Libros de Cátedra

Introducción a la Química Medicinal

Luciana Gavernet (coordinadora)

FACULTAD DE
CIENCIAS EXACTAS

exactas

edulp
EDITORIAL DE LA UNLP

UNIVERSIDAD
NACIONAL
DE LA PLATA

INTRODUCCIÓN A LA QUÍMICA MEDICINAL

Luciana Gavernet

(coordinadora)

Facultad de Ciencias Exactas

UNIVERSIDAD NACIONAL DE LA PLATA

edulp
EDITORIAL DE LA UNLP

Al Prof. Dr. Luis Bruno Blanch.

Agradecimientos

Los autores de este libro queremos agradecer al Profesor Extraordinario Dr. Luis Bruno-Blanch por su contribución a la formación en Química Medicinal de los alumnos de la carrera de Farmacia de la Facultad de Ciencias Exactas (UNLP) durante ininterrumpidos 40 años dedicados a la docencia universitaria. Los continuos esfuerzos del Dr. Bruno-Blanch constituyen un aporte fundamental para el desarrollo de las Ciencias Químicas y Farmacéuticas tanto a nivel científico, como a nivel de formación de docentes-investigadores. Su devoción a la formación de recursos humanos se ha traducido, recientemente, en la creación del Laboratorio de Investigación y Desarrollo de Bioactivos (LIDeB) de la UNLP.

Índice

CAPÍTULO 5
Métodos indirectos. Búsqueda racional de fármacos

Lucas N. Alberca y Alan Talevi

1. Cribado virtual

Las etapas tempranas del descubrimiento de fármacos implican la identificación de nuevos *hits* contra un blanco molecular específico y las sucesivas optimizaciones para mejorar las propiedades farmacológicas de dichos compuestos. Como ya se ha mencionado en el Capítulo 2, la industria farmacéutica recurre principalmente al Cribado Farmacológico de Alto Rendimiento (o *High Throughput Screening* - HTS*)* como estrategia para identificar en tiempos breves nuevos andamios (*scaffolds*) moleculares con la actividad deseada. Sin embargo, las tecnologías necesarias para la implementación de campañas de HTS requieren una inversión muy grande que solo las compañías farmacéuticas y algunos centros académicos de países desarrollados pueden afrontar. Al costo del equipamiento en sí debe sumarse, además, los altísimos costos operativos que deben afrontarse para su utilización y mantenimiento.

Una metodología alternativa, también mencionada previamente como más racional y económica, es el Cribado virtual (en adelante, CV), también conocido como *Screening Virtual* o *Screening in silico* o Tamizado Virtual. Algunos autores describen al CV como el "uso de la informática de alto rendimiento para analizar grandes bases de datos de compuestos químicos con el fin de identificar posibles candidatos a drogas*".* Otros autores amplían esta definición considerando al CV como un "conjunto de técnicas computacionales que permiten, a partir de representaciones de la estructura molecular de los compuestos químicos almacenados en grandes bases de datos, identificar compuestos potencialmente interesantes desde el punto de vista farmacológico*".* De manera más simple, podemos definir al CV como la utilización de modelos o algoritmos computacionales para predecir cuáles compuestos de una biblioteca de compuestos químicos podrían tener una cierta actividad deseada.

A continuación, se enumeran algunas de las características que hacen a la racionalidad de esta estrategia:

a) **Eficiencia en cuanto al costo y tiempo necesarios para su realización:** Actualmente existen una gran cantidad de bases de datos de compuestos químicos que se pueden descargar y manipular de manera gratuita y que son actualizadas/expandidas de manera periódica o continua. Entre ellas pueden mencionarse ChEMBL, que a la fecha compila más de 1,8 millones de

compuestos químicos y más de 15 millones de datos de actividad biológica; PubChem, con más de 96 millones de compuestos y más de 230 millones de datos de bioactividad y; DrugBank que contiene alrededor de 10000 compuestos utilizados terapéuticamente o que se encuentran actualmente cursando estudios clínicos, de los cuales cerca de 9000 son pequeñas moléculas tipo fármaco. Además, muchos de los programas que se utilizan para aplicar las técnicas de CV son de libre acceso o se pueden conseguir licencias académicas gratuitas como el programa Chimera y los diferentes paquetes de ChemAxon. También existen servidores online en los cuales se pueden realizar CV sin necesidad de descargar programas. Más aún, dependiendo de la técnica de CV a utilizar, es posible realizar el cribado con una simple computadora de escritorio. En este sentido, el CV supera ampliamente al HTS, que en cualquiera de los casos requiere tecnología de alto costo. Finalmente, puede mencionarse que, en tanto seleccionados mediante técnicas con fundamento teórico, los *hits* que emergen del CV tienen mayores probabilidades de demostrar actividad biológica contra el blanco de interés que un compuesto químico que hubiera sido elegido al azar, sin criterio racional.

b) **Carácter teórico:** No es necesario disponer de una muestra física de los compuestos químicos a analizar ya que el proceso se realiza en base a representaciones digitales de las moléculas. Además, se pueden considerar colecciones de compuestos que aún no han sido sintetizados de manera de sintetizar o adquirir sólo aquellos que presenten altas probabilidades de poseer la actividad de interés.

c) **Carácter bioético:** Las metodologías *in silico* reducen la cantidad de compuestos a evaluar tanto *in vitro* como *in vivo* (ciñéndose a los conceptos de reemplazo/reducción del uso de modelos animales) y aumenta las probabilidades de resultados positivos en esas instancias. El CV cumple con los principios internacionales para la investigación biomédica que involucra animales (*International Guiding Principles for Biomedical Research involving Animals*, CIOMS), que propone anteponer simulaciones computacionales y modelos *in vitro* a los estudios en animales de laboratorio siempre que sea posible.

En general, las estrategias para la búsqueda y diseño racional de fármacos asistida por computadora son clasificadas en dos categorías: Métodos indirectos (también llamados, métodos basados en el ligando), en los cuales no es necesario conocer la estructura tridimensional del blanco molecular y métodos directos (o métodos basados en la estructura), en los cuales es indispensable poseer un modelo tridimensional de la macromolécula. Subrayamos aquí, además, que cuando pensamos en diseño de fármacos siempre estamos considerando la eventual obtención de un compuesto químico novedoso que no hubiera sido sintetizado hasta el momento. En cambio, el CV, mayormente se enfoca en colecciones de compuestos que ya se conocen, que fueron previamente sintetizados o aislados a partir de fuentes naturales.

El presente capítulo se centra en los métodos indirectos para la búsqueda racional de fármacos.

2. Métodos indirectos

Existe una gran cantidad de blancos moleculares de los cuales no se encuentra disponible su estructura tridimensional y no es posible realizar un modelo tridimensional confiable mediante técnicas de modelado comparativo. En esos casos, la información estructural de uno o más ligandos activos en el blanco molecular seleccionado puede utilizarse para identificar cuáles son las características estructurales responsables de la actividad biológica de los mismos.

La hipótesis implícita de los métodos indirectos es que las moléculas similares exhibirán propiedades de unión similares con respecto a un dado blanco molecular, en tanto la actividad biológica dependerá de la estructura química del mismo.

