UNIVERSIDAD: Universidad Nacional de La Plata

NUCLEO DISCIPLINARIO/COMITÉ ACADEMICO/OTROS TEMAS: Matemática Aplicada

TITULO DEL TRABAJO: **DESARROLLO DE MODELOS MATEMÁTICOS PARA PREDECIR NEUROTOXICIDAD Y BIODISPONIBILIDAD DE FÁRMACOS ANTICONVULSIVOS MEDIANTE DESCRIPTORES DERIVADOS DE LA TEORÍA DE GRAFOS.**

AUTOR(ES): Alan Talevi, Carolina Bellera, Eduardo Castro, Luis E. Bruno-Blanch

CORREOS ELECTRÓNICOS DE LOS AUTORES: atalevi@biol.unlp.edu.ar; cbellera@hotmail.com; lbb@biol.unlp.edu.ar

PALABRAS CLAVES: Teoría de grafos - Diseño de fármacos – Neurotoxicidad
Teoría dos grafos – Planejamento dos fármacos - Neurotoxicidade

INTRODUCCIÓN Y OBJETIVOS

El desarrollo de nuevos fármacos es un proceso sumamente costoso. Las diferentes etapas que debe superar un fármaco durante el proceso de desarrollo previamente a su uso en humanos insumen entre 500 y 2000 millones de dólares y un tiempo promedio de 12 años [1,2]. En este escenario, el uso de modelos matemáticos para predecir la actividad y propiedades biológicas de estructuras químicas novedosas se ha convertido en una estrategia habitual y fructífera de la Química Medicinal. La creación de modelos predictivos reduce los costos económicos asociados al desarrollo de una droga. Al maximizar las probabilidades de éxito en los ensayos biológicos, los modelos permiten reducir (aunque no reemplazar) los ensayos en animales y humanos necesarios para identificar estructuras bioactivas, lo cual también implica considerables ventajas desde el punto de vista ético.

La Teoría de Grafos, iniciada por Euler en 1736, utiliza estructuras matemáticas conocidas como grafos para modelar relaciones entre pares de objetos de una colección de objetos. Recientemente, la Teoría de Grafos ha encontrado un gran número de aplicaciones en disciplinas muy variadas, entre ellas las ciencias económicas y sociales, la ingeniería de los materiales, la química, la biología, la lingüística y la criptografía [3-5]. En el caso de la química, la colección de objetos que se desea estudiar es el grupo de átomos que componen una molécula. A partir del grafo que representa una molécula pueden generarse una serie de matrices que, en base a la forma en que se conectan esos átomos, describen distintas características de la estructura molecular (ver sección Metodología); mediante operaciones algebraicas sobre estas matrices se derivan lo que se conocen como índices o descriptores topológicos, que pueden correlacionarse con propiedades químicas o biológicas de interés, como ser la actividad farmacológica de un compuesto químico [6].

Las metodologías QSAR (Quantitive Structure-Activity Relationships) tienen por objetivo identificar, en un grupo de compuestos que manifiestan una actividad o propiedad biológica de interés, patrones estructurales favorables a la actividad, mediante la aplicación de modelos matemáticos que vinculen las características estructurales con el valor experimental de la actividad biológica [7]. Los métodos QSAR involucran, en general, los siguientes pasos: **1)** selección de un conjunto de estructuras químicas (conjunto de entrenamiento) cuyo valor de la propiedad o actividad que se desea predecir ha sido experimentalmente determinado; **2)** cálculo del valor de un conjunto de descriptores (variables discretas o continuas que reflejan alguna característica estructural) para las estructuras del conjunto de entrenamiento; **3)** aplicación de alguna metodología estadística que permita correlacionar el valor de un subconjunto de descriptores (variables independientes incluidas en el modelo) con los valores de actividad o propiedad biológica experimentalmente determinada (variable dependiente); **4)** Validación del modelo generado

para evaluar su capacidad de predicción y robustez y; **5)** aplicación del modelo para predecir la actividad biológica de compuestos cuya actividad biológica se desconoce.

