Pure Appl. Chem., Vol. 75, Nos. 11–12, pp. 2375–2388, 2003.

Workshop 1.2

Regulatory application of SAR/QSAR for priority setting of endocrine disruptors: A perspective*

Weida Tong[1,‡], Hong Fang[2], Huixiao Hong[2], Qian Xie[2], Roger Perkins[2], Jeanne Anson[1], and Daniel M. Sheehan[1]

[1]FDA's National Center for Toxicological Research (NCTR), Jefferson, AR 72079, USA; [2]Logicon ROW Sciences, Jefferson, AR 72079, USA

Abstract: Some seven years have passed since the U.S. legislature mandated the Environmental Protection Agency (EPA) to develop and implement a screening and testing program for chemicals that may disrupt the delicate endocrine system. The envisioned EPA program has evolved to incorporate a tiered scheme of in vitro and in vivo assays, and considered QSAR as a viable method to set testing priorities. At the U.S. Food and Drug Administration's (FDA) National Center for Toxicological Research (NCTR), the Endocrine Disruptor Knowledge Base Project has developed models to predict estrogen and androgen receptor binding. Our approach rationally integrates various QSAR models into a sequential "Four-Phase" scheme according to the strength of each type of model. In four hierarchical phases, models predict the inactive chemicals that are then eliminated from the pool of chemicals to which increasingly precise but more time-consuming models are subsequently applied. Each phase employs different models selected to work complementarily in representing key activity-determining structure features in order to absolutely minimize the rate of false negatives, an outcome we view as paramount for regulatory use. In this paper, the QSAR models developed at NCTR, and particularly how we integrated these models into the "Four-Phase" system will be discussed for a number of datasets, including 58 000 chemicals identified by the U.S. EPA.

INTRODUCTION

A large number of environmental chemicals are suspected of disrupting endocrine function by mimicking or antagonizing natural hormones in experimental animals, wildlife, and humans. There is growing concern among the scientific community, government regulators, and the public that these endocrine-disrupting chemicals (EDCs) in the environment are adversely affecting human and wildlife health [1,2]. Adverse outcomes have been observed in experimental animals and wildlife; potential effects on humans include reproductive and developmental toxicity, carcinogenesis, immunotoxicity, and neurotoxicity, among others [3]. EDCs may exert adverse effects through a variety of mechanisms, such as estrogen receptor (ER)-mediated mechanisms of toxicity.

The scientific debate surrounding EDCs has grown contentiously, in part owing to the fact that some suspected EDCs are produced in high volume, and many chemicals are economically important. These public and regulatory concerns led to government regulatory actions [4] and expanded research across Europe, Japan, and North America. The U.S. Congress in 1996 mandated that the Environmental

*Report from a SCOPE/IUPAC project: Implication of Endocrine Active Substances for Human and Wildlife (J. Miyamoto and J. Burger, editors). Other reports are published in this issue, *Pure Appl. Chem.* **75**, 1617–2615 (2003).

‡Corresponding author

Protection Agency (EPA) should develop a strategy for screening and testing a large number of chemicals found in drinking water and food additives [4] for their endocrine disruption potential. In response to Congressional action, the EPA established the Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC), which includes scientific expertise from government, academia, and industry. EDSTAC recommended a two-tier (Tier 1: screening and Tier 2: testing) strategy to screen and test estrogenic, androgenic, and thyroidal activities for a large number of chemicals. To accomplish this, chemicals will be screened (Tier 1) using a multiple endpoint strategy that includes more than 20 different in vitro and in vivo assays recommended by EDSTAC [5]. Although more than ~87 000 chemicals were initially selected for evaluation, many are polymers, leaving about ~58 000 chemicals for evaluation in Tier 1. The number that will progress to the testing step (Tier 2) [6], is not known. Processing chemicals through both tiers will require many years and extensive resources. Hence, the EPA has adopted an approach requiring priority setting before Tier 1 (<www.epa.gov/scipoly/oscpendo/>).

Among the types of hormonal activities, estrogenic activities have been most widely studied. Estrogenic endocrine disruption can result from a variety of biological mechanisms. We found a strong linear correlation among a diverse group of chemicals between binding affinities with the ER from the rat uterine cytosol and those with the human ER of the α-subtype [7]. Furthermore, the rat ER binding data also correlates strongly with the results from assays measuring estrogenicity using downstream events, i.e., the yeast-based reporter gene and the MCF-7 cell proliferation assays. Importantly, chemicals positive in uterotrophic responses (in vivo estrogenic activity) are also positive in the ER binding assay, indicating that binding affinity is a good predictor of in vivo activity with few false negatives observed [8]. These findings demonstrate that ER binding is the major determinant for estrogenic EDCs, and therefore the prediction of the rat ER binding affinity provides an important piece of information for priority setting.

Structure–activity relationship (SAR) and/or quantitative structure–activity relationship (QSAR) models have proven their utility, from both the pharmaceutical and toxicological perspectives, for identification of chemicals that might interact with ER. While QSAR models (QSAR is used hereafter to encompass models that predict activity on either an ordinal or categorical scale rather than only on a quantitative scale) in pharmaceuticals identify high-affinity ligands, they are particularly effective in toxicology in separating active and inactive chemicals, and in rank-ordering chemicals according to potency. Developing a useful QSAR model heavily depends on many factors, particularly, including the quality of biological data, the descriptor selection, and the choice of statistical approaches. Since any QSAR approach has strengths and weaknesses, the careful selection of a specific model or a combination of models needs to be done in accordance with the intended application.

This review first summarizes our motives and efforts to set up a robust training set (the NCTR dataset) for developing ER QSAR models. The development and validation of the QSAR models to predict ER binding is then described, as well as the rationale for integrating models into a hierarchical scheme for use in priority setting of potential estrogenic EDCs. The review will conclude with a discussion of some key issues for applying QSAR models for regulatory purposes.

