International Journal of
Environmental Research and Public Health

Article

A Rat α-Fetoprotein Binding Activity Prediction Model to Facilitate Assessment of the Endocrine Disruption Potential of Environmental Chemicals

Huixiao Hong [1,*,†], Jie Shen [1,†], Hui Wen Ng [1], Sugunadevi Sakkiah [1], Hao Ye [1], Weigong Ge [1], Ping Gong [2], Wenming Xiao [1] and Weida Tong [1]

[1] Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR 72079, USA; jieeshen@gmail.com (J.S.); Huiwen.Ng@fda.hhs.gov (H.W.N.); Suguna.Sakkiah@fda.hhs.gov (S.S.); haoye.ecust@gmail.com (H.Y.); weigong.ge@fda.hhs.gov (W.G.); Wenming.Xiao@fda.hhs.gov (W.X.); Weida.Tong@fda.hhs.gov (W.T.)

[2] Environmental Laboratory, U.S. Army Engineer Research and Development Center, 3909 Halls Ferry Road, Vicksburg, MS 39180, USA; Ping.Gong@usace.army.mil

* Correspondence: Huixiao.Hong@fda.hhs.gov; Tel.: +1-870-543-7296; Fax: +1-870-543-7854

† These authors contributed equally to this work.

Academic Editor: Paul B. Tchounwou
Received: 27 January 2016; Accepted: 22 March 2016; Published: 25 March 2016

Abstract: Endocrine disruptors such as polychlorinated biphenyls (PCBs), diethylstilbestrol (DES) and dichlorodiphenyltrichloroethane (DDT) are agents that interfere with the endocrine system and cause adverse health effects. Huge public health concern about endocrine disruptors has arisen. One of the mechanisms of endocrine disruption is through binding of endocrine disruptors with the hormone receptors in the target cells. Entrance of endocrine disruptors into target cells is the precondition of endocrine disruption. The binding capability of a chemical with proteins in the blood affects its entrance into the target cells and, thus, is very informative for the assessment of potential endocrine disruption of chemicals. α-fetoprotein is one of the major serum proteins that binds to a variety of chemicals such as estrogens. To better facilitate assessment of endocrine disruption of environmental chemicals, we developed a model for α-fetoprotein binding activity prediction using the novel pattern recognition method (Decision Forest) and the molecular descriptors calculated from two-dimensional structures by Mold^2 software. The predictive capability of the model has been evaluated through internal validation using 125 training chemicals (average balanced accuracy of 69%) and external validations using 22 chemicals (balanced accuracy of 71%). Prediction confidence analysis revealed the model performed much better at high prediction confidence. Our results indicate that the model is useful (when predictions are in high confidence) in endocrine disruption risk assessment of environmental chemicals though improvement by increasing number of training chemicals is needed.

Keywords: model; prediction; α-fetoprotein; endocrine; disruption; binding; assessment

1. Introduction

Endocrine disruptors (EDs) are exogenous compounds that affect the endocrine system of humans and other vertebrates. Endocrine activity of environmental or foreign chemicals has the potential to cause numerous adverse outcomes, including disrupting the physiologic function of endogenous hormones and altering homeostasis. The known EDs include polychlorinated biphenyls (PCBs), the synthetic estrogen diethylstilbestrol (DES), dichlorodiphenyltrichloroethane (DDT) and other pesticides. For example, DES was approved the Food and Drug Administration (FDA) for treatment of menopausal symptoms, gonorrheal vaginitis, atrophic vaginitis, postpartum lactation suppression, and prostate cancer [1,2]. DES was shown to disrupt the endocrine system causing vaginal tumors in girls and women and other adverse medical complications [3] and thus was withdrawn from the market by the FDA. Concern about EDs has invigorated intense discussion and debate over the past two decades in the scientific community [4,5] and promoted the legislation for regulation of environmental chemicals mandated by the Environmental Protection Agency (EPA) and development of the Endocrine Disruptor Screening Program (EDSP) to screen potential EDs in the environment [6].

EDs can disrupt the endocrine system through different mechanisms [7–12]. One of the well-known mechanisms is mediated by the hormone receptors such estrogen receptor (ER) and androgen receptor (AR), in which EDs exhibit their estrogenic and androgenic effects through binding to the ER and AR in the target cells [13–16]. Therefore, a huge amount of estrogenic and androgenic activity data of structurally diverse chemicals have been generated and organized in sophisticated databases such as the FDA's Endocrine Disruptors Knowledge Base (EDKB) [17] and Estrogenic Activity Database (EADB) [18]. These databases have been used for the development of a diverse set of quantitative structure-activity relationship (QSAR) models for predicting estrogenic and androgenic activity and to assist evaluation of endocrine disruption potential of environmental chemicals [19–29].

ED binding to hormone receptors in target cells is the key mechanism to display endocrine disruption. However, the affinity of binding to ER is not the sole criterion to determine EDs' potential to disrupt the endocrine system. For example, EDs cannot bind to ER or AR in the target cells if they cannot pass the cell membrane. Therefore, *in vitro* ER and AR binding data of chemicals may not reflect well their *in vivo* endocrine activity, even for chemicals with high *in vitro* binding affinity. To accurately estimate the endocrine disruption potential of environmental chemicals, it is necessary to have both their binding activities to hormone receptors and to competing serum proteins such as alpha-fetoprotein (AFP) [30,31] and human sex hormone-binding globulin (SHBG) [32].

There are different transporter proteins in serum, including albumin, globulin, fibrinogen, and others. The transporter proteins can transport hormones, vitamins and other chemicals within and between cells and organs. SHBG is one of the major transporter proteins that bind to hormones and other chemicals in human serum [33]. AFP is a major transport protein in rat and was first discovered approximately 60 years ago [34]. It is a serum biomarker of Down's syndrome and neural tube defects in the clinical practice and alters the growth of fetal and cancer cells [35,36]. Entrance of AFP into cells through receptor-mediated endocytosis was observed in fetal cells of different species including rat [37], mouse [38], human [39] chicken [40] and baboon [41]. Elevated AFP level was observed in maternal circulation through transplacental passage from the fetal circulation and amniotic fluid by the placental or allantois [42–45]. This protein competes with ER to bind estrogens in the blood and thus inhibits EDs access to the target cells [46,47]. It has been found that diverse chemicals bind AFP [30,48–52].

A huge amount of *in vitro* binding assays data have been generated for the targets such as ER and AR involved in the endocrine system. However, available *in vivo* bioactivity data related to endocrine disruption potential are relatively less than the *in vitro* data. Moreover, most of the *in vivo* data are obtained using rats uterotrophic assays [17,18,53]. To better assess endocrine disruption potential of environmental chemicals, we measured rat AFP binding affinity for 125 chemicals with diverse structures using a competitive binding assay according to the methods published in our previous study [30]. Our rat AFP binding data represent the largest such data set to date. Compared with the experimental data on the hormone receptors such as ER and AR, there are fewer chemicals with experimental AFP binding data, hindering the risk assessment of environmental chemicals in terms of endocrine disruption potential. Therefore, for an enhanced risk assessment it was necessary to obtain AFP binding data for those environmental chemicals lacking AFP binding data. To this end, we developed an *in silico* model for prediction of AFP binding activity of environmental chemicals using our previously reported data [30]. The performance of the model was internally evaluated through cross validations and permutation tests. It was also validated externally using the AFP binding activity data curated from the literature. We demonstrated that the model has suitable predictive power and is expected to better assist endocrine disruption assessment of environmental chemicals.

2. Materials and Methods

2.1. Study Design

The study design is depicted in Figure 1 and the detail explanation for each step is described in the following sections. Briefly, the 125 chemicals and their rat AFP finding activity (53 binders and 72 non-binders) from our previous study [30] were used as the training data set. First, 5-fold cross validations were conducted to evaluate the performance of Decision Forest (DF) model as illustrated in the bottom left part of Figure 1. More specifically, the training data set were randomly divided into five equal portions of chemicals. Four portions were used for training the DF model and the remaining portion was used for testing the DF model. The process was repeated five times so that each of the five portions was used as test data set to challenge the models that were constructed from the other four portions. The prediction results from the five DF models were averaged to estimate the models' performance. To reach a statistically robust estimation of the DF models' performance, the 5-fold cross validation process was iterated 1000 times. The resultant data from the 1000 iterations of 5-fold cross validation were used for prediction confidence analysis and identification of informative molecular descriptors that are important for AFP binding. Then, permutation tests were conducted to affirm that the prediction accuracy observed in the 5-fold cross validations was not achieved by chance, as illustrated in the top part of Figure 1. In brief, the binding activity data (binder or non-binder) of the 125 chemicals in the training data set were permutated first and a 5-fold cross validation was carried on the resultant permutated data set. The permutation test was repeated 1000 times to make sure that the permutation tests result is statistically robust. Finally, the whole training data set was used to train a DF model that was validated using an external data set. The external validation data set was curated from the literature [48–50].

