Comparison of the performance of an artificial neural network and multiple linear regression in the prediction of the biological activity of cocaine analogues from molecular descriptors

Luis Puerta , Henry Labrador , Mario Arnías *

Departamento de Química, FACYT, Universidad de Carabobo, Apartado 2005, Valencia, Estado Carabobo. Valencia, Venezuela.

https://doi.org/10.54139/revinguc.v29i3.285

Abstract.- The objective of this investigation was to compare the performance of artificial neural networks against multiple linear regression in predicting the biological activity of cocaine analogues from molecular descriptors. For this purpose, a set of 14 molecular descriptors grouped into quantum chemical descriptors and descriptors of the three-dimensional structure of the molecule were selected and their values were calculated theoretically for 65 cocaine analogue structures, followed by the construction of the artificial neural networks model and multiple linear regression for the prediction of biological activity expressed as affinity (IC_{50}). It was found that the artificial neural networks had an R^2 of 0,8651 while the linear multiple regression had an R^2 value of 0,039, showing that artificial neural networks perform better than linear multiple regression in the prediction of the biological activity of cocaine analogues from the selected molecular descriptors, and that the effect of these descriptors on the biological activity is non-linear in nature.

Keywords: biological activity; cocaine; artificial neural networks; multiple linear regression.

Comparación del desempeño de una red neural artificial y regresión lineal múltiple en la predicción de la actividad biológica de análogos de la cocaína a partir de descriptores moleculares

Resumen.- El objetivo de la presente investigación, fue comparar el desempeño de las redes neurales artificiales con la regresión lineal múltiple en la predicción de la actividad biológica de los análogos de la cocaína a partir de descriptores moleculares. Para esto, se seleccionó un conjunto de 14 descriptores moleculares agrupados en descriptores químicos cuánticos y descriptores de la estructura tridimensional de la molécula y se calcularon sus valores de forma teórica, para 65 estructuras análogas de la cocaína, realizándose luego la construcción del modelo de redes neurales artificiales y regresión lineal múltiple, para la predicción de la actividad biológica expresada como afinidad (IC_{50}). Se encontró que las redes neurales artificiales presentaron un R^2 de 0,8651, mientras que la regresión múltiple lineal presentó un valor de R^2 de 0,039, lo que indica que las redes neurales artificiales tienen un mejor desempeño que la regresión múltiple lineal en la predicción de la actividad biológica de los análogos de la cocaína a partir de los descriptores moleculares seleccionados, y que el efecto de los descriptores sobre la actividad biológica es de naturaleza no lineal.

Palabras clave: actividad biológica; cocaína; redes neurales artificiales; regresión lineal múltiple.

Recibido: 31 de octubre, 2022.
Aceptado: 27 de febrero, 2023.

1. Introdución

La cocaína es un alcaloide extraído de las hojas de las plantas del género *Erythroxylum*. Diferentes especies dentro de este género han sido usadas a lo largo de la historia como purgantes, astringentes y estimuladores del sistema nervioso central [1]. La cocaína contiene un esqueleto de 8-metil-8-azabiciclo[3.2.1]octano. Los análogos de cocaína retienen el grupo 3β-Benzoiloxi o grupos funcionales similares y poseen estructuras químicas similares a las de la cocaína [2]. La actividad biológica de los análogos de la cocaína

* Autor para correspondencia:
*Correo-e:*marniasfacyt@gmail.com (M. Arnías)

ha sido determinada experimentalmente como la concentración de una droga o fármaco, que es capaz de inhibir algún proceso biológico o bioquímico en un 50 % (IC_{50}) y es un parámetro de interés debido a que su estudio tiene el potencial de poder encontrar un nuevo compuesto que interactúe en el sitio activo de la cocaína pero con un espectro de actividad alterado o reducido, pudiéndose hallar incluso nuevos fármacos, con la capacidad de revertir o bloquear la actividad de la cocaína a nivel de su receptor [3].

Meyers [4] usaron descriptores moleculares, para caracterizar la estructura tridimensional de una molécula, el momento principal de inercia, y el plano de mejor ajuste, y encontraron que esta estructura debe ser considerada para el diseño de fármacos. Un descriptor molecular es el desenlace de una secuencia matemática y lógica, que transforma los datos químicos almacenados en una representación de la estructura de una molécula en un valor numérico útil o el producto de un ensayo normalizado [5].

