Ecotoxicology and Environmental Safety 80 (2012) 273–279

Contents lists available at SciVerse ScienceDirect

Ecotoxicology and Environmental Safety

journal homepage: www.elsevier.com/locate/ecoenv

Docking and QSAR study on the binding interactions between polycyclic aromatic hydrocarbons and estrogen receptor

Fei Li [a], Huifeng Wu [a,*], Lianzhen Li [a], Xuehua Li [b], Jianmin Zhao [a], Willie J.G.M. Peijnenburg [c,d]

[a] Key Laboratory of Coastal Zone Environmental Processes, Yantai Institute of Coastal Zone Research (YIC), Chinese Academy of Sciences (CAS); Shandong Provincial Key Laboratory of Coastal Zone Environmental Processes, YICCAS, Yantai Shandong 264003, PR China
[b] Key Laboratory of Industrial Ecology and Environmental Engineering (MOE), School of Environmental Science and Technology, Dalian University of Technology, Linggong Road 2, Dalian 116024, PR China
[c] National Institute of Public Health and the Environment, Laboratory for Ecological Risk Assessment, P.O. Box 1, 3720 BA Bilthoven, The Netherlands
[d] Institute of Environmental Sciences (CML), Leiden University, Leiden, The Netherlands

ARTICLE INFO

Article history:
Received 29 November 2011
Received in revised form
15 March 2012
Accepted 18 March 2012
Available online 13 April 2012

Keywords:
Polycyclic aromatic hydrocarbons (PAHs)
Estrogenic activity
Molecular docking
Quantitative structure–activity relationship (QSAR)
Applicability domain

ABSTRACT

Little is known about the estrogenic activities of polycyclic aromatic hydrocarbons (PAHs) and the underlying mechanisms on estrogenic activities are still unclear. Molecular docking and quantitative structure–activity relationship (QSAR) were used to understand the relationship between molecular structural features and estrogenic activity, and to predict the binding affinity of PAHs to estrogen receptor α (ERα). From molecular docking analysis, hydrogen bonding as well as hydrophobic and π interactions were found between PAHs and ERα. Based on the docking results, appropriate molecular structural parameters were adopted to develop a QSAR model. Five descriptors were included in the QSAR model, which indicated that the estrogenic activity was related to molecular size, van der Waals volumes, shape profiles, polarizabilities and electropological states were significant parameters explaining the estrogenicity. Comparatively, the developed QSAR model had good robustness, predictive ability and mechanistic interpretability. Moreover, the applicability domain of the model was described.

1. Introduction

Polycyclic aromatic hydrocarbons (PAHs) are ubiquitous environmental pollutants (Cao et al., 2011; Dachs et al., 2011). Humans and animals are exposed to PAHs from environmental (air, soil, water), dietary and occupational sources, and also from cigarette smoke (vanSchooten et al., 1997; Watson and Brandt, 2003). Hydroxy-substituted polycyclic aromatic hydrocarbons (HO-PAHs) are formed from the corresponding PAHs in the presence of cytochrome P450 enzymes (CYPs) in humans and in animals, as well as chemically in the atmosphere. Many PAHs are carcinogenic in human and laboratory animals and the principal concern regarding exposure to PAHs is that they increase the risk of cancer (Hayakawa et al., 2007; Ellsworth et al., 2008).

As for PAHs, antiestrogenic activity was observed in the yeast assay system (Tran et al., 1996) and estrogenic activity was found in MCF-7 cells (Charles et al., 2000). The structural similarity of several HO-PAHs to 17β-estradiol, which binds to human estrogen receptor (ER), might account for their estrogenic or antiestrogenic activities. A common structure of estrogenic compounds is a phenol basic structure with a hydrophobic moiety at the *para*-position and no bulky group at the *ortho*-position (Nishihara et al., 2000).

Abbreviations: PAHs, polycyclic aromatic hydrocarbons; 2-OHFl, 2-hydroxyfluorene; 2-OHPh, 2-Hydroxyphenanthrene; Frt, Fluoranthene; 3-OHFR, 3-hydroxyfluoranthene; Py, Pyrene; 1-OHPy, 1-hydroxypyrene; BaA, benzo[a]anthracene; 1-OHBaA, 1-hydroxybenzo[a]anthracene; 2-OHBaA, 2-hydroxybenzo[a]anthracene; 3-OHBaA, 3-hydroxybenzo[a]anthracene; 4-OHBaA, 4-hydroxybenzo[a]anthracene; 5-OHBaA, 5-hydroxybenzo[a]anthracene; 9-OHBaA, 9-hydroxybenzo[a]anthracene; 10-OHBaA, 10-hydroxybenzo[a]anthracene; 11-OHBaA, 11-hydroxybenzo[a]anthracene; BcPh, Benzo[c]phenanthrene; 1-OHBcPh, 1-hydroxybenzo[c]phenanthrene; 2-OHBcPh, 2-hydroxybenzo[c]phenanthrene; 3-OHBcPh, 3-hydroxybenzo[c]phenanthrene; 4-OHBcPh, 4-hydroxybenzo[c]phenanthrene; 5-OHBcPh, 5-hydroxybenzo[c]phenanthrene; Ch, Chrysene; 1-OHCh, 1-hydroxychrysene; 2-OHCh, 2-hydroxychrysene; 3-OHCh, 3-hydroxychrysene; 4-OHCh, 4-hydroxychrysene; 6-OHCh, 6-hydroxychrysene; 10-OHBbFR, 10-hydroxybenzo[b]fluoranthene; 3-OHBbFR, 3-hydroxybenzo[b]fluoranthene; 9-OHBbFR, 9-hydroxybenzo[b]fluoranthene; 1-OHBaP, 1-hydroxybenzo[a]pyrene; 3-OHBaP, 3-hydroxybenzo[a]pyrene; 8-OHBaP, 8-hydroxybenzo[a]pyrene; 4-OHBaP, 4-hydroxybenzo[a]pyrene; 11-OHBgCh, 11-hydroxybenzo[g]chrysene; 13-OHBgCh, 13-hydroxybenzo[g]chrysene; E_{binding}, binding energy; QSARs, quantitative structure-activity relationships; PLS, partial least squares; R^2, squared correlation coefficient; Q^2_{CUM}, the fraction of the total variation of the dependent variables that can be predicted by all the extracted components; Q^2_{EXT}, external explained variance; *SE*, standard error

* Corresponding author. Fax: +86 535 2109000.
E-mail address: hfwu@yic.ac.cn (H. Wu).

