Relación cuantitativa entre estructura y actividad

Betsaida L. Covarrubias

2023-08-22

# Relación cuantitativa entre estructura y actividad

La relación cuantitativa entre estructura y actividad (QSAR, por sus siglas en inglés) es un enfoque de modelado en el campo de la química y la bioquímica que busca establecer relaciones cuantitativas entre las características estructurales (descriptores) de compuestos químicos y su actividad biológica.

Principales objetivos de los estudios QSAR:

* Predicción de la actividad biológica de los compuestos.
* Optimización de los compuestos iniciales prometedores.
* Diseño de nuevos compuestos activos.
* Predicción de riesgo y toxicidad.
* Modelización del perfil farmacocinético y farmacodinámico de nuevas entidades químicas.
* Descubrimiento de compuestos con la actividad deseada mediante la búsqueda en bases de datos químicas o bibliotecas virtuales.

Flujo de trabajo para análisis QSAR (ver Figura 1):

1. Recopilación de datos y preparación de datos (limpieza, selección, tratamiento de valores faltantes, etc.)
2. Cálculo y preprocesamiento de parámetros moleculares (tamaño, forma, carga eléctrica, grupos funcionales, aspectos estructurales y químicos, etc) para obtener descriptores numéricos.
3. Generación de conjunto de datos (conjuntos de entrenamiento y prueba)
4. Selección de descriptores
5. Construcción de modelos
6. Validación interna y externa
7. Desarrollo del modelo

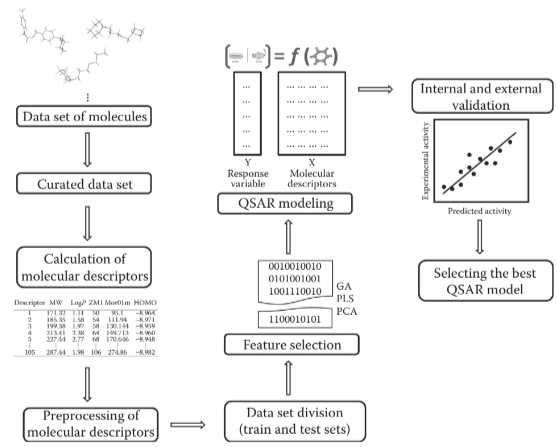


Figura 1. Representación esquemática del flujo de trabajo QSAR.

# Datos

Los datos, al ser objetos propensos a errores en cualquier entorno computacional, deben ser manejados con cuidado. Los datos utilizados para construir un modelo QSAR generalmente se extraen de información experimentalmente publicada en la literatura después de una curación adecuada, como la eliminación de valores atípicos, para mejorar la capacidad predictiva del modelo QSAR. Esta omisión de puntos de datos debe estar respaldada por una buena explicación.

**Recomendaciones sobre el número total de compuestos y el rango de valores de actividad biológica y desviación estándar**

Cuando se trata del número de compuestos estudiados:

* No debe ser demasiado bajo para evitar la correlación aleatoria y el sobreajuste.
* No debe ser excesivamente alto para prevenir restricciones computacionales.
* Se requieren al menos 20 moléculas para llevar a cabo un análisis QSAR.
* El rango mínimo aceptable de actividades biológicas en un conjunto de datos equivale a una orden de magnitud.
* Para lograr una interpolación efectiva de actividades, se suele recomendar un rango de alrededor de 3,0 veces en escala logarítmica.
* La desviación estándar también debe cumplir con el criterio de al menos 1,0 unidad logarítmica.

**Valores de punto final**

Los valores de punto final son datos numéricos que representan la actividad biológica de los compuestos

Es recomendable que los datos biológicos utilizados para el desarrollo de un modelo QSAR sean derivados del mismo procedimiento experimental. Por ejemplo, las actividades biológicas recopiladas o los valores de punto final deben obtenerse de la misma especie utilizando los mismos protocolos. Además, las unidades utilizadas para definir las actividades biológicas deben cumplir con el Sistema Internacional de Unidades (SI), conocido como sistema métrico. Por ejemplo, todos los valores para IC50 deben expresarse en unidades molares y no en unidades de peso.