Los métodos indirectos de CV pueden utilizar características obtenidas de distintos niveles de representación de las estructuras moleculares de los ligandos, habitualmente representaciones bidimensionales (2D) y/o tridimensionales (3D). En los siguientes puntos de este capítulo se describirán tres grandes grupos de métodos indirectos: los basados en la similitud molecular, los basados en el farmacóforo y, las metodologías basadas en descriptores moleculares (QSAR).

2.1 Métodos basados en la similitud molecular

La similitud molecular se enfoca en ciertas características estructurales de los compuestos tales como presencia, ausencia, o frecuencia de determinadas subestructuras químicas o determinados grupos funcionales, para determinar qué tan similares son dos moléculas entre sí. Para realizar esta comparación se requieren tres componentes básicos:

1- Una **representación de la molécula** cuyos componentes codifiquen las características químicas y/o moleculares relevantes de la misma.

2- Un **sistema de ponderación** (asignación cuantitativa de peso/importancia) de tales características. Es decir, se requiere de un esquema que permita determinar cuáles características tienen la misma importancia y cuales características son más relevantes.

3- Un **coeficiente de similitud** que transforme la información contenida en la representación estructural de la molécula a un valor numérico. En general, este número valor se encuentra entre 0 y 1, donde 0 indica que ninguna de las características estructurales comparadas está presente en ambas moléculas y 1 significa que por el método seleccionado se verifica la identidad completa de la molécula. Esto último no necesariamente implica que dos compuestos comparados sean efectivamente idénticos. Hay métodos de cuantificación de similitud que son “ciegos” a ciertas diferencias entre dos moléculas: por ejemplo, un método que caracterice los grupos funcionales como dadores o aceptores de enlaces de hidrógeno será incapaz de distinguir entre una amina primaria y un hidroxilo.

Una forma frecuente de comparar cuantitativamente dos moléculas se basa en la comparación de secuencias de bits (denominadas *molecular fingerprints* o huellas digitales moleculares)

de las estructuras a comparar. Típicamente, las huellas digitales moleculares se construyen asociando cada bit a la presencia o ausencia de una propiedad estructural, las cuales pueden contener información bidimensional o tridimensional.

En la Figura 5.1 se presenta un proceso simple de comparación de tres moléculas utilizando como coeficiente de similitud el coeficiente de Tanimoto en su forma binaria (no considera el número de veces que aparece una subestructura o característica molecular en una dada molécula, tan solo su ocurrencia o no ocurrencia). Este coeficiente es uno de los más utilizados para cuantificar similitud y se define por la Ecuación 5.1:

$$S = \frac{c}{a + b - c}$$

Ecuación 5.1: Índice de Tanimoto.

Donde a y b representan el número total de subestructuras (o propiedades estructurales) presentes en cada una de las moléculas comparadas y, c indica el número de subestructuras comunes entre el par de moléculas a comparar. En el ejemplo de la **Figura 5.1** solo se consideraron ocho subestructuras para construir las huellas digitales moleculares de las moléculas A, B y C. La molécula A presenta seis de las ocho subestructuras consideradas, mientras que la molécula B presenta siete de estas subestructuras y la molécula C presenta sólo cinco. A partir de la comparación de a pares de las huellas digitales obtenidas, pueden calcularse los correspondientes coeficientes de Tanimoto. En el caso del ejemplo, los coeficientes asumen los valores S_{AB} = 0,625, S_{AC} = 0,571 y S_{BC} = 0,500, entonces, las estructuras A y B son las más parecidas de acuerdo con este método.

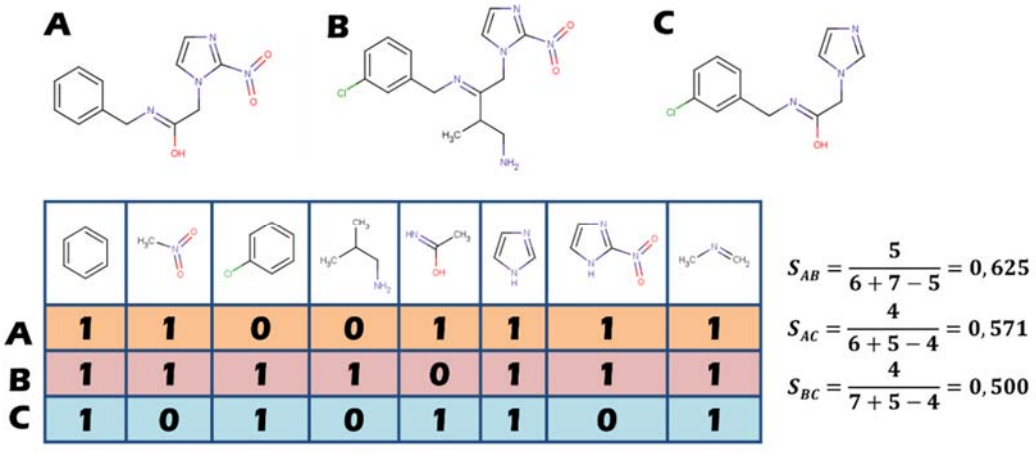


Figura 5.1: Esquema del proceso de comparación de tres moléculas por medio del coeficiente de Tanimoto.

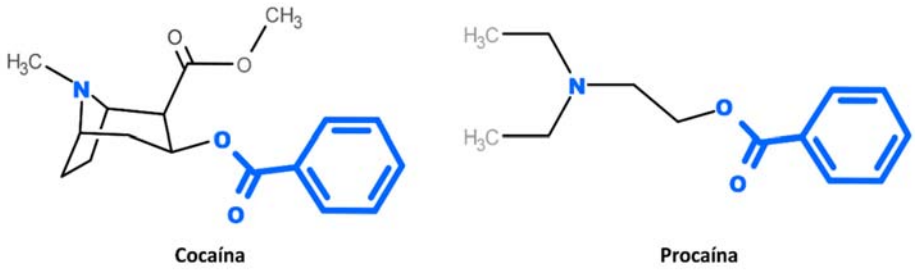
Los métodos de similitud molecular son una buena estrategia para la realización de CV cuando el número de ligandos activos conocidos de un blanco molecular es limitado, o en el peor de los casos,

cuando se conoce un único ligando. Las técnicas de CV basado en la similitud consisten en cuantificar el grado de similitud entre la estructura que presenta la actividad deseada y las estructuras de los compuestos que integran la base sometida al CV. Una ventaja interesante de este tipo de técnicas es el bajo costo computacional que requieren, permitiendo realizar el cribado de grandes bases de datos en poco tiempo. Hoy en día, el uso de estas técnicas es casi trivial en tanto la mayoría de los repositorios químicos online utilizados para cribado incorporan funciones de búsqueda por similitud molecular. Por tal motivo, es raro actualmente el uso de estas técnicas de manera aislada, y se suele asociar a otros métodos indirectos o directos siempre que sea posible.

