Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Molecular Graphics and Modelling 24 (2006) 383–395
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Quantitative structure–pharmacokinetic relationships for drug clearance by using statistical learning methods

C.W. Yap [a], Z.R. Li [a,b], Y.Z. Chen [a,c,*]

[a] *Department of Computational Science, National University of Singapore, Blk SOC1, Level 7, 3 Science Drive 2, Singapore 117543, Singapore*
[b] *College of Chemistry, Sichuan University, Chengdu 610064, PR China*
[c] *Shanghai Center for Bioinformation Technology, Shanghai 201203, PR China*

Received 27 July 2005; accepted 4 October 2005
Available online 14 November 2005

Abstract

Quantitative structure–pharmacokinetic relationships (QSPkR) have increasingly been used for the prediction of the pharmacokinetic properties of drug leads. Several QSPkR models have been developed to predict the total clearance (CL_{tot}) of a compound. These models give good prediction accuracy but they are primarily based on a limited number of related compounds which are significantly lesser in number and diversity than the 503 compounds with known CL_{tot} described in the literature. It is desirable to examine whether these and other statistical learning methods can be used for predicting the CL_{tot} of a more diverse set of compounds. In this work, three statistical learning methods, general regression neural network (GRNN), support vector regression (SVR) and k-nearest neighbour (KNN) were explored for modeling the CL_{tot} of all of the 503 known compounds. Six different sets of molecular descriptors, DS-MIXED, DS-3DMoRSE, DS-ATS, DS-GETAWAY, DS-RDF and DS-WHIM, were evaluated for their usefulness in the prediction of CL_{tot}. GRNN-, SVR- and KNN-developed models have average-fold errors in the range of 1.63 to 1.96, 1.66–1.95 and 1.90–2.23, respectively. For the best GRNN-, SVR- and KNN-developed models, the percentage of compounds with predicted CL_{tot} within two-fold error of actual values are in the range of 61.9–74.3% and are comparable or slightly better than those of earlier studies. QSPkR models developed by using DS-MIXED, which is a collection of constitutional, geometrical, topological and electrotopological descriptors, generally give better prediction accuracies than those developed by using other descriptor sets. These results suggest that GRNN, SVR, and their consensus model are potentially useful for predicting QSPkR properties of drug leads.

Keywords: Clearance; Consensus models; Computational ADME; General regression neural network; k-nearest neighbour; QSAR; Support vector regression

1. Introduction

Drug clearance is measured by a quantity, total clearance (CL_{tot}), which is a proportionality constant describing the relationship between a substance's rate of transfer, in amount per unit time, and its concentration, in an appropriate reference fluid [1]. Drug clearance occurs by perfusion of blood to the organs of extraction, which are generally the liver and the kidney [2]. The CL_{tot} value of a drug is an important pharmacokinetic parameter because it is directly related to bioavailability and drug elimination and can be used to determine the dosing rate and steady-state concentration of a drug [3]. Thus, it is important to predict the CL_{tot} value of drug leads during drug discovery so that compounds with acceptable metabolic stability can be identified and those with poor bioavailability can be eliminated.

Traditionally, the CL_{tot} value of a drug candidate is obtained via in vivo and in vitro studies [4–7], which tends to be time-consuming and costly. Therefore, an in silico method, quantitative structure–pharmacokinetic relationship (QSPkR) modeling, has recently been explored for predicting the CL_{tot} value of drug candidates [8–12] in an effort to eliminate undesirable agents in a fast and cost-effective manner. An initial PLS study conducted by Karalis et al. [8] using 272 structurally unrelated compounds failed to find any correlation between CL_{tot} and a large variety of molecular descriptors used in that study. Karalis et al. [9] then developed a partial least square (PLS) model and non-linear regression model for CL_{tot} by using 23 cephalosporins. The r^2 and q^2 values of the PLS-developed model are 0.775 and 0.731, while the r^2 value of the

* Corresponding author. Tel.: +65 6874 6877; fax: +65 6774 6756.
E-mail address: yzchen@cz3.nus.edu.sg (Y.Z. Chen).

