Chemistry Central Journal

Poster presentation

Open Access

In silico prediction of aqueous solubility – classification models

C Kramer*[1,2], B Beck[2] and T Clark[1]

Address: [1]Computer-Chemie-Centrum and Interdisciplinary Center for Molecular Materials Friedrich-Alexander Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, 91052 Erlangen, Germany and [2]Boehringer-Ingelheim Pharma GmbH&Co KG, Department of Lead Discovery, Birkendorferstr. 65, 88397 Biberach, Germany

* Corresponding author

from 3rd German Conference on Chemoinformatics
Goslar, Germany. 11-13 November 2007

Published: 26 March 2008

Chemistry Central Journal 2008, **2**(Suppl 1):P23 doi:10.1186/1752-153X-2-S1-P23

This abstract is available from: http://www.journal.chemistrycentral.com/content/2/S1/P23

Solubility is a very important parameter in pharmaceutical research, especially for the early phase of drug discovery in fully automatized high throughput screening, compound pool extension and SAR and ADME-Tox parameter measurement. In recent years a multitude of models has been published concerned with the exact prediction of aqueous solubility. Still, almost all in the meantime commercially available tools suffer from comparably bad R^2y values for the prediction of solubility of pharmaceutically relevant molecules [1]. First, this might be attributed either to a bad data situation, as the reaction conditions for obtaining solubility data published in the literature are quite different. Second, many compounds with solubility values extracted from literature are not druglike. But even with high quality data measured in one lab, R^2y values derived from that data with the latest high-end algorithms are often not satisfying. In a very careful study recently published by Müller et al, with a Gaussian process model they got an R^2y value of 0.53 on a separate dataset derived from inhouse shake-flask experiments [1].

However, knowing the exact value is not really important for many applications; it is rather important to know whether a certain compound will be insoluble under the used test-conditions and should thus be excluded from the experiment.

In order to address this question we built classification models based on two datasets measured inhouse at Boehringer-Ingelheim at pH 7.4: one kinetic set of solubility measurements based on nephelometry and one thermodynamic set of solubility measurements based on shake-flask experiments. The datasets were divided into three classes, one well soluble class, one insoluble class and a buffer class in between to compensate for noisy data. For these datasets, we built classification models using support-vector machines (SVM) and Bayesian regularized neural networks (BRANN), trying several different descriptor sets. In each case, MOE2D descriptors and a SVM model gives the best raw results with an overall accuracy of ~70% for triple crossvalidation. Leaving out the predictions for and of the buffer class i.e. only considering strong outliers, the overall accuracy is ~88.5 %.

We evaluated classifier fusion and model applicability domain (MAD) considerations for this dataset. Applying these, we achieved accuracies of ~93 % for ~80 % of the dataset.

References

1. Schwaighofer A, Schroeter T, Mika S, Laub J, ter Laak A, Sülzle D, Ganzer U, Heinrich N, Müller K-R: *J Chem Inf Model* 2007, **47:**407-424.