Educación Química

"Relaciones cuantitativas estructura-actividad/propiedad en dos dimensiones empleando el programa R",
Guillermo Goode-Romero, Rodrigo Aguayo-Ortiz y Laura Domínguez
Volumen 30 | Número 2 | Páginas 27-40 | Abril 2019
DOI: 10.22201/fq.18708404e.2019.2.67211

RELACIONES CUANTITATIVAS ESTRUCTURA-ACTIVIDAD/PROPIEDAD EN DOS DIMENSIONES EMPLEANDO EL PROGRAMA R

Resumen

Las relaciones cuantitativas estructura-actividad (*QSAR*) y estructura-propiedad (*QSPR*) son modelos matemáticos aplicados a la predicción de actividades biológicas o propiedades de un grupo de compuestos. Estos modelos son generados y validados por análisis estadístico a partir de un grupo de moléculas con una actividad biológica o propiedad conocida. En este trabajo se explica la metodología general para realizar un análisis *QSAR/QSPR* utilizando el lenguaje de programación de *R*, analizando como caso de estudio la predicción del transporte a través de la barrera hematoencefálica.

Palabras clave: QSAR/QSPR, RStudio, Barrera hematoencefálica

TWO-DIMENSIONAL QUANTITATIVE STRUCTURE-ACTIVITY/PROPERTY RELATIONSHIPS USING R SOFTWARE

Abstract

Quantitative structure-activity/property relationships (*QSAR/QSPR*) are mathematical models applied to the prediction of biological activities or properties of a series of compounds. These models are generated and validated by statistical analysis from a group of molecules with a known biological activity or property. This paper describes the general methodology to perform a *QSAR/QSPR* study using the R software, employing experimental information of the transport through the blood-brain barrier as a case of study.

Keywords: QSAR/QSPR, RStudio, Blood-brain barrier

Autores: Guillermo Goode-Romero[a], Rodrigo Aguayo-Ortiz[a], and Laura Domínguez*[a]

[a] Facultad de Química, Departamento de Fisicoquímica, Universidad Nacional Autónoma de México, Mexico.
*Autor para correspondencia: lauradd@unam.mx

Volumen 30 | Número 2 | Páginas 27-40 | Abril 2019
DOI: 10.22201/fq.18708404e.2019.2.67211

RELACIONES CUANTITATIVAS ESTRUCTURA-ACTIVIDAD/ PROPIEDAD EN DOS DIMENSIONES EMPLEANDO EL PROGRAMA R

Introducción

La idea central de los estudios de relación estructura actividad/propiedad (QSAR/QSPR, por sus siglas en inglés) consiste en relacionar diferencias moleculares con una actividad o propiedad. Estos estudios se basan en la hipótesis de que moléculas similares tienen actividades similares. Estas relaciones, establecen el nexo entre la estructura molecular y su actividad farmacológica de manera cuantitativa (IUPAC, 1997). Asimismo, QSPR asocia la estructura de las moléculas con sus propiedades fisicoquímicas o propiedades no biológicas (Kunal Roy, Kar, & Das, 2015). De tal forma, en los estudios de QSAR/QSPR se busca correlacionar matemáticamente las propiedades estructurales, fisicoquímicas, topológicas, electrónicas y/o geométricas de un conjunto de moléculas utilizando parámetros cuantitativos experimentales o calculados, denominados descriptores (K. Roy & Das, 2014).

En los modelos QSAR/QSPR se establecen las relaciones empíricas estructura-actividad de la manera más simple y predictiva posible, como un modelo lineal de la forma $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_j x_j + \varepsilon$, donde y es la actividad o propiedad de las moléculas; x_j son los descriptores moleculares; β_j es el coeficiente asociado al descriptor y ε es su error asociado (Montgomery CD, 2002). Los descriptores moleculares pueden describir características inherentes a la conectividad, grupos funcionales presentes y/o geometría molecular, por lo que pueden ser bidimensionales (2D) o tridimensionales (3D). La naturaleza dimensional de los descriptores, depende del tipo de algoritmo empleado para calcularlos (K. Roy & Das, 2014).

Un modelo QSAR/QSPR está basado en la hipótesis de que los compuestos evaluados ejercen su actividad biológica mediante un mismo mecanismo, de lo contrario las correlaciones estructura-actividad serán bajas. De esta forma, cuando los mecanismos de acción no se conocen con precisión, usualmente se asume que los miembros de una sucesión de compuestos relacionados estructuralmente comparten el mismo mecanismo (Dearden & Cronin, 2006).

En la actualidad existen una gran variedad de programas y servidores en línea que permiten realizar estudios QSAR/QSPR de forma automatizada; por ejemplo, BuildQSAR (De Oliveira & Gaudio, 2001), Chembench (Capuzzi et al., 2017), BioPPSy (Enciso, Meftahi, Walker, & Smith, 2016), entre otros. Sin embargo, dicha automatización limita a los estudiantes en el proceso de aprendizaje y comprensión de las metodologías empleadas durante el desarrollo de un modelo de QSAR/QSPR. En los planes de estudios de Química de diferentes Universidades se aborda la enseñanza de herramientas estadísticas y matemáticas, en este trabajo se describen las herramientas para que el alumno pueda generar un estudio QSAR/QSPR completo con programas estadísticos de acceso libre; por ejemplo, el programa R (R-Development-Core-Team, 2008) R es un lenguaje de programación y un programa de distribución libre para cómputo estadístico y gráfico, el cual ha sido implementado para los sistemas operativos Windows, MacOS y Linux. El programa provee una amplia variedad de técnicas estadísticas y gráficas, facilitando la manipulación, el cálculo y la visualización de datos.

