Available online at www.sciencedirect.com
ScienceDirect
European Journal of Medicinal Chemistry 43 (2008) 2442–2452
EUROPEAN JOURNAL OF MEDICINAL CHEMISTRY
http://www.elsevier.com/locate/ejmech

Original article

Explorations into modeling human oral bioavailability

Zhi Wang [a], Aixia Yan [a,*], Qipeng Yuan [a], Johann Gasteiger [b,c]

[a] State Key Laboratory of Chemical Resource Engineering, Department of Pharmaceutical Engineering, P.O. Box 53, Beijing University of Chemical Technology, 15 BeiSanHuan East Road, Beijing 100029, PR China
[b] Computer-Chemie-Centrum and Institut für Organische Chemie, Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany
[c] Molecular Networks GmbH, Henkestrasse 91, D-91052 Erlangen, Germany

Received 6 March 2008; received in revised form 5 May 2008; accepted 15 May 2008
Available online 28 May 2008

Abstract

Explorations into modeling human oral bioavailability started with a whole dataset of 772 drug compounds. First, training set and test set were chosen based on Kohonen's self-organizing Neural Network (KohNN). Then, a quantitative model of the whole dataset was built using multiple linear regression (MLR) analysis. This model had limited predictability emphasizing that a variety of pharmacokinetic factors influence human oral bioavailability. In order to explore whether better models can be built when the compounds share some ADME properties, four subsets were chosen from the whole dataset to build quantitative models and better models were obtained by MLR analysis. These studies show that, indeed, good models for predicting human oral bioavailability can be obtained from datasets sharing certain pharmacokinetic properties.

Keywords: Quantitative structure–activity relationships (QSAR); Bioavailability; Multiple linear regression (MLR); Leave-one-out cross-validation; Kohonen's self-organizing Neural Network (KohNN)

1. Introduction

In the design of new drugs, one of the most important considerations is to increase the oral bioavailability of a drug candidate. It is reported that 95% of lead compounds fail in the developmental stages, and 50% of these failures were shown to be due to unfavorable absorption, distribution, metabolism, and excretion (ADME) properties [1,2].

Among pharmacokinetic properties, an inadequate bioavailability is one of the main reasons that many promising drug candidates fail before further development. Bioavailability represents the percentage of an oral dose which is able to produce a pharmacological activity, in other words, the fraction of the oral dose that reaches the arterial blood in an active form. Oral bioavailability is related to several factors, such as gastrointestinal transition and absorption, intestinal membrane permeation, and intestinal/hepatic first-pass metabolism. Moreover, in the access of absorption, many authors have suggested that gut wall CYP3A4 and P-glycoprotein act in a concerted manner to control the absorption of their substrates [3–8]. The obvious method to maximize oral absorption would be to design the characteristics of a molecule which make it a substrate of P-glycoprotein and CYP3A4 [9].

Predicting oral bioavailability in silico may be guided by Lipinski's 'Rule of Five', which could be used to predict the absorption and permeability of drug molecules qualitatively [10]. In 1994, Hirono and colleagues [11] reported a study on the quantitative structure–bioavailability relationship for 188 noncongeneric organic drugs; the drugs were separated into three groups: nonaromatics, aromatics, and heteroaromatics, then based on each group a quantitative model was developed. In 2000, Yoshida and colleagues published a classification model for human oral bioavailability [12]. This model can get a correct rate of classification of 60% for the test group using three distribution related descriptors and 15 structure

* Corresponding author. Tel.: +86 10 64421335; fax: +86 10 64416428.
E-mail addresses: aixia_yan@yahoo.com, yanax@mail.buct.edu.cn (A. Yan).

0223-5234/$ - see front matter © 2008 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2008.05.017

descriptors which were considered to be related to some metabolic processes. The descriptors were chosen by analyzing bioavailability in relation to physicochemical and structural factors by the ORMUCS (ordered multicategorical classification method using the simplex technique) [12].

In subsequent work, other researchers have introduced rules-of-thumb which can increase the chances of drug compounds being well absorbed. In 2002, Veber and colleagues reported studies on rat bioavailability data for 1100 drug candidates [13]. It was found that the drug molecules having fewer than 10 rotatable bonds and less than 140 $Å^2$ PSA (polar surface area) (or an H-bond count less than 12) usually showed more than 20% rat oral bioavailability. In 2004, Lu and colleagues investigated the relationship between the number of rotatable bonds and PSA for rat oral bioavailability using 434 molecules [14]. Compared to Veber's work [13], Lu reported that the prediction results were dependent on the calculation methods [14].

