Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias Fisico-Matemáticas

Propuesta de modificación metodológica QSAR

Tesis presentada al

Colegio de Matemáticas

como requisito parcial para la obtención del título de

Licenciado en Matemáticas

por

Eduardo Hernández Montero

asesorado por

Dr. Andrés Fraguela Collar
Facultad de Ciencias Físico Matemáticas, BUAP

Dr. Thomas R. F. Scior Jung
Facultad de Ciencias Químicas, BUAP

Puebla Pue.

Enero de 2012

Agradezco a:

mi familia por todo el apoyo;

mi director y codirector de tesis por permitirme trabajar con ellos;

Jorge Lozano Aponte, licenciado en Farmacia y colaborador indispensable para el desarrollo de este trabajo, por el apoyo y las atenciones académicas prestadas;

la Dra. Lidia Aurora Hernández Rebollar y los doctores Alexander Grebenikov y José Jacobo Oliveros Oliveros por tomarse aceptar ser parte del jurado para la defensa de esta tesis, por la amabilidad y atención con que lo hicieron;

la Dra. Lucía Cervantes, el Mto. en ciencias Julio E. Poisot y el Dr. Mario Aurelio Rodríguez por el apoyo académico y personal durante mi formación básica en matemáticas;

todos los buenos amigos con los que tengo el gusto de contar y colaborar, en y para lo que considero impotante;

...todos, porque con lo poco que conocí de mi en los tropiezos que tuve durante el desarrollo de este trabajo, estoy convencido de que no fue fácil permanecer ahí. Gracias.

Índice general

Índice de tablas

Índice de figuras

Introducción

Para aclarar algunas nociones básicas al lector, recordamos que un fármaco o medicamento es una sustancia que es introducida en un organismo y produce un efecto terapéutico, entre los objetivos de la farmacología se encuentra, por supuesto, el diseño de nuevos fármacos eficientes en el combate de padecimientos y enfermedades específicas.

El diseño *in silico* de compuestos es, el diseño con herramientas de cómputo de compuestos químicos teóricos con síntesis viable, que se espera puedan ser empleados posteriormente en el tratamiento de las enfermedades que motivan su creación. La importancia del cambio de escala y el correcto uso de las herramientas teóricas de la matemática, queda de manifiesto cuando se realiza un diseño molecular *in silico*, esperando que una vez que sea sintetizada la molécula teórica, ésta logre producir el efecto terapéutico deseado.

El presente texto es el resultado de la colaboración con los especialistas en farmacología Dr. Thomas Scior y C.M. Lic. Jorge Lozano Aponte, en un esfuerzo que inicialmente tuvo por principal objetivo, realizar conjuntamente una cuidadosa revisión de una de las metodologías más empleadas en la investigación y diseño *in silico* de compuestos, que se proponen como candidatos para convertirse en nuevos medicamentos.

Conforme avanzábamos en el desarrollo de nuestro trabajo, poco a poco entendimos que el primer objetivo planteado era demasiado ambicioso y general, que los objetos y métodos de estudio de la farmacología de los que nos ocupamos, son ejemplos de fenómenos en los que se estudian diversos procesos físicos, químicos y biológicos relacionados entre si, que ocurren en escalas considerablemente distintas y para los cuales el cambio entre una escala y otra es un aspecto que debe considerarse con más cuidado en la metodología de investigación.

Primero hay que pasar de una escala hasta cierto punto atómica, a una escala molecular en donde existen propiedades y características de una molécula que en lo individual no observan los átomos que la conforman, en un segundo momento se transita a una escala en la que el "ligando" (compuesto químico que es candidato a fármaco) intaractúa, por ejemplo, con una enzima intracelular, es decir, una escala donde la interacción observada es del tipo *molécula-célula*, el nivel siguiente es en el que la interacción de interés ocurre entre una molécula y un órgano o sistema de un organismo biológico complejo y, por último, la interacción que se observa es entre un organismo complejo y una concentración muy pequeña de un compuesto químico ¿Cómo poder anticipar en alguna de estas escalas los valores de mediciones relevantes de las posibles interacciones del ligando?

Observando con atención, una molécula diseñada *in silico* no es más que una representación tridimensional de un arreglo de átomos de uno o más elementos químicos, que

mantienen relaciones de proximidad espacial mediante la presencia de enlaces covalentes entre ellos; cómo obtener información, conociendo sólo el diseño *in silico*, sobre el efecto que una cierta dosis tendrá en un organismo. La respuesta no es trivial, al grado en que aún no existe un procedimiento general y estandarizado de diseño que siempre permita disponer de medicamentos lo más eficientes posibles, que además no causen efectos secundarios no deseados y su costo de producción sea mínimo, entre otras cosas.

Una de las respuestas que la farmacología ha dado, es la metodología que aquí estudiamos, propia de la farmacología y referenciada con las siglas en ingles **QSAR** (*Quatntitative Structure-Activity Relationships*); tiene sus orígenes hace aproximadamente 60 años, desde entonces ha estado sujeta a un constante desarrollo, simplificaciones e hipótesis *a priori* sobre la forma de interacción de los candidatos a nuevos medicamentos.

Desde ahora comentamos al lector que existe una diferencia importante entre metodología y análisis QSAR, el análisis es una de las partes medulares de la metodología, pero de ninguna forma es equivalente a ella. Un análisis QSAR es una sucesión de métodos estadísticos y ajuste de parámetros para el análisis de datos sobre un conjunto de mediciones tanto de las características estructurales y fisicoquímicas del ligando, así como de la actividad o respuesta biológica observada en el sistema.

La principal simplificación a partir de la que se realizan los análisis QSAR, proviene de la dificultad de realizar mediciones experimentales de lo que se conoce como actividad biológica, el tipo de mediciones para esos análisis se dice que se realizan *in vitro*: a nivel de laboratorio y como consecuencia de la observación del cambio en los estados termodinámicos de un sistema inmerso en un medio solvente (M), en el que los únicos compuestos presentes son el ligando , el "receptor celular" (biomolécula en el organismo con la que debe interactuar el ligando), y los necesarios para medir una respuesta biológica de interés, de tal forma que el sistema sea "simple" en términos termodinámicos y consecuentemente las reacciones químicas que se producen en él lo sean también.

El medio solvente M es una simplificación de las múltiples fases de un sistema biológico, por lo general la actividad se encuentra expresada en función de la concentración de ligando en un estado determinado del sistema. La forma en que se entiende actualmente la actividad biológica, con apoyo de la termodinámica, es la diferencia observada en la energía libre de Gibbs, al pasar de un estado en el que entre ligando y receptor no existen interacciones, al estado en que se ha formado el complejo *ligando-receptor*, en una concentración específica que se requiera para los fines particulares de investigación (*ver* [5]).

El conjunto muestral de moléculas con las que se realiza el análisis, debe garantizar ser un subconjunto propio de la familia química de compuestos orgánicos, sintetizables, que queda determinada por la afinidad no nula con el receptor o blanco biológico previamente especificado para el análisis.

Un análisis QSAR incluye la selección de las características estructurales mensurables y estadísticamente significantes, utilizando métodos como análisis factorial o de componentes principales, la reducción en la cantidad de las características seleccionadas, un ajuste por mínimos cuadrados de un modelo general y la validación experimental del mismo (*ver* [17] y [24]).

En años recientes Scior T. expuso en [24] una revisión crítica de los análisis QSAR, sobre los cuidados especiales que deben procurarse a este tipo de trabajos para evitar

incurrir en errores graves de procedimiento; ocasionados por una sobre valoración de los procesos estadísticos, en ausencia de información sobre las propiedades fisicoquímicas de los compuestos; o enraizados en una significativa debilidad de la hipótesis de linealidad sobre la que, por lo general, se sustenta un análisis QSAR.

Scior T., señala que los no despreciables errores en las predicciones de los modelos QSAR pueden tener su origen en errores cometidos durante el proceso de selección de los descriptores utilizados, así como su cantidad; no obstante, en la validación de modelos lineales se han observado graves fallos predictivos pese a que tales procedimientos previos sean confiables.

Observando el fenómeno que busca describir un modelo QSAR y las ocasiones en que estos modelos presentan fallos graves en sus predicciones, conjeturamos que tales errores pueden residir principalmente en: modelos incompletos o insuficientes del fenómeno que describen (insuficiencia de parámetros), o una falsa premisa de dependencia lineal de la respuesta como función de un conjunto de descriptores propios de una familia específica de ligandos. Las hipótesis que sustentan su desarrollo pueden ser una excesiva simplificación del problema, por lo que en cada análisis QSAR es necesario verificarlas o modificarlas de forma pertinente.

Por otro lado, las ecuaciones de la termodinámica exhiben a la energía libre de Gibbs, a temperatura constante, como una combinación lineal de dos componentes energéticas de las que pueden realizarse mediciones experimentales, entalpía y entropía; sin embargo, este hecho no es considerado en el momento de realizar los ajustes del modelo general, siendo tal omisión una posible causa de modelos poco congruentes con el fenómeno real, Scior T. sugiere revisar a detalle.

A grandes rasgos, la metodología QSAR puede dividirse en cuatro etapas fundamentales:

- Propuesta de modelos generales: modelación matemática de las relaciones relevantes para la farmacología, entre fenómenos que ocurren a escalas distintas del proceso biofisicoquímico que un medicamento desencadena para producir un efecto terapéutico.

- Análisis QSAR: obtención de datos experimentales que permitan ajustar los parámetros de los modelos generales para una familia específica de ligandos.

- Optimización: procesos de búsqueda de los ligandos que, de acuerdo con el modelo particular que resulte del ajuste de parámetros y criterios adicionales de selección, sean los “mejores candidatos a fármacos”.

- Validación y valoración de resultados: proceso de revisión teórica y experimental de los resultados del análisis QSAR y la optimización de sus modelos.

Conforme avanzábamos en el desarrollo de nuestro trabajo, se hizo claro que el objetivo principal propuesto de origen era sumamente general y se encontraba fuera del alcance de una tesis de licenciatura. Poco a poco logramos acotar nuestro campo de trabajo hasta llegar a lo que aquí ofrecemos al lector.

Nuestro interés se localiza sólo en la etapa del análisis QSAR y desde un enfoque relativamente tradicional, para el cual los modelos que se proponen son de carácter local y tipo lineal. Nos interesan la herramienta y las premisas del ajuste de parámetros para el modelo matemático que un análisis QSAR busca construir, para poder realizar predicciones sobre la respuesta biológica que producirán moléculas que aún no han sido probadas experimentalmente.

Como ya hemos dicho, el trabajo que se presenta no es de carácter general, nos concentramos en análisis QSAR en los que la actividad biológica se entiende como la inhibición del efecto catalítico de una enzima sobre su "sustrato natural", también conocido como *ligando endógeno*.

El enfoque y la mayoría de las propuestas de modificación que se presentan pueden ser extendidas, con el debido cuidado, a diversos tipos de fármacos; sin embargo, el desarrollo que presentamos está motivado y enfocado a fármacos que actúan inhibiendo algún tipo de actividad enzimática intracelular, es decir, fármacos que interactúan con enzimas en el interior de una célula, los antibióticos por lo regular son de este tipo de fármacos.

Lograr modelos matemáticos confiables es una de las necesidades prácticas de la farmacología, ya que las pruebas experimentales tienen un muy elevado coste económico y la demanda de medicamentos eficientes, con precios asequibles para el público en general es un tema ineludible hoy día.

La metodología de investigación QSAR alcanza ya un periodo de vida de alrededor de 50 años, con una popularidad que se refleja en los millares de publicaciones académicas referentes a modelos QSAR para diversas familias de fármacos, pero hay un problema, son mínimas las publicaciones que emplean definiciones estadarizadas que permitan caracterizar y evaluar formalmente esta metodología.

En general no se ha evaluado la forma de utilizar la herramienta matemática más empleada para pasar del análisis a los modelos QSAR, regresión lineal múltiple (RLM), tampoco existen definiciones claras y un criterio o una forma estandarizada de llegar a criterios de optimalidad que permitan discriminar qué fármaco es mejor en un conjunto de ellos.

En aproximadamente medio siglo de análisis QSAR, los cientos de casos particulares ha desembocado en igualmente cientos de definiciones de descriptores de naturaleza distinta, corresponden a enfoques distintos de representación de una molécula como un objeto matemático, corresponden a distintas teorías físicas y químicas que han evolucionado con el paso del tiempo. La acentuada heterogeneidad de los casos particulares QSAR supone para todos un esfuerzo adicional para lograr una comunicación efectiva entre las distintas disciplinas. Por lo regular las definiciones y desarrollo de resultados en matemáticas suelen observar una relación inversa: tanto más general, menos amigable para el lector.

Capítulo 1

Metodología QSAR

En las secciones correspondientes se emplean objetos y resultados propios de las teorías de álgebra lineal, estadística y probabilidad, la mayoría de ellos sólo cuentan con una presentación de notación en este texto por ser herramientas básicas de amplia difusión. Se sugiere el libro de S. H. Friedberg *et ál.* [8] al lector interesado en los detalles de las demostraciones y definiciones aquí omitidos de álgebra lineal, el libro de T. Apostol [1] para análisis matemático, y los libros de J. S. Rosenthal. *et ál.* [21] y Hoog [13] en lo referente a probabilidad y estadística.

La metodología que aquí estudiamos, propia de la farmacología para el diseño *in silico* de nuevos medicamentos, se fundamenta en premisas empíricas que por lo general en la literatura se formulan haciendo uso de términos como medible o mensurable, cercanía, proximidad, relación o función; sin embargo, el significado de palabras como estas no necesariamente está dado por una rigurosa definición matemática, tener esto en mente es importante para el primer acercamiento, con este tema, del lector acostumbrado a textos especializados en matemáticas.

Tan pronto como nos sea posible, dedicaremos atención y cuidado en brindar definiciones formales, nuestro objetivo, lograr transmitir al lector con formación básica en matemáticas, la importancia de la modelación en el proceso metodológico de investigación, toda vez que este involucre el uso de herramienta propia de la matemática.

1.1. Introducción a los análisis QSAR

Un medicamento o fármaco es una molécula que al ser introducida en un sistema biológico produce un efecto terapéutico en el tratamiento clínico de enfermedades o padecimientos específicos, dicho efecto terapéutico es causado por la interacción química entre el fármaco y biomoléculas (Estructuras proteicas, pueden ser intracelulares o no, de alto peso molecular, como receptores celulares, enzimas y diferentes tipos de proteínas entre otros) presentes en los seres vivos. El *sitio activo* de estas biomoléculas, es la región donde ocurre la interacción Fármaco-Receptor (conocida como *ligando-receptor* en el desarrollo de fármacos) y consiste en interacciones químicas débiles entre ambas entidades.

Cuando una molécula arbitraria ς es introducida en un sistema biológico, se llama "blanco biológico" a cualquier subestructura del sistema para la cual no sea nula la probabilidad de interacción química con ς.

El blanco biológico de una molécula puede no existir y cuando lo hace no es necesariamente único, todo dependerá del sistema en el que sea introducido; sin embargo, cuando se encuentra dado el sistema, de forma que ς tenga por lo menos un blanco biológico, y para un observador el interés farmacológico se centre en la interacción entre ς y sólo uno de sus posibles blancos, entonces éste último se conoce como "receptor molecular" o simplemente como "receptor", en tanto que ς se denomina como "ligando".

La afinidad de un ligando con su receptor es la probabilidad de que, al encontrarse espacialmente muy próximos entre si, observen una interacción química. Uno de los requisitos y propiedades de los receptores es la "especificidad", propiedad que les permite distinguir entre una molécula y otra para interactuar de forma selectiva y sólo producir los efectos biológicos esperados.

L y R son usualmente las notaciones empleadas en la literatura para referirse respectivamente a un ligando y su receptor; sin embargo, emplearemos las notaciones $\mathfrak{L}$ y $\mathfrak{R}$, para evitar posibles confusiones con las también comunes notaciones para transformación lineal (L), coeficiente de determinación estadística (R), la relación general entre dos elementos de un conjunto (aRb) y lo que posteriormente se definirá como un sustituyente molecular (R).

Cuando se estudia un receptor molecular en particular y una familia, Ξ, de ligados con alta afinidad y especificidad respecto de dicho receptor, se conoce por "descriptor molecular", simplemente "descriptor" en un contexto adecuado, a toda función que ponga en correspondencia a cada ligando de la dicha familia con un valor real característico del ligando para una propiedad de la familia Ξ. Ejemplos de descriptores moleculares son el peso molecular, la carga promedio, el máximo orden entre los vértices del grafo molecular, etcétera.

Para una pareja dada de ligando y receptor, la "actividad biológica del ligando", o simplemente "actividad" cuando por contexto no exista ambigüedad, es un escalar en función de una característica del ligando, de la que puedan obtenerse mediciones experimentales en el sistema en que se observa su interacción con el receptor.

La diferencia entre un descriptor molecular y la actividad biológica, radica en la dependencia de la interacción entre ligando y receptor, a diferencia de la actividad biológica, un descriptor molecular no depende del receptor, del sistema ni de la cantidad en la que se le observe, depende de una propiedad que sólo tiene sentido en una escala molecular.

Es común que en la literatura sean empleados los términos "actividad " y "respuesta" como sinónimos; no obstante, haremos aquí una distinción entre ellos. De nuevo, para una pareja de ligando y receptor, la "la respuesta biológica causada por el ligando", será una característica biológica cuantitativa de un estado determinado del sistema en el que se observe la interacción entre ligando y receptor, característica definida por una propiedad biológica cuya observación sea de interés para la farmacología. Como antes, toda vez que sea posible el término "respuesta " reemplazará a "respuesta biológica causada por el ligando".

Otro concepto básico del que nos ocuparemos posteriormente es el "efecto terapéutico" ´, por ahora aceptaremos que es un concepto que no necesita definición, que depende

del tiempo, el ligando como medicamento y el cambio de alguna forma observable del estado de salud de un organismo. Aceptamos también, dicho en forma burda, que el efecto terapéutico es "parcialmente comparable" para cualesquiera dos medicamentos que combaten la misma enfermedad mediante los mismos mecanismos; es decir, que partiendo del efecto terapéutico, es posible inducir una relación de orden parcial en el conjunto de todos los medicamentos existentes con tales características, de tal forma que dicha relación de orden se congruente con las nociones médicas y farmacológicas que en la practica dan sentido a la frase "el medicamento a es mejor que el medicamento b".

En adelante, nos referiremos indistintamente a la interacción química entre un ligando y su blanco biológico pro, interacción *ligando-receptor* o nociones relacionadas con la conformación del complejo *ligando-receptor*.

La discriminación que hacemos entre actividad, respuesta y efecto se relaciona con la naturaleza de las propiedades que se observan, la distinción es de carácter epistemológico y en virtud de las propiedades emergentes [1] asociadas con el cambio de escala física e incremento en la complejidad del sistema biológico en el que ocurre la interacción *ligando-receptor*.

El desarrollo de nuevos fármacos está relacionado con sistemas biológicos de una gran complejidad en los que se observa, entre otras cosas, múltiples estructuras y fases, presencia de anticuerpos y mecanismos específicos de transporte. Es por este motivo que en general tener una medición confiable a nivel celular o molecular de la actividad o la respuesta biológica no es una tarea fácil.

Las mediciones experimentales *in vitro* de la actividad biológica (llamadas bioensayos *in vitro*) se llevan a cabo en sistemas biológicos que intentan imitar las condiciones en las que se lleva a cabo la interacción *ligando-receptor*, y en los cuales, se conoce y se puede variar la concentración de cada uno. Se realizan así para poder observar de forma aislada la interacción entre ligando y receptor para la conformación del complejo *ligando-Receptor*, con la mayor cantidad posible de propiedades termodinámicas constantes y descartando la interacción del ligando (compuesto a prueba) con otros blancos biológicos presentes en los seres vivos. El *medio solvente* es una imitación simplificada del medio acuoso de un sistema biológico en el cual se llevan a cabo los procesos químicos (o metabólicos).

Cuando se desea comparar la la actividad, respuesta o efecto de un fármaco respecto de otro (de un ligando respecto de otro), o se desea observar la respuesta de un compuesto que se presenta como un posible nuevo medicamento, las primeras mediciones experimentales se realizan a nivel de laboratorio, en sistemas por mucho más simples que preservan la mayor cantidad de propiedades fundamentales del sistema biológico que realmente importa. Se realizan bioensayos de naturaleza más compleja en la etapa final de la valoración de un compuesto, diseñado *in silico*, para determinar su viabilidad como medicamento, donde el sistema en el que se estudia la interacción del fármaco es un organismo biológico (bioensayos *in vivo*), es justamente en estos experimentos en donde adquiere significado la frase "observación del efecto terapéutico".

La expresión de la actividad puede referirse en términos de si existe o no una respuesta biológica o en términos de la dosis requerida para obtener dicha respuesta. De acuerdo

[1] Para clarificar el significado de "propiedad emergente de un sistema complejo", se sugiere el libro de divulgación [23], lectura asequible y comprensible para un lector con conocimiento básico en matemáticas.

con esto, existen dos tipos de datos biológicos: los que determinan la dosis para una respuesta fija (DRF) o los que determinan la respuesta para una dosis fija (RDF). Los más adecuados considerados en la literatura especializada para establecer las relaciones estructura-actividad son los primeros, entre los cuales se encuentra la concentración de inhibición de la actividad al 50 por ciento, IC_{50}, del inglés 50 % *inhibitory concentration*. Las razones para tal elección son simples, las concentraciones iniciales de los compuestos, principalmente del ligando, siempre son variables de control, a diferencia de la actividad, la respuesta y el efecto, además, por lo regular, con relación a la actividad lo único que experimentalmente puede medires in vitro, es la concentración del ligando que no interactua químicamente con el receptor ni con los organismos suceptibles de este tipo de bioensayo (células).

El valor que toma IC_{50} depende en general de lo que signifique la respuesta en un conjunto de bioensayos, lo cierto es que en la es la forma usual y estandarizada para definir a la actividad.

La metodología precursora de la que aquí estudiamos data de la primera mitad del siglo XX, basada en lo que entonces se definió como "Relación Estructura-Actividad", que se establece cuando un conjunto de propiedades de una serie de compuestos explica su actividad, respuesta o efecto. Para finales de la década de 1950 no existía una distinción entre actividad, respuesta y efecto en las metodologías que estudiaban las relaciones Estructura-Actividad, y es de esa forma que aquí entenderemos a una metodología de ese tipo.

Al realizar bioensayos para cada elemento de una familia de compuestos, en sistemas como los descritos recientemente, las mediciones pueden realizarse garantizando concentraciones constantes del medio solvente y el receptor respecto de la molécula a prueba, bioensayos *in vitro*, la experiencia empírica bajo estas condiciones es que existen medicamentos que requieren una mayor concentración para producir la misma actividad, respuesta efecto que otros.

Si las únicas variantes en las realizaciones de las mediciones son el ligando y la concentración del mismo, entonces es claro que la diferencia entre las mediciones, en estos casos, sólo puede depender de las diferencias entre las características propias de cada una de las moléculas respecto de propiedades moleculares que resulten relevantes para la interacción con el receptor.

La garantía de existencia e identificación de las propiedades que resultan relevantes para la formación del complejo *ligando-receptor* representan ya un problema importante para la farmacología. Puede llegar a ser un dolor de cabeza la simple identificación de una propiedad y cómo es que, individualmente o de forma coordinada con otras propiedades, se relaciona con la respuesta biológica.

En el desarrollo de nuevos fármacos es primordial entender cuál es la relación entre la actividad de un ligando y sus propiedades estructurales (Relación Estructura-Actividad, SAR por sus siglas en ingles *Structure-Activity Relationships*), si se logra comprender esta relación, entones es posible intentar desarrollar *in silico* una molécula con síntesis viable y realizar una predicción del comportamiento que observará la actividad biológica, sin necesidad de una gran inversión en bioensayos con pobres resultados. La aceptación axiomática de la existencia de relaciones Estructura-Actividad es uno de los principios de la medicina convencional, por lo que en realidad no es necesario profundizar en ello.

A principios de la década de 1960 Hansch propuso lo que hoy se conoce como los primeros intentos en el desarrollo de una metodología de análisis de relaciones cuantitativas entre descriptores moleculares y actividad biológica, esta metodología así como los modelos matemáticos que se desprenden de ella se identifican en la literatura por sus siglas en inglés **QSAR** (*Quantitative Structure-Activity Relationships*). Entre las primeras publicaciones en que se postula la existencia este tipo de relaciones cuantitativas destacan las realizadas por A. Crum Brown y T. Frazer, entre 1968 y 1969.

De forma análoga, cualquier metodología en que las relaciones estudiadas por la farmacología sean entre propiedades tanto cualitativas como cuantitativas es referida por las siglas en inglés **SAR**.

En los primeros años los modelos de los análisis que antecedieron a lo que hoy se entiende por un análisis QSAR adolecieron de suficiente formalidad matemática, no se daban definiciones o procedimientos que caracterizaran adecuadamente, tanto a f como a ϕ; entre los avances logrados durante casi cinco décadas de desarrollo se encuentra una forma, suficientemente general, de entender a la respuesta biológica a través de la termodinámica en el caso de bioensayos *in vitro*.

La metodología QSAR consiste, *grosso modo*, en elegir un compuesto probado en la práctica como medicamento (fármaco o estructura base, ***scaffold*** para la literatura en inglés) y considerar conjuntos de moléculas o compuestos orgánicos para las cuales la composición estructural del fármaco base sea una subestructura en la composición de cada una de las moléculas dadas; de forma que, en un sentido conveniente, pueda decirse que las composiciones estructurales de todas ellas son "suficientemente parecidas" entre si, como para que el comportamiento de la respuesta pueda explicarse como un cambio de escala (múltiplo por un escalar) de una función de la actividad.

El principio fundamental en la construcción de un modelo matemático de este tipo de fenómenos biofisicoquímicos es que existen la actividad y respuesta biológica, que la constitución como las propiedades biofisicoquímicas que se observan son mensurables o cuantificables de alguna forma y existe una relación funcional del tipo

$$f(x) = \phi; \tag{1.1}$$

donde para un natural n, x es es un vector en $\mathbb{R}^n$ que almacena las mediciones de la constitución y las propiedades fisicoquímicas del ligando, mientras que ϕ es el valor de medición asociado a la actividad o la respuesta. Se espera que al realizar mediciones experimentales suficientes sea posible identificar y obtener aproximaciones de f.

Hoy día, una gran cantidad de análisis y métodos de la farmacología se concentran en entender y describir lo mejor posible las relaciones cuantitativas entre los descriptores y la actividad o la respuesta de un conjunto de moléculas con un receptor común. Será de nuestra competencia un caso particular y parte del proceso de este tipo de metodología, la revisión del análisis QSAR, el conjunto de herramientas y procesos que comienzan en la selección de un conjunto de descriptores y concluyen en la determinación de un modelo lineal que explica a la actividad como función de tales descriptores moleculares.

La figura 1.1 ilustra a grandes rasgos la estructura cíclica que involucra la totalidad del proceso metodológico QSAR .

En la metodología QSAR se dice que un compuesto orgánico es *candidato a fármaco*

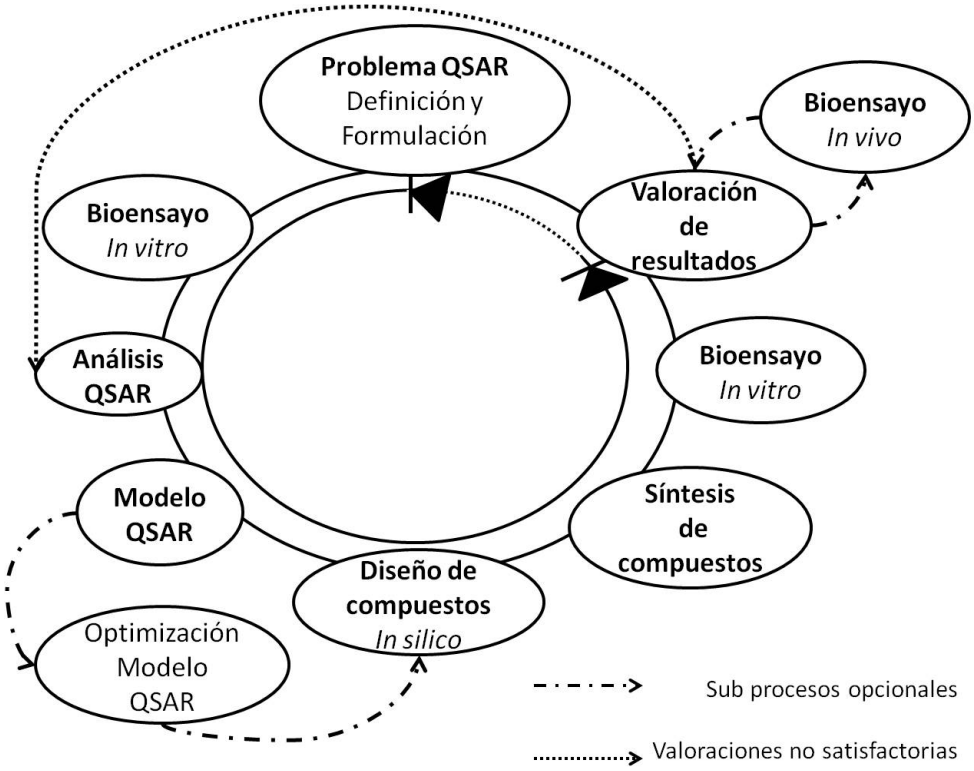

Figura 1.1: Representación esquemática del proceso metodológico QSAR.

o *medicamento* si se tienen pocas dudas de que sea posible que llegue a ser empleado como medicamento, por su alta probabilidad de causar un efecto terapéutico deseado al interactuar con el receptor que se ha elegido para realizar el análisis.

Lo que un análisis QSAR exige como mínimo para trabajar con los resultados de bioensayos de un conjunto muestral de moléculas, es que compartan una subestructura molecular común, que estructuralmente difieran entre sí sólo por uno o por pocos sustituyentes en la estructura base. La razón para ello es garantizar que no se perderá afinidad ni especificidad por parte del receptor con cada miembro de la familia de ligandos. Estas pequeñas diferencias entre ligandos permite inferir que comparten el mismo blanco biomolecular y que es similar su forma de interacción con el mismo.

Un conjunto de compuestos seleccionados de esta forma es lo que por ahora podemos considerar como una familia molecular determinada por el receptor y la estructura base. El desarrollo detallado de lo que entenderemos por familia de moléculas para un análisis QSAR se encuentra contenido en el apéndice A.

Consideramos importante recalcar que, la primara hipótesis básica de un análisis QSAR es, que el efecto la actividad, respuesta y terapéutico o farmacológico que un medicamento causa son de alguna forma características comparables, al igual que lo que se entiende por estructura constitucional de un compuesto químico y que, tanto en el conjunto de todos los posibles compuestos orgánicos como en el conjunto de todos los posibles efectos causados por alguno de ellos, pueden inducirse nociones de proximidad que modelan adecuadamente al fenómeno real.

La premisa restante, no menos importante, es que el efecto farmacológico de un compuesto orgánico se encuentra en función de las propiedades del compuesto, entre las que se cuentan todas las derivadas de su estructura constitucional y propiedades fisicoquímicas, donde, para una familia de compuestos, se espera que pequeñas diferencias en las características de los elementos produzcan diferencias observables en las características cuantitativas relevantes de la actividad o respuesta[2], dicho de otra forma, para cada familia de compuestos orgánicos la función referida es de alguna forma continua.

Los conceptos clave de la metodología QSAR, además de algunos ya comentados, son **actividad**, **eficacia**, **potencia**, **eficiencia**, **respuesta** y **efecto**; conceptos que en ese orden se encadenan para lograr formular la hipótesis de que, para los fármacos es posible establecer una secuencia de cambio de escalas espaciales que explique las relaciones esenciales que son el principio de la medicina convencional, relaciones que explican el cómo es que concentraciones de compuestos, de orden entre pico y micro molar[3], son capaces de producir cambios significativamente distinguibles en una escala tan distinta como lo es aquella en la que se observan los signos y síntomas de una persona afectada por un padecimiento clínico.

En la siguiente sección presentaremos definiciones formales, por ahora solo comentamos al lector que la eficacia puede entenderse como la máxima actividad que puede lograr un fármaco, dentro de un umbral de tolerancia que no resulte en efectos negativos en el organismo final; la potencia se relaciona con la cantidad mínima de fármaco que se

[2] Actividad y respuesta biológica en general no son equivalentes salvo que así se les defina en un caso particular, o cuando como en una interacción molecular se pierde la noción de respuesta.

[3] $1\,\mu M/L = 1 \times 10^{-6} Mol/L$, unidad micromolar;
$1\,pM/L = 1 \times 10^{-12} Mol/L$, unidad picomolar.

requiere para lograr una actividad deseada;la eficiencia, por otro lado, es un concepto de naturaleza cinética, que refleja en algún sentido cuán rápido es un fármaco para lograr una actividad específica.

Con lo dicho hasta ahora, es claro que el objetivo de la metodología QSAR es lograr el diseño *in silico* de candidatos a fármaco para los cuales puedan realizarse predicciones confiables sobre propiedades cuantitativas relevantes para el efecto terapéutico.

Como se comentó en el apartado introductorio de este texto, la metodología QSAR se divide en cuatro etapas básicas: modelación matemática, análisis QSAR, optimización de los modelos QSAR que resulten del análisis y la validación y valoración de de resultados teóricos y experimentales. En el presente capítulo nos concentraremos en explicar la importancia de no trivializar la etapa de modelación, observaciones y propuestas de ligeras modificaciones metodológicas para los análisis QSAR de un tipo particular y, la importancia de un claro vinculo entre estas dos etapas.

1.1.1. Alcances y limitaciones

El conjunto de moléculas con las que en la última parte de este trabajo se realiza un análisis QSAR, con pequeños cambios en la metodología, corresponde a una familia de posibles medicamentos en el tratamiento de la tuberculosis, con la dihidrofolato reductasa (DHFR) por receptor común y estudiadas por Seydel *et ál.* en [6]. El motivo principal es que el receptor es una enzima que ha sido aislada, para la cual es posible considerar una actividad biológica en función de un potencial termodinámico, energía libre de Gibbs, debido a que las interacciones entre un ligando y el receptor pueden ser observadas de forma aislada.

En estos casos la actividad, respuesta y el efecto se relacionan con la inhibición de la función enzimática en el organismo, derivada de la ocupación de los sitios activos en el receptor por el ligando. Nos concentraremos en atender a los análisis QSAR en los que, la actividad se observa en una reacción química con una cinética que se deriva del esquema simple propuesto en 1913 por L. Michaelis y M. L. Menten para una reacción con una enzima por catalizador, que simultáneamente es inhibida por la presencia del ligando.

Los bioensayos

En adelante nos concentraremos en bioensayos *in vitro*, principalmente en aquellos en los que el sistema observado consta sólo del medio solvente $\mathfrak{M}$ (con volumen constante), el receptor $\mathfrak{R}$ (en concentración inicial también constante) y el ligando $\mathfrak{L}$ en concentración inicial como variable controlada. Para estos bioensayos en el medio no existen compuestos o sustancias afines al ligando o receptor. Diremos que el sistema observado en estas pruebas experimentales es del tipo $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$.

Con menor detalle también llegamos a considerar bioensayos en los que, a un sistema $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$ se incorpora una concentración del sustrato natural del receptor, constante respecto de las concentraciones de ligando. Estos bioensayos corresponden a simplificaciones de los sistemas biológicos en los que ocurre la acción enzimática que se desea inhibir. En estos casos el medio debe incluir todos aquellos compuestos que son necesarios para la actividad enzimática, coenzimas por ejemplo. La forma general en que nos

referiremos a los sistemas observados en este tipo de bioensayos es $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R}, \mathfrak{A})$.

Comentaremos también algunas características de bioensayos *in vitro* poco más complejos que los anteriores, en tal caso será sólo como una herramienta para explicar, mediante un ejemplo, la relevancia de la modelación matemática en la primera etapa de la metodología QSAR y su relación con el resto de las etapas restantes, así como con la propuesta de modificación metodológica de los análisis QSAR relativa a la naturaleza de la actividad biológica.

En muchas ocasiones, para el tipo de análisis QSAR que atenderemos aquí, el receptor es una proteína intracelular que no está en contacto con el fármaco sino hasta que éste es capaz de traspasar la membrana celular, reaccionando con el receptor tan rápido y en la mayor cantidad que su afinidad se lo permita, después de lo cual causará un efecto funcional en la célula, dependiendo de la función específica que realice la enzima receptora y el efecto que cause el fármaco. Este escenario es el que en última instancia importa, sobre todo el cambio observable en la actividad o función de la célula. Conviene pues ejemplificar esto.

Detallaremos en extenso tres ejemplos de bioensayos que están estrechamente relacionados, ejemplos en los que nos apoyaremos para mostrar cómo es que la modelación matemática permite simplificar y expresar las relaciones entre observables en escalas distintas de un sistema biológico.

Como usuario final, el efecto terapéutico que se espera de un antibiótico[4] tras las formación del complejo *ligando-receptor* es por ejemplo:

1. Inhibición de la división celular, cuando el antibiótico inhibe el crecimiento de los gérmenes. Antibióticos que observan este efecto se conocen como bacteriostáticos;

2. muerte celular. En este caso se dice que el antibiótico es bactericida;

3. efecto mixto, antibiótico bacteriostático y bactericida.

Dicho en términos más generales, lo que se espera de un antibiótico es que afecte la dinámica de la cantidad de individuos de la bacteria patógena con respecto del tiempo t y en pro de la salud del organismo hospedero.

Por otro lado, es sabido que existen distintos tipos de patógenos bacterianos que, para nuestros fines, pueden diferenciarse en los siguientes aspectos: la forma de interacción entre ligando y patógeno en una escala molécula-célula, si se ven afectados o no por un ligando específico y por la dinámica de sus poblaciones en idénticas condiciones controladas, es decir, por el tipo de comportamiento que puede observarse en un bioensayo *in vitro* en una escala molécula-célula.

Por lo anterior notamos que, desde una perspectiva muy simple y colara sobre los criterios básicos de comparación bajo los que se juzgará si un fármaco es mejor que otro, la perspectiva de un consumidor final (médico o paciente), el tipo de bioensayos en los que estos criterios pueden tener sentido, es cuando el sistema que se observa es uno como el que se observa una interacción aislada entre una concentración de ligando y una población de del patógeno bacteriano.

[4] Las moléculas con las que trabajaremos en el estudio de caso son candidatos a antibióticos.

Hasta aquí, hemos esbozado criterios de comparación de fármacos y un tipo de bioensayo relevante, ahora detallaremos esto como un ejemplo.

Ejemplo 1. *La forma in vitro usual de establecer una medida de la respuesta celular ante la acción del compuesto es colocando en varias cajas de Petri (o matraces) distintas concentraciones del compuesto a prueba en una solución adecuada para la especie del agente patógeno, con una profundidad de acuerdo a los estándares correspondientes, posteriormente en cada una de ellas se inocula idéntica concentración del patógeno, debe existir por lo menos una muestra en la que el compuesto se encuentre en una concentración nula, dicha muestra se conoce como muestra patrón.*

De nuevo, de acuerdo con protocolos y estándares para cada especie de patógeno, se establece un rango de tiempo durante el cual se permite crecer a la población en presencia del ligando; transcurrido dicho intervalo de tiempo se estima el área de la superficie que en cada muestra ocupa la población del patógeno.

En este tipo de pruebas experimentales, respecto del ligando, lo único que puede observarse como medición indirecta de su interacción con el receptor intracelular es la concentración libre, es decir, para este ejemplo, la concentración del ligando que no ha cruzado la membrana celular de alguno de los individuos que conforman la población del patógeno. Para estos bioensayos, se reporta a la actividad del ligando como la ´medición de la concentración libre es una medición indirecta de la cantidad de receptores inhibidos por el ligando en el momento de la medición, que no es necesariamente suficiente para la valoración del ligando como fármaco.

Para antibióticos con pruebas como las anteriores, una muestra de cultivo observa una respuesta biológica al $100\alpha\,\%$*,* $\alpha \in [0,1]$*, cuando el área de la superficie ocupada por el patógeno al cabo del intervalo de tiempo,* A*, corresponde al* $100(1-\alpha)\,\%$ *del área ocupada en la muestra patrón,* A_0*, bajo las mísmas condiciones, es decir, cuando ocurre* $A = (1-\alpha)A_0$*.*

Por comodidad en la notación y convenciones en el uso de porcentajes, suele considerarse a α *en el intervalo* $[0, 100]$*, sustituir la relación* $A = (1-\alpha)A_0$ *por* $A = (1-\frac{\alpha}{100})A_0$ *y decir solamente "respuesta al* α *%". Er rango de* α *quedará usualmente definido aqui por contexto, no siendo en adelante necesario hacer constantes aclaraciones de notación, para los términos y conceptos que se deriven de esta elección de representación notacional.*

Debemos recordar dos aspectos relevantes, que existen errores de medición tanto para las concentraciones de ligando, para el tamaño de la población inicial inoculada y para el área ocupada por el patógeno; además, una bacteria es un ser vivo con mecanismos bioquímicos para la conservación de su propia salud, lo que significa que, para un valor dado de α*, la concentración del ligando que causa una respuesta al* α *% no tiene por qué ser única en general, tampoco tiene por que serlo la concentración libre del ligando en el momento en que se mide la respuesta.*

En este tipo de pruebas es común que se reporten como resultado de las mismas alguno de los siguientes datos experimentales:

1. *Concentración inicial o dosis al 50 %, denotada aquí por* C_{50}*: promedio de un conjunto de valores de las concentraciones iniciales muestrales que, bajo criterios estandarizados, son aquellas para las que se obsrvaron las repuestas más proximas al 50 %,* ***no siempre se reporta****.*

2. *Concentración de inhibición al 50 %, denotado en general en la literatura por IC_{50} (siglas del ingés para 50 % inhibitory concentration): promedio del conjunto de valores muestrales de la concentración libre de ligando en los bioensayos correspondientes al conjunto de valores de la concentración inicial que define a C_{50},* ***siempre se reporta****.*

3. *actividad al 50 %: la evaluación en IC_{50} de una función uno a uno, conocida, incluso puede ser la identidad, es decir, en ocaciones IC_{50} se considera directamente como el valor de actividad del ligando.*

Pese a que la notación no lo exhibe explicitamente, C_{50} e IC_{50} son valores en función del volumen del medio, el ligando y la población inicial de patógeno que es inoculada.

Para que en un análisis QSAR y una familia de ligandos sean comparables los distintos valores de la actividad o la dosis al 50 %, respecto del ligando, debe garantizarse que en lo posile las mediciones de la respuesta en los múltiples bioensayos se realicen en condiciones idénticas, lo qu incluye condiciones de presión, temperatura, equipo de medición, medio solvente, volumen del medio solvente, población inicial de patogeno inoculada, punto o puntos de inhoculación en el medio, intervalos de tiempo entre distintos momentos de interés y demás condiciones experimentales relevantes; incluyendo el día en que se realizan las pruebas y al personal de laboratorio que realice la medición.

Las razones de las exigencias anteriores son en realidad muy simples: el patógeno es un organismo unicelular, escala biológica en que los mecanismos de comunicación genética entre individuos observan comportamientos tanto lineales como horizontales, es fundamental el cuidado que debe tenerse, para garantizar que son lo más parecidas entre si las características biológicas de las distintas poblaciones de patógeno que participan en cada bioensayo.

El tipo de bioensayos como los recién descritos, son hasta cierto comunes en la realización de análisis QSAR, y es también común que para estos análisis se considere sólo un criterio de comparación entre entre fármacos, usualmente es IC_{50}, ya sea por simplificación metodológica o porque es el único valor reportado por los laboratorios que realizan los bioensayos.

Continuando con lo que se puede esperar del efecto terapéutico que se desea observar para un antibiótico, no solo se espera un efecto con respecto a si el medicamento es capaz de erradicar una población bacteriana en un organismo hospedero, la rapidez con la que lo haga también es un factor relevante. En la práctica, de los nuevos medicamentos se espera que actúen lo más rápido y con la mayor contundencia que sea posible, de otra forma se corre un riesgo elevado que el nuevo fármaco tenga un corto periodo de vida útil como medicamento, derivado del desarrollo de mecanismos más eficientes de respuesta al fármaco por parte de los patógenos que se ven afectados por el mismo. En el caso de enfermedades contagiosas, un fármaco capaz de producir un efecto terapéutico en un paciente enfermo no garantiza contribuir en la reducción de contagios si demora en demasiado en producir el efecto esperado.

En general, la velocidad con la que un medicamento actúa es tan relevante como la potencia con que lo hace. Significa que el comportamiento de la velocidad con la que un fármaco actúa es también un criterio de comparación entre medicamentos, criterio

que, de acuerdo con las necesidades específicas de atención de una enfermedad, debe considerarse conjuntamente con la potencia del fármaco.

Desde ahora, ya debería existir sospecha sobre los alcances de resultados la capacidad predictiva de modelos QSAR que se desprenden de análisis de la misma metodología en los que sólo se considera un valor único tipo de actividad para ajustar parámetros de un modelo lineal, trataremos de explicarnos más claramente.

Hemos aceptado que, por lo menos en bioensayos como los descritos en el ejemplo 1, la respuesta biológica depende principalmente de la inhibición del receptor intracelular causada por el ligando, es decir, de la cantidad de receptores con sitio activo ocupado por una molécula de ligando, en el interior de cada individuo por supuesto, y la forma en que esa ocupación por individuo se refleja en la dinámica de la población.

Para un conjunto de bioensayos realizados para una familia de ligandos, Ξ, de forma que en cada prueba experimental las únicas variables sean el ligando, como elemento de Ξ, y la dosis o concentración inicial del ligando, aceptamos que la respuesta en cada una de las muestras depende de dichas variables del ligando y la dosis del mismo.

Vamos a suponer también, de forma teórica, que en idénticas condiciones experimentales la respuesta es una función bien definida respecto del estado inicial de observación, es decir, no es posible obtener dos valores distintos de la respuesta biológica para cualesquiera dos bioensayos distintos pero realizados bajo condiciones idénticas. En particular, la respuesta será, con esta hipótesis, una función bien definida respecto del ligando, la dosis y el tiempo de observación como variables.

A partir de aquí comenzaremos a utilizar un lenguaje más propio de la matemática, para clarificar la intención y el desarrollo de las distintas secciones de este texto.

Continuando en el marco de referencia del ejemplo 1 y recuperando la esencia de la ecuación (1.1), Ξ denotará como hasta ahora y en adelante a una familia de ligandos a estudiar mediante la metodología QSAR, $\mathfrak{L}$ un elemento de Ξ, D a la dosis o concentración inicial de ligando, P a la población del patógeno respecto del tiempo (P) y t denotará la variable temporal (t_0 el instante inicial de observación, t_1 el instante de medición de la respuesta y $\Delta t = t_1 - t_0$).

