Chemistry Central Journal

Poster presentation

Open Access

Comparison of applicability domains of QSAR models: application to the modelling of the environmental toxicity against *Tetrahymena pyriformis*

Igor V Tetko*[1], Alexander Tropsha[2], H Zhu[2], E Papa[3], P Gramatica[3], T Öberg[4], D Fourches[5] and A Varnek[5]

Address: [1]GSF - National Research Centre for Environment and Health, Institute for Bioinformatics, Ingolstaedter Landstrasse 1, D-85764 Neuherberg, Germany, [2]UNC School of Pharmacy, Medicinal Chemistry and Natural Products, Beard Hall, Room 327A, Chapel Hill, NC 27599-7360, USA, [3]Varese, Italy, [4]Kalmar, Sweden and [5]Université Louis Pasteur de Strasbourg, Institut de Chimie, 4, rue B. Pascal, F-67000 Strasbourg, France

* 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):P14 doi:10.1186/1752-153X-2-S1-P14

This abstract is available from: http://www.journal.chemistrycentral.com/content/2/S1/P14

The estimation of the applicability domains and the accuracy of predictions are the critical problems in QSAR modelling. In this presentation we propose a new approach to compare different estimates of applicability domains based on "distances to models" as well as their significance. The "distance to model" is a measure of similarity between training and test sets compounds, which could be expressed in many different ways (e.g., Tanimoto coefficient, Euclidian distance, leverage, etc.). The main idea of our analysis is that the errors can be described by a mixture of Gaussian distributions rather than a single Gaussian distribution. Here, we analyse 12 QSAR models for aqueous toxicity against *Tetrahymena pyriformis* obtained with different machine-learning methods and various types of descriptors. The models were obtained and tested for a dataset of 1093 compounds measured in the same laboratory [1-4]. The dataset has been divided into training and test sets using a diversity sampling algorithm. The majority of examined distances to models provided significantly lower statistical scores (bootstrap test, $p<0.05$) for description of errors when using a mixture of Gaussian distributions. The use of the mixture distributions fitted to the training set affords a significantly better assessment of the test set errors as compared to the assumption of identical distribution of errors in training and test sets.

References

1. Schultz TW, Netzeva TI: **Development and evaluation of qsars for ecotoxic endpoints: The benzene response-surface model for Tetrahymena toxicity.** In *Modeling environmental fate and toxicity* Edited by: *M. T. D. Cronin and D. J. Livingstone. Boca Raton, FL, USA*; 2004:265-284.
2. Schultz TW, Hewitt M, Netzeva TI, Cronin MTD: **Assessing applicability domains of toxicological QSARs: Definition, confidence in predicted values, and the role of mechanisms of action.** *QSAR Comb Sci* 2007, **26:**238-254.
3. Schultz TW, Netzeva TI, Cronin MT: **Evaluation of QSARs for ecotoxicity: A method for assigning quality and confidence.** *SAR QSAR Environ Res* 2004, **15:**385-397.
4. Aptula AO, Roberts DW, Cronin MT, Schultz TW: **Chemistry-toxicity relationships for the effects of di- and trihydroxybenzenes to tetrahymena pyriformis.** *Chem Res Toxicol* 2005, **18:**844-854.