Poster presentation

Frequent substructure mining of GPCR ligands

E van der Horst*, A Bender and AP IJzerman

Address: Division of Medicinal Chemistry, Leiden/Amsterdam Center for Drug Research, Leiden University, Einsteinweg 55, 2333CC Leiden, The Netherlands

* Corresponding author

In this study, we conducted frequent substructure mining to find the structural features that discriminate between ligands that either do or do not bind to G protein-coupled receptors (GPCRs). Finding which substructures are rare and which are common in GPCR ligands will help in the design of new ligands and for prioritizing compounds for screening. Besides the normal 2D structure notation, three other chemical representations were used. The first 'elaborate' representation used a special type for aromatic bonds, the second also added a special type for any aromatic atom, and the third representation used a special notation for planar, not necessarily aromatic, structures. In all but the normal representation, wildcards were used for halogens and aliphatic heteroatoms with an extra label indicating the atom type. A set of 16 k GPCR ligands was compared against a roughly equal number from a screening set of compounds (Chembridge). For analysis of the results, two decision trees were constructed, one for the most-common substructure for GPCR ligands and one for the most-common substructure in the screening set. The alkylamine substructures were most discriminating for GPCR ligands as compared to the Chembridge set. This reflects the presence of aminergic receptor ligands in the GPCR dataset. Carboxamide substructures were most common in the Chembridge dataset. This is probably due to particular reaction types used to construct the screening library. The 'normal' representation mode led to the most significant substructure for GPCR ligands; the aromatic bonds representation yielded the most significant substructure for the screening compounds. In conclusion, frequent substructure mining is a useful approach for characterizing heterogeneous ligand datasets.