0147-6513/$ - see front matter © 2012 Elsevier Inc.. All rights reserved.
http://dx.doi.org/10.1016/j.ecoenv.2012.03.009

Several classes of hydroxy-substituted aromatic hydrocarbons have been linked to estrogenicity (Schultz et al., 1998; Schultz and Sinks, 2002). Previous studies have been conducted to evaluate the estrogenic and antiestrogenic activities of HO-PAHs such as hydroxybenzo[*a*]pyrene (HO-BaP). Nishihara et al. (2000) tested 12 HO-BaPs isomers (1- through 12- HO-BaPs) using the competition binding assay to ERα, and found that 1-, 2-, 3-, and 9-HO-BaPs were estrogenic and 8-HO-BaPs was antiestrogenic. Besides, 2-hydroxyfluorene (2-HO-Fl), 2- and 3-hydroxyphenanthrenes, and 1-hydroxypyrene (1-HO-Py) and *n*-propyl-p-hydroxybenzoate in cigarette smoke condensate were also determined as estrogenic compounds (Hayakawa et al., 2007). These results suggested that the activities of HO-PAHs depend strongly on their structures. Hence, it is important to clarify whether these compounds have any estrogenic effects. However, little is known about the estrogenic activities of PAHs/HO-PAHs and the underlying mechanisms on the estrogenic activities are still unclear.

The experimental methods for determining xenoestrogens summarized by the Organisation for Economic Co-operation and Development (OECD), include methods such as the yeast-based assay (Routledge and Sumpter, 1997), E-Screen (Soto et al., 1995), the MCF-7 cells proliferation test (Gierthy et al., 2003), the rat uterotrophic assay (Kanno et al., 2001) and the Hershberger assay (Yamasaki et al., 2004). However, the number of exhaustive virtual chemicals is much bigger than the number of estrogenic chemicals that we could test. Therefore, given the factors such as time and expense as well as the large number of compounds that may bind to the receptors, there appeared increasing interests in developing computational methods (*in silico*) to predict affinity of compounds with the receptors, including the methodology of quantitative structure–activity relationships (QSARs) (Du et al., 2008; Valadares et al., 2007).

According to the OECD and the US Environmental Protection Agency (US EPA) (OECD, 2007), QSARs are promising tools for modeling and predicting estrogenic activities of xenoestrogens. The mechanistic interpretation is of paramount importance as QSAR models with a clear mechanistic underpinning usually have high credibility, succinctness, and definite boundaries for their applicability domain (Chen et al., 2008). Thus, the molecular structural descriptors selected for constructing QSAR models should be based on analysis of the underlying mechanisms, and facilitate a mechanismtic interpretation.

The initial step for chemicals in the mode of action is their binding to an intracellular receptor (Kavlock et al., 1996). Hence, molecular docking and virtual screening have become an integral part of many modern structure-based computational simulations of chemicals (Martinez et al., 2008). Docking methodologies utilize the knowledge of three-dimensional structure of a receptor in an attempt to optimize the bound ligand or a series of molecules into the active site. Combinational use of docking with QSAR can provide more information on the interaction between the ligand and the receptor (Sippl, 2002; Soderholm et al., 2005). For this purpose, molecular docking was employed in some previous studies to observe the interactions between ligands and receptors (Celik et al., 2008; Amadasi et al., 2009). For example, Celik et al. (2008) found that selected polychlorinated biphenyls (PCBs), plasticizers and pesticides could bind in the steroid binding cavity, interacting with at least one of the two hydrophilic ends of the steroid binding site.

In this study, molecular docking was performed to define a model for the comprehension of the binding interactions between PAHs and ERα, which facilitated the selection of appropriate molecular parameters to characterize the interactions in the QSAR studies. Based on the docking analysis, molecular structural parameters were selected and adopted to construct a QSAR model. From the developed QSAR model, critical molecular structural features related to their estrogenic activities were identified. Furthermore, the developed model was externally validated and the applicability domain was depicted.

2. Materials and methods

2.1. Data compilation and the chemical domain

The relative binding affinity (RBA) of 36 PAHs and HO-PAHs to hERα were determined by Hayakawa et al. using the yeast two-hybrid assay (Hayakawa et al., 2007). Then the RBA values were converted into the form of log *RBA*, which ranged from −3.00 to −0.39 log unit.

2.2. Molecular docking

The binding mode for the compounds to ERα was studied by CDOCKER, which has been incorporated into Discovery Studio 2.5 (Accelrys Software Inc.) through the Dock Ligands protocol. CDOCKER is an implementation of the docking tool based on the CHARMm force field that has been proven to be viable (Wu et al., 2003). The crystal structure of ERα (PDB entry code: 1 × 7 R) was extracted from the Brookhaven Protein Database (PDB http://www.rcsb.org/pdb). In CDOCKER, random ligand conformations are generated through molecular dynamics, and a variable number of rigid-body rotations/translations are applied to each conformation to generate the initial ligand poses. The conformations are further refined by grid-based simulated annealing in the receptor active site, which makes the results accurate. In addition, the electrostatic potentials of the ligand binding site for ERα were calculated by the electrostatic protocol that has been incorporated into Discovery Studio (Ver. 2.5). From the docking analysis, insights into the interactions between the ligands and the receptor were gained, which facilitated the selection of appropriate molecular parameters to characterize the interactions in the subsequent QSAR studies.

2.3. Mechanism consideration and molecular structural parameters selection

As proposed by the OECD guideline (OECD, 2007), QSAR models should be developed based on the mechanism of action. It was hypothesized that the estrogenic activities of PAHs were dependent on the following two processes: (a) The partition of the compounds between water and the biophase, and (b) The interaction between the ligands and the receptor ERα. Molecular structural descriptors that describe hydrophobic, electronic and steric properties of molecules were selected to describe the interaction between PAHs and ERα, which was calculated using the DRAGON 2.1 (Todeschini and Consonni, 2000) and Gaussian 09 packages (Frisch, 2009).