Nuestro grupo de trabajo ya ha aplicado anteriormente estas metodologías, con buenos resultados, en la búsqueda de nuevos fármacos antiepilépticos y antichagásicos, a través del método conocido como Análisis Lineal Discriminante [8-10], seleccionando compuestos promisorios de una base de más de 450.000 estructuras químicas obtenidas de ZINC database [11] y verificando incluso la actividad biológica de algunos de los compuestos seleccionados mediante ensayos en modelos animales [12,13]. Sin embargo, que un compuesto químico reúna las características estructurales esenciales para manifestar la actividad farmacológica de interés es **condición necesaria pero no suficiente** para generar a partir de él un nuevo medicamento. Se desea, idealmente, que el fármaco tenga buena biodisponibilidad[1] y que no presente efectos tóxicos significativos.

En el presente hemos desarrollado mediante Regresión Lineal Múltiple dos modelos matemáticos basados en descriptores topológicos para predecir la Neurotoxicidad y la capacidad de permeación a través de la barrera hematoencefálica[2] de estructuras químicas. Ambos modelos han sido validados interna y externamente y aplicados (a manera de tamices sucesivos) en la selección de aquellos compuestos libres de neurotoxicidad y biodisponibles a nivel del sistema nervioso central de entre 2.649 compuestos clasificados como anticonvulsivos por los modelos desarrollados anteriormente.

1. DESARROLLO

1.1 Materiales y métodos.

Selección de los conjuntos de entrenamiento.

Se desea que los conjuntos de entrenamiento reúnan ciertas características:

- Diversidad estructural: para que la predicción que se haga luego respecto a los valores de la propiedad/actividad sea válida para un conjunto de estructuras pertenecientes a diversas familias de estructuras químicas. Los rangos de valores que asumen las variables independientes del modelo para las estructuras del conjunto de entrenamiento definen lo que se denomina **dominio de aplicabilidad del modelo**. Cuanto más diverso el conjunto de entrenamiento, mayor la diversidad de compuestos cuya actividad podrá ser predicha con exactitud (el modelo identificará patrones estructurales generales y no particulares).
- Homogeneidad en la determinación del valor experimental de la variable dependiente para el conjunto de entrenamiento: si la variable dependiente es una actividad biológica,

[1] El término biodisponibilidad se refiere a que un fármaco, luego de la administración del medicamento, alcance el sitio del organismo en el que ejercerá su acción farmacológica, en cantidades y velocidad adecuadas.

[2] La barrera hematoencefálica es una estructura formada por las células endoteliales de los vasos capilares del cerebro y la médula espinal, destinada a regular, bidireccionalmente, el pasaje de sustancias químicas desde la sangre al cerebro.

los datos del ensayo biológico de los fármacos del conjunto de entrenamiento deben provenir de la misma especie animal e idéntica vía de administración.

- Idealmente, se desea que la propiedad que se quiere modelar y predecir posea una buena distribución a lo largo de entre 3 y 4 órdenes logarítmicos. De esta manera, el modelo puede identificar características/patrones estructurales favorables y desfavorables a la propiedad modelada.

El conjunto de entrenamiento de neurotoxicidad está formado por 30 estructuras con valores conocidos de Dosis Tóxica 50 (DT_{50}) en el ensayo biológico conocido como Rotorod test, utilizado para determinar la capacidad de producir ataxia/falta de coordinación motora en roedores[3]. En este ensayo se observa si un roedor (ratón o rata) es capaz de mantener el equilibrio, durante un minuto, en un cilindro rotatorio que gira a 6 revoluciones por minuto. La incapacidad para mantener el equilibrio indica que el fármaco administrado ha producido ataxia (Figura 1). Se ha utilizado como variable dependiente el log DT_{50} (expresada en µmoles/kg de peso de ratón y determinada en ratón tras administración intraperitoneal). El mismo muestra, para los 30 compuestos, una buena distribución en el rango [1.8-3.6] (Figura 2). No se ha encontrado en literatura datos de DT_{50} en el rotorod test fuera de ese rango.

