Chapter 5

In Silico Prediction of Chemically Induced Mutagenicity: How to Use QSAR Models and Interpret Their Results

Enrico Mombelli, Giuseppa Raitano, and Emilio Benfenati

Abstract

Information on genotoxicity is an essential piece of information gathering for a comprehensive toxicological characterization of chemicals. Several QSAR models that can predict Ames genotoxicity are freely available for download from the Internet and they can provide relevant information for the toxicological profiling of chemicals. Indeed, they can be straightforwardly used for predicting the presence or absence of genotoxic hazards associated with the interactions of chemicals with DNA.

Nevertheless, and despite the ease of use of these models, the scientific challenge is to assess the reliability of information that can be obtained from these tools. This chapter provides instructions on how to use freely available QSAR models and on how to interpret their predictions.

Key words Mutagenicity, Ames test, QSAR, Predictive reliability, Structural alerts

1 Introduction

The assessment of information on mutagenicity represents an important component for the evaluation of the toxicological characteristics of chemicals [1]. For instance, in the field of drug discovery the detection of mutagenic potential of a chemical can result in the rejection of a promising chemotype owing to the deleterious consequences that the introduction of gene mutations can elicit. In addition, the characterization of genotoxicity is required for the regulatory qualification of impurities in drug substances [2] and it is a mandatory requirement for all the different tonnage bands defined by the overarching REACH regulation [3].

Mutagenic effects caused by chemical agents can be detected by the Ames test that was devised by Bruce Ames during the 1970s [4]. This test is still commonly in use in many toxicological laboratories around the world because of its good interlaboratory reproducibility, aptitude at testing different agents, cost-effectiveness, and structure–activity analysis [5]. The remarkable juxtaposition of these attributes has brought the Ames test to the forefront

Emilio Benfenati (ed.), *In Silico Methods for Predicting Drug Toxicity*, Methods in Molecular Biology, vol. 1425,
DOI 10.1007/978-1-4939-3609-0_5, © Springer Science+Business Media New York 2016

of modern toxicology. Indeed, this test is a paradigm for the development of nowadays in vitro toxicology and it has been nicknamed "the stethoscope of genetic toxicology for the twenty-first century" [5] given that testing strategies for carcinogenicity rely on the Ames test as an essential first-tier assay [5, 6].

This test is based upon the ability of *Salmonella typhimurium* and *Escherichia coli* auxotrophic strains to recover the ability to synthesize an essential amino acid (histidine for *S. typhimurium* and tryptophan for *E. coli*) as a consequence of the mutagenic effect of chemicals to which they are exposed. The design of the experimental protocol enables the detection of bacterial colonies that can grow in the absence of essential amino acids as a result of a back mutation that restores their biosynthetic capabilities. The detection of this back mutation to wild type has the potential to identify point mutations that are caused by the substitution, addition, or deletion of one or few DNA base pairs. At least five bacterial strains should be used when testing a chemical [7], including strains that are sensitive to oxidizing mutagens, cross-linking agents and hydrazines (*E. coli* WP2 or *S. typhimurium* TA102, *see* **Note 1**).

Anyhow, it is important to note that, as stated in the OECD guideline [7], mammalian tests may be more appropriate when evaluating certain classes of drugs. For example, the Ames test is not the most appropriate choice for chemicals displaying a high bactericidal activity such as certain antibiotics, topoisomerase inhibitors, and some nucleoside analogs.

The interlaboratory reproducibility of the Ames test is estimated at 85–90 % [8, 9] and these percentages represent the upper limit of predictive performance that can be expected from QSAR models for the same endpoint. Indeed, these models are derived from data obtained by means of the same protocol. In other words, these findings mean that 10–15 % of the chemicals that were experimentally tested gave different results when analyzed in different laboratories. Therefore, this experimental uncertainty in terms of false negative or positive predictions is transposed into the semi-empirical QSAR models that cannot be expected to be more reliable than their experimental counterpart.

Consequently, one key issue that should be given attention when judging the reliability of a QSAR model predicting Ames genotoxicity is whether or not this model predicts with a reliability that is comparable to the reproducibility of the test. It is worth mentioning that this comparison has to be critically assessed as a function of the number and chemotypes of the chemicals that compose the external test set that was adopted in order to validate the model. For example, if the external test set does not include all the chemotypes that are covered by the training set, the estimated predictive performance of the model will only be representative of a subset of chemical structures.

One final word of caution should be added with respect to models whose alleged performance is much higher than the experimental test they are meant to replace. This special situation could indicate a potential overfitting of the model and its lack of ability to provide reliable prediction for new cases (i.e. molecules that are not included in its training set).

The theory of electrophilic reactivity by Miller and Miller [10] adequately describes the molecular mechanisms that control the genotoxicity of chemicals as detected by the Ames test. Indeed, this theory has proved to be in agreement with the observations ever since it was formulated in the late 1970s. According to this theory, the vast majority of known chemical carcinogens are also genotoxic since they are (or are metabolized to) reactive electrophiles that react with nucleic acids. The (Q)SAR models described in this chapter (*see* Subheading 2.6) conform to this theory by identifying structural fragments that trigger electrophilic reactions as formalized by E-state values and fragments (e.g. CAESAR) and by structural alerts (SA) validated by experts (e.g. Toxtree SA) or automatically extracted by learning algorithms (e.g. SARpy).

Because of the complementary nature of these tools, this chapter illustrates the practical application of models covering the three main categories of in silico tools for the prediction of the mutagenic potential of chemicals: (Q)SAR models that are based on numerical descriptors (e.g. partition coefficients, topological descriptors, functional group counts), rule-based expert systems that are based on structural alerts (molecular fragments that are associated with the occurrence of adverse outcomes), and hybrid models combining these two approaches. Models based on all these approaches are implemented within the freely available VEGA platform (version 1.0.8): CAESAR, SARpy, and ToxTree-VEGA (TT-VEGA) (*see* **Note 2**). A brief description of the models is given in the following paragraphs and more detailed information can be found in the literature therein cited.

2 Materials

2.1 Performance Characterization of (Q)SAR Models

The performance of models predicting the presence and absence of toxicological hazards is usually described by Cooper statistics [11] that characterize the predictive capabilities of diagnostic tests: sensitivity, specificity, and accuracy (or concordance). Sensitivity is the ability to identify a chemical that presents a toxicological hazard as toxic; specificity is the ability to correctly identify chemicals that do not present toxicological hazards as safe; and accuracy describes the overall concordance between predicted and experimental values. Their mathematical definitions are the following:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Concordance} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}$$

where TP = number of true positive predictions, TN = number of true negative predictions, FP = number of false positive predictions, FN = number of false negative predictions.