2.2 Métodos basados en el farmacóforo

El termino farmacóforo se refiere a un arreglo tridimensional de la molécula con las características moleculares mínimas para asegurar las interacciones con un blanco molecular especifico, logrando de esa forma que se produzca la respuesta biológica. Complementariamente, la IUPAC define a un farmacóforo como un conjunto de propiedades estéricas y electrónicas para asegurar las interacciones supramoleculares óptimas entre la molécula y un blanco biológico especifico activando o bloqueando su respuesta. En la Figura 5.2 se representa el farmacóforo que le da propiedades anestésicas a la cocaína, el cual fue utilizado para el posterior desarrollo del anestésico local procaína.

Figura 5.2:



Ejemplo de farmacóforo (en azul) entre los anestésicos cocaína y procaína.

En este capítulo se presenta a los modelos farmacofóricos como aproximaciones basadas en el ligando, sin embargo, se debe tener en cuenta que este tipo de modelos cuentan con la versatilidad de poder ser abordados desde ambas perspectivas (receptor y ligando). La estrategia basada en el ligando consiste en superponer las estructuras tridimensionales de un conjunto de moléculas que comparten un blanco molecular y un sitio de unión (denominado conjunto de entrenamiento), con el objetivo de extraer aquellas propiedades químicas comunes que se presentan como esenciales para manifestar dicha actividad, en tanto que la perspectiva basada en el receptor propone postular los puntos de interacción necesarios en los ligandos y su disposición espacial en base a la complementariedad con las características conocidas del sitio activo. Un punto interesante para destacar es que

las características consideradas en un modelo farmacofórico no están exentas de un cierto grado de abstracción. Esencialmente, se consideran grupos aromáticos, lipofílicos, aceptores y dadores de hidrógeno, grupos ionizados positiva y negativamente. De este modo, una gran diversidad de funciones/grupos químicos puede cumplir con un dado requisito farmacofórico. Por ejemplo, un grupo amino primario o secundario o un grupo alcohólico pueden funcionar perfectamente como dadores de hidrógeno. Un sustituyente piridilo o un fenilo pueden, indistintamente, cumplir con el requisito de un anillo aromático en determinada posición espacial.

El modelado farmacofórico basado en el ligando es una opción computacional muy interesante (y, por cierto, gráfica y fácil de interpretar) cuando no se dispone de la estructura del blanco molecular. La generación de las hipótesis farmacofóricas basadas en múltiples ligandos comprende dos etapas principales:

1- Barrer el espacio conformacional para los ligandos del conjunto de entrenamiento de manera de representar la flexibilidad de cada uno de ellos.

2- Alinear las moléculas en aquella conformación que resulte en la superposición geométrica de la mayor cantidad de propiedades importantes para la actividad biológica.

La flexibilización conformacional de los ligandos y el alineamiento molecular son al mismo tiempo la base de la técnica y su principal dificultad. La etapa de alineamiento resulta ser la más costosa computacionalmente y la que representa un mayor desafío a la hora de diseñar algoritmos. Algunos métodos de alineamiento incorporan algoritmos donde el tiempo de cómputo crece exponencialmente con el número de propiedades analizadas, lo que limita la posibilidad de utilizar las hipótesis sobre grandes bases de datos y acota su campo de aplicación, con el equipamiento computacional actual, a moléculas pequeñas o a pequeñas bibliotecas. La etapa de búsqueda del modelo se facilita enormemente si se dispone de un análogo rígido activo, con escasa o nula libertad conformacional, que por lo tanto define la conformación activa o provee un fuerte indicio sobre la misma.

Las propiedades moleculares compartidas entre las moléculas del conjunto de entrenamiento que confieran la actividad (es decir, que estén implicadas en interacciones fundamentales con el blanco molecular) constituirán el farmacóforo. Si se desea utilizar este modelo para una campaña de CV, se busca en bibliotecas de compuestos químicos cuáles de ellos cumplen los requisitos establecidos. La desventaja de utilizar esta técnica para el CV es que todas las moléculas de la biblioteca a cribar requieren ser optimizadas conformacionalmente y superpuestas con el farmacóforo, un proceso que dependiendo de la cantidad de moléculas de la biblioteca puede ser muy costoso en términos de tiempo de cálculo.

Adicionalmente, los modelos farmacofóricos, a diferencia de otros métodos indirectos, tienen la ventaja de que pueden ser utilizados tanto para la búsqueda como para el diseño de nuevos fármacos. Una vez identificado el farmacóforo, es posible proponer modificaciones en un compuesto activo que permitan mejorar la actividad de los compuestos sobre el blanco molecular escogido.

2.3 Métodos basados en descriptores moleculares

Como se describió en el Capítulo 4, un descriptor molecular puede describirse como una variable numérica que representa algún aspecto o característica de la estructura molecular como el tamaño, el volumen, la cantidad de enlaces de hidrógeno disponibles, la lipofilicidad, la forma, la distribución electrónica, etc. En este punto se describirá cómo es posible encontrar relaciones entre una cierta actividad o propiedad de las moléculas y su estructura molecular codificada en descriptores moleculares.

2.3.1 Relaciones Cuantitativas Estructura-Actividad (QSAR)

Existen diferentes métodos estadísticos que permiten encontrar relaciones entre una variable dependiente (la propiedad o actividad que se desea modelar) y uno o más descriptores moleculares. Estas relaciones son conocidas como Relaciones Cuantitativas Estructura-Actividad o **QSAR** (del inglés, *Quantitative Structure-Activity Relationships*) y pueden ser descriptas por la siguiente Ecuación 5.2:

$$\boldsymbol{A = f(d_1, d_2, d_3, \dots, d_d)}$$

Ecuación 5.2: Forma general de las ecuaciones QSAR.

Donde $\boldsymbol{A}$ es la actividad biológica (o una propiedad cualquiera) de un compuesto químico, que es considerada como una función matemática de ciertas características estructurales cuantificadas mediante descriptores moleculares (d_1, d_2). $\boldsymbol{A}$ puede pensarse como la variable dependiente, la "respuesta" (*response*) o la "salida" (*output*) del modelo.

La relación $\boldsymbol{f}$ puede ser obtenida por diferentes métodos estadísticos, métodos evolutivos, etc. Una vez obtenida esta relación, la actividad $\boldsymbol{A}$ de un compuesto nuevo o no testeado frente al blanco molecular seleccionado puede ser predicha a partir de su estructura molecular mediante el cálculo del valor que asumen los descriptores moleculares ($\boldsymbol{d_1}$, $\boldsymbol{d_2}$, etc) (incluidos en el modelo QSAR) para ese compuesto.

2.3.2 Clasificación de los modelos QSAR

Existen diferentes formas de clasificación de los modelos QSAR:

a) Si se tiene en cuenta la ***naturaleza de la variable dependiente*** considerada, los modelos QSAR pueden clasificarse en:

Modelos QSAR cuantitativos: estos darán como resultado (o respuesta) una variable dependiente contínua, es decir, predecirán el valor numérico de la actividad (o propiedad) modelada. Por ejemplo, este tipo de modelos podría dar como resultado el valor de la constante de inhibición (Ki) de

un compuesto frente a un blanco molecular, la concentración del compuesto que genera un 50% de inhibición del blanco molecular (IC_{50}), el valor del logP de un compuesto, etc.

