QSAR Models Using a Large Diverse Set of Estrogens

Leming M. Shi,[†,‡] Hong Fang,[†] Weida Tong,*,[†] Jie Wu,[†] Roger Perkins,[†] Robert M. Blair,[§]
William S. Branham,[§] Stacy L. Dial,[§] Carrie L. Moland,[§] and Daniel M. Sheehan[§]

R.O.W. Sciences, Inc., Jefferson, Arkansas 72079, and Division of Genetic and Reproductive Toxicology,
National Center for Toxicological Research (NCTR), Jefferson, Arkansas 72079

Received June 17, 2000

Endocrine disruptors (EDs) have a variety of adverse effects in humans and animals. About 58 000 chemicals, most having little safety data, must be tested in a group of tiered assays. As assays will take years, it is important to develop rapid methods to help in priority setting. For application to large data sets, we have developed an integrated system that contains sequential four phases to predict the ability of chemicals to bind to the estrogen receptor (ER), a prevalent mechanism for estrogenic EDs. Here we report the results of evaluating two types of QSAR models for inclusion in phase III to quantitatively predict chemical binding to the ER. Our data set for the relative binding affinities (RBAs) to the ER consists of 130 chemicals covering a wide range of structural diversity and a 6 orders of magnitude spread of RBAs. CoMFA and HQSAR models were constructed and compared for performance. The CoMFA model had a $r^2 = 0.91$ and a $q^2_{\mathrm{LOO}} = 0.66$. HQSAR showed reduced performance compared to CoMFA with $r^2 = 0.76$ and $q^2_{\mathrm{LOO}} = 0.59$. A number of parameters were examined to improve the CoMFA model. Of these, a phenol indicator increased the q^2_{LOO} to 0.71. When up to 50% of the chemicals were left out in the leave-*N*-out cross-validation, the q^2 remained significant. Finally, the models were tested by using two test sets; the q^2_{pred} for these were 0.71 and 0.62, a significant result which demonstrates the utility of the CoMFA model for predicting the RBAs of chemicals not included in the training set. If used in conjunction with phases I and II, which reduced the size of the data set dramatically by eliminating most inactive chemicals, the current CoMFA model (phase III) can be used to predict the RBA of chemicals with sufficient accuracy and to provide quantitative information for priority setting.

INTRODUCTION

Experimental and epidemiological studies suggest that some man-made and naturally occurring chemicals released to the environment have the potential to interrupt normal functioning of the endocrine systems of humans and wildlife.[1,2] These chemicals, termed EDs, may pose serious threats to the reproductive capability of humans and wildlife and are thought to be the cause of declines in some wildlife populations.[3,4] In response to scientific and public concerns on EDs, the U.S. Congress in 1996 mandated that the Environmental Protection Agency (EPA) develop a strategy for screening and testing a large number of chemicals (∼58 000) found in drinking water, food additives, and other sources for their endocrine disruption potential.[5] Because of the high cost associated with screening and testing, it is crucial that priorities be set to ensure that compounds with the highest predicted or measured activities be given first priority for entry into the screening procedure.[6]

Several types of hormonal activities, including, but not limited to, estrogenic, androgenic, and thyroidal, are believed to contribute to endocrine disruption.[1] Recently, we proposed an integrated "four-phase" approach for priority setting of potential estrogenic EDs.[7] The general approach is expected to be equally applicable to other endocrine disruption mechanisms, such as androgen and thyroid hormonal activities. The four-phase approach integrates a suite of computational models that can be used in setting priorities for a large number of chemicals. Phase I uses four Lipinski "rule of 5"-type simple rejection filters to eliminate the chemicals that are most unlikely to bind the ER.[8] The chemicals surviving phase I are then classified as active or inactive in phase II on the basis of the presence or absence of three key 2D structural alerts, seven pharmacophore features, and the predictions of two classification models using K-nearest neighbor (KNN) and classification and regression tree (CART) methods. In phase III, QSAR models are used quantitatively to predict activity of the chemicals categorically predicted to be active in phase II. In phase IV of the integrated system, the phase II and III predictions are combined with other available information, such as human exposure level, environmental fate, and production volume, to determine a chemical's priority for testing. The feasibility and application of the first two phases were assessed for priority setting of the ∼58 000 chemicals identified by Walker et al.[9] as candidates for entry into screening and testing. Some 9100 chemicals were predicted to bind to ER. Of these, only 3600 were expected to bind to ER at binding affinity up to 100 000-fold less than that of the endogenous hormone, 17*β*-estradiol (E_2), which might need to be assessed in phase III using QSAR models. Here, we report two QSAR

* Corresponding author. Telephone: (870) 543-7142. Fax: (870) 543-7382. E-mail: wtong@nctr.fda.gov.
[†] R.O.W. Sciences, Inc.
[‡] Current address: BASF Corp. P.O. Box 400, Princeton, NJ 08543.
[§] NCTR.

10.1021/ci000066d CCC: $20.00
Published on Web 12/15/2000

methods that were thoroughly evaluated for application in phase III.

QSAR methods have proven successful in molecular design and drug discovery.[10] The endocrine disruptor screening and testing advisory committee (EDSTAC) organized by EPA considers QSAR as an important part of its priority setting process, as described in its final report.[6] In the past few years, a number of QSAR models have been developed for ligand binding to the ER.[11−19] Most of these QSAR models were constructed using the comparative molecular field analysis (CoMFA). Although a predictive CoMFA model is dependent on a number of factors, a training set with a broad representation over the chemistry space is critical to ensure its predictive capability for a large number of diverse chemicals. Unfortunately, most CoMFA models for ER binding developed previously were based on data sets available in the literature, which to date had been small data sets with limited structural diversity.[11−19] Although these models yield good statistical results and explain some structural determinants for ER binding, they have limited applicability in predicting the ER−ligand binding affinity of chemicals that, in fact, cover a wide range of structural diversity.

To obtain an adequate training set to develop a more robust QSAR model for regulatory purpose, a rat ER binding assay was developed and validated in the U.S. FDA’s National Center for Toxicological Research (NCTR) to provide a large data set for model development.[20,21] The resulting NCTR data set contains chemicals that were selected to cover the structural diversity of chemicals that bind to ER with an activity distribution ranging over 6 orders of magnitude, which is an essential requirement for a robust predictive model for structurally diverse estrogens. The selection process was highly interdisciplinary, involving computational chemists, biologists, and experimental toxicologists and has resulted in the steady improvement in performance of the QSAR models.[22] To the best of our knowledge, the NCTR data set is the best consistent data set available to develop models for estrogens and was the primary basis for construction of the phase I and II models in the integrated “four-phase” approach.

The rat uterine cytosol ER competitive binding assay is the gold standard for in vitro ER assays. When compared to results from other ER binding assays, there is a general consistency between relative ER activities across different assay methods and species.[23] For example, we found a high linear correlation for ER binding affinities among a diverse group of chemicals assayed with ER from rat uterine cytosol and hERα. Further, we also found that ER assay results correlated very well with those from a yeast-based reporter gene assay and MCF-7 cell proliferation assay. These findings demonstrate that ER binding is the major determinant across three levels of biological complexity (receptor binding, a yeast reporter gene response, and cell proliferation) of estrogen action. Moreover, 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.[24] Therefore, the prediction of ER binding activity provides an important piece of information for priority setting.

We have previously evaluated three different techniques for the generation of QSAR models−CoMFA, CODESSA (comprehensive descriptors for structural and statistical analysis), and HQSAR (hologram QSAR)−for their utility (predictivity, speed, accuracy, and reproducibility) to screen a large number of compounds for ER binding activity.[13] Common to the three QSAR methods in the derivation of a regression model is the use of PLS; the differences among these QSAR techniques are primarily in the type of the descriptors used to represent chemical structure. Specifically, CoMFA employs steric and electrostatic field descriptors that encode detailed information concerning intermolecular interaction in three dimensions. CODESSA calculates molecular descriptors on the basis of 2D and 3D structures and quantum-chemical properties; whereas HQSAR uses molecular holograms constructed from counts of substructural molecular fragments. For three relatively small data sets under investigation, the QSAR models generated using CoMFA and HQSAR techniques demonstrated comparable high quality for potential usage to identify ER ligands that may act as EDs. In this paper, these two techniques are further investigated and compared, particularly for their predictivity, by using the NCTR data set.