In the presence of skewed data sets (e.g. a data set including a majority of non-mutagenic chemicals), Cooper statistics are not fully reliable. It is therefore more appropriate to compute the Matthews correlation coefficient (MCC) which is defined as follows:

$$\text{MCC} = \frac{\text{TPTN} + \text{FPFN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

The MCC ranges from −1 to +1. A MCC value of +1 represents a total agreement between experimental results and predictions; a value of 0 no better than random prediction, and a value of −1 indicates a total disagreement between predicted and observed values.

2.2 Software Requirement

TTVEGA, CAESAR, and SARpy models are embedded within the standalone software application VEGA (v. 1.0.8) that allows for a secure in-house execution of the three models without the need to send information to any external server [12]. VEGA can be also used for batch processing of multiple chemical structures. The software application can be freely downloaded for the VEGA website [12] and it can be installed and used on any operative system supporting JAVA.

2.3 Optional Software

Any software application that allows to draw chemical structures and convert them into two types of chemical file formats supported by VEGA: “Simplified Molecular Input Line Entry specification” (SMILES) [13] or “Structure Data Format” (SDF) can be used in order to generate input structures. Several chemical drawing programs can perform this task: VEGA ZZ [14], ACD/ChemSketch [15], MarvinSketch [16], and the OECD QSAR Toolbox [17] (for SMILES formats only).

This list is not exhaustive and these applications are subjected to different software licenses and terms and conditions of use.

2.4 VEGA: The Workflow

VEGA has a simple workflow which is schematically depicted in Fig. 1. Basically, a user types or pastes a SMILES string in the blank space at the top of the user interface and then adds it to a working list of molecules to be analyzed. Once that a SMILES string is added at the working list, it is possible to highlight it and visually

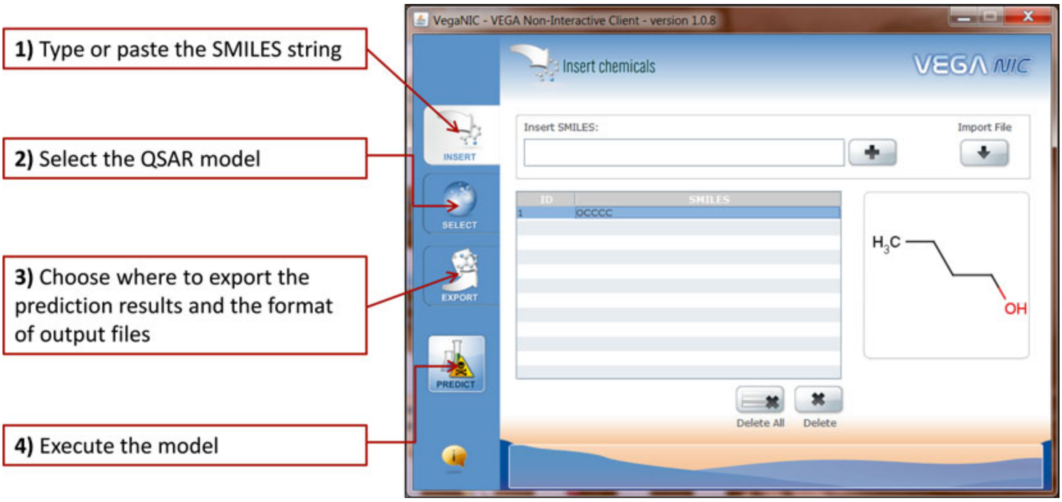

1) Type or paste the SMILES string
2) Select the QSAR model
3) Choose where to export the prediction results and the format of output files
4) Execute the model
VegaNIC - VEGA Non-Interactive Client - version 1.0.8
Insert chemicals
VEGA NIC
INSERT
SELECT
EXPORT
PREDICT
Insert SMILES:
Import File
ID
SMILES
1
OCCCC
H_3C
OH
Delete All
Delete

Fig. 1 Workflow of VEGA. *The SMILES string corresponding to 1-butanol was used as input structure*

check the two-dimensional structure encoded by the text line. This checkpoint is crucial. Indeed, several structural inaccuracies can take place at this stage and compromise the reliability of the predictions [18].

If needed, users can also input multiple molecules at once (“import File” button at the top right of the user interface). In this case the file contains a list of SMILES codes saved in “txt” or “smi” format.

Thanks to the “Select” button it is then possible to choose the model(s) of interest, to specify the desired output format (PDF or csv files), and to indicate where the prediction reports should be saved (“Export” button). Finally, the selected model(s) can be executed by clicking on the “Predict” button.

2.5 Applicability Domain

All the models that will be described in the following paragraphs adopt the VEGA definition of applicability domain [19] (*see* **Note 3**). According to this definition, the degree of membership of a query chemical to the applicability domain of the model is described by an Applicability Domain Index (ADI) with values that range from 0 (no membership) to 1 (full membership). Chemicals characterized by ADI values that are less than 0.7 are to be regarded as potentially not belonging to the AD. ADI values that are within the range 0.7–0.9 represent a critical region since the query chemical could be out of the applicability domain. Finally, ADI values that are greater than or equal to 0.9 indicate chemicals that should be regarded as belonging to the applicability domain of the model.

These reference values represent a general guideline and they should be interpreted in the light of a thorough inspection of the sub-indexes that compose the ADI: the similarity index, the concordance index, the accuracy index, and the atom-centered fragments index. If, as in the case of the CAESAR model, the chemical structures are characterized by numerical descriptors the ADI takes also into account a check of the ranges in descriptor values (*see* **Note 4**).

These critical factors should always be analyzed when interpreting results and they will be described in the following paragraphs.

2.5.1 Similarity Index

This index takes into account the degree of similarity between the query chemical and the three most similar chemicals. Values close to 1 indicate that the chemotype of the query chemical is well represented by the training set of the model (*see* **Note 5**). On the other hand, lower values could indicate that the prediction is an extrapolation since the query chemicals is located in regions of the chemical space that are scarcely populated. In this case the prediction cannot be supported by the evaluation of similar chemicals. This does not mean that the prediction is wrong. It means that the user should gather further elements to support the model results. In particular, additional models should be run to get support.

2.5.2 Concordance Index

This index provides information on the concordance between the predicted value for the query chemical and the experimental values of the three most similar chemicals. Values that are close to zero may indicate an unreliable prediction and the possible identification of a region in the chemical space whose structure–toxicity behavior is not adequately described by the model. Therefore, a careful inspection of chemicals that give rise to conflicting predictions is requested. Indeed, one or more structural analogs can be characterized by experimental values that are at odds with the prediction for the target compound.

For instance, a visual inspection may easily identify the presence of a specific toxic SA within the structure of the structural analog(s).

Consequently, two compounds that are similar from a chemical point of view may differ for the presence/absence of structural alerts, and this fact can explain differences in their property values.

If the user does not recognize SA, it is possible to run VEGA on the similar compound with the conflicting value; VEGA will list the SA, which can then be compared.