NCTR ER DATASET: A ROBUST TRAINING SET FOR QSAR DEVELOPMENT

Although an effective QSAR model depends on a number of factors, the most critical is a training set with high-quality biological data. It is desirable that the training set has the biological data coming from the same assay protocol, and contains a sufficiently large number of chemicals with diverse structure. Most importantly, both the biological activity/potency data and structural characteristics in terms of various descriptors are evenly distributed as far as possible in a range reflecting the domain in which chemicals to be predicted are located.

A number of QSAR models for ER binding developed several years ago [9–13], including some of our early works [14–17], were based on datasets available in the literature. These datasets were too small and/or lacked structural diversity [11,12,14]. Although these models yield good statistical results

in the training and cross-validation steps and explain some structural determinants for ER binding, they had limited applicability in predicting the ER-ligand binding affinity of chemicals that, in fact, cover a wide range of structural diversity.

In order to obtain an adequate training set to develop more robust QSAR models, we developed and validated a rat ER binding assay and assayed 232 chemicals [18,19]. The ER binding activity is represented by relative binding affinity (RBA), where the RBA value for the endogenous ER ligand, 17β-estradiol (E2), was set to 100. This NCTR dataset contains chemicals that were selected to cover the structural diversity that bind to ER with an activity distribution ranging over six orders of magnitude. Figure 1 compares the chemistry space/domain of the NCTR dataset with other literature datasets used for early ER QSAR models.

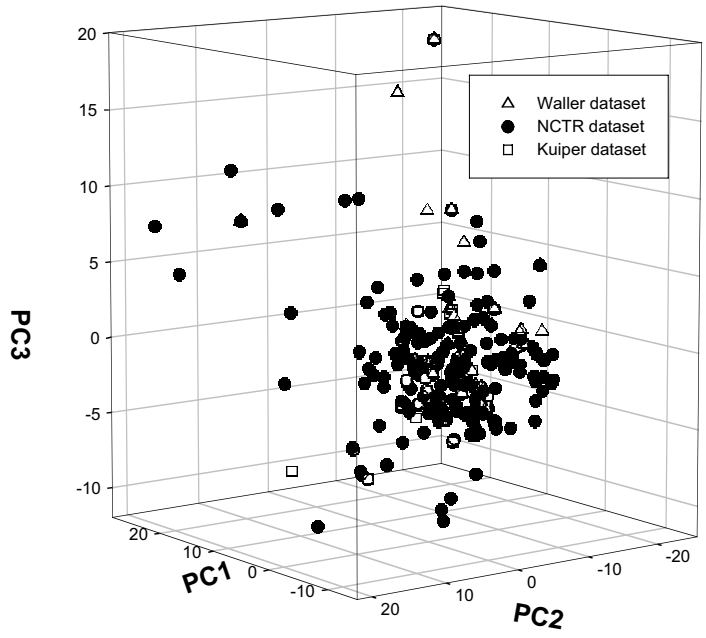

Fig. 1 Comparison of structural diversity of NCTR, Waller, and Kuiper datasets in a chemistry space defined by three principal components (PC) of structural descriptors.

DEVELOPMENT OF QSAR MODELS FOR ER BINDING

We evaluated a number of QSAR approaches useful for predicting ER binding affinity ranging from simple rejection filters often used for drug-like chemical identification to more sophisticated QSAR models often used in the lead optimization. The methods selected for the four-phase approach are discussed below.

Rejection filters

Rejection filters are useful to quickly exclude chemicals from further evaluation. We investigated various physicochemical parameters to use as rejection filters. The ideal filter to be used in the earliest stage of modeling should (1) not generate any false negatives and (2) be able to significantly reduce the number of chemicals for further evaluation. Two rejection filters, molecular weight range and lack of a ring motif, were found to satisfy the two criteria. Chemicals matching any one of these two filters were excluded from subsequent models. The first rejection filter is a molecular weight range, set to <94 and

>1000. The molecular weight of phenol, 94, was considered as the lowest limit for a chemical to bind to ER, whereas a molecular weight of 1000 was considered as the upper limit. The second rejection filter requires that an ER ligand contains at least one ring structure of any size based on the finding from a large literature survey that no known estrogens were found to lack a ring [20].

Structural alerts

Structural alerts are key 2D structural fragments associated with ER binding. Figure 2 depicts the three structural alerts, i.e., the steroid, DES, and phenolic skeletons, selected to identify potential ER binders. Each alert independently characterizes a unique structural feature important for the ER binding. Any chemicals containing any one of these structural alerts are considered to be a potential ER binder.

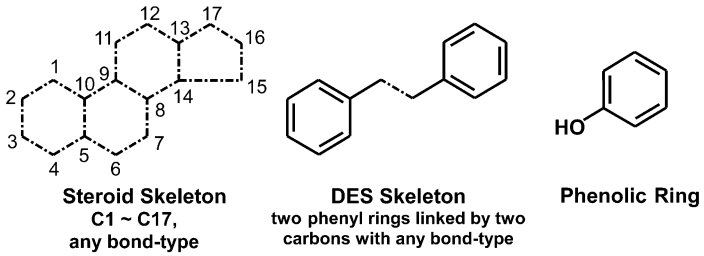

Fig. 2 Three structural alerts defined by 2D substructural features commonly observed for ER ligands.