Las redes neurales artificiales (RNA), son parte de un conjunto de metodologías de relación cuantitativa estructura-actividad (RCEA), que consiste en un conjunto de modelos computacionales basados en las interconexiones neuronales de los organismos naturales y están inspiradas en los patrones de procesamiento de información de los sistemas nerviosos biológicos. Las RNA poseen funciones de transferencia estímulo-respuesta, que aceptan un conjunto de estímulos de entrada y producen un conjunto de respuestas de salida. En los estudios RCEA, las RNA tienen como ventaja el no requerir de un modelo previo de conexión entre las entradas y las salidas, y poseen la capacidad de adaptarse a relaciones no lineales altamente complejas [6]. En el pasado, se ha intentado predecir la actividad biológica de los análogos de la cocaína, a partir de descriptores moleculares de su estructura tridimensional utilizando RNA, alcanzándose una exactitud aceptable pero con baja precisión [7].

La regresión lineal múltiple (RLM), es una de las técnicas pioneras en el desarrollo de modelos de RCEA y consiste en una versión extendida de la regresión lineal simple. Esta técnica produce una línea, que pasa por el centro del conjunto total de datos, también conocida como línea de tendencia. Esta técnica ha sido exitosa en el modelado de muchas actividades biológicas, excepto aquellas que son no lineales en naturaleza [8]. En el presente trabajo de investigación, se comparó el desempeño de una RNA con el de la RLM en la predicción de la actividad biológica de análogos de cocaína a partir de descriptores moleculares.

2. Metodología

Análogos de la Cocaína

Se seleccionó la cocaína junto con 64 análogos estructurales para un total de 65. La actividad biológica de estos compuestos expresada como IC_{50} o afinidad de enlace, fue tomada del trabajo realizado por Singh [2] y normalizada para obtener valores entre 0 y 1.

Descriptores Moleculares

:

Se seleccionaron 14 descriptores moleculares para los análogos de la cocaína, los cuales se dividieron en descriptores químicos cuánticos (QChD) y descriptores de la estructura tridimensional de la molécula. Los descriptores químicos cuánticos, fueron el Momento dipolar (μ), LUMO, Rotación óptica (Rot α), lipofilicidad molecular (MolLogP) y el área superficial topológica polar (TPSA). Los descriptores de la estructura tridimensional de la molécula fueron el momento principal de inercia en diferentes ejes (PMI1, PMI2, PMI3), las relaciones entre los momentos principales de inercia (NPR1, NPR2), la asfericidad (ΩA), excentricidad (ε), radio de giro (Rg), y el índice de esfericidad (ΩS). Los últimos 4 descriptores relacionados con la geometría de la molécula fueron agrupados colectivamente bajo la denominación 4Geo. Todos los cálculos estructurales fueron realizados en el software Gaussian 09, revisión B.01 y fueron normalizados para obtener valores entre 0 y 1.

Predicción de la actividad biológica mediante Redes Neurales Artificiales

Los descriptores moleculares y sus posibles combinaciones fueron utilizados como entradas

en una RNA para la predicción de la actividad biológica, implementada como un perceptrón multicapa basado en un algoritmo de aprendizaje por retropropagación resiliente, utilizando el software Encog Java 3.4 ©2019 por Heaton Research Inc. y se utilizó la validación cruzada K-fold para evaluar la efectividad de los diferentes conjuntos formados por los descriptores moleculares, obteniéndose los valores reportados por Puerta [7].

Predicción de la actividad biológica mediante Regresión Lineal Múltiple

Los descriptores moleculares μ, LUMO, Rot α, MolLogP, TPSA, PMI1, PMI2, PMI3, NPR1, NPR2, ΩA, ε, y Rg fueron utilizados como variables independientes en la regresión lineal múltiple para predecir la actividad biológica (variable dependiente). Los cálculos de la regresión lineal múltiple y la evaluación del desempeño de la regresión en forma de coeficiente de correlación y R^2 fueron realizados en el software Microsoft Excel 2016.

3. Resultados

Los resultados de la predicción de la actividad biológica de los análogos de la cocaína a partir de descriptores moleculares empleando redes neurales, tomada de Puerta [7] se muestra en la Tabla 1

Los coeficientes de cada variable luego de la regresión lineal múltiple se muestran en la Tabla 2:

Los parámetros estadísticos y de diagnóstico de la regresión lineal múltiple se muestran en la Tabla 3:

4. Discusión de resultados

Al analizar los coeficientes de las variables independientes, que se muestran en la Tabla 2, se puede apreciar que la variable con un mayor valor absoluto, y, por tanto, con mayor incidencia en la actividad biológica, es la relación entre momento principal de inercia NPR1. Las relaciones entre momentos principales de inercia normalizadas son descriptores moleculares independientes del tamaño, de baja complejidad computacional, y permiten la proyección en un diagrama ternario de fácil visualización. Las relaciones entre momentos principales de inercia, permiten asignar la estructura general de una molécula en forma de varilla, disco o esfera [9].

