Universitat Ramon Llull



# TESIS DOCTORAL

| | |
|---|---|
| Título | **Herramientas de cribado virtual aplicadas a inhibidores de tirosina quinasas. Contribución al desarrollo del programa PRALINS para el diseño de quimiotecas combinatorias.** |
| Realizada por | Obdulia Rabal Gracia |
| en el Centro | Escola Tècnica Superior IQS |
| y en el Departamento | Química Orgánica y Bioquímica |
| Dirigida por | Dr. Jordi Teixidó i Closa |

*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*

# Agradecimientos

En primer lugar, quisiera agradecer al Dr. Jordi Teixidó, director de esta tesis, el apoyo prestado, tanto en el ámbito computacional como en el personal, en el que no han faltado ánimos y amistad para seguir adelante. En plan más "materialista" no puedo dejar de agradecer la cantidad de medios que ha puesto siempre a mi disposición, así como su ayuda en esas tareas burocráticas de última hora (aunque reconocerás que he ido progresando).

Este agradecimiento se extiende a todos los miembros del GEM, especialmente al Dr. Ignacio Borrell por su colaboración e interés mostrado y por permitirme realizar la tesis en este proyecto. Asimismo, quisiera agradecer a varios profesores del IQS que a lo largo de la carrera o el doctorado me han prestado su apoyo de modos diversos: Dr. José Javier Molins, Dr. Santiago Nonell, Dra. Pepa Blanco, Dr. Xavier Tomás y Dr. Alberto Barrera. Al Dr. Jordi Cuadros, gracias por tu amistad y ayuda, especialmente en nuestro período de ocupación de la sección de Estadística.

A todos los miembros de TICS (Sergi, Joaquín, Javi y Susana) con los que hemos compartido tantas comidas y buenos ratos. Especialmente, Sergi, tu trabajo, paciencia y ayuda con las máquinas ha sido impresionante.

A todos los compañeros de la sección que han ido pasando a lo largo de estos años y que me hicieron pasar buenos momentos. A Rosalía: en el TFC te dije que pondría "sin la cual este proyecto no es el que es", bien, me mantengo, sin el previo de PRALINS no podría haber aprendido a programar. Ya sabes que además te agradezco multitud de favores y tu inestimable amistad. A Oscar (El Rey) por transmitirme tanto optimismo, alegría y serenidad en tus consejos, ¡qué hueco dejaste cuando te fuiste! A Violeta por tu colaboración en este trabajo y por tú amistad, ayudándome en los momentos difíciles del tramo final de la tesis. A Roger por su colaboración, porque de todo se aprende. Asimismo, gracias a todos los compañeros de las secciones de síntesis, fotoquímica y esteroides por animar los momentos en el IQS.

Al Dr. Gisbert Schneider por permitirme realizar una estancia en su grupo en Frankfurt. A todos los amigos que allí hice, que tan cariñosamente me acogisteis y con los que tantas cervezas bebimos: Teresa, Tina, Swetlana, Domingo, Karin, Philip, Carlos, Stephen, Lutz, Andreas, Leyla, Micha, Michi, Norbert…y así hasta el final del grupo MODLAB.

A las de toda la vida: Teresa, María, Mepi, Yolanda y Sandra y a toda la gente del Moko (vamos para 15 años…). A todos los amigos que he conocido en Barcelona, en especial a Núria y Kike, Ángeles, César, María, Dani, Marc y Pere…A la pandilla "política" de Pamplona.

Iñigo, una suerte haberte conocido. Aunque esta tesis cada vez nos ha ido quitando más tiempo, espero devolvértelo.

A toda mi familia, con mucho cariño para mi abuela Irene. A Jorge y Fina, os agradezco lo mucho que hicisteis por mí en mis comienzos en Barcelona. A la familia de Iñigo por el aprecio y todas las atenciones que tenéis para conmigo.

Antonio, aunque eres el hermano pequeño tus consejos y saber hacer me ayudan día a día, eso por no hablar de las sesiones de risas y cachondeo…Ya que no pongo cita, imagina que está escrita tu sugerencia de Cálico electrónico.

A mis padres, Antonio y Obdulia, a los que dedico esta tesis por todo el esfuerzo que siempre habéis hecho por mí y mi hermano.

Esta tesis se ha realizado gracias al apoyo económico de la Generalitat de Catalunya mediante una beca FI (2003FI00934) y una bolsa de viaje (2005BE00124). También agradezco los medios que el Instituto Químico de Sarriá ha puesto a mi disposición.

# Abreviaciones y acrónimos

| | |
|---|---|
| ACD | *Available Chemical Directory* |
| ADMET | Absorción, Distribución, Metabolismo, Eliminación y Toxicidad |
| AMBER | *Assisted Model Building with Energy Refinement* |
| AMP-PNP | Adenililimidodifosfato |
| BBB | *Blood brain barrier*. Barrera hematoencefálica |
| BLAST | *Local Alignment Search Tools* |
| BLOSUM | *Blocks Substitution Matrix* |
| CATS | *Chemically Advanced Template Search* |
| CDK | *Cyclin Dependent kinase*. Quinasa dependiente de ciclina |
| C.I.D | *Cell-integral-diversity criterion* |
| CG | *Conjugate Gradients*. Gradiente conjugado |
| CMC | *Comprehensive Medicinal Chemistry* |
| CoMFA | *Comparative Molecular Field Analysis* |
| CONGEN | *CONformation Generator* |
| CV | *Correlation-vector*. Vector de correlación |
| DAG | 1,2-diacilglicerol |
| D.I | *Diversity integral criterion* |
| ECL | *Extracellular Loop*. *Loop* extracelular (GPCRs) |
| EGFR | *Epidermal Growth Factor Receptor.* Receptor del factor de crecimiento epitelial |
| EP | *Evolutionary programming*. Programación evolutiva |
| ER | *Estrogen receptor*. Receptor de estrógeno |
| Erk | *Extracellular signal-regulated kinase.* Quinasa regulada por señales extracelulares |
| ES | *Evolution strategies.* Estrategias evolutivas |
| FAST | *Fast Alignment* |
| FEP | *Free Energy Perturbation* |
| FGFR | *Fibroblast Growth Factor Receptor*. Receptor del factor de crecimiento de fibroblastos |
| FR | Fletcher-Reeves |
| GA | *Genetic algorithm*. Algoritmo genético |
| GAFF | *General Amber Force Field* |
| GAP | Proteínas activadoras de GTPasa |
| GB | *Generalizad Born*. Generalizado de Born |
| GP | *Genetic programming.* Programación genética |
| GPCR | *G protein coupled receptor*. Receptor acoplado a proteína G |
| GRIND | *GRind INdependent Descriptors* |
| HGFR | *Hepatocyte Growth Factor Receptor*. Receptor del factor de crecimiento de hepatocitos |
| HS | Hestene-Stiefel |
| HTS | *High-throughput screening* |
| ICL | Intracellular *loop*. *Loop* intracelular/citoplasmático (GPCRs) |
| IF | *Interaction Fingerprint*. *Fingerprint* de interacción |
| IFbAP | *Interaction Fingerprint Based on Atom Pairs* |
| Ig | Inmunoglobulina |
| $IP_3$ | Inositol 1,4,5-trisfosfato |
| IR | *Insulin receptor*. Receptor de insulina |
| JAK | *Janus kinase*. Quinasa Janus |
| JNK | *c-Jun N-terminal kinase*. Quinasa c-Jun N-terminal. |
| KID | *Kinase insert domain*. Dominio inserto quinasa |
| LFDs | *Local feature densities* |
| LGA | *Lamarckian genetic algorithm*. Algoritmo genético lamarckiano |
| LIE | *Linear Interaction Energy* |
| LS | *Local Search.* Búsqueda local |
| MAPK | *Mitogen-Activated Protein Kinase*. Proteína quinasa activada por mitógenos |
| MCR | *Multicomponent reaction*. Reacción multicomponente |
| M-CSF | *Macrophage colony-stimulating factor*. Factor estimulador de formación de colonias de macrófagos |
| MD | *Molecular dynamics*. Dinámica molecular |
| MDDR | *MACCS-II Drug Data Report* |
| MDS | *Multidimensional scaling*. Escalado multidimensional |

| | |
|---|---|
| MEP | *Molecular Electrostatic Potential.* Potencial electrostático molecular. |
| MIP | *Molecular Interaction Potential*. Potencial de interacción molecular. |
| MLR | *Multiple linear regression*. Regresión lineal múltiple |
| MM | *Molecular mechanics.* Mecánica molecular |
| MM-PBSA | *Molecular Mechanics Poisson-Boltzmann Surface Area* |
| MM-GBSA | *Molecular Mechanics Generalized-Born Surface Area* |
| MOE | *Molecular Operating Environment* |
| MOGA | *MultiObjective Genetic Algorithm*. Optimización multiobjetivo algoritmos genéticos. |
| MOP | *Multiobjective Optimisation Problems*. Problemas de optimización multiobjetivo |
| MTC | Monte Carlo |
| MW | *Molecular weight*. Peso molecular |
| NGFR | *Nerve Growth Factor Receptor*. Receptor del factor de crecimiento neuronal |
| NNRTI | *Non-nucleoside reverse transcriptase inhibitor*. Inhibidor no nucleósido de transcriptasa reversa. |
| NPT | Colectivo isotermo-isobárico |
| NR | Newton-Raphson |
| NRTK | *Non-receptor tyrosine kinase*. Tirosina quinasa no receptora |
| NVE | Colectivo microcanónico |
| NVT | Colectivo canónico |
| PAM | *Accepted Point Mutation per 100 residues* |
| PB | Poisson-Boltzmann |
| PBC | *Periodic boundary conditions*. Condiciones periódicas de contorno |
| PC | *Principal component*. Componente principal |
| PCA | *Principal component analysis*. Análisis de componentes principales |
| PCH | *Polarity-Charge-Hydrophobicity* |
| PCR | *Principal Component Regression*. Regresión con componentes principales |
| *pdf* | *Probability density functions.* Funciones de densidad de probabilidad |
| PDGFR | *Platelet-derived Growth Factor Receptor*. Receptor del factor de crecimiento derivado de las plaquetas |
| Perl | *Practical Extraction and Report Language* |
| PI | *Predictive index* |
| $PIP_2$ | Fosfatidilinositol 4,5-bisfosfato |
| PI-3K | Fosfatidilinositol 3'-quinasa |
| PK | Polak-Riviere |
| PKA | *cAMP-dependent protein kinase*. Proteína quinasa dependiente de cAMP |
| PKB | *Protein kinase B*. Proteína quinasa B |
| PKC | *Protein kinase C*. Proteína quinasa C |
| PLC-γ | *Phospholipase C-γ.* Fosfolipasa C-γ |
| PLS | *Partial Least Squares.* Mínimos cuadrados parciales. |
| PPP | *Potencial Phamacophore Points*. Puntos potenciales farmacofóricos |
| PRALINS | *Program for Rational Analysis of Libraries in Silico* |
| pSIFt | *Profile Structural Interaction Fingerprint* |
| PTB | *Phosphotyrosine binding*. Dominio de unión a fosfotirosinas. |
| PTK | *Protein tyrosine kinase*. Proteína tirosina quinasa |
| QM | *Quantum mechanics.* Mecánica Cuántica |
| QSAR | *Quantitative Structure-Activity Relationships* |
| RESP | *Restrained ElectroStatic Potencial* |
| RMN | Resonancia Magnética Nuclear |
| RT | *Reverse transcriptase*. Transcriptasa reversa |
| RTK | *Receptor Tyrosine kinase*. Receptor tirosina quinasa |
| SA | *Simulated Annealing* |
| SAPK | *Stress-activated protein kinase*. Proteína quinasa activada por estrés |
| SAR | *Structure–activity relationship*. Relación estructura-actividad |
| SD | *Stepeest Descent* |
| SIFt | *Structural Interaction Fingerprint*. *Fingerprint* de interacción estructural |
| SH2 | *Src homology 2 domain*. Dominio de homología Src tipo 2. |
| SQUID | *Sophisticated Quantification of Interaction Distributions* |
| STAT | *Signal transducer and activator of transcription*. Transductor de señal y activador de transcripción |
| SVL | *Scientific Vector Language* |

| | |
|---|---|
| TCL | *Tool Command Language* |
| TI | *Thermodynamic Integration*. Integración termodinámica |
| TM | *Transmembrane*. Segmento transmembrana |
| vdW | van der Waals |
| VEGFR | *Vascular endothelial Growth Factor Receptor*. Receptor del factor de crecimiento vascular endotelial |
| VS | *Virtual Screening*. Cribado virtual |
| VTFM | *Variable Target Function Method* |
| WDI | *World Drug Index* |

# Sumario

La aplicación de métodos de **cribado virtual** cobra cada vez más importancia en el proceso de descubrimiento de fármacos, complementando a las técnicas de *High-throughput screening* con el fin de facilitar y contribuir a la comprensión de los mecanismos bioquímicos de actuación de los fármacos, agilizar y reducir el coste del proceso.

En particular, el interés farmacológico de la presente tesis es la inhibición de **receptores de tirosina quinasas**. Estos enzimas participan en múltiples procesos de señalización celular, por lo que tanto la disfunción de las mismas o su papel privilegiado en los mecanismos del ciclo celular las convierten en diana farmacológica de enfermedades como el cáncer y otras relacionadas con desórdenes hiperproliferativos, migratorios, del desarrollo embrionario y enfermedades vasculares. Una de las estrategias de inhibición más usuales es el bloqueo del sitio de unión del ATP a través de moléculas orgánicas como las piridopirimidinas, heterociclos especialmente interesantes para el grupo de investigación en el que se desarrolla este trabajo por su amplia experiencia sintética en dichos sistemas.

En la presente tesis se exploran y validan gran parte de las técnicas de cribado virtual con el objetivo de establecer una **secuencia jerarquizada de filtros** que permitan evaluar aquellos compuestos candidatos a ser sintetizados. Los sucesivos pasos de filtrado incluyen la selección de compuestos de una quimioteca virtual a partir de la **diversidad** o representatividad del espacio químico, la aplicación de **búsquedas de similitud y modelos farmacofóricos** construidos a partir de inhibidores conocidos, un filtrado mediante ***docking*** o acoplamiento de los inhibidores en la cavidad de unión de estas proteínas y métodos de **predicción de la afinidad** de unión de una serie de ligandos. La jerarquía de estas etapas se impone a partir de la diferencia de recursos computacionales que requiere cada una de ellas, siendo éstos cada vez superiores. Los métodos han sido validados retrospectivamente en bases de datos formadas por compuestos activos recopilados de la bibliografía. Una vez validadas, han permitido la caracterización prospectiva de los candidatos sintéticos.

Se ha diseñado un ***fingerprint* de interacción estructural proteína-ligando** basado en el concepto de pares atómicos (IFbAP) destinado a facilitar el postprocesado de los resultados de *docking*, aplicándose como filtro en un cribado virtual. Su capacidad para discriminar entre compuestos activos e inactivos se analiza para tres dianas: el receptor de estrógeno, el receptor del factor de crecimiento de fibroblastos y la transcriptasa reversa del HIV.

Paralelamente, se ha continuado con el desarrollo del programa **PRALINS** (Program for Rational Analysis of Libraries *in Silico*), programa dirigido al diseño de quimiotecas combinatorias virtuales que incorpora los principales criterios de selección basados en diversidad. En el contexto de las quimiotecas combinatorias focalizadas, se propone un nuevo método (**Direct**), cuya capacidad de focalización se ha testado frente a los métodos tradicionales, también implementados en PRALINS. Asimismo se incorporan y analizan métodos de evaluación de diversidad, sugiriéndose un método (***cell-integral-diversity criterion***) destinado a superar las desventajas de los métodos tradicionales. Se incorporan los algoritmos genéticos en PRALINS como técnica de optimización, tanto de un único criterio de diversidad/similitud como para realizar **optimizaciones multiobjetivo**.

En el ámbito de otra línea de investigación del grupo dirigida hacia el desarrollo de inhibidores del proceso de fusión del HIV, se estudia el modo de unión de dos **antagonistas de CXCR4** y **CCR5**, receptores celulares de la familia de las GPCRs implicados en dicha etapa del ciclo del virus.

# Índice

# Introducción

En el proceso de descubrimiento de fármacos, el primer paso crítico es la identificación de un buen cabeza de serie o *lead* (*lead discovery)*. Se considera un buen *lead* a aquellos que producen una inhibición del 50% de la actividad *in vitro* ($IC_{50}$) a una concentración alrededor de 10 μM. Una vez identificado dicho *lead*, comienza el proceso de *lead optimization*, cuyo objetivo es mejorar su eficacia terapéutica: incremento de su potencia frente a una diana o *target* (normalmente la $IC_{50}$ se rebaja a valores del rango de 1 a 10 nM), selectividad frente a dianas relacionadas, farmacocinética, minimización de su toxicidad y efectos secundarios.[1,2]

Las técnicas de *High-throughput screening* (HTS) se convierten, desde la década de los 90, en la principal fuente de obtención de nuevos *leads*. El HTS requiere una quimioteca de cientos de miles de compuestos y un método de ensayo de actividad.[3] Además, la introducción de la química combinatoria ha permitido que el tamaño de estas quimiotecas se incremente al orden de millones de compuestos. Por otra parte, la publicación del genoma[4] amplía el espectro de dianas biológicas susceptibles de ser moduladas por un fármaco. Todo ello conduce a que frente a las rutas tradicionales empleadas en química médica para el diseño de fármacos, aparezca la posibilidad de optar por la estrategia de testar experimentalmente todos los posibles candidatos frente a todas las posibles dianas.

Sin embargo, la realidad es que a pesar del uso de estas técnicas a gran escala, la tasa de descubrimiento de *leads* ha decaído[5] y pocos son los fármacos procedentes directamente de los resultados de HTS.[6] En un experimento de HTS, normalmente realizado en formato de dosis única-único experimento, los compuestos que resultan positivos (HTS *hits*) son nuevamente testados para confirmar actividad y estructura (debido a los problemas de pureza inherentes al uso de química combinatoria). Esta etapa de identificación de HTS *hits* tiene un éxito inferior al 0.1%. De cada 2000 HTS *hits*, aproximadamente 1200 se confirman como activos reales (HTS *actives*), ya que existe un gran número de falsos positivos que interfieren con los ensayos biológicos, de agregantes promiscuos y de interferencias causadas por los tintes y compuestos fluorescentes utilizados. Cuando se identifica un gran número de HTS *actives* pertenecientes a una misma familia química, se considera que se ha identificado una serie de *leads*. Cuando es posible optimizar estos *leads*, se habla de *drug candidate*. Típicamente, 1 de cada 10.000 HTS *actives* alcanza este nivel y únicamente 1 de cada 10 *drug candidates* supera las pruebas clínicas convirtiéndose en *drug*. En la Figura I.1 se detallan estas etapas junto con su factor de éxito.[7]

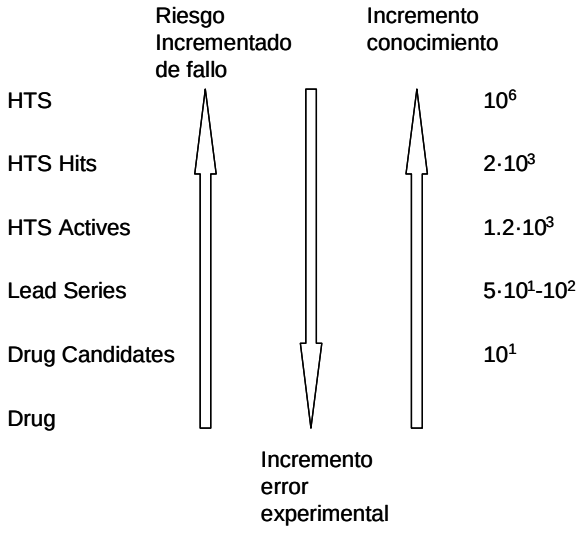



**Figura I.1**. Tasa de éxito y decaimiento en los protocolos de descubrimiento de fármacos.

Esta baja tasa de éxito, junto con el coste de estas técnicas, ha hecho que se replantee la aportación del HTS, perdiendo parte del protagonismo de la década pasada a la vez que las técnicas de diseño de fármacos asistido por ordenador cobran importancia.

Hansch y Leo[8] desarrollan, durante la década de los 60, los primeros estudios de QSAR (*Quantitative Structure-Activity Relationships*), aunque es durante los años 80 cuando se introduce el diseño racional en el proceso de diseño de fármacos. Esto coincide con el desarrollo teórico de técnicas de modelización molecular y la aparición de ordenadores personales. La contribución computacional en esta época se basa principalmente en optimizar y refinar los compuestos a partir de la información extraída de la estructura de complejos cristalinos con la estructura del receptor diana.
Posteriormente, la introducción del HTS hace que también en química computacional se comience a trabajar a nivel de quimiotecas. Aparecen así, en 1997, las técnicas de cribado virtual o *Virtual Screening* (VS), con la finalidad de seleccionar/identificar aquellas moléculas biológicamente activas frente a dianas particulares o dianas pertenecientes a una misma familia. Estás técnicas requieren inevitablemente que se conozca la actividad de algunos compuestos o bien la estructura de la diana biológica.
En los últimos cinco años, se ha incrementado el empleo de VS y, aunque sigue siendo menos usado que HTS durante el proceso de *lead discovery*, se afirma que llegará a reemplazarlo eventualmente. Con ello, cada vez se confía menos únicamente en la suerte o *serendipity* en el descubrimiento de fármacos, aunque siempre hay excepciones como el caso del sildenafil (Viagra®).[9]

Más que una alternativa, el VS supone un complemento al HTS y un criterio para la priorización de la síntesis y la adquisición de quimiotecas. Los posibles *hits* determinados por HTS son reales, aunque por sí solos, sin recurrir a biología molecular, no contribuyen a ampliar el conocimiento acerca del modo de interacción con su diana farmacológica. Por otra parte, el VS propone potenciales *hits* que ni siquiera tienen porqué, a no ser que se consideren otras restricciones, ser fácilmente accesibles sintéticamente. Sin embargo, aporta información acerca del modo de interacción fármaco-diana. Además, estas técnicas son relativamente baratas (ahorran la adquisición de reactivos y robotización), rápidas y permiten considerar un número de compuestos *in silico* del orden de billones, cifra prohibitiva experimentalmente. Típicamente, en una cascada de VS, una quimioteca virtual que contiene unas $10^6$-$10^{12}$ estructuras es sucesivamente filtrada y reducida a una colección de unos 100-1000 candidatos.

En la Figura I.2, se muestra la estructura de una cascada de *in silico screening* con los diferentes pasos de filtrado aplicados y la reducción de compuestos que conlleva cada uno de ellos. La aplicación secuencial de cada una de las técnicas se basa en el nivel de requerimientos computacionales que utiliza cada uno de los pasos y en la complejidad de la información aportada como entrada para cada uno de ellos. En el transcurso de la introducción se describen cada uno de estos pasos.

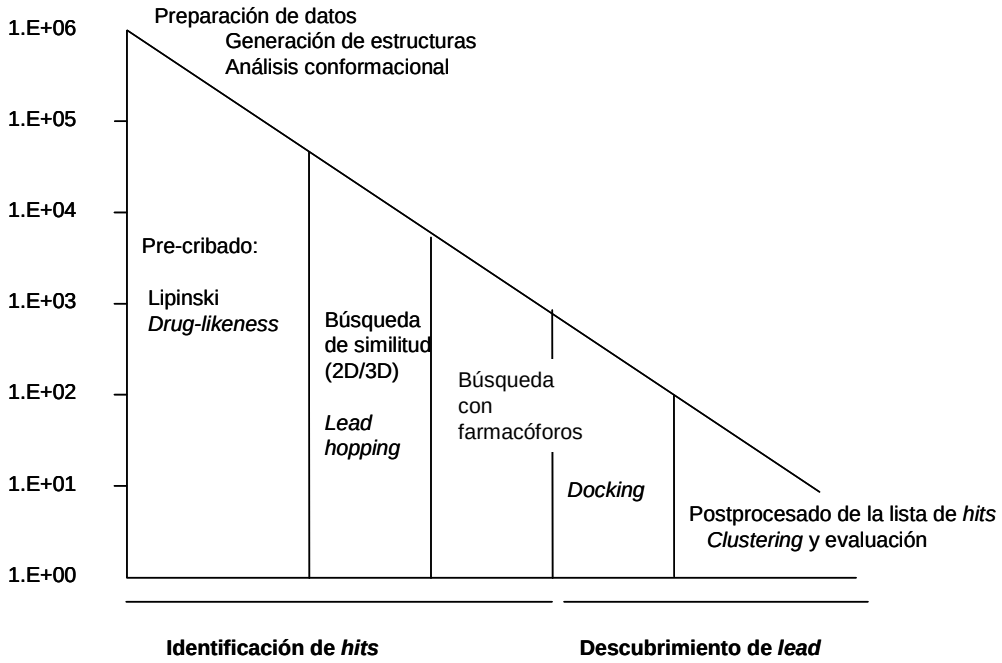



**Figura I.2.** Esquema de una cascada de cribado virtual. Adaptado de [10].

Las técnicas de selección de compuestos derivadas de cribado virtual se clasifican tradicionalmente en dos grandes grupos, dependiendo de cómo extraen la información que requieren. Aquellas que se basan en la estructura de inhibidores ya determinados se las denomina métodos indirectos o basados en la estructura del ligando (*Ligand-Based Virtual Screening*), mientras que los métodos que utilizan la estructura del receptor se denominan directos o basados en la estructura del receptor (*Structure-Based* ó *Receptor-Based Virtual Screening*).
Dentro de las aproximaciones basadas en ligandos se encuentran las búsquedas de similitud a compuestos activos, la obtención de modelos farmacofóricos y el QSAR. Por otra parte, el *docking*, que modela el acoplamiento entre proteína y ligando desde el punto de vista estructural y energético, y el diseño *de novo (de novo design*) corresponden a métodos directos.

La principal restricción de los métodos de VS es que, como se ha comentado, necesitan algún tipo de información previa acerca de los factores responsables de la actividad del fármaco. Sin embargo, cada vez se dispone de un mayor número de compuestos activos frente a familias de proteínas, se determinan secuencias de genes ligadas a determinadas proteínas y se incrementa el número de estructuras de proteínas resueltas experimentalmente, por rayos-X o por resonancia magnética nuclear (RMN).
Además, desde un punto de vista teórico, se sigue avanzando en la comprensión de las interacciones proteína-ligando, aunque todavía los métodos no se encuentran lo suficientemente desarrollados. Básicamente, el problema reside en la implementación de modelos físicos válidos para analizar en un tiempo asequible cientos de miles de posibles compuestos.

Finalmente, los ensayos de VS se pueden realizar tanto de manera prospectiva como retrospectiva. En este último caso, se construye una base de datos que contiene compuestos activos y estos se “diluyen” en una base de datos con *presuntos* inactivos. Este tipo de ensayo se realiza para ajustar los distintos parámetros requeridos en los métodos y en estudios de validación de los mismos. Uno de los principales problemas es que la inactividad de los compuestos se presume, ya que normalmente no se tienen datos de actividad que confirmen dicha inactividad frente a la diana biológica particular. Por otra parte, el ensayo prospectivo está dirigido al descubrimiento real de nuevos *leads*.

## I.1. Pre-filtrado: filtros *Drug-Likeness*

En el primer paso de una cascada de VS, se utilizan filtros generales, inespecíficos de la diana farmacológica, para eliminar aquellas estructuras que posean propiedades de no-fármaco. Es decir, consideran si la molécula está dentro de los estándares de relevancia biológica en cuanto a los grupos funcionales que presenta y sus propiedades físicas (*Drug-Likeness*). Así, se habla y distingue entre compuestos *drug-like* y los *non-druglike.* Sin embargo, estos criterios no son del todo objetivos, de modo que no todos los fármacos actuales satisfacen completamente estos criterios.

Los diversos filtros se establecen a partir del análisis estadístico de bases de datos que incluyen fármacos: *Comprehensive Medicinal Chemistry* (CMC)[11], *MACCS-II Drug Data Report* (MDDR)[12], *World Drug Index* (WDI)[13]; y otras bases de las que se extraen *supuestos* no fármacos, entre la que destaca el *Available Chemical Directory* (ACD)[14].

Entre los diversos filtros establecidos, destacan[15]

i) Establecer márgenes de propiedades. La "regla de los cinco" de Lipinski[16], se considera uno de los primeros pasos del VS para detectar moléculas con una pobre absorción (Figura I.3). 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. Además, normalmente esta regla se extiende con la condición de que el número de enlaces rotables sea inferior a 10. Sin embargo, se ha encontrado que los márgenes de Lipinski son demasiado estrictos y normalmente se aplican valores de corte algo superiores, principalmente en lo referente al peso molecular y a la lipofilia. Otros estudios, como el realizado por Oprea, establecen márgenes de variabilidad de éstos y otros descriptores.[17]

ii) Basados en la presencia de grupos funcionales característicos de fármacos establecidos, se asigna a cada molécula un *score* o puntuación por la presencia de ellos.[18]

iii) Filtros que eliminan grupos funcionales tóxicos o demasiado inestables, como los incluidos en el programa REOS (*Rapid Elimination of Swill*).

iv) Otros estimadores más sofisticados, utilizan, árboles de decisión[19], redes neuronales[19,20] y algoritmos genéticos[21] para clasificar los compuestos de bases de datos como *drug-like* o no. Sin embargo, estos métodos tienen la desventaja de que están muy influenciados por la base de datos utilizada, por lo que es difícil extraer reglas generales útiles para la discriminación.

La inclusión de la predicción 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..., son factores que cada vez se incluyen más en las etapas previas de VS, en un intento de optimizar simultáneamente la potencia y la farmacocinética.[22]

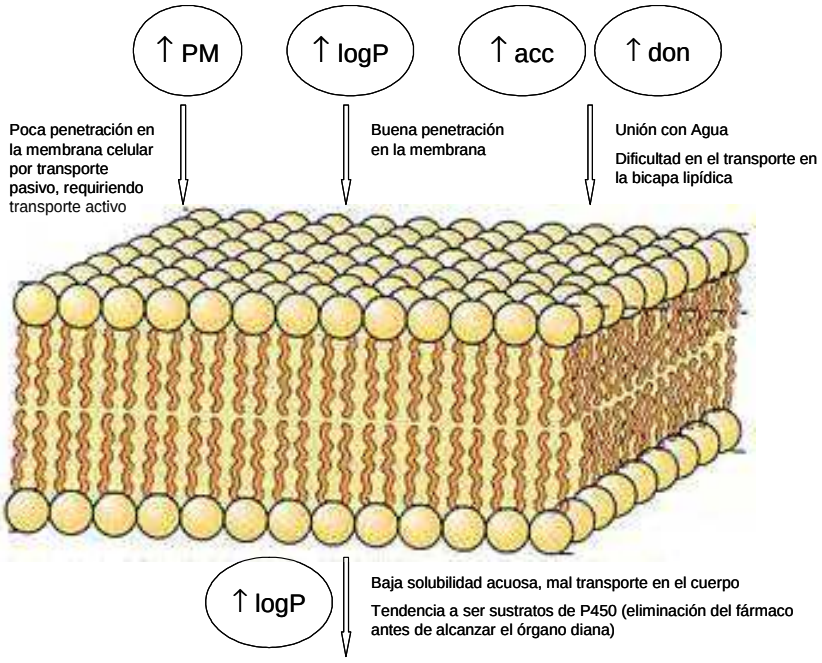



**Figura I.3.** Influencia de las propiedades determinadas en las reglas de Lipinski en la absorción.

## I.2. Cribado virtual basado en ligandos (*Ligand-Based VS*)

La aproximación *ligand-based* se justifica a partir del principio de similitud de Maggiora (*similarity-property principle*), que postula que moléculas estructuralmente relacionadas *deberían* mostrar actividades biológicas similares.[23] Pese a que este criterio permanece no demostrado y existen puntos críticos[24, 25], como el hecho de que a veces pequeños cambios estructurales conducen a un gran cambio en la actividad del compuesto o que moléculas similares a veces muestren modos de unión diferentes, es uno de los criterios centrales en la química médica.

La búsqueda de similitud y diversidad en quimiotecas virtuales, el QSAR y el 3D-QSAR parten del principio de similitud de Maggiora. Estos métodos se han mostrado de gran utilidad cuando no se dispone, o se prescinde, de la información contenida en la estructura del receptor. Dado su bajo coste computacional, frente a los métodos basados en el receptor, se utilizan en los estadios iniciales de las cascadas de VS.

### *Similitud*

La búsqueda de similitud se aplica para diseñar y seleccionar las denominadas quimiotecas focalizadas (*focused libraries*). Este tipo de quimiotecas están orientadas hacia una diana farmacofórica, una clase estructural o un farmacóforo conocido. El diseño de las quimiotecas es cada vez más focalizado a medida se avanza en las etapas de descubrimiento de fármacos, cobrando gran importancia en la fase de optimización de un *lead*.

El procedimiento básico para buscar similitud parte de una o varias estructuras diana (*focus compounds*) y su descripción por uno o más descriptores estructurales, junto con la de los compuestos candidatos contenidos en la quimioteca virtual.[26,27] Así, los dos factores que participan en una búsqueda por similitud son los descriptores utilizados, con su correspondiente peso asignado, y la métrica empleada para establecer la comparación entre pares de moléculas.

Muchos de los descriptores usados en un cribado por similitud proceden de las búsquedas de subestructura (*substructure searching*) en bases de datos. Sin embargo, este tipo de búsquedas únicamente permiten decidir si la subestructura requerida (por ejemplo, un anillo bencénico) se encuentra contenida o no en las estructuras de los compuestos a testar, resultando en una partición binaria del espacio, a no ser que se incluyan otros parámetros. En la búsqueda de similitud, se calcula una medida de similitud entre la estructura diana y cada uno de los compuestos presentes en la base de datos, por lo que posteriormente se pueden ordenar por similitud decreciente. Los primeros de la lista (*nearest neighbours*) se convierten en los candidatos seleccionados por el VS.

Tradicionalmente, los descriptores utilizados para caracterizar quimiotecas virtuales han sido clasificados como *1D*, que 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.[28] Hay alrededor de tres mil descriptores posibles de naturaleza diferente: *número de distintos tipos atómicos*, *fisicoquímicos*: con información de las características estéricas, lipófilas y electrónicas de la molécula tales como la superficie accesible al solvente, el logaritmo del coeficiente de partición octanol-agua, energías HOMO y LUMO, momento dipolar…; *índices topológicos*: calculados a partir de grafos y que codifican información como las estructuras cíclicas, anillos, orden de enlace…; *descriptores basados en fragmentos 2D* : pares atómicos agrupados según tipo de átomo y enlace, relaciones geométricas entre puntos farmacofóricos, búsqueda de grupos funcionales determinados y fragmentos 2D específicos...; y los *basados en fragmentos 3D*, que en muchos casos contienen la misma definición que los correspondientes 2D, aunque en este caso las distancias se miden en el espacio Euclídeo en lugar de tratarse de distancias topológicas.

Otro punto a considerar es la codificación de los descriptores. Aunque normalmente, en los paquetes de *software* cada uno de los posibles tipos se encuentra codificado de una manera particular, la naturaleza y la codificación de cada descriptor son problemas independientes, ya que normalmente es posible codificar un descriptor determinado de diversas maneras.

Los descriptores fisicoquímicos e índices topológicos suelen codificarse en vectores de dimensión constante de valores reales, conocidos como *dataprints*.

Otro tipo de codificación, muy usada con los descriptores basados en fragmentos 2D y 3D, se basa en cadenas de bits de dimensión constante, en las que se indica la ausencia (0) o presencia (1) de una determinada característica, denominados huellas digitales o *fingerprints*. También se pueden usar cada uno de los bits para representar un posible valor de entre un rango de los valores permitidos para variables discretas con varias posibilidades, como el número de ocurrencias, o identificar cada bit con un rango de valores que puede adoptar un descriptor continuo (*binning*).
A su vez, existen tres tipos de construcción de *fingerprints*: i) directos, ii) las llaves estructurales (*structural keys*) o iii) *hashed fingerprints*.[29]

Las llaves estructurales, originalmente desarrolladas para la búsqueda de subestructuras, utilizan un diccionario de fragmentos para asignar cada uno de los bits a un posible fragmento, de manera que se codifica su ausencia o presencia. El principal problema es que la información recopilada en la llave estructural está limitada por el tamaño y tipo de los fragmentos contenidos en el diccionario, por lo que la elaboración de dicho diccionario es la parte clave. Las MACCS *keys*, un subconjunto del set *MDL Information Systems*, son de las más usadas de este tipo.[30]

Para superar esta dependencia y la falta de generalización, se crearon los *hashed fingerprints* para codificar todo tipo de fragmentos o motivos (*patterns*). En lugar de asignar un bit a cada fragmento, se utiliza un algoritmo pseudoaleatorio para codificar cada fragmento, reconocido a partir de un recorrido comprendido entre uno y un número predefinido de átomos conectados en una molécula, en un entero que se traslada a una cadena de bits de tamaño predefinido.

Aunque se reduce la precisión, ya que diferentes fragmentos pueden redundar en un mismo entero, son *fingerprints* más generalizables. Un ejemplo de este tipo de codificación es el desarrollado por *Daylight Chemical Information Systems* Inc. (Daylight)[31] y Tripos Inc. (Unity).[32]

Otro tipo de codificación similar a los *fingerprints* es la basada en vectores de correlación (CV, *correlation-vector*). Este tipo de codificación, introducida por Broto y Moreau a mediados de los '80[33], genera vectores numéricos de dimensión fija a partir de diferentes características moleculares (puntos farmacofóricos o propiedades fisicoquímicas). Los CVs corresponden a histogramas o correlogramas, donde cada columna corresponde a un valor de un rango de distancias entre pares de puntos farmacofóricos (descriptores CATS2D y CATS3D, *Chemically Advanced Template Search*[34]), entre pares de nodos correspondientes a un campo de interacción molecular (descriptores GRIND, *Grind Independent Descriptors*[35]) o entre pares de descriptores fisicoquímicos (electronegatividades, polarizabilidades atómicas y cargas parciales)[36].
La principal ventaja de este tipo de codificación es que los descriptores generados no requieren el alineamiento explícito de las moléculas para ser comparadas (*alignment-free*), lo cual agiliza los cálculos, principalmente si se compara con la obtención de modelos farmacofóricos (véase abajo). Además, también hay que tener en cuenta que la superposición de moléculas, en el modo en que se supone que actúan sobre el receptor, no es trivial. En la Figura I.4 se esquematiza el proceso de derivación de *fingeprints* farmacofóricos y su correspondiente correlograma.

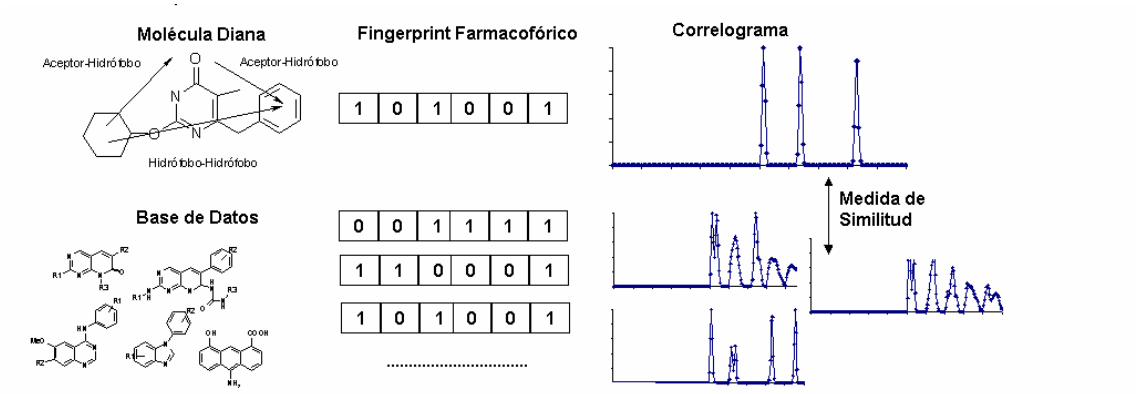



**Figura I.4.** Esquema de derivación de *fingerprints*, correlograma y búsqueda de similitud en una base de datos a partir del correlograma obtenido para la molécula diana (*focus*).

Referente a la medida de similitud, pese a que se han propuesto y comparado diferentes coeficientes de similitud y distancia[26], no existe un criterio unitario ni una definición exacta de similitud. Así, las diferentes métricas se comportan mejor o peor en función del conjunto de descriptores utilizado y de las moléculas a comparar. Esta falta de consenso se traslada también a los descriptores. Existen diferentes estudios dirigidos a establecer una combinación descriptores/coeficiente óptima para la búsqueda de similitud[29] o criterios para la validación de dichos descriptores.[37]

Esta falta de definición de similitud y su medida, es otro de los puntos controvertidos del principio de Maggiora, ya que la lectura "moléculas similares" ha de trasladarse apropiadamente a "moléculas representadas químicamente de manera similar", cosa no trivial.

Diversos programas comerciales que calculan descriptores moleculares son: MOE[38], Cerius2-Descriptor+[39], DRAGON[40], Molecular Modeling Pro[41] y ChemOffice/ChemSAR [42].

Existen aplicaciones prospectivas donde el cribado virtual basado en similitud ha identificado *leads*, en este caso mediante los descriptores farmacofóricos CATS, para bloqueadores de canal de calcio[43], antagonistas del receptor purinérgico (A2A)[44] e inhibidores de la quinasa Glicógeno Sintasa 3 (GSK-3)[45].

### *Obtención de Farmacóforos*

Cuando se dispone de una serie de compuestos activos, la identificación de modelos farmacofóricos es otra de las técnicas estándar para el diseño de quimiotecas focalizadas.

La derivación de modelos farmacofóricos parte de la aproximación del activo análogo (*Active Analog Aproach*), cuyo objetivo primario es 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.[46]

Usualmente, el proceso para derivar un modelo farmacofórico parte del alineamiento de estas 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 dicho modelo, se puede utilizar para buscar en bases de datos otras moléculas que contengan el mismo farmacóforo, para explicar relaciones de estructura-actividad o como punto de partida para el diseño de nuevas moléculas potencialmente activas.

Uno de los problemas asociados a la construcción del modelo farmacofórico, es el tratamiento de la flexibilidad molecular de los activos de partida (plantillas) y su superposición. En este sentido, aparte de la generación manual de hipótesis (*pharmacophoric queries*), como la implementada en el programa MOE[38], se han desarrollado programas para derivar automáticamente hipótesis, basados en superposiciones y alineamientos múltiples (DISCO[47], CATALYST[48], GRASP[49], ALADDIN[50]). Sin embargo, no seleccionan una única mejor propuesta, sino que sigue siendo necesaria la intervención del usuario. Estos programas difieren entre sí en los algoritmos usados para el alineamiento y en el tratamiento de la flexibilidad molecular.

Otro punto interesante es el de la conservación y tolerancia de grupos farmacofóricos. Tradicionalmente, los métodos de identificación obligan a que todas, o un número definido por el usuario, de las características farmacofóricas esté presente en todas o parte de las moléculas alineadas. La incorporación de una estrategia para relajar la tolerancia se contempla como un modo de introducir el concepto de lógica difusa (*fuzziness*) en la generación de modelos farmacofóricos.

En un segundo paso, durante el cribado en bases de datos, la necesidad de alinear las moléculas frente a la hipótesis farmacofórica seleccionada, supone otra desventaja de las técnicas tradicionales de obtención de farmacóforos.

Recientemente, el grupo del profesor Schneider ha desarrollado la metodología SQUID (*Sophisticated Quantification of Interaction Distributions*)[51]. Ésta, permite establecer un nexo entre los modelos farmacofóricos tradicionales y el VS basado en búsqueda de similitud con descriptores farmacofóricos codificados como vectores de correlación. Las principales ventajas de la metodología SQUID son i) la inclusión de información difusa (*fuzzy*) sobre la conservación y la tolerancia en un conjunto de moléculas activas y ii) la codificación de la información en descriptores independientes de alineamiento, aumentando así la eficacia del VS. En el apartado 1.7.2 se detalla en profundidad esta metodología.

El VS basado en modelos farmacofóricos es uno de los métodos que con más éxito ha descubierto *leads* para diferentes dianas biológicas.[52-54] Destaca el uso masivo del programa CATALYST[48], comercializado por Accelrys, y que además de incorporar el módulo HipHop para derivar modelos farmacofóricos, contiene el módulo HypoGen, que utiliza datos cuantitativos de actividad para establecer la hipótesis farmacofórica.

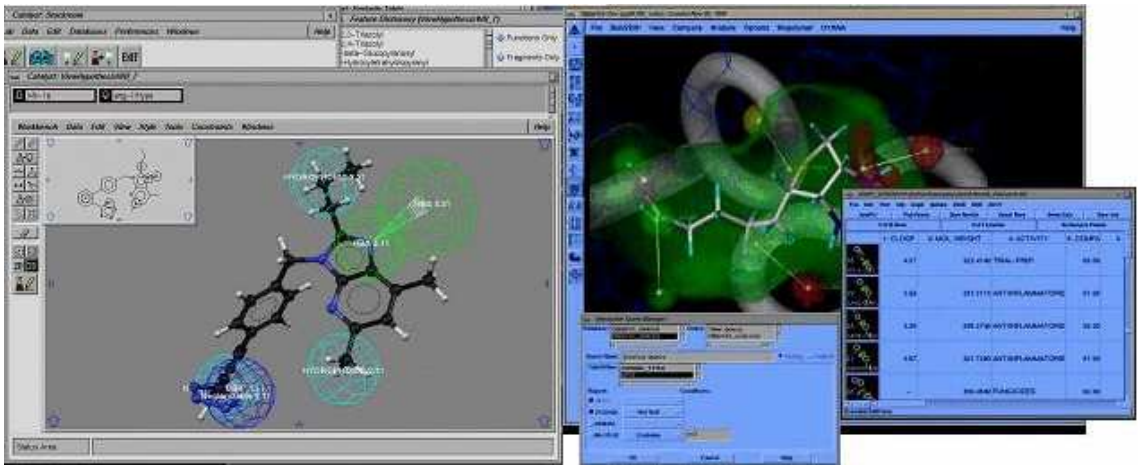


**Figura I.5.** Programas que integran el cálculo de farmacóforos. De izquierda a derecha: CATALYST y Sybyl.

## *QSAR y 3D-QSAR*

Las relaciones estructura-actividad 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 *partial least squares* (PLS). 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. Cuando se mide la interacción entre una sonda protón y la función de onda de la molécula, se habla del potencial electrostático molecular (*Molecular Electrostatic Potential*, MEP).
Los métodos más usados son el CoMFA[55] (*Comparative Molecular Field Analysis*) y GRID[56]/GOLPE[57], que se diferencian principalmente en los MIP que derivan. CoMFA utiliza un MIP estérico y otro electrostático, mientras que en GRID/GOLPE se puede calcular el MIP de distintas sondas químicas implementadas en el programa GRID.

## I.3. Cribado virtual basado en Receptor (*Structure-Based VS*)

La idea de diseñar compuestos a partir de la complementariedad con la estructura del receptor surge a mediados de los 70[58,59] y se generaliza en los 80, como respuesta a la cantidad de estructuras cristalinas de complejos y proteínas resueltas disponibles gracias a los avances en la cristalización de proteínas, la difracción de rayos-X y la resonancia magnética nuclear (RMN). El *Protein Data Bank*[60], creado en 1977, alberga en la actualidad algo más de 34400 estructuras, aunque muchas de ellas corresponden a diversas formas cristalinas de una misma macromolécula, con lo que el número de plegamientos diferentes es mucho menor.

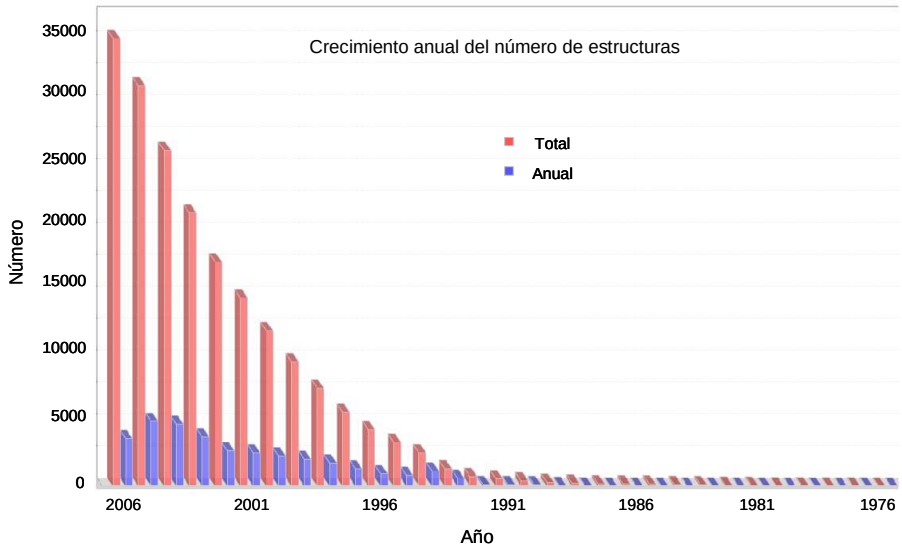



**Figura I.6.** Crecimiento del número de estructuras depositadas en el *Protein Data Bank*. Adaptado de [60].

La estructura del receptor se usa para explorar el espacio químico identificando ligandos de bases de datos de compuestos orgánicos, mediante técnicas de *docking* o bien para diseñar compuestos *de novo* que encajen en el sitio de unión de la proteína.[61]

### *Docking*

Una quimioteca de compuestos orgánicos se posiciona en el sitio de unión y se evalúa la actividad potencial de estos compuestos a partir de la energía de interacción proteína-ligando. Aquellos ligandos con mayor actividad calculada son candidatos a síntesis o pueden comprarse. Este tipo de aproximación es bastante frecuente cuando se dispone de la estructura resuelta de la proteína o bien de un modelo de la misma construido por homología. Así, se han identificado ligandos para más de 50 receptores, tanto de estructura conocida[62-67] como a partir de modelos teóricos [68-69]. Otro de los usos establecidos del *docking* es 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 utiliza para identificar el sitio de unión (*blind docking*)[70].

Un protocolo de *docking* se caracteriza tradicionalmente por dos aspectos: el *docking* en sí mismo, es decir, el 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.[71]
Existen diferentes implementaciones de algoritmos para encontrar configuraciones proteína-ligando (denominadas *poses*) próximas a la conformación nativa del complejo cristalizado (una RMSD inferior a 2 Å es el criterio de aceptación de una *pose* particular)[72].

Actualmente, todos los algoritmos modernos de *docking* modelan el ligando como flexible, dejando de lado las aproximaciones más primitivas en las que el ligando se consideraba rígido (*docking rígido*).

Los métodos más comunes son: *fast shape matching* (DOCK[73], EUDOCK[74], LIGANDFIT[75]), construcción incremental del ligando en la cavidad de la proteína (FLEXX[76], HAMMERHEAD[77]), búsquedas tabú (PRO_LEADS[78], SFDOCK[79]), algoritmos genéticos (GOLD[80], AUTODOCK3.0[81], GAMBLER[82]), algoritmos genéticos acoplados a búsqueda local o Lamarckianos (AUTODOCK3.0), programación evolutiva[83], *simulated annealing* (AUTODOCK2.4[84], GLIDE[85]), métodos de Monte Carlo (MCDOCK[86], QXP[87], ICM-DOCK[88]) y geometría de distancias (DOCKIT[89]). También existen combinaciones de estos métodos.

Los métodos *fast shape matching*, como el implementado en DOCK, caracterizan el sitio activo del receptor mediante esferas, cuyos centros se ajustan a los centros del ligando (átomos pesados o esferas) sobre la base de una comparación de las distancias internas ligando-ligando y receptor-receptor (Figura I.7). Los métodos de construcción incremental del ligando utilizan en muchos casos, como HAMMERHEAD, una caracterización del sitio activo similar a la de los métodos *fast shape matching*. En este caso, acoplan progresivamente fragmentos del ligando que contengan como mínimo dos enlaces rotables, explorando para cada uno de ellos las conformaciones posibles. Los métodos que utilizan algoritmos heurísticos de optimización parten de una o varias conformaciones iniciales, modificando los grados de libertad de rotación y traslación según las particularidades de cada algoritmo. Estos métodos se discuten en términos generales en el apartado 1.9.

