QSAR Modeling for the Predictions of Androgen Receptor Pathway Activity

Q Zang[1], N Kleinstreuer[1], D Allen[1], W Casey[2], R Judson[3]

[1]ILS, RTP, NC, USA; [2]NIH/NIEHS/DNTP/NICEATM, RTP, NC, USA;
[3]EPA/ORD/NCCT, RTP, NC, USA

Introduction

- U.S. (7 U.S.C. 136, 110 Stat 1613) and international regulations require that certain chemicals be tested to detect activity that could potentially affect the function of the estrogen, androgen, and thyroid hormone pathways. As many as 30,000 chemicals may require such testing, with several hundred new chemicals being added to this total each year (EPA 2011).
- The U.S. Environmental Protection Agency (EPA) ToxCast chemical research program (Dix et al. 2007; Judson et al. 2010) and the Tox21 U.S. federal partnership (Tice et al. 2013) include multiple in vitro high-throughput screening (HTS) assays that measure the effects of chemicals on endocrine hormone pathways.
- Following an approach used to model the estrogen receptor pathway (EPA and NICEATM 2014), we constructed a mathematical model to predict chemically induced androgen receptor (AR) activity based on nine ToxCast/Tox21 HTS assays that map to the AR pathway.
- We used this model to train and test quantitative structure–activity relationship (QSAR) models for both binary classification and potency prediction for putative AR antagonists. We then made AR activity predictions for a larger chemical set.

High-Throughput Screening Data

- Nine AR pathway assays (**Table 1)** were used to generate data on 1855 chemicals during ToxCast Phases I and II.
- All concentration–response assay data were analyzed using the ToxCast data analysis pipeline (http://actor.epa.gov/), which automates the processes of baseline correction, normalization, curve-fitting, hit-calling, and AC50 (half-maximal activity) determination, and detects a variety of potential confounders (annotated as “caution flags”).

Table 1 Assays Used in the AR Pathway Model

ID	Assay Name	Source	Gene	Species	Type
A1	NVS human AR	Novascreen	AR	Homo sapiens	Cell-free biochemical radioligand receptor binding
A2	NVS chimpanzee AR	Novascreen	AR	P. troglodytes	Cell-free biochemical radioligand receptor binding
A3	OT_AR_ARSRC1_0480	Odyssey Thera	AR;SRC	Homo sapiens	AR-SRC1 cofactor recruitment
A4	OT_AR_ARSRC1_0960	Odyssey Thera	AR;SRC	Homo sapiens	AR-SRC1 cofactor recruitment
A5	ATG_AR_TRANS	Attagene	AR	Homo sapiens	RNA TA reporter gene
A6	Tox21_AR_BLA_Agonist_ratio	NCGC	AR	Homo sapiens	Beta-lactamase TA reporter gene
A7	Tox21_AR_LUC_MDAKB2_Agonist	NCGC	AR	Homo sapiens	Luciferase TA reporter gene
A8	Tox21_AR_BLA_Antagonist_ratio	NCGC	AR	Homo sapiens	Beta-lactamase TA reporter gene
A9	Tox21_AR_LUC_MDAKB2_Antagonist	NCGC	AR	Homo sapiens	Luciferase TA reporter gene