Si ϕ denota en general a la función de actividad o de la respuesta de interés para un análisis QSAR, para ser consistentes con la noción de inhibición de la respuesta, entonces para un bioensayo definimos la respuesta observada como $1 - \frac{A}{A_0}$; a mayor área ocupada por el patógeno menor la inhibición del crecimiento de la población, es decir, $\phi = \alpha$, una respuesta ϕ será en estos casos equivalente a una inhibición inhibición de la respuesta al $\phi\,\%$. Sustituyendo a ϕ por $1 - A/A_0$ y a x por $(x, P_0, D, \Delta t)$ en (1.1), entonces se tiene la siguiente relación:

$$f(x, P_0, D, \Delta t) = 1 - \frac{A}{A_0}. \tag{1.2}$$

Recordamos al lector con formación no matemática, que si f es una función, no necesariamente uno a uno y sobreyectiva, la imagen inversa de f para el valor ϕ, $f^{-1}(\phi)$, es el conjunto de elementos en el dominio de f que son transformados en ϕ bajo la acción de f, en otras palabras, $f^{-1}(\phi)$ es el conjunto de todos los x para los que se satisface la ecuación $f(x) = \phi$. La imagen inversa en general no tiene por que ser un valor único y tampoco tiene que existir, el ejemplo clásico de ello es la función real de variable real

definida por la expresión $f(x) = x^2$ ($f^{-1}(1) = -1, 1$, $f^{-1}(-1) = \emptyset$). Solamente cuando la función es biyectiva, sólo entonces f^{-1} denota a la función inversa de f.

La primera observación es que tanto C_{50} como IC_{50} son variables no controladas que no necesariamente pueden considerarse como mediciones equivalentes como criterios de comparación en la elección de los "mejores" fármacos en una familia de ellos. Un análisis de actividad QSAR en el que no se incluyan parejas de valores IC_{50} y C_{50} es fácilmente cuestionable sobre la confianza que puede depositarse en sus criterios de comparación y optimización de fármacos probados en este tipo de bioensayos.

Para explicar lo anterior, pedimos al lector que se permita el lujo de no pensar en las limitaciones tecnológicas o los millones de posibles variables biofisicoquímicas que pueden citarse para descartar un posible escenario idóneo como el que vamos a presentar. Lo que haremos, es ubicarnos en un experimento idealizado que será el mejor de los casos en los que podríamos situarnos para una análisis QSAR, para evidenciar que aún en tales condiciones, hay razones suficientes para sospechar de modelos QSAR que sustenten todo su desarrollo en reportes incompletos de laboratorio.

En el tipo de bioensayos expuestos en el ejemplo 1 supondremos momentáneamente que la función de actividad se comporta bastante bien, al grado en que la concentración libre del ligando en el momento de realizar la medición experimental de la respuesta, $Cl_{\mathfrak{L}}$, es inversamente proporcional a la dosis $D_{\mathfrak{L}}$, es decir, para cada ligando, $\mathfrak{L}$, en una familia estudiada, existe un valor real $k_{\mathfrak{L}}$ mayor que 0 y menor que 1, tal que:

$$k_{\mathfrak{L}} D_{\mathfrak{L}} = Cl_{\mathfrak{L}}.$$

Supondremos también, que no se comete error de medición alguno y que en el conjunto conjunto de bioensayos, se observó para cada ligando una muestra en la que la dosis inicial logró una exacta inhibición de la actividad al 50 %. Se infiere que $k_{\mathfrak{L}}$ corresponde a la fracción de la dosis inicial que observa una interacción con la población del patógeno en el instante en que se mide la respuesta, como $k_{\mathfrak{L}}$ es constante para cada ligando, significa que hemos aceptado también que dosis distintas producen valores distintos de IC_{50}, y que para cada valor del mismo existe un único valor de C_{50} para el que, como funciones del ligando, IC_{50} y C_{50} se relacionan como sigue:

$$k_{\mathfrak{L}} C_{50}(\mathfrak{L}) = IC_{50}(\mathfrak{L}).$$

Ahora, como en la práctica las familias de ligandos son finitas, desde que es finita la cantidad de elementos químicos conocidos y las moléculas no tienen pesos infinitos; entonces, podemos construir un mejor escenario todavía, uno en el que la familia estudiada tiene exactamente m miembros y que como función del ligando, el vector x que se observa en la ecuación (1.1) es distinto par cada miembro de la familia de ligandos.

Lo dicho anteriormete puede ser mejorado aún, nos permitimos suponer además, por decirlo de algún modo, que la respuesta que se observa en el momento de la medición no tiene memoria de todo el proceso biológico, que depende única y proporcionalmente de la cantidad de moléculas del receptor que, en el momento de la observación, se encuentran en interacción química con el ligando, en el interior de alguna de las células que conforman la población del patógeno.

En este punto punto aún tenemos una situación biológica que nos incomoda pero que aún podemos despreciar. Ocurre que experimentalmente sólo puede medirse la cantidad de ligando en el exterior de la célula, pero eso no significa que la cantidad de restante de ligando se encuentre en el interior de la célula, todo depende de los mecanismos de absorción y expulsión de la membrana celular. Sin perder la esenia del fenómeno, vamos a suponer que membrana tiene grosor nulo, es decir, que si una molécula del ligando no se encuentra en el exterior de todas las moléculas de la población del patógeno, entonces se encuentra en el interior de alguna de ellas, sin puntos medios.

Pues bien, con lo dicho hasta aquí, la forma de expresarlo en una ecuación es como sigue.

Si $\{\mathfrak{L}_1, ..., \mathfrak{L}_m\}$ es la familia completa de ligandos a estudiar[5], denotaremos por x_i al vector que alberga las características cuantitativas relevantes del ligando $\mathfrak{L}_i$, $i = 1, ..., m$ y por ϕ_i a la respectiva medición de la respuesta. Nuestras hipótesis para el caso ideal dicen que existe un real λ que no depende de la dosis ni el ligando, tal que la ecuación (1.1) se reescribe como:

$$\begin{aligned}(1 - k_{\mathfrak{L}_i})\lambda C_{50} &= \lambda(C_{50}(\mathfrak{L}_i) - IC50(\mathfrak{L}_i)) \\ &= \phi_i \\ &= f(x_i),\ i = 1, ..., m.\end{aligned}$$

Obviando la dependencia del ligando,

$$\lambda(1 - k)C_{50} = \phi. \tag{1.3}$$

Pues bien, evidente en este caso, la ecuación (1.3), aún si es conocido ϕ y C_{50}, tiene dos grados de libertad correspondientes a λ y k, donde k es dependiente de de IC_{50} y C_{50}. Lo que significa, es que si queremos identificar de forma única a λ, aún en el mejor de los casos se requiere de dos mediciones experimentales para hacerlo.

Mas aún y con toda franqueza, conocer el valor de IC_{50} puede considerarse incluso irrelevante si no se conocen C_50 o una condición equivalente e independiente de IC_50. Aún en el mejor de los casos no parece tener mucho sentido comparar entre dos fármacos para saber cuál producirá una mayor inhibición de la respuesta si el único parámetro de comparación es equivalente a comparar considerando sólo a IC_{50} como valor de actividad.

Resumiendo un poco, un valor pequeño de IC_{50} hablará en estos caso de un buen fármaco sólo si observa también un muy pequeño de C_{50} de otro modo, significa que la velocidad promedio de acción del fármaco al 50 por ciento requiera de una pequeña cantidad de medicamento, es decir, el medicamento es más potente en este caso. Por otro lado, aún en el marco de los antibióticos y principalmente de antituberculosos, un valor elevado de C_{50}, para un valor pequeño de IC_{50}, habla de un antibiótico no deseable, requiere mayores cantidades de medicamento para producir la misma velocidad promedio de inhibición al 50 % del crecimiento poblacional del patógeno.

En le transcurrir de estas páginas se verá poco a poco que incluso en una escala molecular la situación es la misma para el caso idóneo en análisis donde los bioensayos

[5] m es un natural que se considera tan grande como sea necesario.

in vitro se realizan para sistemas en los que puede observarse una interacción del tipo molécula-molécula. El argumento es muy simple, en estos análisis el receptor es una macro molécula en comparación con el ligando, la modulación e idealización del problema es equivalente cuando no necesariamente se acepta que el el sitio activo en el receptor es único. Luego, el caso que consideraremos, aunque se detallará con más cuidado, es un caso particular.

Análisis QSAR como un problema inverso: modelación matemática, clasificación de familias moleculares y necesidad de regularización.

En esta sección continuaremos explotando al ejemplo 1 y la ecuación (1.2) como parte de la revisión metodológica QSAR, evidenciando problemas sobre el posible mal uso de las herramientas estadísticas y aquellas que en general la matemática propone con total conocimiento de sus limitaciones.

Ya hemos supuesto que para cada cuarteta $(\mathfrak{L}, D, \Delta t)$ existe un único valor de f, dado que A_0 es constante, significa que también A está bien definida como función; sin embargo, en ningún momento hemos dicho algo acerca del posible comportamiento del área como función de la dosis y/o el tiempo, toda vez que se fije al ligando $\mathfrak{L}$.

En la práctica no se conocen los valores exactos del tiempo, de la dosis, del área y en general tampoco de los descriptores moleculares. Si tildamos cada variable para la que es casi seguro que sólo puede conocerse con errores de medición, una formulación de (1.2) que refleje de mejor forma la situación real de los análisis QSAR es como sigue:

$$\begin{aligned} f(x, P_0, D, \Delta t) &= 1 - \frac{A}{A_0} \\ f(x, P_0, D, \Delta t) &\approx 1 - \frac{\tilde{A}}{A_0} \\ f(\tilde{x}, \tilde{P}_0, \tilde{D}, \tilde{\Delta t}) &\approx 1 - \frac{\tilde{A}}{\tilde{A}_0} - \epsilon; \end{aligned} \tag{1.4}$$

donde

$$\epsilon = f(x, P, D, \Delta t) - f(\tilde{x}, \tilde{P}, \tilde{D}, \tilde{\Delta t}).$$

Hasta ahora, restringido al ejemplo que estamos desarrollando, de la relación en (1.2) se sigue que entre las hipótesis fundamentales de un análisis QSAR, se supone que con suficientes mediciones experimentales es posible identificar plenamente, o al menos suficientemente a f, es decir, el problema se centra en cómo aproximar al funcional f (*ver* sec.1.3 y [1]), si son conocidas las relaciones:

$$\begin{aligned} f(x, P_0, D, \Delta t) &= \phi \\ f(x, P_0, D, \Delta t) &\approx \tilde{\phi} \\ f(\tilde{x}, \tilde{P}_0, \tilde{D}, \tilde{\Delta t}) &\approx \phi + \delta\phi,\ \delta\phi = \phi - \tilde{\phi}; \\ f(x_i, P_0, \tilde{D_{\xi,i}}, \Delta t_\xi) &\approx \tilde{\phi_{\xi,i}}; \end{aligned} \tag{1.5}$$

donde ξ es elemento de índices que indican la cantidad de bioensayos realizados para cada ligando con el propósito de lograr reportar IC_50 y C_50.

Una primera limitarte, que no podemos dejar pasar, es que entre las dificultades que plantea la metodología QSAR es la cantidad de información de la que puede disponerse, realizar grandes cantidades de bioensayos para una misma familia de compuestos, es algo que en el diario acontecer de la farmacología no se observa con la frecuencia con que nos gustaría, y en muchos casos, ni siquiera con la que se necesitaría para utilizar con seguridad herramientas estadísticas como lo es la regresión lineal múltiple, herramienta por demás utilizada en los análisis QSAR.

El problema QSAR como se plantea en (1.5) es lo que se conoce como un problema inverso, con perturbaciones tanto en los datos de entrada como en los de salida.

Como hemos esbozado antes, si bien no exigimos un escenario idóneo, lo mínimo que necesitamos para poder utilizar con tranquilidad la mayoría de métodos estadísticos y herramientas del análisis con la poca información experimental que puede obtenerse es que, el problema QSAR sea una ejemplo de lo que se conoce como un **problema inverso bien planteado**, es decir:

1. Además de estar bien definida, f debe ser una relación uno a uno, de forma que se posible restringir su dominio de una forma adecuada y garantizar que existe su función inversa f_{-1};

2. el error de aproximación debe depender de forma continua del error de medición en los datos, es decir, que f^{-1} se una función continua. Este requisito sólo dice que errores pequeños de medición devuelven errores pequeños de aproximación. Es incluso preferible si es derivable.

La continuidad de f^{-1} se requiere por lo siguiente. Pensemos en ϕ como una función que conocemos explícitamente y cuyo dominio es la familia Ξ como un conjunto arbitrario de objetos, esto es algo que si ocurre, simplemente porque existe un protocolo que nos indica cómo es que deben ser realizadas las mediciones de actividad.

Luego vamos a considerar que por la delicadeza del proceso se han extremado precauciones, logrando que por

1. forma en que está definida ϕ,

2. la forma en que desde la farmacología y la matemática se define a la familia Ξ (*ver apéndice* A),

3. y el cuidado de los especialistas en la elección del conjunto de moléculas que se envían a un laboratorio para realizar los bioensayos correspondientes;

entonces, si se se entrega una medición experimental, con poca dificultad puede conocerse de que elemento en la de moléculas sometidas a bioensayos, dicho en otras palabras, que si Ξ_0 es el conjunto muestral de elementos de Ξ, restringiendo el dominio de ϕ a Ξ_0 existe y es conocida función inversa ϕ^{-1}.

Por otro lado, el hecho de que en el lenguaje técnico se reporten valores experimentales que se digan suficientemente pequeños como para que las mediciones sean significativamente distinguibles, no habla de que todo el proceso se lleva a cabo de forma que como

función ϕ^{-1} también existe en proximidades adecuadas de cada uno de los elementos en el conjunto Ξ_0, en "vecindades suficientemente pequeñas".

Lo que acabamos de pedir es razonable no solo desde el punto de vista de la matemática, es en realidad una formulación de los cuidados que Scior T. enfatiza en [24] y que se encuentran en las diversas publicaciones descriptivas de la metodología QSAR.

Después, si δx modela el error de medición cometido para las características cuantitativas de un ligando y $\delta\phi$ de forma análoga para la medición de la respuesta, para un error $\delta\phi$ suficientemente pequeño se cumple que

$$f(x) \approx \phi + \delta\phi$$

implica que, pensando en la continuidad usual en los espacios $\mathbb{R}^n$, existe un real positivo δ para el que se satisface, por la continuidad de f^{-1} en un problema bien planteado:

$$f^{-1}(\phi + \delta\phi) - \delta < x < f^{-1}(\phi + \delta\phi) - \delta;$$

donde δ se tiende a 0 cuando el error $\delta\phi$ lo hace.

Que se pida la continuidad de la función inversa de f en un problema bien planteado sólo quiere decir que, por lo menos localmente, para los elementos de la familia muestral, podemos recuperar aproximaciones tan buenas como se quiera de f^{-1}, siempre que se garantice que el error de medición se pueda hacer tan pequeño como sea necesario.

Se sabe que bajo las condiciones pedidas en un problema bien planteado se obliga a f a ser continua, así, un problema bien planteado nos dice que locamente, en vecindades de la de las moléculas en la muestra, podemos recuperar a f^{-1}; conociendo esto, existe ya muchísimo trabajo hecho en matemáticas para recuperar localmente a f siendo su función inversa conocida. Ejemplos claros y accesibles son los resultados conocidos como "teorema de la función implícita" y "teorema de la función inversa" (*ver* [1]).

Tristemente, como explica Hansen con más detalle en [10], un problema inverso bien planteado es difícil de encontrar. Incluso si el problema es teóricamente bien planteado, ocurre que para fines prácticos no lo es desde el punto de vista de las aproximaciones numérica, donde no siempre es posible hacer tan pequeño como se desee el error de medición. El caso de los análisis QSAR es un ejemplo de ello.

Ahora, es posible abordar el problema de distints formas, lo ideal sería disponer de un modelo global, dependiente de una mínima cantidad de parámetros que describa la relación cuantitativa entre descriptores y actividad o respuesta, pero no es en general el caso. La forma en que trabajaremos aquí es como usualmente se hace, de forma local, para dominios de diámetro relativamente pequeño sobre los cuales sea aceptable hacer aproximaciones lineales, por simplicidad.

Nuestro problema ahora, es que particularmente para los análisis QSAR, disponer cuantiosos conjuntos maestrales de datos no es algo que ocurra a menudo, entre otras cosas porque no la síntesis y los bioensayos suelen tener costos elevados.

En el ejemplo la actividad puede entenderse como una medida en un sentido formal o como una medición (valor real resultado de un bioensayo) relacionada con la afinidad, con la cinética de la conformación y disociación del compuesto *ligando-receptor*. Lo usual es considerar funciones del cociente *velocidad de conformación/velocidad de disociación*, constante de proporcionalidad que se conoce como constante de equilibrio de la reacción.

Pasando a la eficacia del compuesto, para el ejemplo, es una forma de medir la capacidad del ligando de producir una diferencia entre las propiedades fisicoquímicas del receptor y el complejo *ligando-receptor* que resulte relevante y se refleje en la respuesta biológica. Para los antibióticos está relacionada con la forma de medir individualmente la inhibición de las funciones vitales de la célula, cuando se entiende que los receptores en ella están en constante interacción con moléculas de ligando.

La eficacia, explica Pazos en [22], tiene dos enfoques, el más simple es el *ocupacional*, establece que la eficacia intrínseca de un ligando es una propiedad común y cuantificable de las moléculas del ligando que puede ser caracterizada por una constante, ϵ, y tal que si Act es la actividad del ligando, entonces la potencia α se encuentra en función del producto ϵAct, en otras palabras, la potencia observada de la inhibición causada por el fármaco se encuentra en función de un valor proporcional a la actividad. La eficacia es desde este enfoque una función real dependiente del ligando, por lo que es comparable en el sentido usual en el conjunto de reales.

El otro enfoque de la eficacia corresponde al enfoque *operacional*, que de forma muy general explicaremos como una variación del enfoque ocupacional en el que la ϵ no es un escalar sino un objeto bien definido en un conjunto totalmente ordenado, como función del ligando y la actividad como función de medida, la modificación que se hace en relación a la potencia es que ahora estará en función de la el ligando, le actividad y la eficiencia. Si desde este enfoque la eficacia entre dos ligados es comparable en algún sentido dependerá del conjunto en el que se le encuentre y el orden definido en éste.

Nuestro trabajo se ciñe principalmente a la actividad de los ligandos, la eficacia y la respuesta celular son temas que no abordaremos, supondremos como caso ideal que son conocidas. La respuesta se entenderá como un tipo particular de actividad comprendido en la definición 1.2.2.

1.1.2. Sobre las reacciones químicas

Las reacciones químicas en condiciones controladas se entienden también como un sistema termodinámico, entre los primeros requisitos que debemos cubrir para dar sentido a nuestro trabajo, es considerar sistemas termodinámicos que resulten de incorporar una concentración del *sustrato natural* o *ligando endógeno* del receptor en un sistema $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$, sistemas $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R}, \mathfrak{A})$.

El ligando endógeno es un compuesto presente en el organismo con alta afinidad con el receptor, la interacción entre ligando endógeno y receptor debe ser una pieza fundamental de alguna función celular. El receptor será una enzima especializada, por lo regular el catalizador de la reacción química mediante la que se transforma el sustrato natural en un producto vital para la célula.

La formas principales reacciones químicas que estudiaremos tendrán por tanto ecuaciones químicas de la forma:

$$\mathfrak{L} + \mathfrak{R} \underset{k_2}{\overset{k_1}{\rightleftharpoons}} \mathfrak{LR}; \tag{1.6}$$

$$\mathfrak{A} + \mathfrak{R} \underset{k_4}{\overset{k_3}{\rightleftharpoons}} \mathfrak{AR} \overset{k_5}{\rightarrow} \mathfrak{R} + \mathfrak{P} \tag{1.7}$$

$$\mathfrak{P} \overset{k_6}{\rightarrow} \mathfrak{A}. \tag{1.8}$$

En una reacción química en la que intervengan los compuestos $C_1, ..., C_m, C_{k+1}, ..., C_n$, de forma que se corresponda con la ecuación química:

$$C_1 + \cdots + C_m \underset{k_d}{\overset{k_c}{\rightleftharpoons}} C_{m+1} + \cdots C_n;$$

se conoce a las constantes k_c y k_d como "constante de conformación" y "constante de disociación" de la reacción, respectivamente.

En las ecuaciones químicas (1.6)-(1.8), las constantes k_1, k_3, k_5 y k_6 son constantes de conformación de la correspondiente reacción; mientras que k_2 y k_4 son constantes de disociación. Cuando en una sección o bloque trabajemos con sólo con una ecuación con la forma de (1.6) y por contexto no existe lugar a confusión, emplearemos las notaciones k_c y k_d de forma sistemática.

Las ecuaciones (1.7)-(1.8) se estudiarán de forma independiente y acoplada según convenga. Partimos de suponer que el receptor ha sido aislado con éxito, permitiendo observar el tipo de reacciones que describimos.

En general utilizaremos la notación $[\cdot]_0$ para referirnos a la concentración inicial del compuesto o complejo químico que se encuentre se coloque entre los corchetes. Como ya se mencionó, consideraremos sólo bioensayos en los que $[\mathfrak{LR}]_0 = [\mathfrak{AR}] = [\mathfrak{P}] = 0$.

Respecto al tipo de la función de actividad biológica (*ver detalles de formalidad en la sección 1.2*), será la concentración libre de $\mathfrak{L}$ en el momento t, que se entiende como la cantidad de moles por unidad de volumen de ligando que no participan en la formación del complejo $\mathfrak{LR}$, simplemente una forma de referirse a $[\mathfrak{L}]$ cuando se requiera enfatizar su relevancia, la concentración activa del ligando es simplemente $[\mathfrak{L}]_0 - [\mathfrak{L}]$.

La función de respuesta será para este tipo de sistemas la concentración del complejo $\mathfrak{LR}$, $[\mathfrak{LR}]$, Cuando el sitio activo en el receptor sea único, ocurre que la función de respuesta coincide con la concentración activa de ligando, no se pierda de vista que tal afirmación no es verdadera si un receptor puede observar una reacción química con dos o mas moléculas de ligando.

Cuando en el sistema de interés sea posible que en algún momento no sea nula la concentración del producto de la reacción enzimática ($\mathfrak{P}$ en la ecuación (1.8)), entonces el efecto se entenderá como el comportamiento de dicha concentración respecto de cualquier

variable temporal, física, química o biológica que observe cambios respecto de variaciones en la concentración de ligando, $[\mathfrak{L}]$.

Los bioensayos se realizan garantizando condiciones que permitan el completo estudio de la reacción, entre las que se incluye que se verifique la **ley de acción de masas**:

> *La velocidad a la que ocurre una reacción química en el instante de tiempo t es directamente proporcional al producto algebraico de potencias de las sustancias reaccionantes en ese instante.*

Y también el **principio de independencia**:

> *Si en un sistema transcurren simultáneamente varias reacciones químicas, entonces cada una de ellas es independiente del resto y su velocidad se evalúa usando la Ley de Acción de Masas*

La unidad de medida para la concentración de una sustancia es **mol por unidad de volumen** y todas las potencias referidas por la ley de acción de masas serán en este caso idénticamente 1.

La forma convencional para denotar concentraciones molares es mediante corchetes cuadrados, $[\mathfrak{L}_t]$ denota la concentración de ligando en el sistema en el instante t, por simplicidad de notación se omitirá la dependencia del tiempo en la notación, como ya lo hemos hecho en párrafos anteriores, salvo que sea necesaria.

Hemos aceptado sitio activo único por cada receptor, entonces en cualquier instante de tiempo se satisfacen las igualdades:

$$\begin{aligned} [\mathfrak{L}\mathfrak{R}] &= [\mathfrak{L}]_0 - [\mathfrak{L}]\,; \\ [\mathfrak{A}\mathfrak{R}] &= [\mathfrak{A}]_0 - [\mathfrak{A}] - [\mathfrak{P}]\,; \\ [\mathfrak{R}] &= [\mathfrak{R}]_0 - [\mathfrak{L}]_0 - [\mathfrak{A}]_0 + [\mathfrak{L}] + [\mathfrak{A}] + [\mathfrak{P}]\,. \end{aligned} \tag{1.9}$$

Que las concentraciones de tres de los seis compuestos involucrados en las reacciones se expresen como combinaciones lineales de los tres concentraciones restantes, aunado a una formulación de la ley de acción de masas y el principio de independencia para las reacciones en las ecuaciones (1.7)-(1.8), desde un enfoque cinético permite describir el comportamiento de las concentraciones de los compuestos en cada instante mediante un sistema de ecuaciones diferenciales (*ver* [7]):

$$\begin{aligned} \frac{d\,[\mathfrak{L}]}{dt} &= k_1([\mathfrak{L}]_0 - [\mathfrak{L}]) - \\ &\quad k_2\,[\mathfrak{L}]\,([\mathfrak{R}]_0 - [\mathfrak{L}]_0 - [\mathfrak{A}]_0 + [\mathfrak{L}] + [\mathfrak{A}] + [\mathfrak{P}]); \end{aligned} \tag{1.10}$$

$$\begin{aligned} \frac{d\,[\mathfrak{A}]}{dt} &= k_3([\mathfrak{A}]_0 - [\mathfrak{A}] - [\mathfrak{P}]) + k_4\,[\mathfrak{P}] \\ &\quad -k_5\,[\mathfrak{A}]\,([\mathfrak{R}]_0 - [\mathfrak{L}]_0 - [\mathfrak{A}]_0 + [\mathfrak{L}] + [\mathfrak{A}] + [\mathfrak{P}]); \end{aligned} \tag{1.11}$$

$$\frac{d\,[\mathfrak{P}]}{dt} = k_6([\mathfrak{A}]_0 - [\mathfrak{A}] - [\mathfrak{P}]) - k_4\,[\mathfrak{P}]\,. \tag{1.12}$$

El sistema de ecuaciones (1.10)-(1.12) describe cualquiera de las ecuaciones (1.6)-(1.8) de forma aislada o acoplada, basta con hacer 0 las variables y parámetros adecuados.

Se sabe de la química y la matemática que los sistemas descritos tienden a un equilibrio químico, a un estado en el que las concentraciones de cada uno de los compuestos que intervienen permanecen constantes, los tipos de actividad y respuesta para los análisis QSAR dependerán en estos casos de dicho estado de equilibrio, denotaremos por $[\cdot]_{eq}$ a la concentración del compuesto químico entre corchetes en el estado de equ8ilibrio del sistema en el que se le observe.

De forma análoga a lo expuesto en relación al ejemplo 1, para bioensayos en los que se observe un sistema $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$, diremos que un ligando observa una inhibición de la respuesta al $100\alpha\,\%$, α en el intervalo $[0, 1]$, si se verifica la relación:

$$\alpha = 1 - \frac{[\mathfrak{R}]_{eq}}{[\mathfrak{R}]_0}.$$

Para estos bioensayos, IC_α denotará a la mínima concentración libre de ligando que se observa en un estado de equilibrio caracterizado por una inhibición de la respuesta al $100\alpha\,\%$; siendo C_α la dosis o conccentración inicial de ligando para la cuál se observa una inhibición de la respuesta al $100\alpha\,\%$ (*definición* 1.2.5).

Nuestro objetivo es hacer una revisión de la obtención de parámetros para modelos QSAR, las soluciones de los sistemas de ecuaciones diferenciales tampoco se abordará, figuran en estas páginas porque los parámetros o conjuntos de ellos que caracterizan a estos sistemas deben ser el resultado de un análisis QSAR, estamos, por el momento, más interesados entre la relación que guardan los parámetros que definen un modelo QSAR que en las soluciones de las ecuaciones diferenciales. Una argumentación más detallada será presentada en cuanto se presenten las definiciones y acuerdos de notación suficientes.

1.2. Formulación matemática

El objetivo de esta sección y el texto en general no es un exhaustivo desarrollo ni revisión de la teoría termodinámica. Presentaremos al lector las definiciones conceptuales y de notación necesarias para la lectura de este trabajo. Los detalles teóricos referentes a los objetos y conceptos físicos pueden consultarse en cualquier texto pensado para la enseñanza de la termodinámica, por ejemplo [11].

Nuestra intención en esta sección lograr transmitir al lector con formación en biología o química el objetivo y la relevancia y profundidad, conceptual como metodológica, que involucra una análisis QSAR, incluso cuando se suponen relaciones lineales entre valores relevantes y se emplean herramientas de ajuste de parámetros como regresión lineal múltiple.

Queremos mostrar la importancia de conocer cómo es que el poco cuidado con el que se manejan herramientas matemáticas en una investigación puede implicar fallos graves tanto en los resultados numéricos como en la lectura de los mismos. También queremos mostrar cómo es que las propias hipótesis de una metodología como QSAR conllevan algunas definiciones naturales de actividad.

Haremos uso, sin demostración ni enunciación, de resultados sumamente conocidos del álgebra lineal en dimensión no finita, geometría y topología. Nos basaremos en definiciones y nociones en la mayoría de los casos de carácter elemental, se recomienda al lector el libro de [16] para definiciones de espacio vectorial topológico en el sentido en que lo utilizaremos aquí, cualquier libro sobre espacios de Hilbert es también una alternativa.

Como usualmente se hace en la literatura matemática, $(\mathbb{R}^n, B(\mathbb{R}^n))$ denotará al espacio mensurable que resulta de dotar al espacio euclidiano $\mathbb{R}^n$ de la σ-álgebra de Borel, [3]. Para este trabajo un sistema termodinámico $(Esp, B(Esp))$ formalmente se entenderá como un subespacio de $(\mathbb{R}^n, B(\mathbb{R}^n))$, en el que el espacio de estados Esp es una región acotada de $\mathbb{R}^n$.

Por la naturaleza de los análisis QSAR, no es difícil argumentar que cualquier descriptor molecular es un funcional acotado, con un espacio normado de dimensión finita sobre el campo de los racionales y en el cual se satisface la desigualdad del paralelogramo. Un elemento de un espacio de Hilbert incluso ($L^p(\Omega)$, Ω una región acotada de $\mathbb{R}^n$).

Lo anterior se justifica en tanto, independientemente del descriptor que sea considerado como función definida en el espacio físico, lo que aceptamos dicho espacio, al menos en el que nos movemos cotidianamente, es que puede ser modelado geométricamente con los primeros cuatro postulados de Euclides, mismos que Hilbert expresó como sistema axiomático en alrededor de la primera mitad del siglo pasado. Es decir, es un espacio donde existen las nociones de punto, recta, congruencia, semejanza y ángulo.

Además de las nociones geométricas todos habitualmente recurrimos a nociones de continuidad. Se sabe de los cursos de geometría que como sistemas axiomáticos, la geometría euclidiana y la de Riemman son sistemas categóricos, es decir, que es suficiente con estudiar uno de sus modelos para entender lo que ocurre en cualquier otro modelo de la misma geometría.

Así, cualquier característica es una función del tipo *conjunto-punto*, que pone en correspondencia a conjuntos de un espacio normado de dimensión finita con un conjunto arbitrario al que se conoce como conjunto de características, que en el caso de los análisis QSAR, es el campo sobre el que se define el espacio físico como espacio vectorial. En el mismo sentido físico, los descriptores son funcionales acotados.

En este trabajo diremos que una función de respuesta es una función de actividad donde el sistema que la caracteriza corresponde a aquel en que se realizan un bioensayo, *in vitro* o *in vivo*, en los que el observable de interés corresponde a una característica biológica del sistema. Desde este punto de vista actividad y respuesta parten de una definición común pero son conceptualmente ajenas. Definidas así, las funciones de "actividad" como las de "respuesta", son también funcionales definidas sobre el mismo espacio físico. El tiempo siempre puede considerarse como una dimensión adicional donde cualquier característica para uno de los subespacios que lo conforman es nula para cualquier conjunto y cualquier característica del espacio físico.

Con la introducción dada, vamos a denotar a $\mathcal{E}$ como el espacio de funcionales acotados que van del espacio físico al conjunto de números racionales o reales, según convenga.

Consideremos a Ξ como una familia de moléculas o candidatos a fármaco, que desde el punto de vista de la farmacología interactúan con un receptor común, con mecanismos fisicoquímicos similares y producen respuestas comparables.

El problema básico de la metodología QSAR puede presentarse como sigue:

La respuesta (r) y la actividad (ϕ) son elementos de $\mathcal{E}$, y para una familia química de moléculas Ξ existen f en $\mathcal{E}$, $f_r : Im(f) \longrightarrow \mathbb{R}$ y un descriptor molecular x, tales que f, f_r y x son funciones acotadas, funciones de medida casi siempre (característica mensurable respecto de la σ-álgebra de Borel inducida por la norma en el dominio), que para el conjunto de moléculas Ξ satisfacen:

$$\begin{aligned} f(x(\varsigma)) &= \phi(\varsigma); \\ f_r(\phi(x(\varsigma))) &= r(\varsigma); \qquad (1.13) \\ f_e(r(x(\varsigma))) &= \mathbf{y}; \qquad (1.14) \end{aligned}$$

para cada molécula ς en Ξ, donde $\mathbf{y}$ denota un valor o conjunto de valores asociados al afecto terapéutico y f_e es una función que queda definida por la tercera ecuación en 1.14.

El objetivo de la metodología QSAR es, como ya se ha dicho, recuperar información sobre f_e partiendo de conocer aproximaciones de los valores que toman x y ϕ para un conjunto muestral de moléculas, $\Xi_0 \subset \Xi$, que son sometidas a bioensayos.

Digamos que $\Xi_0 = \{\varsigma_1, ..., \varsigma_m\}$, $\{x(\varsigma_1) = x_1, ..., x(\varsigma_m) = x_m\}$ y $\{\phi(\varsigma_1) = \phi_1, ..., \phi(\varsigma_m) = \phi_m\}$. Nos concentraremos en el análisis QSAR que queda caracterizado por sólo los la primera ecuación en 1.14, un problema inversos con la formulación siguiente.

$$\begin{aligned} f(x) &= \phi \\ f(\tilde{x}_i) &\approx \tilde{\phi}_i, \quad i = 1, ..., m; \qquad (1.15) \end{aligned}$$

donde $\tilde{[\cdot]}$ indica una aproximación del valor real. Implícitamente aceptamos que errores pequeños en las mediciones de los descriptores causan pequeños errores bajo la evaluación de f, de forma que siguen siendo buenas aproximaciones de valores de mediciones con pequeños errores de ϕ.

En este problema es que centraremos esta primera revisión de la metodología QSAR, que completaremos con la sugerencia de **entender a la actividad también como un vector y no como un escalar**.

La hipótesis de linealidad QSAR, afirma que para los intereses que persigue la farmacología, $f(x)$ es un funcional lineal para familias determinadas por su estructura base y sólo un sustituyente (*ver* A), es decir, que para los fármacos base, la actividad de los candidatos a fármacos que se deriven de él por un sustituyente es esencialmente una característica molecular en una escala distinta.

....

Bajo estas hipótesis es que, en la literatura especializada, la principal herramienta de los análisis QSAR es la **regresión lineal múltiple** (**RLM**), si ninguna de las características moleculares conocidas explican linealmente a la actividad de la familia de

compuestos, entonces se buscan las combinaciones lineales de descriptores linealmente independientes para aproximar al descriptor mediante una base conveniente de un subespacio del espacio de descriptores. Dedicaremos un poco de tiempo a esto en secciones posteriores.

Definiciones

Cualquier sistema termodinámico considerado en este texto, salvo explicita excepción, estará determinado por la energía total del sistema, el volumen del mismo y la cantidad, en moles, de cada uno de los compuestos químicos en él presentes.

En lo tocante a los compuestos químicos presentes en el sistema, estarán limitados, como ya se ha establecido, a un ligando $\mathfrak{L}$, un sustrato $\mathfrak{A}$, la enzima $\mathfrak{R}$ (receptor para $\mathfrak{L}$ y ligando endógeno de $\mathfrak{A}$) y al compuesto $\mathfrak{P}$, el producto de la reacción *enzima-sustrato*.

En los sistemas considerados se entiende que tanto el ligando y el ligando endógeno compiten por ocupar el sitio activo de la enzima, siendo éste el mecanismo de inhibición de la acción enzimática. Toda vez que exista interacción química entre una molécula del receptor y una del ligando, ninguna molécula del ligando endógeno podrá ser transformada en una molécula del producto por la enzima ocupada. Salvo excepciones indicadas, el sitio activo en el receptor es único y no existe afinidad entre ligandos; no es posible la existencia de un complejo químico constituido por moléculas del ligando y el sustrato natural del receptor, ni por ligandos en la misma familia molecular del análisis QSAR.

Trabajaremos principalmente en dos tipos de sistemas que quedan determinados por su espacio de estados, ambos considerados inmersos en el mismo medio solvente $\mathfrak{M}$: sistemas $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$, con espacio de estados

$$(U, V, N_{\mathfrak{L}}, N_{\mathfrak{R}}, N_{\mathfrak{L}\mathfrak{R}});$$

y los sistemas $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R}, \mathfrak{A})$, con espacio de estados respectivo:

$$(U, V, N_{\mathfrak{L}}, N_{\mathfrak{R}}, N_{\mathfrak{A}}, N_{\mathfrak{L}\mathfrak{R}}, N_{\mathfrak{A}\mathfrak{R}}, N_{\mathfrak{P}}).$$

U denota la energía total del sistema, V el volumen del sistema, $N_{\mathfrak{L}}$ la cantidad de moles de ligando y análogamente la notación para las componentes restantes, que denotan la cantidad de moles del resto de los compuestos químicos en el sistema.

Definición 1.2.1. Cuando se considera a un conjunto arbitrario de moléculas, digamos Ξ, y una propiedad p observable en cada uno de los elementos de Ξ, la función $x_p : \Xi \longrightarrow \mathbb{R}$ se llama *descriptor molecular*, si para cualesquiera ς_1 y ς_2 elementos de Ξ se cumple que $x_p(\varsigma_1) = x_p(\varsigma_2)$ si y solamente si ς_1 y ς_2 observan la misma característica respecto de la propiedad p.

Definición 1.2.2. Para un conjunto de ligandos Ξ, con afinidad química con un receptor común, y para el espacio de estados, Esp,de un sistema como los recién denotados, definimos de forma general la actividad o respuesta biológica como una función real de la siguiente forma:

$$Act : \Xi \times Esp^n \to \mathbb{R}^q,\ n, q \in \mathbb{N},$$

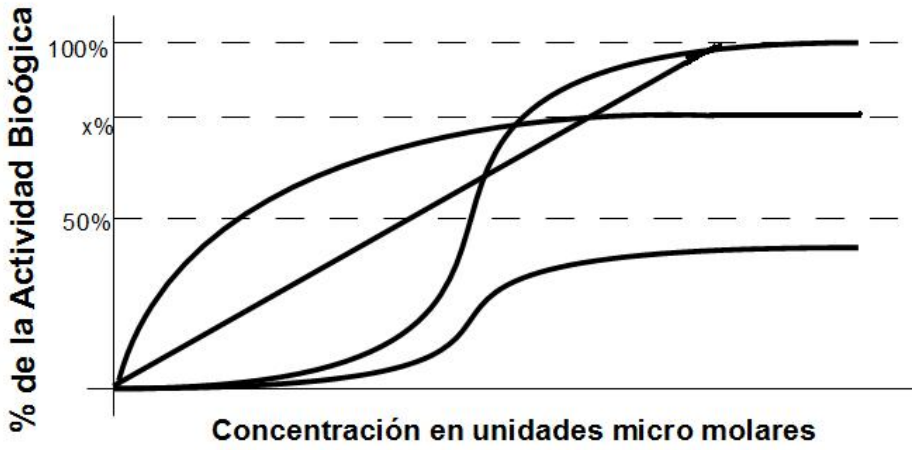

Figura 1.2: Gráfico de la forma general de cuatro funciones de actividad biológica pertenecientes a distintos tipos de actividad.

tal que $Act(\mathfrak{L}, E_S)$ es constante respecto de $\mathfrak{L}$; dado un vector de dimensión n, $\vec{S} = (S_1, ..., S_n)$, de posibles valores de entropía para el sistema y $E_{\vec{S}} = (E_1, ...E_n) \in Esp^n$ de forma que para cada E_j, $j = 1, ..., n$, es nula la coordenada correspondiente a la cantidad de moles de ligando, mientras que S_j es el correspondiente valor de entropía.

1.2.1. Concentraciones de inhibición y Energía libre de Gibbs

Dependiendo de la definición de la actividad y respuesta biológica la concentración de inhibición IC_{50} se define generalizando la definición dada para el ejemplo de antibióticos presentado en la sección 1.1.1, poniendo un poco de cuidado en la posibilidad de no inyectividad de la función de actividad.

En primer lugar nos ocuparemos de funciones de actividad con dominio de la forma $\Xi \times Esp$, donde $Esp = \mathbb{R}^n_+$, con $n = 2$ o $n = 5$, un estado del sistema en general será denotado por E. El tipo particular de la función de actividad se desprende del hecho de que las funciones de actividad usuales para la farmacología dependen de un estado inicial y un estado final, como es el caso de las concentraciones de inhibición.

En general se entiende por estado inicial de un sistema el estado en que se encuentran las coordenadas distintas de la entropía en el instante en que comienza la observación de dicho sistema, mientras el estado final del sistema se caracteriza como el estado que guardan el mismo conjunto de coordenadas al haber transcurrido un periodo de tiempo Δt desde el momento t_0.

Para nuestros fines, un estado inicial E_i de un sistema $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$ o $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R}, \mathfrak{A})$ se entenderá como un conjunto de condiciones iniciales para el sistema de ecuaciones diferenciales (1.10)-(1.11)-(1.12); es decir, valores dados para las cantidades $N_{\mathfrak{L}}$, $N_{\mathfrak{A}}$ y $N_{\mathfrak{R}}$

en el tiempo t_0 en que se inicia la observación de la reacción. Un estado final del sistema, E_f, será un estado asequible a partir de las condiciones iniciales dadas y que será determinado por el observador.

Volviendo al ejemplo de los bioensayos de actividad celular para antibióticos, el estado inicial es aquel en que se inocula una bacteria en una solución con una concentración específica de ligando, el estado final en tales pruebas es aquel que se observa transcurrido un lapso de tiempo estandarizado después de inoculada la cepa.

Para el tipo de reacciones químicas que aquí competen, el estado inicial para una familia de ligados se entenderá como aquél de completa disociación, $N_{\mathfrak{LR}} = N_{\mathfrak{AR}} = N_{\mathfrak{P}} = 0$. El estado final será el de equilibrio.

Ahora daremos las definiciones que manejaremos de "intensidad de la actividad" o "porcentaje de inhibición", "concentración inicial de inhibición" y "concentración de inhibición en un porcentaje α".

Definición 1.2.3. Sea, Ξ una familia química, Esp_i el conjunto de todos los posibles estados iniciales para un sistema $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$ o $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R}, \mathfrak{A})$ con $\mathfrak{L} \in \Xi$, E_f el estado final del sistema, dependiente en estos sistemas sólo del estado inicial y la constante de equilibrio, y $Act : \Xi \times Esp_i \longrightarrow \mathbb{R}$ una función de actividad dada.

Para Ξ se define un estado patrón inicial, E_p como aquel observado en el sistema caracterizado por $\mathfrak{L}$ y tal que $N_{\mathfrak{L}} = N_{\mathfrak{LR}} = 0$, de otra forma, un estado patrón inicial se entenderá como el estado inicial de un sistema en total ausencia de ligando.

Se define

$$\begin{array}{rcl} \alpha_{Act} : \Xi \times Esp_i & \longrightarrow & \mathbb{R} \\ (\mathfrak{L}, E_i) & \longmapsto & 1 - \frac{Act(\mathfrak{L}, E_f)}{Act(\mathfrak{L}, E_p)}. \end{array}$$

La función α_{Act} se dirá la *función de inhibición inducida por la función de actividad* Act, la notación $\alpha_{\mathfrak{L}}$ reemplazará por comodidad a $\alpha_{Act}(\mathfrak{L}, E_i)$ cuando por contexto no sea necesario especificar a la función Act y el estado inicial.

Definición 1.2.4. En el marco de la definición 1.2.3, para un escalar dado α definimos el conjunto de estados iniciales de inhibición con potencia α para el ligando $\mathfrak{L}$ como

$$\mathfrak{E}_\alpha(\mathfrak{L}) = \{E_i \in Esp_i : \alpha_{Act}(\mathfrak{L}, E_i) = \alpha\}.$$

Se define además la restricción de $\mathfrak{E}_\alpha(\mathfrak{L})$ respecto de una cantidad inicial de receptores dada, $N_{\mathfrak{R},0}$, como el conjunto de elementos de $\mathfrak{E}_\alpha(\mathfrak{L})$ para los que se verifica $N_{\mathfrak{R}} = N_{\mathfrak{R},0}$, dicha restricción quedará denotada por $\mathfrak{E}_{\alpha, N_{\mathfrak{R},0}}(\mathfrak{L})$.

Por último, sólo para efectos de notación emplearemos $[\cdot]_E$ para referirnos a la concentración del compuesto entre corchetes en el estado E. Dado el conjunto $\mathfrak{E}_{\alpha, N_{\mathfrak{R},0}}(\mathfrak{L})$, se definen respectivamente a los conjuntos de concentraciones iniciales y de concentraciones con ínfima potencia de inhibición α como:

$$\mathfrak{C}_\alpha(\mathfrak{L}) = \{[\mathfrak{L}]_{E_i} : E_i \in \mathfrak{E}_{\alpha_1, N_{\mathfrak{R},0}}(\mathfrak{L}),\ \alpha_1 \geq \alpha\}$$

y

$$\mathfrak{IC}_\alpha(\mathfrak{L}) = \{[\mathfrak{L}]_{E_f} : E_i \in \mathfrak{E}_{\alpha_1, N_{\mathfrak{R},0}}(\mathfrak{L}),\, \alpha_1 \geq \alpha\}.$$

Las notaciones $\mathfrak{C}$ y $\mathfrak{IC}$ no hacen referencia a la dependencia que cada uno de estos conjuntos guarda con la cantidad inicial de receptores por no ser en general necesaria una alusión explícita, el contexto será siempre suficiente en este texto.

Ahora, estamos en condiciones de definir una *concentración inicial con porcentaje de inhibición* α respecto de una cantidad inicial de receptores en el sistema.

