A combined QSAR model for mutagenicity

T. Ferrari, G. Gini — *DEI, Politecnico di Milano, Italy*
A. Roncaglioni, E. Benfenati — *Istituto Mario Negri, Milan, Italy*
email: tferrari@elet.polimi.it

INTRODUCTION

Mutagenicity has to be evaluated for each chemical within the REACH legislation. Several experimental methods exist, and also *in silico* methods. Some *in silico* models use human knowledge, which has been codified into rules (expert systems), while other models extracts the information using data mining approaches.

Within the CAESAR project we developed a QSAR model for mutagenicity, with the purpose to produce a model suitable for REACH.

MATERIALS AND METHODS

In order to produce models for regulatory purposes it is important to address:

- data quality control
- model reproducibility
- reduction of *false negatives*
- model validation

The Data

A large data set of compounds (more than 4000) was used -as described by Katzius, McGuire and Bursi[1]- and each chemical structure was individually checked within the CAESAR project to increase the model robustness.

For each compound, 37 molecular descriptors were calculated using **MDL QSAR** software[2]: 4 of them are global descriptors and the others are small 2D substructures counts. The resulting data were finally normalized in the [-1,1] range.

Toxtree[3] was used on the same data set.

To provide a good basis for validation, the model was developed using a training set of about 80% of the compounds and the other compounds were left for testing.

The Algorithms

To achieve a model more suitable for REACH, two different cascading tecniques have been arranged: a machine learning algorithm, to build an early model with the best statistical accuracy, then an expert system to refine its predictions.

In the machine learning starting phase, a model was practiced on the training set using ***LibSVM***[4], an open source java implementation of *Support Vector Machines* (SVM) algorithm, with *Radial Basis Function* (RBF) kernel. The optimal parameterization for the kernel function was been automatically performed by a grid-search in the parameter space. This provided a basic model with very good performances.

Then, a further scan for structural alerts for genotoxic carcinogenicity on compounds predicted* as negative (*non-mutagenic*) by SVM model was carried out by **Toxtree**. An analysis on joined results permitted to identify two different subsets of Toxtree rules that better teams up with SVM predictions on two purposes: improving the accuracy and avoiding *false negatives*.

The final model, integration between the two approaches, can be designed like in figure.

