Contents lists available at ScienceDirect

Journal of Chromatography A

journal homepage: www.elsevier.com/locate/chroma

Evaluating the performances of quantitative structure-retention relationship models with different sets of molecular descriptors and databases for high-performance liquid chromatography predictions

Chunlei Wang[a], Michael J. Skibic[b], Richard E. Higgs[b], Ian A. Watson[b], Hai Bui[b], Jibo Wang[b], Jose M. Cintron[b,*]

[a] *Department of Chemistry and Biochemistry, The University of Texas at Arlington, Arlington, TX 76019, USA*
[b] *Lilly Research Laboratories, Eli Lilly and Company, Indianapolis, IN 46285, USA*

ARTICLE INFO

Article history:
Received 20 March 2009
Received in revised form 17 April 2009
Accepted 21 April 2009
Available online 3 May 2009

Keywords:
Quantitative structure-retention relationships
Molecular descriptors
Log *D*
Partial least squares
Random forests
Root-mean-square error of prediction
Database selection

ABSTRACT

Quantitative structure-retention relationship (QSRR) models were studied for two databases: one with 151 compounds and the other with 1719 compounds. In both cases, the three modeling methods employed (multiple linear regression, partial least squares, and random forests) provided similar prediction results with regard to root-mean-square error of prediction. The reversed-phase retention related seven molecular descriptors provided better models for the smaller dataset, while the use of over 2000 molecular descriptors generated better models for the larger dataset. The QSRR models were then validated with a mixture of an active pharmaceutical ingredient and its four process/degradation impurities. Finally, classification of compounds based on similar log *D* profiles before QSRR modeling improved chromatographic predictability for the models used. The results showed that database composition had a desirable effect on prediction accuracy for certain input molecules.

1. Introduction

With ever more diverse stationary phases available, it becomes more challenging for analysts to select a suitable chromatographic system or even a starting point for method development. This is especially the case in HPLC method development for the pharmaceutical industry, where methods are needed to control many related impurities at an ever faster pace. The employment of smaller diameter particles and higher pressure HPLC systems are currently the popular approaches to improve efficiencies in method development. In addition, column characterization and classification methods are well studied [1–5] and used to guide analysts to choose either similar or dissimilar stationary phases depending on development needs. However, despite the aforementioned tools, the inevitable and sometimes time and resource intensive task of screening multiple HPLC conditions is necessary to achieve an adequate methodology. Thus, in an effort to minimize the number of experiments needed, quantitative structure-retention relationship (QSRR) methods are intensively studied for chromatographic retention predictions [6–8]. Ideally, by comparison of predicted retention results across available chromatographic conditions, one would be able to pick the best condition as a starting point for method development.

In QSRRs, as its name carries, retention is modeled as a function of molecular descriptors, which are the numeric transforms of molecular structures. The descriptors can be either experimentally determined or theoretically computed. In the former case, QSRRs are more widely known as the linear solvation energy relationships (LSERs) [9–13]. When building LSER models, only those molecules having experimental molecular descriptors available can be used. On the other hand, molecular descriptors of any molecule can be computed for QSRR modeling, and there are many thousands of descriptors defined in the literature [14]. Despite the large number of descriptors available and numerous studies carried out [6–8], "...a suitable translation, which would reveal the properties of compounds encoded into their structure in a reliable manner, is still lacking" [8]. Multiple linear regression (MLR) was the first statistical tool widely used for QSRR models [6,15]. With the introduction of additional molecular descriptors, many new chemometric modeling tools have been applied to QSRR modeling, such as genetic algorithms on MLR (GA-MLR) [16,17], partial least squares regression [18–21], artificial neural networks [13,21–23], support vector

[*] Corresponding author.
E-mail address: cintron_jose_m@lilly.com (J.M. Cintron).

0021-9673/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2009.04.064

machines [24–27], classification and regression trees [18,28,29], and random forests [18].

Successful QSRR models for specific classes of compounds have been constructed for gas chromatography, HPLC, supercritical fluid chromatography, and micellar electrokinetic chromatography [6–8,30,31]. With respect to general method development, Schefzick et al. built QSRR models from 62 structurally diverse compounds for 75 chromatographic conditions using GA-MLR models selecting from over 1000 molecular descriptors [16]. Baczek et al. demonstrated the combination of QSRR modeling for two different gradient slopes and linear solvent strength theory to predict the optimized gradient profiles for unknown molecules using 15 training molecules and 3 molecular descriptors [22,32–34].

In this study, the performance of QSRR models were compared using two different sets of molecular descriptors, three modeling methods, and two databases differing significantly in size. A real world example was then used to validate the models constructed using two sets of chromatographic conditions, and the importance of database composition was demonstrated. While it would be better to disclose the compounds used in this study, it was not possible due to proprietary reasons. However, to demonstrate the diversity of the datasets, the distribution of nearest neighbor distances was included to give the readers an understanding of the scope of the investigation.