Modelos QSAR cualitativos (modelos clasificadores): darán como resultado una respuesta cuyo valor estará asociado a una categoría previamente establecida. Por ejemplo, las categorías pueden ser "inhibidores" y "no inhibidores" de un cierto blanco molecular, "tóxicos" y "no tóxicos" para células humanas, etc. Esta estrategia es muy útil cuando se entrenan los modelos con datos de diferentes fuentes (datos experimentales medidos en distintos laboratorios, por ejemplo) ya que un modelo clasificador permite amortiguar, al menos parcialmente, la variabilidad experimental presente en la base de datos.

b) De acuerdo a la **dimensionalidad de los descriptores moleculares** utilizados en los modelos, estos se pueden clasificar en:

Modelos QSAR de baja dimensionalidad: son aquellos modelos que utilizan descriptores moleculares independientes de la conformación. Por lo tanto, estos modelos se obtienen a partir de descriptores moleculares 0D, 1D y 2D. Los modelos constituidos enteramente por descriptores independientes de la conformación tienen la ventaja de que no requieren la optimización conformacional de las moléculas, por lo tanto, son poco demandantes computacionalmente. En los siguientes puntos de este capítulo se detallarán algunas estrategias útiles para la obtención de modelos de este tipo.

Modelos QSAR de alta dimensionalidad: son aquellos modelos que utilizan descriptores dependientes de la conformación de las moléculas. En esta categoría se agrupan los descriptores 3D y 4D. Las metodologías QSAR de alta dimensionalidad son más complejas y costosas computacionalmente respecto a las metodologías QSAR de baja dimensionalidad ya que requieren de la optimización conformacional 3D de cada una de las moléculas del conjunto de entrenamiento. La obtención de las conformaciones 3D de mínima energía de estas moléculas es el factor más importante para obtener modelos fiables ya que a partir de sus estructuras tridimensionales se calcularán los descriptores moleculares. La aplicación de modelos QSAR de alta dimensionalidad en CV requiere un tiempo considerablemente mayor que cuando se utilizan modelos QSAR de baja dimensionalidad ya que las bibliotecas de moléculas para el CV deben ser optimizadas conformacionalmente antes de la aplicación de los modelos. Los métodos QSAR-3D más populares, entre los que se destacan CoMFA, CoMSIA y GRID/GOLPE, se basan en colocar cada molécula optimizada en una grilla y calcular el valor que diferentes campos de fuerza (por ejemplo, coulómbico, estérico) ejercen sobre los vértices de la misma (colocando generalmente un átomo o molécula de prueba en cada vértice), estableciendo qué campo, y en qué zona de la grilla, es significativo para la actividad biológica.

c) Finalmente, si se tiene en cuenta en cuenta la manera en que se seleccionan las variables independientes y la forma en que se correlacionan con la actividad de interés, los modelos pueden ser de dos tipos:

Modelos QSAR Lineales: Se basan en la suposición de la existencia de una relación lineal entre las variables independientes y la variable dependiente o respuesta. Pueden ser cuantitativos, como cuando se generan mediante Regresión Lineal Múltiple (RLM), o cualitativos, cuando se obtienen mediante técnicas como el Análisis Lineal Discriminante (ALD).

Modelos NO Lineales: En estos tipos de modelos, los descriptores se asignan principalmente a un espacio relacional no lineal y ayudan a superar algunas limitaciones de los métodos lineales. Entre los métodos más utilizados para la construcción de modelos no lineales se encuentran los Arboles de Decisión, la Redes Neuronales Artificiales y las Maquinas de Soporte Vectorial para regresión (en inglés, *Support Vector Machines*).

2.3.2.1 Etapas involucradas en la construcción de modelos QSAR 2D clasificatorios (cualitativos) basados en descriptores[2]La construcción de ecuaciones QSAR, que relacionan la estructura de los compuestos químicos con una determinada propiedad biológica, conlleva un protocolo específico que implica varias etapas como puede apreciarse en la Figura 5.3.

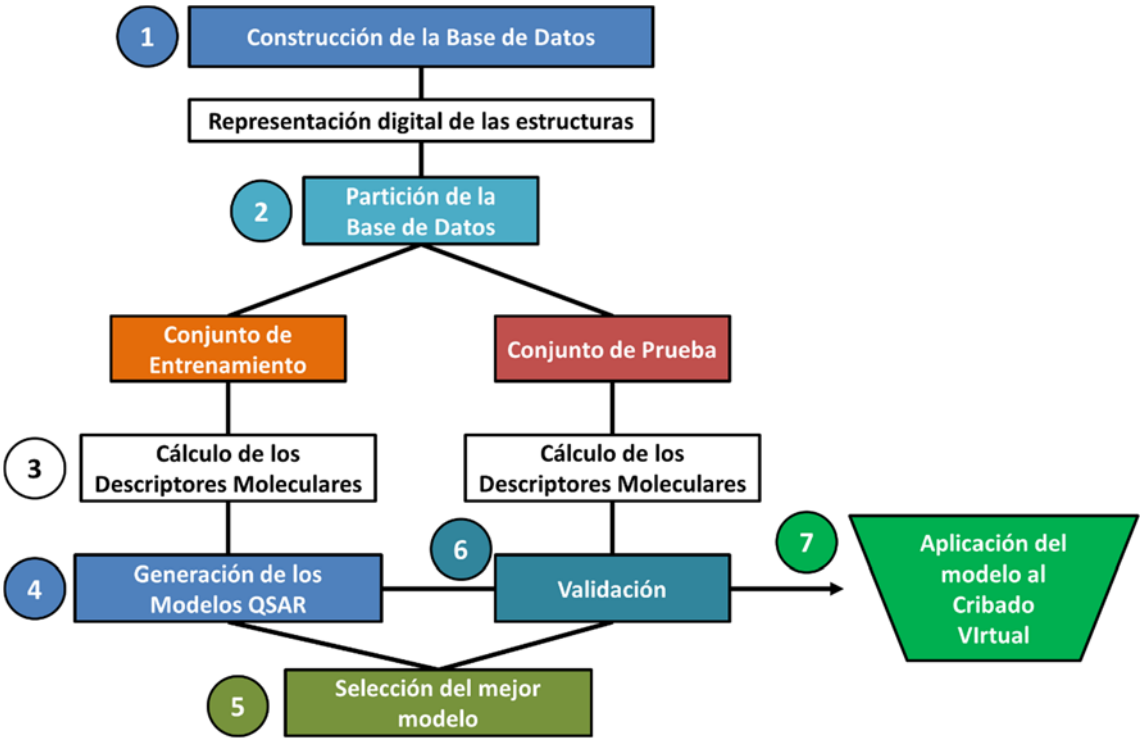


Figura 5.3: Esquema de las etapas vinculadas al desarrollo de un Modelo QSAR y su aplicación al CV.

[2] Las etapas generales para los modelos 3D son similares a las que se detallaran en este apartado, sólo que en todos los casos antes del cálculo se descriptores se agrega una instancia de búsqueda conformacional, y en algunos casos (por ej. CoMFA) una de alineamiento.

1) Compilado o construcción de la base de datos, set de datos o *dataset:*.