MATERIAL AND METHODS

Data Sets. The training set for QSAR model development was comprised of 130 diverse chemicals (Table 1).[20,21] The binding affinities of the chemicals in the data set were determined by a competitive ER binding assay with $[^3H]E_2$, using rat uterine cytosol.[20] RBA, which is defined as 100 times the ratio of the molar concentrations of E_2 and the competing chemical required to decrease the receptor-bound radioactivity by 50%, was used for QSAR model development. Each experimental datum is replicated at least twice. The mean value was used for modeling.

The constructed QSAR models were validated using external test data sets. There are a number of experimental data sets reported in the literature for estrogenic activity. Some of them have been used to develop various types of QSAR models either by us or by other researchers.[11−19] Our earlier analysis on comparison of various in vitro assays demonstrates that there is generally a good linear correlation among activity measurements from the ER binding assay, the yeast-based reporter gene assay and the E-SCREEN assay.[23] The data sets from other ER binding assays were used as primary sources for the test sets. To select appropriate test data sets to validate the models, the following general criteria were applied:

1. Since the QSAR models developed in this study were primarily used to predict the activity of xenoestrogens, the test data sets should contain a substantial portion of xenoestrogens.

2. A literature survey revealed a great variability in the absolute activity value of a chemical obtained from different assays, or conducted on different species, or with the same assay performed by different labs.[23] For these reasons, the activity value of the selected literature data set was normalized to the NCTR data set (the training set). The detailed procedure of normalization is reported in our previous publication.[23] Briefly, the selected data set was first correlated

Table 1. Experimental, CoMFA-Calculated, and HQSAR-Calculated log RBA for 130 Chemicals

name	expt	CoMFA	HQSAR
diethylstilbestrol (DES)	2.60	1.880	0.591
meso-hexestrol	2.48	1.975	1.545
ethynyl estradiol	2.28	1.455	0.928
4-OH-tamoxifen	2.24	1.702	1.077
17*β*-estradiol	2.00	0.855	0.404
4-OH-estradiol	1.82	1.154	0.669
α-zearalenol	1.63	1.548	−0.009
ICI 182 780	1.57	1.102	1.791
dienestrol	1.57	1.938	0.070
α-zearalanol	1.48	1.437	0.180
2-OH-estradiol	1.47	0.475	0.898
diethylstibestrol monomethyl ether	1.31	1.141	0.286
3,3′-dihydroxyl hexestrol	1.19	1.116	2.161
droloxifene Citrate	1.18	0.930	0.978
ICI 164 384	1.16	1.137	1.769
dimethylstibestrol	1.16	0.487	−0.206
moxestrol	1.14	1.6	1.007
17-deoxyestradiol	1.14	0.444	0.294
2,6-dimethylhexestrol	1.11	−0.171	2.079
estriol	0.99	0.324	0.038
monomethyl ether hexestrol	0.97	1.221	1.240
estrone	0.86	0.900	−0.112
p-(α,*β*-diethyl-*p*-methyl-phenethyl)-*meso*-phenol	0.60	0.931	1.533
17α-estradiol	0.49	0.465	0.404
dihydroxymethoxychloroolefin	0.42	−0.556	−0.624
mestranol	0.35	−0.125	0.649
zearalanone	0.32	0.209	0.170
tamoxifen	0.21	1.220	0.720
toremifene citrate	0.14	0.747	0.714
α,α-dimethyl-*β*-ethyl allenolic acid	−0.02	0.885	0.166
coumestrol	−0.05	0.164	−1.187
4-ethyl-7-OH-3-(*p*-methoxyphenyl)-dihydro-1-benzopyran-2-one	−0.05	−0.255	−0.223
Clomiphene	−0.14	0.232	−0.123
nafoxidine	−0.14	−0.252	1.560
6α-OH-estradiol	−0.15	0.934	0.452
b-zearalanol	−0.19	−0.137	0.180
3-hydroxy-estra-1,3,5(10)-trien-16-one	−0.29	0.144	0.073
3-deoxyestradiol	−0.30	−0.573	−0.099
7,3′,4′-trihydroxyisoflavone	−2.35	−0.913	−1.473
3,6,4′-trihydroxyflavone	−0.35	−0.900	−2.896
genistein	−0.36	−1.786	−1.673
4,4′-dihydroxystibene	−0.55	−0.699	−1.496
HPTE	−0.60	−1.096	−0.778
monohydroxy methoxychloroolefin	−0.63	−0.162	−0.930
2,3,4,5-tetrachloro-4′-biphenylol	−0.64	−1.023	−2.170
norethynodrel	−0.67	0.670	−0.614
2,2′,4,4′-tetrahydroxybenzil	−0.68	−0.048	−0.986
β-zearalenol	−0.69	−0.661	−0.009
equol	−0.82	−0.564	−1.078
6,4′-dihydroxyflavone	−0.82	−1.126	−2.858
monohydroxy methoxychlor	−0.89	−0.845	−1.083
3*β*-androstanediol	−0.92	−0.711	−2.144
bisphenol B	−1.07	−1.229	−1.562
phloretin	−1.16	−1.565	−1.473
diethylstibestrol dimethyl ether	−1.25	−0.690	−0.020
4,2′,4′-trihydroxychalcone	−1.26	−1.362	−1.983
2,5-dichloro-4′-biphenylol	−1.44	−2.103	−2.618
4,4′-(1,2-ethanediyl)bisphenol	−1.44	−1.128	−1.160
16b-hydroxy-16-methyl-3-methyl ether estradiol	−1.48	−1.470	−0.680
aurin	−1.50	−1.593	−0.653
nordihydroguaiaretic acid	−1.51	−1.338	−0.071
nonylphenol	−1.53	−1.769	−1.847
apigenin	−1.55	−2.436	−2.691
kaempferol	−1.61	−2.245	−2.729
daidzein	−1.65	−1.075	−1.764
3-methylestriol	−1.65	−1.257	−0.240
4-dodecylphenol	−1.73	−1.506	−1.401
ethylhexylparaben	−1.74	−2.105	−2.304
4-tert-octylphenol	−1.82	−2.229	−2.138
phenolphthalein	−1.87	−2.673	−2.143
kepone	−1.89	−1.803	−1.779
heptylparaben	−2.09	−2.430	−2.329
bisphenol A	−2.11	−2.079	−2.773
naringenin	−2.13	−2.737	−2.577
4-chloro-4′-biphenylol	−2.18	−2.024	−2.822
3-deoxyestrone	−2.20	−0.683	−0.615
octylphenol	−2.31	−2.063	−1.995
fisetin	−2.35	−2.298	−2.577
biochanin A	−2.37	−2.693	−1.979
4′-hydroxychalcone	−2.43	−2.958	−2.444
2,2′-methylenebis(4-chlorophenol)	−2.45	−2.607	−2.824
4,4′-dihydoxybenzophenone	−2.46	−1.773	−2.626
benzylparaben	−2.54	−2.471	−2.22
4-hydroxychalcone	−2.55	−2.040	−2.417
2,4-dihydroxybenzophenone	−2.61	−2.209	−2.708
4′-hydroxyflavanone	−2.65	−2.118	−2.911
3α-androstanediol	−2.67	−2.565	−2.144
4-phenethylphenol	−2.69	−2.221	−1.500
doisynoestrol	−2.74	−1.736	−1.239
5,4′-dihydroxy-7-methoxyiso-flavone (prunetin)	−2.74	−2.812	−1.973
myricetin	−2.75	−2.249	−2.343
2-chloro-4-biphenylol	−2.77	−2.370	−2.685
triphenylethylene	−2.78	−2.217	−0.822
3′-hydroxyflavanone	−2.78	−3.782	−2.987
chalcone	−2.82	−2.839	−2.742
o,p′-DDT	−2.85	−2.348	−1.644
4-heptyloxyphenol	−2.88	−2.173	−3.456
dihydrotestosterone (DHT)	−2.89	−3.141	−2.321
formononetin	−2.98	−1.935	−2.069
bis(4-hydroxyphenyl)methane	−3.02	−2.928	−2.518
p-phenylphenol	−3.04	−2.638	−2.800
6-hydroxyflavanone	−3.05	−2.170	−3.091
4,4′-sulfonyldiphenol	−3.07	−3.154	−3.410
butylparaben	−3.07	−2.805	−2.714
diphenolic acid	−3.13	−3.478	−1.625
1,3-diphenyltetramethyldisiloxane	−3.16	−2.508	−3.288
ethylparaben	−3.22	−3.184	−2.777
propylparaben	−3.22	−2.899	−2.760
3,3′,5,5′-tetrachloro-4,4′-biphenyldiol	−3.25	−2.914	−2.021
phenol red	−3.25	−3.765	−3.347
4-*tert*-amylphenol	−3.26	−3.019	−2.242
baicalein	−3.35	−2.683	−2.983
morin	−3.35	−2.809	−2.594
4-*sec*-butylphenol	−3.37	−3.078	−2.416
4-chloro-3-methylphenol	−3.38	−3.723	−3.107
6-hydroxyflavone	−3.41	−2.560	−3.208
3-phenylphenol	−3.44	−4.062	−2.878
4-benzyloxphenol	−3.44	−3.638	−3.446
methylparaben	−3.44	−3.297	−2.957
2-*sec*-butylphenol	−3.54	−3.904	−2.744
4-*tert*-butylphenol	−3.61	−3.177	−2.922
2,4′-dichlorobiphenyl	−3.61	−3.350	−3.046
2-chloro-4-methylphenol	−3.66	−4.305	−3.025
phenolphthalin	−3.67	−3.351	−2.446
4-chloro-2-methylphenol	−3.67	−4.182	−3.218
7-hydroxyflavanone	−3.73	−3.944	−3.035
3-ethylphenol	−3.87	−3.641	−3.011
rutin	−4.09	−4.235	−5.620
4-ethylphenol	−4.17	−4.220	−2.948
4-methylphenol	−4.50	−4.435	−3.037

