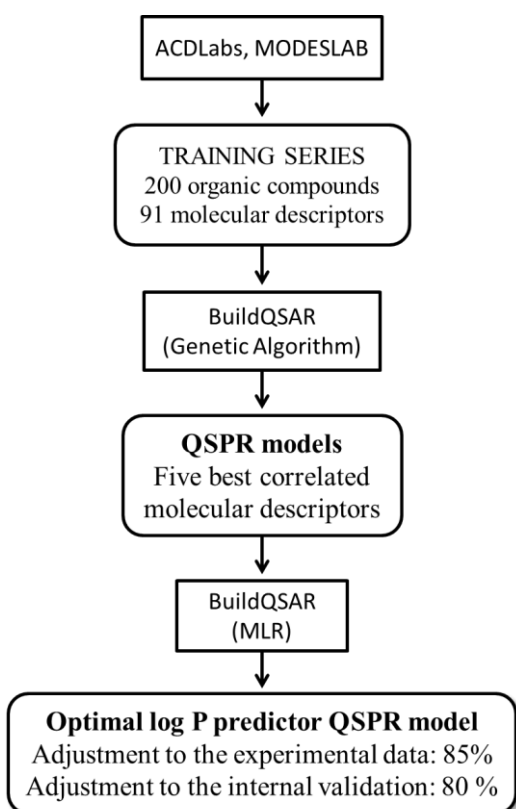


A QSPR model for the prediction of the partition coefficient of organic compounds of pharmaceutical interest

Tam Luong Minh ^a, Luis Alberto Torres Gómez (luistg@ifal.uh.cu) ^a,
Juan Carlos Polo Vega (polo@ifal.uh.cu) ^a, Laura Machín Galarza ^a

^a Department of Pharmacy, Institute of Pharmacy and Food Sciences, University of Havana, Cuba.

Graphical Abstract



Abstract.

The distribution coefficient ($\log P$) is an important molecular characteristic that allows us to estimate the lipophilicity of chemical compounds and predict how a drug will behave, fundamentally against the processes of absorption and excretion. The experimental determination of this and other properties of interest has several limitations, such as the high time invested and the consumption of considerable amounts of sample. In recent years, the development of new drugs has been supported by computational tools that allow a theoretical prediction of their properties from the information collected by their molecular descriptors, their design being much faster and cheaper. This paper shows the results of a structure-property relationship (QSPR) study aimed at finding a predictive mathematical model of the distribution coefficient of organic compounds of pharmaceutical interest. Through the computer programs ACDLabs (simplified molecular representations and calculation of $\log P$) and MODESLAB (calculation of molecular descriptors) a training series consisting of 200 compounds classified in ten pharmacological groups was formed. Using the BuildQSAR computer program, an optimal prediction model of $\log P$ was obtained, considering the five molecular descriptors that best correlated with this property as independent variables. The model obtained showed a percentage of adjustment to the experimental data of 85%, as well as a standard error of the estimate lower than the logarithmic unit. Its internal validation showed an adjustment percentage of 80%.

Introduction

The ability of a substance to cross biological membranes is one of the fundamental aspects among the pharmacokinetic characteristics of a drug, since it allows it to access the site of action and thus generate the desired effect. Among the many physicochemical properties that can affect this faculty, lipophilicity is the one of greatest interest, due to the direct relationship it presents with the distribution of drugs in the body.

The log P partition coefficient is considered as the main parameter to estimate the lipophilicity of chemical compounds and determine their pharmacokinetic properties¹. Like other physicochemical properties of chemical compounds of pharmaceutical interest, it can be estimated by several experimental techniques, but all of them present important disadvantages such as the high time invested, the consumption of large quantities of samples, difficulties for the quantification of extremely lipophilic or hydrophilic compounds and expensive equipment².

These limitations have motivated the development of methodologies such as the so-called *Quantitative Structure-Property Relationship (QSPR)*, based on obtaining theoretical models that numerically relate chemical structures to the properties of substances, through a set of computational techniques related to the design and virtual spatial visualization of molecules, calculation of molecular physicochemical properties, bioinformatics and statistics. As it exists only in a dematerialized virtual (*in silico*) environment of infrastructure needs, its application to the theoretical design of possible new drugs is much cheaper and faster²⁻⁷.

