótesis.

Al igual que en Moe, Discovery Studio posee aceptores de puente de hidrógeno, dadores de puente de hidrógeno, grupos funcionales hidrofóbicos, región ionizable positiva, región ionizable negativa, carga positiva, carga negativa y anillo aromático como características farmacofóricas (*pharmacophore features*). Asimismo dispone del protocolo *Steric Refinement with Excluded Volumes* el cual utiliza el algoritmo Catalyst HypoRefine para añadir volúmenes de exclusión a un farmacóforo generado.

La generación de farmacóforos en Discovery Studio requiere primero del cálculo de las conformaciones 3D tanto de los ligandos utilizados para establecer la hipótesis como de los pertenecientes a la base de datos muestreada. Ello se realiza mediante los protocolos *Diverse Conformation Generation* (el cual utiliza el algoritmo Catalyst Confirm) y *Build 3D Database* para bases de datos con gran número de compuestos. Seguidamente se construye el modelo farmacofórico con *Common Feature Pharmacophore Generation* tal y como se ha mencionado anteriormente. Para ello se definen el número de hipótesis a obtener, el número máximo/mínimo y el tipo de restricciones que pueden contener las hipótesis, y los volúmenes de exclusión si se utilizan, entre otras cosas.

Una vez se obtiene el modelo farmacofórico, se realiza la búsqueda sobre la base de datos multiconformacional, comparando/superponiendo cada una de las conformaciones sobre la *query* de manera rígida o flexible según opción (mediante *Ligand Pharmacophore Mapping* o *Search 3D Database* para bases de datos muy extensas). Cada una de las conformaciones dispone de la opción *Principal*, la cual permite indicar si el ligando es activo (2), moderadamente activo (1) o inactivo (0), lo cual influye a la hora de aplicar la hipótesis. Además la opción *Maximum omitted features* permite indicar cuantas características del farmacóforo pueden no cumplirse para cada molécula alineada a la hipótesis. Si se han de cumplir todas las características para considerar que una molécula está bien superpuesta al modelo farmacofórico, se asigna el valor 0. El resultado de la búsqueda proporciona una puntuación (*fit value*) para las moléculas que encajan en la *query* ordenadas según coincidan de mejor a peor con las características de la hipótesis escogida.

1.7. Técnicas de *shape matching*

En esta tesis se han utilizado las técnicas de *shape matching* (superposición de forma) implementadas en los programas Parasurf/Parafit [266, 267], Hex [255, 267] y Rocs [268].

Para entender las técnicas de *shape matching* o solapamiento de forma, cabe tener bien presente la definición de forma. Se puede decir que dos entidades tienen la misma forma si sus volúmenes corresponden exactamente. Ahora bien, dos objetos pueden tener el mismo volumen y no poseer la misma forma. El volumen se puede definir como una función que tiene un valor escalar en cada punto del espacio.

$$V(volumen) = \int f(x,y,z)dv \qquad \textbf{[1. 23]}$$

Por lo tanto, una definición de similitud en cuanto a forma puede venir dada por la siguiente ecuación:

$$S = \sqrt{\int \left[f(x,y,z) - g(x,y,z)\right]^2 dV} \qquad \textbf{[1. 24]}$$

donde *f* y *g* son dos funciones de volumen. Si el valor de la integral es cero, implica que *f* y *g* son la misma función y por lo tanto corresponden a la misma forma. Multiplicando los términos de la integral se obtiene:

$$S^2 = \int f(x,y,z)^2 dV + \int g(x,y,z)^2 dV - 2\int f(x,y,z)g(x,y,z)dV \qquad \textbf{[1. 25]}$$

