Influence of the Structural Diversity of Data Sets on the Statistical Quality of Three-Dimensional Quantitative Structure–Activity Relationship (3D-QSAR) Models: Predicting the Est...

ARTICLE *in* CHEMICAL RESEARCH IN TOXICOLOGY · NOVEMBER 2002

Impact Factor: 3.53 · DOI: 10.1021/tx0255875 · Source: PubMed

4 AUTHORS, INCLUDING:

Susan Keenan
University of Northern Colorado
38 PUBLICATIONS **792** CITATIONS
SEE PROFILE

Weida Tong
U.S. Food and Drug Administration
244 PUBLICATIONS **9,278** CITATIONS
SEE PROFILE

William Welsh
Rutgers, The State University of New Jersey
139 PUBLICATIONS **3,667** CITATIONS
SEE PROFILE

Influence of the Structural Diversity of Data Sets on the Statistical Quality of Three-Dimensional Quantitative Structure−Activity Relationship (3D-QSAR) Models: Predicting the Estrogenic Activity of Xenoestrogens

Seong Jae Yu,[†] Susan M. Keenan,[†] Weida Tong,[‡] and William J. Welsh*[,†]

Department of Pharmacology, University of Medicine & Dentistry of New Jersey, Robert Wood Johnson Medical School, 675 Hoes Lane, Piscataway, New Jersey 08854, and National Center for Toxicological Research, 3900 NCTR Road, Jefferson, Arkansas 72079

Received July 30, 2002

Federal legislation has resulted in the two-tiered in vitro and in vivo screening of some 80 000 structurally diverse chemicals for possible endocrine disrupting effects. To maximize efficiency and minimize expense, prioritization of these chemicals with respect to their estrogenic disrupting potential prior to this time-consuming and labor-intensive screening process is essential. Computer-based quantitative structure−activity relationship (QSAR) models, such as those obtained using comparative molecular field analysis (CoMFA), have been demonstrated as useful for risk assessment in this application. In general, however, CoMFA models to predict estrogenicity have been developed from data sets with limited structural diversity. In this study, we constructed CoMFA models based on biological data for a structurally diverse set of compounds spanning eight chemical families. We also compared two standard alignment schemes employed in CoMFA, namely, *atom-fit* and *flexible field-fit*, with respect to the predictive capabilities of their respective models for structurally diverse data sets. The present analysis indicates that *flexible field-fit* alignment fares better than *atom-fit* alignment as the structural diversity of the data set increases. Values of log(RP), where RP = relative potency, predicted by the final *flexible field-fit* CoMFA models are in good agreement with the corresponding experimental values. These models should be effective for predicting the endocrine disrupting potential of existing chemicals as well as prospective and newly prepared chemicals before they enter the environment.

Introduction

Certain man-made and naturally occurring chemicals have been shown to bind to the estrogen receptor (ER)[1] (*1*). As a member of the superfamily of nuclear hormone receptors, the ER functions as a DNA transcription factor. Activation of estrogen response elements (EREs) by estrogen receptor homodimers leads to the synthesis of new gene products important for most notably sexual reproduction and differentiation (*2*). Unregulated ER activation by exogenous chemicals has been shown to disrupt the delicate endocrine balance of humans and of other species as divergent as reptiles and fish (*3*). Many of the tens of thousand of chemicals in use today require screening for possible endocrine disrupting properties. This number will surely continue to increase by virtue of modern technologies such as combinatorial chemistry (*4*) and high throughput screening (*5*) associated with industrial drug-discovery programs (*6*, *7*).

The classic and still standard methods for measuring estrogenicity are based on time-consuming and labor-intensive in vitro and/or in vivo assays which are not suitable for routine screening of large numbers of chemicals. Without a more expeditious in vitro protocol for the analysis of endocrine disruption, methods must be developed to predict the endocrine disrupting potential of chemicals and thus enable scientists to establish a priority list to direct the further testing of potential endocrine disrupting compounds (EDCs). Molecular modeling has emerged as a valuable tool for the prediction of biological activities. In particular, use of quantitative structure−activity relationship (QSAR) models to predict the endocrine disrupting properties of compounds will yield significant savings in time and expense and ultimately provide improved environmental protection (*8*, *9*).

Estrogenic EDCs are structurally diverse and span a wide range of chemical families; therefore, QSAR models need to be equally inclusive in terms of structural diversity. Previous CoMFA models for predicting estrogenicity have been constructed largely from data sets with limited structural diversity (*9*). One exception is the CoMFA model developed by Waller et al. for a series of structurally diverse estrogenic compounds (*10*) using the SEAL (Steric and Electrostatic Alignment) (*11*) alignment scheme. In this study, we present CoMFA models based on a structurally diverse data set that spans eight

* To whom correspondence should be addressed. E-mail: welshwj@umdnj.edu.
[†] Robert Wood Johnson Medical School.
[‡] National Center for Toxicological Research.

