Molecular Informatics 2015

GTM-Based QSAR Models and Their Applicability Domains

Gaspar H., Baskin I., Marcou G., Horvath D., Varnek A.
Kazan Federal University, 420008, Kremlevskaya 18, Kazan, Russia

Abstract

© 2015 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim. In this paper we demonstrate that Generative Topographic Mapping (GTM), a machine learning method traditionally used for data visualisation, can be efficiently applied to QSAR modelling using probability distribution functions (PDF) computed in the latent 2-dimensional space. Several different scenarios of the activity assessment were considered: (i) the "activity landscape" approach based on direct use of PDF, (ii) QSAR models involving GTM-generated on descriptors derived from PDF, and, (iii) the k-Nearest Neighbours approach in 2D latent space. Benchmarking calculations were performed on five different datasets: stability constants of metal cations Ca2+, Gd3+ and Lu3+ complexes with organic ligands in water, aqueous solubility and activity of thrombin inhibitors. It has been shown that the performance of GTM-based regression models is similar to that obtained with some popular machine-learning methods (random forest, k-NN, M5P regression tree and PLS) and ISIDA fragment descriptors. By comparing GTM activity landscapes built both on predicted and experimental activities, we may visually assess the model's performance and identify the areas in the chemical space corresponding to reliable predictions. The applicability domain used in this work is based on data likelihood. Its application has significantly improved the model performances for 4 out of 5 datasets.

http://dx.doi.org/10.1002/minf.201400153

Keywords

Activity landscape, Dimensionality reduction, Generative topographic mapping, GTM descriptors., QSAR