An evaluation of selected (Q)SARs/expert systems for the Prediction of Skin Sensitization Potential

Jeremy Fitzpatrick[1], David W Roberts[2], and Grace Patlewicz[1]

1. US EPA, National Center for Computational Toxicology, Research Triangle Park, NC
2. School of Pharmacy, Liverpool John Moores University, Byrom Street, Liverpool, UK

ORCID ID:0000-0002-5401-9706

Jeremy Fitzpatrick 1 Fitzpatrick.Jeremy@epa.gov

Abstract

Predictive testing to characterize substances for their skin sensitization potential has historically been based on animal models such as the Local Lymph Node Assay (LLNA) and the Guinea Pig Maximization Test (GPMT). In recent years, EU regulations have provided a strong incentive to develop non-animal alternatives – both in vitro and in silico. Here we selected three expert systems: Derek Nexus, TIMES-SS, and VEGA , and evaluated their performance using two large sets of animal data, one of 1249 substances (354 sensitizers and 895 non-sensitizers) and a second of 515 substances (329 sensitizers and 186 non-sensitizers). We considered a model to be successful at predicting skin sensitization if it had at least the same balanced accuracy as the LLNA and the GPMT had in predicting the outcomes of one another, which ranged from 79% to 86% depending on the dataset. We found that none of the expert systems evaluated was able to achieve such a high balanced accuracy in their global predictions, with balanced accuracies ranging from 56% to 65%. However, for substances within the domain of TIMES-SS, balanced accuracies were found to be 79% and 82% for the 2 datasets respectively, in line with the animal data.

Aims

Gather available data for skin sensitization and assess its accuracy

- Retrieve a large data set of substances assessed for skin sensitization in vivo using the OECD's eChemPortal
- Determine which substances had been tested in both the GPMT and LLNA
- Compare the outcomes of substances which have been tested in both the GPMT and the LLNA

Determine the performance metrics for three expert systems

- Process all chemical structures though the three selected models, determining balanced accuracy, accuracy, sensitivity and specificity using two data sets collected, that arising from the OECD eChemPortal and a second from NICEATM

Assess the performance of the models

- Compare the global performance metrics of all models against the animal data and each other
- Compare the local performance metrics of all models against the animal data and each other

Gather data using the OECD eChemPortal

Results are reported for all animal data outcomes from the OECD eChemPortal which had a registered structure in DSSTox.

In parentheses are the total results of the substances with three conflicting results that were used in the analysis. It was necessary to separate these because substances with three conflicting results would be counted three times instead of only once, in a two dimensional table.

Results on the diagonal may or may not be from a single study.

Animal dataset gathered using the OECD eChemPortal

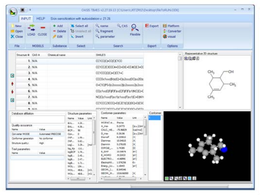
	Sensitizing LLNA result	Non-sensitizing LLNA result	Sensitizing GPMT result	Non-Sensitizing GPMT result
Sensitizing LLNA result	174			
Non-Sensitizing LLNA result	8 (13)	385		
Sensitizing GPMT result	37 (44)	3 (7)	143	
Non-Sensitizing GPMT result	9 (15)	35 (38)	16 (21)	475

Accuracy of the Animal Data

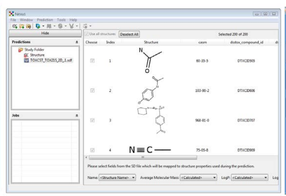
	Accuracy	Balanced Accuracy	Sensitivity	Specificity
LLNA predicting GPMT	86% (79%)	86% (79%)	93% (86%)	80% (72%)
GPMT predicting LLNA	86% (79%)	86% (80%)	80% (75%)	92% (84%)

An accuracy and balanced accuracy of 86% was found when comparing the results of the LLNA to the GPMT and vice versa. These metrics decreased to ~79% when data with 3 different outcomes was taken into account. The comparison of the 118 substances from eChemPortal were similar to the examination of 126 substances by ICCVAM which reported an accuracy of 86% for the GPMT and LLNA predicting one another, and a balanced accuracy of 84% for the GPMT predicting LLNA results while the LLNA predicting the GPMT had a balanced accuracy of 81%. The animal test results provide a range of balanced accuracy values, 79% to 86%, to use as a benchmark when assessing the performance of the selected expert systems.

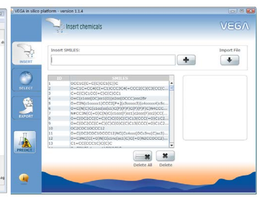
Selected Expert Systems



The TImes MEtabolism Simulator platform for predicting skin sensitization (TIMES-SS) is a hybrid expert system since it encodes structure-toxicity and structure-skin metabolism relationships, some of which are underpinned by mechanistic 3D QSARs.



Derek Nexus (Lhasa Ltd) is a knowledge based expert system containing ~ 89 alerts for skin sensitization. An alert consists of a toxicophore, associated literature references, comments and examples.



