LA APLICACIÓN DE LA TEORÍA QSAR/QSPR EN LA PREDICCIÓN DE ACTIVIDADES BIOLÓGICAS Y PROPIEDADES FISICOQUÍMICAS.

I- INTRODUCCIÓN Y PROPÒSITOS GENERALES

Rafael Villamayor, Pablo R. Duchowicz, Eduardo A. Castro*

INIFTA, División Química Teòrica, Suc.4, C.C. 16, La Plata 1900, Buenos Aires, Argentina

Resumen: En esta serie de trabajos se propone ofrecer una descripción panorámica acerca de los actuales empleos de la Teoría de las Relaciones Cuantitativas de Estructra Actividad (Propiedad) (QSAR/QSPR) para predecir actividades biológicas y propiedades fisicoquímicas de las sustancias químicas. En esta primera parte se brinda una introducción general al tema, señalando las principales características de esta metodología así como sus campos de aplicación.

Abstract: In this series of articles we give an overview on the present applications of the Quantitative Structure Activity (Property) Relationships (QSAR/QSPR) to predict biological activities and physical chemistry properties of chemical substances. In this first part we offer a general introduction to this theme, pointing out the main features of this methodology as well as its application fields.

* Autor Correspondiente (eacast@gmail.com)

1. La Teoría QSAR/QSPR

El continuo interés por lograr predecir las distintas propiedades fisicoquímicas, biológicas y farmacológicas en sistemas reales conduce indudablemente a la aplicación de métodos derivados de la Mecánica Cuántica, con el fin de representar adecuadamente el fenómeno involucrado. Esto se traduce en la necesidad de tener en cuenta todas las interacciones presentes en el sistema físico de partículas, lo cual hoy por hoy parece ser una tarea harto dificultosa, en vista de que los cálculos mecano-cuànticos actuales sólo pueden resolverse con buena aproximación cuando el sistema involucra unos pocos átomos no-interactuantes. Si bien el uso de aproximaciones matemáticas permite resolver el problema de partículas interrelacionadas entre sí, debido a la incertidumbre de dicho método, no siempre será posible justificar la calidad de los resultados encontrados. Por otro lado, los cálculos mecanocuánticos ayudan a la comprensión de los aspectos mecanísticos que originan a las propiedades en cuestión, pero no resultan la herramienta adecuada para el estudio del efecto que tiene la estructura molecular sobre las propiedades macroscópicas de las sustancias químicas.

La Teoría QSAR/QSPR, (Relaciones Cuantitativas Estructura-Actividad/Estructura-Propiedad) ofrece una alternativa a la hora de calcular las propiedades de una colección de moléculas. Cuando nos detenemos a observar un conjunto de estructuras moleculares junto con sus propiedades experimentales medidas, la pregunta inmediata que surge es ¿existirá una correlación directa entre la propiedad y la estructura de estas sustancias? La respuesta es afirmativa, y es la hipótesis principal de la Teoría QSAR/QSPR. La misma es una hipótesis matemática fundamentada en el hecho de que la estructura de una molécula es la principal responsable de sus propiedades químicas, fisicoquímicas, biológicas o farmacológicas[1-3]. Quizás una de las premisas fundamentales de la teoría es el Principio de Similaridad Estructural, que establece que estructuras moleculares similares poseen propiedades similares, mientras que estructuras moleculares diferentes manifiestan propiedades diferentes[4]. Si bien es conocido desde hace mucho tiempo el hecho de que distintas sustancias tienen diferentes efectos biológicos, el avance en la determinación de estructuras permitió establecer relaciones estructura-actividad (SAR), las cuales evidencian

ciertos efectos en las actividades biológicas a partir del cambio en la estructura química de un determinado compuesto.

Los modelos QSAR también nacen en el campo de la Toxicología. De hecho, los intentos por cuantificar relaciones entre la estructura química y la toxicidad aguda han sido parte de la literatura toxicológica por más de 100 años. Las primeras evidencias se remontan al año 1863, cuando en la defensa de su tesis en la University of Strasbourg, Strasbourg, Francia, J. Cross notó las relaciones existentes entre la toxicidad de alcoholes alifáticos primarios y su solubilidad en agua. Esta relación demuestra el axioma central del modelado de la relación estructura-toxicidad. Por lo tanto, existen interrelaciones entre estructura, propiedades y toxicidad. Casi un siglo después Corwin Hansch et al[5]. publicó su famoso artículo sobre la actividad biológica de grupos de compuestos congéneres y con ello sentó la base para el desarrollo de la actual Teoría QSAR/QSPR.