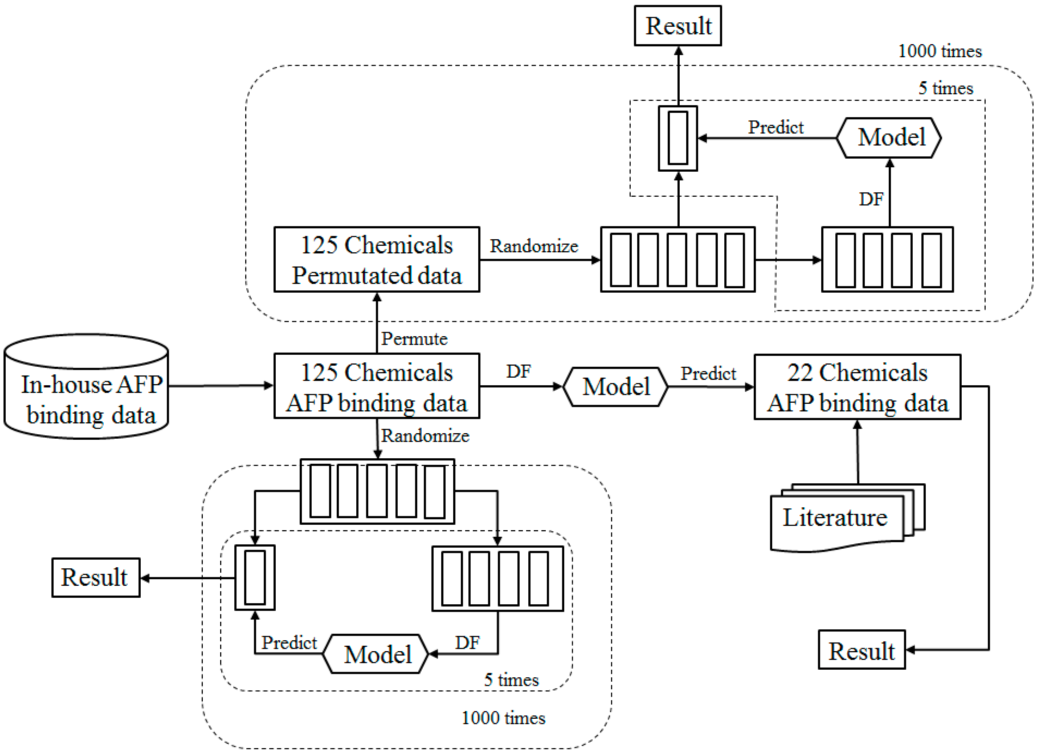

Figure 1. Overview of the study design.

2.2. Data Sets

The 125 structurally diverse chemicals with rat AFP competitive binding assay results published earlier [30] were used as the training data set. Of the 125 chemicals, 53 chemicals displayed binding affinities to rat AFP. The IC_{50} values of the 53 chemicals are in the range of 0.0065 to 590 nM. All of the 53 chemicals were defined as AFP binders in this study. The rest 72 chemicals did not show binding affinity to rat AFP and were determined to be AFP non-binders. In this study, binders were represented by "1" and non-binders by "0" in the model constructions and predictions. The two-dimensional (2D) structures of the 125 chemicals were generated according to our previous study using Marvin Sketch (http://www.chemaxon.com/) and saved in a single 2D SDF (structure-data file) format file [30].

For validation of AFP binding activity prediction model, we curated an external data set through literature search for AFP binding activity. First, the chemicals with AFP binding activity data were collected from the literature. After removing the chemicals that were presented in the training data set, 22 chemicals with known AFP binding activity data from other studies [48–50] were used as the external validation set. The structures of the 22 chemicals were drawn according to the literature using Marvin Sketch and saved in a single 2D SDF format file.

2.3. Molecular Descriptors

QSAR models are developed based on different types of molecular descriptors. The molecular descriptors of the chemicals in both training and external validation data sets were generated using Mold^2 [54,55]. Mold^2 is a free software which calculates molecular descriptors from 2D chemical structures. This software is very fast because it adopts the extremely rapid algorithm for cyclic structure recognition [56] and uses the efficient chemical structure representation system [57,58] that has shown high efficiency in the system for chemicals structure elucidation based on infrared [59] and nuclear magnetic resonance (NMR) spectra [60–62]. Mold^2 has been demonstrated to be reliable for developing QSAR models [63,64]. In brief, 777 Mold^2 descriptors were first calculated for each of the chemicals in the training and external validation data sets. Then, the descriptors were cleaned up by removing those with constant values across all the chemicals in the data sets. Finally, the remaining 512 Mold^2 descriptors were scaled to the values between 0 and 1.

2.4. Prediction Model

Prediction models can be developed using different QSAR methods such as pharmacophore modeling [65–68], molecular docking [69,70] and machine learning methods [71–73]. In this study, the prediction models were built using the Mold^2 descriptors and the pattern recognition algorithm DF that was developed previously by our group [74,75]. DF is a free software for public use [76] that employs a consensus modeling technique by combining multiple decision tree models. It uses a unique procedure to construct different decision tree models to ensure heterogeneous models when combined. Besides, variable selection process is wrapped in the model construction process, which simplifies the model development. In addition to QSAR, the DF algorithm were applied for the development of predictive models based on the genomics data [77,78] and proteomics data [79]. The DF models in this study were constructed using the following algorithmic parameters: the number of trees is set to 5; the minimum size of node to be split is 10; the maximum levels to be pruned to is 3; and the method for node splitting is Gini's diversity index. The tree building and pruning processes were guided by achieving the minimum number of misclassified compounds.

2.5. Cross Validations

To assess the performance of the DF model, 5-fold cross validations were conducted as illustrated in Figure 1. In one 5-fold cross validation, the 125 chemicals of the training data set were randomly divided into five equal portions. Four of the five portions were used to construct a DF model, which was then used to predict AFP binding activity for the chemicals in the remaining one portion. This process was repeated sequentially so that each of the five portions was left out once and only once as the testing set. The prediction results from the five testing sets were then averaged as an estimate of the DF model performance using accuracy, sensitivity, specificity, Matthews correlation coefficient (MCC) and balanced accuracy. These performance metrics were calculated using Equations (1)–(5) through comparison of the predictions with the actual AFP binding activity data:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4}$$

$$Balanced \cdot Accuracy = \frac{TP(TN + FP) + TN(TP + FN)}{2(TP + FN)(TN + FP)} \tag{5}$$

In Equations (1)–(5), true positive (*TP*) is the number of AFP binders that were predicted as binders by the DF models, true negative (*TN*) is the number of AFP non-binders that were predicted as non-binders, false negative (*FN*) is the number of AFP binders that were predicted as non-binders, and false positive (*FP*) is the number of AFP non-binders that were predicted as binders.

2.6. Permutation Tests

Permutation analysis is a common approach to determine whether the model performance estimated from cross validations is due to chance correlations. As shown in Figure 1, in one permutation test, the qualitative AFP binding activity data (1 for binder and 0 for non-binder) of the 125 chemicals in the training data set were randomly shuffled while the Mold^2 descriptors values (the independent variables) remained unchanged to generate a permutated data set. A 5-fold cross validation described above was then conducted on the permutated data set and the cross validation results were compared with the results from the cross validations on the real training data set. The permutation test was repeated 1000 times by using different randomly shuffled AFP binding activity data to reach a statistically significant robust comparison with the 1000 times of 5-fold cross validations using the real training data set.

2.7. Prediction Confidence Analysis

In the cross validations, the AFP binding activity prediction from a DF model for a chemical is a continuous value, p, that is used to forecast the qualitative AFP binding activity of the chemical as AFP binder ($p \geqslant 0.5$) or non-binder ($p < 0.5$). This value indicates the likelihood of the chemical to be a AFP binder or AFP non-binder and represents the confidence for the prediction. A good prediction model is expected not only to show accuracy but also to predict most unknown chemicals with high confidence level. Furthermore, the predictions with a higher prediction confidence level should be more accurate than the predictions at a lower prediction confidence level. We analyzed the relationship between the prediction confidence and the corresponding prediction accuracy of the DF models in the 1000 iterations of 5-fold cross validations using the training data set. The prediction confidence was calculated for each of the predictions from the 1000 times of 5-fold cross validations using Equation (6).

$$Confidence = \frac{|p - 0.5|}{0.5} \tag{6}$$

The calculated prediction confidence is a value between 0 and 1. The larger the value, the more reliable is the prediction. The predictions of the 5-fold cross validations were placed into 20 groups with even confidence bins. For each of the 20 groups of predictions, prediction performance metrics such as sensitivity, specificity and accuracy were calculated by comparing the predictions with the actual AFP binding activity data. At last, the performance of the DF models at difference confidence levels was analyzed.

