MDPI SCIforum

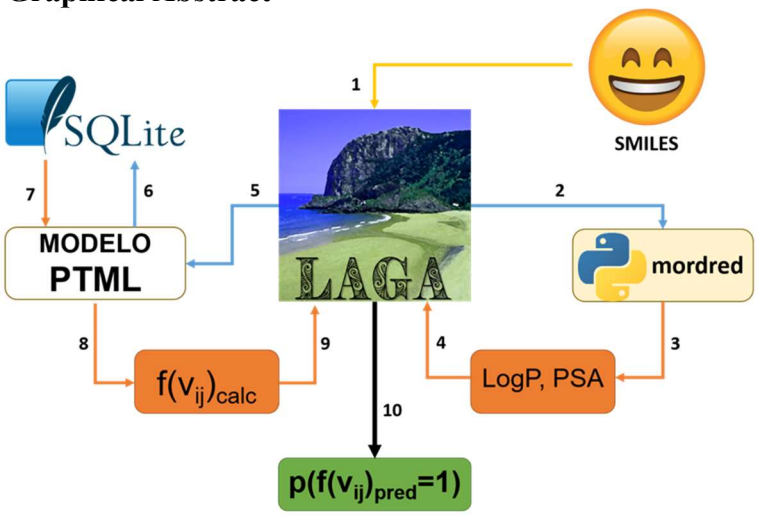
MOL2NET, International Conference Series on Multidisciplinary Sciences
Insert the title of the workshop here

LAGA: New software for new drug design using Perturbation Theory and Machine Learning techniques

Jon Collados [a], Harbil Bediaga [b]

[a] *Department of Organic Chemistry II, University of Basque Country UPV/EHU, 48940, Leioa, Spain.*
[b] *Department of Physical Chemistry, University of Basque Country UPV/EHU, 48940, Leioa, Spain.*

Graphical Abstract



Abstract.

The main objective of this project is the development of a useful computational tool for future preclinical trials. The implemented model will be a PTML-LFER model, for the prediction of the pharmacological activity of a certain molecule, or list of molecules, under multiple test conditions.

Introduction

In silico methods, which are based on computer simulations to obtain models capable of predicting whether a new compound is going to be active or not. Nowadays, more and more research groups are working on these types of methods that, with the improvement of the data calculation and processing capacity of computers, are increasingly accurate and effective.

One type of model is the Quantitative Structure-Activity Relationship (QSAR) models, which mathematically relate the molecular structure of the drug to be studied to its activity. These models are mathematical relationships, both linear and non-linear, of multiple variables. Within the QSAR models, there are the so-called Linear Free Energy Ratio (LFER) models. On the other hand, Perturbation-Theory Machine Learning (PTML) type QSAR models have also been developed that combine ideas taken from Perturbation Theory (PT) and Machine Learning (ML) methods. Finally, the PTML-LFER methods combine both approaches for the treatment of preclinical data with multiple assay specifications. In many cases, they require a long time to perform manual calculations, making it necessary to develop new specialized software that allows these models to be used quickly and easily.

Materials and Methods

Hansch-Fujita models are extratermodynamic methods that use thermodynamic and other input variables (structural, external, etc.) and that assume a linearity between the input variables and the $f(v_i)$ value of the response variable, which quantifies the efficiency v_i of a certain drug i under study.
Typically, models use molecular lipophilicity, which is quantified by the partition coefficient. Also, they often use variables such as molecular retroactivities (MR), acid constants in logarithmic terms (pKa) and other physicochemical parameters to quantify the molecular properties. An example of this type of Hansch-Fujita model is the following:

$$f(v_i) = a_1 \log P + a_2 pK_a + a_3 MR + b_2 (\log P)^2 + b_2 (pK_a)^2 + b_3 MR^2 + e_0$$

From the physical point of view of organic chemistry, Hansch-Fujita models are LFER models. The designation as LFER models comes from the use of Gibbs free energy dependent parameters, such as the equilibrium constants Ki, since the changes in the values of this potential during a process are proportional to the product of temperature and logarithm of said equilibrium constants.
These models are very useful for predicting the values of the aforementioned output function $f(v_i)$ that quantifies the efficiency of binding of a molecule (M_i) with a receptor. Taking into account that the input variables of the Hansch-Fujita models are sets of N molecular descriptors $\{D_k\}_{k=1}^{N}$, a general expression of the equation can be defined as follows:

$$f(v_i) = \sum_{k=1}^{N} a_k D_k + \sum_{k=1}^{N} b_k {D_k}^2 + e_0$$

Results and Discussion

Once the database has been built, a selection is made of inputs that will be used for parameterization, leaving the rest for model validation. With the STATISTICA software an LDA is performed to obtain the coefficients a_0, a_1, a_2 and a_3, establishing an a priori probability of classifying an input value as a value $f(v_{ij})_{pred} = 1$ of 0.8 ($\pi_1 = 0.8$). The equation finally obtained is the following:

$$\begin{aligned} f(v_{ij})_{calc} = & -5.939153 + 14.803823 \cdot f(v_{ij})_{ref} \\ & - 0.108663 \cdot \Delta D_1(\boldsymbol{c_j}) \\ & + 0.006869 \cdot \Delta D_2(\boldsymbol{c_j}) \end{aligned}$$

