SciMatics SciQSAR model for binding to the human Estrogen Receptor alpha (hERalpha) *in vitro* (Japanese METI data, balanced training set)

1. QSAR identifier

1.1 QSAR identifier (title)

SciMatics SciQSAR model for binding to the human Estrogen Receptor alpha (hERalpha) *in vitro* (Japanese METI data, balanced training set), Danish QSAR Group at DTU Food.

1.2 Other related models

Leadscope Enterprise model for binding to the human Estrogen Receptor alpha (hERalpha) *in vitro* (Japanese METI data, balanced training set), Danish QSAR Group at DTU Food.

MultiCASE CASE Ultra model for binding to the human Estrogen Receptor alpha (hERalpha) *in vitro* (Japanese METI data, balanced training set), Danish QSAR Group at DTU Food.

1.3. Software coding the model

SciQSAR version 3.1.00.

2. General information

2.1 Date of QMRF

January 2015.

2.2 QMRF author(s) and contact details

QSAR Group at DTU Food;

Danish National Food Institute at the Technical University of Denmark;

http://qsar.food.dtu.dk/;

qsar@food.dtu.dk

Eva Bay Wedebye;

National Food Institute at the Technical University of Denmark;

Nikolai Georgiev Nikolov;

National Food Institute at the Technical University of Denmark;

Marianne Dybdahl;

National Food Institute at the Technical University of Denmark;

Sine Abildgaard Rosenberg;

National Food Institute at the Technical University of Denmark;

2.3 Date of QMRF update(s)

2.4 QMRF update(s)

2.5 Model developer(s) and contact details

Eva Bay Wedebye;

National Food Institute at the Technical University of Denmark;

Jay Russel Niemelä;

National Food Institute at the Technical University of Denmark;

Nikolai Georgiev Nikolov;

National Food Institute at the Technical University of Denmark;

Danish QSAR Group at DTU Food;

National Food Institute at the Technical University of Denmark;

http://qsar.food.dtu.dk/;

qsar@food.dtu.dk

2.6 Date of model development and/or publication

January 2014.

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

Contrera, J.F., Matthews, E.J., Kruhlak, N.L., and Benz, R.D. (2004) Estimating the safe starting dose in phase I clinical trials and no observed effect level based on QSAR modelling of the human maximum recommended daily dose. *Regulatory Toxicology and Pharmacology*, 40, 185 – 206.

SciQSAR (2009) Reference guide: *Statistical Analysis and Molecular Descriptors*. Included within the SciMatics SciQSAR software.

2.8 Availability of information about the model

The training set is non-proprietary and was compiled from the Japanese Ministry of Economy, Trade and Industry (METI) report (see 9.2, METI 2002). The model algorithm is proprietary from commercial software.

2.9 Availability of another QMRF for exactly the same model

3. Defining the endpoint

3.1 Species

Human (a cell-free assay containing the human Estrogen Receptor alpha, hERalpha).

3.2 Endpoint

QMRF 4. Human Health Effects

QMRF 4.18.a. Endocrine Activity. Receptor-binding (human Estrogen Receptor alpha)

3.3 Comment on endpoint

There is increasing evidence that a variety of environmental chemicals have the potential to disrupt the endocrine system by mimicking or inhibiting endogenous hormones such as estrogens and androgens. These endocrine disrupting chemicals (EDCs) may adversely affect development and/or reproductive function.

Natural estrogens are involved in the development and adult function of organs of the female genital tract, neuroendocrine tissues and the mammary glands; their role in reproduction spans from maintenance of the menstrual cycle to pregnancy and lactation. These effects are mediated through the estrogen receptors (ERs), members of the nuclear receptor superfamily. When estrogen binds to the ER in the cytoplasm a receptor-hormone complex dimer is formed. This dimer translocates to the nucleus, where it recruits co-factors to form the active transcription factor (TF) complex. The active TF complex binds to the estrogen response element (RE) upstream to the target gene. This binding activates transcription of mRNA and subsequent translation to the proteins that exert the hormone effects. Two isoforms of the ER exists in humans, alpha and beta, and both are widely expressed in different tissue types although there are some differences in their expression pattern.
Exogenous compounds able to bind to and activate the ERs (i.e. ER agonists) have the ability mimic natural estrogens and cause adverse effects to the reproductive system. Likewise, exogenous compounds that bind to the ERs without subsequent activation (i.e. ER antagonists) can potentially disturb the effect of the natural estrogens by blocking the receptors.