All the initial geometries of the compounds were optimized by semi-empirical method PM3, then optimized at the hybrid Hartree-Fock DFT B3LYP/6-31 G(d,p) level. Solvent (water) effects were taken into consideration implicitly, including the integral equation formulation of the polarized continuum model (IEFPCM). The frequency analysis was performed on the optimized geometries to ensure that the systems had no imaginary vibration frequencies.

The optimized molecular structures were imported to Dragon 2.1 (Talete Srl, Milano, Italy), and 1481 diverse descriptors (different functional groups, constitutional, geometrical, topological, Whim 3D, electronic, etc.) for each molecule were calculated. The quantum chemical descriptors, including the energy of the highest occupied molecular orbital (E_{HOMO}), the energy of the lowest unoccupied molecular orbital (E_{LUMO}), the most positive hydrogen atom in the molecule ($q\mathrm{H}^+$), and the most negative formal charge in the molecule (q^-) were computed by Gaussian 09 programs (Frisch, 2009). The quantum chemical descriptors like E_{HOMO}, E_{LUMO}, $q\mathrm{H}^+$, and q^- were proved successful in many QSAR studies for characterizing intermolecular electrostatic interactions (Colosi et al., 2006). Forward stepwise regression was performed to screen significant molecular descriptors, as also done by Morales et al. (2006).

2.4. Model development and validation

The 36 PAHs and HO-PAHs were randomly divided into a training set (80%) and a validation set (20%), as listed in Table 1. Partial least squares (PLS) regression was performed for the model development as PLS can analyze data with strongly collinear, noisy and numerous predictor variables (Wold et al., 2001).

Simca-S (Version 6.0, Umetri AB & Erisoft AB) was employed for the PLS analysis. Simca-S adopts leave-many-out cross validation to determine the number of PLS components (*A*). Cross-validation simulates how well a model predicts new data, and gives a statistical Q^2_{CUM} (the fraction of the total variation of the dependent variables that can be predicted by all the extracted components) for the model. The PLS analysis was performed repeatedly so as to eliminate redundant molecular structural parameters, as done in our previous studies (Li et al., 2009; Li et al., 2010a, b).

Table 1
Logarithm of the observed and predicted binding affinity (log *RBA*) of the considered compounds and molecular descriptors in the developed QSAR model.

Compounds	log *RBA*			E_{1s}	$MATS_{1v}$	L_{3s}	Mor_{12v}	RDF_{020e}
	Obs.	Pred.	Residuals					
2-OHFl	−0.893	−0.766	−0.127	0.452	0.126	0.090	0.495	1.877
2-OHPh	−0.635	−0.525	−0.110	0.421	0.120	0.000	0.526	1.655
Frt	−2.301	−2.617	0.316	0.271	0.065	0.000	0.539	1.515
3-OHFR	−0.684	−0.790	0.106	0.374	0.128	0.000	0.593	2.446
Py	−2.398	−2.542	0.144	0.264	0.073	0.019	0.351	3.206
1-OHPy	−0.959	−1.064	0.105	0.345	0.126	0.025	0.408	3.317
BaA	−2.301	−2.305	0.004	0.339	0.061	0.000	0.612	1.849
1-OHBaA	−1.328	−1.352	0.024	0.319	0.114	0.000	0.497	2.453
2-OHBaA[a]	−0.863	−1.136	0.273	0.376	0.114	0.001	0.716	1.864
3-OHBaA	−0.642	−0.618	−0.024	0.450	0.114	0.001	0.704	1.936
4-OHBaA	−0.532	−0.875	0.343	0.417	0.114	0.001	0.734	1.989
5-OHBaA[a]	−1.237	−1.217	−0.020	0.360	0.114	0.000	0.685	1.973
9-OHBaA	−0.387	−0.574	0.187	0.453	0.114	0.001	0.678	1.943
10-OHBaA	−0.530	−0.539	0.009	0.463	0.114	0.001	0.717	1.919
11-OHBaA	−1.569	−1.332	−0.237	0.354	0.114	0.001	0.733	2.535
BcPh	−2.699	−2.357	−0.342	0.312	0.061	0.000	0.433	2.443
1-OHBcPh	−3.000	−3.085	0.085	0.270	0.114	0.189	0.673	2.503
2-OHBcPh[a]	−0.733	−1.134	0.401	0.356	0.114	0.001	0.530	2.492
3-OHBcPh	−0.462	−0.496	0.034	0.445	0.114	0.000	0.505	2.592
4-OHBcPh	−0.678	−0.625	−0.052	0.431	0.114	0.001	0.532	2.643
5-OHBcPh	−1.357	−1.018	−0.339	0.366	0.114	0.000	0.481	2.562
Ch	−2.523	−2.410	−0.113	0.326	0.061	0.000	0.629	1.842
1-OHCh	−0.900	−0.847	−0.053	0.414	0.114	0.001	0.686	1.858
2-OHCh[a]	−0.391	−0.577	0.185	0.440	0.114	0.001	0.564	2.289
3-OHCh	−0.860	−1.060	0.200	0.390	0.114	0.000	0.745	1.898
4-OHCh	−1.921	−1.563	−0.358	0.324	0.114	0.001	0.710	3.568
6-OHCh[a]	−1.046	−1.382	0.336	0.318	0.114	0.001	0.531	2.133
10-OHBbFR	−0.939	−0.779	−0.160	0.418	0.120	0.000	0.738	2.653
3-OHBkFR	−0.824	−0.851	0.027	0.424	0.120	0.007	0.835	2.208
9-OHBkFR	−0.688	−0.410	−0.278	0.476	0.120	0.000	0.810	1.893
1-OHBaP[a]	−0.728	−0.895	0.167	0.371	0.120	0.001	0.544	1.721
3-OHBaP	−0.405	−0.682	0.277	0.404	0.120	0.000	0.571	1.641
8-OHBaP	−0.425	−0.588	0.163	0.410	0.120	0.001	0.499	1.760
4-OHBeP[a]	−0.932	−1.238	0.307	0.329	0.120	0.000	0.650	0.922
11-OHBgCh	−0.706	−0.818	0.112	0.417	0.110	0.001	0.522	3.291
13-OHBgCh	−0.870	−0.926	0.056	0.389	0.110	0.001	0.446	2.896

[a] Compounds in the validation set.