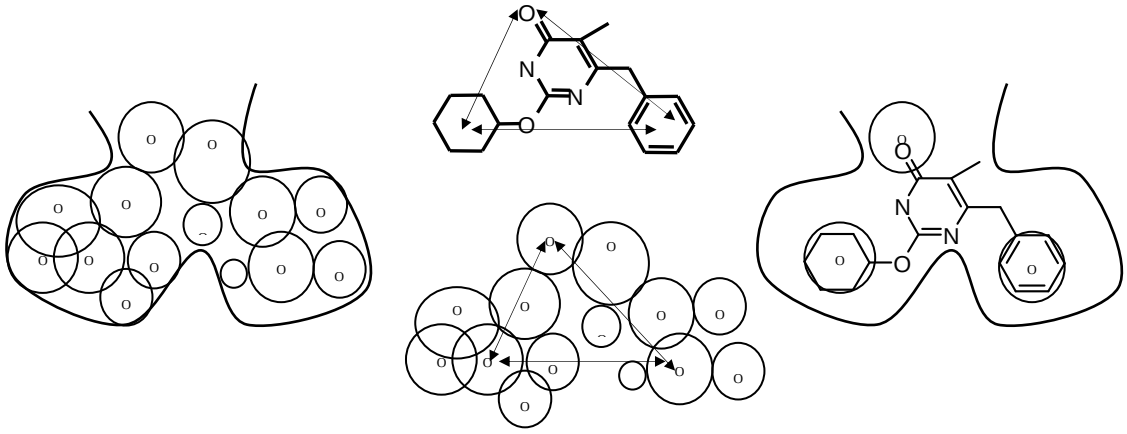



**Figura I.7.** Esquema del método de muestreo en *docking* de los métodos *fast shape matching* (DOCK).

El *docking* es la parte que requiere más tiempo computacional, por lo que los algoritmos que tardan más de tres minutos por ligando por procesador, se consideran demasiado lentos para ser utilizados en VS.

La parte más conflictiva es la función de *scoring* para predecir la afinidad de la unión proteína (o cualquier otra macromolécula)–ligando. Las funciones tradicionalmente aplicadas se clasifican en:[90]

i) Basadas en campos de fuerza (*Force field-Based*): a partir de mecánica molecular, aproximan la energía libre de unión a partir de la suma de interacciones electrostáticas y de van der Waals. Frecuentemente, incluyen también términos empíricos que incluyan la entropía y solvatación. Destacan las funciones DOCK[73] y CHARMm[91].

ii) Empíricas (*Empirical*): estiman la energía libre de unión sumando términos de interacción derivados de la contribución ponderada de parámetros estructurales (número de puentes de hidrógeno, interacciones iónicas, contactos apolares, entropía,…). Los pesos de cada parámetro se obtienen por ajuste a constantes de unión experimentales de un conjunto de complejos proteína-ligando. Las más conocidas son LUDI[92], CHEMSCORE[93], SCORE[94], FRESNO[95], FLEXX[76], PLP[83], AUTODOCK[81] Y GOLDSCORE[80].

iii) *Knowledge-Based*: representan la afinidad como suma de interacciones de pares de átomos proteína-ligando. Estos potenciales se derivan a partir de complejos de estructuras conocidas del *Protein Data Bank*, donde las distribuciones de probabilidad de distancias interatómicas entre diferentes pares de tipos de átomo proteína-ligando se convierten, asumiendo distribuciones energéticas tipo Boltzmann, en funciones de potencial. La energía libre de interacción se calcula sumando las contribuciones de los pares de átomos dentro de una cierta distancia. Destacan PMF[96], DrugScore[97] SMOG[98], BLEEP[99] y SMoG2001[100].

Las funciones empíricas son las más usadas en los programas de diseño de fármacos, aunque no hay ninguna función superior al resto, ya que diferentes funciones se comportan mejor para determinados complejos proteína-ligando. De hecho, debido a la falta de fiabilidad general, normalmente se utiliza una combinación de funciones (*consensus scoring*)[82]. Con ello, se combinan varias funciones y solo aquellas conformaciones (*poses*) que reciben altos *scores* por dos o más funciones de *scoring* son consideradas favorables.

A pesar de que, por fundamento teórico, el *docking* es uno de los filtros más precisos de VS, existen tres grandes problemas asociados a él[101,102]:

i) El gran número de posibles ligandos y sus posibles orientaciones y conformaciones exceden la capacidad computacional.

ii) La flexibilidad se introduce totalmente únicamente en el tratamiento del ligando, mientras que el receptor se considera rígido en la mayor parte de casos. El tratamiento flexible del receptor tiene un coste computacional todavía demasiado alto, por lo que se usan aproximaciones como el introducir movilidad en las cadenas laterales de algunos aminoácidos a partir de librerías de rotámeros (GOLD), uso en simulaciones paralelas de distintos confórmeros de la proteína o la construcción de una geometría difusa que engloba distintas conformaciones (módulo FlexE de FlexX[103], AUTODOCK[104]). Sin embargo, numerosas proteínas muestran fenómenos de inducción (*induced-fit effects*) de las cadenas laterales y cambian de forma y estructura del solvente tras la unión del ligando.

iii) El cálculo de la afinidad proteína-ligando no es, ni muchos menos, exacto, principalmente en lo referente al cálculo de energías de solvatación y a la consideración de cambios en la entropía.

Pese a la reducción del espacio conformacional proteína-ligando muestreado al considerar el receptor rígido, se asume que los algoritmos de *docking* funcionan adecuadamente en esta parte del *docking*[105], siendo la función de *scoring* la parte más débil.

Uno de los problemas de que adolece el VS con *docking*, es la gran cantidad de falsos positivos identificados debido a errores en la medida de afinidad. Se ha comprobado, que las técnicas de *consensus scoring* reducen notablemente esta cifra, tanto en ensayos de VS como en la detección del modo de unión nativo.

Estos falsos positivos proceden en gran parte de ligandos promiscuos ("*frequent hitters*" o "*promiscuos binders*"), y suponen un problema recurrente tanto en el VS como en HTS. Estos compuestos se detectan como *hits* en diferentes resultados de VS y ensayos biológicos dirigidos contra un amplio margen de dianas farmacológicas. Esto sucede por dos razones: 1) la actividad del compuesto no es específica de la diana o 2) el compuesto altera el ensayo o el método de detección. En cualquier caso, estas moléculas no suelen ser válidas como puntos iniciales de los programas de optimización de *leads*.[106]

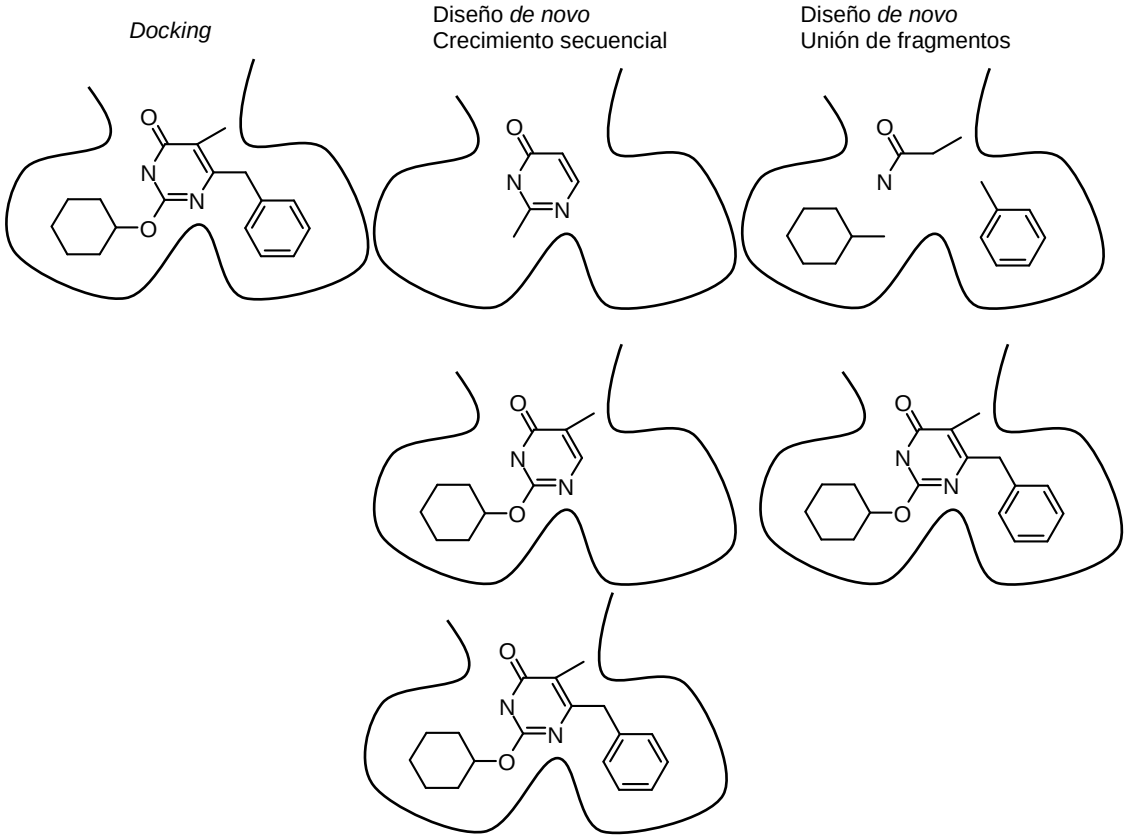



**Figura I.8**. Esquema de *docking*, diseño *de novo* por crecimiento secuencial y por unión de fragmentos (de izquierda a derecha).

## *Diseño de novo*

Estas técnicas permiten diseñar inhibidores/moduladores "*from scratch"* a partir del sitio de unión en la diana o del farmacóforo, es decir, de información de la ordenación espacial de puntos de interacción receptor-ligando relevantes (Figura I.8).
De hecho, los programas de *docking* se pueden utilizar a este propósito si se acoplan con un generador de estructuras, aunque se han desarrollado programas especializados para construir los ligandos dentro del sitio de unión por combinación o ensamblaje de átomos y fragmentos moleculares que se adecuen a los sitos de interacción encontrados.[107]

La generación de un conjunto suficientemente diverso, uno de los problemas originales de estos métodos, no supone actualmente una dificultad, aunque sí lo es el considerar la accesibilidad sintética de los ligandos propuestos.
En función de qué se ensambla se distinguen dos aproximaciones: basadas en átomos (*atom-based methods*), que construyen una molécula átomo a átomo, y en fragmentos (*fragment-based*), que utilizan bases de datos de conectores predefinidos (*building blocks*), conectados por un esquema sintético virtual.
Además, se clasifican en función de cómo es el proceso de ensamblaje en: construcción incremental del ligando (*incremental-growth*) y *construct-and-score*. En el primer caso, se añaden y modifican los fragmentos/átomos, calculando el *score* intermedio, hasta construir la molécula final. En la segunda opción, directamente se construye la molécula completa y se evalúa la afinidad.[108]

Al igual que en los métodos de *docking*, las funciones de *scoring* más comunes son las empíricas y las *knowledge-based*.

Entre los programas más usados destacan LUDI[109], BUILDER[110], CAVEAT[111] y SPROUT[112]. En la referencia [106] se recoge un compendio del *software* destinado a diseño *de novo*.

### *Cálculo de la Afinidad de Unión*

El cálculo de energías libres de unión aparece a comienzos de los ’80, basado en simulaciones de mecánica molecular con dinámica molecular y métodos de Monte Carlo.[113-115] Los dos grandes métodos: el de perturbación de energía libre (*Free Energy Perturbation*, FEP) e Integración Termodinámica (*Thermodynamic Integration*, TI), se presentaron como posibilidades fáciles y fiables. Se basan en que los cambios de energía libre relacionados con pequeñas perturbaciones de un sistema molecular se pueden determinar a partir de una simulación. Así, estos métodos realizan un tratamiento riguroso de todos los grados de libertad de complejos ligando-proteína, incluyendo modelos de solvatación adecuados. A partir del ciclo termodinámico de la Figura I.9, se calcula la diferencia de energía libre entre dos procesos (unión de dos ligandos distintos, *X* e *Y* a la proteína *P*) a partir de introducir mutaciones/perturbaciones que transforman el ligando *X* en *Y*. Así, la necesidad de calcular las ramas horizontales del ciclo se sustituye por la de calcular las ramas verticales del ciclo, es decir, la transformación de *X* e *Y* en entorno acuoso y en la proteína. Durante la mutación gradual, se generan especies químicas inexistentes.

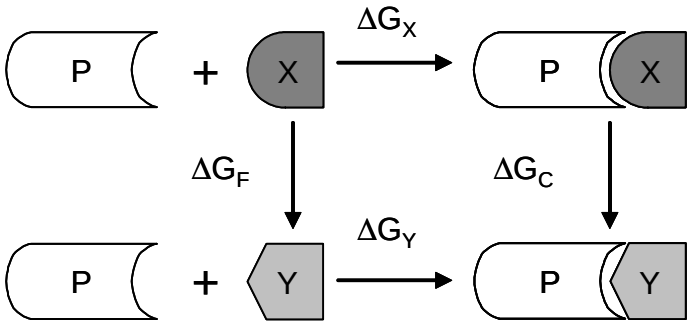



**Figura I.9**. Ciclo Termodinámico considerado en el método FEP. Adaptado de [2].

Hasta los ’90, no se disponía, de manera generalizada, de la capacidad de cálculo para realizar la predicción de manera correcta. Actualmente, también es demasiado costoso computacionalmente para la aplicación al cálculo de miles de compuestos en experimentos de VS, a lo hay que sumarle el hecho de tener que calcular sobre estructuras inexistentes y restringirlo a ligandos muy similares, por lo que han quedado relegadas del mundo del VS.[116]

En el otro extremo, se sitúan las funciones de *scoring* aplicadas en los métodos de *docking*. Como se ha comentado, estas funciones son demasiado simples en su evaluación, ya que están diseñadas para el tratamiento de miles de compuestos.

En un intento de combinar precisión y rapidez, recientemente se han desarrollado varias aproximaciones. Destacan el método LIE[117] (*Linear Interaction Energy*) y el método MM-PBSA[118] (*Molecular Mechanics Poisson-Boltzmann Surface Area*).
El método LIE asume que la energía libre de unión de un ligando a un receptor es la combinación lineal de unas energías ponderadas de interacción electrostática y de van der Waals. Los pesos asignados a cada término son parámetros empíricos.
En MM-PBSA, la energía libre de un sistema se evalúa a partir de la combinación de mecánica molecular, una estimación de la energía electrostática mediante Poisson-Boltzmann, un término de energía de solvatación calculado a partir del área de superficie accesible y un término entrópico. Más detalles de este cálculo aparecen en el apartado 1.4.2.

Al igual que FEP y TI, la evaluación se realiza sobre un conjunto de conformaciones o *snapshots* obtenidos con dinámica molecular. Las funciones de *scoring* utilizadas en *docking* realizan el cálculo a partir de una única conformación o *pose*. En este sentido, MM-PBSA y LIE requieren más tiempo de cálculo que éstas, aunque son más asequibles que FEP y TI.

En las primeras aplicaciones publicadas, estos métodos se han aplicado para el cálculo de energías libres de unión de un reducido conjunto de moléculas (oscilando en torno a 10-20 moléculas), para las que el modo de unión está bien establecido a partir de estructuras cristalográficas de complejos o bien para extraer conclusiones estructurales de conformaciones preferentes e isomería.[119-124]
Sin embargo, recientemente se ha validado el uso de MM-PBSA en VS.[125] En lugar de realizar una dinámica molecular, se evalúa una única conformación del complejo proteína-ligando mediante MM-PBSA. Esta aproximación, aunque controvertida con conclusiones de otros autores[121], deja la puerta abierta a la inclusión de estas técnicas en VS.

### *Modelización de Proteínas*

Como se ha comentado, para utilizar métodos de VS directos, cuando no se dispone de una estructura resuelta de la proteína, se pueden utilizar modelos teóricos, normalmente obtenidos por modelización por homología (*Comparative Modeling, Homology Modeling).*

La diferencia entre el número de secuencias de proteína conocidas y el número de estructura resueltas es cada vez mayor, ya que la secuenciación crece a un ritmo exponencial y la velocidad de determinación estructural no se incrementa a este ritmo. De hecho, aunque se pudiera resolver la estructura de todas estas proteínas, se ha estimado que el tiempo necesario para ello sería de unos quinientos años.[126] Sin embargo, el número de plegamientos estructurales que una proteína adopta es limitado[127], y se prevé que en menos de diez años se tendrá una estructura resuelta, como mínimo, representante de la mayor parte de tipos de plegamiento.[128]

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). Esta aproximación se fundamenta en el hecho de que un pequeño cambio en la secuencia de una proteína, normalmente resulta en un pequeño cambio en su estructura. Así, la condición para modelar por homología es que exista suficiente similitud (entorno al 30-40% de similitud de secuencia se considera el límite inferior[129]) entre la secuencia diana y la(s) secuencia(s) de la plantilla(s). Actualmente, esta técnica es el mejor método de predicción de modelos, ya que es el único que puede predecir estructuras con una exactitud comparable a la obtenida para estructuras a baja resolución con rayos-X.[130]

Quizás el programa de uso más general es MODELLER[131], tanto en su versión gratuita académica, como en su distribución comercial en el paquete INSIGHT[132]. Además, existen varios servidores Web, en los que únicamente es necesario enviar la secuencia de la proteína diana y la estructura de la(s) plantilla(s), como son el servidor SWISS-MODEL[133], CPHmodels[134] y WHAT IF[135], por citar algunos.

Los otros dos grandes métodos de predicción, *ab initio* y por reconocimiento de plegamiento, se utilizan principalmente para la descripción y predicción de fenómenos estructurales de las proteínas, ya que todavía están en desarrollo.

Los métodos *ab initio*[136] 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. Estos métodos se implementan para la modelización de *loops* posterior a la homología, como es el caso del MODELLER, aunque también existen programas íntegros *ab initio*, como el servidor ROSETTA.[137]

Finalmente, 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. Destaca el programa THREADER[138] que ajusta la estructura primaria a plegamientos contenidos en una base de datos y la evalúa mediante funciones desarrolladas estadísticamente. (Figura I.10).

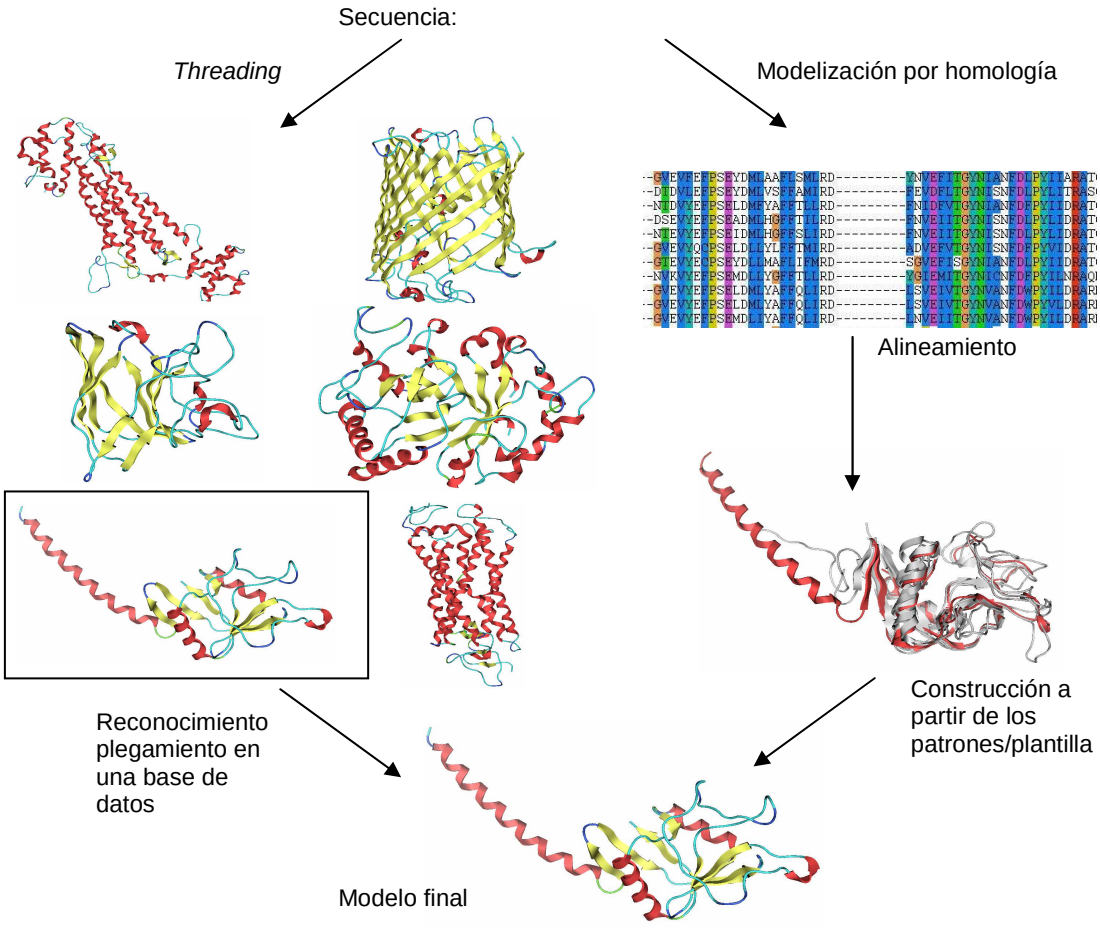



**Figura I.10.** Esquema de los métodos de modelización por homología y *threading*.

### *Definición del Sitio de Unión*

Los métodos de *docking* están optimizados para encontrar el modo de unión, pero no están dirigidos, en principio, a determinar el sitio de unión.
Cuando se dispone del receptor complejado con diversos ligandos, la definición del sitio de unión es fácil, a partir de los residuos comprendidos dentro de una distancia umbral (*cutoff*) desde el ligando. Sin embargo, cuando únicamente se dispone de la estructura tridimensional de la apoproteína (sin ligandos), conviene disponer de información como la función de la proteína o la derivada de experimentos de mutagénesis dirigida.

Existen programas que intentan identificar cavidades en la superficie de la proteína mediante algoritmos denominados *flood-filling*. Básicamente, rellenan el espacio que no está ocupado por la proteína con puntos y eliminan aquellos que no serían "borrados" al deslizar una esfera de un determinado radio por la superficie de la proteína.[139] Además, se han desarrollado otros métodos para priorizar la localización del sitio de unión cuando éste no es una cavidad.[140]

## I.4. Combinación de métodos basados en estructura y en ligandos

A menudo se combinan ambas aproximaciones, la basada en el receptor y la basada en ligandos, de manera que se intentan superar las limitaciones particulares de cada uno de ellos. No únicamente a través de la comparación/complementación de los resultados obtenidos por cada uno de ellos, sino también incorporando la información procedente de uno en la metodología del otro.

Una posibilidad es la de introducir información en el *docking* acerca del modo de unión al receptor, extraída de complejos co-cristalizados con otros ligandos, de la cavidad de la proteína, de átomos prueba o de grupos funcionales. En estos casos se habla de *docking* dirigido directamente (*direct guided-docking*).[141] Normalmente, se puede aceptar que el modo de unión se conserva entre distintos ligandos, aunque no siempre esta afirmación se cumple, como se verá en ciertos casos en este trabajo. Esto permite reducir en gran parte la búsqueda conformacional y eliminar aquellos ligandos para los que la unión sería físicamente imposible (por ejemplo, si se sabe que la interacción se establece a través de un aceptor de puente de hidrógeno, se puede prescindir de intentar realizar el *docking* de una molécula que carezca de grupos aceptores). La introducción de restricciones de interacción se ha mostrado útil en el VS de los receptores acoplados a la proteína G (GPCR), una familia particularmente complicada ya que, a excepción de la rodopsina bovina, no se dispone de la estructura resuelta de ellos, por lo que se trabaja sobre modelos construidos por homología.[142]

Por otra parte, se pueden reconocer modelos farmacofóricos por complementariedad a la estructura tridimensional del sitio activo de un receptor, especialmente si se dispone de complejos ligando-proteína co-cristalizados. Tras el análisis del sitio activo, se genera un mapa de interacción de grupos farmacofóricos deseables (dadores de puente de hidrógeno, aceptores de puente de hidrógeno y sitios lipofílicos) que el ligando debería satisfacer. Con dicho mapa de interacción, se generan varias hipótesis para cribar una quimioteca virtual. Este procedimiento, ha sido integrado en el módulo SBF (*Structure-Based Focusing*) del programa Cerius2.[143] Wang y colaboradores identificaron inhibidores de la proteasa del HIV-1 mediante farmacóforos basados en la estructura del receptor.[144]

En otros casos, se utilizan perfiles de interacción proteína-inhibidor para realizar la búsqueda en bases de datos. Las mejores configuraciones obtenidas por *docking* de todas las moléculas de la base de datos en la estructura de un receptor se trasladan a un *fingerprint* de interacción estructural (*Structural Interaction Fingerprint*, SIFt).
Por otra parte, a partir de un conjunto de complejos receptor-inhibidor, se genera un perfil de interacción, que codifica la probabilidad de encontrar una determinada interacción en una determinada posición (*profile structural interaction fingerprint*, pSIFt). Finalmente, se ordenan los compuestos por similitud decreciente entre los SIFt y los pSIFt.[145]

## I.5. Diversidad

Hasta este punto, la exposición se ha centrado en la estrategia del cribado virtual hacia el diseño de quimiotecas focalizadas a una diana farmacológica en particular. Sin embargo, otra alternativa del diseño de quimiotecas es la selección de un conjunto basándose en la diversidad de los compuestos que la componen, de manera que el diseño final sea representativo de la quimioteca inicial total, disminuyéndose la probabilidad de que existan regiones inexploradas.

Este criterio de diversidad se suele aplicar a quimiotecas generales de compuestos con el fin de identificar un mayor número de *scaffolds* diferentes, por lo que se aplica en las etapas iniciales de descubrimiento de *hits*. Estas quimiotecas (*diversity library* o *random library*) están orientadas a ser testadas frente a un amplio rango de dianas biológicas. Este tipo de cribado va particularmente unido a la química combinatoria: ante la posibilidad de sintetizar en paralelo miles de compuestos, es necesaria una selección racional de éstos o bien de los reactivos que aportarán un determinado sustituyente en una determinada posición de manera que el subconjunto escogido maximice la variabilidad de las propiedades moleculares de los productos. Estas selecciones de carácter más general, no contemplan información estructural de los inhibidores conocidos, por lo que son de utilidad para la identificación de posibles *hits* cuando no se dispone de la información requerida en los métodos descritos anteriormente.

Más allá de la identificación de nuevos *hits*, la competitividad y la presión por explorar, tan pronto como sea posible, grandes regiones de espacio químico cuantificadas en términos de diversidad química, motivan también la aplicación de este enfoque a la hora de seleccionar posibles candidatos a ser sintetizados o a complementar los catálogos disponibles en una empresa.

La selección de grupos de compuestos diversos con la intención de cubrir un mayor espacio químico de actividad y así descubrir islas de actividad, cobró más importancia que el diseño focalizado en los primeros tiempos de introducción de la química combinatoria. Sin embargo, no se cumplieron las expectativas de identificación de nuevos *hits*, por lo que se comenzó a complementar el criterio de diversidad con la optimización de requisitos estructurales, ganando cada vez más importancia el diseño focalizado. Algunas empresas utilizan aproximaciones mixtas en las que se selecciona un pequeño conjunto de compuestos basándose en diversidad y una vez analizados y establecidas las tendencias, se realiza un diseño focalizado para la selección de nuevos compuestos.[146]

La diversidad o disimilitud es el complemento de la similitud, por lo que las medidas de diversidad se efectúan sobre el espacio químico definido sobre los mismos tipos de descriptores y métricas que los utilizados en las búsquedas de similitud.

Para seleccionar un conjunto diverso de moléculas, representativo de todo el espacio químico, existen a grosso modo tres grandes aproximaciones: i) basada en análisis de conglomerados o *clusters*, ii) métodos de partición (*partition methods*) y iii) los métodos basados en distancias o disimilitud (*dissimilarity-based methods*) (Figura I.11).[147, 148]

i) En el las técnicas de *clustering*, las moléculas se agrupan de manera que aquellas pertenecientes a un mismo *cluster* compartan un alto grado de similitud entre sí y sean distantes de las situadas en otros *clusters*. Seleccionando moléculas pertenecientes a cada uno de los grupos, se obtiene la máxima representatividad del espacio químico. Por otra parte, si se desea focalizar la selección a un compuesto, se escogen aquellas moléculas incluidas en el *cluster* de dicho compuesto.

ii) Los métodos de partición también clasifican el espacio químico para posteriormente seleccionar un candidato de cada grupo, pero en este caso lo hacen a partir de celdas (*bins*) generadas por división recursiva de los rangos de todas las propiedades que describen el espacio químico. Este tipo de aproximación es mucho más rápida y requiere menos recursos de memoria que los métodos de *clustering*, por lo que se aplican en quimiotecas de compuestos de tamaño medio y grande.

iii) Finalmente, en los métodos basados en distancias, los compuestos se escogen, normalmente mediante algoritmos heurísticos, de manera que sean lo más disimilares a los ya seleccionados.

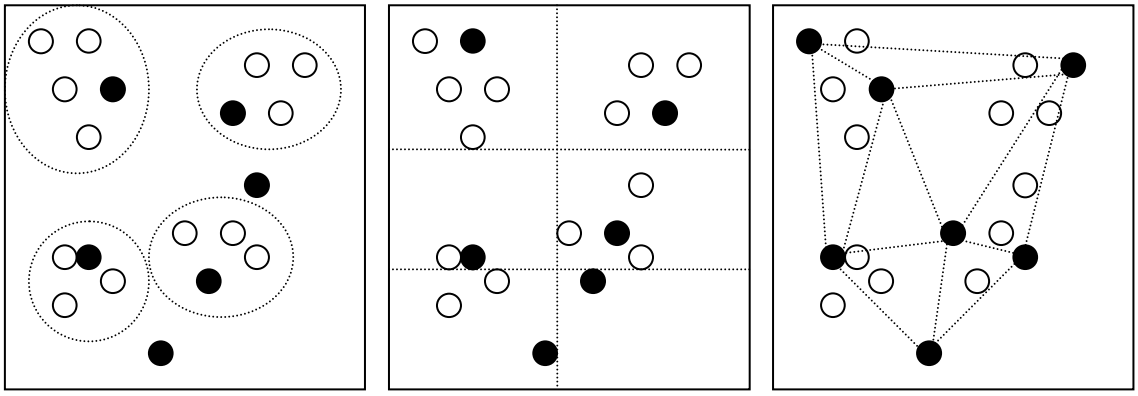


**Figura I.11.** Esquema de selecciones basadas en *clustering*, *bins* y métodos basados en distancia, de izquierda a derecha.

## I.6. Quimiotecas Combinatorias

La selección de compuestos, tanto en versión focalizada como diversa, se puede aplicar a bases de datos generales, como las colecciones propias de una empresa o catálogos públicos como el ACD, *Available Chemicals Directory*, o a quimiotecas virtuales combinatorias, es decir, donde se han generado todas las posibles combinaciones de productos a partir de un número de reactivos, tal y como se obtendrían sintéticamente por química combinatoria (Figura I.12).

Previo a la selección en estas quimiotecas virtuales, éstas se tienen que construir. Para ello, se tiene que considerar la elección de una química accesible, es decir, la elección del espacio químico de interés. Desde un punto de vista sintético, las reacciones multicomponente (MCR) permiten la combinación de tres o más puntos de diversidad, con lo que se facilita la construcción de quimiotecas combinatorias grandes con una amplia variedad de funcionalidades químicas.

Los reactivos se extraen de catálogos de casas comerciales o de bases de datos generales como el ACD y son sometidos a filtros similares a los aplicados en las etapas de pre-filtrado de productos del VS. Además, se incluyen factores como el precio, la accesibilidad comercial de dichos reactivos y las posibles interferencias que puedan generar en la reacción química establecida. Dada la falta de bases de datos que recojan aquellos reactivos no aptos para una determinada reacción, este último criterio se suele más bien realizar basándose en intuición y conocimientos sintéticos que con el uso de filtros automatizados.[149]

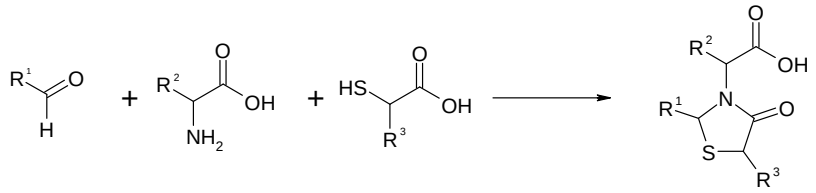


**Figura I.12.** Esquema de una quimioteca combinatoria con tres puntos de variación. La combinación de $N_1$ aldehídos con $N_2$ aminoácidos y $N_3$ tioles genera una quimioteca de $N_1 \times N_2 \times N_3$ productos.

Tanto en bases de datos generales como en quimiotecas combinatorias, se puede aplicar una selección *cherry picking* o *sparse array*, esto es, seleccionando *n* productos de los *N* totales de manera que cumplan el criterio de diversidad o similitud requeridos, pero sin imponer una restricción combinatoria sobre los reactivos de los que proceden, en el caso de trabajar sobre quimiotecas combinatorias. Este tipo de selección, presenta, aunque no necesariamente, el inconveniente de que se incrementa el número de reactivos necesarios y con ello el coste. El número mínimo de reactivos necesarios para sintetizar *n* productos en una reacción *k*-componente es $k \cdot n^{1/k}$. El número máximo corresponde a $k \cdot n$, al que se tiende en el diseño *cherry picking*. Además, en este diseño se generan problemas en la robotización de la síntesis combinatoria.[150]

En las quimiotecas combinatorias, además, pueden aplicarse otras dos estrategias. La primera de ellas, basada en reactivos (*reagent-based*), selecciona directamente un conjunto de reactivos de cada uno de los puntos de variación disponibles, basándose en lo que Gillet bautizó como hipótesis de diversidad (*diversity hypothesis*).[150] Dicha hipótesis asume que si es posible identificar un conjunto de reactivos de máxima diversidad, entonces su uso resultará en la generación de una quimioteca combinatoria de productos diversos. Así, supone que las propiedades derivadas de los reactivos son transferibles, para ciertos descriptores, a los productos. Con ello, el conjunto seleccionado es combinatorio, evitándose los inconvenientes de la selección *cherry picking*. Al prescindir de la construcción virtual o enumeración de todos los productos de la quimioteca, es menos costosa computacionalmente, pero se ha demostrado que esta simplificación es menos eficaz en la selección de conjuntos diversos que la aproximación desarrollada posteriormente, la basada en productos (*product-based*).[151,152]

El diseño *product-based full array*, se ideó para superar las desventajas de los otros dos formatos comentados: pérdida de representatividad de los productos (*reagent-based*) y formato no combinatorio del conjunto escogido (*cherry picking*). La selección se realiza sobre el espacio de los productos, pero de manera que sean la combinación de un subconjunto de reactivos. En este caso, el número de reactivos requeridos tiende al valor mínimo de $k \cdot n^{1/k}$.

En la Figura I.13, se esquematizan las tres alternativas en la selección de un conjunto de 16 compuestos de una quimioteca combinatoria de 49 productos, resultado de la reacción de *4 x* con *4 y* reactivos. Se destaca la diferencia en la necesidad de enumeración: selección (*product-based*) ó selección-enumeración (*reagent-based*) y la naturaleza combinatoria: (*full array*) o no (*cherry picking*) del conjunto seleccionado (*en rojo*).

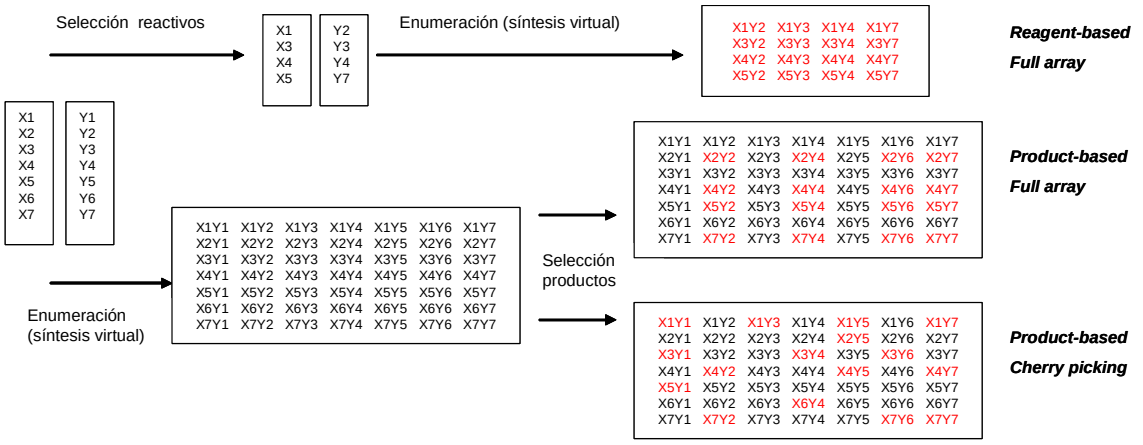



**Figura I.13.** Esquema de selecciones basadas en reactivos (*reagent-based*) frente a las basadas en productos (*product-based*) y de selecciones *cherry picking* frente a *full array*.

La selección *full array* es un problema combinatorio, tipo NP-completo. Las posibilidades de selección para una quimioteca con *k* puntos de variación que presentan $N_i$ reactivos, para la que se desea escoger $n_i$ reactivos vienen dadas por $\prod_{i=1}^{k}\binom{N_i}{n_i}$. Así, para una quimioteca con 4 pasos o puntos de variación con 10 reactivos asequibles en cada uno de ellos, para los que se desea escoger 3 reactivos, el número de selecciones posibles es de $10^8$. La naturaleza combinatoria de las selecciones *cherry picking* es mucho mayor, ya que las posibilidades de selección de *n* productos de un total de *N* corresponden al número combinatorio $\binom{N}{n}$. Comparativamente, en el caso de la selección anterior de 81(3^4) reactivos de un total de 10000 (10^4), el número de selecciones posibles es de $10^{203}$. Mientras que la mayoría de métodos *cherry picking* tienen un carácter determinista, la naturaleza de la restricción combinatoria obliga al uso de técnicas de optimización.

En la última década, han surgido una variedad de referencias que proponen distintos algoritmos de optimización, tanto heurísticos: algoritmos genéticos (programas SELECT[153], GALOPED[154], HARPick[155]…)[156-159] y *simulated annealing*[160,161] como aproximaciones más deterministas[162-164]. Destacan aquellas que eliminan la necesidad de construir/enumerar toda la quimioteca de compuestos, ya que de manera iterativa seleccionan subconjuntos de reactivos a partir de los cuales generan productos hasta encontrar el óptimo.[163]
Además, existen diferentes paquetes integrados dirigidos a la construcción y selección de quimiotecas combinatorias: el módulo CombiChem de Cerius2[39], el paquete Sybyl[165] y MOE[38,] son algunos ejemplos.

En particular, en el Grupo de Ingeniería Molecular, GEM, en el IQS, se está desarrollando el programa PRALINS (*Program for Rational Analysis of Libraries in Silico*). En este programa, iniciado por R. Pascual, se han implementado y adaptado una gran parte de las metodologías de selección y algoritmos de optimización de quimiotecas diversas, tanto en formato *cherry picking* como *full array*.[166]

Otra de las direcciones hacia las que ha evolucionado el diseño de quimiotecas virtuales ha sido hacia la selección de compuestos combinando múltiples criterios como diversidad/similitud, coste, propiedades ADMET, etc[149,167] (selecciones multiobjetivo).

# Objetivos

El presente trabajo se encuentra enmarcado en la línea de investigación del Grupo de Ingeniería Molecular, GEM, en el IQS, dirigida hacia el diseño, síntesis y evaluación de inhibidores potenciales de tirosina quinasas, particularmente el receptor del factor de crecimiento epitelial (EGFR), el receptor del factor de crecimiento de fibroblastos (FGFR) y el receptor del factor de crecimiento derivado de plaquetas (PDGFR). En el marco de esta línea y en el contexto del diseño molecular, el objetivo principal es establecer y validar un protocolo de evaluación de compuestos con potencial actividad inhibidora de tirosina quinasas que permita priorizar los candidatos a ser sintetizados. Dicho protocolo incluye varias de las estrategias expuestas en la introducción, jerarquizadas en función de su requerimiento computacional. Éstas son:

- Diseño y selección de quimiotecas combinatorias basadas en criterios de diversidad.
- Aplicación de filtros *ligand-based*: búsquedas de similitud y farmacóforos.
- Aplicación de filtros *structure-based* basados en *docking*.
- Implementación de un nuevo *fingerprint* de interacción estructural como herramienta en el postprocesado de *docking* y aplicable como filtro en un cribado virtual.
- Evaluación de la afinidad de unión de una serie de inhibidores de tirosina quinasas mediante MM-GBSA/MM-PBSA. Aplicación de esta metodología en el cribado virtual.

Paralelamente, se continúa desarrollando el programa PRALINS. Por un lado, dado el creciente interés hacia el diseño de quimiotecas focalizadas, se exploran e implementan diversas estrategias de selección en formato *full array*. Asimismo, se revisan los criterios de evaluación de diversidad y se habilita la posibilidad de realizar selecciones multiobjetivo basadas en algoritmos genéticos.

Finalmente, se estudia el modo de unión de antagonistas de los receptores CXCR4 y CCR5, co-receptores implicados en la entrada del virus del HIV a las células.

# Capítulo 1.

# Fundamentos teóricos

## 1.1. Modelización Molecular

Los modelos teóricos empleados en la química computacional para estudiar la estructura y reactividad de las moléculas, se distinguen típicamente en *modelos cuánticos*, basados en mecánica cuántica (QM) y *modelos clásicos* derivados de mecánica molecular (MM).

- En la química cuántica, la distribución electrónica se incluye explícitamente mediante su codificación en la función de onda ($\Psi$), relacionada con la energía ($E$) a través de la ecuación de Schrödinger independiente del tiempo:

  $$H\Psi(r) = E\Psi(r) \qquad [1.1]$$

  Donde el operador hamiltoniano ($H$) incluye la energía cinética y potencial de núcleos y electrones. Para resolver esta ecuación, es necesario introducir diversas aproximaciones (Born-Oppenheimer, combinación linear de orbitales atómicos…). No contemplan ningún tipo de parametrización empírica "externa", por lo que también se conocen como métodos *ab initio*.

- Los métodos semiempíricos, en los que sí hay una parametrización empírica para la descripción de los electrones internos (*core*) mientras que los electrones externos se caracterizan mediante funciones de onda cuánticas.

- Los métodos de mecánica molecular consideran la molécula como un conjunto de esferas (átomos) conectados mediante muelles (enlaces), cuyo movimiento se puede describir por las leyes de la física clásica a través de funciones de energía potencial. La simplificación más importante de estas funciones de energía potencial es que sólo consideran los núcleos de los átomos y no existe un tratamiento explícito de los electrones (éstos están considerados implícitamente en los enlaces). El tratamiento de átomos y enlaces se define con los campos de fuerza o *force field*, en el que se incluyen los parámetros y ecuaciones que los describen.

Por otra parte, se han desarrollado también modelos mixtos (QM/MM) que tratan el sistema parcialmente de forma cuántica y clásica.

Aunque los modelos cuánticos son más precisos, su elevado coste computacional los restringe a moléculas con un número de átomos del orden de decenas, resultando inviable el tratamiento cuántico total de macromoléculas. Por otra parte, su uso es obligado en el estudio de reacciones que impliquen la ruptura y formación de enlaces. Asimismo, la modelización de los compuestos de quimiotecas virtuales se realiza básicamente en el entorno de la mecánica clásica, aunque existen aplicaciones de descriptores mecanocuánticos a quimiotecas con un número limitado de compuestos.[168]

Otro tipo de sistemas lo componen los métodos basados en reglas o "*rule-based systems*" que permiten obtener una estructura tridimensional razonable para compuestos orgánicos a partir de la información topológica de las moléculas, expresada mediante una tabla de conexiones.

Para ello, utilizan bases de datos tabuladas para las longitudes de enlace, ángulos, conformaciones de anillos a la par que extienden al máximo los fragmentos acíclicos. Destacan los programas CONCORD[169] y CORINA[170]. El objetivo de estos programas es el de acelerar al máximo la generación de estructuras tridimensionales de compuestos en bases de datos.

### 1.1.1. Mecánica Molecular

El *force field* define los parámetros usados en la descripción de los átomos y enlaces y el tratamiento matemático que los relaciona. Así, en primer lugar asigna a cada átomo (bola) un tipo (*atom type*) en función de su hibridación, carga y átomos a los que está unido. A cada uno de los tipos atómicos les corresponde un grupo de parámetros: constantes de fuerza, datos atómicos (radios atómicos, carga, masa…) y valores estructurales de equilibrio. Estos parámetros se suelen obtener a partir de valores experimentales o bien se derivan de cálculos mecanocuánticos.
Finalmente, 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. Así, 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, ángulo, torsiones…

Aunque la ecuación matemática varía entre distintos *force fields*, de manera general se incluyen los siguientes términos:

- *Interacciones no enlazantes*: intervienen átomos no unidos de manera directa por enlaces. Comprenden los términos electrostáticos y de interacciones de van der Waals.

    - La interacción electrostática se calcula según la ley de Coulomb, a partir de las cargas parciales (*partial charges*) asignadas a cada átomo, en las que se aproxima el efecto de la distribución electrónica. Existen diversos métodos de cálculo de cargas parciales: desde una aproximación topológica basada en los átomos y en cómo están unidos, como las cargas Gasteiger-Marsili[171], al método RESP (*Restrained ElectroStatic Potencial*)[172] que ajusta a cada uno de los átomos la distribución del potencial electrostático molecular, calculado a nivel *ab initio* (HF/6-31G*). El método RESP, aunque más caro computacionalmente que las aproximaciones topológicas, es más refinado desde un punto de vista teórico. Recientemente, se ha desarrollado el método AM1-BCC[173], con el objetivo de reproducir la precisión del método RESP a la par que disminuir el tiempo de cálculo. Del análisis de la distribución de cargas en el hamiltoniano AM1 se realizan correcciones aditivas sobre el enlace (*additive bond charge corrections*, BCCs).

    - Las interacciones de van der Waals vienen tradicionalmente dadas por el potencial electrostático de Lennard-Jones 12-6 (véase ecuación [1.2]), aunque también se pueden formular con otros exponentes en el coeficiente de la interacción repulsiva, como 9 o 10, la decimosegunda potencia se prefiere por la facilidad de cálculo (el cuadrado de la sexta potencia).

- *Interacciones enlazantes*: intervienen átomos unidos por enlaces químicos. Se trata de los términos de estiramiento de enlace, doblamiento de ángulos, ángulos diedros y ángulos impropios (en sistemas planares de cuatro átomos, en los que uno de ellos en posición central está unido al resto). Además, se pueden incluir términos de interacciones cruzadas que reflejan el acoplamiento entre las coordenadas internas: acoplamientos ángulo-enlace, ángulo-ángulo, enlace-enlace…

En la Figura 1.1 se esquematizan los principales términos junto con la representación gráfica de la ecuación que los representa. La vibración de los enlaces y ángulos se modela a partir de un potencial harmónico cuadrático, como en el caso de los muelles. Las torsiones (propias e impropias) se representan mediante funciones periódicas, ya que la rotación atraviesa barreras periódicas. Finalmente, los términos no enlazantes son funciones de potencia inversa de la distancia.

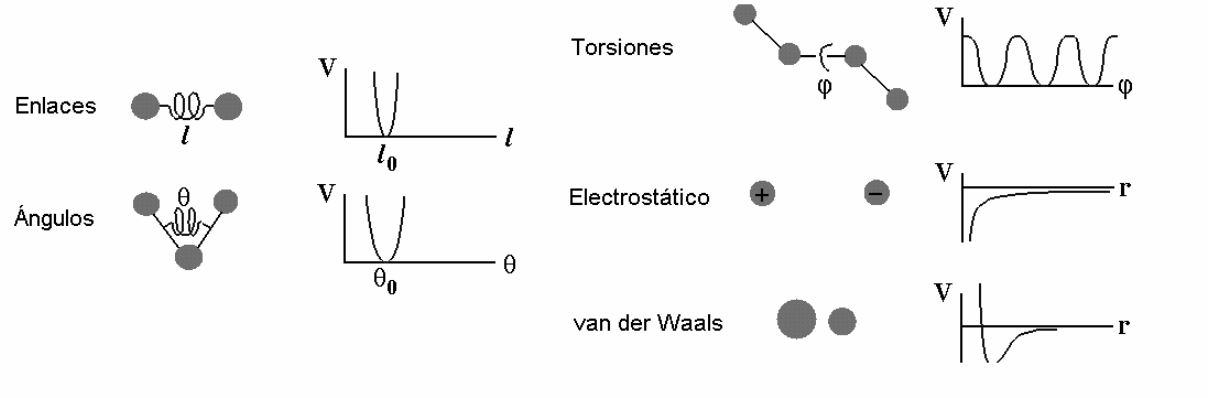



**Figura 1.1.** Modelo y representación gráfica de los términos habituales en un *force field.*

Existe una gran variedad de *force fields* creados en función de los grupos de moléculas empleados como referencia en la parametrización y a los que va destinado. Desde los aplicados a moléculas orgánicas pequeñas y medianas (MM2[174], MM3[175], MM4[176], TRIPOS[177], MMFF94[178], UFF[179], GAFF[180]) a los dirigidos a macromoléculas (AMBER[181], CHARMM[182], GROMOS[183], OPLS[184]). En la referencia [185] se puede encontrar una revisión de los distintos *force fields* aplicados a proteínas.

En el presente trabajo, se ha empleado básicamente los *force fields* AMBER, GAFF y MMFF94, por lo que se detallan sus correspondientes funciones.

La ecuación [1.2] corresponde al *force field* AMBER, donde los dos primeros términos penalizan el alejamiento de los enlaces y ángulo de su valor de equilibrio ($r_o$ *y* $\theta_o$, respectivamente) mediante un potencial harmónico simple (con constantes de fuerza $k_r$ y $k_\theta$, respectivamente). El potencial de torsión se representa mediante una serie de Fourier truncada, donde $V_n$ es el potencial en el máximo, $n$ es la periodicidad y $\phi_o$ es la fase. Los términos no enlazantes se calculan según las ecuaciones tradicionales comentadas.

$$E_{pot} = \sum_{enlaces} k_r (r - r_o)^2 + \sum_{ángulos} k_\theta (\theta - \theta_o)^2 + \sum_{diedros} \frac{V_n}{2}\left[1 + \cos(n\phi - \phi_o)\right] + \sum_{\substack{no\ enlazantes \\ van\ der\ Waals}} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right] + \sum_{\substack{no\ enlazantes \\ electrostáticas}} \frac{q_i \cdot q_j}{4 \cdot \pi \cdot \varepsilon \cdot r_{ij}} \qquad [1.2]$$

La ecuación del *force field* AMBER original[186] incluye también términos de interacción por puente de hidrógeno que se han eliminado en posteriores versiones implementadas en las versiones 7 y 8 del programa AMBER[187].

En dicho programa, a partir de la versión 7, se ha implementado también el *force field* GAFF (*General Amber Force Field*)[180], diseñado para ser compatible con el *force field* AMBER, dirigido a proteínas y ácidos nucleicos. GAFF contiene parámetros para la mayor parte de moléculas orgánicas típicas en química médica, compuestas por hidrógeno, carbono, nitrógeno, oxígeno, azufre, fósforo y halógenos.