In this model data from a cell-free *in vitro* assay measuring the binding affinity of a chemical to the human Estrogen Receptor alpha (hERalpha) were used. The assay does not say anything about if the binding induces transcription of the target genes (i.e. if the chemical is an ER agonist or antagonist).

3.4 Endpoint units

No units, 1 for positives and 0 for negatives.

3.5 Dependent variable

Binding to the human Estrogen Receptor alpha (hERalpha) *in vitro*, positive or negative.

3.6 Experimental protocol

The experimental protocol is described in METI (2002). Briefly, an assay to detect binding of chemicals to the estrogen receptor was established and performed by the Japanese Ministry of Economy, Trade and Industry (METI). The assay measures the binding affinity of chemicals to the human Estrogen Receptor alpha (hERalpha) by measuring the extent of the competition reaction with a reference hormone labelled with a radioisotope (RI). The hERalpha was produced from *Escherichia coli* by a genetic engineering method and was used for the receptor binding assay with RI-labelled estradiol as the reference ligand.
The chemical concentration that inhibits 50% of the binding of the reference hormone to the receptor is measured and defined as IC_{50}. The relative binding affinity (RBA) is the ratio between the IC_{50} value of the chemical tested and the IC_{50} for the natural hormone, which is set to 100.

3.7 Endpoint data quality and variability

The training set data originate from a single source (METI 2000) and therefore no or only very limited variability in the experimental protocol is expected.

4. Defining the algorithm

4.1 Type of model

This is a categorical (Q)SAR model based on calculated molecular descriptors, and if available the modeller's own or third-party descriptors or measured endpoints can be imported and used as descriptors.

4.2 Explicit algorithm

This is a categorical (Q)SAR model made by use of the non-parametric discriminant analysis (DA) k-nearest-neighbor (kNN) (K=1, Mahalanobis distance used to determine proximity) method (see 4.5). The specific implementation is proprietary within the SciQSAR software.

4.3 Descriptors in the model

Molecular connectivity indices

Molecular shape indices

Topological indices

Electrotopological (Atom E and HE-States) indices

Electrotopological bond types indices

SciQSAR software provides over 400 built-in molecular descriptors. Additionally, SciQSAR makes it possible to import the modeller's own or third-party descriptors or use measured endpoints as custom descriptors.

4.4 Descriptor selection

The initial descriptor set is manually chosen by the model developer from the total set of built-in descriptors. Furthermore, the set of descriptors applied in the modelling by the program is on top of this selection determined by thresholds for descriptor variance and number of nonzero values likewise defined by the model developer.

49 descriptors were selected from the initial pool of descriptors by the system and used to build the model.

4.5 Algorithm and descriptor generation

For a binary classification problem SciQSAR uses discriminant analysis (DA) to make a (Q)SAR model. SciQSAR implements a broad range of discriminant analysis (DA) methods including parametric and non-

parametric approaches. The classic parametric method of DA is applicable in the case of approximately normal within-class distributions. The method generates either a linear discriminant function (the within-class covariance matrices are assumed to be equal) or a quadratic discriminant function (the within-class covariance matrices are assumed to be unequal). When the distribution is assumed to not follow a particular law or is assumed to be other than the multivariate normal distribution, non-parametric DA methods can be used to derive classification criteria. The non-parametric DA methods available within SciQSAR include the kernel and *k*-nearest-neighbor (kNN) methods. The main types of kernels implemented in SciQSAR include uniform, normal, Epanechnikov, bi-weight, or tri-weight kernels, which are used to estimate the group specific density at each observation. Either Mahalanobis or Euclidean distances can be used to determine proximity between compound-vectors in multidimensional descriptor space. When the kNN method is used, the Mahalanobis distances are based on the pooled covariance matrix. When the kernel method is used, the Mahalanobis distances are based on either the individual within-group covariance matrices or the pooled covariance matrix. (Contrera *et al*. 2004)

If the data outcome is continuous, regression analysis is used to build the predictive model. Within SciQSAR several regression methods are available: ordinary multiple regression (OMR), stepwise regression (SWR), all possible subsets regression (PSR), regression on principal components (PCR) and partial least squares regression (PLS). The choice of regression method depends on the number of independent variables and whether correlation or multicollinearity among the independent variables exists: OMR is acceptable with a small number of independent variables, which are not strongly correlated. SWR is used under the same circumstances as OMR but with greater number of variables. PSR is used for problems with a great number of independent variables. PCR and PLS are useful when a high correlation or multicollinearity exist among the independent variables. (SciQSAR 2009)

To test how stable the developed models are, SciQSAR have built-in cross-validation procedures (see 6.).