These three substructures were selected as structural alerts because the length and width of both steroid and DES skeletons conform well to the dimension of the ER binding pocket. In addition, while most endogenous hormones contain the steroid skeleton, most strong estrogens have two benzene rings separated by two carbon atoms [20]. It has been long understood that the phenolic ring is often associated with estrogenic activity [21]. In fact, the contribution of the phenolic ring in ER binding is much more significant than any other structural feature [20]. By overlaying the crystal structure of ER complexes of E2, 4-hydroxytamoxifen, raloxifene, and DES based on their common protein residues at the binding site, we found that the phenolic rings of all four of these ligands are closely positioned at the same location as shown in Fig. 3, allowing hydrogen bond interactions with Glu 353 and Arg 394 of the receptor and a water molecule [22].

Pharmacophore queries

The pharmacophore is represented as a 3D arrangement, in which such molecular features as H-donor, H-acceptor, and hydrophobic center are positioned and combined with a certain geometry, that is supposedly needed for a molecule to exhibit a certain type of biological activity [23]. The pharmacophoric arrangements are used as the queries. A query-matched chemical is considered positive and segregated for further evaluation. One of the advantages of the 3D pharmacophore searching is that it can identify chemicals whose 3D structures are similar to the template structure (normally, a highly active chemical) that may not be discernable by chemists in 2D.

The bound ligand-ER crystal structures [24,25] guided our selection of pharmacophore queries. Using 3D structures of E2, raloxifene, 4-hydroxytamoxifen, and DES in the ER-bound conformation as templates, all possible molecular features as well as molecular shapes were delineated. For each template, any of three to six features were combined to form pharmacophore queries. These queries were sorted according to their discriminatory power to separate active from inactive chemicals in the NCTR dataset. A chemical with any of multiple 3D conformations (up to 100) [26–28] matching the query was considered to be active. Queries with high discriminatory power were further evaluated for their bio-

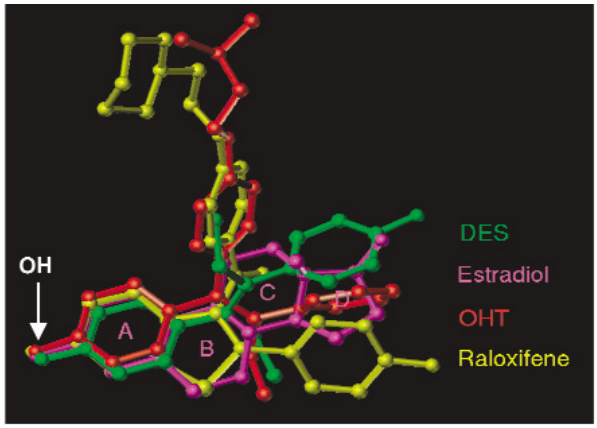

Fig. 3 Relative positions of estradiol (E2), DES, raloxifene, and 4-hydroxytamoxifen (OHT) in the complex with ER.

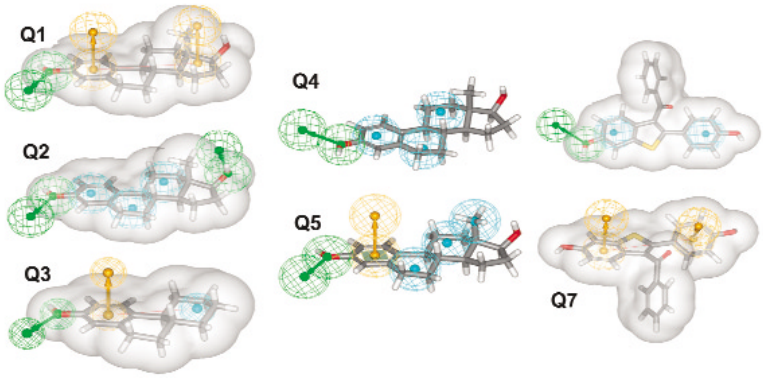

Fig. 4 Seven pharmacophore queries where green mesh balls represent H-bond acceptor sites, blue mesh balls represent hydrophobic centers and yellow mesh balls represent aromatic centers. Solid white surfaces represent shape constraint.

logical relevance based on careful SAR examinations of a large number of chemicals for their binding affinities to ER [20] in conjunction with analysis of the ligand-ER crystal structures [24,25]. Finally, the Tanimoto similarity score was used to determine the uniqueness of each query. Through this process, we identified seven queries that provide unique pharmacophoric signatures for the ER binding (Fig. 4). A chemical could match none, a few, or many of the seven separate queries. We generally found that the number of matches increased in direct proportion to measured activity among the training-set chemicals. Thus, the number of pharmacophore matches could be used to rank-order chemicals in accordance with the potential activity.

CLASSIFICATION MODELS

Classification is a supervised learning technique that provides categorical prediction (e.g., active/inactive). A number of classification methods were evaluated to categorize chemicals as ER binders or nonbinders. While the methods differ in a number of ways, they generally produce similar results [29]. We found that the nature of the descriptors used, and more specifically the effectiveness in which descriptors encode the structural features of the molecules related to the ER binding activity, is far more important than the specific method employed. The selection of biologically relevant descriptors is the crit-

ical step to develop a robust model. We found the Genetic Algorithm to be the preferred method to identify the most biologically relevant descriptors from a large set of descriptors.

For example, using the best 10 descriptors selected by the Genetic Function Approximation approach [30,31] from among 153 descriptors, we were able to construct a Decision Tree model for the classification. The model displays a series of YES/NO rules to classify chemicals into active and inactive categories based on five meaningful descriptors: the Phenolic Ring Index (0 or 1), log *P*, two types of parameters related to the charged partial molecular surface area, and a geometric descriptor connected with the breadth of the molecule. The model identified the Phenolic Ring Index as the most important descriptor for ER binding. If chemicals contained a phenolic moiety but also had log *P* values larger than 1.49, they were more likely to be ER binders. In contrast, chemicals without a phenolic moiety were less likely to be ER binders unless they had relatively large hydrophobicity and charged surface area, and breadth of the structure.