2. Experimental

2.1. Dataset

Two datasets were used for this study: dataset 1 with 151 pharmaceutical compounds; dataset 2 with 1719 compounds. The chromatographic data of dataset 1 compounds were collected using an Agilent 1200SL HPLC system (Santo Clara, CA, USA), equipped with an autosampler, a micro vacuum degasser, a binary pump, a column thermostat, a diode array detector, and ChemStation software for data processing. Dataset 2 compounds were analyzed on an Agilent 1100 HPLC system with a similar configuration to the Agilent 1200SL system described above. Both acidic and basic gradient chromatographic conditions were studied on an X-bridge C18 75 × 4.6 mm, 5 μm (Waters, Milford, MA, USA) column for dataset 1, whereas the basic condition was studied for dataset 2. The aqueous phases (solvent A) were 0.1% aqueous trifluoroacetic acid (TFA) (prepared with ≥99.5% pure TFA from Thermo Scientific, Rockford, IL, USA) solution and 10 mM ammonium bicarbonate pH 10 buffer (Mays Chemical Company, Indianapolis, IN, USA) for acidic and basic gradients respectively. Neat acetonitrile (Omni Solv, Gibbstown, NJ, USA) was used as the organic solvent (solvent B) in both conditions. In both acidic and basic gradients, %B was increased linearly from 5 to 100% in 10 min. All HPLC experiments were run at 2 mL/min, at room temperature (∼22 °C).

2.2. Calculation of molecular descriptors

A total of 2352 numeric molecular descriptors were calculated for each specific chemical structure input in the SMILES format by using in-house software and commercial software. These descriptors include log *P* from CLOGP software (BioByte, Claremont, CA, USA), Lilly internal implementation and extension of published two-dimensional (2D) descriptors such as estate-related descriptors [35], Molconn [36], MACCS keys [37], CATS [38], Ghose-Crippen [39,40], three-dimensional (3D) descriptors such as CPSA [41], CoMMA [42] and MoRSE [43], and other 2D and 3D pharmacophore shape related descriptors. The 3D structure generator CORINA [44] was used to calculate 3D structures for each molecule. In a word, these 2352 descriptors cover different constitutional, topological, geometrical, electrostatic, physical, and shape descriptors. In addition, log *D* values at pH 7.4 and other pHs (from 0 to 14) were computed with Marvin software by ChemAxon (Budapest, Hungary).

2.3. Modeling methods

Three different modeling methods were employed for this study: MLR, partial least squares (PLS), and random forests (RF). These modeling methods have been used in QSRR before, and are readily available from standard software packages such as Matlab. However, our modeling was done with Lilly in house software packages. MLR is attractive for its simplicity relative to more complex models like PLS and RF. PLS has the potential to construct more predictive models when there is correlation in the molecular descriptors (i.e., latent factors representing hydrophobic, dispersive, and polar interactions derived from a combination of molecular descriptors may be more predictive than a pre-specified set of descriptors representing these interactions). RF has the potential to model non-linear relationships as well as statistical interactions (e.g., a dataset containing multiple binding modes of small molecules to a biomolecular target). The MLR and PLS methods were implemented in the SAS system (PROC REG and PLS, respectively) and used an information criterion (predictive determination coefficient) to control for overfitting [45,46]. The predictive determination coefficient (Rcv^2) is similar to other information criteria like the Akaike (AIC) and Bayesian (BIC) information criteria, but penalizes more relative to AIC or BIC to avoid overfitting and has been shown empirically to be a reasonable surrogate for future prediction error [45]. For the MLR method, forward variable selection was done with the predictive Rcv^2 used to determine how many terms to include in the linear model. A similar strategy was taken with PLS, using the predictive Rcv^2 to determine the number of latent variables to include in the PLS model.

2.4. Model evaluations

Each dataset was randomly split into two sets, 70% as the training set and 30% as the testing set. By comparing the predicted and experimental retention time for the 30% testing set, root-mean-square error of prediction (RMSEP) and Rcv^2 were computed. Twenty random splits were performed for each database with the distribution and mean RMSEP and R^2 values used to estimate the predictability and to permit comparison between different models.

3. Results

3.1. Characterizing dataset 1

A group of 151 representative drug-like compounds was selected for initial QSRR modeling. All compounds were identified from an in-house database, and fell within drug-like boundaries according to Lipinski's Rule of Five [47]. In order to assess the internal diversity of the dataset, a composite molecular fingerprint [48] was generated for each dataset using in-house tools. Those fingerprints were used to determine the nearest neighbor (NN) distance (1.0 – similarity, also known as Soergel distance [49]) for each molecule in the set, relative to all other molecules in that set. In a very diverse set of molecules, this distribution would be expected to tend towards longer distances, while a non-diverse set of molecules, like an SAR series, or a combinatorial library, would exhibit shorter distances on average. Since just the nearest neighbor distances are used, the distribution is for 151 points. Dataset 1 shows a distinct lack of short distance near neighbor relationships (Fig. 1), indicating very good internal diversity within this set—average NN distance of 0.34.