La Teoría QSAR/QSPR busca cuantificar las relaciones SAR a través del desarrollo de modelos, y combina métodos de la Estadística Matemática con la Química Computacional. Tales modelos se vuelven vitales a la hora de predecir el valor de la propiedad de una sustancia si ésta es desconocida por resultar difícil de adquirir, sea por su inestabilidad quìmica, su toxicidad respecto de la salud humana, su costo económico, etc. Así también la teoría ha sido ampliamente utilizada para el diseño y optimización de compuestos tipo-droga, y hasta fue fructíferamente empleada para inferir resultados sobre los mecanismos de reacción de compuestos orgánicos.

Con el fin de establecer un modelo de cuantificación apropiado, un requisito indispensable es disponer de un conjunto de moléculas para las cuales se conocen perfectamente los valores experimentales de la propiedad estudiada. El diseño de un modelo implica su calibración y su posterior validación. La calibración establece con exactitud la correspondencia entre la estructura y la propiedad analizada a través de la creación del modelo y determinación de los parámetros ajustables de los que depende. La función matemática (lineal o no-lineal) que cuantifica la relación estructura-propiedad se elige de forma arbitraria y la simplificación del modelo matemático dependerá de aquella expresión que determine las mejores predicciones. La validación certifica la veracidad del modelo obtenido, es decir, verifica si posee o no un poder predictivo sobre moléculas no

contempladas en el ajuste del modelo, y que también deben poseer información conocida de la propiedad experimental.

Pero, ¿cómo representar fielmente las relaciones entre la estructura y la propiedad? Desafortunadamente no existe una vinculación directa entre ambas características, por lo cual la teoría se vale de distintos índices numéricos que codifican la información estructural y ayudan a establecer las relaciones buscadas, estos índices son los denominados descriptores moleculares.

2. Los Descriptores Moleculares

Más estrictamente, un descriptor molecular es el resultado final de una lógica y de un procedimiento matemático que transforma la información química codificada dentro de una representación simbólica de una molécula en un número útil o el resultado de algún experimento estandarizado[6]. Estas variables pueden ser teóricas o experimentales, pueden describir a la molécula como un todo (descriptores globales) o sólo pueden representar un fragmento presente en ella (descriptores fragmentos). Generalmente, un gran número de descriptores moleculares surgen de diferentes teorías, tales como la Teoría de Orbitales Moleculares, la Teoría de Grafos y la Mecánica Cuántica, entre otras.

Ahora bien, puede suceder que una combinación apropiada de números describa adecuadamente la propiedad en cuestión, pero que no dejen de ser eso, sólo "simples números". Así, es requisito fundamental que los descriptores posean algún tipo de interpretación química, y si ese no fuera el caso, que sí puedan derivarse en base a la estructura. Un ejemplo clásico de descriptor lo constituye el número de átomos de una especie química en la molécula, como la cantidad de átomos de carbono en una familia de bencenos o él numero de átomos de cloro en especies clorofluorocarbonadas; la cantidad de enlaces C-C puede ser otro ejemplo de descriptor. Otros descriptores relacionados con propiedades fisicoquímicas pueden ser el índice de refracción, las entalpías de vaporización (ΔH_v), el coeficiente de partición octanol/agua (K_{ow}), los puntos de ebullición, los volúmenes molares, etc.

A continuación describiremos los rasgos más relevantes de algunas de las familias de descriptores moleculares más frecuentemente utilizados en la representación de la estructura molecular. No profundizaremos en detalle en cada una de ellas debido a lo amplio y extenso del tema. El lector interesado en este asunto podrá consultar la bibliografía señalada en este análisis.

2.1. Descriptores de la Teoría de Grafos Química

La Teoría de Grafos es una rama de la Matemática Discreta relacionada a la topología y a la combinatoria, y está vinculada con la manera en que los objetos están conectados. Un grafo es una representación bidimensional de la molécula. Estructuralmente, un grafo puede verse como un conjunto de vértices o nodos, unidos por medio de aristas o arcos, en la representación molecular los nodos serían los átomos y las aristas los enlaces. Por ejemplo en el benceno los átomos C son los nodos y los enlaces C-C las aristas.

Los descriptores que se obtienen a partir de la Teoría de Grafos sólo proporcionan información de constitución y conectividad y, por tanto, no pueden discernir isómeros de una misma molécula. Se pueden definir diversos tipos de índices topológicos y entre los más conocidos tenemos a los dos siguientes:

Indice de Wiener (W)[7]

$W = \frac{1}{2}\sum_{ij} d_{ij}$ donde d_{ij} representa la distancia topológica entre los vértices v_i y v_j, si se considera el camino de longitud más corta. La longitud u orden del camino es el número de aristas que lo componen.