3D-QSAR/CoMFA model

We evaluated three different techniques for the QSAR modeling—CoMFA, CODESSA (COmprehensive DEscriptors for Structural and Statistical Analysis), and HQSAR (Hologram QSAR)—for their utility (predictivity, speed, accuracy, and reproducibility) to quantitatively predict the ER binding activity [16,22]. For three relatively small datasets under investigation, the CoMFA and HQSAR procedures were of a comparable high performance [16]. These two procedures were further investigated and compared, particularly for their predictivity, by using the NCTR dataset and two other test sets [22]. We found that CoMFA performed better for the training set as well as for predicting two different test sets.

To develop a CoMFA model, the molecules of interest must first be aligned to maximize superposition of their steric and electrostatic fields in 3D. The critical and difficult aspect of the CoMFA procedure is choosing the most appropriate set of alignment rules for a structurally diverse training set. Fortunately, crystal structures of the ER complex of four ligands were published [24,25] that aided the derivation of rational CoMFA alignment rules. The CoMFA model based on the crystal structure-guided alignment is statistically robust. With the conventional $r^2 = 0.91$ (r: the correlation coefficient) and the cross-validated q^2 (leave-one-out) = 0.66, (q: the predictive correlation coefficient) indicating that it is both internally consistent and highly predictive.

Model validation

Concordance, specificity, and sensitivity [32] are commonly used to assess the quality of a classification model, while a quantitative regression model is assessed using r^2 [14]. The current challenge in QSARs is no longer in constructing a model that is statistically sound to predict the activity within the training set, but in developing a model with the capability to accurately predict the activity of untested chemicals. Most experts in the QSAR field, as well as the present authors, assert that a model's predictive capability at least needs to be demonstrated using some sort of cross-validation procedure. All models developed in our laboratory are validated using the leave-one-out technique for quantitative models and leave-10-out for qualitative models. It is worthwhile to point out that cross-validation methods only assess the interpolation within the training set, and have a limited ability to validate the prediction of untested chemicals that are structurally different from the training-set chemicals.

When additional data are available, the models are validated by predicting chemicals, not used in the training set but with the known activity data (the test set). The major difference between the cross-validation and external validation is that the chemicals selected in the latter case are in a sense of randomness, providing a more rigorous evaluation of the model's predictive capability for untested chemicals. We strongly believe that the confidence in a model's predictive capability can be tested and validated when predictability has been demonstrated on an external test set.

It is usual that the QSAR training set barely contains enough chemicals to develop a statistically robust model such that setting aside chemicals for use in external validation (10~20 % of the dataset is recommended) is not possible. One approach for selecting a test set is to identify datasets in the literature with the same type of activity. In such cases, care must be taken to avoid interlaboratory and assay variability among different data sources. It is desirable that the potential test set has activity data measured in an assay protocol as similar as possible to that for the training set to keep the variability at minimum. Another important consideration in selecting the test set is to ensure that chemicals in the dataset relate to the real problem in question. Based on these considerations, the datasets reported by Kuiper et al. [33], Waller et al. [10], and Nishihara et al. [34] were selected as test sets. In Kuiper's study, the pure human ERα was used, whereas the mouse uterine cytosol that primarily contains ERα was used in Waller's data. The Nishihara dataset contains 517 chemicals tested with the yeast two-hybrid assay, of which over 86 % are pesticides and industrial chemicals.

NCTR "FOUR-PHASE" SYSTEM AS A PRIORITY SETTING TOOL FOR REGULATORY APPLICATION

The objective of priority setting is to rank order, from most to least importance, a large number of chemicals to undergo more resource-intensive experimental evaluations. The strategy of QSAR is highly dependent on the application domain and goals. The QSAR used in drug discovery is to increase the chance of finding active chemicals or "hits" that may become "lead" chemicals, and false positives are of great concern. In contrast, a good priority setting method for the regulatory application should have minimal false negatives. False negatives constitute a crucial error, because they will be assigned a lower priority for subsequent evaluation. In addition, the methods should provide reasonable quantitative accuracy for true positives, as those with higher affinities will generally be of higher priority.

The QSAR procedures described in the previous section have strengths as well as weaknesses, and they all have a degree of prediction error. All procedures and particularly those that only provide active/inactive predictions can be optimized to minimize either the overall prediction error or the false negative or positive rate. Decreasing false negatives is achieved at a cost of increasing false positives and vice versa. Because selecting a single procedure is problematic, we adopted an approach of rationally combining different QSAR procedures into a sequential "Four-Phase" scheme according to the strength of each type of procedures. A progressive phase paradigm is used to screen out chemicals and thus reduce the number of chemicals to be considered in each subsequent phase. The four phases work in a hierarchical manner, incrementally reducing the size of the dataset while increasing precision of the prediction during each phase. Within each phase, different models are selected to work complementarily in representing key activity-determining structure features in order to minimize the rate of false negatives. The overall architecture of the NCTR "Four-Phase" system for identification of ER ligands is illustrated below and in Fig. 5.

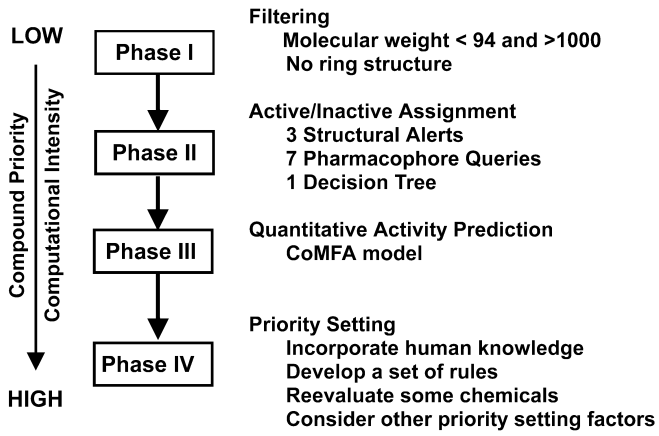

Fig. 5 NCTR "Four-Phase" system as a priority-setting tool.