with the NCTR data set on the basis of the shared compounds in both data sets, and then the activity value for each compound not in the NCTR data set was normalized to the NCTR data set on the basis of the correlation equation. To ensure a valid normalization, the selected data set should include a sufficient number of chemicals shared with the

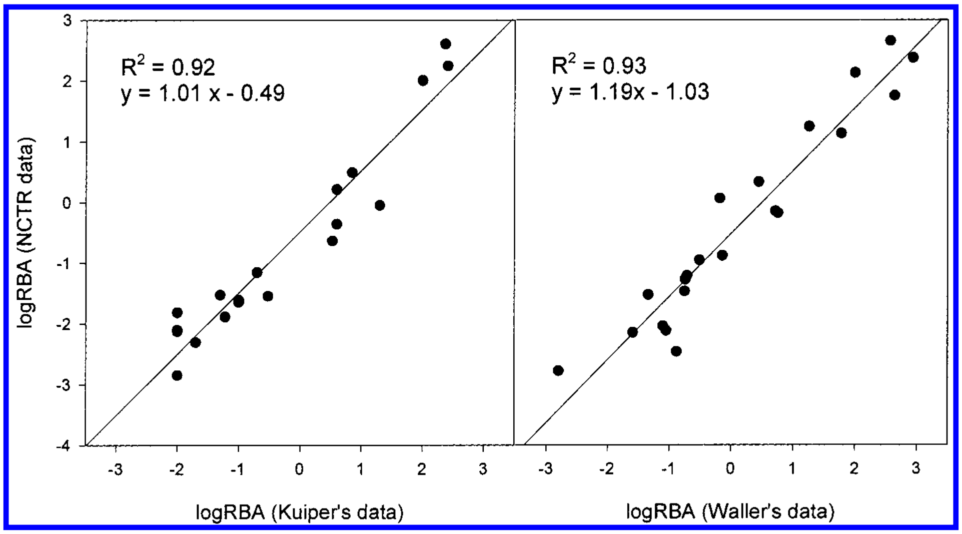

Figure 1. Correlation of the NCTR data with Kuiper's and Waller's data.

NCTR data set to establish a statistically significant linear correlation. The normalized activity value of a chemical can then be compared with the model-predicted value to assess the performance of the QSAR model.

Two data sets reported by Kuiper[25] and Waller,[15] respectively, were selected. The activity values of both data sets were obtained from ER competitive binding assays: 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. Kuiper's data set contains 60 chemicals, of which 35 chemicals are assayed in the NCTR data set. Of 35 common chemicals, 19 chemicals are active in both data sets. A nice linear correlation was observed between the Kuiper and NCTR data sets based on these 19 chemicals (Figure 1). Similarly, Waller's data set contains 58 chemicals with 33 common chemicals. The linear correlation between the Waller and NCTR data sets was also observed for 21 common active chemicals (Figure 1). Both Kuiper's and Waller's data sets contain 25 chemicals that are not assayed in the NCTR data set, which are used to test the QSAR models.

Molecular Modeling. All molecular modeling and statistical analyses were performed using Sybyl 6.5 (Tripos, St. Louis, MO).

As shown in Figure 2, the basic chemical structures for the study consisted of several categories: (1) steroidal compounds; (2) two benzene rings separated by two carbons (DES derivatives and most phytoestrogens); (3) two benzene rings separated by one carbon (DDTs and bisphenol A derivatives); (4) biphenyls (PCBs); (5) chemicals with a single phenolic ring (alkylphenols and parabens); (6) miscellaneous chemicals. The putative bioactive molecular conformation for each chemical used for CoMFA and its rules for structural alignment were determined by selecting a starting conformation that was followed by energy optimization using the standard Tripos force field and parameter settings. The crystal structures of four ligands, E_2, raloxifene (Ral), diethylstilbestrol (DES), and 4-hydroxytamoxifen (OHT), binding to the ERα were used to determine the starting conformation of steroids and DES derivatives, as well as for antiestrogens.[26,27] The chemical structures of phytoestrogens are obtained or modified from the Cambridge Structural Database. The starting conformation of the rest of the chemicals was determined using a systematic search tool for torsion angles.

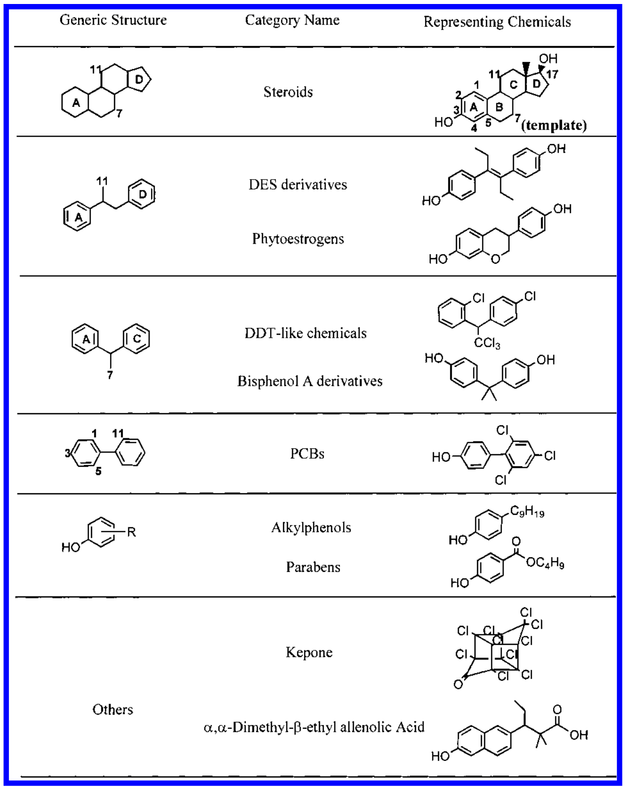

Figure 2. Structural categories and representative chemicals of the NCTR data set.