2.5.3 Accuracy Index

When assessing the reliability of predictions, it is important to understand how well a model predicts the toxicity in the region of the chemical space where the query chemical is located. This index informs on such a local reliability by taking into account the classification accuracy of the three most similar chemicals. Low values for this index should warn about a lack of predictive accuracy. In this case, additional models should be run, to see if they have better accuracy.

2.5.4 Atom-Centered Fragments (ACF) Index

This index takes into account the presence of one or more fragments that are not found in the training set, or that are rare fragments. An index value equal to 1 implies that all atom-centered fragments of the target compound were found in the training set. On the other hand, a value that is less than 0.7 implies that a prominent number of atom-centered fragments of the target compound have not been found in the compounds of the training set or are rare fragments of the training set.

Also in this case, it is recommended to run additional models, because each model can bring new information as a function of its own training set.

2.5.5 Model Descriptors Range Index

Computed only for the CAESAR model, this index checks if the descriptors calculated for the predicted compound are inside the range of descriptors of the training and test set. The index has value 1 if all descriptors are inside the range, 0 if at least one descriptor is out of the range.

2.6 Models Description

To compare the performance of three VEGA models, we applied them to the same evaluation set. This data set counts more than 6000 compounds evenly distributed between mutagens and non-mutagens and was used within the European LIFE project ANTARES for the evaluation of different QSAR models [20].

In the next paragraphs, for each model we report the statistical values referred to the entire evaluation set (6064 compounds) and to the molecules belonging to the applicability domain that are out of its training set.

2.6.1 Benigni-Bossa Mutagenicity (TT-VEGA)

TT-VEGA (version 1.0.0-DEV) is based on a series of rules defined by Benigni and Bossa that detects mutagenic chemicals [21]. This rulebase was originally implemented within the Toxtree application freely distributed by the European Joint Research Center [21].

Toxicity Data Source

Data were extracted from the ISSCAN database [22] and includes 730 compounds, 350 of which are mutagenic.

Description of the Model

Toxtree is a rule-based system that includes alerts for genotoxic carcinogenicity and non-genotoxic carcinogenicity. Genotoxic carcinogenicity alerts can be considered as a valuable tool for the detection of compounds that yield positive results during an Ames test. The version of Toxtree implemented within the VEGA platform offers the same compilation of rules as the original version [21]. This model offers a compilation of SA that refers mainly to knowledge on the mechanism of action for genotoxic carcinogenicity (i.e. they are also pertinent for mutagenic activity in bacteria). The SAs detecting non-genotoxic carcinogens are not to be taken into account when applying this model since non-genotoxic carcinogens cannot, by definition, be detected by the Ames test.

Model Statistics

- **Global performance** (calculated on 6064 compounds):
- Accuracy = 0.75, Specificity = 0.65, Sensitivity = 0.83, MCC = 0.49.
- **Performance in ADI out of training** (calculated on 1419 compounds with ADI > 0.9):
- Accuracy = 0.87, Specificity = 0.75, Sensitivity = 0.94, MCC = 0.72.

Interpretation of the Output

TT-VEGA classifies query chemicals as mutagenic when one or more SAs are detected within their molecular structure or as a non-mutagenic if no SA is identified.

2.6.2 Mutagenicity Model (CAESAR) (Version 2.1.12)

Toxicity Data Source

The CAESAR model [23] was developed on the basis of 4204 chemicals (2348 mutagenic and 1856 non-mutagenic) extracted from the Bursi data set [24]. This initial set was then split into training set (3367 chemicals, 80 % of the entire data set) and external test set (837 chemicals, 20 % of the entire data set) [24].

Description of the Model

The algorithm of the model is described in Ferrari and Gini [23]. CAESAR-VEGA automatically calculates chemical descriptors for the chemicals of interests and contains a subset of Toxtree rules (*see* previous paragraph) to enhance the sensitivity of the model. The model integrates two complementary predictive approaches in series (statistical and rule-based): a support vector machine (SVM) algorithm coupled to two sets of structural alerts rules aimed at reducing the false negative rate. In order not to inflate the false positive rate a chemical which is identified as negative during the first two steps (SVM output and first SA filter) and positive by the second set of rules is flagged as a suspicious mutagenic chemical.

If the user wants only the results of the statistical model, (s)he can check if the model identifies SA and discard this approach.

Model Statistics

- **Global performance** (calculated on 6064 compounds):
- Accuracy = 0.81, Specificity = 0.69, Sensitivity = 0.91, MCC = 0.63.
- **Performance in ADI out of training** (calculated on 942 compounds with ADI > 0.9):
- Accuracy = 0.79, Specificity = 0.61, Sensitivity = 0.93, MCC = 0.57.

During this evaluation, compounds predicted as suspicious mutagens were considered as mutagens.

Interpretation of the Output

CAESAR-VEGA classifies chemicals as mutagenic, non-mutagenic, and suspicious mutagenic. Suspicious chemicals are associated with higher predictive uncertainty.

2.6.3 Mutagenicity SARpy Model (Version 1.0.6—DEV)

Toxicity Data Source

The data set employed for rule extraction was retrieved from the CAESAR model for Ames mutagenicity (*see* previous paragraph). This model and VEGA CAESAR share the same training set.

Description of the Model

SARpy (SAR in python) is a QSAR method that identifies relevant fragments and extracts a set of rules directly from data without any a priori knowledge [25]. The algorithm generates substructures; relevant SAs are automatically selected on the basis of their prediction performance for a training set. The application of this modeling approach to the CAESAR data set extended the previous work [25] by extracting two sets of rules: one for mutagenicity (112 rules) and the other for non-mutagenicity (93 rules) (*see* **Note 6**).

The SARpy application is available through a graphic interface or through the VEGA platform.

Model Statistics

- **Global performance** (calculated on 6064 compounds):
- Accuracy = 0.77, Specificity = 0.71, Sensitivity = 0.82, MCC = 0.54.
- **Performance in ADI out of training** (calculated on 880 compounds with ADI > 0.9):
- Accuracy = 0.81, Specificity = 0.67, Sensitivity = 0.92, MCC = 0.62.

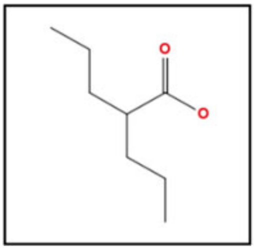
Interpretation of the Output

If the target compound matches one or more mutagenicity rules, the prediction will be “mutagenic”; if the target compound matches one or more non-mutagenicity rules (or no rules), the prediction will be “non-mutagenic.”

3 Methods

A critical assessment of predictions is the most demanding aspect related to the interpretation of the output of (Q)SAR models. VEGA facilitates the interpretability of (Q)SAR predictions by breaking down several critical aspects of the applicability domain as described in Subheading 2.5. Nevertheless, possible misinterpretations can still take place and the following examples will provide further insights into the analysis of (Q)SAR results.