Indice de conectividad molecular (χ)[8]

$\chi = \frac{1}{2}\sum_{ij}\left(\deg_i \deg_j\right)^{-0.5}$ donde $\deg_i$ es el grado de degeneración del vértice v_i y representa el número de vértices adyacentes al mismo.

2.2. Indices de la Teoría de la Información

A menudo sucede que gran cantidad de los índices topológicos calculados poseen alto grado de degeneración. El concepto de degeneración de un descriptor molecular se aplica a aquellos descriptores que posean el mismo valor numérico para estructuras diferentes. La Teoría de la Información ofrece una alternativa para disminuir el grado de degeneración de los descriptores topológicos. La aplicación se basa en darle a la molécula representada por un grafo una cierta distribución de probabilidad respecto a la complejidad que posea, y desde allí aplicar la Teoría de la Información.

2.3. Descriptores para Interacciones Químicas

Estos descriptores caracterizan las interacciones químicas que participan en la molécula tanto a nivel global como local, es decir, refiriéndose a un sector de la molécula o tratándola como un todo. Estas interacciones implican cambios topológicos, geométricos y electrónicos, por lo cual los descriptores suelen combinar algunos de estos aspectos.

2.4. Descriptores del programa Dragon[9]

El programa Dragon® ofrece la posibilidad de calcular un gran número de descriptores moleculares agrupados en diferentes familias. A su vez, la lista de descriptores proporcionados puede ser organizada como cerodimensionales (0D), unidimensionales (1D), bidimensionales (2D), y tridimensionales (3D). Nosotros utilizaremos esta última clasificación para simplificar la descripción. Los descriptores calculados en este trabajo son obtenidos con la aplicación de este programa y son cantidades teórico-definidas y podemos destacar que no se utilizan descriptores experimentales.

Descriptores 0D: describen solamente la constitución de la molécula, pero no dicen nada sobre la conformación ni tipo de conectividad presente. Los más simples son el

número de átomos de un determinado tipo, el número de enlaces y el peso molecular, entre otros.

Descriptores 1D: describen fragmentos de las moléculas constituidos por el agrupamiento de sus átomos constituyentes.

Descriptores 2D: utilizan una función de autocorrelación bidimensional que contiene la topología del grafo, y además representa la distribución de una propiedad atómica determinada en la molécula. La propiedad atómica con la que se pesa/pondera al descriptor depende de los átomos presentes en la molécula con lo cual se pueden seleccionar aquellos átomos que dan mayor peso a la variable. Estos descriptores tienen en cuenta las interacciones inter/intra-moleculares.

Descriptores 3D: esta clase tiene en cuenta los aspectos conformacionales de la estructura molecular, considerando de esta manera las propiedades estereoquímicas de las moléculas. Para su cálculo se utilizan estructuras moleculares previamente optimizadas con métodos convenientes, tales como el Método de Campos de Fuerza de la Mecánica Molecular MM^{+}, en combinación con métodos derivados de la Mecánica Cuántica, sean *ab initio* o Métodos de la Teoría de Orbitales Moleculares Semiempírica. Entre estos descriptores citamos las cargas atómicas, la energía del orbital molecular más alto ocupado (ε_{HOMO}) y la energía del orbital molecular más bajo desocupado (ε_{LUMO}), entre otros.

Un descriptor debe cumplir con un conjunto de características tales como:

i. Cálculo sencillo
ii. Invarianza respecto de la traslación y la rotación
iii. Invarianza respecto a la numeración de los átomos
iv. Buena correlación con la propiedad estudiada
v. Bajo grado de correlación con otros descriptores

3. Sobre el Diseño del Modelo

Durante el diseño de los modelos QSAR/QSPR resulta de fundamental importancia seleccionar los descriptores moleculares más influyentes para predecir la propiedad analizada. Existen dos métodos generales para la selección de descriptores moleculares. El primero de ellos consiste en valerse de la experiencia, de las características observables y perceptibles de las moléculas de estudio y del posible mecanismo subyacente. Por ejemplo, la fotohidrólisis es una de las vías principales para la fotólisis de compuestos aromáticos hidrogenados y así varios descriptores químicocuánticos que caracterizan los enlaces C-X fueron calculados y empleados para el desarrollo de modelos QSAR que describan los rendimientos cuánticos de fotólisis de compuestos halogenados[10, 11]. Por otro lado, el segundo método se basa en realizar un estudio combinatorial de los descriptores estructurales y seleccionar aquellos que sean más predictivos.

La ortogonalización de los descriptores moleculares busca facilitar el desarrollo de un modelo óptimo, reduciendo así el número de descriptores objeto de análisis y la dimensión del problema matemático a tratar, por la eliminación de la intercorrelación existente entre dichas variables. Sin embargo, se ha demostrado que la calidad estadística obtenida con el uso de variables no-ortogonales no difiere de la hallada con variables ortogonales.