Finalmente, el *force field* MMFF94, desarrollado en Merck, está dirigido a un amplio rango de sistemas químicos de interés farmacológico.

$$E = \sum_{enlaces} 143.9325 \cdot \frac{k_r}{2} \cdot (r - r_o)^2 \cdot \left[1 - 2 \cdot (r - r_o) + \frac{7}{3} \cdot (r - r_o)^2\right] +$$

$$\sum_{ángulos} \left| \begin{array}{l} \sum_{ángulos} 0.043844 \cdot \frac{k_\theta}{2} \cdot (\theta - \theta_o) \cdot [1 - 0.4 \cdot (\theta - \theta_o)] \\ 143.9325 \cdot k_\theta \cdot [1 + \cos(\theta - \theta_o)] \end{array} \right| +$$

$$\sum_{\substack{ángulos-enlaces \\ no-lineales}} 2.5121 \cdot \left[k_r \cdot (r - r_o) + k_{r'} \cdot (r' - r'_o)\right] \cdot (\theta - \theta_o) + \sum_{\substack{centros \\ tricoordenados}} 0.043844 \cdot \frac{k_{i\angle}}{2} \cdot \omega_{i\angle}^2 +$$

$$\sum_{diedros} 0.5 \cdot [V_1 \cdot (1 + \cos\phi) + V_2 \cdot (1 - \cos 2\phi) + V_3 \cdot (1 + \cos 3\theta)] +$$

$$\sum_{\substack{no\ enlazantes \\ van\ der\ Waals}} \varepsilon_{ij} \cdot \left[\left(\frac{1.07 \cdot r^*}{r + 1 - 0.7 \cdot r^*}\right)^7 \cdot \left(\frac{1.12 \cdot r^{*7}}{r^7 + 0.12 \cdot r^{*7}} - 2\right)\right] + \sum_{\substack{no\ enlazantes \\ electrostáticas}} \frac{332.071 \cdot q_i \cdot q_j}{\varepsilon_r \cdot (r_{ij} + \delta)^n}$$

[1.3]

La ecuación [1.3] corresponde a la función de energía potencial del *force field* MMFF94. Se observa cómo es más compleja en la definición de los términos que la del AMBER. En general, los *force field* derivados para macromoléculas son los más sencillos en cuanto a la complejidad de las funciones y no suelen incluir términos de interacciones cruzadas, como la del *force field* MMFF94 (término ángulos-enlaces no lineales).
Los términos de estiramiento del enlace y torsión de ángulos se modelan en este caso con una expansión hasta el cuarto orden de la curva de Morse, que se ajusta mejor al modelo de la curva de energía potencial de un enlace que la de la ley de Hooke. El término de torsión de diedros contiene tres términos, a diferencia de AMBER; cada uno de los cuales está dirigido a la explicación de un efecto físico (por ejemplo, el segundo término, refleja el carácter de doble enlace para explicar efectos de conjugación en alquenos). Halgren, autor del *force field*, propuso la forma que adopta la interacción de van der Waals en un intento de mejorar, principalmente, el término de las interacciones repulsivas ($r^{-12}$) de la ecuación tradicional de Lennard-Jones. Finalmente, las interacciones electrostáticas contienen en este caso una constante de *buffering* electrostático ($\delta$) y *n* adopta valores de 1 o 2.

Para disminuir el número de grados de libertad de los sistemas moleculares, se utilizan frecuentemente modelos *united-atom*, en los que los hidrógenos no polares se omiten y los parámetros de interacciones no enlazantes de los átomos a los que está unido se consideran implícitamente en los átomos a los que está unido. En el tratamiento de biomoléculas esto permite reducir el esfuerzo computacional. La función empírica de AUTODOCK incluye un modelo *united-atom* en el tratamiento de la proteína.

## 1.2. Minimización Energética / Optimización Geometría

Una vez introducido el modelo teórico de aproximación a las moléculas, se describen los métodos que permiten buscar las soluciones a la ecuación de la energía potencial en función de las coordenadas atómicas, espacio conocido como superficie de energía potencial. Dentro de esta superficie, son especialmente interesantes los estados estacionarios, en los que la derivada de la energía respecto a las coordenadas (fuerzas) es nula. Particularmente, los mínimos energéticos corresponden a estados estables del sistema. Así, hablar de optimización de

geometría para encontrar esta estructura estable es equivalente a hablar de minimización energética.

Tanto en este caso particular, como en cualquier otro problema de optimización, para moverse por el espacio de búsqueda se pueden adoptar dos posturas diferentes: *explorar* (generando puntos en zonas del espacio que previamente no tienen porqué haber sido visitadas) o *explotar* (explorando también, pero en la cercanía de soluciones ya existentes, sacándoles todo el partido posible). La mayor parte de los algoritmos de búsqueda tratan de establecer un equilibrio entre explotación y exploración, aunque muchos de ellos se inclinan hacia una mayor exploración (aleatoriedad) o explotación (determinismo). Así, en general, los métodos de búsqueda se dividen, a grandes rasgos, en métodos *globales* y *locales*. Los métodos globales tratan de encontrar el mínimo global de un problema, mientras que los locales se concentran en la vecindad de la solución generada inicialmente, por lo que no tienen ninguna garantía de que el mínimo encontrado sea global.

Los problemas de minimización energética se suelen abordar con métodos de búsqueda local, por lo que se describen brevemente en este apartado[188]. Los métodos de búsqueda global, empleados en otras aplicaciones de la química computacional (selección de compuestos en quimiotecas virtuales combinatorias, análisis conformacional aplicado en las búsquedas farmacofóricas y *docking*, superposición de compuestos…) se describen en el apartado 1.9.

Dentro de los métodos de búsqueda local, para variables continuas, son muy comunes los métodos de descenso, de manera que encuentran el mínimo más próximo al punto inicial. Se distinguen en función del orden de la derivada.

### 1.2.1. Métodos no-derivativos o de orden cero

Únicamente utilizan valores de la propia función. Requieren mucho coste computacional, por lo que suelen aplicarse en combinación con otros métodos de optimización más eficientes. Así, en la optimización geométrica, son útiles al inicio, cuando se parte de una configuración muy energética. El más popular es el método simplex. Se genera un simplex, una figura de $M+1$ vértices interconectados, donde $M$ es la dimensionalidad del problema (función de energía). El sistema inicial corresponde a uno de estos vértices, y el resto de vértices se construyen, por ejemplo, imponiendo un incremento a cada una de las variables (coordenadas) de la función. El simplex se mueve sobre la superficie de la energía potencial mediante una serie de reglas (reflexión, expansión, contracción de los vértices), de manera que se asegura que puede explorar la totalidad de la superficie de energía.[188]

### 1.2.2. Métodos derivativos de orden uno o métodos del gradiente

Además de los valores de la propia función, utilizan su primera derivada (gradiente). Son menos robustos que los anteriores, pero más eficientes y con mayor tasa de convergencia. Son los más empleados en mecánica molecular.

Estos métodos iteran la ecuación [1.4]:

$$x_{i+1} = x_i - l_i S_i \qquad [1.4]$$

donde $x_{i+1}$ es la nueva posición en el paso $i+1$, $x_i$ es la posición previa, $l_i$ es el tamaño de paso y $S_i$ es la dirección de este paso. Los diferentes algoritmos varían en cómo definen esta dirección y este paso. La iteración se repite hasta que la variación en la función es menor a un determinado valor umbral.

La longitud del paso se puede determinar con un algoritmo de búsqueda lineal o mediante la aproximación de paso arbitrario.

- La búsqueda lineal localiza el mínimo a lo largo de una dirección especificada (una línea en un espacio multidimensional). Para ello, frecuentemente, se ajustan de manera iterativa funciones polinómicas sobre un conjunto de puntos de la dirección de descenso y se resuelve el mínimo analíticamente. El gradiente en el punto mínimo de la línea de búsqueda es perpendicular a la dirección previa, por lo que el gradiente en la siguiente dirección es ortogonal a la dirección previa.

- En la aproximación del paso arbitrario, el valor del paso tiene un valor predefinido que se incrementa o reduce durante el proceso según si el valor de la función se reduce o incrementa, respectivamente. Este último procedimiento, aunque menos riguroso, suele requerir más pasos para alcanzar el mínimo, pero frecuentemente requiere menos evaluaciones de la función.

Según el modo de escoger la dirección de descenso, destacan:

- *Steepest Descent* (SD) / Dirección del máximo gradiente: La dirección de descenso ($S_i$) corresponde al gradiente ($g_i$) negativo de la función en el punto (ecuación [1.5]).

$$S_i = -g_i / |g_i| \qquad [1.5]$$

  En la optimización geométrica, corresponde a la dirección paralela a la fuerza, con lo que son las mayores fuerzas interatómicas las que determinan la dirección. Así, se trata de un buen método para eliminar rápidamente los peores impedimentos estéricos en una conformación inicial. Sin embargo, en las cercanías del mínimo necesita realizar muchos pasos, ya que al avanzar en direcciones ortogonales a la previa, oscila mucho, reintroduciendo errores ya corregidos en movimientos previos.

- *Conjugate Gradients* (CG) / Gradiente Conjugado: La dirección de búsqueda se establece a partir del gradiente actual y del gradiente del paso anterior (ecuación [1.6]). El conjunto de direcciones generado no es ortogonal y se evita el comportamiento oscilatorio en las cercanías del mínimo, convergiendo más rápido que SD.

$$S_i = -g_i + \gamma_i \cdot S_{i-1} \qquad [1.6]$$

  En función de la relación entre gradientes ($\gamma_i$) se distinguen las diferentes implementaciones del método: Fletcher-Reeves (FR, ecuación [1.7]), Polak-Riviere (PK, ecuación [1.8]) y Hestene-Stiefel (HS, ecuación [1.9]):

$$\gamma_i = \frac{g_i^T \cdot g_i}{g_{i-1}^T \cdot g_{i-1}} \qquad [1.7]$$

$$\gamma_i = \frac{g_i^T \cdot (g_i - g_{i-1})}{g_{i-1}^T \cdot g_{i-1}} \qquad [1.8]$$

$$\gamma_i = \frac{g_i^T \cdot (g_i - g_{i-1})}{S_{i-1}^T \cdot (g_i - g_{i-1})} \qquad [1.9]$$

Generalmente se trabaja con esquemas de minimización, en los que se utilizan combinaciones de estos algoritmos. Así, lo más común es comenzar con una minimización con SD (para eliminar rápidamente los peores impedimentos estéricos) y continuar con CG (para converger rápidamente en un mínimo).

### 1.2.3. Métodos derivativos de orden dos o métodos de Newton

Utilizan las primeras y segundas derivadas de la función. El método Newton-Raphson (NR) es el más simple de ellos. En este caso, se itera la ecuación [1.10], derivada de una expansión en una serie de Taylor en el punto:

$$x_{i+1} = x_i - H_i^{-1}(x_i) \cdot g_i(x_i) \qquad [1.10]$$

Donde $H_i^{-1}(x)$ es la matriz Hessiana inversa. El cálculo de la inversa de esta matriz hace que el método Newton-Raphson requiera más tiempo computacional, por lo que se suele aplicar a sistemas con menos de cien átomos. Además, esta matriz ha de ser definida positiva para impedir que el método se dirija a puntos silla donde la energía se maximiza.

Este ajuste a un modelo cuadrático es más exacto en el mínimo de la función, donde la aproximación harmónica se cumple. Lejos del mínimo la aproximación harmónica es pobre, por lo que la minimización puede volverse inestable. Por ello, NR se aplica normalmente en las cercanías del mínimo, donde previamente se han utilizado métodos más robustos como el método simplex o *Steepest Descent*.

Los métodos Quasi-Newton, para disminuir el tiempo de cálculo de la matriz hessiana, construyen gradualmente la inversa de la hessiana en iteraciones sucesivas a partir de los valores de la función y su gradiente en los puntos previo y nuevo.

## 1.3. Simulación: Dinámica Molecular

La minimización energética no es un método apropiado para explorar un gran número de estructuras de baja energía de macromoléculas. Además de obtener modelos tridimensionales, se aplica como paso previo a los estudios de simulación, como la dinámica molecular. Las simulaciones permiten generar un conjunto de configuraciones representativas de sistemas de los que extraer propiedades estructurales y termodinámicas.

La dinámica molecular[189,190] (MD) permite además estudiar el comportamiento del sistema en función del tiempo, al simular la dinámica del mismo mediante la integración de las ecuaciones de Newton del movimiento para cada átomo. Es un método determinista ya que el estado del sistema en un tiempo posterior se puede predecir invariablemente a partir de su estado actual. Al conjunto de estados accesibles a una molécula se le denomina espacio de fase (*Phase Space*). Se trata de un espacio 6N-dimensional, ya que el estado de un sistema de N átomos queda definido al especificar las 3N coordenadas atómicas y los 3N momentos.

La trayectoria, secuencia de estados resultante en dinámica molecular, se obtiene por integración de la ecuación [1.11] de la segunda ley de Newton:

$$\frac{\delta^2 x_i}{\delta t^2} = \frac{F_{x_i}}{m_i} \qquad [1.11]$$

Donde $m_i$ es la masa de la partícula, $x_i$ es la coordenada y $F_{xi}$ es la fuerza aplicada sobre la partícula en esta dirección. En las funciones de energía potencial la fuerza entre dos átomos o moléculas cambia continuamente con su separación. Dada esta naturaleza continua, la resolución del problema no puede hacerse analíticamente y la integración de las ecuaciones [1.12] y [1.13] se realiza mediante un método de diferencias finitas.

$$r_i(t_2) = r_i(t_1) + \int_{t_1}^{t_2} \frac{p(t)}{m} dt \qquad [1.12]$$

$$p(t_2) = p(t_1) + m\int_{t1}^{t_2} a(t)dt \,^{o} \qquad [1.13]$$

### 1.3.1. Métodos de Integración

Las integrales de las ecuaciones [1.12] y [1.13] se descomponen como suma de pequeñas etapas, cada una correspondiente a un pequeño intervalo de tiempo $\Delta t$ (típicamente, comprendido entre 1 y 10 femtosegundos). En cada paso, se calculan las fuerzas sobre los átomos, asumiéndose que son constantes durante este intervalo de tiempo, y se combinan con las posiciones y velocidades actuales para generar el nuevo estado. Una vez se han movido los átomos a las nuevas posiciones, se actualizan las fuerzas que actúan sobre cada átomo y así hasta generar toda la trayectoria.

Uno de los más conocidos es el algoritmo de Verlet[191]. Su idea básica es escribir dos aproximaciones en serie de Taylor truncadas en el tercer orden para las posiciones del paso nuevo ($t+\Delta t$, ecuación [1.14]) y el previo ($t-\Delta t$¸ ecuación [1.15]):

$$r(t+\Delta t) = r(t) + v(t)\Delta t + (1/2)a(t)\Delta t^2 + (1/6)b(t)\Delta t^3 + O(\Delta t^4) \qquad [1.14]$$

$$r(t-\Delta t) = r(t) - v(t)\Delta t + (1/2)a(t)\Delta t^2 - (1/6)b(t)\Delta t^3 + O(\Delta t^4) \qquad [1.15]$$

Donde $v$ es la velocidad, $a$ la aceleración y $b$ la tercera derivada de las coordenadas respecto al tiempo. Al sumar las dos expresiones se obtiene:

$$r(t+\Delta t) = 2\cdot r(t) - r(t-\Delta t) + a(t)\Delta t^2 + O(\Delta t^4) \qquad [1.16]$$

donde la aceleración se obtiene mediante la ecuación [1.17], a partir de la derivada de la función de energía potencial respecto a las coordenadas:

$$a(t) = -(1/m)\nabla V(r(t)) \qquad [1.17]$$

Se trata de un algoritmo exacto (el error de truncación del algoritmo es del orden de $\Delta t^4$), estable, de fácil implementación y con un coste computacional modesto, lo que explica su gran popularidad en las simulaciones de dinámica molecular.

Su principal problema es que las velocidades no se generan directamente, y aunque no son necesarias para obtener la trayectoria, sí lo son para calcular la energía cinética ($K$) del sistema. Se pueden calcular a tiempo $t$ (ecuación [1.18]) o a mitad del intervalo de tiempo, $t+1/2\Delta t$, (ecuación [1.19]):

$$v(t) = \frac{r(t+\Delta t) - r(t-\Delta t)}{2\cdot\Delta t} \qquad [1.18]$$

$$v(t+\frac{1}{2}\cdot\Delta t)=\frac{r(t+\Delta t)-r(t)}{\Delta t} \quad [1.19]$$

Sin embargo, el error asociado a esta expresión es del orden de $\Delta t^2$, en lugar de $\Delta t^4$. Otro problema es que para inicializar el algoritmo se necesita una alternativa para obtener las posiciones del paso previo ($r(-\Delta t)$). Una posibilidad es realizar la aproximación en serie de Taylor truncado tras el primer paso (ecuación [1.20]):

$$r(-\Delta t)=r(0)-\Delta t\cdot v(t) \quad [1.20]$$

Para superar estas dificultades, se han desarrollado variantes del algoritmo de Verlet que generan exactamente la misma trayectoria, aunque difieren en las variables almacenadas en memoria (posición en paso actual y previo, aceleración en paso actual para Verlet) y los tiempos para los que se calculan.

El algoritmo *Leapfrog*[192], implementado en el módulo SANDER del programa AMBER, calcula explícitamente las velocidades a la mitad del intervalo de tiempo mediante la ecuación [1.21]:

$$v(t+\frac{1}{2}\cdot\Delta t)=v(t-\frac{1}{2}\cdot\Delta t)+\Delta t\cdot a(t) \quad [1.21]$$

y a partir de ellas calcula la posición en el siguiente intervalo de tiempo (ecuación [1.22]):

$$r(t+\Delta t)=r(t)+\Delta t\cdot v(t+\frac{1}{2}\cdot\Delta t) \quad [1.22]$$

Las velocidades a tiempo *t* se calculan con la ecuación [1.23]:

$$v(t)=\frac{1}{2}\cdot\left[v(t+\frac{1}{2}\cdot\Delta t)+v(t-\frac{1}{2}\cdot\Delta t)\right] \quad [1.23]$$

Una de sus desventajas es que las velocidades y posiciones no están sincronizadas, por lo que no se posible calcular la energía cinética (velocidades) al mismo tiempo que la energía potencial (coordenadas), aunque se mantiene la conservación de la energía incluso a intervalos de tiempo mayores.

Existen otros algoritmos de integración como el *velocity Verlet*[193], que obtiene todas las magnitudes sincronizadas aunque con mayor coste de memoria. El *predictor-corrector* de Gear[194] y métodos de Runge-Kutta calculan las velocidades y coordenadas con mayor precisión (utiliza un truncamiento de la serie de Taylor a mayor orden) aunque con mayor coste computacional y requisitos de memoria.

### 1.3.2. Intervalo de tiempo de integración (*Time Step*)

El intervalo de tiempo de integración (*time step*) se establece como un compromiso entre simular la trayectoria "correcta" y cubrir adecuadamente el espacio de fase.
Por una parte interesa tener valores pequeños de *time step*, cuanto más pequeño más se parece la trayectoria de la simulación al resultado de la integración analítica, sin embargo esto conlleva el aumento considerable del coste computacional o la reducción del espacio de fase muestreado. Con *time steps* grandes los átomos pueden colapsar, ocupando las mismas coordenadas

espaciales, se producen violaciones en la conservación de la energía total y del momento lineal o fallos del programa por desbordamiento numérico.

Normalmente, se asume que el límite superior del *time step* es aquel que permite simular bien el movimiento más rápido del sistema (la vibración de un enlace de un átomo de hidrógeno, del orden de 10 fs). Así, típicamente el *time step* es de 1 fs, al menos un orden de magnitud más pequeño que dicho movimiento. Cuando se trabaja a temperaturas por encima de 300 K, el *time step* se suele reducir ya que las energías cinéticas son superiores y los átomos recorren más distancia entre dos evaluaciones de fuerza, pudiendo generar solapamientos de alta energía entre átomos.

Una posible solución para incrementar este valor es el de eliminar del sistema aquellos grados de libertad de mayor frecuencia como lo son los estiramientos de enlace, ya que tienen un efecto mínimo en el comportamiento general del sistema. Para ello, se "congelan" dichas frecuencias al imponer *constraints* sobre estos enlaces (algoritmo SHAKE[195]), permitiendo trabajar con *time steps* de 2 fs.

### 1.3.3. Condiciones de la Dinámica

Siguiendo la hipótesis ergódica, las simulaciones calculan las propiedades macroscópicas como promedio de un conjunto de microestados denominado colectivo (*ensemble*). Las dinámicas moleculares normalmente se realizan bajo condiciones de número constante de partículas (*N*), volumen (*V*) y energía (*E*), conocido como colectivo microcanónico (*microcanonical o constant NVE ensemble*). Sin embargo, se pueden realizar bajo otros colectivos: canónico (*canonical ensemble, NVT*) con número de átomos, volumen y temperatura constantes o el isotermo-isobárico (*isothermal-isobaric ensemble, NPT*). Los resultados de propiedades macroscópicas derivadas de mecánica estadística y obtenidos en un colectivo pueden ser transformados a otro colectivo, aunque estrictamente esto es correcto en el límite de un sistema infinitamente grande.

El primer paso en una dinámica parte de establecer el estado inicial. La conformación inicial puede extraerse de datos experimentales o de modelos teóricos obtenidos con minimización energética. Las velocidades iniciales de los átomos se asignan aleatoriamente de forma que sigan una distribución Maxwell-Boltzmann a la temperatura de interés.

En la *fase de equilibrado* se monitorizan diversas propiedades (energía cinética, potencial y total, temperatura, presión) hasta que se estabilizan. Una vez en equilibrio, es en la *etapa de producción* en la que se muestrea el espacio de fases y se calculan las propiedades macroscópicas.

#### 1.3.3.1. Escalado de la Temperatura

Las dinámicas se realizan a una temperatura determinada. La etapa de producción suele realizarse en colectivo NVE, en el que la temperatura es una variable, pero antes de ello, se suele llevar al sistema a la temperatura de interés, bajo un colectivo NVT, en el que se realiza un re-escalado de la misma para asegurar su constancia.
La temperatura del sistema está relacionada con la energía cinética promedio mediante la ecuación [1.24]:

$$\langle \kappa \rangle = \frac{3}{2} \cdot N \cdot k_B \cdot T \qquad [1.24]$$

Una manera sencilla de mantener la temperatura constante es multiplicar las velocidades tras cada paso de integración por un factor $\lambda$ que relaciona la temperatura actual ($T_a$) con la requerida ($T_r$), ecuación [1.25]:

$$\lambda = \sqrt{T_r / T_a} \qquad [1.25]$$

Sin embargo, el factor de escalado más común procede de acoplar el sistema a un baño calefactor externo fijado a la temperatura de interés, conocido como algoritmo de Berendsen[196] o *weak-coupling* (ecuación [1.26]):

$$\lambda = 1 + \frac{\Delta t}{\tau} \cdot \left( \frac{T_{baño}}{T(t)} - 1 \right) \qquad [1.26]$$

Un mayor valor de la constante de acoplamiento $\tau$, permite un acoplamiento débil entre el baño y el sistema, por lo que se permite fluctuar al sistema entorno a la temperatura objetivo. Estos dos algoritmos únicamente aseguran que la energía cinética total es apropiada a la temperatura de trabajo, pero no que la temperatura esté igualmente distribuida entre todas las partes de la molécula, por lo que la aproximación no es estrictamente canónica. En condiciones de solvente explícito (véase apartado 1.3.6.2), las colisiones entre átomos pueden ayudar a mantener una distribución apropiada de la temperatura entre solvente y soluto, aunque también se puede llegar a una situación en la que la temperatura del soluto sea inferior a la del solvente (‘*hot solvent, cold solute*’).

Además del algoritmo de Berendsen, AMBER incorpora también el esquema de acoplamiento de Andersen o *stochastic collisions*[197], en el que a una partícula, aleatoriamente seleccionada cada un cierto número de pasos, se le reasigna una velocidad aleatoria que cumpla la distribución de Maxwell-Boltzmann a la temperatura deseada. En el período entre colisiones el colectivo es microcanónico, con energía constante, de manera que si la tasa de colisiones es muy baja, el sistema no se comporta realmente como canónico. Si la tasa es excesivamente alta, se reduce la rapidez de muestreo del espacio de fases.
Este tipo de escalado es aconsejable para dinámicas realizadas en condiciones de solvente implícito (véase apartado 1.3.6.3).

### 1.3.3.2. Escalado de la Presión

Cuando se aplican condiciones periódicas de contorno (véase apartado 1.3.4), es necesario ajustar la densidad del sistema durante el proceso de equilibrado, para terminar de empaquetar correctamente el solvente alrededor del soluto en los límites de la caja periódica y evitar la posible formación de huecos de vacío generados por agregación de moléculas de solvente en condiciones de volumen constante.
Para ello, primero se equilibra la temperatura del sistema a volumen constante (colectivo NVT) y después se continúa con una dinámica a presión constante.

En una simulación en condiciones NVE, la presión fluctúa mucho más (varios cientos de bares) que el resto de magnitudes debido a que está relacionada con el virial, que se obtiene como el producto de las posiciones y la derivada de la función de energía potencial. Este producto ($r_{ij} \partial v(r_{ij}) / \partial r_{ij}$) cambia más rápidamente con la posición de lo que lo hace la energía interna. Sin embargo, el valor promedio a lo largo de muchos pasos puede ser próximo a la presión objetivo.

Del mismo modo que un sistema macroscópico, una simulación en el colectivo NPT isotermo-isobárico mantiene la presión constante cambiando el volumen. La fluctuación en el volumen está relacionado con la compresibilidad isoterma ($\kappa$), según la ecuación [1.27]:

$$\kappa = -\frac{1}{V} \cdot \left( \frac{\partial V}{\partial P} \right)_T \qquad [1.27]$$

El algoritmo de acoplamiento de la presión utilizado en AMBER es del tipo "*weak-coupling*", análogo al de la baño calefactor de Berendsen[196]. Se aplica un "baño de presión" que mantiene la presión constante mediante reescalado del volumen de la caja periódica con un factor λ (ecuación [1.28]):

$$\lambda = 1 - \kappa \cdot \frac{\Delta t}{\tau_P} \cdot (P_{baño} - P(t)) \qquad [1.28]$$

Donde $\tau_P$ es la constante de acoplamiento del baño. Reescalar el volumen con un factor λ es equivalente a reescalar cada una de las coordenadas atómicas multiplicándolas por un factor $\lambda^{1/3}$ (ecuación [1.29]):

$$r'_i = \lambda^{1/3} \cdot r_i \qquad [1.29]$$

Esta expresión puede aplicarse isotrópicamente (aconsejado para solutos disueltos en agua) o anisotrópicamente (en sistemas anisotrópicos como simulaciones de membranas, en los que las tensiones superficiales difieren con la dirección).

### 1.3.4. Límites del Sistema (*boundaries*)

Los sistemas simulados en dinámica molecular, de miles y decenas de miles de átomos, son relativamente pequeños a escala macroscópica, por lo que un elevado porcentaje de los átomos se encuentra rodeado de vacío. Esto genera efectos frontera (*boundary effects*), es decir, desviaciones del comportamiento de los átomos en los límites del sistema respecto a los que se encuentran en el centro, que en el caso límite conducen a la "evaporación" del sistema.

Existen dos alternativas posibles para afrontar este problema:

- El uso de condiciones periódicas de contorno (*periodic boundary conditions*, PBC) en las que se simula un sistema infinito al generar réplicas de la celda del sistema en todas las direcciones (Figura 1.2). En un sistema tridimensional, cada celda tendrá 26 celdas vecinas. Las coordenadas de las partículas en las celdas imagen se obtienen sumando/restando múltiplos enteros de los lados de la caja y de manera que si una partícula de la celda abandona la celda durante la simulación, ésta es reemplazada por una partícula imagen que penetra por el otro lado. Existen diferentes geometrías de celda: cúbica, octaedro truncado, prisma hexagonal, dodecaedro rómbico, cuyo tamaño puede ser fijado por el usuario. El módulo SANDER de AMBER está adaptado para el tratamiento de todas estas celdas, aunque en los módulos de generación (XLEAP) y análisis (PTRAJ) únicamente hay implementadas dos geometrías de celda: paralepípedo rectangular y octaedro truncado.

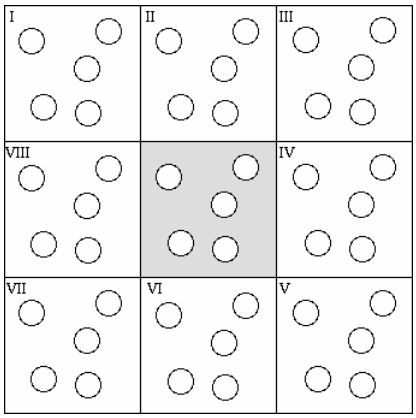



**Figura 1.2.** Condiciones periódicas de contorno.

- *Stochastic boundary conditions* que imponen restricciones al movimiento de los átomos más externos, como generar un "muro" repulsivo o restringir las posiciones de dichos átomos mediante potenciales harmónicos.

Estos últimos métodos son más difíciles de implementar que las simulaciones periódicas y pueden conducir a resultados anómalos, por lo que las simulaciones periódicas PBC siguen siendo el modo más seguro y tradicionalmente recurrido en dinámica molecular.

La elección de tratamiento de los límites va ligada al tipo de solvente aplicado, por lo que se retoma este tema en el apartado 1.3.6, donde se detallan los dos esquemas aplicados en el presente trabajo.

### 1.3.5. Interacciones de largo alcance

En principio, en un sistema de $N$ átomos habría que calcular $N^2$ interacciones no enlazantes entre todos los pares de átomos. En el caso de las interacciones de corto alcance como van der Waals, el potencial de Lennard-Jones decae muy rápidamente con la distancia ($r^{-6}$), por lo que no se justifica el cálculo de dicha interacción para pares de átomos alejados.
En minimización y dinámica molecular, se puede establecer un valor umbral (*cutoff*) por encima del cual no se calculan las interacciones no enlazantes. Éste suele adoptar valores comprendidos entre 8 y 12 Å.
En el caso de simulaciones periódicas, el *cutoff* de las interacciones no enlazantes tiene que ser menor que la mitad de la longitud del lado más corto de la caja, aplicándose el modelo de *minimum image convention*. Así, cada átomo únicamente "ve" como mucho una única imagen de cada átomo del sistema.

En la aplicación de *cutoffs* se utilizan listas de vecinos no enlazados (*non-bonded neighbour list*), en las que se almacenan, para cada átomo, todos los átomos potencialmente vecinos (los situados a una distancia inferior al *cutoff* + átomos que ligeramente sobrepasan esta distancia y que podrían interaccionar por debajo del *cutoff* en pasos sucesivos). Esta lista se actualiza con una determinada frecuencia a lo largo de la dinámica, de manera constante cada 10-20 pasos o mediante un algoritmo automatizado.

Para eliminar las discontinuidades introducidas por el *cutoff* en la función de energía potencial y en las fuerzas en la región de corte, se pueden aplicar *shifted potential*s ó *switching functions*. En las primeras, se desplaza la función de potencial al restársele un término constante (alternativamente también uno lineal). El problema es que al modificar este potencial, las propiedades macroscópicas no son directamente calculables. Las *switching functions* son

funciones polinómicas de la distancia que multiplican a la función de energía potencial, suavizando también el gradiente en la región del *cutoff*.

Sin embargo, la aplicación de *cutoffs* resulta inapropiada para el tratamiento de interacciones de largo alcance como las electrostáticas, que decaen con la inversa de la distancia. Especialmente en sistemas periódicos, aquellas interacciones que decaen no más rápido que $r^{-n}$, donde $n$ es la dimensionalidad del sistema, resultan problemáticas, ya que su rango de interacción es frecuentemente superior que la mitad del tamaño de la celda. Así, se han desarrollado diferentes aproximaciones para el tratamiento de las interacciones de largo alcance: tratamientos del campo de reacción (*reaction fields*), método de los multipolos (*cell multipole method*) y el método de las sumas de Ewald. Éste último es el que está implementado en AMBER en el tratamiento de sistemas periódicos, por lo que se describe brevemente.

#### 1.3.5.1. Método de sumas de Ewald (*Ewald Summation Method*)

En este método[198], una partícula electrostática interacciona no solo con las partículas en la celda de simulación, sino también con sus imágenes en un sistema periódico infinito de celdas, según la expresión de Coulomb correspondiente de la ecuación [1.30]:

$$v = \frac{1}{4 \cdot \pi \cdot \varepsilon_o} \sum_{|n|=0} \left( \sum_{i=1}^{N-1} \sum_{J=I+1}^{N} \frac{q_i \cdot q_j}{|r_{ij} + n|} \right) \qquad [1.30]$$

Donde $N$ es el número de cargas contenido en cada celda, $q_i$ y $q_j$ son dichas cargas, $r_{ij}$ la distancia que las separa y $n$ corresponde a los vectores de una red periódica $n=(n_x \cdot L_x,\ n_y \cdot L_y,\ n_z \cdot L_z)$, siendo $L$ la longitud de cada dimensión de la celda.
La suma de la ecuación [1.30] es condicionalmente convergente (su resultado depende del orden en que los términos son sumados) y tiene una convergencia lenta.

El método de sumas de Ewald, cuyo modelo matemático se muestra en la ecuación [1.31], convierte la suma en dos series, cada una de las cuales converge mucho más rápidamente:

$$\frac{1}{r} = \frac{f(r)}{r} + \frac{1 - f(r)}{r} \qquad [1.31]$$

De este modo, divide la interacción culómbica en un término de corto alcance y otro de largo alcance.

La primera suma, realizada en el espacio real, equivale físicamente a rodear cada carga puntual en el sistema por una distribución neutralizante de cargas de igual magnitud y signo contrario. Esta distribución es típicamente una gaussiana. Este término converge rápidamente y es responsable de las interacciones de corto alcance.
El segundo término compensa la distribución neutralizante del primer término, mediante una distribución imaginaria de cargas de signo opuesto a las del espacio real. Esta suma se realiza en el espacio recíproco y también converge mucho más rápidamente que la suma original. Se trata de una serie que varía muy suavemente con la distancia, por lo que puede aplicarse su transformada de Fourier mediante un número de vectores recíprocos. En la Figura 1.3 se esquematizan las dos distribuciones de carga utilizadas en el método de sumas de Ewald.

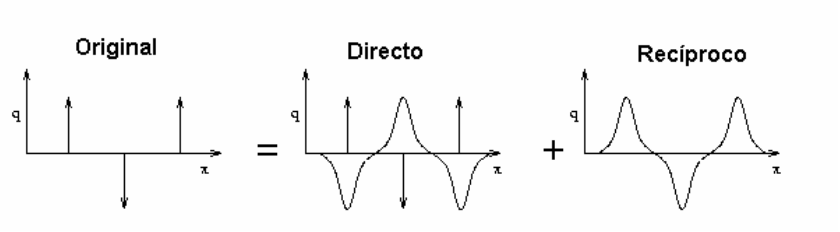



**Figura 1.3.** Distribuciones de carga en el espacio real y recíproco respecto al sistema original utilizadas en el método de sumas de Ewald.

La ecuación [1.32] muestra la energía potencial final obtenida por el método de sumas de *Ewald*, donde el primer y segundo términos corresponden a las sumas en el espacio directo y recíproco, respectivamente

$$\nu = \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left\{\sum_{|n|=0}^{\infty}\frac{q_i q_j}{4\pi\varepsilon_o}\cdot\frac{erfc\,(\alpha|r_{ij}+n|)}{|r_{ij}+n|} + \right.$$
$$\sum_{k\neq 0}\frac{1}{\pi\cdot L^3}\cdot\frac{q_i q_j}{4\pi\varepsilon_o}\cdot\frac{4\pi^2}{k^2}\exp(-\frac{k^2}{4\alpha^2})\cos(K\cdot r_{ij}) - \qquad [1.32]$$
$$\left.\frac{\alpha}{\sqrt{\pi}}\sum_{k=1}^{N}\frac{q_k^2}{4\pi\varepsilon_o}+\frac{2\pi}{3L^3}\left|\sum_{k=1}^{N}\frac{q_k}{4\pi\varepsilon_o}\cdot r_k\right|^2\right\}$$

Donde *erfc* es la función de error complementario (ecuación [1.33]):

$$erfc\,(x)=\frac{2}{\sqrt{\pi}}\int_x^{\infty}\exp(-t^2)\,dt \qquad [1.33]$$

y $K$ son los vectores recíprocos dados por $K = 2\pi n/L$, y $L$ es la dimensión de la celda.

La amplitud de la gaussiana viene determinada por el valor de $\alpha$. Se escoge de manera que sea lo suficientemente grande para que muchos de los términos en la serie del espacio real sean despreciables por encima de un determinado *cutoff* y lo suficientemente pequeña para que se reduzcan el número de términos en el espacio recíproco.

El tercer término de la ecuación [1.32] se añade para eliminar la interacción de cada gaussiana consigo misma realizada en la suma en el espacio real. El último término se añade si el entorno es vacío (con $\varepsilon_o$=1).

La ecuación [1.32] es la manera más exacta de incluir todos los efectos de fuerzas de largo alcance, aunque es computacionalmente cara de implementar. Formalmente es del orden de $O(N^2)$, aunque puede reducirse a $O(N^{3/2})$ si se ajusta adecuadamente la anchura de la gaussiana ($\alpha$), el número de vectores $K$ y el truncamiento de las interacciones de los pares en el espacio directo.

Para acelerar la solución del método de sumas de Ewald, se han diseñado diversas aproximaciones basadas en mallas (*particle mesh-based approaches*). Todas ellas utilizan una transformada rápida de Fourier (FFT) para calcular la suma en el espacio recíproco, para lo cual hay que discretizar los valores. Para discretizar los valores, en lugar de trabajar con una densidad de cargas continua, se aproxima a un modelo de cargas distribuidas en una malla construida sobre el espacio cartesiano sobre el que se realiza la dinámica molecular. A partir de la distribución de cargas en la malla, se obtiene el potencial debido a las distribuciones

gaussianas en los puntos de la malla, que vuelven a interpolarse para generar el potencial en las posiciones de las partículas. En este caso, el algoritmo es de orden *O*(*Nlog(N)*).
En particular, AMBER, para el tratamiento de las interacciones electrostáticas de largo alcance en PBC, utiliza el denominado *particle-mesh Ewald method* (PME) desarrollado por Darden[199]. Este método difiere de otros *particle-mesh* en que la interpolación la realiza mediante ajuste de *splines*.

Las gaussianas de la suma directa se calculan como en la suma de Ewald: por encima de un valor *cutoff* no se calculan y utilizan también una lista de vecinos como las expuestas anteriormente.

### 1.3.6. Modelos de solvente

Las primeras simulaciones de proteínas y complejos de proteínas-ligandos ignoraban todas las moléculas de solvente debido a la limitación computacional. Así, trabajaban *in vacuo*, considerando el sistema biológico en fase gas. Este tipo de simulaciones es problemática, ya que en los límites del sistema se tiende a minimizar la superficie y las moléculas pequeñas adoptan conformaciones más compactas debido a que las interacciones no enlazantes intramoleculares son más favorables.

Actualmente, existen diferentes modelos para el tratamiento del solvente en sistemas biológicos. En este trabajo, únicamente se han considerado sistemas biológicos en los que el solvente es el agua, sin tratar en ningún caso la descripción de membranas biológicas. En un entorno polar como el agua, la contribución principal a la solvatación procede de las interacciones electrostáticas entre soluto y solvente.[200] La elevada polarizabilidad del agua, la gran diferencia entre la constante dieléctrica del agua y las proteínas y la incertidumbre en la localización y magnitud de las cargas parciales, hacen que el término electrostático de la función de potencial sea uno de los más difíciles de representar. Se citan los métodos empleados en el presente trabajo:

#### 1.3.6.1. Métodos Empíricos

Los métodos empíricos tratan el solvente a un coste computacional muy bajo, para simular el apantallamiento que produce el campo de reacción del solvente en las interacciones electrostáticas entre átomos de la molécula. Para solventes homogéneos y disoluciones muy diluidas, este efecto puede representarse mediante la constante dieléctrica (ε=80, para el caso del agua). Sin embargo, en sistemas biológicos la constante dieléctrica efectiva depende de la distancia (*r*) entre grupos cargados, que suele modelarse con dependencia lineal (ecuación [1.34]):

$$\varepsilon_r = EPS \cdot r \qquad [1.34]$$

Donde *EPS* es un factor constante con valores generalmente comprendidos entre 1 y 4.5. En otros casos se utiliza una dependencia exponencial o sigmoidea con la distancia.
Estos métodos se aplicaron ampliamente en las primeras dinámicas moleculares y se siguen utilizando en los programas de *docking*. Actualmente, la implementación de modelos implícitos para el solvente, que aportan información acerca de la solvatación de cada elemento individual del sistema, está reemplazando su uso en dinámica molecular.

#### 1.3.6.2. Solvente Explícito

La inclusión de solvente *de forma explícita*, de manera que se trata a nivel atómico, es una de las formas más exactas, pero también más costosas computacionalmente.

En la mayor parte de los casos, se tratan en condiciones periódicas de contorno (PBC): la molécula de soluto se sitúa en el centro de la celda y el espacio vacío en ella se rellena con moléculas de solvente. En este caso, AMBER realiza el tratamiento de las interacciones de largo alcance con el método de sumas de Ewald.

Otra forma de solvatar explícitamente consiste en rodear la molécula con una capa (*cap*) de moléculas de solvente y sin tratamiento de condiciones periódicas de contorno. En este caso, el número de moléculas de agua requeridas es menor que en PBC, por lo que resulta más asequible computacionalmente que la solvatación explícita periódica. Para prevenir la evaporación de las aguas en el límite solvente-vacío, se aplican *stochastic boundary conditions* mediante la restricción de un potencial harmónico.
En la versión de AMBER 8[187] se ha implementado un modelo alternativo de solvatación para el tratamiento de esta capa de aguas respecto a versiones anteriores de AMBER. Así, se incluye una corrección para el campo de reacción de las aguas que están situadas tras la capa (*cap*), calculado mediante el método de diferencias finitas de Poisson-Boltzmann.[201] No se trata de un modelo de solvatación implícito, como los que se presentan posteriormente, ya que no trata la generalidad del sistema mediante este modelo.
Las regiones interiores al radio de la capa de aguas (soluto+solvente explícito) se detallan a nivel atómico y el resto se trata como un medio continuo. Se destaca que en versiones anteriores de AMBER, se permitía la inclusión de una *cap* de aguas que solvatase parcialmente el sistema (normalmente la región activa). En AMBER 8[187], ya que modela como un continuo todo aquello más allá del radio de la capa, la esfera de aguas ha de englobar a todo el soluto.

#### 1.3.6.3. Solvente Implícito

La descripción exacta del entorno acuoso puede resultar computacionalmente cara: la solvatación explícita de una proteína de tamaño medio requiere miles de moléculas de agua. Actualmente, la alternativa de reemplazar estas aguas discretas por un sistema "virtual" de aguas está cobrando gran popularidad. Así, se modela un medio infinito continuo con las propiedades dieléctricas e hidrofóbicas del agua. Se trata de los modelos de solvente implícito, basados en la teoría clásica de Poisson-Boltzmann (PB). En ellos, el soluto se detalla a nivel atómico, mientras que las moléculas de solvente y posibles electrolitos, se tratan como un continuo sin estructura, caracterizado por una constante dieléctrica del solvente ($\varepsilon_s$). En el interior de la cavidad del soluto, la constante dieléctrica toma valores característicos de proteínas ($\varepsilon_{int}$=2-8) o 1.

Aparte del coste computacional más reducido, estos modelos implícitos presentan una serie de ventajas frente a la representación explícita del agua como evitar el equilibrado del sistema (temperatura y presión); el soluto puede explorar más rápidamente el espacio de fases debido a la ausencia de viscosidad asociada a los modelos explícitos; se modela la solvatación en un volumen infinito, evitándose artefactos del sistema periódico y se facilita la estimación de energías de estructuras solvatadas.
Sin embargo, por otra parte se pierde también la posibilidad de analizar interacciones estructurales soluto-solvente, como la formación de puentes de hidrógeno.

##### 1.3.6.3.1. Ecuación de Poisson-Boltzmann

La ecuación de Poisson resuelve el potencial electrostático ($\phi(r)$) generado por una distribución de cargas moleculares ($\rho(r)$) dentro de un medio con una determinada constante dieléctrica ($\varepsilon(r)$). Si además se considera la presencia de iones, la distribución de los mismos se incluye en la ecuación de Poisson mediante una distribución de Boltzmann, resultando en la ecuación de Poisson-Boltzmann (PB). Para simplificar, únicamente se muestra la ecuación linearizada de PB, adecuada para el tratamiento en soluciones con una fuerza iónica baja. Otras formulaciones de esta ecuación se pueden encontrar en las referencias [200,202]:

$$\nabla \cdot \varepsilon(r)\nabla\phi(r) - \kappa'\phi(r) = -4\pi\rho(r) \qquad [1.35]$$

La constante dieléctrica es dependiente de la distancia ($\varepsilon(r)$): equivale a la del agua en zonas alejadas del soluto y desciende rápidamente con la distancia en las zonas límite soluto-solvente. El parámetro (κ'), inverso de la longitud de Debye-Hückel, adopta valores de 0.1Å$^{-1}$ en condiciones fisiológicas. Una vez se calcula el potencial ($\phi(r)$), la contribución electrostática a la energía de solvatación viene dada por:

$$\Delta G_{elec} = \frac{1}{2}\sum_i q_i(\phi(r_i)_{|solv} - \phi(r_i)_{|vacío}) \qquad [1.36]$$

Donde $q_i$ es la carga parcial del átomo en la posición $r_i$ que genera la densidad molecular y ($\phi(r)_{vacío}$) es el potencial electrostático calculado para la misma distribución de cargas pero en ausencia de límites dieléctricos (en vacío, en el que se utiliza una dieléctrica de 1 tanto en la cavidad del soluto como fuera de ella).

La ecuación de Poisson-Boltzmann no es de fácil resolución para los sistemas de interés, por lo que se tienen que aplicar métodos númericos. Entre ellos, el método de diferencias finitas en una malla (*finite-difference method*, FDPB) es el más usual. Este método se ha implementado en diversos programas como DELPHI[203], MEAD[204] y UHBD[205] y en la versión de AMBER8[187].

En el método FDPB, se superpone una malla de puntos sobre el soluto y el solvente, asignándose a cada punto de la *grid* los valores de potencial electrostático, densidad de carga, constante dieléctrica y fuerza iónica (Figura 1.4). Cada una de las cargas parciales se distribuye sobre los ocho puntos de la malla que la rodean mediante una ecuación trilineal.

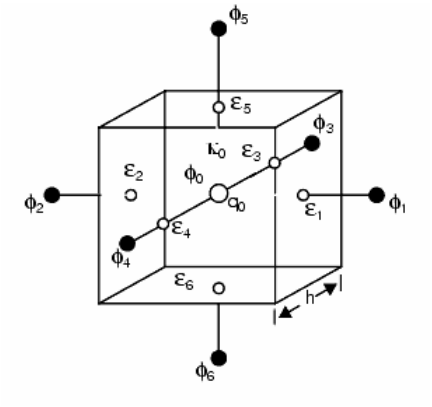



**Figura 1.4.** Esquema de la malla utilizada en el programa DELPHI para resolver la ecuación de Poisson-Boltzmann mediante el método de diferencias finitas. Extraído de [206].

El potencial en cada punto de la malla ($\phi_o$) se obtiene según la ecuación [1.37]:

$$\phi_o = \frac{\sum_{i=1}^{6}\varepsilon_i\phi_i + 4\pi\frac{q_o}{h}}{\sum_{i=1}^{6}\varepsilon_i + N\kappa_o'^2 h^2} \qquad [1.37]$$

Donde el sumatorio *i* se realiza sobre los seis puntos de la grid que rodean al punto con carga $q_o$, de manera que el potencial en cada punto afecta y es afectado por sus vecinos. Esto se traduce en una resolución del sistema de manera iterativa, hasta que alcanza convergencia. El valor *h* corresponde a la arista del cubo, κ' se calcula a partir de la fuerza iónica y *N* adopta el valor de 0 cuando la fuerza iónica es nula, 1 para la ecuación lineal o es equivalente a la expansión en serie $(1+\phi_o^2/6+\phi_o^4/120+...)$ para la ecuación no lineal.

Uno de los puntos de variación entre implementaciones es la aplicación de un modelo dieléctrico para definir el límite de constantes dieléctricas entre el agua y soluto, que puede coincidir con la superficie molecular o la superficie accesible (*Richards surface*, *van der Waals surface* o la *superficie gaussiana de exclusión de solvente*).

Además, se tienen que asignar potenciales en los límites de la superficie de la malla, en condiciones no-periódicas (técnicas de focalización, *focusing*).
Las cargas atómicas y los radios de van der Waals, utilizados para calcular la superficie accesible, se extraen de parámetros del *force field*, aunque también existen parametrizaciones especiales para resolver la ecuación de PB (como el set PARSE desarrollado por Sitkoff [207]).

En aplicaciones en dinámica molecular, la ecuación de Poisson-Boltzmann tiene que ser resuelta cada vez que la conformación de la proteína cambia, por lo que no resulta factible su uso en dinámica. Sin embargo, como se ha comentado, los autores de AMBER han desarrollado una resolución de dicha ecuación[201] para el tratamiento del campo de reacción de una capa de aguas, utilizado en este trabajo.
Así, la ecuación de Poisson-Boltzmann se ha aplicado tradicionalmente en el cálculo de propiedades electrostáticas de configuraciones "estáticas": cálculo del potencial electrostático, potenciales redox de solvatación, desplazamientos conformacionales inducidos por el solvente, flexibilidad de proteínas…) y en la determinación del $pK_a$ de grupos en proteínas.
En este trabajo, se aplica dicha ecuación para resolver el término de energía de interacción electrostática correspondiente a la solvatación, aplicado al método MM-PBSA descrito en el apartado 1.4.2. Para ello, se ha usado la resolución implementada en AMBER8.

#### 1.3.6.3.2. Modelo Generalizado de Born

El método analítico generalizado de Born (GB) supone otra alternativa para el cálculo del término electrostático de la energía libre de solvatación. Debido a su menor coste computacional, comparado con PB, esta metodología se ha convertido en un método bastante popular en dinámica molecular, para el reemplazo del solvente explícito.

A cada átomo de la molécula le corresponde una esfera de radio $\alpha_i$ con carga $q_i$ centrada en el núcleo. En el interior del átomo, se asume un material dieléctrico de constante 1. La molécula está envuelta de un solvente de alta permitividad dieléctrica (80 para el agua a 300K). La energía libre electrostática se determina a partir de la solvatación individual de Born para cada átomo, corregida por la perturbación del resto de átomos, según la ecuación [1.38]:

$$\Delta G_{elec} = -\frac{1}{2}\left(1-\frac{1}{\varepsilon}\right)\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j}{f_{GB}(r_{ij},R_i,R_j)} \qquad [1.38]$$

Donde $r_{ij}$ es la distancia entre cargas, y $R_i$ y $R_j$ son los denominados radios de Born (*effective Born radii*).

Uno de los algoritmos más comunes para $f_{GB}$ es la función desarrollada por Still y colaboradores[208] (ecuación [1.39]):

$$f_{GB} = \left[r_{ij}^2 + R_i R_j \exp(-r_{ij}^2 / 4R_i R_j)\right]^{\frac{1}{2}} \qquad [1.39]$$

Uno de los parámetros más importantes es el valor de los radios de Born, ya que no son propiedades atómicas intrínsecas, sino que dependen de la conformación del soluto, por lo que se han de recalcular tras cada cambio conformacional. Reflejan el grado de enterramiento de un átomo en el interior del soluto: para un átomo cercano a la superficie, los radios de Born son más pequeños, pudiendo igualarse al radio de van der Waals para aquellos átomos de cadena

laterales totalmente expuestas al solvente. El cálculo de estos radios se deriva de los radios de van der Waals implementados en el *force field* o de valores experimentales.

A partir de este modelo, se han generado diversas modificaciones que afectan a la forma de la función $f_{GB}$ y/o al modo en que se calculan los radios efectivos. En este sentido, cada vez se tiende a obtener funciones analíticas más rápidas, transferibles y que funcionen bien en sistemas biológicos.

AMBER8 dispone del modelo de pares de Hawkins-Cramer-Trular ($GB^{HTC}$)[209] y de un nuevo modelo desarrollado por Onufriev-Bashfor-Case ($GB^{OBC}$)[210]. Éste último, ha sido diseñado para calcular el radio efectivo de átomos enterrado en macromoléculas, para los que el modelo $GB^{HTC}$ los subestima. El modelo $GB^{OBC}$ es el utilizado en el método MM-GBSA (véase apartado 1.4.2) para calcular la energía electrostática de solvatación.

### 1.3.7. *Constraints* y *Restraints*

En el apartado 1.3.2, se ha comentado la posibilidad de aplicar constricciones o *constraints*, como las aplicadas por el algoritmo SHAKE[195] o restricciones (*restraints*) en una dinámica. Mientras que las primeras congelan unas coordenadas internas específicas, forzando al sistema a que cumpla una restricción determinada, las restricciones son funciones que penalizan la desviación de las coordenadas respecto a un valor deseado, por lo que se permite el movimiento dentro de un margen.

