Evaluation of the Predictive Accuracy of QSAR Models and Alerts for Genotoxicity Using a Newly Compiled Experimental Dataset

Prachi Pradeep[1,2], Richard Judson[2], Grace Patlewicz[2]

[1]Oak Ridge Institute for Science and Education, Oak Ridge, Tennessee

[2]National Center for Computational Toxicology, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina

Prachi Pradeep | pradeep.prachi@epa.gov | ORCID iD: 0000-0002-9219-4249 | Phone: 919-541-5150

www.epa.gov

BACKGROUND

Carcinogenicity and mutagenicity are among the toxicological end points that pose the highest concern for human health and are subject to regulatory assessment. Here, the predictive accuracy of several publicly available genotoxicity QSARs and structural alerts was assessed using a large new dataset (~45K entries) compiled from a number of sources including TOXNET, COSMOS, eChemPortal, ECVAM. Data from assays that detect bacterial mutagenicity (Ames) or chromosomal aberrations (CA) were evaluated using a conservative IATA to derive a call for genotoxic potential referred to as GeneTox Call using the classification scheme by Williams et al., 2019. The IATA assigned a chemical as genotoxic, if any single assay was positive. QSAR tools, Toxicity Estimation Software Tool (TEST) and Lazar (Lazy structure–activity relationships), and the OECD Toolbox structural alerts/profilers (e.g. OASIS DNA alerts for Ames, CA) were used to make *in silico* predictions for genotoxicity. A new scheme was derived for GeneTox classifications and consensus models were developed to predict genotoxicity. The (*in silico*) predictions were compared against Williams et al., 2019 and newly derived genetox classifications.

DATA PREPARATION

Assay harmonization: A major effort was undertaken to harmonize naming of assay types and assay calls. Assays types were then aggregated into 3 categories:

1. Ames assays
2. Clastogen assays (e.g. mouse lymphoma, micronucleus assay), and
3. Others

Chemical Structure Curation: DSSTox

GeneTox Call: The conservative IATA using Williams et al., 2019 was used to derive an overall GeneTox call per substance.

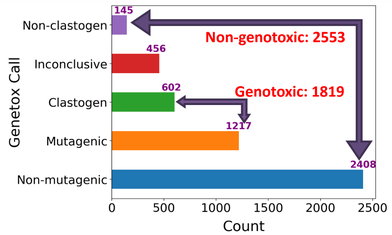


Figure 1: Distribution of GeneTox calls using Williams et al., 2019 classification scheme.

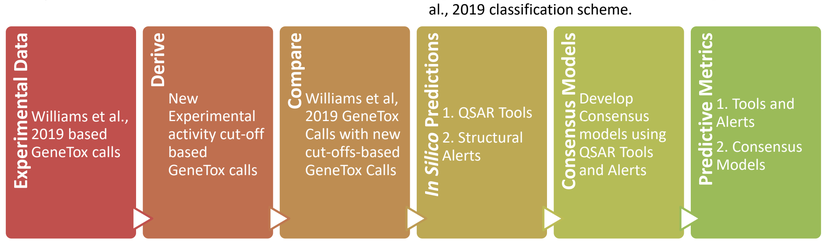


Figure 2: Outline of the workflow adopted in this work.

In Silico Tools	Tool Type	Label	Details
Toxicity Estimation Software Tool (TEST)	QSAR	T1	Ames
Lazy structure–activity relationships (Lazar)	QSAR	T2	Salmonella typhimurium
OECD Toolbox	Alerts	A1-A5	A1: DNA alerts for AMES by OASIS A2: Alerts for CA and MNT by OASIS A3: Protein binding alerts for Chromosomal aberration by OASIS A4: in vitro mutagenicity (Ames test) alerts by ISS A5: in vivo mutagenicity (Micronucleus) alerts by ISS

Figure 2: Summary of *in silico* tools used to make genotoxicity predictions.

U.S. Environmental Protection Agency
Office of Research and Development

GENETOX CLASSIFICATION ANALYSIS

Ames (% active, count)	Williams et al., 2019 Calls		
	Genotoxic	Non-genotoxic	Inconclusive
< 50% (3295)	16.1% (530)	73.1% (2408)	10.8% (357)
≥ 50% (1054)	100% (1054)	0% (0)	2.9% (33)
Inconclusive (300)	15.3% (46)	0% (0)	91.2% (208)

Table 2: (a). Summary statistics of chemical classification relating Ames activity and Williams et al., 2019 classification schemes. Note that ~16% chemicals active in <50% Ames assays were classified as genotoxic and ~15% chemicals with inconclusive Ames data were classified as genotoxic.

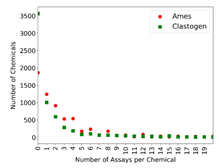
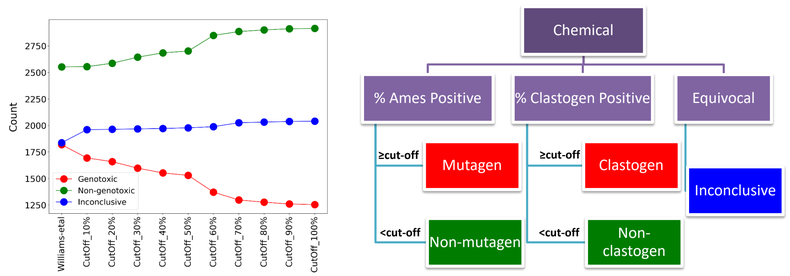


Figure 3: Distribution of number of assay data per chemical.