[1] Abbreviations: CoMFA, comparative molecular field analysis; DHT, dihydrotestosterone; E_2, 17β-estradiol; EDC, endocrine disrupting chemical; ER, estrogen receptor; ERE, estrogen response element; PC, principal component; PDB, Protein DataBank; PXR, pregnane xenobiotic receptor; QSAR, quantitative structure−activity relationship; RP, relative potency; SXR, steroid and xenobiotic receptor.

10.1021/tx0255875 CCC: $22.00
Published on Web 10/01/2002

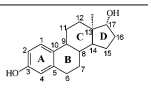
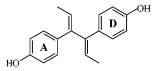
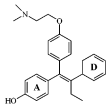
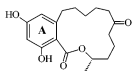
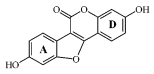
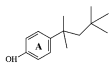
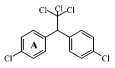
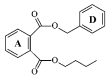
Chemical Class	Representative Compound	Structure
Steroid 15(9)	17β-Estradiol	
Synthetic Estrogen 5(5)	Diethylstilbestrol	
Antiestrogen 3(2)	Tamoxifen	
Lactone 5(4)	Zearalenone	
Phytoestrogen 6(5)	Coumestrol	
Alkylpheol 3(3)	4-tert-octylphenol	
Organochlorine 11(10)	DDT	
Other chemicals 5(2)	Butylbenzylphthalate	

Figure 1. Structures and basic alignment scheme of chemicals in the training set. The first number in column 1 refers to the total number of representatives of that particular chemical class in the data set, while the number in parentheses refers to those representatives that exhibited measurable biological activity and, thus, were used for model development.

chemical families (Figure 1). In addition, we explore the influence of the size and structural diversity of the data set on the statistical quality and predictive capability of the resulting 3D-QSAR models, using the two most frequently employed alignment schemes of *atom-fit* and *flexible field-fit*. The present results for estrogenic EDCs indicate that *flexible field-fit* exhibits improved performance over *atom-fit* as the size and structural diversity of the data set increase. The advantage of field-based over atom-based alignment schemes is a logical consequence of the known promiscuity of ER α and *β*. Other receptors that share this propensity for binding ligands from disparate chemical families, such as the human steroid and xenobiotic receptor (SXR) (*12*, *13*) found in humans and its pregnane xenobiotic receptor (PXR) orthologue found in other mammalian species (*14*), would be expected to follow a similar pattern in terms of their corresponding 3D-QSAR models. Such models that can maintain their statistical quality and predictive ability even for structurally diverse data sets are highly desirable in these cases. This is particularly critical in risk assessment scenarios where QSAR and 3D-QSAR models are expected to predict the endocrine disrupting effects of chemicals with virtually zero tolerance for false negatives.

Experimental Procedures

Data Sets for Analysis. The data set consisted of 53 compounds from 8 structurally diverse chemical families (Figure 1). Values of the relative potency (RP) of the chemicals in the data set were obtained from a yeast-based reporter gene assay. The recombinant yeast cell bioassay (RCBA) utilized in this study is a highly sensitive human estrogen receptor-based assay

Table 1. Chemical Classes Represented in the Respective Training Sets for CoMFA Models 1−5

CoMFA models	chemical class
model 1	steroids
model 2	steroids and lactones
model 3	steroids, lactones, and phytoestrogens
model 4	steroids, lactones, phytoestrogens, and organochlorines
model 5	all of the above chemical classes plus two "other" compounds

the likes of which have been used previously for the detection of xenoestrogens (*15*) and in various estrogen receptor studies (*16*). While the RCBA yields potency values (relative to E_2) similar to literature values, the assay exhibits greater sensitivity than other in vitro and in vivo assays. By virtue of its increased sensitivity, the assay was able to measure detectable values of partial agonist activity of the ER antagonists tamoxifen and 4-hydroxytamoxifen. Consequently, we decided to include both tamoxifen and 4-hydroxytamoxifen in the data set for model building. Another justification for inclusion of these two compounds is due to the fact that the RCBAs are based on the expression of the human ER; thus, the present 3D-QSAR models take into account the potential estrogenic effects these compounds have on humans. RP was defined as 100 times the ratio of the concentration of 17*β*-estradiol (E_2) giving 50% induction in *β*-galactosidase activity (EC_{50}) and the EC_{50} of the tested compounds. Using this scheme, the RP of E_2 equals 100 (*17*). As is standard in QSAR studies, we excluded eight compounds from the data set as these compounds showed undetectable estrogenic activity under the assay conditions. This pruned data set consisted of 10 steroids, 5 synthetic estrogens, 2 antiestrogens, 5 lactones, 6 phytoestrogens, 3 alkylphenols, 11 organochlorines, and 2 "other" chemicals not belonging to any of these classes. We divided the data set into a training set containing 40 compounds (Table 2) used for model development, and a test set of 4 randomly selected compounds (Table 4) used for model validation.