El algoritmo SHAKE, ampliamente usado en dinámica molecular, se aplicó inicialmente para establecer enlaces rígidos, basado en el esquema de integración de Verlet. Consta básicamente de dos etapas: i) inicialmente se permite el movimiento de todos los átomos del sistema, sin imponer *constraints* según el algoritmo de integración y ii) en un segundo paso la desviación de cada longitud de enlace se utiliza para calcular la *constraint* correspondiente que corrige dicho enlace (ecuación [1.40]). Dado que la corrección de un enlace puede afectar al resto, se resuelve de manera iterativa. Así, una vez corregidos todos los enlaces, se compara aquella distancia con la mayor desviación, si ésta supera una tolerancia determinada ($10^{-4}$-$10^{-8}$), el proceso se repite hasta cumplir la convergencia.

$$G_{ij} \approx \frac{\mu_{ij}\left(d_{ij}^{2} - d_{ij}^{'2}\right)}{2\Delta t^{2} d_{ij}^{o} d_{ij}^{'}} d_{ij}^{o} \qquad [1.40]$$

La *constraint* ($G_{ij}$) tiene la forma de la ecuación [1.40], donde $\mu_{ij}$ corresponde a la masa reducida entre dos átomos, $d_{ij}^{o}$ y $d_{ij}^{'}$ y son los vectores de enlace inicial e intermedio y $d_{ij}$ es la distancia de enlace impuesta por la *constraint*.
Los ángulos se incorporan a este esquema a partir de una *constraint* de distancia adicional: en un modelo triatómico como el agua, se impone que los dos átomos en los dos extremos estén a una determinada distancia. Sin embargo, normalmente se trabaja únicamente con *constraints* impuestas a las distancias, ya que la congelación de enlaces reduce la exploración del espacio conformacional.

Junto con la aplicación del algoritmo SHAKE para la congelación de la vibración de los enlaces, las *constraints* se aplican en dinámica molecular cuando sólo resulta de interés el comportamiento de una parte del sistema, como el sitio activo.

Las *restraints* tienen la forma de una ecuación harmónica (ecuación [1.41]), en el que $k$ corresponde a la constante de fuerza y $R_{ij}$ representa la posición de las conformación de partida para el par *i–j* restringido.

$$E = k(R_{ij} - R_{ij\ INICIAL})^2 \qquad [1.41]$$

Además de restringir el sistema de manera general a las coordenadas cartesianas en las que se encuentra, se puede restringir parámetros particulares como la distancia, ángulo y diedros. Estas últimas restricciones están dirigidas a la introducción de datos experimentales obtenidos por RMN en el refinado de los modelos obtenidos.

## 1.4. Cálculo de Energías Libres de Unión Proteína-Ligando

En esta sección se describen los métodos y funciones utilizadas en este trabajo para evaluar la afinidad proteína-ligando. Como se ha introducido, por una parte se encuentran las funciones de *scoring* aplicadas en *docking*, con simplificaciones en su formulación, y por otra los métodos propiamente dedicados a calcular la energía libre de interacción.

### 1.4.1. Funciones de *Scoring*

Las funciones de *scoring* utilizadas en este trabajo para el VS con *docking* son todas empíricas. Se trabaja con los programas AUTODOCK[81] y GOLD[80], éste último incorpora las funciones GOLDSCORE[80] y CHEMSCORE[93].

Estos métodos utilizan la aproximación de una "*master equation*" (ecuación [1.42]), formulada por Ajay y Murcko[211], que asume el carácter aditivo de los componentes de la energía libre:

$$\Delta G = \Delta G_{vdw} + \Delta G_{puenteH} + \Delta G_{elec} + \Delta G_{conform} + \Delta G_{tor} + \Delta G_{sol} \qquad [1.42]$$

Donde los cuatro primeros términos corresponden a los términos típicos de mecánica molecular que consideran la interacción de van der Waals, formación de puentes de hidrógeno, interacción electrostática y desviaciones de la geometría covalente, respectivamente. $\Delta G_{tor}$ modela la traslación y rotación globales y $\Delta G_{sol}$ incluye la desolvatación tras la unión del ligando y el efecto hidrofóbico.

#### 1.4.1.1. Función de *Scoring* de AUTODOCK

La *master equation* de AUTODOCK 3.0, basada en el ciclo termodinámico de Wesson y Eisenberg (Figura 1.5), consta de cinco términos (ecuación [1.43])[81]:

$$\begin{aligned}\Delta G = {} & \Delta G_{vdw} \cdot \sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right) \; + \\ & \Delta G_{puenteH} \cdot \sum_{i,j} E(t) \cdot \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} + E_{puenteH}\right) \; + \\ & \Delta G_{elec} \cdot \sum_{i,j} \frac{q_i \cdot q_j}{\varepsilon(r_{ij}) \cdot r_{ij}} \; + \\ & \Delta G_{tor} \cdot N_{tor} \; + \\ & \Delta G_{sol} \cdot \sum_{i_C,j} S_i \cdot V_j \cdot \exp\left(-\frac{r_{ij}^2}{2 \cdot \sigma^2}\right)\end{aligned} \qquad [1.43]$$

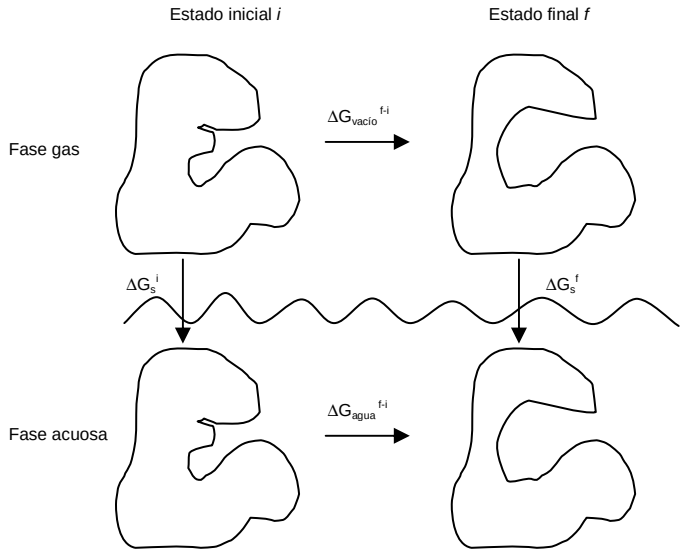



**Figura 1.5.** Esquema del ciclo termodinámico de Wesson y Eisenberg.

Los coeficientes ($\Delta G$) se determinaron empíricamente, por regresión lineal sobre un conjunto de 30 complejos proteína-ligando depositados en el *Protein Data Bank* cuya constante de inhibición ($K_i$) es conocida.
Las contribuciones en fase gas corresponden al potencial 12-6 de Lennard-Jones, a un potencial 12-10 para los puentes de hidrógeno, que tiene en cuenta la dependencia angular del enlace mediante el término *E(t)*, y al potencial electrostático de Coulomb, considerando una constante dieléctrica dependiente de la distancia de tipo sigmoideo ($\varepsilon(r)$). El cuarto término corresponde a la contribución entrópica desfavorable de unión del ligando, proporcional al número de enlaces $sp^3$ en el ligando, $N_{tor}$. Finalmente, el término de desolvatación se calcula mediante una variante del método de Souten *et al*[212] basado en ocupaciones atómicas, en este caso restringido a los carbonos alifáticos y aromáticos del ligando. Para cada uno de estos átomos, se evalúa el porcentaje de volumen alrededor de este átomo que está ocupado por átomos de la proteína y se pondera con el parámetro de solvatación atómica de dicho átomo, obteniéndose la energía de desolvatación. Además, se añade la constante $E_{puenteH}$ en el término de puentes de hidrógeno, para modelar la desolvatación de los átomos polares.
Las sumas se realizan para todos los pares de átomos del ligando (*i*) y los átomos de la proteína (*j*) así como para todos los pares de átomos en el ligando que están separados por tres o más enlaces.
La energía de interacción intramolecular del ligando no se incluye en el cálculo de la energía libre de unión, pero sí se considera en la energía total de la conformación, que es la función objetivo que dirige el proceso de búsqueda del *docking*.

Para evaluar rápidamente la energía, se precalculan potenciales de afinidad atómica para cada tipo de átomo presentes en el ligando. La proteína se sitúa en una malla o *grid* tridimensional y se coloca un átomo sonda en cada punto de la malla, calculándose dicho mapa de afinidad, donde cada punto de la malla almacena así la energía experimentada por la sonda debida a todos los átomos en la macromolécula. El potencial electrostático se obtiene típicamente mediante una sonda de carga puntual +1, aunque también se puede calcular por resolución de la ecuación de Poisson-Boltzmann, según se describe en el apartado 1.3.6.3.1. La energía de cada conformación del ligando se calcula por interpolación trilineal de los valores de afinidad de los ocho puntos de la malla que rodean a cada átomo en el sustrato.

### 1.4.1.2. Función de *Scoring* GOLDSCORE

GOLDSCORE[80,213], ecuación [1.44], incluye tres términos principales: la energía de contribución de puentes de hidrógeno entre ligando y proteína (*external H-bond,* $S_{hb\_ext}$), la energía de van der Waals proteína-ligando (*external vdW,* $S_{vdw\_ext}$) y la energía interna del ligando (*internal strain,* $S_{vdw\_int}$). Opcionalmente, se puede incluir también la energía por puentes de hidrógeno intramolecular ($S_{hb\_int}$).

$$GOLD\ Fitness = S_{hb\_ext} + S_{vdw\_ext} + S_{vdw\_int} + S_{hb\_int} \qquad [1.44]$$

Las características de los átomos (aceptor o dador de puente de hidrógeno, carácter hidrofóbico) se extraen a partir de la asignación de los tipos atómicos (normalmente los usados en Sybyl[165]), basados en la correcta conectividad de la molécula. A diferencia de AUTODOCK, no utiliza cargas parciales o formales. Así, deduce si un átomo está cargado contando el orden de enlace de los enlaces que forma y comparando el resultado con la valencia normal del átomo.

- El término *external H-bond* resulta de la suma de todas las energías de enlace de puente de hidrógeno encontradas de todas las posibles combinaciones entre átomos dadores de puente de hidrógeno del ligando y aceptores de la proteína y las combinaciones entre aceptores del ligando y dadores de puente de hidrógeno de la proteína. La contribución de un determinado par depende de: i) los tipos atómicos del aceptor y dador, que determinan la energía máxima ideal del par en el caso de una geometría de puente de hidrógeno ideal y ii) la ponderación que atenúa este valor máximo dependiendo del grado de distorsión respecto a la geometría ideal. Este peso (*w*) consiste de dos términos, uno que incluye la desviación de distancia (*dist_wt*) y otro para la desviación del ángulo (*angulo_wt*), ecuación [1.45]:

  $$w = dist\_wt \times angulo\_wt \qquad [1.45]$$

  Inicialmente, las energías máximas del par ($E_{par}$) se derivaron mediante cálculos en fase gas sobre modelos utilizando un modelo de cargas Mulliken. Para incluir la desolvatación, esta energía se calcula como la suma de las energías optimizadas para el par aceptor-dador ($E_{D\cdots A}$) y entre aguas ($E_{W\cdots W}$) menos las energías sumadas de los enlaces dador-agua ($E_{D\cdots W}$) y aceptor-agua ($E_{A\cdots W}$).
  Posteriormente, se introduce un modelo más simplista, sin perder precisión en el cálculo. En éste, los pares de puente de hidrógeno entre iones tienen un valor de -10 kcal/mol, para pares neutros es de -2 or -4 kcal/mol y de -6 kcal/mol cuando sólo una de los grupos que interaccionan está cargado.
  La geometría ideal D···A corresponde a una distancia de 2.9 Å y un ángulo de 0 ó 180 grados. La penalización de la distancia (*dist_wt*) se incrementa linealmente con el alejamiento de la distancia del valor ideal y la penalización del ángulo (*angulo_wt*) se asigna en función de la naturaleza del aceptor del grupo.

- El término de energía de van der Waals entre proteína y ligando resulta de la suma de las contribuciones de cada par *ij*, según un potencial 8-4, ecuación [1.46]:

  $$E_{ij} = \frac{A}{r_{ij}^{8}} - \frac{B}{r_{ij}^{4}} \qquad [1.46]$$

  Además del potencial más suave 8-4 que el típico 12-6, se aplica también un *cutoff* de manera que a distancias muy cortas la energía únicamente se incrementa linealmente. De este modo se permiten interacciones no enlazantes a distancias relativamente cortas, para compensar que no se introduce flexibilidad en la proteína.

Este término se multiplica por un factor (1.375) para incrementar la importancia de las interacciones hidrofóbicas.

- El término de energía interna del ligando se estima a partir de las funciones de van der Waals y contribuciones torsionales incluidas en el *force field* TRIPOS[177]. Finalmente, el término de energía de puente de hidrógeno intramolecular del ligando se calcula del mismo modo que el término externo correspondiente.

La validación de esta función se realizó sobre una base de datos de 100 complejos, aunque no se aplicaron técnicas de regresión lineal de predicción de energías de unión experimentales para entrenarla. Así, esta función ha sido optimizada para la predicción del modo de unión de ligandos más que para la predicción de afinidades de unión. Sin embargo, ésta última se puede calcular a partir de los términos de contribuciones externas, según la ecuación [1.47]:

$$\Delta G = S_{hb\_ext} + 1.375 \cdot S_{vdw\_ext} \qquad [1.47]$$

### 1.4.1.3. Función de *Scoring* CHEMSCORE

A diferencia de GOLDSCORE, CHEMSCORE fue derivada especialmente por Eldridge *et al*[214] para la predicción de la afinidad de la unión proteína-ligando, parametrizándose por regresión lineal en un conjunto de 82 complejos proteína-ligando de constante de inhibición conocida e implementándose incialmente en el programa PRO_LEADS[78].

La *master equation* original (ecuación [1.48]), contiene cinco términos:

$$\begin{aligned}\Delta G_{binding\_original} = {} & \Delta G_o + \\ & \Delta G_{puenteH} \cdot \sum_{DA} f(\Delta r_{DA}, \Delta r_1, \Delta r_2)\, f(\Delta\alpha_{DA}, \Delta\alpha_1, \Delta\alpha_2) + \\ & \Delta G_{metal} \cdot \sum_{MA} f(r_{MA}, r_{m,1}, r_{m,2}) + \\ & \Delta G_{lipo} \cdot \sum_{LL} f(r_{LL}, r_{l,1}, r_{l,2}) + \\ & \Delta G_{rot} \cdot H_{rot}\end{aligned} \qquad [1.48]$$

Los coeficientes ($\Delta G$) resultan de la regresión lineal, donde $\Delta G_o$ corresponde a una línea de base independiente del ligando.
El segundo término, correspondiente a las interacciones de puentes de hidrógeno, se computa para cada combinación dador(D)-aceptor(A) mediante dos funciones dependientes de la distancia y del ángulo, respectivamente, que evalúan la desviación de dichos parámetros ($\Delta r_{DA,}$ $\Delta\alpha_{DA}$) respecto a un valor ideal y un valor máximo. El tercer término modela las interacciones de coordinación entre cada par establecido metal(M)-aceptor(A) y el cuarto computa las interacciones lipofílicas (LL) de todos los pares de átomos lipófilos entre proteína y ligando. De nuevo, evalúan la desviación de la distancia del par ($r_{MA}$,$r_{LL}$) respecto a un valor ideal ($r_{m,1}$ , $r_{l,1}$) y máximo ($r_{m,2}$ , $r_{l,2}$). En los tres casos, se definen como funciones en bloque como la de la ecuación [1.49]:

$$B(x, x_{ideal}, x_{max}) = \begin{cases} 1 & \text{si } x < x_{ideal} \\ 1 - \dfrac{x - x_{ideal}}{x_{ideal} - x_{max}} & \text{si } x_{ideal} \le x \le x_{max} \\ 0 & \text{si } x > x_{max} \end{cases} \quad [1.49]$$

El último término, $H_{rot}$, modela la pérdida de entropía conformacional por restricción de los enlaces rotables del ligando tras la unión.

La implementación de CHEMSCORE en el programa GOLD contiene mejoras en el cómputo de cada uno de los términos respecto a la función original.
Así, en GOLD las funciones en bloque se obtienen mediante gaussianas (ecuaciones [1.50] y [1.51]), para suavizar el efecto en los extremos de los rangos de distancias:

$$B'(x, x_{ideal}, x_{max}, \sigma) = \frac{\int_{-\infty}^{\infty} B(x-u, x_{ideal}, x_{max}) g(u, \sigma) du}{\int_{-\infty}^{\infty} g(u, \sigma) du} \quad [1.50]$$

$$g(u, \sigma) = \exp(-u^2 / 2\sigma^2) \quad [1.51]$$

Además, se incluye un término que penaliza los contactos entre proteína-ligando con impedimento estérico ($E_{imped}$) y la energía interna del ligando ($E_{int}$), en un esquema similar al de la implementación de la función CHEMSCORE original en PRO_LEADS[78]. Se incorpora también un término para el tratamiento de interacciones covalentes proteína-ligando, en los casos en que se produzca una unión covalente, ($E_{cov}$), resultando en la ecuación final [1.52]:

$$\Delta G_{binding\_ChemScore\_GOLD} = \Delta G_{binding\_original} + E_{imped} + E_{int} + E_{cov} \quad [1.52]$$

El término $E_{imped}$ se calcula para todos los pares de átomos distintos de hidrógeno entre proteína y ligando según la ecuación [1.53], donde $r$ es la distancia del par y $r_{imped}$ es la distancia a la que colapsa el par. Cuando $r > r_{imped}$ es nula.

$$E_{imped} = \sum \varepsilon_{imped}(r, r_{imped})$$

$$\varepsilon_{imped}(r, r_{imped}) = \begin{cases} (20 / \Delta G_{puenteH}) \cdot (r_{imped} - r) / r_{imped} & \text{pares donor – aceptor} \\ (20 / \Delta G_{metal}) \cdot (r_{imped} - r) / r_{imped} & \text{pares metal – aceptor} \\ 1 + 4 \cdot (r_{imped} - r) / r_{imped} & \text{resto de pares} \end{cases} \quad [1.53]$$

El término $E_{int}$ corresponde a la suma del término rotacional y de impedimento estérico entre átomos del ligando unidos al menos cuatro enlaces. Finalmente, el término de interacción covalente contiene una parte torsional y una parte de acoplamiento enlace-ángulo, calculado sobre las torsiones ($\theta_{CB}$) y enlaces ($\varphi_{BA}$) que participan en la interacción covalente según la ecuación [1.54]:

$$E_{cov} = \sum_{CB} \varepsilon_{tors}(\theta_{CB}) + C_{cov} \sum_{BA} k_{BA} (\varphi_{BA} - \varphi_{o,BA})^2 \quad [1.54]$$

En este trabajo se utilizan las versiones 2.1 y 3.0 de GOLD. Se destaca que esta última versión, a diferencia de las anteriores, sí considera los átomos de carbono como dadores de puente de hidrógeno en interacciones CH···O en la función deCHEMSCORE[215], interacciones que se ha demostrado contribuyen a la estabilidad de diferentes complejos proteína-ligando[216], como las tirosina quinasas. Una validación más reciente de GOLD[215] se realizó sobre una base de datos de 224 complejos.

### 1.4.2. *Molecular Mechanics-Generalized Born Surface Area* (MM-GBSA) *Molecular Mechanics-Poisson Boltzman Surface Area* (MM-PBSA)

Este método fue desarrollado por Srinivassan y Kollman en 1998. Está basado en mecánica estadística, conteniendo los distintos términos fisicoquímicos que intervienen en el proceso de unión de un ligando a una proteína, fenómeno esquematizado en la Figura 1.6.

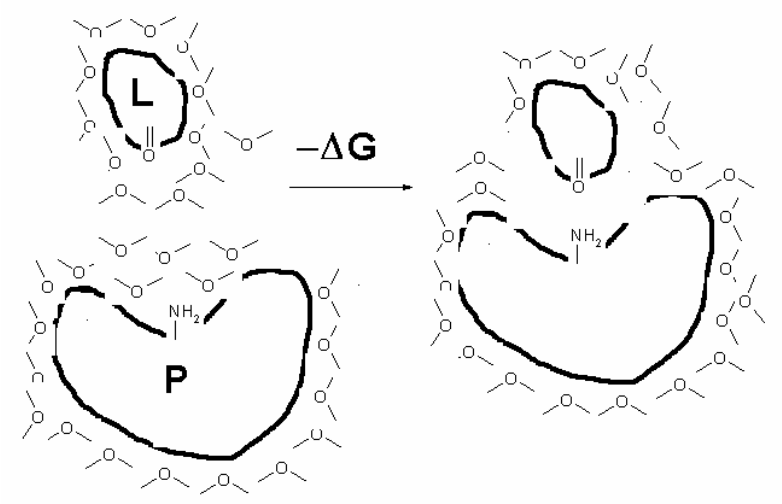



**Figura 1.6.** Esquema de unión de un ligando y proteína, con el desordenamiento de aguas que producen el efecto hidrofóbico.

Inicialmente, la proteína y el ligando se hayan solvatados por moléculas de agua. Tras la unión, las interacciones intermoleculares no enlazantes (suponiendo que no hay unión covalente), estabilizan el complejo. El cambio entrópico asociado al proceso es debido a la reducción de libertad conformacional del ligando (supone una reducción de entropía) y por el denominado efecto hidrofóbico producido por el desordenamiento de las moléculas de agua, inicialmente ordenadas en torno al ligando y receptor, contribuyendo positivamente al cambio entrópico. Termodinámicamente, corresponde a la ecuación [1.55], donde las interacciones intermoleculares establecen la variación entálpica.

$$\Delta G_{binding} = \Delta H - T\Delta S \qquad [1.55]$$

De manera similar al método FEP (véase Introducción), el método MM-PBSA/GBSA utiliza un ciclo termodinámico para calcular $\Delta G_{binding}$. Este ciclo, esquematizado en la Figura 1.7, calcula la energía de unión a partir de las energías de solvatación de cada una de las especies químicas implicadas ($\Delta G_{solv}{}^{L}$, $\Delta G_{solv}{}^{P}$, $\Delta G_{solv}{}^{LP}$) y de la energía libre de formación del complejo en fase gas $\Delta G_{gas}$, ecuación [1.56]:

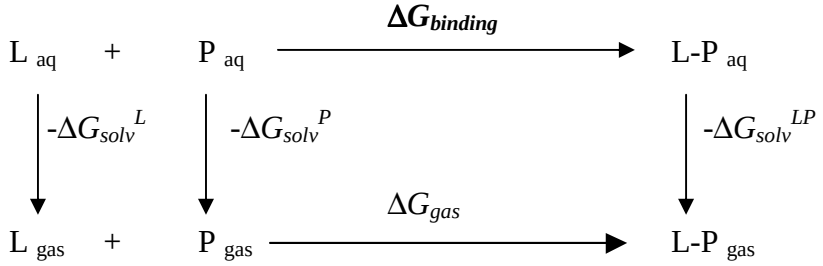



**Figura 1.7.** Ciclo termodinámico para el cálculo de la energía de unión Proteína–Ligando.

$$\Delta G_{binding} = \Delta G_{gas} - \Delta G^{L}_{solv} - \Delta G^{P}_{solv} + \Delta G^{LP}_{solv} \quad [1.56]$$

$$= \Delta H_{gas} - T\Delta S - \Delta G^{L}_{PB/GBSA} - \Delta G^{P}_{PB/GBSA} + \Delta G^{LP}_{PB/GBSA} \quad [1.57]$$

donde:

$$\Delta H_{gas} \approx \Delta E_{gas} = \Delta E_{internal} + \Delta E_{electros} + \Delta E_{vdw} \quad [1.58]$$

$$\Delta G_{PBSA/GBSA} = \Delta G_{PB/GB} + \Delta G_{SA} \quad [1.59]$$

$$\Delta\Delta G_{PB/GB} = \Delta G^{LP}_{PB/GB} - (\Delta G^{L}_{PB/GB} + \Delta G^{P}_{PB/GB}) \quad [1.60]$$

$$\Delta\Delta G_{SA} = \Delta G^{LP}_{SA} - (\Delta G^{L}_{SA} + \Delta G^{P}_{SA}) \quad [1.61]$$

$\Delta G_{gas}$ puede escribirse como la suma de la variación entálpica ($\Delta H_{gas}$) y entrópica ($-T\Delta S_{gas}$) (ecuación [1.57]). A su vez, la entalpía puede escribirse como la energía del potencial en fase gas que adopta la ecuación del *force field* ($\Delta E_{gas}$), donde $\Delta E_{internal}$ representa el potencial de las interacciones de enlace (ángulos, diedros...), $\Delta E_{electros}$ corresponde a la variación en las interacciones electrostáticas y $\Delta E_{vdw}$ a las interacciones de van der Waals en fase gaseosa. El término de energía interna ($\Delta E_{internal}$) se desprecia al asumirse que la energía intramolecular del ligando no varía significativamente tras la unión, de manera que se facilita el cálculo de energías de unión absolutas y relativas. La variación entrópica se puede calcular con diferentes aproximaciones.

La energía de solvatación de un compuesto viene dada por la contribución electrostática, las interacciones de van der Waals y el término de cavitación (debido a la formación de la cavidad en el solvente para albergar al soluto), ecuación [1.62]:

$$\Delta G_{solv} = \Delta G_{ele} + \Delta G_{vdw} + \Delta G_{cav} \quad [1.62]$$

En solvente polares como el agua, la contribución de las fuerzas dispersiva-repulsivas es moderada, inferior al término de energía de cavitación. Estos dos términos, referidos como contribuciones no polares, se estiman conjuntamente ($\Delta G_{SA}$).
El término electrostático supone la contribución más importante, debido a la fuerza de las interacciones soluto-solvente. Este término no solo incluye estas interacciones, sino también el trabajo necesario para generar el campo de reacción del solvente inducido por la distribución de cargas del soluto. $\Delta G_{ele}$ equivale a la mitad de la energía de interacción soluto-solvente. Esta contribución electrostática se evalúa a partir de modelos continuos del solvente (véase apartado 1.3.6.3): bien a partir de la resolución de la ecuación de Poisson-Boltzmann mediante el método de diferencias finitas ($\Delta G_{PB}$, y el método se denomina MM-PBSA) o mediante un modelo Generalizado de Born ($\Delta G_{GB}$, MM-GBSA). En teoría, los resultados obtenidos por MM-GBSA

ó MM-PBSA son similares[116], aunque el modelo generalizado de Born es más rápido. En resumen, la ecuación [1.56] puede reformularse como [1.57], a partir de la ecuación [1.59].

En versiones de AMBER anteriores a la 8, hay que recurrir a programas como DELPHI, UHB o MEAD para la resolución de PB, destacando el uso masivo de DELPHI. A partir de AMBER8, se incorpora un método de resolución de PB. Por otro lado, el modelo GB se resuelve a partir de los implementados en AMBER.

El término de interacciones no-polares es proporcional a la superficie accesible al solvente (*solvent accesible surface area, SA*), que describe el área sobre la cual se produce contacto ligando–proteína, según la ecuación [1.63]:

$$\Delta G_{SA} = \gamma\, SA + \beta \qquad [1.63]$$

La superficie accesible se determina a partir de la posición del centro de una sonda esférica (que representa una molécula de solvente, de radio 1.4 Å) que rueda sobre la superficie de van der Waals de la proteína. Incrementando el valor de los radios de van der Waals por el radio de la sonda, se obtienen los radios denominados *expandidos* (*expanded atom radii*). En la Figura 1.8 se esquematiza este proceso:

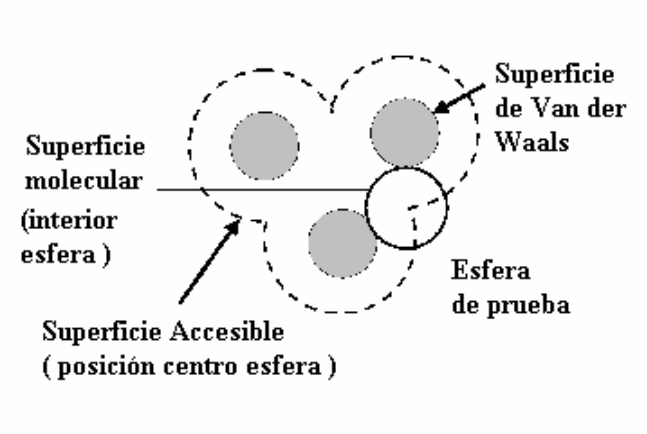



**Figura 1.8.** Representación de la superficie accesible de una molécula. Extraído de [217].

AMBER8 contempla dos posibilidades para el cálculo de SA, i) con el programa *molsurf* de Beroza que implementa el algoritmo de Connolly[218] o bien ii) con el modelo de combinaciones lineales de solapamientos entre pares (*Linear Combinations of Pairwise Overlaps*, LCPO)[219]. En este trabajo, únicamente se utiliza el primer modelo.
Los valores de los parámetros de tensión superficial, $\gamma$ y $\beta$, dependen de la parametrización de los radios utilizada para calcular la superficie, ligada al modelo de cálculo de interacción electrostática, según se muestra en la Tabla 1.1.

**Tabla 1.1.** Constantes para el cálculo de $\Delta G_{SA}$ en función de la parametrización de los radios.

| | **γ (kcal/Å$^2$)** | **β (kcal/mol)** |
|---|---|---|
| Radios Parse (Poisson-Boltzmann, DELPHI) | 0.00542 | 0.92 |
| Radios optimizados de AMBER8 (Poisson-Boltzmann, AMBER) | 0.00500 | 0.00 |
| Radios *mbondi2* (AMBER8) Generalizado de Born | 0.00720 | 0.00 |

Una característica del método MM-PBSA/MM-GBSA es que no utiliza parámetros empíricos, por lo que puede aplicarse directamente en la estimación de las energías de unión. En este sentido, es más versátil que el método LIE citado en la introducción.

La evaluación de cada uno de los términos que intervienen en la ecuación [1.57] se toma como el valor promedio de una serie de *snapshots* ("fotos") de las estructuras tomadas de la trayectoria de una dinámica molecular realizada en solvente explícito. Se quiere puntualizar que los modelos implícitos de solvente se aplican únicamente sobre estas estructuras individuales.

Existen dos protocolos posibles para aplicar el método MM-PBSA/GBSA: i) todos los *snapshots* para ligando, proteína y complejo se extraen de una única simulación del complejo y ii) los *snapshots* del complejo se extraen de una dinámica del complejo, los de la proteína, de una dinámica de la proteína y los *snapshots* del ligando, de una dinámica sobre él.
La primera opción asume que la trayectoria que adoptan la proteína y el ligando en el complejo es de energía libre equivalente a la que adoptarían en una trayectoria por separado. Requiere menos simulaciones, lo que la ha convertido en una alternativa muy generalizada[116,122,220,221] a la par que se sugiere que se trata de una aproximación suficientemente correcta. Sin embargo, otros estudios [121,222] inciden en que se debería tomar con más precaución cuando se aplica a proteínas cuya flexibilidad y estructura varían de forma significativa tras la unión del ligando.

## 1.5. Modelización de Proteínas por homología

Como se ha introducido, la modelización por homología de proteínas se aplica cuando la proteína diana (*target*) comparte un alto grado de similitud secuencial con otras proteínas cuya estructura está resuelta, sirviendo estas últimas de plantilla (*template*). Así, el primer paso consiste en la búsqueda y selección de estas plantillas. Una vez alineadas la secuencia diana frente a la secuencia de las plantillas, se construye el modelo, cuya validez se evalúa a partir de criterios estructurales (por ejemplo, mapas de Ramachandran) y datos experimentales, como los procedentes de experimentos de mutagénesis dirigida, receptores quiméricos o información del acoplamiento de alguno de sus ligandos.

### 1.5.1. Búsqueda de estructuras y secuencias relacionadas con la secuencia objetivo

Generalmente, se utilizan métodos que comparan la secuencia de la proteína objetivo con las secuencias recopiladas en una base de datos. Destacan los paquetes BLAST[223] (*Basic Local Alignment Search Tools*) y FASTA[224] (*Fast Alignment*). Ambos contienen una serie de programas basados en los algoritmos con sus mismos nombres y accesibles a través de servidores Web. Se trata de algoritmos de alineamiento heurísticos, no garantizan encontrar el mejor alineamiento entre la secuencia y las secuencias de la base de datos, ya que priorizan la rapidez del cálculo frente a otros algoritmos más exactos. Deben presentar un balance entre sensibilidad, es decir la capacidad de detectar el máximo número de verdaderos positivos y especificidad, de manera que se rechacen el máximo número posible de falsos positivos. La significancia de las secuencias encontradas se establece a partir de una serie de parámetros estadísticos, expresados mediante valores de corte. El concepto de alineamiento y la base de estos algoritmos se describen en el apartado 1.5.2.

Respecto a las bases de datos, estos programas están directamente conectados a aquellas más importantes, tanto de proteínas como de ácidos nucleicos. En el caso de proteínas, estas bases de datos pueden contener entradas para la translación de genes, secuencias de proteínas y/o proteínas con estructura tridimensional publicada y bases de datos de patentes.

El *Brookhaven Protein Data Bank*[60], que contiene únicamente estructuras resueltas de proteínas por rayos-X o RMN, supone la referencia para la selección de plantillas en modelización por homología. Inicialmente, contenía también modelos teóricos, pero desde julio de 2002 éstos se encuentran depositados de forma separada de las estructuras experimentales. Sin embargo, la búsqueda de secuencias relacionadas con la diana cuya estructura no ha sido resuelta puede resultar también de interés para determinar la familia o subfamilia de la proteína diana, el grado de conservación de residuos, etc. En este caso, se suele recurrir a bases de datos como SWISS-PROT[225], en la que se indexan todas las proteínas secuenciadas y que contiene múltiples referencias a otras bases de datos. En la Tabla 1.2 se recoge un compendio de las más conocidas, utilizadas en este trabajo.

La referencia [231] es una revisión donde se compila información acerca de la mayor parte de estas bases de datos.

Una vez se ha buscado en las bases de datos, se debe revisar individualmente cada plantilla, no solo para asegurar una buena significancia estadística, sino también para seleccionar aquellas más apropiadas según factores como pertenencia a una misma subfamilia, que compartan un mismo entorno (solvente, ligandos, pH…), la calidad de la resolución de la estructura, etc. Tampoco se trata de seleccionar una única plantilla, ya que el uso de varias de ellas generalmente incrementa la calidad del modelo.

Los receptores acoplados a proteínas G (GPCRs) son una de las familias para las que más se recurre a la modelización por homología, debido a su importancia como dianas terapéuticas y a la dificultad de su cristalización. Únicamente se dispone de la estructura resuelta por rayos-X de la rodopsina bovina por Palczewski[232] en el año 2000.

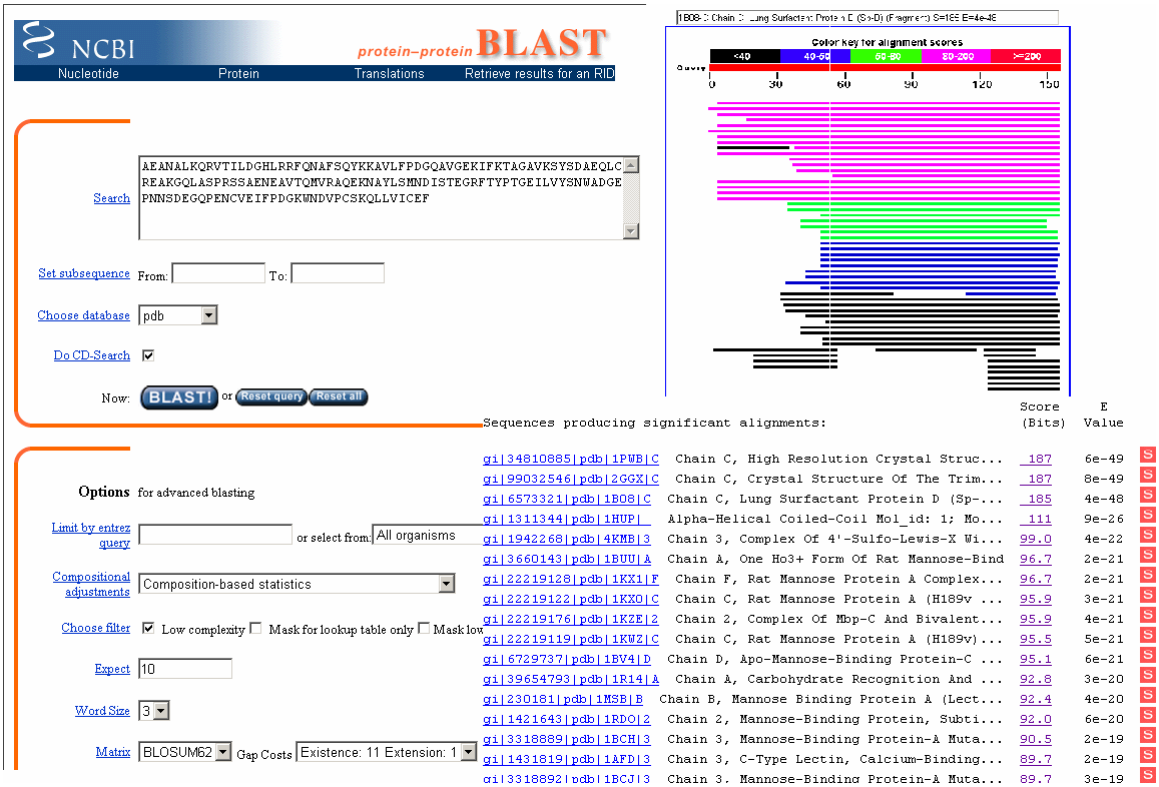



**Figura 1.9.** Ejemplo de búsqueda de estructura en el *Protein Data Bank* mediante BLAST.

**Tabla 1.2.** Bases de datos de secuencias.

| BASE DATOS | DESCRIPCIÓN | CENTRO DE MANTENIMIENTO ACCESO |
|---|---|---|
| SWISS-PROT[225] | Secuencias de proteínas. Múltiples referencias a otras bases de datos. No redundante (problema de esto es que no todas las secuencias aparecen). | Universidad Ginebra (1986)<br>Swiss Institute of Bioinformatics (SIB)//<br>European Bioinformatics Institute(EBI)<br>http://us.expasy.org/sprot/ |
| TrEMBL[225] | Suplemento de SWISS-PROT, contiene translaciones de las secuencias de nucleotidos de la base EMBL. | Universidad Ginebra (1986)<br>Swiss Institute of Bioinformatics (SIB)//<br>European Bioinformatics Institute(EBI)<br>http://us.expasy.org/sprot/ |
| PIR[226]<br>Protein<br>Sequence<br>Database (PSD) | Secuencias de aminoácidos. Intenta ser una mezcla entre una base de datos completa y no redundante, así está organizada en cuatro secciones:<br>- PIR1: no redundante, sólo contiene una entrada por proteína.<br>- PIR2+PIR3+PIR4: redundante, es muy completa, pero acepta incluso entradas no clasificadas o aceptadas. | Margaret Dayhoff (1984).<br>Protein Identification Resource<br>http://pir.georgetown.edu/pirwww/ |
| PDB[60]<br>Protein<br>Data Bank | Información sobre estructuras resueltas (NMR, rayos-X), los modelos teóricos están en otro dominio desde julio 2002.<br>En principio no redundante, sólo se mantiene la mejor determinación, pero se encuentran múltiples estructuras para una molécula, debido a resoluciones parciales, inclusiones de cofactores… | Brookhaven National Laboratory<br>http://www.rcsb.org/pdb/ |
| nr[223]<br>*non-redundant* | Mezcla de las anteriores (PDB, PIR y SwissProt, translaciones de GenBank). Se utiliza por defecto para las búsquedas con BLAST. Las entradas con secuencias absolutamente idénticas se han fusionado. | National Center for Biotechnology Information (NCBI)<br>http://www.ncbi.nlm.nih.gov/ |
| OWL[227] | No redundante, compuesta básicamente a partir de SWISS-PROT y PIR. | Bleasby (1990). University of Manchester Bioinformatics Education and Research (UMBER)<br>http://umber.sbs.man.ac.uk/dbbrowser/OWL/ |
| UniProt<br>Knowledgebase<br>(Universal<br>Protein<br>Resource)[228] | Unión de UniProtKB/Swiss-Prot, UniProtKB /TrEMBL y PIR-PSD. Se propone como una de las bases de datos que indexan más información para cada entrada, con bajo nivel de redundancia. | Apweiler (2003)<br>Swiss Institute of Bioinformatics (SIB) //<br>European Bioinformatics Institute(EBI)<br>http://www.ebi.ac.uk/swissprot/access.html |
| GPCRDB[229] | Contiene información (secuencia, alinea-mientos, filogenia ...) sobre las GPCRs | Horn (1998). Center for Molecular and Biomolecular Informatics<br>http://www.gpcr.org/7tm/ |
| Protein Kinase Resource[230] | Contiene información (secuencia, estructuras alineamientos, filogenia...) sobre las Proteína Quinasas. | University of California. 1997<br>San Diego SuperComputer Center at UCSD<br>http://www.kinasenet.org/pkr/Welcome.do |

### 1.5.2. Alineamiento de Secuencias

El alineamiento es una de las herramientas más importantes de la bioinformática, utilizado en numerosas tareas: reconstrucción de cadenas de DNA largas a partir de solapamientos de fragmentos, comparación de secuencias para encontrar similitudes o elementos estructurales característicos, búsqueda en bases de datos, etc.

El modelo más simple para trabajar con alineamientos parte del concepto de *Edit distance* entre dos secuencias, como el mínimo número de operaciones (inserciones, deleciones y sustituciones) necesarias para transformar una secuencia en otra. En general, la *Edit Distance*, se evalúa a partir de una función *w*, que describe los costes de todas estas operaciones, de forma que el coste de un alineamiento de dos secuencias *S* y *T* es la suma de los costes de cada operación. El alineamiento óptimo será aquel que muestre el mínimo coste entre todos los posibles alineamientos.
Existen diferentes modelos para esta función *w*: *Hamming Distance, Levenshtein Distance o Unit Cost Model* y los modelos basados en matrices de sustitución. Éstos últimos son los más sofisticados, ya que consideran el significado biológico de las sustituciones.
Además de las sustituciones, las eliminaciones e inserciones generan *indels* (aminoácidos de una secuencia se alinean frente a espacios en blanco) en alguna de las secuencias. Cada serie de espacios consecutivos en el alineamiento define un *gap*, caracterizado por su longitud. Cada *gap* se entiende como una unidad, ya que ayuda a la búsqueda de mayor significado biológico (en un único evento mutacional pueden aparecen inserciones/deleciones de una subsecuencia). Existen muchas maneras de evaluar los *gaps* (*gap penalty models*), pero en general se penaliza de manera diferente la abertura de un nuevo *gap* en el alineamiento (*gap opening penalty*) y el hecho de extenderlo (*gap extensión penalty*).

Cuando se realiza una búsqueda en base de datos o se realiza el alineamiento particular de un conjunto de secuencias, los parámetros que lo definen son básicamente: el algoritmo y tipo de alineamiento, la matriz de sustitución empleada y el modelo de penalización de los *gaps*.

#### 1.5.2.1. Alineamiento de Secuencias

Se han desarrollado numerosos algoritmos de alineamiento, en función del tipo de alineamiento que se pretende realizar:

- <u>*Alineamientos globales*</u>: de pares de proteínas con longitud similar a lo largo de ella, generalmente relacionadas por un ancestro común. El algoritmo por excelencia es el de Needleman-Wunsch[233], de programación dinámica permite encontrar el alineamiento óptimo sin tener que enumerar explícitamente todas las posibilidades. En las modificaciones actuales del mismo se permite la introducción de *gaps*.
  Se construye una matriz de $M \times N$, donde $M$ representa a los aminoácidos de la proteína $A$ y $N$ a los aminoácidos de la proteína $B$. Cada elemento $H_{ij}$ de esta matriz corresponde a un *score* óptimo del alineamiento de dos subsecuencias ($1 \ldots i$, para $A$ y $1 \ldots j$, para $B$) de forma que ($1 \leq i \leq M$, $1 \leq j \leq N$). El algoritmo avanza desde el elemento superior-izquierdo hasta el elemento inferior-derecho de la matriz (en la versión original es al contrario). El valor asignado a cada elemento de la matriz se obtiene según la ecuación [1.64]:

$$H_{i,j} = \max \begin{cases} H_{i-1,j-1} + w_{A_i,B_j} \\ H_{i-1,j} + w_{A_i,\Delta} \\ H_{i,j-1} + w_{\Delta,B_j} \end{cases} \qquad [1.64]$$

  Donde $w_{Ai,Bj}$ corresponde al *score* de alineamiento de los aminoácidos procedentes de cada proteína (según la matriz de sustitución) y $w_{Ai,\Delta}$ y $w_{\Delta,Bj}$ corresponden a la penalización por

alineamiento de un aminoácido frente a un *gap*. Se introduce una fila y una columna $H_{0,0}$ con un espacio y unas condiciones base:

$$H_{i.0} = \sum_{k=0}^{i} w_{A_k,\Delta} \; ; H_{0,j} = \sum_{k=0}^{j} w_{\Delta,B_j} \qquad [1.65]$$

Una vez asignado el valor del último elemento ($H_{M,N}$), éste representa el valor del *score* global del alineamiento. El alineamiento final se determina recorriendo en sentido contrario la matriz y escogiendo los elementos de la matriz con valores mayores. El trazado de subíndices indica el alineamiento final resultante.

- <u>*Alineamientos locales*</u>: a partir de dos secuencias de proteínas, se intenta encontrar las subsecuencias de máxima similitud entre ellas, ya que es muy frecuente que dos proteínas únicamente muestren similitud en regiones locales. El más utilizado es el de Smith-Waterman[234], que también es de programación dinámica. En lugar de mirar cada secuencia en su globalidad, compara segmentos de todas las longitudes posibles y escoge cualquiera que maximice la medida de similitud. Corresponde esencialmente al algoritmo de Needleman-Wunsch, aunque se añade un cero y se modifican las condiciones de partida para la fila y columna adicionales, para evitar que se obtengan similitudes negativas (ecuaciones [1.66] y [1.67]):

$$H_{i,j} = \max \begin{cases} H_{i-1,j-1} + w_{A_i,B_j} \\ H_{i-1,j} + w_{A_i,\Delta} \\ H_{i,j-1} + w_{\Delta,B_j} \\ 0 \end{cases} \qquad [1.66]$$

$$H_{i.0} = 0 \; ; H_{0,j} = 0 \;\; \forall i, j \qquad [1.67]$$

El par de segmentos con máxima similitud lo determina el recorrido inverso desde el elemento de la matriz con mayor valor $H_{i,j}$ hasta llegar a un elemento con valor nulo. A partir del segundo máximo valor de $H_{i,j}$ se deriva el siguiente par de segmentos y así sucesivamente.

- <u>*Alineamientos de final libre* (*Ends free Alignment*)</u> de pares de proteínas. El caso más común es cuando una secuencia es relativamente corta respecto a otra y se intenta encontrar aquella subunidad de la otra que mejor alinea con la primera. Este algoritmo puede obtenerse rápidamente a partir del algoritmo de Needleman-Wunsch, ya que principalmente supone el permitir introducir un número de *indels* necesarios en una secuencia sin que tengan ningún tipo de penalización.

- <u>*Alineamientos múltiples*</u> de miembros de una familia de proteínas. Un alineamiento múltiple es más fiable que uno de un par de secuencias, ya que es más sencillo detectar tendencias y evitar posibles artefactos. Este es el punto que presenta mayor variedad de teorías acerca de cómo implementarse. Por una parte, la ampliación del algoritmo de Needleman-Wunsch a $N$ secuencias, aunque posible, en la práctica se ha adaptado únicamente para el alineamiento de un máximo de tres secuencias.
  Una de las aproximaciones más comunes es realizar una aproximación de *clusters* jerárquicos. En principio, se generan todos los alineamientos de parejas posibles y éstos se agrupan según un análisis de *clusters* jerárquicos. Conforme se asciende por el dendrograma, se realizan alineamientos de alineamientos previos frente a pares y alineamientos previos. Para evaluar el valor de la sustitución en este caso, se utilizan matrices dependientes de la distancia (*profiles*) obtenidas promediando los valores de sustitución de todos los aminoácidos en una posición determinada.

En este trabajo se ha utilizado el módulo de alineamiento del programa MOE[38]. Éste realiza los alineamientos múltiples en cuatro niveles (Figura 1.10): i) Inicialmente, estima un primer alineamiento a partir de un esquema en árbol. Para ello, precalcula todos los alineamientos de parejas posibles, y comienza a agruparlos sucesivamente según el que tenga un mayor *score*. ii) Sobre este alineamiento inicial, se aplican realineamientos *Round-robin* (planificación por turno aleatorio), en los que cada cadena es sucesivamente extraída del alineamiento total y realineada de nuevo frente a las restantes. iii) Dado que este segundo paso es dependiente del orden de realineamiento de las secuencias, se aplican una serie de realineamientos aleatorios en los que se parte el global en dos grupos y éstos dos se vuelven a realinear. Si el resultado mejora, se acepta este nuevo alineamiento, de lo contrario se rechaza. iv) Finalmente, en un cuarto paso, se puede incluir la estructura de las proteínas para las que se tienen coordenadas (en este caso, las plantillas), de manera que éstas se realinean. Para ello, se genera una matriz de similitud basada en las coordenadas relativas del esqueleto de carbonos alfa obtenidas por superposición de las mismas. El realineamiento se repite hasta no se mejora la RMSD de la superposición. Entonces, se introduce el bloque de cadenas alineadas por estructura, tratándose a partir de entonces como una única unidad, con el resto de proteínas sin estructura y se repiten los pasos desde i) hasta iii).

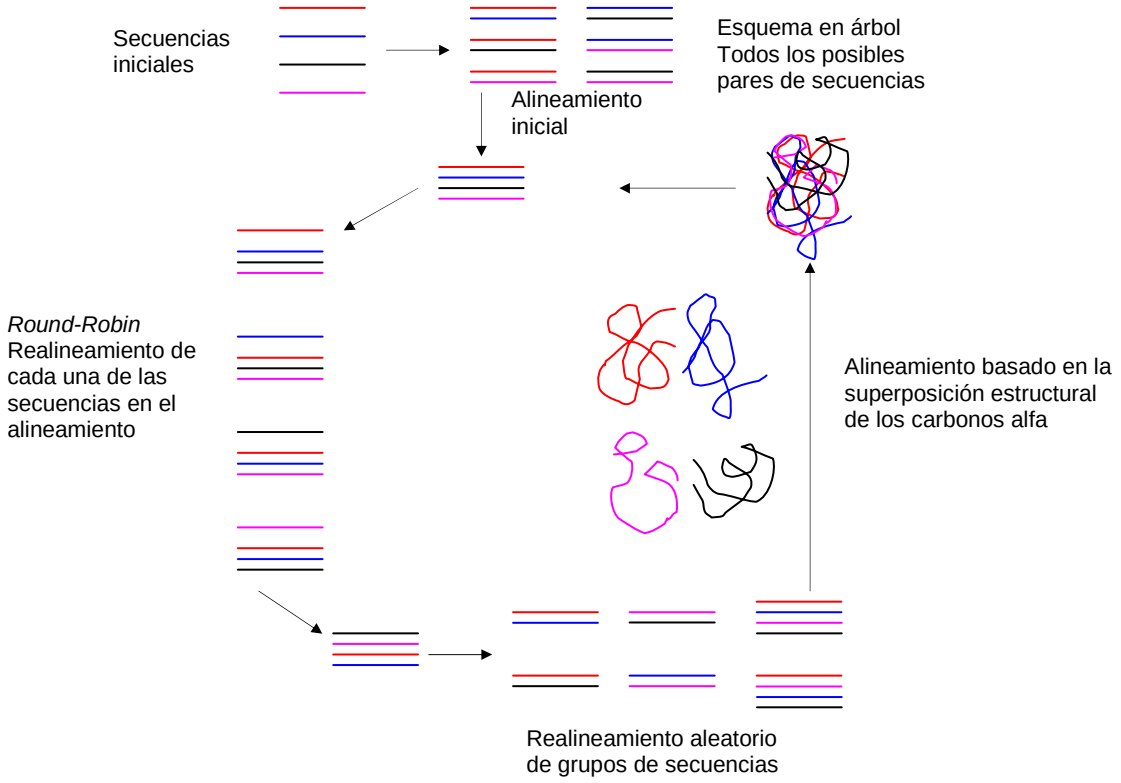



**Figura 1.10.** Esquema del algoritmo de alineamiento múltiple disponible en MOE.

- *Alineamientos Heurísticos* Corresponden a los métodos para realizar búsquedas en bases de datos, cuyo objetivo no es la búsqueda del alineamiento óptimo entre secuencias, sino la identificación de secuencias similares en un intervalo de tiempo razonable y buena sensibilidad.

