	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Sarah Nexus - Mutagenicity**
	Printing Date: **18 May 2020**

1.QSAR identifier

1.1.QSAR identifier (title):

Sarah Nexus - Mutagenicity

1.2.Other related models:

None

1.3.Software coding the model:

Sarah Nexus makes predictions for mutagenicity using fragment-based structural hypotheses derived from a statistically learned self-organising hypothesis network (SOHN) built using bacterial reverse mutation test data.

2.General information

2.1.Date of QMRF:

12 May 2016

2.2.QMRF author(s) and contact details:

Alex Cayley Lhasa Limited Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS, UK

2.3.Date of QMRF update(s):

18 May 2020

2.4.QMRF update(s):

Alex Cayley, Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds,
LS11 5PS, UK
1.3, 2.3, 2.4, 2.6, 2.8, 4.6, 4.7, 5.3, 6.2

2.5.Model developer(s) and contact details:

Lhasa Limited Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS

2.6.Date of model development and/or publication:

Sarah Nexus 3.1 was released on 11 June 2020

2.7.Reference(s) to main scientific papers and/or software package:

[1]Hanser T, Barber C, Rosser E, Vessey JD, Webb SJ & Werner S (2014). Self organising hypothesis networks: a new approach for representing and structuring SAR knowledge. Journal of Cheminformatics 6:21

[2]Barber C, Cayley A, Hanser T, Harding A, Heghes C, Vessey JD, Werner S, Weiner SK, Wichard J, Giddings A, Glowienke S, Parenty A, Brigo A, Spirkl HP, Amberg A, Kemper R & Greene N (2016). Evaluation of a statistics-based Ames mutagenicity QSAR model and interpretation of the results obtained. Regulatory Toxicology and Pharmacology 76, 7-20.

[3]Barber C, Amberg A, Custer L, Dobo KL, Glowienke S, Van Gompel J, Gutsell S, Harvey J, Honma M, Kenyon MO, Kruhlak N, Muster W, Stavitskaya L, Teasdale A, Vessey J, Wichard J (2015). Establishing best practise in the application of expert review of mutagenicity under ICH M7. Regulatory Toxicology and Pharmacology 73, 367-377.

2.8.Availability of information about the model:

Sarah Nexus is a proprietary, statistical system for the prediction of mutagenicity. It employs a self-organising hypothesis network (SOHN) of structural fragments to make predictions for mutagenicity [Hanser et al,

2014] . The fragments in the SOHN (referred to as hypotheses) are associated with activity or inactivity depending on the distribution of compounds containing this fragment in the training set of compounds with associated bacterial reverse mutation data. The SOHN is derived automatically from the training data using a set of rules relying on the statistical distribution of positive and negative results for each structural fragment in the training set. An overall prediction for a query compound is derived based on resolving the results from the different hypotheses it activates. A quantitative confidence value is also associated with each hypothesis based on the activity of the nearest neighbours in the training set to the query compound. An overall confidence in the prediction is then derived by combining these confidences for individual hypotheses. Predictions are supported by displaying the relevant hypotheses associated with the query compound as well as compounds in the training set used to derive these hypotheses in order of similarity to the query compound. Detailed strain information on each training set compound along with CAS identification numbers and references to the primary literature are also provided where available.
A strain profile for each hypothesis is generated based on the strain information from the individual compounds belonging to it. Sarah Nexus has a domain of applicability. A compound is deemed to be within the applicability domain of the model if all of the fragments present in the query structure have been adequately represented in the training set of the model. If they have not then the relevant fragment will be highlighted in the query structure and it will be assigned as out of domain. As well as positive and negative Sarah Nexus can also give equivocal results. These are produced when a prediction of only a low confidence can be made by the model. By reporting this information to the user, Sarah provides highly transparent predictions.

2.9.Availability of another QMRF for exactly the same model:

No

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Predictions are made for the domain of bacteria.

3.2.Endpoint:

TOX 7.6.1. Genetic toxicity in vitro

3.3.Comment on endpoint:

The Sarah Nexus model for mutagenicity is developed from bacterial reverse mutation data.

3.4.Endpoint units:

Sarah Nexus makes an overall qualitative prediction for mutagenicity. This overall prediction is based on the combination of evidence from the hypotheses and nearest neighbours to the query compound in the training set. and is provided with a quantitative percentage measure of confidence in the prediction (again based on the hypotheses and nearest

neighbours in the training set). Confidence levels have been shown to correlate with predictivity [Barber et al, 2016 [2]]. Multiple data sources (e.g. toxicity data from multiple reverse mutation assays) are synthesised into the training set that underpins Sarah Nexus predictions. While the confidence in a prediction is quantitative overall predictions are not and, as a result, do not include units.