El primer paso para la obtención de modelos QSAR consiste en la búsqueda y recolección de compuestos químicos con sus correspondientes datos experimentales de la actividad o propiedad modelada. Con los datos obtenidos de esta búsqueda se genera un set de datos que posteriormente debe ser adecuadamente curado. Por ejemplo, si nuestro objetivo es generar un modelo que permita diferenciar entre inhibidores y no inhibidores de una enzima, entonces buscaremos en bibliografía y/o bases de datos como ChEMBL qué compuestos químicos han sido ensayados sobre esa enzima, cómo han sido evaluados y qué resultados se obtuvieron. Para generar modelos con buena capacidad de predicción es esencial realizar un curado de la base de datos que hemos obtenido; esto implica eliminar los compuestos duplicados, verificar la compatibilidad de los datos que provengan de diferentes tipos de ensayos de actividad, eliminar los compuestos con datos de actividad faltantes, etc.

El curado de la base de datos compilada es un paso crítico y esencial para la obtención de modelos QSAR fiables.

2) Partición de la base de datos:

Típicamente, para la construcción de modelos QSAR, la base de datos obtenida en el punto anterior es dividida en dos conjuntos:

Conjunto de entrenamiento: Conjunto de moléculas con las cuales se generarán/inferirán los modelos.

Conjunto de prueba: Conjunto de moléculas independiente del conjunto de entrenamiento con las que se validarán externamente los modelos (se estimará su capacidad de generalización o capacidad predictiva). Los compuestos del conjunto de prueba son “desconocidos” para los modelos ya que ellos fueron excluidos del proceso de desarrollo de los mismos.

Para realizar la partición de la base de datos en estos dos conjuntos se pueden utilizar diferentes estrategias. Uno de los métodos más utilizados es el de la **asignación al azar** de los compuestos del set de datos a cada uno de los conjuntos; sin embargo, con este método no se puede garantizar que los conjuntos de entrenamiento y de prueba sean representativos del conjunto de datos. Esto podría ocasionar que los compuestos del conjunto de prueba no sirvan para validar los modelos generados con el conjunto de entrenamiento, o viceversa, que los compuestos del conjunto de entrenamiento no produzcan modelos capaces de predecir acertadamente la variable dependiente para los compuestos del conjunto de prueba. Se ha demostrado que este tipo de partición es una aproximación adecuada cuando se seleccionan conjuntos de entrenamiento y de prueba de tamaño similar; en cambio, cuando los conjuntos de prueba son relativamente pequeños, se requieren aproximaciones de muestreo más racionales para proveer mejores resultados. Las mismas utilizan algoritmos de partición racional que intentan dividir la base de datos de una forma más representativa. Por ejemplo, existen numerosos métodos de agrupamiento (en inglés*, clustering*) racional basados en huellas digitales (*fingerprints*) y en descriptores

moleculares que pueden ser divididos en dos clases: los métodos jerárquicos como el agrupamiento de Ward y los métodos no jerárquicos como el de Jarvis-Patrick y k-means. La explicación de estos métodos excede el contenido de este libro, pero la idea fundamental es agrupar subconjuntos de compuestos con características moleculares comunes para luego muestrear representativamente de cada subconjunto una fracción de los mismos para el conjunto de entrenamiento y la fracción remanente para el conjunto de prueba.

Características del conjunto de entrenamiento: El conjunto de entrenamiento es una muestra de compuestos químicos de la cual se infiere una relación entre la estructura molecular y la actividad o propiedad de interés (el modelo QSAR). Se espera que esta relación sea generalizable a la "población" de compuestos químicos de la cual el conjunto de entrenamiento es representativo. Por lo tanto, la selección de los compuestos del conjunto de entrenamiento es crítica ya que define la región del espacio químico dentro de la cual la aplicación del modelo tendrá validez. El valor predicho por el modelo de la actividad/propiedad biológica modelada será confiable si y sólo si el compuesto predicho posee cierta similitud con alguna o varias de las estructuras incluidas en el conjunto de entrenamiento. El espacio químico dentro del cual la aplicación del modelo QSAR es confiable se denomina "dominio de aplicabilidad" o "dominio de aplicación" del modelo. Entonces, si se desea aplicar el modelo generado en campañas de CV sobre bibliotecas químicas que comprenden compuestos de gran heterogeneidad estructural, es conveniente que el conjunto de datos generado sea lo más abarcativo posible.

Otra característica que debe reunir el conjunto de entrenamiento es que los valores observados de la propiedad estudiada deberían presentar una buena distribución, idealmente uniforme en un rango de entre tres y cuatro órdenes logarítmicos de la propiedad modelada. El requisito de distribución uniforme es difícil de cumplir y se acepta como válida una distribución aproximadamente normal. La necesidad de una distribución uniforme tiene un fundamento lógico: por ejemplo, si tenemos un conjunto de entrenamiento formado mayormente por compuestos que poseen alta y mediana actividad, el modelo QSAR generado sólo va a predecir con exactitud valores de compuestos muy activos y moderadamente activos, o incluso estará sesgado hacia la asignación de altos valores de actividad, pero no nos permitirá diferenciar compuestos activos de inactivos. En otras palabras, el modelo no identificará qué características estructurales son desfavorables a la actividad por no estar los compuestos poco activos e inactivos representados dentro del conjunto de entrenamiento.

Un último requisito que se le pide tradicionalmente a los compuestos del conjunto de entrenamiento es que posean un mecanismo de acción común, esto es, interaccionen con el mismo receptor o blanco molecular y presenten el mismo sitio de unión. No obstante, se han modelado exitosamente en muchas oportunidades respuestas fenotípicas "ciegas" al mecanismo de acción considerado.

3) Cálculo de descriptores moleculares:

Una vez generados ambos conjuntos de trabajo, se procede al cálculo del valor que asumirán los descriptores moleculares para las moléculas del conjunto de entrenamiento. Como ya

se mencionó, dependiendo del tipo de los descriptores que se calculen podremos obtener modelos QSAR dependientes o independientes de la conformación. Sin embargo, hay que tener en cuenta que para el cálculo de los descriptores dependientes de la conformación es necesaria la optimización geométrica previa de las moléculas. Entre los softwares disponibles para el cálculo de descriptores moleculares podemos nombrar el Dragon, RDKIT, PaDEL y WEKA, entre muchos otros.

4) Generación del Modelo QSAR:

La generación de modelos QSAR implica dos etapas

a) La selección de descriptores adecuados para la predicción de la actividad o propiedad analizada.

b) La obtención del modelo matemático óptimo que correlacione la actividad o propiedad específica con los descriptores seleccionados.

Como ya se mencionó, los métodos de **selección de variables** (en este caso, descriptores moleculares) se presentan en dos grupos, los métodos lineales y los enfoques no lineales. En la Tabla 5.1 se enumeran algunos de los métodos más usados para la selección de variables tanto lineales como no lineales. En todos los métodos las variables se introducen en el modelo a través de una forma algorítmica y una función de aptitud (o de "fitness") o ciertos criterios de selección determinan qué variable debe ser retenida o eliminada del modelo.

