chemistry proceedings

MDPI

Proceeding Paper

Development of Quantitative Structure–Anti-Inflammatory Relationships of Alkaloids [†]

Cristian Rojas [1,*], Doménica Muñoz [2], Ivanna Cordero [3], Belén Tenesaca [3] and Davide Ballabio [4]

[1] Grupo de Investigación en Quimiometría y QSAR, Facultad de Ciencia y Tecnología, Universidad del Azuay, Av. 24 de Mayo 7-77 y Hernán Malo, Cuenca 010107, Ecuador
[2] Unidad Académica de Salud y Bienestar, Universidad Católica de Cuenca, Av. De las Américas y Humboldt, Cuenca 010101, Ecuador; domenicav.munoz@gmail.com
[3] Facultad de Medicina, Universidad del Azuay, Av. 24 de Mayo 7-77 y Hernán Malo, Cuenca 010107, Ecuador; ivannacordero@es.uazuay.edu.ec (I.C.); mbtenesacacas@es.uazuay.edu.ec (B.T.)
[4] Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano-Bicocca, P.za della Scienza 1, 20126 Milano, Italy; davide.ballabio@unimib.it
[*] Correspondence: crojasvilla@gmail.com

[†] Presented at the 28th International Electronic Conference on Synthetic Organic Chemistry (ECSOC 2024), 15–30 November 2024; Available online: https://sciforum.net/event/ecsoc-28.

Abstract: Alkaloids are naturally occurring metabolites with a wide variety of pharmacological activities and applications in science, particularly in medicinal chemistry as anti-inflammatory drugs. Since they could be labelled as active or inactive compounds against the inflammatory biological response, the aim of this work was the calibration of quantitative structure-activity relationships (QSARs) based on machine learning classifiers to predict anti-inflammatory activity based on the molecular structures of alkaloids. The dataset of 100 alkaloids (58 active and 42 inactive) was retrieved from two systematic reviews. Molecules were properly curated, and the molecular geometries of the compounds were optimized by the semi-empirical method (PM3) to calculate molecular descriptors, binary fingerprints (extended-connectivity fingerprints and path fingerprints) and MACCS (Molecular ACCess System) structural keys. Then, we calibrated QSAR models based on well-known linear and non-linear machine learning classifiers, i.e., partial least squares discriminant analysis (PLSDA), random forests (RF), adaptive boosting (AdaBoost), *k*-nearest neighbors (*k*NN), *N*-nearest neighbors (N3) and binned nearest neighbors (BNN). For validation purposes, the dataset was randomly split into a training set and a test set in a proportion of 70:30. When using molecular descriptors, genetic algorithms-variable subset selection (GAs-VSS) were used for the supervised feature selection. During the calibration of the models, a five-fold venetian blinds cross-validation was used to optimize the classifier parameters and to control the presence of overfitting. The performance of the models was quantified by means of the non-error rate (*NER*) statistical parameter.

Keywords: alkaloids; anti-inflammatory activity; molecular descriptors; machine learning classifiers; QSAR

Citation: Rojas, C.; Muñoz, D.; Cordero, I.; Tenesaca, B.; Ballabio, D. Development of Quantitative Structure–Anti-Inflammatory Relationships of Alkaloids. *Chem. Proc.* **2024**, *6*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s): Name

Published: 15 November 2024

1. Introduction

Inflammation is a crucial homeostatic and defense mechanism in the human body, triggered by mechanical, chemical, or microbial stimuli that affect vascularized tissues. It is a complex process that begins when damaged tissue cells release pro-inflammatory mediators like histamine and prostaglandins. These mediators cause vasodilation, increasing blood flow to the affected area, leading to redness, heat, pain, edema, and loss of function. Vascular permeability also increases, and neutrophils are attracted to the injury site, where they phagocytize pathogens. After the harmful agents are removed, the anti-inflammatory mediators promote tissue repair [1]. Although inflammation is a protective response, dysregulation at various stages can lead to chronic inflammation and tissue damage, which can contribute to diseases such as arthritis and cancer. While there are many anti-inflammatory agents, non-steroidal anti-inflammatory drugs (NSAIDs) that inhibit the cyclooxygenase enzymes, COX-1 and COX-2 have proven efficacy; however, they have adverse side effects [2]. Medicinal chemists within the pharmaceutical industry are searching for new and effective anti-inflammatory agents.

