Leadscope Enterprise model for binding to the human pregnane X receptor (hPXR) *in vitro* (NIH data)

1. QSAR identifier

1.1 QSAR identifier (title)

Leadscope Enterprise model for binding to the human pregnane X receptor (hPXR) *in vitro* (NIH data), Danish QSAR Group at DTU Food.

1.2 Other related models

MultiCASE CASE Ultra model for binding to the human pregnane X receptor (hPXR) *in vitro* (NIH data), Danish QSAR Group at DTU Food.

SciMatics SciQSAR model for binding to the human pregnane X receptor (hPXR) *in vitro* (NIH data), Danish QSAR Group at DTU Food.

1.3. Software coding the model

Leadscope Predictive Data Miner, a component of Leadscope Enterprise version 3.1.1-10.

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

Marianne Dybdahl;

National Food Institute at the Technical University of Denmark;

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

Roberts, G., Myatt, G. J., Johnson, W. P., Cross, K. P., and Blower, P. E. J. (2000) LeadScope: Software for Exploring Large Sets of Screening Data. *Chem. Inf. Comput. Sci.*, 40, 1302-1314.

Cross, K.P., Myatt, G., Yang, C., Fligner, M.A., Verducci, J.S., and Blower, P.E. Jr. (2003) Finding Discriminating Structural Features by Reassembling Common Building Blocks. *J. Med. Chem.*, *46,* 4770-4775.

Valerio, L. G., Yang, C., Arvidson, K. B., and Kruhlak, N. L. (2010) A structural feature-based computational approach for toxicology predictions. *Expert Opin. Drug Metab. Toxicol.*, 6:4, 505-518.

2.8 Availability of information about the model

The training data set is non-proprietary and was kindly provided by Dr. Sunita Shukla (NIH Chemical Genomics Center, National Institutes of Health, Bethesda, Maryland) (see 9.2, Shukla *et al.* 2009). 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 pregnane X receptor (hPXR)).

3.2 Endpoint

QMRF 4. Human Health Effects

QMRF 4.18.a. Endocrine Activity. Receptor-binding (hPXR)

3.3 Comment on endpoint

The pregnane X receptor (PXR) is a member of the nuclear receptor superfamily and has a key role in regulating the metabolism and transport of structurally diverse endogenous and exogenous compounds. PXR is primarily expressed in the liver, intestine and kidney, but its expression is also seen in lung, stomach, the blood–brain barrier, placenta, bone marrow, and specific regions of the brain. The ligand binding domain (LBD) of PXR contains a ligand binding pocket and a ligand dependent AF-2 region that binds transcriptional cofactors. The PXR-LBD is unusually large and flexible and can change its shape to accommodate molecules that vary in size from 300 Da to over 800 Da in molecular weight.
In general PXR antagonists are interacting with the AF-2 site of the LBD while agonists bind to the ligand binding pocket.
When a ligand binds to PXR-LBD, PXR forms a heterodimer with the retinoid X receptor (RXR) in the nucleus. This dimer then binds to a PXR response element (RE) in the promotor region of the target gene and transcription is initiated. PXR target genes include cytochrome P-450 enzymes (e.g. CYP3A4), conjugation enzymes and transporters. These proteins are not only important in metabolism and elimination of xenobiotics such as drugs and environmental chemicals, but also affect the metabolism of many endogenous compounds, e.g. bile acids and hormones. Therefore induction of PXR by a xenobiotic can potentially interfere with normal physiological functions. Identification of potential PXR agonists would be of importance for the evaluation of health risk of environmental chemicals and drugs.
In this model data from a human PXR-LBD binding *in vitro* assay was used (Shukla *et al.* 2009). The classification of active and inactive compounds is described in Shukla *et al.* (2009). Briefly, compounds yielding a concentration-response curve (CRC) assigned to the highest confident classes (class 1 and 2) were considered active. Compounds showing no concentration-response relationship were considered inactive.