Volumen 30 | Número 2 | Páginas 27-40 | Abril 2019
DOI: 10.22201/fq.18708404e.2019.2.67211

En este trabajo se establece la metodología general para la construcción de un modelo QSAR/QSPR empleando el programa R. Para el desarrollo metodológico se propuso como caso de estudio la predicción del logaritmo de la constante de reparto encéfalo-sangre *(logBB*) de un grupo de 202 fármacos que actúan a nivel del sistema nervioso central.

Metodología y datos

Programa RStudio

La interfaz gráfica del programa R, llamado RStudio, puede ser instalado en los sistemas operativos Windows, MacOS y Linux (https://www.rstudio.com/products/rstudio/) (Figura 1).

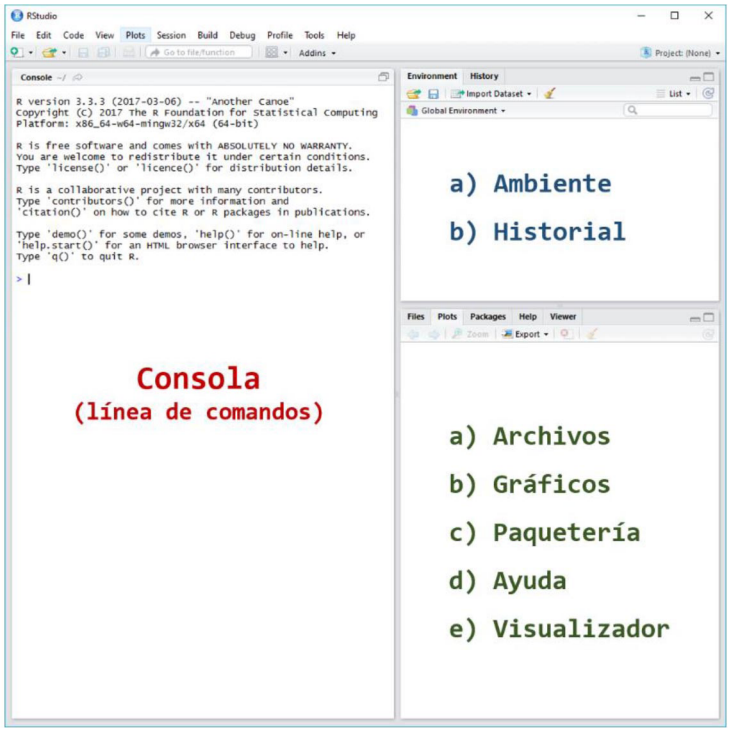

Figura 1. Interfaz gráfica de RStudio.

Paqueterías de R

Las paqueterías o librerías disponibles para análisis de regresión son numerosas, las empleadas en este estudio QSAR son:

i) *rcdk*. Cálculo de descriptores moleculares (Guha, Charlop-Powers, & Schymaski, 2018).

ii) *reshape2*. Manipulación de datos (Wickham, 2017).

iii) *dplyr*. Manejo de conjuntos de datos (Wickham, François, Henry, Müller, & RStudio, 2018).

iv) *car*. Análisis de regresión (John et al., 2018).

v) *cvq2*. Análisis de validación del modelo de regresión (Thaltheim, 2013).

vi) *lattice*. Graficador de relaciones multivariantes (Sarkar, 2017)

Modelo QSAR/QSPR

Las herramientas matemáticas empleadas en QSAR incluyen: análisis de regresión lineal y no lineal, análisis de componentes principales (PCA), redes neuronales artificiales (ANN), análisis comparativo de campos moleculares (CoMFA), y análisis comparativo de índices de similitud molecular (CoMSIA), etc. (Dearden & Cronin, 2006). La metodología de este trabajo se enfoca en el empleo de la regresión lineal múltiple, donde la actividad o propiedad es la variable dependiente, y los descriptores son las variables independientes.

Las instrucciones o comandos empleados para llevar a cabo el estudio QSAR/QSPR en R se especifican en el Anexo 1 y los ocho pasos centrales del diagrama de flujo para la generación del modelo se describen a continuación.

Paso 1. Librerías. Cargar las librerías de R requeridas para el estudio.

Paso 2. Documentos. Cargar la información de actividad o propiedad de las moléculas a analizar en un archivo de texto arreglado en columnas (ver Figura 2A) y el archivo de coordenadas de los compuestos en un formato utilizado en bases de datos, por ejemplo el formato ".sdf" o ".smi" (ver Figura 2B).

A

```
ID,logBB
1,0.03
2,0.03
3,0.03
4,0.63
5,0.68
6,0.44
7,0.69
8,0.52
9,0.67
10,0.97
11,0.86
12,0.98
13,1.05
14,1.01
15,0.9
16,1.04
17,0.11
```

B

```
1
                    3D
 Structure written by MMmdl.
  6  5  0  0  1  0            999 V2000
    1.0220   -0.0710    0.0470 N   0  0  0  0  0  0
    2.4580   -0.0770    0.0710 N   0  0  0  0  0  0
    0.6540    0.5800    0.7470 H   0  0  0  0  0  0
    0.6540   -0.9850    0.3260 H   0  0  0  0  0  0
    2.8260    0.8370   -0.2080 H   0  0  0  0  0  0
    2.8260   -0.7280   -0.6290 H   0  0  0  0  0  0
  1  2  1  0  0  0
  1  3  1  0  0  0
  1  4  1  0  0  0
  2  5  1  0  0  0
  2  6  1  0  0  0
M  END
> <s_m_entry_id>
1
```

Figura 2. Ejemplo del contenido de los archivos (A) En columnas, separadas por comas ".csv" y (B) de coordenadas ".sdf".

