

AN INTEGRATED “4-PHASE” APPROACH FOR SETTING ENDOCRINE DISRUPTION SCREENING PRIORITIES— PHASE I AND II PREDICTIONS OF ESTROGEN RECEPTOR BINDING AFFINITY*

L. SHI[‡]^a, W. TONG[†]^a, H. FANG^a, Q. XIE^a, H. HONG^a, R. PERKINS^a, J. WU^a, M. TU[§]^a,
R.M BLAIR^b, W.S BRANHAM^b, C. WALLER^c, J. WALKER^d and D.M. SHEEHAN^b

^aR.O.W. Sciences, Inc., 3900 NCTR Road, MC910, Jefferson, AR 72079, USA; ^bDivision of Genetic and Reproductive Toxicology, National Center for Toxicological Research (NCTR), Jefferson, AR 72079, USA; ^cSphinx Pharmaceuticals, A Division of Eli Lilly and Company, Research Triangle Park, NC 27709, USA; ^dTSCA Interagency Testing Committee (ITC), US EPA (7401), Washington, DC 20460, USA

(Received 19 September 2000; In final form 26 July 2001)

Recent legislation mandates the US Environmental Protection Agency (EPA) to develop a screening and testing program for potential endocrine disrupting chemicals (EDCs), of which xenoestrogens figure prominently. Under the legislation, a large number of chemicals will undergo various *in vitro* and *in vivo* assays for their potential estrogenicity, as well as other hormonal activities. There is a crucial need for priority setting before this strategy can be effectively implemented. Here we report an integrated computational approach to priority setting using estrogen receptor (ER) binding as an example. This approach rationally integrates different predictive computational models into a “Four-Phase” scheme so that it can effectively identify potential estrogenic EDCs based on their predicted ER relative binding affinity (RBA). The system has been validated using an in-house ER binding assay dataset for 232 chemicals that was designed to have both broad structural diversity and a wide range of binding affinities. When applied to 58,000 chemicals identified by Walker *et al.* as candidates for endocrine disruption screening, some 9100 chemicals were predicted to bind to ER. Of these, only 3600 were expected to bind to ER at RBA values up to 100,000-fold less than that of 17 β -estradiol. The method ruled out 83% of the chemicals as non-binders with a very low rate of false negatives. We believe that the same integrated scheme will be equally applicable to endpoints of other endocrine disrupting mechanisms, e.g. androgen receptor binding.

Keywords: Endocrine disrupting chemicals; Estrogens; Priority setting; Estrogen receptor binding; QSAR

Abbreviations: Endocrine Disrupting Chemicals, EDCs; Estrogen Receptor, ER; Relative Binding Affinity, RBA; Endocrine Disruptor Screening and Testing Advisory Committee, EDSTAC; US Environmental Protection Agency, EPA; US Food and Drug Administration, FDA; National Center for Toxicological Research, NCTR; Quantitative Structure–Activity Relationship, QSAR; K-Nearest Neighbors, KNN; Classification and Regression Tree, CART; Diethylstilbestrol, DES; Comprehensive Descriptors for Structural and Statistical Analysis, CODESSA; Comparative Molecular Field Analysis, CoMFA; Tier 1 Screening, T1S; Tier 2 Testing, T2T; High Throughput Pre-Screening, HTPS; Hologram QSAR, HQSAR; Hydrogen Bond Donor, H-Donor

*Presented at the 9th International Workshop on Quantitative Structure–Activity Relationships in Environmental Sciences (QSAR 2000), September 16–20, 2000, Bourgas, Bulgaria.

[†]Current address: American Cyanamid Co., American Home Products, Princeton, NJ 08543, USA.

[‡]Corresponding author. E-mail: wtong@nctr.fda.gov

[§]Current address: Pfizer Inc., Groton, CT 06340, USA.

INTRODUCTION

Reminiscent of the environmental movement spawned in the 1960s regarding chemicals in the environment causing human cancers, the potential for endocrine disrupting chemicals (EDCs) to cause a broad range of adverse effects has created concern, if not alarm, among the public and governments worldwide [1]. Adverse effects such as compromised reproductive fitness, learning disabilities, cancer and immune disorders have been reported widely in the popular press [2]. The resulting public concern has led to government regulatory actions [3,4] and expanded research across Europe, Japan and North America. EDCs are chemicals that may mimic endogenous hormones, alter their pharmacokinetics or mechanisms of action among other possibilities. The scientific debate has escalated, fueled in part by the fact that some suspected EDCs are high-volume, economically important chemicals.

The US Congress passed laws that resulted in the EPA developing and implementing a strategy for screening and testing chemicals for estrogen, androgen and thyroid endpoints [5]. A two-tiered, multiple-endpoint strategy, which incorporates more than 20 different *in vitro* and *in vivo* assays [6], was recommended by EPA's Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC). As many as 87,000 chemicals may need to be screened for endocrine-disruption potential [7]. The large number of chemicals and assays makes it difficult for each chemical to be run through these assay batteries in a reasonable time. There is a crucial need for priority setting to identify the chemicals most likely to possess endocrine disrupting activity for early entry into screening.

Priority setting using computational approaches is widely applied in the process of drug discovery. The objective of priority setting in pharmaceutical industry is to increase the chance of finding active compounds or "hits" that are more likely to be developed into "leads". Hence, false positives are of great concern. In contrast, minimizing false negatives is critical for regulatory purpose because chemicals labeled as inactive are dropped into a lower priority category. For this purpose, we developed an integrated computational system that rationally combines different computational models into a sequential "Four-Phase" scheme according to the strength of each type of model (Fig. 1). In Phase I, several simple rejection filters or rules are used to exclude those chemicals that are most unlikely to exhibit estrogenic

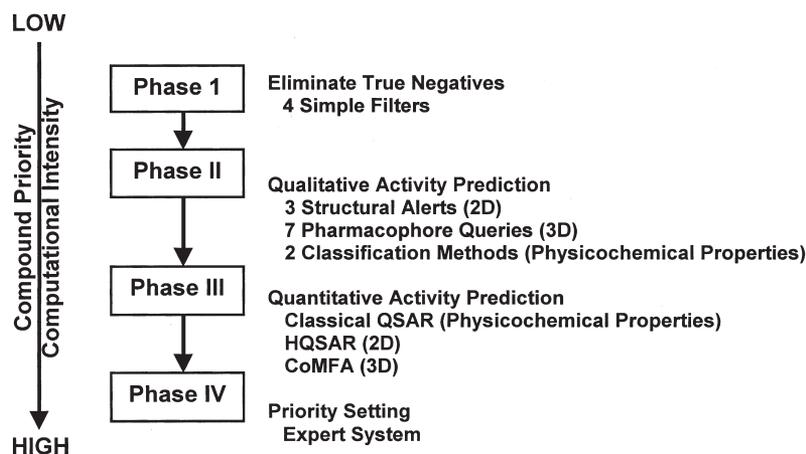


FIGURE 1 Overview diagram of the "Four-Phase" approach for priority setting. Different phases are hierarchical; different methods within each phase are complementary.

activity. Phase II uses three different types of models (structural alerts, pharmacophores and classification methods) to make a qualitative activity predictions. In Phase III, multiple quantitative structure–activity relationship (QSAR) models are used quantitatively to predict activity. In Phase IV, an expert system is recommended to combine Phase II and Phase III predictions with exposure, fate and other data to set priorities. In this scheme, each Phase is used as a screen to reduce the number of compounds to be considered in the subsequent Phase. Therefore, these four Phases work in a hierarchical way to reduce the size of a dataset incrementally while with increasing precision of predictions simultaneously. Within each Phase, different complimentary models have been selected to represent key activity-determining structure features and to minimize the rate of false negatives.

Previously, we have evaluated the performance of a number of QSAR models in Phase III [8,9]. This paper reports results from Phase I and II of the integrated approach for 232 chemicals (the NCTR dataset) assayed by a validated ER competitive binding assay [10] and a subset of ~58,000 chemicals identified by Walker *et al.* [11] from a total list of 87,000 chemicals.