El conjunto de entrenamiento para la permeabilidad a través de la barrera hematoencefálica está formado por 75 estructuras con logaritmo del coeficiente de reparto cerebro/sangre (log BB[4]) experimentalmente determinado. El log BB se encuentra bien distribuido en el rango [-2.15 – 1.71], es decir, abarca desde compuestos con muy baja permeabilidad a través de la barrera hasta compuestos con permeabilidad muy alta (Figura 3).

Cálculo de descriptores moleculares de los conjuntos de entrenamiento.

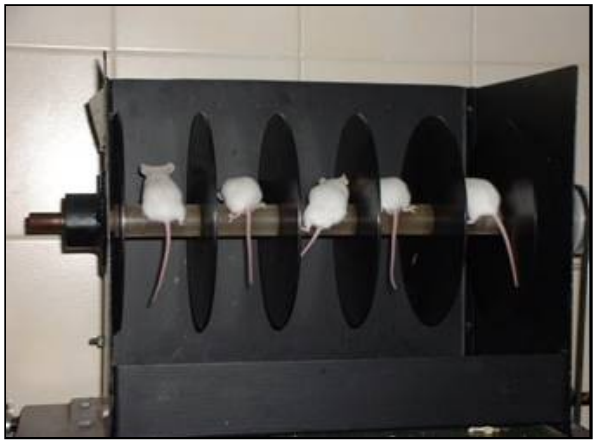
Se calcularon 877 definiciones de descriptores moleculares (variables independientes potenciales de los modelos que codifican información sobre la estructura química de las moléculas). Se decidió utilizar para el modelado descriptores constitucionales (derivados simplemente de la fórmula molecular del compuesto químico) y topológicos. En el grafo que representa las moléculas cada átomo se representa como un vértice del grafo y cada enlace químico se representa como un eje.

A partir del grafo pueden derivarse una serie de matrices. Entre las más comunes encontramos la matriz de conectividad, la matriz distancia y la matriz de distancias

[3] Se define la DT_{50} como la dosis del fármaco que producirá el efecto tóxico (en este caso ataxia) en la mitad de la población.

[4] El log BB es el logaritmo de la relación entre las concentraciones de equilibrio en cerebro y sangre. Un log BB = 0 indica que la concentración de equilibrio en el sistema nervioso central (SNC) es idéntica a la que existe en la sangre; un log BB << 0 indicará que el fármaco no tiene buena llegada al SNC.

topológicas máximas, y otras matrices derivadas de operaciones algebraicas sobre las anteriormente mencionadas.



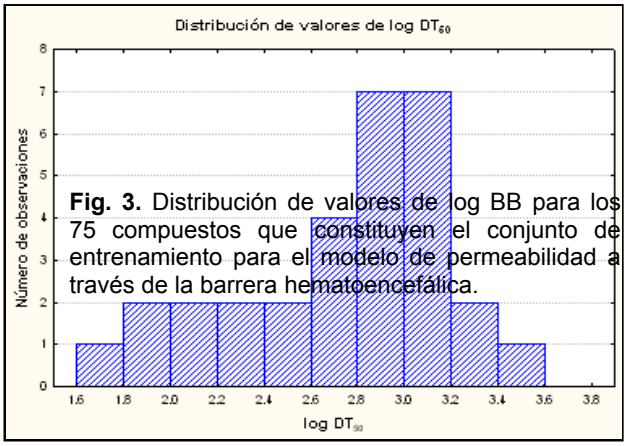
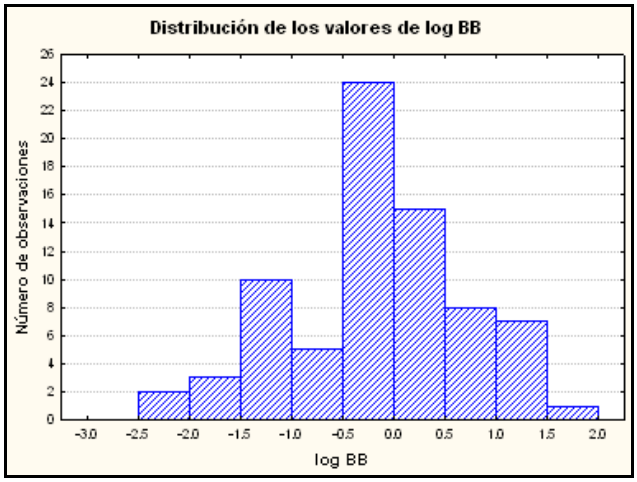


Fig. 3. Distribución de valores de log BB para los 75 compuestos que constituyen el conjunto de entrenamiento para el modelo de permeabilidad a través de la barrera hematoencefálica.

