The *In Silico* Model for Mutagenicity

Giuseppina Gini[1], Thomas Ferrari[1] and Alessandra Roncaglioni[2]

[1]DEI, Politecnico di Milano, Italy; [2]Istituto di Ricerche Farmacologiche Mario Negri, Milano, Italy

Summary

Mutagenic toxicity is the capacity of a substance to cause genetic mutations. This property is of high public concern, because it has a close relationship with carcinogenicity and reproductive toxicity.
In experiments, mutagenicity is assessed by the Ames test on Salmonella. The estimated inter-laboratory reproducibility rate of Salmonella test data is only 85%. This shows the intrinsic limitation of the in vitro test and opens the road to other assessments like an in silico model.
So far a widely used method is to check for the presence of structural alerts (SAs). However the presence of SAs alone is not a definitive method to prove the mutagenicity of the compound; the substituents can change the classification.
So statistically based methods are developed with the final target of obtaining a cascade of systems with tailored properties.
The integrated system has been developed on a set of a few thousand molecules.

Keywords: in silico methods, QSAR, intelligent systems

1 Introduction and questions

Mutagenic toxicity is the capacity of a substance to cause genetic mutations. This property is of high public concern, because it has a close relationship with carcinogenicity and reproductive toxicity (Benigni et al., 2008): most mutagenic substances are suspected carcinogenic substances in case a genotoxic mechanism is considered.

In experiments, mutagenic toxicity can be assessed by various test systems. One crucial point was the creation of cheap and short-term alternatives to the rodent bioassay, the main tool of the research on chemical carcinogens. With this intent Bruce Ames created a series of genetically engineered *Salmonella typhimurium* bacterial strains, each strain being sensitive to a specific class of chemical carcinogens (Ames, 1984). The Ames test is an *in vitro* model of chemical mutagenicity and carcinogenicity and consists of a range of bacterial strains that together are sensitive to a large array of DNA-damaging agents (Ashby, 1985).

An interesting point is the reliability of such experimental tests: as discussed in other papers (Piegorsch and Zeiger, 1991) the estimated inter-laboratory reproducibility rate of Salmonella test data is 85%. This observation will be taken into account in our model that will make use of the available data without using new *in vitro* testing.

Today regulators request the availability of mutagenic potency to correctly label and restrict mutagens/carcinogens and the exposure to them. Another important use of mutagenicity testing is in drug discovery, where mutagens should be stopped in the development of drugs. In environmental protection regulators need to understand the mutagenic potential of chemicals in order to control or limit their use.

Our aim is to develop a QSAR (Quantitative Structure Activity Relationship) model based on available data and to develop it for regulatory purposes; i.e. to reduce the number of false negatives. There is an argument that, if the main aim of QSAR modelling is simply prediction, the attention should be focused on model quality and not on its interpretation. Another argument is that it is dangerous to attempt to interpret models, since correlation does not imply causality, as discussed in (Livingstone, 2000).

On this basis, we can differentiate predictive QSARs, where the focus is best prediction quality, from descriptive QSARs, where the focus is descriptor interpretability. Our first aim is predictive QSAR; however our model is also quite interpretable.

The first step in making a QSAR model is the calculation of molecular descriptors. We limited our models to descriptors computed using the MDL software, including general descriptors and fingerprints. Fingerprints are used to encode structural characteristics of a chemical compound into a fixed bit vector (Durant et al., 2002).

QSARs have already been developed for mutagenicity. The availability of large data sets of non congeneric compounds, the most notable provided and analysed by Kazius et al. (2005), makes it possible to construct more robust models. A few other papers have been already published on this core data set (Liao et al., 2007; Zheng et al., 2006).

2 Materials and Methods

Two methods for predictions are used.
The first one consists in detecting in the molecule the particular structural fragments already known to be responsible for the toxic property under investigation. In the mutagenicity/carcinogenicity domain, the key contribution in the definition of such toxicophores comes from Ashby (Ashby, 1985) and is grounded on the electrophilicity theory of chemical carcinogenesis developed by Miller and Miller (1981). Every subsequent effort starts from knowledge collected by Ashby to derive more specific rules. It is important to mention that so far the mutagenicity structural alerts (SAs) are sound hypotheses that derive from chemical properties and have a sort of mechanistic interpretation; however their presence alone is not a definitive method to prove the mutagenicity of the compound towards Salmonella; the substituents present in some cases are able to change the classification. SAs are in practice rules that state the condition of mutagenicity given the presence and the absence of peculiar chemical substructures.