Molecular Modeling. All molecular modeling and statistical analyses were performed on a Silicon Graphics O_2 workstation running under the IRIX 6.5 operating system using Sybyl 6.7 (Tripos, St. Louis, MO) (*18*). The crystal structures of the ER-bound ligands E_2 (1A52) (*19*) and diethylstilbestrol (3ERD) (*20*), extracted from the Protein DataBank (PDB), were used as structural templates to build the steroids and synthetic estrogens. Molecular structures for the remaining compounds were constructed from the Sybyl 6.7 fragment database (*18*), after which they were energy-minimized to the putative global low-energy conformation. Using the standard Tripos molecular force field with a distance-dependent (1/*r*) dielectric function, molecules were first geometry-optimized to the nearest local minimum-energy conformation until an energy difference of 0.001 kcal/mol between successive iterations was achieved. All rotatable (single) bonds were then systematically searched in 10° increments, and, after setting each torsion angle to its minimum-energy conformation, the molecule was energy minimized a final time. Atomic partial charges were computed using the Gasteiger−Hückel method (*21*).

CoMFA Alignment. The initial and arguably the most important step in CoMFA model development is the alignment of molecules to a template. The compounds must be aligned to ensure maximal superimposition of their steric and electrostatic fields. As our data set is structurally diverse and contains compounds belonging to several different chemical families, there are many possible alignment schemes. In this study, we compared two different alignment schemes, namely, *atom-fit* and *flexible field-fit*, in terms of the statistical quality and predictive ability of their respective CoMFA models. Both alignment schemes used E_2 as the template molecule. Using *atom-fit*, which aligns a set of molecules by a rigid least-squares fit of preselected common atoms in each molecule, and *flexible field-fit*, which aligns a set of molecules by seeking each

Table 2. Comparison of Experimentally Observed and CoMFA-Predicted Activities [log(RP)] Using Model 5, Based on *Atom-Fit* and *Flexible Field-Fit* Alignment Schemes for the 40 Training Set Compounds

		CoMFA-predicted (model 5)			
chemical name	exptl	*atom-fit*	residual	*field-fit*	residual
17β-estradiol	2.00	0.79	1.21	0.91	1.09
17β-estradiol-3(β-D-glucuronide)	−0.50	−0.60	0.10	−0.84	0.34
17β-estradiol-3-sulfate	−2.00	−1.99	−0.01	−2.07	0.07
17α-estradiol	0.72	0.86	−0.14	0.77	−0.05
estriol	−0.20	0.88	−1.08	0.62	−0.82
testosterone	−3.00	−2.06	−0.94	−2.51	−0.49
androstenediol	−1.64	−2.10	0.46	−2.34	0.70
dehydroepi-androsterone	−2.74	−2.76	0.02	−2.41	−0.33
D-norgestrel	−3.40	−2.98	−0.42	−3.46	0.06
17α-ethylnylstradiol	1.95	1.18	0.77	1.64	0.31
mestranol	0.86	1.34	−0.48	1.63	−0.77
diethylstilbestrol	1.87	2.37	−0.50	1.85	0.02
hexestrol	1.49	1.78	−0.29	1.37	0.12
dienestrol	1.40	1.45	−0.04	1.26	0.15
tamoxifen	−2.33	−2.10	−0.23	−2.44	0.11
4-hydroxytamoxifen	−2.14	−1.83	−0.31	−2.47	0.33
α-zearalenol	0.94	0.69	0.25	1.03	−0.09
β-zearalenol	−1.18	−0.28	−0.90	−1.18	−0.00
α-zearalanol (zeranol)	0.11	0.28	−0.17	0.08	0.03
β-zearalanol	−0.34	−0.62	0.28	−0.29	−0.05
coumestrol	−0.17	−0.35	0.18	−0.04	−0.13
equol	−1.07	−1.42	0.35	−0.89	−0.18
daidzein	−2.89	−2.32	−0.57	−2.41	−0.48
formononetin	−2.25	−2.08	−0.17	−2.23	−0.03
genistein	−1.31	−1.97	0.66	−1.50	0.19
4-nonylphenol	−2.66	−2.59	−0.07	−3.02	0.36
4-octylphenol	−2.52	−2.52	−0.00	−2.77	0.25
4-*tert*-octylphenol	−3.44	−2.43	−1.01	−2.93	−0.51
DDT	−4.52	−4.38	−0.14	−4.27	−0.25
o,p′-DDT	−3.96	−4.30	0.34	−4.17	0.21
o,p′-DDE	−4.40	−3.90	−0.50	−4.03	−0.37
2,3,7,8-tetrachloro-dibenzo-*p*-dioxin	−0.59	−0.49	−0.10	−0.53	−0.05
4′-chloro-4-biphenylol	−1.22	−1.98	0.76	−1.53	0.31
2′-chloro-4-biphenylol	−2.43	−2.50	0.07	−2.26	−0.17
2′,5′-dichloro-4-biphenylol	−0.21	−0.78	0.57	−0.34	0.13
2′,4′,6′-trichloro-4-biphenylol	0.00	−0.85	0.85	0.05	−0.05
2′,3′,4′,5′-tetra-chloro-4-biphenylol	−0.09	−0.79	0.70	0.11	−0.19
3,3′,5,5′-tetrachloro-4,4′-biphenyldiol	−1.80	−1.51	−0.29	−1.52	−0.28
bisphenol A	−2.30	−2.82	0.52	−2.72	0.42
butylbenzylphthalate	−3.40	−3.66	0.26	−3.50	0.10

