A Comparative QSAR Study Using CoMFA, HQSAR, and FRED/SKEYS Paradigms for Estrogen Receptor Binding Affinities of Structurally Diverse Compounds

Chris L. Waller*

Pfizer Global Research and Development, Ann Arbor, Michigan 48105

Received November 3, 2003

The three-dimensional quantitative structure−activity relationship (QSAR) technique of comparative molecular field analysis (CoMFA) has demonstrated the ability to provide accurate predictions for diverse chemical compounds when trained with molecules of diverse chemical type. Although predictive, the derivation and utilization of models of this type are quite computationally and person power intensive. It is this intensity that pragmatically limits the widespread implementation of these models as predictive tools. In this study, two newer QSAR techniques were evaluated as possible alternatives to CoMFA based QSAR models for the purpose of rapidly identifying estrogen receptor ligands from diverse collections of molecules. The first of these is Hologram QSAR, or HQSAR. HQSAR utilizes Tripos molecular fingerprints as descriptors in conjunction with partial least squares (PLS) regression and cross-validation routines. The HQSAR technique demonstrated the ability to rapidly develop QSAR models independent of the intense user input (i.e. geometry optimization, conformational analysis, and molecular superposition were not required). Second, a newly developed QSAR paradigm that utilizes Molecular Design Limited (MDL) substructure keys (SKEYS) as descriptors in combination with an evolutionary algorithm, *F*ast *R*andom *E*limination of *D*escriptors (FRED), was evaluated. By utilizing the FRED/SKEYS algorithm, a simple substructure-based QSAR model was derived that was comparable in statistical robustness and predictive ability to both CoMFA and HQSAR derived models. A comparison of the utility of these three approaches as computational tools for the rapid identification of estrogen receptor ligands as potential endocrine disruptors as assessed by model predictive ability will be described.

INTRODUCTION

The need to rapidly assess the toxicological potential of large numbers of compounds has dramatically increased over the past few years. In the pharmaceutical industry this has been primarily due to the advent and widespread implementation of combinatorial and high-speed parallel synthetic chemistry and high-throughput screening for biological activity. Prior to this current revolution in the drug discovery process, synthesis of organic materials and subsequent assaying for desired pharmacological effects stood as bottlenecks in the pathway to development of a clinical candidate. An effective means of rapid assessment of toxicological potential has yet to be implemented in the drug discovery process.

A similar, and complimentary, scenario is beginning to occur within the human health and ecological effects setting. The necessity to quickly identify compounds capable of producing undesirable effects in members of the ecosystem has been stimulated by the recent reports on the ability of certain compounds used in environmental applications (i.e., pesticides[1] and those known to enter the ecosystem i.e., industrial chemicals[2]) to disrupt the “normal” functioning of the endocrine systems of a variety of species.

A great deal of effort has been expended to understand the mechanisms of toxic activity of a variety of the so-called “endocrine disrupting chemicals and mixtures”. It is possible for compounds of this type to disrupt the normal functioning of the endocrine system at numerous levels in the biological process including (1) altering the synthesis/catabolism of endogenous endocrines via interaction with enzymes, (2) interfering with the transport of endogenous endocrines by plasma transport proteins, (3) competition with endogenous endocrines for their respective receptors, and (4) altering the metabolism of endogenous endocrines via interaction with enzymes.[1] Of these possible mechanisms, competition for endogenous receptors exists as a primary means of endocrine disruption, and the abilities of many compounds of environmental importance have been evaluated for their affinities for several endocrine receptors (i.e., estrogen receptor,[2] androgen receptor,[1] etc.). In this paper, a set of diverse compounds that have been assessed for their ability to compete with estradiol[3] for the estrogen receptor have been used as a training set for a variety of computational paradigms aimed at the development of quantitative structure−activity relationship (QSAR) models that can be used to rapidly and accurately identify potential endocrine disruptors from large collections of virtual compounds.

METHODS

Comparative Molecular Field Analysis (CoMFA). The methods implemented in the CoMFA[4] study have been previously reported.[3] In short, all molecules were constructed using the molecule building and conformational analysis tools in the SYBYL[5] molecular modeling suite of programs. All rotatable bonds were systematically searched at 5° increments. The lowest energy conformer of each molecule was

* Corresponding author e-mail: Chris.waller@pfizer.com.