In this context, alkaloids have gained significant importance in medicinal chemistry due to their anti-inflammatory activity. It has been demonstrated that alkaloids reduce the production of pro-inflammatory cytokines such as TNF-α and interleukins (IL-1, IL-6), which inhibit the NF-κB pathway, decrease prostaglandin synthesis, and reduce cellular infiltration into inflamed tissues, thereby preventing various aspects of the inflammatory process [3]. Alkaloids constitute a large group of organic compounds derived from the secondary metabolism of plants, fungi, and microorganisms, which have a common characteristic of the presence of nitrogen atoms in their structure. These compounds are generally colorless, odorless, bitter, crystalline, and, in some cases, amorphous or liquid at room temperature [4].

A well-known strategy to study and predict the anti-inflammatory activity of compounds involves the development of cheminformatic models based on quantitative structure-activity relationships (QSARs) [5]. In this framework, we calibrated diverse QSAR models for the anti-inflammatory activity of 100 alkaloids. Molecules were properly represented by diverse molecular features, which were modelled using six well-known machine learning classifiers. QSAR models were developed following the principles stated by the Organization for Economic Cooperation and Development (OECD) [6].

2. Materials and Methods

2.1. Alkaloids Database Description

For the development of the anti-inflammatory alkaloids database, we considered two systematic reviews [3,7]. These authors listed alkaloids with the corresponding anti-inflammatory activity (active, inactive, or weakly active) measured in different experimental assays. Alkaloids labelled as weakly active were merged into the inactive class. Since only 42 non-active alkaloids were available in the database, we randomly selected a balanced number of active compounds (58 molecules) to complete a database of 100 molecules. The chemical information of the molecules was verified in the PubChem open access library [8], where we also retrieved the Chemical Abstracts Service (CAS) registry number and PubChem CID.

2.2. Molecular Structure of Alkaloids and Feature Representation

The chemical structures of the alkaloids were designed in the HyperChem software [9]. Then, the geometries were optimized with the PM3 semiempirical method until the gradient vector became less than 0.01 kcal (Å mol)$^{-1}$. The molecular structures of alkaloids were curated to check for potential errors by using the diverse tools implemented in the alvaMolecule software [10,11], along with generation of the canonical SMILES (Simplified Molecular Input Line Entry System). Table S1 in the supplementary material shows the information of the curated database. Then, the alvaDesc software [10,11] was used to compute diverse sets of molecular features [5]: (1) 5633 molecular descriptors (MDs); (2) 166 MACCS (Molecular ACCess System) structural keys; (3) 1024 extended-connectivity fingerprints (ECFPs); and (4) 1024 Path Fingerprints (PFPs). When working with MDs, we excluded non-informative features. Subsequently the V-WSP unsupervised variable reduction method [12] was used to further reduce the presence of molecular descriptors with multicollinearity, redundancy and noise.