Definición 1.2.5. Dada una familia de ligandos Ξ, una función de actividad Act, una cantidad inicial de de receptores $N_{\mathfrak{R},0}$ y un escalar α tales que $\mathfrak{C}_\alpha(\mathfrak{L})$ es no vacío para cada $\mathfrak{L}$ en Ξ, entonces definimos la concentración inicial con potencia de inhibición α como la función:

$$\begin{array}{rccl} C_\alpha : \Xi & \longrightarrow & \mathbb{R}_+ \\ \mathfrak{L} & \longmapsto & inf\left(\mathfrak{C}_\alpha(\mathfrak{L})\right). \end{array}$$

Bajo las mismas condiciones y siempre que $\mathfrak{IC}_\alpha(\mathfrak{L})$ es no vacío para cada $\mathfrak{L}$ en Ξ, entonces se define también la *concentración de inhibición con factor* α como

$$\begin{array}{rccl} IC_\alpha : \Xi & \longrightarrow & \mathbb{R}_+ \\ \mathfrak{L} & \longmapsto & inf\left(\mathfrak{IC}_\alpha(\mathfrak{L})\right). \end{array}$$

Por definición hemos pedido que el estado final se encuentre en función del estado inicial, pero en general no se conoce relación explícita entre C_α e IC_α.

Lo importante de las definiciones anteriores es que pueden aplicarse a cualquier análisis QSAR con mínimos cambios y que, bajo la definición que hemos dado de función de actividad, la actividad y la respuesta biológica, son casos particulares que enfatizan el tipo de mediciones experimentales si se observa una interacción entre sólo dos tipos de moléculas sintetizadas o aisladas, o si se observa la interacción de un ligando con su receptor en una forma no aislada (en el interior de una bacteria por ejemplo).

La forma en que se define la actividad y su comportamiento cualitativo en general depende del tipo de reacciones químicas que se observen. En la figura 1.2 se encuentra una representación gráfica del comportamiento cualitativo que en la farmacología se observa para distintas definiciones de la actividad.

Nos interesarán principalmente tres tipos de magnitudes estrechamente relacionadas, son ambas funciones de una restricción particular de la misma función de actividad de acuerdo con nuestra definición general, dos de ellas son IC_{50} y C_{50}.

En la práctica la forma y los errores involucrados en la generación de valores experimentales de C_α e IC_α hacen de estos datos variables aleatorias para cada porcentaje de inhibición α. Una hipótesis adicional, es que cuentan con distribuciones de probabilidad simétrica tales que, para cualquier $r > 0$, la medida de probabilidad de $(x - r, x + r)$ decrece conforme x se aleja de la respectiva esperanza. Así, por fines prácticos no se hace distinción entre la esperanza de cada variable aleatoria y las definiciones de IC_α y C_α.

En la química se reconoce que ligandos distintos pueden corresponder a tipos significativamente distintos de actividad biológica pese a existir sólo pequeñas diferencias en la probabilidad que tienen de interaccionar con $\mathfrak{R}$ bajo condiciones iguales en el sistema

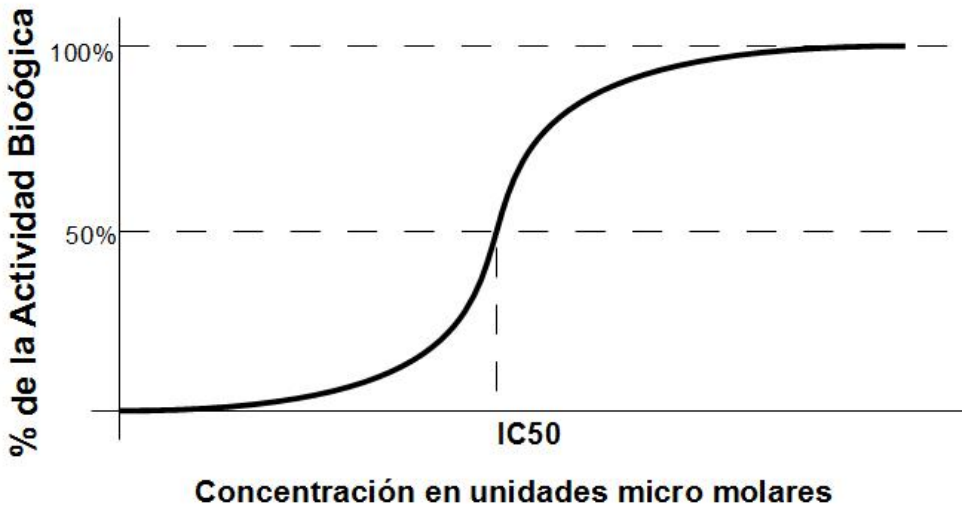

Figura 1.3: Representación gráfica de IC50

antes de ser incorporados los ligandos; con esto nos referimos a que, incluso para dos ligandos $\mathfrak{L}_1$ y $\mathfrak{L}_2$, el comportamiento cualitativo de $IC_\alpha(\mathfrak{L}_1)$ y $IC_\alpha(\mathfrak{L}_2)$, como funciones de α, puede ser significativamente distinto pese a que entre el par de moléculas sean consideradas pequeñas las diferencias tanto constitucionales como fisicoquímicas. En la figura 1.2 se muestra un ejemplo gráfico de cuatro diferentes tipos de comportamiento de la inhibición como función de la concentración.

El objetivo final de estos modelos es brindar herramientas en el diseño de medicamentos eficientes, que logren efectos farmacológicos deseados; información más completa sobre esto podría incluir conjuntos de datos más nutridos, sobre el tiempo y distintos porcentajes de inhibición; aunque, como se ha dicho, no es siempre posible. Lo más común para considerar a un compuesto como un candidato a fármaco es esperar una reducción cuando menos del 50 % y de ahí que sea este estado el que se considera como deseable (*ver fig. 1.3*).

Antes de continuar debemos hacer una observación sobre los descriptores moleculares. De acuerdo con nuestra definición, la constante de equilibrio en cualquiera de los sistemas considerados; así como la afinidad, actividad y eficacia de un ligando respecto de su receptor son descriptores moleculares; esto significa que no estamos entendiendo a un descriptor como un observable de una molécula como sistema termodinámico aislado.

Los descriptores moleculares bajo nuestra definición son sólo variables que se encuentran en función de una molécula entendida como un sistema termodinámico, un grafo, una matriz, un conjunto de matrices, un objeto físico, etcétera. Respecto de ello, hacemos una aclaración: cuando usemos el término "descriptor" nos referiremos a una característica molecular, una función real que no depende del receptor; si un compuesto tiene múltiples receptores posibles, el valor del descriptor para el compuesto es siempre el mismo inde-

pendientemente del receptor con el que se le observe interactuar y las condiciones en las que lo hagan.

Denotaremos por Ξ a una familia de moléculas orgánicas, mientras que se empleará el símbolo $\bar{\Xi}$ para representar al conjunto de todas las moléculas orgánicas y ς se reservará para una molécula orgánica en general.

Sobre la energía libre de Gibbs

El segundo tipo de actividad que motiva nuestro quehacer es la energía libre de Gibbs G, una propiedad de los sistemas termodinámicos que es la energía isotérmicamente utilizable para producir trabajo. Si consideramos como estado inicial del sistema el de total disociación y como estado final aquel en que se ha alcanzado el equilibrio químico con una inhibición de la actividad al $100\alpha\,\%$, entonces la diferencia de energía libre de Gibbs de un estado a otro, ΔG, proporciona información sobre la espontaneidad de la reacción.

Ocurre que IC_α y ΔG se encuentran relacionados de forma que cualquiera de ellos puede expresarse como función del otro, relación que J. S. Tokarski y A. J. Hopfinger muestran en [31]:794. La relación anterior depende también de la concentración disponible del receptor y la potencia observada, particularmente en el caso de IC_{50} la forma de esta relación es muy útil y sencilla.

Para reacciones reversibles a temperatura y presión constantes, sistemas $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$, y sobre todo en las concentraciones tan pequeñas como las que se utilizan experimentalmente (micromolar, 10^{-6} mol/l), una de las relaciones más importantes para los análisis QSAR es la diferencia observada en la energía libre de Gibbs del sistema respeto de un estado inicial de total disociación y el estado de equilibrio como estado final:

$$\Delta G = -KT \ln \left(\frac{[\mathfrak{LR}]_{eq}}{[\mathfrak{L}]_{eq} [\mathfrak{R}]_{eq}} \right); \tag{1.16}$$

donde K es la constante de los gases, T la temperatura del sistema, en unidades absolutas o Kelvins, y las concentraciones en el argumento del logaritmo corresponden al estado de equilibrio del sistema. Una explicación detallada de esta relación puede consultarse en [19].

Valores negativos de ΔG indican que el sistema realizó trabajo a partir de energía interna, significando que la reacción ocurrió de forma espontánea; por otro lado, valores positivos de la diferencia en la energía libre de Gibbs nos dicen que la reacción espontánea ocurre en realidad en la dirección opuesta, cuando se intercambian los estados inicial y final; implica que debe realizarse trabajo externo para lograr ir de un estado a otro.

Para la farmacología son de gran utilidad los ligandos que observan valores negativos y de la mayor magnitud posible para el cambio en la energía libre, en tanto menor sea esta variación mayor será la eficiencia de un fármaco. En este punto es importante recordar que la medición de ΔG no se entiende como una variable que dependa del tiempo. Dos compuestos con el mismo valor negativo para ΔG lograrán una reacción espontánea en el sistema, no obstante pueden ser distintos los tiempos mínimos en que uno y otro compuesto producen que el respectivo sistema alcance el mismo estado deseable. De esta

forma, para los análisis QSAR un segundo tipo de actividad es la medición de ΔG para cada ligando específico en un sistema $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$.

En el tipo de análisis en el que nos concentramos se conoce una relación entre los dos tipos de actividad IC_α y ΔG. Veremos que para cada sistema o reacción determinada por el par de ecuaciones químicas (1.7) y (1.6), o simplemente por (1.6), en el estado de equilibrio los valores IC_α y $\alpha_{\mathfrak{L}}$ dependen del total de ligando disponible en el sistema, de $[\mathfrak{L}]_0$, y de los parámetros k_1 y k_2 en la ecuación diferencial (1.10); ΔG será por tanto una función de dichos parámetros.

Para un estado inicial E_i considerando a $\alpha = 1 - [\mathfrak{R}]_{eq} / [\mathfrak{R}]_0$, con $[\mathfrak{R}]_0 = [\mathfrak{R}]_{E_i}$. Por la unicidad del sitio activo se tiene la relación

$$[\mathfrak{L}\mathfrak{R}] = [\mathfrak{R}]_0 - [\mathfrak{R}] = (1-\alpha)\,[\mathfrak{R}]_0\,;$$

así que, al sustituir $[\mathfrak{L}\mathfrak{R}]$ y $[\mathfrak{R}]$ por $(1-\alpha)\,[\mathfrak{R}]_0$ y $\alpha\,[\mathfrak{R}]_0$ en (1.16) se llega a:

$$\begin{aligned} \Delta G &= -KT \ln\left(\frac{1-\alpha}{\alpha\,[\mathfrak{L}]_{eq}}\right) \\ &= KT\left(\ln([\mathfrak{L}]_{eq}) - \ln(\frac{1}{\alpha} - 1)\right). \end{aligned} \tag{1.17}$$

La diferencia en la energía libre de Gibbs entre el estado final y el inicial, ΔG, es una variable real que depende de la concentración libre de ligando en el estado final y el correspondiente porcentaje de inhibición α. Consideraremos por tanto a la energía libre como una función de actividad para el espacio de estados con sentido físico y sus correspondientes estados finales de equilibrio, en este sentido (1.17) nos dice que:

$$\Delta G = KT\left(\ln(IC_\alpha) - \ln(\frac{1}{\alpha} - 1)\right). \tag{1.18}$$

Por convención, si x_0 y x son elemento de un espacio métrico con distancia no acotada relativo a un observador x es una aproximación de x_0, entonces se dice que el error cometido en la aproximación de x_0 mediante x es de orden p, siempre que p sea el mínimo natural que verifica la pertenencia de x a la bola centrada en x_0 y radio 10^p; esta convención es el motivo por el que lo usual en la literatura es reescribir en términos de log_{10} la relación existente entre $IC_{\alpha_{\mathfrak{L}}}$ y ΔG, ambas como funciones del estado inicial.

Dado α en $(0, 1]$ se define el *índice de inhibición al* 100α *%* como $pIC_\alpha = \log_{10}(IC_\alpha)$. Se sigue de (1.18), que para un α dado y una familia de molecular para la que IC_α se encuentre bien definido:

$$\Delta G = KT\ln(10)\left(pIC_\alpha - \log_{10}\left(\frac{1}{\alpha} - 1\right)\right); \tag{1.19}$$

en particular

$$\Delta G = \ln(10) KT pIC_{50},\ \alpha = 1/2.$$

Es claro que por su definición IC_α como la potencia de inhibición son funciones de actividad toda vez que sea dado el conjunto de estados iniciales (o finales) para un análisis QSAR. Por otro lado, la ecuación (1.16) establece que ΔG también es una función de actividad. La relevancia de (1.19) radica justamente en la relación que exhibe entre las funciones y conceptos relevantes que son reportados por las laboratorios como resultado de los bioensayos realizados a un conjunto de moléculas.

La potencia de inhibición tiene sentido para una análisis QSAR arbitrario y en realidad, lo más común es que en un análisis se conozca a $pIC_{50}(\mathfrak{L})$ como la *actividad de* $\mathfrak{L}$, (1.16) nos dice que para sistemas $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$ es 1 la correlación estadística entre la actividad de un ligando y el cambio de la energía libre de Gibbs en el sistema al pasar de una estado de total disociación a un estado de equilibrio para un estado inicial adecuado.

Suele no hacerse distinción entre ΔG y pIC_{50} en el proceso de buscar combinaciones lineales de descriptores que expliquen la actividad, desde que para $\mathfrak{L}$ en condiciones adecuadas se satisface $Cor(\Delta G, pIC_\alpha) = 1$. Serán indistintos para nosotros también cuando se haga referencia a la actividad de un ligando.

De lo dicho en este bloque comentamos al lector que en un análisis QSAR donde los tipos básicos de funciones de actividad para sistemas $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$ o $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R}, \mathfrak{A})$ son:

$$\begin{array}{rcl} [\mathfrak{R}] : \Xi \times Esp & \longrightarrow & \mathbb{R}_+ \\ (\mathfrak{L}, E) & \longmapsto & [\mathfrak{R}]_E \, ; \end{array} \tag{1.20}$$

$$\begin{array}{rcl} \alpha : \Xi \times Esp & \longrightarrow & \mathbb{R}_+ \\ (\mathfrak{L}, E) & \longmapsto & \alpha_{[\mathfrak{R}]}(\mathfrak{L}, E); \end{array} \tag{1.21}$$

$$\begin{array}{rcl} C_\alpha : \Xi \times Esp & \longrightarrow & \mathbb{R}_+ \\ (\mathfrak{L}, E) & \longmapsto & C_{\alpha_{[\mathfrak{R}]}(\mathfrak{L}, E)}; \end{array} \tag{1.22}$$

$$\begin{array}{rcl} IC_\alpha : \Xi \times Esp & \longrightarrow & \mathbb{R}_+ \\ (\mathfrak{L}, E) & \longmapsto & IC_{\alpha_{[\mathfrak{R}]}(\mathfrak{L}, E)}; \end{array} \tag{1.23}$$

$$\begin{array}{rcl} \Delta G : \Xi \times Esp & \longrightarrow & \mathbb{R}_+ \\ (\mathfrak{L}, E) & \longmapsto & ln(10) K T p IC_{50}. \end{array} \tag{1.24}$$

Como se verá más adelante, en los sistemas aquí considerados, existe una relación uno a uno entre IC_α y C_α siempre que ambos valores se encuentren bien definidos; más que eso, bajo tales condiciones y dado un conjunto de estados iniciales existe un único E_i tal que $C_\alpha = [\mathfrak{L}]_{E_i}$, implicando un relación biyectiva entre los estados inicial y final.

C_α será entonces, dados $[R]_0$ y V_0 el volumen en que se encuentra contenido el sistema, la concentración inicial de ligando que logra un estado de equilibrio en el que el ligando ocupa una fracción α de los receptores disponibles. IC_α y C_α quedarán determinados por el conjunto de estados iniciales que se elija.

En este punto el lector merece una disculpa por lo rebuscado de las definiciones anteriores, la forma en que aquí se enuncian es consecuencia de la incapacidad del autor en lograr explicar de forma sencilla y suficientemente general, lo que en diversas publicaciones de análisis QSAR se entiende por la respuesta biológica, la actividad y particularmente por concentración de inhibición al 50 %.

Por lo general los análisis QSAR se clasifican de acuerdo con la naturaleza de los descriptores moleculares involucrados en el análisis; sin embargo, distinguiendo entre un descriptor molecular como una función de una molécula arbitraria de ligando, y la actividad como una función dependiente de la interacción de una concentración de ligando en un sistema biológico, entonces es pertinente preguntarse si deberíamos considerar nuevas clasificaciones de acuerdo con la naturaleza de la función de actividad, más allá de la que se hace para la familias moleculares en el apéndice A en donde es una premisa que un análisis QSAR responde adecuadamente a las necesidades de investigación.

1.3. Métodos de ajuste de parámetros

A lo largo de las siguientes páginas trabajaremos en espacios vectoriales con producto interior como los espacios euclidianos $\mathbb{R}^n$ y los espacios de funciones funciones L^p. En esta sección se presentan definiciones para las que se requiere establecer una notación común de notación que permita el objetivo de la comunicación. Se prescinde aquí de definiciones como la de espacio vectorial, producto interior y función entre otras.

1.3.1. Definiciones y herramienta básica

Conforme al uso regular de notación en la literatura matemática, se denotará por $M_{m\times n}(\mathbb{R})$ al espacio vectorial de matrices de orden $m \times n$ sobre el campo $\mathbb{R}$ y con componentes en el mismo. Si X es una matriz real de orden $m \times n$ $[X]_{ij}$, $i = 1, ..., m$ y $j = 1, ..., n$, denotará la $ij-$ésima componente de la matriz X; $[X]_{i\cdot}$ y $[X]_{\cdot j}$ serán sus respectivas $i-$ésima fila y $j-$ésima columna.

Siempre que sea empleada una letra latina mayúscula será para denotar un elemento de $M_{m\times n}(\mathbb{R})$, y si no existe otra especificación particular, la respectiva letra minúscula con subíndice ij también se empleará para denotar a la ij-ésima componente de dicha matriz, con subíndice j para la $j-$ésima columna de la matriz y con subíndice $i\cdot$ para la $i-$ésima fila; por ejemplo, si $X \in M_{m\times n}(\mathbb{R})$, entonces $x_{ij} = [X]_{ij}$, $x_{i\cdot} = [X]_{i\cdot}$ y $x_j = [X]_{\cdot j}$.

La suma y producto matriciales que se emplearan, así como el producto de un vector por un escalar, serán los usuales salvo excepciones indicadas:

Definición 1.3.1. Sean m, n y p naturales arbitrarios, X_1 y X_2 elementos de $M_{m\times n}(\mathbb{R})$, X_3 una matriz que pertenece a $M_{n\times p}(\mathbb{R})$ y λ un escalar en $\mathbb{R}$. Se definen

I. **Producto por un escalar**
$\lambda X_1 \in M_{m\times n}(\mathbb{R})$ tal que $[\lambda X_1]_{ij} = \lambda[X_1]_{ij}$;

II. **Suma matricial**
$X_1 + X_2 \in M_{m\times n}(\mathbb{R})$ tal que $[X_1 + X_2]_{ij} = [X_1]_{ij} + [X_2]_{ij}$;

III. **Producto matricial**
$X_1 \circ X_3 = X_1X_3 \in M_{m\times p}(\mathbb{R})$ tal que $[X_1 + X_2]_{ij} = \sum_{k=1}^{n}[X_1]_{ik}[X_2]_{kj}$.

Los elementos de los espacios $\mathbb{R}^n$ serán considerados como vectores columna, matrices de orden $n \times 1$, con lo que, bajo el producto matricial usual, el producto interior que utilizaremos en estos espacios será también el usual, es decir, si x, y son elementos de $\mathbb{R}^n$, entonces el producto interior $< x, y >$ queda caracterizado por el producto matricial $x^t y$.

Dos vectores son *ortogonales entre si* siempre y cuando su producto interior sea nulo, un conjunto de vectores se dice que es *ortogonal* si sus elementos son ortogonales dos a dos, si cualesquiera dos vectores que se elijan en él son ortogonales siempre que sean distintos entre si.

Sin importar la dimensión de un espacio vectorial con producto interior, $(V, < \cdot, \cdot >)$, emplearemos de forma sistemática la notación $\|\cdot\|$ para referirnos a la norma euclidiana, también conocida como la norma usual en los espacios que trabajaremos:

Definición 1.3.2 (Norma euclidiana)**.** Sea $\mathbb{R}_+$ el conjunto de números reales no negativos, $(V, < \cdot, \cdot >)$ un espacio vectorial con producto interior $< \cdot, \cdot >$, sobre el campo $K = (\mathbb{R}$ ó $\mathbb{C})$. La norma euclidiana en V se define como la función $\|\cdot\|$ con el espacio V por dominio, $\mathbb{R}_+$ como conjunto de llegada y definida por

$$\|x\| = \|\cdot\|(x) =< x, x >^{\frac{1}{2}}, \quad \forall x \in V.$$

En un espacio vectorial con producto interior y norma inducida por el mismo, un conjunto de vectores es *ortonormal* si es un conjunto ortogonal y la norma de cualquiera de sus elementos es 1.

Una matriz X de orden $m \times n$ y componentes reales se dice ortogonal si sus columnas como vectores forman un conjunto ortogonal, X se dirá ortonormal o unitaria si sus columnas forman un conjunto ortonormal.

Para cualquier conjunto propio de $\mathbb{R}^n$, Ω, $gen(\Omega)$ denotará al subespacio vectorial generado por Ω, el conjunto de todas las posibles combinaciones lineales de elementos de Ω. Recordamos al lector un espacio vectorial V se dice de dimensión finita si existe un conjunto de vectores finito linealmente independiente que genere al espacio. La cardinalidad del conjunto β es la dimensión del espacio y se dice que β es una base para el espacio V. Se sabe que la dimensión es única pero β no lo es.

Una función $L : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ se dirá una transformación lineal si

> $L(\lambda x_1 + x_2) = \lambda L(x_1) + L(x_2)$, cualesquiera que sean los vectores x_1 y x_2 en $\mathbb{R}^n$ y el escalar λ.

Si la transformación L es lineal, con dominio y contra dominio como antes, entonces el *núcleo* de L es el conjunto $\{x \in \mathbb{R}^n : L(x) = 0\}$.

Cuando V y W son espacios vectoriales sobre el mismo campo K, L es una transformación lineal de V en W, $\beta = \{\beta_1, ..., \beta_n\}$ es una base para V y $\gamma = \{\gamma_1, ..., \gamma_n\}$ lo es para W, entonces siempre existe una única matriz X de orden $m \times n$ y con componentes en K tal que $L(\beta_j) = \sum_{i=1}^{m} x_{ij}\gamma_i$. Dadas las bases β y γ la relación entre X y L es biunívoca y por ello se dice que X es la matriz asociada a la transformación lineal L y bise versa, en ambos casos respecto de las bases dadas.

Cuando X es conocida L se denota por L_X, notación que presupone bases dadas y conocidas; mientras que cuando es L quien se conoce, entonces X es denotada por $[L]_{\beta}^{\gamma}$. En el caso particular del los espacios $V = \mathbb{R}^n$ y $W = \mathbb{R}^m$ la notación L_X presupondrá las respectivas bases canónicas toda vez que sea empleada.

Optimización de campos escalares

Una función $f : Dom(f) \subseteq \mathbb{R}^n \longrightarrow \mathbb{R}^m$ suele conocerse como *funcional* o *campo escalar* si $m = 1$, *campo vectorial* si $m > 1$ y *operador* cuando su dominio y contra dominio coinciden, $m = n$. Cuando f es una transformación lineal se usan los términos funcional y operador lineal.

Como en la mayoría de los libros de cálculo y análisis matemático $J_f(x)$ se referirá, siempre que tenga sentido, a la matriz jacobiana de f en x, emplearemos simplemente J para simplificar la notación siempre que sea posible; C^1 y C^2 serán los conjuntos de campos vectoriales continuamente y dos veces continuamente diferenciables sobre su dominio.

Para el caso particular de los campos escalares en C^1 o C^2 denotaremos por $\nabla f(x)$ al gradiente de f en x y por $\nabla^2 f(x)$ a la matriz hessiana de f (jacobiana del gradiente de f como campo vectorial) en el punto x. Por el resto de la sección f se empleara para referirse a un campo escalar en C^1 y en algunas ocasiones en C^2.

Consideremos el problema general de optimización, con $\Omega \subset \mathbb{R}^n$,

$$\mathcal{P} \left\{ \begin{array}{ll} min & f(x) \\ & x \in \Omega. \end{array} \right.$$

Es un problema que de forma recurrente se presenta cuando se requiere identificar un modelo matemático del que se presume su forma general, en dependencia de parámetros escalares sujetos a restricciones que en la práctica le den sentido o viabilidad al modelo. El estudio del problema $\mathcal{P}$ ha dado importantes y útiles resultados en los que posteriormente nos apoyaremos.

En lo sucesivo usaremos $\mathbf{int}\,(\Omega)$, $\overline{\Omega}$ y $\mathbf{fr}\,(\Omega)$ para denotar el interior, clausura y frontera de Ω.

Definición 1.3.3. Un vector $d \in \mathbb{R}^n$ es una dirección *tangente positiva* de $\Omega \subset \mathbb{R}^n$ en el punto $x_0 \in \Omega$, si y sólo si existe $\epsilon_0 > 0$ y una función $E : (0, \epsilon_0) \longrightarrow \mathbb{R}^n$ tal que

$$\begin{aligned} x(\epsilon) &= x_0 + \epsilon d + E(\epsilon) \in \Omega, \ \ \forall \epsilon \in (0, \epsilon_0) \\ \lim_{\epsilon \to 0^+} \frac{E(\epsilon)}{\epsilon} &= 0. \end{aligned}$$

Si $E \equiv 0$ d se dirá una dirección *factible*. Cuando E está definida sobre $(-\epsilon_0, \epsilon_0)$ y $\frac{E(\epsilon)}{\epsilon} \to 0$ cuando $\epsilon \to 0$, entonces d será una dirección *tangente*.

$K^+(\Omega, x_0)$, $F(\Omega, x_0)$ y $K(\Omega, x_0)$ serán, respectivamente, los conjuntos de direcciones tangentes positivas, factibles y tangentes.

Los conjuntos $K^+(\Omega, x_0)$, $F(\Omega, x_0)$ y $K(\Omega, x_0)$, son todos conos y es inmediato que cualquier dirección tangente o factible es también una dirección tangente positiva, es decir, $F(\Omega, x_0) \subseteq K^+(\Omega, x_0)$ y $K(\Omega, x_0) \subseteq K^+(\Omega, x_0)$.

Definición 1.3.4. Un punto $x_0 \in \Omega$ es un *mínimo local* del problema $\mathcal{P}$ si existe una vecindad $B(x_0, r)$, $r > 0$, tal que.

$$f(x_0) \leq f(x), \ \ \forall x \in \Omega \cap B(x_0, r).$$

Diremos que x_0 es un mínimo local estricto cuando la desigualdad sea estricta y $x \neq x_0$. Será un mínimo global si la desigualdad se cumple para cada elemento de Ω, $r = \infty$.

En el caso correspondiente x_0 se dirá un máximo si se satisface la desigualdad contraria ($\geq$).

Teorema 1.3.1 (Condiciones necesarias de primer orden)**.** *Si $f \in C^1$ y x_0 es un mínimo local del problema $\mathcal{P}$, entonces*

$$\nabla f(x_0)d \geq 0, \ \ \forall d \in K^+(\Omega, x_0).$$

Corolario 1.3.2. *1. Si d es una dirección tangente, entonces d y $(-d)$ son direcciones tangentes positivas y se tiene que*

$$\nabla f(x_0)d = 0, \ \ \forall d \in K(\Omega, x_0);$$

*2. Si x_0**int** (Ω), entonces $K^+(\Omega, x_0) = K(\Omega, x_0) = F(\Omega, x_0) = \mathbb{R}^n$ y por lo tanto $\nabla f \equiv 0$.*

Teorema 1.3.3 (Condiciones necesarias de segundo orden)**.** *Si $f \in C^2$ y x_0 es un mínimo local del problema $\mathcal{P}$, entonces*

1. *$\nabla f(x_0)d \geq 0$, $\forall d \in K^+(\Omega, x_0)$;*
2. *Para toda dirección factible d tal que $\nabla f(x_0)d = 0$ se tiene que*

$$d^t \nabla^2 f(x_0) d \geq 0.$$

Definición 1.3.5 (conjunto convexo)**.** Un subconjunto de $\mathbb{R}^2$, Ω, es convexo si y sólo si para cada escalar $\lambda \in (0, 1)$ y cualesquiera x_1, x_2 elementos de Ω el vector $\lambda x_1 + (1-\lambda)x_2$ también es un elemento de Ω.

Teorema 1.3.4 (Caracterización de las direcciones factibles de un conjunto convexo)**.** *Sea $\Omega \subset \mathbb{R}^n$ un conjunto convexo y x un elemento de Ω, entonces*

$$F(\Omega, x) = \{d \in \mathbb{R}^n : \exists y \in \Omega, \exists \lambda > 0 : d = \lambda(y - x)\}.$$

Definición 1.3.6 (función convexa)**.** Sea Ω un subconjunto convexo de $\mathbb{R}^n$. Una función $f : \Omega \longrightarrow \mathbb{R}$ se dice *convexa* si para todo $x_1, x_2 \in \Omega$ y $\lambda \in [0, 1]$ se cumple que:

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_2) + (1 - \lambda) f(x_2).$$

Teorema 1.3.5 (Condiciones suficientes para el caso convexo del problema $\mathcal{P}$)**.** *Si $f \in C^1$ es una función convexa en el conjunto convexo $\Omega \subset \mathbb{R}^n$, entonces todo punto $x_0 \in \Omega$ que verifique la condición de primer orden*

$$\nabla f(x_0)d \geq 0, \qquad \forall d \in F(\Omega, x_0)$$

es un mínimo global *de f en Ω.*

Teorema 1.3.6 (Condiciones suficientes de segundo orden para el problema $\mathcal{P}$)**.** *Si $f \in C^2$, Ω es un conjunto convexo y se cumplen las siguientes condiciones*

1. $\nabla f(x_0)d \geq 0, \qquad \forall d \in F(\Omega, x_0)$;
2. $\nabla^2 f(x_0)$ *es semidefinida positiva en una vecindad* $B(x_0, r)$, $r > 0$, *de* x_0.

Entonces x_0 es un mínimo local de f en Ω.

Respecto de los teorema de optimización que necesitaremos más adelante, sólo resta presentar un teorema de primer orden y suficiencia que emplearemos en casos particulares en que $h \in C^1(\mathbb{R}^n, \mathbb{R}^m)$, $g \in C^1(\mathbb{R}^n, \mathbb{R}^p)$ y el conjunto de restricciones es de la siguiente forma:

$$\Omega_{g,h} = \{x \in \mathbb{R}^n : h(x) = 0,\, g(x) \leq 0\}\,.$$

Definición 1.3.7. El punto $x_0 \in \Omega_{g,h}$ es un punto regular si es linealmente independiente el conjunto de vectores gradiente

$$\nabla h_i,\, i = 1, ..., m;\, \nabla g_j, j \in J(x_0);$$

donde

$$J(x_0) = \{j \in [i, ..., p] : g_j(x_0) = 0\}\,.$$

El conjunto $J(x_0)$ se conoce como conjunto de indices activos.

Teorema 1.3.7 (Karush-Kuhn-Tucker)**.** *Sea x_0 un punto de mínimo y regular del problema de optimización que resulta de sustituir Ω por $\Omega_{h,g}$ en $\mathcal{P}$; entonces, la condición necesaria de primer orden, teorema 1.3.1, es equivalente a que existen escalares $\mu_i \in \mathbb{R}$, $i = 1, ..., m$, $\lambda_j \geq 0$, $j = J(x_0)$, tales que:*

$$\nabla f(x_0) + \sum_{i=1}^{n} \mu_i \nabla h_i(x_0) + \sum_{i=1}^{p} \lambda_j \nabla g_p(x_0) = 0.$$

1.3.2. Mínimos Cuadrados Ordinarios, MCO

Uno de los aspectos de la metodología QSAR es la identificación o aproximación de parámetros únicos en un modelo general que explique el comportamiento de una muestra de mediciones de un observable realizadas para un fenómeno de interés. Dicho con mayor formalidad, sean:

1. Ω un subconjunto propio del espacio euclidiano $\mathbb{R}^n$. Este conjunto se entiende como el rango de una variable, la j-ésima componente de los vectores en Ω se dirá la j-ésima variable independiente (variable predictora o de predicción).
2. Λ un subconjunto de $\mathbb{R}^k$ para el que cada elemento en Λ se denomina como un vector de parámetros, cada componente de un vector de parámetros es en sí mismo un parámetro.

A cualquier familia de funciones de la forma $\mathcal{F} = \{f : \Omega \times \Lambda \longrightarrow \mathbb{R}\}$ se le conoce como una familia paramétrica de funcionales con dominio común Ω y espacio de parámetros Λ.

Supongamos que y es una variable para la cual existe un vector de parámetros desconocido, λ_0, tal que $y = f(x; \lambda_0)$, con $x \in \Omega$, y que se conocen m mediciones con error de $y_1, ..., y_m$, correspondientes a respectivos m elementos de Ω, $x_1, ..., x_m$. Cualquier método matemático que resuelva el problema de hallar alguna aproximación de λ_0 a partir de esta información se conoce como un método de ajuste de parámetros.

Para resolver un problema como el recién descrito debe elegirse (con el suficiente cuidado de atender a las características propias de la familia de funciones, el dominio común) el espacio de parámetros y por supuesto los restantes objetos matemáticos que sean vitales para la formulación y desarrollo de cada método. En particular, los métodos que por ahora nos conciernen (por ser un recurso en el desarrollo del texto) parten de una enunciación como un problema de optimización en espacios euclidianos.

Con cada conjunto dado de vectores en Ω ocurre que cada una de las funciones en $\mathcal{F}$ induce una función con variable independiente corriendo en Λ y rango en $\mathbb{R}^m$, a saber:

$$\lambda \longmapsto f_X(\lambda) = (f(x_1, \lambda), ..., f(x_m, \lambda))^t;$$

donde X es la matriz de orden $m \times n$ con los vectores $x_1^t, ..., x_m^t$ como sus respectivas filas.

El método de ajuste de parámetros por *mínimos cuadrados ordinarios*, **MCO**, es la base para muchos otros y consiste en minimizar de forma conjunta, respecto de λ, la distancia entre la imagen de $f(x_i, \lambda)$ y la respectiva medición y_i, $i = 1, ..., m$, que significa: $\tilde{\lambda} \in \Lambda$ es una solución al problema de ajuste de parámetros bajo el método de MCO si es la solución del problema de optimización:

$$\begin{aligned} \tilde{\lambda} &= min_\lambda \left\| f_X(\lambda) - y \right\|; \\ y = (y_1, ..., y_m)^t, &\qquad \lambda \in \Lambda. \end{aligned} \tag{1.25}$$

Un problema de ajuste de parámetros que se resuelve por mínimos cuadrados en general no tiene garantía de existencia de una solución, de solución única ni de que errores pequeños de cómputo o medición generen diferencias pequeñas entre la solución calculada y el vector real de parámetros λ_0. El método general requiere en muchos casos modificaciones que permitan un buen planteamiento del problema: existencia de solución única que depende de forma continua de los errores en los datos, las mediciones correspondientes a las componentes de y.

Por las hipótesis fundamentales de los análisis QSAR el método de ajuste de parámetros por MCO es de gran relevancia para este tipo de análisis de la farmacología, cuando $\mathcal{F}$ es justamente el espacio dual de $\mathbb{R}^n$. Es entonces pertinente invertir algunas líneas en detallar la forma de resolución para este caso particular del método de MCO.

Para distinguir este caso adoptaremos la notación a para un vector arbitrario de parámetros. El espacio de parámetros se considera como el mismo $\mathbb{R}^n$ y cada función en $\mathcal{F}$ queda caracterizada por la regla de asociación:

$$f(x; a) = x^t a, \quad x \in \Omega, \quad a \in \mathbb{R}^n.$$

Luego, el primer requisito es que X en $M_{m\times n}(\mathbb{R})$ tenga rango completo para garantizar la existencia y unicidad de la solución. El problema (1.25) se reformula como

$$a_0 = min_{a\in\mathbb{R}^n}\left\{\|Xa-y\|\right\}. \tag{1.26}$$

Lo primero es observar que para un par arbitrario de vectores u, v en $\mathbb{R}^n$ y $\lambda \in [0,1]$ se verifica la relación

$$\begin{aligned}\|X(\lambda u+(1-\lambda)v)-y\| &= \|\lambda Xu+(1-\lambda)Xv-\lambda y-(1-\lambda y)\| \\ &= \|\lambda(Xu-b)+(1-\lambda)(v-y)\|;\end{aligned}$$

la desigualdad triangular aplicada al extremo derecho de esta cadena de igualdades prueba que el campo escalar cuya regla de asociación está dada por $\|Xa-y\|$, con variable a, es una función convexa (*def* 1.3.6).

El espacio euclidiano de dimensión n es por supuesto convexo; se concluye de lo anterior que (1.26) es un problema de optimización convexo y por ello tiene solución y es única (*ver secc.* 1.3.1). Para encontrar tal solución basta con hallar un punto en el espacio que verifique condiciones suficientes de optimalidad de primer orden.

La siguiente observación es que el problema (1.26) es equivalente a

$$a_0 = min_{a\in\mathbb{R}^n}\left\{\|Xa-y\|^2\right\}; \tag{1.27}$$

por lo que sustituirá en adelante al problema (1.26).

Continuando con la determinación del elemento en el espacio del dominio que verifique las condiciones de primer orden requeridas se tiene:

$$\begin{aligned}\frac{d\|Xa-y\|^2}{\delta a_k} &= 2\sum_{i=1}^{m}\left[\left(\sum_{j=1}^{n}x_{ij}a_j\right)-y_i\right]x_{ik} \\ &= 2\left[\sum_{j=1}^{n}\sum_{i=1}^{m}x_{ij}x_{ki}^t a_j-\sum_{i=1}^{m}x_{ki}^t y_i\right];\end{aligned}$$

es decir,

$$\nabla\|Xa-y\|^2 = 2\left(X^tXa-X^ty\right);$$

cuando a_0 es solución del problema (1.26) $\nabla\|Xa_0-y\|^2$ se anula. La implicación inmediata es que la solución a_0 es la solución del *sistema de ecuaciones normales*[6] que en su forma matricial queda expresado como

$$X^tXa = X^ty. \tag{1.28}$$

Por otro lado, un resultado conocido (*ver* [8])es el siguiente

[6]Por ser $\mathbb{R}^n$ un espacio de producto interior sobre el campo $\mathbb{R}$ se sabe que la matriz adjunta (conjugada transpuesta) de $A \in M(\mathbb{R})_{m\times n}$, A^*, coincide con la matriz transpuesta de A, A^t. Además el producto matricial AA^* es una matriz normal.

Teorema 1.3.8 (Descomposición matricial en valores singulares)**.** *En el espacio vectorial $M_{m\times n}(\mathbb{R})$ con el producto matricial usual, si $A \in M_{m\times n}(\mathbb{R})$ es una matriz de rango r con valores singulares $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r$ y Σ es la matriz de orden $m \times n$ definida por*

$$\Sigma_{ij} = \begin{cases} \sigma_i, si\ i = j \leq r \\ 0,\ en\ otro\ caso. \end{cases}$$

Entonces, existen matrices unitarias $U \in M(\mathbb{R})_{m\times m}$ y $V \in M(\mathbb{R})_{n\times n}$ tales que

$$A = U\Sigma V^t.$$

Se conoce a esta factorización como la descomposición en valores singulares (DVS) de A. Los vectores $u_1, ..., u_m$ que constituyen las columnas de U se conocen como vectores singulares izquierdos de A; análogamente los vectores $v_1, ...v_n$ correspondientes a las columnas de V se dirán los vectores singulares derechos de A.

Si $U\Sigma V^t$ es la DVS de X se sigue que $X^t = V\Sigma^t U^t$, sustituyendo a X y X^t por sus DVS's en (1.28) el sistema de ecuaciones normales se reescribe como

$$(V\Sigma^t U^t)(U\Sigma V^t)a = (V\Sigma^t\Sigma V^t)a = V\Sigma^t U^t y;$$

pero X tiene rango completo, n, y por ello $\Sigma' = V\Sigma^t\Sigma V^t$ es invertible.

Si $\mathbf{e}_{ij}$ es la matriz cuadrada de orden n para la cual la ij-ésima componente tiene valor 1 y es la única no nula, entonces $\Sigma' = \Sigma^t\Sigma = \sum_{j=1}^n \sigma_j^2 \mathbf{e}_{jj}$, implicando:

$$(X^t X)^{-1} = V\Sigma'^{-1}V^t. \tag{1.29}$$

Entendiendo a $\mathbf{e}_{ij}$ en el espacio adecuado se verifican las igualdades:

$$\begin{aligned} \left(X^t X\right)^{-1} V\Sigma^t U^t &= V\left(\sum_{j=1}^n \frac{1}{\sigma_j}\mathbf{e}_{jj}\right)U^t \\ &= \sum_{j=1}^n \frac{1}{\sigma_j} V\mathbf{e}_{jj}U^t; \end{aligned}$$

de donde se desprende la caracterización de la solución al problema de optimización (1.27)

$$a_0 = \sum_{j=1}^n \frac{u_j^t y}{\sigma_j} v_j \quad . \tag{1.30}$$

Ajuste de funcionales lineales por MCO, un problema inverso mal planteado

El problema de optimización que se detalló en la primera parte de esta sección suele emplearse para la elección de un elemento en la familia de modelos lineales $\{xa : x \in \mathbb{R}^n\}$, con vector de parámetros a, que de forma uniforme explique lo mejor posible el comportamiento de un conjunto de mediciones realizadas, vector y, de un fenómeno que se

presupone guarda una relación aproximadamente lineal determinada por el vector de parámetros a y los posibles valores tomados por la variable x.

En la práctica un error es inherente a cualquier medición de un observable en un fenómeno real que se estudia, y por lo regular sólo se puede conocer una cota para su magnitud; a este tipo de error se le conoce como "error de medición" o más generalmente como una "perturbación en los datos reales". Dependiendo de las particularidades de un ajuste por mínimos cuadrados, pequeñas perturbaciones en el vector de datos y puede resultar en la elección de un modelo que diste considerablemente de la solución real en la norma de interés.

Sea $U\Sigma V^t$ la DVS de X y consideremos el caso en que $\tilde{y} = y + \delta y$, donde δy es la perturbación sobre el dato real, en tal caso, pretendiendo resolver el problema (1.26) empleando una aproximación de los datos reales, se sustituye y por $\tilde{y}$ en la ecuación (1.30) para llegar a una estimación $\tilde{a}_0$ de la solución real:

$$\begin{aligned} \tilde{a}_0 &= \sum_{j=1}^{n} \frac{u_j^t \tilde{y}}{\sigma_j} v_j = \sum_{j=1}^{n} \frac{u_j^t y}{\sigma_j} v_j + \sum_{j=1}^{n} \frac{u_j^t \delta y}{\sigma_j} v_j \\ &= a_0 + \sum_{j=1}^{n} \frac{u_j^t \delta y}{\sigma_j} v_j \end{aligned} \tag{1.31}$$

y por lo tanto, desde que U y V son matrices unitarias, se satisface la relación

$$\|a_0 - \tilde{a}_0\| = \left\| \sum_{j=1}^{n} \frac{u_j^t \delta y}{\sigma_j} v_j \right\| = \left(\sum_{j=1}^{n} \left(\frac{u_j^t \delta y}{\sigma_j} \right)^2 \right)^{\frac{1}{2}}. \tag{1.32}$$

Así, por lar relación del orden en las magnitudes de los valores singulares se verifican las relaciones

$$\begin{aligned} \frac{1}{\sigma_1} \left(\sum_{j=1}^{n} (u_j^t \delta y)^2 \right)^{\frac{1}{2}} &\leq \|a_0 - \tilde{a}_0\| = \left(\sum_{j=1}^{n} \left(\frac{u_j^t \delta y}{\sigma_j} \right)^2 \right)^{\frac{1}{2}} \\ &\leq \frac{1}{\sigma_n} \left(\sum_{j=1}^{n} (u_j^t \delta y)^2 \right)^{\frac{1}{2}}. \end{aligned}$$

Como ya se ha dicho, ajustar un modelo lineal por mínimos cuadrados es un problema con solución única y además, de acuerdo con la última cadena de desigualdades, la solución depende de forma continua de los datos de medición; sin embargo, la cantidad de cálculos a realizar para resolverlo hace que en la mayoría de los casos sea necesario emplear métodos numéricos que en la implementación computacional no siempre garantizan tal continuidad.

Si ocurre que los valores singulares de la matriz que determina el sistema homogéneo son casi nulos o muy pequeños en comparación con $\|\delta y\|$, entonces puede ocurrir que

el cómputo de la solución aproximada con perturbación en los datos no sea una buena aproximación o que para efectos de cómputo el rango de X sea menor que n, como puede apreciarse en (1.33). Cuando este es el caso, se dice que la matriz X está mal condicionada, siendo imprescindible regularizar el problema, que significa reformularlo en un problema equivalente que aún numéricamente dependa continuamente de los datos.

1.4. Regresión lineal Múltiple

El tema del problema de regresión lineal puede consultarse en casi cualquier libro de estadística, en cada texto suelen encontrarse demostraciones de los resultados básicos de acuerdo a la intención y los objetivos de cada autor. No será ésta una excepción.

Lo más común es que las demostraciones y formulación de resultados se base en lo que se conoce como una descomposición QR de una matriz. Esto ocurre porque tal descomposición tiene la ventaja de simplificar los cálculos numéricos permitiendo algoritmos más robustos y eficientes para las aplicaciones de esta herramienta estadística.

El objetivo ahora es distinto, las aplicaciones para las que hacemos todo esto incluye preselecciones de variables independientes que resulten en conjuntos relativamente pequeños de ellos y exigencias de investigación que permiten el lujo de no hacer de la velocidad cómputo una prioridad, siempre que se garantice el uso de *software* que utilice algoritmos confiables respecto de los problemas de optimización que deben resolverse numéricamente y sean inherentes al problema de regresión lineal múltiple.

El desarrollo de esta sección corresponde a la necesidad de un criterio para la determinación le la mayor cantidad de grados de libertad que que sean posibles para una ecuación lineal que resultante de un análisis QSAR, que garantice una varianza aceptable para los coeficientes de tal ecuación, como estadísticos de los coeficientes reales.