Esta es la ecuación fundamental utilizada para comparar formas. Si se reescribe de manera simplificada se obtiene:

$$S_{f,g} = I_f + I_g - 2O_{f,g} \quad \textbf{[1. 26]}$$

donde los términos *I* equivalen a la superposición del volumen de una entidad consigo misma (molécula, en el caso de este estudio), y el término *O* representa la superposición entre las dos funciones. Estos tres términos constituyen pues lo necesario para comparar la forma entre dos campos. Los términos *I* son independientes de la orientación, pero *O* no lo es. Encontrar la orientación que maximiza *O*, y por lo tanto minimiza $S_{f,g}$ es equivalente a encontrar la mejor superposición entre los dos objetos. De esta ecuación se puede derivar el coeficiente de Tanimoto como medida de similitud recombinando *I* y *O*:

$$Tanimoto_{f,g} = \frac{O_{f,g}}{I_f + I_g - O_{f,g}} \quad \textbf{[1. 27]}$$

Las moléculas se representan tradicionalmente como un conjunto de esferas fusionadas, lo cual se refiere usualmente como modelo CPK. El volumen molecular se puede ver pues, como una función que posee el valor de 1, al menos dentro de una esfera, y el valor de 0 fuera. El volumen de una única esfera es $4\pi r^3/3$ pero la complicación para dos esferas fusionadas es que hay que contar el volumen compartido y no hacerlo dos veces. La fórmula general para *N* esferas que calcula explícitamente el volumen de cada nivel de solapamiento y su correcta contribución es la siguiente:

$$V = 1 - \int \prod_{i}^{N} (1 - f_i) dv \quad \textbf{[1. 28]}$$

Los intentos de usar la fórmula analítica para superposiciones de orden creciente dieron lugar a programas muy lentos y métodos aproximados. Ahora bien, si se utiliza una suma de funciones continuas, como una gaussiana, el volumen de una esfera sólida puede ser recubierto con mayor precisión. El programa Rocs utiliza esta aproximación para calcular rápidamente superposiciones de formas. Mientras que una esfera tiene un parámetro que la define, su radio, una gaussiana posee dos parámetros, su prefactor *p* y su anchura *w*:

$$pe^{-wx^2} \quad \textbf{[1. 29]}$$

Asimismo, los términos de solapamiento entre dos átomos cualesquiera, y por lo tanto cualquier superposición de orden superior, son todos funciones gaussianas:

$$\int e^{a(x-x_i)^2} e^{b(x-x_i)^2} = \int e^{(a+b)(x-x_i)^2} \quad \textbf{[1. 30]}$$

Por lo tanto, la superposición de dos gaussianas correspondientes a dos átomos produce otra gaussiana, así como la superposición de tres gaussianas correspondientes a tres átomos produce otra gaussiana. La simplicidad de estas fórmulas, así como la fórmula para el volumen de cada gaussiana individual conduce a algoritmos muy eficientes para el cálculo del volumen de una molécula.

Rocs alinea moléculas mediante un proceso de optimización que maximiza la superposición del volumen entre ellas. Tal como se ha mostrado previamente, para llevar a cabo el solapamiento de volúmenes no se trata con volúmenes de esferas rígidas, sino que se utilizan gaussianas parametrizadas que reproduzcan el volumen de esferas rígidas. Dado que la forma y el volumen están muy unidos en este contexto, la maximización del solapamiento de volúmenes permite reconocer formas similares. A pesar de que Rocs es fundamentalmente un método basado en la forma (*shape-based*), se pueden incluir también propiedades químicas (aceptor/dador de puente de hidrógeno, anión/catión, hidrofobicidad/aromaticidad…) en el proceso de análisis de superposición y similitud, lo cual facilita la identificación de compuestos que son similares en cuanto a la forma y a la química.

Parafit y Hex calculan del mismo modo la superposición de formas moleculares. Dado que en esta tesis (Artículo II) se ha implementado una nueva técnica de *shape matching* utilizando el software Parafit, se detalla el fundamento teórico de este programa con más detalle.

ParaFit superpone y compara moléculas utilizando las expansiones esféricas harmónicas (SH) de la superficie molecular y propiedades locales de la superficie calculadas con ParaSurf. Explotando las propiedades rotacionales especiales de las funciones esféricas harmónicas básicas [267], los tiempos de cálculo se pueden reducir varios órdenes de magnitud comparado con los algoritmos de *shape matching* convencionales. Por lo tanto el modulo ParaFit es un componente esencial de la suite Parasurf para estudios de cribado virtual en los que un elevado numero de compuestos ha de ser evaluado.

