Ready Biodegradability (consensus model)

Dataset profile

Biodegradability[1] describes the capacity of substances to be mineralized by free-living bacteria. It is a crucial property in estimating a compound's long-term impact on the environment. Chemicals that do not quickly degrade have the potential to release their toxic effects over a long period; they can therefore pose a greater risk than chemicals with higher acute toxicity, but which are not stable. In order to better characterize readily biodegradable chemicals, the Organization for Economic Cooperation and Development (OECD) made efforts to develop standardized methods. In 1992, test guideline 301 was published, describing six methods of screening chemicals for ready biodegradability under aerobic conditions. The ability to reliably predict biodegradability reduces the need for laborious experimental testing. The various test methods share a number of features: the test substance is incubated in a mineral medium (potassium, sodium phosphate, etc.) and an inoculum (activated sludge, surface soils, etc.) under aerobic conditions in dark or diffuse light. A reference compound (aniline, sodium acetate, or sodium benzoate) is run in parallel as a control. The degradation is then determined by measuring properties such as DOC (dissolved organic carbon), CO_2 production, and O_2 uptake. The test should run for a period of 28 days. The pass levels for readily biodegradability must be reached during a 10-day window within the 28-day test period. Depending on the test method employed, these are:

- 70% DOC: percentage of dissolved organic carbon removed
- 60% ThOD: percentage of the theoretical oxygen demand
- 60% ThCO_2: percentage of the theoretical carbon dioxide yield

The initial biodegradability dataset was collected from three main sources: internal CADASTER dataset comprising measurements extracted from CHRIP (Chemical Risk Information Platform), measurements assembled by [Cheng at al In silico assessment of chemical biodegradability. *J. Chem. Inf. Model.* **2012**, *52*, 655-669]and a dataset with measurements of fragrances collected by Prof. Gramatica group. These data were already classified as "readily biodegradable/ non readily biodegradable" compounds and comprised 1884 compounds, including 37% readily biodegradable ones.

[1]The description is according to Vorberg and Tetko, Modeling the biodegradability of chemical compounds using the Online CHEmical Modeling environment (OCHEM), *Mol. Inform.* 2014, 33(1), 73–85 (Open Access), doi: 10.1002/minf.201300030.

Data preprocessing

All chemical structures were processed using OCHEM cleaning and standardization protocols.

Descriptors

The consensus model was calculate as a simple average of seven ASNN models developed with individual descriptors, namely ALOGPS + Estate, GSFRag, ISIDA fragments, Dragon, Adriana, CDK and ChemAxon. 3D conformations of molecules were generated using CORINA software, which is distributed by Molecular Networks GmbH.

Validation

The model was built using 5-fold cross validation. The dataset of 63 and 38 compounds compiled by [Boethling and Costanza, Domain of EPI suite biotransformation models, *SAR QSAR Environ. Res.* **2010**, *21*, 415-443] and [Steger-Hartmann, et al Incorporation of in silico biodegradability screening in early drug development—a feasible approach? *Environ. Sci. Pollut. Res. Int.* **2011**, *18*, 610-619, doi: 10.1007/s11356-010-0403-2] were used.

Statistical parameters

Prediction accuracy

The basic prediction accuracy parameters according to the 5-fold cross-validation procedure and prediction of test sets are:

Property	# samples	Accuracy	BA	MCC	AUC
Training set	1884	88	88	0.74	0.95
Boethling and Costanza	63	86	71	0.4	0.86

Applicability domain

The prediction accuracy is estimated using PROB-STD distance to model as described in [Sushko et al, Applicability domains for classification problems: Benchmarking of distance to models for Ames mutagenicity set. *J. Chem. Inf. Model.* **2010**, 50(12):2094-2111, doi: 10.1021/ci100253r]