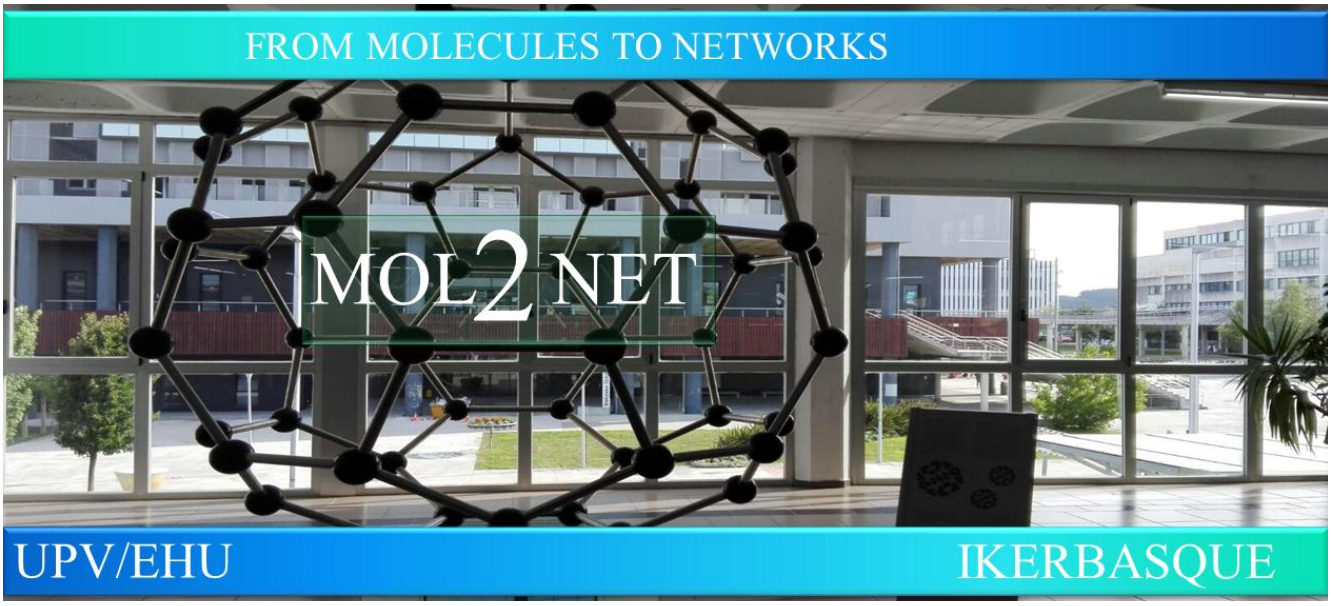
MOL2NET'21, Conference on Molecular, Biomedical & Computational Sciences and Engineering, 7th ed.



QSPR modeling of the logP for drugs potentially active on the central nervous system

Luis A Torres Gómez[a], Juan Carlos Polo Vega[b], Alejandro Almeida Pons[a].

[a] *Institute of Pharmacy and Foods. University of the Havana*
b Center for Genetic Engineering and Biotechnology of Cuba

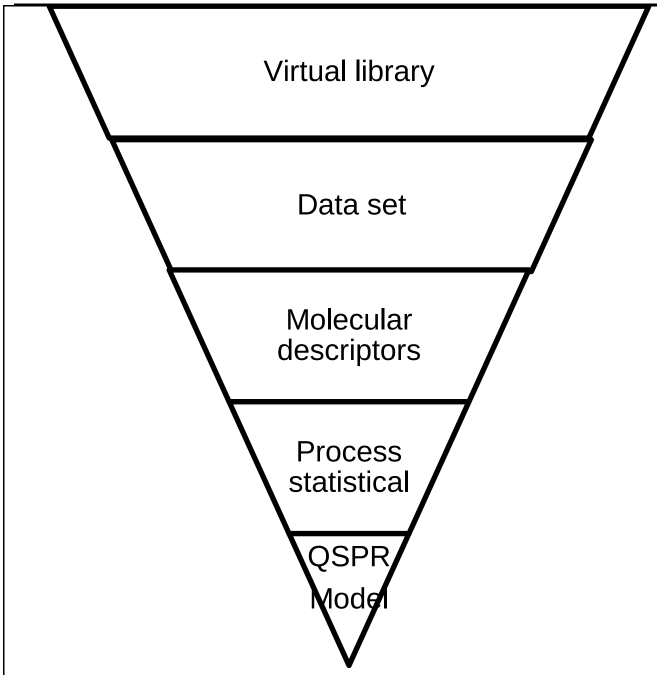
.