ParaFit proporciona tres modos de cálculo principales. En el modo por defecto, ParaFit superpone una o más moléculas "movibles" sobre una molécula de referencia "fija". El programa puede también realizar superposiciones todos-*versus*-todos en las que cada molécula se superpone a todas las demás. En este modo denominado "matrix", se retorna una tabla de comparación de *scores* en un formato adecuado para el posterior análisis de *clusters*, por ejemplo. Además de superponer moléculas, ParaFit también puede ser utilizado para alinear moléculas sobre los ejes de coordenadas (Figura 1.8) con el fin de colocarlas en una orientación estándar o canónica, lo cual es a menudo el primer paso para los estudios de QSAR. Asimismo, puede también aplicar transformaciones arbitrarias de coordenadas a una lista de ParaSurf, VAMP, o Mopac *sd-files* [269].

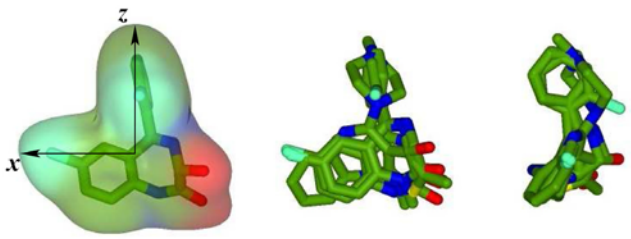


Figura 1.8 Alineamiento de moléculas sobre los ejes principales con Parafit [270].

Las formas esféricas harmónicas (SH) de la superficie molecular (Figura 1.9) se representan como expansiones radiales de la forma:

$$r(\theta,\phi) = \sum_{l=0}^{L} \sum_{m=-l}^{l} a_{lm}\, y_{lm}(\theta,\phi) \qquad \textbf{[1. 31]}$$

donde $y_{lm}(\theta,\phi)$ son las funciones reales esféricas harmónicas normalizadas, a_{lm} son los coeficientes de expansión y L es el orden o la mayor potencia polinomial de la expansión. Los parámetros (θ,ϕ) son las coordenadas esféricas con respecto al centro de la expansión harmónica (CoH). ParaSurf normalmente asigna el CoH al centro molecular de gravedad (CoG).

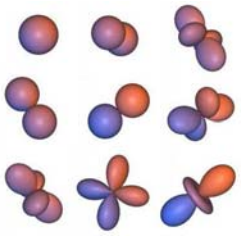


Figura 1.9 Funciones básicas reales esféricas harmónicas [271].

Dado que las funciones esféricas harmónicas forman un conjunto ortonormal, se puede demostrar que se transforman entre sí bajo rotación de acuerdo con:

$$y'_{lm}(\theta',\phi') = \sum_{m'=-l}^{l} R^{(l)}_{m'm}(\alpha,\beta,\gamma)\, y_{lm'}(\theta,\phi) \qquad \textbf{[1. 32]}$$

donde $R^{(l)}{}_{m'm}(\alpha,\beta,\gamma)$ son elementos reales de la matriz de rotación de Wigner expresados en términos de los ángulos de rotación de Euler *z-y-z*, (α,β,γ). Utilizando esta propiedad rotacional es sencillo observar que una expansión SH rotada puede construirse a partir de una expansión no rotada, únicamente rotando los coeficientes de expansión originales [267]:

$$a'_{lm} = \sum_{m'=-l}^{l} R^{(l)}_{mm'}(\alpha,\beta,\gamma)\, a_{lm'} \qquad \textbf{[1. 33]}$$

Para calcular la superposición entre un par de moléculas, ParaFit traslada el CoH de la molécula que se mueve (B) al de la molécula fija de referencia (A) y después busca la rotación que minimiza la "distancia" entre los correspondientes pares de expansiones esféricas harmónicas:

$$D_{EUCLIDEAN} = \int [(r_A(\theta,\phi) - R(\alpha,\beta,\gamma) r_B(\theta,\phi)]^2 d\Omega \qquad \textbf{[1. 34]}$$