Phase I: Filtering – Two filters, the molecular weight range and ring-structure indicator described above, were selected to efficiently eliminate chemicals very unlikely to have ER binding activity. As shown in Table 1, the two rejection filters correctly eliminated 6 and 98 inactive chemicals from the NCTR and Nishihara datasets, respectively. No false negative was introduced with these two rejection filters. For the Nishihara dataset, the data size was reduced by some 21 %. This suggests that, for real-world applications, the rejection filters significantly reduce the number of chemicals for further evaluation with minimum risk of introducing false negatives.

Table 1 Results of two rejection filters for the NCTR, Nishihara and Walker datasets*.

Datasets		Eliminated by				Number (%)
		MW range				of eliminated
		MW		Ring		
Sets	Size	Active	Inactive	Active	Inactive	chemicals
NCTR	232	0	0	0	6	6 (2.6 %)
Nishihara	463	0	28	0	89	98 (21.2 %)
Walker	558 230	16 048		1495		16 689 (28.7 %)

*Table lists the number of chemicals eliminated by either MW range or lack of ring criteria as well as their combination. No active chemical was rejected by these two filters.

Phase II: Active/Inactive Assignment – This phase categorizes chemicals from Phase I as either active or inactive. The three structural alerts, seven pharmacophore queries, and the Decision Tree classification model (11 models in total) discussed above were used in parallel to discriminate between active and inactive chemicals. To ensure the lowest possible false negative rate in Phase II, a chemical predicted to be active by any of the 11 models is presumed active and subsequently evaluated in Phase III, while only those predicted to be inactive by all these models are deemed inactive and eliminated from further evaluation. Since structural alert, pharmacophore, and Decision Tree methods incorporate and weigh differently the various structural features that endow a chemical with the ability to bind the ER, the combined outputs derived from the three approaches are complementary in minimizing false negatives.

All active chemicals in the NCTR, Waller [10], Kuiper [33], and Nishihara [35] datasets were identified by combining the 11 models.

Phase III: Quantitative Predictions – In Phase III, the CoMFA model described above [22] was used to make a quantitative activity prediction for chemicals categorized as active in Phase II.

Chemicals with higher predicted binding affinity are ranked at higher priority for further evaluation in Phase IV.

Phase IV: Rule-Based Decision-Making – In this final stage of the integrated system, we believe that a set of rules needs to be developed as a knowledge-based or expert system to make a priority setting decision. The system is useful only after incorporating accumulated human knowledge and expertise (i.e., rules). This system can make decisions on individual chemicals based on the rules in its knowledge base. Computational chemists, toxicologists, and regulatory reviewers should jointly develop and define the rules. The following are suggestions for such rules:

1. Special attention needs to be placed on the following chemicals, which may need to be reevaluated by assaying or modeling according to the "IF-THEN" scheme depicted in Fig. 6 [20]:

    1) The chemical is predicted to be inactive, but its structure information has been modified during structural preprocessing, e.g., by "correcting" such chemicals as mixtures and organic salts into separate entities.
    2) The chemical whose structure is dissimilar to all those that have been used to train and test the models.
    3) The chemical is active in Phase II, but inactive in Phase III.

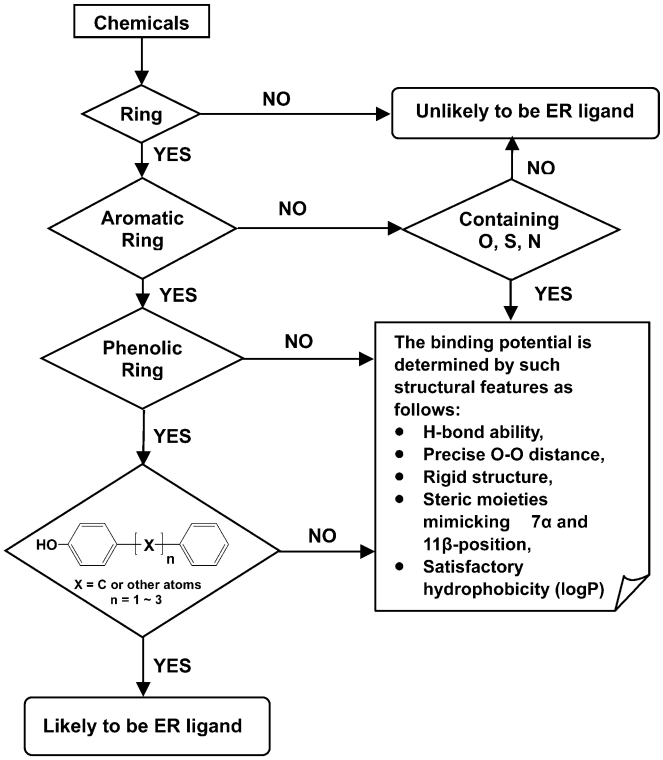

Fig. 6 Identification of ER ligands using a set of "IF-THEN" rules: (a) IF non-cyclic structure, THEN it is unlikely to be an ER ligand; (b) IF a chemical has non-aromatic ring structure, THEN unlikely to be an ER ligand if does not contain an O, S, N or other heteroatoms for H-bonding. Otherwise, binding potential is dependent on the occurrence of key structural features; (c) IF chemical has a non-OH aromatic structure, THEN binding potential is dependent on the occurrence of the key structural features; and (d) IF chemical contains a phenolic ring, THEN it tends to be an ER ligand if it contains any additional key structural features. For the chemicals containing a phenolic ring separated from another benzene ring with the bridge atoms ranging from none to three, it will most likely be an ER ligand.