2.8. Informative Molecular Descriptors Identification

Generally, QSAR models are built for one or both of the purpose of prediction and/or to gain mechanistic understanding of biochemical phenomena [80]. Mechanistic understanding is derived from the ability to interpret the physicochemical meaning of molecular descriptors used in QSAR models. To better understand the chemical aspects that play important roles in the binding interactions with AFP, the molecular descriptors used in the DF models were examined to identify the Mold^2 descriptors that are informative to the DF models. First, the frequency values of the Mold^2 descriptors that were used in the DF models in the 1000 permutation tests were calculated to establish a statistical background. The DF models were constructed from the random data sets obtained by permutation and, thus, the top 5% frequency can be used as the frequency criterion to identify the informative descriptors with a 5% probability for the descriptors being selected due to the random noises, that is at a p-value = 0.05. Then, the frequency values of the Mold^2 descriptors that were used in the DF models in the 5-fold cross validations were computed and compared with the frequency of 0.05 (p-value) that was determined from the permutation tests. The Mold^2 descriptors that had higher frequency values than the frequency of 0.05 (p-value) were identified as the informative descriptors for AFP binding activity prediction.

2.9. External Validation

QSAR models usually perform better on the dataset that was used to construct the models in cross validations than on new data. Validation using external data sets is important and necessary to assess the performance of a predictive model. In this study, 22 chemicals with known AFP binding activity data from the literature were assembled for external validation. The predictive DF model was built on the entire training data set of 125 chemicals and then used to predict the AFP binding activity of these 22 chemicals in the external validation set.

3. Results

3.1. Cross Validations

We conducted 1000 5-fold cross validation cycles using the training data set as shown in Figure 1. The prediction results from the DF models were compared with the actual AFP binding activity data to calculate the metrics for evaluation of the performance of the models. The 5-fold cross validation results were plotted in the boxplots of Figure 2 and are summarized in Table 1. The average values of accuracy, sensitivity, specificity, MCC and balance accuracy are 68.9%, 67.5%, 70.0%, 57.0% and 68.8% respectively. All performance metrics indicate a moderate prediction power of the DF models. The small standard deviation values obtained demonstrated that the DF models are statistically robust.

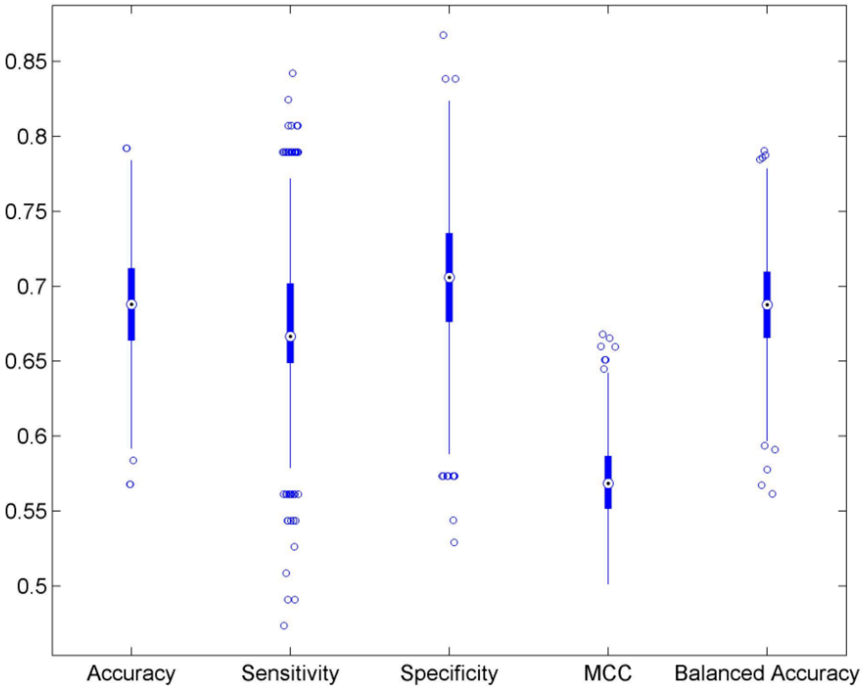

Figure 2. Boxplots for the predictions from the DF models in the 5-fold cross validations. Performance were measured by metrics as indicated on the *x*-axis.

Table 1. Summary of cross validations, permutation tests, and external validation.

Parameter	Cross Validations		Permutation Tests		External Validation
	Mean	STD	Mean	STD	
Accuracy	0.689	±0.034	0.498	±0.049	0.546
Sensitivity	0.675	±0.054	0.427	±0.067	0.412
Specificity	0.700	±0.046	0.558	±0.061	1.000
MCC	0.570	±0.026	0.497	±0.009	0.371
Balanced accuracy	0.688	±0.034	0.492	±0.050	0.706

STD: standard deviation.

3.2. Permutation Tests

Permutation tests were conducted to affirm that the prediction power observed for the DF models in the 5-fold cross validations was not due to chance correlation in the training data set. The prediction results from the DF models that were constructed using the 1000 permutated datasets and were plotted for the distribution of prediction accuracy values as the red line in Figure 3. For comparison, the distribution of the prediction accuracy values from the 1000 times of 5-fold cross validations is represented as the blue line in Figure 3. Obviously, the predictions from the cross validations were significantly more accurate than the predictions from the permutation tests, with a p-value < 0.0001. The same difference were observed for other metrics: the differences between the average values of the cross validations and the permutation tests were 19.0%, 24.7%, 14.2%, 7.3% and 19.5% in overall accuracy, sensitivity, specificity, MCC and balanced accuracy respectively. Therefore, the permutation tests demonstrated that the AFP binding activity predictions of the DF models in the cross validations were not obtained by probability success.

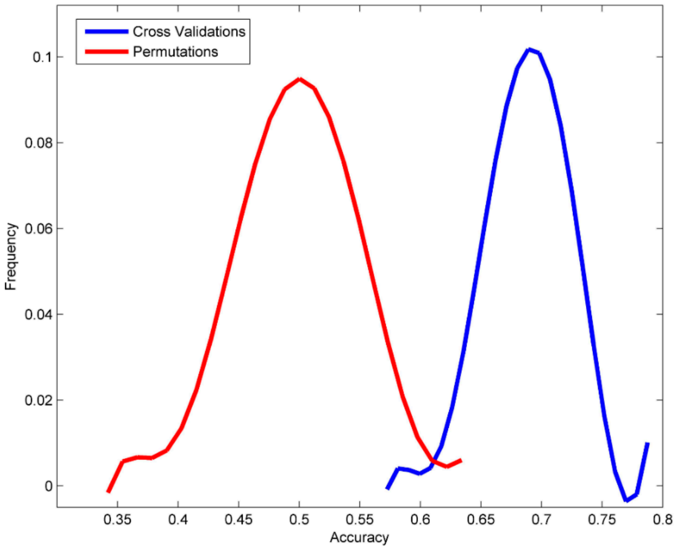

Figure 3. Distributions of the 1000 prediction accuracy values calculated from the DF models in permuation tests (red line) and yielded from the DF models in the cross validations (blue line).

3.3. Prediction Confidence Analysis

We analyzed the prediction confidence using the 1000 times of 5-fold cross validations. The confidence levels of the predictions from the DF models in the 1000 iterations of 5-fold cross validations were calculated and used to place the predictions into 20 groups with even confidence bins. Correct and incorrect predictions were then counted for each of the 20 groups by comparison with the actual AFP binding activity data.

Prediction accuracy was calculated for the predictions in each of the 20 groups. The numbers of predictions, correct predictions and incorrect predictions for the 20 groups were shown as blue, red and green distribution curves respectively in Figure 4.

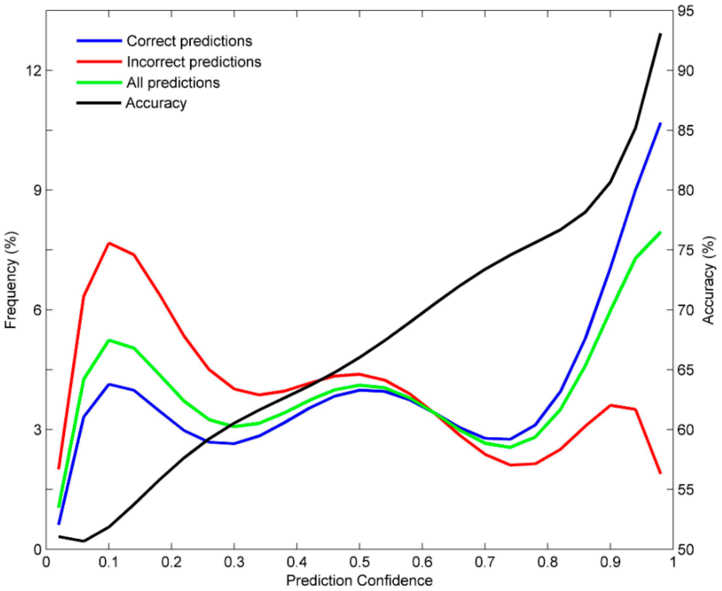

Figure 4. Predictions and accuracy at different confidence levels. The distributions of predictions were given by the left *y*-axis and the prediction accuracy is indicated by the right *y*-axis. Prediction confidence was given at the *x*-axis. Predictions are plotted in green line, correct predictions in blue line, incorrect predictions in red line, and prediction accuracy in black line.