1093-3263/$ – see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmgm.2005.10.004

non-linear regression model is 0.804. These two studies suggest that multiple mechanisms may be involved in CL_{tot} and thus linear methods may not always be suitable for constructing QSPkR models for CL_{tot}. Another study for the prediction of CL_{tot} was done by Turner et al. [10] who used artificial neural network (ANN), which gives a r^2 value of 0.982 for a training set of 16 cephalosporins and a r^2 value of 0.998 for a validation set of four cephalosporins. Subsequently, Turner et al. [11] used a larger training set of 56 compounds to develop an ANN-based QSPkR model, which gives a r^2 value of 0.731 for a validation set of six compounds. These results suggest that non-linear methods may be useful for developing models for CL_{tot} prediction of structurally unrelated compounds. Two QSPkR models for CL_{tot} were developed by Ng et al. [12] by using k-nearest neighbour (KNN) and PLS. The KNN-developed QSPkR model gives a q^2 value of 0.77 for a training set of 38 antimicrobial agents and a r^2 value of 0.94 for a validation set of six antimicrobial agents. There are 68% of the 44 compounds having predicted CL_{tot} within two-fold of actual values. For the PLS-developed QSPkR model, there are only 50% of the 44 compounds having predicted CL_{tot} within two-fold of actual values and the q^2 value of this model is 0.09 for the training set and its r^2 value is 0.35 for the validation set. These results are consistent with the study of Turner [11] and further confirm the usefulness of non-linear methods for developing QSPkR models for predicting CL_{tot}. All of the previous QSPkR models for predicting CL_{tot} have primarily been developed and tested by using a relatively small number of compounds (<70), which is significantly smaller in number and diversity than the number of compounds with known CL_{tot} data. Thus, it is of interest to evaluate the prediction capabilities of QSPkR models that are developed by using much larger and more diverse datasets.

Recently, non-linear statistical learning methods such as KNN [12], general regression neural network (GRNN) [13] and support vector regression (SVR) [14] have shown promising potential for predicting compounds of various pharmacokinetic and pharmacodynamic properties. GRNN has been explored for QSPkR modeling of drug distribution properties [13] and human intestinal absorption [15]. SVR has been applied to blood brain barrier penetration [14] and human intestinal absorption [14]. KNN has been used for the prediction of CL_{tot} [12] as well as metabolic stability of drug candidates [16]. It is of interest to evaluate the usefulness of these methods and other non-linear statistical learning methods for the prediction of CL_{tot}.

This work is intended to evaluate the capability of several statistical learning methods for predicting CL_{tot} by using 503 compounds found from a comprehensive literature search, which is substantially larger in number and more diverse in structure than those used in earlier studies. The methods used include GRNN, SVR and KNN. Different descriptor sets, which encode different combination of the structural and physiochemical properties of a compound, were also compared for their usefulness for constructing QSPkR models to predict CL_{tot}. Consensus modeling strategy has been introduced for developing prediction systems based on multiple models [17,18]. In this work, this strategy was also applied to the development of consensus QSPkR (cQSPkR) models for the prediction of CL_{tot} by using QSPkR models generated from different statistical learning methods.

2. Method

2.1. Dataset

Compounds with known human CL_{tot} values were selected from several sources including *Micromedex* [19], a classic pharmacology textbook [20] and a number of publications [5,6,10–12,21,22]. In order to ensure that experimental variations in determining CL_{tot} do not significantly affect the quality of our data sets, only CL_{tot} values obtained from healthy adult males and from intravenous administration were used for constructing the dataset. In addition, a number of compounds were excluded because they are known to possess certain molecular characteristics which do not permit reliable calculations of the molecular descriptors used in this study [8]. Examples of these compounds are quarternary ammonium compounds, molecules with complex chemical structures like amphotericin-B, aminoglycosides, vancomycin, and compounds containing one or more metal atoms. A total of 503 compounds were selected from this process and these were used as the dataset for this work. The CL_{tot} value for each of these compounds was log-transformed ($\log CL_{tot}$) to normalize the data and to reduce unequal error variances [23].

Representative training set and validation set were constructed from our dataset according to their distribution in the chemical space by using a method used in several studies [24–26]. Here, chemical space is defined by the structural and chemical descriptors used to represent a compound. Each compound occupies a particular location in this chemical space. All possible pairs of these compounds were generated and a similarity score was computed for each pair. These pairs were then ranked in terms of their similarity scores, based on which compounds of similar structural and chemical features were evenly assigned into separate datasets. For those compounds without enough structurally and chemically similar counterparts, they were assigned to the training set. After the dataset separation procedure, the training set and validation set contain 398 and 105 compounds, respectively. The list of compounds with their CL_{tot} values and their allocation into training and validation sets is provided in the supplementary material.

Prediction capability of QSPkR models is known to be strongly affected by the diversity of samples used in the training set [27,28]. Independent validation sets have frequently been used for evaluating the predictive performance of these QSPkR models, and these need also to be sufficiently diverse and representative of the samples studied in order to accurately assess the capabilities of the QSPkR models [27,28]. The diversity of a dataset can be estimated by a diversity index (DI) which is the average value of the similarity between all of the pairs of compounds in that dataset [29]:

$$\mathrm{DI} = \frac{\sum_{i=1}^{n} \sum_{j=1, i \neq j}^{n} \mathrm{sim}(i, j)}{n(n-1)} \quad (1)$$