Fig 1 (izquierda). Ensayo de Rotorod. La incapacidad del ratón para mantener el equilibrio durante 1 minuto indica que el fármaco administrado ha producido falta de coordinación motora. **Fig. 2 (derecha)**. Distribución de los valores de log TD_{50} para los 30 compuestos que componen el conjunto de entrenamiento del modelo de neurotoxicidad.



Los elementos C_{ij} de la matriz de conectividad se definen:

$$C_{ij} = \begin{cases} 1 & \text{si los átomos } i \text{ y } j \text{ están directamente unidos por un enlace químico covalente} \\ 0 & \text{si los átomos } i \text{ y } j \text{ no están unidos por un enlace químico covalente} \end{cases}$$

Los elementos D_{ij} de la matriz distancia se definen:

$$D_{ij} = \begin{cases} 0 & \text{si } i = j \\ n_e & \text{si } i \neq j \end{cases}$$

Donde n_e representa el número de ejes, por el camino más corto, que separan los átomos i y j. La matriz de distancias topológicas máximas se define de forma análoga, con la única diferencia de que n_e es ahora el número de ejes que separan a los dos átomos considerados **por el camino más largo**. A partir de estas y otras matrices pueden derivarse índices o descriptores topológicos que describen diversos aspectos de la estructura

molecular. Por ejemplo, la semisuma de los elementos de la matriz de distancias topológicas máximas da lugar al llamado índice Detour, que caracteriza entre otros aspectos el grado de ciclación de un compuesto y ha sido exitosamente utilizado para caracterizar propiedades tales como el punto de ebullición y la solubilidad de estructuras químicas [14]. En la figura 4 pueden verse, a modo de ejemplo, el grafo y las matrices conectividad y distancia derivados de la estructura química del fenol. La numeración de los vértices del grafo es indistinta, ya que los índices topológicos derivados son invariantes del grafo.

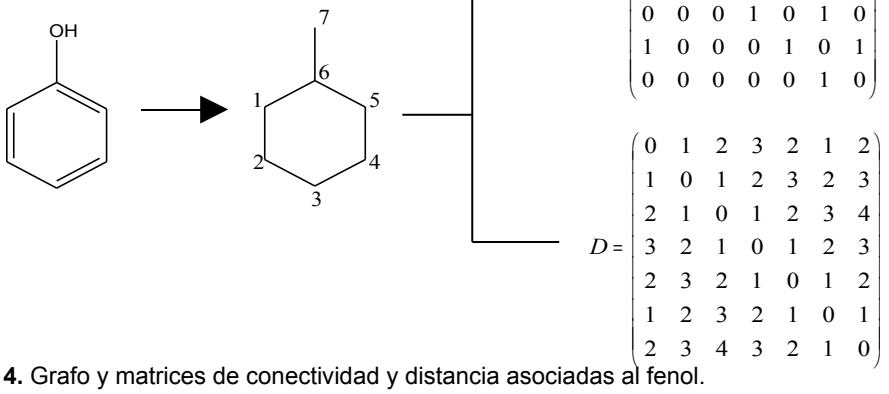


Fig. 4. Grafo y matrices de conectividad y distancia asociadas al fenol.

La ventaja fundamental de los descriptores constitucionales y topológicos es que son independientes de la geometría (disposición tridimensional) de las moléculas analizadas[5].

Generación de los modelos.