En una amplia cantidad de publicaciones sobre métodos QSAR se sugiere que de forma general debe observarse una relación de por lo menos tres a uno entre dichos grados de libertad y la cantidad de moléculas sometidos a bioensayos, estos criterios son de caracter muy general, pese a que la estadística y la teoría de problemas inversos ya consideran formas de aprovechar toda la información disponible.

Conocer información sobre la acotación de la varianza del error de medición y la propia matriz de datos de los descriptores moleculares son por si mismos datos relevantes, que pueden incluirse en el problema de regresión con poca dificultad. Tal formulación es en realidad la consecuencia de ver el problema de regresión desde la perspectiva de la teoría de problemas inversos discretos, para estos tópicos se sugiere al lector el libro de Hansen [10]. Ésta formulación justamente la que se desprende de la que ya se ha empleado para expresar la solución al problema de ajuste de parámetros por **MCO**.

El criterio de selección de descriptores está basado en el estadístico R^2_{ajus}, común en la literatura, y pertenece los criterios de tipo *paso a paso*. Es una forma en que puede generalizarse la hipótesis de linealidad a espacios de dimensión infinita, espacios de Hilbert, resultando en la prueba formal de la que tiene cambiar a la solución usual de los problemas de regresión por aquella que se conoce como la solución de Tikhonov.

Un resultado común en la literatura es que por la naturaleza del problema de investigación que representa la metodología QSAR para la farmacología, es posible establecer

un primer criterio de determinación de la cantidad máxima de variables independientes de un modelo derivado de una prueba débil de hipótesis de linealidad, criterio que será preferible siempre que se disponga de conjuntos muestrales pequeños, esto cuando se supone normalidad para el problema de regresión, que no es poco usual en la práctica.

Se expondrá cómo es que bajo las condiciones de los análisis QSAR la solución de Tikhonov maximiza la bondad de ajuste siempre que las mediciones de los descriptores y los de actividad sean, desde el punto de vista estadístico, estimadores no sesgados de los respectivos valores reales; sin embargo, cuando por cualquier motivo deba recurrirse a la solución tradicional de problema de regresión lineal múltiple, **RLM**, entonces es que se propone una definición de sobre ajuste de parámetros que respecto de la relación entre cantidad de descriptores, n, y tamaño de muestra, m, sólo esté condicionada por la relación $n < m$.

Suponiendo normalidad en el problema de regresión, para la determinación de la relación entre n y m bastará con la prueba de hipótesis mencionada en párrafos anteriores.

Consideremos el caso en que $a_1, a_2, ..., a_n$ son parámetros desconocidos y se dispone de datos muestrales $y_i, x_{i1}, x_{i2}, ..., x_{in}$, para $i = 1, ..., m$, donde y_i es una medición u observación de la variable y dado que $x_1 = x_{i1}, ..., x_n = x_{in}$. En esta situación una forma de estimar el vector de parámetros, $a = (a_1, ..., a_n)^t$, es mediante un ajuste por mínimos cuadrados ordinarios con solución $\tilde{a} = (\tilde{a}_1, ..., \tilde{a}_n)^t$.

Por comodidad en la notación, y por considerar suficiente una advertencia para evitar confusiones posteriores, emplearemos las notaciones $y, x_1, ..., x_n$ de forma indistinta para referirnos a las variables descritas, y a los vectores en $\mathbb{R}^m$ con componentes correspondientes a una ordenación particular de una muestra de tamaño m de la respectiva variable: $y = (y_1, ..., y_m)^t$ y X es la matriz de orden $m \times n$ con x_{ij} por ij-ésima componente y descomposición en valores singulares $X = U\Sigma V^t$.

Por (1.30), el estadístico del vector de parámetros está dado por

$$\tilde{a} = \sum_{j=1}^{n} \frac{u_j^t y}{\sigma_j} v_j.$$

El conjunto de vectores $(y_i; x_{i\cdot}) = (y_i, x_{i1}, ..., x_{in})$ será considerado como una muestra independientemente del modelo lineal que por hipótesis describe al fenómeno observado del vector y, es decir, cada fila de la matriz X junto con su respectiva medición de y asociada a dicha fila constituirán una muestra de tamaño uno del fenómeno de interés, independiente de las filas y mediciones restantes.

Se sigue que $\tilde{a}$ es un estadístico no sesgado con matriz de covarianza dependiente de σ y las características de X como lo muestra el siguiente teorema (*ver* [21]).

Teorema 1.4.1. *Denotando por c_{ij} a la ij-ésima componente de la matriz $(X^tX)^{-1}$ para X, Y y $\tilde{a}$ descritos con anterioridad; entonces*

I. $E(\tilde{a}|X) = a$;

II. $Cov((\tilde{a}_i, \tilde{a}_j)|X) = \sigma^2 c_{ij}$.

Si X es de rango completo y los vectores $(y_i, x_{i1}, ..., x_{im})$ son observaciones independientes.

Denotando por C a $(X^tX)^{-1}$ se tiene el siguiente corolario.

Corolario 1.4.2. *Si $x = (x_1, ..., x_n)$ es un vector de variables aleatorias, entonces*

$$Var(\tilde{a}_1 x_1 + \cdots \tilde{a}_n x_n | X) = \sigma^2(x_1, ..., x_n)C(x_1, ..., x_n)^t.$$

Es común que para el análisis estadístico de variables aleatorias se desconozcan parámetros como la media y la varianza de las mismas, por lo que se requieren estadísticos que las aproximen con una alta probabilidad, preferentemente estadísticos no sesgados; los modelos de regresión lineal $y \approx \tilde{a}_1 x_1 + \cdots \tilde{a}_n x_n$ no son la excepción. Un Estimador no sesgado de σ se presenta a continuación.

Teorema 1.4.3. *Si $(y_i, x_{i1}, ..., x_{in})$ son observaciones independientes del modelo de regresión lineal para $i = 1, ..., m$, $var(y|X) = \sigma$, y s^2 se define por*

$$s^2 = \frac{1}{m-n} \left\| \tilde{y}^t - X\tilde{a}^t \right\|^2;$$

entonces,

$$E(s^2|X) = \sigma^2.$$

A menudo también existe un término independiente en el modelo lineal que se desea ajustar al conjunto de mediciones, en estos casos todos los teoremas y corolarios presentados en las secciones anteriores se verifican sin modificación alguna. Lo convencional en la literatura es considerar a una de las variables x_1 o x_n como la constante idénticamente 1. En adelante serán estos modelos lineales los que se trabajarán, por comodidad asumiremos $x_n \equiv 1$, salvo excepciones debidamente indicadas.

1.4.1. Sobre ajuste en los modelos de regresión

El problema de regresión lineal, como se ha expuesto, consiste en resolver un problema de ajuste de parámetros por MCO, donde la solución del ajuste es una observación (muestra de tamaño 1) de un estadístico del vector de parámetros a.

Un estadístico es una función de la muestra de una variable aleatoria de interés, es en si mismo una variable aleatoria que se espera aglutine en un dato información relevante almacenada implícitamente en la muestra. Cuando se estiman parámetros de una variable estocástica mediante un estadístico se espera que éste observe un buen comportamiento que haga confiable la estimación del parámetro que interesa aproximar. Inherentes a un estadístico existen dos magnitudes que evalúan que tan confiable es el estadístico, sesgo y varianza. Considerando estos valores y sus relaciones cualitativas se consigue información relevante sobre el comportamiento de un estadístico.

Si el sesgo es grande y la varianza es significativamente menor que éste, entonces la probabilidad de aproximar satisfactoriamente al parámetro empleando el correspondiente estadístico es significativamente baja, de modo que no sería prudente confiar la aproximación a ese estadístico. En estos casos se dice que existe *escaso ajuste*[7] por parte del estadístico.

[7] *lowfitting*, en la literatura en inglés

El estadístico tampoco es bueno si su varianza es muy grande pese a tener un sesgo pequeño respecto del parámetro que estima. En este caso la ley de los grandes números nos garantiza una buena aproximación pero siempre y cuando pueda disponerse de una muestra suficientemente grande del estadístico, ello resulta poco útil en la práctica. Se dice que existe "sobre-ajuste"[8] en un estadístico si presenta un sesgo pequeño y varianza significativamente elevada.

En general, si ϕ es un parámetro y $\widehat{\phi}$ es un estadístico de ϕ, suele definirse como el *índice de bondad de ajuste* como $E(||\phi-\widehat{\phi}||^2)+Var(||\phi-\widehat{\phi}||^2)$. Entre menor sea el índice de bondad de ajuste se dice que el estadístico tiene una mejor bondad de ajuste.

Para los problemas de regresión lineal que trabajamos aquí, la hipótesis de que el error de medición tiene una distribución con media 0 conduce a estimadores no sesgados (*teorema* 1.4.1) del vector de parámetros, de lo que debemos ocuparnos entonces, es de proporcionar argumentos que respalden una varianza lo más pequeña posible para el estadístico de regresión obtenido por MCO. Para tales fines disponemos de un corolario del teorema 1.4.1 y es la descomposición en valores singulares de una matriz.

Corolario 1.4.4. *Bajo las hipótesis del teorema 1.4.1 se verifica la relación de orden*

$$\frac{\sigma^2}{\sigma_1^2} \leq Var(\tilde{a}_j) \leq \frac{\sigma^2}{\sigma_n^2}. \tag{1.33}$$

Demostración. Por el teorema 1.4.1:

$$Var(\tilde{a}_j) = \sigma^2 c_{jj};$$

mientras que por el teorema de descomposición en valores singulares se satisface la igualdad

$$c_{jj} = \sum_{k=1}^{n} \frac{v_{jk}^2}{\sigma_j^2}, \quad j = i, ..., n;$$

donde σ es la varianza de δy en el problema de regresión, c_{jj} es el j-ésimo elemento en la diagonal de $(X^tX)^{-1}$, $X = U\Sigma\ V^t$ es la DVS de X y $\sigma - 1 \geq \cdots \geq \sigma_n$ son los valores singulares de X.

El resultado se sigue de las relaciones $\sigma_1 \geq \cdots \geq \sigma_n$ y el hecho de que cada vector singular derecho es unitario. □

El corolario (1.4.4) es consistente con el planteamiento de problema inverso, puesto que si X es, por ejemplo, una matriz ortonormal entonce X^tX es la matriz identidad de orden $n \times n$ y por lo tanto todos sus valores singulares son la unidad real, significando que una varianza pequeña en los datos de medición produce una varianza del mismo orden en cada uno de los estadísticos que estiman los parámetros dados por a.

Casi para terminar con este bloque, resalta que la existencia de sobre-ajuste por parte del vector que resuelve el problema de regresión lineal, conforme al corolario 1.4.4, depende del buen condicionamiento de la matriz X así como de la varianza del error de medición.

[8] *overfitting*, en la literatura en inglés.

Salvo en un primer curso de estadística es poco común encontrar situaciones en las que la varianza teórica del error de medición, σ^2, se conozca; fenómeno que nos indica que, tanto el sesgo como la varianza de un estadístico de regresión, dependerán también de los valores de las mismas características para el estadístico que sea utilizado para aproximar a σ^2. Normalmente la elección es s^2 debido a que es no sesgado, siendo entonces la varianza de s^2 el único valor adicional que debe tomarse en cuenta para prevenir un error por sobre estimación de los parámetros.

En ocasiones, existe información adicional sobre el error de medición δy en un problema de regresión, siendo una hipótesis su buen comportamiento estadístico, es decir, el error de medición tiene un comportamiento simétrico alrededor de su media de tal suerte que su distribución de probabilidad puede considerarse indistinguible de una distribución normal con media 0 y varianza σ^2; por ejemplo, en el caso de los análisis QSAR, esta es una hipótesis que será considerada pues los laboratorios que realizan las mediciones de interés así lo garantizan.

Para nosotros, en consecuencia, un problema de regresión lineal considerará desde ahora la hipótesis de normalidad en el error de medición, a saber:

> La distribución de probabilidad de y, condicionada por los valores de las variables de predicción $x_1, ..., x_n$, es normal con parámetros $\sum_{j=1}^{n} a_j x_j$ y σ^2, dicho de otra forma, $(N(\sum_{j=1}^{n} a_j x_j, \sigma^2))$.

De la inclusión de normalidad para el error de medición se desprenden resultados adicionales (*ver* [21]) que nos permitirán decir algo más sobre cómo evitar el sobre-ajuste del estadístico s^2, y por consiguiente sobre la posible sobre estimación en la aproximación de los parámetros del problema de regresión.

Teorema 1.4.5. *Si la distribución condicional de y dado* $x_1, ..., x_n =$ *es* $N(a_1x_1 + \cdots + a_nx_n, \sigma^2)$, $x = (x_1, ...x_n)$, $a = (a_1, ..., a_n)$; *y los vectores y y* $\tilde{a}$ *así como las matrices* X *y* C *son las ya descritas, entonces*

I. $\tilde{a}_i \sim N(a_i, \sigma^2 c_{ii})$;

II. $\tilde{a}x^t \sim N(ax^t, \sigma^2(xCx^t))$;

III. $(m-n)\frac{S^2}{\sigma^2} \sim \chi^2(m-n)$, *y es es independiente respecto de* $\tilde{a}$;

donde S^2 *es la variable aleatoria condicionada a los valores de* X *y* y *que modela los valores que puede tomar* s^2.

Corolario 1.4.6.

I. $\frac{(\tilde{a}_i - a_i)}{\sqrt{s^2 c_{ii}}} \sim t(m-n)$;

II. $\frac{(\tilde{a}-a)x}{\sqrt{S^2(x^tCx)}} \sim t(m-n)$.

Recordamos también una célebres desigualdad

Teorema 1.4.7 (Desigualdad de Chebychev)**.** *Sea Z una variable aleatoria arbitraria con media μ_Z y varianza σ_Z. Si $\lambda > 0$, entonces*

$$P(|Z - \mu_Z| \geq \lambda) \leq \frac{\sigma_Z}{\lambda^2}.$$

Supongamos ahora que $\epsilon \geq 0$; utilizando el teorema 1.4.5 (consecuencia de la recién incluida condición de normalidad en el error de medición) y la desigualdad de Chabychev, se obtiene:

$$P\left(|s^2 - \sigma^2| \geq \epsilon\right) = P\left((m-n)|\frac{s^2}{\sigma^2} - 1| \geq (m-n)\frac{\epsilon}{\sigma^2}\right) \tag{1.34}$$

$$\leq \frac{2(m-n)\sigma^4}{(m-n)^2\epsilon^2} = \frac{2\sigma^4}{(m-n)\epsilon^2}. \tag{1.35}$$

La desigualdad (1.35) es una relación útil. Volviendo al problema de sobre-ajuste de parámetros y recuperando la acotación para los parámetros del ajuste por MCO expuesta en (1.4.4); entonces, (1.35) ofrece información cualitativa sobre la relación que debe verificarse, para la diferencia entre la cantidad de muestras y la de variables de predicción empleadas.

Para que el estadístico s^2 pueda ser considerado como una buena aproximación del valor teórico de σ^2, siendo posible recuperar, con cierto nivel de confianza $(1-\alpha)$, acotaciones para la varianza de los parámetros ajustados que brinden información cuantitativa acerca del sobre-ajuste de los mismos:

$$m - n \geq \frac{2\sigma^4}{\alpha\epsilon^2}.$$

Se sigue que, si el orden del valor teórico de la varianza del error de medición puede ser considerado como el orden de una buena aproximación de dicha varianza; entonces, la confianza en que s^2 es un buen estadístico recaerá en la diferencia $m - n$. Par $\epsilon = \sqrt{2}\sigma^2$ se satisface, por (1.35):

$$P\left(|s^2 - \sigma^2| \geq \epsilon\right) \leq \frac{1}{m-n};$$

que bajo estas condiciones prueba, por implicación directa del corolario (1.4.4) y si $\lambda \geq 0$ es un escalar que acota superiormente a σ:

$$P\left(\frac{s^2 - \sqrt{2}\lambda^2}{\sigma_1} \leq Var(\tilde{a}_j) \leq \frac{s^2 + \sqrt{2}\lambda^2}{\sigma_n}\right) \geq 1 - \frac{1}{m-n}, \quad j = 1, ..., n. \tag{1.36}$$

Si bien el valor real de σ^2 es un dato desconocido, no necesariamente lo es el orden de su magnitud; dependiendo de la naturaleza del observable y su medición pueden existir acotaciones conocidas de su varianza. Ejemplo de ello es nuestro caso. Recordemos que

hemos supuesto que en un análisis QSAR la aleatoriedad corresponde al error de medición de la actividad biológica, por la hipótesis de linealidad; entonces σ^2 es un valor cuyo orden es menor que la unidad y tal acotación es garantizada por los laboratorios que realizan las mediciones experimentales en observación de los respectivos criterios de calidad. El escalar λ será para nosotros un valor menor que el final estudio de caso.

1.4.2. Hipótesis de linealidad

Los resultados del problema de regresión lineal múltiple se garantizan siempre y cuando se satisfagan las hipótesis principales de este método estadístico pero, cómo saber si las hipótesis realmente se verifican.

Validar un modelo de regresión lineal puede hacerse si existe información determinista o estadística sobre el fenómeno que se modela, cuando se hacen análisis QSAR la validación es de carácter estadístico.

Pensemos en que el observable y depende de las variables $x_1, ..., x_n$ de tal forma que existe un funcional f, en el espacio euclidiano de dimensión n, que cumple con la condición $f(x) = y$, donde $x = (x_1, ..., x_n)$. Por lo general el funcional como modelo matemático del observable sólo tiene sentido en la realidad para un subconjunto acotado $\Omega \subseteq \mathbb{R}^n$ y cuando f es una función acotada, es este el caso en que trabajaremos.

La propiedad de ser acotados del conjunto Ω y del funcional f son suficientes para garantizar que cualquier potencia positiva del valor absoluto de f es de Lesbegue integrable en Ω y, por lo tanto, existe un único vector a_0 en $\mathbb{R}^n$ que caracteriza al funcional lineal que, en la norma uniforme, mejor aproxime a f sobre Ω, es decir,

$$sup_{x\in\mathbb{R}^n}\left\{\left\|a_0^t x - y\right\|\right\} \leq sup_{x\in\mathbb{R}^n}\left\{\left\|a^t x - y\right\|\right\}, \forall a, x \in \nleq.$$

Si $\delta y = f(x) - a_0^t x$, entonces es inmediato que $y = a_0^t x + \delta y$. Por lo general se tiene poca o nula información respecto al comportamiento de f, tampoco de los valores de las componentes del vector de parámetros a_0; así que decidimos, entonces, que δy será una variable aleatoria definida en Ω, con media $\mu_x = f(x) - a_0^t x$ y varianza σ. Hay que recalcar que cualquier error de medición, cometido al momento de calcular el valor de un estado del observable y, se incluye implícitamente en la variable δy; el error de medición aún es considerado como una variable aleatoria independiente de las variables de predicción, con distribución normal y media 0.

Nuevamente llegamos a una formulación estadística de un problema de ajuste de parámetros, la única diferencia es que la esperanza condicionada de y no es una combinación lineal de los valores dados para las componentes de la variable vectorial x. La mecánica del ajuste de parámetros por MCO que se ha desarrollado con anterioridad no se ve alterada en lo más mínimo; dada una matriz de datos de las variables, para las respectivas m muestras del observable y, el estadístico $\tilde{a}$ se se obtiene sin modificación metodológica alguna.

Lo ideal sería poder ser capaces de rechazar la hipótesis nula en una prueba como la siguiente:

$$H_1\text{: } y = a_0^t x \text{ contra } H_0\text{: } y \neq a_0^t x;$$

no obstante, al ser tan general la prueba y suponer que H_0 es verdadera, se convierte en una tarea muy complicada encontrar un estadístico que ofrezca útiles acotaciones para la medida de probabilidad de la región de rechazo. Además, con la prueba de hipótesis formulada así el propio estadístico de prueba no es evidente, será conveniente encontrar una prueba asequible en su tratamiento.

Una forma de atacar el problema es debilitar la prueba al intercambiar las hipótesis entre ellas (H_0 y H_1), buscando posteriormente una prueba equivalente que resuelva nuestro problema de la elección del estadístico de prueba. La hipótesis nula será entonces la dependencia lineal:

$$H_1{:}y \neq a_0^t x \text{ contra } H_0{:}\ y = a_0^t x.$$

En una prueba de hipótesis, con región de rechazo Γ, la regla de decisión bajo la cual se acepta o no la hipótesis alternativa H_1 es como sigue:

> Si el estadístico de prueba es elemento de la región de rechazo Γ, con probabilidad α; entonces se rechaza la hipótesis nula H_0 (se acepta H_1 con un nivel de confianza $1-\alpha$).

El intercambio que hemos realizado entre las hipótesis de la prueba sobre linealidad ha ocurrido en virtud de que la anterior regla de decisión es equivalente a su proposición contrarrecíproca:

> Si H_0 no es rechazada (no se acepta H_1 con un nivel de confianza $1-\alpha$), entonces el estadístico de prueba es elemento de la región de rechazo Γ^c, con probabilidad $1-\alpha$.

Al cambiar la prueba, cambiamos una condición estadística suficiente para validar la hipótesis de linealidad por una condición necesaria; la condición estadística de suficiencia es ahora sobre la hipótesis de no linealidad, característica de la prueba que no debemos olvidar. Si la evidencia estadística no permite rechazar la hipótesis nula no significa que pueda aceptarse la dependencia lineal de y respecto de las variables de predicción. Con una alegoría coloquial, decimos que liberar por insuficiencia de pruebas a una persona indiciada como presunto culpable de un delito no significa que sea inocente, sólo significa que no fue posible hallar las pruebas suficientes para probar su culpabilidad.

Antes de continuar enunciaremos algunos resultados que nos serán de utilidad (*ver* [21]).

Teorema 1.4.8. *Para modelos de regresión lineal con término independiente no nulo se verifica* $\tilde{a}_n = \overline{y} - \sum_{j=1}^{n-1} \tilde{a}_j \overline{x}_j$ *y*

$$\begin{aligned}\sum_{i=1}^{m}(y_i-\overline{y})^2 &= \sum_{i=1}^{m}\left(\tilde{a}_1(x_{i1}-\overline{x}_2)+\cdots+\tilde{a}_{n-1}(x_{i(n-1)}-\overline{x}_{(n-1)})\right)^2 \\ &\quad+\sum_{i=1}^{m}\left(y_i-\tilde{a}_1x_{i1}+\cdots+\tilde{a}_nx_{in}\right)^2;\end{aligned}$$

donde $\overline{y}$ *denota la media muestral del vector* y.

Lema 1.4.9. *Sean l un natural menor que n y $\tilde{a} \in \mathbb{R}^n$ el vector de coeficientes de un modelo de regresión lineal con término independiente no nulo. Si $a_1 = \cdots = a_l = 0$, $\tilde{a}' \in \mathbb{R}^{n-l}$ es el vector de coeficientes correspondiente al modelo de regresión lineal con variables de predicción $x_{l+1}, ..., x_n$, para el mismo grupo de muestras $y_1, ..., y_m$, y*

$$E(y|X) = \beta_{l+1} = a_{l+1}x_1 + \cdots + a_n x_n;$$

entonces se satisface la relación de orden

$$\sum_{i=1}^{m} \left(\tilde{a}_1(x_{i1} - \overline{x}_1) + \cdots + \tilde{a}_{n-1}(x_{i(n-1)} - \overline{x}_{(n-1)})\right)^2 \geq$$
$$\sum_{i=1}^{m} \left(\tilde{a}'_{l+1}(x_{i(l+1)} - \overline{x}_{l+1}) + \cdots + \tilde{a}'_{n-1}(x_{i(n-1)} - \overline{x}_{(n-1)})\right)^2.$$

De forma conjuntan el teorema (1.4.8) y el lema (1.4.9) son una advertencia y no deben pasarse por alto al momento de intentar explicar el comportamiento o la tendencia de un fenómeno. Incluir en un modelo de regresión variables que no expliquen de forma lineal el comportamiento de la tendencia de y, al fijarse los valores del resto de las variables de predicción, puede mejorar el ajuste por mínimos cuadrados, mejorando la correlación estadística pese a que no exista una correlación real entre el modelo de regresión y el observable de interés.

El lema 1.4.9 se presenta, como se ha hecho, para enfatizar el cuidado que merece la selección de adecuadas variables de descripción en la regresión lineal; sin embargo, la desigualdad que guarda su consecuente es válida en un sentido más general, aún cuando a las primeras l variables independientes correspondan coeficientes no nulos en el modelo de regresión. El caso general de dicho lema será necesario más adelante así que nos detendremos un poco en ello.

Primero, por teorema 1.4.8 la desigualdad que sostiene el lema 1.4.9 es equivalente a

$$\|X\delta a - \delta y\|^2 \leq \|X\delta a' - \delta y\|^2;$$

desigualdad que en general tiene sentido, no importando si en el problema de regresión lineal el término independiente es nulo, para esta magnitud enunciamos los resultados siguientes.

Lema 1.4.10. *Sean X una matriz real de orden $m \times n$ y rango completo, $M^{(k)}$ la matriz de orden $m \times k$, con $k \leq n$, tal que las columnas de $M^{(k)}$ son las primeras k columnas de M en el mismo orden, $\tilde{y} \in \mathbb{R}^m$ y $SEC(x_1, ..., x_k) = \left\|\tilde{y} - X^{(k)}\tilde{a}^k\right\|^2$, donde $\tilde{a}^k$ es la solución por MCO al problema de ajuste de parámetros*

$$X^{(k)}a \approx y.$$

Entonces $SEC(x_1, ..., x_j) \geq SEC(x_1, ..., x_k)$, siempre que $j \leq k$.

Demostración. Primero, trabajaremos de forma equivalente con las transformaciones lineales L_k, $k = 1, ..., n$, asociadas respectivamente con las matrices $X^{(k)}$ en los espacios adecuados. Es claro que L_k es una exención de L_j al subespacio $Dom(L_k)$, toda vez que j sea menor o igual que k. Por otro lado, por la forma en que se define el problema de ajuste de parámetros por MCO es inmediato que

$$\left\|\tilde{y} - L_k(\tilde{a}^k)\right\|^2 \leq \|\tilde{y} - L_k(a))\|^2, \quad \forall a \in Dom(L_k);$$

el lema se prueba simplemente notando que $a^j \in Dom(L_j) \subset Dom(L_k)$. □

Lema 1.4.11. *Sean V y W dos espacios vectoriales de producto interior sobre el campo de los números reales y dimensión finita n y m respectivamente, $m \geq n$. Si $L : V \longrightarrow W$ es lineal e inyectiva con valores singulares positivos $\sigma_1 \geq \cdots \geq \sigma_n$; entonces;*

$$\sigma_n = min\{\|L(x)\| : x \in V, \|x\| = 1\} \quad y \quad \sigma_1 = max\{\|L(x)\| : x \in V, \|x\| = 1\}.$$

Demostración. Sean $u_1, ..., u_m \in W$ los vectores singulares izquierdos de L y $v_1, ..., v_n \in W$ los respectivos vectores singulares derechos. Por definición de la norma euclidiana y por las propiedades de los espacios de producto interior (*proceso de ortogonalización de Gram-Schmidt, ver* [8])tenemos, para $x \in V$,

$$\begin{aligned} \|L(x)\|^2 &= \left\| L(\sum_{j=1}^{m} < v_j, x > v_j) \right\|^2 = \sum_{j=1}^{m} < v_j, x >^2 \|L(v_j)\|^2 \\ &= \sum_{j=1}^{n} < v_j, x >^2 \sigma_j^2 \|u_j\|^2 = \sum_{j=1}^{n} < v_j, x >^2 \sigma_j^2. \end{aligned}$$

Del hecho de que los valores singulares decrecen respecto de su subíndice se infiere

$$(\sigma_n||x||)^2 = \sigma_n \sum_{j=1}^{n} < v_i, x >^2 \leq \|L(x)\|^2 \leq \sigma_1 \sum_{j=1}^{n} < v_j, x >^2 = (\sigma_n||x||)^2.$$

El resultado se logra evidenciando que $||u_1|| = \cdots = ||u_j|| = 1$ y que $||L(u_i)|| = \sigma_i$ para $i = 1, ..., m$. □

Corolario 1.4.12. *Para los objetos y bajo las mismas premisas que en el antecedente del lema 1.4.10; si $U\Sigma V^t$ es la descomposición en valores singulares de X y $U^{(k)}\Sigma^{(k)}V^{(k)}t$ lo es respectivamente para $X^{(k)}$, entonces*

$$\sigma_1^{(k)} \geq \sigma_i^{(j)} \geq \sigma_k^{(k)}, \; i = 1, ..., j; \; k = 1, ..., n.$$

Demostración. La prueba es una consecuencia del lema 1.4.11 y del hecho de que, para $k \geq j$ la transformación lineal asociada a $X^{(k)}$ es una extensión o prolongación de la respectiva transformación asociada a $X^{(j)}$. □

Continuando, los términos

$$SCR(x_1,...,x_{n-1}) = \sum_{i=1}^{m} \left(\tilde{a}_1(x_{i1} - \overline{X}_{\cdot 2}) + \cdots + \tilde{a}_{n-1}(x_{i(n-1)} - \overline{X}_{\cdot(n-1)})\right)^2$$

y

$$SEC(x_1,...,x_{n-1}) = \sum_{i=1}^{m} (y_i - \tilde{a}_1 x_{i1} + \cdots + \tilde{a}_n x_{in})^2$$

del teorema (1.4.8) se conocen respectivamente como *suma de cuadrados de regresión* y *suma de errores cuadrados*. Son expresiones empleadas en la validación del modelo de regresión lineal como se verá.

En probabilidad que la correlación de dos variables aleatorias sea 1 es equivalente a que una es la traslación de un múltiplo escalar de la otra, es decir, alrededor de sus respectivas medias el comportamiento de una es exactamente el mismo que un múltiplo escalar de la otra; es por esto que la hipótesis de linealidad de los modelos de regresión lineal es equivalente a que la correlación del modelo con la variable de interés, $\sqrt{\rho}$, es justamente 1.

El coeficiente de determinación del modelo de regresión con término independiente no nulo se define por

$$R^2 = \frac{SCR(x_1,...x_{n-1})}{\sum_{i=1}^{m}(y_i - \overline{y})^2} = 1 - \frac{SEC(x_1,...,x_n)}{\sum_{i=1}^{m}(y_i - \overline{y})^2}.$$

Puede probarse que el coeficiente de determinación de un modelo de regresión lineal múltiple, con término independiente no nulo, es un buen estimador del cuadrado del coeficiente de correlación estadística entre el modelo obtenido por un ajuste de mínimos cuadrados y el vector de observaciones y, así la validación del modelo a partir de pruebas de hipótesis sobre la correlación estadística podría hacerse utilizando a R^2 como estadístico de ρ, sin perder de vista sus limitaciones inherentes.

Del teorema (1.4.8) se aprecia que el coeficiente de determinación es un valor positivo no mayor que 1 cuando se considera a una variable predictora como constante en el modelo de regresión lineal. El coeficiente de determinación guarda una relación inversa con la suma de errores cuadrados y es idéntico a 1 si y solamente si $SEC(X_1,...X_{n-1}) = 0$; por lo tanto el coeficiente de determinación R^2 también suele usarse para la validación del modelo de regresión, validando el modelo para valores de R^2 suficientemente [9] próximos a la unidad.

Como función de la cantidad de variables empleadas para describir el comportamiento del observable y, el estadístico R^2 es monótono no decreciente como consecuencia del lema 1.4.10, lo que puede contribuir a validar un modelo con variables que realmente no explican el comportamiento de y.

Para mejorar este estadístico es de uso corriente el *coeficiente de correlación ajustado*

[9] La suficiencia en la proximidad de R^2 con 1 depende de las características propias del fenómeno que se modela.

$$R^2_{ajus} = 1 - \frac{(m-1)SEC(x_1,...,x_n)}{(m-n)\left\|y - \overline{y}1_m\right\|^2} \tag{1.37}$$

Esta modificación en el coeficiente de determinación garantiza que el nuevo estadístico es 1 si y solamente si R^2 lo es también, logrando además que al considerar una variable de predicción adicional en un modelo de regresión, el coeficiente de determinación ajustado incremente sólo si

$$SEC(x_1,...,x_{n-1}) - SEC(x_1,...,x_n) \geq \frac{SEC(x_1,...,x_{n-1})}{m-n}.$$

La observación obligada sobre este nuevo estadístico será sobre la diferencia entre el coeficiente de determinación y éste:

$$0 \leq R^2 - R^2ajus = \frac{n-1}{m-n}\frac{SEC(x_1,...,x_n)}{\left\|y - 1_m\overline{y}\right\|^2};$$

aquí es muy clara la desventaja de la regresión lineal.

Para garantizar buenos ajustes y confianza en el coeficiente de determinación usual es necesaria una gran cantidad de muestras para lograr asegurar que este coeficiente refleje información real sobre la correlación entre las variables de predicción y la variable dependiente.

Por otro lado, cuando las variables de predicción empleadas reflejan información relevante sobre la correlación y ésta sea elevada, entonces la diferencia entres R^2 y R^2ajus será pequeña, significando que el error de medición es significativamente menor que la varianza de la variable independiente. El coeficiente de determinación ajustado no necesariamente es monótono pero si es siempre menor o igual que el coeficiente de correlación estadístic. Esto es una ventaja debido a que, en cualquier prueba de hipótesis como la que pronto ocupará nuestra atención, si la hipótesis nula $R^2 \leq \rho_0$ es rechazada, entonces una prueba con un mayor nivel de confianza se obtiene si en el criterio de rechazo se sustituye a R^2 por R^2_{ajus}.

Para una muestra de tamaño m y n variables de predicción, $n < m$, incluido el término independiente no nulo cuando sea el caso, y tales que R^2_{ajus} sea una función no decreciente de los modelos con $1,...,n$ variables, entonces

$$\begin{aligned} SEC(x_1) - SEC(x_1,...,x_n) &= \sum_{j=1}^{n-1}(SEC(x_1,...,x_j) - SEC(x_1,...,x_{j+1})) \\ &\geq \sum_{j=1}^{n-1}\frac{SEC(x_1,...,x_j)}{m-j} \geq 0; \end{aligned}$$

de esto se sigue que, sumando $SEC(x_1,...,x_n)$ en cada eslabón de la anterior cadena de desigualdades y por la relación de orden que establece el lema 1.4.10,

$$\begin{aligned} SEC(x_1) &\geq \sum_{j=1}^{n} \frac{SEC(x_1,...,x_j)}{m-j} \\ &\geq SEC(x_1,...,x_n) \sum_{j=1}^{n} \frac{1}{m-j} \\ &\geq SEC(x_1,...,x_n) \sum_{j=m-n}^{m-1} \frac{1}{j} \\ &= SEC(x_1,...,x_n) \int_{j=m-n}^{m-1} \frac{1}{t} dt \\ &= SEC(x_1,...,x_n) \ln\left(\frac{m-1}{m-n}\right); \end{aligned} \tag{1.38}$$

atendiendo sólo a la relación que guardan los extremos de la cadena de relaciones, y multiplicando por la $\ln\left(\frac{m-1}{m-n}\right)$ se obtiene

$$\frac{SEC(x_1)}{\ln\left(\frac{m-1}{m-n}\right)} \geq SEC(x_1,...,x_n),\ n>1,\ m>n. \tag{1.39}$$

Esta acotación revela información valiosa sobre el coeficiente de correlación ajustado. El logaritmo natural diverge cuando su argumento tiende a ∞ y para cualquier natural N es cierto que $\frac{m-1}{m-n} > N$ sólo cuando $m > Nr-1$. Significa que, en un proceso en el que se crean modelos nuevos de regresión a partir de incluir en cada paso una nueva variable de descripción a un conjunto de variables empleado en el paso anterior, $SEC(x_1,...,x_n)$ eventualmente se vuelve casi nulo.

Con todo lo hasta ahora dicho sobre el coeficiente de correlación ajustada vemos que es una opción del estadístico de prueba que necesitábamos para la hipótesis de linealidad. La prueba queda como sigue

$$H_1 : \rho_0 < \rho \text{ contra } H_0\text{: } \rho_0 \geq \rho;$$

para un valor de ρ_0 significativamente próximo a 1, empleando el estadístico de prueba R^2_{ajus} y con una región de rechazo $\{R^2_{ajus} \leq \rho_0\}$.

De acuerdo con nuestro planteamiento, δy tiene media condicional $\mu_x = f(x) - a_0^t$ y varianza σ; además, implícitamente consideramos el error de medición en el funcional f, por lo tanto, suponer verdadera la hipótesis nula ahora se traduce en suponer que las hipótesis del problema de regresión lineal se satisfacen, es decir, media idénticamente nula (independiente de las variables de predicción) y varianza pequeña.

En adelante supondremos verdadera la hipótesis $H_0 : y = \sum_{j=1}^{n} a_{0_j} x_j$.

Lema 1.4.13 (Tamaño de muestra)**.** *Sean (y_i, x_{i1}, x_{in}), $i = 1,...,m$, una muestra independiente del problema de regresión, $y_i \geq y_j$ si $i \geq j$. Si y es una variable acotada y el diámetro de su rango está acotado superiormente por $\lambda > 0$; entonces, para la prueba de hipótesis*

$$H_1 : \rho_0 < \rho \text{ contra } H_0\text{: } \rho_0 \geq \rho;$$

con estadístico de prueba R^2_{ajus} y región de rechazo $RR = \{R^2_{ajus} < \rho\}$, si la región de rechazo tiene probabilidad menor que α, entonces

$$m - n > \frac{3m\chi_\alpha\sigma^2}{2\lambda^2(2m-1)(1-\rho)};$$

donde $\chi_\alpha \geq 0$ es el escalar tal que $P(\omega > \chi_\alpha) = \alpha$, con $\omega \sim \chi^2_{(m-n)}$.

Demostración. Primero,

$$\begin{aligned} \|y - \overline{y}1_m\|^2 &= \sum_{i=1}^{m}(y_i - \overline{y})^2 = \frac{1}{m^2}\sum_{i=1}^{m}\left(\sum_{k=1}^{m} y_i - y_k\right)^2 \\ &= \frac{4}{m^2}\sum_{i=1}^{m}\left(\sum_{k<i} y_i - y_k\right)^2 \leq \frac{4\lambda^2}{m^2}\sum_{i=1}^{m}(m-i)^2 \\ &= \frac{4\lambda^2}{m^2}\sum_{i=1}^{m-1} i^2; \end{aligned} \tag{1.40}$$

pero,

$$\sum_{i=1}^{m-1}(i)^2 = \frac{2(m-1)^3 + 3(m-1)^2 + (m-1)}{6} = \frac{m(m-1)(2m-1)}{6}.$$

Sustituyendo en (1.40) y tomando inversos se logra la relación

$$\frac{1}{\|y - \overline{y}1_m\|^2} \geq \frac{3m}{2\lambda^2(m-1)(2m-1)}. \tag{1.41}$$

Por otro lado, si definimos $s^2_y = \frac{\|y-\overline{y}1_m\|^2}{(m-1)}$:

$$\begin{aligned} \alpha &\geq P(R^2_{ajus} < \rho) = P(1 - \frac{s^2}{s^2_y} < \rho) = P(s^2 > (1-\rho)s^2_y) \\ &= P\left((m-n)\frac{s^2}{\sigma^2} > (m-n)\frac{(1-\rho)s^2_y}{\sigma^2}\right); \end{aligned} \tag{1.42}$$

donde, por el teorema 1.4.5, $(m-n)\frac{s^2}{\sigma^2} \sim \chi^2_{(m-n)}$. Por lo tanto

$$(m-n)\frac{(1-\rho)s^2_y}{\sigma^2} > \chi_\alpha,$$

que empleando la primera parte de esta demostración implica la conclusión de la prueba:

$$\begin{aligned}(m-n) &> \frac{\chi_\alpha \sigma^2}{(1-\rho)s_y^2} = \frac{(m-1)\chi_\alpha \sigma^2}{(1-\rho)\left|y-\overline{y}1_m\right|^2} \\ &\geq \frac{3m\chi_\alpha \sigma^2}{2\lambda^2(2m-1)(1-\rho)}\end{aligned} \tag{1.43}$$

□

El lema 1.4.13 nos dice que la diferencia entre el tamaño de la muestra y la cantidad de variables de descripción, empleadas para la prueba de hipótesis de linealidad, guardan una relación inversa con el cuadrado del diámetro del rango de la variable dependiente. Lo cierto es, que el radio de dicho conjunto no puede ser menor que el radio del conjunto de muestras de la variable y; luego, es evidente que una selección de muestras srá preferible entre mayor sea el diámetro del conjunto $\{y_1, ..., y_n\}$. Razonando como lo hemos hecho el reciente resultado nos otorga un criterio estadístico útil:

Relación entre el tamaño de muestra y la cantidad de variables de predicción para la prueba de hipótesis de linealidad.
Bajo el marco de referencia del lema 1.4.13, si $\lambda_0 = max_{i,j<m}|y_i - y_j|$ y $\sigma_0 \geq \sigma$; la relación

$$m - n > \frac{3\chi_\alpha \sigma_0^2}{4\lambda_0^2(1-\rho)}$$

garantiza una la adecuada realización de la prueba de hipótesis:

$$H_1 : \rho_0 < \rho \text{ contra } H_0\text{: } \rho_0 \geq \rho.$$

Tenemos ahora una nueva restricción sobre la diferencia entre el tamaño de la muestra y la cantidad de variables de predicción empleadas para resolver el problema de regresión, siempre que en lo posible se quiera respaldar un poco la hipótesis de linealidad empleando el coeficiente de determinación ajustado.

Es importante notar que que en el consecuente del lema 1.4.13 es posible es posible, empleando tablas de distribución normal, determinar el mínimo valor α para el cuál no puede rechazarse una la hipótesis nula para una muestra dada y un valor conocido de n, $1 < n < m$, con esto y en el caso de los análisis QSAR en que la selección previa de descriptores hecha por especialistas resulte en un conjunto de descriptores aún mayor que el tamaño de la muestra; se tiene entonces, un proceso de selección de desccripotres en una segunda etapa, que nos interesa ahora.

De acuerdo con el criterio presentado para la selección del tamaño de muestra puede partirse de la determinación de un valor máximo para n, digamos n_{max}, luego:

- Para un valor dado para ρ y α_0 una cota conocida para α, elegir los conjuntos de descriptores para los que no pueda rechazarse la hipótesis débil de linealidad que trabajamos en esta sección;

- despreciar todos aquellos para los que si existe un subconjunto de ellos para el que tampoco se rechaza la hipótesis nula en la prueba de linealidad, pero que observan un valor mayor de R_{ajus} respecto en comparación con el conjunto que los contiene;
- por último, elegir aquellos en que los valores propios de la matriz $X^{(k)}$ muestren una más rápida convergencia a 0, o los que $SEC(x_1, ...x_k)$ observe también una convergencia a 0 más rápida conforme a la ecuación (1.39).

Proceder de esta forma en una segunda etapa es una forma de buscar que el problema de regresión sea bien planteado, o en su defecto un primer intento por garantizar que se cumplen las condiciones que son necesarias para poder llevar a cabo un proceso de ajuste de parámetros desde la teoría de regularización, particularmente mediante lo que se conoce como solución de Tikhonov y que posteriormente comentaremos con más detalle.

El resultado, en cualquier caso, es que se logren modelos más confiables en general, siempre que no se rechace H_0. Todo con criterios y herramientas conocidos por la estadística desde hace 50 años por lo menos, criterios y proceso de selección que aprovechan toda la información disponible en pro de mejores resultados con poca información experimental disponible.

Capítulo 2

Propuesta de modificación metodológica

Sometiendo a escrutinio a la metodología QSAR resalta el carácter reduccionista de la misma, las definiciones presentadas en esta sección resaltan el hecho de que una de las posibles causas del fracaso de un análisis QSAR, con relación a la capacidad predictiva de sus modelos, sería inherente a la incompatibilidad entre la función de actividad (o respuesta biológica) dada y premisas metodológicas radicalmente reduccionistas.

Un análisis QSAR en si mismo está caracterizado no sólo por el sistema que observa, también depende de lo que se desea observar en él y el objetivo por el que se hace, es cierto que IC_α es el valor que en cualquier análisis de este tipo se reporta; sin embargo, depende de la función de actividad. En el caso en que los bioensayos son pruebas de actividad in vitro y en forma asilada la interacción entre el receptor, su sustrato natural y el ligando, ocurre que la única forma estandarizada de comparar la eficacia entre ligandos es, por las relaciones expuestas, ΔG.

En el diseño de fármacos, en general no suele ser ΔG el único criterio que se emplea para decir que un fármaco es más eficaz que otro, en muchas ocasiones el criterio de comparación no está siquiera relacionado con el efecto terapéutico de los ligandos, incluso si para múltiples funciones de actividad, incluida ΔG, para dos fármacos en la misma familia, uno de ellos observa valores de actividad apenas significativamente mejores que el fármaco restante pero el costo económico de la síntesis es sustancialmente mayor para el primero, entonces el criterio de comparación es muy simple, el mejor compuesto es aquel que tenga un menor costo de producción.

Por otro lado, también ocurre que el criterio de comparación sí se relaciona con el efecto terapéutico pero no depende exclusivamente de una análisis QSAR sino de más de un conjunto de ellos, por ejemplo, en el caso de los antibióticos, suelen usarse para ellos dos tipos de bioensayos in vitro de los que se obtienen mediciones independientes de IC_{50} que brindan información sobre el efecto terapéutico que se espera de la familia de ligandos. El primer tipo de actividad es ΔG, mientras que otro tipo de función de actividad es el área, A, de la superficie de la solución en la caja de Petri ocupada por

una población de bacterias en presencia de una concentración libre de ligando al cabo de un lapso de tiempo determinado.

Existe una relación entre estos dos tipos de actividad, como lo es que la eficacia de un ligando como reflejo del efecto terapéutico no puede inferirse sólo de las observaciones que se hagan en el sistema $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R}, \mathfrak{A})$, en términos biológicos como es bien sabido la clave aquí es la membrana celular, que para la farmacología sólo indica que la relación entre las funciones de actividad ΔG y A está mediada por un proceso de absorción que no depende del receptor sino de la célula que lo sintetiza como parte de sus procesos vitales, el metabolismo por supuesto, y en general cada una de las partes de dicha célula que reacciona en presencia del ligando y que afecta significativamente la interacción entre el ligando y su blanco.

No debemos perder de vista que en el caso de antibióticos y desde la perspectiva de un consumidor, un buen fármaco es eficiente y potencia como medicamento (actúa rápido sobre su blanco biológico y requiere de dosis pequeñas para lograr el objetivo terapéutico), porque se sabe qué si el medicamento no erradica suficientemente rápido el agente patógeno, entonces éste puede desarrollar resistencia al fármaco y causar efectos no deseados.

Lo cierto es que esa potencia depende de la rapidez con que una concentración produzca la muerte de una célula y vuelva a estar disponible en el exterior para atacar a una célula distinta, es decir, la potencia está relacionada con ΔG en general, puesto que en realidad la potencia para la función A está relacionada con las potencias de múltiples análisis QSAR en sistemas $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$.