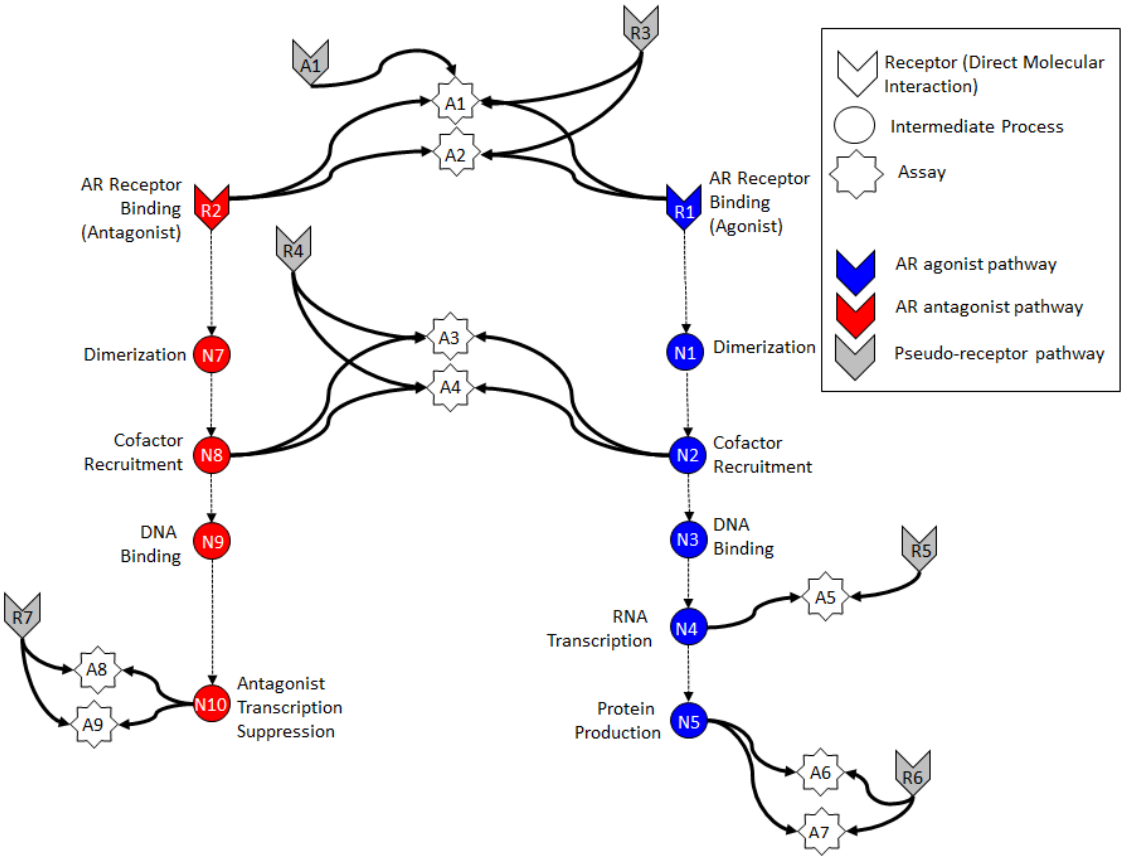
Abbreviations: AR = androgen receptor; NCGC = National Institutes of Health Chemical Genomics Center; SRC = c-Src tyrosine kinase; TA = transactivation.

AR Pathway Model

- **Figure 1** presents a model of the AR pathway and shows how the assays listed in **Table 1** map to the pathway. The model is based on the molecular events that typically occur in a receptor-mediated response**.** The assays listed in **Table 1** are represented in **Figure 1** as white stars.
- The model assumes that a chemical that interacts with the AR will bind in either or both of the agonist or antagonist conformations, triggering activity in the appropriate pathway.
- The agonist pathway starts at the interaction of a chemical with a nuclear AR (receptor node R1), and continues through dimerization (node N1), translocation, co-factor recruitment (N2), DNA binding (N3), mRNA transcription (N4), and protein translation (N5). The antagonist pathway is similarly represented by receptor node R2 and nodes N7, N8, N9, and N10. Dimerization and DNA binding are the only processes not assessed by one or more of the assays listed in **Table 1**.

- Potential assay interference pathways are shown in **Figure 1** as "pseudo-receptors" (gray arrow nodes).