A partir de diversos subconjuntos de los 877 descriptores constitucionales y topológicos calculados, se han derivado modelos de neurotoxicidad y permeabilidad mediante regresión lineal múltiple utilizando los programas de estadística Statistica y BMDP [15,16], utilizándose el desempeño en la validación externa como criterio de selección de los mejores modelos. Para evitar la inclusión en los modelos de pares de descriptores redundantes (de manera de evitar sobrepesar la importancia de una característica estructural determinada) se utilizó un valor de coeficiente de tolerancia β de 0.5, siendo que:

$$\beta = 1 - r^2 \quad (1)$$

donde r^2 es el cuadrado del coeficiente de correlación entre cada una de las variables independientes del modelo y el resto de las variables independientes incluidas en el mismo.

[5] Las moléculas pueden adquirir diferentes disposiciones espaciales (conformaciones) asociadas a diferentes valores de energía; la búsqueda de la conformación más estable es un proceso que supone explorar la superficie de energía de las moléculas en busca de un mínimo global. Como aún moléculas poco complejas pueden asumir virtualmente infinitas conformaciones, la búsqueda exhaustiva del mínimo absoluto de energía es inviable y en general no hay garantías de haber arribado a la conformación más establece entre todas las posibles.

Adicionalmente se considera, a fin de evitar la sobreparametrización que la relación $\frac{\text{número de compuestos del conjunto de entrenamiento}}{\text{número de descriptores incluidos en el modelo}}$ debe ser por lo menos igual a 5.

Validación de los modelos.

Ambos modelos han sido validados mediante validación interna (validación cruzada leave-one-out (LOO) y leave-group-out (LGO) y test de aleatorización de Fisher) y validación externa [17]. La validación cruzada LOO y LGO consisten respectivamente en excluir una o varias de las estructuras del conjunto de entrenamiento, regenerando el modelo con las estructuras remanentes y calculando, con el nuevo modelo, los valores de la propiedad para las estructuras excluidas. Aunque ambos métodos tienden a sobreestimar la capacidad predictiva del modelo, son un buen indicador de cómo se comportará el mismo cuando se lo utilice con fines predictivos y para evaluar su robustez. Los resultados de estas técnicas de validación pueden expresarse mediante los parámetros q^2 y $SEE_{LOO/LGO.}$

$$q^2 = 1 - \sum \frac{(y_{obs} - y_{pre})^2}{(y_{obs} - y_{med})^2} \quad (2)$$

$$SEE_{LOO\,/\,LGO} = \sqrt{\frac{\sum (y_{obs} - y_{pre})^2}{n - d - 1}} \quad (3)$$

donde y_{obs} y y_{med} son los valores experimentales y el valor medio de los valores experimentales de la variable dependiente considerada, en ese orden, y y_{pre} representa al valor de la variable dependiente predicho para cada compuesto del conjunto de entrenamiento cuando el compuesto es excluido del conjunto de entrenamiento durante la validación LOO y LGO. *n* es el número de compuestos que conforman el conjunto de entrenamiento y *d* es el conjunto de variables independientes (descriptores) del modelo. Valores de q^2 mayores a 0.5 indican que el modelo es robusto y tendrá buena capacidad predictiva [18]. La validación LGO se llevó a cabo removiendo sistemáticamente grupos de 5 compuestos del conjunto de entrenamiento en el caso del modelo de neurotoxicidad y grupos de 10 compuestos en el caso del modelo de permeabilidad. El ensayo de aleatorización consiste en asignar arbitrariamente los valores experimentales de la variable dependiente a los compuestos del conjunto de entrenamiento y luego regenerar el modelo (repitiendo este procedimiento un número de veces N). Si el modelo no ha sido generado por correlación azarosa entre las variables independientes, se espera que los modelos obtenidos por aleatorización tengan un valor estadístico muy inferior al modelo original.

Finalmente, la validación externa supone verificar que el modelo prediga adecuadamente los valores de la propiedad para un conjunto de prueba independiente (12 compuestos en el caso del modelo de neurotoxicidad y 40 compuestos en el caso del modelo de permeabilidad).

Aplicación de los modelos para filtrar estructuras neurotóxicas y con poca biodisponibilidad central.

Los modelos validados por los procedimientos anteriores se aplicaron para retener, de entre 2.649 estructuras seleccionadas por su potencial anticonvulsivo de una base de 450.000 estructuras químicas, aquellas que según los modelos biodisponibilidad a nivel del SNC (log BB predicho > -0.5) sin presentar neurotoxicidad (DT_{50} predicha > 300 mg/kg) .