FASTA se basa en la identificación de un motivo de palabras (*word*) conservados entre el par de secuencias para localizar los posibles puntos de similitud antes de realizar una búsqueda optimizada. Estas palabras son de una determinada longitud (*ktup*), normalmente de valor igual a dos para el caso de proteínas, aunque puede ser fijado por el usuario. Así en un primer paso identifica pares de identidades (*ktup*=2) entre dos secuencias (diana y una secuencia procedente de la base de datos) mediante una tabla de búsqueda. A continuación, se unen los pares presentes en una misma diagonal de la matriz de secuencias, se evalúa su similitud mediante una matriz de sustitución (PAM250) y se seleccionan las diez mejores regiones locales, que no tienen porqué pertenecer a la misma diagonal. Cada una de estas regiones corresponde a un alineamiento parcial sin *gaps* que se evalúa de nuevo mediante la matriz PAM250 (*scores* denominados *init1*). Aquellas regiones con *init1* superior a un determinado valor *cutoff* se unen, permitiendo la introducción de *gaps* entre ellos, calculándose de nuevo un *score* total inicial (*initn*) mediante la suma de los *init1* individuales menos una penalización (20) por cada *gap* introducido. Se construye un alineamiento óptimo mediante el algoritmo Needleman-Wunch-Sellers considerando únicamente el segmento comprendido a ±32 residuos de la mejor región inicial (*opt score*). Finalmente, se ordenan todas las secuencias contenidas en la base de datos en función de los *scores* iniciales u optimizados y aquellas *N* mejores se alinean mediante un algoritmo de optimización (Needleman-Wunch-Sellers o Smith-Waterman).

BLAST utiliza también palabras (*words*), en este caso de longitud de tres, identificando aquellas con un *score*, evaluado con una matriz de sustitución, superior a un determinado valor frontera (*T*). Cada *hit* se extiende en ambas direcciones una determinada distancia (*X*) para ver si se pueden unir en un alineamiento mayor (*maximal segment pair*, MSP), que son de nuevo reevaluados. El programa devuelve el conjunto de alineamientos locales que excede un determinado *score* (*S*). La versión Gapped-BLAST introdujo además la opción de contemplar *gaps* en los MSPs. El valor de *S* se establece mediante análisis estadístico basado en la probabilidad de que un aminoácido se encuentre en una posición aleatoriamente y en la distribución Poisson que siguen los *scores* obtenidos en los MSPs. Se obtiene un valor de significancia *p*, que corresponde a la probabilidad de que un determinado segmento se identifique fortuitamente. Su fuerte fundamento estadístico, que le permite asignar cuantitativamente una significancia del resultado, junto con su mayor rapidez, ha convertido a BLAST en uno de los métodos más usados en la búsqueda de secuencias en bases de datos.

#### 1.5.2.2. Matrices de Sustitución

Tal y como se ha comentado previamente, los elementos de cada matriz especifican el coste a asignar a una comparación entre dos aminoácidos. Las más conocidas son:

- *Matriz de Identidad*: $H_{ij} = 1, i = j; H_{ij} = 0, i \neq j$

- *Matriz de código genético* (*Genetic Code Matrix*): El *score* está basado en el mínimo número de cambios en nucleótidos necesarios para convertir un aminoácido en otro (por ejemplo: de Met a Tyr se necesitan que las 3 posiciones de codon varíen para permitir la mutación).

- *Matrices de similitud fisicoquímica*: Se intenta cuantificar propiedades fisicoquímicas de los aminoácidos y arbitrariamente asignar costes basados en las similitudes de los residuos según estas características.

- *Matrices "log odds"*: incluyen información de sustituciones para obtener el alineamiento que mejor refleje la historia evolutiva. El valor *log odds*, $S_{ij}$ corresponde a la proporción

entre la probabilidad de que dos aminoácidos *i* y *j* se alineen por descendencia y la probabilidad de que lo hagan por casualidad (ecuación [1.68]). El valor $q_{ij}$ corresponde a la frecuencia observada en secuencias conocidas en las que se alinearon los aminoácidos (*i* y *j*) y $p_i$ y $p_j$ corresponden a las frecuencias observadas de los aminoácidos en un conjunto de secuencias.

$$S_{ij} = \log \frac{q_{ij}}{p_i p_j} \qquad [1.68]$$

El uso de estas matrices proporciona una predicción de la fiabilidad del alineamiento.
Dentro de las matrices *log odds*, las dos series más comunes son:

- Matrices PAM (*Accepted Point Mutation per 100 residues*)[235]. Denominadas también Dayhoff (ya que fueron creadas por Magaret Dayhoff) o MDM (*Mutation data Matrix*). Las probabilidades de cambio de un aminoácido en otro se derivan a partir de alineamientos globales de secuencias pertenecientes a una familia de proteínas relacionadas y al menos un 85% idénticas. A partir de ellas, se construye una matriz normalizada en valores que expresan la probabilidad de que un aminoácido de cada 100 sufra una mutación (PAM-1). El resto de matrices de la serie, para distancias evolutivas más grandes, se extrapola a partir de las de menor distancia. Así, si se suponen *N* mutaciones independientes, se multiplica la PAM-1 por sí misma *N* veces, obteniéndose las PAM160, PAM250... Existen otras matrices desarrolladas por otros grupos, que han seguido esta metodología o la han mejorado al utilizar otras bases de datos con más ejemplos. Jones y Thornton[236] derivaron las matrices PET91 a partir de 2621 familias de secuencias extraídas de SWISS-PROT, aunque es equivalente a una actualización de la PAM120. Gonnet *et al*[237] desarrollan la matriz GONNET, derivada por un proceso iterativo de alineamiento y refinamiento de la propia matriz. Sin embargo, parece que no se incrementa la habilidad del sistema para encontrar miembros de la mayoría de familias de proteínas[238].

- Matrices BLOSUM (*Blocks Substitution Matrix*)[238]. Creadas por Henikoff. Las probabilidades de sustitución se han derivado a partir de un conjunto de unos 2000 motivos conservados (*blocks*) encontrados en una base de datos de unas 500 proteínas relacionadas. Se trata de alineamientos locales, en los que no se han introducido *gaps*. Para asignar la probabilidad, se generan *clusters* de proteínas, de manera que todos aquellos motivos que presentan un 60% de identidad se agrupan en uno para evaluar las probabilidades de mutación y de allí se deriva la correspondiente BLOSUM60. Todas las matrices se calculan directamente, no se utilizan extrapolaciones. A medida se incrementa el porcentaje de identidad del *cluster*, la habilidad para diferenciar un alineamiento correcto de un alineamiento fortuito (entropía relativa) también se incrementa. Sin embargo, también se desvía el resultado, ya que se focaliza más la probabilidad en aquella proteína más probable. Por ello, la BLOSUM62 representa un compromiso bastante óptimo entre la capacidad de diferenciación y la focalización del resultado.

Las matrices PAM son más sensibles para alineamientos de secuencias con homólogos relacionados evolutivamente. Dentro de ellas, la matriz aconsejada depende del tipo de alineamiento a realizar: para una búsqueda en base de datos (BLAST/FASTA) se aconseja la PAM120 y para alineamientos de dos secuencias la PAM200. Otra posibilidad es la de utilizar combinaciones de ellas. Por ejemplo, para alinear dos secuencias, utilizar la PAM80 y PAM250 conjuntamente o bien la PAM120 y PAM320[239]. La serie de matrices BLOSUM generalmente es mejor que la serie PAM para la búsqueda de similitudes locales[240], ya que es posible encontrar alineamientos entre proteínas divergentes más en acuerdo con su

estructura tridimensional que la correspondiente PAM. La equivalencia entre una serie y otra es:

PAM100 ≥ BLOSUM90
PAM120 ≥ BLOSUM80
PAM160 ≥ BLOSUM60
PAM200 ≥ BLOSUM52
PAM250 ≥ BLOSUM45

Como regla, se mantiene:
- Las PAM*s* menores (PAM120) y BLOSUM*s* altas (BLOSUM80) se utilizan principalmente para alineamientos locales de regiones conservadas de alta similitud. (*Hard matrices*).
- Las PAMS mayores (PAM250) y BLOSUM menores (BLOSUM45) encuentran alineamientos entre regiones más largas y menos conservadas. (*Soft matrices*).

Respecto a los modelos de penalización de *gaps*, todavía no se ha desarrollado una teoría estadística completa acerca de los *gaps* en los alineamientos, por lo que los mejores costes para los *gaps* se han de determinar empíricamente para cada matriz y caso particular, aunque, en general, la penalización para abrir un *gap* es mayor que la de extenderlo.

### 1.5.3. Construcción del Modelo

Uno de los métodos para construir el modelo 3D por homología es el de modelización por satisfacción de restricciones espaciales (*modeling by satisfaction of spatial restraints*), que es el que implementa el programa MODELLER[131] utilizado en este proyecto. El proceso seguido por MODELLER para modelar la estructura tridimensional parte de la generación de un primer modelo crudo obtenido por transferencia de coordenadas entre todos los átomos equivalentes en el alineamiento de la secuencia diana y la proteínas plantilla e interpolación del resto de coordenadas indefinidas. Los métodos de modelización implementan una función potencial (*score*) que pretende ser equivalente a una función de energía (función de pseudoenergía), de manera que el valor mínimo de la misma corresponda con la conformación más probable de la proteína. La función pseudoenergética o función objetivo del MODELLER resulta de considerar una serie de restricciones, de manera que el mejor modelo sea aquel que viole el mímino número de ellas. Una vez se obtiene un modelo, es usual modelar *ab initio* los *loops*, en cuyo caso las coordenadas iniciales del primer modelo se obtienen aleatoriamente y no por transferencia de las coordenadas de las plantillas presentes en el alineamiento, aplicándose posteriormente una optimización de dichos *loops* según las restricciones calculadas para dicho segmento.

Estas restricciones son principalmente de dos tipos:

- Estereoquímicas: se calculan con métodos de mécanica molecular, derivadas a partir del *force field* CHARMM-22[241]. No están basadas en el alineamiento, ya que dependen únicamente del tipo de átomo y/o residuo. Incluyen los términos de enlace, ángulo, ángulo diedro y ángulos diedros impropios, que restringen la planaridad del enlace peptídico, los anillos de las cadenas laterales y los centros quirales y pro-quirales. También se incluyen los términos de interacciones no enlazantes, calculadas a partir de listas dinámicas de átomos vecinos, como las interacciones de van der Waals según un potencial de Lennard-Jones, solapamiento de esferas e interacciones electrostáticas de Coulomb. Además, se pueden incorporar restricciones para forzar conformaciones de hélice alfa, láminas beta y puentes de hidrógeno entre pares de láminas beta.

- Derivadas por homología: se obtienen a partir de las proteínas relacionadas estructuralmente presentes en el alineamiento e incluyen la distancia entre carbonos alfa, la distancia N-O, los ángulos de la cadena principal (ω, Φ, Ψ) y los ángulos de las cadenas laterales ($\chi_i$).

Las restricciones se expresan como funciones de densidad de probabilidad (*pdfs*, $p(x)$) para la propiedad restringida ($x$). La probabilidad finita de que una propiedad $x$ adopte un valor comprendido entre $x_1$ y $x_2$ se obtiene según la ecuación [1.69]:

$$p(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} p(x)dx \qquad [1.69]$$

por lo que es necesario establecer la función de probabilidad que mejor defina cada propiedad. La forma general de esta función de probabilidad corresponde a la ecuación [1.70], que indica que la probabilidad condicional de la propiedad $x$ viene determinado por los valores conocidos que adoptan otras propiedades ($a,b,c$ ...):

$$p(x/a,b,.....,c) \qquad [1.70]$$

Estas propiedades ($a,b,c$ ...), listadas en la Tabla 1.3, se establecieron empíricamente por correlación de características estructurales en una base de datos con 17 familias de proteínas respresentativas de las diferentes clases estructurales (clase α, clase β, clase α+β, clase α/β), alineadas estructuralmente[242]. La combinación es empírica, no tienen porqué tener un sentido físico, de forma que se ajustan las propiedades para definir cuáles de ellas tienen un significado estadístico en el valor que adopta $x$.

**Tabla 1.3.** Propiedades ($a,b,c$..) utilizadas para derivar las probabilidades condicionales de la propiedad $x$.

| |
|---|
| Tipo de residuo (aminoácido) |
| Ángulos diedros de la cadena principal ( Φ,Ψ) |
| Clase de estructura secundaria de este residuo |
| Clase de conformación de la cadena principal de este residuo |
| Contenido fraccional de los residuos que adoptan una determinada conformación de cadena principal |
| Ángulos diedros de la cadena lateral ($\chi_1$ , $\chi_2$, $\chi_3$, $\chi_4$) |
| Clase de conformación de los ángulos diedros |
| Accesibilidad de solvente de este residuo |
| Diferencia de vecindad de residuos entre dos residuos equivalentes en dos proteínas |
| Valor medio de la diferencia de vecindad entre dos proteínas |
| Identidad de secuencia relativa entre dos proteínas |
| Diferencia entre las distancias $C_\alpha$ - $C_\alpha$ entre residuos equivalentes en dos proteínas |
| Valor medio del factor de temperatura isotrópico de un residuo |
| Resolución de la estructura por rayos-X |
| Distancia media de un par de residuos intramoleculares a un *gap* en el alineamiento |

Esta función de probabilidad verdadera se calcula mediante la aproximación mostrada en la ecuación [1.71], donde $W_{x,a,b,c....}$ es la frecuencia relativa de ocurrencia de un valor $x$ de la propiedad espacial a restringir ($x$) en unas ciertas condiciones de valores de las propiedades $a,b,c...$, calculada a partir de la base de datos.

$$p(x/a,b,...,c) \approx W_{x,a,b,...c} \approx f(x,a,b,...c,q) \qquad [1.71]$$

La función analítica $f$ se construye de forma que se ajuste lo mejor posible a la tabla de valores de $W$ :

$$rms = \sqrt{\sum_{x,a,b...c} \left[ W_{x,a,b,...c} - f(x,a,b...,c,q) \right]^2} \qquad [1.72]$$

donde *q* adopta el valor que minimiza la función anterior, ajustada por mínimos cuadrados. La forma normal de estas funciones *f* es la de una gaussiana, aunque existen otras posibilidades como *splines* cúbicos, que pueden ser seleccionadas por el usuario para restricciones especiales.

De este modo, se obtienen las diferentes funciones de densidad de base (*basis pdf*), ya que permiten modelar una característica particular de la secuencia objetivo a partir de una única secuencia homóloga de estructura conocida. Para modelar las características a partir de varias estructuras homólogas, estas *basis pdf* se combinan en lo que se denominan funciones de probabilidad de parámetros o *feature pdfs*. Por ejemplo, en el caso de querer obtener la función de densidad para la distancia entre carbonos alfa ($C_\alpha$-$C_\alpha$) en una determinada proteína de estructura desconocida a partir de dos proteínas de estructura conocida (*A* y *B*), se debe combinar la función de probabilidad de base que describe la distancia *d'* entre los $C_\alpha$ de los residuos equivalentes en el alineamiento de la proteína *A* y la función de probabilidad de base de la distancia equivalente *d''* de la proteína *B*. Además, se deben tener en cuenta las restricciones estereoquímicas, por ejemplo, el criterio de van der Waals (distancia superior a la suma de los radios). En la Figura 1.11 se esquematiza el concepto.

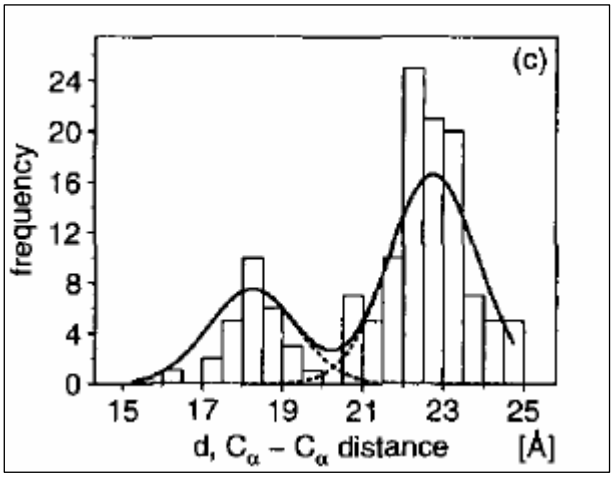



**Figura 1.11.** Derivación de una *feature pdf*. Extraído de [131].

La línea contínua corresponde a la *feature pdf* del parámetro distancia $C_\alpha$-$C_\alpha$ ($p^D(d)$) resultante de la suma de las *basis pdfs* (líneas discontínuas) correspondientes a cada una de las plantillas ($p_1^d(d)$ y $p_2^d(d)$).

Finalmente, se combinan todas las *feature pdf*s de los parámetros en una única función de probabilidad molecular, *molecular pdf*. Se asume que los diferentes parámetros son independientes (aunque es erróneo, porque por ejemplo, el valor de un ángulo Φ viene muy influenciado por el valor del ángulo Ψ), de forma que la *molecular pdf* (*P*) es el producto de las *feature pdfs* [$p^F(f_i)$], ecuación [1.73]:

$$P = \prod_i p^F(f_i) \qquad [1.73]$$

La *molecular pdf* expresa la probabilidad de ocurrencia de cualquier combinación de estos parámetros simultáneamente, a mayor valor de dicha probabilidad, mayor probabilidad de la estructura tridimensional. La optimización de este valor no se realiza sobre la función *P*, sino sobre su logaritmo neperiano, denominado función objetivo (*F*), ecuación [1.74], debido a que es más asequible computacionalmente transformar el productorio en un sumatorio.

$$F = -\ln P = g(f,a,b,c...) \qquad [1.74]$$

$$\frac{\delta F}{\delta f} \circ \frac{\delta}{\delta(x, y, z)} \qquad [1.75]$$

Así, el objetivo de maximizar *P* se convierte en el de minimizar *F*. Esta función *F* se deriva en función de los parámetros (*f*), que a su vez se expresan en función de las coordenadas cartesianas (ecuación [1.75]). Se expresa en kcal/mol, aunque no sea estrictamente correcto, al tratarse de una función de pseudoenergía.

Para optimizar dicha función, se aplica en primer lugar el *Variable Target Function Method* (VTFM) que consiste en una serie de minimizaciones de la función anterior realizadas con gradiente conjugado. La particularidad de dicho método es que parte de unas restricciones "locales", de manera que en cada ciclo de minimización se introducen más y más restricciones de mayor alcance, hasta llegar a la verdadera *molecular pdf*, que incorpora todas las restricciones. Para ello, utiliza un *schedule* (plan) de *n* ciclos, en el que se indica la amplitud del rango de residuos sobre los que actúa cada ciclo de la optimización junto con los factores de escalado de la desviación estándar de cada restricción (esto permite debilitar la importancia de ciertas restricciones frente a otras al aumentar la desviación, la restricción es más potente y una violación mayor es más probable). En la librería del MODELLER existen siete *schedules* diferentes, dependiendo de la exhaustividad con que se pretenda optimizar. Posteriormente, se realiza un *simulated annealing* con dinámica molecular.

#### 1.5.3.1. Construcción de los *loops ab initio*

El problema de la modelización de los *loops* se puede considerar como un problema reducido de plegamiento de proteínas. La conformación correcta de un segmento dado de una cadena polipeptídica tiene que ser calculado principalmente a partir de la secuencia de la propia cadena ya que, por ejemplo, segmentos de más de nueve residuos a veces tienen una conformación totalmente diferente en diversas proteínas.[243] Por lo tanto, la conformación de un segmento dado viene también influenciado el resto de la proteína o la estructura central que une el *loop*.

En general los métodos de modelización de *loops* se basan en los métodos *ab initio* ó de búsqueda en una base de datos. Sin embargo, este último presenta bastantes dificultades, como es el hecho de que sólo los segmentos de siete residuos o menos tienen representantes para cada una de las conformaciones que el segmento puede adoptar.[244]
En este proyecto se ha utilizado la rutina de construcción de *loops* implementada en el programa MODELLER[136], por lo que se describe brevemente.

Como se ha comentado, una vez obtenido un modelo, se pueden construir de manera independiente los *loops*, seleccionándose una serie de segmentos de residuos (*loops*) sobre los que se repite el proceso anterior con algunas modificaciones:

- Sobre los átomos seleccionados como *loops*, se generan las restricciones que actúan sobre ellos. La diferencia es que en este caso se calculan todas las restricciones (incluidas las de los ángulos Φ, Ψ, ω y χ) a partir de una librería y no como derivadas de homología con una plantilla (como es el caso de la modelización estándar por homología). A partir de ellas, se construye la función de pseudoenergía (*F*) que es del mismo tipo que la anteriormente descrita (ecuación [1.76]).

$$F = \sum_{enlaces} k_b (b - \bar{b})^2 + \sum_{angulos} k_\alpha (\alpha - \bar{\alpha})^2 + \sum_{diedros} \left| k_\phi \right| - b_t \cos(n\phi + \delta) + \sum_{impropios} k_i (\theta - \bar{\theta})^2 -$$
$$\sum_{\substack{diedros \\ cadena \\ lateral}} \ln p_s(\chi / R) - \sum_{residuos} \ln p_\omega(\omega / R) - \sum_{residuos} \ln p_m(\Phi, \Psi / R) + \sum_{\substack{atomos \\ enlazados \\ nounidos}} \xi \left[ E(a, a', d, \Delta_i) + S(r, r', d) \right]$$

[1.76]

$b$ = longitud enlace
$\alpha$ = ángulo enlace
$\phi$ = ángulo diedro
Φ, ψ, ω = ángulos de la cadena principal
χ = ángulos de la cadena lateral
$\theta$ = ángulo diedro impropio
$R$ = tipo de residuo
$\xi$ = factor de escalado que aparece en el *schedule* de la optimización
$a$, $a'$ = tipo de átomos en el par
$d$ = distancia entre átomos
$\Delta_i$ = diferencia entre índice de residuo en la secuencia
$r$, $r'$ = radios atómicos van der Waals

Los cuatro primeros términos corresponden a la ecuación del *force field* CHARMM[182] para las distancias de enlace, ángulos, ángulos diedros y ángulos impropios (la parametrización de las constantes de fuerza ($k_i$), valores en el punto de equilibrio, fase y periodicidad de los ángulos diedros también se han extraído de la versión CHARMM-22[241]).
Los tres términos siguientes de la ecuación [1.76] se extraen estadísticamente, de forma similar a lo explicado anteriormente, según la preferencia de cada residuo hacia un valor para los ángulos de la cadena principal y cadena lateral (ω, Φ Ψ, $\chi_i$).
El término energético de interacciones no enlazantes también está derivado estadísticamente, a partir de un potencial medio de fuerza dependiente de la distancia para pares de átomos en proteínas[245] (esta función de *score* es del tipo de las usadas en métodos *ab initio,* derivadas aplicando el teorema de Boltzmann).

- Una vez construidas las restricciones, se borran todas las coordenadas de dichos átomos del modelo de partida, de manera que se construyen aleatoriamente sus coordenadas de tal modo que los extremos N-terminal y C-terminal de cada segmento del *loop* constituyan el punto de anclaje, desde el cual hacer la búsqueda del espacio conformacional del *loop*.

- El conjunto de la optimización se realiza en dos partes: primero, se optimiza como si los átomos del *loop* "no sintieran" el entorno (dado que en la lista de interacciones de no enlace sólo se incluyen aquellos átomos que pertenecen al *loop*) y a continuación se optimiza en el contexto de toda la proteína (en la lista de interacciones se incluyen aquellos átomos que forman pares con átomos del *loop*, los situados a una distancia inferior a un *cutoff* de 4 Å).

## 1.6. Descriptores Moleculares

En una de las primeras publicaciones de estudios QSAR, realizada por Crum Brown y Frazer[246] en 1868, los autores relacionan la acción fisiológica ($\phi$) como una función de la constitución química ($C$), según la ecuación [1.77]:

$$\phi = f(C) \qquad [1.77]$$

Actualmente, el principal escollo en obtener una definición precisa de la función $f$ reside en la caracterización de los cambios en la estructura química que producen una determinada respuesta. La información estructural y propiedades fisicoquímicas se representan

numéricamente en descriptores que codifican a las moléculas. A pesar de la investigación teórica y experimental en este campo, no existe acuerdo acerca de aquel conjunto de descriptores óptimo, y dado que diferentes descriptores codifican distinta información, la estrategia consiste en aplicar aquellos más relevantes según la particularidad del caso de estudio. Los descriptores pueden ser tanto teóricos como experimentales, resultado de la cuantificación de una propiedad o de un procedimiento matemático y lógico que caracterice a una molécula.

Aunque en la introducción se han presentado en el apartado de búsquedas de similitud, los descriptores se utilizan en una amplia variedad de tareas, destacando las técnicas QSAR y predicción de propiedades, clasificación de compuestos, diseño de quimiotecas diversas, búsquedas de estructuras en bases de datos e interpretación de reactividad química y bioquímica.

En el cálculo y selección de descriptores existe básicamente un compromiso entre su eficacia y la eficiencia. La eficacia se entiende como la bondad de un descriptor en términos de diferenciar entre moléculas diferentes, mientras que la eficiencia hace referencia a la velocidad de cálculo asociada al descriptor. En este sentido, en el análisis de diversidad/similitud aplicado a quimiotecas con gran número de productos, descriptores como los basados en campos o los derivados de mecánica cuántica no son eficientes por su elevado coste computacional.

También se han introducido ya dos de los criterios más típicos según los cuales se clasifican los descriptores: el tipo de representación química requerida (1D, 2D, 3D…) y el tipo de codificación matemática. Además, se pueden clasificar en función de:

- La invariabilidad de sus propiedades, es decir, su capacidad para rendir un valor independiente de características particulares de la representación del compuesto. Estas propiedades son la invariabilidad química (tipos de átomos o enlaces), invariabilidad translacional y rotacional (en función del marco de referencia espacial) y la conformación de la representación geométrica. Los descriptores 3D que presentan invariabilidad translacional y rotacional son particularmente útiles, ya que no requieren el alineamiento previo de las moléculas, por lo que se ahorra tiempo de cálculo y se evitan problemas asociados con el alineamiento.

- Su 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 cualquier caso, no existe un único esquema de clasificación de los descriptores, aunque entre las propuestas más aceptadas destacan la de Todeschini[28], cuyo *handbook* se ha convertido en una de las referencias básicas del campo de descriptores. Diferentes esquemas pueden encontrarse también en las referencias [26] y [247].

En el presente trabajo se utilizan una gran variedad de descriptores, principalmente en el diseño de quimiotecas diversas. En estos casos, se suele incluir un gran número de descriptores no correlacionados ya que, al no estar dirigidas hacia una única diana particular, no se contemplan consideraciones específicas, sino todo lo contrario, se desea cubrir un amplio margen de propiedades ante distintas dianas. En este caso, se calculan los descriptores del programa MOE versión 2004.03 que incluye unos 200 descriptores, presentados de forma general en las siguientes secciones. En posteriores versiones de MOE, se ha ampliado este conjunto de descriptores, principalmente con descriptores mecanocuánticos.
Por otra parte, en las búsquedas de similitud se han calculado descriptores farmacofóricos basados en fragmentos 2D y 3D. Dado que se ha profundidazo más en su fundamento y aplicación, se describirán más detalladamente que los anteriores.

### 1.6.1. 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. Entre los más conocidos destacan los índices de conectividad molecular, propuestos por Randic[256] y desarrollados en profundidad por Hall y Kier[258-259]. Son rápidos de calcular y se ha comprobado que correlacionan con un amplio rango de propiedades biológicas.

En la Tabla 1.4 se recogen aquellos utilizados en este trabajo, junto con la palabra clave incluida en MOE para ellos.

**Tabla 1.4.** Índices topológicos utilizados en el trabajo.

| **ÍNDICES TOPOESTRUCTURALES** | |
|---|---|
| Índice de Zagreb[248] (*Zagreb*) | $Zagreb = \sum_i \delta_i^2$ |
| Índice de Wiener[249-251] (*weinerPath*) | $W = \frac{1}{2}\sum_i \sum_j d_{ij}$ |
| Número de Polaridad de Wiener[249] (*weinerPol*) | $\frac{1}{2}\sum_i \sum_j d_{ij}$ ; sobre $d_{ij} = 3$ |
| Índices de forma de Kier[252] de orden uno, dos y tres (*Kier1,Kier2,Kier3*) | ${}^1\kappa = \frac{A\cdot(A-1)^2}{({}^1P)^2}$ ; ${}^2\kappa = \frac{(A-1)\cdot(A-2)^2}{({}^2P)^2}$<br>${}^3\kappa = \frac{(A-3)\cdot(A-2)^2}{({}^3P)^2}$ ; si A es par, (A>3)<br>${}^3\kappa = \frac{(A-1)\cdot(A-2)^2}{({}^3P)^2}$ ; si A es impar (A>3) |
| Índice de Balaban[253,254] (*balabanJ*) | $J = \frac{B}{C+1}\cdot\sum_b (\sigma_i\cdot\sigma_j)^{-1/2}_b$ ; $C = B - A + 1$ |
| Diámetro Topológico (*diameter*) | $D = \max_i \eta_i$ |
| Radio Topológico (*radius*) | $R = \min_i \eta_i$ |
| Índice de Petitjean[255] (*petitjean*) | $I_2 = \frac{D-R}{R}$ $\quad 0 \le I_2 \le 1$ |

| **ÍNDICES TOPOQUÍMICOS** | |
|---|---|
| Índices de Conectividad[256, 257] de orden cero, uno y dos (*chi0,chi1*) | ${}^0\chi = \sum_i \delta_i^{-1/2}$ ; ${}^1\chi = \sum_{enlaces} (\delta_i\cdot\delta_j)^{-1/2}$ ;<br>${}^2\chi = \sum_{k=1}^{2-path} (\delta_i\cdot\delta_j\cdot\delta_k)_k^{-1/2}$ |
| Índices de Conectividad de valencia[258,259] de orden cero, uno y dos (*chi0v,chi1v*) | ${}^0\chi^v = \sum_i (\delta_i^v)^{1/2}$ ; ${}^1\chi^v = \sum_i (\delta_i^v\cdot\delta_j^v)^{1/2}$ ; |

| | |
|---|---|
| | $^{2}\chi^{v} = \sum_{k=1}^{2-path} (\delta_i \cdot \delta_j \cdot \delta_k)_k^{-1/2}$ |
| Índices de Forma de Kier[260] modificados<br><br>(*KierA1*,*KierA2*,*KierA3*) | $^{1}\kappa_\alpha = \frac{(A+\alpha)\cdot(A+\alpha-1)^2}{(^{1}P+\alpha)^2}$ ;<br>$^{2}\kappa_\alpha = \frac{(A+\alpha-1)\cdot(A+\alpha-2)^2}{(^{2}P+\alpha)^2}$<br>$^{3}\kappa_\alpha = \frac{(A+\alpha-3)\cdot(A+\alpha-2)^2}{(^{3}P+\alpha)^2}$ ; si A es par, (A>3)<br>$^{3}\kappa_\alpha = \frac{(A+\alpha-1)\cdot(A+\alpha-2)^2}{(^{3}P+\alpha)^2}$ ; si A es impar (A>3) |
| Índice de Flexibilidad Molecular de Kier[261]<br>(*KierFlex*) | $\phi = \frac{^{1}\kappa_\alpha \cdot ^{2}\kappa_\alpha}{A}$ |

| **ÍNDICES TOPOLÓGICOS BASADOS EN LA TEORÍA DE LA INFORMACIÓN** | |
|---|---|
| Contenido de Información de un sistema con *n* elementos<br>(*a_IC*) | $I_C = \sum_{g=1}^{G} n_g \cdot \log_2 n_g$ |
| Contenido medio de información | $I = n \cdot \log_2 n - \sum_{g=1}^{G} n_g \cdot \log_2 n_g$ |
| Índice de Contenido medio de información de igualdad de adyacencia<br>(*VAdjEq*) | $^{V}\bar{I}_{adj}^{E} = -\frac{2B}{A}\cdot\log_2\left(\frac{2B}{A}\right) - \left(1-\frac{2B}{A}\right)\cdot\log_2\left(1-\frac{2B}{A}\right)$ |
| Índice de Contenido medio de información de magnitud de adyacencia<br>(*VAdjMa*) | $^{V}\bar{I}_{adj}^{M} = -2B\left(\frac{1}{2B}\cdot\log_2\frac{1}{2B}\right) = 1+\log_2 B$ |
| Índice de Contenido medio de información de igualdad de distancia<br>(*VDistEq*) | $^{V}\bar{I}_{D}^{E} = -\sum_{g=1}^{D}\frac{2\cdot^{g}f}{A\cdot(A-1)}\cdot\log_2\frac{2\cdot^{g}f}{A\cdot(A-1)}$ |
| Índice de Contenido medio de información de magnitud de distancia<br>(*VDistMa*) | $^{V}\bar{I}_{D}^{M} = -\sum_{g=1}^{D} {}^{g}f\cdot\frac{g}{W}\cdot\log_2\frac{g}{W}$ |
| Índice de Contenido medio de información de igualdad de adyacencia de arista | $^{E}\bar{I}_{adj}^{E} = -\frac{2N_2}{B^2}\cdot\log\left(\frac{2N_2}{B^2}\right) - \left(1-\frac{2N_2}{B^2}\right)\cdot\log_2\left(1-\frac{2N_2}{B^2}\right)$ |
| Índice de Contenido medio de información de magnitud de adyacencia de arista | $^{E}\bar{I}_{adj}^{M} = 1+\log N_2$ |
| Índice de Contenido medio de información de igualdad de distancia de arista | $^{E}\bar{I}_{D}^{E} = -\sum_{g=1}^{D}\frac{2\cdot^{g}f}{B\cdot(B-1)}\cdot\log_2\frac{2\cdot^{g}f}{B\cdot(B-1)}$ |
| Índice de Contenido medio de información de magnitud de distancia de arista | $^{E}\bar{I}_{D}^{M} = -\sum_{g=1}^{D} {}^{g}f\cdot\frac{g}{^{E}W}\cdot\log_2\frac{g}{^{E}W}$ ;<br>$^{E}W = \frac{1}{2}\sum_{i=1}^{B}\sum_{j=1}^{B} {}^{E}d_{ij}$ Índice de Wiener de aristas |
| Índice de Información total de la composición atómica | $I_{ACT} = -A^h \cdot \sum_g \frac{A_g}{A^h}\cdot\log_2\frac{A_g}{A^h}$ |
| Índice de Información o Entropía de Shanon<br>(*a_ICM*) | $IC_r = \mathrm{H} = -\sum_{g=1}^{G} p_g \log_2 p_g$ |
| Índice de contenido de Información Estructural | $SIC_r = \frac{IC_r}{\log_2 A}$ |

| | |
|---|---|
| Índice de contenido de Información de enlace | $BIC_r = \frac{IC_r}{\log_2\left(\sum_{b=1}^{B}\pi_b^*\right)}$ |
| Índice de Información complementario | $CIC_r = \log_2 A - IC_r$ |

Las distintas definiciones corresponden a:
$\delta_i$: grado de vértice del átomo *i* (número de átomos pesados adyacentes al átomo). $d_{ij}$: distancia o número de aristas del camino más corto entre dos átomos. *A*: número de vértices. *B*: número de enlaces. *C*: número de anillos independientes y no solapados. ${}^mP$: número de trayectorias, caminos sin átomos repetidos de orden *m* que se definen según el número de aristas implicadas. $\sigma_i$: suma de los elementos de la fila *i* de la matriz de distancias, que contiene las distancias entre todos los pares de vértices de un grafo. $\eta_i$: excentricidad atómica, valor máximo de la fila *i* de la matriz de distancias. $\delta^v$: grado de vértice de valencia, definido como el número de electrones de valencia menos el número de átomos de hidrógeno enlazados. $\alpha$: mide la relación entre el radio de covalencia del átomo *i* relativo al radio del carbono en configuración $sp^3$. *G*: número de clases de equivalencia *g* en el sistema a partir de la definición de tipos de relación de elementos del conjunto. $n_g$: número de elementos de la clase *g*. $p_g$: probabilidad de seleccionar aleatoriamente un elemento de la clase *g* ($p_g=n_g/n$). ${}^gf$: número de distancias iguales en la submatriz triangular de distancias. $N_2$: número de trayectorias de orden dos. $A^h$: número total de átomos, incluyendo hidrógenos. $A_g$: número de átomos pertenecientes al mismo elemento químico. $\pi_b^*$: orden del enlace *b*.

### 1.6.2. Descriptores de forma

**Tabla 1.5.** Descriptores de Forma.

| | |
|---|---|
| Globuralidad (*glob*) | Número de condición inverso (menor valor propio/mayor valor propio) de la matriz de covarianza de las coordenadas atómicas. Un valor de uno corresponde a una esfera perfecta y un valor de 0 a una molécula mono- o bidimensional. |
| Momento de Inercia Principal (*pmi*) | $I = \sum_{i=1}^{A} m_i \cdot r_i^2$ ; $m_i$ (masa atómica), $r_i$ (distancia ⊥ del átomo *i* al eje) |
| Radio de Giro (*rgyr*) | $Rg_1 = \sqrt{\frac{\sum_{i=1}^{A} r_i^2}{A}}$ ; $Rg_2 = \sqrt{\frac{\sum_{i=1}^{A} m_i \cdot r_i^2}{MW}}$ |
| Primera, Segunda y Tercera Dimensión Standard (*std_dim1*,*std_dim2*,*std_dim3*) | Raíz cuadrada del primer, segundo y tercer valor propio mayor de la matriz de covarianza de coordenadas atómicas. Equivalente a la desviación estándar a lo largo de los ejes de componentes principales. |
| Superficie Molecular (VSA) | Área de la superficie de van der Waals. Se puede calcular según una representación poliédrica para cada átomo (*VSA*) o mediante una tabla de conexiones (*vdw_area*) |
| Volumen molecular de van der Waals ($VM_{VDW}$) | Volumen delimitado por la superficie molecular. Se puede calcular mediante una aproximación en mallas (*vol*) o una aproximación mediante una tabla de conexiones (*vdw_vol*). |

### 1.6.3. Descriptores de propiedades fisicoquímicas

Estos descriptores son los más aplicados en técnicas QSAR. Se clasifican también tradicionalmente en varias categorías, según describan propiedades hidrofóbicas, estéricas y efectos electrónicos. Además, se incluyen también los descriptores mecanocuánticos (energías del HOMO y el LUMO, entalpía de formación, potencial de ionización, energía electrónica,

energía de solvatación…) o propiedades estructurales (peso molecular, número de enlaces rotables, número de centros quirales…). En la Tabla 1.6 se detallan aquellos utilizados en este trabajo, junto con la palabra clave incluida en MOE para ellos.

**Tabla 1.6.** Descriptores de propiedades fisicoquímicas.

| | |
|---|---|
| Peso Molecular (*Weight*) | Descriptor 0D, reflejo del tamaño molecular y tipo de átomos constituyentes del compuesto. |
| Momento Dipolar (*Dipole*) (AM1_dipole, MNDO_dipole, PM3_dipole) | Descriptor electrónico 3D, codifica el desplazamiento respecto al centro de gravedad de densidad de cargas parciales positivas y negativas. Es el ejemplo más simple de un descriptor libre de alineamiento, ya que no depende de la orientación absoluta en el espacio. |
| Suma de Polarizabilidades atómicas (*apol*) | Descriptor electrónico. La polarizabilidad atómica ($\alpha_i$) corresponde a la relación entre el momento dipolar inducido en un átomo y el campo eléctrico inductor. La suma de polarizabilidades atómicas es una buena aproximación a la polarizabilidad molecular. |
| *bpol* | $bpol = \sum_{i=1}^{A}\sum_{j>i}^{A}\left\|\alpha_i - \alpha_j\right\|$ donde $\alpha_i$ es la polarizabilidad atómica |
| Densidad (*density ó dens*) | Relación tener el peso y el volumen molecular ($VM_{VDW}$) Dependiendo del método de cálculo de $VM_{VDW}$, MOE distingue dos densidades: i) *density*, con *vdw_vol* y ii) *dens* con *vol* |
| Logaritmo del Coeficiente de Partición Octanol/Agua. (*SlogP* ó *logP(o/w)*) | Es el descriptor más recurrido para efectos hidrofóbicos, junto con el parámetro de hidrofobicidad $\pi$ desarrollado por Hansch. Existen varios modelos para calcularlo, normalmente por modelos fragmentales, donde se adicionan contribuciones atómicas, definidas para cada tipo de átomo e hibridación. MOE dispone del método de Ghose-Crippen[262,263] (*SlogP*) y de un modelo lineal basado en tipos atómicos ajustado sobre 1847 moléculas (*logP(o/w)*)[264]. |
| Refractividad Molecular(MR) (*SMR* ó *mr*) | Descriptor estérico, definido por la ecuación de Lorentz-Lorenz: $MR = \frac{n^2-1}{n^2+2}\cdot\frac{MW}{d}$ siendo *n* el índice de refracción. MR es una propiedad aditiva-constitutiva, por lo que se puede calcular mediante modelos aditivos de contribución atómica de Ghose-Crippen (*SMR*[262]) o a partir de un modelo lineal de once descriptores sobre 1947 moléculas (*mr*[265]) |
| Energía del HOMO (*AM1_HOMO*) | Medida de la nucleofilia de un compuesto (reactividad). |
| Energía del LUMO (*AM1_LUMO*) | Medida de la electrofilia de un compuesto (reactividad). |
| Energía total y electrónica de la molécula (*AM1_E, MNDO_E, PM3_E*) (*AM1_Eele,MNDO_Eele,PM3_Eele*) | Calculada en distintos Hamiltonianos (MNDO, PM3, AM1) con el programa MOPAC[267]. |
| Entalpía de Formación (*AM1_HF, MNDO_HF, PM3_HF*) | Medida de la estabilidad térmica de un compuesto. Calculada en distintos Hamiltonianos (MNDO, PM3, AM1) en el programa MOPAC[267]. |
| Potencial de Ionización (*AM1_IP, MNDO_IP,PM3_IP*) | Medida de la estabilidad térmica de un compuesto. Calculada en distintos Hamiltonianos (MNDO, PM3, AM1) en el programa MOPAC[267]. |
| *FCharge* | Carga total de la molécula (suma de cargas parciales) |
| Superficie molecular Accesible (*ASA*) | Descriptor 3D, corresponde a la superfice accesible al agua calculada mediante una esfera-sonda de 1.4 Å. |
| Área de la superficie Polar (*TPSA*) | Descriptor 2D, se calcula a partir de contribuciones de grupo, según la parametrización de Ertl *et al.*[267] |
| Descriptores derivados de la Energía | Desde la propia energía potencial según el *force field* (*E*) a |

| | |
|---|---|
| Potencial o *Force Field* | términos de ella (*E_ang, E_ele, E_nb, E_sol, E_str, E_vdw, E_tor, E_stb*…). |
| Descriptores de Carga Parcial | En MOE, se pueden utilizar las cargas parciales calculadas previamente (serie *Q_**) o calcularse mediante el método PEOE (*Partial Equalization of Orbital Electronegativities*) de Gasteiger[170] (serie de descriptores *PEOE_**), basado únicamente en topología.<br>Estos descriptores comprenden la suma de cargas parciales positivas (*Q_PC+*, *PEOE_PC+*), de cargas parciales negativas (*Q_PC-*, *PEOE_PC-*) y sus correspondientes valores relativos (*Q_RPC+*, *PEOE_RPC+*, *Q_RPC-*, *PEOE_RPC-*). |

**DESCRIPTORES DE CARGA PARCIAL Y ÁREA DE SUPERFICIE**

Combinan la información electrónica y de forma. Dentro de MOE, se diferencian dos subconjuntos dependiendo de cómo cuantifiquen la forma de cada átomo: i) para cada átomo se cuantifica, una superficie de van der Waals ($\nu_i$) según una tabla de conexiones (descripción 2D, serie **_VSA_**) o ii) se calcula el área de la superficie molecular accesible ($SA_i$) a partir de una esfera-sonda de 1.4 Å (descripción 3D, serie **_ASA_**)[268].

| | |
|---|---|
| *Q_VSA_POS*<br>*PEOE_VSA_POS* | Área positiva total de la superficie de van der Waals. Suma de todos los $\nu_i$ cuya carga parcial $q_i>0$. |
| *Q_VSA_NEG*<br>*PEOE_VSA_NEG* | Área negativa total de la superficie de van der Waals. Suma de todos los $\nu_i$ cuya carga parcial $q_i<0$. |
| *Q_VSA_PPOS*<br>*PEOE_VSA_PPOS* | Área positiva polar total de la superficie de van der Waals. Suma de todos los $\nu_i$ cuya carga parcial $q_i > 0.2$ |
| *Q_VSA_PNEG*<br>*PEOE_VSA_PNEG* | Área negativa polar total de la superficie de van der Waals. Suma de todos los $\nu_i$ cuya carga parcial $q_i < -0.2$ |
| *Q_VSA_HYD*<br>*PEOE_VSA_HYD* | Área hidrofóbica total de la superficie de van der Waals. Suma de todos los $\nu_i$ cuya carga parcial absoluta $\|q_i\| \leq 0.2$ |
| *Q_VSA_POL*<br>*PEOE_VSA_POL* | Área polar total de la superficie de van der Waals. Suma de todos los $\nu_i$ cuya carga parcial absoluta $\|q_i\| > 0.2$ |
| *Q_VSA_FPOS*<br>*PEOE_VSA_FPOS* | Área positiva fraccional de la superficie de van der Waals. Relación entre **_VSA_POS* y *VSA*. |
| *Q_VSA_FNEG*<br>*PEOE_VSA_FNEG* | Área negativa fraccional de la superficie de van der Waals. Relación entre **_VSA_NEG* y *VSA*. |
| *Q_VSA_FPPOS*<br>*PEOE_VSA_FPPOS* | Área positiva polar fraccional de la superficie de van der Waals. Relación entre **_VSA_PPOS* y *VSA*. |
| *Q_VSA_FPNEG*<br>*PEOE_VSA_FPNEG* | Área negativa polar fraccional de la superficie de van der Waals. Relación entre **_VSA_PNEG* y *VSA*. |
| *Q_VSA_FHYD*<br>*PEOE_VSA_FHYD* | Área hidrofóbica fraccional de la superficie de van der Waals. Relación entre **_VSA_HYD* y *VSA* |
| *Q_VSA_FPOL*<br>*PEOE_VSA_FPOL* | Área polar fraccional de la superficie de van der Waals. Relación entre **_VSA_POL* y *VSA* |
| *ASA+* | Superficie accesible al agua de todos los átomos con carga parcial positiva. Suma de todos los $SA_i$ cuya carga parcial $q_i>0$. |
| *ASA-* | Superficie accesible al agua de todos los átomos con carga parcial negativa. Suma de todos los $SA_i$ cuya carga parcial $q_i<0$. |
| *ASA_H* | Superficie accesible al agua de todos los átomos hidrofóbicos. Suma de todos los $SA_i$ cuya carga parcial $\|q_i\|<0.2$ |
| *ASA_P* | Superficie accesible al agua de todos los átomos polares. Suma de todos los $SA_i$ cuya carga parcial $\|q_i\|\geq0.2$ |
| *DASA* | Valor absoluto de la diferencia entre *ASA+* y *ASA-* |
| *CASA+* | Superficie accesible al agua ponderada por la carga parcial positiva. Producto de *ASA+* por la máxima carga positiva[268]. |
| *CASA-* | Superficie accesible al agua ponderada por la carga parcial negativa. Producto de *ASA-* por la máxima carga negativa[268]. |
| *DCASA* | Valor absoluto de la diferencia entre *CASA+* y *CASA-*[268]. |
| *FASA+* | *ASA+* fraccional: *ASA+* / *ASA* |

| | |
|---|---|
| *FASA-* | *ASA-* fraccional: *ASA-* / *ASA* |
| *FCASA+* | *CASA+* fraccional: *CASA+* / *ASA* |
| *FCASA-* | *CASA-* fraccional: *CASA-* / *ASA* |
| *FASA_H* | *ASA_H* fraccional: *ASA_H*/ *ASA* |
| *FASA_P* | *ASA_P* fraccional: *ASA_P* / *ASA* |

Además, para una determinada propiedad aditiva-constitutiva, se pueden obtener los descriptores denominados *Subdivided Surface Areas*. Para cada átomo, se calcula una superficie de van der Waals aproximada ($v_i$), según una tabla de conectividades (descriptores 2D). El rango de una propiedad fisicoquímica (*SlogP*, *SMR*, cargas parciales) se divide en distintos *bins* que comprenden un rango de valores y se cuenta la contribución de $v_i$ de todos aquellos átomos cuya contribución atómica a la propiedad en cuestión se encuentre dentro de este rango. Así, en MOE se definen las series: *SlogP_VSA* (*SlogP* distribuido en diez *bins*), *SMR_VSA* (SMR dividido en ocho *bins*) y *PEOE_VSA* (cargas parciales partidas en catorce *bins*).

### 1.6.4. Descriptores *count-based*

Simplemente cuentan instancias de los *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 fisicoquímicas o farmacofóricas (número de enlaces rotables o de aceptores/dadores de puente de hidrógeno). En la Tabla 1.7 se muestran aquellos más relevantes implementados en MOE.

**Tabla 1.7.** Descriptores *count-based*.

| | |
|---|---|
| *Reactive* | Indicador de la presencia de grupos reactivos, basados en el conjunto propuesto por Oprea[17]. |
| *b_count* | Número de enlaces |
| *b_rotN* | Número de enlaces rotables |
| *b_rotR* | Fracción de enlaces rotables |
| *b_1rotR* | Número de enlaces rotables simples (no forma parte de un anillo y no es conjugado) |
| *b_1rotR* | Fracción de enlaces rotables simples |
| *b_ar* | Número de enlaces aromáticos |
| *b_single* | Número de enlaces simples. |
| *b_doble* | Número de enlaces dobles |
| *b_triple* | Número de enlaces triples |
| *a_acc* | Número de aceptores de puente de hidrógeno (incluyendo átomos que actúan tanto como aceptores o dadores, -OH). |
| *a_acid* | Número de átomos acídicos |
| *a_base* | Número de átomos básicos |
| *a_don* | Número de dadores de puente de hidrógeno (incluyendo átomos que actúan tanto como aceptores o dadores, -OH). |
| *a_hyd* | Número de átomos hidrofóbicos |

### 1.6.5. Descriptores Farmacofóricos basados en fragmentos 2D y 3D

El origen de los descriptores o *fingerprints* farmacofóricos se encuentra en las búsquedas de subestructuras en un espacio bidimensional. Así, las dos primeras aplicaciones de búsquedas de similitud aparecieron a mediados de los '80[269,270], donde se introduce el concepto de pares atómicos o *atom pairs*, definidos a partir de un par de tipos atómicos y la distancia entre ellos, definida a partir del mínimo recorrido en el grafo.

En un principio, el tipo atómico se define a partir del elemento atómico, el número de enlaces con átomos pesados y el número de enlaces π. Esta definición de tipo atómico se amplia en sucesivos trabajos de modo que no sea tan restrictiva y específica. Además del concepto de pares atómicos, otros fragmentos 2D típicos de subestructuras son: el átomo aumentado (*augmented Atom*), la secuencia atómica (*atom sequence*), la secuencia de anillo (*ring sequence*) y la torsión topológica (*topological torsion*).

Posteriormente, se amplía el concepto de fragmento 2D a su correspondiente equivalente tridimensional, diferenciándose entre aquellos *fingerprints* basados en distancias o en ángulos.

- Métodos basados en distancias:

  En 1991, Pepperrell *et al*[271] desarrollan un método basado en la distribución de distancias entre pares de átomos. Se realiza una partición en rangos de la distancia, obteniéndose distintos *bins*, y cada una de las posibles distancias en una molécula contribuye con un valor de uno al *bin* que incluye esta distancia. La distribución de frecuencias resultante se utiliza para describir la molécula.