**Dominio de aplicabilidad**

El dominio de aplicabilidad es la región en el espacio de descriptores químicos donde el modelo QSAR es confiable y puede hacer predicciones precisas. Para evaluar la capacidad predictiva del modelo, se emplean herramientas de Análisis de Dominio (AD) que utilizan descriptores y actividades biológicas en los conjuntos de entrenamiento y prueba. Algunos ejemplos de herramientas de AD incluyen:

* Distancia en el espacio de descriptores.
* Gráficos de dispersión.
* Comparación de descriptores.

Si un compuesto estructuralmente similar cae dentro del espacio químico abarcado por el modelo QSAR, la predicción de actividad biológica debe coincidir con el rango de actividades biológicas usadas para entrenar el modelo. Esto asegura predicciones confiables.

Es importante considerar problemas como nombres químicos incorrectos y estructuras duplicadas. Las estructuras químicas repetidas en los conjuntos de entrenamiento y prueba pueden dar lugar a un aumento falso en el rendimiento predictivo de los modelos QSAR. Estas repeticiones pueden surgir de las mismas estructuras con diferentes nombres o de compuestos con diferentes actividades biológicas.

**Servidores de base de datos**

Los datos de entrada, se pueden extraer tanto de la literatura como de bases de datos. Estos tipos de datos generalmente consisten en estructuras moleculares junto con sus correspondientes actividades biológicas.

# Descriptores moleculares

Los descriptores moleculares son características numéricas o datos cuantitativos que describen diversas propiedades de las moléculas químicas.

Estos parámetros se calculan considerando toda la molécula o una parte de ella, a partir de la cual se generan descriptores basados en la molécula completa o en fragmentos.

Los descriptores se pueden clasificar según diferentes parámetros:

* Constitucionales
* Topológicos
* Geométricos
* Termodinámicos
* Electrónicos

También se pueden clasificar según su dimensionalidad:

* 0D
* 1D
* 2D
* 3D
* 4D

Además, se pueden clasificar según su origen:

* No dependen de la estructura (por ejemplo, LogP, pKa, refractividad molar [MR], momento dipolar y polarizabilidad)
* Dependen de la estructura

El cálculo de ciertos valores de los descriptores moleculares está influenciado por la disposición de la molécula. Para lograr resultados precisos, se optimiza la geometría de las estructuras, buscando la mínima energía. Esto se logra utilizando técnicas de mecánica cuántica o mecánica molecular.

**Descriptores comúnmente utilizados**

Los descriptores constitucionales reflejan las propiedades químicas relacionadas con la composición molecular. Algunos ejemplos son:

* Peso molecular (MW)
* Número de átomos
* Número de enlaces
* Número de enlaces giratorios
* Número de enlaces múltiples

Los descriptores topológicos se derivan de la teoría de los grafos y proporcionan información sobre el tamaño, forma, hibridación, ramificación y flexibilidad molecular. Algunos ejemplos son:

* Índice de Wiener: Evalúa la distancia promedio entre átomos y la ramificación molecular.
* Índice de Randic: Evalúa la conectividad y ramificación molecular.
* Índices Kappa: Evalúan la asimetría y complejidad tridimensional.
* Índice de Zagreb: Considera la conectividad y energía π de la molécula.
* Índice Balaban J: Evalúa complejidad y anillos en la molécula.

Los descriptores geométricos son influenciados por la disposición tridimensional (3D) de átomos. Ejemplos son:

* Descriptores moleculares invariantes holísticos ponderados (WHIM). Estos contienen información sobre el tamaño, forma, simetría y distribución de átomos
* Autocorrelación 3D, geometría, topología y ensamblaje de pesos atómicos (GETAWAY)
* Descriptores de representación molecular 3D basados en difracción de electrones (3D-MoRSE).