Paso 3. Cálculo de descriptores. La librería *rcdk* de R puede calcular 265 descriptores disponibles de cinco categorías distintas: [1] híbridos, [2] constitucionales, [3] topológicos, [4] electrónicos y [5] geométricos. La instrucción permite generar los descriptores deseados, omitiendo aquellos cuyos valores no puedan ser asignados. La información obtenida se almacena en una tabla con la instrucción *write* (ver Nota 1 del Anexo 1).

Paso 4. Descriptores colineales. La colinealidad de los descriptores en un mismo modelo produce una variación excesiva cuando se cambian las *i*-ésimas observaciones. La colinealidad implica una dependencia casi lineal entre los descriptores x_j, e impacta significativamente en la estimación de los coeficientes (Montgomery C.D., Peck, & Vining, 2002). La colinealidad entre descriptores se puede evaluar mediante la matriz triangular simétrica de correlación, la cual es de tamaño j×j y contiene los valores de los coeficientes de correlación de Pearson *r* para cada par de los *j*-ésimos descriptores.

El coeficiente de correlación de Pearson *r* es la covarianza normalizada de dos variables, por lo que toma valores en el intervalo de -1 a 1. Cuando el valor absoluto de *r* tiende a uno, existe una alta correlación lineal entre las variables analizadas (Montgomery, Peck, & Vining, 2002). Cuando $r \rightarrow 0$ se dice que no hay correlación. Uno de los criterios empíricos utilizados para considerar que dos descriptores son colineales es la obtención de un valor de $r > 0.6$ (Maran et al., 2007), aunque estrictamente el valor depende de la naturaleza de los datos. Los descriptores colineales no deben combinarse en el modelo.

Es deseable que los descriptores empleados para generar un modelo QSAR/QSPR no provengan de propiedades relacionadas. Por tanto, se evalúa la colinealidad de los descriptores antes de construir el modelo, y se excluye a los pares de descriptores identificados con un coeficiente de correlación $r > 0.6$.

Paso 5. Grupo de entrenamiento y de prueba. Para el ajuste del modelo QSAR/QSPR es necesario dividir el grupo de moléculas en dos subgrupos, el grupo de estimación o de entrenamiento y el grupo de predicción o de prueba. El grupo de entrenamiento de *n* estructuras permitirá generar el modelo de QSAR/QSPR con los descriptores moleculares y la actividad/propiedad, mientras que el grupo de prueba permitirá establecer si el modelo es suficientemente predictivo. Para este estudio, se seleccionó el 80% de los datos para el grupo de entrenamiento y el 20% restante para el grupo de prueba.

Paso 6. Modelo QSPR. Se seleccionan los descriptores x_j no colineales, con los cuales se genera la regresión lineal múltiple (ecuación 1) mediante el método de cuadrados mínimos para el grupo de entrenamiento.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_j x_j + \epsilon \qquad (1)$$

Donde β_j son los coeficientes de los descriptores y ε es el error asociado a la regresión.

El método de cuadrados mínimos busca que la suma de los cuadrados de las diferencias entre los valores observado y predicho sea mínima. El informe de la regresión lineal múltiple resume los valores de: (1) coeficiente de determinación r^2, el cual indica la variabilidad de la variable dependiente que el modelo puede explicar; (2) el coeficiente de determinación ajustado *adj* r^2, que es la variabilidad de la variable dependiente y que es sensible al número de descriptores; (3) el estadístico *F* de Fisher, que indica la proximidad relativa del análisis de regresión al valor *F* de la función de densidad de Fisher correspondiente; y (4) el valor de probabilidad *p*, el cual es el área bajo la curva de la densidad de Fisher y se interpreta como la probabilidad de que los estadísticos presenten los valores obtenidos, o más extremos (Wasserstein, Lazar, Wasserstein, Lazar, & Asa, 2016). Además, se genera el gráfico de actividad observada vs actividad calculada. El sobreajuste de un modelo de regresión puede aumentar la variabilidad explicada al aumentar r^2, pero pierde considerablemente la capacidad predictiva ante nuevas observaciones, por lo que el estadístico *adj* r^2 es de importancia fundamental.

Paso 7. Valores atípicos. Para obtener un mejor ajuste del modelo, los valores atípicos deben detectarse y eliminarse empleando un método estadístico. La eliminación de valores atípicos para cualquier modelo puede realizarse mediante un análisis de residuales estandarizados o de residuales estudentizados, debiéndose evitar la identificación y selección visual de estos valores, ya que pueden afectar notablemente la significancia del modelo. Una vez descartados los valores atípicos, se calcula un nuevo modelo con las mismas variables iniciales.

Paso 8. Validación. Finalmente, se analiza la capacidad predictiva del modelo generado mediante una validación interna y externa. Se consideran los estadísticos r^2, r_0^2, $r'_0{}^2$, que son los coeficientes de determinación para las variables dependientes observada y calculada; y las pendientes del gráfico de actividad observada contra actividad calculada ajustado al origen y viceversa.

Resultados y discusión

Cuando un fármaco ingresa a un organismo debe ser transportado al sitio de acción para desencadenar su actividad biológica. El transporte *in vivo* involucra el reparto a través de membranas lipídicas, las cuales pueden ser modeladas por las constantes de partición (Dearden & Cronin, 2006). La barrera hematoencefálica es un sistema protector que mantiene el correcto funcionamiento neuronal al restringir el transporte de compuestos neurotóxicos provenientes de la circulación y regulando el ingreso de nutrientes y metabolitos según las demandas celulares (Chakraborty, de Wit, van der Flier, & de Vries, 2016).