Aprovechando la ortonormalidad de las funciones base, esta expresión se reduce a:

$$D_{EUCLIDEAN} = |\underline{a}|^2 + |\underline{b}|^2 - 2\,\underline{a}.\underline{b}' \qquad \textbf{[1. 35]}$$

donde $\underline{b}'$ representa el vector de coeficientes de expansión SH rotados de la molécula que se mueve, $\underline{b}$ representa el vector de coeficientes de expansión originales de la molécula que se mueve, y $\underline{a}$ representa el vector de coeficientes de expansión de la molécula fija. ParaFit llama a esto función de distancia Euclídea debido a su analogía con las distancias Euclídeas en el espacio 3D. Esta función de distancia depende claramente del tamaño relativo de las moléculas que se comparan. Sin embargo, cuando se comparan múltiples moléculas, es a menudo conveniente utilizar funciones de similitud normalizadas en las cuales moléculas idénticas retornan como *score* la unidad. ParaFit implementa las siguientes funciones de similitud:

$$S_{CARBO} = \frac{\underline{a}.\underline{b}'}{|\underline{a}|.|\underline{b}|} \quad \textbf{[1. 36]}$$

$$S_{HODGKIN} = \frac{2\,\underline{a}.\underline{b}'}{|\underline{a}|^2 + |\underline{b}|^2} \quad \textbf{[1. 37]}$$

$$S_{TANIMOTO} = \frac{\underline{a}.\underline{b}'}{|\underline{a}|^2 + |\underline{b}|^2 - \underline{a}.\underline{b}'} \quad \textbf{[1. 38]}$$

De las definiciones anteriores se puede observar que:

$$S_{HODGKIN} = 1 - \frac{D_{EUCLIDEAN}}{|\underline{a}|^2 + |\underline{b}|^2} \quad \textbf{[1. 39]}$$

Para cada una de las funciones anteriores, ParaFit permite un *score* compuesto calculado a partir de una combinación arbitraria de propiedades SH locales:

$$S = w_{SURFACE}S^{SURFACE} + w_{MEP}S^{MEP} + w_{IEL}S^{IEL} + w_{EA}S^{EA} + w_{\alpha_L}S^{\alpha_L} \quad \textbf{[1. 40]}$$

donde $w_{SURFACE}$ representa un factor de peso definido por el usuario referente a la forma de la superficie molecular, w_{MEP} es el factor de peso referente al potencial molecular electrostático, w_{IEL} es el factor de peso referente a la energía local de ionización, w_{EA} es el factor de peso referente a la afinidad electrónica y $w_{\alpha L}$ es el factor de peso referente a la polarizabilidad local.

Parafit superpone moléculas utilizando una fuerza bruta en una búsqueda rotacional teniendo en cuenta los tres ángulos de rotación de Euler. Conceptualmente, cada molécula movible se rota respecto a la molécula de referencia fija, y la rotación de Euler que proporciona un *score* correspondiente a la mayor similitud (o menor distancia) se guarda. Esto es esencialmente una búsqueda de correlación por transformada de Fourier en las coordenadas de los ángulos de Euler. Sin embargo, dado que se pueden conseguir buenas superposiciones utilizando únicamente expansiones harmónicas de bajo orden L ≈ 6 (Figura 1.10), no es necesario utilizar técnicas de transformadas rapidas de Fourier (FFT) para acelerar el cálculo. Además la FFT es solo provechosa cuando L>=16, lo cual es considerablemente más elevado que el valor recomendado por defecto en Parafit (L=6).

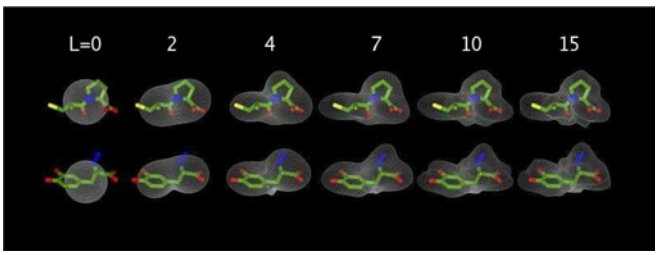


Figura 1.10 Expansiones esféricas harmónicas de orden L [271].