Based on these premises, the present work is aimed at obtaining a QSPR model capable of efficiently predicting log P values of organic compounds of pharmaceutical interest from their chemical structures.

Materials and Methods

Training Series. It was made up of 200 organic compounds of pharmaceutical interest, of wide structural variability, divided into ten pharmacological groups. Using the ACDLabs software version 10.04, the molecular structures of each of the compounds were represented, which were saved as abbreviated representation codes (SMILES). The log P values of the compounds included in the training series were determined using the same software.

Molecular descriptors. The calculation of the molecular descriptors was carried out with the help of the MODESLAB software version 1.5, after importing the SMILES generated by the ACDLabs software. The molecular graphs selected were bond distance (Std), dipole moment (Dip), hydrophobicity (Hyd), polarizability (Pol), atomic radius of van der Waals (Van) and atomic weight (Ato), due to the influence of these parameters in the distribution coefficient. By default, the number of atoms present in the molecule (Θ) is included.

Mathematical modeling. The Genetic Algorithm⁸ procedure of the BuildQSAR statistical software was used for the definition of the three best multiple linear regression models based on the five independent variables (molecular descriptors) that best correlated with the calculated values of log P. Using the Multiple Linear Regression procedure (MLR) of the same software, the best fit model for experimental data was obtained.

Validation. First, cross-validation type "leave one out" was carried out. An internal validation was also developed, where the training series was divided into 4 subgroups, each of which contained 25% of the compounds analyzed. To develop both processes the statistical software BuildQSAR was used.

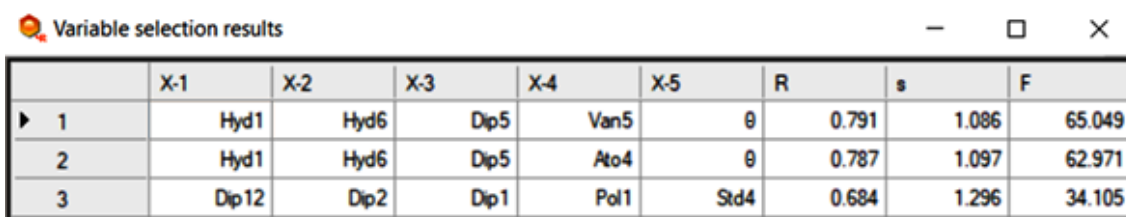
Results and Discussion

The predictive capacity of a model depends largely on the characteristics of the selected compounds for its preparation. The 200 organic compounds included in the training series represent ten pharmacological groups, corresponding to the polyfunctionality that distinguishes the molecules of pharmaceutical interest: anti-inflammatory, antibacterial, antifungal, anthelmintic, anticholinergic, antidepressant, antihypertensive, contraceptive, vasodilator and inhibitors of IMAO, with 20 compounds per group.

Using the TOPS-MODE approach of the MODESLAB software, a matrix formed by the spectral moments from μ^1 to μ^{15} of each of the molecular graphs indicated above (see Materials and Methods) was obtained, so 91 molecular descriptors were calculated for each compound included in the training series.

The inclusion of a large number of variables in a QSPR function may hamper its explanation, so it is recommended to use as few descriptors as possible, capable of providing a reasonable model, with adequate statistical quality and relatively easy to interpret ⁶. For this reason, the Genetic Algorithm procedure of the BuildQSAR statistical software was used to define the five molecular descriptors that best correlated with the calculated values of log P. Table I shows the statistical parameters corresponding to the models generated by this procedure:

Table I. Molecular descriptors (independent variables) and statistical parameters of the generated models. Genetic Algorithm Procedure, BuildQSAR



	X-1	X-2	X-3	X-4	X-5	R	s	F
▶ 1	Hyd1	Hyd6	Dip5	Van5	0	0.791	1.086	65.049
2	Hyd1	Hyd6	Dip5	Ato4	0	0.787	1.097	62.971
3	Dip12	Dip2	Dip1	Pol1	Std4	0.684	1.296	34.105

As can be seen in table I, model 1 is the one with the highest statistical quality, so it was selected for optimization. This model can be represented by equation 1:

$$\text{Log } P = + 0,8251 (\pm 0,1175) \text{Hyd}^1 - 0,0001 (\pm 0,0003) \text{Hyd}^6 - 0,0000 (\pm 0,0001) \text{Dip}^5 + 0,0003 (\pm 0,0003) \text{Van}^5 - 0,0543 (\pm 0,0604) \theta - 0,1780 (\pm 0,5780) \quad \text{eq. 1}$$