.

Graphical Abstract

Abstract.

The degree of lipophilicity of a drug, defined by its partition coefficient, is a key parameter to understand and analyze the activity of said compound in the body. In the case of drugs active on the central nervous system, the value of the partition coefficient or log P indicates their ability to cross the blood-brain barrier and carries out their pharmacological action. However, the experimental determination of this property is accompanied by several drawbacks, since the methods used for this purpose are



usually time-consuming, resource-intensive, and expensive equipment, in addition to being prone to measurement errors. To solve these problems, computational molecular modeling methods are used. Among them are the QSPR studies, with which quantitative relationships are established between the structural characteristics of the analyzed compounds and the property of interest. In the present work, a predictive model was built based on the partition coefficient as a property of interest. For this, a training series of 286 active drugs on the central nervous system was built, obtaining their structures, simplified representations and experimental values of the partition coefficient using the ACDLabs software. The MODESLAB program was used to calculate a total of 91 molecular descriptors and based on these, the BuildQSAR software was used to obtain the predictive model.

Introduction

.

The central nervous system (CNS) has been the object of scientific study for a long time, with emphasis on the diseases of this system and the cure for them. Today there are effective treatments and medications against conditions such as Alzheimer's disease, psychosis and depression. In addition, research is being carried out to develop more effective drugs against these diseases. Said efficacy is closely related to the physical-chemical properties of the compounds used

One of these properties is the partition or distribution coefficient, which provides a direct measure of the degree of lipophilicity of each compound. This property significantly influences the behavior of active drugs on the CNS, since the more lipophilic a drug is, the greater it is its ability to cross the blood-brain barrier and performs its pharmacological function. In particular, quantitative structure-property relationship studies (QSPR) are designed to determine the characteristics, qualities and chemical-physical properties of a compound, through the information collected by their molecular descriptors. Specifically, they create a correlation between the chemical structure of the compound and the property of interest, so that the value of said property can be predicted from its structure, or through some change in it For the adequate construction of predictive models relatively simple and easy to interpret, different software is used: MODESLAB for the determination of molecular descriptors and Build QSAR for the elaboration of the predictive model.

Results and Discussion

In order to choose the best molecular descriptors to include as independent variables, the Genetic Algorithm method of the Build QSAR software was used (Table II). As a result, 3 candidate models were obtained with 5 independent variables each. The following conditions were established in the statisticians as essential requirements for the evaluation of the candidate model to be selected: multiple correlation coefficient R (R>0.6); determination coefficient R2 (R2>0.5); F coefficient of the ANOVA test (F>>1 with p< 0.05); standard error to the estimate s (s<1). Cross-validation statisticians were also taken into account to assess the quality of the independent variables and the predictive ability of the candidate models. The requirements to be met by these parameters were the following: the predictive residual sum of the squared standard deviation (Spress) and the error in prediction of the standard deviation (Sdep) must be similar to the value of the standard error of the estimate (s) and the cross-validation coefficient (Q2) must be greater than 0.5 (Q2>0.5).

Table I. Models obtained after applying the genetic algorithm method. Source Build QSAR/Genetic Algorithm

	X-1	X-2	X-3	X-4	X-5	R	s	F	Q^2	S_{press}	S_{dep}
1	μ(Hyd)1	μ(Hyd)14	μ(Pol)2	μ(Ato)13	μ(Std)1	0.768	1.214	80.485	0.567	1.247	1.236
2	μ(Hyd)1	μ(Std)15	μ(Pol)2	μ(Ato)15	μ(Std)1	0.764	1.223	78.367	0.559	1.259	1.248
3	μ(Hyd)1	μ(Std)12	μ(Pol)2	μ(Ato)14	μ(Std)1	0.763	1.225	77.992	0.557	1.262	1.250

As can be seen in Table II, the first candidate model presented a better fit. It showed a correlation coefficient R value of 0.768 which demonstrates a relatively high correlation between the molecular descriptors and the biological property. The value of the estimated standard error (s) although it is not less than 1, it is a fairly close value. An F value of the simple ANOVA test of 80.458 was obtained, which is much greater than 1, which shows the linear relationship between the set of independent variables and the biological property in question. On the other hand, the values of the cross-validation statisticians also show favorable values. In the case of the squared standard deviation (Spress) and the standard deviation prediction error (Sdep), the similarity in numerical value to the estimated standard error (s) can be observed, which demonstrates an adequate level of stability when compounds are excluded for the construction of the mathematical model. The cross-validation coefficient Q^2 is 0.567, which indicates that the explanatory capacity of the models created by these selections is greater than 50% for the training series. As a result, this model was selected for the study and was named M1 for the purposes of this work. Based on the regression statisticians, it can be seen that the model complied with the proposed statistical requirements, since the value of R was greater than 0.7, the value of F was much greater than 1, and the value of R2 was greater than . 0.5, so it presented a satisfactory fit. However, these values are not entirely high because their ideal values to obtain the best possible model are------. In addition, the value of s, although not very large, is greater than 1, so it does not meet the previously stated requirement of s<1. For these reasons, the model, despite having a moderate fit, had to be optimized.