The second method we propose uses statistics, in particular an effective method to build non-linear models, as developed under the name Support Vector Machines (Vapnik, 1995).

The integration of both methods will give the final result, making the QSAR both predictive and interpretable.

2.1 Data

Following quality checks (IRFMN and CSL) the Kazius database was pruned and modified to 4225 compounds: 2358 classified as mutagens and 1867 classified as non-mutagens by the Ames test. For developing and evaluating the model we split them into a training set (80%) and a test set (20%). For each compound molecular descriptors were calculated with MDL-QSAR software, including both substructures and global descriptors.

A subset of 27 descriptors has been automatically selected with the BestFirst search method, using as subset evaluator the 5-fold cross-validation score on the training set. In short, BestFirst algorithm searches the space of attribute subsets by greedy hill climbing (considering all possible single attribute additions and/or deletions at a given point), with a backtracking facility to explore also non-improving nodes. The same subset of 27 descriptors has been obtained either searching forward, starting from the empty set, or with a bi-directional search starting from the 10 top rated attributes by a single attribute evaluator (Relief), both with 3 steps of backtracking.

The resulting dataset has been normalised by dividing each descriptor column by its maximum absolute value. Table 1 shows the selected substructural descriptors, while Table 2 shows the global descriptors. It is interesting to explain their meaning.

– *Gmin* = the minimum atom_level E-state value in a molecule. The E-State descriptor Gmin is a measure of the most electrophilic atom in the molecule and the polarity of the molecule. Mechanistically, an electrophilic centre is important for covalent bond formation with nucleophilic DNA, and so it is not surprising that Gmin is found to be important in modelling.

Tab. 1: The 23 local descriptors.
Modelling method for a statistical QSAR: Support Vector Machines

Symbol	Definition
$SsCH_3$_acnt	Count of all ($-CH_3$) groups in molecule
$SdCH_2$_acnt	Count of all ($=CH_2$) groups in molecule
$SssCH_2$_acnt	Count of all ($-CH_2-$) groups in molecule
SdsCH_acnt	Count of all (= CH –) groups in molecule
SaaCH_acnt	Count of all (CH) groups in molecule
SsssCH_acnt	Count of all (> CH –) groups in molecule
SdssC_acnt	Count of all (= C <) groups in molecule
SaasC_acnt	Count of all (= CH =) groups in molecule
SaaaC_acnt	Count of all (= CH =) groups in molecule
SssssC_acnt	Count of all (> C <) groups in molecule
$SsNH_2$_acnt	Count of all ($-NH_2$) groups in molecule
StN_acnt	Count of all (≡ N) groups in molecule
SdsN_acnt	Count of all (= N –) groups in molecule
SaaN_acnt	Count of all (= N =) groups in molecule
SsssN_acnt	Count of all (> N –) groups in molecule
SdaaN_acnt	Count of all (= N =) groups in molecule
SsOH_acnt	Count of all (– OH) groups in molecule
SdO_acnt	Count of all (= O) groups in molecule
SssO_acnt	Count of all (– O –) groups in molecule
SaaO_acnt	Count of all (≡ O ≡) groups in molecule
SHsOH_Acnt	Count of all [– OH] groups in molecule
SHother_Acnt	Count of all [other] groups in molecule
SHCHnX_Acnt	Count of all Halogen on C with 1 or 2 H atoms

– *idwbar* = Bonchev-Trinajstic mean information content based on the distribution of distances in the graph
– *LogP* = partition coefficient between octanol and water
– *nrings* = Number of rings in a molecular graph: cyclomatic number (i.e. the smallest number of bonds which must be removed such that no ring remains)

Atom types are classifications based on element and bonding environment. Atom type assignments are used in functional group identification, hydrogen addition, and hydrogen bond identification, and to determine VDW radii.

Except for the first capital "*S*", each lower case letter represents a bond:

– each "s" within an atom type designation represents a single bond to that atom
– each "d" within an atom type designation represents a double bond to that atom
– each "t" within an atom type designation represents a triple bond to that atom
– each "a" within an atom type designation represents an aromatic bond to that atom

We can observe that a few of them match known SAs.