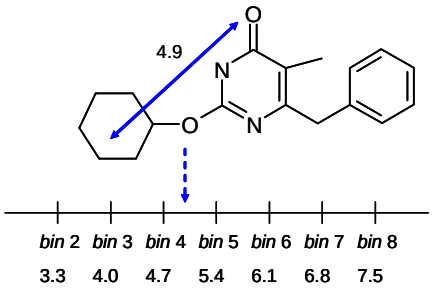



**Figura 1.12.** Ejemplo de asignación de un *atom pair* a un *binning scheme* en función de la distancia medida.

  En 1992, Bemis y Kuntz[272] describen también un método basado en distribución de distancias ampliando el concepto a tripletes de átomos. A partir de la estructura tridimensional del compuesto, se construye la matriz de distancias interatómicas y se analizan cada una de las combinaciones de tres átomos posibles. Para cada triplete, caracterizado por distancias entre sí de $n_1$, $n_2$ y $n_3$ se calcula el valor del perímetro del correspondiente triángulo según la ecuación [1.78], asignándose este valor a un *bin* de una distribución compuesta por 64 celdas.

$$P = n_1^2 + n_2^2 + n_3^2 \qquad [1.78]$$

  Las distribuciones se comparan entre sí mediante el coeficiente de Tanimoto. Esta codificación de la distancia, se amplía posteriormente a combinaciones de dos átomos y cuatro átomos[273].
  Posteriormente, Nilakantan[274] utiliza también las tres distancias interatómicas ($n_1$, $n_2$ y $n_3$) para cada conjunto posible de tres átomos pesados (tripletes), y según las ecuaciones [1.79] y [1.80] calculan un valor entero que caracteriza al triplete. La comparación de códigos entre moléculas la realiza mediante el coeficiente de Dice y un coeficiente de asimetría.

$$n_1 \le n_2 \le n_3 \qquad [1.79]$$

$$n_1 + 1000 \times n_2 + 1000000 n_3 \qquad [1.80]$$

- Métodos basados en ángulos:

  Bath[273] propone dos tipos de medidas basadas en ángulos a partir de la torsión de cuatro átomos: *A-B-C-D* (Figura 1.13). En la primera de ellas, *BNB measure*, se consideran todas

aquellas posibles torsiones del tipo *A-B≈C-D* en las que los pares *A-B* y *C-D* están enlazados, pero no los átomos *B-C*. El índice se establece a partir de la media aritmética de los ángulos *ABC* y *BDC* ($n_1$), el valor absoluto de la torsión ($n_2$) y la distancia interatómica *B-C* ($n_3$), según la ecuación [1.81]:

$$n_1 + 180 \times n_2 + 180^2 \times n_3 \qquad [1.81]$$

De manera análoga a los métodos basados en distancias, se generan todos los posibles índices de todos los fragmentos *BNB* de una molécula referencia y su distribución se compara, mediante el coeficiente de Tanimoto, con la del resto de moléculas de la base de datos.

La otra medida, denominada *NBN measure*, considera todas las posibles torsiones del tipo *A≈B-C≈D*, en la que únicamente está enlazado el par *B-C*. El código en este caso se establece a partir del valor del ángulo diedro ($n_1$) y la suma de las aristas de los triángulos formados por *ABC* ($n_2$) y *ACD* ($n_3$), redondeados a su entero más próximo, según la ecuación [1.82]:

$$n_1 + 10 \times n_2 + 1000 \times n_3 \qquad [1.82]$$

De nuevo, para cada molécula se obtiene la distribución de códigos de cada uno de los posibles fragmentos *NBN*.

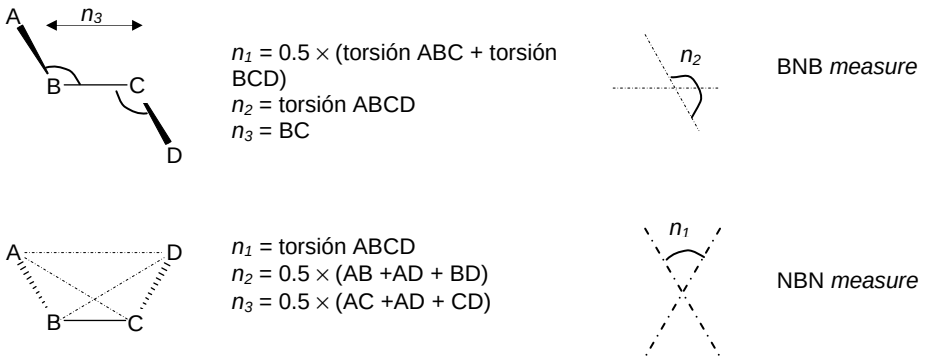



**Figura 1.13.** Descripción de los fragmentos BNB y NBN. Adaptado de [273].

El siguiente paso en el desarrollo de *fingerprints* farmacofóricos consistió en la ampliación de la definición de tipo de átomo según su dependencia a una clase. Estas clases, se establecen a partir de criterios fisicoquímicos o farmacofóricos.

Good y Kuntz[275] proponen la reducción del número de puntos posibles trabajando con cinco tipos atómicos en lugar de todos los átomos constituyentes de la molécula. Construyen tripletes de estos átomos, con las distancias medidas en el espacio Euclídeo y donde cada triplete queda caracterizado por: i) el perímetro del triángulo formado por los tres átomos, almacenado en una partición de 4 *bytes* y ii) la desviación de este triángulo respecto a un triángulo equilátero, cuantificada en términos de la relación del área del triángulo obtenido con el área máxima de un triángulo equilátero. Esta relación se áreas se parte en un espacio de 10 *bytes*. Los cinco tipos atómicos generan un total de 35 posibles tripletes, de forma que el espacio de almacenamiento total por molécula es de 1400 *bytes* (10×4×35).

Sheridan[276] introduce en el año 1996 lo que define como *binding property pair*, en el que cada átomo se clasifica según siete posibles tipos: catión, anión, dador de puente de hidrógeno, aceptor de puente de hidrógeno, polar, hidrofóbico u otro. A partir de ellos, se establecen los pares atómicos medidos en distancia Euclídea.

El rango de distancias se particiona en *bins* e incluye la ponderación de la contribución de cada distancia de un par atómico a cada *bin* en función de su cercanía al centro de los *bins* vecinos. Así, un par atómico puede ocupar más de un *bin*. Por ejemplo, para una partición con un *bin1* centrado en 4.1 Å y otro *bin2* centrado en 4.9 Å, un par atómico con distancia 4.7 Å contribuirá con 0.25 al *bin1* y con 0.75 al *bin2*.

Similarmente, el grupo de Abbot laboratorios[29] desarrolla dos descriptores basados en puntos potenciales farmacofóricos (*potencial phamacophore point*, PPP): PPP-*pairs* y PPP-*triangles*. Definen cinco tipos atómicos según el programa 3D-FEATURES: dadores y aceptores de puente de hidrógeno, átomos positiva y negativamente cargados y átomos hidrofóbicos.
El descriptor PPP-*pairs* codifica la información de las distancias Euclídeas contenidas en todos los posibles pares de PPPs según tres esquemas:

- Una cadena de bits se divide en secciones según los valores de mínimo, máximo y anchura definidos por el usuario.

- Se permite el solapamiento de los *bins*: cada *bin* viene codificado por dos bits. En el primer *bin* se asigna un uno si la distancia medida corresponde al rango de valores que codifica. En el segundo bit del *bin* se coloca un uno si la distancia no cae en los límites del *bin*, si lo hace, entonces se coloca el uno en el segundo bit del *bin* contiguo. El solapamiento se especifica según un porcentaje de la anchura del *bin*.

- En lugar de utilizar una partición en *bins* equifrecuentes, la anchura de éstos se deduce de la distribución de frecuencias de distancias interatómicas en una base de datos. Así, se define la posición de un bit en un *bin* según la ecuación [1.83]:

$$\text{Número\_Bin} = (\text{int})\left( 5 \cdot \tan^{-1}\left( \frac{\text{Distancia\_PPP} - 3}{2} \right) + 6 \right) \qquad [1.83]$$

El descriptor PPP-*triangles* codifica todas las combinaciones de tripletes entre PPPs presentes en una molécula. Cada una de las 35 posibles combinaciones se coloca en un bit de una cadena según su distancia Euclídea. La partición en *bins* se realiza según un valor mínimo (2 Å), máximo (15 Å) y con una anchura de *bin* de 1 Å. Debido al elevado número de bits necesario para codificar cada molécula, la codificación se realiza en forma de *hashed fingerprint*, (véase Introducción) reduciéndose el almacenamiento en memoria.

Otra aproximación basada en tripletes de puntos farmacofóricos 3D es la propuesta por Pickett[277] en el módulo ChemDiverse. En este caso, se definen 6 PPPs (aceptor/dador de puente de hidrógeno, átomos ácidos y básicos, centros aromáticos e hidrofóbicos). La partición de distancias en *bins* comprende seis rangos: 2-4.5, 4.5-7, 7-10, 10-14, 14-19 y de 19-24 Å. Estos descriptores se aplican en el diseño de quimiotecas diversas.

El programa PharmPrint[278] calcula también *fingerprints* farmacofóricos basados en tripletes según el esquema de partición en distancias presentado por Pickett, aunque incorpora dos restricciones para reducir el número de combinaciones posibles: i) la regla del triángulo, de manera que la longitud de cada lado de un triángulo no supere la suma de las longitudes de las otras dos aristas y ii) elimina aquellos farmacóforos redundantes por simetría. Se describe su aplicación hacia el diseño focalizado de quimiotecas.

Finalmente, en 1999 Mason y colaboradores[279] introducen un *fingerprint* basado en 4 puntos, dirigido a ampliar la resolución de los descriptores farmacofóricos y permitiendo la inclusión de la quiralidad. Consideran siete *features* farmacofóricas (aceptores/dadores de puente de hidrógeno, centros ácidos, centros básicos, regiones hidrofóbicas, centros aromáticos y una caracterización extra). Esta caracterización es flexible, permitiendo la definición de puntos

especiales diseñados específicamente para codificar subestructuras privilegiadas sobre dianas particulares. Permite también la generación de *fingerprints* complementarios al sitio activo de una proteína. Incorporan dos esquemas de partición de distancias en siete o diez rangos, cuyos tamaños se fijan según un porcentaje fijo de variación a partir del centro del rango (±15%), de manera que distancias mayores se correspondan a rangos de valores superiores.

En la Figura 1.14 se esquematizan los farmacóforos basados en dos, tres y cuatro puntos. Mientras que un par atómico queda caracterizado por una distancia, un triplete necesita tres distancias y el tetraedro, seis. El elevado número de combinaciones posibles en los tetraedros, generando cadenas de bits con gran requerimiento de memoria, conduce a que el esquema de partición de distancias incluya menos rangos.

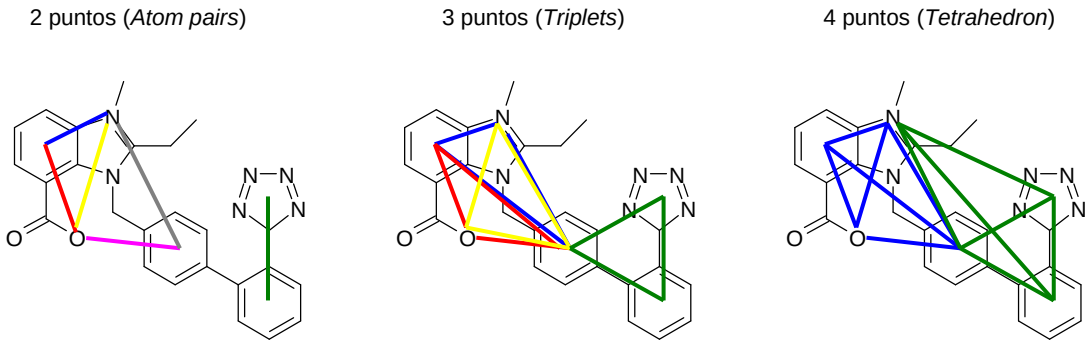



**Figura 1.14.** Representación de *fingerprints* farmacofóricos basados en dos, tres y cuatro puntos.

En el año 2000, Tripos[280] fusiona todas las combinaciones de tupletes en un *fingerprint* múltiple, codificado mediante mapas de bit y permitiendo así un almacenamiento más eficiente.

Desarrollos posteriores, como el del método ToPD[281] (*total pharmacophore diversity*) calculan las distancias entre pares de átomos basados en la *feature* farmacofórica y la forma, calculada a partir de todos los átomos pesados presentes en una molécula. La caracterización farmacofórica no se realiza únicamente midiendo las distancias entre PPPs, sino que se determinan las distancias de cada uno de los PPPs al resto de átomos pesados de la molécula. De este modo, se muestrea la posición relativa de todos los PPPs sobre la forma global de la molécula. En este caso, la codificación no es binaria, sino que se generan representaciones para cada una de las características que posteriormente son descritas según parámetros estadísticos.

Finalmente, Hovarth[282] incluye el concepto de *fuzziness* (“difusión”) en los denominados *Fuzzy Bipolar Pharmacophore Autocorrelograms* (FBPA). Se trata de vectores de números reales, en lugar de *fingerprints* binarios, que se ha mostrado muestran un buen comportamiento de vecindad estructural-biológica.

Los descriptores *fingerprint* farmacofóricos usados en este trabajo son los CATS (*Chemically Advanced Template Search*) desarrollados por Schneider *et al*[43], en un primer momento introducidos como descriptores 2D (CATS2D) y extrapolados a 3D (CATS3D) en versiones posteriores[283].
La versión original considera cinco tipos de átomos generalizados: dador de puente de hidrógeno (D), aceptor de puente de hidrógeno (A), átomo cargado positivamente (P), átomo cargado negativamente (N) y centros lipofílicos (L). La distancia se mide como el número de enlaces a lo largo del camino más corto que conecta dos nodos del grafo (CATS2D). En total, estas distancias están clasificadas en 10 particiones (de un mínimo de 0 enlaces a un máximo de 9 enlaces), por lo que el número de bits necesarios o dimensión del *fingerprint* corresponde a 150 (15 combinaciones de pares × 10 distancias de *binning*). Cada una de las 15 posibles combinaciones de pares (DD, DA, DP, DN, DL, AA, AP, AN, AL, PP, PN, PL, NN, NL, LL) se

escala en función de las ocurrencias totales del par correspondiente. En la Figura I.4 de la Introducción se esquematiza el proceso de derivación típica de los CATS. El vector de correlación obtenido (CV) corresponde a la ecuación [1.84]:

$$\mathrm{CV}_d^{TP} = \frac{1}{A+B} \sum_{i=1}^{A} \sum_{j=1}^{B} \frac{1}{2} \delta_{ij,d}^{TP} \qquad [1.84]$$

Donde *i* y *j* son los átomos, *d* es el rango de distancias, *TP* corresponden a los tipos de átomos del par de átomos *i* y *j*, *A* y *B* son el número total de átomos del tipo de los átomos *i* y *j*, respectivamente, y $\delta_{ij,d}^{TP}$ es la delta de Kronecker, que se evalúa a uno para todos los pares de átomos de los tipos *TP* en el rango de distancia *d*. Los pares de átomos con uno mismo no se consideran, así como tampoco aquellos átomos que no corresponden a ninguno de los tipos atómicos. Cada uno de los *bins* se encuentra escalado según la ocurrencia del número de tipos farmacofóricos $(A+B)^{-1}$. Finalmente, una vez obtenidos todos los *bins* del CV, éstos se normalizan entre cero y uno. Estos descriptores se encuentran implementados en el programa *speedcats*.

CATS3D[283] expresa la distancia como distancia geométrica Euclídea entre los dos átomos. La asignación de los tipos generalizados de átomos se puede realizar mediante la función *PATTY_Type* de MOE, basada en el esquema propuesto por Bush y Sheridan[284], o mediante la función *ph4_aType*[285], también implementada en MOE. En el primer caso[283], se consideran siete tipos generalizados de átomos (catiónico, aniónico, polar, aceptor, dador, hidrofóbico u otros), mientras que la función *ph4_aType*, utilizada en el presente trabajo, define seis tipos de átomos: aceptor, dador, polar, catiónico, aniónico e hidrofóbico. Así, en el primer caso el número de combinaciones de pares es de 28 y en el segundo, de 21. Las distancias se reparten en 20 *bins* equiespaciados [0,20] Å, conduciendo a un CV de dimensión 560 (*PATTY*) o de 420 (*ph4_aType*).

Como se ha introducido, los descriptores CATS se han aplicado con éxito en diferentes procesos de *virtual screening*.

## 1.7. Obtención de Modelos Farmacofóricos

Los dos métodos utilizados para derivar modelos farmacofóricos son: el módulo de farmacóforos implementado en MOE versión 2004.03 y el modelo SQUID (*Sophisticated Quantification of Interaction Distributions*)[51].

### 1.7.1. Modelos Farmacofóricos en MOE

La herramienta implementada en MOE para la derivación de farmacóforos supone una de las aproximaciones más sencillas para la generación de hipótesis, ya que la generación del modelo es manual. Así, no se muestrean automáticamente y de manera exhaustiva todas las posibles combinaciones/alineamientos de las características farmacofóricas, como otros programas más especializados realizan (Catalyst, GASP, DISCO)[287], sino que MOE opera a partir de un alineamiento inicial introducido previamente, el cual permanece rígido. La versión MOE 2005.06, posterior a la realización de esta parte del trabajo, incorpora de forma automatizada la flexibilidad conformacional en la formulación de la hipótesis.

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. 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. Es el usado en MOE por defecto.

- 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 ($sp^2$) o no ($sp^3$) 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.

El proceso para generar la hipótesis o *query* parte de un conjunto de ligandos alineados. Este alineamiento inicial se puede obtener por superposición de las estructuras cristalográficas de los ligandos en el sitio activo de la proteína, mediante algoritmos de alineamiento flexible, como el algoritmo MOE-FlexAlign[287] o incluso, a partir de los resultados de un *docking* en la proteína diana.
La hipótesis incluye restricciones acerca de una *feature* farmacofórica que un punto en el espacio debe satisfacer, dentro de un radio de tolerancia. Esta *feature* puede corresponder a un único punto de anotación del ligando (por ejemplo, que el átomo sea dador) o etiquetarse con una asignación múltiple como combinación lógica de varios (por ejemplo, dador o aceptor). Además, varias restricciones de este tipo se pueden agrupar de manera que se fuerce el cumplimiento de todas ellas por parte de una determinada molécula. MOE permite también la inclusión de restricciones sobre la forma de la molécula mediante la definición de volúmenes. Éstos pueden ser excluyentes (el interior del volumen no puede contener ningún átomo con una determinada característica), incluyentes (se obliga a que en su interior se encuentre al menos un átomo con una característica) o exteriores (fuera del volumen definido, no se sitúa ningún átomo que satisfaga una determinada expresión).
A partir del alineamiento, el usuario define las restricciones de la *query*, ajustando las posiciones, radios de los puntos potenciales farmacofóricos, sus combinaciones y, adicionalmente, volúmenes. La herramienta *Pharmacophore Consensus* sugiere restricciones farmacofóricas, a través de todos los átomos con una anotación equivalente, superpuestos en el espacio dentro de una tolerancia y comunes a un determinado porcentaje de las moléculas presentes en el alineamiento.

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. Entonces, se realiza el emparejamiento exhaustivo de todos los puntos de anotación del ligando con los puntos potenciales farmacofóricos (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 el cuadrado promedio de las distancias (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.

### 1.7.2. SQUID. *Sophisticated Quantification of Interaction Distributions*

Tal y como se ha introducido, el objetivo del desarrollo del método SQUID[51] fue doble: por una parte, incluir información "difusa" (*fuzzy*) sobre la conservación y tolerancia de las características (*features*) farmacofóricas en el conjunto de moléculas activas sobre las que se deriva el modelo y por otra, evitar el alineamiento de las moléculas de la base de datos sobre el modelo farmacóforo obtenido.

SQUID agrupa las *features* farmacofóricas presentes en el alineamiento inicial de moléculas en puntos potenciales farmacofóricos (PPPs), expresados en la forma de funciones de densidad de probabilidad gaussianas. Cada uno de estos PPPs contiene información de i) el tipo farmacofórico de los átomos, ii) su posición en el espacio, iii) la desviación estándar en la posición desde el centro del PPP de todos aquellos átomos pertenecientes a un PPP, lo que equivale al radio del PPP y iv) el grado de conservación de un PPP particular en todas las moléculas presentes en el alineamiento.

En la Figura 1.15 se esquematiza el proceso de derivación de un modelo farmacóforo SQUID, que se detalla a continuación. Se parte de un alineamiento inicial de las moléculas activas, también "fijo" como en el apartado anterior, es decir, no se realizan sucesivas optimizaciones de la superposición de estas moléculas.
El primer paso consiste en la anotación de los átomos de los ligandos según su tipo farmacofórico. SQUID ha sido implementado[51] en lenguaje SVL (*Scientific Vector Language*) en el programa MOE. Así, la asignación de tipos atómicos se realiza mediante los esquemas farmacofóricos implementados en MOE. Inicialmente, se utiliza con este fin la función *ph4_aType*, la misma que la utilizada en los descriptores CATS3D, que define seis tipos: catiónico, aniónico, polar, hidrofóbico, dador y aceptor de puente de hidrógeno.

Seguidamente, se generan los PPPs por agrupamiento de aquellos átomos próximos en el espacio que comparten tipo farmacofórico. Para ello, se calculan las denominadas *local feature densities* (LFDs) para cada átomo *k* de tipo farmacofórico *t*, según la ecuación [1.85]:

$$\text{LFD}\,(\textit{átomo}_k^t) = \sum_i \max\left\{0\ ,\ 1 - \frac{D_2(\textit{átomo}_k^t\ ,\ \textit{átomo}_i^t)}{r_c}\right\} \qquad [1.85]$$

donde *i* recorre todos los átomos del tipo *t* presentes en el alineamiento, $D_2$ corresponde a la distancia Euclídea entre dos átomos *k* e *i*, y $r_c$ es el radio del *cluster* o *cluster radius*. Este *cluster radius* es el parámetro que determina la resolución del modelo, ya que indica el nivel de agrupamiento en *clusters* de las *features* para generar los PPPs, y tiene que ser fijado empírica e independientemente en cada caso particular de estudio. Todos aquellos átomos pertenecientes a un tipo farmacofórico particular situados dentro de una esfera de radio $r_c$ se agrupan alrededor de aquel que presenta una LFD máxima.
La posición central del PPP resultante corresponde al centro geométrico de todos los átomos que comparten *cluster*. La desviación estándar ($\sigma$) se establece a partir de la distancia mediana de todos los átomos del *cluster* al centro del PPP, con un valor mínimo de 0.5. Esta desviación caracteriza la anchura de la distribución de los átomos representados por un PPP y en las ilustraciones gráficas de los modelos farmacóforos SQUID, $\sigma$ equivale al radio de los PPPs.
Finalmente, la conservación de cada PPP se pondera mediante el peso (*w*), calculado según la ecuación [1.86]:

$$w(\mathrm{PPP}) = \sum_{i=1}^{m} \min\left\{ \frac{1}{m} \ , \ \frac{\#\textit{átomos de la molecula}_i \textit{ del } PPP_k}{\#\textit{átomos del } PPP_k} \right\} \qquad [1.86]$$

donde *m* representa el número de moléculas en el modelo. Cuando un PPP representa el mismo número de átomos de todas las moléculas del alineamiento, el peso adopta el valor máximo de uno. El mínimo ($m^{-1}$) corresponde a PPPs que consisten únicamente de átomos presentes en una de las moléculas.

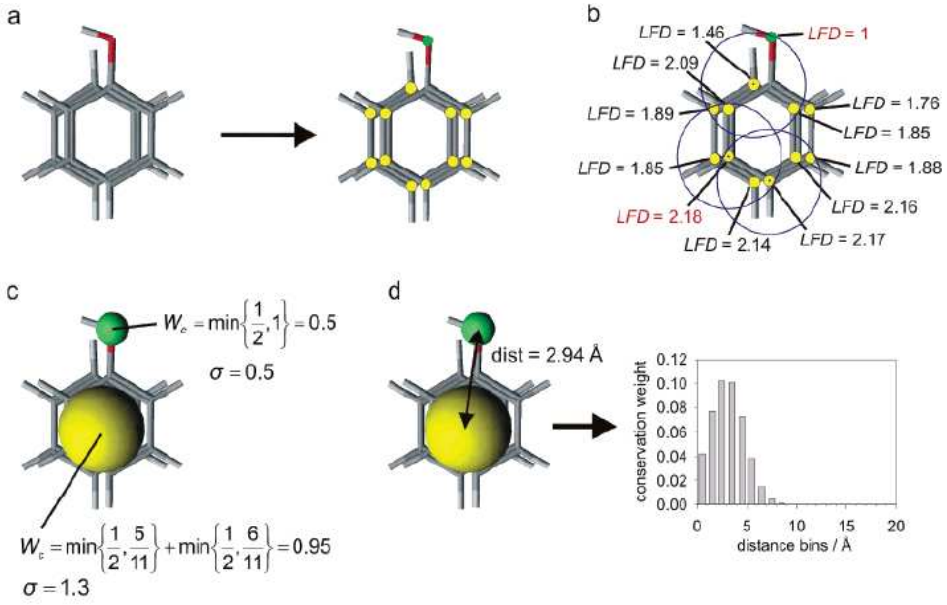



**Figura 1.15.** Esquema de derivación de un modelo SQUID. Extraído de [51].

Una vez obtenido el modelo farmacofórico, éste se codifica en un vector de correlación (CV) para realizarse el VS en una base de datos. El modelo SQUID resultante se encapsula en un vector de dimensión de 420 bits, resultante de la combinación de los 21 pares atómicos (*TP*) y un esquema de partición de las distancias (*d*) en 20 rangos equiespaciados [0,20] Å. La contribución a cada uno de los bits del CV se obtiene según la ecuación [1.87]:

$$\mathrm{CV}_d^{TP} = \frac{1}{\#\ pairs\ (TP)} \sum_{p=1} \sum_{q=1} \frac{1}{2}\, \delta_{pq}^{TP} \left( \frac{w_p w_q}{\sqrt{2\pi}\left(\sigma_p + \sigma_q\right)} \exp\left( -\frac{1}{2} \frac{\left(D_2(p,q) - centre_d\right)^2}{\left(\sigma_p + \sigma_q\right)^2} \right) \right) \qquad [1.87]$$

donde *p* y *q* se refieren a los PPPs de un tipo farmacofórico *T* y *P*, respectivamente. *d* es el rango de distancias, $w_p$ *y* $w_q$ son los pesos de los PPPs *p* y *q*, $\sigma_p$ y $\sigma_q$ son sus desviaciones estándar, $centre_d$ es el centro del rango de distancias *d*, $D_2(p,q)$ es la distancia Euclídea entre los dos PPPs *p* y *q*, $\delta_{pq}^{TP}$ es la delta de Kronecker, que se evalúa como 1 para todos los pares de PPPs del tipo *TP* cuya distancia está comprendida en el rango *d*. Los sumatorios recorren todos los PPPs de un determinado tipo farmacofórico y el factor de 0.5 evita la duplicación en la cuenta de los pares.
De manera similar a CATS (apartado 1.6.5), el valor de cada *bin* se escala según el número de pares *TP* presentes en el modelo y el CV final obtenido se normaliza entre cero y uno.

En la búsqueda de similitud, este CV-SQUID se compara con los CV-CATS3D calculados para cada una de las moléculas contenidas en una base de datos. El uso de estos vectores de correlación, libres de alineamiento, evita la superposición de todas las moléculas frente al

modelo farmacofórico, ahorrándose tiempo de cálculo. La similitud se calcula según el índice de la ecuación [1.88], desarrollado específicamente en este método para permitir la comparación de las gaussianas obtenidas en el CV-SQUID frente a los picos de los CV-CATS3D.

$$S(a,b) = \frac{\sum_{i=1}^{n}(a_i b_i)}{1+\sum_{i=1}^{n}((1-a_i)b_i)} \qquad [1.88]$$

Donde $a_i$ y $b_i$ corresponden al *bin i* del vector de correlación CV-SQUID y CV-CATS3D, respectivamente. Los sumatorios se extienden a lo largo de la dimensión *n* de estos CV (*n*=420).

Durante el cribado virtual, se utilizan pesos adicionales (*feature-type weights*) que ponderan la importancia de cada uno de los tipos farmacofóricos generalizados en el CV. Estos pesos se establecen particularmente para cada caso de estudio, ajustándose empíricamente sobre un subconjunto de moléculas de la base de datos, lo que supone una desventaja de esta metodología.

SQUID ha sido validado tanto retrospectivamente frente a ligandos de la ciclooxigenasa 2 (COX-2) y de la trombina[51], como prospectivamente, en la identificación de nuevos inhibidores de la interacción Tat-TAR RNA[52].

## 1.8. Técnicas Estadísticas de Análisis de Datos

Los datos químicos son normalmente multidimensionales, definiéndose un objeto a partir de varios componentes de datos. Por ejemplo, en el caso de las moléculas éstas se caracterizan a través de una gran variedad de descriptores.
En general, una vez se calcula un conjunto de descriptores éstos no pueden utilizarse directamente para generar un modelo, ya que deben solventarse tres tipos de problemas: i) existe una gran correlación entre las variables, de manera que diferentes descriptores codifican el mismo aspecto estructural, ii) puede existir descriptores que no aporten información relevante al modelo y iii) el número de descriptores es demasiado elevado como para ser tratable computacionalmente y no es representable. Con ello, resulta difícil extraer interrelaciones y asociaciones entre estas variables y los objetos de estudio.
Para evaluar la calidad de un conjunto de descriptores normalmente se analizan dos medidas estadísticas: la varianza y la correlación entre ellos. La varianza permite ver el grado de variación de un descriptor a lo largo del conjunto de datos, de manera que si esta es muy baja, el descriptor aporta muy poca información al conjunto. La correlación entre descriptores aporta información del grado de redundancia interna. Descriptores independientes presentan un coeficiente de correlación nulo, denominándose ortogonales. Se establece que el coeficiente de correlación entre dos descriptores no debe ser superior a 0.6, aunque se acepta trabajar en un margen de 0.4 a 0.9.

Así, se realiza un pre-procesado del conjunto de descriptores reduciéndose la dimensionalidad del problema y obteniéndose un conjunto reducido de descriptores con una mayor densidad de información relacionada con la propiedad objetivo (actividad biológica). Este procesado implica también un autoescalado de los datos, de manera que descriptores con mayor orden de magnitud no tengan más peso en el análisis.

Entre los métodos de reducción de dimensionalidad destacan diferentes métodos basados en el aprendizaje (*Machine Learning Methods*)[288]: desde algoritmos genéticos (GA) que automatizan

el proceso de selección de descriptores a métodos que transforman los descriptores, generándose un conjunto reducido. Entre ellos, destacan el análisis de componentes principales (PCA) y métodos de regresión como el *Partial Least Squares regression* o *Projection to Latent* Structures (PLS), la regresión lineal múltiple (MLR) y la regresión con componentes principales (PCR).

Los métodos de regresión (PLS, MLR, PCR) establecen un modelo predictivo de una o más variables dependientes (actividad) en función de la variables independientes (descriptores), por lo que son ampliamente usados en QSAR (especialmente PLS).

En este trabajo se utilizan técnicas de análisis de datos dirigidas hacia la reducción de la dimensionalidad del problema en quimiotecas combinatorias, y no hacia la regresión de modelos. El PCA es una de las técnicas estándar, aunque otras como el análisis factorial y otras técnicas no lineales son también comunes en el análisis de datos y la visualización.

- Análisis de componentes principales (PCA). Se reduce un conjunto de datos parcialmente correlacionados en un número de nuevas variables ortogonales, los componentes principales, con pérdida mínima en la contribución a la variación. Los componentes principales se establecen como combinación lineal de las variables originales: se aproxima la matriz de los datos *X*, de dimensión *n* (número de objetos, moléculas) × *m* (número de variables, descriptores) mediante dos matrices más reducidas: la matriz de los *scores T* (*n* objetos × *d* variables) y la matriz de los *loadings P* (*d* objetos y *m* variables) según la ecuación [1.89]:

  $$X = TP^T \qquad [1.89]$$

  La matriz de los *loadings* contiene los coeficientes de la combinación lineal, indicando qué variables influencian el modelo y cómo éstas están correlacionadas. La matriz de *scores* recoge la proyección de los objetos en el espacio de las componentes principales.
  Normalmente, antes de realizar un PCA los datos se procesan mediante centrado en la media y escalado.

  Geométricamente, al representar en un espacio *m*-dimensional los *n* objetos, el primer componente principal (PC1) corresponde al vector que representa la máxima varianza dentro de los datos, el segundo componente principal (PC2) es ortogonal al primero y con la siguiente máxima varianza, y así sucesivamente. Estas direcciones ortogonales corresponden a los vectores propios de la matriz $X^TX$ y sus valores propios ($\lambda_m$) corresponden a la varianza asociada a cada uno de ellos.

  En la mayor parte de los casos, con 3 a 5 componentes principales (PCs) se explica la mayor parte de la varianza de los datos, de manera que la representación tridimensional de los tres primeros componentes principales suelen cubrir el 60-80% de la varianza. En el diseño de quimiotecas virtuales se suelen considerar las componentes principales que cubren el 90-95% de varianza.

  La principal ventaja del análisis de componentes principales es que no asume distribuciones de probabilidad de las variables originales, aunque es muy sensible a puntos extremos y datos pobremente distribuidos.

- Análisis factorial. Las variables originales (*X*) se describen como combinaciones lineales de un conjunto menor de factores comunes (*CFs*), que contienen la varianza común a varios descriptores (*communality*). La varianza individual de cada una de los descriptores (*uniqueness*) se estima mediante una función de error (*E*), ecuación [1.90]:

$$X = CF \times V + E \qquad [1.90]$$

Tanto PCA como el análisis factorial asumen una constricción lineal del espacio de entrada, por lo que se comportan mal en espacios altamente dimensionales no lineales. En espacios no lineales, se pueden aplicar técnicas como el escalado multidimensional, los mapas no lineales de Sammon o los mapas de Kohonen, basados en redes neuronales.[289]

- Escalado multidimensional (MDS). Esta técnica permite visualizar objetos a partir de su matriz de similitud o disimilitud. En un conjunto de *n* objetos representados en un espacio *m*-multidimensional, la distancia $d_{ij}$ entre los objetos *i* y *j* viene dada por la ecuación [1.91]:

$$d_{ij}^2 = \sum_{k=1}^{m} (x_{i,k} - x_{j,k})^2 \qquad [1.91]$$

  El objetivo es encontrar unas coordenadas en un espacio reducido (normalmente 2D o 3D) tal que la nueva distancia $\delta_{ij}$ entre el par de objetos *i* y *j* se aproxime a la distancia $d_{ij}$ en el espacio *m*-multidimensional. El ajuste de nuevas coordenadas se realiza de forma iterativa con algoritmos de minimización hasta que se satisface con una cierta tolerancia el criterio de Kruskal, ecuación [1.92]:

$$S = \sqrt{\frac{\sum_{i<j} (\delta_{ij} - d_{ij})^2}{\sum_{i<j} \delta_{ij}^2}} \qquad [1.92]$$

  Normalmente, las proyecciones en un espacio tridimensional pueden cubrir hasta un 80% de la varianza de los datos. Los mapas no lineales, como la proyección de Sammon, también aproximan relaciones geométricas en un gráfico bi- o tridimensional.

## 1.9. Métodos de Optimización Globales

Los métodos de búsqueda global tratan de escaparse de los mínimos locales, explorando con más eficiencia el espacio de búsqueda. Generalmente, añaden algún componente aleatorio a la búsqueda, de forma que, si se encuentra un mínimo local, se salte a otro punto del espacio de búsqueda, donde pueda haber otro mínimo, posiblemente global. En este caso se habla de métodos de optimización heurísticos o estocásticos, aunque también existen métodos globales deterministas, con un elevado coste computacional asociado debido a su exhaustividad.

Como se ha mencionado, los algoritmos heurísticos se emplean en diversos campos de la química: desde la optimización de geometrías de conformaciones de pequeñas moléculas en procesos de *docking*, la superposición de compuestos, la elaboración de modelos para la predicción de propiedades o actividades biológicas, el diseño de moléculas *de novo*, el análisis de la interacción proteína-ligando, la selección de descriptores o la selección de compuestos en quimiotecas combinatorias.[290] Dentro de la variedad de algoritmos de optimización estocásticos, en la mayor parte de aplicaciones se incorporan los métodos de *Simulated Annealing* (SA) o bien los algoritmos evolutivos, aunque también se han implementado otros algoritmos como las búsquedas Tabú en programas de *docking*.

Este tipo de algoritmos opera muy bien en los problemas de optimización combinatoria en los que el conjunto de soluciones posibles es discreto o susceptible de discretizarse. Estos

problemas son normalmente del tipo NP-completo (*NP-complete, non-deterministic polynomial time*) ya que no existe un algoritmo general que pueda determinar la solución global en un orden de tiempo computacional polinómico con el tamaño del problema, $O(n^k)$. Los problemas de optimización combinatoria normalmente se formulan en un espacio discreto, es decir, todas o algunas de las variables de la función objetivo se restringen a asumir únicamente valores discretos como enteros, aunque tanto los algoritmos evolutivos como el *Simulated Annealing* son aplicables a optimizaciones continuas globales.

### 1.9.1. *Simulated Annealing*

Estas técnicas se basan en la analogía física con la técnica de *annealing* en la que un material se calienta a elevadas temperatura y posteriormente se enfría de manera lenta y controlada para incrementar el tamaño de sus cristales y reducir sus defectos, alcanzándose una estructura cristalina de mínima energía. El calor permite que los átomos abandonen su posición inicial, un mínimo local de energía interna, y muestren de forma aleatoria estados de mayor energía. El enfriamiento lento permite que se incrementen las posibilidades de encontrar configuraciones con menor energía interna que la inicial.

Análogamente, cada paso del algoritmo de SA reemplaza la solución actual *e* por otra solución aleatoria próxima *e'*, escogida según una probabilidad que depende de la diferencia entre los valores de la función en los dos puntos y un parámetro global de control *T* (denominado temperatura por correspondencia con el símil), que se reduce gradualmente durante el proceso. La probabilidad de transición se ajusta de manera que a altas temperaturas, las soluciones aleatorias se acepten fácilmente (el algoritmo se mueve "*uphill*"), reduciéndose la probabilidad de aceptación conforme disminuye la temperatura (sentido "*downhill*"). En la formulación original[291], esta probabilidad de transición $P(e,e',T)$ se define según la ecuación [1.93], siguiendo el criterio de Metropolis implementado en las técnicas de Monte Carlo, a su vez basado en la distribución de energías de Boltzmann.

$$P(f(e), f(e'), T) = \begin{cases} 1 & \text{si } f(e') < f(e) \\ \exp\left(\dfrac{f(e) - f(e')}{T}\right) & \text{si } f(e') \geq f(e) \end{cases} \qquad [1.93]$$

Otro parámetro a considerar es el esquema de *annealing* que determina el modo de actualización de la temperatura a partir de un valor inicial elevado y su valor mínimo final. Así, los parámetros que deben definirse en la implementación de SA son un generador aleatorio de estados vecinos para el espacio de soluciones definido, una función de probabilidad de transición (aunque normalmente se mantiene la presentada en la ecuación [1.93]) y el esquema de *annealing*.
Se puede demostrar que para un problema finito, la probabilidad de que un SA determine la solución óptima global se aproxima a uno a medida que se incrementa la duración del proceso de *annealing*. Sin embargo, este tiempo teórico es demasiado grande, por lo que se suelen aproximar esquemas de *annealing* más asequibles.

### 1.9.2. Algoritmos Evolutivos

Estos métodos mimetizan las estrategias evolutivas de la naturaleza: las poblaciones se desarrollan tras muchas generaciones siguiendo el principio de supervivencia de los individuos mejor adaptados al medio. El proceso de búsqueda adaptada se basa en una población de soluciones candidatas sobre el que sucesivas iteraciones conllevan una selección competitiva que elimina aquellas soluciones con un menor valor de la función objetivo o *fitness function*. Aquellas soluciones mejor adaptadas, con mayor valor de la función de *fitness*, se recombinan

con otras soluciones generándose una nueva población, continuándose el proceso hasta que se encuentra una solución óptima.

Dentro de los algoritmos evolutivos, se distinguen la programación evolutiva (EP), las estrategias evolutivas (ES), la programación genética (GP) y los algoritmos genéticos (GA), siendo éstos últimos los más populares, desarrollados en 1970 por Holland.[292]

En la Figura 1.16 se muestra un esquema del funcionamiento de un algoritmo genético. En un primer paso, se inicializa una población de soluciones, codificadas en cromosomas artificiales. Para cada uno de estos cromosomas, se evalúa su función objetivo o de *fitness* y se seleccionan aquellos que serán emparejados para la reproducción. Sobre estos cromosomas seleccionados se aplican los operadores de recombinación o *crossover* y mutación, rindiendo una nueva generación de cromosomas.

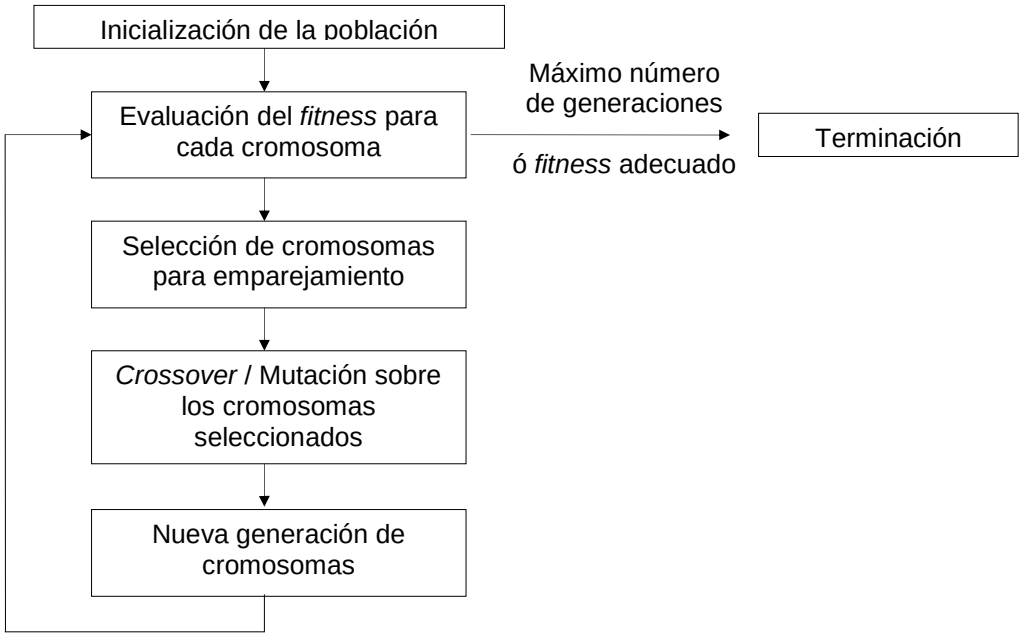



**Figura 1.16.** Esquema de un algoritmo genético.

### 1.9.2.1. Representación y Codificación de los cromosomas

El primer paso en la aplicación de un GA parte de decidir cómo representar las posibles soluciones para un problema determinado. Normalmente, los cromosomas se codifican mediante cadenas (*strings*) que pueden contener valores binarios, valores enteros o incluso valores reales en coma flotante. Cada uno de los cromosomas se divide en genes (representando cada una de las variables de la función objetivo) que a su vez agrupan varios alelos (relacionados con los posibles valores que puede adoptar una variable concreta). La representación en datos binarios, inicialmente introducida, es la más sencilla e interpretable. La codificación del problema es directa en algunas aplicaciones, de manera que los parámetros iniciales de la función (conocido como fenotipo) se trasladan directamente al cromosoma (conocido como genotipo). Sin embargo, en la mayor parte de casos es necesario implementar un sistema de descodificación del genotipo en fenotipo. Así, cada uno de los miembros de una población queda caracterizado por su cromosoma (genotipo), el cromosoma descodificado (fenotipo) y el valor de la función objetivo asociado a esta solución.

### 1.9.2.2. Inicialización de los individuos

La población o número de cromosomas es un parámetro del algoritmo, definido por el usuario y dependiente de la naturaleza del problema, que normalmente permanece constante durante la optimización. Usualmente, la población inicial se genera de forma aleatoria, cubriendo el mayor espacio de soluciones posibles, aunque también se puede inicializar desviándola hacia áreas del espacio con elevada probabilidad de encontrar soluciones óptimas.

### 1.9.2.3. Selección

La selección determina qué individuos se escogen para el apareamiento y cuanta descendencia produce cada uno de ellos. Se basa en una probabilidad establecida según el valor de *fitness* $f_i$ de cada uno de los individuos, los mejor adaptados tienen mayor probabilidad de reproducción. La presión selectiva hace referencia a la probabilidad de que el mejor individuo sea seleccionado en comparación a la probabilidad promedio de selección de todos los individuos de la población. En la ecuación [1.94] se muestra esta probabilidad de selección:

$$pselec_i = \frac{f_i}{\sum f} \qquad [1.94]$$

Sin embargo, la aplicación de este método sobre la función de *fitness* "cruda" conlleva dos problemas: i) la existencia de "superindividuos" seleccionados muy frecuentemente deriva en convergencia hacia su genoma, perdiéndose diversidad en la población con lo que el algoritmo no progresa y la solución final es muy pobre y ii) conforme progresa el algoritmo, las diferencias entre los valores de *fitness* se reducen. De este modo, la probabilidad asociada a las mejores soluciones es casi la misma que la del resto de individuos, con lo que la progresión del algoritmo se transforma en un proceso aleatorio.
Los métodos de selección son mayoritariamente estocásticos, diseñados de manera que también se incluyan individuos con un peor valor de *fitness*. De todos modos, esto no es suficiente para superar los problemas mencionados, por lo que se adoptan dos estrategias de acondicionamiento de la función de *fitness*:

- Escalado de la función de *fitness* (*Proportional Fitness Assignment*)
  Los más comunes son un escalado lineal (ecuación [1.95]), la truncación en sigma ($\sigma$) según la desviación estándar de los valores de *fitness* de la población y su valor promedio (ecuación [1.96]) y un escalado en función potencial (ecuación [1.97]):

  $$f' = a \times f + b \qquad [1.95]$$

  $$f' = f - (\overline{f} - c \times \sigma) \qquad [1.96]$$

  $$f' = f^k \qquad [1.97]$$

  Los coeficientes *a* y *b* de la función de escalado se pueden ajustar en varias formas, aunque normalmente se establecen de manera que el valor máximo de la función escalada sea 1.2 o 2 veces el promedio de la función escalada.
  Además, el acondicionamiento de la función de *fitness* es también necesario cuando se trata de problemas de minimización (el sentido del GA es siempre hacia la maximización de un valor de *fitness*) y cuando se trabaja con valores de *fitness* negativos. El problema de la conversión minimización-maximización se solventa mediante la transformación mostrada en la ecuación [1.98]:

$$f'(x) = \begin{cases} C_{max} - f(x) & \text{si } f(x) < C_{max} \\ 0 & \text{en otros casos} \end{cases} \quad [1.98]$$

donde $C_{max}$ se establece como el mayor valor de *f(x)* en la población actual.
Para el problema de los valores negativos se aplica la conversión de la ecuación [1.99]:

$$f'(x) = \begin{cases} C_{min} + f(x) & \text{si } f(x) + C_{min} > 0 \\ 0 & \text{en otros casos} \end{cases} \quad [1.99]$$

donde $C_{min}$ corresponde al valor absoluto del peor valor de *f(x)* en la población actual.

- Ordenación de los valores de *fitness* (*Rank-based Fitness Assignment*)
  La población se ordena en función del valor de la función de *fitness* de manera que el valor de adaptación asignado a cada individuo depende únicamente de su posición en el rango de individuos y no de su valor de *fitness* real. El valor de *fitness* asignado a un individuo en una posición (*position*) se puede calcular de forma lineal (ecuación [1.100]) o no-lineal (ecuación [1.101]):

$$f'(position) = 2 - SP + 2 \cdot (SP - 1) \cdot (position - 1) / (GA\_population - 1) \quad [1.100]$$

$$f'(position) = \frac{GA\_population \cdot X^{(position-1)}}{\sum_{i=1}^{GA\_population} X^{(i-1)}} \quad [1.101]$$

  donde *GA_population* corresponde al número total de individuos de la población, *SP* corresponde a la presión selectiva fijada (para lineal entre 1 y 2 y para no lineal entre 1 y GA_population-2), y *X* es el resultado de una ecuación polinómica.

Una vez acondicionado el valor de la función objetivo, éste se introduce en el operador selección, aplicándose tantas veces como sea necesario hasta obtener la generación filial. Entre los distintos métodos destacan:

- *Roulette Wheel selection* o *Stochastic Sampling With Replacement*: un segmento unitario se divide en tantas regiones como individuos de tamaño proporcional a su valor de *fitness* acondicionado y un número al azar determina qué segmento/individuo es seleccionado. (Figura 1.17).

- *Tournament selection* escoge aleatoriamente un conjunto de individuos de la población y el mejor individuo entre ellos es seleccionado.

- *Stochastic Remainder Selection Without Replacement* se obtiene el número esperado de copias de un individuo como la relación entre su valor de *fitness* y el valor promedio de la población. Este valor se trunca en el entero más próximo, determinando que el individuo sea seleccionado exactamente este número de veces y una parte fraccional, que es tratada como una probabilidad de que sea seleccionado. Así, un individuo con un número de copias esperado de 1.5 será seleccionado seguramente 1 vez y otra vez con probabilidad de 0.5.

### 1.9.2.4. *Crossover* y Mutación

Se distinguen métodos de *crossover* que pueden ser aplicados tanto a variables binarias como reales y aquellos métodos que quedan restringidos a cromosomas codificados en valores reales. En los primeros, se encuentra el *uniform crossover*, en el que cada elemento del cromosoma hijo generado es elegido aleatoriamente de cada uno de los padres, el *single-point crossover* en el que previo a un punto de corte el cromosoma descendiente procede de uno de los padres y a partir de este punto del otro de los padres o el *multi-point crossover*, análogo al anterior, incluyendo varios puntos de corte (Figura 1.17). La probabilidad de que un par de cromosomas seleccionados se recombinen viene dada por la tasa de *crossover*, parámetro impuesto por el usuario.

Tras la recombinación, los cromosomas descendientes sufren mutación con una probabilidad establecida por la tasa de mutación. La mutación consiste en el cambio de valor de un alelo aleatorio: de 0 a 1 o viceversa en cromosomas binarios, adoptando un valor comprendido en un rango para un cromosoma entero o adicionándole un valor aleatorio pequeño en cromosomas en coma flotante.

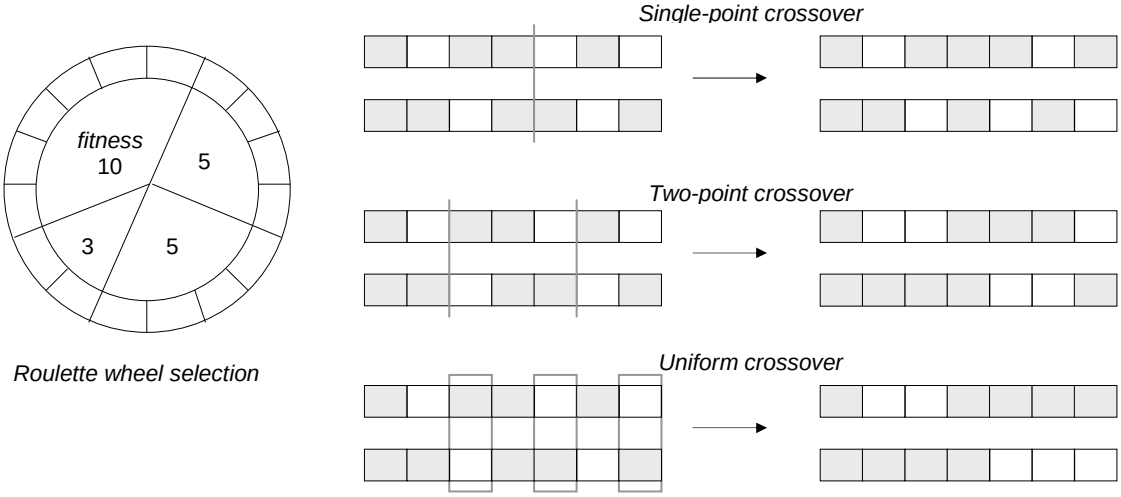



**Figura 1.17.** Esquema del método de selección *Roulette Wheel selection* y de tres métodos de *crossover* para variables binarias.

### 1.9.2.5. *Replacement*

Una vez se dispone de una nueva generación de individuos hijos y se ha evaluado su función de *fitness*, se distinguen dos modelos en función de cómo se realiza el reemplazo de la generación anterior de padres (Figura 1.18):

- En el modelo generacional (*generational replacement*) una generación paternal produce una generación completa filial, de manera que la generación paternal es condicional o incondicionalmente reemplazada por sus hijos.