molecule's conformation that most closely resembles the steric and electrostatic fields of the template molecule, we aligned the training-set compounds with respect to the 3-hydroxylphenyl group of E_2. The *flexible field-fit* procedure sometimes distorts the original molecular structure, and, therefore, each compound was reminimized after field-fitting. The alignment schemes based on *atom-fit* and *flexible field-fit* are illustrated in Figure 2A and Figure 2B, respectively.

Calculation of CoMFA Descriptors. The calculation of CoMFA steric and electrostatic descriptors has been described in detail previously (*8*, *9*). Briefly, following alignment, the molecules were placed in a three-dimensional cubic lattice with 2 Å spacing. Steric (van der Waals) and electrostatic (Coulombic) field descriptors were calculated for each molecule at all lattice points using a probe represented by an sp^3-hybridized carbon atom with a +1.0 charge. The steric and electrostatic energy values were truncated to 30 and ±30 kcal/mol, respectively. The CoMFA field descriptors were scaled using the CoMFA standard scaling method provided in Sybyl 6.7.

Statistical Regression Methods. The biological activity for the 40-compound training set was correlated with the CoMFA-generated steric and electrostatic fields using the statistical method of partial least-squares (PLS) regression (*22*). Using PLS, the large number of steric-electrostatic descriptors was reduced to a few principal components (PCs) that are linear combinations of the original descriptors. The optimum number of PCs was determined by the Leave-One-Out (LOO) cross-validation procedure (*23*). In this method, each compound is systematically excluded once from the training set, after which its activity is predicted by a model derived from the remaining compounds. Using the optimal number of PCs, the final PLS analysis was carried out without cross-validation to generate a predictive QSAR model with a conventional correlation coefficient (*23*).

Results

A data set containing 53 structurally diverse compounds was chosen for the present study with the goal of analyzing the effectiveness of *atom-fit* and *flexible field-fit* alignment schemes for model generation as the size and structural diversity of the data set increase. Unlike most other CoMFA models associated with ER ligands that are based on ligand−receptor binding data, the present models were constructed from a functional assay whose high sensitivity was such that RP values associated with the partial agonist activity of the prototypical antagonists tamoxifen and 4-hydroxytamoxifen were included for model building. Values of the relative potency (RP) for the data set of compounds were converted to log(RP) values for the construction of the 3D-QSAR models. After eliminating compounds which had no estrogenic activity under assay conditions, the data set was divided into a training set of 40 compounds (Table 2) and a test set of 4 randomly selected compounds (Table 4) for model validation. We applied two separate alignment rules for CoMFA model generation. First, we aligned the training set using an *atom-fit* alignment scheme to the 3-hydroxylphenyl group of E_2 (Figure 2A). Second, we applied a *flexible field-fit* alignment scheme by aligning the 3D steric and electrostatic field energies to the E_2 template (Figure 2B).

CoMFA model 1 was developed using only the nine steroid compounds. Although this particular data set is admittedly sparse as compared to typical QSAR applications, it is justified in the present case as a benchmark for assessing the impact of gradually expanding the size and structural diversity of the data set. As all steroid molecules embody a four-ring system (see Figure 1) which is amenable to superimposition using either alignment scheme, it is not surprising that the models developed from *atom-fit* and *flexible field-fit* alignment schemes were almost equally satisfactory insofar as their calculated statistical parameters ($r_{cv}^2 = 0.576$ vs 0.548; $r^2 = 0.893$ vs 0.902). Indeed, both models were successful in predicting the activity of the steroid estrone included in the test set (Table 4). CoMFA model 2 was developed from the same nine steroids and, in addition, the four lactones in the training set. The r_{cv}^2 statistical parameter for CoMFA model 2, using either *atom-fit* or *flexible field-fit* alignment, was diminished compared with CoMFA model 1. Nevertheless, it is noteworthy that *flexible-field fit* fared better than *atom-fit* and performed better in absolute terms ($r_{cv}^2 = 0.490$ vs 0.442). This advantage of *flexible field-fit* over *atom-fit* became more apparent as the size and structural diversity of the training set were gradually expanded by inclusion of 5 phytoestrogens (model 3) and, then, by 10 organochlorines (model 4) (Table 3).