10.1021/ci0342526 CCC: $27.50 © 2004 American Chemical Society
Published on Web 02/19/2004

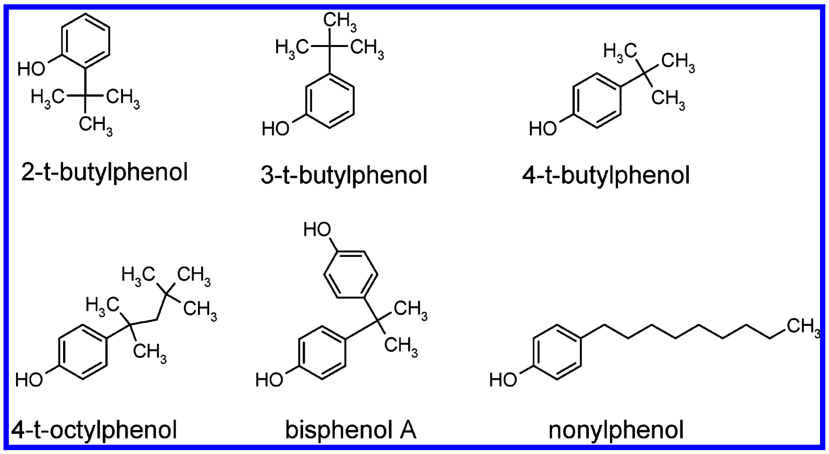

Figure 1. Phenols.

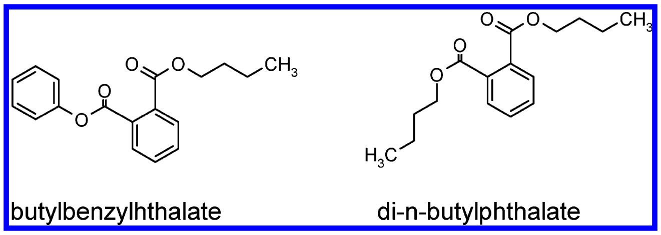

Figure 2. Phthalates.

then fully minimized using the semiempirical program MOPAC[6] and the AM1[7] model Hamiltonian (keywords: AM1 MMOK NOINTER). The automated structure alignment program, *S*teric and *E*lectrostatic *AL*ignment (SEAL),[8] was implemented to derive the mutual superposition, or alignment, rule using estradiol as the template to which all molecules were fitted. Two hundred trial alignments were evaluated for each compound with the alignment best mimicking the potential energy fields of estradiol being included in the QSAR model training set. During the alignment process, the compounds were evaluated for fit on the basis of steric, electrostatic, and hydrophobic fields with relative weightings of 1:2:2.

CoMFA interaction potential energy fields were evaluated on a region (lattice) extending 4 A° in the X, Y, and Z axes beyond the volume defined by the union of all molecules in the training set with a grid spacing of 2 A°. Standard CoMFA steric and electrostatic fields were calculated with energy truncation values of ±30 kcal/mol. *H*ydrophobic *INT*eraction (HINT) (eduSoft, LC) potential energy fields were computed using an exponential distance-dependent function ($1/e^{-r}$) on the same region described above.

The compounds included in this original study have been implemented as the training sets for all QSAR models developed in the present study. The compounds are presented in Figures 1−8 and grouped according to structural/functional classifications.

Hologram QSAR (HQSAR). HQSAR[9] is a technique based on the concept of using molecular substructures expressed in a binary pattern (i.e., fingerprint) as descriptors in QSAR models. In this study, fingerprints were generated for all substructures between 4 and 7 atoms in size for all molecules. The substructure fingerprints were then hashed into hologram bins with lengths of 53, 59, 61, 71, 83, 97, 151, and 199. Information on atoms, bonds, and connections was considered and included in the fingerprint. Hydrogen atoms and chirality flags were not utilized in the present study. The best model was chosen based on the least standard error of cross-validated predictions.

FRED/SKEYS. The details of the original variable elimination code utilized in this study have been reported earlier.[10] In this manuscript, the original FRED code was modified in order for it to perform optimally using binary indicator variables as descriptors. In the following sections, the specific details of the algorithm will be presented.