2.4. Machine Learning Classifiers

Since anti-inflammatory activity is a qualitative discrete response, we used diverse machine learning classifiers to calibrate linear and non-linear quantitative structure-activity relationships for the discrimination between the active and inactive alkaloids. In this work, we used six classifiers: (1) Partial Least-Squares Discriminant Analysis (PLSDA) [13], that combines the properties of partial least squares regression (PLS2-based method) with the ability of a discrimination classifier by calculating latent variables (LVs); (2) *k*-Nearest Neighbors (*k*NN) [14], which is a nonparametric local-based method that classifies a molecule by means of the majority vote of its *k* closest neighbors; (3) *N*-Nearest Neighbors (N3) [15] that classifies an alkaloid considering the class of all the $n - 1$ molecules, in which the contribution to the class assignment is sorted in a vector with the similarity rank; (4) Binned Nearest Neighbors (BNN) [15], which considers the majority vote for a variable number of *k* neighbors defined by similarity intervals (bins) where alkaloids are distributed; (5) Random Forest (RF) [16], which is a non-linear ensemble learning that constructs several decision trees (sub-samples) and uses the averaging prediction by combining them; and (6) Adaptive Boosting (AdaBoost) [17], which is another ensemble classifier that sequentially fits decision trees and then calibrates additional models by adjusting the weights of the misclassified molecules in such a way as the subsequent decision trees pay more attention to them.

2.5. Supervised Feature Selection

To find the most informative subset of molecular features, PLSDA and *k*NN classifiers were coupled with the Genetic Algorithms-Variable Subset Selection (GAs-VSS) [18]. The GA-VSS creates an initial population of models (also called chromosomes) in a random way. Each model consists of a binary vector that indicates the presence/absence of features in the model. Then, new models are created by combining the initial chromosomes (also called crossover) or by a random inclusion/exclusion of MDs (also known as mutation). This evolutionary process optimizes the non-error rate classification parameter in cross-validation.

2.6. Validation Performance

The reliability of QSAR models is related to the application of cross-validation procedures to analyze internal model stability, as well as external validation to quantify the model's predictiveness. To this end, the database of 100 alkaloids was randomly split into training and test sets in a proportion of 70:30, maintaining the class proportions in both sets. Alkaloids of the training set were used to calibrate the models and to measure the internal models' stability in cross-validation by means of the five-fold venetian blinds approach. Finally, test set alkaloids were used to check the predictive ability of the models. For all the classifiers, we analyzed the non-error rate (*NER*) [19].

3. Results and Discussion

The database of 100 alkaloids was randomly split into a training set and test set of 70 and 30 molecules, respectively (refer to Table S1 for splitting assignment). Initially, we used the molecular descriptors to calibrate models based on the PLSDA, *k*NN, RF and AdaBoost classifiers. After the exclusion of non-informative descriptors, 3335 features were reduced by means of V-WSP unsupervised variable reduction using a correlation threshold of *thr* = 0.90. Thus, 936 molecular descriptors were retained to calibrate models in two instances: (1) models that use conformation independent features only; and (2) models with the complete pool of descriptors (0D, 1D, 2D and 3D features). The supervised selection of MDs for the PLSDA and *k*NN classifiers was performed by the GA-VSS. The optimal number of *LVs* and *k* for the PLSDA and *k*NN, respectively; were optimized using five-fold venetian blinds cross-validation. The same process was used to optimize the hyperparameters for the RF (m*inLeafSize*: minimum number of observations per tree leaf and *n_trees*: number of trees to be grown) and AdaBoost (*maxNumSplits*: maximal number of decision splits per tree, *LR*: learning rate and *n_trees*: number of trees) ensemble classifiers. This optimization process led to eight models (Table S2).

The best quantitative-structure–anti-inflammatory relationship based on MDs was obtained with the *k*NN classifier (Table 1). This model used six neighbors and two features: distance/detour ring index of order 9 (*D*/*Dtr09*) and P_VSA-like on mass, bin 2 (*P_VSA_m_2*). Figure 1 shows the chemical space defined by these two descriptors, where a non-linear separation is clearly defined.

Table 1. Non-error rate classification performance of the best QSAR models using MDs and ECFPs.