Table 3. Summary of Statistical Parameters from CoMFA Models 1−5[a]

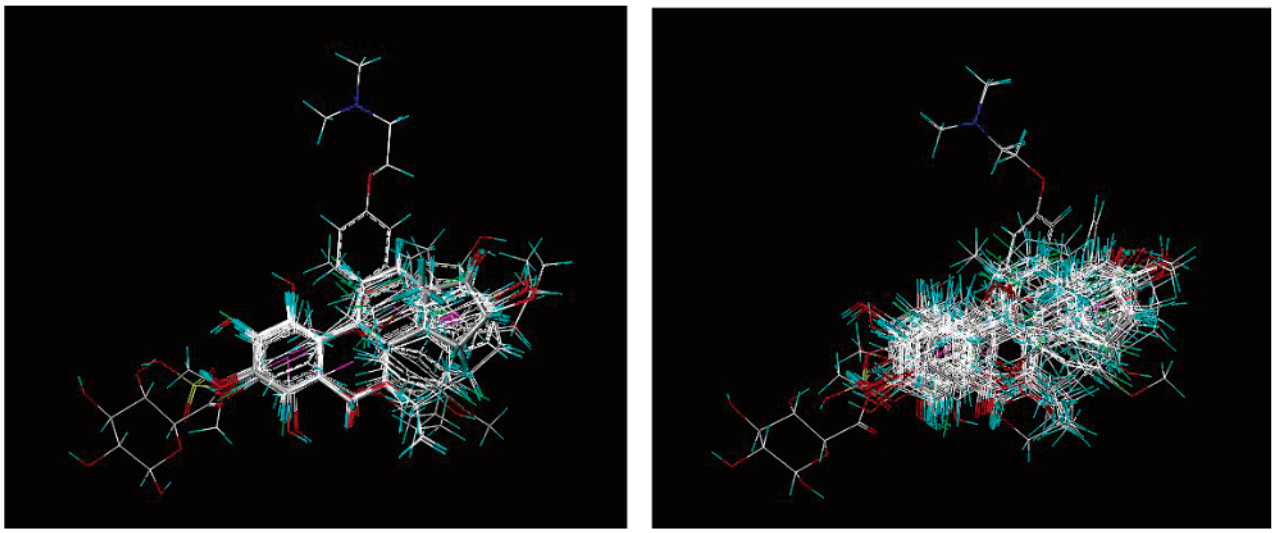
	model 1		model 2		model 3		model 4		model 5	
alignment:	*atom-fit*	*field-fit*	*atom-fit*	*field-fit*	*atom-fit*	*field-fit*	*atom-fit*	*field-fit*	*atom-fit*	*field-fit*
no. of compounds	9	9	13	13	18	18	28	28	40	40
no. of PCs	4	4	5	5	5	5	3	3	5	6
r_{cv}^2	0.576	0.548	0.442	0.490	0.492	0.565	0.508	0.513	0.446	0.533
r^2	0.893	0.902	0.904	0.944	0.868	0.933	0.920	0.895	0.913	0.960

[a] The size and structural diversity of the training sets increase from model 1 through model 5.

Table 4. Values of Experimentally Observed vs CoMFA-Predicted log(RP) for the Test Set Compounds[a]

			log(RP) for test set compounds							
			estrone obs = 0.98		zearalenone obs = −0.59		biochanin A obs = −2.04		methoxychlor obs = −2.48	
	no. of compounds		pred	dev[b]	pred	dev[b]	pred	dev[b]	pred	dev[b]
model 1	9	*atom-fit*	0.69	0.29	(−0.72)	(0.13)	(−0.07)	(−1.97)	(−1.43)	(−1.05)
		field-fit	0.47	0.51	(0.81)	(−1.40)	(−0.97)	(1.07)	(−0.39)	(−2.09)
model 2	13	*atom-fit*	0.57	0.41	−0.48	−0.11	(0.68)	(−2.72)	(−1.51)	(−0.97)
		field-fit	0.63	0.35	0.35	−0.94	(−0.17)	(−1.87)	(−0.48)	(−2.00)
model 3	18	*atom-fit*	0.65	0.33	−0.15	−0.44	−2.09	0.05	(−1.92)	(−0.56)
		field-fit	0.72	0.26	0.39	−0.98	−2.13	0.09	(−0.51)	(−1.97)
model 4	28	*atom-fit*	0.22	0.76	−0.08	−0.51	−2.26	0.22	−4.02	1.54
		field-fit	0.44	0.54	0.10	−0.69	−2.21	0.17	−2.14	−0.34
model 5	40	*atom-fit*	0.41	0.57	0.55	−1.14	−1.69	−0.35	−2.99	0.51
		field-fit	0.79	0.19	−0.26	−0.33	−1.99	−0.05	−2.75	0.27

[a] CoMFA models 1−5, derived from training sets that gradually increase both in size and in structural diversity (see Table 3), were constructed using both *atom-fit* and *flexible field-fit* alignment schemes. [b] The deviation (dev) refers to the difference between the corresponding experimentally observed and CoMFA-predicted log(RP) value obtained from the *atom-fit* and *flexible field-fit* alignment schemes. The values highlighted in parentheses refer to predictions made for compounds that fall outside of the chemical families comprising the training set used to build the respective CoMFA model.



(A) (B)

Figure 2. Orientation of training set compounds with respect to the template compound 17β-estradiol obtained by *Atom-Fit* alignment (A) and *Flexible Field-Fit* alignment (B).

