Mutagenicity (Ames test)

Dataset profile

The Ames mutagenicity data set was published in [Sushko et al. Applicability Domains for Classification Problems: Benchmarking of Distance to Models for Ames Mutagenicity Set] [Hansen at al. Benchmark data set for in silico prediction of Ames mutagenicity. —*J. Chem. Inf. Model.*, 2010, 50 (12), pp 2094–2111].

The The Ames test relies on the determination of the mutagenic effect of a given compound on histidine-dependent strains of Salmonella typhimurium. Thus, the measurable mutagenic ability of a compound may signal its potential carcinogenicity. The Ames test can be used with different bacteria strains and can be performed with or without metabolic activation using liver cells. For this study, all such diverse data were pooled together. According to that approach, a molecule can be considered as active if it demonstrates mutagenic activity for at least one strain.

Thus, considering that the benchmark set molecules were tested with different strains, there may be a significant variance in results. Moreover, different authors used different thresholds to decide whether a given molecule is active or not. As shown in the Results and Discussion section, we estimated the intra- and interlaboratory accuracies of measurements in the Ames mutagenicity data set to be 94% and 90%, respectively. The initial data set was randomly divided into training and external test sets. The training set contained 4361 compounds, including 2344 (54%) mutagens and 2017 (46%) nonmutagens. The external test set contained 2181 compounds (1/3 of initial set) including 1172 (54%) mutagens and 1009 (46%) nonmutagens. These data sets were used for the 2009 Ames mutagenicity challenge, where the external test set was given to the participants for “blind predictions”.

Data preprocessing

All chemical 3D structures were cleaned using OCHEM cleaning protocol. The standardization was performed in OCHEM.
All salt counter ions were removed and resulting ions were neutralized.

Descriptors

This model was built using EState descriptors (electrotopological EState indices) according to OCHEM implementation.

Validation

The model was built using 5-fold cross validation together with an external validation set.

Statistical parameters

Prediction accuracy

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

Data Set	#	Accuracy	Balanced accuracy	MCC	AUC
Training set	4359 records	77.7% ± 0.6	77.5% ± 0.6	0.55 ± 0.01	0.854 ± 0.01
Test set	2181 records	79.6% ± 0.8	79.5% ± 0.9	0.59 ± 0.02	0.875 ± 0.01

Applicability domain

The prediction accuracy is estimated using PROB-STD distance to model and sliding window based accuracy averaging. The detailed technical description of these methodology can be found in a thesis work [Sushko. Applicability Domain of QSAR models. Doctoral thesis. 2011. Technical University of Munich.]. The thesis can be downloaded at http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:bvb:91-diss-20110301-1004002-1-2