The first two examples illustrate predictions characterized by a clear output which is concordant across all VEGA models. On the contrary, the last example is more challenging and it will advise the reader about complex cases. The purpose of this section is to provide an insight into the critical assessment of QSAR predictions and to highlight relevant aspects that should be taken into account when analyzing (Q)SAR outputs.



Systematic Name: (2-Propylvaleric acid)
CAS Registry Number: 99-66-1
VEGA SMILES: O=C(O)C(CCC)CCC
Experimental activity: Non-mutagenic in Ames test

Fig. 2 Valproic acid structure, chemical information, and experimental activity [26]

3.1 Case Study: Valproic Acid (Fig. 2)

- **CAESAR results**: *Prediction is non-mutagenic and the result appears reliable.*

The CAESAR model does not identify any SA linked to mutagenic activity.

Similarity values for the six most similar compounds are very high (ranging from 0.989 to 0.903). Furthermore, experimental and predicted toxicities agree for all the similar molecules that were found in the training set. Indeed, predicted and experimental toxicities systematically designate non-mutagenic chemicals (*see* **Note** 7).

On the basis of this information and in particular thanks to a visual inspection of the first three similar compounds, the predicted substance is considered into the applicability domain of the model (ADI = 0.978) (*see* Fig. 3).

- **SARpy results**: *Prediction is non-mutagenic and the result appears reliable.*

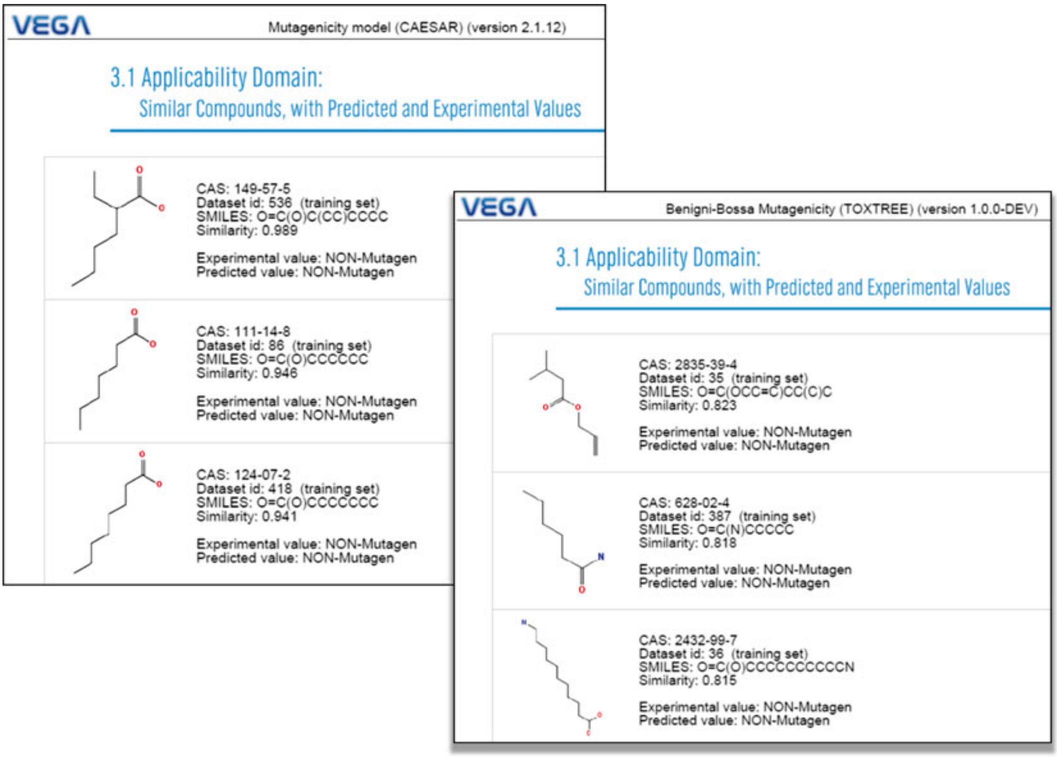
The model finds within the structure of the query chemical only SAs for non-mutagenicity ("Inactive" rules) (*see* Fig. 4).

Also in this case, the query chemical falls into the applicability domain (ADI = 0.978) and the predicted and experimental toxicities for the most similar compounds are the same. This behavior is not completely surprising since CAESAR and SARpy are based on the same training set. Nevertheless, this result corroborates the prediction computed by CAESAR by assessing toxicities according to a complementary analysis executed by a different algorithm.

- **TT-VEGA results**: *Prediction is non-mutagenic and the result appears reliable.*

Similarly to what described for the CAESAR model, Toxtree does not find any SA for mutagenicity.

The most similar compounds shown in the output are different from those of CAESAR and SARpy since the corresponding training sets are different. These structural analogs are characterized by lower similarity values (ranging from 0.823 to 0.773) and this lower degree of similarity is reflected by the ADI (0.906). This degree of overall similarity combined with a lack of identification of SA substantiates the validity of the prediction (*see* **Note 8**).



VEGA Mutagenicity model (CAESAR) (version 2.1.12)

3.1 Applicability Domain:
Similar Compounds, with Predicted and Experimental Values

CAS: 149-57-5
Dataset id: 536 (training set)
SMILES: O=C(O)C(CC)CCCC
Similarity: 0.989

Experimental value: NON-Mutagen
Predicted value: NON-Mutagen

CAS: 111-14-8
Dataset id: 86 (training set)
SMILES: O=C(O)CCCCCC
Similarity: 0.946

Experimental value: NON-Mutagen
Predicted value: NON-Mutagen

CAS: 124-07-2
Dataset id: 418 (training set)
SMILES: O=C(O)CCCCCCC
Similarity: 0.941

Experimental value: NON-Mutagen
Predicted value: NON-Mutagen

VEGA Benigni-Bossa Mutagenicity (TOXTREE) (version 1.0.0-DEV)

3.1 Applicability Domain:
Similar Compounds, with Predicted and Experimental Values

CAS: 2835-39-4
Dataset id: 35 (training set)
SMILES: O=C(OCC=C)CC(C)C
Similarity: 0.823

Experimental value: NON-Mutagen
Predicted value: NON-Mutagen

CAS: 628-02-4
Dataset id: 387 (training set)
SMILES: O=C(N)CCCCC
Similarity: 0.818

Experimental value: NON-Mutagen
Predicted value: NON-Mutagen

CAS: 2432-99-7
Dataset id: 36 (training set)
SMILES: O=C(O)CCCCCCCCCCN
Similarity: 0.815

Experimental value: NON-Mutagen
Predicted value: NON-Mutagen

Fig. 3 A particular of the three on six most similar compounds that are shown in the pdf outputs of the models for Valproic acid. SARpy and CAESAR display the same molecules

