Benchmarking of linear and non-linear approaches for QSPR studies of metal complexation with ionophores

Igor V. Tetko
Institute for Bioinformatics, GSF, Germany and Institute of Bioorganic & Petrochemistry, Kiev, Ukraine, http://www.vcclab.org

Vitaly P. Solov'ev
Institute of Physical Chemistry, Russian Academy of Sciences, Leninskiy prospect 31a, 119991 Moscow, Russia

Alexey V. Antonov
Institute for Bioinformatics, Neuherberg D-85764, Germany

Xiaojun Yao, Jean Pierre Doucet, Botao Fan
Université Paris 7-Denis Diderot, ITODYS-CNRS UMR 7086, 1, rue Guy de la Brosse, Paris 75005, France

Frank Hoonakker, Denis Fourches, Piere Jost, Nicolas Lachiche and Alexandre Varnek*
Laboratoire d'Infochimie, UMR 7551 CNRS, Université Louis Pasteur, 4, rue B. Pascal, Strasbourg 67000, France, e-mail: varnek@chimie.u-strasbg.fr

A benchmark of several popular methods, Associative Neural Networks (ANN), Support Vector Machines (SVM), k Nearest Neighbors (kNN), Maximal Margin Linear Programming (MMLP), Radial Basis Function Neural Network (RBFNN), Multiple Linear Regression (MLR) is reported for quantitative-structure property relationships (QSPR) of stability constants logK_1 for the 1:1 (M:L) and log $_2$ for 1:2 complexes of metal cations Ag^+ and Eu^{3+} with diverse sets of organic molecules in water at 298 K and ionic strength 0.1 M. The methods were tested on three types of descriptors: molecular descriptors including E-state values, counts of atoms determined for E-state atom types and substructural molecular fragments (SMF). Comparison of the models was performed using 5-fold external cross-validation procedure. Robust statistical tests (bootstrap and Kolmogorov-Smirnov statistics) were employed to evaluate significance of calculated models. The Wilcoxon signed-rank test was used to compare the performance of methods. Individual structure – complexation property models obtained with non-linear methods demonstrated significantly better performance than the models built using multi-linear regression analysis (MLRA). However, the averaging of several MLRA models based on SMF descriptors provided as good prediction as the most efficient non-linear techniques. Support Vector Machines and Associative Neural Networks contributed in the largest number of significant models. Models based on fragments (SMF descriptors and E-state counts) had higher prediction ability than those based on E-state indices. The use of SMF descriptors and E-state counts provided similar results, whereas E-state indices lead to less significant models. The current study illustrates difficulties of quantitative comparison of different methods: conclusions based only on one data set without appropriate statistical tests could be wrong.