[P32] A Novel Approach to Quantify Data Set Topology

Annkathrin Weißenborn, Knut Baumann

Institute of Medicinal and Pharmaceutical Chemistry, Braunschweig University of Technology, Beethovenstrasse 55, D-38106, Braunschweig, Germany

Recently, a multitude of different benchmark data sets with various design criteria and widely different screening performance were published for ligand-based and structure-based virtual screening (VS) methods. The results of VS strongly depend on the underlying data sets. Self-similarity of the active compounds and their separation to the inactive compounds are two important factors for the screening performance in retrospective VS studies.[1][2]
The aim of this study is to characterize these two factors for ligand-based VS with novel measures of data set topology. Recently, three different, publicly available benchmark data sets (MUV, DUD and a subset from ChEMBL) were evaluated with various ligand-based VS techniques.[3] These data sets are used to study and characterize different data set topologies. Furthermore, the new measures of topology are correlated with the VS performance.

In addition to the already known Refined Nearest Neighbor Analysis,[4] measures based on Bemis-Murcko scaffolds (BMS) and maximum common substructures (MCS) were developed to characterize the individual data sets.

It is shown that composite measures derived from BMS and MCS correlate well with the screening performance. These results are in accord with previous studies by Rohrer *et al.* in which the topology in the simple descriptor space was related to the VS performance.[4]

[1] A.C. Good; T.I. Oprea J. Comput. Aided. Mol. Des. 22 (2008) 169–178.
[2] M.L. Verdonk; V. Berdini; M.J. Hartshorn; W.T.M. Mooij; C.W. Murray; R.D. Taylor; P. Watson J. Chem. Inf. Comput. Sci. 44 (2004) 793–806.
[3] S. Riniker; G.A. Landrum J. Cheminform. 5 (2013) 26.
[4] S.G. Rohrer; K. Baumann J. Chem. Inf. Model. 48 (2008) 704–718.