SciMatics SciQSAR model for Androgen Receptor (AR) antagonism (human vector) *in vitro*

1. QSAR identifier

1.1 QSAR identifier (title)

SciMatics SciQSAR model for Androgen Receptor (AR) antagonism (human vector) *in vitro*, Danish QSAR Group at DTU Food.

1.2 Other related models

Leadscope Enterprise model for Androgen Receptor (AR) antagonism (human vector) *in vitro*, Danish QSAR Group at DTU Food.

MultiCASE CASE Ultra model for Androgen Receptor (AR) antagonism (human vector) *in vitro*, 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

Jay Russel Niemelä;

National Food Institute at the Technical University of Denmark;

Gunde Egeskov Jensen;

National Food Institute at the Technical University of Denmark;

Eva Bay Wedebye;

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 is composed of experimental data from our own laboratory and additional data from the literature (see references under 9.2). 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 (human androgen receptor in Chinese Hamster Ovary (CHO) cells).

3.2 Endpoint

QMRF 4. Human Health Effects

QMRF 4.18.c. Endocrine Activity. Other (human Androgen Receptor antagonism in a reporter gene assay)

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.

Among the many biological mechanisms that can result in endocrine disruption, one important is the expression of an antiandrogenic response. Chemicals with antiandrogenic activity counteract the effect of the male sex steroid hormones either by affecting their synthesis or metabolism or by blocking the effects of androgens. Androgens such as testosterone and dihydrotestosterone play a crucial role at several stages of male development and in the maintenance of the male phenotype. The development of the male phenotype during gestation is totally dependent on the action of androgens, and interference with the androgen receptor (AR) at this point of development is hypothesized as being linked to the increased frequency of male reproductive disorders such as testicular dysgenesis syndrome. Blocking of the androgen action may be exerted by antagonism of the AR, that is, by direct interaction of a chemical with AR.

The AR is a member of the nuclear receptor superfamily. Upon ligand binding to the AR in the cytoplasm the receptor undergoes a conformational change and the receptor-ligand dimer is transported to the nucleus where it binds to an androgen response element (RE) on the DNA. This binding modulates the transcription of target genes downstream the RE. The structural diversity of chemicals, which can bind to and affect the activity of AR is very broad. *In vivo* assays for the detection of antiandrogenic action are time-consuming, costly, and labour intensive, which makes them impractical for routine screening and testing of a large number of chemicals. Although *in vitro* data for AR antagonism alone are not sufficient to characterize a compound as an EDC, information on the ability of a chemical to antagonize AR *in vitro* provides an important piece of information for priority setting of chemicals for the more elaborate *in vivo* assays.
For this model training set data originates from reporter gene assays using hAR plasmid transfected Chinese Hamster Ovary (CHO) cells. The training set consists of data from our own laboratory (Vinggaard *et al.* 2008) and data compiled from the literature.

3.4 Endpoint units

No units, 1 for positives and 0 for negatives.

3.5 Dependent variable

Human Androgen Receptor (hAR) antagonism *in vitro*, positive or negative.

3.6 Experimental protocol

The experimental protocol for the data obtained in our own laboratory can be found in Vinggaard *et al.* (2008). Briefly, Chinese Hamster Ovary (CHO) cells were transfected with a plasmid containing a gene coding for the human androgen receptor (AR) and a plasmid containing a gene coding for the reporter enzyme Luciferase. The synthetic androgen, R1881, responsible for AR activity, was added, and the response of 0.1 nM R1881 was set to 100%. Chemicals were tested at various concentrations and data was related to the response of 0.1 nM R1881. Cytotoxicity was determined in parallel using CHO cells transfected with a plasmid containing a gene coding for a constitutively active AR lacking the ligand binding domain. The IC_{25}, defined as the concentration of the test compound that caused a 25% inhibition of the luciferase activity induced by R1881, was calculated for each compound.

For the data obtained from the literature different experimental protocols have been employed. We therefore refer to the references in 9.2 for a specific description of the different protocols.

All the AR antagonism data was separated in to two groups: chemicals reaching an IC_{25} at non-cytotoxic concentration ≤10 μM were defined as positives, and chemicals with IC25 > 10 µM or showing no activity were defined as negatives.

3.7 Endpoint data quality and variability

The dataset from our own laboratory is expected to have low data variability. Because multiple different experimental protocols were used for the data obtained from the literature a certain degree of interlaboratory variability in the data is expected. Jensen *et al.* (2012) compared data where different laboratories had tested the same substances and found an agreement of 83% (29/35) in one case and 91% (40/44) in another. Some chemicals were excluded from the training set due to significant discrepancies between data from different sources without other supporting data.

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 parametric discriminant analysis to create a linear discriminant function (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.

83 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 linear 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 83 descriptors were used. The training set consists of 874 compounds. The descriptor/chemical ratio is 1:10.5 (83:874).

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

874 compounds are in the training set: 231 positives and 643 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)): 60.6%
- Specificity (true negatives / (true negatives + false positives)): 93.6%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 84.9%

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)): 56.3%
- Specificity (true negatives / (true negatives + false positives)): 91.1%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 81.9%

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 has an antagonistic effect on the human androgen receptor *in vitro*.