In the parameterization, the model presents a specificity Sp = 90.2%, sensitivity Sn = 70.6% and accuracy Ac = 87.7%. In the validation, it presents values of Sp = 90.1%, Sn = 71.4% and Ac = 87.8%.
In short, the PTML model receives as input variables the descriptors D_1 and D_2 and the condition vector c_j; calls the database, to which the fragment of the next Table 1 belongs, from which it obtains the reference value and the averages of the descriptors corresponding to c_j. For its implementation it is necessary the program to be developed that allows the management and obtaining of the input and output values respectively.

Table 1. Reference and average values of the descriptors corresponding to c_j.

c_0	c_1	c_2	c_3	c_4	$f(v_{ij})_{ref}$	$\langle D_1 \rangle(\boldsymbol{c_j})$	$\langle D_2 \rangle(\boldsymbol{c_j})$
Growth(%)	MCF7	MCF7	Homo sapiens	*Homo sapiens*	0.581	3.731	57.760
Growth(%)	SF-268	SF-268	Homo sapiens	*Homo sapiens*	0.516	3.731	57.760
IC_{30}(μM)	NCI-H460	NCI-H460	Homo sapiens	*Homo sapiens*	0.857	4.041	178.063
IC_{30}(μM)	DLD-1	DLD-1	Homo sapiens	*Homo sapiens*	0.857	4.041	178.063

Conclusions

The LAGA software was intended to develop a tool to implement the multi-condition PTML-LDA model, which predicts the success of preclinical tests using a discriminant function. To make this model accessible to users, a graphical interface has been developed in which to introduce all the input variables and in which, at the end, the results are presented in the form of probability of success. It must be taken into account that LAGA favors the prediction of positive results, since a high a priori probability has been chosen. In this way, possible favorable tests, but close to the set of unfavorable tests, will be classified in the first group, so that good candidates for in vitro tests will not be lost.
On the other hand, it is intended to include in LAGA a new PTML-LFER model based on the Linear Discriminant Analysis (LDA) algorithm, which allows combinations of conditions outside the database. With the inclusion of this model, LAGA will be a complete tool for predicting the success of preclinical trials.

References

[1] K. Roy, S. Kar, and R. N. Das, A Primer on QSAR/QSPR Modeling: Fundamental Concepts (Springer International Publishing, Cham, 2015).
[2] P. Abeijon, X. Garcia-Mera, O. Caamano, M. Yanez, E. Lopez-Castro, F. Romero-Duran, and H. Gonzalez-Diaz, Current Drug Targets 18, 511 (2017).
[3] H. Bediaga, S. Arrasate, and H. González-Díaz, ACS Combinatorial Science 20, 621 (2018).
[4] P. Ambure, A. K. Halder, H. G. Díaz, and M. N. D. S. Cordeiro, Journal of Chemical Information and Modeling (2019).
[5] R. Todeschini and V. Consonni, Handbook of Molecular Descriptors (Wiley-VCH, Weinheim, 2000).
[6] Medicinal, Chemistry and Biochemistry blog from the Sussex Drug Discovery Centre; www.sussexdrugdiscovery.wordpress.com/2015/02/03/not-all-logps-are-calculated-equal-clogp-and-other-short-stories (09/06/2019).
[7] H. Gonzalez-Diaz, S. Arrasate, A. Gomez-Sanjuan, N. Sotomayor, E. Lete, L. Besada-Porto, and J. Ruso, Current Topics in Medicinal Chemistry 13, 1713 (2013).
[8] D. Sunil and P. Kamath, Current Topics in Medicinal Chemistry 17, 959 (2017).

[9] Silberschatz, H. F. Korth, and S. Sudarshan, Database System Concepts (McGraw-Hill, New York, NY, 2006), Capítulos 1-3-4.
[10] F. Nelli, Python Data Analytics 49 (2018)
[11] SQLite3; www.sqlitetutorial.net/sqlite-python (07/06/2019).
[12] Open-source cheminformatics rdkit; http://www.rdkit.org (17/10/2018).
[13] H. Moriwaki, Y.-S. Tian, N. Kawashita, and T. Takagi, Journal of Cheminformatics 10, (2018).
[14] J. E. Grayson, Python and Tkinter Programming (Manning, Greenwich, CT, 2000).
[15] S. A. Wildman and G. M. Crippen, Journal of Chemical Information and Computer Sciences 39, 868 (1999).
[16] P. Ertl, B. Rohde, and P. Selzer, Journal of Medicinal Chemistry 43, 3714 (2000).
[17] T. Hill and P. Lewicki, Statistics: Methods and Applications; a Comprehensive Reference for Science, Industry, and Data Mining (StatSoft, Tulsa, OK, 2006).
[18] Penn State: Statistics Online Courses; www.newonlinecourses.science.psu.edu/stat505/lesson/10