MATERIALS AND METHODS

Datasets

A computational model is generally trained and validated first based on a reasonable amount of reliable data; it then can be expected to provide reliable predictions on new chemicals. To build robust and predictive computational models, it is important to have a reliable training set of chemicals with known biological activity. Because literature data are not sufficient for the purpose, an in-house rat ER binding assay was established to provide data for the model development [10]. Chemicals in the dataset were selected to reflect the structural diversity of EDCs and the distribution of biological activity needed for building robust models. This selection process has been a highly interactive one, involving computational chemists and experimental toxicologists, and has resulted in the steady improvement in performance of our models [12]. Our current dataset consists of 129 active and 103 inactive compounds [10]. This data set, called the NCTR dataset, has been extensively used to build and validate a series of computational models proposed for priority setting. The distribution of the binding activity (calculated as RBA) and chemical classes of the dataset are shown in Figs. 2 and 3, respectively. A chemical with RBA value smaller than a million-fold below 17 β -estradiol is defined as inactive ($RBA < 10^{-4}$), where the RBA value for 17 β -estradiol is set to 100.

As noted by Walker *et al.* [11], about 87,000 chemicals may be screened for potential endocrine disruption. The Walker *et al.* dataset [11] contains a large and diverse collection of known environmental chemicals as well as some food additives and drugs, of which some 8000 chemicals are regulated by the US Food and Drug Administration (FDA). Among the 87,000 chemicals, 57,810 were discrete organic chemicals with specific chemical structure. The molecular structures of these 57,810 chemicals were preprocessed according to the following criteria [13].

1. The records are valid, i.e. they contain the connection table fields and there are no obvious errors in the structure description.
2. Counterions and solvent molecules were removed in order to obtain single compound records.
3. Charges at acidic and basic groups are neutralized by adding or removing protons. This prevented structural differences caused by different protonation states, which might lead to differences in the calculation of the molecular descriptors.

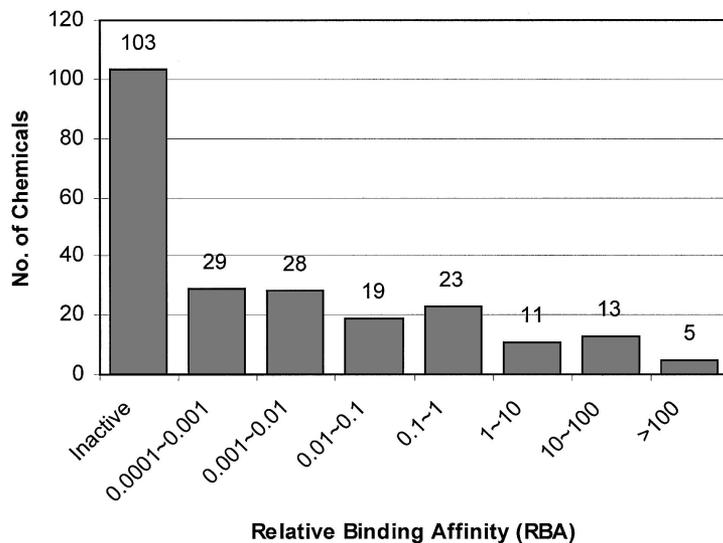


FIGURE 2 Experimental binding activity distribution of the NCTR dataset. The activity is represented as RBA (Relative Binding Affinity). The RBA for the endogenous ligand, 17 β -estradiol, was set to 100.

The comparison of structural diversity between the NCTR and Walker *et al.* data sets was shown in Fig. 4. The NCTR dataset chemicals covered a significant part of the chemistry space defined by the Walker *et al.* dataset. Most active chemicals cluster together, whereas inactive chemicals scattered over the space.

General Approaches and Methodological Considerations

The proposed integrated approach for priority setting is composed of four sequential phases (Fig. 1). Each phase contains a number of rules and/or models to estimate a compound's binding affinity. Briefly,

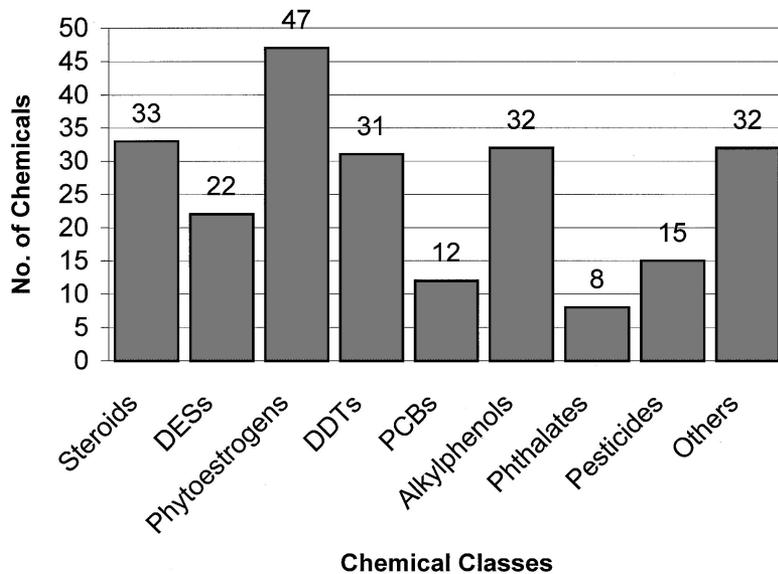


FIGURE 3 Chemical class distribution of the NCTR dataset.

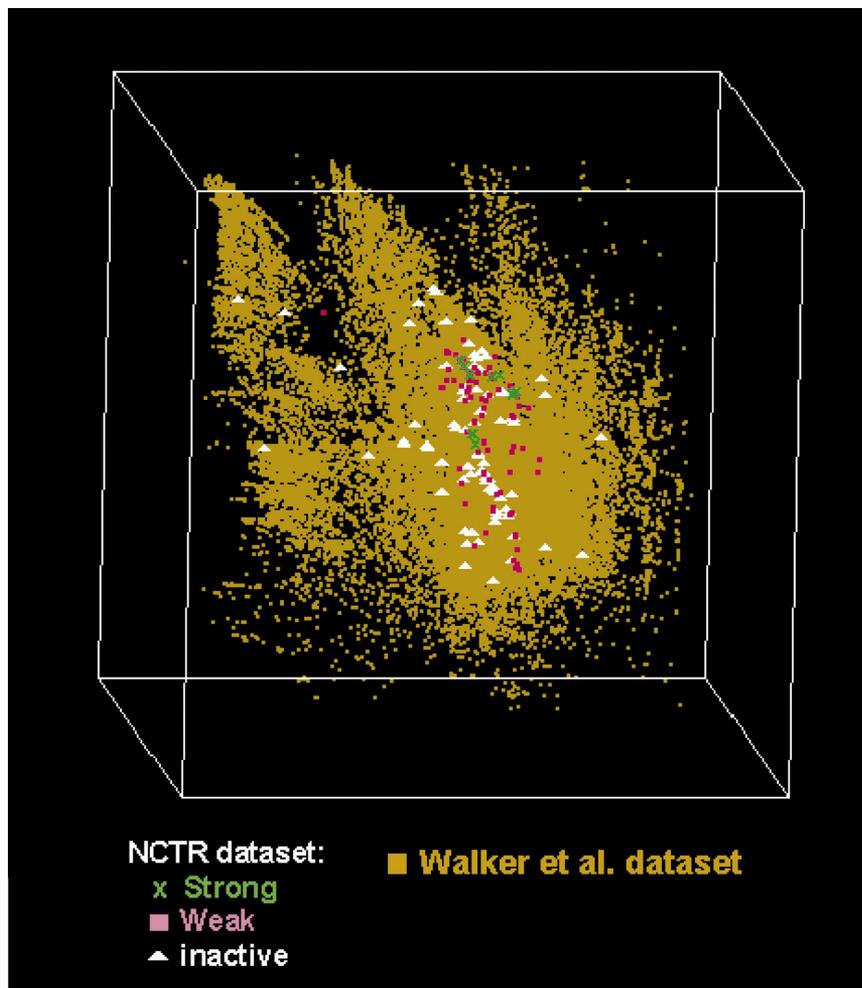


FIGURE 4 Chemistry space of the NCTR and Walker *et al.* Datasets based on BCUT descriptors.

Phase I: Filtering—A set of rejection filters are designed to significantly and with high confidence eliminate those chemicals with extremely low or no ER binding affinity. The key criterion in selecting these filters is to rule out as many true negatives as possible, while having an extremely low probability of passing a false negative.

