[P35] Scaffold analysis of environmental and drug discovery databases

Milan Voršilák[12], Daniel Svozil[1], Igor Tetko[2]

[1] *Laboratory of Informatics and Chemistry, ICT Prague, Technická 5, Prague 6, 16628, Prague, Czech Republic.*
[2] *Institute of Structural Biology, Helmholtz Zentrum Munich, Ingolstädter Landstraße 1 D-85764 Neuherberg, Germany*

Molecular scaffold is the core structure of a molecule. It typically consists of rings connected by linkers [1]. Some scaffolds, such as large fused heterocyclic compounds, are strongly associated with toxicity and carcinogenicity. On the other hand, privileged scaffolds, such as a benzodiazepine core, serve as ligands for various types of receptors. Selection of the non-redundant representative set of compounds covering different scaffolds is, for example, important for the compound inclusion in biological screening programme or for the construction of combinatorial libraries. In this work, we performed the scaffold analysis of >60M molecules from the ChemNavigator database of commercially accessible screening compounds. More than 10 million scaffolds with up to 5 rings were generated, and most representative scaffolds were identified. E-state descriptors were computed for all scaffolds and were used, after normalization, as parameters for data clustering using a Kohonen self-organizing map [2]. Scaffolds from freely available drug and environmental databases DrugBank [3], ChEMBL DrugStore [4] and EINECS (European Inventory of Existing Commercial chemical Substances) were then projected into the Kohonen map, and observed differences in scaffold coverage were further studied. Our results demonstrate that the Kohonen map is a useful tool both for chemical space exploration, and for chemical library comparison and design.

[1] Bemis. GW.; Murcko. MA. J. Med. Chem. 39 (1996) 2887-93.
[2] Kohonen, T., Self-Organizing Maps. 3 ed.; Springer Verlag: New York, 2000.
[3] Knox, C.; Law, V; et al. Nucleic Acids Res. 39 (2011) D1035-1041.
[4] Gaulton A.; Bellis, LJ.; et al. Nucl. Acids Res. 40 (2012) D1100-D1107.