The corresponding prediction accuracy values for the 20 groups were plotted as a black distribution line in Figure 4. As the confidence level increased, the correct predictions increased (blue line) while the incorrect predictions reduced (red line). More importantly, it was found that higher the prediction confidence, the more accurate are the predictions (black line). Moreover, most predictions from the DF models were at high confidence (green line). The prediction confidence analysis demonstrated that the DF models not only had a reasonable prediction power but also gave prediction confidence values that could be utilized to better assist evaluation AFP binding activity of chemicals.

3.4. Identification of Informative Descriptors

The more frequently a descriptor is used in QSAR models, the more informative it is to the QSAR models. The informative molecular descriptors are important for interpretation of QSAR models. To identify the informative descriptors to the DF models in the 5-fold cross validations, we first extracted the Mold^2 descriptors that were actually used in the models. Then, the frequency of each of the 512 Mold^2 descriptors used by the 5000 DF models was calculated. The results were plotted as the solid blue line in Figure 5. Similarly, the frequency of each Mold^2 descriptor used in the 5000 DF models in the permutation tests was calculated. The results were displayed as the solid red line in Figure 5. The top 5% descriptors in the permutation tests were separated by the dotted black line at a frequency of 1680 models in Figure 5. Therefore, the Mold^2 descriptors that were used in more than 1680 DF models in the 5-fold cross validations should be informative to the DF models at the 5% significance level in a statistical view. Using this cut-off, 16 Mold^2 descriptors that were used by more than 1680 DF models were identified as the informative descriptors. Table 2 lists these 16 Mold^2 descriptors, the numbers of DF models, and the descriptor definitions.

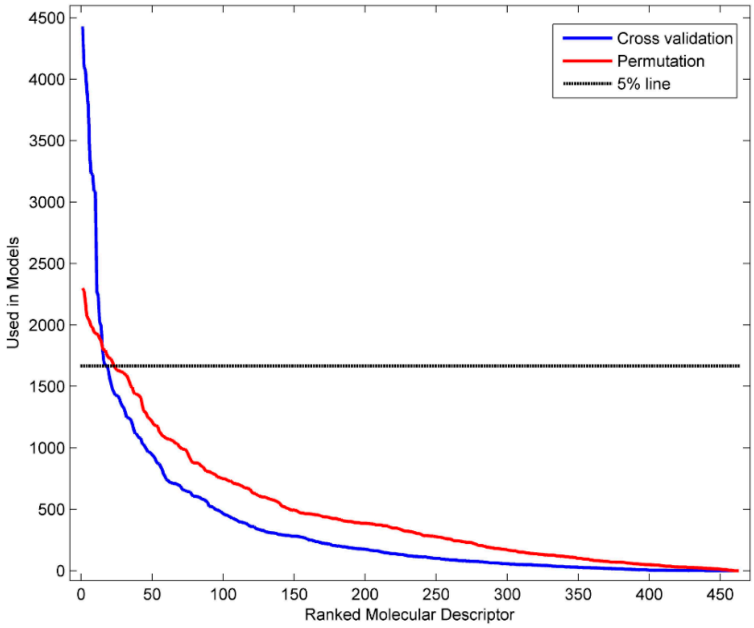

Figure 5. The distribution of descriptors used in the DF models.

The identified informative descriptors are the indices that are related to molecular shape, electronegativity and polarizability of the chemicals. Therefore, the molecular shape of a chemical and its hydrophilic interactions with the ligand binding pocket of AFP are the key structural features that determines if a chemical can bind to AFP. This finding is consistent with our previous structural analysis of AFP ligand binding pocket [32].

Table 2. Informative descriptors identified from the cross validations.

ID	Models	Descriptor Definition
D282	4429	complementary information content (neighborhood symmetry of 2-order)
D281	4099	structural information content (neighborhood symmetry of 2-order)
D450	4075	Geary autocorrelation-lag 4/weighted by atomic masses
D432	3916	Broto-Moreau autocorrelation of a topological structure-lag 2/weighted by atomic Sanderson electronegativity
D458	3770	Geary autocorrelation-lag 4/weighted by atomic van der Waals volumes
D361	3391	ratio of multiple path counts to path counts
D213	3233	valence connectivity index chi-1
D467	3225	Geary autocorrelation-lag 5/weighted by atomic Sanderson electronegativity
D491	3091	Moran autocorrelation-lag 5/weighted by atomic van der Waals volumes
D259	3084	mean information content on the distance degree equality
D496	2272	Moran autocorrelation-lag 2/weighted by atomic Sanderson electronegativity
D478	2238	Geary autocorrelation-lag 8/weighted by atomic polarizabilities
D463	2024	Geary autocorrelation-lag 1/weighted by atomic Sanderson electronegativity
D246	1995	Maximum of the differences between vertex distance and unipolarity
D473	1799	Geary autocorrelation-lag 3/weighted by atomic polarizabilities
D595	1698	highest eigenvalue n. 8 of Burden matrix/weighted by atomic polarizabilities

3.5. Prediction Model and External Validation

The AFP binding activity prediction DF model was constructed using the 125 chemicals of the training data set. The DF model consisted of five decision trees that are illustrated in Figure 6. The trees had eight to ten terminal nodes. The DF model was used to predict AFP binding activity for the 22 chemicals from the external data set. The 22 chemicals, including their names used in the literature, experimental AFP binding data, DF model prediction results and the references are given in Table 3.

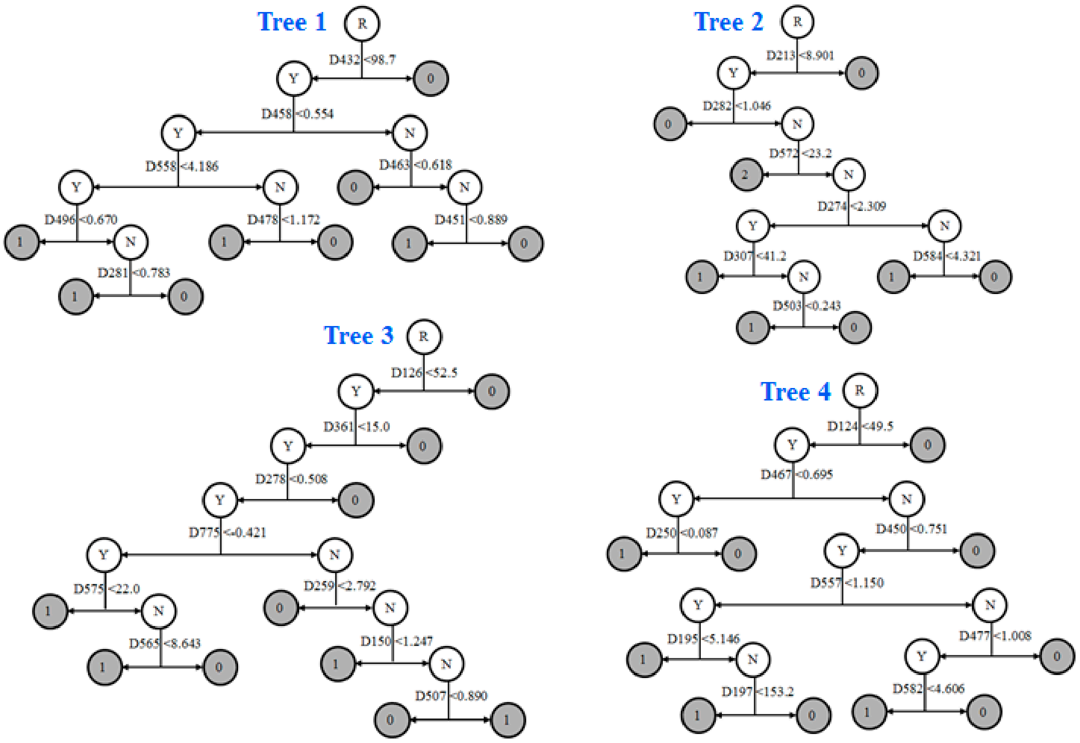

Figure 6. *Cont.*

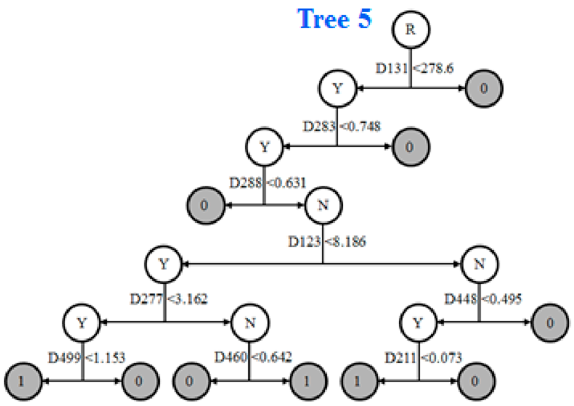

Figure 6. Decision trees of the AFP binding activity prediction DF model. The descriptors and their criteria that were used to split the intermediate nodes are given under the nodes. The left nodes are the sets of chemicals that meet the criteria for splitting their parent nodes; the right nodes represent the sets of chemicals that do not meet the criteria. The root node (whole training data set) and the intermediate nodes are presented in empty/white circles. Letter Y in a circle indicates the chemicals in the node meet the splitting criterion, whereas the letter N means the chemicals do not meet the splitting criterion. The terminal nodes are the leaves of the trees where the AFP binding activity predictions were determined and are shown in grey circles. Number 1 in a circle indicates that the chemicals in the node are predicted as AFP binders while number 0 marks the node where chemicals are predicted as AFP non-binders.