As the training set increased in structural diversity, we calculated log(RP) values for additional test set compounds. The predictive abilities of these models for the test set proved excellent (Table 4). For example, the log(RP) value of the organochlorine methoxychlor according to experiment is −2.48. The corresponding values predicted by CoMFA model 4 are −2.82 (−0.34) using *flexible field-fit* and −0.91 (1.57) using *atom-fit*, where the numbers in parentheses refer to the difference (i.e., residual) between the corresponding experimentally observed and predicted values.

The final CoMFA models, constructed from the entire training set, span eight structurally diverse chemical classes. Scatter plots of the experimentally observed versus CoMFA-calculated log(RP) values for the complete training set of 40 compounds are shown for both the *atom-fit* and *field-fit* alignment schemes (Figures 3A,B, respectively). Comparison of statistical parameters associated with *flexible field-fit* alignment ($r_{cv}^2 = 0.533$; $r^2 = 0.960$) and *atom-fit* alignment ($r_{cv}^2 = 0.446$; $r^2 = 0.913$) reinforces the advantage of field-based alignment over atom-based alignment for structurally diverse data sets. The results for the test set are consistent with these conclusions (Table 4). Whereas the largest residual log(RP) value using *field-fit* alignment is −0.33 (zearalenone), it is −1.14 (zearalenone) using *atom-fit* alignment. The residual values given in parentheses in Table 4 correspond to predictions made for compounds that are

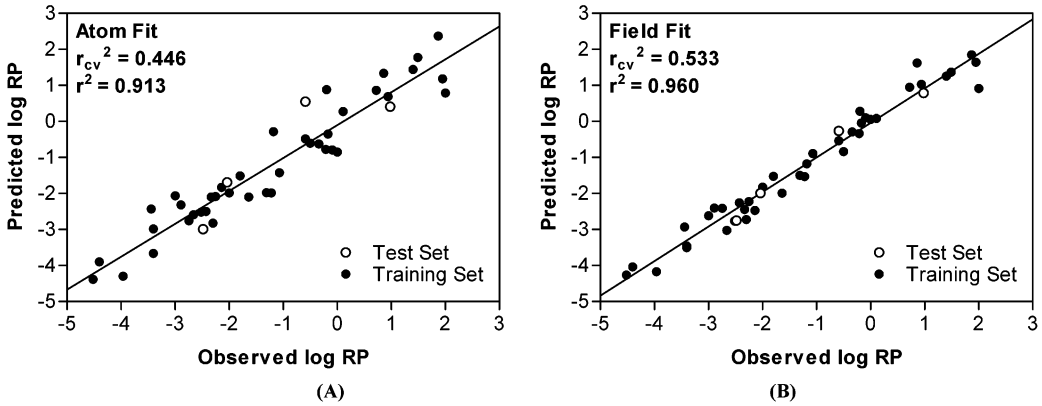


Figure 3. Plot of CoMFA-predicted versus observed values of log(RP) for the 40 training set compounds. (A) *Atom-Fit*; (B) *Flexible Field-Fit*.

not represented in the training set for a particular model. For example, CoMFA model 2 constructed from steroids and lactones performs poorly in predicting the log(RP) of methoxychlor (residual = −0.97). Extrapolation, or predicting values of compounds that fall outside the "chemical space" of a model, often leads to gross inaccuracies and is generally unjustified. This caveat underscores the need for structurally diverse data sets to build statistically robust and predictive models.

Discussion

The present study, in which data sets were systematically expanded by inclusion of structurally diverse compounds spanning eight chemical families (Figure 1), demonstrates that the choice of initial alignment strategy is of great importance in achieving optimal 3D-QSAR models using CoMFA. Another, albeit qualitative, advantage of field-fit over atom-fit approaches stems from practical difficulties in achieving atom-based alignments as the structural diversity of the data set increases. It is intuitively obvious that the process of atom fitting becomes more problematic as the structural diversity of the data set increases, and especially when the data set encompasses different chemical families. In biological applications, this circumstance will occur for receptors either that are inherently promiscuous with respect to ligand selectivity or that become promiscuous upon mutation. Prominent examples ascribing to the first scenario are ER α and β and the steroid and xenobiotic receptor (SXR) (*12*, *13*). The ERs are recognized as targets for EDCs, a large and diverse group of exogenous compounds represented by the present data set. The human SXR and its pregnane xenobiotic receptor (PCR) orthologue in other mammalian species are so-called "orphan" nuclear receptors that function as transcription factors to mediate the metabolism of exogenous chemicals (including drugs) by activation of the cytochrome P450 enzymes (*24*). A prime example following the second scenario is the androgen receptor (*25*) (AR), another nuclear receptor for which key mutations (*26*) (e.g., Thr877Ala) within the ligand binding pocket have been associated with increased affinity for (and activation by) a wider range of endogenous hormones besides the receptor's natural ligands testosterone and dihydrotestosterone (DHT) (*27*, *28*). Compelling experimental findings indicate that this particular mutation is associated with the ultimate failure of androgen ablation therapies for treatment of human prostate cancer (*29*, *30*).