CoMFA Alignment. To develop a CoMFA model, the molecules of interest must first be aligned to maximize superposition of their steric and electrostatic fields. Although a statistically robust CoMFA model is dependent on a number of factors, proper alignment is essential to produce a valid QSAR model. For chemical congeners, the alignment rule is normally defined on the basis of the maximum common substructure among the training set chemicals, which usually leads to a statistically robust CoMFA model. The drawback for such models is predicting activities of chemicals whose structures are not similar to the training set. In contrast, a CoMFA model based on a structurally diverse data set provides more robust predictions. But the

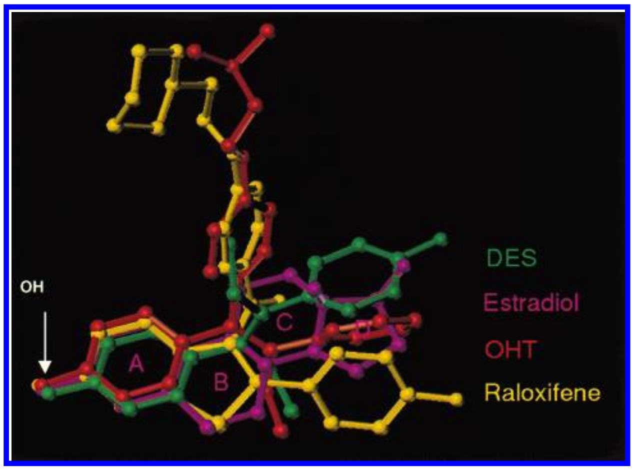

Figure 3. Relative positions of E_2, DES, raloxifene, and 4-OH-tamoxifen in the ER binding site derived from superposition of their bound receptor crystal structures.

critical and difficult aspect for such a CoMFA model, like the one in this study, is choosing the most appropriate set of alignment rules for the structurally diverse training set. Fortunately, crystal structures of four ligands binding to the ER have been published,[26,27] which aided our derivation of rational CoMFA alignment rules. The backbone of the α helices (except helix 12, which shows a dramatic conformational difference when binding to estrogens vs antiestrogens) were used to superimpose the four complexes. The superimposed complex structures revealed the conformational conservation and flexibility of the ER ligand binding domain. Furthermore, by overlapping of the backbone of the α helices, the corresponding superimposition of the four ligands (note: coordinates of these ligands were not used in the superimposition of the complexes), as shown in Figure 3, could be examined in detail with respect to the binding contribution from different structural features, individually and in combination. It appears that the A-rings overlap very nicely, whereas there is considerable flexibility at the D-ring end. This is consistent with the observation of the importance of the A-ring phenolic group in ER-ligand binding, where the 3-OH group forms three hydrogen bond interactions with the ER ligand binding domain and a water molecule.

Systematic studies on the influence of substituents at various positions of E_2 revealed that for most positions the introduction of substituents produces a loss of binding affinity.[16-29] However, introducing a substituent to the 7α or 11β positions generally enhances binding. The degree of increase in activity is strongly dependent on the nature of substituents. Small steric substituents increase the activity,[30] while rather large substituents reduce it, but also give rise to antiestrogenic activity, exemplified by ICI 180 780 and ICI 164 384. The two ethyl groups of DES functionally resemble substituents at the 7α and 11β positions of E_2. One of the ethyl groups is in the precise space that is occupied by the 11β-substituent of E_2.[31] Dimethystilbestrol (DMS) (RBA = 14.50) and 4,4′-dihydroxystilbene (RBA = 0.38) are in the order of one carbon atom less on both side chains. The loss of RBAs for these two chemicals was 28- and 1423-fold compared to that of DES.

The aforementioned evidences indicate that the structural features of E_2 important for ER binding include (1) a phenolic A ring, (2) a D ring, (3) 7α and/or 11β substituents, and (4) a hydrophobic backbone. The combination of any three of the structural features generally yields a strong estrogen. Using E_2 as a template molecule, these four structural features were the basis for the alignment rules. Specifically, six pharmacophoric elements of E_2—the centroids of the A-, B-, C-, and D-rings and 7 and 11 positions—were used for alignment. The corresponding pharmacophoric elements for each structural category are shown in Figure 2 (left panel). The alignment rules employ a least-squares fitting of pharmacophoric elements between E_2 and the molecule from the training set, and specifically the following: (1) Steroids were aligned on the basis of the centers of the A- and D-rings and positions 7 and 11. (2) The corresponding pharmacophoric elements for DES derivatives and phytoestrogens were the centers of the A- and D-rings and position 11. (3) For DDT- and bisphenol A-type chemicals the corresponding aligned positions were the centers of the A- and D-rings and the 11 position of E_2. (4) PCBs were aligned on the basis of the positions 1, 3, 5, and 11 of E_2. (5) Alignment of alkylphenols and parabens was based on the superposition of their phenolic rings to that of E_2. (6) The alignment of the miscellaneous chemicals was determined individually on the basis of an appropriate rationalization consistent with overall alignment strategy.

Calculation of QSAR Descriptors. The calculation of CoMFA steric and electrostatic descriptors, as well as HQSAR holograms, is described in our previous publications.[13] Briefly, there is the following:

CoMFA Descriptors. After alignment, the molecules in the training set were placed in a three-dimensional cubic lattice with 2 Å spacing. The steric (van der Waals) and electrostatic (Coulombic) fields were calculated for each molecule at each mesh point using an sp^3 carbon probe with +1.0 charge. Any calculated steric and electrostatic energies that were greater than 30 kcal/mol were truncated to this value. Column filtering was set to 1 kcal/mol.

HQSAR Holograms. The substructual fragments in the predefined size range of atoms (the default range is 4−7) were generated for each molecule in the training set. The information contained in each fragment is defined by fragment distinction parameters, including atoms, bonds, connections, hydrogen, and chirality. The generated fragments were then hashed into a fixed length array to produce the molecular hologram. The hologram length defines the dimensionality of the descriptor space, which is determined from a range of predefined hologram lengths using a trial-and-error approach evaluated by the smallest error generated in the models. HQSAR descriptors encode all possible molecular fragments (linear, branched, and overlapping). Additional 3D information such as hybridization and chirality are encoded in the molecular holograms.

PLS-QSAR. To form the basis for a predictive statistical model, the method of partial least-squares (PLS) regression[32] was used to correlate variations in the biological activities with variations in the respective descriptors for the NCTR data set. The optimal number of principal components (PCs), corresponding to the smallest standard error of prediction, was determined by the leave-one-out (LOO) cross-validation procedure, which yields a cross-validated q^2_{LOO} to measure the model's predictivity. Using the optimal number of PCs, the final PLS analysis was carried out without cross-validation to generate a predictive QSAR

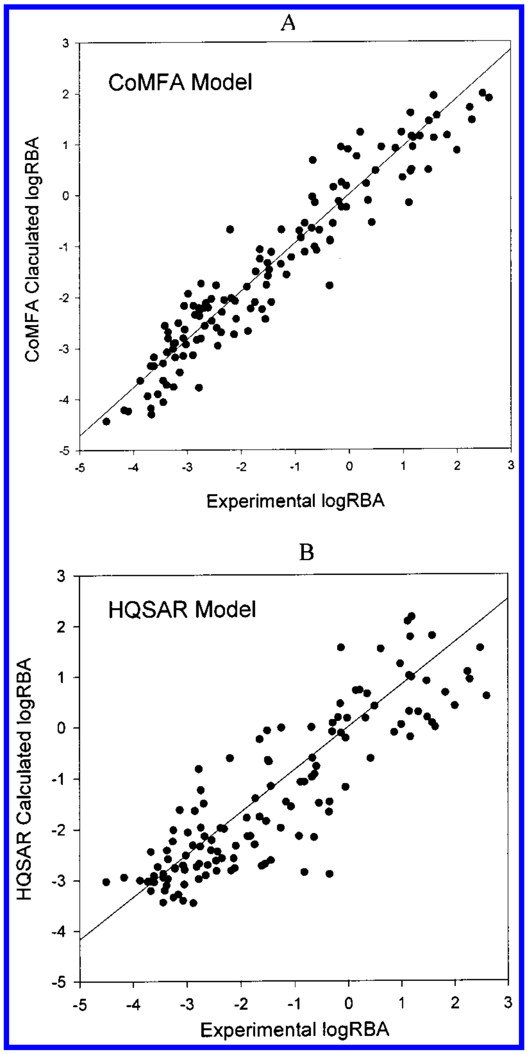

Figure 4. CoMFA (A) and HQSAR (B) models for the NCTR data set.

model with a conventional correlation coefficient, r^2. In addition, the leave-N-out (LNO) cross-validation procedure was employed to further validate the models. In this method, the data set is first randomly divided into N groups with approximately equal numbers of chemicals in each group. Each group is systematically excluded once from the data set, and the activities of the chemicals in the omitted group are predicted by a model derived from the remaining chemicals in the data set. Similar to the q^2_{LOO} value derived from the LOO process, the q^2_{LNO} value for this procedure can also be calculated on the basis of the prediction of the left-out chemicals. In this study, a range of N groups ($N = 2-10$, 13, 20, 50) was used to perform the LNO cross-validation. For each group, the cross-validation was carried 100 times to determine the mean and standard error of q^2_{LNO}.

