UNIVERSIDAD DEL AZUAY

FACULTAD DE MEDICINA

DESARROLLO DE UNA RELACIÓN CUANTITATIVA ESTRUCTURA-ACTIVIDAD PARA LA ACTIVIDAD ANTIINFLAMATORIA DE ALCALOIDES

Trabajo de titulación previo a la obtención del título de Médico General

Ivanna Renata Cordero Jarrín

María Belén Tenesaca Castro

Dr. Cristian Rojas Villa

Cuenca-Ecuador
2023

Resumen

El objetivo de este trabajo fue calibrar diversas relaciones cuantitativas estructura-actividad (QSAR) para la predicción de la actividad antiinflamatoria de alcaloides. La base de datos está compuesta por 58 moléculas activas y 42 inactivas. Su estructura molecular fue modelada por el método de mecánica molecular (MM+) y el método semiempírico PM3. Las estructuras optimizadas se utilizaron para calcular diversos descriptores moleculares, huellas dactilares moleculares y claves estructurales. Se aplicó el método de reducción de variables basado en el algoritmo de Wootton, Sergent, Phan-Tan-Luu (V-WSP) para reducir la presencia de ruido, multicolinealidad y redundancia en el grupo inicial de descriptores. Con la base de datos curada se desarrollaron diversos modelos del aprendizaje automático basados en clasificadores lineales y no lineales. Para la validación del modelo, la base de datos se dividió de forma aleatoria en conjuntos de calibración (70 %) y predicción (30 %), manteniendo la proporción de las clases. Para la selección supervisada de los descriptores, se acoplaron los algoritmos genéticos (GAs-VSS) con los diversos clasificadores. Durante la calibración de los modelos se aplicó la validación cruzada de ventanas venecianas con cinco grupos de exclusión para evitar la presencia de sobreajuste. El desempeño de los modelos se cuantificó mediante la tasa de aciertos (NER). Entre los once modelos calibrados, el clasificador *k*NN exhibió la mejor capacidad predictiva para todas las formas de representación molecular. Estos modelos cumplen con los principios establecidos por la Organización para la Cooperación y el Desarrollo Económicos.

Palabras clave: Acción antiinflamatoria; Alcaloides; Clasificación; Modelos; QSAR.

Revisado por:

Dr. Cristian Rojas Villa

Director de tesis

Abstract

The aim of this work was to calibrate a quantitative structure-activity relationship (QSAR) for the prediction of the anti-inflammatory activity of alkaloids. The dataset was comprised of 52 active and 48 inactive molecules. Their molecular structure was modeled by the molecular mechanics (MM+) method followed by a semi-empirical method (PM3). The optimized structures were used to develop molecular descriptors, binary fingerprints and structural keys. The variable reduction method based on the algorithm of Wootton, Sergent, Phan-Tan-Luu (V- WSP) was used to reduce the presence of noise, multicollinearity and redundancy in the initial pool of descriptors. The curated database was then used to develop several models based on linear and non-linear machine learning classifiers. For validation purposes, the dataset was randomly divided into a training set (70%) and a test set (30%) while maintaining class proportions in both sets. When applying molecular descriptors, genetic algorithms-variable subset selection (GAs-VSS) methods were coupled with classifiers for the supervised descriptor selection. During the calibration of the models, a five-fold cross- validation protocol was used to optimize the classifier parameters and to avoid the presence of overfitting. The performance of the models was quantified by means of the non-error rate (NER) method. Among the eleven calibrated models, the *k*NN classifier model exhibited the best predictive ability for all molecular representations. These models meet the principles stated by the Organization for Economic Co-operation and Development.

Keywords: Anti-inflammatory action; Alkaloids; Classification; Models; QSAR.