Figure 1 AR Pathway Model



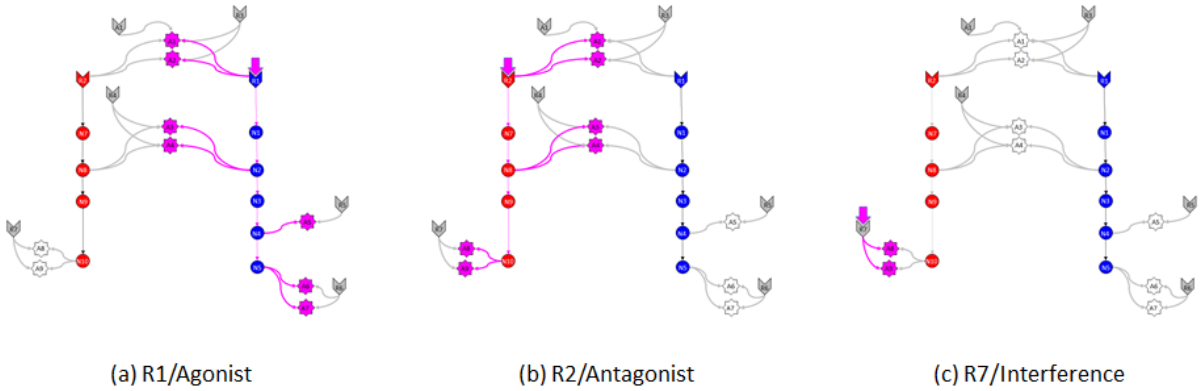
Abbreviations: AR = androgen receptor

Colored arrow nodes (R1/R2) represent "receptors" with which a chemical can directly interact. Colored circles represent intermediate biological processes that are not directly observable. White stars represent the assays that measure activity at the biological nodes. Arrows represent transfer of information. Gray arrow nodes represent biological interference pathways (R3-R7) or technology-specific interference (one example shown, A1).

Chemical Interactions with AR Pathway Model

- Examples of how a specific chemical may interact with the AR pathway are shown in **Figure 2**, in which pink highlighting represents the expected activity from a true agonist (**Figure 2a**), a true antagonist (**Figure 2b**), or a chemical causing assay interference (**Figure 2c**).

Figure 2 Examples of Chemical Interactions with AR Pathway Model

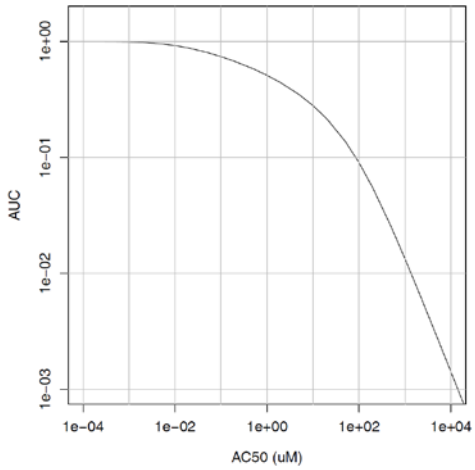


Abbreviations: AR = androgen receptor

Activity of ToxCast Chemicals in the AR Pathway Model

- The activity of a chemical against the AR pathway is represented by a simple linear model that calculates area under the curve (AUC).
- **Figure 3** is a calibration curve to aid interpretation of the AUC distributions, showing for example that an AUC of 0.1 is equivalent to an AC50 against the AR pathway at ~100 µM.
- The AR pathway model AUC scores were previously shown to be 96% predictive against a set of 23 reference chemicals, and correctly identified environmental compounds with known anti-androgenic activity in vivo (Kleinstreuer et al. 2015).

Figure 3 Calibration Curve for AR Pathway Model Scores



Abbreviations: AC50 = half maximal activity concentration; AR = androgen receptor; AUC = area under the curve.

QSAR Predictions of AR Antagonist Activity

- We focused on building QSAR models for AR antagonism because anti-androgenic activity is the response of highest environmental relevance and concern.
- For construction of the QSAR models, chemicals with an AUC antagonist score greater than or equal to 0.01 (~AC50 ≤ 1mM) were considered positives (**Table 2**, color-coded amber), and chemicals with AUC antagonist scores less than 0.01 were considered negatives (**Table 2**, color-coded blue).

Table 2 Data Set Used for the QSAR Model

AUC Antagonist Score	AUC ≥ 0.1	0.1 > AUC ≥ 0.01	0.01 > AUC > 0	AUC = 0
Number of chemicals	132	96	34	1593

- The chemicals in each score class in **Table 2** were partitioned into a training set containing 1243 chemicals (67%) and a test set containing 612 chemicals (33%) to build the models and validate their predictive power, respectively.
- Training and test sets were randomly selected from the larger chemical set while maintaining the distribution of positives and negatives.