Además de utilizar búsquedas de correlación de bajo orden, los cálculos de superposición con ParaFit se pueden acelerar de dos maneras. La primera técnica aprovecha el hecho de que las expansiones harmónicas de orden L no pueden tener más de $(L+1)^2/2$ máximos locales. Por lo tanto, ParaFit inicialmente utiliza ángulos de búsqueda relativamente grandes (*steps* de unos 8^o)

para cubrir el espacio de búsqueda. Con el fin de muestrear el espacio de forma eficiente, este muestreo angular es generado a partir de los vértices de una teselación icosaédrica de la esfera (Figura 1.11). Para un tamaño de paso angular dado, este método aporta alrededor de un 30% menos de puntos de muestreo que una *grid* equiangular [267]. Una vez se ha localizado aproximadamente el máximo de similitud, éste se refina mediante una búsqueda en una *grid* localizada utilizando *steps* de 2^{o}. Ambos pasos angulares de búsqueda pueden ser ajustados por el usuario.

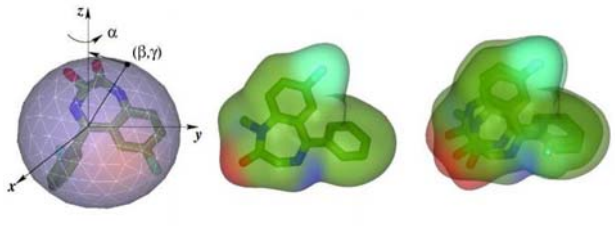


Figura 1.11 Rotaciones de Euler generadas a partir de una teselación icosaédrica de la esfera [267].

La segunda técnica, utilizada cuando se comparan múltiples moléculas, consiste en que en lugar de ir rotando cada una de las moléculas movibles separadamente por turno, es mucho más eficiente rotar las expansiones SH únicamente de la molécula de referencia y comparar éstas respecto a vectores estáticos de coeficientes SH de cada una de las moléculas movibles. De esta forma, la multiplicación relativamente costosa de matrices de rotación SH se aplica únicamente para una en vez de para N moléculas. Una vez se han encontrado las rotaciones óptimas, las moléculas movibles son rotadas posteriormente usando la inversa de las rotaciones de la referencia correspondientes. Utilizando estas técnicas, un par de moléculas se puede superponer en unos 1/20 segundos en un procesador Pentium Xeon de 1,8 GHz, y los tiempos de cálculo se pueden reducir hasta un factor de 5 si se comparan múltiples moléculas en un único cálculo de Parafit.

Una vez se ha determinado la orientación de la rotación, ParaFit retorna cada una de la moléculas movibles rotadas en un nuevo archivo *sd-file* (el cual permite codificar en un solo fichero ascii la estructura tridimensional de cada una de las moléculas, así como descriptores de valores reales). Cada archivo nuevo contiene las coordenadas atómicas originales rotadas y trasladadas, cargas puntuales, multipolos moleculares y atómicos, elementos de la matriz de densidad y los coeficientes de expansión SH. El resto de parámetros del nuevo *sd-file* son copiados sin cambios de los datos originales.

1.8. *De novo design*

Tal y como se ha expuesto en la introducción, las técnicas de diseño *de novo* permiten proponer nuevas moléculas con actividad biológica potencial. El diseño *de novo* de ligandos puede basarse en el ligando o en el receptor. El *de novo design ligand-based* utiliza farmacóforos para guiar la colocación de los fragmentos. Ello da lugar a *hits* que no solo complementan el sitio activo de la proteína sino que también poseen las características farmacofóricas de los ligandos activos. Con esta técnica se pueden crear listas de nuevos compuestos que contengan características conocidas como críticas para la unión a una diana farmacológica específica. Dicha aproximación puede llevarse a cabo en Discovery Studio 2.0 con la herramienta DS De Novo Ligand Builder. El *de novo design structure-based* utiliza un mapa de interacciones creado a partir del sitio activo del receptor. Con ello se pueden crear listas de nuevos compuestos confeccionados únicamente para

encajar en un receptor específico. Dicha aproximación puede llevarse a cabo en Discovery Studio 1.7 y 2.0 con la herramienta DS Catalyst Structure Based Pharmacophore (protocolos De Novo Receptor, De Novo Link, y De Novo Evolution). Asimismo Discovery 2.0 permite integrar las características derivadas de la estructura del receptor con las características derivadas del ligando para crear un modelo más completo de las características necesarias de un ligando para la unión al receptor. Esta aproximación se considera de especial interés cuando la estructura de la proteína con que se trabaja no está del todo definida.

