ADVANTAGES OF A DESCRIPTORS SELECTION METHOD WITH MULTIPLE SOLUTIONS APPLIED TO QSPR MODELS FOR ADMET PROPERTIES PREDICTION

Palomba D1,2, Díaz MF2, Vazquez GE1*

1 Laboratory for Research and Development in Scientific Computing (LIDeCC), DCIC, UNS, Av Alem 1250, Bahía Blanca, Argentina
2 Planta Piloto de Ingeniería Química, CONICET-UNS, Cno. La Carrindanga km. 7, Bahía Blanca, Argentina

INTRODUCTION
QSPR (Quantitative structure-property relationship) methods are widely used for ADMET (Absorption, Distribution, Metabolism, Elimination, Toxicology) properties prediction in drug discovery. However, there are some critical issues, which are difficult to address. The usual problems are the selection of the proper number of descriptors that model a desired property, redundant descriptors, chance correlation, overfitting, and getting complex models with low predictive capacity. A method that avoids these problems and in addition provides several models would allow to interpret them using physicochemical principles, and to choose the most suitable one. This work uses a technique based on evolutionary algorithms (1), which provides different predictive models as output. Validity is analyzed for estimating human intestinal absorption (HIA).

MATERIALS AND METHODS
From the best reported models for HIA prediction in (1)* we selected 16 with cardinality ≤ 10 and MSE (mean square error) average over 50 runs < 0.135. Each model was assessed individually and, in addition, a global analysis of selected descriptors was performed.

RESULTS
The models had an average cardinality of 6 descriptors, a low figure which quickens the analysis. Out of descriptors we found, 75% belong to 7 families which are related to physicochemical properties connected to HIA:
- Molecular properties: LAI (Lipinski Alert Index), ALOGP and TPSA(Tot) were the most chosen; the last one gives information about absorbed fraction that is an usual way to express intestinal absorption.
- 1-D Descriptors (functional group counts): nArCONHR (number of secondary amides-aromatic) was the most chosen in this group, in accordance with the capacity to donate and accept H bonds from the amides, thus contributing to the aqueous solubility (essential to HIA); in the same way the aromatic group contributes to lipophilicity. The nROH (number of hydroxyl groups) descriptor was also prominent, which shows OH group’s importance and its ability to form H bonds.
- Atom-centered fragments: the favourite one was O-056 (alcohol), in line with was explained formerly. It is necessary to note that nROH was not chosen when O-056 was selected, thus avoiding variable intercorrelation.
- 3D Descriptors: 3D-Morse, WHIM, RDF and GETAWAY, which provide data related to molecular size and shape, both of which are properties associated to lipophilicity and thus to HIA. Depending of the needs of the study, the modeller will be able to choose among models with few descriptors and low prediction error (e.g. models #5 or #10 in (1)) or models with more descriptors which come from various families and optimal prediction error (e.g. model #9 in (1)). In all cases, these models improve those previously presented on literature.

CONCLUSIONS
Using a method which provides multiple statistically significant models allows the modeller to choose one based on different criteria, such as the quantity of descriptors used, the prediction error derived, and/or the type of descriptors in each model. In our particular study we demonstrate how to apply it for the prediction of HIA.

* Corresponding author. Tel +54 291 4595101(2626); e-mail: gev@cs.uns.edu.ar

* Supporting Information file sm008.xls - Worksheet Exp-HIA-NSGAII.
REFERENCES.