Tabla 5.1: Ejemplos de métodos de selección de variables.

Métodos Lineales	Métodos No Lineales
Selección por pasos hacia adelante	Algoritmos genéticos
Eliminación por pasos hacia atrás	Programación evolutiva
Selección por pasos (combina los dos métodos anteriores)	Recocido simulado generalizado
Método de sustitución	Sistema de colonia de hormigas
Método de sustitución mejorado	Optimización de enjambre de partículas

La correlación entre la actividad o propiedad de interés y la estructura molecular se basará en métodos de regresión/clasificación que ponderan la contribución de cada una de las variables seleccionadas a la propiedad modelada. Estos métodos también pueden ser lineales y no lineales. La elección del tipo de método a emplear varía en cada caso y no suele ser evidente en primera instancia dado que ninguna técnica es consistentemente mejor que todas los demás. Por regla general, se prefiere ir de los métodos más sencillos a los más complejos.

Entre los métodos que se utilizan para la construcción de los modelos independientes después de la etapa de selección de variables o que incluyen una estrategia de selección de variables se pueden nombrar: Regresión Lineal Múltiple, Regresión de Componentes Principales, Mínimos Cuadrados Parciales, Redes Neuronales Artificiales, Maquinas de Soporte Vectorial, Árboles de Decisión, etc.

Se puede pensar que los errores en los que el modelo incurrirá al predecir la propiedad serán mayores o iguales a los errores presentes en los datos experimentales; por lo tanto, se desea disponer de datos experimentales sin errores significativos.

Cuando el conjunto de entrenamiento es compilado a partir de datos de literatura, se presenta una limitación para desarrollar un modelo QSAR con una variable dependiente continua (por ejemplo, IC_{50} o Ki), dado que los datos reportados no corresponden a las mismas condiciones experimentales (provienen de diferentes ensayos en distintos laboratorios). Para superar esta limitación se puede recurrir a los modelos QSAR clasificatorios lineales como los que se obtienen por la técnica conocida como **Análisis Lineal Discriminante (ALD)** o modelos QSAR clasificatorios no lineales como los obtenidos mediante **Árboles de Decisión**. Debido a su sencillez, en las próximas líneas se presenta una breve descripción de estos dos métodos, sin embargo, cabe la aclaración de que no son los únicos utilizados en para la generación de modelos QSAR cuantitativos.

Análisis Lineal Discriminante:

Esta técnica permite obtener un modelo clasificador lineal capaz de distinguir entre compuestos que presenten la actividad deseada (por ej, inhibidores de cierta enzima) de compuestos sin la actividad deseada (en el ejemplo, no inhibidores de la enzima elegida como blanco molecular).

ALD es un método de aprendizaje supervisado destinado a encontrar una combinación lineal de variables independientes (en nuestro caso, los descriptores moleculares) capaz de diferenciar entre objetos de dos o más categorías (en nuestro anterior, dos: ACTIVOS e INACTIVOS). La función discriminante (FD) obtenida corresponde a un plano en el espacio k-dimensional (siendo k el número de descriptores incluidos en el modelo) que, idealmente, deja a uno y otro lado los compuestos activos e inactivos. La Ecuación 5.3 representa en general a una FD:

$$\boldsymbol{FD = \sum_{i=1}^{k} a_i d_i}$$

Ecuación 5.3: Forma general de una función discriminante.

Siendo $\boldsymbol{a_i}$ el coeficiente de regresión asociado al descriptor molecular $\boldsymbol{d_i}$. Cada categoría de objeto (ACTIVOS o INACTIVOS) se asocia a un valor de una variable dependiente binaria arbi-

traria que funciona como etiqueta de las categorías. Por ejemplo, el valor 1 se asocia a los compuestos ACTIVOS y el -1 a los INACTIVOS. Dado que la función que se busca no predecirá una variable continua sino la categoría a la que pertenece cada elemento, el ALD y otras técnicas destinadas a encontrar modelos clasificadores pueden ser útiles para manejar datos ruidosos.

Árboles de Decisión

Los árboles de decisión son ampliamente utilizados debido a su capacidad predictiva, su facilidad de interpretación y su robustez para trabajar con datos ruidosos. Esta metodología permite obtener árboles de clasificación mediante un método jerárquico divisivo. Para construir un árbol se utilizan reglas de división del tipo: dado un elemento **o** a ser clasificado por una propiedad d, "si $d > x$ entonces el elemento o pertenece a la clase **A** y si $d \leq x$ entonces pertenece a la clase **B**" (siendo d en este caso un descriptor y x el valor de corte para ese descriptor). Se procede a realizar una serie de divisiones binarias de los datos en subconjuntos (nodos internos), hasta llegar a un árbol maximal donde se reparten todas las observaciones en la hojas o nodos terminales (ver Figura 5.4). En cada nodo del árbol, el algoritmo elige el descriptor que más eficazmente divide el conjunto de entrenamiento en subconjuntos enriquecidos en una clase u otra. Para el caso de la clasificación, cuando ninguno de los descriptores proporciona una ganancia de información, se crea un nodo terminal (hoja) donde se asigna la clase que está más representada en el grupo de compuestos asignados al nodo.

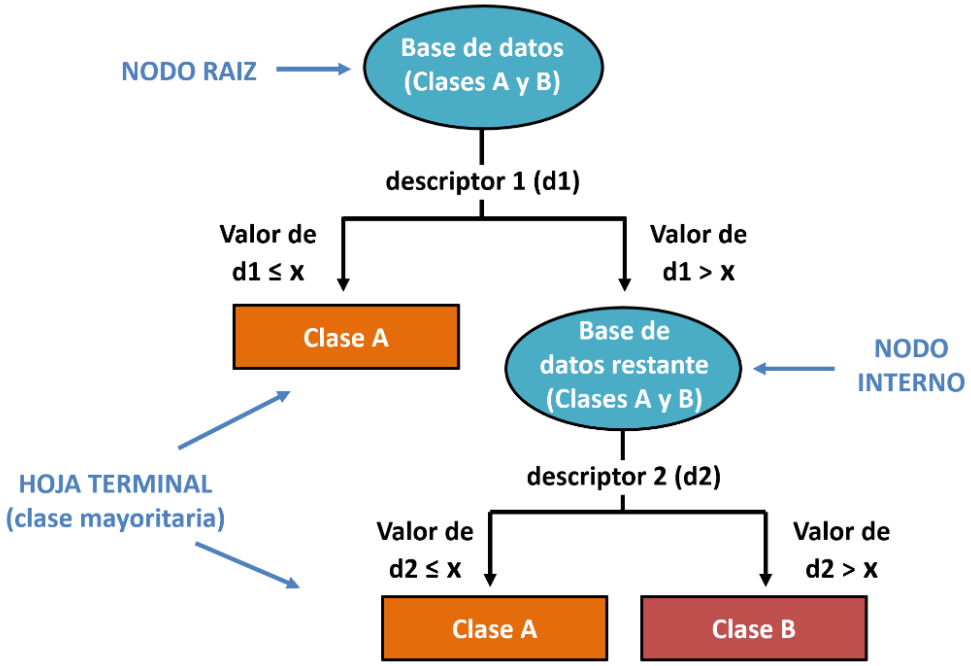


Figura 5.4: Árbol de decisión. "x" e "y" representan el valor de corte del descriptor de cada nodo.