- En el modelo *steady-state*, en cuanto se genera un cromosoma hijo, éste es condicional o incondicionalmente insertado en la generación paterna, sustituyendo al peor de los padres, de manera que el material genético del hijo está disponible inmediatamente para influenciar la producción del siguiente hijo. Como se ha mencionado, normalmente el reemplazo se realiza de manera que el número total de cromosomas resultantes se mantiene constante durante la optimización.

El término reemplazamiento incondicional hace referencia a que la sustitución de los padres por parte de los hijos se produce siempre, independientemente del valor de función de *fitness* que estos presentan, comparativamente frente a los padres. De este modo, la conservación de las soluciones óptimas no se asegura del todo, ya que éstas, si bien seleccionadas frecuentemente, pueden perderse durante el *crossover* y la mutación y ser sustituidas por los nuevos hijos. Por ello, se suelen imponer esquemas condicionales de reemplazo más efectivos, en los que los hijos únicamente se insertan en la población si suponen una mejora de los miembros existentes de la población. Además, se pueden aplicar técnicas de elitismo en los que un determinado número de individuos son insertados incondicionalmente en las siguientes generaciones, aunque también participan en los eventos reproductivos.

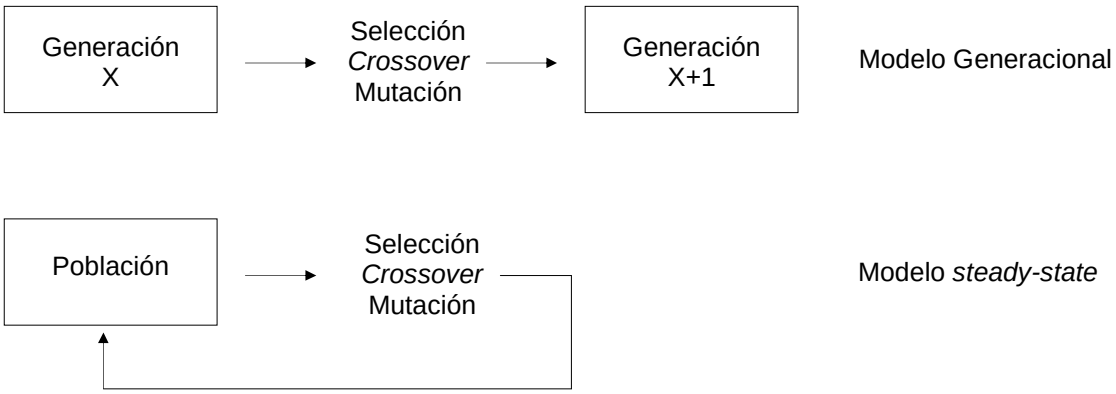



**Figura 1.18.** Modelo generacional *vs* modelo steady-state

Otro problema potencial asociado a los algoritmos genéticos es la deriva génica o especiación, de manera que el proceso se desvía hacia áreas del espacio de búsqueda donde residen agrupaciones de individuos muy próximas, dejando áreas del espacio de búsqueda inexploradas. Para reducir este fenómeno, se pueden aplicar técnicas de *niching*. La primera solución encontrada se posiciona en el centro de un hipervolumen o *niche*. Si las siguientes soluciones caen dentro de un radio de distancias definido próximas a un *niche*, su valor de *fitness* es penalizado, de manera que se limita el crecimiento incontrolado de especies particulares dentro de una población.
Los *island models* mantienen un número de subpoblaciones separadas e introducen el operador migración cada cierto número de generaciones, permitiendo el intercambio de material genético entre ellas. Este tipo de modelos, además de mantener la diversidad de las especies, constituye una estrategia útil en la paralelización de los algoritmos genéticos.

### 1.9.2.6. Otros Algoritmos Evolutivos

La descripción anterior corresponde a los algoritmos genéticos. El resto de algoritmos evolutivos, aunque similares en espíritu, difieren en los detalles de su implementación y naturaleza del problema de aplicación, ya que están dirigidos principalmente a la optimización global de variables contínuas, más que a problemas combinatorios enteros.

En la programación evolutiva (EP), los miembros de una población se contemplan como partes de especies específicas más que miembros de una misma especie, por lo que no existe proceso de recombinación y el único operador es la mutación. El método de selección típico es $(\mu+\mu)$, en el que los $\mu$ padres generan $\mu$ hijos, y entre estos $2\mu$ individuos se seleccionan probabilísticamente los $\mu$ individuos que pasan a la siguiente generación. La codificación típica del cromosoma suele ser en valores reales.

Las estrategias evolutivas (EG), muy similares a EP, operan con vectores de números reales sobre los que el operador primario es la mutación. Ésta, se aplica adicionando un valor aleatorio de una distribución gaussiana cuya desviación estándar se adapta durante la optimización, por lo que se conocen como procesos autoadaptados.

La programación genética (GP), fuertemente desarrollada a partir del año 2000, es una metodología inspirada en la evolución biológica para encontrar aquellos programas que mejor realizan una determinada tarea.

### 1.9.3. Optimización Multiobjetivo

En muchos problemas se presentan simultáneamente varios criterios o parámetros a optimizar que no pueden o no deben combinarse en un único valor objetivo ya que normalmente están en conflicto entre sí. Estos casos se denominan problemas de optimización multiobjetivo o multicriterio (*Multiobjective Optimisation Problems*, MOP). El concepto de óptimo no es evidente en estos casos ya que debe respetarse la integridad de cada uno de los criterios por separado. La noción de óptimo más aceptada en estos casos es la propuesta inicialmente por Edgeworth y generalizada por Pareto en 1896.[293] Un óptimo de Pareto o solución no dominante es aquella en la que una mejora en uno de los criterios resulta en un deterioro en uno o más de los restantes criterios, comparadas frente al resto de soluciones en la población. Así, una solución domina a otra si es equivalente o superior en todos los criterios u objetivos y, estrictamente, si es al menos superior en uno de los objetivos.
Dentro de los distintos algoritmos heurísticos de optimización, es en los algoritmos evolutivos donde la MOP se ha implementado principalmente, y dentro de estos, en los algoritmos genéticos (MOGA, *MultiObjective Genetic Algorithm*).
En general, se pueden considerar dos grandes estrategias de afrontar la MOP en GA: i) aquellas que jerarquizan la población en función de la dominancia de los individuos, según el criterio de Pareto (Pareto ranking), buscando un conjunto de soluciones no dominantes y ii) algoritmos que no incorporan el concepto de óptimo de Pareto, sino que optimizan un único objetivo global resultado de la combinación lineal ponderada de los distintos objetivos.

La primera aplicación del Pareto ranking en quimioinformática se encuentra en la superposición flexible de estructuras 3D[294]. Posteriormente, se introduce en el diseño de quimiotecas combinatorias[159, 167, 295, 296], en la derivación de modelos QSAR[297], en la evolución de moléculas de tamaño medio[298] y en la obtención de múltiples hipótesis farmacofóricas[299].

## 1.10. Diseño de Quimiotecas

### 1.10.1. Medidas de Similitud y Diversidad

La implementación de un método de selección requiere tanto la especificación de los descriptores moleculares como la de las medidas de similitud intermolecular. Muchos de estos métodos parten de las técnicas usadas para la búsqueda y agrupación de bases de datos.

Algunos coeficientes son medidas de la distancia o disimilitud entre moléculas (presentando un valor de 0 para objetos idénticos) mientras que otros miden directamente la similitud (adquieren un valor máximo para objetos idénticos). Además, en la mayor parte de casos el coeficiente adopta valores comprendidos entre 0 y 1 o puede ser normalizado a este rango, por lo que se pueden transformar en su coeficiente complementario por sustracción de la unidad.

Para que un determinado coeficiente sea considerado métrica debe satisfacer las siguientes condiciones: i) sus valores deben ser cero o positivos y la distancia de un objeto consigo mismo

tiene que ser cero, ii) tiene que ser simétrico, iii) debe cumplir la desigualdad triangular y iv) la distancia entre dos objetos no idénticos tiene que ser superior a cero. Se denominan coeficientes pseudométricos a aquellos que presentan tres de estas propiedades y coeficientes no-métricos a aquellos que no cumplen la tercera propiedad.[26,27]

En la Tabla 1.8 se presentan los coeficientes más comunes en quimioinformática usados en este trabajo. En la referencia [300] puede encontrarse una recopilación más amplia.

Los vectores $X_A$ y $X_B$ corresponden a la descripción de las moléculas *A* y *B* a través de *n* atributos, pudiendo ser éstos valores reales o binarios. En el caso de descriptores binarios se definen los valores *a*, *b* y *c* según la ecuación [1.102], donde *a* y *b* corresponden al número total de bits puestos a uno en cada una de las moléculas y *c* al número de bits comunes y puestos a uno en ambas moléculas:

$$a = \sum_{i=1}^{n} x_{iA} \qquad b = \sum_{i=1}^{n} x_{iB} \qquad c = \sum_{i=1}^{n} x_{iA} \cdot x_{iB} \qquad [1.102]$$

**Tabla 1.8.** Descripciones de Métricas de distancia ($D_{AB}$) y coeficientes de similitud ($S_{AB}$).

| | **Variables Continuas** | **Variables binarias** |
|---|---|---|
| Distancia Manhattan, City-Block, Hamming | $D_{A,B} = \sum_{i=1}^{n} \lvert x_{iA} - x_{iB} \rvert$<br>Rango de ∞ a 0 | $D_{A,B} = a + b - 2c$<br>Rango de *n* a 0 |
| Distancia Euclídea | $D_{A,B} = \left[ \sum_{i=1}^{n} (x_{iA} - x_{iB})^2 \right]^{1/2}$<br>Rango de ∞ a 0 | $D_{A,B} = (a + b - 2c)^{1/2}$<br>Rango de *n* a 0 |
| Coeficiente de Tanimoto o Jaccard | $S_{A,B} = \left[ \sum_{i=1}^{n} x_{iA} x_{iB} \right] \Big/ \left[ \sum_{i=1}^{n} x_{iA}^2 + \sum_{i=1}^{n} x_{iB}^2 - \sum_{i=1}^{n} x_{iA} x_{iB} \right]$<br>Rango de –0.333 a 1 | $S_{A,B} = c/(a + b - c)$<br>Rango de 0 a 1 |
| Coeficiente del coseno u Ochiai | $S_{A,B} = \left[ \sum_{i=1}^{n} x_{iA} x_{iB} \right] \Big/ \left[ \sum_{i=1}^{n} x_{iA}^2 \cdot \sum_{i=1}^{n} x_{iB}^2 \right]^{1/2}$<br>Rango de -1 a +1 | $S_{A,B} = c/(a \cdot b)^{1/2}$<br>Rango de 0 a 1 |
| Coeficiente de Dice o Czekanowski | $S_{A,B} = \left[ 2\sum_{i=1}^{n} x_{iA} x_{iB} \right] \Big/ \left[ \sum_{i=1}^{n} x_{iA}^2 + \sum_{i=1}^{n} x_{iB}^2 \right]$<br>Rango de -1 a +1 | $S_{A,B} = 2c/(a + b)$<br>Rango de 0 a 1 |

En cada aplicación, la elección del coeficiente va ligada al conjunto de descriptores utilizado. Así, típicamente se trabaja con la distancia Euclídea en variables reales continuas y con el coeficiente de tanimoto en variables binarias.
En la mayor parte de búsquedas de similitud en VS y en el diseño de quimiotecas focalizadas y diversas se emplean este tipo de medidas. Sin embargo, se han desarrollado otras medidas como coeficientes de correlación entre variables y medidas probabilísticas basadas en la ocurrencia de propiedades en bases de datos.

En los métodos de selección basados en *clusters* y métodos de partición, se evalua la diversidad en función de la fracción de *clusters* o celdas/*bins* que alcanzan un determinado nivel de ocupación. Este grado de recubrimiento se puede evaluar de distintas formas:

- Espacio o *cell-based Fraction*: #*Bins* ocupados / # *Bins* totales de la partición [1.103]
- Fracción de población: $\sum_{i=1}^{\#bins\_ocupados} n_i \Big/ N$ [1.104]
- *Cell-based Chi2*: $\sum_i (n_i - n_{promedio})$ [1.105]
- *Cell-based Entropy*: $-\sum_i (n_i \cdot \log n_i)$ [1.106]
- *Cell-based Density*: $-\sum_i (\frac{n_i}{n_{promedio}} \cdot \log \frac{n_i}{n_{promedio}})$ [1.107]

Donde $n_i$ corresponde al número de compuestos presentes en el *bin i*, *N* es el número total de compuestos totales de la quimioteca y $n_{promedio}$ es el número promedio de compuestos por celda. Los criterios introducidos por *cell-based* fueron implementados en el programa Cerius2 por Jamois y Hassan[152]. En su implementación original, PRALINS dispone de los dos primeros criterios.[166]

### 1.10.2. Diseño de Quimiotecas Diversas: Métodos de selección de compuestos

Estos métodos se pueden aplicar tanto a la selección de un conjunto de reactivos incluidos en un catálogo comercial para luego ser aplicados a síntesis combinatoria (aproximación *reagent-based*), a la selección de compuestos individuales de una base de datos (*cherry picking selection*) o a la selección de una quimioteca combinatoria de productos en formato *full array* (aproximación *product-based*, véase Introducción).

La división de las distintas metodologías en tres grupos (distancias, *clusters* y métodos basados en particiones) expuesta en la introducción no es estricta en el sentido de que diferentes autores han propuesto distintos modelos de clasificación.

Por una parte, Willet[301] y Pérez[148] dividen los métodos de selección de compuestos en cuatro grupos: los tres anteriores y una clasificación adicional reservada para las aproximaciones basadas en métodos de optimización. Éstas abordan el problema de la selección como un problema de optimización combinatoria, incluyendo algunos algoritmos de selección *cherry picking* que requieren el uso de técnicas heurísticas y la adaptación de los tres métodos anteriores en la selección de compuestos en formato *full array*. También se incluyen en este cuarto subgrupo los métodos basados en el diseño de experiencias, como el *D-Optimal Design*.

Otra clasificación alternativa es la propuesta por Pearlman[302] en métodos *cell-based* y *distance-based*. La primera corresponde a los métodos de partición y la segunda incluye el resto de métodos, fundamentándose en que los métodos de *clustering* miden la distancia intermolecular para crear los *clusters*.

Pascual[303] distingue también dos grupos: los basados en distancias y los basados en técnicas de clasificación del espacio, incluyendo en este último los métodos de *clustering* y los de partición, ya que el índice de diversidad determinado en ambos casos es equivalente (ecuaciones [1.103]-[1.107]).

En los siguientes apartados se describen de manera general estos métodos, haciendo hincapié en los implementados en la versión original de PRALINS y en el módulo CombiChem de Cerius2, empleados en este trabajo. En las referencias [301] y [303] puede encontrarse una recopilación histórica de la incorporación de estas metodologías al diseño de quimiotecas diversas.

### 1.10.2.1. Métodos basados en Distancias

El objetivo de estos métodos es la identificación de aquel subconjunto de las *n* moléculas más diversas pertenecientes a una base de datos con *N* compuestos (donde típicamente *n*<<*N*). La diversidad se establece en términos de disimilitudes intermoleculares entre compuestos. Los algoritmos más típicos son los de máxima disimilitud (*maximum dissimilarity*) y los de esferas de exclusión (*sphere exclusion*).

El algoritmo básico de máxima disimilitud es el propuesto por Kennard y Stone en 1969, y aplicado a la selección de compuestos por Lajineess y Bawden. El subconjunto *n* se inicializa transfiriendo un compuesto de la base de datos. A partir de allí, los restantes compuestos añadidos hasta obtener un tamaño *n* se escogen de forma que sean lo más disimilares a los ya presentes. El compuesto inicial puede ser escogido aleatoriamente, puede corresponder al más disimilar de la base de datos o puede ser un compuesto próximo al centro de la base de datos. Por ejemplo, en la correspondiente implementación de este algoritmo en el programa MOE se escoge invariablemente el primer compuesto de la base de datos. Por otra parte, la disimilitud se define siguiendo típicamente el criterio MaxMin que maximiza la mínima distancia intermolecular en el conjunto (ecuación [1.108]) o el criterio MaxSum, que maximiza la suma de distancias de cada compuesto con los restantes (ecuación [1.109]):

$$\max\left\{\min_{i\neq j;\, j\subset n} d_{i,j}\right\} \qquad [1.108]$$

$$\max\left\{\sum_{j=1}^{n} d_{i,j}\right\} \qquad [1.109]$$

Una variante del método MaxSum maximiza la suma de distancias de cada compuesto con un centroide, molécula ficticia situada en el centro del conjunto seleccionado, permitiendo reducir el orden de tiempo de $O(n^2N)$ *a* $O(nN)$[304].

Este formato básico no garantiza que se obtenga el subconjunto óptimo, ya que es un proceso altamente dependiente del punto inicial. Por ello, se introducen posteriormente estas definiciones de disimilitud en combinación con algoritmos de optimización globales como los algoritmos genéticos[150], *Simulated Annealing*[305] o métodos de Monte Carlo[306].
Además, se incorporan diferentes definiciones de disimilitud, como el criterio MaxMin promediado, también implementado en la versión original de PRALINS (ecuación [1.110]):

$$D_{MaxMin\_P} = \sum_{i=1}^{N} \min_{i\neq j;\, c\subset n} d_{i,j} \qquad [1.110]$$

o las funciones *Product* (ecuación [1.111]) y *PowerSum* (ecuación [1.112]), introducidas por Hassan[306] e incluidas en el módulo CombiChem de Cerius2:

$$\max\left\{\left[\prod D_{i,j}^{2}\right]^{\frac{1}{0.5\cdot n(n-1)}}\right\} \qquad [1.111]$$

$$\max\left\{\frac{0.5\cdot n\cdot(n-1)}{\sum 1/D_{i,j}^{2}}\right\} \qquad [1.112]$$

Estos métodos de máxima disimilitud, aplicados inicialmente en selecciones *cherry picking* son extrapolados en 1997 a la selección de subbibiotecas *full array* en combinación con algoritmos genéticos[150, 153] o *Simulated Annealing*[155].

Por otra parte, en los métodos basados en esferas de exclusión a partir de una molécula inicial seleccionada, aleatoriamente o de manera que sea central a la quimioteca, se genera una hiperesfera de un determinado radio. Los restantes compuestos comprendidos a una distancia de este compuesto inferior al radio de la esfera son excluídos. El siguiente compuesto añadido puede ser aquel más disimilar al seleccionado o un compuesto aleatorio, variando según la implementación particular. El proceso se repite hasta completar el tamaño $n$[307]. Otras variantes de este algoritmo, como la implementada en PRALINS, corresponden a métodos de *clustering*, ya que generan agrupaciones de compuestos al incorporarlos a esferas previas si su distancia es inferior al radio o generan nuevas esferas que se convierten en centros de nuevas agrupaciones.

#### 1.10.2.2. Métodos de *Clustering*

El *clustering* es un proceso que divide un grupo de objetos en grupos o *clusters* de objetos, de manera que éstos muestran un alto grado de similitud intra-*cluster* y de disimilitud inter-*cluster*[308]. De este modo, seleccionando un compuesto perteneciente a cada *cluster* se obtiene una muestra representativa de todo el conjunto.

Entre los métodos de *clustering* se aplican mayoritariamente aquellos que no producen solapamiento, es decir, cada molécula es asignada a un único *cluster*. Dentro de ellos, se distinguen los métodos jerárquicos y los no jerárquicos. Los métodos jerárquicos iteran sucesivamente en dos posibles sentidos: a partir de un *cluster* inicial que comprende toda la base de datos éste se divide progresivamente (jerárquicos divisivos) o bien a partir de *clusters* formados por compuestos individuales (*singletons*) éstos se fusionan produciendo *clusters* más grandes que terminan englobando a todo el conjunto (jerárquicos aglomerativos).

En la versión aglomerativa, se parte del cálculo de una matriz de similitud intermolecular entre todos los pares de compuestos, cada uno de los cuales constituye un *singleton*. El par de compuestos más similares se fusiona en un *cluster* formando un único nuevo punto (*cluster* o *singleton*) para el que se calcula su similitud a todos los demás puntos de la base de datos, actualizándose la matriz de similitud. Los distintos métodos difieren en el modo en que se define cuál es el par más similar y cómo este par es fusionado para generar un nuevo *cluster*. En el algoritmo *single linkage* se selecciona la distancia más corta entre las moléculas. Alternativamente, cuando se emplea la distancia más larga entre objetos, se denomina *complete linkage*. Finalmente, si se utiliza la distancia promedio, el método corresponde al *average linkage*.

Así, la matriz de interdistancias se actualiza según la fórmula de Lance-Williams[309] (ecuación [1.113]), cuyas constantes (Tabla 1.9) distinguen el método particular en cuestión. La versión original de PRALINS dispone de los métodos *single linkage, median linkage, complete linkage* y *centroid linkage*. En este trabajo, se implementan las variantes *Group Average* y *Ward* (apartado 8.5.1).

$$d_{k,(i,j)} = \alpha \cdot d_{k,i} + \beta \cdot d_{k,j} + \delta \cdot d_{i,j} + \gamma \cdot \left| d_{k,i} - d_{k,j} \right| \qquad [1.113]$$

**Tabla 1.9.** Constantes de la fórmula de Lance-Williams para las distintas variantes de *clustering* jerárquico aglomerativo. Las variables *i, j* son los *clusters* que se fusionan en el nuevo *cluster k* y $n_i$, $n_j$, $n_k$ corresponden al número de compuestos en los *clusters i, j, k* respectivamente.

| | α | β | δ | γ |
|---|---|---|---|---|
| Complete linkage (Furthest Neighbour) | 0.5 | 0.5 | 0 | 0.5 |
| Median linkage | 0.5 | 0.5 | -0.25 | 0 |
| Single linkage (Nearest Neighbour) | 0.5 | 0.5 | 0 | -0.5 |
| Centroid | $\frac{n_i}{n_i + n_j}$ | $\frac{n_j}{n_i + n_j}$ | $\frac{-n_i \cdot n_j}{(n_i + n_j)^2}$ | 0 |
| Average linkage (unweighted) | 0.5 | 0.5 | 0 | 0 |
| Average linkage (weighted) o Group Average | $\frac{n_i}{n_i + n_j}$ | $\frac{n_j}{n_i + n_j}$ | 0 | 0 |
| Ward | $\frac{n_i + n_k}{n_i + n_j + n_k}$ | $\frac{n_j + n_k}{n_i + n_j + n_k}$ | $-\frac{n_k}{n_i + n_j + n_k}$ | 0 |

Dado que el objetivo es seleccionar un compuesto representativo de cada *cluster*, el proceso de fusión se repite hasta obtener un número de *clusters* igual al tamaño de la selección *n*.

Estas técnicas presentan una complejidad $O(N^2)$ en tiempo y espacio de memoria para la creación de la matriz de interdistancias y orden $O(N^3)$ en tiempo de realización del *clustering*, por lo que su aplicación está limitada a bases de datos de decenas de miles de compuestos.

Por otra parte, los métodos de *clustering* no jerárquicos exigen menos demanda computacional que los jerárquicos. Dentro de la variedad de algoritmos posibles, destacan los métodos *single-pass*, los de *relocation* y los de *nearest-neighbour*:

- *Single-pass*: son sencillos de implementar y muy rápidos. En una única vuelta sobre la base de datos asignan los compuestos a *clusters* y según una tolerancia de similitud deciden si se asigna el siguiente compuesto a un *cluster* existente o se utiliza para generar un nuevo *cluster*.

- *Relocation*: asignan los compuestos a un número de *clusters* semilla e iterativamente reasignan los compuestos a otros *clusters* durante un número de iteraciones o hasta que ningún compuesto migra de un *cluster* a otro. Dentro de ellos, destaca el método *K-means*, también implementado en PRALINS. El problema principal que presentan es que son muy propensos a detectar óptimos locales y no es generalmente posible determinar si realmente se ha alcanzado una clasificación óptima. En estos métodos, el usuario determina el número de *clusters* iniciales.

- *Nearest-neighbour*: como su nombre indica, agrupan aquellas moléculas vecinas al entorno de cada compuesto. El más extendido es el algoritmo de Jarvis-Patrick, que identifica los *K* compuestos más próximos para cada compuesto *N* de la base de datos. Una vez se ha construido esta lista para todos los compuestos, dos moléculas se agrupan en un *cluster* si ellas son vecinas recíprocamente y adicionalmente, si comparten en común un número

mínimo de vecinos $K_{min}$ (*similarity threshold*). Este valor de $K_{min}$ es el que determina principalmente la partición. El proceso de agrupar los pares se repite hasta que no se identifica un nuevo par a agrupar. Este algoritmo presenta la desventaja de que identifica un gran número de *clusters* compuestos de muy pocas moléculas o *singletons* y también la imposibilidad de especificar *a priori* el número de *clusters* finales requeridos. También se encuentra implementado en la versión original de PRALINS.

En general, las técnicas de *clustering* son apropiadas para el tratamiento de datos con elevada dimensionalidad, aunque quedan bastante restringidos a su aplicación en bases de datos de tamaño medio. Otra ventaja es que realizan una partición natural de los datos, aunque la adición de nuevos compuestos obliga a repetir la clasificación de nuevo.

#### 1.10.2.3. Métodos de Partición

Para cada una de las propiedades o descriptores que definen el espacio químico se subdivide su rango en subrangos cuyo producto combinatorio define un conjunto de celdas hipercúbicas o *bins*. Cada molécula se asigna a aquella celda que comprende el rango de propiedades que presenta dicha molécula. Las distintas técnicas difieren en el criterio seguido para definir el rango.

PRALINS dispone del algoritmo de *Optimum Binning* que iterativamente divide en dos aquel rango o segmento con un mayor intervalo de valores hasta que se obtiene un número de celdas ocupadas equivalente o superior al tamaño de selección deseado. En caso de que sea superior, se retiene la partición previa de manera que el número de celdas ocupadas no supere el número de moléculas a seleccionar. De este modo, los *bins* o celdas tienden a presentar lados iguales.

La partición *Optimum Binning* se encuentra también implementada en el módulo CombiChem de Cerius2, que presenta además otros esquemas de partición como son: *Binning Uniforme*, donde cada una de las dimensiones se divide en un número determinado de particiones de igual tamaño o se establece el tamaño del segmento para todos los subrangos posibles de una propiedad; *Binning basado en la desviación estándar*: cada eje de propiedades se divide en tres intervalos según un número *n* de desviaciones estándar: i) desde el mínimo a la media menos *n* desviaciones estándar, ii) desde el punto anterior a la media más *n* desviaciones estándar y iii) desde el punto anterior hasta el máximo del valor de la propiedad del eje; *Binning ponderado por la población*: la división del eje en un número específico de *bins* se realiza de manera que todos los *bins* resultantes estén igualmente poblados. Alternativamente, también se permite la generación manual de un *binning*.

Este tipo de métodos son particularmente útiles para comparar bases de datos diferentes, siempre que se trabaje sobre el mismo conjunto de descriptores y para identificar agujeros de diversidad (celdas no ocupadas). Además, la adición de nuevos compuestos no fuerza la repetición de la partición, por lo que se aplican en la complementación de quimiotecas con quimiotecas externas. Su última ventaja reside en su baja complejidad de cálculo, del orden de *O(N)*, lo que los convierte en métodos accesibles a quimiotecas del orden de centenares de miles de compuestos.
Por el contrario, quedan restringidos a espacios químicos de baja dimensionalidad, dada la explosión combinatoria del número de celdas generadas en espacios de alta dimensión. Además, la arbitrariedad en la definición de los límites de las celdas provoca efectos frontera (*edge effects*) ya que dos compuestos muy cercanos pueden quedar incluidos en distintas celdas, tratándose entonces como compuestos disimilares. Este fenómeno se recoge en este trabajo en el capítulo 8.

Tanto en el caso de los métodos de *clustering* como en los de partición, en el caso de realizar una selección *sparse* o *cherry picking* basada en diversidad, se escoge un producto representante de cada uno de los *clusters* o *bins.*

En el caso de las selecciones *full array* es necesario acoplar un algoritmo de optimización global que escoja aquel subconjunto combinatorio que maximize alguno de los criterios implementados en las ecuaciones [1.103]-[1.107]. La versión original de PRALINS dispone de los métodos de Monte Carlo y *Simulated Annealing* y también el algoritmo de *Local Search* para tal fin.

### 1.10.3. Diseño de Quimiotecas Focalizadas: Métodos de selección de compuestos

En el diseño focalizado, el objetivo es maximizar la similitud del subconjunto de compuestos seleccionados *C* frente a un compuesto activo o *lead* o a una familia de ellos. Esta similitud se define normalmente como la distancia promedio de un compuesto a su *lead* más próximo[162], ecuación [1.114]:

$$S(C) = \frac{1}{n}\sum_{i=1}^{n} \min_{j=1}^{F}(d_{ij}) \qquad [1.114]$$

siendo *n* es la cardinalidad del conjunto *C*, *F* es el número de *leads* y $d_{ij}$ es la distancia entre el compuesto *i* de la base de datos y el *lead j*. Normalmente, se trabaja focalizando a un único *lead*, como en las búsquedas de similitud mediante *fingerprints* farmacofóricos.

Así en un diseño *sparse* o *cherry picking*, simplemente se evalúan las distancias, se ordena la base de datos y se escoge aquel subconjunto que maximiza la similitud. Alternativamente, se puede realizar una clasificación del conjunto según métodos de *clustering* y escoger aquellos compuestos pertenecientes al *cluster* de moléculas activas.

Sin embargo, en las selecciones sobre quimiotecas combinatorias en formato *full array* es necesario imponer un algoritmo de optimización que identifique aquellos productos combinatorios que minimizen dicho criterio. Como en el caso de las selecciones diversas, los más aplicados corresponden a algoritmos genéticos y a *Simulated Annealing*.

De hecho, las primeras implementaciones de procesos heurísticos en selecciones *full array* proceden del diseño focalizado y se generalizan posteriormente al diseño diverso. En 1995, Sheridan[156] y Weber[157] publican sendas aplicaciones de los algoritmos genéticos en las que la codificación del problema se realiza según valores enteros. Posteriormente, en 1997, Brown y Martin presentan el programa GALOPED, que difiere en la codificación en valores binarios[154]. Las implicaciones asociadas a ambas codificaciones se discuten en el apartado 8.1, donde se describe la implementación de GA en PRALINS. Sheridan utiliza inicialmente[156] descriptores farmacofóricos basados en *atom pairs* y en 1999 amplia el estudio a los valores de *scoring* obtenidos según diversas funciones[158].

Zheng *et al* en el programa Focus-2D, incorporan las técnicas de SA en el diseño de quimiotecas combinatorias focalizadas, analizando las frecuencia de aparición de los distintos *building blocks* en el conjunto seleccionado para identificar a los candidatos más probables[160].
Tanto SA como GA son adaptados en el programa Cerius2 por Jamois[161] para realizar optimizaciones "*on the fly*", evitando la enumeración y descripción de toda la quimioteca, especialmente útil en el caso de quimiotecas con centenas de millones de compuestos.

Previamente, en el año 2000, Agrafiotis y Lobanov desarrollan dos algorimos con una mayor componente determinista. El primero de ellos, definido como "*ultrafast greedy algorithm*"[162] comienza con una selección aleatoria *full array* y de manera secuencial para cada punto de diversidad, selecciona aquellos reactivos que maximizan la función objetivo. Para cada punto de diversidad, construyen tantas quimiotecas como reactivos disponibles para este punto, combinando la estructura correspondiente con el resto de listas de reactivos seleccionadas para

el resto de puntos de diversidad. Una vez que el proceso se ha repetido para todos los puntos de diversidad, se termina el ciclo y la similitud de la selección *full array* se compara con el valor del ciclo previo. Si se mejora el resultado, el algoritmo continúa, de lo contrario, termina. Los autores concluyen que el algoritmo presenta una mejor convergencia que las técnicas heurísticas de optimización, alcanzando los mismos valores. El tiempo de preprocesado escala linealmente con el tamaño de la quimioteca virtual, mientras que el tiempo de refinado escala linealmente con el número total de reactivos disponibles.

El otro algoritmo desarrollado por Agrafiotis[163] está diseñado para evitar la enumeración y descripción de toda la quimioteca, rindiendo una solución óptima o *quasi* óptima en un orden de tiempo razonable. Para ello, selecciona una fracción aleatoria *sparse* de productos de la quimioteca. Éstos se enumeran y describen, ordenándose por similitud decreciente a la estructura objetivo. Aquellos que presentan mayor similitud se deconvolucionan en sus *building blocks*, denominados reactivos preferenciales ("*preferred reagents*"). Estos reactivos preferenciales se combinan posteriormente, produciendo una quimioteca *full array* de productos, sobre la que, una vez enumerada y descrita, se evalua la similitud. La selección final se establece sobre aquellos compuestos con una mayor similitud al compuesto *lead*. Debido a su naturaleza estocástica en la selección de compuestos aleatoria inicial, el proceso se repite varias veces, combinándose los resultados por consenso.

El programa PLUMS[310] genera selecciones combinatorias focalizadas imponiendo una serie de restricciones en ciertas propiedades (reglas de Lipinski, satisfacción de un modelo farmacofórico…). Inicialmente clasifica todas las moléculas de la quimioteca virtual en *virtual hits*, aquellas que satisfacen las restricciones, y no *virtual hits*. Seguidamente, genera una quimioteca *full array* por combinación de los monómeros presentes en los *virtual hits*. La función objetivo a optimizar se construye como un balance ponderado de eficiencia y efectividad. En cada iteración, se elimina el peor monómero, es decir, aquel que eliminado permite la obtención de una selección de menor tamaño con un mejor valor de la función. El proceso continúa hasta que se eliminan todos los monómeros desfavorables.

Finalmente, en el año 2003, Young[311] presenta un algoritmo alternante (*Alternanting algorithm*) destinado hacia el diseño de quimiotecas combinatorias focalizadas y generales. Partiendo de una selección *full array* aleatoria del tamaño y configuración deseados, analiza sucesivamente cada uno de los puntos de diversidad. Así, para el punto *i*, con un número de reactivos seleccionados aleatoriamente $n_i$, se añade el mejor de los reactivos disponibles no presentes en la lista de seleccionados y se elimina el peor del nuevo conjunto $n_i+1$. A continuación, salta al siguiente punto de diversidad *j* y una vez recorridos todos, si el resultado mejora el inicial por encima de una tolerancia, se retoma el proceso. De lo contrario, el algoritmo se detiene. Análogamente al algoritmo de Lobanov[163], debido al carácter estocástico del proceso, el cálculo se repite varias veces, reteniendo el mejor de los resultados.

### 1.10.4. Evaluación y Comparación de los métodos de selección

Como se ha comentado, las técnicas computacionales se evalúan en términos de su efectividad y eficiencia. La eficiencia computacional de los distintos métodos se ha comentado brevemente, más detalles se pueden encontrar en las referencias [301] y [303].

Respecto a la efectividad, el primer criterio de evaluación de la efectividad de los distintos métodos del diseño de quimiotecas generales o diversas es la comparación de los resultados obtenidos frente a los alcanzados mediante selecciones aleatorias. En este sentido, pese a que los primeros análisis indicaron que no existían diferencias en la distribución de los compuestos seleccionados por ambas aproximaciones, posteriormente un mayor número de estudios concluyeron que las selecciones racionalizadas son una mejor aproximación que las aleatorias[29, 150, 306].

Dado que el diseño diverso está dirigido a la identificación de compuestos activos frente a varias dianas, algunos de estos estudios, partiendo de bases de datos con varias clases de actividad biológica, comparan el grado de recubrimiento de cada una de las clases según una selección diversa con el obtenido aleatoriamente. En el estudio de Brown y Martin[29], se comparan distintos métodos de *clustering* en función de su capacidad para agrupar los compuestos activos en un mismo *cluster* y separarlos de de los inactivos, identificando así el *active cluster subset*. Concluyen que el método de *clustering* de Ward es superior al resto de métodos de *clustering* testados.

Otro criterio utilizado para evaluar la efectividad de los distintos algoritmos frente a un diseño aleatorio y compararlos entre sí es el basado en medir el grado de diversidad en el espacio de propiedades alcanzado en cada uno de ellos, es decir, hasta qué punto la selección queda "extendida" en el espacio químico. En este sentido, es necesario disponer de métodos que permitan comparar bases de datos, evaluando el recubrimiento alcanzado por ambas de manera independiente al método de selección aplicado y en un mismo marco de referencia.

Como se ha mencionado, los métodos de partición son especialmente adecuados para la comparación de bases de datos ya que son independientes de los datos incluidos, por lo que es uno de los criterios más ampliamente seguidos[312, 313]. Si se comparan dos subconjuntos de una misma base de datos, una partición establecida según métodos de *clustering* es igualmente válida para el análisis.

Otros métodos, como el del centroide, facilitan la expresión de la diversidad como suma de las distancias intermoleculares incluidas en una quimioteca. La combinación de los centroides de dos bases de datos rinde una medida cuantitativa del cambido en diversidad resultante de la fusión de las dos bases de datos.[314]

El *diversity integral criterion* difunde un determinado número de puntos aleatorios en el espacio químico definido por las dos bases de datos comparadas (Figura 1.19). La diversidad de cada quimioteca se establece como la suma de las distancias para cada punto y su molécula más próxima. Aquella quimioteca con un menor índice de diversidad está más extendida en el espacio químico. Esta técnica, se encuentra implementada en el módulo CombiChem de Cerius2.[39]

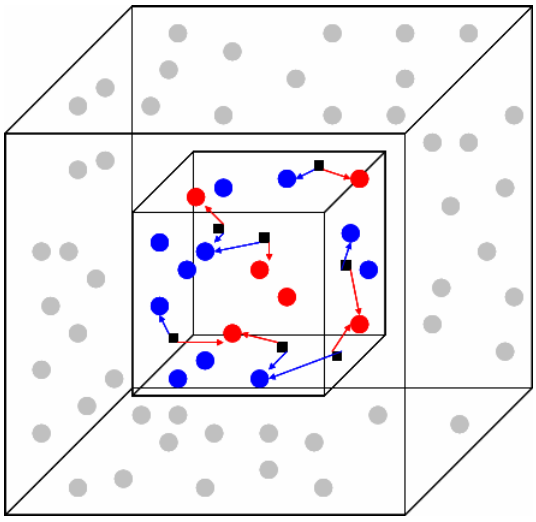


**Figura 1.19**. Representación del *diversity integral criterion*. Para las dos selecciones *A* (*azul*) y B (*rojo*) se extienden puntos aleatorios exclusivamente en el espacio químico definido por ambas (*cuadrados negros*) y se cuantifica la distancia de cada uno de ellos al compuesto más cercano de cada selección. La quimioteca total se representa por los puntos grises.

# Capítulo 2.

# Tirosina Quinasas

## 2.1. Proteína Tirosina Quinasas

El interés farmacológico de esta tesis es la inhibición de receptores tirosina quinasas. En este capítulo se describe el papel que estos receptores juegan en los mecanismos de señalización intracelular, su interés terapéutico, su caracterización estructural y los diferentes mecanismos de inhibición desarrollados.

Uno de los mecanismos fundamentales por los que las células eucariotas se comunican es mediante la unión de ligandos a la superficie de receptores celulares que actúan directamente como enzimas o están asociados a enzimas. Entre ellos, la mayor parte corresponden a proteína quinasas: tirosina quinasas o serina/treonina quinasas, que fosforilan determinados residuos de tirosina, serina o treonina de proteínas señal intracelulares o bien están asociados a proteínas que tienen actividad tirosina quinasa.

La importancia de la fosforilación de proteínas en la regulación de la vida celular eucariótica se refleja en el hecho de que en un 2% de los genes eucariotas se encuentran dominios con actividad quinasa.[315] Así, las proteína quinasas se han convertido en el segundo grupo de dianas farmacológicas, tras los receptores acoplados a proteínas G (GPCRs), cubriendo el 20-30% de los proyectos de descubrimiento de fármacos en muchas compañías farmaceúticas.[316]

El *quinoma* humano contiene 518 proteínas quinasa, de las cuales 478 pertenecen a una única superfamilia cuyos dominios catalíticos están relacionados en secuencia. Éstos se pueden agrupar en 7 grupos, 20 familias y subfamilias, con creciente similitud de secuencia y función bioquímica.[317] Las proteína tirosina quinasas (PTKs) forman un único grupo, correspondiendo los seis restantes a serina/treonina quinasas. Además, se han secuenciado 40 quinasas "atípicas" que no comparten similitud secuencial con el resto, pero cuya actividad enzimática y/o plegamiento estructural es conocido o previsto similar al de una proteína quinasa. El árbol del quinoma humano se encuentra accesible a través de diferentes servidores *web* como son el *Protein Kinase Resource*[230] , *Cell Signaling Technology, Inc*[318] y *Evolutionary Bioinformatics and Sugen, Inc*[319]. Esta clasificación rebasa la previamente utilizada, propuesta por Hanks y Quinn en el año 1991[320].

En la Figura 2.1 se muestra el árbol filogenético del quinoma humano correspondiente al grupo de las proteína tirosina quinasas (PTKs), en el que se ha centrado el trabajo. La reacción específica catalizada por las PTKs es la transferencia del fosfato γ del ATP al grupo hidroxilo de la tirosina de la proteína diana. Las PTKs se diferencian tradicionalmente en dos subgrupos:

- Los receptores tirosina quinasa (RTKs): son glicoproteínas transmembrana que se activan por la unión de sus ligandos y transducen la señal extracelular al citoplasma mediante autofosforilación y posterior fosforilación de proteínas intracelulares. Esta familia incluye los receptores de insulina y muchos receptores de factores de crecimiento como el factor de crecimiento epitelial (EGF), los factores de crecimiento de los fibroblastos (FGF), el factor de crecimiento derivado de las plaquetas (PDGF), el factor de crecimiento vascular endotelial (VEGF), el factor de crecimiento de los hepatocitos (HGF), el factor de crecimiento neuronal (NGF) y el factor estimulador de la formación de colonias de macrófagos (M-CSF). Estos receptores se componen de un dominio extracelular, implicado en la unión del ligando y la dimerización del receptor (véase abajo), un único

dominio transmembrana y un dominio citoplasmático que contiene el dominio catalítico tirosina quinasa, así como diversas secuencias reguladoras.

- La familia de tirosina quinasas no receptoras (NRTKs): componentes integrales de las cascadas de señalización iniciadas por las RTKs y otros receptores de la superficie celular como las GPCRs y los receptores del sistema inmunológico. La mayor parte se localizan en el citoplasma, aunque algunas se encuentran ancladas en la membrana celular. Se incluyen la familia Src, la familia Janus (Jaks) y otras como Tec, Fes, Abl, FAK y Syk.

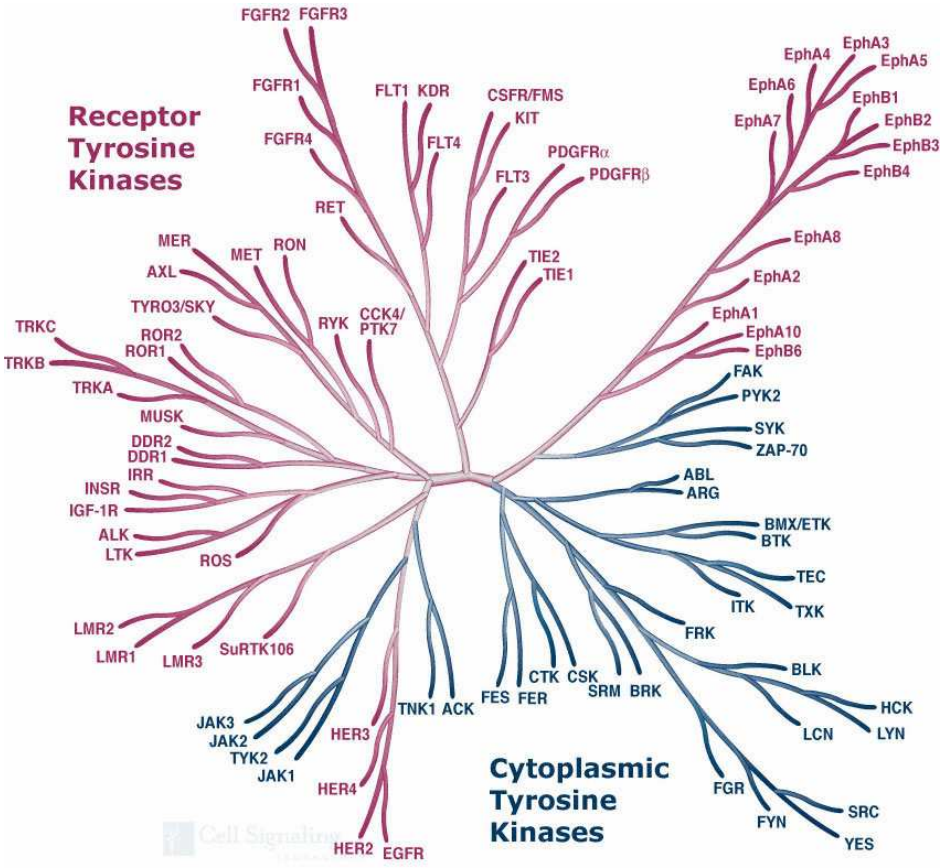



**Figura 2.1.** Grupo de PTKs del quinoma humano. Extraído de [230].

## 2.2. Señalización Celular en Tirosina Quinasas

Las RTKs activan, en respuesta a los factores de crecimiento, numerosas vías de señalización que generan respuestas celulares tales como la mitogénesis y proliferación, diferenciación, migración, la supervivencia celular, la prevención o inducción de apoptosis, el reordenamiento del citoesqueleto y cambios metabólicos.

Esta variedad de respuestas ante un mismo estímulo puede depender del tipo celular y más genéricamente de las diversas condiciones fisiológicas a las cuales estén sometidas las células. Así, en cultivos celulares estas respuestas pueden depender de la densidad celular de los cultivos, del tipo de matriz extracelular a la que estén adheridas las células o de la presencia en el medio de otros factores de crecimiento u hormonas, ya que normalmente estos actúan en combinaciones específicas. Por ejemplo, un número pequeño de factores de crecimiento pueden servir, en combinaciones diferentes, para regular selectivamente la proliferación de cada una de las diferentes clases de células de un animal superior.

Los factores de crecimiento pueden presentar una especificidad amplia (como EGF, FGF y PDG) o reducida (NGF). Mayoritariamente se encuentran implicados en regulaciones paracrinas (mediadores locales), aunque algunos están presentes en la circulación.

### 2.2.1. Activación de los Receptores de Tirosina Quinasa

La mayor parte de las RTKs existen como monómeros en la membrana celular, siendo las dos principales excepciones la familia de receptores de insulina (tetrámeros $\alpha_2\beta_2$) y la familia Met. La unión del ligando a los receptores monoméricos induce la dimerización de éstos, produciendo un acercamiento de sus extremos que permiten que los dominios TK interaccionen y se autofosforilen (por trans-fosforilación, se ha descartado la posibilidad de una cis-fosforilación), conduciendo a su activación.

El mecanismo de dimerización difiere entre distintas RTKs. Se pueden unir ligandos monoméricos bivalentes, homodímeros o heterodímeros. Éstos pueden ser factores solubles o estar unidos a la membrana (receptores Eph), o pueden requerirse factores adicionales como los heparina sulfato proteoglicanos, en el caso de FGFR. En algunos casos, la dimerización por sí sola no es capaz de activar toda la funcionalidad posible, necesitándose oligomerizaciones (caso de los receptores Eph). Además, los dímeros formados pueden ser homodímeros o heterodímeros, compuestos por dos cadenas de RTKs de una misma familia[321], aunque no todas las configuraciones diméricas de un receptor son capaces de señalizar.

La activación por autofosforilación no solo aumenta la actividad del dominio catalítico, sino que se hacen accesibles determinados sitios de unión con tirosinas autofosforiladas, normalmente fuera del dominio TK, que reclutan proteínas para ser fosforiladas, continuándose la cascada de señalización.
Estas proteínas reclutadas poseen dominios no catalíticos altamente conservados SH2 (*Src homology 2 domain*) o dominios PTB (*phosphotyrosine binding*). Los dominios SH2 se unen específicamente a secuencias de aminoácidos definidas por 1-6 residuos C-terminales a una fosfotirosina. Por su parte, los dominios PTB reconocen secuencias de 3-5 aminoácidos N-terminales a una tirosina, fosforilada o no. Estas proteínas pueden ser de dos tipos: i) proteínas adaptadoras, sin actividad catalítica, que pueden reclutar a otras proteínas transductoras, o ii) factores o enzimas directamente transductores/as que tras unirse al receptor son fosforilados por éste, pasando de un estado inactivo a otro activo. Las proteínas adaptadoras poseen también dominios SH3 ó WW que reconocen motivos ricos en prolinas, permitiendo el ensamblaje de complejos de proteínas a través de uniones SH2 y SH3[322].
Además, existen proteínas de reclutamiento (*docking proteins*) con dominios señal dirigidos a los fosfolípidos de la membrana celular (como el dominio PH, *pleckstrin homology domains*) que permiten la translocación a la membrana de proteínas de señalización, dominios SH2 para unirse a estas proteínas y dominios PTB que se unen al receptor. Destacan las familias IRS y FRS como *docking proteins* de los receptores IR (receptor de insulina) y FGFR.

Por lo tanto, mediante estos reclutamientos y/o fosforilaciones se producen cambios conformacionales y/o cambios en la localización intracelular de estas proteínas señalizadoras, siendo así capaces éstas de transmitir sus mensajes a otros componentes de las diversas rutas intracelulares de transducción de señales. Otras NRTKs utilizan dominios específicos de la subfamilia para mediar las interacciones proteína-proteína.

### 2.2.2. Mecanismos de Señalización Intracelular

De manera general, los mecanismos intracelulares normalmente terminan en el núcleo celular, resultando en la activación de distintos factores de transcripción que regulan la expresión génica. En la Figura 2.2 se recogen los distintos mecanismos de señalización activados por RTKs.

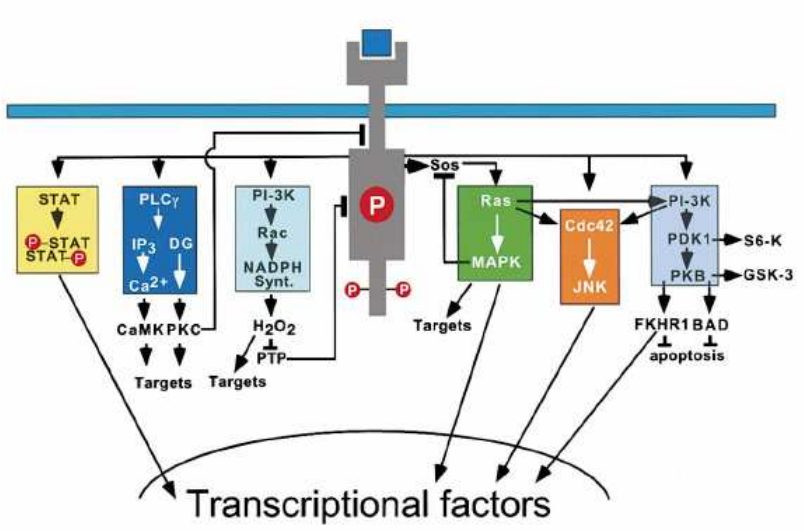



**Figura 2.2.** Mecanismos de señalización activados por RTKs. Extraído de [321].

- Cascada MAPK (*Mitogen Activated Proteins Kinases*) / Erk (*Extracellular Signal Regulated Kinases*). Implicada en el crecimiento y diferenciación celular. Requiere la activación por transferencia de GTP de proteínas GTPasas monoméricas como las proteínas de la familia Ras (Ras y Rap1). La activación de Ras es mediada por el factor de liberación de nucleótidos de guanina SOS, que a su vez transloca a la membrana por formación del complejo Grb2-SOS. Grb2 es una proteína adaptadora que puede interaccionar directamente por sus dominios SH2 con los RTKs (como en el caso del EGFR) o alternativamente interaccionar indirectamente con otras proteínas adaptadoras acopladas a los RTKs (como Shc en EGFR, unida vía dominios PTB) o *docking proteins* como FRS2α en FGFR (directa o indirectamente por acoplamiento a la fosfatasa Shp2).
  La activación de Ras inicia la cascada MAP/Erk, que consta de tres serina/treonina quinasas secuenciales (Raf, Mek, Erk). Una vez que las ERK1/2 son activadas, éstas pueden fosforilar a diferentes proteínas dianas localizadas en la membrana plasmática y en el citoplasma, dando lugar a la activación de otras vías de señalización o translocarse al núcleo y fosforilar diversos factores de transcripción como son, entre otros, c-Myc, c-Jun, c-Fos, Elk-1 y p62TCF, produciendo así la activación o la represión transcripcional de determinados genes.