(*) 10-folds cross-validation on the training set

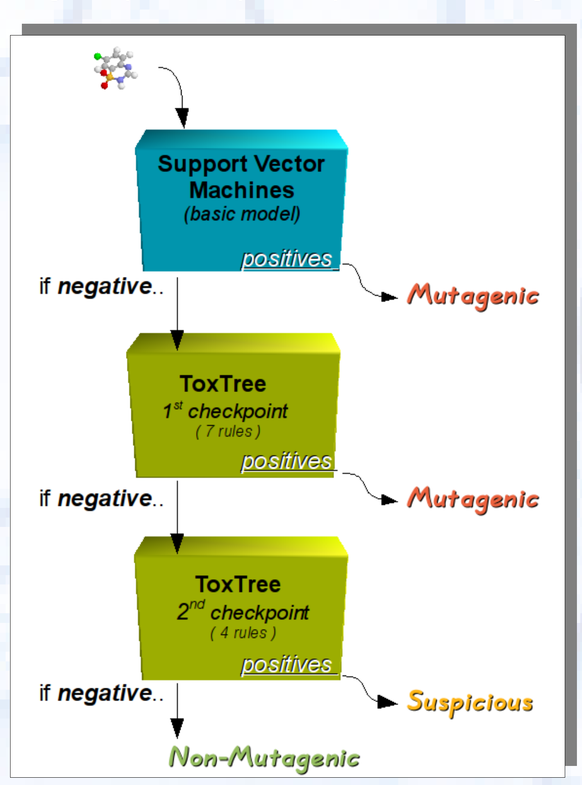


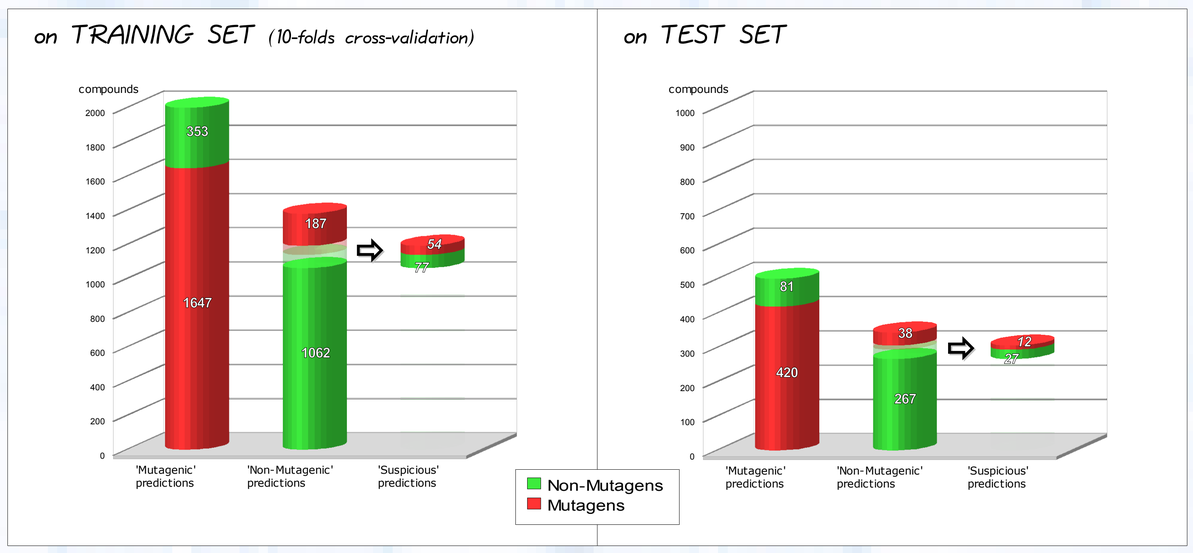
Figure. *Compounds evaluated as positive by SVM (■) are immediately classified as **Mutagenic**, the presumed negative are further tested by two consecutive checkpoints based on structural alerts (■) . The first has the aim to enhance the prediction accuracy attempting a precise identification of misclassified FN, the second goes on with the FN removal as much as this doesn't noticeably downgrade the original accuracy (generating too many FP as well). To point out this distinction, compounds picked out by the former checkpoint are classified as **Mutagenic**, and those picked out by the latter one are classified as **Suspicious**. Unaffected compounds are finally classified as **Non-Mutagenic**.*

RESULTS AND DISCUSSIONS

The combined model has three possible ouputs: *Mutagenic*, *Non-Mutagenic* or *Suspicious.* The *Suspicious* prediction it's a warning: it just denotes a candidate mutagen, since it has fired a structural alert but not a so specific one. The choice for the final binary classification is left to the end-user, depending on his/her purposes. This leads to two different statistic results with different features : *best accuracy* or *low false negative rate*.

Results of combined model validation are presented in the table below.

VALIDATION OF THE MODEL :



