SIMILARITY-BASED DESCRIPTORS (SIBAR): A TOOL FOR SAFE EXCHANGE OF CHEMICAL INFORMATION?

Dominik Kaiser[a], Barbara Zdrazil[a], and Gerhard F. Ecker[a]

[a] Department of Medicinal/Pharmaceutical Chemistry, University of Vienna,
Althanstrasse 14, 1090 Wien, Austria

Recently we published the sucessful application of a set of new descriptors on similarity values, denoted as SIBAR-descriptors (Similarity Based SAR). These descriptors are based on calculation of similarity (on basis of euclidian distances) for each compound of the data set to each compound of a reference set, using common descriptors. These euclidian distances (= similarity values) are then further used for QSAR-studies. Both the reference set as well as the descriptors used for calculating the SIBAR-values are tailored to the specific QSAR-problem. Best results have been obtained when targeting ADMET-problems. In any case it needs the knowledge of the reference set to retrieve the corresponding descriptors. Assuming that only the descriptors for calculating the SIBAR-values, but not the structures of the reference compounds are available, it should be impossible to trace back the chemical structure of the original compounds of the training set.

For this, compounds GPV0005, Diazepam and Estriol were used as query compounds to search for similar compounds in a large database. For the description of the structures three different descriptor sets were used. The 32 VSA-descriptors [1], a set of ADME-realted descriptors (weight, TPSA, SMR, SlogP, APOL, number of rotable bonds and the number of H-bond acceptors and donors) and a set of autocorrelations vectors of the PETRA-descriptors [2]. The reference compounds needed for applying SIBAR were a set of 20 highly diverse compounds from the SPECs-library. As target database a compound collection was build up with more than 1.5 million compounds via merging the databases from ChemDiv, SPECS, Maybridge, and Enamine. Additionally, the database was spiked with 200 compounds similar to the query structures.

In the first run the euclidian distances for the three reference compounds to each compound of the database were calculated and the first 20 similar compounds for the three reference structures were extracted from the database. Subsequently, the step was repeated using the SIBAR-approach.

The results show, that, using only the euclidian distance, the top ranked compounds retrieved by the similarity search are structurally very closed to the reference compounds. Doing the additional step of calculating the SIBAR-values the number of similar structures retrieved is decreased.

[1] Chemical Computing Group www.chemcomp.com

[2] Molecular Networks www.mol-net.de