2. Information on the level of human exposure and production, environmental fate, and other public health related parameters can be used independently or jointly incorporated for priority setting.

The NCTR "Four-Phase" system has been validated using a number of existing datasets, including the E-SCREEN assay data [36], the yeast two-hybrid reporter-gene assay data [35], and other datasets [10,37–41]. The system has produced no false negative for these test datasets. The system was recently applied to two environmental datasets recognized by EPA as important chemical subsets of potential EDCs:

***1. HPV-Inerts dataset*,** containing 623 high-production-volume inerts (HPV-Inerts), which is a subset of the Toxic Substances Control Act (TSCA) Inventory. The EPA is including HPV-Inerts in version 2 of the Endocrine Disruption Priority Setting Database (EDPSD2), and there was a need to prioritize HPV-Inerts for further experimental evaluation [42]. Of 623 chemicals, 166 chemicals were either mixtures or their structures were not available, leaving 457 chemicals for prediction.

2. Walker dataset [43], containing a large and diverse collection of known pesticides and industrial chemicals as well as some food additives and drugs. The database contains 92 964 chemicals, of which the Chemical Abstract Service (CAS) Registry number is available, that will probably have to be evaluated for their potential endocrine disruption. After eliminating chemicals for which structures were not available [43] and/or 3D structures could not be generated [32], a final dataset of 58 391 chemicals were predicted.

Table 2 summarizes the priority setting results for these two datasets using the NCTR "Four-Phase" system. Even when only the Phase I and II protocols are used, the system dramatically reduces the number of potential estrogens by some 80~85 %, demonstrating effectiveness in eliminating the most unlikely ER binders from further expensive experimentation. The Phase III CoMFA model further reduces the data size by about 5~10 %. Importantly, the quantitative binding affinity prediction from Phase III also provides an important rank-ordering value for priority setting.

Table 2 Size reduction of two environmental datasets processed by the NCTR "Four-Phase" system.

	HPV-Inerts	Walker
Original data size	457	58 391
After Phases I and II	15.7 %	12.0 %
After Phases III	9.8 %	–

QSAR APPLICATION IN PERSPECTIVE

The QSAR procedures are applied extensively to a wide range of scientific disciplines including chemistry, biology, and toxicology [44,45]. In both drug discovery and environmental toxicology [46], QSAR models are now regarded as a scientifically credible tool for predicting and classifying the biological activities of untested chemicals. However, analogous to misinterpretation of experimental results, QSAR results can also be misleading when limitations of the procedures are not well understood. Presented below are our experientially based suggestions of the salient aspects in applying QSARs for toxicology and regulation.

1. **Limitation in the fundamental principle of the QSAR** – Any QSAR model will produce some degree of error. This is partially due to the inherent limitation to predict a biological activity solely based on the chemical structure. One can argue from the principles of chemistry that the molecular structure of a chemical is key to understanding its physicochemical properties and ultimately its biological activity and the influence on organisms. Since both molecular structure and physico-

chemical properties are associated with the chemical itself, the relationship between structure and physicochemical properties should be apparent and, therefore, more accessible using the QSARs. In contrast, the biological activity of a chemical is an induced response that is influenced by numerous factors dictated by the levels of biological complexity of the system under investigation. The relationship between structure and activity is thus more implicit and thereby poses a more challenging problem in QSAR applications.

2. **Limitation in extrapolation of QSAR models** – A chemical can be represented in three distinct, but also related, structural representations such as 2D substructures, 3D pharmacophores, and physicochemical properties. If a biological mechanism is mainly related to the chemical structure (probably in the case for receptor binding), QSARs become meaningful using the aforementioned structural features. However, we often find that, even for a simple mechanism such as ER-binding, some features may well represent binding dependencies for one structural class, whereas other features will better represent binding dependencies for a different structural class [20]. In such cases, caution is warranted in interpreting QSAR results for the chemical classes that are not well represented in the training set. In other words, no matter how rigorous is the validation procedure used, the model may give incorrect predictions for some chemicals since the entire chemistry space of active chemicals is unknown. This realization is especially important when a QSAR model is used for regulatory application since it could lead to false negatives.
3. **Limitation in processing chemical structure** – Most computational chemistry programs accept only discrete organic chemicals for the descriptor calculation and QSAR modeling. Unfortunately, most toxicological databases contain chemicals that are not necessarily discrete organic chemicals, but sometimes are mixtures. Thus, it is necessary to process molecular structures of a toxicological dataset by "correcting" chemicals with separate entities (e.g., mixtures, organic salts, the presence of H_2O and HCl, etc.), and eliminating those chemicals whose descriptors cannot be calculated. Such procedures might lead to a prediction that might not reflect the real activity of the preprocessed chemical.
4. **Limitation in quality and transformation of biological data** – Predictions from any model are intrinsically no better than the experimental data employed to calibrate the model. Any limitations of the assay used to generate the training data apply equally to the model's predictions. Furthermore, additional limitations could be introduced by transforming the biological data for use in the QSAR modeling. When developing quantitative models, inactive chemicals are usually either not used because their exact values cannot be determined or included by assigning an arbitrary value. Either case will tend to decrease the model sensitivity in prediction of chemicals with low activity. For qualitative models, false negatives and false positives depend on the defined cut-off value to distinguish active from inactive. As the cut-off value is lowered, it is likely that the error will increase even for a well-designed and executed assay. The increased experimental error in close proximity to the cut-off value will be transferred to the QSAR model, which in turn will increase false prediction for chemicals with activity in this region. For example, as defined by EPA, the cut-off log RBA value to distinguish ER binders from non-ER binders is set to –3.0. This cut-off was used to develop the models in the NCTR "Four-Phase" system. There are 31 chemicals exhibiting the binding affinity in our assay that were assigned to be inactive because their RBA is below this cut-off value. Among these 31, 14 chemicals have the RBA within 0.3 log units of the cut-off value ("within 0.5 log units" is expected for a good assay). Since the real activity classification for these chemicals are unknown, assigning them arbitrarily as inactive to train QSAR models introduces errors in prediction for chemicals with similar structures. Thus, caution must be taken in interpreting QSAR prediction results for these chemicals. Apparently, a high confidence in the prediction can only be expected for chemicals with the log RBA more positive than –2.5.