Los descriptores termodinámicos brindan información sobre interacciones bioquímicas:

* Lipofilicidad (LogP)
* Refractividad molar
* Calor de formación

Los descriptores electrónicos se derivan de la distribución electrónica y se calculan con teoría mecano-cuántica:

* Orbitales moleculares más altos ocupados (HOMO)
* Orbitales moleculares más bajos desocupados (LUMO)
* Momento dipolar
* Polarizabilidad
* Cargas atómicas

Esta variedad de descriptores es esencial para caracterizar diferentes aspectos de las moléculas y su comportamiento.

**Descriptores moleculares basados en fragmentos**

El uso de descriptores moleculares basados en fragmentos ofrece numerosas ventajas, como la simplicidad en el cálculo y la interpretación más clara del efecto local de las propiedades estructurales en relación con la actividad biológica. Estos descriptores se derivan de diversos fragmentos moleculares, incluyendo átomos, enlaces, cadenas, torsiones, subestructuras comunes y subgrafos básicos. Aquí hay un ejemplo de dos tipos de descriptores basados en fragmentos:

* Constante electrónica (σx): Esta constante está asociada con la densidad electrónica alrededor de átomos específicos dentro de la estructura molecular. Proporciona información sobre cómo los electrones están distribuidos en la molécula.
* Análisis Free-Wilson: Este análisis evalúa cómo diferentes fragmentos o sustituyentes en una molécula influencian su actividad biológica. Permite identificar qué partes de la molécula contribuyen significativamente a su actividad.

Una forma de usar descriptores basados en fragmentos es con el recuento de subestructuras, lo que significa que se seleccionan ciertos fragmentos y se cuentan cuántas veces aparecen en un conjunto de datos. Estos recuentos de subestructuras se convierten en números que representan cuántas veces aparece cada tipo de subestructura en una molécula en particular. Luego, estos números se utilizan como “descriptores” para caracterizar cada molécula. En otras palabras, en lugar de describir una molécula en detalle, se puede describir de manera más concisa utilizando los recuentos de subestructuras como características importantes.

**Software para cálculo de descriptores moleculares**

Existen varios software y herramientas disponibles para el cálculo de descriptores moleculares. Estos programas ayudan a calcular una amplia variedad de descriptores basados en la estructura molecular de compuestos químicos.

# Selección de descriptores

**Normalización**

En el proceso de normalización de los descriptores moleculares, se utilizan dos métodos: la normalización automática o estándar, y la normalización de rango.

En la normalización automática, los valores completos de los descriptores se transforman de manera que tengan una media de cero y una varianza de uno. En el caso de la normalización de rango, se logra al escalar los valores de los descriptores dentro de un rango específico, como por ejemplo [0, 1] o [-1, 1]. Esto se realiza dividiendo cada valor por la diferencia entre los valores máximo y mínimo de los descriptores y luego ajustando el rango según sea necesario

La razón principal para normalizar los descriptores moleculares es la existencia de diferencias significativas entre los diferentes descriptores en términos de su rango de valores, lo que conduce a dificultades para determinar el efecto ponderado de cada descriptor en el modelo. En otras palabras, los descriptores con valores más pequeños pueden quedar opacados por aquellos con valores más grandes.

**Colinealidad**

Colinealidad en QSAR se refiere a la fuerte correlación entre descriptores moleculares. Cuando dos descriptores están altamente relacionados, es difícil determinar cuál influye en la actividad.

La alta colinealidad entre descriptores puede dificultar la interpretación correcta de los resultados, dar lugar a un modelo menos confiable y estable, y aumentar la complejidad innecesaria del modelo. Por lo tanto, es importante identificar y manejar la colinealidad para obtener resultados más confiables y útiles en el análisis de QSAR.

**División de entrenamiento y prueba**

Uno de los pasos previos recomendados es dividir el conjunto de datos en conjuntos de entrenamiento y prueba.

El principal desafío en esta división recae en el número y la distribución de los compuestos entre los conjuntos de entrenamiento y prueba. Los compuestos en el conjunto de prueba se utilizan para la validación externa durante la construcción del modelo. En relación al tamaño del conjunto de prueba, se sugiere que represente entre el 15 % y el 20 % del conjunto de datos total.

