Chemistry Central Journal

Poster presentation

Open Access

On some aspects of validation of predictive QSAR models

K Roy*, PP Roy and JT Leonard

Address: Drug Theoretics and Cheminformatics Lab, Division of Medicinal and Pharmaceutical Chemistry, Department of Pharmaceutical Technology, Jadavpur University, Kolkata 700 032, India

Email: K Roy* - kunalroy_in@yahoo.com

* 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):P9 doi:10.1186/1752-153X-2-S1-P9

This abstract is available from: http://www.journal.chemistrycentral.com/content/2/S1/P9

Quantitative structure-activity relationships (QSARs) represent predictive models derived from application of statistical tools correlating biological activity (including therapeutic and toxic) of chemicals (drugs/toxicants/environmental pollutants) with descriptors representative of molecular structure and/or property. The success of any QSAR model depends on accuracy of the input data, selection of appropriate descriptors and statistical tools, and most importantly validation of the developed model. Validation is the process by which the reliability and relevance of a procedure are established for a specific purpose. Leave one-out cross-validation generally leads to an overestimation of predictive capacity, and even with external validation, no one can be sure whether the selection of training and test sets was manipulated to maximize the predictive capacity of the model being published. In this paper, we present some representative examples of validation of QSAR models in order to explore possible importance of the method of selection of training set compounds, setting training set size and impact of variable selection for training set models for determining the quality of prediction. The major conclusions from the study are: (1) *K*-means cluster based division of training and prediction sets can be used as a reliable method of division of data set into training and test sets for developing predictive QSAR models; (2) the training set size should be set at an optimal level so that the model is developed with proper training (learning) process and the developed model is able to satisfactorily predict the activity values of the test set compounds; (3) choice of variables for regression based only on Q^2 value may not be optimum. Furthermore, predictive R^2 value may not be considered as the only criterion to indicate external predictability of a model.