The substructural keys were taken from the ISIS/MACCS database system (MDL Information Systems, Inc.). One hundred sixty-six molecular substructure keys are user-accessible and can be used in specifying complex structural queries in ISIS/MACCS. The keys represent structural features ranging in complexity from the presence of singular atoms (i.e., key 161 = N), functional groups (i.e., key 139 = OH), ring systems (i.e., key 96 = five-membered ring system), to complex patterns (i.e., key 91 = QHAAACH2A, where Q = not C, H and A = not H). The keys generated by ISIS/MACCS for the molecules in this study were extracted using ISISHost[11] and expressed in SYBYL as indicator variables (i.e., 0 and1) populating 166 columns in a Tripos molecular spreadsheet (MSS).

Two preliminary variable reduction routines are presently available prior to the generation of the first set of offspring models. The first of these is a variable filter routine that performs simple statistics on each set of descriptor variables. Variables can be eliminated from the available descriptor pool if there is no or little variation in the values within a given set. Elimination of *zero variance* descriptors is highly recommended and greatly enhances the efficiency of the algorithm. Although it is also possible to eliminate variables on the basis of a minimal variance, this option was not implemented in the present study due to the discrete nature of the descriptors themselves. For example, the presence of a unique substructure in the series will be retained if variables are eliminated under the zero variance criterion, while it may be eliminated if a higher threshold variance is selected.

Additionally, it is possible to eliminate those variables across the series that are covariant. The FRED algorithm implements a simple forward stepwise comparison routine that eliminates downstream variables that do not meet a user-defined collinearity threshold. In this routine, a full pairwise correlation matrix is not generated. Rather, the descriptors are compared sequentially using a nested loop algorithm.

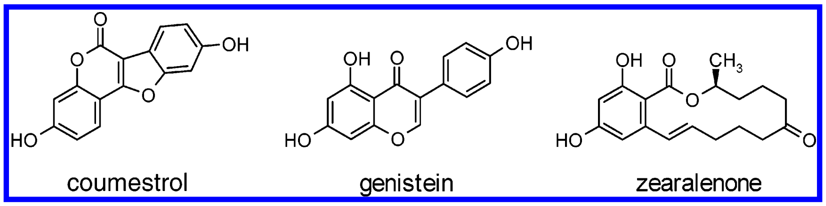

Figure 3. Phytoestrogens.

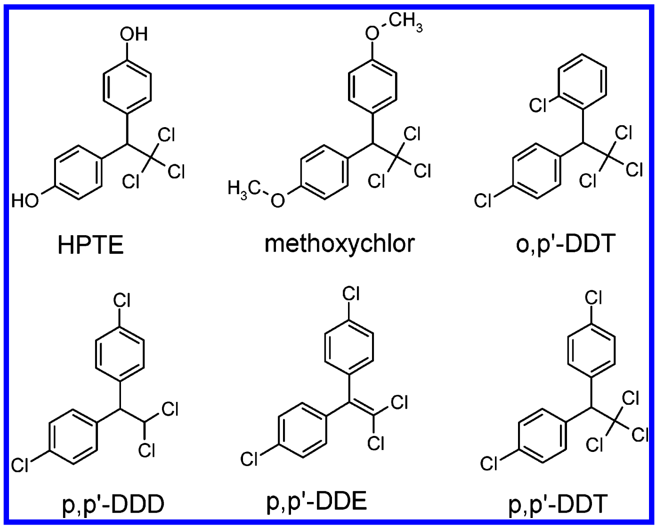

Figure 4. DDTs.

After the available descriptor pool has been reduced by eliminating invariant and collinear variables, the number of allowable descriptors per model is computed. A simple rule-of-thumb approach was implemented in that for each descriptor (*m*), or independent variable, one must have at least six observations (*n*). This factor varies by opinion; however, it is generally agreed that between five and six observations are required per descriptor in order to minimize the chance of over-fitting.[12] This conservative approach ensures that insignificant variables have little spurious effect on the resultant offspring models during the evolutionary process.

A simple random number generation algorithm is utilized in the selection of descriptors for the offspring models. An oversampling (progeny) factor is implemented to ensure that all descriptors are selected during this process. This number is variable and effects the algorithm in the following manner. If the number of descriptors to be included in each offspring model is set to the maximal number of allowed descriptors (defined above) so that each descriptor may be chosen only once per offspring and each descriptor is only allowed to be selected once per generation, then the minimal number of models which would be required to sample all the descriptors in the total descriptor pool can be calculated as the total number of descriptors divided by the maximum number of descriptors allowed per model. The algorithm as implemented does not enforce this last restriction that effectively means that it is possible that one or more descriptors can be either included in or excluded from all models in a generation. An empirical oversampling factor has been utilized which takes this minimal number of models calculated above and factors it upward. A default fixed value of 20 was utilized.