Translated by:

Ivanna Renata Cordero Jarrín

María Belén Tenesaca Castro

1. Introducción

La inflamación es un proceso biológico muy importante debido a que es el signo más común cuando existe algún tipo de enfermedad en el organismo (1). La inflamación es una reacción del organismo a noxas, cuya intensidad es lo suficientemente grande como para provocar daños en la estructura y la función de los tejidos expuestos. El propósito de la inflamación es la neutralización de la noxa, la localización y limitación del daño y su reparación. Las noxas pueden ser de naturaleza física o química, así como microorganismos, parásitos, trastornos inmunológicos y neoplasias (2). Los signos y síntomas cardinales son: rubor, calor, dolor, tumor (hinchazón) (3) y la pérdida de funcionalidad ("*functio laesa*") (4). La acción antiinflamatoria farmacológica se basa por lo general en la inhibición de las ciclooxigenasas (COX-1 y COX-2) necesarias para la síntesis de las prostaglandinas (efecto de los antiinflamatorios no esteroidales o AINE) (5), el bloqueo de los receptores de los leucotrienos con fármacos como montelukast, zafirlukast (6) y la interferencia en la transcripción génica (glucocorticoides) (7).

Por otra parte, los alcaloides son compuestos químicos naturales que en su mayoría contienen átomos de nitrógeno y anillos aromáticos. También se denominan alcaloides a algunos compuestos de origen sintético con estructura similar a los alcaloides naturales (8). Existen alcaloides que son tóxicos para el ser humano, debido a que tienen actividades fisiológicas muy fuertes. Sin embargo, existen otros derivados beneficiosos para el ser humano; por ejemplo, actúan a nivel del sistema nervioso central con especial atención a la acción de los neurotransmisores acetilcolina, adrenalina, noradrenalina, ácido *γ*-aminobutírico (GABA), dopamina y serotonina (9). Por consiguiente, se han identificado diversos alcaloides que presentan variadas actividades farmacológicas, como por ejemplo antiinflamatoria, antihipertensiva, antiarrítmicos, anticáncer, entre los de mayor relevancia en las ciencias médicas (10).

Una estrategia para estudiar la capacidad antiinflamatoria es recurrir a modelos quimioinformáticos basados en las relaciones cuantitativas estructura-actividad

(QSAR) (11, 12). El objetivo de los modelos QSAR es el desarrollar modelos matemáticos (lineales o no lineales) entre la estructura química de las moléculas codificada mediante descriptores moleculares, y sus actividades farmacológicas (por ejemplo, la actividad o inactividad antiinflamatoria). Por lo tanto, un modelo basado en las relaciones cuantitativas estructura-actividad pueden ser de utilidad para la búsqueda de nuevos alcaloides con potencial antiinflamatorio (blancos o dianas moleculares) mediante la predicción a priori de su actividad biológica.

Con estos antecedentes, en el presente trabajo de titulación se ha desarrollado un modelo quimioinformático basado en las relaciones cuantitativas estructura-actividad (QSAR) para 100 alcaloides que presentan actividad e inactividad antinflamatoria. El modelo ha sido desarrollado siguiendo los cinco principios estipulados por la Organización para la Cooperación Económica y Desarrollo (OECD) (13) para asegurar la aplicabilidad del mismo. Se ha usado una representación molecular dependiente de la conformación mediante la optimización con el método semiempírico PM3 para el cálculo de los descriptores moleculares, huellas dactilares moleculares y claves estructurales. Posteriormente, se usaron diversos métodos de clasificación para encontrar el modelo QSAR que presente la mejor capacidad predictiva. La estabilidad y la predictividad del modelo se ha analizado mediante validación interna y validación externa, respectivamente, utilizando la tasa de aciertos (NER) como índice global de calidad.

2. Materiales y Métodos

2.1 Descripción de la base de datos y curado de la información

Para el desarrollo del modelo in silico, se ha utilizado una base de datos de 100 alcaloides a partir de dos publicaciones científicas “Anti-inflammatory activity of alkaloids: A twenty-century review” (14) y “Anti-Inflammatory Activity of Alkaloids: An Update from 2000 to 2010” (15). La respuesta reportada para cada compuesto es la actividad (58 moléculas) e inactividad (42 moléculas) con respecto a su capacidad antiinflamatoria. La medición de la actividad biológica ha sido determinada en

diversos blancos moleculares (humanos, ratones, ratas y pollos). La estructura química de cada alcaloide fue verificada en las bibliotecas químicas PubChem (16), ChemSpider (17) y NIST Chemistry WebBook (18), usando como criterio de entrada el nombre químico o el número de registro CAS.

