Mining the Openfoodtox database for *in silico* modelling

E. Benfenati[1], F. Como[1], M. Marzo[1], D. Gadaleta[1], and J.L. Dorne[2]

[1]*IRCCS-Istituto di Ricerche Farmacologiche Mario Negri. Department of Henvironmental Health Sciences. Laboratory of Environmental Chemistry and Toxicology. Via La Masa 19, 20156 Milan (Italy).*

[2]*European Food Safety Authority. Scientific Committee and Emerging Risks Unit (SCER). Department of Risk Assessment and Scientific Assistance. Via Carlo Magno 1A, 43126 Parma (Italy).*

Background

Pesticides are applied on plants to control pests and diseases and ensure food production, and may induce resistances in target organisms and adverse effects in non-target species. Computational tools are of great value for risk assessors to predict toxicity of such pesticides as parent compounds, impurities, degradation products and related compounds, in multiple species of environmental relevance.

The strategy

1. Exploiting Openfoodtox

Over the last 5 years, EFSA has been developing the Openfoodtox database, providing toxicological data for critical effects of over five thousand chemicals peer reviewed by EFSA and Member States' experts, including pesticides. Openfoodtox provides a tool with many potential applications including the development of new in silico models, explored here within an EFSA funded contract.

The strategy

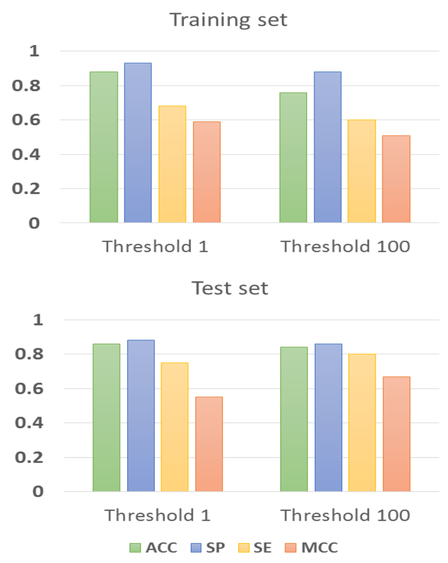
2. Developing in silico methods

CASE A: Bee Toxicity

1. Data: Acute toxicity data for contact exposure (LD50) in bees were extracted from Openfoodtox, Terrestrial US - EPA ECOTOX database present in the OECD QSAR Toolbox and the DEMETRA project. When conflicting values were observed, the values were rejected. Overall, the dataset covered 256 chemicals, which were the basis to develop the in silico model. Three toxicicty classes based on LD50 values were defined with chemicals above, below and between 1 and 100 µg/bee.

2. Algorithm: The algorithms applied were the *k*-nearest neighbors (*k*-NN) models which were built using the in-house software istkNN, previously described (Manganaro et al., 2016).

3. Results: We report a first case study on the classification of honeybees. Figure 1 shows the statistics of the model. The model has been described (Como et al., 2017). The validation of K-NN models for honeybees suggests their suitability to reliably predict toxicity of structurally diverse pesticides and their use for the screening and prioritization of new pesticides in the future. Another advantage of this model is that it introduces practical thresholds. Furthermore, the model is freely available at the VEGA website: www.vega-qsar.eu, providing immediate result.



CASE B: Algae Toxicity

1. Data: Acute toxicity data (EC50) in algae for 230 pesticides were extracted from Openfoodtox. We defined two toxicity classes, with chemicals above and below 1 mg/L. The dataset was split into a Training set (TS) (80%) and a Validation set (VS) (20%) with an equal distribution of toxic and non-toxic compounds.

2. Algorithm

A. Statistical-based method: Random Forests (RF) implemented in KNIME were applied for model derivation. Different techniques for feature selection were used on DRAGON descriptors for derive five RF models:

a. All descriptors (ALL)
b. Genetic algorithm (GA)
c. Random Forest (RF)
d. Random Forest + Backward Feature Elimination (RF+BFE)
e. Genetic Algorithm + Backward Feature Elimination (GA+BFE)

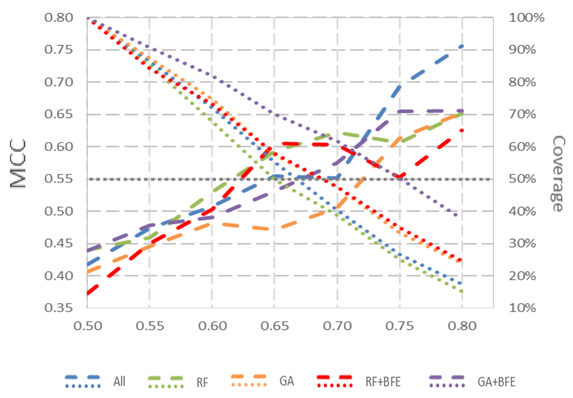


Figure 2

Out-of-bag statistics on TS were used for selecting the best model and the optimal prediction confidence threshold returning the best compromise on classification performance (dashed lines) and dataset coverage (dotted lines). The RF derived from the RF+BFE pool (red lines) with a 0.65 threshold was selected as best model (Figure 2).

B. Knowledge-based method: IstChemFeat software by Kode s.r.l. (www.chm.kode-solutions.net) was used to extract relevant chemical features (functional groups and atom centered fragments) for discriminating toxic and non-toxic compounds.

Selected features were classified for relevance on the basis of the ratio between toxic and non-toxic compounds fired within the TS.

C. Integrated strategy: The two approaches were combined into a multi-step integrated strategy:

Step 1: predictions using features with perfect discrimination ability

step 2: predictions of RF with confidence ≥ 0.65.

step 3: predictions with other features

step 4: predictions with both confidence < 0.60 and little reliable features were rejected.

3. Results: performance in internal (TS) and external (VS) validation are in Figure 3:

A: RF with confidence ≥ 0.65
B: Features model
C: Integrated strategy, steps 1-3
D: Integrated strategy, steps 1-4

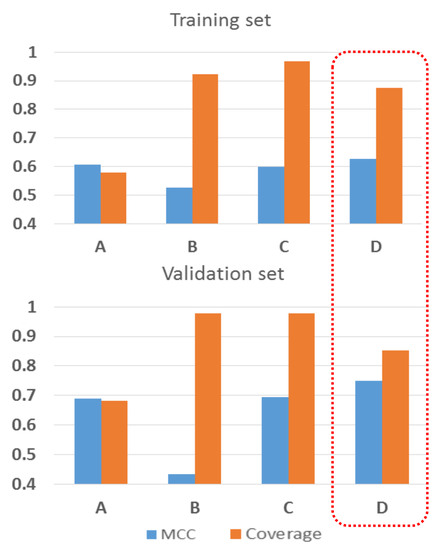


Figure 3

Conclusions

The availability of a large number of acute toxicity values for a large number of substances, as implemented in Openfoodtox, allows the development of new in silico models, which provide tools to predict toxicity while extracting all relevant information in a structured manner. Here the development of new in silico models for acute toxicicty endpoints have been exploited in bees and algae for which very limited or no model available so far.

References

Como F, Carnesecchi E, Volani S, Dorne J L, Richardson J C, Bassan A, Pavan M, Benfenati E. Chemosphere; 166 : 438-444, 2017

Manganaro, A., Pizzo, F., Lombardo, A., Pogliaghi, A., Benfenati, E., 2016. Chemosphere 144, 1624e1630

We acknowledge the financial contribution of EFSA, contract NP/EFSA/AFSCO/2016/01.