Molecular Feature Type	Classifier	Optimal Parameters	Training Set	Cross-Validation	Test Set
Molecular descriptors	*k*NN	*k* = 6; *MDs* = 2	0.71	0.76	0.77
Extended-connectivity fingerprints	BNN	*α* = 0.9	0.75	0.75	0.81

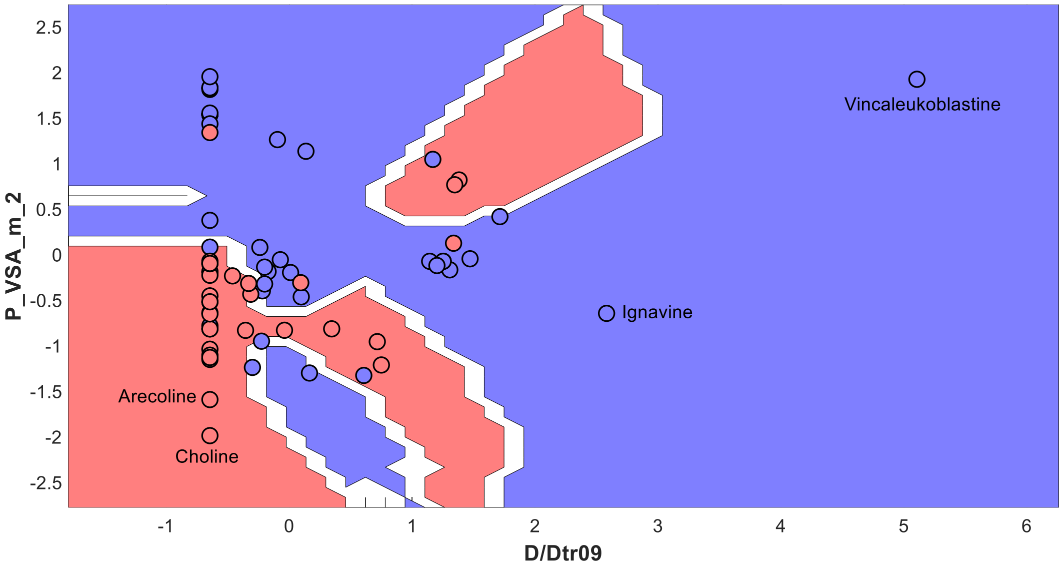
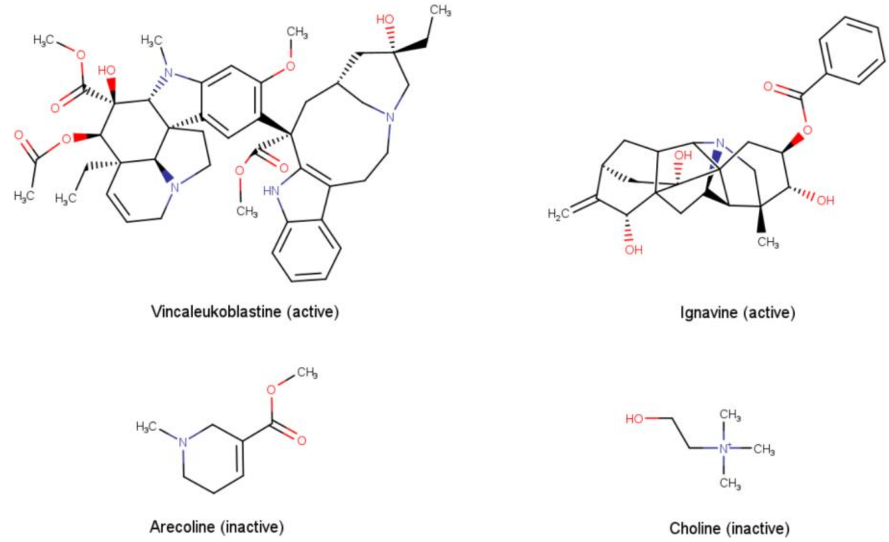


Figure 1. Chemical space of the quantitative structure–anti-inflammatory activity of alkaloids for the training set. The *D*/*Dtr09* and *P_VSA_m_2* features are autoscaled. Blue and Red indicate the class spaces related to active and inactive anti-inflammatory alkaloids, respectively.