Para el filtrado se programó un diagrama de flujo en el programa KNIME (19), el cual permite realizar filtrados, comparaciones e incluso transformaciones de los datos seleccionados, esto mediante nodos instalados. Un aspecto importante previo al cálculo de los descriptores es comprobar que la estructura química de los compuestos sea correcta. El curado se realizó con el programa AlvaMolecule (20) y se enfocó en 5 pasos:

1. Eliminación de estructuras químicas que no pueden procesarse por algunos programas de cálculo de descriptores.
2. Conversión y limpieza de estructuras químicas.
3. Estandarización y normalización de quimiotipos específicos.
4. Eliminación de compuestos repetidos.
5. Control manual final.

Durante el curado de las estructuras químicas, se identificaron incompatibilidades entre la estructura reportada en la literatura y aquella provista por las bibliotecas para los siguientes compuestos: iso-coridina, corituberina, coclaurina, tetrahidroalstonina, holsteina, staurosporine, isodelfinina, isotetrandrina, tetrandrina, melatonina, hipaconitina. Por lo tanto, para estos compuestos se han considerado la estructura provista por las librerías previamente indicadas.

2.2 Optimización de geometrías y cálculo de descriptores moleculares

Los alcaloides fueron diseñados en el programa HyperChem (21) para posteriormente optimizar las geometrías mediante la mecánica molecular (MM+) y refinadas usando el método semiempírico PM3. El criterio de convergencia

empleado es que el elemento máximo del vector gradiente de la energía total del sistema sea menor a 0.01kcal·(A·mol)-1). Seguidamente, la representación conformacional de los alcaloides se ha usado para calcular 5239 descriptores moleculares en el programa AlvaDesc (22), así como 1024 huellas dactilares moleculares de conectividad ampliada (ECFPs) en donde los parámetros que se usaron fueron 2 Bits por patrón, longitud mínima de 0 y máxima de 2, 1024 huellas dactilares de trayecto (PFPs) con 2 Bits por patrón, longitud mínima de 0 y máxima de 6 y 166 claves moleculares del sistema de acceso molecular (MACCS). Los descriptores moleculares son el resultado final de un procedimiento lógico y matemático que transforma la información química codificada en una representación simbólica de una molécula en un número útil o el resultado de algún experimento estandarizado (12). Posteriormente se excluyeron los descriptores con valores constantes y aquellos con al menos un valor faltante, obteniendo 3074 descriptores moleculares.

2.3 Desarrollo del modelo QSAR

2.3.1 Modelos de clasificación

En este trabajo se han utilizado diversos clasificadores, los que se pueden agrupar en dos categorías: métodos locales y métodos lineales. En la primera categoría se tiene a los *k*-vecinos más cercanos (*k*NN) (23), que es un método no lineal y no paramétrico que clasifica en función de analogías, es decir, un compuesto es clasificado según las clases a las que pertenecen la mayoría de las *k* moléculas más cercanas en el espacio de los descriptores. Para ello, el algoritmo calcula y analiza la matriz de distancias entre los objetos. El valor *k* se optimiza mediante validación cruzada de tal forma que brinde la mayor tasa de aciertos (NER). El modelo estará constituido por los objetos del grupo de calibración, el valor óptimo de *k* y la matriz de distancias para la predicción de la clase de nuevos compuestos. Otro método no lineal es el de los *N*-vecinos más cercanos (N3) (24), que es un método parecido a *k*NN con la variante de que, en lugar de utilizar los *k* vecinos, utiliza todos los n-1 elementos en base a un vector de clasificación que mide la contribución de los

vecinos desde el más cercano al más alejado de acuerdo a un factor de ponderación que está en función de dicha contribución. Por otra parte, los vecinos más cercanos agrupados (BNN) (24) utilizan un valor variable de *k* para cada objeto de acuerdo con un entorno definido, es decir, para clasificar un objeto toma en cuenta a todos los vecinos que posean similitud. Para esto se definen intervalos de similitud (denominados bins) y todos los vecinos que estén dentro de este intervalo serán considerados para asignar la clase a cada objeto.