The optimization process of model 1 was carried out using the Multiple Linear Regression (MLR) procedure of the same BuildQSAR statistical software. Tables II and III show the main results of this analysis:

Table II. Non-standardized coefficients and t test of significance of the intercept and the slopes for model 1. MLR Procedure, BuildQSAR

MLR results

QSAR Model Fitting analysis Dataset Graphics Full analysis

Coefficient analysis:

	Coef.	Stdev	95% Conf.	t-ratio	p	Comment
▶ Constant	-0.1780	0.2890	0.5780	-0.6160	0.5386	Not signif.
Hyd1	0.8251	0.0587	0.1175	14.0494	0.0000	
Hyd6	-0.0001	0.0001	0.0003	-0.9471	0.3447	Not signif.
Dip5	0.0000	0.0001	0.0001	-0.1202	0.9044	Not signif.
Van5	0.0003	0.0001	0.0003	2.6034	0.0099	
θ	-0.0543	0.0302	0.0604	-1.8005	0.0733	Not signif.

Table III. Correlation matrix for model 1. MLR procedure, BuildQSAR

Correlation matrix:

	Hyd1	Hyd6	Dip5	Van5	θ
▶ Hyd1	1	0.077	0.096	0.131	0.308
Hyd6	0.077	1	0.766	0.940	0.745
Dip5	0.096	0.766	1	0.820	0.703
Van5	0.131	0.940	0.820	1	0.884
θ	0.308	0.745	0.703	0.884	1

The results of the t-test (see table II) indicate that neither the intercept nor the slopes of the Hyd⁶, Dip⁵ and θ descriptors contribute significantly to the value of the function, so the model can be expressed much more simplified according to:

$$\text{Log } P = 0.8251 (\pm 0.1175) \text{Hyd}^1 + 0.0003 (\pm 0.0003) \text{Van}^5 \quad \text{eq. 2}$$

As table III shows, the independent variables Hyd¹ and Van⁵ are weakly correlated with each other (correlation coefficient: 0,131), so the orthogonality principle is met. The statistical parameters corresponding to this model are:

$$n = 200; R = 0,787; s = 1,086; F = 160,748 (p < 0,0001)$$

When comparing these parameters with those of the model represented by equation 1 (see table 1), it is observed that the coefficient of determination (R) decreases slightly, the standard deviation (S) remains constant and the coefficient F of ANOVA increases significantly. Together, these values allow us to affirm that, in addition to being much simpler, the new model must exhibit a better fit to the experimental data.

The decrease in the value of R may be due to the existence of atypical cases (outliers) among the compounds included in the training series. Once the outliers were identified and eliminated, the repetition of the MLR analysis produced the following model:

$$\text{Log } P = 0.8006 (\pm 0.0839) \text{Hyd}^1 + 0.0002 (\pm 0.0000) \text{Van}^5 \quad \text{eq. 3}$$

The statistical parameters corresponding to the model represented by equation 3 are:

$$n = 190; R = 0,849; s = 0,900; F = 242,043 (p < 0,0001)$$

The coefficient of determination (R) indicates an 85% adjustment to the experimental data, the standard error of the estimate (S) is lower than the logarithmic unit and the increase in F expresses a significant linear relationship between the values of the molecular descriptors included in the model and the value of log P for the compounds that make up the training series. All this guarantees the superior statistical quality of the QSPR model represented by equation 3⁹.