3.4 Endpoint units

No units, 1 for positives and 0 for negatives.

3.5 Dependent variable

Binding to the human PXR-LBD *in vitro*, positive or negative.

3.6 Experimental protocol

The experimental protocol is described in Shukla *et al.* (2009). Briefly, data was obtained using a time resolved fluorescence resonance energy transfer (TR-FRET) assay, which measured binding to the human PXR in a cell free assay. The TR-FRET assay reports on the ability of a test ligand to displace a fluorescein-labeled tracer molecule from the nuclear receptor.

3.7 Endpoint data quality and variability

As data originates from a single source the variability is assumed low. Shukla *et al.* (2009) alleviated batch-to-batch variations in PXR performance through reformulation of PXR storage buffer. A test for experimental artifacts was also performed (Shukla *et al.* 2009).

4. Defining the algorithm

4.1 Type of model

A categorical (Q)SAR model based on structural features and numeric molecular descriptors.

4.2 Explicit algorithm

This is a categorical (Q)SAR model made by use of partial logistic regression (PLR). The specific implementation is proprietary within the Leadscope software.

4.3 Descriptors in the model

structural features,

aLogP,

polar surface area,

number of hydrogen bond donors,

Lipinski score,

number of rotational bonds,

parent atom count,

parent molecular weight,

number of hydrogen bond acceptors

4.4 Descriptor selection

Leadscope Predictive Data Miner is a software program for systematic sub-structural analysis of a chemical using predefined structural features stored in a template library, training set-dependent generated structural features (scaffolds) and calculated molecular descriptors. The feature library contains approximately 27,000 pre-defined structural features and the structural features chosen for the library are motivated by those typically found in small molecules: aromatics, heterocycles, spacer groups, simple substituents. Leadscope allows for the generation of training set-dependent structural features (scaffold generation), and these features can be added to the pre-defined structural features from the library and be included in the descriptor selection process. It is possible in Leadscope to remove redundant structural features before the descriptor selection process and only use the remaining features in the descriptor selection process. Besides the structural features Leadscope also calculates eight molecular descriptors for

each training set structure: the octanol/water partition coefficient (alogP), hydrogen bond acceptors (HBA), hydrogen bond donors (HBD), Lipinski score, atom count, parent compound molecular weight, polar surface area (PSA) and rotatable bonds. These eight molecular descriptors are also included in the descriptor selection process.

Leadscope has a default automatic descriptor selection procedure. This procedure selects the top 30% of the descriptors (structural features and molecular descriptors) according to X^2-test for a binary variable, or the top and bottom 15% descriptors according to *t*-test for a continuous variable. Leadscope treats numeric property data as ordinal categorical data. If the input data is continuous such as IC_{50} or cLogP data, the user can determine how values are assigned to categories: the number of categories and the cut-off values between categories. (Roberts *et al.*2000).

When developing this model, intermediate models with application of different modelling approaches in Leadscope were tried:

1. 'Single model' using only the Leadscope pre-defined structural features, i.e. no scaffolds, and calculated molecular descriptors for descriptor selection.
2. 'Single model' using both the Leadscope pre-defined structural features and the training set dependent features (scaffolds generation) as well as the calculated molecular descriptors in the descriptor selection.
3. 'Single model' using both Leadscope pre-defined structural features and the training set dependent features (scaffolds generation), with subsequent removal of redundant structural features, and calculated molecular descriptors for descriptor selection.
4. 'Composite model' using only the Leadscope pre-defined structural features, i.e. no scaffolds, and calculated molecular descriptors in the descriptor selection.
5. 'Composite model' using both Leadscope pre-defined structural features and the training set dependent features (scaffolds generation) as well as the calculated molecular descriptors in the descriptor selection.

Based on model performance as measured by a preliminary cross-validation the model developed using approach number 1. was chosen.

In this model the descriptors were automatically selected among the pre-defined structural features and the eight molecular descriptors.