Model Prediction. The QSAR models were further validated for prediction of the external validation data sets. A statistical measure, q^2_{pred}, was used to compare the predictive capability among different QSAR models. The predictive q^2_{pred} is calculated from[33]

$$q^2_{\mathrm{pred}} = 1 - \mathrm{PRESS/SD}$$

where PRESS is the sum of squared differences between the "actual activity" (normalized to the NCTR data set in this study) and the predicted activity data for each molecule in the test set and SD is the sum of squared deviations between the "actual activity" data for each molecule in the test set and the mean activity of the training set.

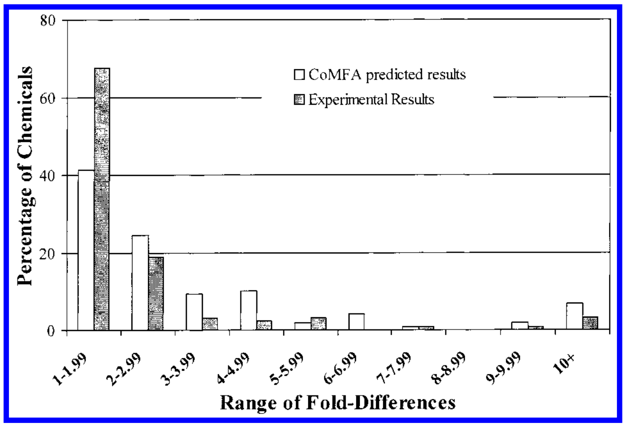

Figure 5. Fold differences for experimental measurements and CoMFA calculated results.

Table 2. Summary of Statistical Results for the CoMFA and HQSAR Models

statistics	CoMFA	HQSAR	statistics	CoMFA	HQSAR
q^2_{LOO}	0.655	0.585	SE	0.568	0.901
r^2	0.908	0.756	PCs	6	5

RESULTS

CoMFA Model. The CoMFA-calculated versus experimental RBAs (as logs) for the training-set compounds are plotted in Figure 4A and listed in Table 1. The conventional r^2 and cross-validated q^2_{LOO} were 0.91 and 0.66 (Table 2), respectively, indicating that the CoMFA model was both internally consistent and highly predictive. The steric and electrostatic field contributions to the CoMFA model were 43% and 57%, respectively, which were similar to those we reported for our earlier CoMFA models for much smaller ER ligand data sets.[11-13]

Each experimental datum for the NCTR data set is replicated at least twice. The experimental fold-difference for each data point can be calculated by dividing the highest value by the lowest value. Figure 5 shows two different distributions: (1) the range of fold-differences for experimental replications; (2) the range of fold-differences for CoMFA predicted and experimental means. It is apparent that the CoMFA prediction error is in a similar range as the experimental deviation. Predictions fell within 5-fold of experimental values for some 85% of the chemicals.

On the basis of the evaluation of 30 DES congeners, Sadler et al. reported significant improvement of the CoMFA model by applying the cross-validated r^2-guided region selection method.[18] It provides more weight on the CoMFA regions that show greater influence on the standard CoMFA PLS model, as determined by the PLS coefficients and CoMFA field variations for particular grids. The weighted CoMFA fields are calculated and used to correlate with biological activity. A similar approach, called region focusing, implemented in the Sybyl software was also applied for the NCTR data set. It appears that region focusing did not significantly improve the statistical results of the model (data not shown).

Table 3. Leave-*N*-Out Cross-Validation Results for CoMFA

no. of CV groups	cmpds left (%)	min q^2_{LNO}	max q^2_{LNO}	mean q^2_{LNO}	SD of q^2_{LNO}	mean no of PCs[a]
2	50	0.339	0.684	0.569	0.0540	4.4
3	33.3	0.512	0.671	0.607	0.0372	4.8
4	25	0.534	0.706	0.623	0.0350	5.0
5	20	0.528	0.676	0.623	0.0335	5.16
6	16.7	0.557	0.693	0.634	0.0288	5.2
7	14.3	0.587	0.683	0.637	0.0220	5.3
8	12.5	0.579	0.690	0.641	0.0229	5.4
9	11.1	0.595	0.689	0.645	0.0209	5.5
10	10	0.600	0.690	0.649	0.0188	5.5
13	7.7	0.583	0.686	0.647	0.0185	5.6
20	5	0.617	0.680	0.652	0.0126	5.5
50	2	0.630	0.674	0.656	0.0074	5.6
130	0.77	0.655	0.655	0.655	NA	6

[a] The maximum PC = 6.

Table 4. Leave-*N*-Out Cross-Validation Results for HQSAR

no. of CV groups	cmpds left (%)	min q^2_{LNO}	max q^2_{LNO}	mean q^2_{LNO}	SD of q^2_{LNO}	mean no. of PCs	mean HL[a]
2	50	0.276	0.652	0.503	0.0718	5	347
3	33.3	0.430	0.655	0.546	0.0535	5.4	380
4	25	0.445	0.637	0.565	0.0399	5.5	384
5	20	0.466	0.647	0.574	0.0343	5.6	382
6	16.7	0.444	0.640	0.572	0.0338	5.6	387
7	14.3	0.481	0.633	0.575	0.0298	5.5	386
8	12.5	0.496	0.645	0.576	0.0309	5.6	389
9	11.1	0.464	0.628	0.584	0.0271	5.5	394
10	10	0.524	0.631	0.584	0.0219	5.6	392
13	7.7	0.513	0.634	0.587	0.0216	5.5	397
20	5	0.514	0.620	0.585	0.0189	5.5	397
50	2	0.563	0.608	0.589	0.0098	5.35	401
130	0.77	0.585	0.585	0.585	NA	5	401

[a] Hologram lengths (HL) were set to "53 59 61 71 83 97 151 199 257 307 353 401", with fragment length 4−7, and only atom and bonds flags are turned on.

Table 5. Statistical Results of the CoMFA Models with or without Inclusion of the Phenolic Ring Indicator and the log *P* Descriptors

statistics	CoMFA	CoMFA with log *P* descriptors	CoMFA with phenolic indicator
q^2_{LOO}	0.655	0.648	0.707
r^2	0.908	0.880	0.903
SE	0.568	0.635	0.570
contributions (%)			
steric	43	41.4	44.4
electrostatic	57	56.3	48.8
log *P* and/or PhOH	NA	2.3	6.8
PCs	6	6	6

For a given set of aligned molecules, CoMFA results (q^2_{LOO}) have been reported to be largely dependent on the way in which these aligned molecules are placed in the CoMFA region.[34] To explore all the possible orientations and placements of aligned molecules in the CoMFA region, Wang et al. recently developed all-orientation search (AOS) and all-placement search (APS) methods.[35] These have been shown to be able to generate CoMFA models with q^2_{LOO} significantly higher than those obtained with standard CoMFA. However, no significant improvement of q^2_{LOO} was observed by applying AOS and APS to the NCTR data set.

To increase the precision and predictive ability of the QSAR models, several factors were considered, such as reducing the lattice spacing and/or the value of the filtering energy. Reducing these two parameters did improve the q^2_{LOO} value slightly but not enough to justify the extra computing time, which was consistent with our previous studies for several smaller data sets.[12]

HQSAR Model. The performance of an HQSAR model can be optimized by varying the fragment type and length parameters. The fragment type parameters determine the compositional and topological structural information encoded in the molecular hologram, while the fragment length parameter controls the minimum and maximum length of fragments to be included in the hologram. Through systematic investigation of these parameters, we used only elemental and bond-type information with the fragment length between 4 and 7 to construct the final model. The model results are shown in Figure 4B and Table 1. The model's q^2_{LOO} and r^2 values were lower than those for the CoMFA model.