The SdsN descriptor (for the nitrogen atom type \-N\=) is associated with the azo group, a structural alert. Molecules with larger SdsN descriptor values tend to have larger calculated output values.

SsssN is the atom count of all tertiary nitrogens in molecules. Tertiary nitrogen group alerts occur when the nitrogen is attached to either an aromatic or partially unsaturated ring. SaasC counts aromatic carbons with an attached substituent atom. It is not an alert per se; however, it reflects the nature of structural alerts attached to the ring system.

2.2. The statistical model

We created the statistical model using Support Vector Machines (SVM). SVM can use linear models to implement non-linear class boundaries. The input space is mapped into a higher (maybe infinite) dimensional space by a function ϕ, and a linear model constructed in the new space can represent a non-linear decision boundary in the original space. In the transformed space, the algorithm calculates the maximum margin hyperplane, i.e. the linear model that gives the greatest separation between the classes. The instances that are closest to it are called *support vectors*.

In our model we chose the *Radial Basis Function* as the kernel. A complete environment to develop SVM models is the open source LibSVM library[1], containing C++ and Java implementation of SVM algorithms with high-level interfaces (Python, Weka and more).

The optimal parameterisation of the model can be fully automated by one of the scripts included in LibSVM. With this tool it is possible to perform an almost exhaustive grid-search in the 2-dimensional parameter space of the objective function, using as evaluation criterion a cross-validation on the training set: the best assignment found was $(C, \gamma) = (8, 8)$.

With this parameterisation a model was trained on the training set, and its prediction ability was evaluated on the test set, normalised with the same scale factors used for the training set. Moreover, its robustness was assessed by a stratified 10-fold cross-validation. Table 3 reports the accuracy of the obtained model.

The accuracy of the model is very high. However, for the scope of CAESAR, we may try to reduce the FN rate. To this end we can apply another check and see if some widely known SAs can be used to detect other mutagens.

2.3. A model using SAs

As stated above, the available knowledge on mutagenicity is expressed in terms of SAs. A recent compilation of those alerts has been chosen as the knowledge base of our approach. To this end we considered the set of 30 SAs for mutagenicity derived by Benigni and Bossa from several literature sources. This rulebase is implemented as a module of Toxtree. (Developed by Ideaconsult Ltd. under the terms of a JRC contract. Software available at http://ecb.jrc.ec.europa.eu/qsar/: a java open source wrapper for structure-based predictions inclusive of a few other plugins on some toxicological endpoints.) In this realisation, the SAs are coded into SMARTS (SMiles ARbitrary Target Specification) strings, and the compounds in SMILES strings. Therefore the SAs detection is accomplished basically as a SMARTS[2] matching task. SMARTS strings are a text representation of substructures. To be matched, both the SMILES and the SMARTS strings are translated into graphs and the two graphs are compared.

The Benigni/Bossa rulebase was evaluated on the same set of 4225 chemical structures of CAESAR. The prediction ability on the entire data set, compared with the respective Ames Test results, is summarised in Table 4.

With respect to the SVM model, the SA model shows an increase of 48% of FP and an increase of 5% of FN. However, even if good sensitivity is exhibited by a low False Negative (FN) rate, the toxicity is often overestimated (low specificity), compromising the overall performance. This highlights the apparent drawback of the SAs: their compilation is engineered to individuate candidate mutagens by detecting the presence

Tab. 2: The 4 global descriptors

MDL code	Definition
MDL187	Smallest atom E-State value in molecule
MDL198	Bonchev-Trinajstic mean information content
MDL226	Calculated value of LogP
MDL230	Number of rings (cyclomatic number) in a molecular graph

Tab. 3: Accuracy of the SVM model

TRAINING SET *3380 chemicals*	**Predicted** ***mutagen***	**Predicted** ***non-mutagen***
mutagen	***1766***	***122***
non-mutagen	*137*	***1355***
Correct classification rate: ***92.3%***		
TRAINING SET *845 chemicals*	**Predicted** ***mutagen***	**Predicted** ***non-mutagen***
mutagen	***407***	***63***
non-mutagen	*79*	***296***
Correct classification rate: ***83.2%***		

[1] available at http://www.csie.ntu.edu.tw/\textasciitilde{}cjlin/libsv

[2] Daylight Theory Manual available at http://www.daylight.com/

of (presumed) toxicophores; just the remainder is labelled as non-mutagen.