9.2 Bibliography

Andersen, H.R., Vinggaard, A.M., Rasmussen, T.H., Gjermandsen, I.M., and Bonefeld-Jørgensen, E.C. (2002) Effects of currently used pesticides in assays for estrogenicity, androgenicity, and aromatase activity *in vitro. Toxicology and Applied Pharmacology*, 179, 1-12.

Araki, N., Ohno, K., Takeyoshi, M., and Iida, M. (2005a) Evaluation of a rapid in vitro androgen receptor transcriptional activation assay using AR-EcoScreen^{TM} cells. *Toxicology In Vitro*, 19, 335-352.

Araki, N., Ohno, K., Nakai, M., Takeyoshi, M., and Iida, M. (2005b) Screening for androgen receptor activities in 253 industrial chemicals by in vitro reporter gene assays using AR-EcoScreen^{TM} cells. *Toxicology In Vitro*, 19, 831-842.

Jensen, G.E., Nikolov, N.G., Wedebye, E.B., Ringsted, T., and Niemela, J.R. (2011) QSAR models for anti-androgenic effect--a preliminary study. *SAR and QSAR in Environmental Research*, 22:1-2, 35-49.

Jensen, G.E., Nikolov, N.G., Dreisig, K., Vinggaard, A.M., and Niemelä, J.R. (2012) QSAR Model for Androgen Receptor Antagonism - Data from CHO Cell Reporter Gene Assays. *J Steroids and Hormonal Science*, S2:006, doi:10.4172/2157-7536.S2-006.

Kawamura, Y., Mutsuga, M., Kato, T., Iida, M., and Tanamoto, K. (2005) Estrogenic and Anti-Androgenic Activities of Benzophenones in Human Estroge and Androgen Receptor Mediated Mammalian Reporter Gene Assays. *J Health Sci,* 51:1, 48-54.

Kojima, H., Iida, M., Katsura, E., Kanetoshi, A., Hori, Y., and Kobayashi, K. (2003) Effects of a diphenyl ether-type herbicide, chlornitrofen and its amino derivative on androgen and estrogen receptor activities. *Environ. Health Perspect*, 11:4, 497–502.

Kojima, H., Takeuchi, S., Uramaru, N., Sugihara, K., Yoshida, T., and Kitamura, S. (2009) Nuclear hormone receptor activity of polybrominated diphenyl ethers and their hydroxylated and methoxylated metabolites in transactivation assays using Chinese hamster ovary cells. *Environ Health Perspect*, 117:8, 1210-1218.

Körner, W., Vinggaard, A.M., Térouanne, B., Ma, R., Wieloch, C., Schlumpf, M., Sultan, C., and Soto, A.M. (2004) Interlaboratory comparison of four in vitro assays for assessing androgenic and antiandrogenic activity of environmental chemicals. *Environ Health Perspect*, 112:6, 695-702.

Satoh, K., Ohyama, K., Aoki, N., Iida, M., and Nagai, F. (2004) Study on anti-androgenic effects of bisphenol a diglycidyl ether (BADGE), bisphenol F diglycidyl ether (BFDGE) and their derivatives using cells stably transfected with human androgen receptor, AR-EcoScreen. *Food Chem Toxicol,* 42, 983-993.

Satoh, K., Nonaka, R., Ohyama, K., Nagai, F., Ogata, A., and Iida, M. (2008) Endocrine disruptive effects of chemicals eluted from nitrile-butadiene rubber gloves using reporter gene assay systems. *Biol Pharm Bull,* 31:3, 375-379.

Takeuchi, S., Iida, M., Kobayashi, S., Jin, K., Matsuda, T., and Kojima, H. (2005) Differential effects of phthalate esters on transcriptional activities via human estrogen receptors alpha and beta, and androgen receptor. *Toxicology*, 210, 223-233.

Takeuchi, S., Takahashi, T., Sawada, Y., Iida, M., Matsuda, T., and Kojima, H. (2009) Comparative study on the nuclear hormone receptor activity of various phytochemicals and their metabolites by reporter gene assays using Chinese hamster ovary cells. *Biol Pharm Bull*, 32:2, 195-202.

Vinggaard, A.M., Bonefeld-Joergensen, E.C., and Larsen, J.C. (1999) Rapid and sensitive reporter gene assays for detection of antiandrogenic and estrogenic effects of environmental chemicals. *Toxicol Appl Pharmacol*, 155, 150-160.

Vinggaard, A.M., Nellemann, C., Dalgaard, M., Jørgensen, E.B., and Andersen, H.R. (2002) Antiandrogenic effects in vitro and in vivo of the fungicide prochloraz. *Toxicol Sci*, 69, 344-353.

Vinggaard, A.M., Niemelä, J.R., Wedebye, E.B., and Jensen, G.E. (2008) Screening of 397 Chemicals and Development of a Quantitative Structure - Activity Relationship Model for Androgen Receptor Antagonism. *Chem. Res. Toxicol*., 21, 813-823.

9.3 Supporting information