Leave-*N*-Out Cross Validation. In addition to the standard LOO validation, extensive LNO validation was conducted for both CoMFA and HQSAR models. The predictive capability of a QSAR model for CoMFA and HQSAR is generally determined by q^2_{LOO} using the LOO cross-validation procedure. The q^2_{LOO} is primarily considered a measure of the ability of the model to interpolate within the training set population. Compared to the LOO procedure, the LNO procedure allows more chemicals to be omitted for prediction to test the model's stability. Thus, q^2_{LNO} accounts for more extrapolation of the model than does q^2_{LOO}. In contrast to q^2_{LOO}, q^2_{LNO} is varied for a selected *N* group in each run because of the random nature of selection of chemicals in the process. It is necessary to run LNO multiple times (100 times in the study) for each random *N* group for valid statistical analysis. Consequently, the standard deviation (SD) of q^2_{LNO} can be used to assess the model's stability of prediction for diverse chemicals. As shown in Tables 3 and 4, the mean q^2_{LNO} values decreased to about the same extent for both CoMFA and HQSAR models as the left-out compounds were increased. CoMFA generated better models than HQSAR for all left-out groups. Even when randomly leaving 33% of the training compounds out, the worst CoMFA model selected from randomly trying 100 times still gave a highly robust $q^2_{LNO} > 0.5$. Moreover, the SD of q^2_{LNO} was smaller for CoMFA than for HQSAR, indicating that CoMFA provided much more robust QSAR models for prediction of structurally diverse chemicals.

Enhancement of the CoMFA Model. The initial comparison between CoMFA and HQSAR for the NCTR data set demonstrated that the CoMFA model was more statistically robust. To further improve the quality of the CoMFA model, the effects of hydrophobic/hydrophilic characteristics and of a phenolic ring indicator were also evaluated in combination with the standard CoMFA field descriptors.

(i) **Inclusion of log *P*.** While the CoMFA descriptors encode information for steric/electrostatic distribution on the molecular surface, the absence of explicit representations that encode for hydrophobic/lipophilic balance reduces its ability to model many in vitro systems. To address this limitation, the CoMFA model was modified by supplementing values of log *P* (log of the octanol−water partition coefficient) with the CoMFA descriptors. As shown in Table 5, the model was not improved with inclusion of log *P* values. The

Table 6. Prediction Results of the CoMFA Models with and without PhOH and the HQSAR Model for Kuiper's Dataset

name	normali zed log RBA	CoMFA		CoMF A with PhOH		HQSAR	
		predicted	residual	predict ed	residual	predicted	residual
2,2′,3,3′,4′,5,5′-heptachloro-4-biphenylo l	−1.498	−2.031	0.533	−1.792	0.294	−1.686	0.188
2,2′,3,3′,4′,5-hexachloro-4-biphenylol	−1.650	−2.439	0.789	−1.920	0.270	−1.943	0.293
2,2′,3′,4,4′,5,5′-heptachloro-3-biphenylo l	−1.549	−2.164	0.615	−1.710	0.161	−1.770	0.221
2,2′,3,4′,5,5′6-heptachloro-4-biphenylol	−1.498	−2.194	0.696	−2.184	0.686	−1.640	0.142
2,2′,3,4′,5,5′-hexachloro-4-biphenylol	−2.023	−2.175	0.152	−2.219	0.196	−2.012	0.011
2,2′,3′,4′,5′-pentachloro-4-biphenylol	−1.498	−0.939	0.559	−0.820	0.678	−2.098	0.600
2,2′,3′,4′,6′-pentachloro-4-biphenylol	−1.014	−1.211	0.197	−1.293	0.279	−1.999	0.985
2,2′,3′,5′,6′-pentachloro-4-biphenylol	−1.549	−1.696	0.147	−1.187	0.362	−1.979	0.430
2,2′,4′,6′-tetrachloro-4-biphenylol	−1.014	−1.464	0.450	−1.376	0.362	−2.263	1.249
2′,3,3′,4′,5′-pentachloro-4-biphenylol	−1.458	−0.917	0.541	−1.076	0.382	−2.097	0.639
2,3,3′,4′,5-pentachloro-4-biphenylol	−2.023	−2.537	0.514	−2.536	0.513	−2.255	0.232
2′,3,3′,4′,5-pentachloro-4-biphenylol	−2.507	−2.252	0.255	−1.894	0.613	−2.212	0.295
2′,3,3′,4′,6′-pentachloro-4-biphenylol	−1.387	−1.264	0.123	−1.125	0.262	−1.971	0.584
2′,3,3′,5′,6′-pentachloro-4-biphenylol	−1.720	−1.500	0.220	−1.291	0.429	−1.947	0.227
2′,3,4′,6′-tetrachloro-4-biphenylol	−1.236	−1.506	0.270	−1.615	0.379	−2.238	1.002
2,4,6-trichloro-4′-biphenylol	−0.106	−1.604	1.498	−1.496	1.390	−2.307	2.201
5-androstenediol	−0.489	−0.658	0.169	−0.790	0.301	−2.450	1.961
16a-bromoestradiol	1.408	0.332	1.076	0.868	0.540	0.278	1.130
16-ketoestradiol	−0.378	0.582	0.960	1.104	1.482	0.264	0.642
17-epi-estriol	0.984	−0.158	1.142	−0.112	1.096	0.038	0.946
2-OH-estrone	−0.187	0.358	0.545	0.305	0.492	0.382	0.569
Raloxifene	1.367	−0.236	1.603	−0.521	1.888	−1.840	3.207
Zearalenone	0.368	−0.121	0.489	0.210	0.158	0.003	0.365
4,4′-biphenol	<−2.510	−1.805		−2.622		−2.473	
Ipriflavone	<−2.510	−4.252		−5.358		−2.712	
predictive q^2_{pred}		0.63		0.62		0.15	

Table 7. Prediction Results of the CoMFA Models with and without PhOH and the HQSAR Model for Waller's Data Set

name	normalized logRBA	CoMFA		CoMFA with PhOH		HQSAR	
		predicted	residual	predicted	residual	predicted	residual
2-*tert*-butylphenol	−4.546	−3.831	0.715	−3.946	0.600	−3.393	1.153
3-*tert*-butylphenol	−4.819	−3.225	1.594	−3.181	1.638	−3.009	1.810
2,4,6-trichloro-4′-biphenylol	−0.158	−1.604	1.446	−1.496	1.338	−2.307	2.149
2-chloro-4,4′-biphenyldiol	−0.610	−1.486	0.876	−1.532	0.922	−2.359	1.749
2,6-dichloro-4′-biphenylol	−1.110	−2.406	1.296	−1.905	0.795	−2.488	1.378
2,3,5,6-tetrachloro-4,4′-bipheny ldiol	−2.180	−0.815	1.365	−0.572	1.608	−1.528	0.652
2,2′,3,3′,6,6′-hexachloro-4-biph enylol	−2.739	−3.055	0.316	−1.917	0.822	−1.852	0.887
2,2′3,4′,6,6′-hexachloro-4-biph enylol	−2.596	−2.479	0.117	−1.985	0.611	−1.850	0.746
2,2′,3,6,6′-pentachloro-4-biphe nylol	−1.966	−3.073	1.107	−2.280	0.314	−2.031	0.065
2,2′,5,5′-tetrachlorobiphenyl	−2.667	−2.737	0.070	−3.956	1.289	−2.806	0.139
2,2′,4,4′,5,5′-hexachlorobiphe nyl	−2.834	−1.522	1.312	−3.282	0.448	−2.568	0.266
2,2′,4,4′,6,6′-hexachlorobiphe nyl	−1.870	−1.826	0.045	−2.982	1.111	−2.282	0.411
2,2′,3,3′,5,5′-hexachloro-6′-bip henylol	−2.691	−3.008	0.317	−2.176	0.515	−2.158	0.533
4′-deoxyindenestrol	−1.371	−0.526	0.845	−0.010	1.361	2.281	3.652
4′-deoxyindenestrol	−0.230	0.111	0.341	0.629	0.859	2.281	2.511
5′-deoxyindenestrol	−0.587	−0.999	0.413	−0.382	0.204	2.129	2.715
5′-deoxyindenestrol	0.353	−0.591	0.944	0.267	0.086	2.129	1.776
indenestrol A (*R*)	1.078	0.288	0.790	0.473	0.605	2.637	1.559
indenestrol A (*S*)	2.386	0.622	1.764	0.993	1.393	2.637	0.251
R5020	−1.811	−0.703	1.108	−1.413	0.398	−2.306	0.495
Zearalenone	0.912	−0.121	1.033	0.210	0.702	0.003	0.909
DACT	NA[a]	−5.255		−6.258		−3.642	
Hydroxyflutamide	NA	−1.041		−3.224		−4.049	
M1	NA	−2.550		−3.526		−3.761	
M2	NA	−4.353		−5.672		−3.766	
predictive q^2_{pred}		0.68		0.71		0.22	