1.2 RESULTADOS Y DISCUSIÓN.

Se presentan los dos modelos generados y los valores de los parámetros estadísticos:

$\log DT_{50} = 2.42 - 0.34 \times Hy - 0.80 \times O\text{-}059 - 0.13 \times nCIR + 0.04 \times (D/D)^{1/3} + 0.00136 \times TIE$

N=30 $r^2 = 0.83$ F = 24.53

$q^2_{LOO} = 0.77$ $SEE_{LOO} = 0.24$ $q^2_{LGO} = 0.71$ $SEE_{LGO} = 0.27$ %

logBB = 2.3723 + 0.5041*GATS6v + 0.2506*nX + 0.2072*MLOGP + 0.2009*H-053 - 0.1255*PHI - 1.1460*IC1

N=75 $r^2 = 0.76$ F = 34.97

$q^2_{LOO} = 0.70$ $SEE_{LOO} = 0.49$ $q^2_{LGO} = 0.62$ $SEE_{LGO} = 0.53$

en ambos casos se ha retenido la nomenclatura de descriptores del software Dragon para el cálculo de descriptores moleculares [19], excepto en el caso de $D/D^{1/3}$, un descriptor definido por nuestro grupo de trabajo como la raíz cúbica de la relación entre el índice Detour y el número de átomos dadores de enlaces de hidrógeno [14]. Ambos modelos tienen una muy buena relación $\frac{\text{número de compuestos del conjunto de entrenamiento}}{\text{número de descriptores incluidos en el modelo}}$ (6 y 12.5 respectivamente) y buenos resultados en la validación interna LGO y LOO (q^2 > a 0.5 y desviación estándar en la predicción (SEE) << 1.0, en ambos casos). Esto indica que en general los modelos predecirán con un residual mucho menor a una unidad logarítmica. La validación externa confirmó estos resultados (83% de los compuestos del conjunto de prueba fueron predichos con un residual menor a 0.5 unidades logarítmicas en el caso del modelo de neurotoxicidad y 87.5% fueron predichos con un residual menor a una unidad logarítmica en el caso del modelo de permeabilidad, con un 60% predicho con un residual menor a 0.5 unidades logarítmicas). Las figuras 5 y 6 presentan los gráficos de: valor predicho versus valor observado para las dos propiedades estudiadas, y los valores de r^2 vs q^2_{LOO} para los modelos originales y los modelos generados por aleatorización (30 modelos generados por aleatorización para cada propiedad modelada).

De las 2.649 estructuras a las que se aplicó el modelo se descartaron 841 compuestos predichos como neurotóxicas en dosis anticonvulsivas (DT_{50} < 300 mg/kg) y/o con baja capacidad de atravesar la barrera hematoencefálica (log BB < -0.5).

2. CONCLUSIONES

La asociación de química y estadística para generar modelos multivariable con la capacidad de predecir diversas propiedades químicas y biológicas se ha convertido en una estrategia de suma utilidad en las diversas ramas de la Química. El área de la Química Medicinal, cuyo objetivo fundamental es la identificación de nuevos agentes terapéuticos, se ha beneficiado particularmente de esta asociación interdisciplinaria. El modelado de propiedades biológicas, identificando patrones favorables y desfavorables a una propiedad biológica determinada, es un campo en creciente expansión debido al potencial para reducir los tiempos y recursos necesarios para introducir nuevos fármacos en el mercado y a la tendencia global de reducir tanto como sea posible, por cuestiones bioéticas, el ensayo en animales y humanos. Hemos presentado dos nuevas aplicaciones de estas metodologías, complementando modelos anteriores para identificar un conjunto de nuevos fármacos anticonvulsivos con modelos para seleccionar, de ese conjunto, subconjuntos con los perfiles de biodisponibilidad y toxicidad buscados.