Molecular Descriptors and Structural Fingerprints

- The molecular structures of the chemicals were represented in SMILES (Simplified Molecular Input Line-Entry System) format.
- SMILES strings were converted into three-dimensional (3D) structures using Molecular Operating Environment (MOE).
- 188 molecular descriptors were obtained from MOE.

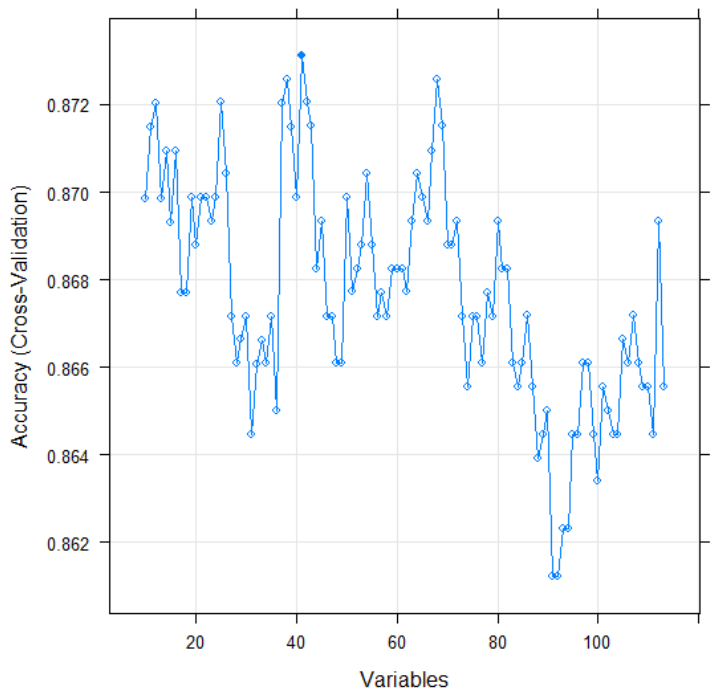
Binary Classification for AR Data

Variable Selection

- An algorithm called recursive feature elimination (RFE) is used for variable/feature selection. RFE selects model variables by repeatedly considering smaller and smaller sets of molecular descriptors and evaluating cross-validated binary prediction performance against the training set.

- **Figure 4** depicts the relationship between the number of variables (molecular descriptors) and the cross-validation accuracy. The optimal subset of molecular descriptors is 41, the number of descriptors yielding the highest accuracy (0.873)

Figure 4 Feature Selection of MOE Descriptors



Classification Modeling and Performance

- We developed a series of binary QSAR models to classify the chemicals as active or inactive according to their anti-androgenic activity. Random forest (RF) and support vector machine (SVM) approaches were used to classify the chemicals based on the most significant molecular descriptors from RFE.
- We investigated the influence of various methods for handling the imbalanced data:
    - Cost-sensitive algorithm (considers misclassification costs based on the data distribution)
    - Oversampling (repeat sampling of positives)
    - Undersampling (sub-sampling of negatives)
- An ideal classification model should maximize accuracy on both sides of the active and inactive classes. G-mean, the geometric mean of sensitivity and specificity, was used to emphasize the joint performance of sensitivity and specificity.

- **Table 3** lists the classification results according to the four performance metrics, i.e., sensitivity, specificity, accuracy, and G-mean.

Table 3 Prediction Performance from SVM and RF Models

SVM Model

Model/ Imbalance Handling	Data Set	TP	FN	SE (%)	TN	FP	SP (%)	Accuracy (%)	G-mean (%)
Cost-sensitive	Training	139	12	92.1	834	220	79.1	80.7	85.4
	Test	60	15	80.0	390	140	73.6	74.4	76.7
Under-sampling	Training	142	9	94.0	142	9	94.0	94.0	94.0
	Test	61	14	81.3	378	152	71.3	72.6	76.1
Over-sampling	Training	995	59	94.4	834	220	79.1	80.0	86.4
	Test	60	15	80.0	388	142	73.2	74.1	76.5