The *D*/*Dtr09* descriptor is an index calculated as a quotient of the distance and detour matrices (**D**/**Δ**), which is a symmetric matrix whose off-diagonal elements are the ratio of the lengths of the shortest over the longest path between any pair of vertices. The row sums have been proposed as local invariants with high capability to discriminate between branching vertices (small row sums) and bridging vertices (large row sums) [5,20]. In addition, the *P_VSA_m_2* feature defines the amount of van der Waals surface area (VSA) having the mass property withing the values of the range for bin two: [1,1.2). This descriptor corresponds to a partition of the molecular surface area conditioned by the atomic values of the mass [5,21]. For instance, the Vincaleukoblastine (*D*/*Dtr09* = 878.3 and *P_VSA_m_2* = 241.8) and Ignavine (*D*/*Dtr09* = 492.8 and *P_VSA_m_2* = 100.9) active anti-inflammatory alkaloids are the largest molecules having diverse cycles, including both aromatic and internal. In contrast, the inactive alkaloids Arecoline (*D*/*Dtr09* = 0.0 and *P_VSA_m_2* = 49.0) and Choline (*D*/*Dtr09* = 0.0 and *P_VSA_m_2* = 27.3) are characterized by small molecules and the low presence of cycles in their scaffolds (Figure 2).



Vincaleukoblastine (active)

Ignavine (active)

Arecoline (inactive)

Choline (inactive)

Figure 2. Alkaloids with positive (Vincaleukoblastine and Ignavine) and negative (Arecoline and Choline) anti-inflammatory activity.

In a second step, we used a pool of 166 MACCS, 1024 ECFPs and 1024 PFPs to calibrate models using the *k*NN, N3 and BNN local-based classifiers. The Jaccard-Tanimoto distance was used to quantify the similarity between pairs of alkaloids. Table S3 shows the classification performances of the calibrated quantitative structure–anti-inflammatory relationships. In this case, the best model corresponded to the ECFPs and the BNN classifier (Table 1). In fact, ECFPs exhaustively enumerated all fragments in the chemical scaffold into a fixed-length vector, which presented an effective way for molecular representation and a similarity search. Thus, the anti-inflammatory activity prediction of alkaloids depends on the use of a variable number of *k* neighbors to contribute to the majority vote to the class assignment.

4. Conclusions

In this work, quantitative structure-anti-inflammatory relationships based on machine learning classifiers has been developed to discriminate between active and inactive alkaloids. The chemical structure of alkaloids was represented by diverse molecular features. When using molecular descriptors, a two-feature model with the *k*NN classifier emerged as the optimal one. On the other hand, when using binary features, the best model was achieved using the ECFPs and the BNN classifiers. It is important to note that anti-inflammatory activity prediction depends on locally based classifiers. Thus, the models developed here could assist medicinal chemists to better understand the mechanism involved in the chemical structure of alkaloids, as well as the prediction of novel natural or synthetic alkaloids to be used as anti-inflammatory agents.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Table S1: Details of the database of active/inactive anti-inflammatory alkaloids; Table S2. QSAR models for the anti-inflammatory activity using molecular descriptors; Table S3. QSAR models for the anti-inflammatory activity using binary structural keys and fingerprints.

Author Contributions: Conceptualization, C.R.; methodology, C.R., D.M., I.C., B.T. and D.B.; software, C.R. and D.B.; validation, C.R. and D.B.; formal analysis, C.R., I.C., B.T. and D.B.; investigation, C.R., D.M. and D.B.; data curation, C.R., D.M., I.C., B.T.; writing—original draft preparation, C.R., D.M. and D.B.; writing—review and editing, C.R., D.M. and D.B.; visualization, C.R. and D.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Institutional Review Board Statement:

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in [3,7].