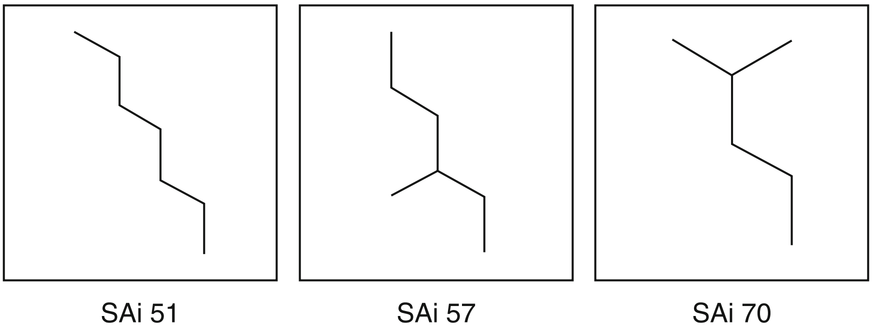


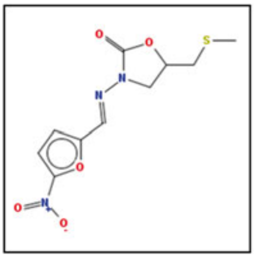
Fig. 4 Inactive SAs identified by SARpy in Valproic acid molecule

- **Overall evaluation**: *for this case there is agreement between the three models, and each model is corroborated by the high ADI value.*

3.2 Case Study: Nifuratel (See Fig. 5)

- **CAESAR results:** *Prediction is mutagenic and the result appears reliable.*

The model identifies one fragment related to mutagenic activity included within the Benigni-Bossa rulebase [21]: Nitro aromatic, SA27 (*see* Fig. 6).



Systematic Name: (2-Oxazolidinone, 5-((methylthio)methyl)-3-(((5-nitro-2-furanyl) methylene)amino)-
CAS Registry Number : 4936-47-4
VEGA SMILES: O=C2OC(CN2(N=Cc1oc(cc1)[N+](=O)[O-]))CSC
Experimental activity: Mutagenic in Ames test

Fig. 5 Nifuratel structure, chemical information, and experimental activity [27]

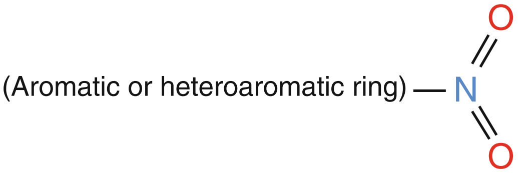


Fig. 6 Nitro aromatic structural alert no. 27

In addition to the six most similar molecules found in the training set, the model shows the three most similar compounds having the same fragment (*see* Fig. 7).

The similarity index is high, 0.9. The concordance for similar molecules and the accuracy index are both equal to 1.

For these reasons the predicted substance is considered into the applicability domain (ADI = 0.948).

- **SARpy results:** *Prediction is mutagenic and the result appears reliable.*

In this case the identified fragments are four and all linked to mutagenic activity (*see* Fig. 8).

SARpy also shows the most similar compounds that are characterized by the presence of the identified fragments. In this case predictions and experimental values agree for all the structural analogs.

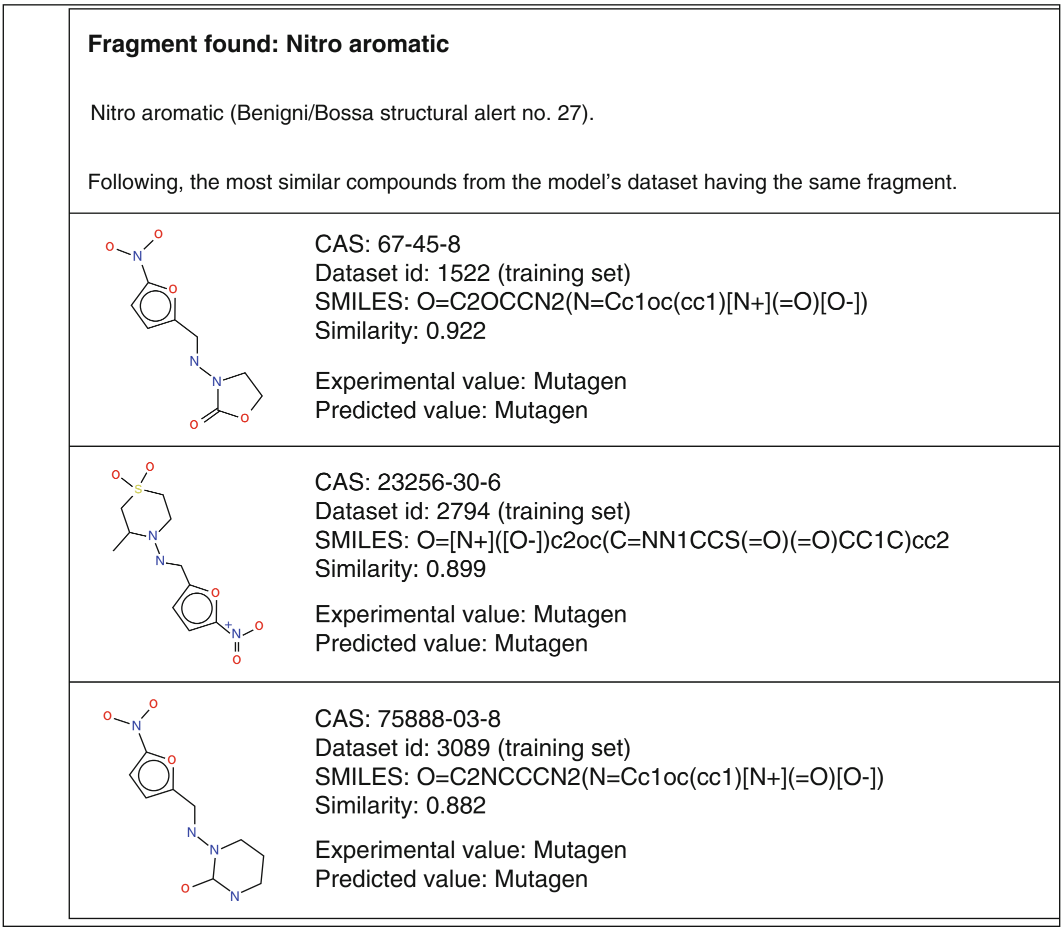
This prediction is characterized by the same ADI (and his subindexes) as the prediction computed by the CAESAR model.

- **Toxtree results**: *Prediction is mutagenic and the result appears reliable.*

As explained in Subheading 2.6, CAESAR contains a subset of Toxtree rules and both models identify the same nitro aromatic fragment that plays a key role in supporting the prediction.