Figure 1 shows the correlation between the log P values observed and those calculated by this model for the training series:

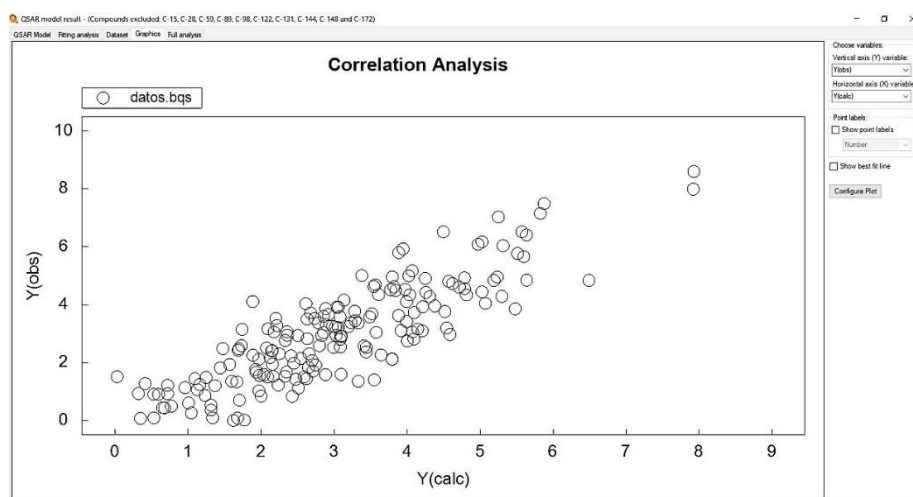


Fig. 1. Correlation between the log P values observed and those calculated by the model of equation 3

The high correlation between experimental and predicted values, together with its statistical quality and relative simplicity, suggest the relevance of its use in the prediction of log P for organic compounds of pharmaceutical interest.

Finally, the predictive model was validated. The prediction coefficient resulting from the cross-validation, $Q^2 = 0,712$ was similar to the value obtained previously for the complete series, $R^2 = 0,849$ (see statistical parameters of the model represented by equation 3), which reinforces the quality of the selected model. To perform the internal validation, the training group was divided into 4 subgroups (I-IV), each of which contained 25% of the compounds analyzed. Three of the four subgroups (I, II and III), (I, II and IV), (II, III and IV), (I, III and IV) were used as a training group, with the fourth subgroup (IV) remaining, (III), (I), (II) as a test group. Table IV summarizes the results of this process:

Table IV. Internal validation. MLR procedure, BuildQSAR

Training group	Test group	R (training)	R _{pred} (test)
I, II, III, IV	—	0,791	—
II, III, IV	I	0,822	0,815
I, III, IV	II	0,807	0,805
I, II, IV	III	0,804	0,795
I, II, III	IV	0,774	0,779
Average	—	0,801	0,799

As can be seen, for the training group the values of the correlation coefficient R of the complete group and the average of the four series are very similar (0,791 and 0,801 respectively), being very similar to the average value for the test group (0,799).

Conclusions

These results indicate that the optimal model selected (equation 3) is predictive and stable even when 25% of the compounds are eliminated.

References

1. Kujawski J, Bernard MK, Janusz A, Kuźma W. Prediction of log P – ALOGPS Application in Medicinal Chemistry Education. *J. Chem. Educ.* 2012; 89: 64-67.
2. Tetko I, Poda G, Ostermann C, Mannhold R. Accurate In Silico log P Predictions: One Can't Embrace the Unembraceable. *QSAR Comb. Sci.* 2009; 28: 845-849
3. Dearden JC. The use of topological indices in QSAR and QSPR modeling. Roy, K, ed. *Advances in QSAR Modeling, Challenges and Advances in Computational Chemistry and Physics*. Springer International Publishing. 2017.
4. Lokendra O, Rachana S, Mukta B. Modern drug design with advancement in QSAR: A review, *International journal of research in BioSciences*. 2013; 2 (1): 1-12.
5. Patel H, Noolvi M, Sharma P, Jaiswal V, Bansal S. Quantitative structure - activity relationship (QSAR) studies as strategic approach in drug discovery. *Med. Chem. Res.* 2014.
6. Polishchuk PG, Kuzmin VA, Artemenko AG, Muratov EN. Universal Approach for Structural Interpretation of QSAR/QSPR Models. *Mol. Inf.* 2013; 32: 843 – 853
7. Torres LA. Relationship studies structure property of pharmaceutical interest. QSPR methods applied to pharmaceutical analysis. *Spanish Academic Editorial*; 2016.
8. Rodríguez-Piñero, PT. Introduction to genetic algorithms and their applications. Rey Juan Carlos University. Publications Service. 2003.
9. Cherkasov A, Muratov EN, Fourches D, Varnek A, Baskin II. QSAR Modeling: Where have you been? Where are you going to? *J MedChem.* 2014; 57: 4977-10.