The model predictability was evaluated by external validation. The performance of external validation was characterized by the external explained variance (Q^2_{EXT}) and the standard errors (*SE*), which were described as following (Schüürmann et al., 2008).

$$Q^2_{EXT} = 1 - \sum_{i=1}^{n_{EXT}} (y_i - y_i^{pred})^2 / \sum_{i=1}^{n_{EXT}} (y_i - \bar{y}_{EXT})^2 \tag{1}$$

$$SE = \sqrt{\sum_{i=1}^{n} (y_i - y_i^{pred})^2 / n - 1} \tag{2}$$

where y_i and y_i^{pred} are the observed and predicted values for the *i*-th compound, respectively. $\bar{y}_{EXT}$ is the average response value of the validation set. n and n_{EXT} stand for the number of compounds in the training and validation sets, respectively.

2.5. QSAR applicability domain

The Applicability Domain (AD) is determined from the Williams plot of standardized residuals versus leverage (Hat diagonal) values (h_i). The leverage approach for defining the AD has been described in detail previously (Eriksson et al., 2003; Tropsha et al., 2003). The leverage (h_i) value of a chemical in the original variable space is defined as:

$$h_i = x_i^T (X^T X)^{-1} x_i (i = 1, \ldots, n) \tag{3}$$

where x_i is the descriptor vector of the considered compound and X is the model matrix derived from the training set descriptor values.

The warning leverage value (h^*) is defined as $3(K+1)/n$, where K is the number of predictor variables. When the h value of a compound is lower than h^*, the probability of accordance between predicted and actual values is as high as that for the compounds in the training set. A chemical with $h_i > h^*$ will reinforce the model if the chemical is in the training set. But such a chemical in the validation set implies that it is structurally distant from compounds in the training set and its predicted data may be unreliable. However, this chemical may not appear to be an outlier because its residuals may be low. Thus the leverage and the standardized residual should be combined for the characterization of the AD.

3. Results and discussion

3.1. Structural analysis of docking

The docking view of three representative HO-PAHs (1-HO-BaP, 3-HO-BaP, and 8-HO-BaP) in the binding site of ERα is shown in Fig. 1. Hydrogen bonding, hydrophobic and π interactions are observed to be the characteristic interactions between the HO-PAHs and ERα. H-bond formation was found to occur between the hydroxyl oxygen of the HO-PAHs and the hydrogen of the imidazole of His524 (Fig. 1B), and H-bonds were also found between the hydroxyl oxygen of the HO-PAHs and the carbonyl oxygen of Glu353 which strengthens the binding interaction (Fig. 1C).

Acting as an 'anchor', the hydrogen-bonding intensely determines the 3D space position of the benzene ring in the binding pocket, and facilitates the hydrophobic interaction of the HO-PAHs with the side chain of Leu346, Leu387, Ala350, Phe404, Met421, and Ile424 (Fig. 1). As shown in Fig. 1C, there are also σ–π interactions between the benzene rings of the HO-PAHs and Phe404.

Fig. 2 shows the electrostatic potential of the ligand-binding site for ERα. The binding site has positive potentials. It can thus be

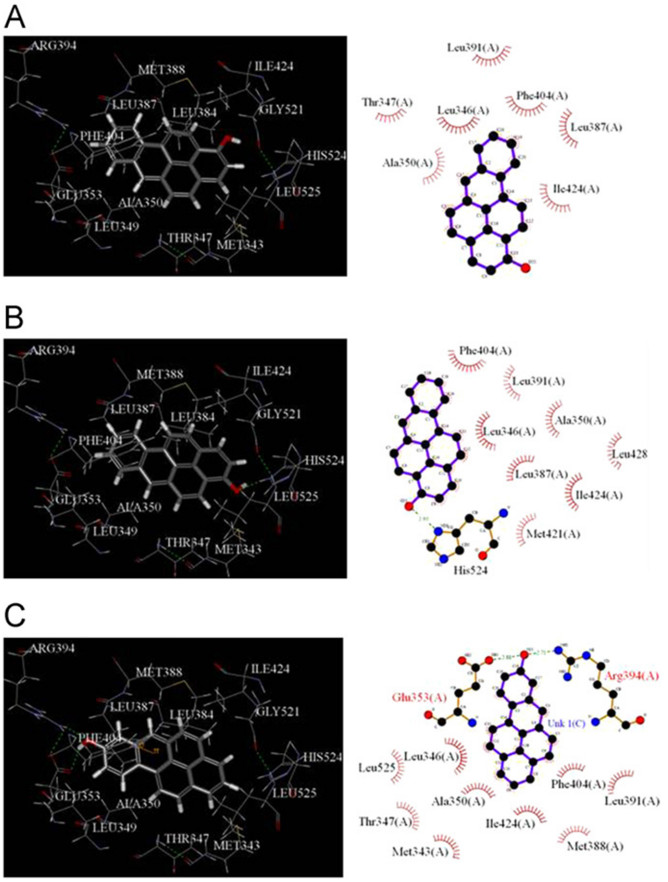

Fig. 1. Docking views of (A) 1-OHBaP, (B) 3-OHBaP, and (C) 8-OHBaP and hydrophobic interaction between HO-PAHs and ERα in the binding site. Green dotted line shows H-bonds. Carbon is colored in grey, oxygen red, and nitrogen blue. ●━● Ligand bond, ●━● Non-ligand bond, ●- 3.0 -● Hydrogen bond and its length, His 53 Non-ligand residues involved in hydrophobic contacts, ● Corresponding atoms involved in hydrophobic contacts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

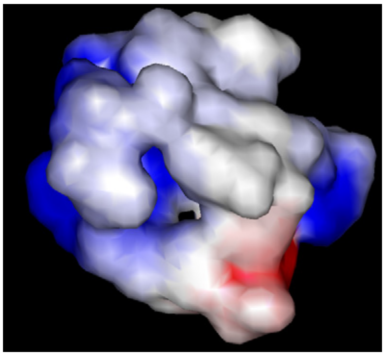

Fig. 2. Electrostatic potential of the ligand binding site for ERα. Further details are needed, like the meaning of the various colors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

concluded that the negative potentials of the molecules facilitate them to bind with ERα.