- Cascadas MAPK alternativas como la JNK/SAPK (*c-Jun N-terminal kinase, Stress-activated protein kinases*), que interviene en respuesta a numerosas situaciones de estrés medioambiental, también puede iniciarse tras la activación de Ras por una RTK. También la cascada p38 MAPK, se activa en respuesta a factores de crecimiento como el factor de crecimiento neuronal o el tipo-insulina.

- Activación de las proteínas activadoras de GTPasa (GAP) que se unen directamente a los RTKs e incrementan la velocidad de hidrólisis del GTP unido a Ras, inactivándolo.

- Activación de la fosfolipasa C-γ (PLC-γ) por unión directa de sus dominios SH2 al RTK. Este enzima hidroliza el fosfatidilinositol 4,5-bisfosfato ($PIP_2$) generando inositol 1,4,5-trisfosfato ($IP_3$) y 1,2-diacilglicerol (DAG). Tanto el $IP_3$ como el DAG son potentes mensajeros secundarios. El $IP_3$ es un efector de canales de calcio localizados en la membrana del retículo endo/sarcoplásmico que estimulan la liberación de $Ca^{2+}$. Este $Ca^{2+}$ se une a la calmodulina, activando la familia de quinasas dependientes de calmodulina. Además, el DAG y $Ca^{2+}$ activan la proteína quinasa C (PKC). Además de una serie de respuestas intracelulares, como puede ser el reordenamiento del citoesqueleto mediado por $Ca^{2+}$, los efectos de DAG y $Ca^{2+}$ se transducen en la activación de ciertos factores de transcripción.

- Activación de la fosfatidilinositol 3'-quinasa (PI-3K). Las PI-3K de la clase I son heterodímeros compuestos de una subunidad reguladora, p85, con dominios SH2 y SH3 y una subunidad catalítica, p110. PI-3K puede interaccionar directamente a través de sus dominios SH2 con el receptor (caso del receptor ErbB3 de la familia EGFR) o bien hacerlo con *docking proteins*, como la proteína Gab1 en EGFR y FGFR. La PI-3K activada fosforila el $PIP_2$, generándose fosfatidilinositol 3,4,5-trisfosfato ($PIP_3$). Éste es un potente efector que se une a proteínas que contienen dominios PH, interviniendo en la translocación de una variedad de proteínas de señalización y su activación:

  - Una de ellas es la serina/treonina quinasa PKB/Akt (PKB, por *protein kinase B*; y Akt, por ser homóloga de la oncoproteína v-Akt). La PKB/Akt activada fosforila a multitud de proteínas sustrato generando, entre otras, señales de supervivencia celular que previenen la aparición de apoptosis. Por una parte, inactiva caspasas (procaspasa 9), suprime la expresión de genes proapoptóticos e inhibe la formación del complejo apoptótico BAD-Bcl2.
  - En el receptor de insulina, la activación de PI-3K conduce a la translocación de los transportadores de glucosa a la membrana celular.
  - Interviene en la generación de $H_2O_2$ inducida por factores de crecimiento. $H_2O_2$, entre otras respuestas, inactiva a la fosfatasa PTP, que desfosforila la EGFR activada.

- Activación de la cascada JAK/STAT (*signal transducers and activators of transcription*). La fosforilación de las quinasas de Janus (JAK) permite el acoplamiento y fosforilación de STATS, que dimerizan y migran al núcleo, aumentando la expresión del inhibidor del ciclo celular p21WAF1/CIP1, quedando así bloqueado el mismo, y de la caspasa 1, proteasa implicada en apoptosis. Esta parada de la proliferación celular e inducción de apoptosis, opuesta a la respuesta anterior, se realiza dependiendo del estado de la célula (células tumorales).

- Finalmente, los factores de crecimiento inducen la transcripción de genes tardíos como los de las ciclinas y las quinasas dependientes de ciclina (CDKs), que intervienen en la progresión de las células desde la fase G1 del ciclo celular a la fase S.[323]

Por otra parte, las NRTKs, además de estar integradas en los mecanismos iniciados por las RTKs, intervienen en el funcionamiento del sistema inmunológico. La familia Jak está asociada a receptores de citoquinas (como el interferón γ), cuya activación conduce a la transcripción de genes específicos mediante el sistema JAK/STAT. La quinasa Lck, un miembro de la familia Src, está constitutivamente asociada a los receptores CD4 y CD8 de los linfocitos T, que una vez estimulados, transducen la señal a través de la quinasas Lck y ZAP-70, que finalmente deriva en la activación transcripcional de genes de citoquinas que intervienen en la activación de las células T. Análogamente, en la activación de las células B intervienen las NRTKs Lyn y Syk.

En la Tabla 2.1 se recogen las actividades más representativas en que participan los tres receptores de factores de crecimiento con los que se trabaja: EGFR, PDGFR y FGFR.

**Tabla 2.1.** Familias de receptores de factores de crecimiento estudiados en el trabajo, sus ligandos, receptores y funciones representativas.

| | **Ligandos** | **Receptores** | **Actividades Representativas** |
|---|---|---|---|
| Familia EGFR[324] | EGF, TFG-α, BTC, HB-EGF | EGFR (ErbB1) ErbB2 (Neu) ErbB3 y ErbB4. | Diferenciación, proliferación de muchos tipos celulares. Señal inductora durante el desarrollo embrionario. Apoptosis. |
| Familia PDGFR[325] | PDGF-AA PDGF-AB PDGF-BB PDGF-CC PDGF-DD | PDGFR-α PDGFR-β | Mitogenicidad de células del tejido conjuntivo. Quimiotaxis. Reordenamiento filamentos de actina. Movilización del $Ca^{2+}$. Inhibición apoptosis (PDGFR-ββ) |
| Familia FGFR[326] | 22 FGFs distintos. | Cuatro genes (FGFR1-FGFR4) que generan distintas isoformas por *splicing* alternativo | Durante el desarrollo embrionario desempeñan un papel crítico en la morfogénesis, regulando la proliferación, diferenciación y migración celulares. En los organismos adultos, intervienen en el control del sistema nervioso central, la reparación de tejidos y la angiogénesis tumoral. |

## 2.3 Proteína Tirosina Quinasas / Implicación Terapéutica

Las proteína quinasas han surgido como dianas farmacológicas en numerosas enfermedades, bien porque se encuentran sobreexpresadas y/o muestran una disfunción en un órgano o tejido particulares, o por el papel que desempeñan en mecanismos del ciclo celular implicados en distintas enfermedades.

Entre estas enfermedades, el cáncer focaliza la mayor parte de estudios, no solo por su predominancia en la población occidental, sino también porque su estudio permite avanzar en el conocimiento de las pautas de comportamiento de las células en organismos pluricelulares.

Las células cancerosas se caracterizan por una proliferación celular incontrolada y porque invaden y colonizan territorios normalmente reservados para otras células.[327] En la mayor parte de cánceres, las anomalías que presentan estas células se transmiten a su progenie gracias a que son debidas a cambios genéticos (alteraciones en la secuencia de DNA), aunque también pueden tener un origen epigenético (cambios en la pauta de expresión génica, sin que exista ningún cambio en la secuencia de DNA).
El análisis de las alteraciones genéticas en células cancerosas ha revelado un gran número de genes que codifican proteínas implicadas en el control de la proliferación celular. Por una parte, genes cuyos productos ayudan a estimular la proliferación celular, de manera que al mutar se sobreexpresan o se vuelven hiperactivos, denominándose *oncogenes* (siendo el alelo normal un proto-oncogén). Por otra, los genes que inhiben la proliferación celular sufren mutaciones que los inactivan, denominándose entonces *genes supresores de tumores*. Los primeros presentan un fenotipo dominante, únicamente se requiere la activación de una única copia del proto-oncogén, mientras que la mutación de los genes supresores de tumores tienen un efecto recesivo, las dos copias del gen en la célula deben estar inactivadas o delecionadas.

La identificación de proto-oncogenes en células normales ha sido posible en muchos casos gracias al estudio de retrovirus que los transforman en oncogenes y actúan de vectores, insertándolos en el DNA de una célula huésped. Sin embargo, existen otros agentes carcinógenos externos que producen mutagénesis, tales como radiaciones ionizantes o productos químicos.
Los proto-oncogenes incluyen ejemplos de prácticamente todos los tipos de proteínas que intervienen en los mecanismos de señalización molecular descritos en el apartado anterior: factores de crecimiento y sus receptores RTK, proteínas quinasa citoplasmáticas (Src), proteínas Ras, proteínas serina/treonina (cascadas MAPK) y proteínas nucleares de regulación génica (Myc, Fos, Jun).

En particular, la sobreexpresión y/o alteración estructural de las RTKs están frecuentemente asociadas a cánceres en humanos y numerosas células tumorales utilizan los mecanismos de transducción mediados por RTK para conseguir el crecimiento tumoral, la angiogénesis (las células tumorales estimulan la generación de vasos sanguíneos que les aporten nutrientes y oxígeno) y metástasis (propagación de un tumor a sitios diferentes del origen). Así, las RTKs activan muchas proteínas de señalización que no están *per se* implicadas en el proceso de proliferación tumoral (activado por las cascadas dependientes de Ras), pero que también contribuyen a la oncogénesis (cambios en el citoesqueleto, movilidad celular, angiogénesis tumoral y cambios en la supervivencia celular).[328]

EGFR se utiliza como un marcador tumoral en numerosos tipos de cáncer en los que se encuentra sobreexpresado (colon, cabeza y cuello, páncreas, ovario, mama, riñón, gliomas). Por otra parte, se encuentran alteraciones en PDGF y su receptor en cánceres como el de pulmón, próstata, renal, glioblastoma y la leucemia crónica monomielocítica. Además, este PDGFR tiene una gran importancia en la angiogénesis tumoral.

Además del cáncer, los RTKs intervienen en distintas enfermedades asociadas con desórdenes hiperproliferativos, migratorios, del desarrollo embrionario y enfermedades vasculares, tales como arteriosclerosis, la psoriasis, la artritis reumatoide, la retinopatía diabética, homeostasis del fosfato, displasias esqueléticas o fibrosis[325,326]. En la referencia [329] puede encontrarse una recopilación de las distintas enfermedades en las que están implicadas las quinasas humanas.

## 2.4. Caracterización Estructural de los Receptores de Tirosina Quinasa: dominio Tirosina Quinasa

Como se ha comentado, las RTKs consisten de una porción extracelular, una hélice transmembrana y una porción catalítica.

La porción extracelular contiene típicamente un conjunto diverso de dominios globulares, como los dominios tipo inmunoglobulinas (Ig), dominios de tipo fibronectina III, dominios ricos en cisteína y dominios tipo EGF.

La porción intracelular es más simple, compuesta de una región yuxtamembrana situada tras la región transmembrana, seguida del dominio catalítico tirosina quinasa y de una región carboxi-terminal. Algunos receptores, principalmente los de la familia del receptor PDGFR poseen además una inserción de unos 100 aminoácidos, denominado KID (*kinase insert domain*), que no es conservado entre los distintos receptores y pese a que no es necesario para la actividad catalítica, es un sitio de autofosforilación e interacción con otras proteínas (como Grb2 y PI-3K).
Los dominios yuxtamembrana y carboxi-terminal varían en longitud en distintas RTKs y, junto con KID, contienen residuos de tirosina que son autofosforilados tras la activación de la actividad catalítica del dominio tirosina quinasa.

En la Figura 2.3 se esquematizan las estructuras de distintos RTKs.

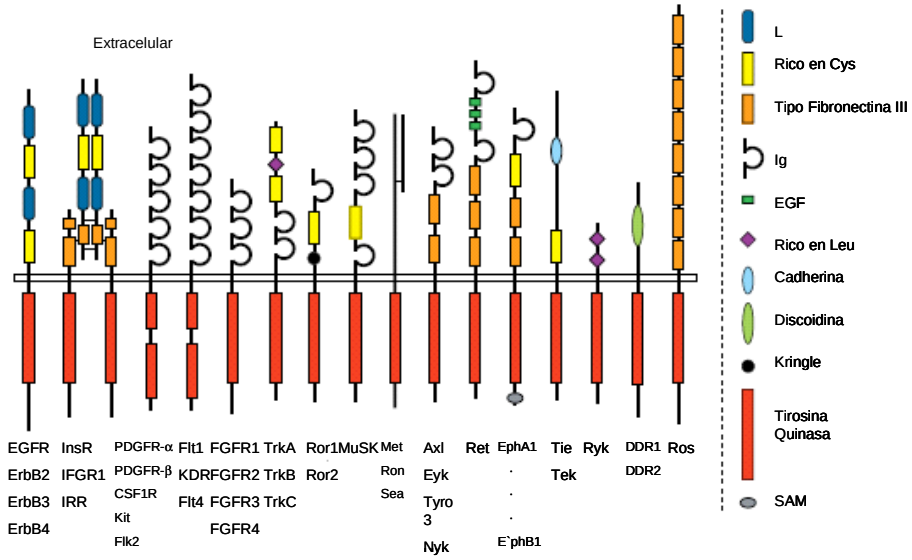



**Figura 2.3.** Organización de dominios en RTKs. El dominio KID se muestra como una línea negra que divide el dominio tirosina quinasa. Adaptado de [330].

Los dominios quinasa comprenden entre 250 y 300 aminoácidos, con un peso alrededor de 30kD. Se trata de un dominio muy similar entre serina/treonina quinasas y tirosina quinasas, aunque existen diferencias a nivel de secuencia que caracterizan cada familia, de manera que permiten distinguir si una secuencia putativa es de un tipo u otro.

A nivel de secuencia, Hanks[331] estableció once subdominios conservados (I-XI) a partir del alineamiento múltiple de quinasas, separados por regiones menos conservadas, donde se encuentran *gaps* e insertos. Así, el dominio KID de PDGFR, CSF1R y Kit aparece entre los subdominios V y VI.

La arquitectura general del dominio quinasa es bilobular: con un lóbulo N-terminal y un lóbulo C-terminal. El lóbulo N-terminal comprende cinco láminas antiparalelas β (β1-β5) y una hélice α (αC), previa a la hélice αC se encuentra otra hélice (αB), aunque ésta última no está tan conservada entre quinasas. El lóbulo C-terminal, más grande que el anterior, está formado por dos láminas β (β7, β8) y siete hélices α (αD, αE, αEF, αF-αI). También se puede encontrar en algunas quinasas otra lámina β (β9). El lóbulo N-terminal está asociado a la unión del ATP, mientras que el extremo C-terminal lo está con la catálisis y la unión del sustrato (véase Figura 2.4).

La notación de la estructura secundaria es la aceptada e introducida por primera vez por Knighton[333], en la publicación de la primera estructura resuelta de una proteína quinasa, la PKA (*cAMP-dependent protein kinase*). Posteriormente, la resolución de muchos otros dominios quinasa ha permitido comprobar que, aunque el porcentaje de identidad de secuencia total entre todas ellas no es muy elevado (evidentemente, dentro de una familia éste se incrementa), el plegamiento general se conserva. De hecho, existen varios residuos conservados, nueve de ellos invariablemente y el resto altamente conservados entre las quinasas, implicados en la funcionalidad del dominio. En la Tabla 2.2 se detallan los correspondientes residuos conservados y su funcionalidad siguiendo la secuencia de PKA, ya que no existe una nomenclatura de residuos, aunque estos se conservan en los alineamientos múltiples de dominios quinasa como el presentado en las referencias [320] y [331].

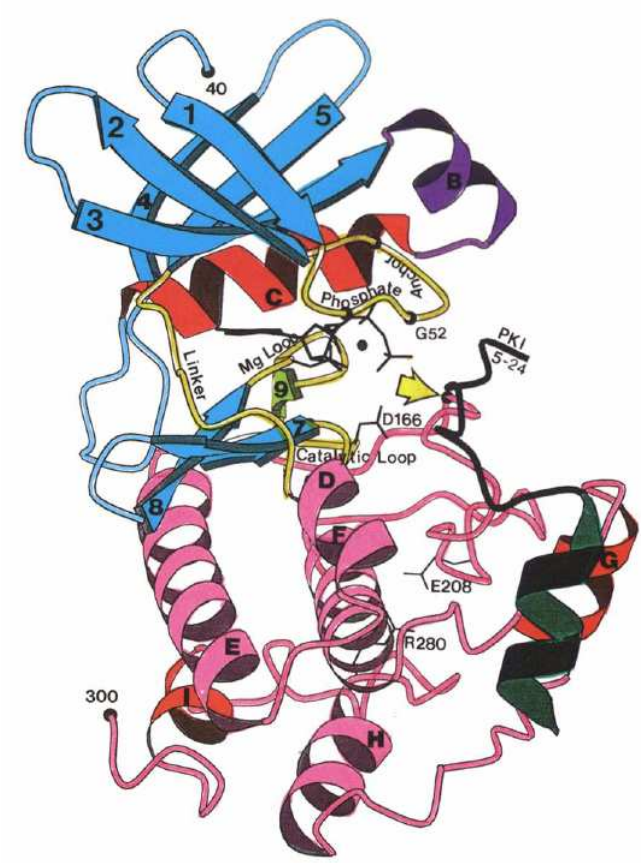



**Figura 2.4.** Esquema de la estructura del PKA. Extraído de [332].

**Tabla 2.2.** Residuos conservados (notación secuencial de PKA) en el dominio catalítico quinasa. Adaptado de [332].

| Residuo PKA | Función | Número de dominio de Hanks | Situación en la estructura secundaria |
|---|---|---|---|
| $Gly^{50}$ | *Loop* que ancla el β-$PO_4$ de ATP | I | *Loop* de unión del ATP entre β1 y β2 |
| $Glu^{52}$ | *Loop* que ancla el β-$PO_4$ de ATP | I | *Loop* de unión del ATP entre β1 y β2 |
| $Gly^{55}$ | *Loop* que ancla el β-$PO_4$ de ATP | I | *Loop* de unión del ATP entre β1 y β2 |
| $Val^{57}$ | Alinea el sitio de unión de la adenina del ATP | | β2 |
| $Lys^{72}$ | Forma un par iónico con α-$PO_4$ y β-$PO_4$ de ATP | II | β3 |
| $Glu^{91}$ | Forma un par iónico con la Lys72 | III | αC |
| $Asp^{166}$ | Base catalítica | VIb | *Loop* catalítico entre β6 y β7 |
| $Lys^{168}$ | Interacciona con γ-$PO_4$ de ATP | VIb | *Loop* catalítico entre β6 y β7 |
| $Asn^{171}$ | Quela $Mg^{2+}$ en PKA | VIb | *Loop* catalítico entre β6 y β7 |
| $Asp^{184}$ | Quela $Mg^{2+}$ en PKA | VII | Inicio del *loop* de activación (tras β8) |
| $Phe^{185}$ | | | Inicio del *loop* de activación (tras β8) |
| $Gly^{186}$ | | | Inicio del *loop* de activación (tras β8) |
| $Glu^{208}$ | Forma un par iónico con Arg280 | VIII | Extremo del *loop* P+1 |
| $Asp^{220}$ | Estabiliza el *loop* catalítico | IX | αF |
| $Arg^{280}$ | Forma un par iónico con Glu208 | XI | *Loop* entre αH y αI |

Por su papel en la fosforilación, destacan las regiones[330-333]:

- *Loop* de unión del nucleótido (*nucleotide-binding loop*). Corresponde al sitio de unión del ATP situado en la hendidura situada entre los dos lóbulos, de manera que el nucleótido queda coordinado por los residuos de las láminas β1-β2 del lóbulo N-terminal. Se encuentra un motivo de glicinas: Gly-X-Gly-X-X-Gly, también muy conservado entre proteínas que unen nucleótidos. Además, en muchas TKs, 14 residuos antes de la primera Gly del motivo consenso, se encuentra un motivo WE que estabiliza la estructura en el lóbulo N-terminal y que parece demarcar el límite entre el dominio quinasa y la región yuxtamembrana precedente. También se encuentra, casi invariablemente, una valina situada a dos posiciones del extremo carboxi del motivo Gly-X-Gly-X-X-Gly y que se posiciona en la parte superior de la adenina del ATP.

- *Loop* catalítico (*catalytic loop*): En el dominio VIb se encuentra el *loop* que interviene en la transferencia de fosfato. De hecho, su secuencia permite determinar si se trata de una serina/treonina quinasa o de una tirosina quinasa. En el primer caso, la secuencia corresponde a Asp-Leu-Lys-Pro-Glu-Asn (como en el caso del PKA entre $Asp^{166}$ y $Asn^{171}$), mientras que en las secuencias Asp-Leu-Arg-Ala-Ala-Asn o Asp-Leu-Ala-Ala-Arg-Asn indican especificidad por tirosina en la fosforilación.

- *Loop* de activación (*activation loop*). Ocupa los subdominios VII-VIII. El motivo invariante Asp-Phe-Gly ($Asp^{184}$-$Phe^{185}$-$Gly^{186}$ en PKA) se encuentra al comienzo del *loop* de activación y está implicado en la unión de Mg-ATP. El *loop* termina con un motivo conservado: Ala-Phe-Glu. Las RTKs poseen de 1 a 3 tirosinas en el *loop* de activación quinasa. La fosforilación de estos residuos es crítica para estimular la actividad catalítica y biológica de una gran parte de RTKs, como IR, FGFR, VEGFR, PDGFR y Met (*hepatocyte growth factor receptor*). Una gran excepción supone el EGFR, ya que la mutación de los residuos Tyr del *loop* de activación por Phe no afecta a las propiedades de señalización de dicho receptor[329]. Es un *loop* con una gran movilidad. Los residuos que preceden al *loop* de activación se denominan *hinge residues*.

- *Loop* P+1: Situado en el dominio VIII e incluido en el *loop* de activación. En su extremo se encuentra el triplete consenso Ala-Pro-Glu ($Ala^{206}$-$Pro^{207}$-$Glu^{208}$ en PKA). El $Glu^{208}$ forma un par iónico, conservado entre quinasas, con una arginina del dominio XI ($Arg^{280}$ en PKA). Las secuencias que preceden a este triplete, también son indicativo de la especificidad de la quinasa (si es tirosina o serina/treonina quinasa). Es el *loop* que reconoce el residuo contiguo al residuo diana del sustrato peptídico. En general, el sitio de unión del sustrato peptídico se extiende al final del *loop* de activación y muestra mayor variabilidad de secuencia que el sitio de unión del ATP y el *loop* catalítico.

La orientación relativa de los dos lóbulos muestra una considerable variabilidad entre proteína quinasas. La forma apo desfosforilada se encuentra en una conformación más abierta, que se cierra tras la activación. El mecanismo de autoinhibición observado en estructuras cristalográficas no fosforiladas sugiere que el *loop* de activación bloquea el sitio de unión del ATP y/o el sitio de unión del sustrato, y que tras la autofosforilación, dicho *loop* se estabiliza en una conformación no-inhibitoria, sufriendo un gran cambio conformacional.[330]

## 2.5. Inhibidores de Tirosina Quinasas

Se han desarrollado distintas estrategias para prevenir la activación de los RTKs: desde anticuerpos monoclonales que se unen selectivamente a su porción extracelular (como por ejemplo para el EGFR y VEGFR) bloqueando su unión con el ligando natural, hasta fármacos que inhiben la actividad quinasa del receptor. En este apartado, se describen aquellos compuestos diseñados para interferir en el sitio de unión del ATP.

Actualmente, superado un escepticismo inicial, el sitio de unión del ATP se considera una diana farmacológica, a pesar de las dos desventajas asociadas a él: i) la necesidad de obtener una potencia suficiente como para competir con la gran concentración de ATP intracelular *in vivo* y ii) la naturaleza ubicua del sitio de unión del ATP, con los problemas asociados de selectividad que conlleva. Normalmente, los inhibidores están dirigidos a la conformación activa de la proteína, aunque resultan más interesantes aquellos que se dirigen a la conformación inactiva, ya que es más fácil conseguir especificidad para el sitio de unión del ATP en esta situación.

De hecho, en 2004, veinte inhibidores se encontraban en fase clínica y tres habían sido aprobados: Gleevec® (STI-571, *imatinib mesylate*), dirigido contra c-Kit / PDGFR; gefitinib (ZD1839, IRESSA®), dirigido contra EGFR y erlotinib (CP358,774, Tarceva®), dirigido contra EGFR.[334]

A continuación, se muestran aquellos *scaffolds* para los que se ha encontrado actividad:

**Quinazolinas**

Se han realizado numerosos estudios SAR (*structure–activity relationship*) y pruebas biológicas sobre este *scaffold*, encontrándose los compuestos **1**-**2** en pruebas clínicas[335] y **3** (erlotinib), para el tratamiento de cáncer de pulmón y páncreas (Figura 2.5).

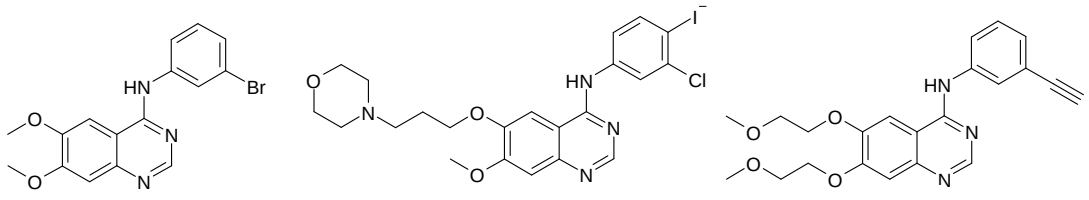
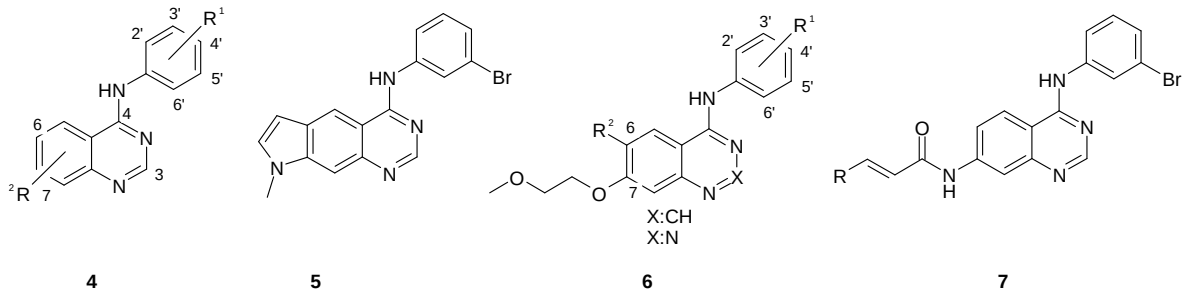



**Figura 2.5.** Compuestos representativos de quinazolinas en fase clínica y aprobados.

Entre las diversas derivaciones destacan las sustituciones en las posiciones 3-, 4-, 6-, o 7-, (**4, 6**) así como análogos de quinazolinas tricíclicos (**5**) (Figura 2.6).

Muchas de ellas son inhibidores del EGFR en el orden submicromolar y nanomolar, con un buen perfil de selectividad. En la referencia [336] puede encontrarse una revisión del SAR de estos compuestos frente a EGFR. Por otra parte, también se han diseñado quinazolinas con mayor selectividad hacia otras dianas como Raf, CSF-1R y VEGFR[337].




**4** **5** **6** **7**

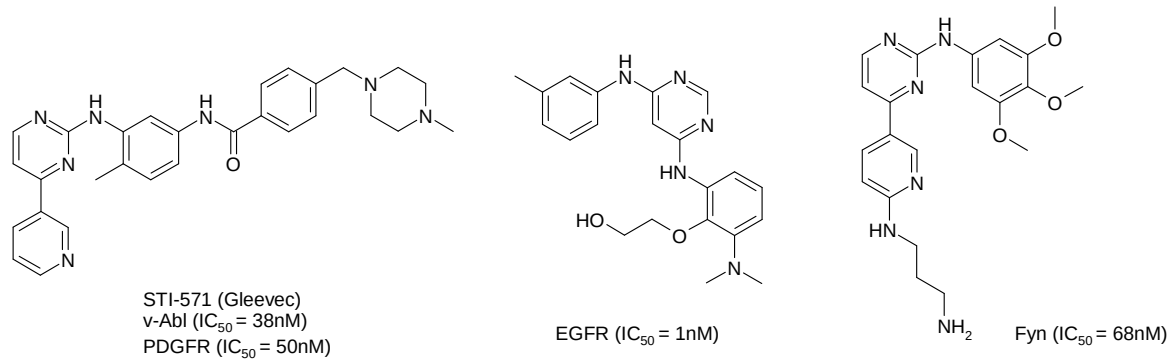
**Figura 2.6.** *Scaffolds* de quinazolina derivatizados.

La derivatización de las posiciones 6 y 7 responde principalmente a un aumento de solubilidad, aunque C-6 es más restrictivo en lo referente a los sustituyentes. Por otra parte, la sustitución en el anillo de anilina en 3-bromo o 3-cloro y 4-flúor produce un aumento de la actividad, frente a otras sustituciones.

Estos inhibidores, son ATP-competitivos reversibles. Sin embargo, también se han desarrollado una nueva clase de inhibidores irreversibles[338] con potencia subnanomolar para los receptores EGFR y erbB-2. Éstos, representados por el compuesto **7**, contienen un aceptor de Michael en la posición 6- ó 7- del anillo de quinazolina, de manera que se unen irreversiblemente a una cisteína ($Cys^{773}$) del sitio de unión del ATP en el EGFR, que es única para esta familia de quinasas, lo que les confiere una gran selectividad frente a otras quinasas.

**Fenilaminopirimidinas**

En este grupo se encuentra STI-571 o Gleevec® (**8**) (Figura 2.7), que inhibe a v-Abl y PDGFR. Aunque inicialmente se identificaron como inhibidores del receptor de PDGF y de PKC, la selectividad por PDGFR se consiguió mediante la introducción del grupo metilo en la posición 6-del fenilo. La potencia frente a v-Abl se obtuvo derivatizando los sustituyentes del fenilo.[339] También se han descrito 4,6-dianilinopirimidinas (**9**) como inhibidores de EGFR y 2-anilinopirimidinas (**10**) como inhibidores de Lck, Fyn, ZAP-70, Csk, EGFR y PKC.




**8** **9** **10**

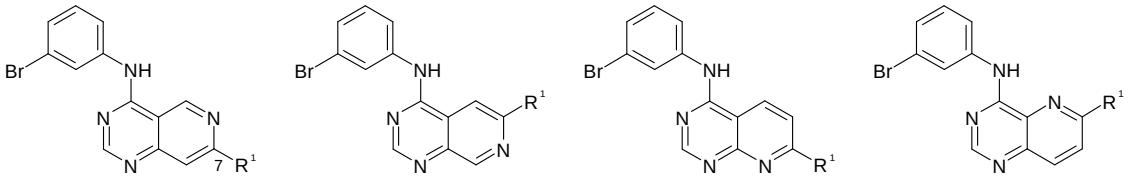
**Figura 2.7.** Representantes de fenilaminopirimidinas.

**Piridopirimidinas y pirimidopirimidinas**

Se han descrito pirido[4,3-*d*], pirido[3,4-*d*], pirido[2,3-*d*] y pirido[3,2-*d*]pirimidinas como inhibidores de una gran cantidad de quinasas (Figura 2.8).

En un estudio inicial SAR realizado por Rewcastle *et al*[340] se compararon estos cuatro *scaffolds* según su capacidad de inhibir el receptor de EGF, encontrándose que las series [3,4-*d*] (**12**) y

[4,3-*d*] (**11**) eran las más activas, seguidas de [3,2-*d*] (**14**) y siendo los compuestos [2,3-*d*] (**13**) los menos potentes, para los compuestos sintetizados.

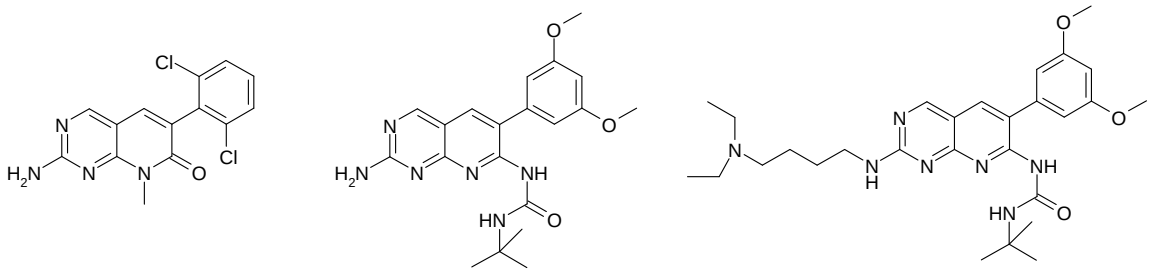



**11** **12** **13** **14**

**Figura 2.8.** Series de piridopirimidinas testadas por Rewcastle frente a EGFR.

Además, la introducción de sustituyentes básicos débiles en la posición 7- de la serie [4,3-*d*] (**11**), permite aumentar no solo la solubilidad de los compuestos, sino también su potencia. Estas series se han ampliado con estudios de pirido[5,4-*d*]pirimidinas, identificándose compuestos con una $IC_{50}$ del orden nanomolar frente a EGFR.

Por otra parte, durante el *screening* de quimiotecas, el equipo de Parke–Davis/Warner Lambert[341] identificó derivados pirido[2,3-*d*]pirimidínicos activos frente a PDGFR, FGFR y $pp60^{c\text{-}src}$. A partir del compuesto **15** (Figura 2.9), se realizaron distintos SARs[342-345] modificando los sustituyentes en las posiciones C-2, C-6, C-7, y N-8, y junto con la información extraída por rayos-X de estructuras de quinasas unidas a inhibidores, permitieron elucidar un modelo de unión para esta clase de compuestos.[346]




**15** **16** **17**

PD166866
FGFR ($IC_{50}$ = 60nM)

PD173074
optimización de **16**

**Figura 2.9.** Optimización de pirido[2,3-*d*]pirimidinas inhibidoras de FGFR.

En este modelo, se propone un motivo de puentes de hidrógeno de unión similar al de la olomoucina (véase abajo), según el cual los nitrógenos N-3 y el nitrógeno exocíclico del grupo 2-amino forman un puente de hidrógeno bidentado con dos aminoácidos de la región *hinge*. El grupo 6-fenilo interacciona en una cavidad próxima a este sitio.

De este modo, se han obtenido compuestos selectivos para FGFR variando los sustituyentes que penden del grupo 6-fenilo (**16**) (Figura 2.9). Finalmente, este compuesto se optimizó mediante la sustitución de la cadena de la amina, incrementando su solubilidad (**17**). Otros estudios recogen la optimización de las posiciones N-8 y C-6, lo que permitió la identificación de un inhibidor de PDGFR.[347]

Finalmente, del mismo modo que para las anilinoquinazolinas, las piridopirimidinas se han derivatizado con aceptores de Michael para obtener inhibidores irreversibles, así, se han preparado 6-acrilamido pirido[3,4-*d*]pirimidinas y 6-acrilamido pirido[3,2-*d*]pirimidinas.

**Pirrolopirimidinas y pirrolo[2,3-*b*]piridinas**

El *scaffold* pirrolopirimidina ha sido derivatizado por varias compañías farmacéuticas para encontrar inhibidores ATP-competitivos frente a EGFR y c-Src.

Análogamente al caso anterior, se ha desarrollado un modelo farmacofórico de interacción para la serie 7*H*-pirrolo-[2,3-*d*]pirimidinas (**18**) frente a EGFR[348-349]. Se postuló que el NH(7) del anillo de pirrol y el N(1) de la pirimidina forman un puente de hidrógeno bidentado con la Gln[767] y la Met[769], similar al que forma el ATP en el EGFR, donde el anillo *m*-clorofenil reemplaza a la ribosa en el bolsillo del azúcar. Para mejorar la potencia y farmacocinética, se realizaron modificaciones en las posiciones C-4 y C-6, introduciéndose sustituyentes que aumentaran el número de contactos de van der Waals con la región hidrofóbica formada por los residuos Thr[766] y Thr[860] (**19-22**) (Figura 2.10). También se han desarrollado series de pirrolo[3,2-*d*] y [2,3-*d*]pirimidinas como inhibidores de pp60$^{c\text{-}src}$.[335]

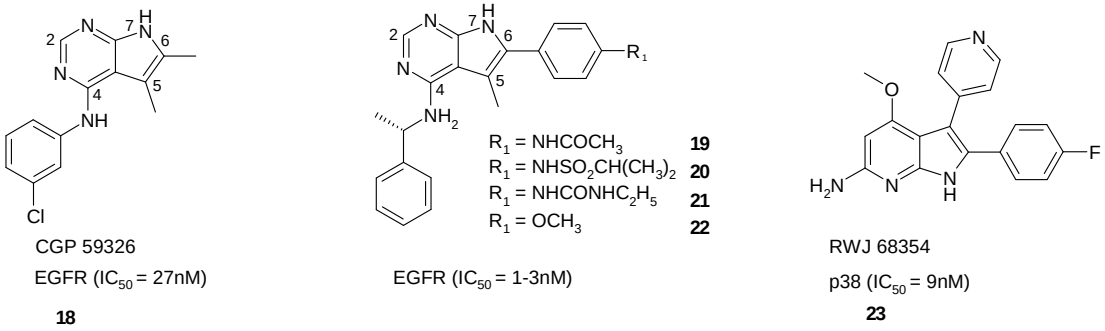



**Figura 2.10.** Series de pirrolo[2,3-*d*]pirimidinas testadas frente a EGFR.

Por otro lado, se han identificado pirrolo[2,3-*b*]piridinas[350], como el compuesto **23**, como inhibidores de la quinasa p38.

**Pirazolopirimidinas y pirazolopiridinas**

Pfizer identificó en 1996 los compuestos PP1 (**24**) y PP2 (**25**) (Figura 2.11), representativos de una serie de 4-aminopirazolo[3,4-*d*]pirimidinas, como inhibidores selectivos de las quinasas Lck y FynT.[351] Se estudiaron distintas sustituciones del anillo aromático en el nitrógeno y en la posición C-3 del anillo de pirazol.

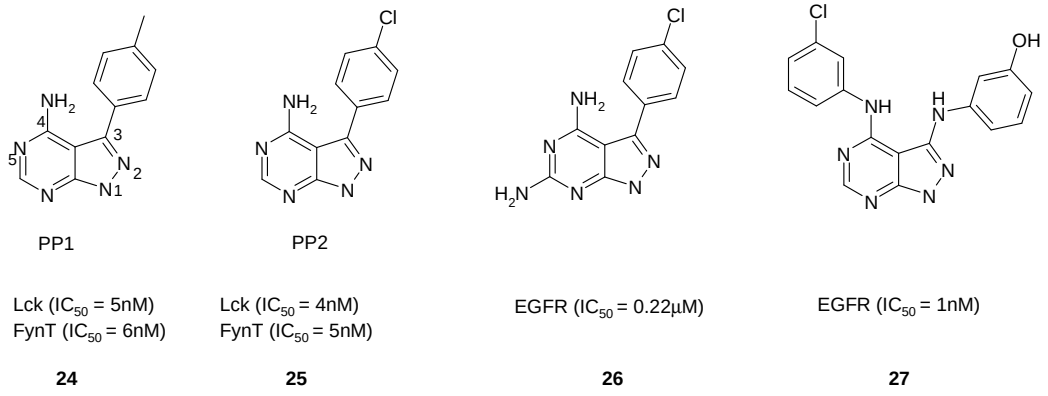


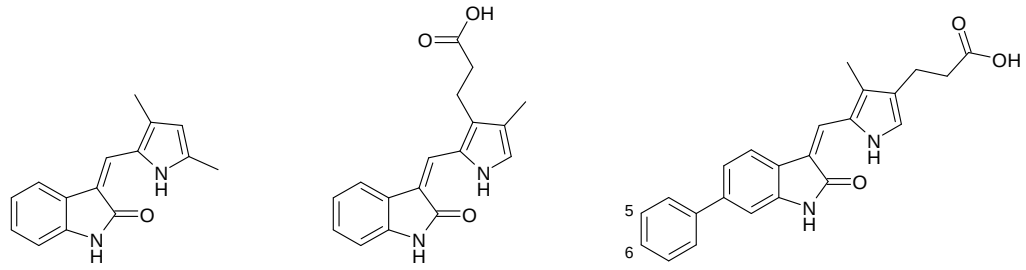

**Figura 2.11.** Series de pirazolo[3,4-*d*]pirimidinas testadas frente a Lck, FynT y EGFR.

Posteriormente, en un *screening* rutinario de quimiotecas se descubrió la potencia del compuesto **26** frente a EGFR. Sobre esta base y a partir del modelo farmacofórico establecido

para las pirrolo[2,3-*d*]pirimidinas[348], se optimizó una serie de 4-(fenilamino) pirazolo-[3,4-*d*]pirimidinas[352], diseñándose compuestos como **27**, con una mayor potencia, explicada a partir de interacciones adicionales por puente de hidrógeno del grupo hidroxilo del fenilo.

**Indolin-2-onas**

Sugen ha desarrollado varias series indolin-2-onas sustituidas en la posición 3, como inhibidoras de VEGFR, FGFR, EGFR, Her-2 y PDGFR en el orden submicromolar (**28** y **29**) (Figura 2.12). En función de los resultados SAR, se pueden establecer distintos criterios para obtener selectividad frente a las distintas dianas. También se han estudiado sustituciones en las posiciones 5- y 6- (**30**).[353-355]



SU5416 SU5402

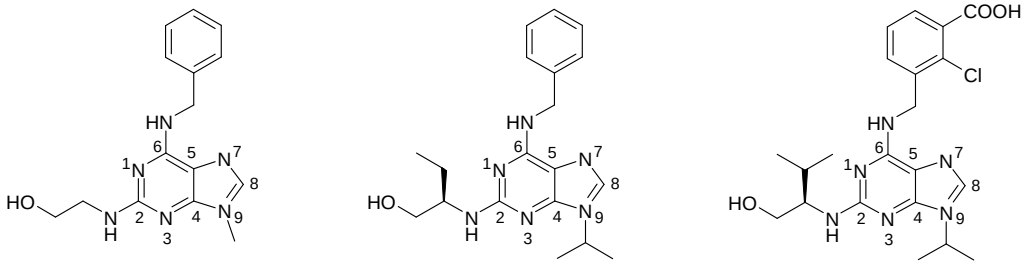
KDR / flk-1 ($IC_{50}$ = 1μM)

**28** **29** **30**

**Figura 2.12.** Ejemplos de indol-2-onas.

**Purinas**

Las purinas se han testado frente a un gran número de quinasas, especialmente las serina/treonina quinasas y, dentro de ellas, para las quinasas dependientes de ciclina (CDKs). A partir de las inicialmente descritas, olomoucina (**31**) y roscovitina (**32**), se han estudiado análogos por modificación de las posiciones 2-, 6- y 9-, conduciendo a compuestos con mayor potencia y selectividad dentro de esta familia, como el purvalanol B (**33**).



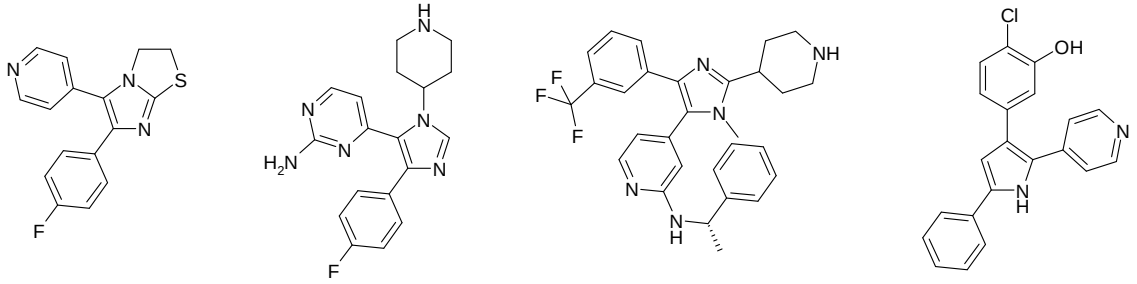
Olomoucina Roscovitina Purvalanol B

CDKs ($IC_{50}$ = 3-7 μM) CDKs ($IC_{50}$ = 0.2-0.7 μM) Cdk2-ciclinaA ($IC_{50}$ = 6 nM)

**31** **32** **33**

**Figura 2.13.** Ejemplos de análogos de purinas.

**Piridinilimidazoles, pirimidinilimidazoles y fenilbenzimidazoles**



SKF-86002 — **34**

SB-220025 — p38 ($IC_{50}$ = 19nM) — **35**

p38 ($IC_{50}$ = 0.19nM) — **36**

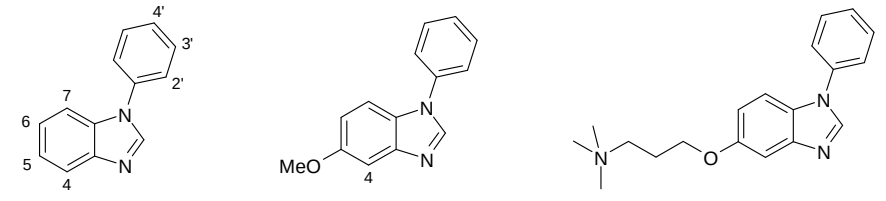
L-779450 — Raf ($IC_{50}$ = 2.0nM) — **37**

**Figura 2.14.** Ejemplos de análogos de piridinilimidazoles y pirimidinilimidazoles como inhibidores de la quinasa p38.

El compuesto SKF-86002 (**34**) (Figura 2.14), derivado piridinilimidazol, es el primer compuesto de esta serie identificado como inhibidor de la quinasa p38. Sin embargo, el grupo piridina genera efectos secundarios por su interacción con el citocromo P450, por lo que se sustituyó este anillo por pirimidinas, como el compuesto **35**.

Por otra parte, Merck, a partir de estudios SAR ha identificado los compuestos **36** y **37** como inhibidores potentes y selectivos de las quinasas p38 y Raf, respectivamente.[356]

En 1998, Palmer y colaboradores presentaron una serie de 1-fenilbenzimidazoles como inhibidores ATP-competitivos del PDGFR[357], con el compuesto **38** como cabeza de serie (Figura 2.15). A partir de un estudio SAR inicial, concluyeron que las modificaciones en 4'- y 3'- del anillo de fenilo, aunque toleradas, no mejoran significativamente la actividad, mientras que las sustituciones en 2'- abolen su actividad. Las sustituciones en las posiciones 2-, 4- y 7- del anillo de benzimidazol también eliminan la capacidad inhibitoria de esta serie. Sin embargo, las sustituciones en las posiciones 5- y 6- mantienen o incrementan la actividad, encontrando para el compuesto **39** una actividad máxima frente a PDGFR. Posteriormente[358], este estudio SAR se amplió para la posición 5-, encontrándose que la sustitución por grupos catiónicos solubilizantes aumenta la potencia de esta serie (**40**).



PDGFR ($IC_{50}$ = 9.3 μM) — **38**

PDGFR ($IC_{50}$ = 0.43 μM) — **39**

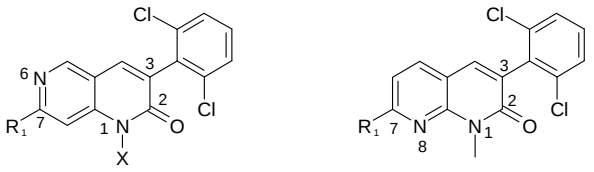
PDGFR ($IC_{50}$ = 0.15 μM) — **40**

**Figura 2.15.** Ejemplos de análogos de 1-fenilbenzimidazoles inhibidores de PDGFR.

**Naftiridin-2(1*H*)-onas**

Thompson y colaboradores[359], presentan en 2000, estudios SAR sobre los *scaffolds* 3-(2,6-diclorofenil) 1,6-naftiridin-2(1*H*)-onas (**41**, **42**) (Figura 2.16) y 3-(2,6-diclorofenil)

1,8-naftiridin-2(1*H*)-onas (**43**), con actividad submicromolar frente a pp60$^{c\text{-}src}$, FGFR y PDGFR (en menor grado).
La serie 1,6-naftiridona presentan actividades semejantes a la serie pirido[2,3-*d*]pirimidina, mientras que la serie 1,8-naftiridona es mucho menos activa. Como en otros *scaffolds*, la sustitución con cadenas laterales básicas en la posición C-7, incrementa la potencia.



X = Me **41**
X = H **42**

**43**

**Figura 2.16.** Series de 1,6-naftiridin-2(1*H*)-onas y 1,8-naftiridin-2(1*H*)-onas testadas como inhibidores de pp60$^{c\text{-}src}$ y FGFR.

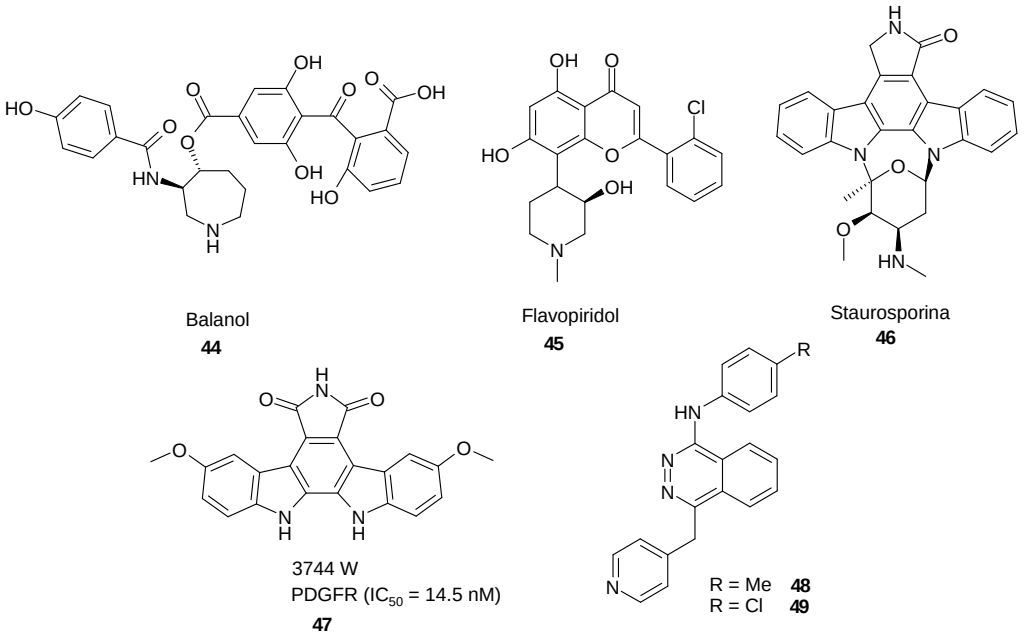
## Otras clases estructurales

El balanol (**44**) (Figura 2.17) es un producto natural aislado del hongo *verticullium balanoides*, inhibidor específico de serina/treonina quinasas, que apenas muestra actividad frente a tirosina quinasas. Se han realizado diversas modificaciones sobre su estructura para incrementar su actividad en ensayos celulares.

El flavopiridol (**45**) es un flavonoide inhibidor de muchas quinasas dependientes de ciclina.

La staurosporina (**46**) es un alcaloide microbiano, inhibidor muy potente aunque no específico de proteína quinasas. De hecho, se ha tomado como estructura de partida para preparar derivados con mejores perfiles de selectividad, como el 3744W (**47**), que inhibe la autofosforilación de PDGFR.

Finalmente, Novartis ha descrito también derivados de ftalazinas como inhibidores de los receptores de VEGF y PDGF. Los compuestos **48** y **49** muestran actividad submicromolar frente a estos enzimas.



Balanol
**44**

Flavopiridol
**45**

Staurosporina
**46**

3744 W
PDGFR ($IC_{50}$ = 14.5 nM)
**47**

R = Me **48**
R = Cl **49**

**Figura 2.17.** Series de inhibidores derivados de productos naturales.

Recientemente, se ha profundizado en los mecanismos que puedan explicar la elevada inespecificidad de determinados inhibidores con actividad micromolar frente a varias tirosina quinasas. Compuestos como el índigo, indirrubina y bisindolilmaleimida inhiben a enzimas varios, aparte de quinasas, a través de la formación de agregados.[360]

Tal y como se ha ido comentando, el uso de la información estructural obtenida por difracción de rayos-X, junto con la modelización por homología de dominios tirosina quinasa ha permitido el diseño de inhibidores de quinasas. En este sentido, destaca también la contribución de los modelos farmacofóricos y las búsquedas de similitud a compuestos activos ya conocidos.