3 Final results

So far we have seen how it is possible to achieve good prediction ability through a statistical approach, and also quite good results in predicting Salmonella mutagenicity through Toxtree software.

How can be these models be made more suitable for regulatory purposes? An answer is to address with special care the reduction of FNs, the hazardous compounds predicted as safe. There are various tricks to implement such enhancement in learning algorithms simply by throwing the model off centre, but all attempts in this direction will unavoidably raise several new FPs for each FN removed, since a just trained model is already in its best equilibrium. From here arose the idea of a trained classifier supervised by an expert layer: the aim is to refine the good statistical separation between classes supplied by SVM, not by introducing a perturbation in the optimality of the model, but by applying a complementary knowledge-based filter in order to allow an accurate identification of misclassified mutagens (FNs), even if isolated. In other words, instead of making the model more (or too much) sensible to mutagens, our intent is to equip it with an additional device to be applied to non-mutagenic predictions, skilled in what it has had difficulty to learn.

In practice, although every SA should be trusted if evaluated on a random set of molecules, here we are considering only that portion of compounds already presumed non-mutagenic by SVM, i.e. cleaned for the most part of mutagens. This means that while the FP rate spawned by each rule is unchanged (since all the non-mutagens should still be present), the rate of caught true positives (TPs) will decrease, because just a few mutagens are left. Hence a selection is needed to extract just a subset of the rulebase skilled in finding the mutagens potentially subject to misclassification by the SVM model. Having such a large data set, the above selection can be carried out in a straightforward way. The predictions obtained by cross-validating the model on the training set (3380 compounds) shall be representative of its general prediction ability, so a filter fixing the inaccuracies of such a meta-model will probably provide even for defects of the original one.

An integrated model was arranged cascading the two techniques: a trained SVM classifier with an additional expert facility for FNs removal based on SAs. The SVM classifier is the one described previously, while the rulebase for the expert filter was extracted from the Benigni/Bossa SAs set after an analysis of their individual effect, evaluated on those structures of the training set labelled non-mutagenic by 10-fold cross-validating the model.

This spotlights two different subsets of SAs (see Appendix). The former (10 SAs) is the set of "good" rules: each of them showed a balance of more FNs caught than FPs spawned once evaluated on the cross-validated predictions. Their supposed capacity to refine the SVM model prediction ability is confirmed by the proof on the test set. The latter (5 SAs) is the set of "suspicious" rules, i.e. those ones with a still remarkable FNs removal power but a higher misclassification rate. As can be seen in Table 5, the FNs removal carried out by the first set of rules improved both sensitivity and accuracy similarly, either in the calibration on the training set or the validation on the test set. 13% of FNs are cleaned from the "safe" prediction on the test set and this benefits even the correctly classified prediction rate with a slight increment.

By applying the second set of rules (Tab. 6), the performances in classification accuracy are not noticeably downgraded if compared with those of the basic SVM model, but about a third of FNs (32%) are removed from the overall predictions, boosting the sensitivity over 90%.

The statistics on the final model are described in Table 7.

A global overview of the performances of the combined model is illustrated in Figure 1, where an interpretation of the set of "suspicious" rules is given: it can extract the more suspect

Tab. 4: Results of the SA model on the test set

TEST SET *845 chemicals*	Predicted *mutagen*	Predicted *non-mutagen*
mutagen	***404***	***66***
non-mutagen	*117*	***258***
Correct classification rate: ***78.3%***		

5: Confusion matrix on the test set after using the first set of 10 rules

TEST SET *845 chemicals*	Predicted *mutagen*	Predicted *non-mutagen*
mutagen	***427***	***43***
non-mutagen	*109*	***267***
Correct classification rate: ***82.1%***		

Tab. 6: Confusion matrix after applying both sets of rules

TEST SET *845 chemicals*	Predicted *mutagen*	Predicted *non-mutagen*
mutagen	***415***	***55***
non-mutagen	*86*	***289***
Correct classification rate: ***83.3%***		

Tab. 7: Statistics about the cascade model

CAESAR test set	**Suspicious taken as non-mutagenic**	**Suspicious taken as mutagenic**
accuracy:	***83.3%***	***82.1%***
sensitivity:	***88.3%***	***90.9%***
specificity:	***77.1%***	***71.2%***

compounds from "safe" prediction with good accuracy, if related to the very low number of real mutagens still present. The so obtained global model for mutagenicity prediction has been released through the portal of the CAESAR project.

