

Comparison of the results of QSAR models for Mutagenicity



Chiara Milan, Nazanin Golbamaki Bakhtyari, Alessandra Roncaglioni,
Antonio Cassano, Giuseppa Raitano, Emilio Benfenati

Istituto di Ricerche Farmacologiche Mario Negri, Laboratory of Environmental Chemistry and Toxicology
Via Giuseppe La Masa 19, 20156 Milan, Italy



Introduction

There is a number of models for **MUTAGENICITY**, which have been implemented.
The LIFE+ project **ANTARES** listed many of them.



<http://www.antes-life.eu/software.php>

We checked the performance of **eight of these models**, considering a large set of chemicals. The results have been also evaluated separately for the chemicals inside and outside the training set of each model, when possible, and inside and outside the applicability domain, when defined by the model itself.

Experimental

Experimental data have been taken from *Hansen et al. (2009) - Benchmark Data Set for in Silico Prediction of Ames Mutagenicity (version 2)*.

Table 1 lists the models we used.

| MODELS | FREE / COMMERCIAL | STATISTICAL / RULES BASED |
|---------|-------------------|---------------------------|
| ACD | COMMERCIAL | STATISTICAL |
| ADMET | COMMERCIAL | STATISTICAL |
| DEREK | COMMERCIAL | RULES BASED |
| SARPY | FREE | RULES BASED / STATISTICAL |
| TEST | FREE | STATISTICAL |
| TOPKAT | COMMERCIAL | STATISTICAL |
| TOXTREE | FREE | RULES BASED |
| VEGA | FREE | STATISTICAL |

Results

Figure 1 shows the results of the eight models for the total set of 6,512 compounds. We have to remember that the reproducibility of the Ames test is 85%, thus this value represents a kind of threshold for the overall accuracy of the model. The *a priori* split between mutagens and not is quite balanced. From this picture the models which are statistically based, such as ACD, T.E.S.T. and VEGA-CAESAR, seem to give better results than models based on the codification of human knowledge, such as DEREK and Toxtree.

A further evaluation has been done considering *chemicals which are inside or outside the training set*. This information is not available for all models, because the model may be based on rules (such as in the case of DEREK and Toxtree), or because the data are confidential. **Figure 2** shows the results.

Values are lower for the compounds not in the training set, as common, but results are still reasonable. The best models are the same as in **Figure 1**. Chemicals in the test sets ranged from 2,567 for VEGA and SARpy to 969 for ACD.

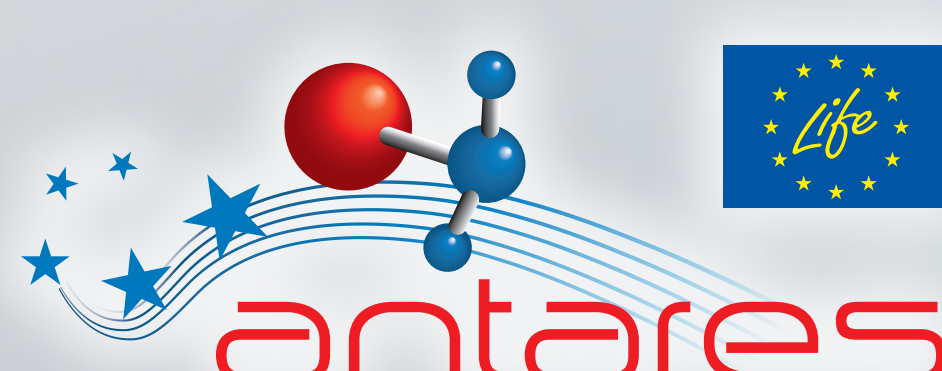
Figure 3 shows the results considering the chemicals *inside the applicability domain (AD) of the model*. This is a quite valuable information, since the user can appreciate the expected reliability of the model according to the declared boundaries of the model. Chemicals inside the AD ranged from 6,490 for T.E.S.T. to 5,178 for VEGA.

An even more interesting evaluation can be obtained for the *compounds which are in the test set, and inside the applicability domain*. This is shown in **Figure 4**. Comparing the overall results of the different models there are about 40 false negatives, for all eight models, and about 50 false positives. This figure is quite low, for the large number of chemicals we used (6,512). It is preferable to use more than one model, to reinforce the reliability.

Conclusions

This evaluation shows that there are models with good performance for mutagenicity, Ames test. The evaluation has been done considering chemicals not in the training set, and within the applicability domain of the model. The models have different algorithms and thus may provide different results. It may be useful to use more than one model and compare results.

Financial support from
ANTARES (EU Project LIFE08 ENV/IT/000435)
is gratefully acknowledged



www.antes-life.eu