For this model, the quadratic method was used.

4.6 Software name and version for descriptor generation

SciQSAR version 3.1.00.

4.7 Descriptors/chemicals ratio

In this model 49 descriptors were used. The training set consists of 595 compounds. The descriptor/chemical ratio is 1:12.1 (49:595).

5. Defining Applicability Domain

5.1 Description of the applicability domain of the model

The definition of the applicability domain consists of two components; the definition in SciQSAR and the in-house further refinement algorithm on the output from SciQSAR to reach the final applicability domain call.

1. SciQSAR

The first criterion for a prediction to be within the models applicability domain is that all of the descriptor values for the test compound can be calculated by SciQSAR. If SciQSAR cannot calculate each descriptor value for the test chemical no prediction value is given by SciQSAR and it is considered outside the model's applicability domain.

2. The Danish QSAR group

The Danish QSAR group has applied a stricter definition of applicability domain for its SciQSAR models. In addition to the applicability domain definition made by SciQSAR a second criterion has been applied for predictions generated from (Q)SAR models with a binary endpoint. For each prediction SciQSAR calculates the probability (p) for the test compound's membership in one of the two outcome classes (positive or negative). The probability of membership in a class is a measure of how well training set knowledge is able to discriminate a positive prediction from a negative prediction within the nearest space of the subject compound-vector. The probability of membership value is also a measure of the degree of confidence of a prediction. The Danish QSAR group uses this probability for a prediction to further define the model's applicability domain. Only positive predictions with a probability equal to or greater than 0.7 and negative predictions with a probability equal to or less than 0.3 are accepted. Positive predictions with a probability between 0.5 and 0.7 as well as negative predictions with a probability between 0.3 and 0.5 are considered outside the model's applicability domain. When these predictions are wed out the accuracy of the model in general increases at the expense of reduced model coverage. Furthermore, as SciQSAR does not define a structural domain, only predictions which were within either Leadscope structural domain (defined as at least one training set chemical within a Tanimoto distance of 0.7) or CASE Ultra structural domain (no unknown fragments for negatives and maximum 1 unknown fragment for positives) were defined as being inside the SciQSAR applicability domain.

5.2 Method used to assess the applicability domain

The system does not generate predictions if it cannot calculate each descriptor value for the test compound.

Only positive predictions with probability equal to or greater than 0.7 and negative predictions with probability equal to or less than 0.3 were accepted.

5.3 Software name and version for applicability domain assessment

SciQSAR version 3.1.00.

5.4 Limits of applicability

The Danish QSAR group applies an overall definition of structures acceptable for QSAR processing which is applicable for all the in-house QSAR software, i.e. not only SciQSAR. According to this definition accepted structures are organic substances with an unambiguous structure, i.e. so-called discrete organics defined as: organic compounds with a defined two dimensional (2D) structure containing at least two carbon atoms, only certain atoms (H, Li, B, C, N, O, F, Na, Mg, Si, P, S, Cl, K, Ca, Br, and I), and not mixtures with two or more 'big components' when analyzed for ionic bonds (for a number of small known organic ions assumed not to affect toxicity the 'parent molecule' is accepted). Structures with less than two carbon atoms or containing atoms not in the list above (e.g. heavy metals) are rendered out as not acceptable for further QSAR processing. Calculation 2D structures (SMILES and/or SDF) are generated by stripping off accepted organic and inorganic ions. Thus, all the training set and prediction set chemicals are used in their non-ionized form. See 5.1 for further applicability domain definition.

6. Internal validation

6.1 Availability of the training set

Yes

6.2 Available information for the training set

CAS

SMILES

6.3 Data for each descriptor variable for the training set

No

6.4 Data for the dependent variable for the training set:

All

6.5 Other information about the training set

595 compounds are in the training set: 284 positives and 311 negatives.