Phase II: Active/Inactive Assignment—Three different methods, i.e. structural alert, pharmacophore searching and classification/clustering, are used in parallel in this phase to distinguish active from inactive compounds. Each method incorporates and weighs differently the various structural features that endow a chemical with the ability to bind the ER. In order to further reduce false negatives, a chemical predicted to be active by any of these methods is subsequently evaluated in the Phase III. At Phase II, the combined output derived from the three approaches could be used for initial categorical priority setting.

Phase III: Quantitative Predictions—A number of QSAR methods are used for quantitative prediction of the binding affinity of the compounds identified as active in Phase II [8,9]. Compounds with higher predicted binding affinity are given higher priority for further evaluation.

Phase IV: Rule-Based Decision-Making System—In this final stage of our integrated priority setting approach, we propose to use a rule-based (or knowledge-based) decision-making system to foster definitive decision making. This system will be useful only after we incorporate our accumulated human knowledge and expertise, i.e. rules, into the system.

The importance of the incorporation of multiple models is twofold. First, different techniques have different strengths and weaknesses in their ability to correlate and encode specific chemical structure features that endow a chemical with its activity. Multiple models enable the system to aggregately encode these features in a complementary manner. Second, some techniques are fast, but others are moderate to very time-consuming. Hence, hierarchical sequencing of the models allows faster models to be used to eliminate the majority of inactive chemicals with an extremely low rate of false negatives. Subsequently, the progressively more time-consuming but more precise models can be used to refine predictions for an increasingly smaller number of remaining chemicals. The application of the more refined models further eliminates true negatives as well as false positives from earlier models.

Phase I: Rejection Filters

Four simple rejection filters were used in Phase I. The first filter ruled out chemicals with molecular weight (M_w) < 94 or > 1000 . The M_w of phenol (94) was considered the lowest limit for a compound to bind ER, whereas a MW of 1000 was considered the upper limit of ER ligands, as suggested by the EDSTAC. The second filter was “number of rings = 0”, which implies that an active compound needs at least one ring of any type, e.g. aromatic, alicyclic, etc. This is based on the lack of known estrogens without a ring. The third filter was “number of carbons = 0”. It is believed that a chemical containing no carbon would be too lipophilic to cross a membrane, even though it might exhibit binding activity. The fourth filter ruled out inorganic compounds because their probability of being estrogenic was quite low. To our best knowledge, there are no known inorganics that exhibit endocrine disrupting potential via the mechanism of ER-binding.

Phase II: Structural Alerts, Pharmacophore Queries and Classification Models

Phase II comprises three types of models for qualitative activity prediction: structural alerts, pharmacophore searching and classification. The structural alerts were three 2D substructures that were identified as key 2D structural features for most estrogens. The pharmacophore search used seven different 3D queries based on known 3D structural features for ER binding. Two classification models were developed based on KNN and CART methods to qualitatively categorize compounds into active and inactive subsets on the basis of their similarity in physicochemical properties. These 12 models are complementary and were designed to distinguish active from inactive chemicals. A prediction of the active or inactive was first made by applying each of the 12 models to each chemical. The results of the individual model predictions were then combined, and only chemicals identified as inactive by all 12 models were eliminated from further evaluation in Phase III.

Structural Alerts

Figure 5 shows three structural alerts that were designed to identify as potentially active chemicals with any of these substructural features. The steroid skeleton alert was designed to make sure that all steroidal compounds, which include most endogenous hormones, would be passed to more advanced computational prediction models. The DES skeleton, two phenyl rings separated by two carbons bound with any bond-type, was used to make sure that

compounds similar to diethylstilbestrol (DES), one of the most active synthetic estrogens, would not be missed. The third structural alert was the phenolic ring. The precise overlapping of the A-rings of crystal structures for ER-estradiol, ER-raloxifene, ER-DES and ER-4-hydroxytamoxifen [14,15], as shown in Fig. 6, and our knowledge of the structural requirements for ER-ligand binding, all suggest the importance of the A-ring phenolic structural feature for ER-binding. Hence, compounds with steroid, DES, or phenolic skeletons will be classified as active and moved to Phase III for quantitative activity prediction, regardless of the results of pharmacophore searching and classification models.

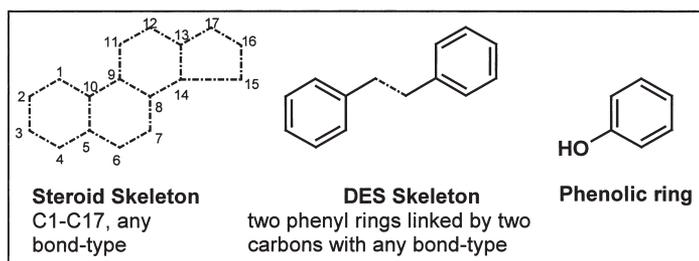


FIGURE 5 Structural alerts for priority setting. Compounds with any of these substructural features that are commonly seen in ligands to ER should be passed to Phase III (Multiple QSAR predictions) of the integrated priority setting approach. It is to ensure that compounds with these features will receive proper attention in priority setting.

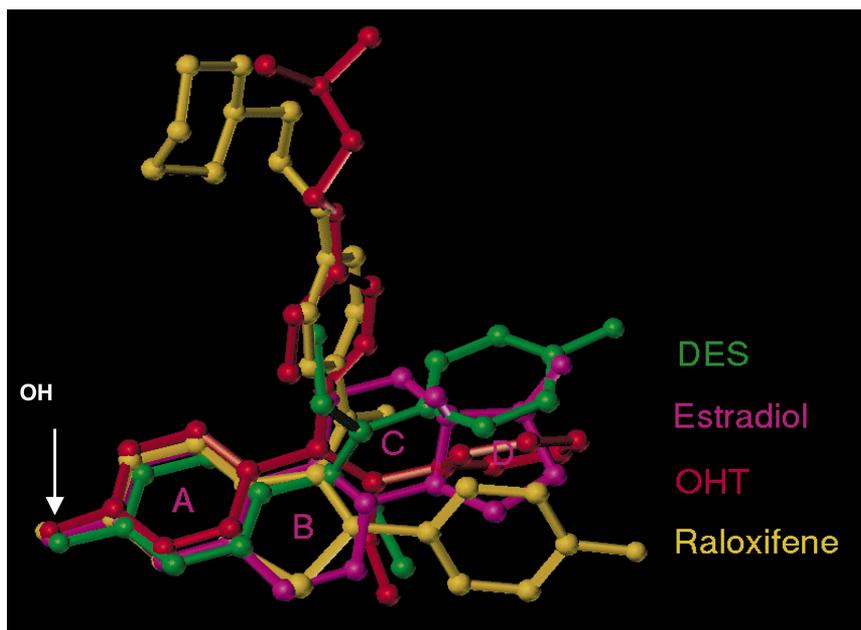


FIGURE 6 Superimposition of ligands bound to ER. The phenolic A-ring is very important for ER-binding.

Pharmacophore Query Construction and Database Searching

A pharmacophore is a combination of a few molecular features (e.g. H-donor, H-acceptor, hydrophobic centers and associative geometry) needed for a molecule to exhibit a certain type of biological activity [16]. A training set of molecules may be used to automatically generate a number of pharmacophore queries that specify the minimum requirement for

binding based on a predefined number of interactions between a receptor and its ligands. Alternatively, pharmacophore queries may be constructed manually [17], guided by a ligand–receptor crystal structure, as was done in the present study. In the latter case, one usually starts with template molecules that either are derived from bound receptor–ligand crystal structure or from the 3D structure of highly active chemicals. Molecular features are delineated from those templates and combined to form pharmacophore queries, where the 3D distance tolerance among these features can be adjusted for optimal performance. The construction and validation of the pharmacophores in the present study consisted of several steps [18]. First, the crystal structures of 17β -estradiol and raloxifen bound to ER were selected for the templates. Second, all possible pharmacophore elements (the molecular features) for the templates were identified using Feature Selection Function in CATALYST. Seven pharmacophore elements identified for 17β -estradiol are shown in Fig. 7, including two H-bond donor/acceptors sites (3- and 17β -hydroxyl groups), four hydrophobic centers (the centers of A–D rings) and shape constrain. At last, the pharmacophore queries were constructed based on any combination of 3–6 elements. We initially generated over 20 pharmacophore queries based on the knowledge of our careful SAR examination of a large number of chemicals for their binding affinities to ER [19] in conjunction with study of the recently reported ligand–ER crystal structures [14,15]. A Tanimoto similarity score was used to estimate the quality of the queries, and the correlation between two queries. A good query should have maximum discrimination power to separate active from inactive chemicals, and also provide unique contribution for prediction. On the basis of this consideration, seven queries were selected for final application (Fig. 8). The developed 3D queries were in turn used to search the Walker *et al.* dataset of 3D-chemical structures for “hits” that contain these queries. The 3D structures of chemicals were prepared using catDB command in CATALYST to generate up to 100 conformations per chemical [20–22]. The hits were presumably active or estrogens.

