SciMatics SciQSAR model for human Thyroid hormone Receptor alpha (hTRa) binding *in vitro*

1. QSAR identifier

1.1 QSAR identifier (title)

SciMatics SciQSAR model for human Thyroid hormone Receptor alpha (hTRa) binding *in vitro*, Danish QSAR Group at DTU Food.

1.2 Other related models

Leadscope Enterprise model for human Thyroid hormone Receptor alpha (hTRa) binding *in vitro*, Danish QSAR Group at DTU Food.

MultiCASE CASE Ultra model for human Thyroid hormone Receptor alpha (hTRa) binding *in vitro*, Danish QSAR Group at DTU Food.

1.3 Software coding the model

SciMatics SciQSAR version 2.3.0.0.12.

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;

Trine Klein Reffstrup;

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

Sine Abildgaard Rosenberg;

National Food Institute at the Technical University of Denmark;

Marianne Dybdahl;

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;

Eva Bay Wedebye;

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 published literature (see 6.6 and 9.2 for more details). 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 (cell-free assay containing the human Thyroid hormone Receptor alpha, hTRa).

3.2 Endpoint

QMRF 4. Human Health Effects

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

3.3 Comment on endpoint

Endocrine disruption not only involves the sex hormone system, but also includes disruption of the thyroid hormone (TH) system. Exposure to chemicals can potentially disrupt the TH system by a number of different mechanisms, one of them being binding to the thyroid hormone receptor (TR). Thyroid disrupting chemicals (TDCs) can affect important physiological processes such as metabolism, growth and development, including development of the brain, and are therefore of high concern (OECD 2006; Murk *et al.*2013).

Thyroid hormones are produced in the thyroid gland and exert a wide array of effects through binding of the active hormone triiodothyronine (T3) to TR. The TR is a member of the nuclear receptor superfamily and two isoforms of the receptor exist in humans, TR alpha (TRa) and beta (TRb). The TR usually heterodimerize with the retinoid X receptor (RXR), another nuclear receptor, and binds to a thyroid response element (TRE) on the DNA. At low levels of T3, a nuclear receptor transcriptional co-repressor is bound to the activation function 2 (AF-2) domain in the Ligand Binding Domain (LBD) of TR and represses the basal transcription through chromatin deacetylase activity. When the thyroid hormone level is high, T3 binds to the TRs LBD causing conformational changes in the LBD that leads to release of the co-repressor from the AF-2 domain and the basal transcriptional activity is restored. Subsequently a nuclear receptor transcriptional co-activator (SRC-1, SRC-2 and others) bind to the AF-2 domain. This binding causes a destabilization of the chromatin and enhances the transcriptional activity through histone acetylation and contacts with the basal transcriptional machinery. Together the binding of T3 and a co-activator to TR leads to an increased transcription of the genes downstream the TRE.

Multiple assays have been established for different mechanisms in the TH system in order to identify TDCs. For this (Q)SAR model data compiled from *in vitro* assays for binding to the human TR alpha (hTRa) have been used to make a model that within the defined applicability domain (see 5.) can predict if a chemical binds to the LBD of hTRa. The *in vitro* binding assay is a cell free assay that detects a compounds affinity to hTRa by determining its ability to compete with radioactive [$_{125}$I]triiodothyronine ([^{125}I]T3) for binding to hTRa. Using the concentration response curve an IC50 value can be determined for each compound as the concentration of compound measured in µM required to inhibit 50% of the binding of [^{125}I]T3 to hTRa. The assay does not say anything about if the binding induces transcription of the target genes (i.e. if the chemical is an hTRa agonist or antagonist).

3.4 Endpoint units

-log(IC50), where IC50 is in µM.

3.5 Dependent variable

Binding affinity to human Thyroid Receptor alpha *in vitro*, –log(IC50) (µM).

3.6 Experimental protocol

Data for the training set was obtained from studies using a hTRa-binding *in vitro* assay to measure affinity of a compound to hTRa. Currently this assay does not have an internationally agreed guideline. The assay has been described in Greenidge et al. (1998) The assay is known to have a potential for a high rate of false negatives (Murk *et al.*2013).

3.7 Endpoint data quality and variability

All data in the final training sets originated from the same laboratory (Karo Bio). From the publications it is expected that the tests were performed by the same protocol, possibly with minor justifications, however it is not explicitly stated. The overall data variability is assumed to be low due to the fact that all data points originate from the same laboratory.

4. Defining the algorithm

4.1 Type of model

A continuous (Q)SAR model based on calculated molecular descriptors, and if available own or third-party descriptors or measured endpoints can be imported and used as descriptors.

4.2 Explicit algorithm

As the endpoint for this training set was continuous the (Q)SAR model was made by use of Partial Least Squares (PLS) regression method (see 4.5). The derived algorithm 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 own or third-party descriptors or use measured endpoints as custom descriptors.

4.4 Descriptor selection

Genetic algorithm (GA) analysis was used to select descriptors to make the best PLS regression model.

In this model the Initial population size was set to 80, the Tournament size was set at 8 and the number of iterations was 50,000. The remaining settings for the GA were the default settings in SciQSAR.