(a) (b)

Figure 4: (a). Count of chemicals re-classified as genotoxic (red) and non-genotoxic (green), (b). Cut-off based classification scheme defined by the percentage of Ames and Clastogen assays a chemical is active in. The thresholds for cut-offs range from 10-100%.

PREDICTIVE ACCURACY OF TOOLS

Predictor	Accuracy (%)	Sensitivity (%)	Specificity (%)	Balanced Accuracy (%)	Kappa
	GeneTox Call: Williams et al., 2019				
Lazar (n = 3011)	73.17	75.27	71.95	73.61	0.45
TEST (n = 3106)	73.08	78.70	70.42	74.56	0.44
OECD A1 (n = 3074)	69.52	91.64	63.91	77.78	0.37
OECD A2 (n = 3074)	69.42	89.68	64.00	76.84	0.37
OECD A3 (n = 3074)	58.49	70.57	56.50	63.53	0.14
OECD A4 (n = 3074)	73.26	75.63	71.67	73.65	0.46
OECD A5 (n = 3074)	63.57	58.47	72.74	65.60	0.28
	GeneTox Call: CutOff_50%				
Lazar (n = 3091)	75.29	72.32	77.00	74.66	0.48
TEST (n = 2999)	76.43	76.68	76.31	76.50	0.50
OECD A1 (n = 3060)	73.56	90.79	69.21	80.00	0.42
OECD A2 (n = 3060)	72.99	87.78	69.05	78.41	0.41
OECD A3 (n = 3060)	61.56	66.18	60.82	63.50	0.14
OECD A4 (n = 3060)	75.52	73.00	77.21	75.10	0.50
OECD A5 (n = 3060)	63.05	54.64	78.06	66.35	0.29

Table 3: Performance metrics of predictive (in silico) tools against Williams et al., 2019 and GeneTox cut-off based calls.

CONSENSUS MODELING

Figure 5: Heatmap depicting correlation between different predictors. Based on the coefficient of correlation (>0.8) alert 2 was dropped from the consensus model. The total number of tools in the model are 6 (2 QSAR and 4 alerts).

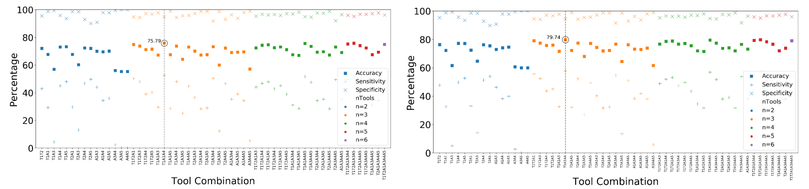
Algorithm: Each tool has a binary (genotoxic or non-genotoxic) prediction. A combination of tools is considered as a unique tool leading to 2^6= 64 unique tool combinations. The posterior probability of a chemical being genotoxic or non-genotoxic is calculated using the Bayes Theorem.
Prediction: The posterior probability of the prediction combination for a new (test) chemical is compared to a cut-off (=50%) to make the final prediction.

Two combinations were calculated:

$$Combination1 = T1 + T2 + A1 + A2 + A3 + A4 + A5$$
$$Combination2 = 2^5 \times T1 + 2^4 \times T2 + 2^3 \times A1 + 2^2 \times A3 + 2^1 \times A4 + 2^0 \times A5$$

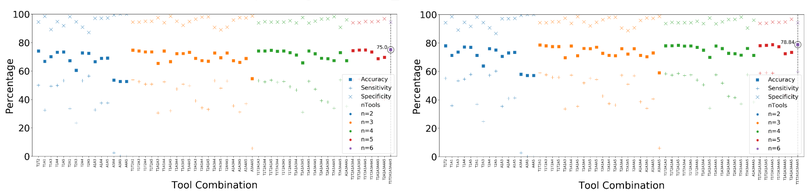
1	1	1	1	1	1	C1	
32	16	8	4	2	1		C2
T1	T2	A1	A3	A4	A5		
0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	1
0	0	0	0	1	0	1	2
0	0	0	0	1	1	2	3
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
1	1	1	1	0	0	4	60
1	1	1	1	0	1	4	61
1	1	1	1	1	0	5	62
1	1	1	1	1	1	6	63

Table 4: Depiction of prediction combinations from different tools and the resultant prediction combination (C1: Combination, C2: Combination2).



(a) (b)

Figure 6: Performance metrics of different consensus tools based on Combination 1. (a). As compared against Williams et al., 2019 GeneTox calls. (b) As compared against CutOff_50% calls.



(a) (b)

Figure 7: Performance metrics of different consensus tools based on Combination 2. (a). As compared against Williams et al., 2019 GeneTox calls. (b) As compared against CutOff_50% calls.

SUMMARY

- Williams et al., 2019 scheme for chemical classification as genotoxic or non-genotoxic is conservative in nature as compared to experimental activity cut-off based classification scheme.
- The individual QSAR tools and alerts have similar predictivity with balanced accuracies ranging from 64-80% with the models performing better against the cut-off based classification scheme.
- The consensus models using various combination of tools do not result in significant differences in the overall prediction across various combinations.
- The predictivity of individual tools and the consensus models is slightly improved using the experimental activity cut-off based classification scheme as opposed to the Williams et al., 2019 scheme for GeneTox calls.