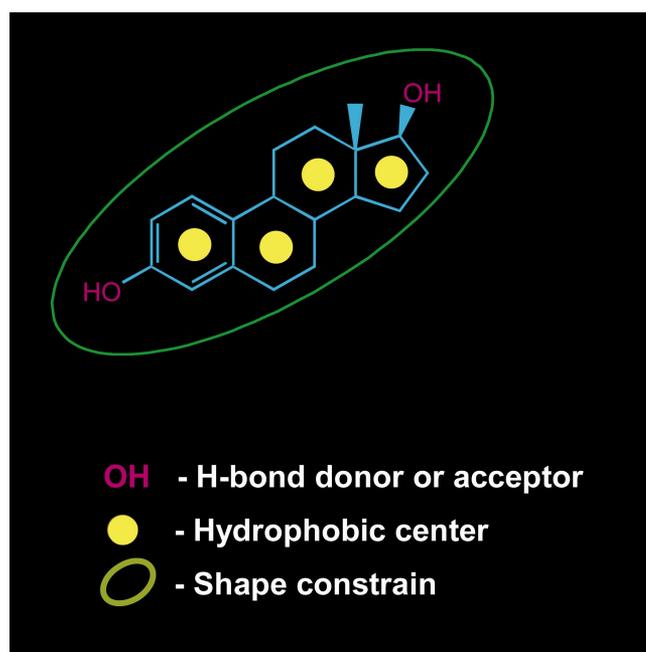


FIGURE 7 Structural features of 17β -estradiol used to construct pharmacophore queries.

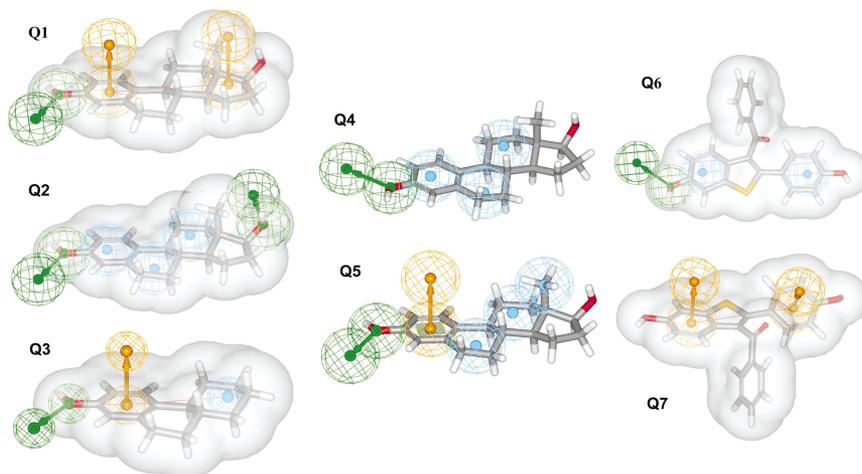


FIGURE 8 Seven pharmacophore queries. The green mesh balls represent H-bond acceptor sites. The blue mesh balls represent hydrophobic centers. The yellow mesh balls represent aromatic centers. The solid white surfaces represent shape constrain.

K-nearest Neighbors

K-nearest neighbors (KNN) is a widely used pattern recognition technique that can categorize an unknown chemical based on its proximity to samples already placed in categories [23]. Specifically, the predicted class, i.e. active or inactive in this study, of an unknown chemical depends on the distribution of class assignment of its *K* nearest neighbors in the training set, which accounts for the name of the technique. The nearness is generally measured by an Euclidean distance metric in an *N*-dimensional space of molecular descriptors, although other distance metrics can be applied. In a fashion analogous to polling, each of the *K* nearest training set samples votes once for its class; the unknown chemical is then assigned to the class with the most votes. With a chosen distance metric, the most important part of the KNN process is to determine an optimal *K* value for the final model development, which was selected by the leave-one-out cross validation in this study. The standard KNN process was implemented as follows: (1) remove a chemical from the dataset; (2) calculate the distance metric between the omitted chemical and all remaining chemicals in the dataset; (3) select *K* chemicals nearest (or similar) to the left-out chemical according to the calculated distances; (4) assign the left-out chemical the activity class to which a majority of the *K* chemicals belongs; (5) count the misclassification by comparing the predefined class with the predicted class of the left-out chemical; and (6) repeat steps (1)–(5) until each chemical in the dataset is left out once for prediction and the rate of misclassification is summarized for the predefined *K* value. Theoretically, the *K* value can vary from 1 to $N - 1$, where *N* is the size of the dataset. In our application, the steps (1)–(6) were repeated for each *K* value from 3 to 10, and the optimal *K* value, which was 3 for the final model, was determined according to the smallest rate of misclassification.

Classification and Regression Tree

Classification and regression tree (CART) uses a decision tree to determine how a chemical may be classified or predicted through a series of rules based on selection of variables (or descriptors). These rules are operated by using “if ... then ...” expressions. Since tree-construction methods are recursive in nature, CART is also called the recursive partitioning

method for pattern recognition in drug discovery [24]. Depending on the nature of activity data (endpoint), the tree can be constructed for either regression or classification. Each terminal node of the regression tree gives a quantitative prediction, while the classification tree gives a categorical prediction. The classification tree is used most commonly in data analysis, where the endpoint is usually binomial. In the present application, the tree method used to classify chemicals into active and inactive categories is described by Clark Pregibon [25] and implemented in the S-Plus software. The development of a tree model involves two processes: tree construction and pruning. In the tree construction process, a parent population is split into two children nodes that become parents for further splits. The split is selected to distinguish maximally the response variable in the left and the right branches. Splitting continues until nodes are pure or data are too sparse. To avoid over-fitting the training data, the tree needs to be cut down to the desired size using tree cost-complexity pruning. In this study, the number of the terminal nodes was set to 10, which corresponds to about 23 compounds per node.

General Computational Modeling

The Phase I filters and Phase II substructural search were done using ISIS Base (MDL Information System, Inc., San Leandro, CA). Log P was calculated using the atom/fragment contribution method [26]. The S-Plus software (MathSoft, Inc., Seattle, Washington) was used to develop KNN and CART models [27]. The pharmacophore searching was performed with the CATALYST package (Molecular Simulations Inc., San Diego, CA).

RESULTS

The feasibility of the “Four-Phase” approach for priority setting was tested on the NCTR and Walker *et al.* datasets. The NCTR dataset was used as the training set to develop all the models. These models were then integrated together to form the “Four-Phase” system for priority setting of the Walker *et al.* dataset. The results of the first two phases are summarized in Tables I and II. There were no true positives (of the 129 chemicals in the NCTR dataset) that were predicted to be non-binder (Table I). About 21% of the chemicals in the NCTR dataset were non-ER binders and no false negatives were introduced in either Phase I or II. As shown in Table II, a significant number of chemicals in Walker *et al.* dataset was eliminated after first two phases’ analysis. About 80% of chemicals of the Walker *et al.* dataset were non-ER binders.

Phase I: Filtering

Several simple rejection filters were applied to eliminate the chemicals with very high- or very low-molecular weight, or other characteristics that make a chemical unlikely to bind to the ER. Table I shows that Phase I filters correctly excluded seven inactive compounds, or

TABLE I Summary of Phase I and II results for the NCTR dataset

| Analysis | Active chemicals | Inactive chemicals | Predicted as non-binders |
|-----------------|------------------|--------------------|--------------------------|
| Before analysis | 129 | 103 | – |
| After Phase I | 129 | 96 | 7 |
| After Phase II | 129 | 54 | 49 |

TABLE II Reduction in number (percentage) of chemicals for potential endocrine disruption screening as a result of Phase I and II analysis

| | <i>NCTR dataset</i> | <i>Walker et al. dataset</i> |
|---------------|---------------------|------------------------------|
| Original size | 232 (100%) | 57,810 (100%) |
| Reduced size | | |
| Phase I | 225 (97%) | 39,822 (69%) |
| Phase II | 183 (79%) | 9,810 (17%) |

about 3% of the original NCTR dataset. This percentage is as would be expected for a dataset designed to cover active and inactive compounds in the chemical space around ER binders. In real situations, like the Walker *et al.* dataset, a much higher percentage of true negatives would be excluded in this phase. As shown in Table II, more than 30% of the Walker *et al.* dataset is expected to be excluded at Phase I. The total number of the Walker *et al.* dataset compounds that passed to Phase II was about 40,000. Most importantly, there were no false negatives introduced in this phase using these filters based on the NCTR dataset.