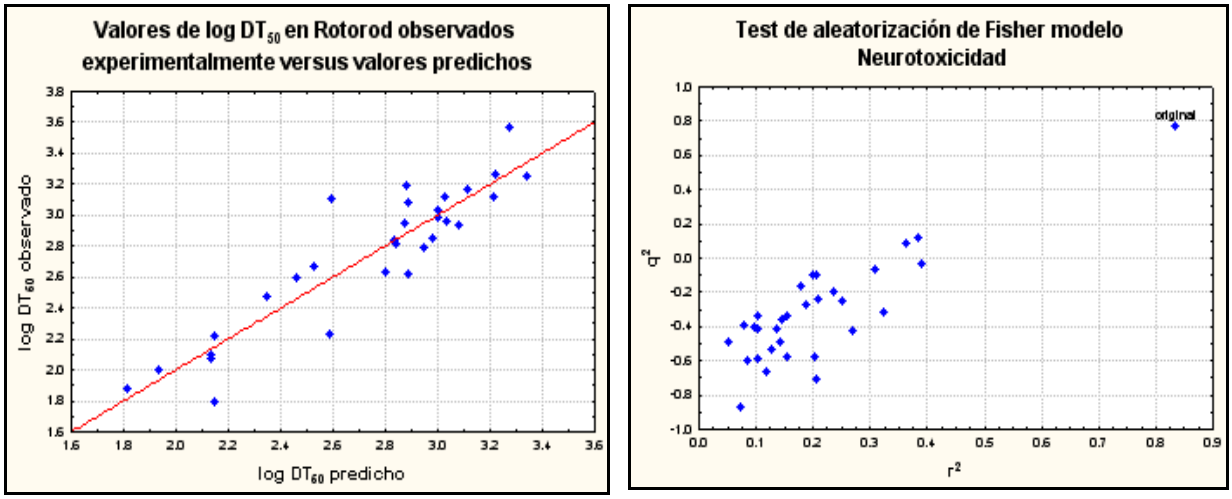


Fig. 5. Valores observados versus predichos de log TD_{50} (ensayo de Rotorod, ratones, administración intraperitoneal) y resultados del estudio de aleatorización de Fisher.

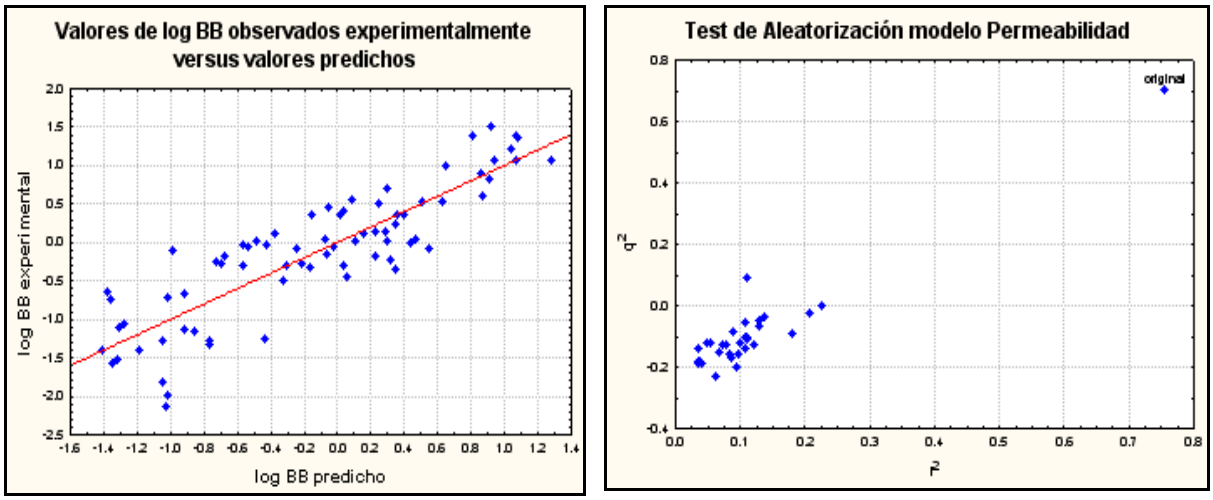


Fig. 6. Valores observados versus predichos de log BB y resultados del estudio de aleatorización de Fisher.