5) Selección del Mejor modelo Generado:

Resultados de la validación externa: como se verá en el siguiente paso, la validación externa consiste en predecir la actividad de los compuestos del conjunto de prueba. Con la cantidad total de aciertos en la predicción, y la cantidad total de compuestos del conjunto de prueba se puede calcular el **% de buenas clasificaciones**, que es un buen estimador de la capacidad predictiva del modelo.

Principio de Parsimonia: también conocido como Navaja de Occam, es un principio metodológico y filosófico, según el cual "*No ha de presumirse la existencia de más cosas que las absolutamente necesarias*" es decir, cuando dos explicaciones logran explicar una misma observación, la más sencilla es la que se prefiere a priori. En este sentido, frente a la situación de encontrar varios modelos con buenos desempeños estadísticos, se optará por aquel que incluya menor número de descriptores (el más sencillo).

Sobreajuste (*overfitting*): se refiere al efecto de sobreentrenar un algoritmo de aprendizaje, es decir, que el modelo "memorice" las características del conjunto de entrenamiento (aumentando su poder explicativo o descriptivo) a expensas de sacrificar poder predictivo/capacidad de generalización (capacidad predictiva sobre compuestos no utilizados para el entrenamiento). Para los modelos lineales se exige una relación de al menos 10 entre el número de ejemplos de entrenamiento y el número de descriptores incorporados al modelo, de modo de reducir la probabilidad de sobreajuste.

En el caso de los árboles de decisión, se utilizan **algoritmos de poda** para evitar el sobreajuste. Estos remueven varias hojas y ramas del árbol, obteniéndose un árbol que tenga mejor poder de predicción como así también menor tamaño y complejidad. Básicamente hay 2 métodos de poda:

- post poda, que se lleva a cabo una vez que se finalizó la construcción del árbol
- poda en línea, que se realiza a medida que el árbol es inducido.

El principio que gobierna al proceso de poda es comparar la cantidad de errores que un árbol de decisión comete antes y después de cada posible procedimiento de poda, de modo de reducir al máximo ese error.

6) Validación del modelo:

Un proceso clave en el desarrollo de modelos QSAR es la validación de los mismos. Los procedimientos por los que se evalúan la robustez y capacidad predictiva de los modelos y, por lo tanto, su capacidad de predecir la actividad biológica (o la categoría) de compuestos aún no evaluados componen las técnicas de validación. Los métodos de validación se pueden clasificar en dos categorías: ***Validación interna***, que utilizan los compuestos del conjunto de entrenamiento para realizar la validación, y ***Validación externa***, que utiliza para evaluar la capacidad predictiva compuestos que no fueron utilizados para la generación de los modelos. Es de suma importancia utilizar ambos métodos para seleccionar modelos de buena calidad.

Validación interna

Validación cruzada (*cross-validation*): Las estrategias por excelencia son las técnicas denominadas *Leave One Out* (LOO, "Dejar uno afuera") y *Leave Group Out* (LGO, "Dejar un grupo afuera"). Estas estrategias permiten determinar la robustez de los modelos QSAR generados identificando si entre los compuestos del conjunto de entrenamiento hay algunos que influyen en mayor medida en el modelo generado (descriptores incluidos, valor de los coeficientes de regresión). La técnica consiste en la remoción de un número n (n igual a 1 corresponde a la técnica de LOO) de compuestos del conjunto de entrenamiento, la generación del modelo con los compuestos restantes, y la evaluación del nuevo modelo en el conjunto de compuestos removidos. Este proceso se repite hasta que todos los compuestos del conjunto de entrenamiento han sido removidos y evaluados por lo menos una vez.

Test de Aleatorización de Fisher: este ensayo se utiliza para descartar la probabilidad de correlación azarosa entre las variables independientes (los descriptores) y la variable dependiente (la propiedad o actividad). El método supone aleatorizar los valores de la variable dependiente en el conjunto de entrenamiento, cancelando por tanto cualquier relación que pudiera existir entre la estructura y la actividad. Posteriormente, se generan nuevos modelos para evaluar si existe probabilidad de correlación azarosa entre las variables independientes y la variable dependiente. Idealmente se espera que los modelos generados por aleatorización sean estadísticamente inferiores al modelo elegido.

Validación Externa: Esta validación consiste en evaluar la capacidad que poseen los modelos en la predicción de la actividad de un conjunto de moléculas que no hayan sido utilizadas para la generación de los modelos. Aunque las validaciones internas hayan obtenido buenos resultados, la gran mayoría de los especialistas coinciden en que el poder predictivo de los modelos QSAR sólo se puede establecer si el modelo se aplica con éxito para predecir los compuestos de un conjunto de prueba externo de tamaño adecuado. Para realizar esta validación, se aplican los modelos obtenidos en el conjunto de entrenamiento para predecir la actividad de los compuestos del conjunto de prueba.

Evaluación del desempeño de los modelos QSAR y selección del valor de corte.

Dos indicadores relevantes para estimar el desempeño de un modelo QSAR son su sensibilidad ($\boldsymbol{Se}$, la tasa de verdaderos positivos) y su especificidad ($\boldsymbol{Sp}$, la tasa de verdaderos negativos). Estos factores se definen por las Ecuaciones 5.4 y 5.5:

$$\boldsymbol{Se} = \frac{\boldsymbol{VP}}{\boldsymbol{VP+FN}}$$

Ecuación 5.4: Definición de Sensibilidad como indicador de desempeño de un modelo QSAR.

$$\boldsymbol{Sp} = \frac{\boldsymbol{VN}}{\boldsymbol{VN+FP}}$$

Ecuación 5.5: Definición de Especificidad como indicador de desempeño de un modelo QSAR.

Donde FN indica los falsos negativos (en nuestro caso, compuestos ACTIVOS que son clasificados por el modelo como INACTIVOS), FP indica los falsos positivos (compuestos INACTIVOS que son clasificados por el modelo como ACTIVOS), VN indica los verdaderos negativos (compuestos INACTIVOS clasificados como INACTIVOS) y VP indica los verdaderos positivos (compuestos ACTIVOS clasificados como ACTIVOS).

Cuando se utilizan modelos QSAR con fines clasificatorios cada molécula del conjunto de datos que se está evaluando obtendrá un resultado numérico único (o *score*) luego de la aplicación del modelo (volcado en el eje de las ordenadas, Figura 5.5). Se debe establecer un valor de corte del *score* a partir del cual se considerarán ACTIVOS e INACTIVOS los compuestos calculados. Por medio de la modificación de este valor de corte se puede observar que Se y Sp evolucionan de forma opuesta, por lo tanto, no es posible optimizar ambos parámetros de manera simultánea y se debe encontrar un balance adecuado entre los mismos. Para tal fin, se utilizan las curvas Receiver Operating Chracteristics (ROC), que son representaciones gráficas de la Se versus 1- Sp con las que es posible establecer el valor de corte óptimo de las funciones discriminantes (Figura 5.5).