Dentro de los métodos lineales más utilizados en el modelado QSAR se tiene al análisis discriminante por mínimos cuadrados parciales (PLSDA) (25), el cual es un método que combina las propiedades de los mínimos cuadrados parciales en regresión con la capacidad de un discriminante lineal. Para ello realiza una transformación de las variables en variables latentes que tienen la característica de ser ortogonales entre sí. El segundo método lineal es el de los bosques aleatorios (RF) (26, 27), que son un algoritmo de aprendizaje automático que combina la salida de múltiples árboles de decisión para llegar a un único resultado mediante el consenso. El algoritmo es una extensión del método de bagging, ya que utiliza tanto el ensacado como la aleatoriedad de características para crear un bosque de árboles de decisión no correlacionado. El último clasificador usado en este trabajo es Boosting Adaptativo (AdaBoost) (28), que es una técnica que tiene como objetivo combinar múltiples clasificadores débiles para construir un clasificador fuerte. Es posible que un solo clasificador no pueda predecir con precisión la clase de un objeto, pero cuando se agrupan varios clasificadores débiles, y cada uno aprende progresivamente de los objetos clasificados incorrectamente, se puede construir un modelo robusto.

2.3.2 Reducción y selección de descriptores moleculares

Una etapa fundamental durante el desarrollo de una relación cuantitativa estructura-actividad es la selección de un modelo que incluya un número reducido y parsimonioso de descriptores. Para tal efecto, se ha realizado una reducción inicial del número de descriptores en la base de datos mediante el método no supervisado

V-WSP (29), el cual selecciona un subconjunto representativo de variables de tal forma que muestren mínima correlación (por ejemplo, un umbral de 0,95) entre ellas dentro de un espacio multidimensional definido. Asimismo, para el desarrollo de los modelos de clasificación basados en descriptores moleculares (*k*NN y PLSDA) se han utilizado los algoritmos genéticos (GAs) (30) como estrategia de selección de descriptores. Los GAs son un método de selección de variables basados en la teoría de la evolución de Darwin, en el que se genera una población inicial de vectores binarios de p descriptores (cromosomas). Posteriormente, un proceso evolutivo genera nuevos cromosomas mediante la reproducción (crossover) y mutación de los cromosomas presentes en la población, de tal forma que generen un máximo en la tasa de aciertos en validación cruzada (NERcv).

2.3.3 Validación del modelo

Para la validación de los modelos de clasificación, la base de datos se dividió de forma aleatoria y proporcional a la numerosidad de las clases en conjuntos de calibración (70% de los alcaloides) y predicción (30% de moléculas). De esta forma, el grupo de calibración se utilizó para ajustar el modelo y las moléculas del grupo de predicción para evaluar la capacidad de predicción del mismo. La calidad del modelo se evaluó mediante el índice global de calidad denominado tasa de aciertos (*NER*), así como los índices primarios especificada (*Sp*), sensibilidad (*Sn*) y precisión (*Pr*) (31). La sensibilidad indica la capacidad del modelo para reconocer correctamente las muestras pertenecientes a una clase; mientras que la especificidad representa la capacidad que tiene una clase para rechazar las muestras de todas las otras clases. De igual manera, se utilizó la validación cruzada de dejar-varios-fuera (Imo) de ventanas venecianas con cinco grupos (32). Finalmente, la capacidad predictiva del modelo se obtuvo mediante la predicción de la actividad de las 30 moléculas del grupo de predicción.

2.4 Programas y algoritmos quimioinformáticos

Se utilizó el programa HyperChem (21) para diseñar y optimizar la geometría de los alcaloides, AlvaMolecule (20) para el curado de las moléculas. Los descriptores moleculares fueron calculados por el programa AlvaDesc (22). En el programa de MATLAB (33) se utilizó el V-WSP variable reduction (29), para realizar la reducción de los descriptores, la calibración del modelo y el análisis del dominio de aplicabilidad. Los algoritmos genéticos junto con los clasificadores *k*NN, N3, BNN, RF y AdaBoost fueron programados en MATLAB.