Inspection of the log(RP) values for the test-set compounds obtained from our final CoMFA models (Table 4) reveals that *flexible field-fit* alignment is predictive equally well for compounds with weak potency (e.g., methoxychlor) as for strong potency (e.g., estrone). Of note, both the *atom-fit* and *flexible field-fit* models correctly predicted the low activational activity of tamoxifen and 4-hydroxytamoxifen (Table 2) despite their high affinity for ER. It is worth noting that both alignments made no "false negative" predictions. A false negative refers to an active (e.g., estrogenic) compound that is predicted as inactive, whereas a "false positive" refers to an inactive compound predicted as active. False negatives are especially problematic in risk assessment scenarios. Unlike the inclusion of a false positive which would ultimately be excluded with additional testing, "false negative" compounds that are predicted not to exert endocrine disrupting effects would receive no further scrutiny insofar as their environmental and/or toxicological impact.

In summary, the final CoMFA model (model 5) is based on a structurally diverse data set and demonstrates both internal and external predictive ability regardless of the choice between atom-fit and field-fit alignment. Nevertheless, the present study demonstrates that field-based alignment is generally preferred over atom-based alignment as the size and certainly the structural diversity of the data set increase. This conclusion is consistent with the CoMFA study by Waller et al. (*10*), who employed the field-based SEAL alignment scheme for a series of structural diverse estrogenic compounds. This guidance should prove useful in developing robust 3D-QSAR models when dealing with structurally diverse data sets. In the present application on estrogenic compounds that are known to span several chemical classes, CoMFA models using field-based alignment approaches are recommended for the prioritization of chemicals as to their endocrine disrupting effects prior to in vitro and in vivo screening.

Acknowledgment. We thank Dr. Vladyslav Kholodovych for his invaluable assistance in the preparation of the manuscript. W.J.W. acknowledges financial support for this research from the U.S. Environmental Protection Agency's Science To Achieve Results (STAR) program. Although the research described in this article has been funded in part by the U.S. Environmental Protection Agency's Science To Achieve Results (STAR) program through Grant GAD R826133, it has not been subjected

to any EPA review and, therefore, does not necessarily reflect the views of the Agency, and no official endorsement should be inferred.

References

(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., 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.

(2) Gorski, J., and Hou, Q. (1995) Embryonic estrogen receptors: Do they have a physiological function? *Environ. Health Perspect.* **103 (Suppl. 7)**, 69−72.

(3) Guillette, L. J., Jr., Crain, D. A., Rooney, A. A., and Pickford, D. B. (1995) Organization versus activation: The role of endocrine-disrupting contaminants (edcs) during embryonic development in wildlife. *Environ. Health Perspect.* **103 (Suppl. 7)**, 157−164.

(4) Warr, W. A. (1997) Combinatorial chemistry and molecular diversity. An overview. *J. Chem. Inf. Comput. Sci.* **37** (1), 134−140.

(5) Broach, J. R., and Thorner, J. (1996) High-throughput screening for drug discovery. *Nature* **384 (6604 Suppl.)**, 14−16.

(6) DeLisle, R. K., Yu, S. J., Nair, A. C., and Welsh, W. J. (2001) Homology modeling of the estrogen receptor subtype beta (er-beta) and calculation of ligand binding affinities. *J. Mol. Graph. Model.* **20** (2), 155−167.

(7) Grese, T. A., Sluka, J. P., Bryant, H. U., Cullinan, G. J., Glasebrook, A. L., Jones, C. D., Matsumoto, K., Palkowitz, A. D., Sato, M., Termine, J. D., Winter, M. A., Yang, N. N., and Dodge, J. A. (1997) Molecular determinants of tissue selectivity in estrogen receptor modulators. *Proc. Natl. Acad. Sci. U.S.A.* **94** (25), 14105−14110.

(8) Tong, W., Lowis, 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** (4), 669−677.

(9) 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. *Endocrinology* **138** (9), 4022−4025.

(10) Waller, C. L., Oprea, T. I., Chae, K., Park, H. K., Korach, K. S., Laws, S. C., Wiese, T. E., Kelce, W. R., and Gray, L. E., Jr. (1996) Ligand-based identification of environmental estrogens. *Chem. Res. Toxicol.* **9** (8), 1240−1248.

(11) Kearsley, S. K., and Smith, G. M. (1990) An alternative method for the alignment of molecular structures: Maximizing electrostatic and steric overlap. *Tetrahedron Comput. Methodol.* **3**, 615−633.

(12) Xie, W., Barwick, J. L., Downes, M., Blumberg, B., Simon, C. M., Nelson, M. C., Neuschwander-Tetri, B. A., Brunt, E. M., Guzelian, P. S., and Evans, R. M. (2000) Humanized xenobiotic response in mice expressing nuclear receptor sxr. *Nature* **406** (6794), 435−439.