3.2. Development and validation of the QSAR model for the logRBA

Five descriptors with *F* statistics > 3.84 (significance level = 0.05) were selected as the predictive variables for model development. The molecular structural descriptors are listed in Table 2, along with their physical–chemical meaning.

PLS analysis with the log *RBA* as the dependent variable and the molecular structural parameters as predictor variables resulted in the following optimal QSAR model:

$$\log RBA = -5.18 + 6.90E_{1s} + 1.91 \times 10^{1} MATS_{1v} - 6.53L_{3s} - 8.97 \times 10^{-1} Mor_{12v} - 4.35 \times 10^{-2} RDF_{020e}$$

$$n(\text{training set}) = 29,\ A = 2,\ R^2 = 0.941,\ Q^2_{CUM} = 0.846, SE = 0.188(\text{training set}),$$

$$n(\text{validation set}) = 7,\ Q^2_{EXT} = 0.578,\ SE = 0.905(\text{validation set}),\ p < 0.0001$$

where *p* is the significance level.

The predicted log *RBA* values and residuals for the compounds selected, are listed in Table 1. The R^2 value of the QSAR model was 0.941, indicating a good goodness-of-fit of the model. Q^2_{CUM} of the QSAR is as high as 0.846, implying good robustness of the model. The differences between R^2 and Q^2_{CUM} (0.095) did not exceed 0.3, indicating no over-fitting in the model (Golbraikh and Tropsha, 2002). As shown in Fig. 3, the predicted log *RBA* values were consistent with the observed values for both the validation and training sets. The model revealed acceptable predictability with Q^2_{EXT} = 0.578, *SE* = 0.905. In summary, the developed QSAR model shows satisfactory performance.

3.3. Applicability domain of the developed QSAR model

The distribution of residuals is shown in Fig. 4. Application of the Kolmogorov–Smirnov test for normality (at the 95% confidence level) confirms that the distribution of residuals is a distinctive bell-shaped pattern associated with a normal distribution (mean = 0.05, standard deviation = 0.20), which implies that the residuals are non-systematic and the applicability domain of the developed QSAR model can be visualized by the Williams plot.

Table 2
Physical–chemical meanings of the descriptors used in the developed QSAR model.

Descriptor	Chemical meanings
E_{1s}	1st component accessibility directional WHIM index/ weighted by atomic electrotopological states
$MATS_{1v}$	Moran autocorrelation—lag 1/weighted by atomic van der Waals volumes
L_{3s}	3rd component accessibility directional WHIM index/ weighted by atomic electrotopological states
Mor_{12v}	3D-MoRSE—signal 12/weighted by atomic van der Waals volumes
RDF_{020e}	Radial Distribution Function—2.0/weighted by atomic Sanderson electronegativities

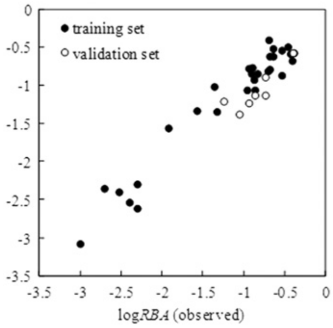

Fig. 3. Plot of observed versus predicted log *RBA* values for the training and validation.

As shown in the Williams plot (Fig. 5), h_i values of all the compounds in the training and validation sets were lower than the warning value ($h^*=0.621$), and all the compounds in both the training and validation sets were in the domain. 1-OHBcPh in the training set was found to have large leverage values ($h > h^*$). This chemical was predicted correctly, indicating that the developed QSAR model has good extrapolation ability. For all the compounds in the training and validation sets, their standardized residuals were smaller than 2.5 standard deviation units (2.5σ), and there were no outliers for the developed QSAR model. Thus, the developed QSAR model can be used to predict the log *RBA* of PAHs and HO-PAHs.

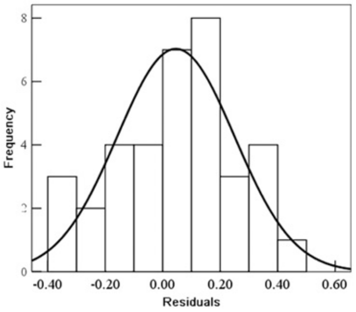

Fig. 4. Distribution of the residuals for log *RBA* values.

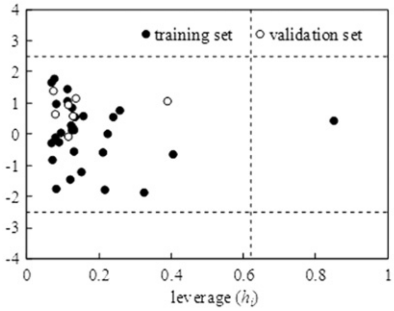

Fig. 5. Plot of standardized residuals versus leverages. Dash lines represent ±2.5 standardized residual, dotted line represents warning leverage ($h^*=0.621$).

Table 3
VIP values and PLS weights for the optimal PLS model.

	VIP	*W**c[1]	*W**c[2]
E_{1s}	1.506	0.703	0.147
$MATS_{1v}$	1.298	0.605	0.289
L_{3s}	0.693	−0.313	−0.346
Mor_{12v}	0.664	0.146	−0.876
RDF_{020e}	0.352	−0.142	0.243

3.4. Mechanistic implications of the developed QSAR model

The developed PLS model extracted 2 PLS components which were loaded primarily upon 5 predictor variables. Values of the variable importance in the projection (*VIP*) and PLS weights (*W**) are listed in Table 3. The *W** values can be used to estimate how the predictor variables and the response variables combine in the projections (PLS components), and how they relate to each other.