3. Resultado y discusión

La base de datos se encuentra conformada por 100 alcaloides (58 activos y 42 inactivos) con relación al proceso inflamatorio. Para efectos de validación de los modelos, la base de datos se ha dividido en grupos de calibración y predicción en una relación 70:30, en donde al mantener la numerosidad de las clases se tomaron del grupo activos 40 moléculas para el grupo de calibración y 18 para el grupo de predicción, mientras que del grupo inactivo 29 fueron seleccionadas para el grupo de calibración y 13 a predicción. A continuación, se han desarrollado once modelos QSAR basados en el aprendizaje supervisado de clasificación, considerando las 4 formas de representación de la estructura molecular y los seis clasificadores (los resultados se presentan en la Tabla 1). Se observa que el clasificador que presenta el mejor desempeño es el *k*NN con tres formas de representación molecular. Con las huellas dactilares moleculares de conectividad ampliada alcanza una capacidad predictiva del 80%, mientras que con las huellas dactilares moleculares de trayecto muestra una predictividad del 79%. Asimismo, al utilizar los descriptores moleculares se obtiene una predictividad del 77%. Por el contrario, los clasificadores que presentan la predictividad más baja es el *k*NN con los descriptores tridimensionales (40%), así como RF y AdaBoost con los descriptores moleculares.

Tabla 1. Parámetros globales de calidad de los modelos QSAR para la predicción de actividad antiinflamatoria

Modelo	**Parámetro Opt.**	**Tasa de Aciertos (NER)**		
		Cal	**Val**	**Pred**
Desc2D+*k*NN	*k*=6; d=2	0.71	0.76	0.77
Desc 2D+AdaBoost	N° max part=3; LR=0.5; N° árboles=100	1	0.79	0.59
Desc2D+RF	N° min obs hoja=1; N° árboles=75	1	0.81	0.62
Desc3D+AdaBoost	N° max part=3; LR=0.1; N° árboles=100	1	0.79	0.59
Desc3D+RF	N° min obs hoj=1; N° árboles=250	1	0.78	0.58
Desc3D+*k*NN	*k*=2; d=2	0.77	0.75	0.40
MACCS+*k*NN	*k*=3	0.64	0.67	0.77
MACCS+N3	α=1	0.72	0.72	0.69
MACCS+BNN	α=0.75	0.66	0.67	0.61
ECFPs+*k*NN	*k*=2	0.73	0.73	0.80
PFPs+*k*NN	*k*=2	0.72	0.72	0.79

4. Discusión

Para una explicación del fenómeno que desencadena el proceso antiinflamatorio, se ha tomado el modelo basado en el clasificador *k*NN y constituido por dos descriptores moleculares independientes de la conformación y seis vecinos. El espacio químico definido por los descriptores D/Dtr09 (índice anillo distancia/detour de orden 9) y P_VSA_m_2 (descriptor tipo P_VSA de entorno 2 ponderado por la masa) se presenta en la Figura 1. Claramente se ve que es un caso de clasificación asimétrico, en el cual no se logra identificar una frontera de separación bien definida

para los alcaloides activos e inactivos. Esta puede ser una justificación por la que el clasificador *k*NN funciona muy bien (a pesar de algunas superposiciones).