For each offspring model generated during a generation, a PLS regression analysis with full cross-validation and the number of components set equal to the number of maximum number of allowable descriptors is performed. The leave-one-out cross-validated r^2, or q^2_{LOO}, statistic was utilized as the fitness function.

Prior to the generation of the first set of offspring, the user is prompted to supply a so-called "kill factor". This factor is used in the algorithm in the following manner. Once all the offspring models have been generated for a given generation, the models are sorted according to fitness. The kill factor is then used to select a given percentage of models from the tails of the distribution of the offspring model population. The descriptors for the models from the lower distribution (i.e. less fit models) are then compared to the descriptors for the models from the upper distribution (i.e. more fit models). At this point, the algorithm implements a simplified version of taboo search.[13] Those descriptors from the lower distribution not found in the set from the upper distribution are considered detrimental to the fitness of an entire generation of offspring models and are considered taboo. Unlike some previous implementations of taboo search, the variable is not always deselected at this point; however, it is placed on a taboo list. The algorithm is allowed to select this descriptor for incorporation into the makeup of subsequent generations of offspring models on the premise that it may not contribute to a suboptimal model given a second chance in combination with other descriptors. However, once a variable appears on the taboo list in a subsequent but not necessarily sequential generation, it is removed from the allowable descriptor pool. This interpretation of strategic forgetting provides additional insurance that a given descriptor is truly detrimental. A kill factor of 5% was utilized.

Several termination criteria are available for use in the FRED algorithm. The fitness function chosen must be considered when selecting a termination criterion. In this implementation of the FRED algorithm, a fitness function of q^2_{LOO} was selected as described above. A termination criterion was designed that examines the standard deviation of the fitness values (i.e. q^2_{LOO}'s) for an entire population in a generation. If this value is less than a set threshold value, or minimum sigma, then the algorithm will terminate. It possible to force the algorithm to terminate with a singular fit individual model with a fixed set of characteristics by setting the termination criterion to a smaller minimum sigma value. In instances where q^2_{LOO} is implemented as the fitness function and numerous offspring models are desired, a minimum sigma value of 0.1 or greater is recommend. A minimum sigma value of 0.025 or smaller is generally sufficient to yield a singular offspring model at algorithm termination. This latter value was utilized herein.

Predictive Power of the CoMFA, HQSAR, and FRED-Derived QSAR Models. To fully assess and compare the general utility of the models derived using each of these methods, the training set was divided into eight (8) subclasses of structurally related molecules: phenols, phthalates, phytoestrogens, DDTs, PCBs, pesticides, DESs, and steroids (see

