European Journal of Medicinal Chemistry 46 (2011) 218–228

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

Prediction of drug intestinal absorption by new linear and non-linear QSPR

Alan Talevi [a], Mohammad Goodarzi [b], Erlinda V. Ortiz [c], Pablo R. Duchowicz [b,*], Carolina L. Bellera [a], Guido Pesce [d], Eduardo A. Castro [b], Luis E. Bruno-Blanch [a]

[a] Medicinal Chemistry, Department of Biological Sciences, Faculty of Exact Sciences, National University of La Plata (UNLP), 47 y 115, 1900 La Plata, Argentina
[b] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (UNLP, CCT La Plata-CONICET), Diag. 113 y 64, C.C. 16, Suc.4, 1900 La Plata, Argentina
[c] Facultad de Tecnología y Ciencias Aplicadas, Universidad Nacional de Catamarca, Av. Maximio Victoria 55, 4700 Catamarca, Argentina
[d] Department of Pharmacology, National Institute of Medicines (INAME), National Administration of Food, Medicines and Medical Technology (ANMAT), Buenos Aires, Argentina

ARTICLE INFO

Article history:
Received 8 February 2010
Received in revised form
31 October 2010
Accepted 1 November 2010
Available online 9 November 2010

Keywords:
QSPR theory
Molecular descriptors
Replacement method
Drug intestinal absorption
Model's applicability domain
ADME properties

ABSTRACT

In order to minimize the high attrition rate that usually characterizes drug research and development projects, current medicinal chemists aim to characterize both pharmacological and ADME profiles at the beginning of drug R&D initiatives. Thus, the development of ADME High-Throughput Screening *in vitro* and *in silico* ADME models has become an important growing research area. Here we present new linear and non-linear predictive QSPR models to predict the human intestinal absorption rate, which are derived from a medium sized, balanced and diverse training set of organic compounds. The structure–property relationships so obtained involve only 4 molecular descriptors, and display an excellent ratio of number of cases to number of descriptors. Their adjustment of the training set data together with the performance achieved during the internal and external validation procedures are comparable to previously reported modeling efforts.

1. Introduction

The former paradigm of drug development has focused on the optimization of the molecule in order to gain potency and selectivity. As a consequence, drug development has been characterized by a high attrition rate, with about nine from ten drugs that have entered clinical trials which have not made it into the pharmaceutical market, mostly because of toxicity issues or the inability of the drug to reach its pharmacological target [1,2]. Moreover, modern Combinatorial Chemistry and High-Throughput Screening (HTS) technologies have tended to produce novel entities with poor ADME properties [3]. In order to reduce the rate of failure of drug development programs due to ADMET issues at late stages of the research (i.e. clinical trials), the modern paradigm of drug development has moved towards finding a balance between potency, bioavailability and safety from the very beginning of the project. This modern paradigm may be synthesized under the expression "to fail early is to fail cheap", and is implemented by including at early stages parallel ADMET filters to discard chemical entities with unfavorable pharmacokinetic and toxicity profiles.

The oral route is generally considered the most convenient route of administration because of production costs, stability of the drug and ease of administration and transport. For this reason, some of the most important physicochemical properties currently assessed at the beginning of a novel drug project are aqueous solubility, intestinal permeability and oral bioavailability. A possible way for assessing ADMET-related physicochemical properties is to rely on *in vitro* assays. For instance, Caco-2 or MDCK cells and parallel artificial membrane permeation assays (PAMPA) [4], coupled with high-throughput liquid chromatography–mass spectrometry [5], and liquid chromatography using special stationary or mobile phases (e.g. the immobilized artificial membrane – IAM – technique) [6], have proved successful to simulate transport processes. Although there have been significant innovations in the area of high throughput *in vitro* ADME screening in the last few years [7,9], *in vitro* assays are laborious, expensive, time consuming and demand certain drug quantities, usually more than what is produced in a standard combinatorial library synthesis. Thus, *in vitro* ADME assays are still not entirely compatible with HTS technologies [8]: compounds are currently synthesized and pharmacologically screened much faster than the speed by which experimental ADME studies can be carried out, and ADME studies have become a bottleneck during modern drug discovery efforts. *In silico* models thus constitute an inexpensive and faster option to apply at early

* Corresponding author. Fax: +54 221 425 4642.
E-mail addresses: prduchowicz@yahoo.com.ar, pabloducho@gmail.com (P.R. Duchowicz).

0223-5234/$ – see front matter © 2010 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2010.11.005

stages of the screening process, and are completely compatible with HTS. These can be also applied to estimate ADME properties of a compound *prior* to synthesis, providing a reference frame to guide an 'informed synthesis' and improving the chances of landing in better chemical space [8–10]. When a large number of compounds have been synthesized, *in silico* ADME models can be build from experimental data of a representative subset of the compounds and then used to predict the attributes of the remainder [8–10]. *In silico* models of ADMET-related properties may rely on either theoretical or experimental descriptors to establish a quantitative relationship between the target property and the molecular structure of a set of compounds. Quantitative Structure Property–Activity Relationships (QSPR-QSAR) may be derived through a wide range of linear or non-linear models [11,12].