Existe una gran variedad de fármacos que actúan a nivel del sistema nervioso central (SNC) denominados neurofármacos y psicofármacos. Los neurofármacos son esenciales para el tratamiento de enfermedades neurodegenerativas (Alzheimer, Parkison, etc.), y los psicofármacos para patologías psiquiátricas (depresión, ansiedad, etc.) (Nageshwaran, Ledingham, & Wilson, 2017). Para el desarrollo de nuevos fármacos de acción central, es necesario evaluar su capacidad de atravesar la barrera hematoencefálica para ejercer su acción. Es por esto que para el presente estudio se evaluó el logaritmo de la constante de reparto encéfalo-sangre (*logBB*), donde se busca generar un modelo predictivo del transporte a través de la barrera hematoencefálica (Chakraborty et al., 2016; Clark, 1999).

Base de datos

Para este caso de estudio se siguieron los ocho pasos presentados en la metodología, empleando los valores experimentales de *logBB* de 202 compuestos recopilados por Abraham y colaboradores (Abraham, Ibrahim, Zhao, & Acree, 2006). Las estructuras químicas de los compuestos se descargaron en formato ".sdf" a partir de los SMILES de la base de datos de PubChem.

Cálculo de descriptores

El programa permitió calcular 278 descriptores para las 202 moléculas. Para la predicción de *logBB* se emplearon como descriptores del modelo: (1) el logaritmo de la constante de reparto 1-octanol-agua calculado (*XLogP*) y (2) el área superficial polar topológica (*TopoPSA*). Análogamente, la ecuación de Clark (Clark, 1999) emplea el *logP* calculado (*clogP*) y el área superficial polar (*PSA*). La sustitución de *PSA* por *TopoPSA* está justificada por su alto coeficiente de correlación (r = 0.99 con n = 34810) (Ertl, Rohde, & Selzer, 2000).

Descriptores colineales

En este caso, los descriptores *XLogP* y *TopoPSA* no presentaron una relación colineal (*r* = 0.10), y son los más significativos en un modelo lineal múltiple.

Modelo QSAR/QSPR

Se generó el modelo QSAR/QSPR con el grupo de entrenamiento, el cual se obtuvo de manera aleatoria. Al escoger los datos de esta manera, es posible que se seleccionen aquellas que disminuyen la correlación y el ajuste en el modelo. Es necesario mencionar que existen algoritmos específicos para la división de datos (Montgomery, Peck, & Geoffrey, 2001), además de la selección empírica.

La instrucción *lm(logBB ~ XLogP + TopoPSA, entrenamiento)* efectuó la regresión lineal múltiple a partir de los datos de 161 moléculas en el grupo de entrenamiento, el 80% del total. En la Figura 3 se muestra el resumen proveniente del análisis de regresión lineal múltiple, mostrando los valores de los coeficientes (*Estimate*), sus errores estándar (*Std. Error*), su significancia (*Pr*), y los estadísticos de la regresión. También se reporta el valor del residual mínimo y máximo y los cuartiles (1Q, mediana o 2Q y 3Q). Estos últimos son útiles para el análisis de normalidad de los residuales.

```
> summary(modelo , correlation=TRUE)

Call:
lm(formula = logBB ~ XLogP + TopoPSA, data = entrenamiento)

Residuals:
     Min       1Q   Median       3Q      Max
-0.92442 -0.23180 -0.04849  0.22742  1.20184

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.1769796  0.0708405   2.498   0.0135 *
XLogP        0.1147821  0.0197577   5.809 3.35e-08 ***
TopoPSA     -0.0096086  0.0008159 -11.777  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3866 on 158 degrees of freedom
Multiple R-squared:  0.5529,    Adjusted R-squared:  0.5472
F-statistic: 97.68 on 2 and 158 DF,  p-value: < 2.2e-16

Correlation of Coefficients:
        (Intercept) XLogP
XLogP   -0.80
TopoPSA -0.53        0.14
```

Figura 3. Resumen de la regresión lineal múltiple.

Grados de libertad

Es conocido que el cociente de dos varianzas s^2_1 y s^2_2 de dos muestras independientes tiene una distribución F de Fisher (Weimer, 2002), con v_1 y v_2 grados de libertad respectivamente. En el análisis de regresión lineal, la variación dentro del grupo de descriptores y la variación entre éste y el error constituyen las varianzas del análisis, y se conocen como raíces de los cuadrados medios.

En nuestro caso, los grados de libertad del análisis de varianza son $v_1 = 2$ (*j*) para la regresión y v = 158 (*n-j-1*) para el error; donde *n* es el número de componentes (compuestos) y *j* el número de variables independientes (descriptores).

Cuadrado medio residual (CMres)

El valor esperado de *CMres* es un estimador insesgado de la varianza del modelo, lo que significa que es un estimado adecuado de ésta. Si el valor de *CMres* es considerablemente mayor al cuadrado medio de la regresión, disminuye la significancia global del modelo.

En nuestro modelo, el valor del cuadrado medio residual fue de *CMres* = 0.3866, que no influyó de manera notable en la significancia.

Significancia de los descriptores

Se realizó una prueba de hipótesis para la regresión que busca examinar la relación entre la actividad o característica biológica y los descriptores. La hipótesis nula H_0 (ecuación 2) asume que los coeficientes no tienen significancia en el modelo.