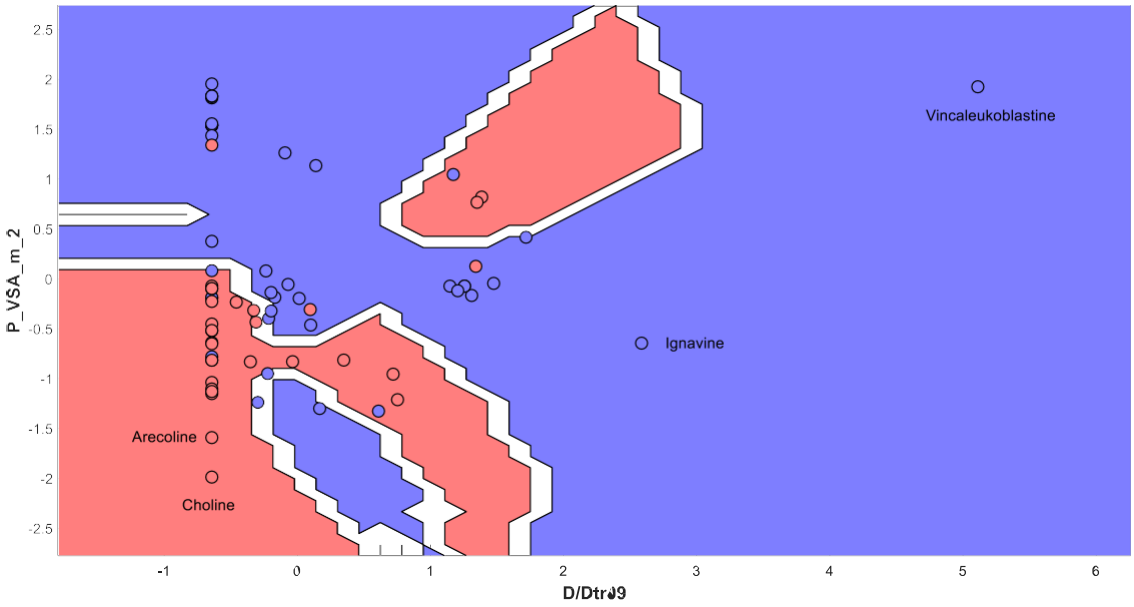

Figura 1. Distribución de los valores de los descriptores autoescalados D/Dtr09 y P_VSA_m_2 en los alcaloides activos e inactivos

Para un mejor entendimiento de la estructura molecular se han tomado los dos alcaloides activos que se encuentran distantes de los demás Ignavine y Vincaleukoblastine y dos alcaloides inactivos ubicados en el otro extremo Choline y Arecoline. Su estructura molecular se presenta en la Figura 2. Se puede evidenciar que la propiedad antiinflamatoria tiene una relación positiva en cuanto a la dimensión de cada alcaloide, es decir, mientras más grande sea su estructura y, por ende, este conformada por más anillos (entre estos los bencénicos) y más ciclos internos, mayor actividad antiinflamatoria va a poseer.

Al observar la Figura 2 se pone de manifiesto que la molécula Choline está conformada por una estructura básica al igual que la Arecoline. Por otro lado, se observa que la Ignavine se vuelve un mejor antiinflamatorio al tener ciclos internos

y un anillo bencénico presentando características similares que la Vincaleukoblastine que es la molécula que mostró mayor actividad antiinflamatoria y al igual que su análoga está constituida por varios ciclos internos y dos anillos bencénicos.

