PREDICTING AQUEOUS SOLUBILITY FROM STRUCTURE

Jörg Wichard, Ronald Kühne
Leibniz Institut für Molekulare Pharmakologie, Robert-Rössle-Str. 10, D-13125 Berlin

Die hinreichende Wasserlöslichkeit kleiner organischer Moleküle ist entscheidend bei der Entwicklung neuer Arzneimittel. Viele Projekte der Wirkstoffentwicklung scheitern in relativ späten Entwicklungsphasen wegen unzureichender Löslichkeit in Wasser. Aus diesem Grunde ist die Vorhersage der Wasserlöslichkeit aus der 2-D Struktur wünschenswert. Wir stellen ein Modell zur Vorhersage der Wasserlöslichkeit vor, welches wir mit Methoden des maschinellen Lernens auf chemischen Datenbanken entwickelt haben.

1. Introduction

The aqueous solubility of small organic molecules is a key physical property for successful drug development because it affects directly the bioavailability. Poor solubility has been identified as a main problem in many drug development projects. It is therefore desirable to determine the solubility of the drug candidates as early as possible and there is much interest in the development of models that predict aqueous solubility directly from the structure [1].

Based on a data set containing experimentally determined aqueous solubilities of more than 14.000 compounds we build a machine learning model that uses 2D-molecular descriptors to predict solubility. We prefered simple countable molecular descriptors that could be calculated without knowledge of the 3-D structure and having a simple chemical interpretation (functional group counts, element counts, molecular properties).

We built ensembles of regression trees which are known to be robust against overfitting and that work quite fast. In order to test our model on “unseen” data, we split the entire data set in a training set that was used for model training and a test set that was used to validate the trained model.

2. Database

We extracted 24.949 measurements of aqueous solubility for 14.186 unique compounds from the Beilstein [2] and the Physprop database [3]. In the case of multiple measurements we took the median as a robust estimate of the solubility.

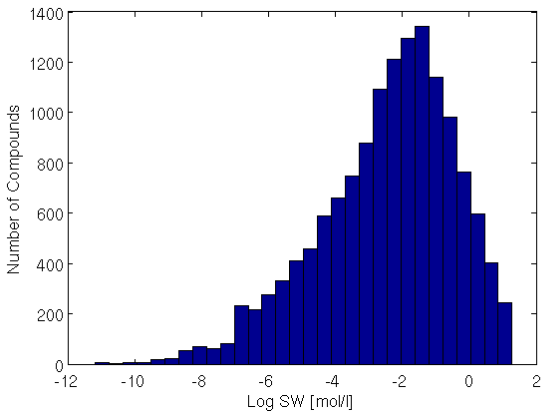

Figure 1: Distribution of log SW for the database.

The unit of the solubility measurement is mol/l and we took the logarithm of these values (log SW). The distribution of the log SW for the data set is shown in figure 1.

3. Molecular Descriptors and Feature Selection

According to Todeschini and Consonni a molecular descriptor is the final result of a logical and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number or the result of some standardized experiment [4].

For characterization of physical and chemical properties of the compounds in the data base we had to select some meaningful and interpretable candidates among the rich variety of molecular descriptors. The most relevant descriptors are listed in Table I with the ranking based on the variable importance. The final model included discrete and countable descriptors like element counts, molecular property counts, Ghose-Crippen AlogP counts [6] and electrotopological state counts from Kier and Hall (ES-Counts) [7]. All molecular descriptors were calculated using the Pipeline Pilot software [5].

Our feature selection approach follows in principle the method of variable importance as proposed by Breiman [8]. The underlying idea is to select descriptors according to their prediction errors after random permutation of these descriptors. Briefly, a regression model is trained which uses all descriptors as input variables and the prediction error on a hold out data set is calculated. In a second step, the same is done after the successive permutation of the descriptor values. The relative increase of the prediction error calculated using permutated descriptor values compared to that found for the original descriptor set is a measure of the variable importance following the idea that the most discriminative descriptors are the most important ones.

