Pharmaceuticals ecotoxicity: data curation and QSAR modeling

PAOLA GRAMATICA, ALESSANDRO SANGION,STEFANO CASSANI

QSAR Research Unit in Environmental Chemistry and Ecotoxicology, DiSTA, University of Insubria, Via J.H.Dunant 3, 21100 Varese, Italy, paola.gramatica@uninsubria.it, http://www.qsar.it

In the last years pharmaceuticals have became a class of potential emerging pollutants since they are increasingly present in waste and surface water, air and soil. Due to their presence in the environment, the European regulatory agency on pharmaceuticals, EMEA (European Medicines Evaluation Agency), published a guideline for the environmental risk assessment (RA) of human pharmaceuticals, (EMEA 2006) that shall drive the application for marketing authorisation. This guideline requires, for each chemical, a large amount of data as (eco)toxicity, environmental fate, consumption and so on.

Unfortunately, there is lack of (eco)toxicological data, available for human pharmaceuticals in literature and different databases. *In silico* approaches, like those based on QSARs, are valuable tools to maximize the information contained in existing experimental data and to predict missing information, filling the data gap. The global quality of a QSAR model, in terms of robustness and predictivity, depends on the quality of the data input, according to the general rule of computer science GIGO (Garbage in, Garbage out). Online databases like ECOTOX (US-EPA 2015) together with literature datasets were checked, looking for consistent (eco)toxicological data on a set of more than 400 pharmaceuticals. In addition to good quality data, we found thousands of values for unspecified species, unspecified measured effect, unspecified time of exposure, undefined test condition, often with a wide range of experimental responses (even more than three order of magnitude) for the same chemical.

A data curation was thus necessary, and in this study we also present how to prepare consistent datasets of pharmaceuticals with a clearly defined end-point, according to the first OECD principle for QSAR modeling (OECD 2004). Then, we used these refined datasets to develop local QSARs for different species representatives of a simplified aquatic ecosystem at different trophic levels (alga, daphnia and fish), in order to define the potential aquatic toxicological profile of pharmaceuticals. Toxicity was modeled by multiple linear regression (MLR) and the Genetic Algorithm was used to select the relevant molecular descriptors by the MLR-Ordinary Least Squares (OLS) method by using our software QSARINS (QSAR-Insubria) (GRAMATICA et al. 2013, 2014). The best models were validated for their robustness using leave-one-out, leave more out and the scrambling of the responses. External validation was also performed demonstrating the high predictive ability of the models. Finally, predictions by different models were combined by Principal Component Analysis (PCA) to verify the toxicity trend and to screen and prioritize the most environmentally hazardous pharmaceuticals.