Our final cascade model is described in Figure 2. Its logic flow is very clear, and its statistical analysis defined.

4 Discussion

Our hypothesis that a QSAR approach was a good method to build models of non-congeneric compounds has been proven. The two-step method proposed in our CAESAR model for mutagenicity demonstrated that the QSAR method is more apt to screen the data set than the SAs approach.

In our implementation, besides improving accuracy, we are also biased toward reducing the number of false negatives, as required by regulators. As we have seen, the first screening is based on statistical correlation between small fragments and the mutagenicity property, and we only check the presence of SAs upon a negative outcome. Since regulators use this method manually, we are so able to provide them with a similar check.

However, there are important differences between our method and the traditional SAs. As we already pointed out, the known SAs are biased toward pollutants and carcinogenic molecules; in our approach we can use only the first screening to deal with drugs and other families of compounds. SAs are a fixed list of substructures, while the MDL keys used in the QSAR phase are in a number automatically selected to give better performance to the correlation with the endpoint. So the keys are automatically derived from the reduction algorithm, not extrapolated from human experts. They can cover or not the known SAs. In this sense we are really performing data mining and deriving a set of keys that can become, in principle, new SAs in case the chemical classes considered are new and an interpretation about reactivity is available.

In terms of accuracy our model, which uses powerful algorithms, can reach accuracy very near to the rate of the reproducibility of the experimental data in different laboratories.

In terms of interpretability of our model, the first step can be understood in terms of the few global descriptors used and the MDL keys. However we should remember that the interpretability of non-linear models does not depend on simple relations between input and output, and the mix of the descriptors cannot be translated into rules. The second step is obviously defined in terms of rules, stating that the presence of any of the SAs and the absence of external conditions would put the compound in the mutagenicity class.