4. Ensembles of Trees

Trees are conceptually simple but powerful tools for classification and regression. For our purpose we use the classification and regression trees (CART) as described in Breiman et al. [9]. The main feature of the CART algorithm is the binary decision role that is introduced at each tree node with respect to the information content of the split. In this way the most discriminating binary splits are near the tree root and they are forming a hierarchical decision scheme. It is known that trees have a high variance, so they benefit from the ensem-

ble approach [10]. A particular kind of tree ensemble is also known as random forest. The free parameters of the tree models are the number of splits, the impurity measure and the split criterion.

We built ensembles of regression trees in order to improve the accuracy of the final solubility model. These ensembles differ from the well known random forest by the training procedure. Our model training scheme is a mixture of bagging [8] and cross-validation. Bagging or bootstrap aggregating improves the final model by combining several trees that were trained on randomly generated subsets of the entire training set. We extended this approach by applying a cross-validation scheme for selecting the best performing tree on each subset and subsequently we combined these selected trees to an ensemble. In K-fold cross-validation, the data set is partitioned into K subsets. From these K subsets, a single subset is retained to test the predictive power of the trees. The remaining K-1 subsets are used for model training. The cross-validation process is repeated K times with each of the K subsets used only once as the validation data. The training set and the validation set contain randomly drawn subsets of the data without replications. The size of each validation set was 25% of the entire data set. The size of the test set was 50% of the entire dataset.

In every CV-fold we train several different trees with a variety of model parameters (i.e. the number of splits and the split criterion). In each fold we select only one tree to become a member of the final ensemble, namely the best tree with respect to the modelling error on the validation set.

The test set for the final model validation is held out from the entire training as shown in figure 2.

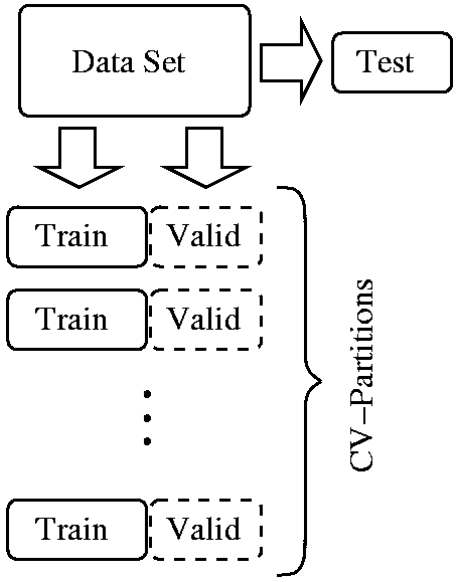

Figure 2: The data is divided in the training and the validation subsets for the cross-validation. The test set is held out from the training.

5. Results

After feature selection we ended up with 20 descriptors used to train the solubility model. The descriptors are listed in table I. We split the data set in two parts each containing 7093 samples and used the first part for training and the second part for testing the model. We trained a tree ensemble with 151 trees and calculated the root mean squared prediction errors (RMSE) for both data sets. Further we calculated the amount of compounds with a prediction error below 1 log-unit. The results are listed in table II.

Score	Descriptor (number of …)
3,59	atoms
2,99	bonds
2,73	H
2,05	molecular weight
1,56	C
1,32	O
0,99	C in CH3
0,82	aromatic rings
0,77	aromatic bonds in R--CH--R
0,69	ES Count ssCH2
0,65	CH2R2
0,45	H attached to C0sp3
0,33	ES Count ssO
0,32	CH3 attached to a heteroatom
0,30	ring bonds
0,24	Hetero atoms in R--CX—R
0,22	H attached to heteroatom
0,19	N
0,19	H acceptors
0,19	O in alcohol

Table I: The 20 descriptors of the solubility model. The score was calculated with the feature selection scheme as described in section 3.

	Training Set	Test Set
RMSE	0.8706	1.1745
Percent in ± 1 log-unit	80.53	66.54

Table II: The root mean squared prediction error (RMSE) and the amount of compounds with a prediction error below 1 log-unit.