RF Model

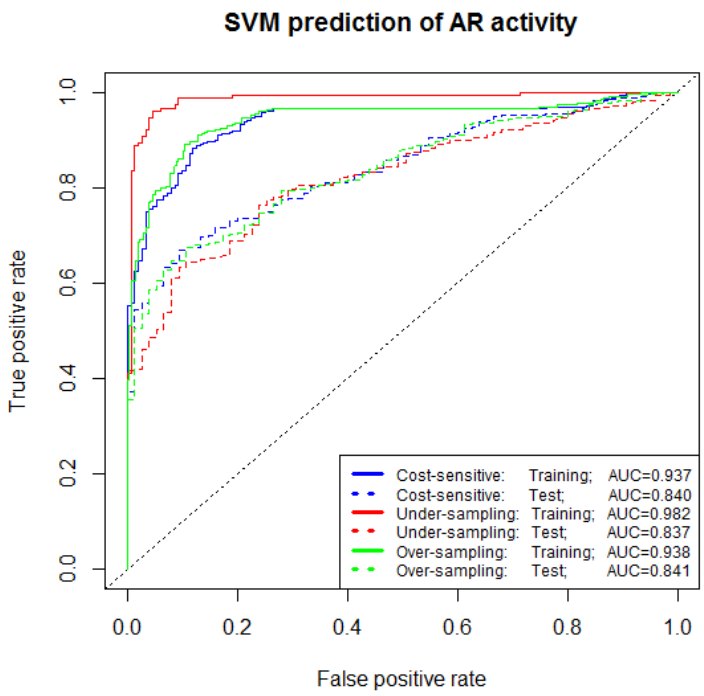
Model/ Imbalance Handling	Data Set	TP	FN	SE (%)	TN	FP	SP (%)	Accuracy (%)	G-mean (%)
Cost-sensitive	Training	149	2	98.7	1054	0	100	99.8	99.3
	Test	10	65	13.3	523	7	98.7	88.1	36.2
Under-sampling	Training	151	0	100	151	0	100	100	100
	Test	62	13	82.7	362	168	68.3	70.1	75.2
Over-sampling	Training	1054	0	100	1052	2	99.8	99.9	99.9
	Test	20	55	26.7	504	26	95.1	86.6	50.4

Abbreviations: FN = false negative (active chemicals); FP = false positive (inactive chemicals); RF = random forest; SE = sensitivity (active chemicals); SP = specificity (inactive chemicals); SVM = support vector machines; TN = true negative (inactive chemicals); TP = true positive (active chemicals).

- A receiver operating characteristic (ROC) curve can help visualize the performance of a classifier by providing a graphical representation of the trade-off between the true positive rate (y-axis) and the false positive rate (x-axis). **Figure 5** shows the ROC curves for SVM classifier models based on cost-sensitive modeling, oversampling, and undersampling.

- The best binary classification model was used to make predictions for ~30,000 chemicals in the broader chemical universe eligible for assessment in the EPA Endocrine Disruptor Screening Program. Based on this preliminary model, 20.60% (6475/31428) of these chemicals were predicted to have some AR antagonist activity. However, this initial result is very likely an overestimate that may be confounded by very weak activity or cytotoxicity. Further refinement of the model should improve specificity.

Figure 5 ROC Curves for SVM Classifier Models of AR Activity



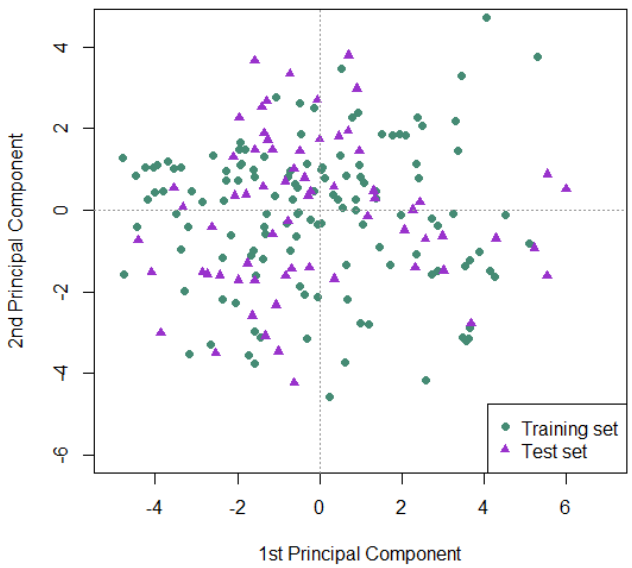
Abbreviations: AR = androgen receptor (antagonism); AUC = area under the curve; ROC = receiver operating characteristic; SVM = support vector machines.