3.5.Dependent variable:

Data from bacterial reverse mutation assays are used to define the model.

3.6.Experimental protocol:

The model is based primarily on data from the bacterial reverse mutation assays and precise experimental protocol may vary between data points.

3.7.Endpoint data quality and variability:

A large data set of bacterial reverse mutation data from the public domain is used to build the training set for Sarah Nexus. The structures in the dataset are standardised according to the Lhasa Limited protocol (outlined at

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Statistical model for mutagenicity (2D SARs).

4.2.Explicit algorithm:

Sarah Nexus uses a self organising hypothesis network (SOHN) to generate structure fragment-based hypotheses which are used to make predictions. More detail on this approach can be found in Hanser et al 2014 [1].

4.3.Descriptors in the model:

2D structural fragments

4.4.Descriptor selection:

There is an a priori assumption that the presence of certain structural features in a compound can be directly reactive or produce reactive species capable of reacting with DNA and causing mutations. These structural features can be encoded with 2D structural fragment descriptors which are then used to model toxicity within Sarah Nexus.

4.5.Algorithm and descriptor generation:

Binary (positive/negative) data from bacterial reverse mutation assays along with the associated chemical structure is provided to the program. The chemical structures are fragmented according to a proprietary algorithm and recursive partitioning based on structure and activity information is used to determine whether individual fragments will be considered for inclusion in the self organising hypothesis network (SOHN). For those that are considered acceptable for use in the SOHN this network is produced using a set of rules for inclusion of nodes based on signal strength (a predominance for association with positive or negative compounds) and signal strength gain compared to related nodes in the tree. Once built this SOHN is used to make predictions for query compounds which are fragmented in the same way and then processed through this network. A prediction is made by combining local models

based on the fragments present in the query structure and represented in the SOHN network; these local models are derived from the nearest neighbours in the training set that contain these fragments (local kNN model). Neighbours are selected using a Tanimoto similarity derived from a fingerprint based on the fragments generated from the training set. The process is described in more detail in Hanser et al 2014.

4.6.Software name and version for descriptor generation:

Sarah Nexus 3.1.

4.7.Chemicals/Descriptors ratio:

11774 chemicals (comprising 5780 mutagens and 5994 non-mutagens) are used to build the model producing a network of 346 hypotheses (258 unique hypotheses).

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The applicability domain of Sarah Nexus is defined by comparing the structural fragments present in the training set with those present in the query compound. If all of the atoms in the query compound are covered by structural fragments found in the Sarah Nexus training set then the query compound is considered inside the applicability domain of the model. If one or more of the atoms in the query structure is not represented by a fragment in the training set then the query structure is considered out of the applicability domain of the model. In this case the fragments from the query structure associated with the out of domain atoms are highlighted on the query structure and the results of the prediction are also shown. This allows the user to carry out an expert interpretation of the result in order to assess the out of domain fragment using knowledge not contained in the model which may in turn allow them to resolve this out of domain prediction into a positive or negative one.

5.2.Method used to assess the applicability domain:

Structural fragments are used to assess the applicability domain. Chemical structures from the training set are fragmented according to a pre-defined set of rules and the fragments are stored. The query structure is fragmented in the same way and the fragments from the query structure are checked against the training set fragments. If one or more atoms can’t be found in a fragment of the structure also present in the fragments of the training set they are considered outside the applicability domain of the model. This indicates a previously unseen structural environment and the whole structure is labeled outside the applicability domain.

5.3.Software name and version for applicability domain assessment:

Sarah Nexus 3.1.

5.4.Limits of applicability:

Applicability is limited by the structural fragments present in the Sarah Nexus training set.

6.Internal validation - OECD Principle 4

6.1.Availability of the training set:

Yes

6.2.Available information for the training set:

CAS RN (where available), Ames test strain detail (where available), Reference (where available)

CAS RN: No

Chemical Name: No

Smiles: Yes

Formula: No

INChI: Yes

MOL file: Yes

NanoMaterial: No

6.3.Data for each descriptor variable for the training set:

No

6.4.Data for the dependent variable for the training set:

No

6.5.Other information about the training set:

Access to the training set is available through exposure of relevant chemical structures in the predictions within Sarah Nexus. The training set of Sarah Nexus can also be used to build supplemented models within the software. The entire training set is not available, due to the proprietary nature of Sarah Nexus.