The ADI value (0.933) is slightly lower than what observed for CAESAR and SARpy; this is related only to the index of similarity (0.871) while the other indices are all excellent.

- **Overall evaluation**: *all models agree, and there are good examples of similar compounds suggesting the predictions.*



Fragment found: Nitro aromatic

Nitro aromatic (Benigni/Bossa structural alert no. 27).

Following, the most similar compounds from the model's dataset having the same fragment.

CAS: 67-45-8
Dataset id: 1522 (training set)
SMILES: O=C2OCCN2(N=Cc1oc(cc1)[N+](=O)[O-])
Similarity: 0.922

Experimental value: Mutagen
Predicted value: Mutagen

CAS: 23256-30-6
Dataset id: 2794 (training set)
SMILES: O=[N+]([O-])c2oc(C=NN1CCS(=O)(=O)CC1C)cc2
Similarity: 0.899

Experimental value: Mutagen
Predicted value: Mutagen

CAS: 75888-03-8
Dataset id: 3089 (training set)
SMILES: O=C2NCCCN2(N=Cc1oc(cc1)[N+](=O)[O-])
Similarity: 0.882

Experimental value: Mutagen
Predicted value: Mutagen

Fig. 7 Part of the CAESAR output in Nifuratel prediction: three of the most similar compounds within training set that have the same SA27 fragment found in the target

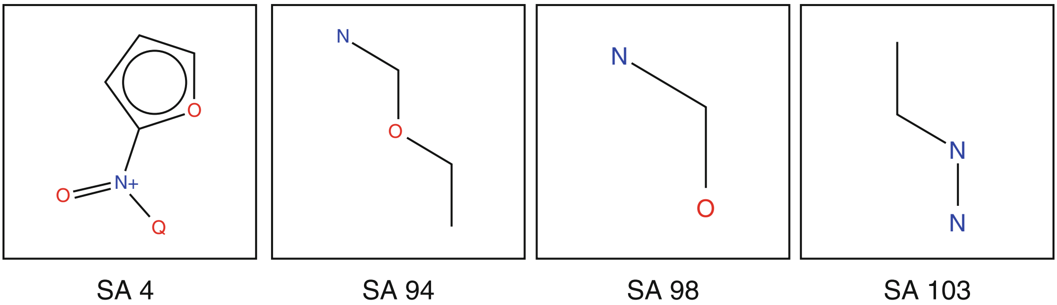
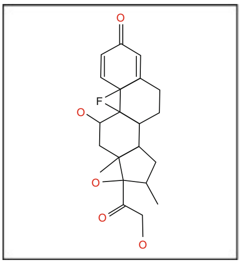


Fig. 8 Four active fragments identified by SARpy

3.3 Case Study: Dexamethasone (See Fig. 9).

Unlike the previous examples, in this case the output is equivocal because the prediction models are in disagreement and show very low values of ADI.

- **CAESAR results**: *Prediction is non-mutagenic but the result may not be reliable.*



Systematic Name: Pregna-1,4-diene-3,20-dione, 9-fluoro-11,17,21-trihydroxy-16-methyl-, (11beta,16alpha)-
CAS Registry Number: 50-02-2
VEGA SMILES:
O=C1C=CC3(C(=C1)CCC2C4CC(C)C(O)(C(=O)CO)C4(C)(CC(O)C23(F)))(C)
Experimental activity : Non mutagenic in Ames test

Fig. 9 Dexamethasone structure, chemical information, and experimental activity [28]

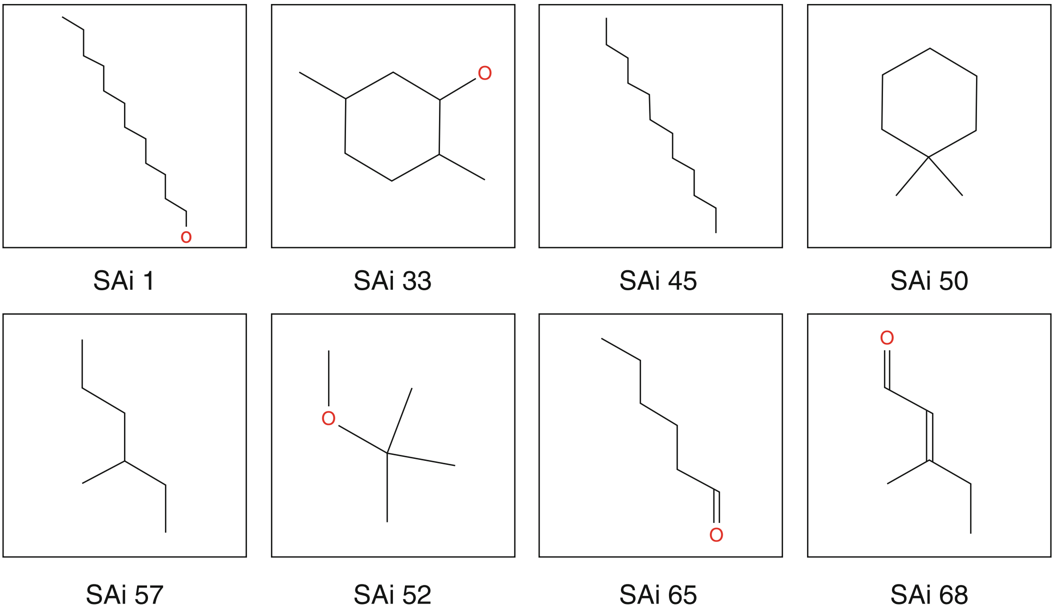


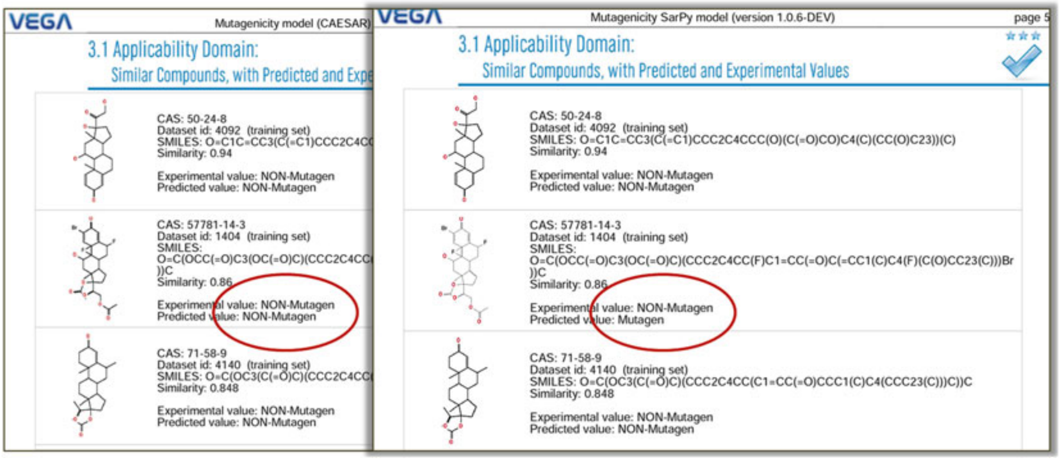
Fig. 10 Examples of inactive fragments identified by SARpy

Although similarity, concordance, and accuracy indices are high (respectively 0.875, 1 and 1), ADI is equal to 0.795, therefore Dexamethasone could be out of the applicability domain of the model. This lack of reliability is caused by a low (0.85) value of the ACF index.