Para lograr distribuciones uniformes en ambos conjuntos, es esencial que los compuestos representativos en términos de sus valores finales y descriptores moleculares estén equilibradamente presentes en ambas particiones. El equilibrio implica asegurarse de que los conjuntos de entrenamiento y prueba contengan compuestos con una variada distribución de actividades y características moleculares. El objetivo es obtener una representación adecuada y diversa de compuestos en ambos conjuntos, lo que garantiza que el modelo sea sólido y eficaz en su capacidad para realizar predicciones precisas en diversas situaciones. Se emplean diversos algoritmos para llevar a cabo la división del conjunto de datos, basados en criterios específicos. La selección puede ser:

* Aleatoria: No se considera ningún patrón específico ni criterio predeterminado al seleccionar los compuestos.
* Racional: Utilizan el rango completo de actividades y/o descriptores moleculares como criterio de división.

Los métodos de selección racionales previamente mencionados se basan en algoritmos generales, que incluyen:

* Agrupamiento k-means: Los compuestos se dividen en k grupos según su similitud, evaluada a partir de los valores promedio en diferentes características para cada grupo. Luego, se elige un subconjunto específico de compuestos (por ejemplo, el 20%) de cada grupo para formar el conjunto de prueba, mientras que los restantes conforman el conjunto de entrenamiento
* Kennard-Stone: Este método implica seleccionar compuestos de manera iterativa de modo que estén lo más distantes posible en términos de sus características. Comienza eligiendo dos compuestos distantes y luego agrega compuestos adicionales, asegurándose de que estén lo más alejados posible de los ya seleccionados. Esto garantiza que el conjunto de prueba represente una amplia diversidad de propiedades moleculares, esencial para evaluar la precisión de un modelo QSAR.
* Exclusión de esferas: En este método, se parte de un compuesto y se crea una “esfera” imaginaria alrededor de él. Luego, se excluyen los compuestos dentro de esta esfera, eligiendo el compuesto más distintivo, el cual se añade al conjunto de prueba. Este proceso se repite, excluyendo compuestos en esferas adicionales.
* Mapa autoorganizado de Kohonen: El algoritmo busca la distancia más cercana entre los datos de los compuestos para crear un mapa. En este mapa, los compuestos similares se agrupan juntos en áreas cercanas. Esto implica que en el mapa, las zonas con compuestos similares están ubicadas cerca una de la otra.
* Diseño molecular estadístico: Este método implica realizar un análisis de componentes principales y generar gráficos de puntuación basados en dicho análisis. Estos gráficos muestran cómo los compuestos se agrupan según sus características moleculares. Los grupos identificados en los gráficos representan compuestos con similitudes en términos de descriptores. Estos grupos se eligen como el conjunto de entrenamiento, ya que se considera que capturan la diversidad de compuestos presentes en el conjunto de datos original. Esto se debe a que el análisis de componentes principales tiende a resaltar las principales fuentes de variación en los datos, lo que puede corresponder a diferencias fundamentales entre los compuestos en términos de sus características moleculares.
* Métodos de selección de conjuntos de prueba orientados a la extrapolación: En este método, se selecciona un par de compuestos con la mayor distancia euclidiana en términos del espacio de descriptores y se trasladan al conjunto de prueba. Esto se repite hasta alcanzar el tamaño deseado para el conjunto de prueba.

**Métodos de selección de descriptores**

El proceso de selección de descriptores es crucial para reducir el conjunto amplio de características moleculares y resaltar las más pertinentes. Se busca elegir aquellos descriptores más relevantes para el análisis.

Dentro de las estrategias de selección, se encuentran los métodos de envoltura, filtro e híbridos.

Los métodos de envoltura se basan en dos procesos matemáticos: el uso de una función objetivo y un algoritmo de búsqueda de optimización para elegir parámetros importantes en la generación del modelo QSAR. La selección de parámetros depende del algoritmo de clasificación utilizado. Estas metodologías se clasifican en métodos clásicos, basados en inteligencia artificial y métodos híbridos.