In 2007, Hou and colleagues collected a dataset of 773 compounds with experimental human oral bioavailability values [15]. Then, Veber's rules [13] were used for the entire dataset to see if these rules could be applied for the prediction of human oral bioavailability. Afterwards, the correlations between several important molecular descriptors and human oral bioavailability were examined. They conjectured that there are no simple rules based on molecular descriptors that can be used to predict human oral bioavailability truly well compared to the rules based on analyzing rat oral bioavailability data [15].

It is clear that powerful descriptors related to carrier-mediated transport and first-pass metabolism are needed for building a useful prediction model for human oral bioavailability. However, it seems that until now, for a diverse dataset of drug compounds such as the above 773 drug compounds, no set of calculated descriptors was able to represent the complicated relationship between human oral bioavailability and structure.

Hence, in this study, we intended to explore the relationships between human oral bioavailability of drugs and their structures, by building quantitative models from two sides: (1) by including experimental HIA (human intestinal absorption) value as an input descriptor as this property is considered to be related to absorption and metabolism processes; (2) by building individual quantitative models for three different groups of drug compounds, which exhibit similar structures or pharmacological activities.

In this work five different sets of compounds were investigated. (1) Initially, the whole dataset of Hou et al. [15], including 772 drug compounds was studied in Models 1. It was investigated whether a quantitative prediction model could be built based on the 141 available descriptors. Then, four other prediction models were constructed for four different groups. (2) In Models 2, 161 drug compounds, for which experimental HIA values were available, were studied using the HIA value as a descriptor to increase the prediction power of the model. (3) Models 3 and Models 4 were built for specific types of compounds, i.e., for 51 sulfonamide and for 29 β-lactam drug compounds, respectively. (4) Models 5 were constructed based on 58 central nervous system (CNS) drug compounds, which were considered as having similar pharmacological activity. CNS drugs which can restrain or excite the action of central nervous system are considered to have similar absorption and metabolism profiles.

To compare the performance of descriptors from ADRIANA.Code [16] and from those of Cerius^2 [17], individual models were built with the descriptors from ADRIANA.Code, from Cerius^2 and with a combination of them.

2. Methods

2.1. Dataset of human oral bioavailability

In this work, the human oral bioavailability dataset of 772 compounds was taken from http://modem.ucsd.edu/adme, which was collected from 185 literatures, by Hou and colleagues [15]. Experimental human oral bioavailability data is defined as the fraction of the oral dose that reaches the arterial blood in an active form. The original dataset included 773 compounds, however, one compound in this dataset lacked an experimental human oral bioavailability value and was therefore removed. All the structures of the dataset (especially chirality) were checked through Cambridge Chemfinder Database [18] and Chemical Information Specialized Information Services of National Library of Medicine of US [19].

Experimental HIA values for 161 compounds used in the subset for Models 2 were also taken from a dataset collected by Hou et al. [20].

2.2. Molecular descriptors

Two program packages ADRIANA.Code [16] and Cerius^2 [17] were used to calculate the 141 molecular descriptors of this study.

2.2.1. Descriptors calculated by ADRIANA.Code

Seventy-four descriptors including molecular weight (MW), topological polar surface area (TPSA), mean molecular polarizability (MMP) and 2D property-weighted autocorrelation were calculated using ADRIANA.Code [16].

TPSA was calculated using the parameters originally reported by Ertl and colleagues [21]. Mean molecular polarizability (MMP) can be estimated from additive contributions of atoms as shown by Miller [22]. The 2D property-weighted autocorrelation uses the molecule's 2D structure and atom pair properties as a basis for obtaining vectorial molecular descriptors [23,24]. The atom pair properties are summed up for certain topological distances which count the number of bonds on the shortest path between two atoms. Thereby a single value for each topological distance is derived that is one entry in the resulting 2D autocorrelation vector. The 2D molecular autocorrelation vectors [25] were calculated based on the following seven atomic properties: σ charge (SigChg) [26,27], π charge (PiChg), total charges (TotChg), σ electronegativity