4.5 Algorithm and descriptor generation

For a binary classification problem SciQSAR uses discriminant analysis (DA) to make a (Q)SAR model. SciQSAR implements the entire range of 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 not assumed to 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. In general, 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.

If the data outcome is continuous, regression analysis is used to build the predictive model. Within SciQSAR several regression methods are available: for independent variables ordinary multiple regression (OMR) (only for a small number of independent variables), stepwise regression (SWR) or all possible subsets regression (PSR) is useful, and for analysis of variables with high correlation or multicollinearity regression on principal components (PCR) or partial least squares regression (PLS) should be used. For the above mentioned regression methods a built-in cross validation procedure tests how stable the built models are.

In SciQSAR descriptors for regression analysis are selected with the use of genetic algorithm (GA) analysis. The GA method sequentially generates sets of descriptors. Selection of the best descriptors is accomplished through an algorithm which simulates mutation and genetic cross-over. Each set of descriptors (generation) is evaluated and its ‘‘goodness of fit’’ determined by a set of criteria. The algorithm makes use of the whole descriptor pool to select a set of descriptors with good regression statistics (high R^2 and Q^2). The performance of each candidate model is assessed using an automated cross validation process within SciQSAR. (Contrera et al. 2004)

4.6 Software name and version for descriptor generation

SciMatics SciQSAR version 2.3.0.0.12.

4.7 Descriptors/chemicals ratio

11 PLS components were used to make this model.

5. Defining Applicability Domain

5.1 Description of the applicability domain of the model

Domain of applicability of a (Q)SAR model is partly a function of the molecular coverage of the test molecule relative to the molecules in the training data set. If a test molecule is not well-represented in the training data molecular library, the test molecule will be considered outside of the domain of the model.

5.2 Method used to assess the applicability domain

For a (Q)SAR model on a binary endpoint the probability (p) of test compounds membership in one of the two (low or high risk) classes is calculated and used for determining whether the test chemical is within the models domain of applicability. The probability of membership in a class is a measure of how well training set knowledge is able to discriminate compounds with high risk from those with low risk 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.

For a continuous (Q)SAR model the test chemical is considered out of domain if SciQSAR cannot calculate each descriptor value for the test chemical.

Predictions outside the $-\log IC_{50}$ interval [-5.2;1.39] are considered out of domain.

5.3 Software name and version for applicability domain assessment

SciMatics SciQSAR version 2.3.0.0.12.

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 Leadscope. 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). Calculation 2D structures (SMILES and/or SDF) are generated by stripping off ions (of the accepted list given above). Thus, all the training 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

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

118 compounds are in the training set.

6.6 Pre-processing of data before modeling

Data used to develop the (Q)SAR model was compiled from two public databases, ChEMBL and BindingDB, on the internet . Data in the two databases originate from published literature, and in this case all the data were made by Karo Bio (Karo Bio AB, Novum, Huddinge S-141 57, Sweden) (Carlsson et al. 2002, Ye et al. 2003, Hangeland et al. 2004, 2005, Hedfors et al. 2005, Collazo et al. 2006, Koehler et al. 2006, Li et al. 2006, Garg et al. 2007, Malm et al. 2007). The initial data sets consisted of results from various *in vitro* assays measuring different endpoints related to the TH system. Therefore a thorough manual review of the data sources was performed, and data originating from other assays than the hTRa binding assay were removed.

Only structures acceptable for the commercial software were used in the training set, that is only discrete organic chemicals as described in 5.4 were used. Subsequently, duplicates were identified and removed according to defined criteria: IC50 values for duplicates were compared and in case the difference between the IC50 values were more than 10 fold both data points were removed. If the difference was less than 10 fold the data point with the lowest IC50 value was kept (conservative approach). For this model no duplicates had a 10 fold or more difference in the IC50 values. IC50 values were transformed into –logIC50.

6.7 Statistics for goodness-of-fit

The R-squared (internal performance) was calculated by SciQSAR and gave 0.64.

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

Not performed. (It is not a preferred measurement for evaluating large models).

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

A cross-validation was performed in SciQSAR using SqiQSARs own validation procedure with default settings. The resulting Q-squared (predictive performance) was 0.57.

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 global model identifies structural features and molecular descriptors which in the model development was found to be statistically significant associated with effect. Many predictions may indicate modes of action that are obvious for persons with expert knowledge for the endpoint.

8.2 A priori or posteriori mechanistic interpretation

A posteriori mechanistic interpretation. The identified structural features and molecular descriptors may provide 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 a chemicals binding affinity to the human Thyroid hormone Receptor alpha *in vitro*. The outcome from the prediction is –logIC50 and this can be transformed to make an estimate for the IC50 (μM) value for the chemical.

9.2 Bibliography

Carlsson, B., Singh, B.N., Temciuc, M., Nilsson, S., Li, Y.L., Mellin, C., and Malm, J. (2002) Synthesis and Preliminary Characterization of a Novel Antiarrhythmic Compound (KB130015) with an Improved Toxicity Profile Compared with Amiodarone. *J. Med. Chem*., 45:3, 623 – 630.