Debemos prestar atención en que la definición de IC_α atiende a que, en general, en un bioensayo in vitro o in vivo, nunca es posible observar directamente la interacción entre un ligando y su receptor ya que, incluso en sistemas $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$, un ligando es una micromolécula mientras que los receptores biológicos son en general proteínas, macromoléculas en las que no es posible distinguir experimentalmente entre un receptor no ocupado por el ligando y uno que si lo está.

Si aceptamos que la forma de crecimiento radial del área ocupada por la bacteria en el bioensayo determina la forma de interacción entre la concentración del ligando y la frontera; entonces, los protocolos y estándares de la farmacología obligan a que cualquier definición de actividad, respuesta, eficacia o potencia, en los bioensayos con función de actividad A, tiene por característica depender de las cantidades $N_{\mathfrak{L}}$, $N_{\mathfrak{L}_0}$, $N_{\mathfrak{R}_0}$, $N_{\mathfrak{A}_0}$ y cada una de las parejas (k_c, k_d), en función de la presión, no necesariamente constante, que caracterizan a las familias de ecuaciones diferenciales que, por la ley de acción de masas, describen la dinámica de las concentraciones de los compuestos químicos en los sistemas del tipo $(\mathfrak{M}, \mathfrak{R}_{\mathfrak{L}}, \mathfrak{L})$ y $(\mathfrak{M}, \mathfrak{R}, \mathfrak{A})$, donde la membrana celular es el primer blanco común de la familia de ligandos y $\mathfrak{R}$ es el blanco biológico intracelular último.

Por lo anterior parece sensato invertir tiempo en una forma de relacionar ambas funciones de actividad.

Un primer intento de modelación sería pensar en la célula como el sistema en el que se observa interactuar al ligando y receptor, que en el interior de la célula son constantes presión y temperatura. pensar que el fármaco puede inocularse directamente en el interior de la célula en un momento dado y que no existe actividad de la membrana celular, es decir, ningún compuesto entra o sale de la célula.

En un modelo tan simple, y por tratarse de antibióticos, el principal efecto esperado de la acción del fármaco es la muerte celular, pero no sólo se busca un fármaco que logre dicho efecto con la menor concentración posible del medicamento, se busca un fármaco que, además de matar a un individuo del agente patógeno con concentraciones bajas, lo haga rápido.

Lo menos que queremos es que el fármaco demore demasiado tiempo como para permitir que la bacteria tenga la oportunidad de comenzar a desarrollar resistencia al fármaco y sobre todo, queremos que no tenga la oportunidad de interactuar con otro individuo en un posesos de transmisión genética horizontal.

El criterio de eficiencia en este caso está relacionado entonces con la velocidad con la que el fármaco comienza el proceso de inhibición enzimática y con el periodo de tiempo de vida del complejo *ligando-receptor*, el criterio debe reflejar en un primer momento una constante de la mayor magnitud posible y un constante de disociación pequeña, por ejemplo.

Aquí, la función de respuesta biológica debe ser una característica mensurable de la célula que refleje el cambio en la propiedad cualitativa enmarcada en la frase "El individuo del agente patógeno permanece con vida", y que pueda expresarse como una función de la actividad del ligando, en función de la concentración del complejo $\mathfrak{LR}$.

Lo que se ha observado es que el comportamiento cualitativo e incluso cuantitativo de la actividad de lligando y la respuesta del blanco biológico es sensible a la magnitud de gradientes de calor, temperatura y distintos tipos de variables termodinámicas relacionadas con la energía interna del complejo $\mathfrak{LR}$.

Para este tipo de análisis QSAR se supone que el proceso de muerte celular no involucra cambios de presión o temperatura, la célula no muere por causas mecánicas, también hemos supuesto que la temperatura puede considerarse constante durante el periodo de observación, lo que podemos formular como hipótesis es que que, la vida celular es sensible respecto de cambios en la magnitud del gradiente de la entalpía de la célula como sistema.

Aceptando la hipótesis QSAR de linealidad, un segundo tipo de actividad, independiente al cambio en la energía libre de Gibbs del sistema $(\mathfrak{M}, \mathfrak{LR}, \mathfrak{A})$, es el cambio en la entalpía de dicho sistema. Lo que estamos diciendo es que ΔH se presenta como una opción de actividad para este análisis, en el que escalando el comportamiento del potencial termodinámico dH del sistema que se observa *in vitro*, puede explicarse el cambio de la respuesta biológica o el efecto terapéutico respecto del bioensayo *in vivo*, en el que la célula es el sistema de interés.

Supondremos por un momento que se conocen ya un modelo general basado en lo anterior y que conocemos cada uno de los parámetros que se requieren para la modulación, excepto las constantes k_c y k_d para el sistema $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$; entonces la primera complicación es que estaríamos ante la necesidad de resolver un problema de condiciones iniciales para recuperar una aproximación de tales parámetros. Significa todo esto que si queremos lograr mejores criterios de comparación entre moléculas respecto de su efecto terapéutico y realizando la menor cantidad de análisis QSAR posibles, entonce requerimos una función de actividad con rango de dimensión 2 por lo menos, dicho de otro modo, es necesario robustecer las funciones de actividad tradicionales.

Un radical reduccionismo, complementado con una incompleta descripción termod-

inámica y cinética, que serán desarrolladas en las secciones 2.2 y 2.1, repercute en la confianza que que puede depositarse en los resultados de un análisis QSAR aún cuando se tuviesen los cuidados necesarios para determinar la pertinencia de los mismos. Esto en la literatura puede apreciarse en el hecho de que la correlación estadística mínima que se exige a un modelo QSAR respecto de los datos reales, para considerar como viable dicho modelo, por lo regular no excede al 60 %.

Nuestro trabajo se centrará en presentar ligeras y sencillas modificaciones a los análisis QSAR aceptando sus hipótesis fundamentales de linealidad de la actividad, en una versión debilitada, respecto de las características moleculares mensurables de una familia de ligandos. Funciones de actividad más complejas están fuera de los alcances de este trabajo y se mencionan sólo porque hemos considerado que es importante hacerlo y tenerlo en cuenta con el fin de observar la debida diligencia y responsabilidad en el uso de la herramienta matemática.

2.1. Actividad vectorial para un análisis QSAR

Los análisis QSAR recuperan la esencia de la ecuación (1.1) desglosándola además en un conjunto de premisas aceptadas *a priori* que determinan la forma general del modelo que explica las relaciones cuantitativas entre actividad y descriptores moleculares:

> Para cada familia de compuestos Ξ existe una cantidad finita de descriptores $x_1, ...x_{n_0}$ y un vector a en el espacio euclidiano $\mathbb{R}^{n_0}$ tal que la respuesta de $\mathfrak{L}$ puede aproximarse, con un error aceptable, por el operador lineal definido por el producto interior de a con el vector formado por las correspondientes mediciones $x_1(\mathfrak{L}), ..., x_{n_0}(\mathfrak{L})$. La ecuación 1.1 se re formula como sigue

$$f(x_1, ...x_{n_0}) = a_1x_1+, ...ax_{n_0} + b. \tag{2.1}$$

Para los modelos QSAR se acepta además, que compuestos con estructuras "parecidas" producen respuestas con comportamiento cualitativo parecido, esto es entre otras cosas lo que define a una familia de moléculas de acuerdo con la definición que T. Scior comenta en su crítica (*ver* [24]).

Por ahora ignoraremos la hipótesis de linealidad de los análisis QSAR. Por el tipo de sistemas aceptaremos, solamente y por ahora, la hipótesis de que para una familia molecular QSAR existe una función que pone en correspondencia a cada molécula en ella con el valor de su actividad. Si aceptamos este hecho implícitamente estamos aceptando que cualquier cambio cualitativo o cuantitativo entre el comportamiento de cualquier observable entre los sistemas que define cada receptor queda en función del conjunto de moléculas.

Desde el momento en que se decide ver a una reacción como un sistema termodinámico se sigue que un análisis exitoso debe resultar en modelos que expliquen el comportamiento de la mayor cantidad de características relevantes para la farmacología y, en el mismo sentido, comparables entre elementos de familias de sistemas termodinámicos[1],

[1]Forma de entender familias de reacciones químicas de pendientes de un complejo variable en las condiciones iniciales de la reacción.

con definición por comprensión dependiente de una familia de ligandos Ξ y con alguna de las dos formas siguientes:

$$\{(\mathfrak{M}, \mathfrak{L}, \mathfrak{R}) : \mathfrak{L} \in \Xi\}$$

o

$$\{(\mathfrak{M}, \mathfrak{L}, \mathfrak{R}, \mathfrak{A}) : \mathfrak{L} \in \Xi\}$$

Por simplicidad aceptaremos que, durante el periodo en que se realizan las mediciones de actividad, es nula o despreciable la velocidad a la que el producto $\mathfrak{P}$ se transforma de nuevo en un ligando endógeno. Esto es algo que se observa en la práctica y por ello tal simplificación no significa una cambio en el tipo de análisis QSAR estudiado. Hecho esto, del sistema de ecuaciones diferenciales con que se ha caracterizado la cinética de las reacciones químicas se infiere que, en el estado de equilibrio, las concentraciones del receptor, el ligando y el complejo *ligando-receptor* son idénticas al estado de equilibrio del correspondiente sistema $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$. Será entonces suficiente con estudiar sólo el caso de los sistemas más simples.

En un sistema a temperatura y presión constantes, como en el que son probados las moléculas de un análisis QSAR, se sabe por los resultados en termodinámica que la diferencia en la energía libre de Gibbs de un estado a otro, puede ser expresada como una combinación lineal de las respectivas diferencias de entalpía, ΔH, y entropía, ΔS, como sigue:

$$\Delta G = \Delta H - T\Delta S; \tag{2.2}$$

donde T es la temperatura del sistema.

La entalpía puede comprenderse como el contenido calorífico en el sistema, valores negativos indican un mayor contenido calorífico en el estado inicial que en el final, la magnitud de ΔH en este caso indica la cantidad de calor que es transferida al medio por la conformación del complejo *ligando-receptor* (reacción exotérmica). Valores positivos por otra parte indican que la transición de un estado al otro implicó que el nuevo compuesto $\mathfrak{LR}$ requiriese absorber calor del medio $\mathfrak{M}$ para su conformación, (reacción endotérmica).

La entropía, por otro lado, establece una medición de la cantidad de energía isotérmicamente no utilizable por unidad de temperatura, en un proceso irreversible, por ejemplo, la diferencia de entropia será siempre positiva, no necesariamente así para procesos reversibles.

La espontaneidad de una reacción, es decir, el signo negativo de ΔG no es equivalente a que la reacción sea exotérmica pese a que pueda parecer así cuando la magnitud del valor de $T\Delta G$ sea significativamente menor que la magnitud, con signo negativo, de ΔH. Con esto, cabe mencionar que la energía libre de Gibbs por si misma parece omitir información relevante sobre el fenómeno de la conformación del complejo *ligandos y receptores*, información cualitativa que como se expone en la sección A puede resultar útil para definir criterios más detallados de lo que se entenderá por una familia molecular QSAR o simplemente para la comparación de eficacia, potencia y eficiencia entre dos fármacos.

2.1.1. Por qué no es suficiente ΔG

Primero, para un sistema $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$ el espacio de estados es un subconjunto de $\mathbb{R}^5$, con coordenadas de la forma

$$(S, V, N_{\mathfrak{L}}, N_{\mathfrak{R}}, N_{\mathfrak{LR}});$$

donde S y V corresponden respectivamente a la entropía y al volumen del sistema. Pero, es claro que las relaciones presentadas en (1.9), para las concentraciones de los distintos compuestos químicos del sistema, se cumplen también cuando se sustituye la concentración del respectivo compuesto por la cantidad de moles del mismo:

$$\begin{aligned} N_{\mathfrak{LR}} &= N_{\mathfrak{L},0} - N_{\mathfrak{L}}; \\ N_{\mathfrak{AR}} &= N_{\mathfrak{A},0} - N_{\mathfrak{A}} - N_{\mathfrak{P}}; \\ N_{\mathfrak{R}} &= N_{\mathfrak{R},0} - N_{\mathfrak{L},0} - N_{\mathfrak{A},0} + N_{\mathfrak{L}} + N_{\mathfrak{A}} + N_{\mathfrak{P}}; \end{aligned} \tag{2.3}$$

denotando por $N_{\mathfrak{L},0}$ y $N_{\mathfrak{A},0}$ a las cantidades iniciales[2] de ligando y receptor respectivamente.

Las coordenadas de los estados del sistema son entonces de la forma:

$$(S, V, N_{\mathfrak{L},0} - N_{\mathfrak{LR}}, N_{\mathfrak{R},0} - N_{\mathfrak{LR}}, N_{\mathfrak{LR}}).$$

Por último, el volumen V también puede aproximarse mediante una transformación afín dependiente sólo de $N_{\mathfrak{LR}}$. Si $\nu_{\mathfrak{L}}$, $\nu_{\mathfrak{R}}$ y $\nu_{\mathfrak{LR}}$ denotan al respectivo volumen molecular de ligando, receptor y el complejo *ligando-receptor*, y ϵ_V como función de t modela un error de aproximación, entonces

$$\begin{aligned} V &= \nu_{\mathfrak{L}} N_{\mathfrak{L}} + \nu_{\mathfrak{R}} N_{\mathfrak{R}} + \nu_{\mathfrak{LR}} N_{\mathfrak{LR}} + \epsilon_V \\ &= \nu N_{\mathfrak{LR}} + \nu_{\mathfrak{L}} N_{\mathfrak{L},0} + \nu_{\mathfrak{R}} N_{\mathfrak{R},0} + \epsilon_V; \end{aligned} \tag{2.4}$$

con $\nu = \nu_{\mathfrak{LR}} - \nu_{\mathfrak{L}} - \nu_{\mathfrak{R}}$.

Así, si ϵ_V es despreciable o depende de S y $N_{\mathfrak{LR}}$, el espacio de estados de un sistema $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$ se encuentra contenido en un subespacio de dimensión 2, pues sus coordenadas, para fines prácticos, son de la forma

$$(S, \nu N_{\mathfrak{LR}} + \nu_{\mathfrak{L}} N_{\mathfrak{L},0} + \nu_{\mathfrak{R}} N_{\mathfrak{R},0} + \epsilon_V, N_{\mathfrak{L},0} - N_{\mathfrak{LR}}, N_{\mathfrak{R},0} - N_{\mathfrak{LR}}, N_{\mathfrak{LR}}).$$

Luego, una función de actividad en estos casos tendrá por rango a una pareja de características termodinámicas del sistema que sean linealmente independientes, lo que bastará para buscar modelos matemáticos que recuperen información relevante del efecto terapéutico que causará un ligando en el organismo final, la intención claro está es lograr criterios de selección de buenos fármacos haciendo la menor cantidad de análisis QSAR posibles. si esto se logra entonces tendrá sentido hablar de optimizar los modelos QSAR.

[2] Se considera al mol como unidad de medida.

Sobre la información no apreciable en los análisis tradicionales trabajaremos con dos tipos distintos de ella: la propiedad en una reacción de ser exotérmica y la relacionada con la cinética de la reacción, que se desprende de la ley de acción de masas. El motivo de ello existen ya protocolos y métodos confiables para la obtención de valores experimentales.

Sobre la cinética de la reacción

En principio se ha dicho que para estudiar la actividad biológica de un ligando con su blanco biológico, experimentalmente se estudia el sistema $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$ a temperatura y presión constantes; además de lo cuál se procuran las condiciones que permitan observar una reacción reversible de primer orden (*ver* [33]), con representación esquemática

$$\mathfrak{L} + \mathfrak{R} \underset{k_d}{\overset{k_c}{\rightleftharpoons}} \mathfrak{LR};$$

en donde k_c y k_d son constantes desconocidas relacionadas con la velocidad a la que ocurre la reacción en la respectiva dirección, es decir, son las constantes de proporcionalidad que caracterizan la formulación de la ley de acción de masas en su formulación como una ecuación diferencial no lineal [7], a saber:

$$\frac{d[\mathfrak{LR}]}{dt} = k_c[\mathfrak{L}][\mathfrak{R}] - k_d[\mathfrak{RL}],\ k_c > 0,\ k_d > 0. \tag{2.5}$$

En realidad, la ecuación diferencial es un modelo continuo que extrapola un fenómeno en el que puede ocurrir que el espacio de estados no sea continuo bajo las condiciones dadas (rango no conexo de las funciones de concentración). La apreciación de ello es muy simple, cualquier concentración es de la forma $\frac{n}{N}$, donde n la medida de una característica observable particular para una componente de interés en el sistema y N es la medida de una segunda característica observable de interés para la totalidad del sistema, incluso par la totalidad del sistema, es evidente que la continuidad de la concentración $\frac{n}{N}$, en función de cualesquiera otras variables, dependerá del comportamiento que observen el numerador como el denominador en el cociente que la define.

La medición experimental de ΔG se lleva acabo cuando la concentración de $[\mathfrak{LR}]$ deja de ser monótona creciente y simultáneamente se observa un porcentaje de inhibición de actividad deseado, es decir, $c_1 = C_\alpha$ y en el momento en que el sistema se encuentra en equilibrio u oscilando alrededor del mismo. La intención es recuperar las concentraciones que caracterizan a la solución de equilibrio 2.5), la forma de obtención de $\tilde{\Delta G}$ consiste en tomar la media aritmética de una serie de mediciones de $[\mathfrak{L}]$ cuando el sistema se encuentra en las condiciones descritas. La dificultad y elevados costos de cada medición de $[\mathfrak{L}]$ conlleva una pequeña cantidad de ellas, tres por lo regular.

De lo anterior se desprende que en la práctica no debe esperarse que en algún momento el sistema realmente se encuentre en equilibrio, lo que se espera y observa en general es que a partir de un tiempo determinado de comenzar la reacción el estado del sistema oscile al rededor de la solución de equilibrio de para la ecuación (2.5).

En la figura 2.1 se presenta el resultado de una simulación computacional sobre el comportamiento de las concentraciones de los distintos compuestos en un bioensayo como

los que se reportan en [6]; donde se observa la interacción de concentraciones de ligando y receptor hasta que la reacción alcanza el equilibrio, por último se añade una concentración específica del sustrato natural y se observa la actividad enzimática.

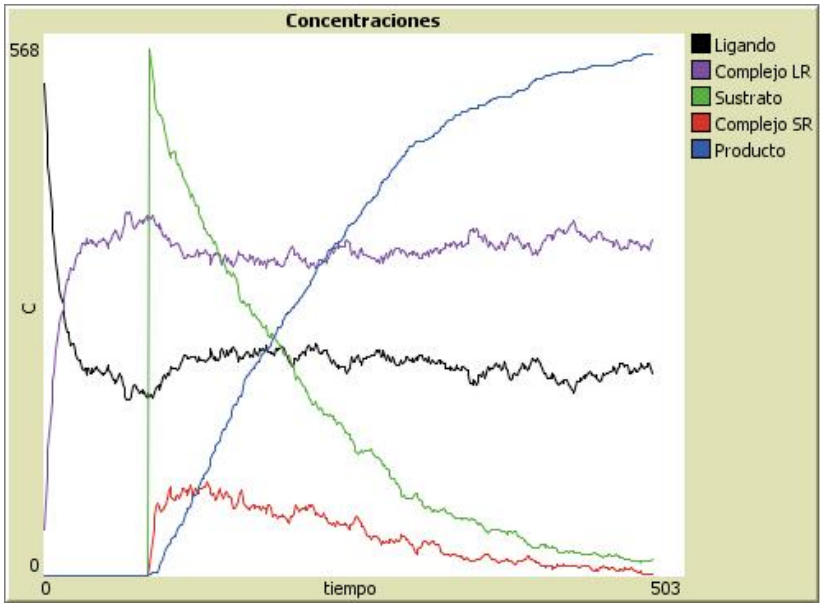

Figura 2.1: Grafico generado con una modificación del modelo de inhibición enzimática en la librería de Net Logo, versión 4.1.3 (*ver* [37] y [26])

Las mediciones deben realizarse con el suficiente cuidado para lograr una distribución del error inherente a ellas que observe una distribución de tendencia central en le valor real de la concentración teórica de equilibrio, sólo de esta forma adquiere validez reportar un promedio aritmético como medición de la actividad.

Por comodidad haremos un breve cambio de notación: $\phi = [\mathfrak{LR}]$ y $\phi_0 = [\mathfrak{LR}]_0$ mientras que $c_1 = [\mathfrak{L}]_0$ y $c_2 = [\mathfrak{R}]_0$ denotarán las concentraciones iniciales de ligando y receptor, en ese orden; pese a que la notación no lo exhibe, ϕ depende de la pareja $(\mathfrak{L}, t)$ para una familia de ligados con receptor común.

Retomando las relaciones de igualdad entre las concentraciones expuestas en (1.9) y la ecuación química (2.5), definimos $k_{eq} = \frac{k_c}{k_d}$ y por la ecuación (2.5). La reacción estará en equilibrio cuando $\frac{d\phi}{dt} = 0$:

$$\begin{aligned} 0 &= k_c(c_1 - \phi)(c_2 - \phi) - k_d\phi \\ &= k_c(\phi^2 - (c_1 + c - 2 + k_{eq}^{-1})\phi + c_1c_2); \end{aligned}$$

las soluciones de la ecuación cuadrática, por fórmula general están dadas por

$$\phi_{eq} = \frac{k_c(c_1 + c_2 + k_{eq}^{-1}) \pm \sqrt{(c_1 + c_2 + k_{eq}^{-1})^2 - 4c_1c_2}}{2}.$$

Ocurre que $(c_1 + c_2 + k_{eq}^{-1})^2 - 4c_1c_2 = (c_1 - c_2)^2 + k_{eq}^{-1}(c_1 + c_2 + k_{eq}^{-1})$, motivo por el que algebraicamente ϕ_{eq} puede tener dos soluciones reales, aunque sólo una de ellas tiene sentido físico. $c_1 + c_2 + k_{eq}^{-1}$ es mayor que cualquiera de las concentraciones iniciales de ligando y receptor, la concentración del complejo $\mathfrak{LR}$ es menor o igual que cualquiera de ellas, así que

$$\phi_{eq} = \frac{k_c(c_1 + c_2 + k_{eq}^{-1}) - \sqrt{(c_1 + c_2 + k_{eq}^{-1})^2 - 4c_1c_2}}{2}. \tag{2.6}$$

Bajo las hipótesis que hemos aceptado, principalmente la unicidad delo sitio activo en el receptor, siempre que sean conocidos los valores de c_1 c_2 y la concentración libre de ligando en el estado de equilibrio, entonces es posible tener una medición de ϕ_{eq}. Luego, dado que en el estado de equilibrio se satisface

En el equilibrio de la reacción se satisface:

$$k_c(c_1 - \phi_{eq})(c_2 - \phi_{eq}) = k_d\phi_{eq};$$

equivalente a:

$$\frac{k_c}{k_d} = \frac{\phi_{eq}}{(c_1 - \phi_{eq})(c_2 - \phi_{eq})};$$

de donde se infiere que también se conoce, bajo las condiciones dadas , una medición de k_{eq}.

En función de c_1, c_2 y ϕ_{eq}, definimos

$$\begin{aligned} a &= c_1 + c_2; \\ b &= 2\phi_{ec} + \sqrt{(c_1 + c_2 + k_{eq}^{-1})^2 - 4c_1c_2}. \end{aligned}$$

En las condiciones en que estamos trabajando debemos notar que c_1 es justamente C_α para algún α, en tanto que ϕ_{eq} es una función de c_1, c_2 e IC_α.

Los valores a y b pueden ser entendidos entonces como funciones del vector $(C_\alpha, IC_\alpha, [\mathfrak{R}]_0)$, de la pareja (C_α, IC_α) cuando, como en nuestro caso, la concentración inicial de receptor se supone constante para cada uno de los bioensayos que se realicen para una familia de moléculas. Se sigue que (2.6) se reescribe como

$$a(C_\alpha, IC_\alpha)k_c + k_d = b(C_\alpha, IC_\alpha). \tag{2.7}$$

Por otro lado, por la relación expuesta en (1.16), en el estado de equilibrio también se verifica la relación:

$$\Delta G = -KT\ln(k_{eq});$$

donde K es la constante de los gases y $k_{eq} = \frac{k_c}{k_d}$.

Tomando en cuanta esta relación entre la energía libre de Gibbs y los parámetros de la ecuación diferencial que estamos atendiendo, notamos que:

$$k_c = e^{-\frac{\Delta G}{KT}} k_d. \tag{2.8}$$

La energía libre de Gibbs por si misma no nos ofrece información sobre la velocidad a de la reacción para dos fármacos que observen actividades muy semejantes, sólo brinda una relación de proporcionalidad entre las constantes de conformación y disociación, k_c y k_d. Para dos ligandos con actividad semejante una análisis QSAR tradicional no permitiría saber cuál de ellos se aproxima más rápido al estado de equilibrio o que tan amplia se espera que sea la oscilación alrededor de tal estado.

Uno de los fallos reportados para los análisis tradicionales puede radicar aquí, al realizar ligeras modificaciones a una estructura base se espera que en general la actividad cambie preservando propiedades cualitativas relevantes, como en este caso las características de la cinética para la reacción química. En la figura 2.2 se muestra el comportamiento de la concentración ϕ para un valor de actividad fijo y distintas velocidades de conformación.

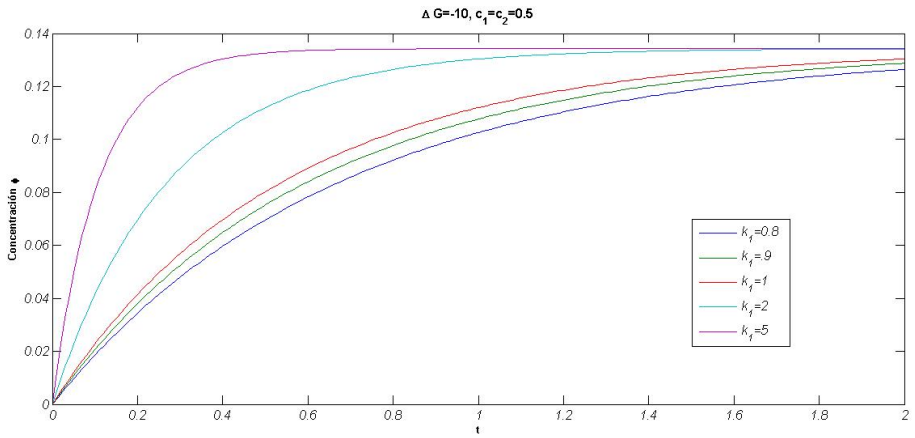

Figura 2.2: Ejemplo de posibles comportamientos de la concentración como función del tiempo para una medición de actividad dada.

Para el caso de la cinética de nuevo encontramos que sólo el valor de ΔG no es suficiente para explicar el comportamiento de la reacción respecto del tiempo.

De las relaciones de igualdad expresadas en las ecuaciones (2.7) y (ec:relación k1-k1-DG), es claro que no es necesario invertir en pruebas experimentales adicionales para entender no sólo la espontaneidad de una reacción sino también los parámetros que determinan sus velocidad de conformación y disociación, es suficiente con pedir reportes detallados de cada bioensayo realizado por un laboratorio para poder determinar aproximaciones de los valores k_c y k_d, sin recurrir a herramienta matemática distinta a la contenida aquí para la realización de una análisis QSAR.

Por ejemplo, si se realizan m, con $m > 1$, bioensayos y se registran las mediciones $\tilde{a}_1, ..., \tilde{a}_m$ y $\tilde{b}_1, ..., \tilde{b}_m$, correspondientes a la relación en (2.7); entonces, para recuperar aproximaciones de los valores de las constantes e conformación y disociación se requiere resolver el problema inverso:

$$\begin{pmatrix} \tilde{a}_1 & 1 \\ \vdots & \vdots \\ \tilde{a}_m & 1 \end{pmatrix} \begin{pmatrix} k_c \\ k_d \end{pmatrix} = \begin{pmatrix} \tilde{b}_1 \\ \vdots \\ \tilde{b}_m \end{pmatrix}.$$

Este es un punto muy importante, sucede que los laboratorios reportan datos insuficientes, en general las concentraciones inicial y final se miden con errores, al igual que la concentración inicial, los laboratorios suelen no reportar ni siquiera la forma en que se estimó el valor de pIC_{50} o si se realizó algún proceso de regularización de regularización para reducir el sesgo y la varianza de los errores de medición.

En [7] y [22] se muestra cómo es que después de todo, las constantes de equilibrio y la concentración inicial de receptor, pueden ser calculadas resolviendo un problema de ajuste de parámetros para una función lineal, justo el problema que atendemos aquí como problema inverso.

La solución inmediata, solicitar información detallada a los laboratorios que realizan los bioensayos *in vitro* y resolver recuperar la medición de C_{50} y pIC_{50} resolviendo el problema de regularización para el ajuste de parámetros, tal y como se hace aquí con los datos de actividad y los descriptores moleculares.

De nuevo dos posibilidades, formular criterios para una familia la caracterización de una familia QSAR en términos de k_c y k_d o establecer una nueva exigencia para la elección de descriptores moleculares, lo que significa que puede elegirse como actividad vectorial:

$$(C_\alpha, IC_\alpha);$$

con α un escalar dado conforme a las necesidades de investigación.

Un conjunto de descriptores que resulte de un análisis QSAR de actividad vectorial, para el que se garanticen modelos que expliquen el comportamiento de cada una de las componentes de actividad (C_α, IC_α), es entonces un conjunto de descriptores que explica el comportamiento de cualquier actividad relevante que pueda definirse para tales sistemas.

Sobre la transferencia de calor

Consideremos dos moléculas distintas en la misma familia de un análisis QSAR con respectivas mediciones de actividades ΔG_1 y ΔG_2, y supongamos también que se observa la relación $\Delta G_1 = \Delta G_2$, bajo el entendido de temperatura y presión constante para el sistema durante la reacción, la implicación inmediata es:

$$\Delta H_2 = T\Delta S_2 + \Delta G_1.$$

La gráfica de la ecuación anterior, con coordenadas $(\Delta S_2, \Delta H_2)$, es una recta con pendiente T que intersecta a los ejes en $(0, \Delta G_1)$ y $(-\Delta G_1, 0)$, esto significa que dado

un compuesto con una medición de actividad de ΔG puede observar posibles comportamientos cualitativos distintos que dependerán del cuadrante en el que se encuentre $(\Delta H, \Delta G)$.

Como se ha dicho ya por lo general se trabaja con compuestos que observan valores negativos para ΔG, en estos casos, para una diferencia de entropía positiva la reacción puede ser endotérmica o exotérmica. Intuitivamente esto nos dice que si un conjunto de descriptores moleculares explica el cambio en la energía libre de Gibbs y sólo eso, realmente no puede esperarse demasiado respecto de posibles predicciones de eficacia o respuesta celular, pese a que estadísticamente proporcione un modelo confiable.

Que dos ligandos puedan observar la misma medición de ΔG y comportamientos termodinámicos significativamente distintos no es algo que deba tomarse a la ligera pues para una análisis QSAR no se consideran los mecanismos que permiten al fármaco traspasar la membrana celular ni cómo es que el compuesto afectará a los procesos metabólicos de una célula, mismos que por supuesto importan para determinar la eficacia del fármaco y que se relacionan directamente con la cinética y todas las variables termodinámicas de la reacción química de conformación y disociación del compuesto *ligando-receptor*.

Ante esto tenemos al parecer dos alternativas, considerar una clasificación adicional de familia QSAR que dependa del sentido en que ocurre la transferencia de calor, o bien exigimos a los descriptores moleculares que además de explicar la actividad usual expliquen el valor de ΔH como una función adicional de actividad, con lo que estamos entonces ante lo que será un análisis de actividad vectorial, donde la actividad ya no será entendida como una función escalar, sino como una función vectorial:

$$(\Delta G, \Delta H).$$

Relación entre $(\Delta G, \Delta H)$ y (k_c, k_d)

Distinguiremos por V a volumen del sistema $(\mathfrak{M}, \mathfrak{L}.\mathfrak{R})$, y por V' al volumen en el que se encuentran contenido el sistema. En general, para cualquier sistema termodinámico se también se satisface (*ver* [11]), la siguiente relación entre los potenciales termodinámicos dU, dH y dV:

$$dU = dH + PdV.$$

Es posible realizar pruebas experimentales para obtener una medición del calor de conformación ΔH, a presión constante, en un bioensayo para un sistema $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$, inclusos existen a la fecha distintas herramientas de cómputo que permiten realizar pruebas *in silico*. El cambio de én la entalpía del sistema de un estado de total disociación al estado de equilibrio será para nosotros una constante

Recordamos que en (2.4) expresamos la relación

$$V = \nu N_{\mathfrak{L}\mathfrak{M}} + \nu_{\mathfrak{L}} N_{\mathfrak{L},0} + \nu_{\mathfrak{R}} N_{\mathfrak{R},0} + \epsilon_V;$$

se sigue:

$$\begin{aligned}\frac{dV}{dt} &= \nu\frac{d\left(N_{\mathfrak{LR}}+\frac{\epsilon_V}{\nu}\right)}{dt} = \nu V'\frac{d\left([\mathfrak{L}]+\frac{\epsilon_V}{\nu V'}\right)}{dt}\\ &= \nu V'\frac{d\left(\phi+\frac{\epsilon_V}{\nu V'}\right)}{dt} = \nu V'\left(\frac{d\phi}{dt}+\frac{d\epsilon_V}{\nu V' dt}\right)\\ &= k_1(\phi^2-(c_1+c_2+k_{eq}^{-1})\phi+c_1c_2)+\frac{d\epsilon_V}{\nu V' dt}.\end{aligned}$$

Por lo tanto,

$$\frac{dU}{dt}-P(k_1(\phi^2-(c_1+c_2+k_{eq}^{-1})\phi+c_1c_2)+d\frac{\epsilon_V}{\nu V'})=\frac{dH}{dt};$$

y con ello, integrando respecto de t en el intervalo $[t_0,t_1]$, donde t_1 es cualquier instante en que el sistema se encuentre en equilibrio:

$$\Delta U-P(\phi_{eq}^3-(c_1+c_2)\phi_{eq}^2+c_1c_2\phi_{eq})+P\phi_{eq}^2k_2=\Delta H-\frac{\epsilon_V}{\nu V'}. \tag{2.9}$$

La información relevante en la ecuación (2.9) es que, cuando en un bioensayo sean conocidos los valores de c_1, c_2 y tres mediciones adecuadas de ϕ_{eq} como de ΔH, se sigue que mediante la regularización de un problema inverso mal planteado de ajuste de parámetros para un funcional lineal, de $\mathbb{R}^3$ en $\mathbb{R}$, será posible recuperar aproximaciones numéricas de las constantes de conformación y disociación, así como del cambio en la energía interna del sistema.

De nuevo, reportes detallados de cada bioensayo realizado por un laboratorio son necesarios para lograr información termodinámica y cinética completa de una reacción.

La ventaja de que tienen los reportes detallados para la investigación son múltiples, por ejemplo, si para cada bioensayo se reportan las tres mediciones que se realizan de la concentración libre de ligando en el estado de equilibrio, ya hemos visto que es suficiente para plantear tener aproximaciones de los principales parámetros cinéticos y las magnitudes termodinámicas. Por lo que si se realizan con sumo cuidado los bioensayos, entonces es posible tener una predicción de las concentraciones de inhibición al 50 %, abriendo la posibilidad de aprobechar el resto de los bioensayos en un rango de concentraciones más proximo a las de inhibición al 50 %, logrando así información estadística más útil.

Si además se hace eso para cada bioensayo, al final, con m bioensayos realizados se dispone de un conjunto de m vectores de la forma $(k_c, k_d, \Delta G, \Delta S, \Delta H)$ para cada molécula probada, con lo que para la muestra puede observarse el comportamiento de cualquier parámetro cinético u observable termodinámico que pueda resultar relevante para el diseño *in silico* de nuevos fármacos.

Lo importante es que como se ha expuesto, es preferible conocer toda la información sobe las mediciones y los posibles errores cometidos que conocer sólo el valor de IC_{50}. Conocer esta información es el punto clave y de partida para lograr determinar el tipo de herramienta matemática que más conviene para cada análisis QSAR.

Se concluye que $(\Delta G, \Delta H)$ es una medición indirecta de (k_c, k_d) y por ello es que dedicaremos lo que resta de estas páginas a la pareja de actividad $(\Delta G, \Delta H)$. Además de presentarse la energía y sus componentes como posibles respuestas biológicas.

2.1.2. Eficiencia

Hemos comentado que las velocidades de conformación y disociación importan en todo esto. Las constantes de conformación y disociación son entonces parámetros relevantes que si bien no son por si mismas equivalentes a la respuesta o el efecto, si son importantes en la determinación de criterios de optimalidad que permitan distinguir un *mejor candidato a fármaco* en una familia molecular QSAR.

Uno de los objetivos de este trabajo era lograr proponer un modelo lineal QSAR resultante un análisis que incluyera las propuestas de modificaciones a la metodología tradicional, objetivo no alcanzado entre otras razones por la no existencia de criterios de comparación suficientemente estandarizados o que justifiquen una relación con la noción de orden respecto del lo que en cada caso particular se entiende por efecto biológico.

Por ahora nos limitaremos a proponer una definición de un tipo de función de actividad que, desde el punto de vista teórico es lo que buscamos como criterio de optimalidad en un análisis QSAR.

Para una función de actividad arbitraria Act, definimos

- β_{Act}, como la función que pone en correspondencia a cada elemento en el conjunto de posibles estados iniciales de Act con el ínfimo tiempo que se requiere par ir del estado inicial correspondiente estado final, caracterizados para la función de actividad Act. β se dirá la función de rapidez inducida por Act.

- ι_{Act}, como la función que pone en correspondencia a cada elemento en el dominio de Act, E, y un instante de tiempo t, con la variable aleatoria que, definida sobre dicho estado e instante, modela el tiempo de vida de una unidad del complejo $\mathfrak{LR}$, ι_{Act} se dirá la función de esperanza de vida del complejo $\mathfrak{LR}$ inducida por la función de actividad Act.

El único criterio que en la practica distingue un análisis QSAR como base de comparación entre distintos ligandos es la potencia, que para los análisis tradicionales significa que para ligandos distintos en estados iniciales iguales, el mejor fármaco es aquel que logre una mayor inhibición de la actividad enzimática, equivalente a valores negativos de la mayor magnitud de la medición de ΔG. En un análisis de actividad múltiple tal criterio no es suficiente.

Sólo en el caso de la familia de ligandos que que utilizamos como ejemplo de análisis con actividad múltiple, una familia de candidatos a medicamentos para el tratamiento de tuberculosis, la forma en que se observa que interactuan los ligandos con su receptor durante el periodo de observación involucra una elevada esperanza de vida del complejo *ligando-receptor* y una ágil conformación del mismo. Es decir, valores pequeños de la constante de disociación y grandes para la constante de conformación; es por este ejemplo que hemos definido a ι_{Act} y β_{Act} como lo hemos hecho.

Lo que pudimos rescatar de la forma en que se busca mejorar los fármacos es que deben procurarse tales criterios de optimalidad para lograr una metodología robusta y completa. La forma teórica que proponemos aquí es mediante funciones de eficiencia.

Toda función, f_{efi}, que ponga en correspondencia a la terna $(\alpha_{Act}, \beta_{Act}, \iota_{Act})$ con un conjunto parcialmente ordenado se dirá una **función de eficiencia** si es tal que la

noción de orden en el rango de f_{efi} es consistente con la noción de orden para el efecto terapéutico de los compuestos.

Determinar una función de eficiencia para un análisis QSAR es en si mismo una labor de investigación y modelación no trivial y que no es muy atendida en general. Sin embargo, es justamente una función de este tipo la que de forma implícita caracteriza a los criterios de optimalidad para distinguir los mejores candidatos a fármacos en la debida etapa del proceso.

Decir que la eficiencia de los fármacos importa es equivalente a decir que además de lo que hagan en el organismo, nos interesa que lo hagan en los mejores tiempos y formas posibles. Una observación al margen del trabajo, no relacionar a un fármaco más que con su actividad en la forma tradicional es uno de los tantos puntos débiles de la metodología QSAR.

2.2. Modelos locales basados en la relación $\Delta G = \Delta H - T\Delta S$

Conforme a la notación y formulación general en que hemos presentado nuestro problema, la hipótesis QSAR tradicional se reescribe como:

Hipótesis de linealidad tradicional[3]

Existe $\lambda \in \Lambda \subset \mathbb{R}^n$ tal que $\Delta G = \lambda^t x$; implicando

$$\mathcal{F} = \left\{ \begin{array}{rcl} f : \Omega \times \Lambda & \longrightarrow & \mathbb{R} \\ (\lambda, x) & \longmapsto & \lambda^t x \end{array} \middle| \, \Lambda \subset \mathbb{R}^n \right\}.$$

En un análisis tradicional se acepta que existe un funcional lineal que asocia a cada elemento de Ω y en el rango de vector de descriptores el valor real de la diferencia de energía libre que le corresponde; sin embargo, no nos dice algo sobre el comportamiento de la entalpía o la entropía para los elementos de la familia molecular, pese a que existen hoy día métodos experimentales que permiten obtener mediciones independientes tanto de la entalpía como de la energía libre de Gibbs para familias moleculares como las que aquí importan, garantizando mediciones de la entropía toda vez que se conoce la temperatura del sistema en (2.2).

Omitir la información disponible sobre mediciones de entalpía (consecuentemente de entropía) en un análisis tradicional, desde el punto de vista de la matemática, es una forma de pasar por alto una advertencia sobre un modelo general que no es congruente con la naturaleza real del problema.

Para hacer asequible lo anterior incluiremos premisas adicionales, que posteriormente serán retomadas para presentar una versión ligeramente modificada del problema QSAR.

Volviendo a la formulación general de $\mathcal{F}$ sea $g(x) = f(x, \lambda)$, λ constante, tal que $\Delta G = g$ toda vez que el argumento de g corresponda a un vector de descriptores moleculares. De la formulación tradicional preservaremos que para cada elemento de $\mathcal{F}$ sus proyecciones sobre Ω y Λ sean continuamente derivables en el respectivo interior de su dominio.

[3] Recordar que ΔG y los elementos de Ω están en función de la familia molecular Ξ.

La hipótesis tradicional, implícitamente afirma que existe un funcional continuamente diferenciable en un conjunto abierto que contiene a cada uno de los vectores de descriptores moleculares de una familia estudiada, respetando esto, extenderemos esta hipótesis a las componentes de la energía libre, suponiendo que el comportamiento de las componentes de la energía libre respecto de los descriptores que la explican es en lo individual de la misma naturaleza. Con otras palabras, que existen descriptores $x_1, ..., x_p, ..., x_{p+q}, ..., x_n$ y funcionales continuamente diferenciables, h y s, definidos con el mismo dominio que g (inducido por los vectores de descriptores moleculares $x = (x_1, ...x_n)^t$) y tales que el comportamiento de la entalpía se explica sólo por cambio en los descriptores con subíndices de 1 a $p-1$ y de q a n, mientras que la entropía queda explicada por los descriptores con subíndices de p a n:

$$\begin{aligned} \frac{dh}{\delta x_i} &\equiv 0, \quad si \quad i = p, p+1, ..., q-1; \\ \frac{ds}{\delta x_i} &\equiv 0, \quad si \quad i < p. \end{aligned} \tag{2.10}$$

En lo tocante a ΔH, como un descriptor en si mismo, aceptaremos que también existe un elemento h en la familia paramétrica de funcionales $\mathcal{F}_H = \{h_\gamma : \Omega \longrightarrow \mathbb{R} | \gamma \in \Gamma \subset \mathbb{R}^p\}$ tal que $\Delta H = h$ cuando el argumento de h sea el vector de descriptores de un elemento de la familia Ξ, donde Γ es un conjunto de vectores de parámetros que definen a la familia. El conjunto Ω contiene al rango de un vector de descriptores de los que depende el cambio en la entalpía para la familia molecular. Lo único que se exige a Ω es que sea la clausura de un conjunto abierto y conexo que contenga al rango de los vectores de descriptores que lo definen, y que no incluyan a ΔG, ΔH ni a ΔS entre tales descriptores. Análogamente se define a la familia de funcionales $\mathcal{F}_S$, a la que pertenece s.

Estamos aceptando que para cada punto en el rango de los respectivos vectores de descriptores para la energía libre y la entalpía existe una vecindad de él y funcionales continuamente diferenciables en cada una de estas vecindades, que explican el comportamiento de las deferencias de energía en función de los valores que tomen los descriptores moleculares correspondientes.

Un análisis QSAR tradicional no contempla la existencia de $\mathcal{F}_\mathcal{H}$, por consiguiente tampoco las características mínimas que deba observar. Por simplicidad sólo vamos a extrapolar a esta familia de parámetros las características de diferenciabilidad continua que hemos establecido para g con el fin de debilitar la hipótesis tradicional de linealidad. Aceptamos que es suave, no sólo continuo, el cambio que producen en la deferencia de entalpía pequeñas diferencias en las propiedades estructurales y fisicoquímicas de un compuesto.

Ahora, apelando a la fórmula de Taylor de primer grado, las nuevas condiciones del problema permiten expresar las relaciones

$$g(x_0 + d) = h(x_0 + d) - Ts(x_0 + d); \tag{2.11}$$

$$h(x_0 + d)) = (\nabla h(x_0))^t d + h(x_0) + o(x_0, d); \tag{2.12}$$

$$s(x_0 + d)) = (\nabla s(x_0))^t d + s(x_0) + o(x_0, d); \tag{2.13}$$

donde x_0 y x pertenecen simultáneamente a Ω o Ω_H en el caso correspondiente, con x_0 fijo, $d = x - x_0$ y $\frac{o(x_0,d)}{\|d\|}$ tiene a 0 cuando $\|d\|$ lo hace.

No debemos perder de vista que la hipótesis tradicional de linealidad es un caso particular de la ecuacion (2.11), cuando $o(x_0, \cdot)$ es idénticamente nula en (2.13) y (2.13), en este caso el término independiente del modelo lineal queda determinado por $g(x_0) - \nabla g(x_0)^t x_0$. Las ecuaciones (2.11), (2.12) y (2.13) son por tanto una generalización y extensión del enfoque tradicional QSAR.

Como se aclarará en breve, las modificaciones que hemos hecho no requieren de herramientas matemáticas o teóricas distintas a las que hasta ahora hemos empleado, lo único que hemos hecho es limitar el estudio de un análisis QSAR a familias con menos moléculas, esto con la intención de mostrar cómo es que los fallos significativos en las predicciones de los modelos QSAR puede relacionarse con la poca pertinencia de sus hipótesis.