Simple filters are commonly employed in drug discovery to eliminate compounds for further testing. The filters used vary according to application, but are generally derivatives of the well-known Lipinski's "rule of 5" [28]. Typically, criteria are applied for molecular weight, hydrophobicity and number of H-donors and acceptors that make a chemical "drug-like". The rules are aimed at ruling-out candidate compounds lacking "drug-likeness" properties, since false positives waste time and resources. When we applied similar rules to the NCTR dataset, many false negatives resulted (i.e. many estrogens lacked good drug-like properties). In contrast to drug discovery, false negatives are an unacceptable outcome for regulatory purposes. Our approach to filter design embodied in Phase I is to minimize the number of false negatives while keeping a lower rate of false positives. Phase I rules reflect our aggregate knowledge of the structural requirements for estrogenic activity via the mechanism of ER-binding. These filters were designed to significantly and confidently reduce, without false negatives, the number of compounds to be analyzed further by other more advanced and time-intensive computational models in the remaining phases.

Phase II: Active/inactive Assignment

In principle, the biological activity of a chemical is determined by its structure that can be encoded in three distinct, but also related structural representations: 2D substructures, 3D pharmacophores and physicochemical properties. A 2D substructure is a structural fragment of a molecule, which can be often used as a strong indicator of a particular activity (a structural alert). A 3D pharmacophore is a portion of a chemical's 3D structure that is considered essential in eliciting the biological activity of interest. A physicochemical property of a molecule is a measure of one property of a whole molecule represented by a single value. For example, $\log P$ measures a chemical's hydrophobicity. The biological activity of a chemical is related to the aforementioned structural features, but for a particular mechanism, the functional dependence is better represented by some features than by others. That is, a feature important for one mechanism may be less relevant for a different mechanism. Similarly, for a single mechanism such as ER-binding modeled here, some features may well represent binding dependencies for one structural class, while other features will better represent binding dependencies for a different structural class. Phase II encompasses multiple representations of structural features among the different structural alerts, pharmacophore and classification/clustering models. Consequently, when used in

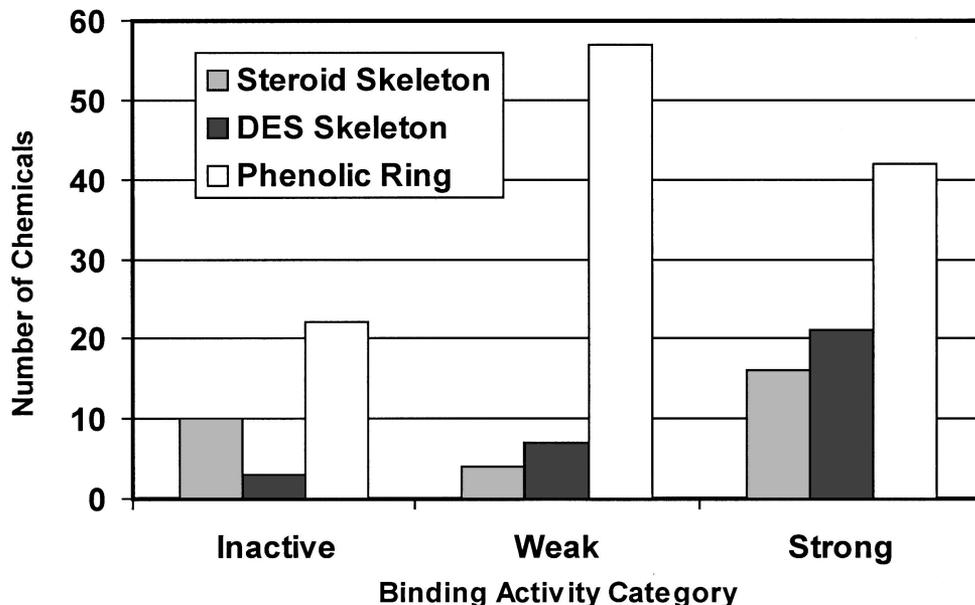


FIGURE 9 Performance of three structural alerts, the steroid skeleton, DES skeleton and the phenolic ring, on the NCTR dataset. The activity categories were arbitrary defined as follows: the strong estrogens had RBA larger than 0.1, the weak estrogens between 0.1 and 10^{-4} and inactive chemicals less than 10^{-4} .

parallel, these differing approaches will in a complementary manner encompass the diversity of structural features relating with activity. This approach is particularly critical for regulation where minimizing false negatives is a paramount concern. To attain minimal false negatives, we developed the rule that only chemicals predicted to be inactive by *all* models were categorized as inactive and eliminated. Any chemical predicted to be an ER-binder by any one or more models is passed to Phase III for quantitative QSAR evaluation.

Three structural alerts were used in this phase. Each was designed to represent 2D substructural features highly associated with estrogens. Figure 9 shows the performances of these alerts when applied to the NCTR dataset. Most chemicals matching the steroid and DES skeletons were strong estrogens. This is consistent with the observation that a common structural feature of many strong estrogens is the presence of two rings (one of them usually a phenolic ring) separated by two carbons [19,29]. In contrast, the phenolic ring alert was a less specific structural indicator of activity. Chemicals matching this alert were structurally diverse, of which about 80% were active chemicals for the NCTR dataset.

In the drug design industry, 3D pharmacophores have proven valuable as queries for lead discovery, whether applied alone or in conjunction with 2D substructure queries. A pharmacophore query is defined by specifying distance and/or angular constraints to characterize activity. A query-matched chemical is considered positive and segregated for further evaluation. One of the advantages of pharmacophore searching is that it can identify chemicals similar to the template in a 3D sense that may not be discernable by chemists in a 2D sense. Seven pharmacophore queries were developed using 17β -estradiol and DES as templates. Chemicals matching any of these queries were labeled as active. A chemical could match none, a few, or many of the seven separate queries, and the number of matches should increase in direct proportion to probability of activity. This so-called “pharmacophore hit frequency” could be used to rank order chemicals in accordance with potential activity. Figure 10 shows that the chemicals in the NCTR dataset with a hit frequency >2 were mostly

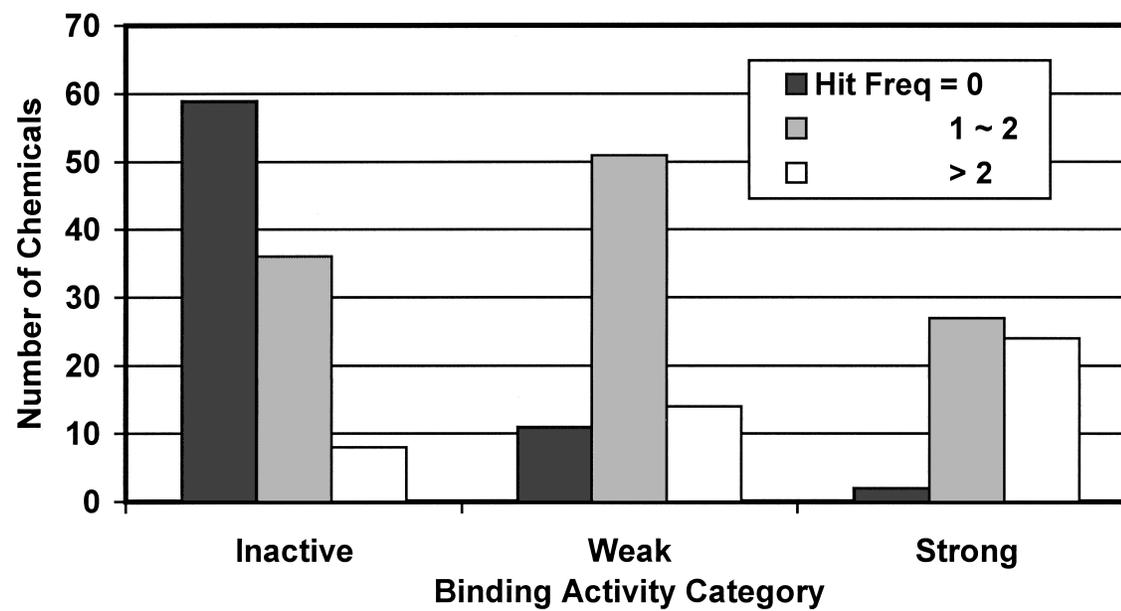


FIGURE 10 Ranking order the NCTR dataset based on pharmacophore hit frequency. The activity categories are defined in Fig. 6.

strong estrogens, while those with a hit frequency ≤ 2 were less active. The majority of the chemicals not matched in any query were inactive.