The first PLS component was loaded primarily on the 3 descriptors, E_{1s}, $MATS_{1v}$ and L_{3s}. (Table 3). E_{1s} and L_{3s} belong to the WHIM descriptors and are weighted by atomic electrotopological states (Todeschini and Consonni, 2000). E_{1s} remarkably governs log*RBA* since its *VIP* is the largest among all the predictor variables. $MATS_{1v}$ is a 2D autocorrelation descriptor that is weighted by atomic van der Waals volumes (Roy and Kadam, 2006). *W**[1] and the coefficients in the developed QSAR model indicate that E_{1s} and $MATS_{1v}$ are positively correlated with the log *RBA* values, whilst the L_{3s} was negatively correlated with the log *RBA* values. The observation is reasonable since $MATS_{1v}$ correlates with $q\mathrm{H}^+$ positively ($r=0.955$, $p < 0.001$), and PAHs with large $q\mathrm{H}^+$ values tend to form Hydrogen bonding easily, leading to large log *RBA* values.

The second PLS component also extracts 3 descriptors, L_{3s}, Mor_{12v} and RDF_{020e}. Mor_{12v} belongs to 3D-MoRSE descriptors, which is the representation of the 3D structure of a molecule and is weighted by atomic van der Waals volumes (Gasteiger et al., 1996). RDF_{020e} is a RDF descriptor, which could provide information about bond lengths, ring types, planar and nonplanar systems, atom types and molecular weight (Ferreira et al., 2009). RDF_{020e} is weighted by atomic Sanderson electronegativities (Todeschini and Consonni, 2000). The negative *W**[2] and coefficient of Mor_{12v} in the QSAR model indicated the positive correlation between Mor_{12v} and log *RBA*, and the log *RBA* is correlated with RDF_{020e} positively. In general, the current QSAR model indicated the log *RBA* value was related to molecular size, van der Waals volumes, shape profiles and reactivity parameters such as polarizabilities and electropological states.

3.5. Comparison with other QSAR models

In Table 4, the current QSAR model was compared with 3 published QSAR models for RBA to the ER. Asikainen et al. (2004)

Table 4
Comparison with current QSAR models.

No.	Endpoint	Algorithm[a]	Goodness-of-fit, robustness and predictivity[b]							AD[c]	References
			Training set				Validation set				
			n	*K*	R^2	Q^2_{CUM}	*m*	Q^2_{EXT}	*RMSE*		
1	log *RBA*	kNN*	61	185	0.889	0.790	22	0.740	0.680	N	Asikainen et al. (2004)
2	log *RBA*	BP-ANN*	132	49	0.920	0.710	NM	NM	NM	N	Marini et al. (2005)
3	log *RBA*	HQSAR**	130	NM	0.756	0.585	23	0.150	**1.088**	N	Shi et al. (2001)
4	log *RBA*	PLS***	29	5	0.941	0.846	7	0.578	0.905	Y	This study

Note. The bold-faced values were not listed in the references and calculated for comparison using the supplementary data. Estrogenic activity is recorded as the *RBA*, which defined as the ratio of the molar concentration of E_2 to that of the competing chemical required to decrease radiolabeled E_2-receptor binding by 50%, which is then multiplied by 100.

[a] The transparency of different statistical methodology were asterisked: *=lower transparency, **=medium transparency, and ***=higher transparency.

[b] *n* and *m* are the numbers of compounds in the training and validation sets, respectively; *K* is the number of molecular descriptors; R^2 is the squared correlation coefficient between observed and predicted values; Q^2_{CUM} is the fraction of the total variation of the dependent variables that can be predicted by all the extracted components, Q^2_{EXT} is squared correlation coefficient for the validation set. "NM" means it was not mentioned in the reference.

[c] Y and N denote the model is assessed with or without AD discussion, respectively.

employed the kNN method and 185 molecular structural descriptors to develop a QSAR model. The Asikainen model has good robustness and predictivity ($Q^2_{LOO}=0.790$, $Q^2_{EXT}=0.740$), however it is difficult to interpret the model since the involvement of so many descriptors. A BP-ANN model was well developed by Marini et al. (2005), with $Q^2_{LOO}=0.710$. Nevertheless the ANN is like a black box, which does not facilitate mechanism interpretations (Liu et al., 2006). Furthermore, external validation was not mentioned. Shi et al. (2001) developed QSAR models using HQSAR methods. For the HQSAR model, $Q^2_{LOO}=0.585$, $Q^2_{EXT}=0.150$, $RMSE=1.008$, indicative of low robustness and poor external predictivity. All the 3 comparative QSAR models did not discuss ADs.

4. Conclusion

Hydrogen bonding, hydrophobic and π–π interactions between ligands and ERα govern the estrogenic activities of the PAHs/HO-PAHs. Comparatively, the developed QSAR model has good robustness, predictive ability and mechanism interpretability. Compounds with higher E_{1s} and $MATS_{1v}$ values tend to have higher estrogenic activity. The model can be applied to predict the estrogenic activity of other PAHs/HO-PAHs. Comprehension of the binding interactions between the ligands and the receptor through docking analysis is necessary for development of mechanism-based QSAR models.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (21107136), The 100 Talents Program of the Chinese Academy of Sciences, the Key Laboratory of Industrial Ecology and Environmental Engineering and China Ministry of Education (KLIEEE-10-09).

References

Amadasi, A., Mozzarelli, A., Meda, C., Maggi, A., Cozzini, P., 2009. Identification of xenoestrogens in food additives by an integrated in silico and in vitro approach. Chem. Res. Toxicol. 22, 52–63.

Asikainen, A.H., Ruuskanen, J., Tuppurainen, K.A., 2004. Consensus kNN QSAR: A versatile method for predicting the estrogenic activity of organic compounds in silico. A comparative study with five estrogen receptors and a large, diverse set of ligands. Environ. Sci. Technol. 38, 6724–6729.

Cao, J., Han, X., Zhou, N.Y., Cui, Z.H., Ma, M.F., Li, L.B., Cai, M., Li, Y.F., Lin, H., Li, Y., Ao, L., Liu, J.Y., 2011. Association between urinary Polycyclic aromatic hydrocarbon metabolites and sperm DNA damage: a population study in chongqing, China. Environ. Health Perspect. 119, 652–657.