El hecho de que la relación entre momento principal de inercia NPR1 represente la variable con mayor incidencia en el valor de la actividad biológica puede explicarse si se toma en cuenta que la afinidad del análogo de la cocaína hacia el sitio activo de la misma tiene un carácter altamente relacionado con la geometría molecular en general. Por lo general, el sitio activo y el ligando poseen simetría estructural, con los sitios activos siendo un poco más globulares que los ligandos, y la aproximación frecuente en diseño de fármacos es la complementariedad entre el ligando y el sitio activo, por lo que el ajuste geométrico es un prerrequisito para la formación de un enlace. La coincidencia entre la estructura tridimensional del fármaco es un factor que conduce al reconocimiento molecular, y la relación entre momento principal de inercia NPR1, es un factor que está íntimamente relacionado con la estructura tridimensional del análogo de cocaína en general [10].

Por otra parte, el momento dipolar (μ) fue la variable con menor incidencia en el valor de la actividad biológica. Se ha reportado que el μ, aunque puede explicar parte de la actividad biológica, no es un factor exclusivamente responsable de la misma. En casos donde no se encontró una correlación entre μ y actividad biológica, se ha determinado que otros factores, como solubilidad, interacciones celulares, permeabilidad, y acoplamiento molecular, pueden explicar la actividad biológica observada y tienen un mayor efecto sobre la misma. Probablemente, este sea el caso de los análogos de la cocaína, en los que la estructura tridimensional cobra una mayor importancia, que la polaridad general de la molécula [11].

Al comparar el desempeño de la predicción de la actividad biológica mediante RNA y la predicción de la actividad biológica mediante regresión lineal múltiple (MLR), se puede observar que el valor de R^2, que mide la exactitud del modelo, es

Tabla 1: Resumen de los resultados de la validación cruzada K-fold y R^2 de la predicción de la actividad biológica a partir de descriptores moleculares [7]

Descriptor Molecular	Puntuación de la validación cruzada	Error de entrenamiento	Error de validación	R^2
PM1 PM2 PM3	0,0192	0,0193	0,0511	0,7098
QChD	0,0361	0,0202	0,4755	0,3288
QChD PMI1 PMI2 PMI3	0,0245	0,0191	0,0059	0,8528
QChD PMI1 PMI2 PMI3 NPR1 NPR2	0,0421	0,0093	0,0563	0,7875
4Geo	0,0346	0,0087	0,0551	0,7906
QChD 4Geo	0,0344	0,0011	0,0529	0,7974
4Geo PMI1 PMI2 PMI3 NPR1 NPR2	0,0486	0,0059	0,0521	0,805
4Geo PMI1 PMI2 PMI3 NPR2	0,0139	0,0142	0,0504	0,7192
QChD 4Geo PMI1 PMI2 PMI3 NPR1 NPR2	0,0368	0,001	0,0749	0,7247
QChD 4Geo PMI1 PMI2 PMI3	0,0448	0,0012	0,0568	0,7848
QChD ΩA PMI1 PMI2 NPR2	0,0181	0,008	0,0355	0,8651

Tabla 2: Coeficientes de las variables independientes utilizadas en la regresión lineal múltiple para predecir la Actividad Biológica de los análogos de la cocaína

Variable	Coeficientes
Intercepto	3,54
Area superficial topológica polar (TPSA)	-0,56
Lipofilicidad molecular (MolLogP)	-0,30
Asfericidad (ΩA)	-0,92
Excentricidad (ε)	-2,09
Momento Principal de Inercia 1 (PMI1)	0,85
Momento Principal de Inercia 2 (PMI2)	0,07
Momento Principal de Inercia 3 (PMI3)	-0,87
Relación entre Momento Principal de Inercia 1 (NPR1)	-3,32
Relación entre Momento Principal de Inercia 2 (NPR2)	0,18
Radio de Giro (Rg)	0,28
Momento Dipolar (μ)	-0,04
LUMO	-0,24
Rotación Óptica (Rot α)	0,32

Tabla 3: Estadísticas de la regresión lineal múltiple para predecir la actividad biológica de la cocaína a partir de descriptores moleculares

Parámetro	Valor
Coeficiente de correlación	0,484
Coeficiente de determinación	0,234
R^2 ajustado	0,039
Error típico	0,184
Observaciones	65

mayor para las RNA que para la MLR. Esto ocurre para cualquier combinación de descriptores moleculares, lo que indica que las RNA tienen un mejor desempeño en la predicción de la actividad biológica de los análogos de la cocaína que la MLR. Esto implica que la actividad biológica de los análogos de la cocaína depende de los descriptores moleculares seleccionados de una forma no lineal, la cual es modelada de forma más exacta a partir de métodos no lineales como las RNA, las cuales tienen alta capacidad de predicción, son confiables y robustas [12, 13].