[a] NA = no activity.

contribution of log P to the model was insignificant (2.3%), which was consistent with our previous observation for several smaller estrogen data sets.[12]

(ii) Inclusion of Phenol Indicator (PhOH). The crystal structure of the E_2−ER complex reveals that the hydroxyl group at the 3 position of the A-ring forms hydrogen bonds with Glu 353 and Arg 394 and a water molecule in the receptor binding site,[26] thus stabilizing the binding conformation. Moreover, crystal evidence suggests this binding action is exactly the same for the other three ligands, DES, 4-OH-tamoxifen, and raloxifene. That is, the crystal structure data are consistent with historical observation that a chemical with a phenolic ring structure is likely to exhibit estrogenic activity. Accordingly, a phenolic indicator (PhOH) for the presence or absence of a phenolic ring of a molecule was included in conjunction with the basic CoMFA descriptors. As shown in Table 5, the inclusion of PhOH significantly enhanced the quality of the model. Although the steric/

electrostatic field contributed predominately to the ER binding, the 6.8% contribution from PhOH confirms the appropriateness of inclusion of appreciable contribution of the phenolic functional group.

Prediction of the Test Sets. Estrogenic EDs cover a wide range of structurally diverse chemicals. The current challenge in developing QSAR models for ER binding is no longer in constructing a statistically robust model but in developing a model with the capability to accurately predict the activity of such structurally diverse estrogens. The three QSAR models, i.e., the CoMFA models with and without phenolic indicator as well as the HQSAR model, were compared in their ability to predict RBAs of two test sets. The predicted vs normalized activities for Waller's and Kuiper's data sets as well as the predictive $q^2_{\rm pred}$ results are listed in Tables 6 and 7, respectively. The CoMFA models generally provided better prediction than the HQSAR model, which was confirmed by much lower $q^2_{\rm pred}$ values of the HQSAR model for both test sets compared to those of the CoMFA models. Of 44 active chemicals, the number of predictions with the residuals larger than 1.0 (a 10-fold difference between predicted and normalized activity value) was 13, 11, and 16 for the CoMFA, CoMFA with PhOH, and HQSAR models, respectively. Moreover, 6 out of 16 predictions were off 100-fold or more for the HQSAR models. The CoMFA models with and without PhOH were comparable for prediction of active chemicals. However, the CoMFA with PhOH was better in discriminating inactive from active chemicals. For 6 chemicals that were reported to be inactive or to have undetectable activity in the original references, the CoMFA model with PhOH provided best estimation of their activities. 4,4′-Biphenol shows undetectable activity from the maximum experimentally determined limit in Kuiper's data set, while it is active with log RBA = −1.7 in the ER binding assay using mouse uterine cytosol.[18] All three models predicted it to be active with an average log RBA around −2.1.

DISCUSSION

Two QSAR methods, CoMFA and HQSAR, were evaluated for their ability to predict ER binding of chemicals based on a larger structurally diverse data set. The resultant QSAR models were compared with respect to the statistical measures from the leave-one-out and leave-*N*-out cross-validation processes. The CoMFA yielded the best QSAR models in terms of self-consistency and ability to predict test chemicals. Furthermore, the results showed that it was advantageous to include the phenolic indicator (PhOH) in the CoMFA model. The CoMFA model with PhOH was not only internally robust but also provided the best activity predictions for both active and inactive chemicals.

The rate of false positives and false negatives should be considered for application of QSAR models for priority setting of a large number of environmental chemicals. The positives or negatives are defined for the chemicals whose activity values are larger or less than a predefined cutoff activity value, respectively. The criteria to use for determining a cutoff are dependent on the nature of application. For drug discovery, false positives are of primary concern because of the cost to bring such a chemical with a low probability of being efficacious to the development phase. A relativly higher cutoff value could be used for such application. In contrast, minimizing false negatives is more important for regulatory purpose. Once a toxicant is labeled as a low priority, it will cause more potential threat to public health than the one without such a label, even though it might show active at lower doses. With application of the present QSAR models for the prediction of 42 unique active chemicals from Waller's and Kuiper's data sets, no false negatives were observed, even for the chemicals with activities 1 million-fold below that of E_2. This indicates that the QSAR models reported in the study, particularly the CoMFA model with PhOH, have potential utility for regulatory purposes.

The utility of CoMFA has been demonstrated in a wide range of applications.[36−39] Since CoMFA employs steric and electrostatic field descriptors that encode detailed information concerning intermolecular interaction in three dimensions, it is able to provide the best model by capturing the salient features associated with molecular recognition in ER binding. However, aligning molecules requires substantial chemical and biological knowledge and may be too time-consuming to allow processing of a large number of diverse structures in CoMFA. This difficulty limits the potential usage of CoMFA for screening tens of thousands of chemicals. In the current form of our integrated four-phase approach, the CoMFA model with PhOH is being applied in phase III to provide quantitative predictions for chemicals passing through phases I and II. To optimize the efficiency of phase III, a suite of computational models in phase I and II with rapid screening capability was used to eliminate the majority of chemicals that are the most unlikely to bind to the ER. Care was exercised to minimize false negatives when constructing the first two phases. The reduced data set is now tractable for CoMFA prediction in phase III.

The use of CoMFA and other QSAR models can likely be used to predict activity of chemicals that may act by other mechanisms, such as androgen or other receptor binding, or mechanisms with more biological complexity. This provides the possibility of developing a suite of models for predicting multiple activities of a single chemical and/or linking models through a mechanism sequence to predict activity at a downstream event. This more comprehensive suite of models provides an additional alternative for priority setting of potential EDs.

ACKNOWLEDGMENT

The authors acknowledge the Chemical Manufacturers Association (CMA) for partial financial support of the project. H.F. and R.M.B. gratefully acknowledge the Oak Ridge Institute for Science and Education Program, supported by the U.S. Department of Energy (DOE) and the U.S. Food and Drug Administration (FDA), for postdoctoral support. This work was also partially supported by the FDA's Office of Women's Health.

REFERENCES AND NOTES

(1) 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.; Tilson, H. A. 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.* **1996**, *104, Suppl 4,* 715−740.

(2) Cooper, R. L.; Kavlock, R. J. Endocrine disruptors and reproductive development: a weight-of-evidence overview. *J. Endocrinol.* **1997**, *152*, 159−166.

(3) Colborn, T. Environmental estrogens: health implications for humans and wildlife. *Environ. Health Perspect.* **1995**, *103, Suppl 7*, 135−136.

(4) Colborn, T.; vom Saal, F. S.; Soto, A. M. Developmental effects of endocrine-disrupting chemicals in wildlife and humans [see comments]. *Environ. Health Perspect.* **1993**, *101*, 378−384.

(5) U.S. Congress. The Food Quality Protection Act (FQPA) and the Safe Drinking Water Act (SDWA), 1996.

(6) EDSTAC. http://www.epa.gov/opptintr/opptendo/finalrpt.htm.

(7) Shi, L. M.; Tong, W.; Fang, H.; Perkins, R.; Wu, J.; Tu, M.; Blair, R.; Branham, W.; Waller, C.; Sheehan, D. An integrated “Four-Phase” approach for priority setting of endocrine disruptors−Part 1: Phase I and II for prediction of potential estrogenic endocrine disruptor. *SAR QSAR Environ. Res.* **2000**, submitted for publication.