Acknowledgments: The authors thank Wayne R. Hanson for his valuable revision of the manuscript and for providing some useful comments for improving the technical quality.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kulinsky, V.I. Biochemical Aspects of Inflammation. *Biochemistry* **2007**, *72*, 595–607.
2. Mitchell, J.A.; Kirkby, N.S. Eicosanoids, prostacyclin and cyclooxygenase in the cardiovascular system. *Br. J. Pharmacol.* **2019**, *176*, 1038–1050.
3. Souto, A.L.; Tavares, J.F.; Da Silva, M.S.; Diniz, M.d.F.F.M.; de Athayde-Filho, P.F.; Filho, J.M.B. Anti-inflammatory Activity of Alkaloids: An Update from 2000 to 2010. *Molecules* **2011**, *16*, 8515–8534.
4. Roberts, M.F.; Wink, M. Introduction. In *Alkaloids: Biochemistry, Ecology, and Medicinal Applications*; Roberts, M.F., Wink, M., Eds.; Springer: Boston, MA, USA, 1998; pp. 1–7.
5. Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; WILEY-VCH: Weinheim, Germany, 2009.
6. OECD. *Guidance Document on the Validation of (Quantitative)Structure-Activity Relationships [(Q)SAR] Models*; 2014.
7. Barbosa-Filho, J.M.; Piuvezam, M.R.; Moura, M.D.; Silva, M.S.; Lima, K.V.B.; da-Cunha, E.V.L.; Fechine, I.M.; Takemura, O.S. Anti-inflammatory Activity of Alkaloids: A twenty-century Review. *Rev. Bras. Farmacogn.* **2006**, *16*, 109–139.
8. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B. PubChem 2023 Update. *Nucleic Acids Res.* **2023**, *51*, D1373–D1380.
9. Hypercube Inc. HyperChem™ Professional Version 8.0. Available online: http://www.hyper.com (accessed on).
10. Alvascience Software Solutions for Cheminformatics and QSAR Research. 2023. Available online: https://www.alvascience.com (accessed on).
11. Mauri, A.; Bertola, M. Alvascience: A New Software Suite for the QSAR Workflow Applied to the Blood–Brain Barrier Permeability. *Int. J. Mol. Sci.* **2022**, *23*, 12882.
12. Ballabio, D.; Consonni, V.; Mauri, A.; Claeys-Bruno, M.; Sergent, M.; Todeschini, R. A Novel Variable Reduction Method Adapted from Space-filling Designs. *Chemom. Intell. Lab. Syst.* **2014**, *136*, 147–154.
13. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A Basic Tool of Chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.
14. Kowalski, B.; Bender, C. *k*-Nearest Neighbor Classification Rule (Pattern Recognition) Applied to Nuclear Magnetic Resonance Spectral Interpretation. *Anal. Chem.* **1972**, *44*, 1405–1411.
15. Todeschini, R.; Ballabio, D.; Cassotti, M.; Consonni, V. N3 and BNN: Two New Similarity Based Classification Methods in Comparison with Other Classifiers. *J. Chem. Inf. Model.* **2015**, *55*, 2365–2374.
16. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
17. Freund, Y.; Schapire, R.E. A Desicion-Theoretic Generalization of On-line Learning and an Application to Boosting. In Proceedings of the Computational Learning Theory, Barcelona, Spain, 13–15 March 1995; pp. 23–37.
18. Leardi, R. Genetic Algorithms in Chemistry. In *Comprehensive Chemometrics: Chemical and Biochemical Data Analysis*, Second ed.; Brown, S., Tauler, R., Walczak, B., Eds.; Elsevier: Oxford, UK, 2020; Volume 1, pp. 617–634.
19. Ballabio, D.; Grisoni, F.; Todeschini, R. Multivariate Comparison of Classification Performance Measures. *Chemom. Intell. Lab. Syst.* **2018**, *174*, 33–44.
20. Randić, M. On Characterization of Chemical Structure. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 672–687.
21. Labute, P. A Widely Applicable Set of Descriptors. *J. Mol. Graph. Model.* **2000**, *18*, 464–477.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.