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_j = 0 \qquad (2)$$

El rechazo de H_0 implica que al menos uno de los *j*-ésimos descriptores contribuye al modelo en forma significativa (Mendenhall & Sincich, 1997a; Montgomery C.D. et al., 2002).

En una prueba de hipótesis, se calcula un estadístico de prueba que sigue una distribución conocida y se contrasta con el valor crítico de esta distribución para el cual el área bajo la curva en su dominio es el nivel de confianza, generalmente 0.95 o 95%. El resto del área bajo la curva se conoce como área de significancia. Un estadístico de prueba que se encuentre en el área de significancia indica que existe una diferencia estadísticamente significativa en el contraste que se está realizando. El área bajo la curva en el intervalo del estadístico de prueba y el límite superior del dominio es el valor *p*. Éste valor indica la probabilidad de obtener un resultado tan extremo como el obtenido considerando que el modelo no es significativo (Wasserstein, 2016).

El cociente del coeficiente del descriptor y su error estándar siguen una distribución t de Student. Para los descriptores del modelo lineal, los valores p fueron $p_{XLogP} = 3.35\times10^{-8}$ y $p_{TopoPSA} < 2\times10^{-16}$, que son menores al área de significancia de 0.05, por lo que son significativos para el modelo. Este hallazgo es de esperarse ya que el descriptor *XLogP* describe el predominio de la solubilidad en medio lipídico o acuoso, siendo liposoluble cuando es mayor a 0.3 (Clark, 1999). El descriptor *TopoPSA* es una medida del área polar de la molécula, y un valor bajo favorece el transporte a través de la barrera hematoencefálica (Ertl et al., 2000).

Estadístico F de Fisher

En este caso de estudio, el valor del estadístico *F* del modelo con 2 y 158 grados de libertad es $F = 97.68$, y el valor $p < 2.2\times10^{-16}$. Lo cual indica en la prueba de hipótesis global de regresión que el modelo es significativo.

Coeficiente de determinación

El coeficiente de determinación r^2 es la proporción de la varianza de la observación explicada por los descriptores (Montgomery et al., 2002). En general, aumenta al incluir un descriptor más al modelo, sin considerar su contribución. El estadístico r^2 ajustado

(*adj* r^2) sólo aumenta cuando el descriptor reduce el valor del *CMres*. Al igual que el estadístico *F* de Fisher, r^2 ajustado evita el sobreajuste del modelo, esto es, la adición de términos que son innecesarios.

Para este modelo lineal, el coeficiente de determinación es $r^2 = 0.5529$ y el coeficiente ajustado es *adj* $r^2 = 0.5472$. Los cuales no concuerdan con los valores esperados dada la significancia de los descriptores en el modelo, por lo que se debe a la dispersión de los valores respecto a la línea recta de ajuste (r = 0.7435).

Al graficar con la opción *xyplot*, se puede observar que la relación de los valores calculados y los valores observados (Figura 4) presentan una tendencia aproximadamente lineal, con múltiples pares ordenados dispersos.

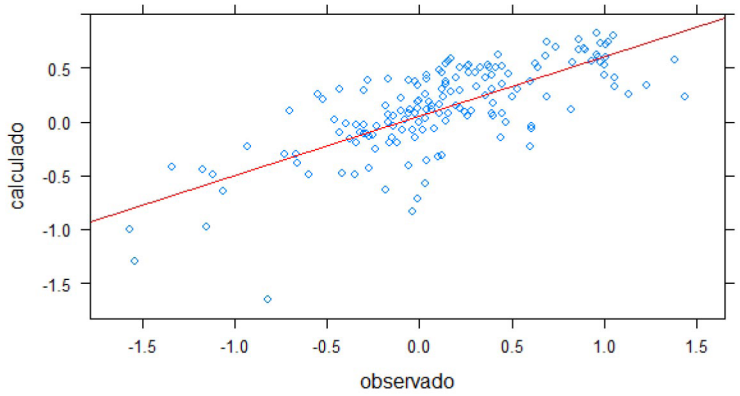

Figura 4. Gráfico de valores calculados contra valores observados.

Se espera que la relación entre las variables sea aproximadamente lineal, y la dispersión excesiva de los valores de los descriptores x_j afecta el modelo QSAR al disminuir el valor de r^2 y aumentar la varianza del *j*-ésimo descriptor.

Valores atípicos

Residuales estandarizados y estudentizados

Los residuales (e_i) son la diferencia de los valores observados y_i con los valores calculados $\hat{y}_i$ para cada observación.

Los residuales estandarizados son el resultado de dividir cada residual por la raíz del cuadrado medio residual. Los residuales estudentizados no consideran esta aproximación y emplean la desviación exacta del *i*-ésimo residual, y pueden tener mayor confiabilidad en la detección de valores atípicos (Montgomery, Peck, & Vining, 2002). Se puede considerar el intervalo de -2 a 2 de los residuales distribuidos normalmente como región de confianza (aproximadamente el 95.4%) para su aceptación en el modelo.

Es importante mencionar que existen otros métodos para la detección de valores atípicos, como las distancias de Cook (Weisberg & Cook, 1985) y los intervalos de confianza (Mendenhall & Sincich, 1997b).