Algoritmos de clasificación:

* Metodos clasicos
  + Selección hacia adelante: Se agregan características en cada paso, seleccionando la característica que más mejora la métrica de rendimiento (como R2 o el error) en el conjunto de datos de entrenamiento. Este método es intuitivo y simple, pero podría no considerar interacciones entre características.
  + Eliminación hacia atrás: En cada paso, se elimina la característica menos relevante hasta que el rendimiento no mejora significativamente. Esto puede ser útil para reducir la complejidad del modelo, pero podría perder características importantes. Selección paso a paso: Se combina la idea de la selección hacia adelante y hacia atrás. Comienza con un modelo sin características y luego realiza pasos iterativos para agregar o eliminar características una a una
  + Método de selección y modelado de variables basado en la predicción: Participan dos criterios : “coeficiente de interrelación entre los pares de descriptores (R int) y coeficiente de correlación (q^2) obtenido utilizando la técnica de validación cruzada de exclusión uno a uno”
* Métodos basados en inteligencia artificial
  + Algoritmo genético
  + Simulado
  + Optimización por enjambre de partículas
  + Sistema de colonia de hormigas
  + Bosque aleatorio
  + Máquinas de soporte vectorial
* Métodos híbridos

Los métodos de filtro eliminan descriptores poco relevantes basándose en parámetros estadísticos como la baja variabilidad y correlaciones entre parámetros . Estos métodos son apropiados para conjuntos de datos con un gran número de características (descriptores). Son sencillos, rápidos y no requieren de un clasificador para su aplicación.

Ejemplos de métodos de filtro:

* Criterios estadísticos
  + Prueba T: compara las medias de dos grupos para ver si hay una diferencia significativa entre ellos.
  + Prueba Chi-cuadrado: se utiliza para evaluar si hay una relación entre dos variables categóricas.
  + Prueba Fisher: se utiliza para comparar las varianzas de dos grupos.
  + Evaluacion Kolmogorov-Smirnov: se utiliza para comparar distribuciones de datos y determinar si provienen de la misma distribución.
* Métodos basados en correlación
  + Coeficiente de correlacion de Pearson: Mide la relación lineal entre dos variables.
* Teoría de la información: mide la cantidad de información compartida entre diferentes características, es decir, si la información de una variable a su vez da información de la otra variable.

Los estadísticos, métodos de correlación y teoría de la información se aplican de manera independiente, lo que significa que se usan tal como son sin tener en cuenta si sería útil aplicar algún tipo de transformación a los datos antes de usar estos métodos, por ejemplo, normalizar.

Por último, los métodos híbridos realizan primero un preprocesamiento para eliminar características irrelevantes (métodos basados en Filtro) y luego utilizan los métodos de Envoltura para seleccionar descriptores.

**Herramientas y software**

Algunos algoritmos de selección de características están disponibles como paquetes para entornos programables (como R y MATLAB®); sin embargo, también existen software independientes con interfaces gráficas de usuario personalizadas (GUIs).

En resumen, los principales desafíos en este procedimiento son el uso de descriptores colineales, la utilización de un número excesivo de descriptores en el QSAR, la falta de escalado automático de descriptores y la selección inadecuada de conjuntos de entrenamiento/prueba.

# Construcción del modelo

Se eligen descriptores importantes usando diferentes algoritmos en QSAR. El objetivo es crear un modelo efectivo con pocos descriptores que expliquen muchos datos de actividad. Los datos de entrenamiento son esenciales para este proceso, y la elección del mejor modelo predictivo se basa en su rendimiento en un conjunto de pruebas, evaluado mediante análisis estadísticos.

Los métodos de construcción de modelos se dividen en lineales y no lineales, según cómo se relacionan los descriptores con las actividades biológicas.