While the results presented in this article clearly show both the feasibility and utility of using QSARs for various applications, it is important to realize that any QSAR model has to be considered as a *living* model that will be improved whenever new data is available. In other words, the model development should be a recursive process that alternates between incorporating new data in the model and using the model to choose new chemicals for assay [47,48]. As depicted in Fig. 7, the process starts with an initial set of chemicals from the literature for QSAR modeling [14–16]. Next, the preliminary QSAR models are used prospectively to define and rationalize a set of chemicals that may further improve the model's robustness and predictive capability. These new chemicals are assayed, and the data are then used to challenge and refine the QSAR model. Thus, the process results in the iterative increase of the chemistry space of the training set.

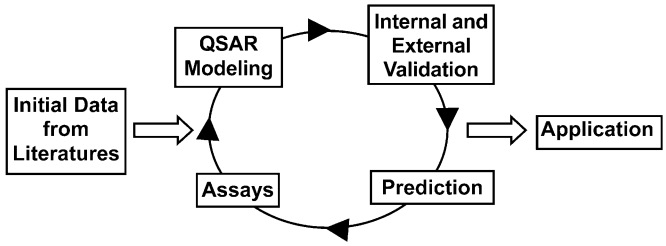

Fig. 7 The recursive process to develop QSAR models.

Several benefits accrue from the integration of the experimental and modeling efforts. Immediate feedback can be given to the experimentalists so that suspected assay problems can be rapidly investigated. Also, as the models evolve, the modelers can select the chemicals for subsequent testing, based on considerations of structural diversity and activity range. Each new assay data point coming from the lab becomes a challenge to the evolving model; the result is either further confirmation of its validity, identification of a limitation, or an outlier prediction. Failure of the model also provides important information, such as identification of the need for new data based on a rational understanding of the dependence of activity on structure. Regardless of the cause of model failure, a research hypothesis is spawned during each iteration process that should lead to new data and/or an improved training set, and, ultimately, an improvement to the living model.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the U.S. Environmental Protection Agency (EPA) and the American Chemistry Council for financial support.

REFERENCES

1. B. Hileman. *Chem. Eng. News* **72**, 19–23 (1994).
2. B. Hileman. *Chem. Eng. News* **75**, 24–25 (1997).
3. R. J. Kavlock, G. P. Daston, C. DeRosa, P. Fenner-Crisp, L. E. Gray, S. Kaattari, G. Lucier, M. Luster, M. J. Mac, C. Maczka, R. Miller, J. Moore, R. Rolland, G. Scott, D. M. Sheehan, T. Sinks, H. A. Tilson. *Environ. Health Perspect.* **104** (Suppl. 4), 715–740 (1996).
4. US-Congress, The Food Quality Protection Act (FQPA) and the Safe Drinking Water Act (SDWA) (1996).
5. L. E. Gray, Jr. *Toxicol. Lett.* **102–103**, 677–680 (1998).
6. M. Patlak. *Environ. Sci. Technol.* **30**, 540A–544A (1996).
7. H. Fang, W. Tong, R. Perkins, A. Soto, N. Prechtl, D. M. Sheehan. *Environ. Health Perspect.* **108**, 723–729 (2000).