VEGA's skin sensitization model was derived using an adaptive fuzzy partitioning algorithm based on eight descriptors. The algorithm assigns substances into two classes; sensitizers and non-sensitizers.

Global performance of the expert systems

	eChemPortal dataset				NICEATM dataset			
	Accuracy	Balanced Accuracy	Sensitivity	Specificity	Accuracy	Balanced Accuracy	Sensitivity	Specificity
Derek	71%	65%	53%	76%	61%	61%	55%	66%
TIMES-SS	67%	63%	55%	71%	61%	61%	65%	57%
VEGA	44%	56%	80%	32%	57%	58%	76%	40%

For the global performance assessment, we sought to exclude any substances that were part of the training sets of any of the three models. This resulted in a set of 903 substances for the OECD eChemPortal set and the 180 substances for the NICEATM dataset. (Only 13 substances were in common to both datasets).

Derek Nexus gave rise to the highest accuracy (71%) and balanced accuracy (65%) for the eChemPortal data set. This balanced accuracy was 14 points lower than the balanced accuracy observed when comparing the animal tests. None of the expert systems performed as well as the animal tests on the basis of these 2 datasets.

Derek Nexus and TIMES-SS give rise to a significantly better performance than VEGA. This was because VEGA was the most conservative of the three systems, as evidenced by the high sensitivity but very low specificity values. This was not surprising given that the VEGA developers in their own evaluations had noted that all of the incorrect predictions made were false positives.

Examining Selected False Positive Predictions

CAS	Animal Test	Explanations of Incorrect Predictions
190965-45-8 	Negative	Alert for resorcinol or precursor. Resorcinol formation is deactivated by electron with drawing COAr group which would also deactivate resorcinol's oxidation.
313680-92-1 	Negative	Alert for epoxide as SN2. Epoxide is probably not reactive enough per Roberts et al (2017).
13504-15-9 	Negative	Alert for Michael acceptor. The metabolic activation is not significant in cutaneo.
6973-09-7 	Negative	Alert for imine methide formation and aromatic primary or secondary amine. Oxidation is de-activated by the electronegative SO2Me.

Local performance of the expert systems

	eChemPortal dataset				NICEATM dataset			
	Accuracy	Balanced Accuracy	Sensitivity	Specificity	Accuracy	Balanced Accuracy	Sensitivity	Specificity
TIMES-SS								
In Domain	80%	79%	75%	82%	85%	82%	86%	77%
Out of Domain	66%	63%	57%	69%	58%	58%	63%	53%
VEGA								
Good	40%	59%	98%	20%	71%	66%	91%	40%
Moderate	31%	51%	82%	19%	62%	54%	96%	11%
Low	51%	57%	71%	42%	55%	55%	75%	36%

All three expert systems have different approaches of characterizing their applicability domains and how this impacts the prediction derived.

Derek Nexus identifies alerts and assigns a level of confidence to the prediction from nine possible options. Substances that are predicted as plausible or higher are considered sensitizers. A level of confidence of certain or probable indicate that there is available experimental sensitization data in humans or other mammals within the system. TIMES-SS provides a convenient flag to indicate whether a substance is outside of its structural domain although there are other domain components included such as mechanistic, metabolic domains. VEGA provides a rating of good, moderate, or low to qualify the confidence in a prediction.

If VEGA has a prediction rated as good and the substance is predicted to be a non-sensitizer, it is almost certain to be a non-sensitizer. Most substances fall outside the domain of TIMES-SS, however those that lie within its domain do have a balanced accuracy that is comparable with the animal tests.

Conclusions

Accuracy of animal data for skin sensitization

- The balanced accuracies of the GPMT and LLNA data when compared to one another ranged from 79% to 86%.

Performance of the three expert systems

- Derek Nexus and TIMES-SS were most successful at predicting skin sensitizers from a global performance perspective.
- Substances that lay within the TIMES-SS domain gave predictions that had the same balanced accuracy as the animal data.

References

Check Out Our Paper: J. M. Fitzpatrick, D. W. Roberts & G. Patlewicz (2018) An evaluation of selected (Q)SARs/expert systems for predicting skin sensitisation potential, SAR and QSAR in Environmental Research, 29:6, 439-468, DOI: 10.1080/1062936X.2018.1455223

Derek Nexus: https://www.lhasalimited.org/products/derek-nexus.htm

TIMES-SS: http://oasis-lmc.org/products/software/times.aspx

VEGA: http://www.caesar-project.eu/

DSSTox information: https://comptox.epa.gov/dashboard

ICCVAM Report Comparing LLNA to GPMT: Sailstad, D. & Hattan, D. G. The Murine Local Lymph Node Assay: A Test Method for Assessing the Allergic Contact Dermatitis Potential of Chemicals/Compounds. National Toxicology Program, 1999 p. 13

Roberts et al (2017) CRT 30: 524-531

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