CHOICE
by end-user

Aimed at **performances** :
↳ *Suspicious compounds taken as **Non-Mutagenic***

OR

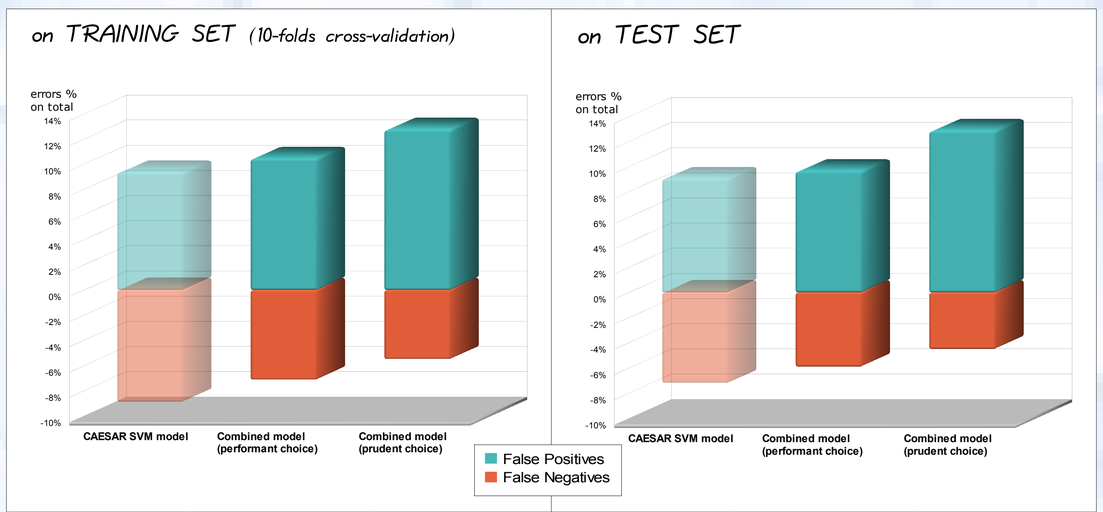
Aimed at **prudence** :
*Suspicious compounds taken as **Mutagenic*** ↵

Validation on the **TRAINING** set:	Validation on the **TEST** set:
SVM in 10-folds cross-validation + 7 Toxtree rules	SVM + 7 Toxtree rules
accuracy: 82.43%	*accuracy:* 84.50%
sensitivity: 87%	*sensitivity:* 89%
specificity: 76%	*specificity:* 78%

Validation on the **TRAINING** set:	Validation on the **TEST** set:
SVM in 10-folds cross-validation + 11 Toxtree rules	SVM + 11 Toxtree rules
accuracy: 81.75%	*accuracy:* 82.72%
sensitivity: 90%	*sensitivity:* 92%
specificity: 71%	*specificity:* 71%

Combined model has good accuracy and a low *false negative* rate. This effect is powered by the 'expert' *Toxtree layer* supervision on compounds at first predicted as *Non-Mutagenic* by the former *SVM layer*, as explicated in the table on the right.

ERROR-SHIFTING ANALYSIS :



SVM basic model:	SVM + 1 Toxtree checkpoint:	SVM + 2 Toxtree checkpoints:	SVM basic model:	SVM + 1 Toxtree checkpoint:	SVM + 2 Toxtree checkpoints:
Balanced errors	- 20 % FN + 12 % FP	- 38 % FN + 37 % FP	Balanced errors	- 18 % FN + 12 % FP	- 38 % FN + 50 % FP
Good accuracy (81.80 %)	+ 0.66 % accuracy	- 0.05 % accuracy (stable)	Good accuracy (84.26 %)	+ 0.24 % accuracy	- 1.54 % accuracy

CAESAR SVM basic model has higher performances than Toxtree only. However, when used as a further screening tool, Toxtree was useful to avoid a certain number of *false negatives*. This shows the utility of integrating models based on different approaches: SVM is based on data mining and chemical descriptors, while Toxtree is based on knowledge from human experts and chemical fragments.

We remember that the reproducibility of the experimental results is about 85%.

CONCLUSIONS

CAESAR SVM model gave good results on mutagenicity, with accuracy similar to the reproducibility of the experimental data. Strict quality check of the data has been done, and validation with about 850 compounds. The combination of CAESAR SVM model and Toxtree further reduced *false negative* rate.

REFERENCES

(1) J. Med. Chem. **2005**
(2) http://www.mdl.com/
(3) From from ECB, JRC. Software available at http://ecb.jrc.ec.europa.eu/qsar/
(4) Chang, CC. and Lin, CJ. LibSVM: a library for support vector machines, **2001**. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

We acknowledge Q. Chaudry, J. Cotterill (CSL, UK) and A. Chana (Mario Negri, Italy) for checking chemical structures.
We acknowledge financial contribution of the EC (project CAESAR).