- **SARpy results**: *Prediction is non-mutagenic but the result may not be reliable.*

The model identifies nine inactive fragments. Some of these fragments are depicted in Fig. 10.

Even if the values of similarity and ACF indexes are the same than what observed when using CAESAR, the ADI (0.721) value is lower because the accuracy does not reach the minimal recommended threshold (0.676) (*see* Fig. 11).



VEGA Mutagenicity model (CAESAR)

3.1 Applicability Domain:
Similar Compounds, with Predicted and Expe

CAS: 50-24-8
Dataset id: 4092 (training set)
SMILES: O=C1C=CC3(C(=C1)CCC2C4CC
Similarity: 0.94

Experimental value: NON-Mutagen
Predicted value: NON-Mutagen

CAS: 57781-14-3
Dataset id: 1404 (training set)
SMILES:
O=C(OCC(=O)C3(OC(=O)C)(CCC2C4CC
))C
Similarity: 0.86

Experimental value: NON-Mutagen
Predicted value: NON-Mutagen

CAS: 71-58-9
Dataset id: 4140 (training set)
SMILES: O=C(OC3(C(=O)C)(CCC2C4CC
Similarity: 0.848

Experimental value: NON-Mutagen
Predicted value: NON-Mutagen

VEGA Mutagenicity SarPy model (version 1.0.6-DEV) page 5

3.1 Applicability Domain:
Similar Compounds, with Predicted and Experimental Values

CAS: 50-24-8
Dataset id: 4092 (training set)
SMILES: O=C1C=CC3(C(=C1)CCC2C4CCC(O)(C(=O)CO)C4(C)(CC(O)C23))(C)
Similarity: 0.94

Experimental value: NON-Mutagen
Predicted value: NON-Mutagen

CAS: 57781-14-3
Dataset id: 1404 (training set)
SMILES:
O=C(OCC(=O)C3(OC(=O)C)(CCC2C4CC(F)C1=CC(=O)C(=CC1(C)C4(F)(C(O)CC23(C)))Br
))C
Similarity: 0.86

Experimental value: NON-Mutagen
Predicted value: Mutagen

CAS: 71-58-9
Dataset id: 4140 (training set)
SMILES: O=C(OC3(C(=O)C)(CCC2C4CC(C1=CC(=O)CCC1(C)C4(CCC23(C)))C))C
Similarity: 0.848

Experimental value: NON-Mutagen
Predicted value: NON-Mutagen

Fig. 11 The *red circle* indicates the different predictions of CAESAR and SARpy for the second most similar compound. Since the prediction computed by SARpy does not match the experimental activity, its accuracy is lower than what observed when using CAESAR

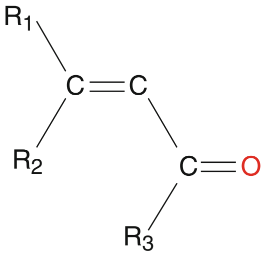


Fig. 12 SA10 (α, β unsaturated carbonyl)

- **TT-VEGA results**: *Prediction is mutagenic but the result may not be reliable.*

The model identifies the presence of the SA10 as a cause of mutagenicity of the target compound (*see* Fig. 12).

Conversely to what observed for Nifuratel, the predictions yielded by CAESAR and TT-VEGA are in disagreement since CAESAR does not contain the SA10 fragment in its subset of rules (see above).

The unreliability of the TT-VEGA prediction is highlighted by the poor value of its ADI (0) that is determined by low values of the concordance, accuracy, and ACF indices (0, 0, and 0.6 respectively).

Indeed, even if the prediction yielded by TT-VEGA is characterized by a similarity index which is greater (0.922) than the corresponding index of CAESAR and SARpy, the experimental and the predicted values are in disagreement for all the similar compounds in the output.

Difficult cases such as this example could benefit from tools such as ToxRead (*see* Chapter 13) that can perform read-across analysis while providing p-values calculated by using the Fisher's test and accuracies for each structural alert. In this case ToxRead could provide an insight into the analysis of the SA10 fragment by showing its low accuracy (0.49) and *p*-value (0.015).

On the contrary, the nine fragments identified by SARpy have accuracies ranging from 0.7 to 0.9 and *p*-values $<10^{-6}$.

The examples detailed in the previous paragraphs highlight the fact that a thorough analysis of all the factors that influence the predictive accuracy of a model should be taken into account instead of simply relying on the final prediction. Several potential pitfalls can be prevented by analyzing all the sub-indices that compose the ADI and by a visual inspection of the input molecule versus all the identified structural analogs. Particularly, the pertinence of such a visual inspection can be corroborated by the recognition of SA within the query chemical and/or its structural analogs.

It is also important to point out that QSAR and read-across predictions are not mutually exclusive and that such a synergy can potentially provide relevant information in difficult cases that are characterized by fuzzy QSAR predictions (e.g. the case of Dexamethasone). Indeed, an expert can always compare the results computed by a model with its own read-across prediction on the basis of the identified analogs. These concepts will be discussed in Chapter 13.

4 Notes

1. The predictive models discussed in this chapter do not predict for a specific *S. typhimurium* strain. On the other hand, ADMET predictor (Absorption, Distribution, Metabolism, Elimination, and Toxicity of chemical substances), a commercial tool, includes ten different models for different strains of *S. typhimurium* with and without microsomal activation [29]. We notice that the performance of the "general" mutagenicity models was superior compared to the strain-specific models, when tested in a large set of compounds [20].