6.6.Pre-processing of data before modelling:

Structural standardisation is carried out before building the Sarah Nexus model and some assessment of the biological data is also made. Tautomers and resonance forms are standardised, appropriate salts and mixture components are removed, metals are represented consistently and stereochemistry is removed from the structures. Any structures with conflicting results from different data sources are also assessed. If the compounds have a result from a trusted data source (e.g. Vitic Nexus) this result overrules the others. In cases where there is no result from a trusted data source and the overall results conflict then the compounds are removed from the training set.

6.7.Statistics for goodness-of-fit:

Average Cooper statistics from a leave many out 5-fold cross validation of the Sarah Nexus are provided along with the model in the software. A ROC curve and accuracy statistics against confidence are also plotted and provided with these results.

6.8.Robustness - Statistics obtained by leave-one-out cross-validation:

Not available.

6.9.Robustness - Statistics obtained by leave-many-out cross-validation:

Cooper statistics from a leave many out 5-fold cross validation of the Sarah Nexus are provided along with the model in the software. The variance in these statistics can be used to assess the robustness of the model.

6.10.Robustness - Statistics obtained by Y-scrambling:

Not available.

6.11.Robustness - Statistics obtained by bootstrap:

Not available.

6.12.Robustness - Statistics obtained by other methods:

Not available.

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

No

7.2.Available information for the external validation set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

NanoMaterial: No

7.3.Data for each descriptor variable for the external validation set:

No

7.4.Data for the dependent variable for the external validation set:

No

7.5.Other information about the external validation set:

External validation of Sarah Nexus has primarily been carried out using proprietary data sets. Results from the validation of Sarah Nexus version 1.2 can be found in Barber et al 2016 [2]. These are currently the most up to date published validation statistics for the model. Fourteen proprietary data sets from nine different pharmaceutical companies and one data sharing group have been used in the validation of Sarah Nexus version 1.2.

7.6.Experimental design of test set:

Proprietary data sets were sought.

7.7.Predictivity - Statistics obtained by external validation:

Results from the external validation of Sarah Nexus 1.2 can be found in Barber et al 2016 [2].

7.8.Predictivity - Assessment of the external validation set:

Thirteen data sets were provided by Lhasa Limited pharmaceutical company members and one was gathered as part of an intermediates data sharing initiative. The compounds in the datasets are primarily small and medium-sized chemicals and so are representative of the structures used to build the model. More detail on the size, proportion of positives and negatives and the origin of each data set can be found in Barber et al 2016 [2].

7.9.Comments on the external validation of the model:

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The presence of certain structural features in a compound that can be directly reactive or produce reactive species capable of reacting with DNA may lead to mutations and positive results in reverse mutation assays. These structural features can be encoded with 2D structural fragment descriptors which are then used to model toxicity within Sarah Nexus.

8.2.A priori or a posteriori mechanistic interpretation:

The mechanistic interpretation of the structural features which may potentially lead to mutagenicity can be made posteriori by the expert user by assessing the structural fragments which have been associated with activity and are present in the query structure.

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

[1]Hanser T, Barber C, Rosser E, Vessey JD, Webb SJ & Werner S (2014). Self organising hypothesis networks: a new approach for representing and structuring SAR knowledge. Journal of Cheminformatics 6:21

[2]Barber C, Cayley A, Hanser T, Harding A, Heghes C, Vessey JD, Werner S, Weiner SK, Wichard J, Giddings A, Glowienke S, Parenty A, Brigo A, Spirkl HP, Amberg A, Kemper R & Greene N (2016). Evaluation of a statistics-based Ames mutagenicity QSAR model and interpretation of the results obtained. Regulatory Toxicology and Pharmacology 76, 7-20.

[3]Barber C, Amberg A, Custer L, Dobo KL, Glowienke S, Van Gompel J, Gutsell S, Harvey J, Honma M, Kenyon MO, Kruhlak N, Muster W, Stavitskaya L, Teasdale A, Vessey J, Wichard J (2015). Establishing best practise in the application of expert review of mutagenicity under ICH M7. Regulatory Toxicology and Pharmacology 73, 367-377.

9.3.Supporting information:

Training set(s)Test set(s)Supporting information

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

To be entered by JRC

10.2.Publication date:

To be entered by JRC

10.3.Keywords:

To be entered by JRC

10.4.Comments:

To be entered by JRC