(8) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(9) Walker, J. D.; Waller, C. W.; Kane, S. *The Endocrine Disruption Priority Setting Database (EDPSD): A Tool to Rapidly Sort and Prioritize Chemicals for Endocrine Disruption Screening and Testing*; Walker, J. D., Ed.; SETAC: 2000.

(10) Hansch, C.; Leo, A. Exploring QSAR−Fundamentals and applications in chemistry and biology, American Chemical Society: Washington, DC, 1995.

(11) Tong, W.; Perkins, R.; Strelitz, R.; Collantes, E. R.; Keenan, S.; Welsh, W. J.; Branham, W. S.; Sheehan, D. M. Quantitative structure−activity relationships (QSARs) for estrogen binding to the estrogen receptor: predictions across species. *Environ. Health Perspect.* **1997**, *105*, 1116−1124.

(12) Tong, W.; Perkins, R.; Xing, L.; Welsh, W. J.; Sheehan, D. M. QSAR models for binding of estrogenic compounds to estrogen receptor alpha and beta subtypes. *Endocrinology* **1997**, *138*, 4022−4025.

(13) Tong, W.; Lowis, D. R.; Perkins, R.; Chen, Y.; Welsh, W. J.; Goddette, D. W.; Heritage, T. W.; Sheehan, D. M. Evaluation of quantitative structure−activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 669−677.

(14) Bradbury, S.; Mekenyan, O.; GT, A. Quantitative structure−activity relationships for polychlorinated hydroxybiphenyl estrogen receptor binding affinity- An assessment of conformer flexibility. *Environ. Toxicol. Chem.* **1996**, *15*, 1945−1954.

(15) Waller, C. L.; Oprea, T. I.; Chae, K.; Park, H. K.; Korach, K. S.; Laws, S. C.; Wiese, T. E.; Kelce, W. R.; Gray, L. E., Jr. Ligand-based identification of environmental estrogens. *Chem. Res. Toxicol.* **1996**, *9*, 1240−1248.

(16) Wiese, T. E.; Polin, L. A.; Palomino, E.; Brooks, S. C. Induction of the estrogen specific mitogenic response of MCF-7 cells by selected analogues of estradiol-17 β: a 3D QSAR study. *J. Med. Chem.* **1997**, *40*, 0, 3659−3669.

(17) Xing, L.; Welsh, W. J.; Tong, W.; Perkins, R.; Sheehan, D. M. Comparison of estrogen receptor alpha and beta subtypes based on comparative molecular field analysis (CoMFA). *SAR QSAR Environ. Res.* **1999**, *10*, 215−237.

(18) Sadler, B. R.; Cho, S. J.; Ishaq, K. S.; Chae, K.; Korach, K. S. Three-dimensional quantitative structure−activity relationship study of nonsteroidal estrogen receptor ligands using the comparative molecular field analysis/cross-validated r^2-guided region selection approach. *J. Med. Chem.* **1998**, *41*, 2261−2267.

(19) Zheng, W.; Tropsha, A. A novel variable selection QSAR approach based on the k-nearest neighbor principle. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 185−194.

(20) Blair, R.; Fang, H.; Branham, W. S.; Hass, B.; Dial, S. L.; Moland, C. L.; Tong, W.; Shi, L.; Perkins, R.; Sheehan, D. M. Estrogen receptor relative binding affinities of 188 natural and xenochemicals: structural diversity of ligands. *Toxicol. Sci.* **2000**, *54*, 138−153.

(21) Branham, W. S.; Dial, S. L.; Moland, C. L.; Hass, B.; Blair, R.; Fang, H.; Shi, L.; Tong, W.; Perkins, R.; Sheehan, D. M. Phytoestrogen and mycoestrogen binding to rat uterine estrogen receptor. *Am. J. Nutr.* **2000**, in press.

(22) Perkins, R.; Anson, J.; Branham, W.; Fang, H.; Tong, W.; Welsh, W.; Chen, Y.; Meehan, J.; Jackson, M.; Nossaman, R.; Shi, L.; Sheehan, D. *The Estrogen Knowledge Base (EKB), A Prototype Toxicological Knowledge Base for Endocrine Disrupting Compounds*; Walker, J. D., Ed.; SETAC: 2000.

(23) Fang, H.; Tong, W.; Perkins, R.; Soto, A.; Prechtl, N.; Sheehan, D. M. Quantitative comparison of in vitro assays for estrogenic activity. *Environ. Health Perspect.* **2000**, *108*, 723−729.

(24) Zacharewski, T. Identification and assessment of endocrine disruptors: limitations of in vivo and in vitro assays. *Environ. Health Perspect.* **1998**, *106, Suppl 2*, 577−582.

(25) Kuiper, G. G.; Lemmen, J. G.; Carlsson, B.; Corton, J. C.; Safe, S. H.; van der Saag, P. T.; van der Burg, B.; Gustafsson, J. A. Interaction of estrogenic chemicals and phytoestrogens with estrogen receptor beta. *Endocrinology* **1998**, *139*, 4252−4263.

(26) Brzozowski, A. M.; Pike, A. C.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; Engstrom, O.; Ohman, L.; Greene, G. L.; Gustafsson, J. A.; Carlquist, M. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* **1997**, *389*, 753−758.

(27) Shiau, A. K.; Barstad, D.; Loria, P. M.; Cheng, L.; Kushner, P. J.; Agard, D. A.; Greene, G. L. The structural basis of estrogen receptor/coactivator recognition and the antagonism of this interaction by tamoxifen. *Cell* **1998**, *95*, 927−937.

(28) von Angerer, E. *The estrogen receptor as a target for rational drug design*; R.G. Landes Co.: Georgetown, TX, 1995.

(29) Anstead, G. M.; Carlson, K. E.; Katzenellenbogen, J. A. The estradiol pharmacophore: ligand structure-estrogen receptor binding affinity relationships and a model for the receptor binding site. *Steroids* **1997**, *62*, 268−303.

(30) Bucourt, R.; Vignau, M.; Torelli, V. New Biospecific adsorbents for the purification of estradiol receptor. *J. Biol. Chem.* **1978**, *253*, 8221−8228.

(31) Fang, H.; Tong, W.; Shi, L.; Blair, R.; Perkins, R.; Branham, W. S.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. Structure activity relationship for a large diverse set of natural, synthetic and environmental chemicals. *Chem. Res. Toxicol.* **2000**, in press.

(32) Cramer, R. D. I.; Bunce, J. D.; Patterson, D. E. Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies. *QSAR* **1988**, *7*, 18−25.

(33) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(34) Cho, S. J.; Tropsha, A. Cross-validated R^2-guided region selection for comparative molecular field analysis: a simple method to achieve consistent results. *J. Med. Chem.* **1995**, *38*, 1060−1066.

(35) Wang, R.; Gao, Y.; Lin, L.; Lai, L. All orientation search and all-placement search in Comparative Molecular Field Analysis. *J. Mol. Model.* **1998**, *4*, 276−283.

(36) Tong, W.; Collantes, E. R.; Chen, Y.; Welsh, W. J. A comparative molecular field analysis study of *N*-benzylpiperidines as acetylcholinesterase inhibitors. *J. Med. Chem.* **1996**, *39*, 380−387.

(37) Tong, W.; Collantes, E. R.; Welsh, W. J.; Berglund, B. A.; Howlett, A. C. Derivation of a pharmacophore model for anandamide using constrained conformational searching and comparative molecular field analysis. *J. Med. Chem.* **1998**, *41*, 4207−4215.

(38) Welsh, W. J.; Tong, W. D.; Collantes, E. R.; Chickos, J. S.; Gagarin, S. G. Enthalpies of sublimation and formation of polycyclic aromatic hydrocarbons (PAHs) derived from comparative molecular field analysis (CoMFA)-Application of moment of inertia for molecular alignment. *Thermochim. Acta* **1997**, *290*, 55−64.

(39) Welsh, W.; Tong, W.; Collantes, E. Heats of sublimition and formation of polycyclic aromatic hydrocarbons (PAHs) derived from comparative molecular field analysis (CoMFA): Application of moment of inertia for molecular alignment. *Thermochim. Acta* **1996**, *290*, 55−64.

CI000066D