Volumen 30 | Número 2 | Páginas 27-40 | Abril 2019
DOI: 10.22201/fq.18708404e.2019.2.67211

Reconstrucción del modelo de QSPR

Al excluir los valores atípicos de los datos iniciales, se vuelve a crear un modelo lineal con las mismas variables, del cual se espera que aumenten los coeficientes de determinación y el estadístico *F*. Así, una alta determinación y un valor *p* mínimo justifica el modelo propuesto. Al eliminar los valores atípicos, tomando como base la información de los residuales estudentizados, se obtuvo el siguiente modelo:

$$\begin{aligned} logBB = {} & 0.1453(\pm 0.0131)XLogP \\ & - 0.0108(\pm 0.0006)TopoPSA \\ & + 0.0834(\pm 0.0450) \end{aligned} \tag{3}$$

Con $n = 129$, *CMres* = 0.2280, $r = 0.8989$, $r^2 = 0.8048$, *adj* $r^2 = 0.8017$, $F = 259.8$, $p < 2\times10^{-16}$. Los coeficientes β_j siguen siendo significativos para el modelo ($p_{Descriptor} < 2\times10^{-16}$).

El nuevo modelo (3) presenta una mejora significativa con el inicial y con el modelo propuesto por Clark y colaboradores (Clark, 1999), el cual se representó de la siguiente forma:

$$\begin{aligned} logBB = {} & 0.152(\pm 0.036)clogP \\ & - 0.0148(\pm 0.001)PSA \\ & + 0.139(\pm 0.073) \end{aligned} \tag{4}$$

Con $n = 55$, $r = 0.887$, $r^2 = 0.787$, $s = 0.354$, $F = 95.8$

Validación del modelo

La validación permite explicar la capacidad predictiva del modelo, y puede ser interna o externa. La primera emplea la información de los integrantes del modelo para establecer la robustez del mismo, mientras que para la segunda se emplea un grupo de compuestos externos de los cuales se conoce su actividad/propiedad pero no fueron incluidos durante la construcción del modelo, es decir, el grupo de prueba.

Validación interna

La validación interna cruzada consiste en el cálculo del estadístico q^2, que mide en forma aproximada cuánto cabe esperar que el modelo explique la variabilidad de las nuevas observaciones (Montgomery et al. 2002). Para el cálculo del estadístico se puede utilizar el método de "dejar uno fuera" (*LOO*, por sus siglas en inglés), que calcula el valor de la *i*-ésima observación a partir de un modelo generado excluyendo esa misma *i*-ésima observación ($\hat{y}_{i,LOO}$) (Dearden & Cronin, 2006; Kiralj & Ferreira, 2009; Montgomery et al., 2001) y también se utilizan los métodos "dejar un subconjunto fuera" y mediante remuestreos.

El valor de los estadísticos de la validación interna son $q^2 = 0.7988$ y $r^2_{int} = 0.8048$. Se espera que el valor de q^2 sea mayor a 0.5 para considerar una buena predicción (Veerasamy et al., 2011).

Volumen 30 | Número 2 | Páginas 27-40 | Abril 2019
DOI: 10.22201/fq.18708404e.2019.2.67211

Validación externa

Cuando no es posible reunir nuevos datos para fines de validación, un procedimiento razonable es dividir los datos disponibles en dos grupos, los datos de estimación o entrenamiento y los datos de predicción o de prueba. El grupo de entrenamiento se usa para formar el modelo lineal, y el grupo de predicción se usa para analizar la capacidad predictiva del modelo (Montgomery et al., 2001). Si en el grupo de prueba hay presencia de valores atípicos que impacten en los estadísticos de validación externa, se debe realizar el procedimiento de generación del modelo para este grupo, así como detección y eliminación de valores atípicos y generación de un nuevo modelo, de manera análoga a como se realizó en el grupo de entrenamiento.

La validación externa puede ser cruzada o no. Para la validación externa cruzada se calcula el estadístico r^2_{ext} para el grupo de prueba generado. Para la validación externa no cruzada se predice la actividad o característica ($\hat{y}_{i,pred}$) de un grupo de moléculas no considerado inicialmente (Kiralj & Ferreira, 2009).

Coeficientes de determinación

Los coeficientes de determinación del modelo ajustado al origen r_0^2 y $r_0'^2$ indican la correlación de propiedad observada y la propiedad calculada. El estadístico r_0^2 evalúa la determinación de un modelo con la propiedad observada y calculada. Por su parte, el estadístico $r_0'^2$ evalúa la determinación de un modelo con la propiedad calculada y observada (Tropsha & Golbraikh, 2010). Si la recta de ajuste parte del origen (0,0), la asignación de las variables propiedad observada y calculada en el eje de ordenadas o abscisas no debe influir en la interpretación de la validación.

De las rectas de regresión que parten del origen se obtienen los valores de la pendiente *k* de la variable calculada contra la observada, y la pendiente *k'* de la variable observada contra la variable calculada. Los criterios esperados para los estadísticos de determinación y las pendientes (ecuaciones 5 a 8) para un modelo predictivo (Tropsha & Golbraikh, 2010) son:

$$q^2 > 0.5 \quad (5)$$

$$r^2_{ext} > 0.6 \quad (6)$$

$$\frac{(r^2 - r_0^2)}{r^2} < 0.1 \quad (7)$$

$$0.85 \leq m \leq 1.15 \quad \text{ó} \quad 0.85 \leq m' \leq 1.15 \quad (8)$$

En la validación externa del modelo, los estadísticos obtenidos fueron r^2_{ext} = 0.6032, r_0^2 = 0.6022, r'^2_0 = 0.5868, *k* = 0.5848 y *k'* = 1.0491. El cociente (7) es menor a 0.1, y sólo el valor k es menor al límite de 0.85. Estos resultados indican que el modelo tiene capacidad predictiva siguiendo el principio *LOO* y razonablemente para la predicción de un grupo de moléculas externas.