Optimization of the M1 model and obtaining

the M2 model To optimize the M1 model, atypical outlier compounds were determined and eliminated using the Build QSAR software. After excluding a total of 41 of these values, a model was obtained from 245 compounds with a higher statistical quality, which was called M2. Their regression statisticians presented the following values: R = 0.912; R2= 0.8322; s = 0.712; F=237.1094; p < 0.0001. In this model, an increase in the value of R up to 0.912 is observed. The standard error of the estimate (s) is reduced to 0.712, lower value 1, with which the statistical criterion is met. As for the values of R2 and F, they increase to the figures of 0.8322 and 237.1094, respectively, which represents a considerable increase with respect to the values of the same parameters in the M1 model. These results indicate that model M2 has a better fit than model M1, and is represented by the following mathematical function: Model 2 (M2

$$\begin{aligned} Log\ P = {} & +0{,}0064\ (\pm\ 0{,}0006)\ \mu\ (Hyd)1 - 0{,}0000\ (\pm 0{,}0000)\ \mu\ (Hyd)14 \\ & + 0{,}0000\ (\pm 0{,}0000)\ \mu\ (Pol)2 - 0{,}0000\ (\pm 0{,}0000)\ \mu\ (Ato)13 \\ & + 0{,}0009\ (\pm\ 0{,}0009)\ \mu\ (Std)1 + 0{,}4565\ (\pm 0{,}2823) \end{aligned}$$

Analysis of the M2 model

The t test of significance of the slopes in the multiple linear regression shows whether the independent variables contribute, significantly or not, to the variation of the physicochemical property chosen as the dependent variable (log P). Table II shows the results of this test.

Table II. Significance t-test of the slopes and coefficients of the M2 model. Source BuildQSAR/RLM/Fitting analysis

	Coefficients	**St. Dev.**	**95%**	**t-ratio**	**p**	Commentary
Const.	0,4565	0,1412	0,2823	3,2346	0,0014	significant
μ(Hyd)1	0,0064	0,0003	0,0006	22,2505	0,0000	significant
μ(Hyd)14	0,0000	0,0000	0,0000	-1,2924	0,1975	not significant
μ(Pol)2	0,0000	0,0000	0,0000	4,2182	0,0000	significant
μ(Ato)13	0,0000	0,0000	0,0000	-4,1342	0,0000	significant
μ(Std)1	0,0009	0,0004	0,0009	2,2329	0,0265	significativa

The results of the significance t-test of the slopes show that the coefficients of the independent variables, except µ(Hyd)14, are markedly different from 0. This indicates that any variation in the molecular descriptors µ(Std)1, µ(Hyd)1, µ(Pol)2 and µ(Ato)13 contribute significantly to the variation of the log P physicochemical property. The molecular descriptor µ(Hyd)14 has a coefficient much closer to 0, so it does not has a significant relationship with the value of log P

.

.

References

- A Primer on QSAR/QSPR Modeling Fundamental Concepts by Kunal Roy, Supratik Kar, Rudra Narayan Das (z-lib.org) 2015
- Aponte LJ, Scior T. ¿Qué sabe usted acerca de QSAR/QSPR? Revista Mexicana de Ciencias Farmacéuticas. 2012; 43.
- Apostol TM. Calculus. USA: Blaisdell Publishing Co. 1969.
- Blake JF. Chemo informatics - predicting the Physicochemical Properties of Drug-like Molecules. CurrOpinBiotech. 2000; 11: 104-7.
- Bun S. Predicción y evaluación de la solubilidad de los compuestos orgánicos de interés farmacéutico (tesis de diploma en Ciencias Farmacéuticas). Universidad Central de Las Villas, Marta Abreu. 2009.
- Cabrera MA, Bermejo M, Ramos L, Grau R, Pérez M, et al. A topological sub-structural approach for predicting human intestinal absorption of drugs. EuropeanJournal of Medicinal Chemistry. 2004; 39: 905-16.

.

.