Celik, L., Lund, J.D.D., Schiott, B., 2008. Exploring interactions of endocrine-disrupting compounds with different conformations of the human estrogen receptor alpha ligand binding domain: a molecular docking study. Chem. Res. Toxicol. 21, 2195–2206.

Charles, G.D., Bartels, M.J., Zacharewski, T.R., Gollapudi, B.B., Freshour, N.L., Carney, E.W., 2000. Activity of benzo[a]pyrene and its hydroxylated metabolites in an estrogen receptor-alpha reporter gene assay. Toxicol. Sci. 55, 320–326.

Chen, J.W., Li, X.H., Yu, H.Y., Wang, Y., Qiao, X.L., 2008. Progress and perspectives of quantitative structure–activity relationships used for ecological risk assessment of toxic organic compounds. Sci. China Series B-Chem. 51, 593–606.

Colosi, L.M., Huang, Q.G., Weber, W.J., 2006. Quantitative structure–activity relationship based quantification of the impacts of enzyme–substrate binding on rates of peroxidase-mediated reactions of estrogenic phenolic chemicals. J. Am. Chem. Soc. 128, 4041–4047.

Dachs, J., Cabrerizo, A., Moeckel, C., Ojeda, M.J., Caballero, G., Barcelo, D., Jones, K.C., 2011. Ubiquitous net volatilization of polycyclic aromatic hydrocarbons from soils and parameters influencing their soil–air partitioning. Environ. Sci. Technol. 45, 4740–4747.

Du, J., Qin, J., Liu, H.X., Yao, X.J., 2008. 3D-QSAR and molecular docking studies of selective agonists for the thyroid hormone receptor beta. J. Mol. Graph. Model. 27, 95–104.

Ellsworth, D.L., Weyandt, J., Ellsworth, R.E., Hooke, J.A., Shriver, C.D., 2008. Environmental chemicals and breast cancer risk—a structural chemistry perspective. Curr. Med. Chem. 15, 2680–2701.

Eriksson, L., Jaworska, J., Worth, A.P., Cronin, M.T.D., McDowell, R.M., Gramatica, P., 2003. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ. Health Perspect. 111, 1361–1375.

Ferreira, I.C.F.R., Abreu, R.M.V., Queiroz, M.J.R.P., 2009. QSAR model for predicting radical scavenging activity of di(hetero)arylamines derivatives of benzo[b]thiophenes. Eur. J. Med. Chem. 44, 1952–1958.

Frisch, M.J., Trucks, G.W., Schlegel, H.B., Scuseria, G.E., Robb, M.A., Cheeseman, J.R., Scalmani, G., Barone, V., Mennucci, B., Petersson, G.A., Nakatsuji, H., Caricato, M., Li, X., Hratchian, H.P., Izmaylov, A.F., Bloino, J., Zheng, G., Sonnenberg, J.L., Hada, M., Ehara, M., Toyota, K., Fukuda, R., Hasegawa, J., Ishida, M., Nakajima, T., Honda, Y., Kitao, O., Nakai, H., Vreven, T., Montgomery Jr., J.A., Peralta, J.E., Ogliaro, F., Bearpark, M., Heyd, J.J., Brothers, E., Kudin, K.N., Staroverov, V.N., Kobayashi, R., Normand, J., Raghavachari, K., Rendell, A., Burant, J.C., Iyengar, S.S., Tomasi, J., Cossi, M., Rega, Millam, N.J., Klene, M., Knox, J.E., Cross, J.B., Bakken, V., Adamo, C., Jaramillo, J., Gomperts, R.E., Stratmann, O., Yazyev, A.J., Austin, R., Cammi, C., Pomelli, J.W., Ochterski, R., Martin, R.L., Morokuma, K., Zakrzewski, V.G., Voth, G.A., Salvador, P., Dannenberg, J.J., Dapprich, S., Daniels, A.D., Farkas, O., Foresman, J.B., Ortiz, J.V., Cioslowski, J., Fox, D.J., 2009. Gaussian 09, Revision A.1. Gaussian, Inc., Wallingford CT.

Gasteiger, J., Sadowski, J., Schuur, J., Selzer, P., Steinhauer, L., Steinhauer, V., 1996. Chemical information in 3D space. J. Chem. Inf. Comput. Sci. 36, 1030–1037.

Gierthy, J.F., Arcaro, K.F., Li, A., Silber, P., Lloyd, S., Yang, Y., 2003. Optimization and validation of the MCF-7 focus assay for estrogen modulators. Toxicol. Sci. 72, 154–155.

Golbraikh, A., Tropsha, A., 2002. Beware of q^2!. J. Molecul. Graph. Model 20, 269–276.

Hayakawa, K., Onoda, Y., Tachikawa, C., Hosoi, S., Yoshita, M., Chung, S.W., Kizu, R., Toriba, A., Kameda, T., Tang, N., 2007. Estrogenic/Antiestrogenic activities of polycyclic aromatic hydrocarbons and their monohydroxylated derivatives by yeast two-hybrid assay. J. Health Sci. 53, 562–570.

Kanno, J., Onyon, L., Haseman, J., Fenner-Crisp, P., Ashby, J., Owens, W., 2001. The OECD program to validate the rat uterotrophic bioassay to screen compounds for *in vivo* estrogenic responses: Phase 1. Environ. Health Perspect. 109, 785–794.

Kavlock, R.J., Daston, G.P., DeRosa, C., FennerCrisp, P., Gray, L.E., Kaattari, S., 1996. Research needs for the risk assessment of health and environmental effects of endocrine disruptors: A report of the US EPA-sponsored workshop. Environ. Health Perspect. 104, 715–740.

Li, F., Chen, J.W., Wang, Z., Li, J., Qiao, X.L., 2009. Determination and prediction of xenoestrogens by recombinant yeast-based assay and QSAR. Chemosphere 74, 1152–1157.

Li, F., Li, X.H., Shao, J.P., Chi, P., Chen, J.W., Wang, Z.J., 2010a. Estrogenic activity of anthraquinone derivatives: *in vitro* and *in silico* studies. Chem. Res. Toxicol. 23, 1349–1355.