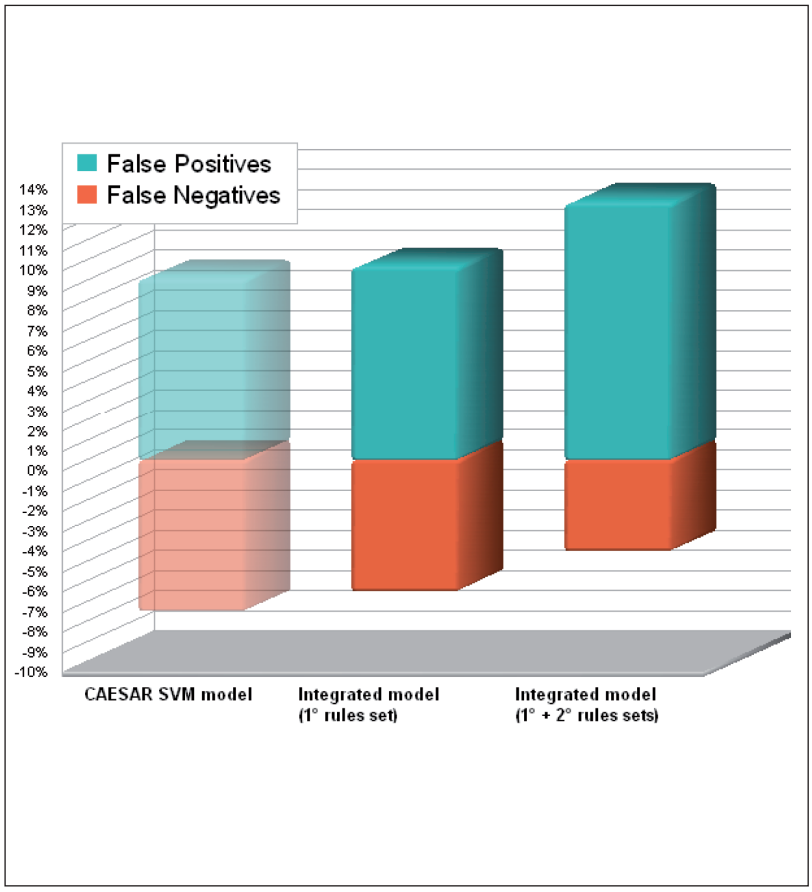


Fig. 1: Reduction of false negatives

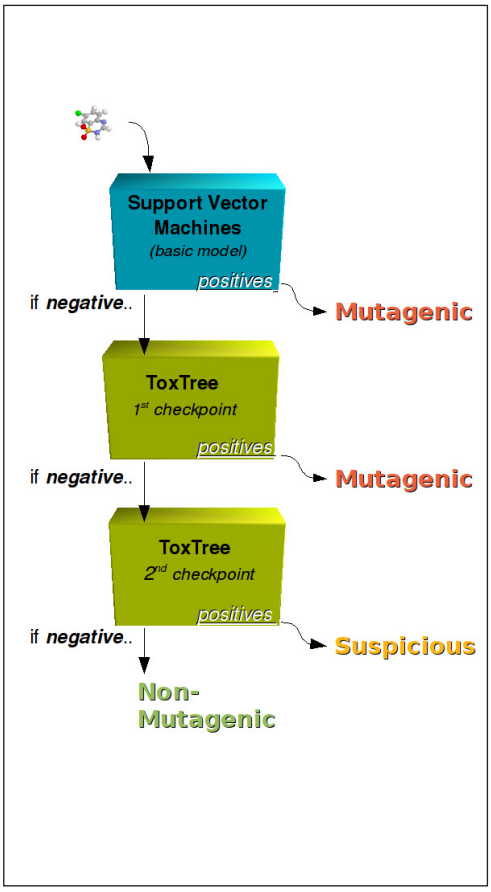


Fig. 2: The cascade model

References

Ames, B. N. (1984). The detection of environmental mutagens and potential. *Cancer 53*, 2030-2040.

Ashby, J. (1985). Fundamental structural alerts to potential carcinogenicity or noncarcinogenicity. *Environ. Mutagen 7*, 919-921.

Benigni, R., Bossa, C., Jeliazkova, N. G. et al. (2008). The Benigni/Bossa rulebase for mutagenicity and carcinogenicity – a module of toxtree. Technical Report EUR 23241 EN, European Commission – Joint Research Centre.

Chang, C.-C. and Lin, C.-J. (2001). LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Hsu, C.-W., Chang, C.-C., and Lin, C.-J. A practical guide to support vector classification. Download from http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.

Kazius, J., Mcguire, R. and Bursi, R. (2005). Derivation and validation of toxicophores for mutagenicity prediction. J. Med. Chem. 48(1), 312-320.

Durant, J. L., Leland, B. A., Henry, D. R. and Nourse, J. G. (2002). Reoptimization of mdl keys for use in drug discovery. J. Chem. Inf. Comput. Sci. 42, 1273-1280.

Liao, Q., Yao, J. and Yuan, S. (2007). Prediction of mutagenic toxicity by combination of recursive partitioning and support vector machines. *Molecular Diversity 11*, 59-72.

Livingstone, D. J. (2000). The characterization of chemical structures using molecular properties: a survey. *J. Chem. Info Comput. Sci. 40(2),* 195-209.

Miller, J. A. and Miller, E. C. (1981). Searches for ultimate chemical carcinogens and their reactions with cellular macromolecules. *Cancer 47*, 2327-2345.

Piegorsch, W. W. and Zeiger, E. (1991). Measuring intra-assay agreement for the ames salmonella assay. In L. Hotorn (ed.), *Statistical Methods in Toxicology, Lecture Notes in Medical Informatics*. (35-41). Springer-Verlag.

Vapnik, V. (1995) *The Nature of Statistical Learning Theory*. Springer-Verlag.

Zheng, M., Liu, Z., Xue, C. et al. (2006). Mutagenic probability estimation of chemical compounds by a novel molecular electrophilicity vector and support vector machine. *Bioinformatics 22(17),* 2099-2106.

Acknowledgements

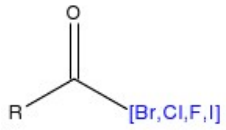
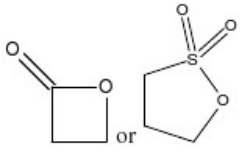
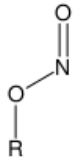
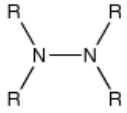
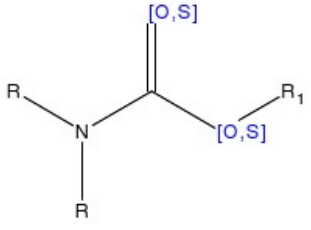
Partial support for his work has been provided through the CAESAR project of the European Union.