Table 3. The experimental and predicted AFP binding activity of the external data set.

Chemical Name	Experiment	Prediction	Reference
17-α-Ethynylestradiol	1	1	[49]
11-β-Ethyloxyestradiol	1	0	[48]
11-β-Methoxyestradiol	1	1	[48]
Compound **7b**	1	0	[49]
16-α-Fluoroestradiol (FES)	1	1	[48]
Compound **8b**	1	0	[49]
Compound **8c**	1	1	[49]
Compound **3**	1	1	[48]
Compound **1**	1	0	[48]
Compound **2**	1	0	[48]
Compound **7c**	1	1	[49]
11-β-Ethyl-17-α-ethynylestradiol	1	0	[49]
11-β-Ethylestradiol	1	0	[49]
Compound **8a**	1	0	[49]
17-α-Ethynyl-11-β-Methoxyestradiol	1	0	[49]
Compound **7a**	1	0	[49]
4-Nonylphenoxyacetic acid (NP1EC)	1	1	[50]
4-*tert*-Butylphenol (BP)	0	0	[50]
Igepal	0	0	[50]
2,4'DDT	0	0	[50]
2,4'-DDE	0	0	[50]
Kepone	0	0	[50]

AFP binding data: 1 represents binder and 0 indicates non-binder.

The predictive performance of the DF model on the external validation set was measured using five different metrics: overall prediction accuracy, sensitivity, specificity, MCC and balanced accuracy. The calculated performance metrics for the external validation are listed Table 1. Slightly lower performance was observed for the external validation compared to the performance of the 5-fold cross validations.

4. Discussion

AFP is a protein in the plasma that binds to estrogens with high affinity. It can sequester EDs in the plasma and thereby reduces the concentration of EDs that can enter into the target cells. Thus, AFP can protect EDs in maternal circulation. Hence, AFP binding activity of chemicals is important information for assessment of endocrine disruption potential. If a chemical does not bind to AFP but binds to hormone receptors such as AR and ER, it can bypass AFP protection and has the potential to disrupt the endocrine system. In contrast, if a chemical binds to AFP, AFP could protect against endocrine disruption even if it has the potential to bind AR or ER. However, a very limited number of chemicals have been experimentally assayed for their AFP binding activity. Thus, we previously measured AFP binding activity for 125 structurally diverse chemicals using the competitive assay developed from rat amniotic fluid [30]. The number of chemicals with AFP binding activity data is still much smaller than the chemicals having ER and AR binding activity, hampering comprehensive assessment of endocrine disruption potential for environmental chemicals. Therefore, in this study, we developed and extensively validated AFP binding activity prediction models using the data published in the literature including our in-house data set. Our model showed a reasonable predictive power and robustness and could be expected to help assess endocrine disruption potential of environmental chemicals.

The DF prediction model was constructed using rat AFP binding data. It could be used for prediction of rat AFP binding activity for the environmental chemicals that have no experimental data. However, the limitation of current model should be noticed when applying the model in applications of human risk assessment of environmental chemicals because the human AFP is not completely homologous to the rat AFP.

Prediction confidence analysis showed that the DF models predicted AFP binding activity very accurately for some chemicals but not so well for other chemicals. The higher the prediction confidence, more likely the prediction is accurate as demonstrated in Figure 4. Therefore, we suggest that the AFP binding activity prediction (binder or non-binder) should be combined with the prediction confidence to better apply the DF model in assessment of endocrine disruption potential of environmental chemicals.

Though AFP was identified long time ago and has been extensively studied, no three-dimensional structure (3D) of AFP or complexes of AFP bound to ligands has been determined by X-ray crystallization. The structural features of this protein, especially in its ligand binding domain, were understood based only on the experimental binding activity data. Therefore, a homology model of rat AFP was constructed and the ligand binding interactions of this protein were elucidated using molecular docking and molecular dynamics simulations in our previous study [31]. The computationally constructed 3D structure of rat AFP and the *in silico* elucidated ligand binding interactions are expected to help the estimated AFP binding activity of environmental chemicals. Our previous study identified two different binding pockets in rat AFP, consistent with the two putative estrogen binding sites in AFP [81]. The ligand binding interactions of rat AFP contribute from residues Glu206, Glu209, Gly210, Leu213, Lys236, His260, Try306 and His310 in the first binding site and from residues Leu233, Gln239 and Glu312 in the second binding site [31]. Most of these amino acids have charged or have polar residues. Thus, hydrophilic and electrostatic interactions are important for a chemical to bind to AFP. Furthermore, the binding pockets were found to be different in size and shape. In this study, 16 Mold^2 descriptors (Table 2) were identified as the informative descriptors to the DF prediction models. Therefore, these molecular descriptors represent the important structural features that are determinant to AFP binding activity of chemicals. The 16 Mold^2 descriptors are the structural features of the chemicals interacting with AFP related to molecular shape, electronegativity, and polarizability of chemicals indicating molecular shape, hydrophilic and electrostatic interaction capability. These molecular characteristics are used to differentiate AFP binders from non-binders. The informative descriptors identified in this study confirmed the reliability of our previously constructed 3D structure of rat AFP and the elucidated ligand binding interactions.

Recently EPA considered utilization of high throughput screening assays and computational models in the endocrine disruptor screening program [82]. EPA led CERAPP project to develop QSAR models for prediction of estrogenic activity and the models were used for prioritize environmental chemicals for Tier-2 testing [83]. With binding data of transporter proteins obtained from experiments or *in silico* predictions, it is speculated that better priority setting the environmental chemicals for testing would be yielded.

The DF prediction models showed lower prediction accuracy than the DF model we previously developed for prediction of ER binding activity [20]. The less predictive power of the AFP binding activity prediction models may be partially due to the relatively small sample size. We expected more accurate DF prediction models would be constructed when AFP binding activity is experimentally measured for more chemicals that can be used as training samples. Another speculation on the cause of the relatively low prediction accuracy is the multiple binding sites in AFP. The 125 chemicals bind AFP in different interaction regions. The first ligand binding site in rat AFP lies in the region of amino acids 419–433 and the second ligand binding site consists of amino acids 450–464. The chemicals that displayed rat AFP binding activity in our previous study are structurally diverse [30]. The existence of two distinct ligand binding sites in AFP indicates that prediction of binding activity of a chemical depends on the AFP site where the chemical binds [84,85]. Therefore, we assume separate prediction models should be developed, each for one of the two ligand binding sites, to improve the performance of AFP binding activity prediction model. Our previous study demonstrated competitive modeling based on molecular docking may perform better than the DF modeling for AFP binding prediction. Lack of knowledge on the binding sites for chemicals and the limited number of experimental binding data available is a major impediment in the development of such separate prediction models. Our results indicated that simple predictive models such as the DF models in this study sometimes yield inaccurate predictions, especially when the system in modeling is not simple. Even though a moderate prediction power has been shown for the AFP binding activity prediction DF model, caution is warranted in application of the DF model in assessment of endocrine disruption potential of environment chemicals, especially when a prediction has a low prediction confidence. Nonetheless, the rat AFP binding activity predictions of high confidence from the DF models should be useful for assistance in estimation of rat AFP binding activity of environmental chemicals.

5. Conclusions

Using a set of structurally diverse chemicals whose rat AFP binding activity data were measured in our previous study, a DF model for prediction of the AFP binding activity was developed in this study. Internal cross validations and external validations were conducted to demonstrate the accuracy and robustness of the models. Our results showed a moderate prediction performance of the models. More importantly, the DF model provides prediction confidence that is very useful when applying the model in assessment of endocrine disruption potential of environment chemicals.

Acknowledgments: This research was supported in part by an appointment to the Research Participation Program at the National Center for Toxicological Research (Jie Shen, Hui Wen Ng, Sugunadevi Sakkiah and Hao Ye) administered by the Oak Ridge Institute for Science and Education through an inter-agency agreement between the U.S. Department of Energy and the U.S. Food and Drug Administration.

Author Contributions: Huixiao Hong conceived and designed the study; Jie Shen curated the experimental data; Huixiao Hong, Jie Shen, Sugunadevi Sakkiah, Hui Wen Ng and Hao Ye conducted the calculations and developed the models. Weigong Ge, Ping Gong, Wenming Xiao and Weida Tong contributed to the data analysis. Huixiao Hong and Jie Shen wrote the first draft of the manuscript. All authors contributed to writing the manuscript.