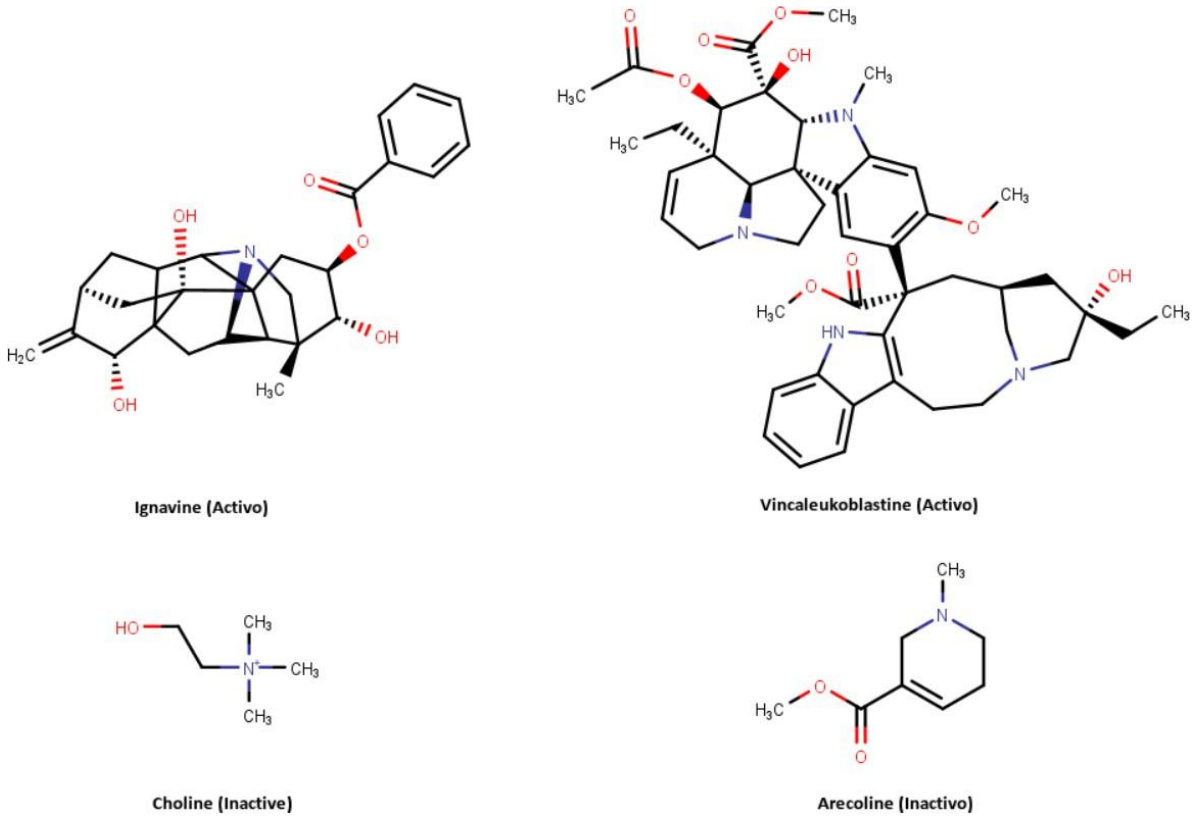

Figura 2. Comparación de la estructura química de los alcaloides activos e inactivos

La implementación de este modelo, y de los otros dos basados en el mismo clasificador y las huellas dactilares (ECFPs y FPFs) tiene varias ventajas. En primer lugar, acelera el proceso de descubrimiento de fármacos, ya que permite la identificación temprana de alcaloides prometedores con potencial antiinflamatorio. Además, reduce la necesidad de recursos y tiempo al eliminar la etapa de prueba y error en la selección de compuestos para estudios posteriores. Sin embargo, es

importante validar las predicciones obtenidas mediante ensayos experimentales adicionales antes de considerar los alcaloides como candidatos terapéuticos reales, así como la evaluación de toxicidad previo al uso como fármacos de uso terapéutico en humanos. Finalmente, el modelo debería actualizarse a medida que se encuentran más blancos moleculares con actividad e inactividad antiinflamatoria, de esta manera se cubre un mayor espacio químico de predictividad.

5. Conclusiones

En este trabajo se ha desarrollado un modelo QSAR basado en métodos de clasificación lineales y no lineales para discriminar alcaloides con capacidad antiinfamatoria de aquellos que no presentan esta actividad. Los valores experimentales tomados de la literatura fueron verificados previo al desarrollo del modelo. El modelo se desarrolló siguiendo varias estrategias quimiométricas. Los mejores modelos se obtienen con el clasificador *k*NN con tres formas diversas de representación molecular. Este modelo contribuye a la predicción de alcaloides antiinflamatorios, así como a la búsqueda racional de nuevos compuestos basados en modelos in silico, lo que constituye un avance significativo en el campo de la investigación en química medicinal.

Referencias bibliográficas

1. Kulinsky VI. Biochemical aspects of inflammation. Biochemistry (Moscow). 2007;72(6):595-607.

2. Wilting J, Becker J, Buttler K, Weich H. Lymphatics and Inflammation. Current Medicinal Chemistry. 2009;16(34):4581-92.

3. Scott A, Khan M, Cook JL. Inflammation What is "inflammation"? Are we ready to move beyond Celsus? Different definitions of inflammation are a cause for concern. British Journal of Sports Medicine. 2004.

4. Medzhitov R. Inflammation 2010: new adventures of an old flame. Cell. 2010;140(6):771-6.

5. Mitchell JA, Kirkby NS. Eicosanoids, prostacyclin and cyclooxygenase in the cardiovascular system. British Journal of Pharmacology. 2019;176(8):1038-50.

6. Lipworth BJ. Leukotriene-receptor antagonists. Lancet (London, England). 1999;353(9146):57-62.

7. Li Z, Hadlich F, Wimmers K, Murani E. Glucocorticoid receptor hypersensitivity enhances inflammatory signaling and inhibits cell cycle progression in porcine PBMCs. Frontiers in Immunology. 2022;13.