En esta tesis (Apéndice III) se ha utilizado Discovery Studio 1.7, con lo que se ha llevado a cabo la metodología *de novo design structure-based*, la cual se detalla un poco más a continuación.

Se han utilizado dos procedimientos *de novo receptor-based*: búsqueda de fragmentos que encajen en el mapa del sitio activo del receptor (protocolo De Novo Receptor) y búsqueda de fragmentos que se unan a una estructura base (*scaffold*) dada situada ya en el mapa del receptor (protocolos De Novo Link y De Novo Evolution, véase Figura 1.12). La selección de fragmentos en ambos procedimientos se basa en los compuestos que posean mayor Ludi *score* [251]. Asimismo, el protocolo De Novo Evolution se puede utilizar en tres modos diferentes: *full evolution mode*, el cual selecciona supervivientes de generaciones de compuestos ranqueados según Ludi *score* (en cada generación se fusiona un fragmento al *scaffold*), *quick mode*, el cual sugiere el compuesto con mejor *score* después de cada generación (en cada generación múltiples fragmentos se fusionan al *scaffold*), y *combinatorial mode*, el cual enumera todas las combinaciones de fragmentos unidas a los puntos del *scaffold* especificado.

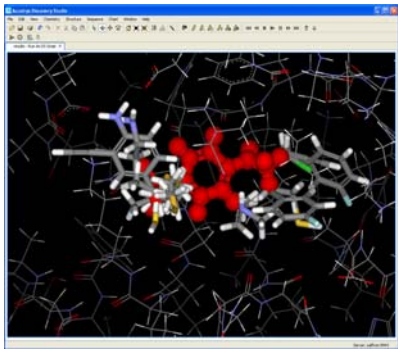


Figura 1.12 Diseño *de novo structure-based* mediante el protocolo De Novo Evolution de Discovery Studio. El *scaffold* inicial se muestra en el sitio activo de la proteína representado en *balls and sticks* de color rojo y los fragmentos encontrados, los cuales son fusionados covalentemente al *scaffold,* se muestran en representación de *sticks* de color de los elementos. La proteína se muestra en representación de *lines*.

Para encontrar fragmentos que encajen en el receptor, se debe definir el sitio activo de la proteína mediante una esfera de radio determinado, y utilizar una librería de fragmentos pre-generada, ya sea la librería de fragmentos por defecto (librería estándar Ludi), otra librería comercial, o una creada previamente mediante *De novo Library Generation*. Para encontrar fragmentos que se unan a un *scaffold* dado, se debe colocar el *scaffold* de manera correcta en el sitio activo de la proteína, definir el sitio activo, definir los sitios de unión (*link points*) en que se unirán los nuevos fragmentos al *scaffold*, y utilizar una librería de fragmentos por defecto (librería estándar Ludi Link), otra librería comercial, o una creada previamente mediante *De novo Library Generation*.

Durante el proceso de encaje de los fragmentos en el receptor, los fragmentos se pueden mantener rígidos o se pueden rotar uno o dos enlaces a la vez para generar nuevas conformaciones de fragmentos. Alternativamente, la flexibilidad conformacional se puede tener en cuenta incluyendo directamente en la librería de fragmentos diversos confórmeros para un fragmento. La librería Ludi estándar contiene aproximadamente 1000 entradas con estructuras de 5 a 30 átomos. La librería Ludi Link estándar contiene aproximadamente 1100 entradas (900 entradas *one-link*, 150 entradas *two-link*, y 50 entradas *three-link*).