Conflicts of Interest: The authors declare no conflict of interest. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Food and Drug Administration as well as the U.S.Army Corps of Engineers.

Abbreviations

The following abbreviations are used in this manuscript:

2D	Two-Dimensional
3D	Three-Dimensional
AFP	Alpha-Fetoprotein
AR	Androgen Receptor
EADB	Estrogenic Activity Database
ED	Endocrine Disruptor
EDKB	Endocrine Disruptors Knowledge Base
EDSP	Endocrine Disruptor Screening Program
EPA	Environmental Protection Agency
DF	Decision Forest
ER	Estrogen Receptor
FDA	Food and Drug Administration
FN	False Negative
FP	False Positive
MCC	Matthews Correlation Coefficient
NMR	Nuclear Magnetic Resonance
QSAR	Quantitative Structure-Activity Relationship
SDF	Structure-Data File
SHBG	Sex Hormone-Binding Globulin
STD	Standard Deviation
TN	True Negative
TP	True Positive

References

1. Dodds, E.C.; Goldberg, L.; Lawson, W.; Robinson, R. Estrogenic activity of certain synthetic compounds. *Nature* **1938**, *141*, 247–248. [CrossRef]
2. Huggins, C.; Hodges, C.V. Studies on prostatic cancer. I. The effect of castration, of estrogen and androgen injection on serum phosphatases in metastatic carcinoma of the prostate. *CA Cancer J. Clin.* **1972**, *22*, 232–240. [CrossRef] [PubMed]
3. National Cancer Institute. Diethylstilbestrol (DES) and Cancer. Available online: http://www.cancer.gov/about-cancer/causes-prevention/risk/hormones/des-fact-sheet (accessed on 8 March 2016).
4. Zoeller, R.T.; Brown, T.R.; Doan, L.L.; Gore, A.C.; Skakkebaek, N.E.; Soto, A.M.; Woodruff, T.J.; Vom Saal, F.S. Endocrine-disrupting chemicals and public health protection: A statement of principles from the endocrine society. *Endocrinology* **2012**, *153*, 4097–4110. [CrossRef] [PubMed]
5. Nohynek, G.J.; Borgert, C.J.; Dietrich, D.; Rozman, K.K. Endocrine disruption: Fact or urban legend? *Toxicol. Lett.* **2013**, *223*, 295–305. [CrossRef] [PubMed]
6. Willett, C.E.; Bishop, P.L.; Sullivan, K.M. Application of an integrated testing strategy to the U.S. EPA endocrine disruptor screening program. *Toxicol. Sci.* **2011**, *123*, 15–25. [CrossRef] [PubMed]
7. Lisse, T.S.; Hewison, M.; Adams, J.S. Hormone response element binding proteins: Novel regulators of vitamin D and estrogen signaling. *Steroids* **2011**, *76*, 331–339. [CrossRef] [PubMed]
8. Anand-Ivell, R.; Ivell, R. Insulin-like factor 3 as a monitor of endocrine disruption. *Reproduction* **2014**, *147*, 87–95.
9. Martinez-Arguelles, D.B.; Papadopoulos, V. Mechanisms mediating environmental chemical-induced endocrine disruption in the adrenal gland. *Front. Endocrinol.* **2015**, *6*, 29. [CrossRef] [PubMed]
10. Jégou, B. Paracetamol-induced endocrine disruption in human fetal testes. *Nat. Rev. Endocrinol.* **2015**, *11*, 453–454. [CrossRef] [PubMed]
11. Rieger, S.; Zhao, H.; Martin, P.; Abe, K.; Lisse, T.S. The role of nuclear hormone receptors in cutaneous wound repair. *Cell Biochem. Funct.* **2015**, *33*, 1–13. [CrossRef] [PubMed]

12. Sharma, P.; Grabowski, T.B.; Patiño, R. Thyroid endocrine disruption and external body morphology of Zebrafish. *Gen. Comp. Endocrinol.* **2016**, *226*, 42–49. [CrossRef] [PubMed]
13. Ng, H.W.; Perkins, R.; Tong, W.; Hong, H. Versatility or promiscuity: The estrogen receptors, control of ligand selectivity and an update on subtype selective ligands. *Int. J. Environ. Res. Public Health.* **2014**, *11*, 8709–8742. [CrossRef] [PubMed]
14. Beekmann, K.; de Haan, L.H.; Actis-Goretta, L.; Houtman, R.; van Bladeren, P.J.; Rietjens, I.M. The effect of glucuronidation on isoflavone induced estrogen receptor (ER)α and ERβ mediated coregulator interactions. *J. Steroid Biochem. Mol. Biol.* **2015**, *154*, 245–253. [CrossRef] [PubMed]
15. Pellegrini, M.; Bulzomi, P.; Lecis, M.; Leone, S.; Campesi, I.; Franconi, F.; Marino, M. Endocrine disruptors differently influence estrogen receptor β and androgen receptor in male and female rat VSMC. *J. Cell. Physiol.* **2014**, *229*, 1061–1068. [CrossRef] [PubMed]
16. Kharlyngdoh, J.B.; Pradhan, A.; Asnake, S.; Walstad, A.; Ivarsson, P.; Olsson, P.E. Identification of a group of brominated flame retardants as novel androgen receptor antagonists and potential neuronal and endocrine disrupters. *Environ. Int.* **2015**, *74*, 60–70. [CrossRef] [PubMed]
17. Ding, D.; Xu, L.; Fang, H.; Hong, H.; Perkins, R.; Harris, S.; Bearden, E.D.; Shi, L.; Tong, W. The EDKB: An established knowledge base for endocrine disrupting chemicals. *BMC Bioinform.* **2010**, *11*, S5. [CrossRef] [PubMed]
18. Shen, J.; Xu, L.; Fang, H.; Richard, A.M.; Bray, J.D.; Judson, R.S.; Zhou, G.; Colatsky, T.J.; Aungst, J.L.; Teng, C.; *et al.* EADB: An estrogenic activity database for assessing potential endocrine activity. *Toxicol. Sci.* **2013**, *135*, 277–291. [CrossRef] [PubMed]
19. Ng, H.W.; Shu, M.; Luo, H.; Ye, H.; Ge, W.; Perkins, R.; Tong, W.; Hong, H. Estrogenic activity data extraction and *in silico* prediction show the endocrine disruption potential of bisphenol a replacement compounds. *Chem. Res. Toxicol.* **2015**, *28*, 1784–1795. [CrossRef] [PubMed]
20. Ng, H.W.; Doughty, S.W.; Luo, H.; Ye, H.; Ge, W.; Tong, W.; Hong, H. Development and validation of decision forest model for estrogen receptor binding prediction of chemicals using large data sets. *Chem. Res. Toxicol.* **2015**, *28*, 2343–2351. [CrossRef] [PubMed]
21. Ng, H.W.; Zhang, W.; Shu, M.; Luo, H.; Ge, W.; Perkins, R.; Tong, W.; Hong, H. Competitive molecular docking model for predicting estrogen receptor agonists and antagonists. *BMC Bioinform.* **2014**, *15*, S4. [CrossRef] [PubMed]
22. Hong, H.; Fang, H.; Xie, Q.; Perkins, R.; Sheehan, D.M.; Tong, W. Comparative Molecular Field Analysis (CoMFA) model using a large diverse set of natural, synthetic and environmental chemicals for binding to the androgen receptor. *SAR QSAR Environ. Res.* **2003**, *14*, 373–388. [CrossRef] [PubMed]
23. Shi, L.; Tong, W.; Fang, H.; Xie, Q.; Hong, H.; Perkins, R.; Wu, J.; Tu, M.; Blair, R.M.; Branham, W.S.; *et al.* An integrated "4-Phase" approach for setting endocrine disruption screening priorities—Phase I and II predictions of estrogen receptor binding affinity. *SAR QSAR Environ. Res.* **2002**, *13*, 69–88. [CrossRef] [PubMed]
24. Hong, H.; Tong, W.; Fang, H.; Shi, L.; Xie, Q.; Wu, J.; Perkins, R.; Walker, J.D.; Branham, W.; Sheehan, D.M. Prediction of estrogen receptor binding for 58,000 chemicals using an integrated system of a tree-based model with structural alerts. *Environ. Health Perspect.* **2002**, *110*, 29–36. [CrossRef] [PubMed]
25. Tong, W.; Hong, H.; Xie, Q.; Shi, L.; Fang, H.; Perkins, R. Assessing QSAR limitations-a regulatory perspective. *Curr. Comput.-Aided Drug Des.* **2005**, *1*, 195–205. [CrossRef]
26. Hornung, M.W.; Tapper, M.A.; Denny, J.S.; Kolanczyk, R.C.; Sheedy, B.R.; Hartig, P.C.; Aladjov, H.; Henry, T.R.; Schmieder, P.K. Effects-based chemical category approach for prioritization of low affinity estrogenic chemicals. *SAR QSAR Environ. Res.* **2014**, *25*, 289–323. [CrossRef] [PubMed]
27. Devillers, J.; Bro, E.; Millot, F. Prediction of the endocrine disruption profile of pesticides. *SAR QSAR Environ. Res.* **2015**, *26*, 831–852. [CrossRef] [PubMed]
28. Wang, Y.; Bai, F.; Cao, H.; Li, J.; Liu, H.; Gramatica, P. A combined quantitative structure-activity relationship research of quinolinone derivatives as androgen receptor antagonists. *Comb. Chem. High Throughput Screen.* **2015**, *18*, 834–845. [CrossRef] [PubMed]
29. Niinivehmas, S.P.; Manivannan, E.; Rauhamäki, S.; Huuskonen, J.; Pentikäinen, O.T. Identification of estrogen receptor α ligands with virtual screening techniques. *J. Mol. Graph. Model.* **2016**, *64*, 30–39. [CrossRef] [PubMed]