2. There are several commercial or freely available software programs that can predict mutagenic hazards. In addition to the VEGA platform, other examples of free models are T.E.S.T. (Toxicity Estimation Software Tool) [30] and Toxtree (Estimation of Toxic Hazard—A Decision Tree Approach) by Ideaconsult Ltd. [31].
3. VEGA calculates the applicability domain through a program which is different from the (Q)SAR model predicting the value of interest.
4. The ADI measurement within VEGA is composed of a series of sub-indices which vary depending on the (Q)SAR model.
5. For the models embedded within the VEGA platform, the expression "training set" refers to the set of molecules used during the calibration of the models and their internal validation. The membership of the most similar structural analogs of the query chemical (training or test set) is specified in the output provided by the software. The output format is different for TEST. In this case the output shows the most similar structural analogs of the query chemical that are found in the test set and, if prompted by the user, it also shows the most similar compounds identified in the training set.
6. SARpy adopts SAs but these fragments are not based on "a priori" knowledge of the biochemical mechanism of action like for the rules-based systems (such as Toxtree and DEREK); it is more correct to refer to SARpy as a statistical model, which is highly transparent and communicates the extracted knowledge by means of rules. Another major difference between SARpy and the rule-based models is that SARpy shows rules associated with lack of toxicity. These fragments are most frequently present in the non-mutagenic compounds of the training set. However, considering the SA for mutagenicity there are strong similarities with rule-based models.
7. The evaluation on the similar compounds carried out by using VEGA can be regarded as a kind of read-across approach. The user may also apply VEGA for read across, without considering the prediction done by the model.
8. Please notice that each model in VEGA has its own data set. Also the ADI is based on this data set. It may be that the same chemical is characterized by conflicting properties value (mutagenic or non-mutagenic) depending on the data set.

Acknowledgements

We acknowledge the LIFE+CALEIDOS project.

References

1. Moore MM, Myers MB, Heflich RH (2000) Mutagenesis and genetic toxicology. In: Williams PL, James RC, Roberts SM (eds) Principles of toxicology: environmental and industrial applications, 2nd edn. Wiley-Interscience, New York, pp 239–264
2. Sutter A, Amberg A, Boyer S et al (2013) Use of *in silico* systems and expert knowledge for structure-based assessment of potential mutagenic impurities. Regul Toxicol Pharmacol 67:39–52
3. Regulation (EC) No. 1907/2006 of the European Parliament and of the Council, of December 18, 2006 concerning the Registration, Evaluation, Authorization and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/45/EC and repealing Council Regulation (EEC) No 793/93 and Commission
4. Ames BN, McCann J, Yamasaki E (1975) Methods for detecting carcinogens and mutagens with the salmonella/mammalian-microsome mutagenicity test. Mutat Res 31: 347–364
5. Claxton LD, Umbuzeiro GA, DeMarini DM (2010) The *Salmonella* mutagenicity assay: the stethoscope of genetic toxicology for the 21st century. Environ Health Perspect 118: 1515–1522
6. Benigni R, Bossa C, Battistelli CL et al (2013) IARC classes 1 and 2 carcinogens are successfully identified by an alternative strategy that detects DNA-reactivity and cell transformation ability of chemicals. Mutat Res 758:56–61
7. OECD (1997) Test No. 471: bacterial reverse mutation test, OECD guidelines for the testing of chemicals, section 4. OECD Publishing, Paris
8. Piegorsch W, Zeiger E (1991) Measuring intra-assay agreement for the Ames Salmonella assay. In: Hotorn L (ed) Statistical methods in toxicology, lecture notes in medical informatics, vol 43. Springer-Verlag, Berlin, pp 35–41
9. Sushko I, Novotarskyi S, Körner R et al (2010) Applicability domains for classification problems: benchmarking of distance to models for Ames mutagenicity set. J Chem Inf Model 50:2094–2111
10. Miller JA, Miller EC (1977) Ultimate carcinogens as reactive mutagenic electrophiles. In: Hiatt HH, Watson JD, Winston JA (eds) Origins of human cancer, mechanisms of carcinogenesis, Book B. Cold Spring Harbor Laboratory, New York, pp 605–627
11. Cooper JA, Saracci R, Cole P (1979) Describing the validity of carcinogen screening tests. Br J Cancer 39:87–89
12. Virtual models for evaluating the properties of chemicals within a global architecture. http://www.vega-qsar.eu/download.html. Accessed 19 Jun 2015
13. SMILES, Simplified Molecular Input Line Entry System. http://www.daylight.com/smiles/index.html. Accessed 19 Jun 2015
14. Drug Design Laboratory. http://nova.disfarm.unimi.it/cms/index.php?Software_projects. Accessed 19 Jun 2015
15. ACD/ChemSketch for Academic and Personal Use. http://www.acdlabs.com/resources/freeware/chemsketch/. Accessed 19 Jun 2015
16. Marvin, intuitive applications and API for chemical sketching, visualization and data exploration. http://www.chemaxon.com/products/marvin/. Accessed 19 Jun 2015
17. The OECD QSAR Toolbox. http://www.oecd.org/chemicalsafety/risk-assessment/theoecdqsartoolbox.htm. Accessed 19 Jun 2015
18. Mombelli E, Devillers J (2010) Evaluation of the OECD (Q)SAR Application Toolbox and Toxtree for predicting and profiling the carcinogenic potential of chemicals. SAR QSAR Environ Res 21:731–752
19. Benfenati E, Pardoe S, Martin T (2013) Using toxicological evidence from QSAR models in practice. ALTEX 30:19–40
20. Bakhtyari NG, Raitano G, Benfenati E et al (2013) Comparison of *in silico* models for prediction of mutagenicity. J Environ Sci Health C Environ Carcinog Ecotoxicol Rev 31:45–66
21. Benigni R, Bossa C, Jeliazkova N et al (2008) Benigni/Bossa rulebase for mutagenicity and carcinogenicity—a module of toxtree. JRC scientific and technical reports. https://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology/doc/EUR_23241_EN.pdf. Accessed 19 Jun 2015
22. Istituto Superiore di Sanità. "Chemical carcinogens structures and experimental data" (ISSCAN). http://www.epa.gov/ncct/dsstox/sdf_isscan_external.html. Accessed 19 Jun 2015
23. Ferrari T, Gini G (2010) An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. Chem Cent J 4:S2
24. Kazius J, McGuire R, Bursi R (2005) Derivation and validation of toxicophores for mutagenicity prediction. J Med Chem 48:312–330

25. Ferrari T, Cattaneo D, Gini G et al (2013) Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction. SAR QSAR Environ Res 24: 365–383
26. Hansen K, Mika S, Schroeter T et al (2009) Benchmark data set for in silico predictions of Ames mutagenicity. J Chem Inf Model 49:2077–2081
27. Byeon WH, Hyun HH, Lee SY (1976) Mutagenicity of nitro furan nitroimidazol and nitrothiazole derivatives on salmonella microsome system. Kor J Microbiol 14:151–158
28. Singh H, Singh JR, Dhillon VS et al (1994) In vitro and in vivo genotoxicity evaluation of hormonal drugs II. Dexamethasone. Mutat Res 308:89–97
29. Simulations Plus, Inc. http://www.simulations-plus.com. Accessed 19 Jun 2015
30. Quantitative structure activity relationship, Toxicity Estimation Software Tool (TEST). http://www.epa.gov/nrmrl/std/qsar/qsar.html. Accessed 19 Jun 2015
31. Toxtree—toxic hazard estimation by decision tree approach. http://toxtree.sourceforge.net. Accessed 19 Jun 2015