Regression Analysis for the Prediction of AR Potency

Training and Test Sets

- There are 225 active chemicals in the ToxCast data set with MOE descriptors available.
- To develop a model for prediction of AR antagonist potency, we split these into training and test sets with 150 and 75 chemicals, respectively.
- A principal component analysis (PCA) based on structural descriptors indicated that the chemicals in the test set are representative of the chemicals in the training set, and thus the predictions for the test set using the developed models are reliable (**Figure 6**).

Figure 6 Principal Component Analysis of the AR Potency Model Training and Test Sets



Variable Selection

- To optimize potency model construction and avoid overfitting, we performed variable selection using genetic algorithm (GA), a popular algorithm well-suited for regression models.
- GA is an efficient stochastic optimization tool and randomized search technique. GA analysis is conducted in five steps: (1) encoding the descriptors as chromosomes, (2) generating the initial population of chromosomes, (3) evaluating the fitness function, (4) creating the next generation of chromosomes, (5) terminating the process.
- The initial population was assigned to 200 chromosomes. The crossover probability and mutation rate were set to 50% and 1%, respectively. The coefficient of determination was taken as the fitness score of each chromosome. A series of variables from 10 to 60 were selected from GA.

Regression Models

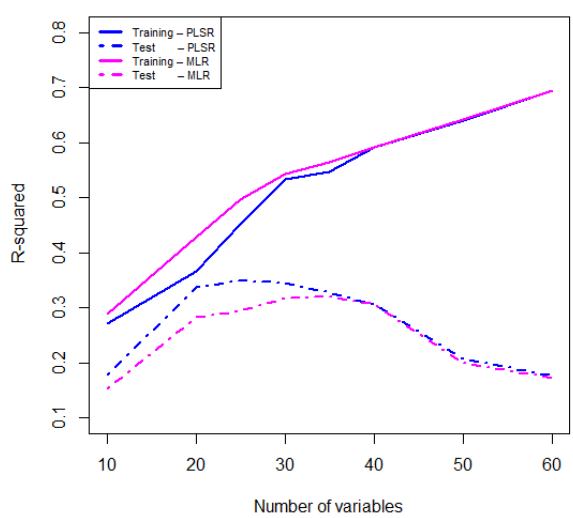
- We used two approaches to develop models of antagonist potency, multiple linear regression (MLR) and partial least squares regression (PLSR).
- MLR produced a linear model describing the relationship between the variables (molecular descriptors) and antagonist potency. The contribution coefficients for each descriptor are determined by regression analysis.

- PLSR employs a set of orthogonal latent variables or principal components (PCs) that are generated through a linear combination of the original molecular descriptors. The PCs are new variables used to build regression models to predict the response variable, i.e., the potency. The optimal number of PCs was determined by 10-fold cross-validation.