REFERENCIAS

1. DIMASI, Joseph A.; HANSEN, Ronald W.; GRABOWSKY, Henry G. The price of innovation: new stimates of drug development cost. Journal of Health Economics, Holanda, 22, 2, 151-185, Mar 2003.
2. ADAMS, Christopher P.; BRANTNER, Van V. Estimating The Cost Of New Drug Development: Is It Really $802 Million? Health Affairs, 25, 2, 420-428, Estados Unidos, Abr 2006
3. YEN, William C.; WANG, Paul P. Information Sciences - Applications (special issue on Graph Theory and its applications), Estados Unidos, 177, 12, 2403-2620, Jun 2007.
4. KELLER, André A. Graph Theory and Economic Models: from Small to Large Siza Applications. Electronic Notes in Discrete Mathematics, 28, 1, 469-476, Mar 2007.
5. VENKATA RAO, R. A Material Selection Model using Graph Theory and Matrix Approaches. Material Graphics and Engineering: A, 431, 1-2, 248-255, Sep 2006.
6. DEVILLERS, James; BALABAN, Alexandru T (editores). Topological Indices and Related Descriptors in QSAR and QSPR, Estados Unidos, Gordon and Breach Science Publishers, 1999.
7. WINKLER, David A. The role of quantitative structure-activity relationships (QSAR) in biomolecular discovery. Briefings in Bioinformatics, Inglaterra, 3, 1, 73-86, Mar 2002.
8. TALEVI, Alan; BELLERA, Carolina L.; CASTRO, Eduardo A.; BRUNO-BLANCH, Luis E. Application of molecular topology in descriptor-based virtual screening for the discovery of new anticonvulsant agents. Drugs of the Future, España, 31, 188, Ago 2006.
9. BELLERA, Carolina L.; TALEVI, Alan; BRUNO-BLANCH, Luis E. Aplicación de análisis lineal discriminante en la búsqueda de drogas antiepileptogénicas. Latin American Journal of Pharmacy, Argentina, 26, 2, Jun 2007.
10. PRIETO, Julián J.; TALEVI, Alan, BRUNO-BLANCH, Luis E. Application of linear discriminant analysis in the virtual screening of trypanothione reductase inhibitors and redox cycling agents. Molecular Diversity, Holanda, 10, 3, 361-375, Ago 2006
11. IRWIN, John J.; SHOICHET, Brian K. ZINC – A free database of commercially available compounds for virtual screening. Estados Unidos, 45, 1, 177-182, Feb 2005
12. TALEVI, Alan; SELLA-CRAVERO, Mariana; CASTRO, Eduardo A.; BRUNO-BLANCH, Luis E. Discovery of anticonvulsant activity of abietic acid through application of linear discriminant analysis. Bioorganic and Medicinal Chemistry Letters, Estados Unidos, 17, 6, 1684-1690, Mar 2007-06-15
13. TALEVI, Alan; BELLERA, Carolina L.; CASTRO, E.A.; BRUNO-BLANCH, L.E. A successful virtual screening application: prediction of anticonvulsant activity in the MES test of widely used pharmaceutical and food preservatives methylparaben and propylparaben. Enviado.
14. TALEVI, Alan; CASTRO, Eduardo A.; BRUNO-BLANCH, Luis E. New solubility models based on descriptors derived from the detour matriz. Journal of the Argentine Chemical Society, Argentina, 94, 1, 129-141, Jun 2006.
15. Statsoft, Inc. STATISTICA v. 7.0. 2004.
16. Biomedical Computer Programs. BMDP v. 2.0. 2001.
17. YASRI, Aziz, HARTSOUGHT, David. Toward an optimal procedure for variable selection and QSAR model building. Journal of Chemical Information and Computer Science, Estados Unidos, 41, 5, Oct 2001.
18. TROPSHA, Alexander; GOLBRAIKH, Alexander. Beware of q^2. Journal of Molecular Graphics and Modeling, Estados Unidos, 20, 4, 269-276, Ene 2002.
19. Milano Chemometrics. DRAGON v. 4.0. 2003.