Recordemos que el gradiente de una función derivable en el punto x_0 es el vector de derivadas parciales de dicha función, así por comodidad en la notación denotaremos a $\nabla h(x_0)$ por a y a $\nabla s(x_0)$ por b.

Ahora, ΔH tiene un rango acotado pues la transferencia de calor de un estado a otro no puede ser mayor que la energía total del sistema, y como ya se expuso con anterioridad ΔG también es una magnitud acotada, de esto se sigue que $o(x_0, d)$ tiene un rango acotado en (2.12) y (2.13).

Para ser consistentes con la lo desarrollado en la sección 1.1, vamos a aleatorizar a $o(x_0, d)$, es decir, por no tener información sobre su comportamiento en una vecindad de x_0; entendiendo de esta forma a $o(x_0, d)$ podemos decir que es una variable aleatoria con media y varianza finita, desde que es acotada. Continuando con las ventajas de entender a la variable del error de la fórmula de Taylor en términos de teoría de probabilidad, los respectivos errores de medición experimental de las componentes serán considerados implícitamente en el error de medición, lo que significa que la característica de convergencia del cociente $o(x_0, d)/\|d\|$ es remplazada por su versión estocástica $E[o(x_0, d)|d]/\|d\|$, donde suponemos que los errores de medición tienen media 0. Así, $o(x_0, d)$ se convertirá en una perturbación en los datos reales.

Para terminar los ajustes de notación, debido a la independencia en las mediciones de ΔG y ΔH es necesario hacer explícita la diferencia entre las funciones de error, se sustituirá $o(x_0, d)$ por $-\delta\Delta H$ para en (2.12) y por $-\delta\Delta S$ en (2.13), logrando, al sustituir d por $(x - x_0)$, reescribir estas ecuaciones como:

$$\begin{aligned} \tilde{\Delta G} &= g(x) + \delta\Delta G = (h - Ts)(x) + (\delta\Delta H - T\delta\Delta S) \qquad (2.14) \\ &= (a - Tb)^t(x - x_0) + g(x_0); \\ \tilde{\Delta H} &= h(x)) + \delta\Delta H = a^t(x - x_0) + h(x_0). \qquad (2.15) \end{aligned}$$

Como puede notarse, en general a y b son vectores desconocidos, si deseamos hacer predicciones locales sobre lo que ocurre con la actividad del ligando y la el cambio de entalpía del sistema en el paso de un estado a otro (inhibición al 0 % y 50 %) alrededor de x_0; entonces son a y b variables cuyos rangos caracterizan a los conjuntos de parámetros

de dos familias paramétricas de funcionales lineles. Se tiene entonces un problema de auste de parámetros para un sistema de ecuaciónes parcialmente acoplado.

Con esta formulación, cada ecuación por separado plantea un problema de ajustes de parámetros para modelos QSAR tradicionales, dos problemas de regresión lineal múltiple con término independiente no nulo. Lo único inusual en (2.15) es que el análisis sería para determinar relaciones cuantitativas entre un conjunto de descriptores moleculares, y la actividad vista como una medida relativa a cuán endotérmica o exotérmica es la reacción de conformación de enlaces entre ligando y receptor, ΔH.

Una lectura oportuna de la relación (2.2), es que la relación existente entre las ΔG, ΔH y ΔS en un sistema isotérmico es quien dotará simultáneamente de un sentido físico, y por lo tanto real, a la ecuación (2.14).

Resolver nuestro problema QSAR en esta etapa radica en encontrar los parámetros para el sistema de ecuaciones:

$$\begin{aligned} (a - Tb)^t(x - x_0) &= \tilde{\Delta G} - g(x_0) \\ (a)^t(x - x_0) &= \tilde{\Delta H} - h(x_0); \end{aligned} \tag{2.16}$$

que es, de nuevo, un problema de ajuste de parámetros del mismo tipo que un el que el usualmente llevado a cabo en los análisis QSAR, sólo que se ha trasladado el origen del sistema de referencias al vector de descriptores moleculares de x_0. Se traducirá, como se explicará en breve, en un problema de optimización con restricciones que dependerán de los fines y las condiciones bajo las que se realice un análisis QSAR, así como de la relación entre p, q y n en el juego de ecuaciones (2.10).

Denotaremos por $\Xi_0 = \{\varsigma_1, ..., \varsigma_m\}$ a un subconjunto de la familia molecular Ξ, con $m > n$ (desigualdad estricta que es consecuencia de la traslación del origen a x_0); X será la matriz real de orden $m \times n$ y rango completo con el vector transpuesto de descriptores moleculares de ς_i como i-ésima fila, $\tilde{y}$ el vector de mediciones experimentales de ΔG para los elementos de Ξ_0 y $\tilde{z}$ el vector análogo de mediciones experimentales de ΔH.

Con todo esto, estimar los valores de las componentes de a y b en el sistema (2.16) se resuelve como un problema de ajuste de parámetros por MCO, con o sin restricciones:

$$\mathcal{P}_0 \left\{ \begin{array}{ll} min & f(a;b) = \|X(a - Tb) - \tilde{y}\|^2 + \|Xa - \tilde{z}\|^2 \\ & (a;b) \in \Gamma. \end{array} \right.$$

La notación $(a;b)$ hace referencia al vector $\begin{pmatrix} a \\ b \end{pmatrix} = (a_1, ..., a_n, b_1, ..., b_n)^t$.

Veremos que en el caso más sencillo, cuando no existe información alguna que ayude a clasificar a los descriptores que explican el comportamiento de cada componente y los que no; entonces, lo que hemos hecho aquí es incluir una condición más para tener mayor certeza de que al menos localemente se cumpla la condición de linealidad y que tal linealidad sea consistente con ele fenómeno real y no sea sólo una forzada selección de descriptores con alta correlación estadística pero no real.

2.2.1. El problema sin restricciones

Hemos aclarado que esta nueva formulación es una generalización y exención de un análisis QSAR tradicional, que se trabaja localmente; sin embargo, la formulación y la manera en que ajustaremos los parámetros no se ve afectada cuando se verifican las condiciones de linealidad para un conjunto de mayor diámetro que sea una vecindad de x_0, incluso si tal premisa se extrapola al comportamiento de ΔH.

Habiendo realizado la anterior aclaración y considerando el caso en que para un conjunto de descriptores moleculares no es posible saber cuáles de ellos no afectan el comportamiento de la entalpía y cuáles de ellos no influyen en los cambios observados en la entropía, es decir, cuando en 2.10 ninguna derivada parcial puede garantizarse que sea idénticamente nula como función del vector de descriptores moleculares, en tal caso el problema $\mathcal{P}_\prime$ se transforma en el problema de optimización sin restricciones

$$\mathcal{P}_1 \left\{ \begin{array}{ll} min & f(a;b) = \|X(a-Tb) - \tilde{y}\|^2 + \|Xa - \tilde{z}\|^2 \\ & (a;b) \in \Gamma = \mathbb{R}^{2n}; \end{array} \right.$$

siendo un problema puramente convexo, bastado así con verificar que se cumplan las condiciones de optimalidad de primer orden en un punto para implicar que es un mínimo global del problema. Primero

$$\begin{aligned} \frac{df}{\delta a_k} &= 2\sum_{i=1}^{m}\sum_{j=1}^{n}[2x_{ik}x_{ij}a_j + Tx_{ik}x_{ij}b_j] + \sum_{i=1}^{m} x_{ik}\left(\tilde{y}_i + \tilde{z}_i\right); \\ \frac{df}{\delta b_k} &= -2T\sum_{i=1}^{m}\sum_{j=1}^{n}[x_{ik}x_{ij}a_j + Tx_{ik}x_{ij}b_j] + \sum_{i=1}^{m} x_{ik}\tilde{y}_i. \end{aligned} \tag{2.17}$$

Si $\nabla f = 0$, cambiando el orden de las sumatorias en el primer sumando a la derecha de las relaciones anteriores, se sigue que $\tilde{a}$ y $\tilde{b}$ son soluciones del problema $\mathcal{P}_1$ si

$$\left\{ \begin{array}{rcl} X^tX(\tilde{a} - T\tilde{b}) &=& X^t\tilde{y} \\ X^tX(2\tilde{a} - T\tilde{b}) &=& X^t(\tilde{y} + \tilde{z}) \end{array} \right. ;$$

que al restar la primera a la segunda ecuación implica

$$\left\{ \begin{array}{rcl} X^tX(\tilde{a} - T\tilde{b}) &=& X^t\tilde{y} \\ X^tX\tilde{a} &=& X^t\tilde{z} \end{array} \right. . \tag{2.18}$$

Por lo tanto

$$\left\{ \begin{array}{rcl} \tilde{b} &=& (X^tX)^{-1}X^t\left(\frac{\tilde{z}-\tilde{y}}{T}\right) \\ \tilde{a} &=& (X^tX)^{-1}X^t\tilde{z} \end{array} \right. . \tag{2.19}$$

Recordando que $a - Tb$ es justamente el gradiente de g en el punto x_0, lo que se obtiene en (2.18) es un juego parcialmente acoplado de las ecuaciones normales que por separado determinan la identificación de parámetros por MCO de lo que pueden considerarse

dos análisis QSAR tradicionales: el primero con ΔG como actividad mientras que en el segundo la actividad se entiende como la medición de ΔH.

Es claro que esta formulación puede abordarse desde la misma perspectiva que los análisis QSAR usuales, validando los modelos resultantes haciendo uso de toda la herramienta desarrollada para el problema de regresión lineal múltiple con una ventaja adicional respecto de la correlación estadística. Partiendo de (2.19) tenemos que $\tilde{a} - T\tilde{b}$ coincide con el ajuste de parámetros del análisis tradicional

$$\tilde{a} - T\tilde{b} = (X^t X)^{-1} X^t \tilde{y};$$

mientras, localmente, buenas estimaciones de los gradientes de h y s implican buenas estimaciones del gradiente de g, sin que el recíproco sea cierto en general:

$$\left\| \nabla g - (\tilde{a} - \tilde{T}b) \right\| = \|\delta a - \delta T b\| \leq \|\delta a\| + \|\delta T b\| \,; \tag{2.20}$$

relación en que se reemplaza la norma por la varianza para el casos estadístico.

Después de todo vemos en 2.19 que para nuestra modificación, cualquier modelo en el que la actividad depende de n descriptores, para ser considerado como aceptable debe garantizar que esos mismos descriptores expliquen con el mismo o mayor nivel de confianza el comportamiento de la entalpía.

Pese a que la herramienta necesaria para desarrollar un análisis QSAR que considere simultáneamente mediciones de "dos tipos de actividad" (ΔG y ΔH) no difiere de aquella que se elige un enfoque tradicional, lo que se pone de manifiesto aquí, es que aún bajo premisas mucho más débiles como la linealidad a nivel local, la congruencia de un modelo QSAR con el fenómeno real no es verosímil, si un conjunto de descriptores no es capaz de explicar cada uno de los observables de interés en el sistema al menos en una vecindad de una un ligando con afinidad probada.

En adelante a este tipo de análisis QSAR lo referiremos como una análisis con actividad vectorial, por el hecho de que es un análisis que entiende a la actividad como el par de mediciones experimentales $(\Delta G, \Delta H)$, en lugar de sólo la primera componente.

Un análisis con actividad vectorial es, como en el caso de un análisis tradicional, un problema inverso mal planeado que depende por supuesto del condicionamiento de la matriz $X^t X$; por ahora, elegir atacarlo con regresión lineal múltiple o alguna técnica de regularización dependerá de las necesidades y el criterio de quien realice el análisis, esta elección estará sujeta por lo regular a la cantidad de muestras de las que se disponga. Ante una cantidad escasa de muestras, regularizar el problema sería la mejor opción, siempre que se factible; como ejemplo, regularizar el problema utilizando el método de Tikhonov requiere que la condición discreta de Picard se verifique para ambos sistemas en (2.18).

2.2.2. El problema con restricciones

Entre mayor la cantidad de información relevante que pueda ser incorporada en un análisis QSAR, mejor. Por conocimiento empírico, hipótesis del investigador, análisis estadísticos de componentes principales o cualquier otra forma, un análisis QSAR puede

incluir restricciones como las expuestas en (2.10), este tipo de restricciones pueden generarse a partir de un análisis convencional, si se observa un coeficiente de determinación R^2_{ajus} significativo y con esos mismos descriptores se realiza segundo análisis con ΔH o ΔS como valor de actividad. La existencia de valores significativamente próximos a 0 para componentes de $\tilde{a}$ o $\tilde{b}$ son una forma de establecer las restricciones comentadas, de tal suerte que del problema $\mathcal{P}_0$ deriva un nuevo problema de optimización para la identificación de parámetros:

$$\mathcal{P}_2 \left\{ \begin{array}{ll} min & f(a;b) = \|X(a-Tb) - \tilde{y}\|^2 + \|Xa - \tilde{z}\|^2 \\ & (a;b) \in \Gamma = \left\{ (a,b) \left| \begin{array}{l} a_p = \cdots = a_{q-1} = 0, \\ b_1 = \cdots = b_{p-1} = 0, \\ 0 < p < q \leq n \end{array} \right. \right\}. \end{array} \right.$$

Recalcamos que en el problema $\mathcal{P}_2$, tal como se exhibe, no es considerado el error de medición o perturbación alguna en los datos; será incluida posteriormente.

En esta ocasión, el conjunto sobre el que se restringe la optimización es un subespacio vectorial, siendo el problema nuevamente un problema puramente convexo. Se garantiza una solución única si se verifican condiciones de optimalidad de primer orden.

Por ser Γ un subespacio, adicionalmente se verifica que coinciden los conos de direcciones factibles, tangentes y tangentes positivas con el mismo Γ, y el punto $(a;b) \in \Gamma$ será el mínimo que buscamos siempre que $d^t \nabla f(a;b) \geq 0$ para cualquier d elemento de Γ. En realidad este problema no difiere sustancialmente del problema sin restricciones, al notarlo podemos ahorrarnos un largo desarrollo algebraico.

En primer lugar, para dos conjuntos de reales no negativos se satisface que el mínimo del conjunto suma es la suma de los mínimos en los respectivos conjuntos, de esta forma, y por corresponder f a un sistema semi acoplado, podemos resolver el problema $\mathcal{P}_2$ minimizando cada una de las normas al cuadrado en los sumandos de f de forma independiente, comenzando por $\|Xa - z\|$, y sustituyendo el valor de a en el sumando restante.

Tanto Xa como Xb son transformaciones lineales cuando a y b son variables vectoriales, como X es de rango completo, entonces Xa es el vector en el dominio que, como combinación lineal de los vectores columna de X, tiene por respectivos coeficientes a las componentes de a, análogamente para Xb y b. Por tal motivo resolver el problema $\mathcal{P}_2$ consiste en encontrar las proyecciones ortogonales de y y z sobre el rango de dos transformaciones lineales, que se distinguen entre si por las restricciones dadas por Γ; pero por lo que recién hemos comentado, una restricción del tipo $a_j = 0$ es equivalente a suprimir esa componente de a y lo mismo para la j-ésima columna de X, es decir, restringiendo el dominio de la transformación de $\mathbb{R}^n$ a $\mathbb{R}^{n-1}$; de igual forma para una restricción del tipo $b_j = 0$.

Para concluir esta sección, sea $X^{(1)}$ la matriz que resulta de cambiar la columna j-ésima de Xpor el vector nulo si una de las restricciones dadas por Γ es $a_j = 0$, $X^{(2)}$ será la matriz análoga que resulte de atender a las restricciones tocantes a b; a' y b' serán los vectores que resulten de hacer nulas en a y b las componentes sujetas a restricciones de la forma $a_j = 0$ y $b_j = 0$. El problema $\mathcal{P}_2$ muta en dos problemas secuenciados sin restricciones secuenciados:

$$min_{a'} \left\| X^{(1)}a' - z \right\|^2 \Rightarrow min_{a'-Tb'} \left\| X^{(2)}(a' - Tb') - y \right\|^2;$$

que por las secciones previas sabemos que es equivalente a

$$a' = (X^{(1)t}X^{(1)})^{-1}X^{(1)t}z \Rightarrow min_{a'-Tb'} \left\| X^{(2)}((X^{(1)t}X^{(1)})^{-1}X^{(1)t}z - Tb') - y \right\|^2;$$

con lo que se resuelve

$$\begin{aligned} a' &= (X^{(1)t}X^{(1)})^{-1}X^{(1)t}z, \\ -Tb' &= (X^{(2)t}X^{(2)})^{-1}X^{(2)t}(y - X^{(2)}((X^{(1)t}X^{(1)})^{-1}X^{(1)t}z) \\ &= (X^{(2)t}X^{(2)})^{-1}X^{(2)t}y - (X^{(1)t}X^{(1)})^{-1}X^{(1)t}z. \end{aligned} \tag{2.21}$$

Al introducir la perturbación en los datos y y z, la relación de desigualdad expuesta en (2.20) se conservará. La elección de la herramienta que se utilizará para franquear los problemas de los errores de medición, métodos de regularización o el problema de regresión, dependerán nuevamente de la elección de quién realice el análisis y por supuesto de que en cada etapa los sistemas verifiquen las condiciones mínimos necesarias de cada método.

A modo de comentario, en un análisis QSAR con doble actividad las hipótesis son de carácter local y del tipo deterministas en primera instancia, pese a que el error de medición en los datos garantice una distribución normal, el error en la fórmula de Taylor no tiene por que observar una distribución de este tipo, aún bajo el supuesto de que los elementos de la familia molecular que conforman la muestra correspondan a un muestreo simple sin restricciones con distribución uniforme.

La regresión lineal pierde terreno en estos casos frente a la regularización del problema, esta última nos brinda acotaciones del error que dependen sólo de conocer cotas iniciales para el error en la fórmula de Taylor y el error de medición; cuando el error de medición corresponde a una variable aleatoria con distribución normal, con varianza σ, entonces una acotación de $\pm 3\sigma$ puede considerarse una cota del error e medición, mientras que para el error de la fórmula de Taylor debe disponerse de más información proporcionada por una especialista en farmacología o información estadística en su defecto.

Con esto queda claro qué un análisis QSAR con actividad vectorial, como lo hemos propuesto, no difiere significativamente de un análisis tradicional en el manejo de los datos, difieren en dos aspectos importantes:

- Una análisis con actividad vectorial, o múltiple, se sustenta en premisas más débiles garantizadas sólo en " pequeños entornos" de una estructura base, como compuesto determinado por un grafo orientado geométricamente, lo que lo hace congruente con las posibles condiciones reales del fenómeno y con familias determinadas por una estructura base y con sólo un susttituyente que se procure suficientemente pequeño. Es más fácil argumentar en favor de las hipótesis de un análisis con doble actividad que en defensa de la hipótesis tradicional de linealidad.

II Cualquier buen modelo QSAR de actividad vectorial como el aquí expuesto incluye una ecuación que en si misma es un buen modelo para el análisis tradicional, realizado con los mismos datos. La afirmación recíproca no es necesariamente cierta, dado un buen modelo tradicional, al emplear los mismos descriptores y conjunto muestral para explicar el comportamiento en el cambio de entalpía, no se infiere en general que el modelo que se obtenga de este "seguno tipo de actividad" observe una alta correlación estadística entre sus predicciones y los datos reales.

2.2.3. Elección de fármacos con actividades óptimas a partir de modelos QSAR

Hasta ahora hemos atendido a la general del modelo QSAR (lineal, con actividad vectorial, de caracter local e incluyendo el tipo de familias moleculares sobre el que tendrá sentido), determinar la forma en que se validará como un buen modelo QSAR y el problema de la determinación de los parámetros para un caso particular; lo que resta ahora antes de pasar al estudio de caso es una breve explicación sobre por qué la forma general el modelo, en este caso, también influye en la determinación de descriptores que correspondan a moléculas en la familia de interés con actividad óptima o con la mejora actividad posible.

Proponer coeficientes para el modelo lineal QSAR así como la determinación de los compuestos que logren una respuesta óptima se reduce a resolver problemas de optimización, por tratarse de un fenómeno fisicoquímico el que se modela, deben existir restricciones a las que debe sujetarse el proceso de optimización, toda vez que se desee reducir la posibilidad de errores de estimación. Tales restricciones quedaran determinadas por la elección de la familia molecular Ξ.

En la forma tradicional o con actividad múltiple, un modelo de actividad QSAR es un funcional lineal determinado por el gradiente ∇g o por el par de gradientes $(\nabla g, \nabla h)$ (*ecuaciones* 2.11, 2.12 y 2.13). Por otro lado, sabemos que entre menor sea el valor de ΔG para un miembro de la familia será considerado como un mejor fármaco, de donde el objetivo en cualquiera de los dos análisis es minimizar al funcional $\nabla g(x_0)(x - x_0)$, donde x es el vector de descriptores, pero como g se ha definido en un conjunto abierto que contiene al rango del vector de descriptores, entonces el problema a resolver ahora es

$$\mathcal{P}_{\Delta G} \begin{cases} min_x & (a - Tb)^t x \\ s.a. & x \in \overline{x(\Xi)} \end{cases}$$

Es aquí donde el problema de opimización se complicará dependiendo de los descriptores, independientemente de si el análisis es tradicional o no. Se comentó ya, que las familias QSAR pueden ser de distintos tipos dependiendo de la información adicional disponible, en la mayoría de los casos, incluyendo lo que delante nos aguarda, las familias moleculares serán determinadas por un receptor, estructura base, su función de actividad y restricciones sobre los descriptores, en esta etapa el problema de optimización no resultará tan sencillo como hasta ahora lo habían sido los problemas con que nos habíamos encontrado.

En el mejor de los caso los descriptores elegidos son continuos (momentos de inercia o la carga total, entre otros) y el conjunto $x(\Xi)$ resulta ser un poliedro; en este caso será un problema de programación lineal. No obstante, si entre los descriptores elegidos se encuentran algunos que solo tomen valores en conjuntos numerables (cantidad de átomos de cierto tipo o el peso molecular entre otros), entonces habrá que recurrir a teorías como programación entera y afines.

La diferencia entre un análisis tradicional y uno de actividad múltiple en esta etapa, es que el problema $\mathcal{P}_{\Delta G}$ en general no ofrece garantía de solución única, el conjunto de soluciones depende por supuesto de de la elección de la familia molecular.

En un análisis tradicional, una vez resuelto el problema $\mathcal{P}_{\Delta G}$, no existe un segundo criterio que permita comparar las posibles soluciones y elegir la mejor. En un análisis de actividad múltiple, en cambio, el segundo criterio existirá y dependerá del funcional lineal $a^t(x - x_0)$.

Sabemos ya que la espontaneidad de una reacción entre ligando y receptor no significa que dicha reacción se exotérmica, la relación entre una reacción exotérmica y su espntaneidad a temperatura y presión constantes dependerá de la relación entre el comportamiento de la entropía respecto de la entalpía.

En un fármaco es deseable que, dicho burdamente, la reacción mediante la cuál forma un enlace químico con su receptor sea lo más exotérmica posible, dicho de otra forma, es deseable que la transferencia de calor del medio al nuevo compuesto se mínimo o que el calor transferido del nuevo compuesto en su conformación al medio sea máxima.

Con la notación que estamos manejando, un segundo criterio de comparación entre fármacos es ΔH, la molécula ς_1 se dirá un mejor fármaco respecto de la entalpía que el fármaco ς_2, ambos en la misma familia, si la medición de ΔH es significativamente menor para ς_1 que para ς_2.

Resumiendo un poco, cada componente de la actividad múltiple del análisis QSAR determina un criterio de comparación, teniendo con ello 4 posibles formas de comparación de actividad entre fármacos, incluyendo el criterio usual. Un buen modelo QSAR con actividad vectorial, uno confiable, permite realizar estimaciones bajo cualquiera de los criterios de optimalidad mencionados.

Criterio usual de comparación de actividad

Continuando en la tónica de los análisis tradicionales, el criterio usual se refleja en el problema $\mathcal{P}_{\Delta G}$. No es difícil probar que dos moléculas en la misma familia observan la misma medición de ΔG, si la diferencia entre sus vectores de descriptores yace en el subespacio para el que el gradiente de g en x_0 es un vector ortoganoal.

La prueba es como sigue, sean ς_1 y ς_2 moleculas en Ξ, con respectivos vectores de descriptores x_1 y x_2 en $x(\Xi)$; entonces , por la relación en (2.11), se verifica

$$\tilde{\Delta G}_{\varsigma_1} = (a - Tb)^t(x_1 - x_0) + g(x_0), \quad \tilde{\Delta G}_{\varsigma_2} = (a - Tb)^t(x_2 - x_0) + g(x_0);$$

por lo que, sus mediciones respecto de la energía libre de Gibbs son iguales sólo su diferencia es nula, equivalente a

$$\begin{aligned} 0 &= (a - Tb)^t(x_2 - x_0) + g(x_0) - \left[(a - Tb)^t(x_1 - x_0) + g(x_0)\right] \\ &= (a - Tb)^t(x_2 - x_1) = (a - Tb)^t x_2 - (a - Tb)x_1. \end{aligned}$$

Otra forma de decirlo es que las mediciones de la energía libre son iguales para dos moléculas en la misma familia, si x_2 pertenece al hiperplano resultante de aplicar la traslación caracterizada por x_1 al subespacio, $\mathfrak{S}_{x_0}$, con vector normal en la dirección del gradiente de g en x_0.

El conjunto $x(\Xi)$ es un conjunto compacto, cerrado y acotado en $\mathbb{R}^n$, la imagen del funcional lineal también es por lo tanto un compacto en $\mathbb{R}$ y se sigue que el problema $\mathcal{P}_{\Delta G}$ tiene solución; digamos sin pérdida de generalidad que $\tilde{x}$ es una solución arbitraria del problema, entonces el conjunto solución de dicho problema será:

$$[\tilde{x} + \mathfrak{S}_{x_0}] \bigcap x(\Xi).$$

Priorizando el criterio usual de comparación entre actividades moleculares y cuando $[\tilde{x} + \mathfrak{S}_{x_0}] \bigcap x(\Xi)$ se no vacío, un análisis de actividad vectorial permite realizar un segundo proceso de selección mediante el problema

$$\mathcal{P}_{\Delta H} \left\{ \begin{array}{ll} min_x & a^t x \\ s.a. & x \in [\tilde{x} + \mathfrak{S}_{x_0}] \bigcap x(\Xi) \end{array} \right.$$

Aplicando los criterios de comparación de forma secuenciada, primero respecto de la energía libre de Gibbs y luego respecto de la entelpía de ser necesario, no se garantiza aún una solución única pero se reduce el conjunto solución. Por un argumento análogo al que se empleo para determinar el conjunto solución del problema $\mathcal{P}_{\Delta G}$, se llega a que el el conjunto solución del problema $\mathcal{P}_H$ es

$$[\tilde{x}_H + \mathfrak{S}_{H,x_0}] \bigcap [\tilde{x} + \mathfrak{S}_{x_0}] \bigcap x(\Xi);$$

donde $\mathfrak{S}_{H,x_0}$ es el subespacio con vector normal en la dirección del gradiente de h en x_0 y $\tilde{x}_H$ es una solución arbitraria del problema $\mathcal{P}_H$.

Otros criterios de comparación

Este apartado incluye a modo de comentario extenso y desde un lúdico punto de vista matemático que no tiene motivos conocidos por el autor para afirmar que se corresponde con una situación real, es por el momento una simple exploración de las posible situaciones (reales o no) que motivarían distintas formas de entender el concepto de un buen medicamento; situaciones que bien podrían ser resultas sin mayor complicación teórica que la de uso corriente en los análisis QSAR.

Un modelo de actividad múltiple permite considerar situaciones más allá de la espontaneidad de la reacción entre receptor y ligando para entender un buen fármaco, por ejemplo, si en un momento determinado el investigador decide que para los efectos que se espera produzca el fármaco es más importante el comportamiento de la transferencia

de calor entre el medio y el nuevo compuesto; suponiendo un escenario como este, la forma en que se buscarían las moléculas con mejor actividad sería invirtiendo la secuencia en que se realizan los procesos de optimización del apartado anterior inmediato; minimizando primero al funcional $a^t x$ sobre el rango de los descriptores y después, sobre el conjunto solución resultante, minimizar al funcional $(a - Tb)^t x$.

Otro criterio que al autor considera sensato, con su mínima formación en farmacología y sólo por los comentarios de quienes le auxiliaron en lo tocante a esa área, es que un fármaco observe un comportamiento equilibrado, que en lo que podríamos llamar una evaluación conjunta de las actividades en un análisis múltiple, (ΔG y ΔH), se minimicen ambas de forma coordinada.

En adelante denotaremos por $(\mathcal{P}_d, \Omega)$, con d un vector de dimensión n y Ω un subconjunto del espació euclidiano con la misma dimensión, al problema de optimización

$$\mathcal{P}_d \left\{ \begin{array}{ll} min_x & a^t x \\ s.a. & x \in \Omega \end{array} \right. ;$$

Siempre que $(\mathcal{P}_d, \Omega)$ tenga solución no vacía $\tilde{x}_(d, \omega)$ denotará una solución arbitraria y $\mathfrak{S}_(d, \Omega)$ será el subespacio con vector normal en la dirección de d.

Ahora, volvido a los criterios de comparación, un buen fármaco sujeto a una evaluación conjunta de las actividades involucradas se haría resolviendo el problema $(\mathcal{P}_{(a-Tb)}, x(\Xi))$. Mientras que un fármaco "ideóneo" sería todo ς elemento de la familia molecular con vector de descriptores $x(\varsigma)$, tal que

$$x(\varsigma) \in \left[\tilde{x}_{(a,x(\Xi))} + \mathfrak{S}_{(a,x(\Xi))}\right] \bigcap \left[\tilde{x}_{(a-Tb,x(\Xi))} + \mathfrak{S}_{(a-Tb,x(\Xi))}\right] \bigcap x(\Xi).$$

Dicho con menos notación, un fármaco "idóneo", si existe, será aquel que sea solución de los problemas, $(\mathcal{P}_a, x(\Xi))$ y $(\mathcal{P}_{(a-Tb)}, x(\Xi))$. Redundando un poco, un fármaco "idóneo" será aquel que sea considerado con actividad óptima para cada uno de los tres criterios que hasta ahora hemos definido: evaluación conjunta, priorizando la medición de la energía libre de Gibbs o priorizando la medición de entalpía.

Un modelo QSAR con actividad múltiple tiene el potencial de ampliar los criterios de comparación que permiten decidir si un fármaco puede decirse mejor que otro. Contando al criterio usual hemo descrito ya los cuatro criterios de comparación que podrían interesarnos en nuestre inmediato quehacer; aunque cada pareja del conjunto $\{\Delta G, \Delta H, \Delta S\}$, procediendo como antes, brinda tres criterios adicionales, suponiendo claro que siempre se preferirá minimizar la entalpía y la energía libre, maximizando así la entropía.

Cuando no siempre se quiera minimizar la energía libre o la entalpía, entonces cada una de las parejas que pueden ser consideradas como parejas de actividad proporciona 7 criterios de comparación adicionales, dependiendo de que componente quiera maximizarse, cuál minimizarse y de la forma en que se prioricen. En total, de forma teórica e irrestricta un análisis QSAR de actividad múltiple proporciona un total de 21 criterios de comparación; pero sólo nos interesará el tradicional y si acaso un criterio de comparación conjunta.

Lo que no debemos olvidar es que cada proceso de optimización arrastra errores de ajuste que proceden de la perturbación en los datos originales, al resolver los problemas

de determinación de descriptores óptimos no debe olvidarse revisar que tan sensible es la solución respecto de los errores de ajuste.

2.3. Problema QSAR de ajuste de parámetros

Las mediciones de la energía libre de enlace de Gibbs se realizan de forma independiente, de tal suerte que el error de medición sea pequeño y que el promedio aritmético de las mediciones pueda hacerse teóricamente tan próximo como se desee al valor real del observable, siempre que se realicen una cantidad suficiente de mediciones del mismo estado del sistema. Pensar en los errores de medición o perturbación en los datos muestrales como variables aleatorias idéntica e independientemente distribuidas con valor esperado cero y varianza pequeña, es el punto de partida en la búsqueda de relaciones cuantitativas entres mediciones de actividad y el valor de los descriptores moleculares.

Bajo estas condiciones y tomando en cuenta que la parte medular de un análisis QSAR será, la determinación de la transformación afín f (funcional lineal mas una constante), definida en una vecindad, Ω, común al rango de los descriptores $\{x_1, ...x_n\}$ que globalmente explique lo mejor posible el comportamiento la actividad biológica de la familia molecular a la que pertenece la muestra como función de los descriptores moleculares, dado que la única información conocida es el conjunto de mediciones $\{(x_j(\varsigma_i), \tilde{y}_i) : i = 1, ..., m, \quad j = 1, ..., n\}$. $\varsigma_1, ..., \varsigma_m$ es el conjunto muestral de moléculas para el análisis y $\tilde{y}_i$ es la medición experimental de la actividad biológica de ς_i.

Denotaremos por y a la variable de la respuesta biológica de una familia de moléculas, y empleando la notación $\tilde{[]} = [] + \delta[]$ para indicar un dato con un error para la variable por la que sea sustituido el par de corchetes, donde $[]$ denota al dato real y $\delta[]$ al error de medición.

También definimos $\tilde{y} = (\tilde{y}_1, ..., \tilde{y}_m)$, $x_{ij} = x_j(\chi_i)$, $x_j = (x_1, ..., x_n)$ y por X a la matriz de orden $m \times n$ tal que $[X]_i j = x_{ij}$ como antes. Luego entonces una forma de abordar el problema, que se detallará posteriormente, es como el problema de optimización

$$\mathcal{P}\ \mathbf{1.} \qquad \begin{cases} min & \|Xa - y\|^2 \\ s.a. & a \in \mathbb{R}^n; \end{cases}$$

con a el vector que caracteriza al la transformación afín. Las definiciones y detalles formales sobre los objetos y resultados teóricos de los que en este bloque se hace uso, se detallan en 1.3 y pueden ser consultados en cualquiera de los libros del área de estadística en las referencias de este trabajo, en lo relacionado con regularización y solución de Tikhonov se sugiere consultar el libro de Hansen [10] como principal referencia de nuestro quehacer y los libros de Tikhonov *et. ál.* para el lector con mayor interés interés en el desarrollo original de la teoría de regularización: [27], [28] y [29].

Conocer los datos exactos de actividad de cada molécula en la muestra sería lo ideal; pero tal cosa en la práctica nunca ocurre, incluso para los descriptores moleculares requeriremos una clasificación que atienda a la naturaleza del proceso mediante el cuál se obtienen los datos para el análisis.

Los descriptores moleculares pueden ser de diversos tipos, alguno de ellos pueden ser determinados sin errores de medición para cada molécula, como los descriptores consti-

tucionales o los topológicos como cantidades de átomos de cierto elemento o cantidad de diversos tipos de enlaces σ o π; pero, también existe una gran cantidad de descriptores cuyos valores también deben ser calculados de forma experimental o mediante aproximaciones teóricas mediante el uso de herramientas de cómputo.

Expondremos a continuación cómo es que la presencia de descriptores con errores de medición en el conjunto $\{x_1, ...x_n\}$ debe ser considerada al momento de elegir la metodología para el ajuste de parámetros.

Lo usual en los análisis QSAR es resolver el problema de optimización como un problema de regresión lineal múltiple, la hipótesis de independencia entre las variables aleatorias implica que la medida de probabilidad condicional de $\tilde{y}$ dadas $\tilde{x_1}, ..., \tilde{x_n}$ coincide con la medida de probabilidad de $\tilde{y}$, que significa $E[\tilde{y}|x_1, ..., x_n) \sim N(y = \sum_j^n a_j x_j, \sigma^2 = \sigma_y^2 + \sum_{j=1}^n \sigma_j^2)$, con lo que se dispone de las condiciones suficientes para realizar un ajuste de parámetros mediante el método estadístico de regresión lineal múltiple, toda vez que como hipótesis básica se verifique lo siguiente:

> *La variable aleatoria de actividad y tiene esperanza y varianza condicionadas por las variables de predicción $x_1, x_2, ..., x_n$,de forma que:*
>
> $$E(y|x_1, ..., x_n) = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$
>
> *y*
>
> $$Var(y|x_1, ..., x_n) = \sigma.$$
>
> .

Entonces, el ajuste se obtiene de determinar el vector $\tilde{\mathbf{a}} = (\tilde{a}_1, ..., \tilde{a}_n)^t$ que minimiza $\left\|\tilde{\mathbf{b}} - \tilde{X}\mathbf{a}\right\|$ respecto de $\mathbf{a}$; verificando:

$$\tilde{\mathbf{a}} = \frac{adj(\tilde{X}^t\tilde{X})}{det(\tilde{X}^t\tilde{X})}\tilde{X}^t\tilde{\mathbf{b}}.$$

Las observaciones anteriores nos dicen que jamás nos encontraremos ante una situación idónea para la recuperación exacta de cada una de las componentes del vector de parámetros del problema, pero que dependiendo de la selección de descriptores que hagamos tendremos variaciones metodológicas que no pueden ser pasadas por alto, atenderemos primero el caso en que el conjunto de descriptores para él ajuste de parámetros puede ser en general obtenido sin error de medición alguno.

La regresión lineal múltiple es una de las herramientas más empleadas en los para determinar los parámetros de un modelo lineal en los análisis QSAR tradicionales. Esta herramienta tiene mucha potencia estadística por la generalidad de su planteamiento; sin embargo, las debilidades en ella se presentan en problemas para los cuales se dispone de un poco más de información, lo que usualmente se plantearía como un problema de regresión, ante nueva información la mejor forma de abordar el problema puede ser mediante herramientas de programación lineal, programación entera o técnicas de regularización.

El problema de regresión lineal es un ejemplo de lo que se conoce como un problema inverso, que dependiendo de sus particularidades puede clasificarse como mal planteado, es decir, que errores pequeños en las mediciones llegan a implicar errores graves de aproximación, haciendo poco confiables los modelos o requiriendo una gran cantidad de mediciones para alcanzar un ajuste confiable.

La principal ventaja de la regresión lineal es que la información previa que se necesita sobre la naturaleza del observable es menor, siendo posible abordar una mayor cantidad de situaciones en las que, con un tono coloquial, puede decirse que el investigador incursiona casi a ciegas. Pero, el amplio horizonte de aplicación de los modelos de regresión lineal tiene un costo fijado sobre la confianza que puede depositarse en ellos: a mayor confiabilidad del modelo, mayor la cantidad mínima necesaria de las mediciones del observable.

La pertinencia de los modelos de regresión lineal, en su forma corriente, dependerá entonces de las posibilidades reales del investigador de realizar la suficiente cantidad de mediciones, de acuerdo con sus exigencias de precisión en las predicciones de los modelos, aunque en general el problema de ajuste de cuadrados por mínimos cuadrados ordinarios se presenta de forma natural lo que se conoce como un problema de regularización tipo Tihonov.

2.3.1. Regresión lineal frente a regularización

El problema de fondo que buscamos resolver es hallar un funcional lineal en $\mathbb{R}^n$ que en un sentido conveniente sea lo suficientemente parecido al funcional de actividad biológica como función del vector de descriptores moleculares, en otras palabras

> $(\mathfrak{X}, \|\cdot\|_\infty)$ denota el subespacio normado sobre $\mathbb{R}$ en el que el conjunto de vectores es el conjunto de funcionales definidos en Ω y con rango acotado en el sentio usual; mientras que $y : \Omega \subseteq \mathbb{R}^n \longrightarrow \mathfrak{R}$ será acotada y medible respecto de la σ-álgebra de Borel restringida a su dominio.
>
> El problema de identificación de parámetros que requiere solución es determinar el vector $\widehat{a}$ que que caracteriza al funcional lineal, $f_{\widehat{a}}$, que resuelve
>
> $$\mathcal{P}\ \mathbf{2.} \qquad \begin{cases} \text{mín} & \|f_a - y\|_\infty^2 \\ sa & a \in \mathbb{R}^n \end{cases}$$
>
> toda vez que la única información dada es la matriz de datos con error $\tilde{X}$ y el vector de mediciones con $\tilde{y}$.

Bajo las condiciones establecidas no es fácil brindar garantía de resolver el problema $\mathcal{P}2$. Lo usual es tratar de obtener una muestra de valores de y para valores conocidos de x y restringir el dominio del funcional a una vecindad común a los a los vectores de la muestra de la variable independiente, con el fin de garantizar que restringido al nuevo dominio las muestras brindan suficiente información global del comportamiento de y, de esta forma el problema $\mathcal{P}1$ brinda aproximaciones locales dependiendo de quién sea la matriz de datos X.

En un problema QSAR existen descriptores moleculares que por su naturaleza no pueden ser evaluados de forma exacta para una molécula dada, ejemplos de ello son el volumen molecular, la refractividad molar y descriptores relacionados con la distribución molecular de cargas entre otros.

Que los descriptores moleculares puedan ser dados con errores de medición significa que en general no tendremos un problema de regresión lineal múltiple de la forma usual. Hemos visto que la solución del problema de optimización por MCO, método empleado en los problemas de regresión lineal, depende del buen condicionamiento de la matriz de datos, y la matriz de datos como hemos dicho, en general se conoce con errores de medición, mismos que no pueden ser ignorados cuando la matriz está mal condicionada. Atenderemos este problema primero desde un punto de vista estadístico y después desde uno determinista que surge de la natural acotación de la función de actividad, factor que no es considerado en los problemas de regresión usuales.

Antes de continuar haremos algunas aclaraciones de notación para esta sección. $\tilde{X} \in M_{m\times n}\mathbb{R}$, Ω y δy serán como antes; se entiende δX_j como la variable aleatoria que denota el error de medición que se comete para el j-ésimo descriptor, δX_{ij} corresponde al error cometido al calcular el valor del $j-$ésimo descriptor para la i-ésima molécula en la muestra.

Las variables aleatorias $\{\delta y, \delta X_1, ...\delta X_n\}$ se consideraran con distribuciones independientes dos a dos. Nuevamente reservamos μ y σ^2 para denotar la media y varianza de δy, mientras que, análogamente μ_i se empleará para referirnos a la esperanza de δX_i, mientras que por haber reservado a lo largo de este trabajo las la notación σ_i para referirnos a valores singulares de una matriz, entonces ocuparemos la notación ζ_i para referirnos a la varianza de δX_i.

Proposición 2.3.1. *Dados, $x, x_{1\cdot}, ..., x_{m\cdot} \in \Omega \subset \mathbb{R}^n$, $\tilde{X}$ y $\tilde{y}$, se satisface*

$$\begin{aligned} E\left(|x^t a - y(x)|^2 \,\middle|\, \tilde{X}, \tilde{y}\right) &= \frac{1}{m}\left\|(\tilde{X} - \boldsymbol{\mu})a - (\tilde{y} - 1^t\mu)\right\|^2 \\ &+ \left\|diag(\vec{\zeta})a\right\|^2 + \sigma^2 \end{aligned}$$

donde 1^t se entiende como el vector de la dimención adecuada con cada coordenada igual a 1 (equivalente 1_m en ese caso), a es un vector arbitrario en $\mathbb{R}^n$, $\boldsymbol{\mu}$ es la matriz con $\mu_j 1_m$ por j-ésima columna y $\vec{\zeta} = (\zeta_1, ..., \zeta_n)^t$.

Demostración. Primero,

$$\begin{aligned} \|Xa - y\|^2 &= \left\|\tilde{X}a - \tilde{y} + \delta y - \delta X a\right\|^2 \\ &= \left\|\tilde{X}a - \tilde{y}\right\|^2 + \|\delta X a - \delta y\|^2 - 2(\tilde{X}a - \tilde{y})^t(\delta X a - \delta y) \\ &= \left\|\tilde{X}a - \tilde{y}\right\|^2 + \|\delta X a\|^2 + \|\delta y\|^2 \qquad (2.22) \\ &\quad -2\left((\delta X a)^t(\delta y + \tilde{X}a - \tilde{y}) - \delta y^t \tilde{X}a + \delta y\tilde{y}\right). \end{aligned}$$

Pensando en las filas de X como vectores elegidos en $\Omega \subseteq \mathbb{R}^n$ de forma aleatoria e independiente tanto de y como de cualquiera de los posible errores de medición cometidos, nos interesa la esperanza condicionada por $\tilde{X}$ y $\tilde{y}$ en el extremo izquierdo de (2.22); pero,

$$\begin{aligned} E(\|\delta X a\|^2 |\tilde{X}, \tilde{y}) &= \sum_{i=1}^{m}\sum_{j=1}^{n} a_j^2 E(\delta x_{ij}^2|\tilde{X},\tilde{y}) + \sum_{i=1}^{m}\sum_{j\neq k} a_j a_k E(\delta x_{ij}\delta x_{ik}|\tilde{X},\tilde{y}) \\ &= m\left(\left\|diag(\vec{\zeta})a\right\|^2 + (\vec{\mu}^t a)^2\right); \end{aligned}$$

$$E(\|\delta y\|^2 |\tilde{X}, \tilde{y}) = \sum_{i=1}^{m} E(\delta y_i^2|\tilde{X},\tilde{y}) = m(\sigma^2+\mu^2);$$

$$E(\delta X \delta y|\tilde{X},\tilde{y}) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_j E(\delta x_{ij}|\tilde{X},\tilde{y})E(\delta y_i|\tilde{X},\tilde{y}) = m\mu(\vec{\mu}^t a);$$

$$\begin{aligned} E((\tilde{X}a)^t(\delta X a)|\tilde{X},\tilde{y}) &= \sum_{i=1}^{m}\sum_{j=1}^{n} a_i a_j \sum_{k=1}^{m} \tilde{x}_{ki} E(\delta x_{ik}|\tilde{X},\tilde{y}) \\ &= \left(1^t \tilde{X} a\right)\left(\vec{\mu}^t a\right); \end{aligned}$$

$$E(\left(\tilde{X}a\right)^t |\tilde{X},\tilde{y}) = = \sum_{i=1}^{m} (\tilde{X}_{i\cdot}.a)E(\delta y_i|\tilde{X},\tilde{y}) = \mu(1^t \tilde{X} a);$$

$$E(\tilde{y}^t \delta X a|\tilde{X},\tilde{y}) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_j \tilde{y}_i E(\delta x_{ij}|\tilde{X},\tilde{y}) = (1^t\tilde{y})(\vec{\mu}^t a);$$

$$E(\tilde{y}^t \delta y|\tilde{X},\tilde{y}) = \sum_{i=1}^{m} \tilde{y}_i E(\delta y_i|\tilde{X},\tilde{y}) = \mu(1^t\tilde{y});$$

Retomando los extremos en (2.22) y por la linealidad de la esperanza matemática, para cada terna $X \in M_{m\times n}(\mathbb{R})$, $a \in \mathbb{R}^n$ y $y \in \mathbb{R}^m$, se verifica

$$
\begin{aligned}
E \quad & \left(\|Xa-y\|^2 \,|\tilde{X},\tilde{y}\right) \\
& = \left\|\tilde{X}a-\tilde{y}\right\|^2 + m\left(\left\|diag(\vec{\zeta})a\right\|^2 + (\vec{\mu}^t a)^2\right) + m(\sigma^2+\mu^2) \\
& +2\left(-m\mu(\vec{\mu}^t a) - (1^t\tilde{X}a)(\vec{\mu}^t a) + \mu(1^t\tilde{X}a) + (1^t\tilde{y})(\vec{\mu}^t a) - \mu(1^t\tilde{y})\right) \\
& = \left\|\tilde{X}a-\tilde{y}\right\|^2 - 2(\mu+\vec{\mu}^t a)1^t\left(\tilde{X}a-\tilde{y}\right) \\
& +m\left((\vec{\mu}^t a)^2 - 2\mu(\mu^2-\vec{\mu}^t a)\right) + m\left(\left\|diag(\vec{\zeta})a\right\|^2+\sigma^2\right) \\
& = \left\|\tilde{X}a-\tilde{y}\right\|^2 - 2(\mu+\vec{\mu}^t a)1^t\left(\tilde{X}a-\tilde{y}\right) + m(\mu+\vec{\mu}^t a)^2 \\
& +m\left(\left\|diag(\vec{\zeta})a\right\|^2+\sigma^2\right) \\
& = \left\|\tilde{X}a-\tilde{y}-1^t(\mu-\vec{\mu}^t a)\right\|^2 + m\left(\left\|diag(\vec{\zeta})a\right\|^2+\sigma^2\right)
\end{aligned}
$$

Por último, para $i=1,...,m$ sabemos que:

$$
\begin{aligned}
E\left(|x^t a - y(x)|^2 \,\middle|\, \tilde{X},\tilde{y}\right) &= \frac{1}{m}\sum_{i=1}^{m} E\left(|x_{i\cdot}^t a - y_i|^2 \,\middle|\, \tilde{X},\tilde{y}\right) \\
&= \frac{1}{m} E\left(\|Xa-y\|^2 \,\middle|\, \tilde{X},\tilde{y}\right)
\end{aligned}
$$

,
Terminando así la prueba. □

La proposición 2.3.1 ofrece información valiosa que comentaremos en extenso, lo primero es que en general no podemos resolver el problema $\mathcal{P}1$ de forma directa; pero, si condicionamos el problema respecto de los errores de medición para la matriz de datos y las mediciones de actividad, entonces podemos fijarnos en:

$$
\mathcal{P}\ \mathbf{3.} \qquad \begin{cases} \text{mín} & \frac{1}{m}\left\|(\tilde{X}-\boldsymbol{\mu})a-(\tilde{y}-1^t\mu)\right\|^2 + \left\|diag(\vec{\zeta})a\right\|^2+\sigma^2 \\ sa & a\in\mathbb{R}^n \end{cases}
$$

para buscar una forma de hallar, conociendo sólo los datos con error, al vector a que brinde una mínima acotación del valor esperado de la variable $\|Xa-y\|^2$.