4.5 Algorithm and descriptor generation

For descriptor generation see 4.4.

After selection of descriptors the Leadscope Predictive Data Miner program performs partial least squares (PLS) regression for a continuous response variable, or partial logistic regression (PLR) for a binary response variable, to build a predictive model. By default the Predictive Data Miner performs leave-one-out or leave-groups-out (in the latter case, the user can specify any number of repetitions and percentage of structures left out) cross-validation on the training set depending on the size of the training set. In the cross-validation made by Leadscope the descriptors selected for the 'mother model' are used when building the validation submodels and they therefore have a tendency to be overfittet and give overoptimistic validation results.

In this model, because of the categorical outcome in the response variable, PLR was used to build the predictive model. For this model 370 descriptors were selected to build the model. These include 9 Leadscope calculated molecular descriptors, 359 hierarchy features, and 2 dynamic features. The 370 descriptors were distributed on 2 PLS factors.

4.6 Software name and version for descriptor generation

Leadscope Predictive Data Miner, a component of Leadscope Enterprise version 3.1.1-10.

4.7 Descriptors/chemicals ratio

In this model 370 descriptors were used and distributed on 2 PLS factors. The training set consists of 629 compounds. The descriptor/chemical ratio is 1:1.7 (370:629).

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 of a structural domain in Leadscope and the in-house further probability refinement algorithm on the output from Leadscope to reach the final applicability domain call.

1. Leadscope

For assessing if a test compound is within the structural applicability domain of a given model Leadscope examines whether the test compound bears enough structural resemblance to the training set compounds used for building the model (i.e. a structural domain analysis). This is done by calculating the distance between the test compound and all compounds in the training set (distance = 1 - similarity). The similarity score is based on the Tanimoto method. The number of neighbours is defined as the number of compounds in the training set that have a distance equal to or smaller than 0.7 with respect to the test compound. The higher the number of neighbours, the more reliable the prediction for the test compound. Statistics of the distances are also calculated. Effectively no predictions are made for test compounds which are not within the structural domain of the model or for which the molecular descriptors could not be calculated in Leadscope.

2. The Danish QSAR group

In addition to the general Leadscope structural applicability domain definition the Danish QSAR group has applied a further requirement to the applicability domain of the model. That is only positive predictions with a probability equal to or greater than 0.7 and negative predictions with probability equal to or less than 0.3 are accepted. Predictions within the structural applicability domain but with probability between 0.5 to 0.7 or 0.3 to 0.5 are defined as positives out of applicability domain and negatives out of applicability domain, respectively. When these predictions are wed out the performance of the model in general increases at the expense of reduced model coverage.

5.2 Method used to assess the applicability domain

Leadscope does not generate predictions for test compounds which are not within the structural domain of the model or for which the molecular descriptors could not be calculated.

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

5.3 Software name and version for applicability domain assessment

Leadscope Predictive Data Miner, a component of Leadscope Enterprise version 3.1.1-10.

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 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

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

629 compounds are in the training set: 299 positives and 330 negatives.

The hPXR binding data in Shukla *et al.* (2009) included compounds from 6 different chemical libraries. Out of the 8,280 compounds screened in the TR-FRET assay, compounds defined as "high confidence" actives and an equal number of randomly selected inactive compounds were used for preparation of the training set (Dybdahl *et al.* 2012).

6.6 Pre-processing of data before modelling

Only structures acceptable for Leadscope 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

Not performed.

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 five times two-fold 50 % cross-validation was performed. This was done by randomly removing 50% of the full training set used to make the “mother model”, where the 50% contains the same ratio of positive and negatives as the full training set. A new model (validation submodel) was created on the remaining 50% using the same settings in Leadscope but with no information from the “mother model” regarding descriptor selection etc. The validation submodel was applied to predict the removed 50% (within the defined applicability domain for the submodel). Likewise, a validation submodel was made on the removed 50% of the training set and this model was used to predict the other 50% (within the defined applicability domain for this submodel). This procedure was repeated five times.