Correspondence to

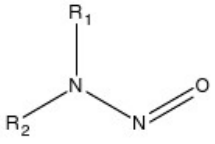
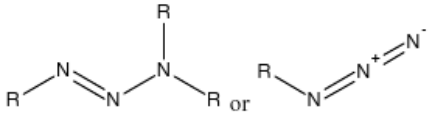
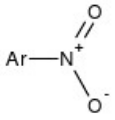
Giuseppina Gini
DEI, Politecnico di Milano
piazza L. da Vinci 32
20133 Milano, Italy
e-mail: gini@elet.polimi.it

The *In Silico* Model for Mutagenicity
Appendix A

1st Ruleset:

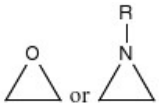
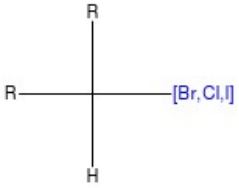
Structural Alert	SMARTS or details
SA_1: Acyl halides  R = any atom/group, except OH, SH	[!$([OH1,SH1])]C(=O)[Br,Cl,F,I]
SA_6: Propiolactones or propiosultones 	[O,S]=C1[O,S]CC1 OR O=S1(=O)(CCCO1)
SA_9: Alkyl nitrite  R= any alkyl group	O=[NX2]OC
SA_13: Hydrazine 	[N+0]!@;-[N+0](=[!O;!N]) OR [N+0]([#1,*])!@;-[N+0]([#1,*])
SA_16: alkyl carbamate and thiocarbamate  R = Aliphatic carbon or hydrogen R1 = Aliphatic carbon	[NX3]([CX4,#1])([CX4,#1])C(=[O,S])[O,S][CX4]
SA_18: Polycyclic Aromatic Hydrocarbons	Three or more fused rings, not heteroaromatic

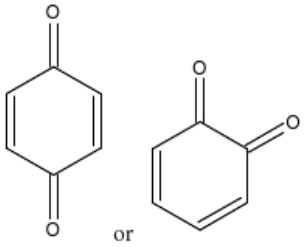
The *In Silico* Model for Mutagenicity
Appendix A

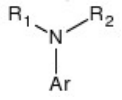
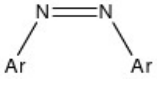
Structural Alert	SMARTS or details
SA_19: Heterocyclic Polycyclic Aromatic Hydrocarbons	Three or more fused rings, heteroaromatic
SA_21: alkyl and aryl N-nitroso groups  R1= Aliphatic or aromatic carbon, R2= Any atom/group	[C,c]N[NX2;v3]=O
SA_22: azide and triazene groups  R= Any atom/group	[N]=[N]-[N] OR [N]=[N]=[N]
SA_27: Nitro-aromatic  Ar = Any aromatic/heteroaromatic ring	• Chemicals with ortho-disubstitution, or with an ortho carboxylic acid substituent are excluded. • Chemicals with a sulfonic acid group (-SO3H) on the same ring of the nitro group are excluded .

The *In Silico* Model for Mutagenicity
Appendix B

2nd Ruleset:

Structural Alert	SMARTS or details
SA_7:Epoxides and aziridines  R = any atom/group	`C1[O,N]C1`
SA_8: Aliphatic halogens  R = any atom/group	

SA_12: Quinones
 
 	
R = any atom/group
	`O=[#6]1[#6]=,:[#6][#6](=O)[#6]=,:[#6]1`
	OR
	`O=[#6]1[#6]=,:[#6][#6]=,:[#6][#6]1(=O)`

SA_28bis: Aromatic mono- and dialkylamine  Ar = Any aromatic/heteroaromatic ring R1 = Hydrogen, methyl, ethyl R2 = Methyl, ethyl	• Chemicals with ortho-disubstitution, or with an ortho carboxylic acid substituent are excluded. • Chemicals with a sulfonic acid group (- SO3H) on the same ring of the amino group are excluded .
SA_29: Aromatic diazo  Ar = Any aromatic/heteroaromatic ring	• Chemicals with a sulfonic acid group (-SO3H) on both rings linked to the diazo group are excluded.