Classification models, which use pattern recognition methods, group compounds on the basis of their similarities in structural features or biological activity patterns. Two types of classification methods, i.e. (1) non-supervised (untrained) and (2) supervised (trained) methods, are widely used in various areas of science and technology to find the regularities and patterns in data sets [30–32]. Their applications are widespread in chemistry, biology and drug discovery [33,34]. Non-supervised classification techniques group a set of chemicals into subsets on the basis of descriptors representing only structure so that the chemicals within a particular subset generally have similar structural patterns. The supervised classification methods identify the structural features that determine the biological activity patterns and refine the model to amplify the importance of these structural determinants. To construct a supervised classification model, a set of molecular representations (descriptors) is first generated for chemicals in the dataset. Next, a chemometric method is applied to correlate these descriptors with their biological activity, usually represented on a categorical scale.

A number of supervised and non-supervised classification methods were evaluated to classify chemicals into two estrogenic categories, active and inactive [35]. While these approaches are different in a number of ways, they generally produced similar classification results. The nature of the descriptors used, and more particularly the effectiveness in which they encode the structural features of the molecules related to the estrogenic activity, was far more critical than the specific method employed. A number of computational codes can generate a large number of descriptors for a chemical, such as CODESSA (Semichem, Shawnee, KS), which can generate over 400 descriptors. The selection of molecular descriptors is paramount for the model development. A close examination of descriptor correlation with activity enabled identification of four simple descriptors, $\log P$, M_w , phenolic indicator and pharmacophore hit frequency, which yielded good supervised classification models. Together, these four descriptors encode the key 2D- and 3D-structural features, as well as the global molecular properties, which determine ER binding. The KNN and CART methodologies were chosen from among many possibilities because of easy implementation and automation of the system. Figure 11 shows that both models' prediction for the NCTR dataset were about 85–90% correct for chemicals in either active or inactive groups, corresponding to about 10–15% false positive and false negative rates.

Each of the 12 models provides some false negatives. When all models were combined, all active chemicals in the NCTR dataset were predicted to be binders by one or more of the models. Hence, the combined models result in no false negatives for the NCTR dataset (Table I), confirming that the models are complementary in identifying structural features for ER binding.

Consensus Ranking for Priority Setting

There was a major overlap between the hit lists from three structural alerts, seven pharmacophore queries and two classification models. A chemical could be predicted to be active by one, a few, or many of these 12 models. Since each model identifies specific structural attributes associated with activity, the number of models predicting a chemical to be active should tend to increase in direct proportion to its actual activity. In other words, chemicals can be ranked based on the number of models in consensus. This so-called “consensus ranking” thus provides a suitable index for priority setting. To test this approach, the NCTR dataset was divided into five priority groups in accordance with the consensus ranking: Groups 1–5 contained chemicals that had the consensus ranking >6 , 5–6, 3–4,

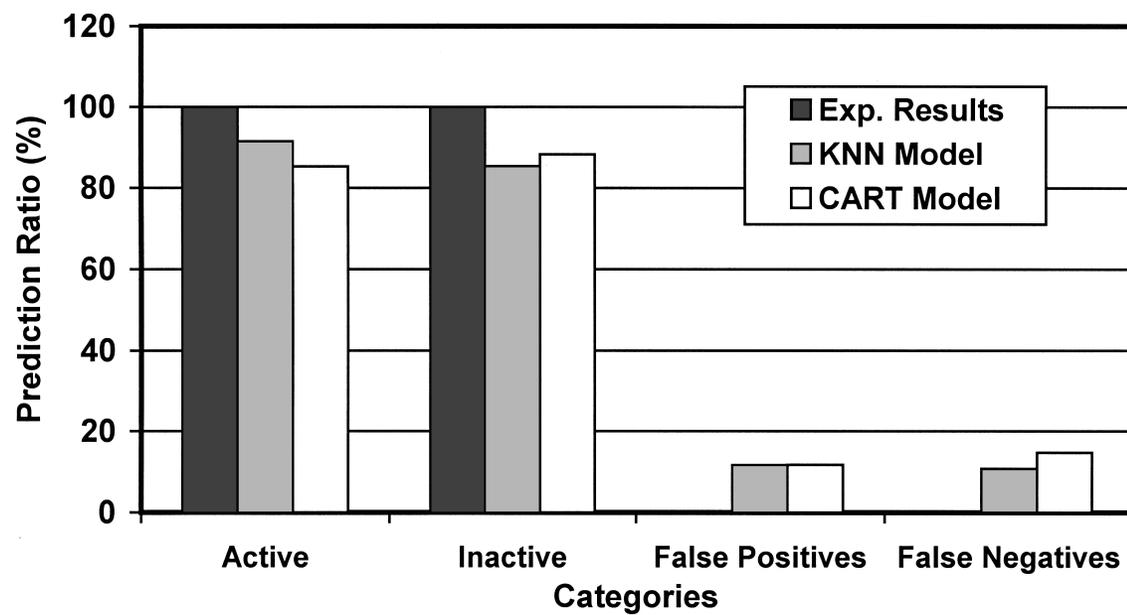


FIGURE 11 Classification results for both active and inactive chemicals in the NCTR dataset using KNN and CART. False positive and false negative rates were the number of misclassified chemicals divided by the total number of chemicals predicted for the inactive and active groups, respectively.

TABLE III Priority setting of the NCTR and Walker *et al.* datasets

| Experiment | | Prediction | | | | |
|---------------------|-----------------------|--------------------|-------------------|----------------------|-----------------|---|
| Assigned categories | RBA range | Prioritized groups | Consensus ranking | NCTR dataset | | Walker <i>et al.</i> dataset No. of compds |
| | | | | Mean RBA* | Active/inactive | |
| A | >10 | 1 | >6 | 1.2 | 32/1 | 124 |
| B | 10^{-1} – 10 | 2 | 5–6 | 1.1×10^{-2} | 44/9 | 317 |
| C | 10^{-3} – 10^{-1} | 3 | 3–4 | 3×10^{-3} | 43/16 | 3,183 |
| D | 10^{-4} – 10^{-3} | 4 | 1–2 | 2.3×10^{-4} | 9/49 | 6,186 |
| E | $<10^{-4}$ | 5 | 0 | $<10^{-4}$ | 0/49 | 30,012 |

*RBA value for inactives was set to 10^{-4} to calculate mean RBA.

1–2 and 0, respectively (Table III). Similarly, the NCTR dataset can also be grouped into five RBA categories: Categories A–E contained chemicals with RBAs >10 , 0 – 10^{-1} , 10^{-1} – 10^{-3} and 10^{-3} – 10^{-4} and $<10^{-4}$, respectively. As expected, the mean RBA values of priority groups decreased as the consensus ranking decreased, and were consistent with the RBA range values assigned for the RBA categories (Table III). The majority of active chemicals (93%) were within priority groups 1–3, and most of these had RBAs no less than 10^{-5} -fold below 17β -estradiol (Fig. 12). Both Table III and Fig. 12 show a clear trend of decreasing inactive chemicals with increasing consensus ranking. Among the approximately 40,000 chemicals of the Walker *et al.* dataset that were not ruled out in Phase I, 9810 chemicals (those in Groups 1–4, Table III) were identified as potential estrogens by at least one model. Of these, some 3600 chemicals were in Groups 1–3. These groups would be expected to contain the majority of active chemicals with $RBA > 10^{-3}$ ($>10^5$ -fold below 17β -estradiol). These chemicals would also be expected to compete for binding to the ER at less than $100 \mu\text{M}$ in our assay because the IC_{50} of 17β -estradiol is about 1 nM.