Para investigaciones futuras, se recomienda estudiar modelos no lineales alternativos para la predicción de la actividad biológica de los análogos de la cocaína, como máquinas de vectores de soporte, redes neurales de k vecinos más cercanos, bosque aleatorio, redes neurales convolucionales o redes neurales profundas [14].

5. Conclusiones

Las redes neurales artificiales mostraron un mejor desempeño que la regresión lineal múltiple en la predicción de la actividad biológica de los análogos de la cocaína a partir de descriptores moleculares químicos cuánticos y de estructura tridimensional.

La relación entre momento principal de inercia NPR1 y el momento dipolar fueron las variables con mayor y menor efecto sobre la actividad biológica de los análogos de la cocaína, respectivamente.

El efecto de los descriptores moleculares seleccionados sobre la actividad biológica de los análogos de la cocaína es predominantemente no lineal.

6. Referencias

[1] S. Bieri, A. Brachet, J.-L. Veuthey, and P. Christen, "Cocaine distribution in wild Erythroxylum species," *Journal of ethnopharmacology* , vol. 103, no. 3, pp. 439–447, 2006. https://doi.org/10.1016/j.jep.2005.08.021

[2] S. Singh, "Chemistry, design, and structure-activity relationship of cocaine antagonists," *Chemical Reviews*, vol. 100, no. 3, pp. 925–1024, 2000. https://doi.org/10.1021/cr9700538

[3] E. S. Lazer, G. J. Hite, K. A. Nieforth, and E. S. Stratford, "Synthesis and biological activity of cocaine analogs. 2. 6H-[2] Benzopyrano [4, 3-c] pyridin-6-ones," *Journal of Medicinal Chemistry* , vol. 22, no. 7, pp. 845–849, 1979. https://doi.org/10.1021/jm00193a018

[4] J. Meyers, M. Carter, N. Y. Mok, and N. Brown, "On the origins of three-dimensionality in drug-like molecules," *Future medicinal chemistry*, vol. 8, no. 14, pp. 1753–1767, 2016. https://doi.org/10.4155/fmc-2016-0095

[5] R. Todeschini and V. Consonni, *Handbook of molecular descriptors*. John Wiley & Sons, 2008.

[6] J. C. Gertrudes, V. G. Maltarollo, R. A. Silva, P. R. Oliveira, K. M. Honorio, and A. B. F. da Silva, "Machine learning techniques and drug design," *Current medicinal chemistry* , vol. 19, no. 25, pp. 4289–4297, 2012. https://doi.org/10.2174/092986712802884259

[7] L. Puerta and C. Gonzalez, "Molecular descriptor to predict biological activity of analogues cocaine," Experiment findings, 2020.

[8] C. Nantasenamat, I.-N.-A. Chartchalerm, and P. Virapong, "Advances in computational methods to predict the biological activity of compounds," *Expert opinion on drug discovery*, vol. 5, no. 7, pp. 633–654, 2010. https://doi.org/10.1517/17460441.2010.492827

[9] S. Wolfgang H. B. and S. Matthias K., "Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity," *Journal of chemical information and computer sciences*, vol. 43, no. 3, pp. 987–1003, 2003. https://doi.org/10.1021/ci025599w

[10] M. Wirth, A. Volkamer, V. Zoete, F. Rippmann, O. Michielin, M. Rarey, and W. H. B. Sauer, "Protein pocket and ligand shape comparison and its application in virtual screening," *Journal of computer-aided molecular design*, vol. 27, no. 6, pp. 511–524, 2013. https://doi.org/10.1007/s10822-013-9659-1

[11] A. Das and B. K. Banik, "26 - Dipole moment in medicinal research: Green and sustainable approach," in *Green Approaches in Medicinal Chemistry for Sustainable Drug Design*, ser. Advances in Green and Sustainable Chemistry, B. K. Banik, Ed. Elsevier, 2020. https://doi.org/10.1016/B978-0-12-817592-7.00021-6

[12] A. Abdolmaleki and J. B. Ghasemi, "Inhibition activity prediction for a dataset of candidates' drug by combining fuzzy logic with MLR/ANN QSAR models," *Chemical Biology & Drug Design*, vol. 93, pp. 1139–1157, 2019. https://doi.org/10.1111/cbdd.13511

[13] P. Žuvela, J. David, X. Yang, D. Huang, and M. W. Wong, "Non-linear quantitative structure–activity relationships modelling, mechanistic study and in-silico design of flavonoids as potent antioxidants," *International journal of molecular sciences*, vol. 20, no. 9, 2019. https://doi.org/10.3390/ijms20092328

[14] X. Lin, X. Li, and X. Lin, "A review on applications of computational methods in drug screening and design," *Molecules*, vol. 25, no. 6, p. 1375, 2020. https://doi.org/10.3390/molecules25061375