Of the 948 entries in Appendix I (METI 2002), 843 contained test information for ERalpha binding. Among these 802 were discrete organic substances with available SMILES (see 5.4 for this step). Data was categorised to either active or inactive for binding to ERalpha based on the given RBA values in Appendix 1 (METI 2001); all chemicals, that were assigned a numerical RBA value, were defined as actives and all chemicals assigned with 'N.B' (Not Bound) or 'N.D.' (Not Determined) were defined as negatives (i.e. for these, whether [N.B] or [N.D], the RBA value was less than 0.001). This gave 284 compounds positive and 518 compounds negative for ERalpha binding. To balance the training set in order to prevent bias on the predictive performance of the model 311 negatives were randomly chosen among the 518 negatives.

6.6 Pre-processing of data before modelling

Only structures acceptable for SciQSAR were used in the final training set. That is, only discrete organic chemicals as described in 5.4 were used. In case of replicate structures, one of the replicates was kept if all the compounds had the same activity and all were removed if they had different activity. No further structures accepted by the software were eliminated (i.e. outliers).

6.7 Statistics for goodness-of-fit

SciQSARs own internal performance test of the model gave the following Cooper's statistics for predictions within the applicability domain as defined by SciQSAR (i.e. the first criterion described in 5.1):

- Sensitivity (true positives / (true positives + false negatives)): 89.4%
- Specificity (true negatives / (true negatives + false positives)): 90.0%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 89.7%

6.8 Robustness – Statistics obtained by leave-one-out cross-validation

Not performed.

6.9 Robustness – Statistics obtained by leave-many-out cross-validation

SciQSAR's own internal 10-fold cross-validation (10*10% out) procedure was used for predictions within the applicability domain as defined by SciQSAR (i.e. the first criterion described in 5.1). As the probability domain was not applied (i.e. the second criterion described in 5.2) the accuracy of the predictions when applying this domain can be expected to be higher than reflected in these cross-validation results. This gave the following Cooper's statistics:

- Sensitivity (true positives / (true positives + false negatives)): 76.1%
- Specificity (true negatives / (true negatives + false positives)): 83.3%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 79.8%

6.10 Robustness - Statistics obtained by Y-scrambling

Not performed.

6.11 Robustness - Statistics obtained by bootstrap

Not performed.

6.12 Robustness - Statistics obtained by other methods

Not performed.

7. External validation

7.1 Availability of the external validation set

7.2 Available information for the external validation set

7.3 Data for each descriptor variable for the external validation set

7.4 Data for the dependent variable for the external validation set

7.5 Other information about the training set

7.6 Experimental design of test set

7.7 Predictivity – Statistics obtained by external validation

7.8 Predictivity – Assessment of the external validation set

7.9 Comments on the external validation of the model

External validation has not been performed for this model.

8. Mechanistic interpretation

8.1 Mechanistic basis of the model

The SciQSAR software provides over 400 calculated physico–chemical, electrotopological E-state, connectivity and other molecular descriptors. The descriptors selected for the model may indicate modes of action that are obvious for persons with expert knowledge about the endpoint.

8.2 A priori or posteriori mechanistic interpretation

A posteriori mechanistic interpretation. The descriptors selected for the model may provide a basis for mechanistic interpretation.

8.3 Other information about the mechanistic interpretation

9. Miscellaneous information

9.1 Comments

The model can be used to predict if a chemical can bind to the human Estrogen Receptor alpha (hERalpha) *in vitro*.

9.2 Bibliography

METI (2002) Current status of testing methods development for endocrine disrupters. In 6th Meeting of the Task Force on Endocrine Disrupters Testing and Assessment (EDTA) 24-25 June 2002, Tokyo. Ministry of Economy, Trade and Industry, Japan. Data can be found in Appendix I. Available online on: http://www.meti.go.jp/english/report/data/g020205ae.html

Jensen, G.E., Niemelä, J.R., Wedebye, E.B., and Nikolov, N.G. (2008) QSAR models for reproductive toxicity and endocrine disruption in regulatory use – a preliminary investigation. *SAR and QSAR in Environmental Research*, 19:7–8, 631–641.

9.3 Supporting information