Las moléculas estándares que constituyen el llamado conjunto de calibración servirán como "moléculas objetivo", pues representan moléculas a las cuales las moléculas de validación deberán imitar, copiar, seguir, aproximarse y lo más deseable, superar en calidad predictiva[12]. Es preciso que las moléculas del conjunto de validación posean estructuras congruentes con las del conjunto de calibración, pues ello influirá directamente en la calidad predictiva del modelo. Una determinada selección de moléculas de calibración y de validación en conjuntos moleculares homogéneos/heterogéneos influenciará considerablemente en los resultados finales que se obtengan con posterioridad con los modelos QSAR/QSPR, y el modelo establecido tendrá algún significado estadístico en la medida que se utilicen conjuntos adecuados.

Finalmente, es esperable que un modelo sencillo que presente error de predicción de la propiedad durante la calibración supere el proceso de validación, en comparación de uno que sea más exacto y sin error de calibración, pues este último se ajusta excesivamente o "memoriza" al conjunto de calibración y de esta manera es incapaz de predecir la propiedad

en cuestión durante la validación. Se busca entonces que los errores cometidos por el modelo en la etapa de calibración sean similares a los encontrados durante la etapa de su validación.

4. Objetivo Específicos

El objetivo principal de la presente serie de artìculos consiste en estudiar diferentes técnicas estadísticas de diseño molecular que permitan el armado de conjuntos moleculares de calibración y validación balanceados, es decir, conjuntos que posean similares errores de predicción de la propiedad. Se busca así seleccionar la metodología que mejor funcione y así implementarla para el trabajo de investigación QSAR/QSPR cotidiano. Para ello, se abordan las técnicas: Análisis de Agrupamiento Jerárquico, Análisis de Componentes Principales, Análisis Discriminante Lineal, Análisis de Agrupamiento k-Medias y k-vecinos más cercanos. La formulación de relaciones estructura-propiedad está basada en la técnica del Análisis de Regresión Lineal, y considera los aspectos multidimensionales de la estructura por medio del análisis de más de mil descriptores moleculares calculados con el programa Dragon. Se compara la bondad de estos métodos clasificadores de objetos sobre tres bases de datos diferentes, a saber: solubilidades acuosas de 166 compuestos orgánicos tipo-droga, 128 actividades anti-SIDA-1 de compuestos orgánicos, y 470 toxicidades acuosas en compuestos alifáticos heterogéneos.

Referencias

1. King, R. B., *"Chemical Applications of Topology and Graph Theory"*. *Studies in Physical*
and Theoretical Chemistry. Elsevier: Amsterdam, 1983.
2. Sexton, W. A., *Chemical Constitution and Biological Activity*. D. Van Nostrand: New York, 1950.
3. Hansch, C., Fujita, T., A quantitative approach to biochemical structure-activity relationships. *Acc. Chem. Res.* **1969**, (2), 232.

4. A. M. Johnson, G. M. M., *Concepts and Applications of Molecular Similarity*. John Willey & Sons: New York, 1990.

5. Hansch, C., Fujita, T., A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc* **1964,** 86, 1616.

6. Roberto Todeschini, V. C., *Handbook of Molecular Descriptors*. WILEY-VCH: Univ. Milano-Bicocca, Italy, 2000.

7. Hosoya, H., A Newly Proposed Quantity Characterizing the Topological Nature of Structural Isomers of Saturated Hydrocarbons. *Bull. Chem. Soc. Jpn* **1971**, (44), 2332.

8. Randic, M., Characterization of molecular branching. *J. Am. Chem. Soc* **1975**, (97), 6609.

9. Dragon 3.0. http://www.disat.unimib.it/chm.

10. Chen J W, Q. X., Schramm K -W, Kettrup A, Yang F L., Quantitative structure-property relationships (QSPRs) on direct photolysis of PCDDs. *Chemosphere* **2000,** 45, (2), 151-159.

11. Free S M, W. J. M., A mathematical contribution to structure-activity studies. *J Med Chem* **1964,** 7, (4), 395－399.

12. Randic, M., Resolution of Ambiguities in Structure-Property Studies by Use of Orthogonal Descriptorst. *J. Chem. If. Comput. Sci* **1991**, (31), 311-320.

13. Duchowicz, P. R., Fernández, F. M., Castro, E. A., Alternative Algorithm for the Search of an Optimal Set of Descriptors in QSAR-QSPR Studies. *MATCH Commun. Math. Comput. Chem.* **2006,** 55, 179-192.