8. T. Zacharewski. *Environ. Health Perspect.* **106** (Suppl. 2), 577–582 (1998).
9. S. Bradbury, O. Mekenyan, A. GT. *Environ. Toxicol. Chem.* **15**, 1945–1954 (1996).
10. C. L. Waller, T. I. Oprea, K. Chae, H. K. Park, K. S. Korach, S. C. Laws, T. E. Wiese, W. R. Kelce, L. E. Gray, Jr. *Chem. Res. Toxicol.* **9**, 1240–1248 (1996).
11. T. E. Wiese, L. A. Polin, E. Palomino, S. C. Brooks. *J. Med. Chem.* **40**, 3659–3669 (1997).
12. B. R. Sadler, S. J. Cho, K. S. Ishaq, K. Chae, K. S. Korach. *J. Med. Chem.* **41**, 2261–2267 (1998).
13. W. Zheng and A. Tropsha. *J. Chem. Inf. Comput. Sci.* **40**, 185–194 (2000).
14. W. Tong, R. Perkins, R. Strelitz, E. R. Collantes, S. Keenan, W. J. Welsh, W. S. Branham, D. M. Sheehan. *Environ. Health Perspect.* **105**, 1116–1124 (1997).
15. W. Tong, R. Perkins, L. Xing, W. J. Welsh, D. M. Sheehan. *Endocrin.* **138**, 4022–4025 (1997).
16. W. Tong, D. R. Lowis, R. Perkins, Y. Chen, W. J. Welsh, D. W. Goddette, T. W. Heritage, D. M. Sheehan. *J. Chem. Inf. Comput. Sci.* **38**, 669–677 (1998).
17. L. Xing, W. J. Welsh, W. Tong, R. Perkins, D. M. Sheehan. *SAR QSAR Environ. Res.* **10**, 215–237 (1999).
18. R. Blair, H. Fang, W. S. Branham, B. Hass, S. L. Dial, C. L. Moland, W. Tong, L. Shi, R. Perkins, D. M. Sheehan. *Toxicol. Sci.* **54**, 138–153 (2000).
19. W. S. Branham, S. L. Dial, C. L. Moland, B. Hass, R. Blair, H. Fang, L. Shi, W. Tong, R. Perkins, D. M. Sheehan. *J. Nutr.* **132**, 658–664 (2002).
20. H. Fang, W. Tong, L. Shi, R. Blair, R. Perkins, W. S. Branham, S. L. Dial, C. L. Moland, D. M. Sheehan. *Chem. Res. Toxicol.* **14**, 280–294 (2001).
21. G. M. Anstead, K. E. Carlson, J. A. Katzenellenbogen. *Steroids* **62**, 268–303 (1997).
22. L. M. Shi, H. Fang, W. Tong, J. Wu, R. Perkins, R. Blair, W. Branham, D. Sheehan. *J. Chem. Inf. Comput. Sci.* **41**, 186–195 (2001).
23. H. Hong, N. Neamati, S. Wang, M. C. Nicklaus, A. Mazumder, H. Zhao, T. R. Burke, Y. Pommier, G. W. A. Milne. *J. Med. Chem.* **40**, 930–936 (1997).
24. A. M. Brzozowski, A. C. Pike, Z. Dauter, R. E. Hubbard, T. Bonn, O. Engstrom, L. Ohman, G. L. Greene, J. A. Gustafsson, M. Carlquist. *Nature* **389**, 753–758 (1997).
25. A. K. Shiau, D. Barstad, P. M. Loria, L. Cheng, P. J. Kushner, D. A. Agard, G. L. Greene. *Cell* **95**, 927–937 (1998).
26. A. Smellie, S. D. Kahn, S. Teig. *J. Chem. Inf. Comput. Sci.* **35**, 285–294 (1995).
27. A. Smellie, S. D. Kahn, S. Teig. *J. Chem. Inf. Comput. Sci.* **35**, 295–304 (1995).
28. A. Smellie, S. L. Teig, P. Towbin. *J. Comput. Chem.* **16**, 171–187 (1995).
29. L. M. Shi, W. Tong, H. Fang, R. Perkins, J. Wu, M. Tu, R. Blair, W. Branham, J. Walker, C. Waller, D. Sheehan. *SAR QSAR Environ. Res.* **13**, 69–88 (2002).
30. D. E. Clark and D. R. Westhead. *J. Comput. Aided Mol. Des.* **10**, 337–358 (1996).
31. S. Forrest. *Science* **261**, 872–878 (1993).
32. H. Hong, W. Tong, H. Fang, L. M. Shi, Q. Xie, J. Wu, R. Perkins, J. Walker, W. Branham, D. Sheehan. *Environ. Health Perspect.* **110**, 29–36 (2002).
33. G. G. Kuiper, J. G. Lemmen, B. Carlsson, J. C. Corton, S. H. Safe, P. T. van der Saag, B. van der Burg, J. A. Gustafsson. *Endocrin.* **139**, 4252–4263 (1998).
34. T. Nishihara, J. Nishikawa, T. Kanayama, F. Dakeyama, K. Saito, M. Imagawa, S. Takatori, Y. Kitagawa, S. Hori, H. Utsumi. *J. Health Sci.* **46**, 282–298 (2000).
35. T. Nishirara, J. Nishikawa, T. Kanayama, F. Dakeyama, K. Saito, S. Imagawa, S. Takatori, Y. Kitagawa, S. Hori, H. Utsumi. *J. Health Sci.* **46**, 282–298 (2000).
36. A. M. Soto, C. Sonnenschein, K. L. Chung, M. F. Fernandez, N. Olea, F. O. Serrano. *Environ. Health Perspect.* **103** (Suppl. 7), 113–122 (1995).
37. K. W. Gaido, L. S. Leonard, S. Lovell, J. C. Gould, D. Babai, C. J. Portier, D. P. McDonnell, *Toxicol. Appl. Pharmacol.* **143**, 205–212 (1997).
38. N. G. Coldham, M. Dave, S. Sivapathasundaram, D. P. McDonnell, C. Connor, M. J. Sauer. *Environ. Health Perspect.* **105**, 734–742 (1997).

39. E. J. Routledge and J. P. Sumpter. *J. Biol. Chem.* **272**, 3280–3288 (1997).
40. E. J. Routledge, J. Parker, J. Odum, J. Ashby, J. P. Sumpter. *Toxicol. Appl. Pharmacol.* **153**, 12–19 (1998).
41. C. A. Harris, P. Henttu, M. G. Parker, J. P. Sumpter. *Environ. Health Perspect.* **105**, 802–811 (1997).
42. J. D. Walker, H. Fang, R. Perkins, W. Tong. *QSAR Comb. Sci*. **22**, 89–105 (2003).
43. J. D. Walker, C. W. Waller, S. Kane. "The Endocrine Disruption Priority Setting Database (EDPSD): A Tool to Rapidly Sort and Prioritize Chemicals for Endocrine Disruption Screening and Testing". In: *Handbook on Quantitative Structure Activity Relationships (QSARs) for Predicting Chemical Endocrine Disruption Potentials*, J. D. Walker (Ed.), SETAC Press, Pensacola, FL (2001).
44. C. Hansch and A. Leo. *Exploring QSAR - Fundamentals and applications in chemistry and biology*, American Chemical Society, Washington, DC (1995).
45. C. Hansch, B. R. Telzer, L. Zhang. *Crit. Rev. Toxicol.* **25**, 67–89 (1995).
46. S. Bradbury. *Toxicol. Lett.* **79**, 229–237 (1995).
47. W. Tong, W. J. Welsh, L. Shi, H. Fang, R. Perkins. *Environ. Toxicol. Chem*. **22**, 1666–1679 (2003).
48. R. Perkins, H. Fang, W. Tong, W. J. Welsh. *Environ. Toxicol. Chem.* **22**, 1680–1695 (2003).