30. Hong, H.; Branham, W.S.; Dial, S.; Moland, C.L.; Fang, H.; Shen, J.; Perkins, R.; Sheehan, D.; Tong, W. Rat alpha-fetoprotein binding affinities of a large set of structurally diverse chemicals elucidated the relationships between structures and binding affinities. *Chem. Res. Toxicol.* **2012**, *25*, 2553–2566. [CrossRef] [PubMed]
31. Shen, J.; Zhang, W.; Fang, H.; Perkins, R.; Tong, W.; Hong, H. Homology modeling, molecular docking, and molecular dynamics simulations elucidated alpha-fetoprotein binding modes. *BMC Bioinform.* **2013**, *14*, S6. [CrossRef] [PubMed]
32. Hong, H.; Branham, W.S.; Ng, H.W.; Moland, C.L.; Dial, S.L.; Fang, H.; Perkins, R.; Sheehan, D.; Tong, W. Human sex hormone binding globulin binding affinities of 125 structurally diverse chemicals and comparison with their binding to androgen receptor, estrogen receptor and a-fetoprotein. *Toxicol. Sci.* **2015**, *143*, 333–348. [CrossRef] [PubMed]
33. Anderson, D.C. Sex-hormone-binding globulin. *Clin. Endocrinol.* **1974**, *3*, 69–96. [CrossRef]
34. Bergstrand, C.G.; Czar, B. Paper electrophoretic study of human fetal serum proteins with demonstration of a new protein fraction. *Scand. J. Clin. Lab. Investig.* **1956**, *9*, 277–286. [CrossRef] [PubMed]
35. Mizejewski, G.J.; MacColl, R. Alpha-fetoprotein growth inhibitory peptides: Potential leads for cancer therapeutics. *Mol. Cancer Ther.* **2003**, *2*, 1243–1255. [PubMed]
36. Mizejewski, G.J. Biological role of alpha-fetoprotein in cancer: Prospects for anticancer therapy. *Expert. Rev. Anticancer Ther.* **2002**, *2*, 89–115. [CrossRef] [PubMed]
37. Alava, M.A.; Sturralde, M.; Lampreave, F.; Pineiro, A. Specific uptake of alpha-fetoprotein and albumin by rat Morris 777 hepatoma cells. *Tumour. Biol.* **1999**, *20*, 52–64. [CrossRef] [PubMed]
38. Uriel, J.; Faivre-Bauman, A.; Trojan, J.; Foiret, D. Immunocytochemical demonstration of alpha-fetoprotein uptake by primary cultures of fetal hemisphere cells from mouse brain. *Neurosci. Lett.* **1981**, *272*, 171–175. [CrossRef]
39. Laborda, J.; Naval, J.; Allouche, M.; Calvo, M.; Georgoulias, V.; Mishal, Z.; Uriel, J. Specific uptake of alpha-fetoprotein by malignant human lymphoid cells. *Int. J. Cancer* **1987**, *40*, 314–318. [CrossRef] [PubMed]
40. Hajeri-Germond, M.; Trojan, J.; Uriel, J.; Hauw, J.J. *In vitro* uptake of exogenous AFP by chicken dorsal root ganglia. *Dev. Neurosci.* **1983**, *6*, 11–15. [CrossRef]
41. Uriel, J.; Trojan, J.; Dubouch, P.; Pineiro, A. Intracellular alpha-fetoprotein and albumin in the developing nervous system of the baboon. *Pathol. Biol.* **1982**, *302*, 79–83.
42. Brock, D.J.; Bolton, A.E.; Monaghan, J.M. Prenatal diagnosis through maternal serum AFP measurement. *Lancet* **1973**, *2*, 293–294.
43. Leek, A.E.; Ruoss, C.F.; Kitau, M.J.; Chard, T. Raised AFP in maternal serum with anenecephalic pregnancy. *Lancet* **1973**, *2*, 385–386. [CrossRef]
44. Benassayag, C.; Migrot, T.M.; Haurigui, M.; Civel, C.; Hassid, J.; Carbonne, B.; Nunez, E.A.; Ferre, F. High polyunsaturated fatty acid, thromboxane A2, and alpha-fetoprotein concentrations at the human feto-maternal interface. *J. Lipid Res.* **1997**, *38*, 276–286. [PubMed]
45. Gross, S.; Catillo, W.; Crane, M.; Espinosa, B.; Carter, S.; DeVeaux, R.; Salafia, C. Maternal serum AFP and HCG levels in women with human HIV. *Am. J. Obstet. Gynecol.* **2003**, *188*, 1052–1056. [CrossRef] [PubMed]
46. Savu, L.; Benassayag, C.; Vallette, G.; Christeff, N.; Nunez, E. Mouse alpha 1-fetoprotein and albumin. Comparison of their binding properties with estrogen and fatty acid ligands. *J. Biol. Chem.* **1981**, *256*, 9414–9418. [PubMed]
47. Uriel, J.; Bouillon, D.; Aussel, C.; Dupiers, M. Alpha-fetoprotein: The major high-affinity estrogen binder in rat uterine cytosols. *Proc. Natl. Acad. Sci. USA* **1976**, *73*, 1452–1456. [CrossRef] [PubMed]
48. Pomper, M.G.; VanBrocklin, H.; Thieme, A.M.; Thomas, R.D.; Kiesewetter, D.O.; Carlson, K.E.; Mathias, C.J.; Welch, M.J.; Katzenellenbogen, J.A. 11 beta-methoxy-, 11 beta-ethyl- and 17 alpha-ethynyl-substituted 16 alpha-fluoroestradiols: Receptor-based imaging agents with enhanced uptake efficiency and selectivity. *J. Med. Chem.* **1990**, *33*, 3143–3155. [CrossRef] [PubMed]
49. VanBrocklin, H.F.; Brodack, J.W.; Mathiss, C.J.; Welch, M.J.; Katzenellenbogen, J.A.; Keenan, J.F.; Mizejewski, G.J. Binding of a 16 a-F18-fluoro-17B-estradiol to alpha-fetoprotein in Sprague-Dawley female rats affects blood levels. *Nucl. Med. Biol.* **1992**, *17*, 769–773.
50. Milligan, S.R.; Khan, O.; Nash, M. Competitive binding of xenobiotic oestrogens to rat alpha-fetoprotein and to sex steroid binding proteins in human and rainbow trout (Oncorhynchus mykiss) plasma. *Gen. Comp. Endocrinol.* **1998**, *112*, 89–95. [CrossRef] [PubMed]