Conclusiones

Los estudios QSAR y QSPR son herramientas de química computacional útiles para el diseño racional de fármacos, y pueden realizarse en su totalidad utilizando el lenguaje R. El modelo de predicción de transporte a través de la barrera hematoencefálica con los descriptores *XLogP* y *TopoPSA* es estadísticamente significativo, y es significativo respecto a las observaciones obtenidas por Clark y colaboradores.

Referencias

Abraham, M. H., Ibrahim, A., Zhao, Y., & Acree, W. (2006). A Data Base for Partition of Volatile Organic Compounds and Drugs From Blood/Plasma/Serum to Brain, and an LFER Analysis of the Data. *Journal of Pharmaceutical Sciences*, *95*(10), 2091–2100.

Capuzzi, S. J., Kim, I. S.-J., Lam, W. I., Thornton, T. E., Muratov, E. N., Pozefsky, D., & Tropsha, A. (2017). Chembench: A Publicly-Accessible, Integrated Cheminformatics Portal. *Journal of Chemical Information and Modeling*, acs.jcim.6b00462. https://doi.org/10.1021/acs.jcim.6b00462

Chakraborty, A., de Wit, N. M., van der Flier, W. M., & de Vries, H. E. (2016). The blood brain barrier in Alzheimer's disease. *Vascular Pharmacology*, *89*, 12–18. https://doi.org/10.1016/j.vph.2016.11.008

Clark, D. E. (1999). Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. *Abstracts of Papers of the American Chemical Society*, *217*(8), U696–U696.

De Oliveira, D. B., & Gaudio, A. C. (2001). BuildQSAR: A New Computer Program for QSAR Analysis. *Quantitative Structure-Activity Relationships*, *19*(6), 599–601.

Dearden, J. C., & Cronin, M. T. D. (2006). Quantitative Structure-Activity Relationships (QSAR) in Drug Design. In H. J. Smith (Ed.), *Smith and Williams' Introduction to the Principles of Drug Design and Action* (Fourth, p. 188,189). Cardiff, Wales: Taylor & Francis.

Enciso, M., Meftahi, N., Walker, M. L., & Smith, B. J. (2016). BioPPSy: An Open-Source Platform for QSAR/QSPR Analysis. *Plos One*, *11*(11), 1–11. https://doi.org/10.1371/journal.pone.0166298

Ertl, P., Rohde, B., & Selzer, P. (2000). Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry*, *43*(20), 3714–3717. https://doi.org/10.1021/jm000942e

Guha, R., Charlop-Powers, Z., & Schymaski, E. (2018). Interface to the "CDK" Libraries.

IUPAC. (1997). *Compendium of Chemical Terminology, 2nd ed. (The "Gold Book").* (A. D. McNaught & A. Wilkinson, Eds.) (XML on-lin). Retrieved from http://goldbook.iupac.org

John, F., Weisberg, S., Price, B., Adler, D., Bates, D., Baud-Bovy, G., … R-Core. (2018). Companion to Applied Regression.

Kiralj, R., & Ferreira, M. M. C. (2009). Basic validation procedures for regression models in QSAR and QSPR studies: Theory and application. *Journal of the Brazilian Chemical Society*, *20*(4), 770–787. https://doi.org/10.1590/S0103-50532009000400021

Maran, U., Sild, S., Mazzatorta, P., Casalegno, M., Benfenati, E., & Romberg, M. (2007). Grid computing for the estimation of toxicity. In *Distributed, High-Performance and Grid Computing in Computational Biology* (p. 66). Eilat: Springer.

Mendenhall, W., & Sincich, T. (1997a). Análisis de regresión lineal múltiple. In *Probabilidad y estadística para ingeniería y ciencias* (Cuarta, p. 603). Prentice-Hall Hispanoamericana, S.A.

Mendenhall, W., & Sincich, T. (1997b). Regresión lineal simple. In *Probabilidad y estadística para ingeniería y ciencias* (Cuarta, pp. 561–563).

Montgomery, D. C., Peck, E. A., & Geoffrey, G. . (2001). Validación de los modelos de regresión. In *Introducción al análisis de regresión lineal* (p. 584). México.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2002a). Comprobación de la adecuación del modelo. In *Introducción al análisis de regresión lineal* (1st ed., pp. 117–122). México.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2002b). Regresión lineal simple. In *Introducción al análisis de regresión lineal* (Primera, p. 13,14). Compañía Editorial Continental.

Montgomery C.D., Peck, E. A., & Vining, G. G. (2002). Regresión lineal múltiple. In *Introducción al análisis de regresión lineal* (Primera, p. 61, 62, 65, 74, 78–80, 82,83). Compañía Editorial Continental.

Nageshwaran, S., Ledingham, D., & Wilson, H. C. (2017). *Drugs in Neurology*. Oxford University Press, 2017.

R-Development-Core-Team. (2008). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Roy, K., & Das, R. N. (2014). A review on principles, theory and practices of 2D-QSAR. *Current Drug Metabolism*, *15*(4), 346–379.

Roy, K., Kar, S., & Das, R. N. (2015). Statistical Methods in QSAR/QSPR. In *A Primer on QSAR/QSPR Modeling* (pp. 37–59). https://doi.org/10.1007/978-3-319-17281-1

Sarkar, D. (2017). Trellis Graphics for R.

Thaltheim, T. (2013). Calculate the predictive squared correlation coefficient.