As part of our ongoing research to build QSAR models of ADME properties to assist drug development projects [13–16], we now develop both linear and non-linear models to estimate Human Intestinal Absorption rate (%HIA). Tables 1 and 2 include a summary of some of the most notable efforts to model Human Intestinal Absorption rate (%HIA), which have been reported in the last 10 years. Table 1 shows a summary of reported quantitative models, aimed to predict exact %HIA values; Table 2 presents a review of classificatory models, aimed to classify a given drug into one of two or more %HIA categories. Only the best model reported in each article, in terms of squared correlation coefficient (R^2) or root mean squared error (*RMSE*) in the training set in Table 1, or overall percentage of good classifications or *RMSE* in Table 2, is included in each of the tables. Note that many of the training and/or test sets of these models are clearly heavily biased towards highly permeable drugs [17–27,29–33], while some models present very low cases to descriptors ratio *N*/*D* (with the resultant risk of overfitting) or have been derived from scarce training sets that limit the applicability domain of the models. Moreover, some of them are based on experimental descriptors, jeopardizing their applicability in HTS campaigns (e.g. the model presented in Deconinck et al. [26] requires the determination of the retention time of each compound in three different HPLC systems). In some cases 3D descriptors are included in the models without a systematic conformational analysis of the structures of the training set [26]. Another limitation of these models is that they have been derived

Table 1
Summary of modeling efforts towards quantitative prediction of %HIA in the last ten years. 'In house or literature' indicates whether the %HIA data have been measured either by the authors or obtained from literature. Only results of the external validation have been summarized in the table, although in some cases the authors performed internal cross-validation procedures.

Authors	*N*	Data source	Descriptor types	Modeling technique	*N*/*d*	R^2 training set	*RMSE* training set	R^2 test set	*RMSE* test set
Wessel et al. [17]	76	Literature	1D–3D theoretical descriptors (functional group counts, topological, geometrical and electronic descriptors)	Combination of GA and ANN	12.7	–	9.4[b]	–	16.0[b]
Norinder et al. [18]	20	Literature	MolSurf theoretical physicochemical descriptors related to oral bioavailability	Principal Components Analysis plus PLS	6.7[c]	0.92	0.49[a]	–	–
Österberg and Norinder [19]	20	Literature	Number of H bond acceptor nitrogens and oxygens, number of H bond donors, log P	PLS	5.0	0.81	1.48[a]	–	–
Agatonovic-Kustrin et al. [20]	76	Literature	0D–3D theoretical descriptors (constitutional, topological, geometrical, quantum-chemical descriptors)	Combination of GA and ANN	5.1	–	0.59 [a]	0.802	0.42 [a]
Klopman et al. [21]	417	Literature	1D theoretical descriptors plus 6 basic parameters related to oral absorption	Combination of Group Contribution and CASE	11.3	0.79	12.3[b]	0.79	12.3[b]
Abraham et al. [22]	127	In house	Abraham's solvation parameters	MLR	25.4	0.80	0.29[a]	–	–
Niwa [23]	67	Literature	0D and 1D theoretical descriptors (constitutional descriptors and count of functional groups and atom types)	Neural networks (ANN)	9.6	–	6.5[b]	–	22.8[b]
Sun [24]	169	In house	Theoretical (Atom types)	PLS	56.3[c]	0.92	–	–	–
Yen et al. [25]	52	Literature	Experimental (chromatographic) descriptor (IAM chromatography) plus Molecular Modeling Pro topological, geometrical and physicochemical descriptors	Multiple Linear Regression (MLR)	5.8	0.68	–	–	–
Deconinck et al. [26]	67	Literature	Experimental (chromatographic) descriptors combined with Dragon and Hyperchem theoretical 1D–3D descriptors plus one of Abraham's solvation parameters.	Non-linear, MARS	9.6	0.93	–	–	–
Yan et al. [27]	380	Literature	Adriana Code and Cerius2 0D–2D theoretical descriptors (constitutional, functional group counts, topological and physicochemical descriptors)	Combination of Genetic Algorithms (GA), Partial Least Squares (PLS) and Support Vector Machine (SVM)	42.2	0.66	12.5[b]	0.77	16.0[b]
Reynolds et al. [28]	567	Literature	ADME Boxes and Algorithm Builder; log P experimental values were used when available	Non-Linear Least Squares (NLS)	43.6	0.93	9.5[b]	–	0.35[a,d] 0.45[a,e]
Guerra et al. [29]	37	Literature	Codes 2D (topological) theoretical descriptors	ANN	12.3	0.93	8.0[b]	–	–
Talevi et al. [this work]	120	Literature	0D–3D Dragon theoretical descriptors	Linear (MLR) and non-linear	30.0	0.80	0.18[a]	0.66	0.21[a]

N refers to the number of compounds in the training set. *N*/*d* refers to the ratio between the number of cases in the training set and the number of independent variables included in the best model.

[a] Expressed in log units (the dependent variable is some log transformation of an experimental variable linked to intestinal absorption, usually %HIA).

[b] Expressed in % units.

[c] The ratio is calculated considering the number of PLS latent variables as the number of independent variables, though these latent variables are combinations of a higher number of descriptors.

[d] *RMSE* reported when logarithm of absorption rate constant is taken as dependent variable.

[e] *RMSE* reported when logarithm of human permeability coefficients taken as dependent variable.