Entre lo que debemos resaltar, es que dado el vector de mediciones $\tilde{y}$ y la matriz $\tilde{X}$, la solución del problema $\mathcal{P}3$ no depende de m ni de σ^2 y por ello, denotando por u_a al vector unitario en la dirección de a, es equivalente al problema

$$
\mathcal{P}\ \mathbf{4.} \qquad \begin{cases} \text{mín} & \left\|(\tilde{X}-\boldsymbol{\mu})a-(\tilde{y}-1^t\mu)\right\|^2 + \left\|diag(\vec{\zeta})u_a\right\|^2\|a\|^2. \\ sa & a\in\mathbb{R}^n \end{cases}
$$

En el mejor de los casos podemos suponer que los errores de medición tanto para las variables independientes como para la variable independiente son bien comportados, el valor esperado de los errores de medición es nulo y las varianzas son pequeñas; aún en este escenario correspondiente a hipótesis más débiles que las del problema de regresión lineal, se advierte que no es suficiente minimizar respecto de a al funcional que se define por $\left\|\tilde{X}a - \tilde{y}\right\|$.

Hasta ahora hemos supuesto que a puede ser cualquier vector en el espacio $\mathbb{R}^n$, pero hemos comentado con insistencia que el valor de la variable independiente es acotado, digamos $|y(x)| \leq \gamma$ para cualquier x en Ω, la restricción inmediata que debe considerarse antes de resolver el problema $\mathcal{P}4$ es $(\tilde{x}_{i.}^t a)^2 \leq \gamma^2$.

Ahora, definiendo

$$\begin{array}{rcl} r : \mathbb{R}^m & \longrightarrow & \mathbb{R}^m \\ r(a) & \longmapsto & \tilde{X}a - \gamma 1_m \end{array} \tag{2.23}$$

se llega a que el problema por resolver es

$$\mathcal{P}\ \mathbf{5.} \qquad \left\{ \begin{array}{ll} \text{mín} & \left\|(\tilde{X} - \boldsymbol{\mu})a - (\tilde{y} - 1^t\mu)\right\|^2 + \left\|diag(\vec{\zeta})u_a\right\|^2 \|a\|^2 \\ sa & a \in \Omega_r = \{a \in \mathbb{R}^n : r(a) \leq 0_m\}. \end{array} \right.$$

Ya hemos aclarado que valores pequeños de m son preferibles, al menos para hallar una acotación más pequeña del valor esperado de $|x^t a - y(x)|$, con x corriendo sobre Ω, y cuando sólo nos interesa encontrar una buena aproximación de la solución del problema $\mathcal{P}1$, independientemente de si la variable independiente y es en una norma adecuada muy semejante a un funcional lineal o no.

También es cierto que estamos interesados en lograr una solución única para la aproximación pues el problema originar es un problema de optimización puramente convexo; es decir, tiene solución única. Para atacar nuestro problema de ajuste buscamos una combinación para m y la selección de filas de para $\tilde{X}$ que hagan de dicha matriz una cuadrada de rango completo, $m = n$.

Con las condiciones recién exigidas para $\tilde{X}$ y sabiendo que $\nabla r_i = 2(x\ t_{i.}a)x_{i.}$, se sigue de inmediato que el conjunto de vectores gradiente $\nabla r_1(a), .., \nabla r_n(a)$ es siempre linealmente independiente, sin importar quién sea a, y por ende, cualquier solución de $\mathcal{P}5$ es un punto regular de Ω_r; además, el conjunto de indices activos es no vacío sólo cuando a satisface $(x^t a)^2 = \gamma^2$.

El teorema 1.3.7 nos garantiza en este caso que cualquier solución de $\mathcal{P}5$, a_0, existen escalares no negativos y no todos nulos $\lambda_1, ...\lambda_n$ tales que a_0 es la solución única del problema de optimización puramente convexo:

$$\mathcal{P}\ \mathbf{6.} \qquad \left\{ \begin{array}{ll} \text{mín} & \left\|(\tilde{X} - \boldsymbol{\mu})a - (\tilde{y} - 1^t\mu)\right\|^2 + \left\|diag(\vec{\zeta})u_a\right\|^2 \|a\|^2 \\ & +(\vec{\lambda}^t(\tilde{X}a - \gamma 1_n))^2 \\ sa & a \in \mathbb{R}^n; \end{array} \right.$$

$\vec{\lambda}^t = (\lambda_1, ..., \lambda_n)$.

Por otro lado, si $a, b \in \Omega_r$ y $k \in [0, 1]$, entonces

$$|x^t((1-\lambda)a + \lambda b)| \leq (1-\lambda)|x^t a| + \lambda |x^t b| \leq k;$$

es decir, $\mathcal{P}5$ es puramente convexo, y por ello existe un único vector $\vec{\lambda}$ que hace equivalentes a $\mathcal{P}5$ y$\mathcal{P}6$.

Definiendo

$$\lambda^2_{a,\vec{\zeta}} = \left\| diag(\vec{\zeta}) u_a \right\|^2 + (\vec{\lambda}^t (\tilde{X} u_a - \frac{\gamma}{\|a\|} 1_n))^2,$$

llegamos a que, en general, el problema de optimización a resolver es de la forma:

$$\mathcal{P}\ \mathbf{7.} \qquad \begin{cases} \text{mín} & \left\| (\tilde{X} - \boldsymbol{\mu}) a - (\tilde{y} - 1^t \mu) \right\|^2 + \lambda^2_{a,\vec{\zeta}} \|a\|^2 \\ sa & a \in \mathbb{R}^n. \end{cases}$$

2.3.2. La hipótesis de linealidad QSAR

Uno de los objetivos de cualquier análisis de farmacología del tipo QSAR o afines, es lograr observar de forma aislada cómo interactúa un grupo de fármacos con respecto a un receptor común, la intención es explicar el comportamiento de la función de actividad o de la respuesta biológica observada a partir de las únicas variables posibles en el sistema, que se desprenden del ligando como variable en los sistemas termodinámicos observados.

Se espera de la actividad o la respuesta biológica, como funciones acotadas, que si los mecanismos de integración entre el grupo de ligados y el receptor pueden considerarse que son los mismos, entonces la principal hipótesis de una análisis QSAR, en un nivel fundamental, es que la actividad y la respuesta se encuentran en función de las características mensurables de ese grupo de compuestos.

Sabemos que tanto la respuesta como los descriptores moleculares y la respuesta biológica son todas funciones reales acotadas, una hipótesis adicional de nuestro análisis será que cada descriptor y función de actividad con las que se trabaje sean elementos de un espacio de Hilbert.

Lo que tenemos hasta ahora es una generalización de la hipótesis de linealidad de una análisis QSAR. Si aceptamos cada una de las premisas anteriores, lo que se hace en un análisis de este tipo es utilizar un conjunto muestral de ligandos para los que se realizan mediciones de actividad o respuesta, a partir de los cuales se utiliza toda la herramienta teórica disponible para buscar entre los cientos de descriptores moleculares conocidos, solo aquellos que pueden ser calculados sin pruebas experimentales, entre aquellos se busca después conjuntos linealmente independientes que pertenezca una base del espacio, digamos, $\{x_1, x_2, ..., x_n, ...\}$ para la cual, al ser ortonormalizada por el procesos de Gram-Schmidt (*ver* [8]) y resultar en $\{x_{\perp,1}, ...\}$, el término general de la serie de Fourier de la función de actividad converja lo más rápido posible a 0:

$$Act = \sum_{j=1}^{\infty} < x_{\perp,j}, Act > x_{\perp,j}.$$

Nuestra intención al generalizar la hipótesis de linealidad de un análisis QSAR es la de tener un marco de referencia que nos permita utilizar distintas herramientas de regularización que ayuden en el proceso de diseñar nuevos medicamentos, con esta generalización pretendemos dar un sentido físico a los coeficientes de los modelos lineales QSAR. Ahora, mediante un cambio de base los coeficientes del modelo lineal son de la forma $a_{\perp,j} =< x_{\perp,j}, Act >$.

Formulado así, aceptamos que los coeficientes del modelo miden una dependencia entre la actividad y los descriptores moleculares que está sujeta a una noción de ángulo entre vectores, una lectura informal pero clarificarte de esto es:

$$cos(\angle(Act, x_{\perp,j})) = \frac{< x_{\perp,j}, Act >}{< Act, Act >} = \frac{a_{\perp,j}}{< Act, Act >}.$$

En general no pedimos que el diámetro del rango de los descriptores $x_1, x_2, ...$ se anule conforme j tiende a ∞, significa que desde que la respuesta y la actividad son acotadas, entonces el valor absoluto de los coeficientes en el término general de la serie de Fourier converge a 0, sin depender del ligando, de otra forma:

$$\begin{array}{rcl} |a_j(\mathfrak{L})| & \longrightarrow & 0 \\ j & \longrightarrow & \infty \end{array} , \forall \mathfrak{L} \in \Xi.$$

Para nuestro problema de optimización significa que si $a_0 = \sum_{j=1}^{n} \frac{u_j^t y}{\sigma_j} v_j$ es la solución exacta del problema, sin errores de medición, $\mathcal{P}1$, entonces debe satisfacerse la condición de Picard:

El valor $|u_j^t|$ converge a 0 más rápido que σ_j, a partir de un cierto natural j_0.

Una forma metodológica de seleccionar descriptores que respalde esto es aquella consistente con una selección del tipo paso a paso, donde el criterio de selección sea un incremento en el estadístico R^2_{ajus}, que como ya se ha expuesto, corresponde a una selección que en ∞ converge a la solución real de la actividad, consistente con la hipótesis de linealidad QSAR en general.

Lo anterior se sigue de hipótesis QSAR debilitadas, significa que es pertinente para un análisis QSAR en general. Son condiciones necesarias para poder llevar a cabo procesos de regularización que como hemos visto se sugieren de una forma natural ante esta formulación.

Solución de Tikhonov

Para un análisis QSAR, entre los requerimientos mínimos está el hecho de que tanto las mediciones del valor de los descriptores como de los respectivos datos de actividad sean confiables, que los errores de medición sean bien comportados. Partimos de confiar en que los procedimientos para las mediciones son los más adecuados y que los instrumentos involucrados en las mediciones se encontraban en óptimas condiciones.

Por otro lado dada la matriz $\tilde{X}$, es decir, dado el vector $\vec{\zeta}$, la relación entre a y $\lambda_{a,z\vec{e}ta}$ es uno a uno, con esto en mente nos permitimos una simplificación de notación: $\lambda = \lambda_{a,z\vec{e}ta}$.

Volviendo al problema, $\mathcal{P}7$, la forma que aquí sugerimos como alternativa para abordarlos es mediante lo que se conoce como *solución regularizada tipo Tikhonov*. Comienza con suponer que λ es un valor dado, denominado *parámetro de regularización*, con lo que el primer paso es resolver el problema

$$\mathcal{P}\ \mathbf{8.} \qquad \begin{cases} \text{mín} & \left\| \tilde{X}a - \tilde{y} \right\|^2 + \lambda^2 \|a\|^2 \\ sa & a \in \mathbb{R}^n. \end{cases}$$

Tomando en cuenta que es un problema convexo basta con verificar condiciones de optimalidad de primer orden y por la linealidad del operador de derivación no es difícil ver la solución se obtiene de forma análoga a como se infirieron las ecuaciones normales para el problema sin el segundo sumando, la solución de Tikhonov, dependiente del parámetro de regularización, es:

$$a_\lambda = \sum_{j=1}^{n} \varphi_j \frac{u_j^t y}{\tilde{\sigma}_j} v_j;$$

donde $\varphi_1, ..., \varphi_n$ son los filtros de Tikhonov, que dependen de los valores singulares de $\tilde{X}$, $\tilde{\sigma}_1, ..., \tilde{\sigma}_n$, y se definen por:

$$\varphi_j = \frac{\tilde{\sigma}_j^2}{\tilde{\sigma}_j^2 + \lambda^2}.$$

El proceso de regularización en este caso depende del criterio que se elija para determinar el mejor valor del parámetro λ. Algunos criterios pueden consultarse en el libro de Hansen [10], en donde además se detalla el motivo por el que la es importante para la pertinencia de la solución de Tikhonov que se observe la condición de Picard, que en el caso discreto, como los que atendemos, implica necesariamente que el valor esperado $y(x)$, dada la matriz de datos sin perturbación X, sea una combinación lineal de los las componentes de x, es decir, que se verifique la primera hipótesis del problema de regresión, sin la hipótesis de que y dada X sea una variable aleatoria de tendencia central.

Para nosotros es suficiente por ahora con justificar el por qué se sugiere también en este trabajo resolver el problema de ajuste de parámetros como una de regularización con solución tipo Tikhonov, lo importante es que esta solución minimiza el error de sobre estimación de parámetros y permite hacerlo sin preocuparse por ´disponer de demasiadas muestras de bioensayos con respecto a la cantidad de descriptores que sean empleados para un análisis QSAR.

2.4. Resumen

La parte formal de revisión de la metodología QSAR, para familias de ligandos cuya función sea la inhibición de algún tipo de actividad enzimática, termina aquí, sólo con un resumen de una sugerencia de pasos metodológicos:

1. Identificar con claridad el tipo de familia que se estudia, para saber si es o no pertinente un análisis QSAR;

2. definir con formalidad una función de actividad y/o de respuesta;

3. proponer modelos sobre la relación entre respuesta y actividad;

4. proponer explicita y formalmente una función de eficiencia, que sea un criterio de optimalidad y comparación entre ligandos;

5. realizar una cuidadosa selección inicial de descriptores para el análisis QSAR, selección realizada por una especialista en farmacología o campos afines, la finalidad es evitar sesgos estadísticos;

6. realizar una análisis QSAR de actividad múltiple si es necesario para determinar de forma única a la función de eficiencia;

7. identificar lo mejor posible el conjunto de restricciones sobre el que se optimizará la función de eficiencia;

8. optimizar la eficiencia en la familia de candidatos a fármacos,

9. realizar el esto de los procesos y pasos de la metodología QSAR.

Para una análisis de actividad múltiple en particular, si la primera selección de descriproes es suficiente, es decir, si resulta en un conjunto de descriptores menor o igual que la cantidad de moléculas probadas en bioensayos, menos una molécula que será a donde se trasladará el origen del espacio en que se obtengan los parámetros del modelo lineal, entonces se propone abordar el problema como un problema de regularización, con solución tipo Tikhonov, siempre que exista confianza o evidencia suficiente de que se satisface la condición discreta de Picard y que los errores de medición son bien comportados.

Si no es suficiente, pero cuando menos los errores de medición son variables aleatorias de tendencia central, entonces se sugiere realizar una segunda selección de descriptores mediante un proceso de tipo *paso a paso*, donde los criterios de selección sean:

- observación de incremento en el estadístico R^2_{ajus} (hipótesis de linealidad en el paso al ∞);

- $\sigma_n >= 1$, donde σ_n es el más pequeño de los valores singulares de la matriz de datos del problema de regresión, de acuerdo con lo cuál se verifica que la varianza del estadístico de parámetros será proporcional a la cantidad de muestras y la varianza del error de medición de la actividad (*ver* 1.31).

Si además se supone normalidad en los errores si medición de la actividad, entonces se puede elegir la relación entre la cantidad de muestras y la cantidad de descriptores mediante la determinación arbitraria de parámetros para la confianza en la prueba de hipótesis de linealidad presentada en este trabajo.

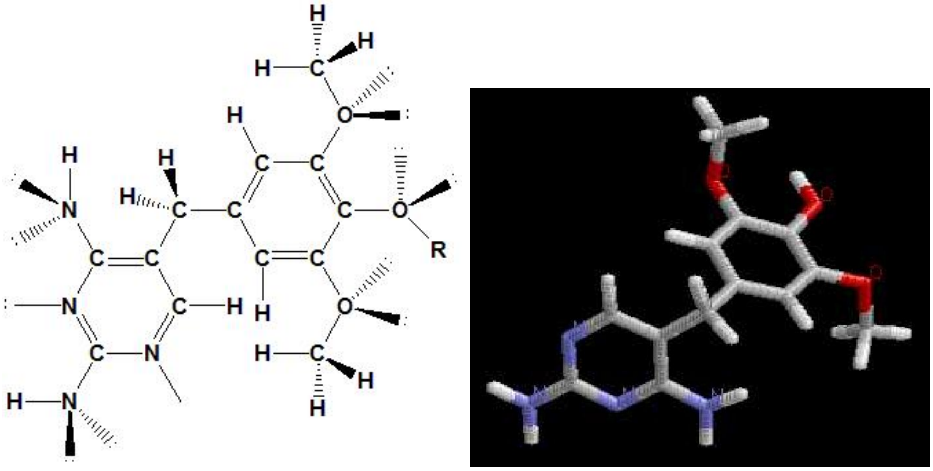

Figura 2.3: Grafos 2D y 3D de la estructura base, **R** indica el lugar del sustituyente.

2.5. Familia molecular y la selección de descriptores para estudio de caso

Por motivos de titulación hemos colocado aquí un ejemple de lo que recién hemos sugerido, con datos reales de un juego de dieciocho moléculas, pero en el trabajo con especialistas de farmacología se han detectado pequeñas inconsistencias o insuficiencias en los datos experimentales que hacen de esta sección sólo un ejemplo académico y no un resultado científico aún, a la fecha estos datos se encuentran bajo revisión.

Con relación a la sección anterior, vamos a trabajar con una familia de antibióticos con una función de actividad para la que existe una concentración que logra una inhibición al 100 % de la actividad del receptor, la estructura base es la que se muestra en la figura 2.3,el conjunto de descriptores resultado de una previa selección por especialistas en farmacología se presenta en la tabla 2.1. Para el desarrollo de este trabajo las definiciones de los descriptores moleculares son las contenida en [30] y [32].

Teniendo receptor común, estructura base, restricciones para la función de actividad y una selección previa de descriptores, el siguiente paso es buscar las restricciones naturales y adicionales de los descriptores para la caracterización del conjunto de restricciones Ω.

Atendemos brevemente a esto porque en dicha búsqueda requerimos evitar en lo posible las dependencias lineales entre descriptores, si somos capaces de identificar a ciencia cierta la parecencia de ese tipo de relaciones entre los descriptores; entonces, seremos capaces de aprovechar nuestra capacidad de cómputo en la revisión de una mayor cantidad de posibles modelos entre los que pueden ser propuestos con m datos muestrales y una cantidad menor de variables de predicción que la cantidad de descriptores previamente seleccionados.

Las primeras restricciones que se imponen sobre el conjunto de descriptores y que no se infieren de la descripción de los mismos, son aquellas que dependen de la información

Desccriptor	**Símbolo**	**Descripsión**
x_1	nC	No. de átomos de carbono.
x_2	nH	No. de átomos de hidrógeno.
x_3	nN	No. de átomos de nitrógeno.
x_4	nO	No. de átomos de oxígeno.
x_5	nS	No. de átomos de azufre.
x_6	nBr	No. de átomos de bromo.
x_7	nF	No. de átomos de Fluor.
x_8	nAT	No. total de átomos.
x_9	nHAcc	No. de átomos aceptores para puentes de hidrógeno (N, O, F).
x_{10}	Sv	Suma de los volúmenes atómicos de van der Waals.
x_{11}	NBT	Número de enlaces totales.
x_{12}	nBO	No. de enlaces no de hidrógeno.
x_{13}	nBM	No. de enlaces múltiples.
x_{14}	RBN	No. de enlaces rotables.
x_{15}	nAB	No. de enlaces aromáticos.
x_{16}	qpmax	Máxima carga positiva.
x_{17}	qnmax	Máxima carga negativa (Valor absoluto de la mínima carga negativa).
x_{18}	Qpos	Carga total positiva.
x_{19}	Qneg	Carga total negativa (valor absoluto de la suma de las cargas negativas).
x_{20}	Qmean	Media del valor absoluto de las cargas.
x_{21}	Q2	Carga cuadrada total (sumatoria de los cuadrados de las cargas).
x_{22}	PSA	Área de la superficie polar .
x_{23}	MLOGP	Coeficiente de partición, octanol-agua, de Moriguchi.
x_{24}	MR	Refractividad molar.
x_{24}	η	Índice de refractividad molar.

Tabla 2.1: Descriptores preseleccionados para el análisis QSAR de la familia Ξ.

adicional de la que se dispone, que en conjunto se conocen como *regla Lipinski (Ro5) en el diseño de nuevos compuestos bioactivos* o también llamada la regla del número 5; indica que una sustancia podría tener problemas de permeabilidad celular si viola o no cumple más de uno de los siguientes requerimientos:

1. Peso molecular (PM) menor a 500 g/mol;
2. número de donadores de enlace de hidrógeno (nOHNH) menor o igual a 5 (suma de -OH y -NH en la molécula);
3. Número de aceptores de enlace de hidrógeno (nON) menor o igual a 10 (suma de -O y -N en la molécula);
4. Log P calculado (C log P) menor a 5.

Otras reglas adicionales consideran a los siguientes parámetros como criterios de selección de sustancias con potencial actividad biológica:

- **Log P** o coeficiente de partición, es un parámetro que mide la hidrofobicidad de la molécula. La hidrofobicidad afecta la absorción, la biodisponibilidad, las interacciones hidrófobas del fármaco y el receptor, el metabolismo, así como la toxicidad. Éste debe de estar entre -0.4 y 5.6, en promedio en 2.52. Este criterio reemplazará para nosotros a la cuarta condición de la regla de Lipinski.
- **TPSA** (Área Polar Molecular Superficial), es un parámetro usado en las propiedades del transporte del fármaco, el área polar superficial es definida por la suma de las superficies de los átomos polares (oxígeno, nitrógeno e hidrógenos unidos a estos átomos) en la molécula. Para la absorción oral los valores normales se encuentran entre 100 y 150 Å. Para atravesar la barrera hematoencefálica debe ser menor o igual a 90 Å.
- Peso Molecular, debe de estar entre 160 y 480 uma, y en promedio 357 uma.
- Número de átomos, entre 20 y 70 átomos, y en promedio 48.
- Número de enlaces rotables, este parámetro topológico simplemente es una medida de flexibilidad molecular. Se ha demostrado ser un parámetro muy útil para la biodisponibilidad oral de los fármacos. Deben de presentar entre 2 y 8 enlaces rotables.
- Número de anillos,entre 1 y 4 anillos.
- Refractividad Molar, descripción del tamaño molecular y la tendencia a participar en interacciones de dispersión. Los valores aceptados se encuentran entre 40 y 130, en promedio 97.

La previa selección de descriptores corresponde a la primera reducción de posibles descriptores para el modelo realizada por especialistas de farmacología de acuerdo con criterios relacionados con los conocimientos empíricos y estadísticos de en su área de

especialidad. Aún es un conjunto que requiere ser reducido para lograr una matriz de datos de rango completo, condición necesaria, no suficiente, para que exista solución única del problema de optimización que consiste en encontrar una aproximación por mínimos cuadrados de los vector a y b en el modelo QSAR, caracterizado por la ecuación (2.1).

Los descriptores preseleccionados serán clasificados por su naturaleza, las relaciones entre ellos rerán exploradas en bloques de descriptores basados en tal clasificación:

Grupo 1. Combinaciones lineales de las cantidades de átomos de cierto tipo presentes en la molécula, descriptores de x_1 a x_{11};

Grupo 2. Combinaciones lineales de las cantidades de enlaces covalentes de cierto tipo presentes en la molécula, descriptores de x_{12} a x_{15};

Grupo 3. Funciones escalares cuyo argumento es el conjunto de cargas de una molécula, generado por el método de Gasteiger y Marsili (*ver definición* A.2.1), descriptores de x_{16} a x_{22};

Grupo 4. Parametrizaciones de propiedades físico-químicas, descriptores x_{23} y x_{25}.

Las relaciones más simples y directas que existen entre el conjunto de descriptores de la preselección son:

$$
\begin{aligned}
x_k &\in \mathbb{N}, \qquad k = 1, ..., 6, 10, ..., 14; \\
0 &\equiv x_2 + 3x_3 + x_6 + x_7 mod(2) \\
x_8 &= \sum_{1 \leq k \leq 7} x_k \\
x_9 &= x_3 + x_4 \\
x_{10} &= \sum_{1 \leq k \leq 6} v_k x_k \\
x_{12} &= x_{12} - x_2 \\
x_{14} &= x_{11} - x_2 - x_6 - x_{13} \\
x_{18} &= x_{19} \\
x_{20} &= \frac{2x_{18}}{\sum_{k \leq 6} x_k} \\
x_{24} &= \frac{x_{25}^2 - 1}{(x_{25}^2 + 1)\rho} \sum_{k \leq 7} p_k x_k \\
0 &\leq 2x_1 - x_2 + x_3 + 3x_5 - x_6 - x_7 - 2x_{0,1} + x_{0,2} - x_{0,3} - 3x_{0,5} + x_{0,6} + x_{0,7} \\
x_8 &\leq max\{x_8(\varsigma) : \varsigma \in \Xi_0\} \\
x_8 &\geq 20 \\
x_9 &\leq 10 \\
2x_{11} &\geq 2[x_{0,10} + \sum_{1 \leq k \leq 7} (x_{0,k} - x_k)] - 2 \\
2x_{11} &\leq 2x_{0,10} + \sum_{1 \leq k \leq 7} eV_k(x_{0,k} - x_k) \\
x_{14} &\geq 2 \\
2x_{18} &\leq \sqrt{x_{21} \sum_{k \leq 7} x_k} \\
x_{20} &\leq x_{16} \\
x_{20} &\geq -x_{17} \\
x_{22} &\leq 150 \\
x_{22} &\geq 100 \\
x_{23} &\leq 5{,}6 \\
x_{23} &\geq -0{,}4 \\
x_{24} &\leq 130 \\
x_{24} &\geq 40 \\
N_A \sum_{k \leq 6} w_k x_k &\leq 500 \\
N_A \sum_{k \leq 6} w_k x_k &\geq 160 \\
nOHNH &\leq 5
\end{aligned}
$$

en donde w_k es el peso en gramos de cada tipo de átomo y N_A es el número de Avogadro. Las restricciones sobre la cantidad de anillo ni sobre los donadores para la formación de puentes de hidrógeno figuran en las relaciones anteriores. Los detalle de la inferencia de las relaciones anteriores se encuentran disponibles al lector en el apéndice A.2.

Con las relaciones de igualdad se pone de manifiesto que cada uno de los descriptores con los que hemos trabajado se encuentran en función de los descriptores en el conjunto

$$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_{10}, x_{12}, x_{14}, x_{15}, x_{16}, x_{17}, x_{20}, x_{21}, x_{22}, x_{24}\};$$

Este conjunto es entonces lo que se conoce como una base de descriptores moleculares[4].

Siempre que sea posible, en el caso de los descriptores que son combinaciones lineales de los elementos de la base de descriptores, sobrará un análisis estadístico de independencia lineal de variables cuando se proceda a resolver el problema de regresión.

Las relaciones expuestas que son sintetizadas en esta sección serán las que determinarán el conjunto de restricciones para cualquier tipo de análisis QSAR que realicemos, en la etapa correspondiente a la optimización de los modelos lineales que sean propuestos para explicar el comportamiento local de las componentes de la energía libre.

2.5.1. Los modelos QSAR

El conjunto muestral de moléculas sobre las que se realizaron las mediciones experimentales de actividad se encuentran listadas en la tabla 2.2.

Nombre	id mol	Nombre	id mol
base	ς_0	kc1308	ς_{10}
kc131	ς_1	kc1310	ς_{11}
kc132	ς_2	kc1311	ς_{12}
kc134	ς_3	kc1314	ς_{13}
kc136	ς_4	kc1315	ς_{14}
kc1005	ς_5	kc1316	ς_{15}
kc1300	ς_6	kc1317	ς_{16}
kc1301	ς_7	kc1318	ς_{17}
kc1303	ς_8	kc1321	ς_{18}
kc1307	ς_9		

Tabla 2.2: Conjunto muestral Ξ_0, conformado por dieciocho elementos de la familia de moléculas de interés.

La presente sección está los resultados de un problema de regresión múltiple QSAR tradicional, que extenderemos empleando el criterio de selección de buenos modelos basados en incrementos de la magnitud del estadístico $R^2 ajus$ tanto los datos de actividad tradicionales ΔG como para los vectores de actividad $(\Delta G, \Delta H)$.

[4]No debe confundirse el concepto de base de descriptores moleculares con los conceptos de bases topológicas ni de bases en espacios vectoriales.

Por la naturaleza y complejidad de los procesos bioquímicos que estudian los análisis QSAR es suficiente un coeficiente de determinación de por lo menos 1/2 para considerar que un modelo lineal describe satisfactoriamente el fenómeno; es por este motivo que no descartaremos la hipótesis de linealidad fundamental de un modelo QSAR tradicional siempre que R^2ajus sea mayor o igual que $1/2$.

A continuación se presenta un ejemplo do los primeros resultados que obtenidos de la inclusión de un segundo dato de actividad para una familia de fármacos para el combate de la tuberculosis. Los datos experimentales ya su viabilidad para una análisis QSAR sigue en camino y es por ello que debe considerarse esto solamente como una ejemplo que ilustro lo desarrollada en este trabajo de tesis.

2.5.2. Regresión lineal múltiple

Lo primero en la obtención de modelos de regresión lineal será el aclarar a qué es lo que entenderemos por un conjunto de descriptores que generan un buen modelo de regresión. Independientemente de si el el error de medición observa normalidad o no, para nosotros los buenos modelos de regresión serán, siempre que por lo menos la media del error de medición sea 0, los que cumplan tres condiciones:

- con un nivel de confianza $1-\alpha$ y región de rechazo $R^2ajus <$,5 no pueda descartarse la hipótesis de dependencia lineal entre la variable de actividad y el conjunto de descriptores que determina al particular modelo de regresión;
- no observa sobre-ajuste de parámetros;
- cumple las dos condiciones anteriores con la menor cantidad de variables posibles.

Algunas observaciones pertinentes sobre nuestro entendido de buenos modelos de regresión son, en primer lugar, que la región de rechazo para la prueba de hipótesis de linealidad se ha elegido así por una convención para los análisis QSAR, para los farmacéuticos un coeficiente de determinación mayor que .5 es suficiente para aceptar la existencia de correlación entre la actividad y el modelo que se proponga. Respecto de la condición de sobre-ajuste, la presencia del mismo dependerá de las necesidades de la investigación y la forma en que en cada problema de regresión se entienda por una varianza pequeña del estadístico de parámetros del modelo. En este trabajo la significación que daremos a "varianza pequeña" se desprende de los corolarios 1.4.2 y 1.4.4.

Sobre ajuste de parámetros

Al momento de trabajar con los datos maestrales debemos partir de suponer que el orden de la varianza en los errores de medición es tolerable, de otra forma ni siquiera tendría sentido utilizarlos para una análisis; luego, al modelo de regresión se le exige que su varianza, dada la matriz de datos X para el conjunto muestral de variables independientes, sea del mismo orden o menor que la del error de medición. El corolario 1.4.2 nos garantiza la igualdad

$$Var(\tilde{a}_1x_1 + \cdots \tilde{a}_nx_n|X) = \sigma^2(x_1, ..., x_n)C(x_1, ..., x_n)^t.$$

Pero como $C = (X^tX)^{-1} = V\Sigma^{-2}V$, donde $X = U\Sigma V^t$ es la DVS de X; se sigue

$$\sigma^2 xCx^t = \sigma^2 (Vx)\Sigma^{-2}(Vx)^t \leq \left(\frac{\sigma}{\sigma_n}\|x\|\right)^2.$$

Con el desarrollo anterior se justifica las formas en que determinaremos cuándo que no existe sobre-ajuste en un modelo de regresión para un análisis QSAR.

Definición 2.5.1 (Criterio de sobre-ajuste de parámetros.)**.** Si $\tilde{a}$ es la solución al problema de regresión lineal $ax \sim \tilde{y}$, con $a, x \in \mathbb{R}^n$, dada X y tal que $E(\delta y) = 0$; entonces, diremos que con factor de escalamiento λ y sobre el conjunto Dom no existe sobre-ajuste del vector de parámetros $\tilde{a}$ siempre que

$$\frac{\lambda max(\{||x|| : x \in Dom\})}{\sigma_n} \leq 1;$$

con $\sigma_1 \geq \cdots \sigma_n$ los valores singulares de X.

Con la definición 2.5.1 realzamos la importancia que para nosotros tiene conocer el conjunto de posibles descriptores para los cuales nos atreveremos a hacer predicciones sobre el valor que observará la actividad de un fármaco, en la familia de interés, que se corresponda con tal vector de descriptores.

Esta definición de sobre-ajuste redunda la importancia que par la estadística tiene, independientemente del tamaño de la muestra, una cantidad lo más pequeña que sea posible. Como hemos visto ya en secciones anteriores, si se construye un modelo de regresión a partir de uno dado mediante la inclusión de variables de predicción adicionales, el la magnitud del menor valor singular de la matriz de datos correspondiente será menor o igual que el correspondiente valor singular para el modelo inicial, no obstante, tal incremento en las variables de predicción genera un incremento en la dimensión del vector x y consecuentemente un natural incremento de $max(||x||Dom)$. En la siguiente sección explicaremos en qué forma entender así el sobre-ajuste de un modelo de regresión lineal es parte de la caracterización de la familia de moléculas, o implica que se verifique la condición de Picard figura entre las condiciones de suficiencia mínimas para un proceso de regularización tipo Thikonov.

La definición 2.5.1 y el corolario 1.4.4 dan una referencia del rango en que debe encontrarse la varianza de cada uno de los estadísticos $\tilde{a}_j$, $j = 1, ..., n$, ser considerada como tolerable, relación que es independiente de la matriz X y se desprendo sólo de la forma en que se encuentre caracterizado el conjunto Dom:

$$Var(\tilde{a}_j) \leq \left\{\frac{\sigma}{\lambda max(\{||x|| : x \in Dom\})}\right\}, \quad j = 1, ..., n.$$

En ausencia de información adicional asumiremos en cada caso que $max(\{||x|| : x \in Dom\}) = max(\{||x_1||, ..., ||x_m||\})$, mientras que *lambda* se elegirá como el supremo del conjunto de escalares que permite construir al menos un modelo de regresión sin sobre-ajuste para el conjunto muestral de ligandos y un subconjunto de los descriptores previamente seleccionados.

Normalidad en los errores y tamaños de muestra.

El conjunto muestral del que disponemos es pequeño, no obstante necesitamos conjuntos de entrenamiento y prueba para poder para tener por lo menos una somera validación interna. Aleatoriamente elegimos 14 de las 18 moléculas como conjunto de entrenamiento:

Conjunto de entrenamiento		
kc131	kc132	kc134
kc136	kc1005	kc1300
kc1301	kc1307	kc1310
kc1311	kc1314	kc1316
kc1317	kc1318	

Para poder proceder con la selección de buenos modelos de regresión lineal resta atender a la relación entre el tamaño de la muestra, m, y la mayor cantidad de variables de predicción que nos serán permitidas utilizar en los modelos de regresión lineal que se propongan como modelos QSAR, n. En cada caso consideramos al término independiente de los modelos lineales no nulo y como una variable de predicción constante.

Aceptamos *a priori* que los error de medición de las componentes de la actividad son variables aleatorias bien comportadas, es decir, son independientes entre si y sus distribuciones son ambas normales con media 0 y varianzas con magnitudes del mismo orden.

Para este análisis QSAR de actividad vectorial no disponemos de evidencia estadística que respalde la afirmación de que algún subconjunto de a lo más dos descriptores de los aquí considerados explique locamente y casi en su totalidad el comportamiento de alguna de las dos componentes de actividad, nuestro problema de regresión para el ajuste de parámetros será sin restricciones.

Por lo expuesto en la sección 2.1, en las ecuaciones matriciales destacadas en (2.19), las soluciones del ajuste de parámetros con actividad vectorial, para un conjunto de descriptores linealmente independientes y con no más de m elementos, serán las soluciones de los problemas de ajustes resueltos de forma independiente para ΔH y ΔS, a saber

$$\begin{cases} \tilde{b} &= (X^tX)^{-1}X^t\tilde{\Delta S} \\ \tilde{a} &= (X^tX)^{-1}X^t\tilde{\Delta H} \end{cases} .$$

La elección inicial de los descriptores que pueden ser considerados para nuestros modelos QSAR es cuenta con más elementos que la cantidad de muestras de actividad, lo que no nos exenta de tener que realizar una selección de los mejores modelos entre las soluciones de todos los problemas de regresión lineal múltiple que pueden ser planteados con los subconjuntos propios de menos de 17 elementos extraídos del conjunto de 25 posibles.

La ventaja operativa que tiene una primera selección de variables de predicción con base en criterios no estadísticos es que, como en este caso el conjunto de posibles variables independientes es relativamente pequeño y hace que el tiempo de cómputo que resolver cada uno de los posibles problemas de regresión sea relativamente pequeño.

El criterio que tenemos para un buen modelo incluye que para un conjunto de descriptores y un nivel de confianza $1-\alpha$, con $\alpha \in (0,1]$, no se rechace la hipótesis de linealidad para ninguno de los modelos resultantes de los problemas de regresión para ΔH y ΔS. Al respecto de este criterio emplearemos como principal herramienta, para la determinación de la cantidad máxima de descriptores en un modelo QSAR de actividad vectorial, a la relación entre el tamaño de muestra y la cantidad de variables de predicción que se desprende del lema 1.4.13, los resultados se muestran en la tabla 2.5.2, tolerando como mínimo un nivel de confianza de ,8.

NivConf	Max n/ΔH	Max n/ΔS
0.999	12.000	0.000
0.971	12.000	1.000
0.966	12.000	2.000
0.960	12.000	3.000
0.953	12.000	4.000
0.945	12.000	5.000
0.935	12.000	6.000
0.924	12.000	7.000
0.910	12.000	8.000

Tabla 2.3: Máxima cantidad de variables de predicción para cada componente de actividad respecto de un nivel de confianza específico para la prueba de hipótesis de linealidad.

En ausencia de acotaciones reportadas para las varianzas de las componentes de la actividad por parte de quienes realizaron las mediciones experimentales, hemos acotado a las mismas con la unidad simplemente por que una varianza mayor significaría que estas mediciones no pueden ser usadas aquí para los fines que se persiguen.

En lo tocante al sobre-ajuste o sobre-estimación de los modelos, de nuestra definición se hace evidente la relatividad del concepto, lo cierto es que en nuestro conjunto de entrenamiento las magnitudes de algunos de los descriptores considerados son del orden de 10^2, por fines de exploración relajaremos los criterios de selección para los modelos lineales; ponderaremos la importancia que de la norma del vector de descriptores con $\lambda = 10^{-1}$, implicando que un buen modelo observará una matriz de descriptores con valores singulares simultáneamente mayores. Sólo en el caso de un modelo dos construido con dos descriptores relajamos aún más el valor de λ por haber destacado notoriamente entre el resto de los modelos con la misma cantidad de descriptores.

Los resultados obtenidos por **RLM** se resumen en la tabla:

En el caso de este conjunto de datos no había información suficiente como para confiar en que se satisfacían las condiciones suficientes para el proceso de regularización mediante la solución de Thikonov, por eso es que concluimos aquí nuestro ejemplo y trabajo de revisión de la metodología QSAR.

No. Vars.	Descriptores	σ_1	σ_n
3	nC nHAcc	64.073567	9.673570
3	nHAcc Molweight	1150.434104	9.264467
5	nAT Molweight PolarArea Q2	1286.849500	9.321802
5	NBT Molweight PolarArea Q2	1287.545766	9.643755
5	Molweight PolarArea Volume Q2	1649.978017	9.944621

No. Vars.	Descriptores	$R^2_{ajus_{\Delta H}}$	$s^2_{\Delta H}$
3	nC nHAcc	0.704023	24007036.539256
3	nHAcc Molweight	0.702239	24151753.315155
5	nAT Molweight PolarArea Q2	0.769599	18688064.549305
5	NBT Molweight PolarArea Q2	0.777137	18076636.102846
5	Molweight PolarArea Volume Q2	0.765893	18988664.373755

No. Vars.	Descriptores	$R^2_{ajus_{\Delta S}}$	$s^2_{\Delta S}$
3	nC nHAcc	0.997557	7.169254
3	nHAcc Molweight	0.985167	43.522112
5	nAT Molweight PolarArea Q2	0.997514	7.293887
5	NBT Molweight PolarArea Q2	0.998243	5.155917
5	Molweight PolarArea Volume Q2	0.997437	7.520538

Tabla 2.4: Selección de conjuntos para posibles modelos QSAR, criterio R^2 y actividad múltiple.

Conclusiones

Comenzaremos por puntualizar que, para el una análisis QSAR, es muy importante solicitar reportes detallados de cada bioensayo realizado por un laboratorio, incluso para una análisis tradicional permite recuperar valiosa información sobre las características cinéticas y termodinámicas de la interacción *ligando-receptor*. La cantidad mínima de información relevante que debe reportarse o de bioensayos que deban realizarse dependerá de la dimensión del espacio de estados del sistema que se observe en cada bioensayo.

Un análisis QSAR con vector de actividad $(\Delta G, \Delta H)$, es asequible en la práctica y con suficiente información sobre los bioensayos, no necesariamente información proporcionada por un conjunto muestral con gran cantidad de elementos, es posible recuperar aproximaciones de los principales observables termodinámicos, incluido el cambio en la energía interna del sistema. Se tiene la hipótesis, aún no no estudiada por nosotros, de que conocer éste tipo de observables es de suma importancia por estar relacionados con los distintos mecanismos de acción que puede observar un fármaco, donde distinguir cuáles son las variables termodinámicas relevantes sería de gran utilidad para tratar de entender y modelar distintos mecanismos de acción entre un ligando y su blanco biológico, conocimiento que confiamos será relevante en el diseño *in silico* de nuevos medicamentos.

Hemos trabajado sólo en la etapa de la metodología QSAR que corresponde al al análisis, considerar a la actividad como un vector y no como un escalar nos abre la puerta a trabajar en las etapas anterior y posterior, con la finalidad de logra obtener, en etapas posteriores, modelos matemáticos basados en los descriptores moleculares y que sean capaces de describir el comportamiento y la acción terapéutica de un fármaco, no solo para sistemas en los que se observe una interacción del tipo *molécula-molécula*.

Por ahora también estará pendiente la optimización de los modelos QSAR recuperados, por requerir algo de teoría de programación entera, herramienta no contemplada en estas páginas; además del hecho de que no logramos, en el periodo de trabajo que comprende esta tesis, realizar el trabajo suficiente sobre la identificación de las relaciones funcionales entre descriptores que para el proceso de optimización harán del modelo lineal una no lineal. También es cierto que en materia de optimización aún no contamos con la seguridad de que sea adecuado el conjunto de restricciones que hemos logrado para hasta ahora.

Siempre que se elija resolver el principal problema de ajuste de parámetros de un análisis QSAR mediante Regresión Lineal Múltiple, la definición de sobre ajuste presentada brindará un criterio de selección de modelos y logrará una menor dependencia de la cantidad máxima de variables independientes que pueden emplearse en el modelo

respecto de la cantidad de ligandos que componen la muestra.

Un análisis de actividad vectorial no difiere esencialmente de tres análisis QSAR independientes con un mismo conjunto de descriptores, no involucra un esfuerzo o dificultad mayor que un análisis tradicional, la herramienta matemática es la misma en general, con la diferencia de que enfatiza la adecuada definición de una familia molecular QSAR, con la intención de evitar fallos graves de predictividad por la omisión de características relevantes sobre los mecanismos de interacción *ligando-receptor*.

Ante la problemática de la dificultad de un análisis QSAR para contar con conjuntos suficientemente grandes de datos maestrales provenientes de bioensayos, aunque no es la única forma, si se tiene suficiente confianza en la hipótesis de linealidad, se considera además que los descriptores pueden se conocidos con perturbación y se conoce una acotación confiable para los datos de actividad; en tal caso, resolver el problema de ajuste de parámetros del modelo lineal QSAR mediante un proceso de regularización con solución de Tikhonov se presenta de una forma natural como alternativa.