Regression Model Validation

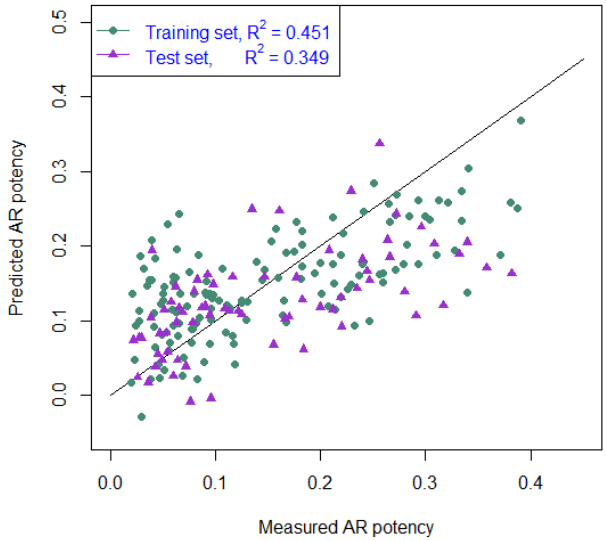
- We evaluated the performance of the regression models by examining the correlation between the experimental and calculated values for both training and test sets, using R^2 (coefficient of determination) and RMSE (root mean squared error).
- R^2 for the training set improved continuously with increasing feature number.
- R^2 for the test set initially increased, attained a maximum at a medium number of variables, and then gradually decreased (**Figure 7**).
    - The best MLR model had an R^2 of 0.320 and RMSE of 0.0619 for the test set (n= 35 variables)
    - The best PLSR model had an R^2 of 0.349 and an RMSE of 0.0541 (n=25 variables).
- **Figure 8** is a scatter plot of predicted versus actual AR antagonist potency (based on AR pathway model AUC) values from PLSR modeling.

Figure 7 Effect of Model Complexity on Coefficient of Determination



Abbreviations: MLR = multiple linear regression; PLSR = partial least squares regression; R-squared = coefficient of determination.

Figure 8 Comparison of Predicted and Actual AR Antagonist Potency by PLSR



Abbreviations: AR = androgen receptor (antagonism); PLSR = partial least squares regression; R^2 = coefficient of determination.

Conclusions

- The AR pathway model provides a biologically based mathematical approach to distinguish assay interference from true agonist or antagonist activity and to prioritize large numbers of environmental chemicals for their potential androgenic or anti-androgenic activity.
- Because the AR pathway model correctly classified AR reference chemicals and identified environmental chemicals with known anti-androgenic activity, it was used to train QSAR models for hazard classification and potency prediction.
- The best performing classification model was obtained using SVM in combination with a subset of descriptors identified via the RFE algorithm (sensitivity = 80.0%; specificity = 73.6% for test set).
- The classification model was used to predict AR antagonist activity for the Endocrine Disruptor Screening Program universe of 30,000 chemicals, predicting that 20.6% (6475/31428) of these chemicals may have AR antagonist activity. This initial result that is certainly an overestimate confounded by very weak activity or cytotoxicity. Further refinement of the model should improve specificity.
- For AR potency prediction, the PLSR regression model had reasonable accuracy with an R^2 of 0.451 for the training set and an R^2 of 0.349 for the test set. Further work will be conducted to expand molecular descriptor libraries for use in the regression model and to use this model to predict activities of chemicals in a broader universe.

References

Dix DJ et al. 2007. Toxicol Sci 95(1):5-12.

EPA. 2011. The Incorporation of *In Silico* Models and *In Vitro* High Throughput Assays in the Endocrine Disruptor Screening Program (EDSP) for Prioritization and Screening. Summary Overview. A Part of the EDSP Comprehensive Management Plan [Internet]. Washington, DC: U.S. EPA.

EPA, NICEATM. 2014. Integrated Bioactivity and Exposure Ranking: A Computational Approach for the Prioritization and Screening of Chemicals in the Endocrine Disruptor Screening Program. Presentation to the December 2014 meeting of the Scientific Advisory Panel for the EPA Federal Insecticide, Fungicide and Rodenticide Act [Internet]. Washington, DC: U.S. EPA.

Food Quality Protection Act of 1996. 7 U.S.C. 136. Public Law 104-170.

Judson RS et al. 2010. Environ Health Perspect 118(4): 485-492.

Kleinstreuer et al. 2015. Toxicol Sci/Toxicologist 144(1):285.

Safe Drinking Water Act Amendments of 1996. 110 Stat 1613. Public Law 104-182.

Tice RR et al. 2013. Environ Health Perspect 121:756–765.

Acknowledgements

The Intramural Research Program of the National Institute of Environmental Health Sciences (NIEHS) supported this poster. Technical support was provided by ILS under NIEHS contract HHSN273201500010C.
The views expressed above do not necessarily represent the official positions of any federal agency. Since the poster was written as part of the official duties of the authors, it can be freely copied.