Collazo, A.M.G., Koehler, K.F., Garg, N., Färnegårdh, M., Husman, B., Ye, L., Ljunggren, J., Mellström, K., Sandberg, J., Grynfarb, M., Ahola, H., and Malm, J. (2006) Thyroid receptor ligands. Part 5: Novel bicyclic agonist ligands selective for the thyroid hormone receptor *β. Bioorg. Med. Chem. Lett*., 16:5, 1240 – 1244.

Garg, N., Li, Y.-L., Collazo, A.M.G., Litten, C., Ryono, D.E., Zhang, M., Caringal, Y., Brigance, R.P., Meng, W., Washburn, W.N., Agback, P., Mellström, K., Rehnmark, S., Rahimi-Ghadim, M., Norin, T., Grynfarb, M., Sandberg, J., Grover, G., and Malm, J. (2007) Thyroid receptor ligands. Part 8: Thyromimetics derived from N-acylated-a-amino acid derivatives displaying modulated pharmacological selectivity compared with KB-141. *Bioorg. Med. Chem. Lett*., 17:15, 4131 – 4134.

Greenidge,P.A., Carlsson, B., Bladh, L.-G., and Gillner, M. (1998) Pharmacophores Incorporating Numerous Excluded Volumes Defined by X-ray Crystallographic Structure in Three-Dimensional Database Searching:

Application to the Thyroid Hormone Receptor. *J. Med. Chem.*, 41*,* 2503 - 2512.

Hangeland, J.J., Doweyko, A.M., Dejneka, T., Friends, T.J., Devasthale, P., Mellström, K., Sandberg, J., Grynfarb, M., Sack, J.S., Einspahr, H., Färnegårdh, M., Husman, B., Ljunggren, J., Koehler, K., Sheppard, C., Malm, J., and Ryono, D.E. (2004) Thyroid receptor ligands. Part 2: Thyromimetics with improved selectivity for the thyroid hormone receptor beta. *Biorg. Med. Chem. Lett.*, 14, 3549 – 3553.

Hangeland, J.J., Friends, T.J., Doweyko, A.M., Mellström, K., Sandberg, J., Grynfarb, M., and Ryono, D.E. (2005) A new class of high affinity thyromimetics containing a phenyl-naphthylene core. *Bioorg. Med. Chem. Lett*., 15:20, 4579 - 4584.

Hedfors, Å., Appelqvist, T., Carlsson, B., Bladh, L.-G., Litten, C., Agback, P., Grynfarb, M., Koehler, K.F., and Malm, J. (2005) Thyroid Receptor Ligands. 3. Design and Synthesis of 3,5-Dihalo-4- alkoxyphenylalkanoic Acids as Indirect Antagonists of the Thyroid Hormone Receptor. *J. Med. Chem*., 48:9, 3114 – 3117.

Koehler, K., Gordon, S., Brandt, P., Carlsson, B., Bäcksbro-Saeidi, A., Appelqvist, T., Agback, P., Grover, G.J., Nelson, W., Grynfarb, M., Färnegårdh, M., Rehnmark, S., and Malm, J. (2006) Thyroid Receptor Ligands. 6. A High Affinity “Direct Antagonist” Selective for the Thyroid Hormone Receptor. *J. Med. Chem*., 49:23, 6635 – 6637.

Li, Y.-L., Litten, C., Koehler, K.F., Mellström, K., Garg, N., Collazo, A.M.G., Färnegård, M., Grynfarb, M., Husman, B., Sandberg, J., and Malm, J. (2006) Thyroid receptor ligands. Part 4: 40-amido bioisosteric ligands selective for the thyroid hormone receptor beta. *Bioorg. Med. Chem. Lett*., 16:4, 884 – 886.

Malm, J., Gordon, S., Brandt, P., Carlsson, B., Agback, P., Bäckbro-Saeidi, A., and Sandberg, J. (2007) Thyroid receptor ligands. Part 7: Indirect antagonists of the thyroid hormone receptor with improved affinity. *Bioorg. Med. Chem. Lett*., 17:7, 2018 – 2012.

Murk, A.J., Rijntjes, E., Blaauboer, B.J., Clewell, R., Crofton, K.M., Dingemans, M.M.L., Furlow, J.D., Kavlock, R., Köhrle, J., Opitz, R., Traas, T., Visser, T.J., Xia, M., and Gutleb, A.C. (2013) Mechanism-based testing strategy using in vitro approaches for identification of thyroid hormone disrupting chemicals. *Toxicology in Vitro*, 27, 1320–1346.

OECD 2006: OECD series on testing and assessment, Number 57: Detailed Review Paper on Thyroid Hormone Disruption Assays (ENV/JM/MONO(2006)24).

Ye, L., Li, Y.-L., Mellström, K., Mellin, C, Bladh, L.-G., Koehler, K., Garg, N., Collazo, A.M.G., Litten, C., Husman, B., Persson, K., Ljunggren, J., Grover, G., Sleph, P.G., George, R., and Malm, J. (2003) Thyroid Receptor Ligands. 1. Agonist Ligands Selective for the Thyroid Receptor β_1. *J. Med. Chem.*, 46:9, 1580 – 1588.

9.3 Supporting information