Apéndice A

Familias moleculares QSAR

Uno de los puntos de partida de los análisis QSAR es la determinación de un receptor biológico y un fármaco base, dados estos elementos constantes es razonable pensar que la variación de la respuesta biológica depende de la variación en el sustituyente. Cada descriptor molecular es una forma de medir una característica molecular, el cambio en cada descriptor de una familia de compuestos estudiado es por tanto una forma de medir un tipo particular de modificación en las moléculas de prueba, refleja un cambio en el sustituyente, las relaciones entre el cambio en los descriptores moleculares y la respuesta biológica de un fármaco son el objeto de estudio de un análisis QSAR.

Un primer problema es por supuesto la elección de los descriptores que realmente se relacionan con el cambio de la respuesta biológica, la cantidad de mediciones de las que se dispone por lo regular no es muy grande y sin embargo, son cientos los descriptores moleculares conocidos por la química. Este problema de selección no nos concierne en este trabajo, nuestra labor comienza en depositar nuestra confianza en la experiencia de especialistas que, con base en el conocimiento de su campo, eligen una pequeña cantidad de descriptores (receptor y estructura base dadas) que consideran son los más relacionados con la variación de la respuesta, disminuyendo así la posibilidad de recuperar modelos con poco sentido real.

Partiendo de esta selección previa de descriptores aún quedan observaciones metodológicas que realizar sobre los modelos QSAR, de las que se desprenderán los distintos modelos que se recuperen, dependiendo de cómo sea definida una familia de compuestos susceptible de un análisis QSAR.

El problema real implica mediciones de la respuesta biológica, que dependen de la ínfima concentración del fármaco que se requiere para causar una respuesta al 50 %, es claro que una concentración menor o igual a 0 es un dato absurdo, significaría que la respuesta biológica no depende del fármaco, luego ni siquiera tiene sentido realizar un análisis. Por otro lado una concentración infinita tampoco tiene sentido, ni siquiera lo tiene que la concentración sea tal que el volumen que ocupe sea mayor que el que ocupa la muestra de tejido en la medición experimental de la respuesta. Que la medición de la respuesta debe estar acotada superior e inferiormente por constantes positivas es una restricción natural.

Así como la respuesta, el rango de los descriptores también está sujeto a restricciones de acotación, algunas de ellas evidentes, como que la cantidad de átomos tiene que ser acotada, o que la cantidad total de átomos de cualquier elemento de una familia de moléculas no puede ser menor que la cantidad de átomos de la estructura base.

Veremos que la acotación para la respuesta y para los descriptores seleccionados son de vital importancia para la búsqueda de fármacos con el mejor efecto terapéutico posible, dentro de la familia en la que son considerados, desde la elección de herramienta matemática que será empleada para ello, hasta la forma en que se entenderá una familia de moléculas en el análisis QSAR.

A.1. Breve clasificación

Comencemos por profundizar en las hipótesis de un análisis QSAR. El primer aspecto es entender la importancia de la caracterización de la familia de compuestos orgánicos que se estudiará; uno de los aspectos de valía de un análisis QSAR tradicional es que busca obtener información del comportamiento de la familia de ligandos prescindiendo de información explícita del receptor, que claramente es importante. La información sobre el receptor es inherente al fármaco que se elige como estructura base, la única información de la que un análisis QSAR dispone sobre el receptor se encuentra condensada en el fármaco base qué por principio es afín al receptor.

Las características constitutivas, electrónicas, topológicas, geométricas y fisicoquímicas tienen garantía de alcanzar una concentración suficiente como para reaccionar con el receptor y producir una respuesta deseada, pese a que la concentración pueda no ser tan pequeña como se quisiera. La estructura base y sus propiedades son en primera instancia el factor que determina las primeras restricciones que se establecen para caracterizar la pertenencia de un compuesto orgánico a una familia de moléculas susceptible de un análisis QSAR.

Haremos distinciones entre conjuntos de moléculas orgánicas a considerar como familias de ligandos en análisis QSAR, los primeros criterios de clasificación son independientes de las relacione inherentes a la definición de los descriptores, son criterios impuestos por el blanco biológico para el cuál se pone en marcha toda la maquinaria de una análisis QSAR y lo que vulgarmente puede llamarse una molécula pivote.

A.1.1. Familias determinadas por un receptor o una estructura base

Tres útiles conceptos matemáticos son asociados a una molécula como un objeto real en el espacio, el grafo, las coordenadas moleculares y la orientación geométrica de la molécula en el espacio. Por ahora y para nuestros fines, una molécula quedará caracterizada por lo que llamaremos su grafo molecular geométricamente orientado (u orientado por un producto interior).

El grafo de una molécula, o su grafo molecular, es una representación de ella que modela sus propiedades topológicas, principalmente la de conexidad, mientras que sus

coordenadas moleculares almacenan la información geométrica del conjunto de núcleos atómicos, las posiciones relativas en un sistema de coordenadas y su orientación.

La orientación por su parte quedará determinada por las coordenadas moleculares, se incluye puesto que dado un conjunto de átomos, la forma en que se encuentran unidos mediante enlaces covalentes y conociendo las distancias euclidianas que guardan los pares de núcleos atómicos entre si, aún con esta información no hay garantía de una caracterización única, en realidad, para cada grafo molecular acompañado de las distancias entre núcleos atómicos, si la estructura de la molécula no es plana, entonces existen exactamente dos moléculas con esa descripción, moléculas quirales entre si.

Definición A.1.1. Un grafo $\mathcal{G}$ es una terna que consiste en un conjunto no vacío de vértices $\mathcal{V}(\mathcal{G})$, un conjunto de aristas $\mathcal{A}(\mathcal{G})$ y una relación $\mathcal{R}(\mathcal{G})$ que asocia a cada arista una pareja de vértices:

$$\begin{aligned} \mathcal{R}(\mathcal{G}) : \mathcal{A}(\mathcal{G}) &\longrightarrow \{\{\upsilon_1, \upsilon_2\} : \upsilon_1, \upsilon_2 \in \mathcal{V}(\mathcal{G})\} \\ \zeta &\longmapsto R(G)_\zeta = \{\upsilon_1, \upsilon_2\} \end{aligned}$$

Al par de vértices que asociados con una arista se les conoce como los extremos de la misma. Dos vértices se dicen adyacentes si son los extremos de la misma arista, emplearemos la notación $\upsilon_1 \leftrightarrow \upsilon_2$ para indicar que los vértices υ_1 y υ_2 son adyacentes.

Un grafo se dice simple si la relación $\mathcal{R}(\mathcal{G})$ es inyectiva, es un multigrafo si no lo es. Recurriremos a la notación $\upsilon_1 \overset{}{\underset{k}{\leftrightarrow}} \upsilon_2$ para indica que existen k aristas asociadas con la pareja de vértices $\{\upsilon_1, \upsilon_2\}$.

Definición A.1.2. El grafo $\mathcal{G}$ se dice conexo si cada vértice en $\mathcal{V}$ es uno de los extremos de por lo menos una arista en $\mathcal{A}$.

El grafo de una molécula es aquel en que el $\mathcal{V}(\mathcal{G})$ es su conjunto de átomos, $\mathcal{A}(\mathcal{G})$ es el conjunto de pares de electrones que conforman los enlaces covalentes localizados (enlaces σ) y $\mathcal{R}(\mathcal{G})$ asigna hace corresponder a cada par de electrones con la pareja de átomos que los comparte. Es claro que cualquier grafo molecular es un grafo conexo con conjuntos finitos de vértices y aristas.

La orientación molecular por otro lado, es un concepto geométrico que quedará caracterizado por una variable con sólo tres posibles valores. Aceptando que la posición relativa de los núcleos atómicos de una molécula respecto del resto es constante, entonces tales núcleos pueden caracterizarse como parejas de la forma (e, z), donde e es símbolo de cada átomo como elemento químico y z son las coordenadas del núcleo respecto de un sistema de referencia dado (coordenadas moleculares).

Dada una molécula ς con grafo molecular $\mathcal{G}_\varsigma$ para el que existe un subgrafo conexo con vértices $v_1, ..., v_4$ tales que v_1 es adyacente a cada uno de los vértices restantes, entonces la molécula tiene orientación positiva (orientación 1) si es positivo el determinante de la matriz asociada con la transformación lineal que envía a la base canónica en el conjunto ordenado $z_2 - z_1, z_3 - z_1, z_4 - z_1$, con z_i las coordenadas moleculares del átomo que caracteriza al vértice v_i. De forma análoga, si el determinante de la matriz es negativo entonces la molécula se dice con orientación negativa (orientación -1). Se dice que la

estructura de una molécula es plana si existe un plano que contenga propiamente al conjunto de coordenadas moleculares, si existen un vector d y una constante k tales $dz = k$ para cada coordenada molecular z. Cualquier molécula con estructura plana se dirá no orientada o con orientación 0.

La orientación de una molécula definida como antes, atiende a la necesidad de una representación matemática de la molécula que permita discriminar entre pares de moléculas que son *quirales* entre si, pares de moléculas que son indistinguibles salvo por el hecho de que la única forma de sobreponer una en la otra mediante movimientos rígidos elementales (reflexiones o composición de pares de ellas en el espacio) es a través de la composición de una cantidad non de reflexiones. El par de manos de una persona es una clara alegoría pera entender la quiralidad molecular, si se pretende usar un guante izquierdo en la mano derecha no hay forma de hacerlo sin que el dorso del guante se corresponda con la palma de la mano o sin que la parte interior del guante se convierta en la exterior.

La orientación que hemos definido para una molécula depende de las nociones de ángulo en un espacio euclidiano, más generalmente depende del producto interior, aunque sin entrar en detalles, la orientación de una molécula depende de la moderación de sus núcleos atómicos como un sistema con una cantidad finita de elementos en un espacio vectorial con producto interior; siempre que exista una noción de coordenadas moleculares como vectores en un espació de dimensión n y una noción de ángulo, entonces será posible definir la orientación molecular a partir de la existencia de un conjunto de $n + 1$ átomos que induzca un subgrafo conexo.

La orientación será 0 si el conjunto de diferencias entre sus coordenadas moleculares genera un subespacio de dimensión menor que n, mientras que las orientaciones positiva y negativa dependerán del determinante de la matriz de cambio base entre una base canónica y una formada por diferencias de coordenadas moleculares del conjunto de átomos elegido. Se comenta la posibilidad de generalización de la orientación molecular para dejar la puerta abierta a una definición en que las posiciones relativas entre los átomos no sean constantes y estén determinadas por variables aleatorias, o si las moléculas son objetos reales considerados en espacios de dimensión finita mayores o menores que tres.

Para una molécula ς definiremos el *grafo orientado por un producto interior* como la pareja $(\mathcal{G}_\varsigma, \vartheta_\varsigma)$, en la que ϑ es la orientación de ς. Un subgrafo molecular orientado en el mismo sentido será la pareja $(\mathcal{G}_\varsigma^{(B)}, \vartheta_\varsigma^B)$, donde B es un subconjunto propio de $\mathcal{V}_(\mathcal{G}_\varsigma)$, $\mathcal{G}_\varsigma^{(B)}$ es el grafo inducido por B y $\vartheta_\varsigma^{(B)}$ es la orientación del conjunto B como conjunto de núcleos en el espacio con posiciones relativas constantes.

En tanto no se indique sólo estará definida para nosotros un tipo de orientación molecular, y es la definición que hemos presentado, por brevedad nos referiremos a la orientación molecular dada por un producto interior simplemente como orientación molecular, e incluso sólo como orientación cuando por contexto sea inmediata la alusión de conceptos.

Cualesquiera dos grafos moleculares orientados se dirán equivalentes si los grafos en su primera componente son isomorfos entres si y observan la misma orientación en la segunda componente que los define.

Lo que en adelante entenderemos por una estructura base inducida por la molécula ς, será un subgrafo molecular orientado, $(\mathcal{G}_{\varsigma}^{(B)}, \vartheta_{\varsigma}^{B})$, tal que el subgrao molecular inducido por B^c no posee subgrafos conexos no triviales.

Para verificar que un conjunto de moléculas orgánicas es una familia molecular como la que se exige para una análisis QSAR el primer paso es establecer la certeza de que es una *familia determinada por su estructura base*: Que la estructura base inducida por la molécula pivote (el fármaco base del análisis) es una subestructura de cada uno de los compuestos, lo que significa que para cada uno de los compuestos pueda elegirse un subconjunto de átomos, A, de forma que puedan asociarse de forma biunívoca, mediante φ, con los átomos del fármaco base, poniendo en correspondencia átomos del mismo tipo (carbono-carbono, hidrógeno-hidrógeno, etcétera) y respetando las propiedades electrónicas originales, en otros términos, si atm_1 y atm_2 son elementos de A, entonces existirá un enlace químico de orden k entre $\varphi(atm_1)$ y $\varphi(atm_2)$ solamente si existe uno equivalente entre atm_1 y atm_2, preservando la orientación.

Considerando a las moléculas como conjuntos de núcleos atómicos en un espacio euclidiano de dimensión tres, en el interior de regiones con interior no vacío (orbitales), entonces ϕ sería un movimiento rígido que conserva la orientación, que puede expresarse como una cantidad par de reflexiones en el espacio o simplemente con una matriz asociada de determinante positivo respecto de la base canónica.

Un sustituyente es un compuesto relacionado con una familia molecular determinada por sus estructura base, Ξ. Digamos que ς es un miembro de la familia Ξ, con grafo molecular $\mathcal{G}_{\varsigma}$; supongamos ahora que $\mathcal{G}_{\varsigma}^{(A)}$ es el grafo inducido por el conjunto de átomos A de los párrafos que anteceden, mientras que existe un subconjunto de átomos, B, contenido en A^c y tal que el subgrafo inducido por B, $\mathcal{G}_{\varsigma}^{\mathcal{B}}$, es conexo y existe una pareja única de átomos $atm_1 \in A$ y $atm_2 \in B$ para la cual la imagen inversa de $\{atm_1, atm_2\}$ bajo la acción de $\mathcal{R}(\mathcal{G})$ es un conjunto no vacío. Un *sustituyente* será considerado aquí como cualquier compuesto con un conjunto de átomos de la misma cantidad y tipo que un conjunto B como el descrito, con un grafo molecular isomorfo a $\mathcal{G}_{\varsigma}^{\mathcal{B}}$ y la misma orientación que el conjunto B. En este texto reservamos la notación $\mathscr{R}$ para indicar un sustituyente como una molécula o compuesto químico.

En los análisis QSAR son importantes los sustituyentes, nos serán de utilidad ahora para realizar subclasificaciones de familias moleculares determinadas por su estructura base; una familia de este tipo, determinada por la estructura base inducida por la molécula ς, se dirá *determinada por k sustituyentes* si el grafo inducido por B^c es un grafo conformado por exactamente k componentes conexas, siendo cada una de ellas un sustituyente.

Otro requisito indispensable en un análisis QSAR es que dado un blanco biológico, este sea un receptor común de los ligandos en la familia molecular considerada, condición que no es necesaria si la familia sólo se determina por su estructura base; la geometría molecular es importante, tanto la de un ligando como aquella del sitio activo del receptor que por lo general se presume desconocida, si la diferencia entre la cantidad de átomos de la estructura base y la de los ligandos de la familia no tiene restricciones, entonces para una estructura base con pocos átomos pueden añadirse una gran cantidad de átomos que que resulte en una configuración del ligando que aísle a la estructura base, como

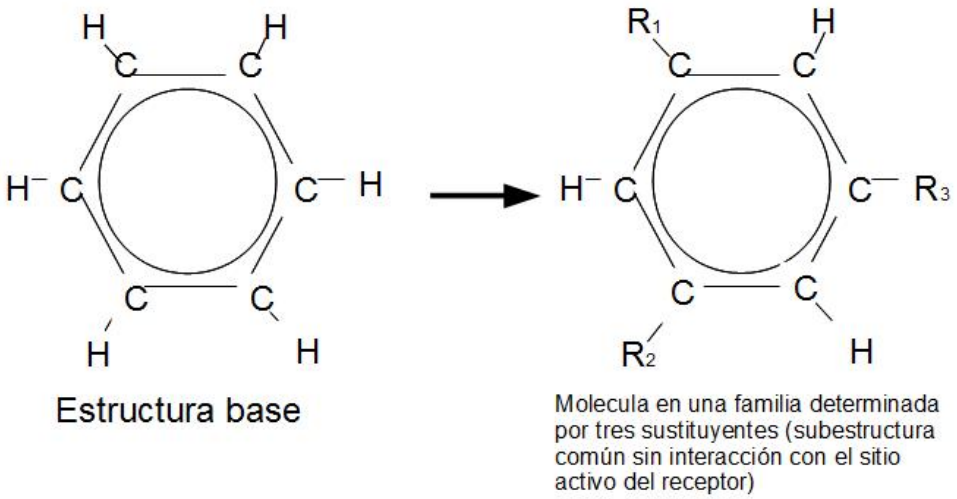

Figura A.1: Ejemplo de compuesto en con estructura base aislada.

subestructura del ligando, imposibilitando toda interacción de esta con el receptor (*figura* A.1.1).

Diremos que una familia de moléculas queda *determinada por el receptor* $\mathfrak{R}$ si $\mathfrak{R}$ es un blanco biológico común a los elementos de la familia.

A.1.2. Familias determinadas por propiedades de su función de actividad

Tradicionalmente, como nos hemos esforzado en recordar, la estructura base almacena implícitamente toda la información del receptor de la que es posible disponer para un análisis QSAR, son los sistemas $(\mathfrak{L}, \mathfrak{R}, \mathfrak{M})$ el único universo conocido; sin embargo, para fortuna de todos, en la practica esta limitada hipótesis no es cierta en general.

Existe información adicional cualitativa y cuantitativa que los análisis QSAR deben incluir de alguna forma en su formulación particular, ejemplos de ello es que los sistemas termodinámicos en los que se realizan las mediciones de la actividad son por mucho más simples que los sistemas biológicos finales en los que un candidato a fármaco será incorporado si en algún momento se desea alcanzar la categoría de medicamento.

La información adicional que en la práctica puede incorporarse en un análisis QSAR puede ser la experiencia de los especialistas de distintas áreas involucrados, lo mínimo que se requiere que un medicamento haga en el sistema final para llegar al receptor deseado, lo que no puede permitirse que un fármaco cause en el sistema final, las relaciones como funcionales no lineales entre los descriptores empleados (no detectables mediante correlación estadística), mediciones adicionales de observables termodinámicos relacionados con la respuesta biológica, incluso la propia formulación del modelo que sea propuesto para explicar las relaciones cuantitativas *estructura-actividad*.

Toda esta información no tiene por qué existir o ser accesible simultáneamente, además del hecho de que la inclusión de cada tipo de ella en el análisis QSAR será por lo general un problema en si mismo.

En lo que resta de este bloque nos concentraremos en continuar la clasificación de familias moleculares para análisis QSAR, pero que en esta ocasión serán determinadas por un tipo particular de información adicional susceptible de ser incorporada en una análisis QSAR a partir de la caracterización de la familia para la cual se ajustaran los parámetros de un modelo general.

Puede ocurrir que no se conozcan todas las particularidades del receptor como la geometría del sitio activo; no obstante, por lo general sabemos que tipo de respuesta esperamos que produzca el ligando al ocurrir la reacción química con el receptor, esperamos reducción de la inflamación de un tejido, mejoría del estado de ánimo o erradicación de una población de parásitos, entre otros. Esto significa para $IC_\alpha(\varsigma)$, en función de α, que esperamos diferencias en el comportamiento cualitativo para moléculas orgánicas con distintos blancos biológicos.

Más importante aún, si se exige un mínimo de rasgos comunes para los elementos del conjunto de funciones

$$\left\{ \begin{array}{rcl} IC.(\varsigma) : [0, \alpha_\varsigma) & \longrightarrow & \mathbb{R}^+ \\ \alpha & \longmapsto & IC_\alpha(\varsigma) \end{array} \middle| \begin{array}{c} \varsigma \in \Xi, \\ \alpha_\varsigma = sup\{\alpha \in [0, 100] : \mathfrak{IC}_\alpha(\varsigma) \neq \emptyset\} \end{array} \right\}.$$

Al elemento $IC.(\varsigma)$ le llamaremos la *función de actividad respecto de la concentración* de la molécula ς en la familia Ξ, la función de actividad de ς de forma breve.

A una familia de antibióticos, por ejemplo, $\alpha_\varsigma = 100$ para toda ς es el primer requisito que de alguna forma debe verificar la familia de ligandos para ser considerada para un análisis QSAR.

Ahora, cuando se establecen un conjunto de propiedades de la función de actividad de una molécula respeto de un receptor determinado, entonces la *familia molecular determinada por el conjunto de características de su actividad* será, el conjunto de moléculas definido por el hecho de que una molecular pertenece al conjunto sólo si su función de actividad cumple con las características establecidas. Nos referiremos a estas familias moleculares simplemente por *familias determinadas por la su función de actividad*. Es claro que cualquier familia molecular determinada por su función de actividad es una familia determinada por el receptor, donde el recíproco no se verifica en general.

Para la familia de ligandos que trabajaremos aquí, se pide que para cada elemento en la familia la actividad, respecto de la concentración, observe un comportamiento monótono creciente y exista una concentración teórica para la que se logre una inhibición al 100 %, es decir, que la función de actividad sea una relación uno a uno y su rango un intervalo compacto.

Trabajaremos con una familia que no se determina solamente por su receptor y estructura base, también por su función de actividad.

A.1.3. Familias determinadas por restricciones sobre los descriptores seleccionados para el análisis

Un compuesto puede ser afín a más de un blanco biológico presente en el sistema final, y la respuesta biológica que producirá al reaccionar con cada uno de ellos será diferente, es posible que un mismo compuesto produzca respuestas antagónicas en el organismo (deseable y no deseable), cualquier medicamento con contraindicaciones o posibles efectos secundarios resaltados por el fabricante es un ejemplo de tal fenómeno.

Aceptando las hipótesis básicas de un análisis QSAR el argumento no es difícil de encontrar, intercambiando los roles entre el receptor y el ligando, en el sistema biológico final todos al conjunto de posibles blancos biológicos de un compuesto dado, se le puede entender como símil de un familia determinada por el receptor, como una familia molecular determinada por el ligando. Ligeras variaciones estructurales en el sitio activo de los receptores no afectaran la afinidad con el ligando dado.

Por otro lado, desde hace por lo menos 20 años se sabe que un buen análisis QSAR debe considerar de algún modo las características del organismo en que en la instancia final se llevará a cabo la reacción de un fármaco, incluso permaneciendo en la categoría de QSAR tradicional (sin más información sobre la constitución del receptor que la proporcionada por la estructura base de la familia).

Además de la unión entre ligando y receptor, las propiedades mensurables (estructurales y fisicoquímicas) tienen una estrecha relación con distintas propiedades farmacocinéticas del compuesto en un organismo, muchas de las cuales son cruciales para la evaluación final de un medicamento. Propiedades como la absorción el transporte la capacidad de unión con componentes sanguíneos, el volumen de distribución, velocidad metabólica y la eliminación se encuentran entre las propiedades que no dependen por lo general del receptor biológico sino del organismo final y el compuesto de interés. Una breve pero sintética y suficientemente detallada aclaración de sobre las relaciones entre propiedades estructurales, fisicoquímicas y farmacocinéticas de los candidatos a fármacos puede revisarse en el trabajo de Seydel J. K. [25].

La experiencia empírica y científica sobre las propiedades fisicoquímicas comentadas antes, sugiere que exciten relaciones de tipo cualitativo y cuantitativo, entre ellas como respecto de las propiedades estructurales y fisicoquímicas de los compuestos. Para un análisis QSAR del que se espera obtener modelos con resultados significativos más allá del sistema simple $(\mathfrak{M}, \mathfrak{L}, \mathfrak{R})$, lo anterior signifique que una vez elegido el conjunto de descriptores que será empleado en el desarrollo del análisis, este conjunto debe someterse a un análisis previo, como parte del preprocesamiento de datos, con la intención de identificar y discriminar, si tiene, con cuáles observa una considerable relación con las propiedades fisicocinéticas comentadas.

Siempre que la revisión previa de los descriptores establezca que las relaciones cuantitativas o cualitativas existen; entonces, serán estas relaciones las que en conjunción establezcan restricciones sobre los descriptores elegidos, adicionales a las restricciones naturales ejemplificadas en los primeros párrafos de la sección .

Para el estudio de caso que realizaremos, para una familia de antibióticos, aparecerán restricciones adicionales a la naturales para los descriptores tal suerte que la familia molecular que los descriptores moleculares de los elementos de la familia deberán ceñirse

a criterios que establecerán un conjunto Ω, en el espacio de $\mathbb{R}^n$, al cuál deberá pertenecer el vector de descriptores para cada elemento de la familia. La intersección de las imágenes inversas de los descriptores moleculares figurará en una de las características en la definición por comprensión de la familia molecular de interés.

La forma en que las restricciones adicionales (relacionadas o no relacionadas con el receptor celular) influyen en un análisis QSAR para la caracterización de la familia molecular que motiva la última categoría de la clasificación que estamos realizando.

Una conjunto de moléculas $\Xi_0 \subset \Xi$ se dirá una familia molecular *determinada por el vector de descriptores,* $x = (x_1, ...x_n)^t$*, y el conjunto de restricciones,* $\Omega \subset \mathbb{R}^n$, si cada x_i es un descriptor bien definido con dominio en Ξ y se verifica

$$\Xi_0 = \bigcap_{i=1}^{n} x_i^{-1}(\Omega).$$

En lo que resta distinguiremos entre una familia química en el sentido usual de la química y una familia molecular QSAR, la segunda será una familia molecular determinada por su receptor común, estructura base, función de actividad y un conjunto de descriptores. Esta distinción en gran medida es de carácter teórico y no siempre puede tenerse evidencia clara de que una conjunto de ligandos sea una familia molecular QSAR. El mismo conjunto de moléculas puede considerarse como dos familias distintas QSAR siempre que se cambie la definición de la función de actividad, el receptor celular de interés o el conjunto de descriptores elegido para caracterizarla.

Lo cierto es que el diseño de la selección de moléculas para una familia QSAR por lo menos observar todas estas condiciones ya que justamente lo que interesa es explicar la actividad mediante un conjunto de descriptores, si la estructura base corresponde a una molécula con geometría plana y la quiralidad puede hacer que una molécula sea afín al receptor y la molécula restante no, entonces los descriptores moleculares por principio deben permitir distinguir entre ellas.

A.2. Preselección de descriptores, relaciones naturales y necesarias para el estudio de caso

A.2.1. Descriptores constitucionales , Grupos 1 y 2

Toda molécula orgánica tiene la característica de estar conformada solamente por combinaciones de átomos de carbono (C), hidrógeno (H), oxígeno (O), nitrógeno (N), azufre (S), bromo (Br), cloro (Cl), flúor (F) o iodo (I); por lo que si nC, nH, nO, nN, nS, nBr, nCl, nF y nI denotan respectivamente las cantidades de átomos de C, H, O, S, Br, Cl, F e I, entonces

nAT=nC+nH+nO+nS+nN+nBr+nI+nF+nCl.

Especialmente en el conjunto muestral Ξ_0 no existen moléculas con átomos de Cl, F o I; la relación de igualdad anterior en consecuencia será para nosotros: nAT=nC+nH+nO+nN+nS+nBr. Las primeras restricciones para los descriptores de la

familia molecular de nuestro análisis QSAR serán: que los descriptores con subíndices en de 1 a 8 toman valores en el conjunto de números naturales y

$$x_8 = \sum_{k=1}^{7} x_k. \tag{A.1}$$

El motivo de la restricción (A.1) es simple, al no disponer de una muestra en que por lo menos una molécula incluya átomos de cloro, flúor o iodo no somos capaces de identificar un cambio en la actividad que pueda relacionarse con la presencia o ausencia de estos elementos químicos en la molécula.

Luego, el descriptor nHAcc es la suma de nO, nN y nF, así que la siguiente relación es por lo tanto

$$x_9 = x_3 + x_4 + x_8. \tag{A.2}$$

La relación entre el volumen de van der Waals y los descriptores anteriores está dada por la propia definición de Sv, si v_i es el el volumen de van der Waals (volumen VdW) de un átomo de carbono, oxigeno, hidrógeno, nitrógeno, azufre y bromo, para $i = 1, ..., 6$ respectivamente, entonces para nuestro caso particular se verifica

$$x_9 = \sum_{k=1}^{7} v_k x_k. \tag{A.3}$$

Los valores v_i son magnitudes experimentales conocidas que dependen del programa de cómputo con el que se realice el cálculo (*tabla* A.2.1).

Símbolo	C	H	N	O	P	S	F	Cl	Br
Vol VdW	1	0.299	0.695	0.512	1.181	1.088	0.41	1	1.384

Tabla A.1: Tabulación de volúmenes de van der Wals. Datos extraídos de [32]

.

A.2.2. Conectividad, Grupo 2

Un enlace covalente entre dos átomos es, *groso modo*, la proximidad espacial constante entre los núcleos atómicos (en un sistema físico estable) que resulta de la existencia de uno o más pares de electrones compartidos por estos átomos. Se dice que un enlace covalente es de orden 1 si sólo un par de electrones es compartido, de orden 2 si se comparten dos pares y de orden 3 si se comparten tres pares.

Puede ocurrir que durante un periodo de tiempo arbitrario un par de átomos de forma constante compartan por lo menos un par de electrones, pese a que la cantidad de los electrones compartidos no sea constante, cambiando entre uno, dos 0 tres pares de electrones comunes. Tal fenómeno se conoce como efecto de resonancia, los enlaces covalentes afectados por él se conocen como enlaces aromáticos y son, por definición, de orden 1.5.

A los enlaces de orden uno se les conoce también como enlaces simples, a los de orden dos como enlaces dobles y los de orden 3 como enlaces triples. Cuando es necesario, para efectos operacionales, un enlace covalente de orden 0 entre dos átomos significará que no existen electrones comunes a ambos.

Para este grupo de descriptores la primera restricción es de tipo lineal; aunque en el sentido modular modular de la teoría de números. Se exige que las moléculas de la familia sean estables, lo que significa que mediante los enlaces covalentes que se forman, cada átomo adquiere una configuración electrónica igual a la de algún gas noble. Además como parte de los requerimientos de estabilidad se exige que para los átomos de oxigeno y nitrógeno ne se observe protonación o ionización, todo esto significa que los enlaces que se formen entre los elementos dependan de los electrones no apareado de los átomos en los orbitales externos y de la hibridación de los orbitales tanto del carbono como del azufre. Motivos que permiten establecer relaciones entre los descriptores de conectividad.

Por el hecho de que los enlace covalentes ocurran por pares de electrones compartidos entre átomos, la suma los electrones no apareado en los orbitales más externos de los átomos debe ser una cantidad par, paridad que dependerá de la suma de dichos electrones sólo para los átomos con valencias impares, por lo que

$$x_2 + 3x_3 + x_6 + x_7 \equiv 0 mod(2). \tag{A.4}$$

Aunado a lo anterior la familia molecular también está determinada por la estructura base, en este caso particular lo que se sustituye en la estructura base para dar paso a los compuestos de la familia es un hidrógeno enlazado a un oxigeno; además, para que la familia este determinada por su estructura hemos dicho que el sustituyente no altera las propiedades electrónicas de la estructura base, significando que la cantidad de átomos y enlaces de cada tipo en la molécula serán siempre mayores o iguales que los ya existentes en la estructura base, excepto tal vez en el caso de los hidrógenos y los enlaces rotables; el resto de los detalle de acotación dependerán del sustituyente pensado como una molécula independiente.

Continuando con las acotaciones para las cantidades de átomos y enlaces, por ser un grafo molecular un grafo conexo y debido a que cada oxigeno puede formar a lo más dos enlace sencillos, en una molécula como de nuestra familia pueden existir a lo sumo dos átomo de oxígeno enlazados a otro de hidrógeno o bromo, de lo contrario cada uno de los oxígenos restantes con esta característica tendría que, en el mejor de los casos, formar un mínimo de dos enlaces adicionales para garantizar la conexidad del grafo molecular, sumando un total de tres pares de electrones para cada uno de estos átomos de oxígeno, hecho que no puede ocurrir por tener el oxígeno valencia 2.

Luego, la cantidad de átomos de hidrógeno, bromo y flúor en el sustituyente está acotada superiormente por la cantidad de átomos de nitrógeno, carbono y azufre, más las posibilidades que ofrecerían dos átomos de oxígeno en los extremos de una estructura molecular sin anillos:

$$(x_{0,2} - x_2) + (x_{0,6} - x_6) + (x_{0,7} - x_7) \leq 2(x_{0,1} - x_1) + (x_{0,3} - x_3) + 3(x_{0,5} - x_5);$$

equivalente a

$$2x_1 - x_2 + x_3 + 3x_5 - x_6 - x_7 - 2x_{0,1} + x_{0,2} - x_{0,3} - 3x_{0,5} + x_{0,6} + x_{0,7} \leq 0. \quad \text{(A.5)}$$

En un grafo el orden de un vértices es la cantidad de aristas que lo tienen como extremo, cada arista tiene exactamente dos vértices por extremos, así que sumando el orden de cada uno de los átomos de una molécula se obtiene el doble de la cantidad de total de aristas; cuando el grafo es un grafo molecular simple, entonce se tiene que el doble de la cantidad de enlaces está acotado inferior y superiormente por la mínima y máxima cantidad de enlaces en que cada uno de los átomos de una molécula tiene que participar.

Para acotar el total de enlace inferiormente notamos que la estructura base establece la primera restricción, luego para el sustituyente, en su grafo cada átomo de hidrógeno y bromo tienen exactamente un enlace, mientras que de los átomos restantes, para garantizar la conexidad del grafo debe ocurrir que cada uno de ellos esté enlazado con al menos un par de los restantes excepto tal vez dos átomos, puesto que un mínimo de enlaces implica una mínima cantidad de anillos, preferentemente ninguno; pero lo átomos de hidrógeno, bromo y flúor deben enlazarse con alguno de los átomos restantes, por lo que al sumar el orden de cada átomo en el caso de suponer un mínimo de enlaces, la suma es equivalente a suponer que cada átomo debe establecer exactamente un mínimo de dos enlaces, excepto tal vez dos átomos que sólo tendrán una enlace cada uno, serán lo vértices correspondientes a los extremo de un grafo que sea una trayectoria, la restricción de acotación inferior es por lo tanto:

$$2x_10 \geq 2[x_{0,10} + \sum_{1 \leq k \leq 7} (x_{0,k} - x_k)] - 2. \quad \text{(A.6)}$$

Para la acotación superior de la cantidad total de enlaces el razonamiento es análogo, en realidad la acotación queda determinada por la valencia de cada elemento químico:

$$2x_10 \leq 2x_{0,10} + \sum_{1 \leq k \leq 7} eV_k(x_{0,k} - x_k); \quad \text{(A.7)}$$

denotando por eV_k al número de electrones no apareados del respectivo elemento en el nivel de energía superior, o las respectivas valencias bajo las consideraciones previas.

En un grafo molecular diremos que un vértices se encuentra mínimamente enlazado si existe una única arista para la que dicho vértice sea un extremo; dicho esto, un enlace rotable es un enlace simple que no se encuentra mínimamente enlazado. Todo átomo de hidrógeno o bromo se identifica con un vértice mínimamente enlazado en el grafo molecular simple, se sigue que ningún enlace con un átomo del tipo H o Br como extremo puede ser rotable. Por otro lado ningún átomo del tipo C, O, N, o S pueden estar mínimamente enlazado y simultáneamente ser el extremo de un enlace rotable, esto es porque para alcanzar una configuración electrónica como la que se desea cada uno de estos átomos debe compartir al meno un par de electrones, por lo tanto, si alguno de estos átomos se encuentra mínimamente enlazado, entonces el único enlace en que se ve involucrado debe ser un enlace múltiple. Se deduce entonces que la cantidad de enlaces rotables está dada por la relación

$$x_{14} = x_{11} - x_2 - x_6 - x_{13}. \tag{A.8}$$

donde $x_{(0,i)}$ es el valor del i-ésimo descriptor evaluado en la estructura base.

A.2.3. Funciones del conjunto de cargas, Grupo 3

Definición A.2.1 (Conjunto molecular de cargas atómicas)**.** Para una molécula orgánica ς conformada por k átomos dotados de un buen orden $\{atom_1, atom_2,..., atom_k\}$ y un procedimiento específico de asignación de cargas atómicas en $\mathbb{R}$, se denomina al conjunto

$$Charge_{\varsigma} = \{charge(\varsigma, atom_i) : i = 1, 2, ..., k\} \subset \mathbb{R}$$

como el conjunto de cargas atómicas de ς generado por el procedimiento de asignación de carga dado; donde $charge(\varsigma, atom_i)$ es la carga que el procedimiento de asignación asocia al átomo $atom_i$.

Observe que el conjunto molecular de cargas atómicas depende del procedimiento de asignación pese a que este hecho no es reflejado por la notación $Charge_{\varsigma}$.

Respetando el orden definido en el conjunto de cargas, este puede entenderse como un vector en un espacio euclidiano cuya dimensión sea la misma que la cantidad de átomos en la molécula; bajo un razonamiento análogo pensaremos particionaremos el conjunto de cargas atómicas en el conjunto de cargas positivas y el conjunto de cargas negativas, para construir los vectores de cargas positivas, q^+, y el de cargas negativas, q^-.

El hecho de que la familia molecular esté compuesta por moléculas estables significa, en términos de los descriptores de carga, que las magnitudes de la carga total positiva y de la carga total negativa son iguales. Denotaremos por $\vec{1}$ al vector horizontal (1,..,1) con dimensión determinada por el contexto en el que sea empleado. De esta forma, los descriptores $Qpos$ y $Qneg$ quedarán definidos como sigue

$$Qpos = \vec{1}q^+, \;\; Qneg = \vec{1}q^-;$$

en tanto que el descriptor Q^2 quedará definido como

$$Q^2 = \left\|q^+\right\|^2 + \left\|q^-\right\|^2;$$

la desigualdad de Cauchy-Schwarz ofrece de inmediato una restricción natural que se complementa con la estabilidad que de la moléculas de la familia en que estamos interesados:

$$x_{18} + x_{19} = 2x_{18} \leq \sqrt{x_8 x_{21}} = \sqrt{x_{21} \sum_{k \leq 6} x_k}. \tag{A.9}$$

De la propia definición del descriptor de carga absoluta media, Qmean, se desprenden las relaciones:

$$x_{20} = \frac{2x_{18}}{x_8} = \frac{2x_{18}}{\sum_{k\leq 6} x_k}; \tag{A.10}$$

$$x_{20} \leq x_{16}; \tag{A.11}$$

$$x_{20} \geq -x_{16}. \tag{A.12}$$

A.2.4. Refractividad Molar y Aproximación de Moriguchi del coeficiente de partición, Grupo 4

El índice de refractividad molar, η, es el cociente entre la velocidad a la que viaja un haz de luz en el aire bajo circunstancias determinadas y la velocidad a la que viaja a través de una sustancia, en este caso una solución del fármaco o compuesto elegido. Si se estima el índice de refractividad de una familia de compuestos con una frecuencia constante para el haz y siempre en un estado determinado el aire, entonces el índice de refractividad es un descriptor.

La refractiviad molar, MR, de un compuesto queda definida por

$$\text{MR}=\frac{\eta^2-1}{(\eta^2+1)\rho}\text{MW};$$

donde MW es el peso molecular y ρ es la densidad, constante, de la sustancia en la que se mide la velocidad del haz. Lo que significa:

$$x_{24} = \frac{x_{25}^2 - 1}{(x_{25}^2+1)\rho}\sum_{k\leq 6} p_k x_k \tag{A.13}$$

con p_k el peso atómico del elemento correspondiente.

En lo tocante al descriptor MLOGP, es una aproximación del coeficiente de partición, es el resultado de un problema de regresión no lineal, fue propuesta por Moriguchi en 1992. Por ahora este descriptor será considerado como un descriptor con perturbación, al igual que los descriptores x_{25} y los descriptores de carga por ser el resultado de un algoritmo recursivo como lo es el método de Gasteiger.

Recuperando las ecuaciones de (A.1) a (A.13) es que se logran una parte de las relaciones entre descriptores moleculares previamente seleccionados para el estudio de caso, presentadas en la sección 2.5.

Bibliografía

[1] Apostol M. Tom, *Análisis matemático*, Reverté, 2da. edición, Barcelona, 1976

[2] Avendaño López María del Carmen, *Introducción a la química farmacéutica*, McGraw-Hill-Interamericana de España, 1a Edición, tercera reimpresión, 1996

[3] Bartle Robert G. , *The elements of integration and Lebesgue measure*, Wiley, New York, 1995

[4] Blanchard Paul, Devaney Robert L. y Hall Glen R., *Ecuaciones diferenciales*, International Thomson Editores, México 1999

[5] Correia Romeiro Nelilma, Girão Albuquerque Magaly y Hopfinger Anton J., *Free-energy force-field three-dimensional quantitative structure-activity relationship analysis of a set of p38-mitogen activated protein kinase inhibitors* en **J Mol Model** 2006, 12, 855-868

[6] Czaplinski, Hänsel W, Wiese M and Seydel JK, *New benxylpyrimifines: inhibition of DHFR from various species, QSAR, CoMFA and PC analysis*, **Eur J Me Chem** 1995, 30:779-787

[7] Faustino Sánchez Garduño, *Matemáticas y Química, una mirada a la cinética química desde la matemática*, S y G Editores, S.A. de C.V., México D.F., 2004

[8] Friedberg Stephen H., Insel Arnold J. and Spence Lawrence E., *Linear algebra*, Prentice Hall of India, New Delhi, 2006

[9] Greenberg, Marvin Jay., *Euclidean and non-Euclidean geometries : development and history*, W.H. Freeman, New York, 1993

[10] Hansen Per Christian, *Discrete inverse problems, Insight and algorithms*, SIAM, Philadelpphia 2010

[11] Herbert B. Callen, *Thermodynamics, and an introduction to thermostatistics*, JHOHN WILEY AN SONS, New York 1985

[12] Hilbert David, *The fundations of Geomety*, tr. Townsend E. J., reprint edition, The open court publishing compsny, Illinois, 1950 http://www.gutenberg.org/files/17384/17384-pdf.pdf

[13] Hogg Robert V. and Craig Allen T., *Introduction to mathematics statistics*, Macmillan Publishing, fourth edition, New York 1978

[14] Hopfinger A. J., Seydel J. K., Lopez de Compadre Rosa L., Koghler M. G. and Emery S., *An extended QSAR aálisis of some 4-Aminoiphenylsulfone antibacterial agents using molecular modeling and LFE-Relationships* **Quant. Struct.-Act. Relat** 1987, 6: 111-117

[15] Ki H. Kim, *Thermodynamic aspects of hydrophobicity and biological QSAR* en **Journal of Computer-Aided Molecular Design**, Kluwer Academic Publishers 2001, 15: 367-380

[16] Kolmogorov A. N., S. V. Fomin, *Elementos de la teoría de funciones y del análisis funcional*, Mir, 2da. edición, Moscu, 1975.

[17] Kyaw Zeyar Myint and Xiang-Qun Xie, *Recent advances in fragment-based QSAR and Multi-Dimensional QSAR methods, review* en **International Journal of Molecular Sciences** 2010, 11: 3846-3866

[18] Livingstone David J., *Data analysis for chemists : applications to QSAR and chemical product design*, Oxford U. Pr., 1995

[19] Martín Alfred N. *Principios de físico-química para farmacia y biología*; tr. P. Sanz Pedrero, Alhambra, 1967

[20] Mc Farlane Mood Alexander, *Introduction to the theory of statistics*, McGraw-Hill, sixth edition, 1974

[21] Michael J. Evans and Jeffrey S. Rosenthal, *Probability and statistics : the science of uncertainty*, W.H. Freeman and Co., New York, 2004

[22] Nelson David L. and Cox Michael M., *Principios de bioquímica*; tr. Claudi M. Cuchillo, Pere Suau León and Josep Vendrell Roca, Ediciones Omega, Barcelona, 2006

[23] Ruelas Barajas Enrique, Mansilla Ricardo y Rosado Javier (Coordinadores), *Las ciencias de la complejidad médica, ensayos y modelos*, Grama Editora, México 2006

[24] Scior T., Medina Franco J.L., et. al., *How to recognize and workaraund pitfalls in QSAR studies: A critical review*, **Current Medicinal Chemestry**, Bentham Science Publishers Ltd. 2009, 16: 4297-4313

[25] Seydel J. K. , *Quatntitative Structure-Pharmacokinetics and their importance in drug design*, **Meth and Find Exptl and Clin Pjarmacol** 1984, 6(10): 571-581

[26] Stieff, M. and Wilensky, U. (2001). NetLogo Enzyme Kinetics model. http://ccl.northwestern.edu/netlogo/models/EnzymeKinetics. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.

[27] Tikhonov A.N., Arsenin V.Y., *Solutions of Ill-Posed Problems*, Winston, New York, 1977

[28] Tikhonov A.N., Goncharsky A.V., *Ill-posed Problems in the Natural Sciences*, Oxford University Press, Oxford, 1987

[29] Tikhonov A.N., Leonov A.S., Yagola A.G., *Nonlinear Ill-Posed Problems*, Chapman and Hall, London, Weinheim, New York, Tokyo, Melbourne, Madras, V. 1-2, 1998

[30] Todeschini Roberto and Consonni Viviana, *Handbook of molecular descriptors*, VILEY-VCH, Weinheim 2000

[31] TokarskiJ. S. y Hopfinger A. J.,*Prediction of Ligand-Receptor Binding Thermodynamics by Free Energy Force Field (FEFF) 3D-QSAR Analysis: Application to a Set of Peptidometic Renin Inhibitors* en **J. Chem. Inf. Comput. Sci.** 1997, 37: 792-811

[32] *Molecular Descriptors Guide, Description of the Molecular Descriptors Appearing in the Toxicity Estimation Software Tool, Version 1.0.2*, 2008 U.S. Environmental Protection Agency

[33] *Quantitative structure-activity relationship (QSAR) models of mutagens and carcinogens*, ed by Romualdo Benigni, CRC Press, Boca Raton, Fla., 2003

[34] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[35] David B. Dahl <dahl@stat.tamu.edu>(2011). xtable: Export tables to LaTeX or HTML. R package version 1.6-0.
http://CRAN.R-project.org/package=xtable

[36] Frank E Harrell Jr <f.harrell@vanderbilt.edu>and with contributions from many other users. (2010). Hmisc: Harrell Miscellaneous. R package version 3.8-3.
http://CRAN.R-project.org/package=Hmisc

[37] Wilensky, U. 1999. NetLogo. http://ccl.northwestern.edu/netlogo/. Center for Connected Learning and Computer-Based Modeling, Northwestern University. Evanston, IL.

Índice alfabético