Predictions within the defined applicability domain for the ten validation submodels were pooled and Cooper’s statistics calculated. This gave the following results for the 58.4% (1836/(5*629)) of the predictions which were within the applicability domain:

- Sensitivity (true positives / (true positives + false negatives)): 80.4%
- Specificity (true negatives / (true negatives + false positives)): 80.4%
- Concordance ((true positives + true negatives) / (true positives + true negatives + false positives + false negatives)): 80.4%

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

Yes

7.2 Available information for the external validation set

SMILES

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

All

7.5 Other information about the training set

Two test sets have been used for external validations of the model made in the former version of the Leadscope software (version 3.04-10, see Dybdahl *et al.* 2012). Only minor differences exist between this and the current version of the model and therefore the external validations are likely to be useful for the current version as well.
The first test set (test set 1) consisted of human PXR activation data (n=145) determined by a reporter gene assay (Khandelwal *et al.* 2008). After removing molecules also present in the training set and two molecules known to be PXR antagonists (fluconazole and ketonazole), the test set consisted of 120 chemicals (68 actives and 52 inactives).
The second test set (test set 2) consisted of the remaining 3351 inactive compounds with respect to PXR binding, which were not used in the balanced training set of the model (Shukla *et al.* 2009). (Dybdahl *et al.* 2012)

7.6 Experimental design of test set

Test set 1 consisted of human PXR activation data determined by a reporter gene assay. See Khandelwal *et al.* (2008) for a description of the experimental protocol.

Test set 2 consisted of human PXR binding data determined by a TR-FRET assay equal to that used to obtain the human PXR binding data for the training set. See 3.6 for a description of the experimental protocol.

7.7 Predictivity – Statistics obtained by external validation

Test set 1:

- Sensitivity (true positives / (true positives + false negatives)): 57.9%
- Specificity (true negatives / (true negatives + false positives)): 83.9%

Test set 2:

- Specificity (true negatives / (true negatives + false positives)): 82.1%

7.8 Predictivity – Assessment of the external validation set

7.9 Comments on the external validation of the model

The lower sensitivity of test set 1 most likely reflects the fact that the external test set and the training set are from two different assay types (reporter gene vs. binding). The reporter gene assay in addition to identifying compound binding to the receptor also identifies compounds activating the receptor through other routes than receptor binding. Therefore this model cannot identify all compounds activating the PXR. (Dybdahl *et al.* 2012)

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 binding to the human pregnane X receptor (PXR) *in vitro*. It should be noted that the model cannot predict if the binding compound have an agonistic (i.e. initiate gene transcription) or antagonistic effect.

9.2 Bibliography

Dybdahl, M., Nikolov, N.G., Wedebye, E.B., Jónsdóttir, S.Ó., and Niemelä, J.R. (2012) QSAR model for human pregnane X receptor (PXR) binding: Screening of environmental chemicals and correlations with genotoxicity, endocrine disruption and teratogenicity. *Toxicology and Applied Pharmacology,* 262:3, 301–309.

Shukla, S.J., Nguyen, D.T., MacArthur, R., Simeonov, A., Frazee, W.J.,Hallis, T.M., Marks, B.D., Singh, U., Eliason, H.C., Printen, P., Austin, C.P., Inglese, J. and Auld, D.S. (2009) Identification of Pregnane X Receptor Ligands Using Time-Resolved Fluorescence Resonance Energy Transfer and Quantitative High-Throughput Screening. *Assay & Drug Development Technologies,* 7:2, 143-169.

Khandelwal, A., Krasowski, M.D., Reschly, E.J., Sinz, M.W., Swaan, P.W., and Ekins, S. (2008) Machine Learning Methods and Docking for Predicting Human Pregnane X Receptor Activation. *Chem. Res. Toxicol.,* 21, 1457–1467.

9.3 Supporting information