8. Harborne JB. Phytochemical methods : a guide to modern techniques of plant analysis: Chapman and Hall; 1998. 302- p.

9. Roberts MF, Wink M. Introduction. In: Roberts MF, Wink M, editors. Alkaloids: Biochemistry, Ecology, and Medicinal Applications. Boston, MA: Springer US; 1998. p. 1-7.

10. Keller WJ. REVIEWS: Introduction to Alkaloids—A Biogenetic Approach. Journal of Pharmaceutical Sciences. 1983;72(3):328-.

11. Gutman I. QSPR/QSAR Studies by Molecular Descriptors By Mircea V. Diudea. Nova Science Publishers: Huntington, NY. 2001. viii+438 pp. ISBN 1-5672-859-0. $97.00. Journal of Chemical Information and Computer Sciences. 2003;43(5):1720-1.

12. Todeschini R, Consonni V. Molecular Descriptors for Chemoinformatics. Second ed. Germany: Wiley-VCH Verlag GmbH & Co; 2009.

13. OECD. Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. 2014.

14. Barbosa-Filho JM, Piuvezam MR, Moura MD, Silva MS, Lima KVB, da-Cunha EVL, et al. Anti-inflammatory activity of alkaloids: A twenty-century review. Brazilian Journal of Pharmacognosy. 2006;16(1):109-39.

15. Souto AL, Tavares JF, Da Silva MS, Diniz MdFFM, De Athayde-Filho PF, Barbosa Filho JM. Anti-Inflammatory Activity of Alkaloids: An Update from 2000 to 2010. Molecules. 2011;16(10):8515-34.

16. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: Improved access to chemical data. Nucleic acids research. 2019;47(D1):D1102-D9.

17. Pence HE, Williams A. ChemSpider: An Online Chemical Information Resource. Journal of Chemical Education. 2010;87(11):1123-4.

18. Linstrom PJ, Mallard WG. The NIST Chemistry WebBook: A Chemical Data Resource on the Internet. 2001. p. 1059-63.

19. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME: The Konstanz Information Miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R, editors. Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization. Germany: Springer; 2008. p. 319-26.

20. Alvascience. alvaMolecule (software to view and prepare chemical datasets) version 2.0.6, https://www.alvascience.com. 2023.

21. Hypercube Inc. HyperChem™ Professional version 8.0. http://www.hyper.com.

22. Alvascience. alvaDesc (software for molecular descriptors calculation) version 2.0.16, https://www.alvascience.com. 2023.

23. Cover TM, Hart PE. Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory. 1967;13(1):21-7.

24. Todeschini R, Ballabio D, Cassotti M, Consonni V. N3 and BNN: Two New Similarity Based Classification Methods in Comparison with Other Classifiers. Journal of Chemical Information and Modeling. 2015;55(11):2365-74.

25. Wold S, Sjöström M, Eriksson L. PLS-Regression: A Basic Tool of Chemometrics. Chemometrics and Intelligent Laboratory Systems. 2001;58(2):109-30.

26. Breiman L. Random Forests. Machine Learning. 2001;45:5–32.

27. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Springer Series in Statistics. 2009.

28. Hu G, Yin C, Wan M, Zhang Y, Fang Y. Recognition of diseased Pinus trees in UAV images using deep learning and AdaBoost classifier. Biosystems Engineering. 2020;194:138-51.

29. Ballabio D, Consonni V, Mauri A, Claeys-Bruno M, Sergent M, Todeschini R. A Novel Variable Reduction Method Adapted from Space-Filling Designs. Chemometrics and Intelligent Laboratory Systems. 2014;136:147-54.

30. Leardi R. Genetic algorithms in chemistry. In: Brown S, Tauler R, Walczak B, editors. Comprehensive Chemometrics: Chemical and Biochemical Data Analysis. 1. Second ed. Amsterdam, The Netherlands: Elsevier; 2020. p. 617-34.

31. Ballabio D, Grisoni F, Todeschini R. Multivariate Comparison of Classification Performance Measures. Chemometrics and Intelligent Laboratory Systems. 2018;172:1-12.

32. Ballabio D, Consonni V. Classification tools in chemistry. Part 1: Linear models. PLS-DA. Analytical Methods. 2013;5(16):3790-8.

33. The MathWorks Inc. MATLAB, http://www.mathworks.com.