51. Nunez, E.A.; Bennassayag, C.; Sava, L.; Vallette, G.; Delorme, J. Oestrogen binding function of AFP. *J. Steroid Biochem.* **1979**, *11*, 237–243. [CrossRef]
52. Garreau, B.; Vallette, G.; Adlercreutz, H.; Wähälä, K.; Mäkelä, T.; Benassayag, C.; Nunez, E.A. Phytoestrogens: New ligands for rat and human alpha-fetoprotein. *Biochim. Biophys. Acta* **1991**, *1094*, 339–345. [CrossRef]
53. Kleinstreuer, N.C.; Ceger, P.C.; Allen, D.G.; Strickland, J.; Chang, X.; Hamm, J.T.; Casey, W.M. A Curated Database of Rodent Uterotrophic Bioactivity. *Environ. Health Perspectect.* **2015**. [CrossRef] [PubMed]
54. Hong, H.; Xie, Q.; Ge, W.; Qian, F.; Fang, H.; Shi, L.; Su, Z.; Perkins, R.; Tong, W. Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. *J. Chem. Inf. Model.* **2008**, *48*, 1337–1344. [CrossRef] [PubMed]
55. Mold2, Descriptors Generator Software. Available online: http://www.fda.gov/ScienceResearch/BioinformaticsTools/Mold2/ (accessed on 13 January 2016).
56. Hong, H.; Xin, X. ESSESA: An expert system for structure elucidation from spectra analysis. 2. A novel algorithm of perception of the linear independent smallest set of smallest rings. *Anal. Chim. Acta* **1992**, *262*, 179–191. [CrossRef]
57. Hong, H.; Xin, X. ESSESA: An expert system for structure elucidation from spectra analysis. 3. LNSCS for chemical knowledge representation. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 116–120. [CrossRef]
58. Hong, H.; Xin, X. ESSESA: An expert system for structure elucidation from spectra analysis. 4. Canonical representation of structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 730–734. [CrossRef]
59. Hong, H.; Xin, X. ESSESA: An expert system for structure elucidation from spectra analysis. 1. The knowledge base of infrared spectra and analysis and interpretation program. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 203–210. [CrossRef]
60. Hong, H.; Xin, X. ESSESA: An expert system for structure elucidation from spectra analysis. 5. Substructure constraints from from analysis of first-order 1H-NMR spectra. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1259–1266. [CrossRef]
61. Hong, H.; Han, Y.; Xin, X.; Shi, Y. ESSESA: An expert system for structure elucidation from spectra. 6. Substructure constraints from analysis of 13C-NMR spectra. *J. Chem. Inf. Comput. Sci.* **1994**, *35*, 979–1000.
62. Masui, H.; Hong, H. Spec2D: A structure elucidation system based on 1H NMR and H-H COSY spectra in organic chemistry. *J. Chem. Inf. Model.* **2006**, *46*, 775–787. [CrossRef] [PubMed]
63. McPhail, B.; Tie, Y.; Hong, H.; Pearce, B.A.; Schnackenberg, L.K.; Ge, W.; Valerio, L.G.; Fuscoe, J.C.; Tong, W.; Buzatu, D.A.; *et al*. Modeling chemical interaction profiles: I. Spectral data-activity relationship and structure-activity relationship models for inhibitors and non-inhibitors of cytochrome P450 CYP3A4 and CYP2D6 isozymes. *Molecules* **2012**, *17*, 3283–3406. [CrossRef] [PubMed]
64. Tie, Y.; McPhail, B.; Hong, H.; Pearce, B.A.; Schnackenberg, L.K.; Ge, W.; Buzatu, D.A.; Wilkes, J.G.; Fuscoe, J.C.; Tong, W.; *et al*. Modeling chemical interaction profiles: II. Molecular docking, spectral data-activity relationship, and structure-activity relationship models for potent and weak inhibitors of cytochrome p450 cyp3A4 isozyme. *Molecules* **2012**, *17*, 3407–3460. [CrossRef] [PubMed]
65. Neamati, N.; Hong, H.; Owen, J.M.; Sunder, S.; Winslow, H.E.; Christensen, J.L.; Zhao, H.; Burke, T.R., Jr.; Milne, G.W.; Pommier, Y. Salicylhydrazine-Containing Inhibitors of Hiv-1 Integrase: Implication for a Selective Chelation in the Integrase Active Site. *J. Med. Chem.* **1998**, *41*, 3202–3209. [CrossRef] [PubMed]
66. Hong, H.; Neamati, N.; Winslow, H.E.; Christensen, J.L.; Orr, A.; Pommier, Y.; Milne, G.W. Identification of Hiv-1 Integrase Inhibitors Based on a Four-Point Pharmacophore. *Antivir. Chem. Chemother.* **1998**, *9*, 461–472. [CrossRef] [PubMed]
67. Neamati, N.; Hong, H.; Sunder, S.; Milne, G.W.A.; Pommier, Y. Potent Inhibitors of Human Immunodeficiency Virus Type 1 Integrase: A Novel Four-Point Pharmacophore Searching of the NCI 3D Database. *Mol. Pharmacol.* **1997**, *52*, 1041–1055. [PubMed]
68. Hong, H.; Neamati, N.; Wang, S.; Nicklaus, M.C.; Mazumder, A.; Zhao, H.; Burke, T.R.; Pommier, Y.; Milne, G.W.A. Discovery of Hiv-1 Integrase Inhibitors by Pharmacophore Searching. *J. Med. Chem.* **1997**, *40*, 930–936. [CrossRef] [PubMed]
69. Luo, H.; Du, T.; Zhou, P.; Yang, L.; Mei, H.; Ng, H.W.; Zhang, W.; Shu, M.; Tong, W.; Shi, L.; *et al*. Molecular docking to identify associations between drugs and class I human leukocyte antigens for predicting potential idiosyncratic drug reactions. *Comb. Chem. High Throughput Screen.* **2015**, *18*, 296–304. [CrossRef] [PubMed]

70. Drake, R.R.; Neamati, N.; Hong, H.; Pilon, A.; Sunthankar, P.; Hume, S.D.; Wilne, G.W.A.; Pommier, Y. Identification of a mononucleotide binding site in human HIV-1 integrase. *Proc. Natl. Accad. Sci. USA* **1998**, *98*, 1495–1500.
71. Hong, H.; Hong, Q.; Perkins, R.; Shi, L.; Fang, H.; Su, Z.; Dragan, Y.; Fuscoe, J.C.; Tong, W. The accurate prediction of protein family from amino acid sequence by measuring features of sequence fragments. *J. Comput. Biol.* **2009**, *16*, 1671–1688. [CrossRef] [PubMed]
72. Liu, J.; Mansouri, K.; Judson, R.; Martin, M.T.; Hong, H.; Chen, M.; Xu, X.; Thomas, R.; Shah, I. Predicting hepatotoxicity using ToxCast *in vitro* bioactivity and chemical structure. *Chem. Res. Toxicol.* **2015**, *28*, 738–751. [CrossRef] [PubMed]
73. Luo, H.; Ye, H.; Ng, H.W.; Shi, L.; Tong, W.; Mendrick, D.L.; Hong, H. Machine learning methods for predicting HLA-peptide binding activity. *Bioinform. Biol. Insights* **2015**, *9*, 21–29. [PubMed]
74. Tong, W.; Hong, H.; Fang, H.; Xie, Q.; Perkins, R. Decision forest: Combining the predictions of multiple independent decision tree models. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 525–531. [CrossRef] [PubMed]
75. Hong, H.; Tong, W.; Xie, Q.; Fang, H.; Perkins, R. An in silico ensemble method for lead discovery: decision forest. *SAR QSAR Environ. Res.* **2005**, *16*, 339–347. [CrossRef] [PubMed]
76. Decision Forest. Available online: http://www.fda.gov/ScienceResearch/BioinformaticsTools/DecisionForest/ (accessed on 13 January 2016).
77. Xie, Q.; Ratnasinghe, L.D.; Hong, H.; Perkins, R.; Tang, Z.Z.; Hu, N.; Taylor, P.R.; Tong, W. Decision forest analysis of 61 single nucleotide polymorphisms in a case-control study of esophageal cancer; a novel method. *BMC Bioinform.* **2005**, *6*, S4. [CrossRef] [PubMed]
78. Hong, H.; Tong, W.; Perkins, R.; Fang, H.; Xie, Q.; Shi, L. Multiclass decision forest-a novel pattern recognition method for multiclass classification in microarray data analysis. *DNA Cell Biol.* **2004**, *23*, 685–694. [CrossRef] [PubMed]
79. Tong, W.; Xie, Q.; Hong, H.; Shi, L.; Fang, H.; Perkins, R.; Petricoin, E.F. Using decision forest to classify prostate cancer samples on the basis of seldi-tof ms data: Assessing chance correlation and prediction confidence. *Environ. Health Perspect.* **2004**, *112*, 1622–1627. [CrossRef] [PubMed]
80. Guha, R. On the interpretation and interpretability of quantitative structure-activity relationship models. *J. Comput. Aided Mol. Des.* **2008**, *22*, 857–871. [CrossRef] [PubMed]
81. Nishi, S.; Shahbazzadeh, D.; Azuma, M.; Sakai, M. Estrogen binding site of rat AFP. *Tumour. Biol.* **1993**, *14*, 234–237.
82. Use of High Throughput Assays and Computational Tools in the Endocrine Disruptor Screening Program. Available online: http://www.epa.gov/endocrine-disruption/use-high-throughput-assays-and-computational-tools-endocrine-disruptor (accessed on 8 March 2016).
83. Mansouri, K. CERAPP: Collaborative estrogen receptor activity prediction project. *Environ. Health Perspect.* **2016**. [CrossRef] [PubMed]
84. Nishi, S.; Matsue, H.; Yoshida, H.; Yamaoto, R.; Sakai, M. Localization of the estrogen-binding site of alpha-fetoprotein in the chimeric human-rat proteins. *Proc. Natl. Acad. Sci. USA* **1991**, *88*, 3102–3105. [CrossRef] [PubMed]
85. Herve, F.; Gentin, M.; Rajkowski, K.M.; Wong, L.T.; Hsia, C.J.; Cittanova, N. Estrogen-binding properties of rat AFP and its isoforms: Investigation of the apparent non-integrality of sites on the unfractionated protein. *J. Steroid Biochem.* **1990**, *36*, 319–324. [CrossRef]