(13) Xie, W., and Evans, R. M. (2002) Pharmaceutical use of mouse models humanized for the xenobiotic receptor. *Drug Discov. Today* **7** (9), 509−515.

(14) Lehmann, J. M., McKee, D. D., Watson, M. A., Willson, T. M., Moore, J. T., and Kliewer, S. A. (1998) The human orphan nuclear receptor pxr is activated by compounds that regulate cyp3a4 gene expression and cause drug interactions. *J. Clin. Invest.* **102** (5), 1016−1023.

(15) Arnold, S. F., Robinson, M. K., Notides, A. C., Guillette, L. J., Jr., and McLachlan, J. A. (1996) A yeast estrogen screen for examining the relative exposure of cells to natural and xenoestrogens. *Environ. Health Perspect.* **104** (5), 544−548.

(16) Lyttle, C. R., Damian-Matsumura, P., Juul, H., and Butt, T. R. (1992) Human estrogen receptor regulation in a yeast model system and studies on receptor agonists and antagonists. *J. Steroid Biochem. Mol. Biol.* **42** (7), 677−685.

(17) Coldham, N. G., Dave, M., Sivapathasundaram, S., McDonnell, D. P., Connor, C., and Sauer, M. J. (1997) Evaluation of a recombinant yeast cell estrogen screening assay. *Environ. Health Perspect.* **105** (7), 734−742.

(18) Tripos, Sybyl 6.7, Tripos Inc., 1699 S. Hanley Rd., St. Louis, MO 63144.

(19) Tanenbaum, D. M., Wang, Y., Williams, S. P., and Sigler, P. B. (1998) Crystallographic comparison of the estrogen and progesterone receptor's ligand binding domains. *Proc. Natl. Acad. Sci. U.S.A.* **95** (11), 5998−6003.

(20) 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** (7), 927−937.

(21) Gasteiger, J., and Marsili, M. (1980) Iterative partial equalization of orbital electronegativity a rapid access to atomic charges. *Tetrahedron* **36** (22), 3219−3228.

(22) Wold, S., Albano, C., Dunn, W. J. I., Edlund, U., Esbensen, K., Geladi, P., Hellberg, S., Johansson, E., Lindberg, W., and Sjostrom, M. (1984) Multivariate data analysis in chemistry. In *Chemometrics: Mathematics and statistics in chemistry* (Kowalski, B., Ed.) Reidel, Dordrecht, The Netherlands.

(23) Cramer, R. D., Bunce, J. D., and Patterson, D. E. (1988) Cross-validation, bootstrapping, and partial least squares compared with multiple regression in conventional qsar studies. *Quant. Struct.-Act. Realat.* **7**, 18−25.

(24) Synold, T. W., Dussault, I., and Forman, B. M. (2001) The orphan nuclear receptor sxr coordinately regulates drug metabolism and efflux. *Nat. Med.* **7** (5), 584−590.

(25) Waller, C. L., Juma, B. W., Gray, L. E., Jr., and Kelce, W. R. (1996) Three-dimensional quantitative structure−activity relationships for androgen receptor ligands. *Toxicol. Appl. Pharmacol.* **137** (2), 219−227.

(26) Sack, J. S., Kish, K. F., Wang, C., Attar, R. M., Kiefer, S. E., An, Y., Wu, G. Y., Scheffler, J. E., Salvati, M. E., Krystek, S. R., Jr., Weinmann, R., and Einspahr, H. M. (2001) Crystallographic structures of the ligand-binding domains of the androgen receptor and its t877a mutant complexed with the natural agonist dihydrotestosterone. *Proc. Natl. Acad. Sci. U.S.A.* **98** (9), 4904−4909.

(27) Matias, P. M., Donner, P., Coelho, R., Thomaz, M., Peixoto, C., Macedo, S., Otto, N., Joschko, S., Scholz, P., Wegg, A., Basler, S., Schafer, M., Egner, U., and Carrondo, M. A. (2000) Structural evidence for ligand specificity in the binding domain of the human androgen receptor. Implications for pathogenic gene mutations. *J. Biol. Chem.* **275** (34), 26164−26171.

(28) Poujol, N., Wurtz, J. M., Tahiri, B., Lumbroso, S., Nicolas, J. C., Moras, D., and Sultan, C. (2000) Specific recognition of androgens by their nuclear receptor. A structure−function study. *J. Biol. Chem.* **275** (31), 24022−24031.

(29) Grigoryev, D. N., Long, B. J., Njar, V. C., and Brodie, A. H. (2000) Pregnenolone stimulates lncap prostate cancer cell growth via the mutated androgen receptor. *J. Steroid Biochem. Mol. Biol.* **75** (1), 1−10.

(30) Buchanan, G., Greenberg, N. M., Scher, H. I., Harris, J. M., Marshall, V. R., and Tilley, W. D. (2001) Collocation of androgen receptor gene mutations in prostate cancer. *Clin. Cancer Res.* **7** (5), 1273−1281.

TX0255875