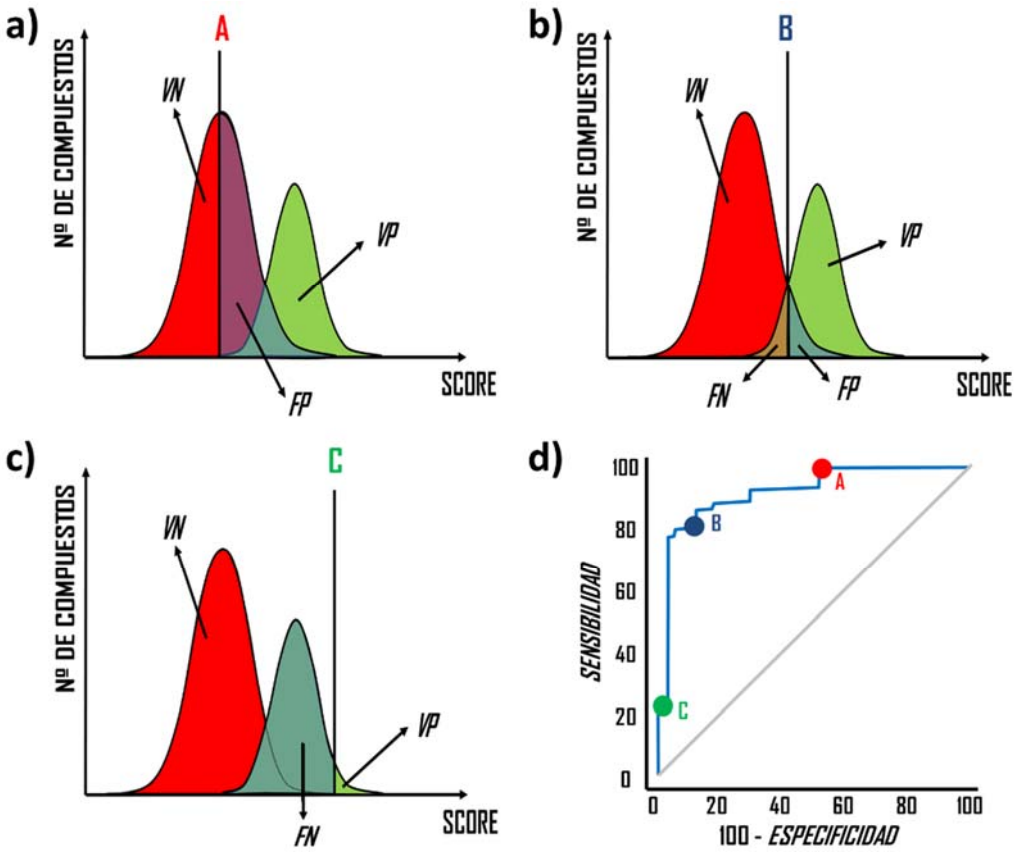


Figura 5.5: Esquema de la construcción de una curva ROC. Las figuras a), b) y c) representan gráficas de número de compuestos (eje y) versus el score obtenido por el modelo (eje x). Se puede observar como varían las proporciones de VN, VP, FN y FP cuando se modifica el valor de corte del score en los puntos A, B y C. d) Curva ROC.

Por otro lado, el **área bajo la curva ROC (AUCROC)** constituye una métrica valiosa para evaluar si el modelo se comporta significativamente mejor que una clasificación al azar y también permite comparar estadísticamente el desempeño de distintos modelos. Como puede observarse en la Figura 5.5, los ejes del gráfico de la curva ROC adoptan valores entre entre 0 y 100, aunque también pueden normalizarse a valores entre 0 y 1, por lo que un modelo ideal (Se y Sp igual a 1) corresponde a una curva que delimita una superficie cuadrada de área = 1. Un modelo clasificador que no difiera de la clasificación al azar obtendrá un AUCROC de 0,5. La capacidad de predicción de los modelos irá por lo tanto incrementándose a medida que el AUCROC aumente (se aleje hacia arriba de la línea diagonal) considerándose como un modelo clasificador perfecto a aquel que obtiene un AUCROC = 1.

La selección de un determinado balance entre Se y Sp no es una cuestión estadística, sino que depende del contexto. Por ejemplo, si pensamos en un laboratorio del ámbito público con bajos recursos económicos se priorizará Sp sobre Se para reducir el número de falsos positivos (compuestos que ensayarían en un test biológico, pero no tendrían la actividad predicha sobre el blanco molecular elegido). Esto significa, no obstante, que para reducir el número de falsos positivos se arriesga a perder andamios/motivos estructurales activos potencialmente valiosos. Lo contrario ocurrirá en laboratorios con altos recursos económicos, seguramente se priorizará la Se frente a la Sp de manera de no perder ningún candidato potencialmente valioso.

7) Cribado Virtual:

Una vez seleccionado y validado el mejor modelo QSAR se procede al CV de la base de datos elegida. Para llevar a cabo esta tarea, primero, para cada una de las moléculas de la base de datos a cribar, se realiza el cálculo de los descriptores moleculares que forman parte del modelo y luego, los valores obtenidos de los descriptores obtenidos se insertan en el modelo QSAR. De esa forma, el modelo obtiene un score (o una probabilidad) para cada uno de los compuestos de la biblioteca. Si el modelo generado era del tipo cuantitativo, el score representará el valor predicho de la actividad, mientras que si el modelo era del tipo QSAR cualitativo el valor del score se compara con el valor de corte seleccionado y de esa forma se predice la categoría o clase de cada compuesto.

Referencias

Bellera, C. L. (2014). *Búsqueda racional de nuevos fármacos antichagásicos inhibidores de la cruzipaína* (Tesis doctoral). Recuperada del repositorio institucional de la Universidad Nacional de La Plata.

Di Ianni, M. E. (2014). *Topología molecular aplicada a la búsqueda de nuevos fármacos para el tratamiento de la epilepsia refractaria* (Tesis doctoral). Recuperada del repositorio institucional de la Universidad Nacional de La Plata.

Gantner, M. E. (2016). *Topología molecular aplicada al reconocimiento de sustratos de la proteína de resistencia del cáncer de mama (BCRP)* (Tesis doctoral). Recuperada del repositorio institucional de la Universidad Nacional de La Plata.

Graham L. Patrick. An Introduction to Medicinal Chemistry. 5th ed. Oxford University Press, New York. 1995. ISBN 0-19-855872-4.

Talevi, A. & Bruno-Blanch, L.E. (2009). Screening virtual: Una herramienta eficaz para el desarrollo de nuevos fármacos en Latinoamérica. *Latin American Journal of Pharmacy*, 28(1), 141–150.

Triballeau, N. et al. (2005). Virtual screening workflow development guided by the "receiver operating characteristic" curve approach. Application to high-throughput docking on metabotropic glutamate receptor subtype 4. *Journal of Medicinal Chemistry*, 48(7), 2534–2547.

Truchon, J.F. & Bayly, C.I. (2007). Evaluating virtual screening methods: Good and bad metrics for the "early recognition" problem. *Journal of Chemical Information and Modeling*, 47(2), 488 508.