Emerging patterns mining and automated detection of contrasting chemical features

Alban Lepailleur[1], Jean-Philippe Métivier[2], Bertrand Cuissart[2], Ronan Bureau[1]

[1] *CERMN, Université de Caen Basse-Normandie, F-14032 Caen, France*
[2] *GREYC, Université de Caen Basse-Normandie, F-14032 Caen, France*

Due to the evolution of the modern information methods and technology, collecting, combining, storing, and mining huge amounts of chemical and biological data can be done very efficiently. The calculation of the frequency of a chemical substructure in a data set is often at the core of the process for the definition of its toxicological relevance. The rationale for using a frequency constraint is that it is unlikely to generalize on a substructure that has been observed on a few chemicals. However, algorithms that enumerate frequent substructures from a set of molecules often lead to the generation of too many such substructures. To limit the number of generated substructures, methods for finding representative and significant structural patterns have been developed in recent years.

The method developed by our group is based on the mining of emerging patterns. It directly operates from the molecular graphs and computes the conjunctions of chemical features whose frequencies of occurrence in a data set are sufficiently discriminative between different subgroups of molecules to be of interest.

A first application dealt with the automated detection of structural alerts from a mutagenicity data set [1] and the last evolution of is concerning the extraction of 2D-pharmacophoric patterns from a biologically annotated data set.

[1] Métivier, J.-P.; Lepailleur, A.; Buzmakov, A.; Poezevara, G.; Crémilleux, B.; Kuznetsov, S. O.; Goff, J. L.; Napoli, A.; Bureau, R.; Cuissart, B. J. Chem. Inf. Model. **2015**, *55* (5), 925–940.