DISCUSSION

The EDSTAC, organized by the EPA, worked for two years on a strategy for identification of potential endocrine disruptors. They recommended two phases: Tier 1 Screening (T1S) and Tier 2 Testing (T2T) phases [5,6]. T1S consists of several *in vitro* and short-term *in vivo* assays to identify hazards, while T2T consists of multi-generation, multi-endpoint animal

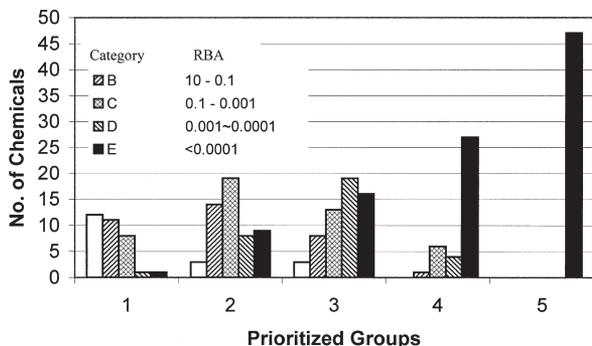


FIGURE 12 Ranking the chemicals of the NCTR dataset according to the consensus ranking derived from structural alerts, pharmacophore queries and classification models. RBA = 100 for 17β -estradiol.

tests for chemicals shown positive in T1S. For a very small portion of the 58,000 chemicals that would undergo some level of assessment, there is enough information to warrant screening and testing. But for most of the chemicals, there is little or no hormonal activity information available. Conduct of these screens and tests on all of the appropriate chemicals will be slow due to either cost, time or test animal expenditures. Therefore, methods for priority setting are essential for identifying the most likely chemicals to possess hormonal activity and thus for efficiency of implementation of the EDSTAC strategy. Transcription-based high throughput pre-screening (HTPS) was recommended by the EDSTAC as the primary source of biological effect information for priority setting. However, in a pilot study undertaken by EPA, the HTPS system did not perform well [36].

Here, we have demonstrated the utility of computer-based methods to develop a scheme to prioritize chemicals in rank order based on estimates of their RBA range. This alternative approach to HTPS for priority setting uses a suite of integrated computational models that include rule-based, pharmacophoric, chemometric and QSAR techniques. There are two critical aspects of the system. First, the suite of models was developed using the NCTR dataset that was designed to calibrate and validate models. Second is the optimal selection and sequencing of appropriate predictive computational models that are used in a hierarchical and complementary manner resulting in both effective and efficient prioritizing of a very large database of chemicals for subsequent assessment.

The advantage of the computational approach to priority setting is the efficiency of scale when applied to a large number of chemicals. When several endpoints are analyzed simultaneously, the efficiency of scale of computation is even more pronounced. Importantly, results from Phases I and II alone, together with information on exposure level, production volume, and environmental persistence of chemicals, may be sufficient to enable initial priority setting for estrogens. Similar procedure appears appropriate for androgen and thyroid hormone mimics.

Phases I and II ruled out as potential estrogens all but 9100 chemicals from the original 58,000 in the Walker *et al.* dataset. Of these, about 3600 were predicted to be in the highest priority groups that identify chemicals with ER binding affinities at an IC_{50} of less than $100\ \mu\text{M}$. How to make full usage of this predictive information is dependent on the application. The highest three priority groups from Phase II might be of equal concern in a regulatory context, where false negatives are the primary issue. The highest three groups would also be suitable for evaluating large combinatorial libraries for adverse activity early in the drug discovery process. In contrast, the chemicals in priority group 1 (Table III) might be important for lead discovery in pharmaceutical industry, because of their requirement for a low incidence of false positives.

The RBAs of the chemicals that were predicted to be active in the first two phases will be quantitatively predicted using multiple QSAR models in Phase III. The performance of several QSAR methods, including CoMFA, CODESSA and HQSAR, developed from literature data were reported previously [8,37–39]. Among these QSAR methods, CoMFA stands out as superior in precision. Currently, we have adopted for Phase III a high-performance CoMFA model developed from the NCTR ER-binding dataset [9]. The model has been validated rigorously, and provides accurate prediction for test chemicals. Furthermore, while time intensive to use, the CoMFA model demonstrates the ability to eliminate a large number of false positives resulting from Phase II. Quantitative predictions of RBAs can be also used for a more refined priority setting. In general, compounds with higher predicted binding affinity should be given a higher priority for earlier assessment in the EDSTAC recommended screening and testing assays.

In Phase IV, a knowledge-based system, or expert system, is proposed to make final decisions on priority setting. We anticipate that the number of chemicals predicted to be

active by the QSAR models in Phase III will be on the order of 4000, or less than 10% of the original Walker *et al.* dataset. Clearly, conducting the EDSTAC battery of assays on 4000 chemicals concurrently is intractable, and there remains the need for prioritizing within this group. The expert system will be useful only after incorporating human knowledge and expertise, or predictive models, i.e. rules, into the system. Then, the expert system will be able to make decisions on individual chemicals based on the rules in its knowledge base.

The expert system's performance will only be as good as the correctness and usefulness of its knowledge base. We suggest that rule determination is inherently a multidisciplinary undertaking, and offer the following general suggestions on the design of priority setting rules, which should be subject to routine changes to accommodate regulatory needs and public concerns.

First, information gained at each phase of the integrated computational approach should be used for setting priorities. For example, the RBAs predicted from Phase III could be converted into the categorical value to rank order chemicals in a manner similar to that used in Phase II. The initial priority setting results from both Phase II and III can then be combined to determine a chemical's priority. Second, structural novelty should be considered. For chemicals with similar predicted RBAs, those with novel structural features should be given higher priority, because chemicals with novel structural features are more likely to be missed, and therefore, they are more likely to cause regulatory problems. However, the nature of structural novelty is difficult to be defined, which should be justified by experts.

The various models presented here as an integrated suite also have utility when applied individually, or when integrated in part in different ways. For example, the Phase II models together have particular utility for application to a large number of chemicals. Alternatively, for a small number of chemicals, the Phase III QSAR models may be preferable since they provide more refined quantitative predictions.

CONCLUDING REMARKS

Various mechanisms are involved for endocrine disruption. The integrated computational approach reported in the paper is based on prediction of chemical binding to the ER, which, in turn, is correlated with numerous estrogenic endpoints. We anticipate that the same scheme will be equivalently applicable to other mechanistic steps (e.g. androgen receptor binding) involved in endocrine disruption, and the associated endpoints. The stringent requirement for developing models for additional mechanisms is appropriately designed training datasets similar to the one employed here for the ER-binding models. Properly validated data allows the structural rules that govern activity to be determined, and used to develop robust predictive models.

While the results presented here clearly show both the feasibility and utility of using a computational approach for priority setting, it is important to note that predictions from any model are intrinsically no better than the experimental data employed for calibration. Any limitations of the assay used to generate the calibration data apply equally to the model's predictions. It is difficult to guarantee that no active chemicals are predicted to be inactive, other than by assaying all chemicals predicted to be inactive. Moreover, 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 error will increase even for a well-designed and executed assay, and false positives and false negatives will both increase. Similarly, more false prediction might be introduced for chemicals with activity close to the cut-off value. The issue for a large number of chemicals of the rate of false positives and false negatives in predicted RBA values must be dealt with experimentally by running assays on a sufficiently

large number of chemicals to characterize the rates. However, using the NCTR dataset there were no false negatives among 129 chemicals that assayed positive, even though there were numerous low-affinity ligands among them.

Acknowledgements

This work was partially supported by the FDA's Office of Women's Health and the American Chemistry Council (ACC). Dr Hong Fang gratefully acknowledges the Oak Ridge Institute for Science and Education Program, supported by the US Department of Energy (DOE) and the US Food and Drug Administration (FDA), for postdoctoral support.