Tropsha, A., & Golbraikh, A. (2010). Predictive Quantitative Structure-Activity Realtionships Modeling. Developmental and Validation of QSAR Models. In J.-. L. Faulon & A. Bender (Eds.), *Handbook of Chemoinformatics Algorithms* (pp. 214–216). Chapman & Hall Book.

Veerasamy, R., Rajak, H., Jain, A., Sivadasan, S., Varghese, C. P., & Agrawal, R. K. (2011). Validation of QSAR Models - Strategies and Importance. *International Journal of Drug Design and Disocovery*, *2*(3), 511–519. https://doi.org/10.1016/j.febslet.2005.06.031

Wasserstein, R. L., Lazar, N. A., Wasserstein, R. L., Lazar, N. A., & Asa, T. (2016). The ASA â€™ s Statement on p-Values : Context , Process , and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Weisberg, S., & Cook, R. D. (1985). *Applied Linear Regression* (Second Edi). New York: Wiley.

Wickham, H. (2017). Flexibly Reshape Data: A Reboot of the Reshape Package.

Wickham, H., François, R., Henry, L., Müller, K., & RStudio. (2018). A Grammar of Data Manipulation.

ANEXO 1

Paso	Instrucción	Descripción
(1)	library(*nombre de la librería*)	(1) Cargar librerías [ej. *library(rcdk)*]
(2)	moleculas <- load.molecules("Compuestos.sdf") datos <- read.csv("Tabla.csv")	(2a) Cargar moléculas en SDF (2b) Cargar datos en archivo CSV
(3)	dc <- get.desc.categories() tipo_descriptores <- unique(unlist(sapply(get.desc.categories(), get.desc.names))) calcdescriptores <- eval.desc(moleculas, tipo_descriptores) descriptores <- calcdescriptores[, colSums(is.na(calcdescriptores))==0] datos_qsar <- cbind(datos,descriptores)	(3a) Cargar descriptores (3b) Cálculo de descriptores (3c) Datos para estudio QSAR[NOTA 1]
(4)	n_muestra <- floor(0.80 * nrow(datos_qsar)) muestra <- sample(seq_len(nrow(datos_qsar)), size = n_muestra) entrenamiento <- datos_qsar[muestra,] prueba <- datos_qsar[-muestra,]	(4a) Tamaño de muestra (4b) Selección aleatoria (4c) Grupo de entrenamiento[NOTA 1] (4d) Grupo de prueba[NOTA 1]
(5)	matriz_correl <- as.matrix(cor(entrenamiento), method=Pearson) pares_correl <- melt(matriz_correl)[melt(lower.tri(matriz_correl))$value,] colineales <- dplyr::filter(pares_correl, abs(value) > 0.6)	(5a) Matriz de correlación (5b) Pares de descriptores (5c) Descriptores colineales[NOTA 1]
(6)	modelo <- lm(logBB ~ XLogP + TopoPSA, entrenamiento) summary(modelo , correlation=TRUE) calculado <- cbind(c(predict(modelo, entrenamiento))) observado <- cbind(entrenamiento[c("logBB")]) cor(calculado, observado) xyplot(calculado ~ observado, entrenamiento, type=c("p", "r"), col.line = "red")	(6a) Regresión lineal múltiple (6b) Reporte del modelo[NOTA 2] (6c) Coeficiente de correlación (6d) Gráfico calculado *vs* observado[NOTA 3]
(7)	residuales <- resid (modelo) shapiro.test(residuales) res_std <- rstandard (modelo) res_stud <- rstudent (modelo) outliers <- ifelse(abs(res_stud)>((2*sd(res_stud))+(mean(res_stud))), 1, 0) entrenamiento2 <- entrenamiento[!outliers,]	(7a) Residuales (7b) Prueba de normalidad[NOTA 3] (7c) Valores atípicos[NOTA 3] (7d) Filtro con residuales estudentizados
(8)	modelo2 <- lm(logBB ~ XLogP + TopoPSA, entrenamiento2) datos_entrenamiento <- entrenamiento2[,c("**logBB**","XLogP","TopoPSA")] looq2(datos_entrenamiento, logBB ~ XLogP + TopoPSA) datos_prueba <- prueba[,c("logBB","XLogP","TopoPSA")] calculado <- cbind(c(predict(modelo2, datos_prueba))) observado <- cbind(datos_prueba[c("logBB")]) analisis <- cbind(observado,calculado) datos_predict <- analisis[,c("logBB","calculado")] looq2(datos_predict, logBB ~ calculado) looq2(datos_predict, logBB ~ calculado + 0) looq2(datos_predict, calculado ~ logBB + 0) summary(lm(calculado ~ 0 + logBB, analisis))$coefficient[, "Estimate"] summary(lm(logBB ~ 0 + calculado, analisis))$coefficient[, "Estimate"]	(8a) Nuevo modelo (8b) Validación interna (LOO): q^2_{int} (8c) Validación externa[NOTA 1] - Predicción Reportes:[NOTA 2] - r^2 - r_0^2 - $r_0'^2$ - m - m'

Tabla 1. Comandos necesarios de RStudio para el análisis QSAR. Las librerías pueden ser instaladas en la ventana de instrucciones (consola) con el comando: *install.packages("nombre de librería").*

Nota 1. Los datos se pueden guardar con la instrucción: write.csv (*variable*,"*nombre.csv*", quote=FALSE, row.names=FALSE).

Nota 2. Los reportes se pueden guardar con la instrucción: capture.output (*variable*, file = "*nombre.txt*").

Nota 3. Los gráficos se pueden exportar a imagen o archivo .pdf en la ventana de "Gráficos".

Recepción: 17 de septiembre de 2018. Aprobación: 28 de enero de 2019