These results are comparable with that of other groups using a different machine learning approach but the same data source [11]. We have to keep in mind that the experimental error in measuring aqueous solubility is believed to be at least 0.5 log units [12] and can reach even more than 1 log unit [1]. From this point of view, a model for aqueous solubility that predicts about two-thirds of all data inside of the 1 log unit is quite good.

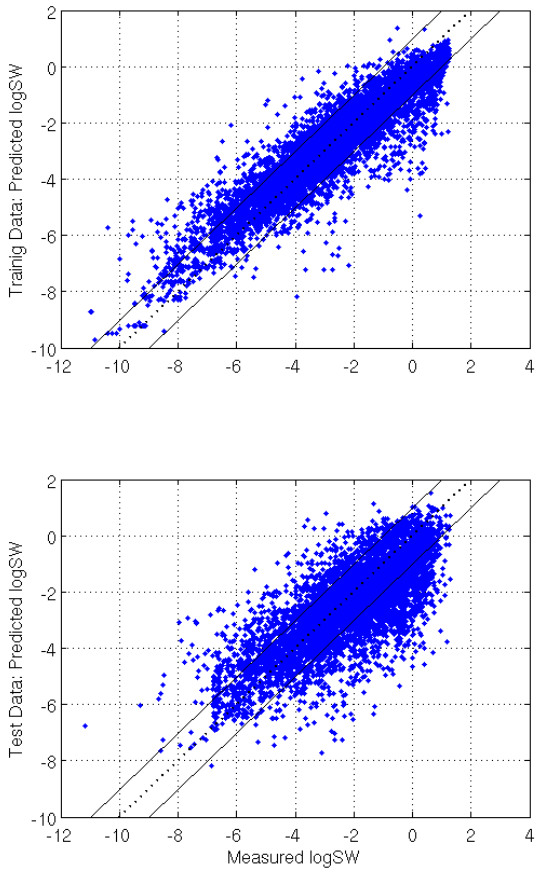

Figure 3: Predicted solubility (y-axis) versus measured solubility (x-axis) for the training set (on the top) and for the test set (on the bottom). The solid lines indicate the region where the prediction lies in the ±1 log unit range from the measured value.

5. Discussion

In this work, we present the use of machine learning models to predict aqueous solubility from the 2 D structure. Based on a database of about 14.000 compounds, we trained a model that achieves good accuracy and performance in an out-of-sample test. As pointed out by several other groups [1,11,12] the collection of well defined, high quality measurements is still the key issue for the development of accurate machine learning models.

References

[1] A. Llinàs, R.C. Glen, J.M. Goodman, J. Chem. Inf. Modeling 2008, 48, 1289-1303.
[2] MDL Information Systems. Beilstein crossfire database, 2005.
[3] NY Syracuse Research Corporation, Environmental Science Center Syracuse. Physical/chemical property database, 2005.
[4] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley-VCH, 2000.
[5] http://accelrys.com/products/scitegic/
[6] A.K. Ghose, V.N. Viswanadhan, J.J. Wendoloski, Journal of Physical Chemistry A, 1998, 102, 2, 3762-3772.
[7] L.H. Hall, L.B.Kier, J. Chem. Inf. and Comput. Sci., 2000, 40, 3, 784-791.
[8] L. Breiman, The Annals of Statistics, 1998, 26, 3, 801-849.
[9] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and Regression Trees, Wadsworth and Brooks, new edition 1993.
[10] L. Breiman, Machine Learning, 1996, 24, 2, 123-140.
[11] A. Schwaighofer, T. Schroeter, S. Mika, J. Laub, A. ter Laak, D. Sülzle, U. Ganzer, N. Heinrich, K.R. Müller, 2007, Journal of chemical information and modeling, 47, 2, 407-424.
[12] K.V. Balakin , N.P. Savchuk , I.V. Tetko, 2006, Curr. Med. Chem. 13, 2, 223-241.