Li, F., Xie, Q., Li, X.H., Li, N., Chi, P., Chen, J.W., Wang, Z.J., Hao, C., 2010b. Hormone activity of hydroxylated polybrominated diphenyl ethers on human thyroid receptor-beta: *in vitro* and *in silico* investigations. Environ. Health Perspect. 118, 602–606.

Liu, H.X., Papa, E., Gramatica, P., 2006. QSAR prediction of estrogen activity for a large set of diverse chemicals under the guidance of OECD principles. Chem. Res. Toxicol. 19, 1540–1548.

Marini, F., Roncaglioni, A., Novic, M., 2005. Variable selection and interpretation in structure–affinity correlation modeling of estrogen receptor binders. J. Chem. Inf. Model. 45, 1507–1519.

Martinez, L., Polikarpov, I., Skaf, M.S., 2008. Only subtle protein conformational adaptations are required for ligand binding to thyroid hormone receptors: Simulations using a novel multipoint steered molecular dynamics approach. J. Phy. Chem. B 112, 10741–10751.

Morales, A.H., Duchowicz, P.R., Perez, M.A.C., Castro, E.A., Cordeiro, M.N.D.S., Gonzalez, M.P., 2006. Application of the replacement method as a novel variable selection strategy in QSAR. 1. Carcinogenic potential. Chemometr. Intell. Lab. 81, 180–187.

Nishihara, T., Nishikawa, J., Kanayama, T., Dakeyama, F., Saito, K., Imagawa, M., Takatori, S., Kitagawa, Y., Hori, S., Utsumi, H., 2000. Estrogenic activities of 517 chemicals by yeast two-hybrid assay. J. Health Sci. 46, 282–298.

OECD, 2007. Guidance document on the validation of (Quantitative) Structure–Activity Relationships [(Q)SARs] models. Available online at〈http://appli1.oecd.org/olis/2007doc.nsf/linkto/env-jm-mono(2007)2〉.

Routledge, E.J., Sumpter, J.P., 1997. Structural features of alkylphenolic chemicals associated with estrogenic activity. J. Biol. Chem. 272, 3280–3288.

Roy, N., Kadam, R.U., 2006. Cluster analysis and two-dimensional quantitative structure–activity relationship (2D-QSAR) of Pseudomonas aeruginosa deacetylase LpxC inhibitors. Bioorg. Med. Chem. Lett. 16, 5136–5143.

Schüürmann, G., Ebert, R.U., Chen, J.W., Wang, B., Kuhne, R., 2008. External validation and prediction employing the predictive squared correlation coefficient—test set activity mean vs training set activity mean. J. Chem. Inf. Model. 48, 2140–2145.

Schultz, T.W., Kraut, D.H., Sayler, G.S., Layton, A.C., 1998. Estrogenicity of selected biphenyls evaluated using a recombinant yeast assay. Environ. Toxicol. Chem. 17, 1727–1729.

Schultz, T.W., Sinks, G.D., 2002. Xenoestrogenic gene exression: Structural features of active polycyclic aromatic hydrocarbons. Environ. Toxicol. Chem. 21, 783–786.

Shi, L.M., Fang, H., Tong, W.D., Wu, J., Perkins, R., Blair, R.M., Branham, W.S., Dial, S.L., Moland, C.I., Sheehan, D.M., 2001. QSAR models using a large diverse set of estrogens. J. Chem. Inf. Comput. Sci. 41, 186–195.

Sippl, W., 2002. Development of biologically active compounds by combining 3D QSAR and structure-based design methods. J. Comput. Aid. Mol. Des. 16, 825–830.

Soderholm, A.A., Lehtovuori, P.T., Nyronen, T.H., 2005. Three-dimensional structure–activity relationships of nonsteroidal ligands in complex with androgen receptor ligand-binding domain. J. Med. Chem. 48, 917–925.

Soto, A.M., Sonnenschein, C., Chung, K.L., Fernandez, M.F., Olea, N., Serrano, F.O., 1995. The E-Screen assay as a tool to identify estrogens—an update on estrogenic environmental-pollutants. Environ. Health Perspect. 103, 113–122.

Todeschini, R., Consonni, V., 2000. Handbook of molecular descriptors. Wiley-VCH, Weinheim, Germany.

Tran, D.Q., Ide, C.F., McLachlan, J.A., Arnold, S.F., 1996. The anti-estrogenic activity of selected polynuclear aromatic hydrocarbons in yeast expressing human estrogen receptor. Biochem. Bioph. Res. Co. 229, 102–108.

Tropsha, A., Gramatica, P., Gombar, V.K., 2003. The importance of being earnest: Validation is the absolute essential for successful application and interpretation of QSPR models. QSAR Comb. Sci. 22, 69–77.

Valadares, N.F., Castilho, M.S., Polikarpov, I., Garratt, R.C., 2007. 2D QSAR studies on thyroid hormone receptor ligands. Bioorgan. Med. Chem. 15, 4609–4617.

vanSchooten, F.J., Moonen, E.J.C., vanderWal, L., Levels, P., Kleinjans, J.C.S., 1997. Determination of polycyclic aromatic hydrocarbons (PAH) and their metabolites in blood, feces, and urine of rats orally exposed to PAH contaminated soils. Arch. Environ. Con. Toxicol. 33, 317–322.

Watson, W.P., Brandt, H.C.A., 2003. Monitoring human occupational and environmental exposures to polycyclic aromatic compounds. Ann. Occup. Hyg. 47, 349–378.

Wold, S., Sjostrom, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemometr. Intell. Lab. 58, 109–130.

Wu, G.S., Robertson, D.H., Brooks, C.L., Vieth, M., 2003. Detailed analysis of grid-based molecular docking: A case study of CDOCKER—A CHARMm-based MD docking algorithm. J. Comput. Chem. 24, 1549–1562.

Yamasaki, K., Sawaki, M., Noda, S., Muroi, T., Takakura, S., Mitoma, H., Sakamoto, S., Nakai, M., Yakabe, Y., 2004. Comparison of the Hershberger assay and androgen receptor binding assay of twelve chemicals. Toxicology 195, 177–186.