References

- [1] Colburn, T., Dumanoski, D. and Myers, J.P. (1996) *Our Stolen Future* (Plume, New York).
- [2] Kavlock, R.J., Daston, G.P., DeRosa, C., Fenner-Crisp, P., Gray, L.E., Kaattari, S., Lucier, G., Luster, M., Mac, M.J., Maczka, C., Miller, R., Moore, J., Rolland, R., Scott, G., Sheehan, D.M., Sinks, T. and Tilson, H.A. (1996) "Research needs for the risk assessment of health and environmental effects of endocrine disruptors: A report of the U.S. EPA-sponsored workshop", *Environ. Health Perspect.* **104**(Suppl. 4), 715–740.
- [3] 104th U.S. Congress (1996). "Food Quality Protection Act 21 U.S.C. 346a(p). The Safe Drinking Water Act (42 U.S.C. (300j-17). Public Law 104–182".
- [4] US-Congress (1996). "The Food Quality Protection Act (FQPA) and the Safe Drinking Water Act (SDWA)".
- [5] EDSTAC. <http://www.epa.gov/opptintr/opptendo/finalrpt.htm>
- [6] Gray, Jr, L.E. (1998) "Tiered screening and testing strategy for xenoestrogens and antiandrogens", *Toxicol. Lett.* **102–103**, 677–680.
- [7] Patlak, M. (1996) "A testing deadline for endocrine disrupters", *Environ. Sci. Technol.* **30**, 540A–544A.
- [8] Tong, W., Lewis, D.R., Perkins, R., Chen, Y., Welsh, W.J., Goddette, D.W., Heritage, T.W. and Sheehan, D.M. (1998) "Evaluation of quantitative structure–activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor", *J. Chem. Inf. Comput. Sci.* **38**, 669–677.
- [9] Shi, L.M., Fang, H., Tong, W., Wu, J., Perkins, R., Blair, R., Branham, W. and Sheehan, D. (2001) "QSAR models using a large diverse set of estrogens", *J. Chem. Inf. Comput. Sci.* **41**, 186–195.
- [10] Blair, R., Fang, H., Branham, W.S., Hass, B., Dial, S.L., Moland, C.L., Tong, W., Shi, L., Perkins, R. and Sheehan, D.M. (2000) "Estrogen receptor relative binding affinities of 188 natural and xenochemicals: structural diversity of ligands", *Toxicol. Sci.* **54**, 138–153.
- [11] Walker, J.D., Waller, C.W. and Kane, S. (2000) In: Walker, J.D., ed, *The Endocrine Disruption Priority Setting Database (EDPSD): A Tool to Rapidly Sort and Prioritize Chemicals for Endocrine Disruption Screening and Testing* (SETAC, Pensacola, FL).
- [12] Perkins, R., Anson, J., Branham, W., Fang, H., Tong, W., Welsh, W., Chen, Y., Meehan, J., Nossaman, R., Shi, L. and Sheehan, D. (2000) In: Walker, J.D., ed, *The Estrogen Knowledge Base (EKB), a Prototype Toxicological Knowledge Base for Endocrine Disrupting Compounds* (SETAC, Pensacola, FL).
- [13] Sadowski, J. and Kubinyi, H. (1998) "A scoring scheme for discriminating between drugs and nondrugs", *J. Med. Chem.* **41**, 3325–3329.
- [14] Brzozowski, A.M., Pike, A.C., Dauter, Z., Hubbard, R.E., Bonn, T., Engstrom, O., Ohman, L., Greene, G.L., Gustafsson, J.A. and Carlquist, M. (1997) "Molecular basis of agonism and antagonism in the oestrogen receptor", *Nature* **389**, 753–758.
- [15] Shiau, A.K., Barstad, D., Loria, P.M., Cheng, L., Kushner, P.J., Agard, D.A. and Greene, G.L. (1998) "The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen", *Cell* **95**, 927–937.
- [16] Hong, H., Neamati, N., Wang, S., Nicklaus, M.C., Mazumder, A., Zhao, H., Burke, T.R., Pommier, Y. and Milne, G.W.A. (1997) "Discovery of HIV-1 integrase inhibitors by pharmacophore searching", *J. Med. Chem.* **40**, 930–936.
- [17] Greenidge, P.A., Carlsson, B., Bladh, L.-G. and Gillner, M. (1998) "Pharmacophores incorporating numerous excluded volumes defined by X-ray crystallographic structure in three-dimensional database searching: application to the thyroid hormone receptor", *J. Med. Chem.* **41**, 2503–2512.
- [18] Qian, X., Tong, W., Fang, H., Hong, H., Tu, M., Perkins, R. and Sheehan, D. (2000). "Pharmacophore model for rapid screening estrogenic chemicals" (in preparation).
- [19] Fang, H., Tong, W., Shi, L., Blair, R., Perkins, R., Branham, W.S., Dial, S.L., Moland, C.L. and Sheehan, D.M. (2001) "Structure activity relationship for a large diverse set of natural, synthetic and environmental chemicals", *Chem. Res. Toxicol.* (in press).
- [20] Smellie, A., Kahn, S.D. and Teig, S. (1995) "An analysis of conformational coverage. 1. Validation and estimation of coverage", *J. Chem. Inf. Comput. Sci.* **35**, 285–294.
- [21] Smellie, A., Kahn, S.D. and Teig, S. (1995) "An analysis of conformational coverage. 2. Applications of conformational models", *J. Chem. Inf. Comput. Sci.* **35**, 295–304.

- [22] Smellie, A., Teig, S.L. and Towbin, P. (1995) "Poling: promoting conformational coverage", *J. Comput. Chem.* **16**, 171–187.
- [23] Livingstone, D. (1995) *Data Analysis for Chemists—Applications to QSAR and Chemical Product Design* (Oxford University Press, New York).
- [24] Hawkins, D.M., Young, S.S. and Rusinko, III, A. (1997) "Analysis of large structure–activity data set using recursive partitioning", *QSAR* **16**, 296–302.
- [25] Clark, L.A. and Pregibon, D. (1992) *Tree-Based Models* (Chambers and Hastie).
- [26] Meylan, W. and Howard, P. (1995) "Atom/fragment contribution method for estimating octanol–water partition coefficients", *J. Pharm. Sci.* **84**, 83–92.
- [27] Venables, W.N. and Ripley, B.D. (1997) *Modern Applied Statistics with S-PLUS* (Springer, Berlin).
- [28] Lipinski, C.A., Lombardo, F., Dominy, B.W. and Feeney, P.J. (1997) "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings", *Adv. Drug Deliv. Rev.* **23**, 3–25.
- [29] Fang, H., Tong, W., Perkins, R., Soto, A., Prechtel, N. and Sheehan, D.M. (2000) "Quantitative comparison of *in vitro* assays for estrogenic activity", *Environ. Health Perspect.* **108**, 723–729.
- [30] Arabie, P., Hubert, L.J. and De Soete, G. (1996) *Clustering and Classification* (World Scientific Pub Co, Singapore).
- [31] Mirkin, B.G. (1996) *Mathematical Classification and Clustering* (Kluwer Academic Publishers, Boston).
- [32] Gordon, A.D. (1981) *Classification: Methods for the Exploratory Analysis of Multivariate Data* (Chapman and Hall, London).
- [33] Willett, P. (1987) *Similarity and Clustering in Chemical Information Systems* (Research Studies Press, Letchworth).
- [34] Shi, L.M., Tong, W., Perkins, R., Chen, Y., Fang, H., Poirier, L. and Sheehan, D. (1998) *Classification Methods for Toxicity Prediction and Anticancer Drug Discovery* (Urbana, Illinois).
- [35] Hong, H., Tong, W., Fang, H., Shi, L., Qian, X., Wu, J., Perkins, R., Walker, J.D., Braham, W. and Sheehan, D. (2000) "Prediction of estrogen receptor binding for 58,000 chemicals using an integrated system of a tree-based model with structural alerts", *Environ. Health Perspect.* **110**, 29–36.
- [36] E/ELetter (1999). "EPA Scientific Advisors: HTPS Needs More Work; Mixtures Low Priority", *Endocrine/Estrogen Letter*.
- [37] Tong, W., Perkins, R. and Sheehan, D.M. (1999) "Perspectives on three-dimensional quantitative structure–activity relationship (3D-QSAR)/comparative molecular field analysis (CoMFA) in determining estrogenic effects", *Jpn. Chem. Today* **2**, 50–57.
- [38] Tong, W., Perkins, R., Strelitz, R., Collantes, E.R., Keenan, S., Welsh, W.J., Branham, W.S. and Sheehan, D.M. (1997) "Quantitative structure–activity relationships (QSARs) for estrogen binding to the estrogen receptor: predictions across species", *Environ. Health Perspect.* **105**, 1116–1124.
- [39] Tong, W., Perkins, R., Xing, L., Welsh, W.J. and Sheehan, D.M. (1997) "QSAR models for binding of estrogenic compounds to estrogen receptor alpha and beta subtypes", *Endocrine* **138**, 4022–4025.