**Modelos lineales**

* Regresión lineal múltiple
* Mínimos cuadrados parciales
* Análisis de componentes principales
* Análisis de wilson libre
* Análisis discriminante lineal

**Modelos no lineales**

* Red neuronal artificial
* k-vecino más cercano
* Árbol de decisión
* Máquinas de vectores soporte

**Regresión lineal múltiple**

Es una extensión de la regresión lineal simple en la que se considera más de una variable independiente para predecir una variable dependiente. La fórmula de regresión lineal múltiple es: . Donde es la variable dependiente, son las variables independientes, y son los coeficientes.

Los coeficientes indican cómo las variables independientes (descriptores moleculares) influyen en la variable dependiente (actividad biológica). Si los valores de los descriptores se normalizan (se escalan para tener la misma escala), entonces la magnitud de estos coeficientes es relevante. Se consideran los siguientes puntos clave:

* Si un coeficiente es positivo para un descriptor, significa que un aumento en ese descriptor se asocia con un aumento directo en la actividad biológica.
* Si un coeficiente es negativo, significa que un aumento en ese descriptor se relaciona inversamente con la actividad biológica.
* Un valor p menor que 0.05 generalmente se considera significativo, lo que significa que el coeficiente tiene una influencia real en el modelo y no es solo el resultado del azar.
* Se sugiere que el número de compuestos (muestras) en el conjunto de datos sea mayor que el número de variables independientes (descriptores). Se recomienda que la relación entre el número de compuestos y el número de variables independientes sea al menos 5:1 según la “relación de Topliss”.
* Los descriptores colineales resultará en un modelo QSAR inexacto en términos de rendimiento.

**Valores atípicos**

Los valores atípicos son compuestos que se desvían de las predicciones del modelo y pueden surgir debido a variaciones estructurales inusuales o errores experimentales en las mediciones. Para categorizar los tipos de valores atípicos, se dividen en:

* Valores atípicos estructurales: ocurren cuando un compuesto tiene una estructura molecular inusual que puede llevar a un comportamiento biológico imprevisto.
* Valores atípicos de actividad: ocurren cuando pequeños cambios en la estructura molecular conducen a cambios drásticos en la actividad biológica, lo que va en contra de la expectativa de que compuestos similares tengan actividades similares.

Para tratar con los valores atípicos, se han utilizado diferentes métodos en la literatura, como:

* Técnicas estadísticas: Pueden incluir cálculos de estadísticas como desviaciones estándar, percentiles o puntuaciones z para determinar qué puntos de datos se consideran atípicos en relación con el resto.
* Análisis de distancia: En el análisis de distancia, se mide la distancia entre cada punto de datos y los demás. Los valores que están significativamente alejados de los demás pueden considerarse valores atípicos. Diferentes medidas de distancia, como la distancia euclidiana, pueden utilizarse según el contexto.
* Máquinas de vectores de soporte (SVM): Las SVM son un tipo de algoritmo de aprendizaje automático que se utiliza para el análisis y la clasificación de datos. En el contexto de valores atípicos, las SVM pueden ayudar a identificar y etiquetar los puntos de datos que se desvían significativamente de la tendencia general.

Eliminar los valores atípicos no siempre es necesario ni garantiza una mejora en el modelo. Es esencial realizar un análisis más detallado para determinar si los valores atípicos son verdaderos y si deben ser eliminados del análisis.

**Herramientas y software**

El creciente interés en la aplicación de estudios QSAR en diferentes aspectos del diseño y desarrollo de medicamentos ha llevado a la introducción de una variedad de algoritmos y herramientas directa o indirectamente adecuadas para los estudios QSAR.

**Desafíos en el desarrollo del modelo**

* Detección y manejo de valores atípicos
* Sobreajuste de datos, qué ocurre cuando se incorporan demasiados descriptores al modelo, lo que puede afectar su capacidad de generalización.
* El uso de descriptores ambiguos y no interpretables
* Proporcionar medidas estadísticas adecuadas al informar los resultados del modelo QSAR para asegurar su validez y aplicabilidad.

# Validación del modelo QSAR