Guide to Mutagenicity Consensus version 1.0.3

Table of Contents

1. Model explanation....2
    1.1 Introduction....2
    1.2 Model details....2
2. Model usage....3
    2.1 Input....3
    2.2 Output....3
3. Differences from previous versions....5
    3.1 VEGA model history....5
        3.1.1 Version 1.0.1....5
        3.1.2 Version 1.0.2....5
        3.1.3 Version 1.0.3....5

1. Model explanation

1.1 Introduction

The model provides a qualitative prediction of mutagenicity on Salmonella typhimurium (Ames test), applying a consensus approach based on the four QSAR models currently available in VEGA. It is implemented inside the VEGA online platform, accessible at: http://www.vega-qsar.eu/

1.2 Model details

The model performs a consensus assessment based on the predictions of the available VEGA mutagenicity models (CAESAR, SarPy, ISS and KNN). The consensus algorithm uses the Applicability Domain assessment of each single model's prediction as its weight, so that the final assessment will be more influenced by the single models that produced more reliable predictions.

The Applicability Domain assessment of each model is converted to a numerical value in the range [0..1] with the following scheme:

AD Assessment	*Value / Weight*
Experimental value	1.0
High reliability	0.9
Moderate reliability	0.6
Low reliability	0.2

For each prediction class (i.e. mutagenic or non-mutagenic), a score is calculated as the sum of the weights for each model that produced that prediction. For the purpose of this consensus approach, the “suspect mutagenic / non-mutagenic” predictions are considered simply as “mutagenic / non-mutagenic” predictions. The calculated score is normalized over the number of used models, so that it has a theoretical [0..1] range. Then, the prediction class with the highest score is used as the final consensus prediction. If at least one model has found an experimental value, only experimental values are considered for the score. In this case, the reported number of models used refers only to the number of models having an experimental value.

With this approach, the score of the final prediction can be used as a measure of the reliability of the produced consensus assessment. Indeed, the score would achieve its maximum value (1) only if one or more models found experimental values and these values are in agreement. In all other cases, the score will result in lower values.

2. Model usage

2.1 Input

The model accepts as input two molecule formats: SDF (multiple MOL file) and SMILES. All molecules found as input are preprocessed before the calculation of molecular descriptors, in order to obtain a standardized representation of compound. For this reason, some cautions should be taken.

- **Hydrogen atoms**. In SDF files, hydrogen atoms should be explicit. As some times SDF file store only skeleton atoms, and hydrogen atoms are implicit, during the processing of the molecule the system tries to add implicit hydrogens on the basis of the known standard valence of each atom (for example, if a carbon atoms has three single bonds, an hydrogen atom will be added such to reach a valence of four). In SMILES molecules, the default notation uses implicit hydrogen. Anyway please note that in some cases it is necessary to explicitly report an hydrogen; this happens when the conformation is not unambiguous. For example, when a nitrogen atom is into an aromatic ring with a notation like "cnc" it is not clear whether it corresponds to C-N=C or to C-[NH]-C, thus if the situation is the latter, it should be explicitly reported as "c[nH]c".

- **Aromaticity**. The system calculates aromaticity using the basic Hueckel rule. Note that each software for drawing and storing of molecules can use different approaches to aromaticity (for instance, commonly the user can choose between the basic Hueckel rule and a loose approach that lead to considering aromatic a greater number of rings). As in the input files aromaticity can be set explicitly (for instance, in SMILES format by using lowercase letters), during the processing of the molecule the system removes aromaticity from rings that don't satisfy the Hueckel rule. Please note that when aromaticity is removed from a ring, it is not always possible to rebuild the original structure in Kekule form (i.e. with an alternation of single and double bonds, like in the SMILES for benzene, C=1C=CC=CC1), in this case all bonds are set to single. Furthermore, please note that aromaticity detection is a really relevant issue, some molecular descriptors can have significantly different values whether a ring is perceived as aromatic or not. For this reason it is strongly recommended:
- Always use explicit hydrogens in SDF file.
- Avoid explicit aromaticity notation in original files; in this way, the perception of aromaticity is left to the preprocessing step and there is no chance of mistakes due to the transformation of rings that were set to aromatic in the original format but not recognized as aromatic in VEGA.

Note that when some modification of the molecule are performed during the preprocessing (e.g. adding of lacking hydrogens, correction of aromaticity), a warning is given in the remark field of the results.

2.2 Output

Results given as text file consist of a plain-text tabbed file (easily importable and processable by any spreadsheet software) containing in each row all the information about the prediction of a molecule, including the assessment from the single used models. Note that if some problems were encountered

while processing the molecule structure, some warnings are reported in the last field (Remarks).

Results given as PDF file consists of a document containing all the information about the prediction. For each molecule, results are organized in sections with the following order:

1 – Prediction summary

Here is reported a depiction of the compound and the final assessment of the prediction (i.e. the prediction made together with the analysis of the applicability domain). Note that if some problems were encountered while processing the molecule structure, some warnings are reported in the last field (Remarks). The number of models used for the consensus prediction is reported, together with the score of the final prediction and the assessment from the single used models.
A graphical representation of the evaluation of the prediction is also provided, using the following elements:

 Compound is classified as non-mutagen

 Compound is classified as mutagen

3. Differences from previous versions

3.1 VEGA model history

3.1.1 Version 1.0.1

Changed weights for Applicability Domain conversion.

3.1.2 Version 1.0.2

Changed the main algortithm when experimental values are found.

3.1.3 Version 1.0.3

Fix in rounding of score values