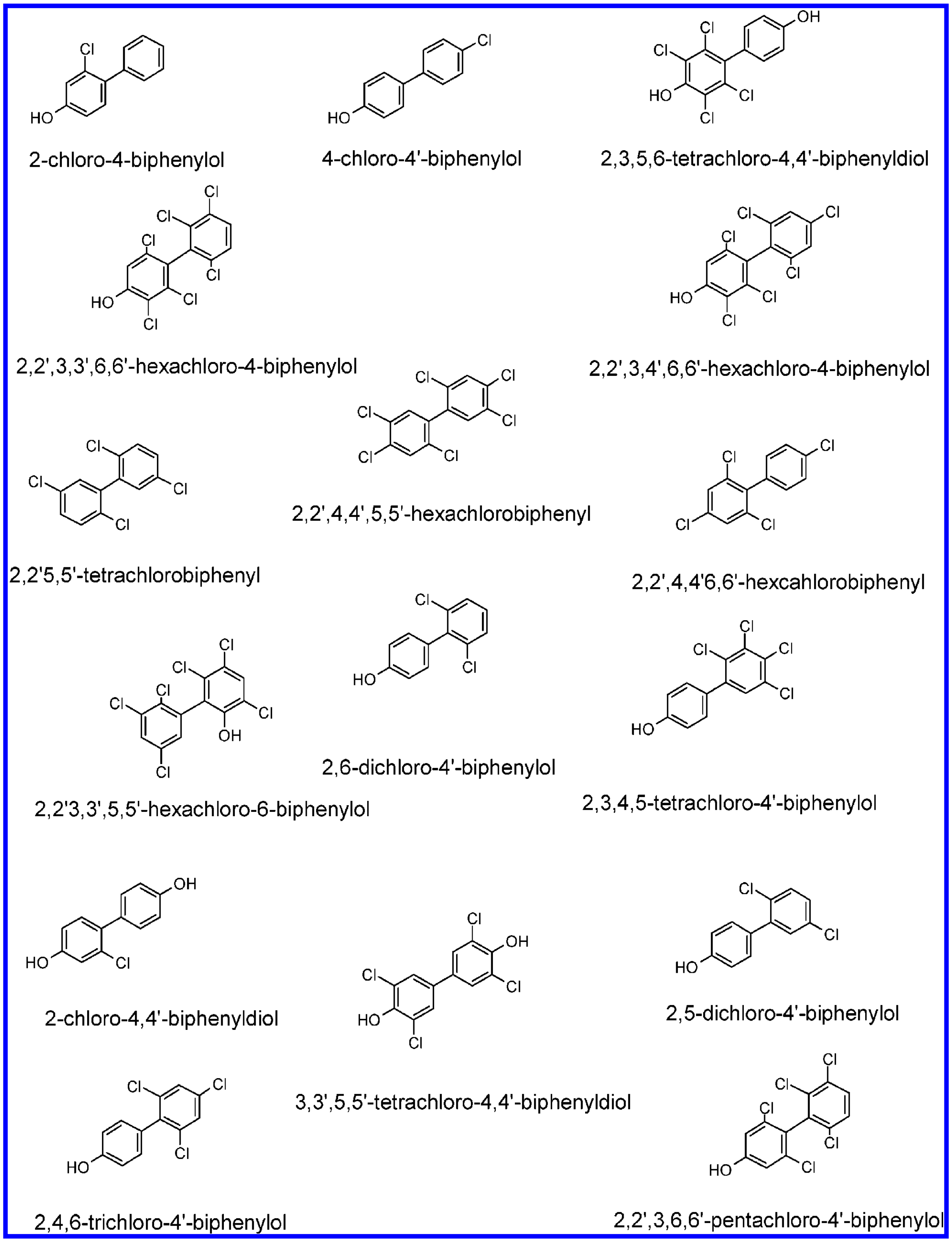
2-chloro-4-biphenylol
4-chloro-4'-biphenylol
2,3,5,6-tetrachloro-4,4'-biphenyldiol
2,2',3,3',6,6'-hexachloro-4-biphenylol
2,2',3,4',6,6'-hexachloro-4-biphenylol
2,2',4,4',5,5'-hexachlorobiphenyl
2,2'5,5'-tetrachlorobiphenyl
2,2',4,4'6,6'-hexcahlorobiphenyl
2,6-dichloro-4'-biphenylol
2,2'3,3',5,5'-hexachloro-6-biphenylol
2,3,4,5-tetrachloro-4'-biphenylol
2-chloro-4,4'-biphenyldiol
3,3',5,5'-tetrachloro-4,4'-biphenyldiol
2,5-dichloro-4'-biphenylol
2,4,6-trichloro-4'-biphenylol
2,2',3,6,6'-pentachloro-4'-biphenylol

Figure 5. PCBs.

Figures 1−8). Eight submodels were generated for each QSAR method by excluding all of the members of each class from the training set. The binding affinity of each member of a particular class was then predicted using the models from which that entire class was excluded. In this manner, a more realistic (termed “unbiased” herein) estimate of the external predictive ability of each model derived using each QSAR method was attained.

RESULTS AND DISCUSSION

The statistical results for all models generated by CoMFA, HQSAR, and FRED/SKEYS for the entire training set are presented in Tables 1−9. The CoMFA models were based on steric, electrostatic, and hydropathic interactions. The optimal CoMFA QSAR model for the entire training set suggested that 3 principal components were sufficient to describe approximately 90% of the variance in the binding data. The corresponding CoMFA field contour plots from the analysis based on the entire data set (not shown, see ref 3) suggest that the core of the ligand should be sterically bulky and hydrophobic with an overall positive electrostatic potential with distinct areas of negative electrostatic potential that are required at either end of a long axis defining the molecule.

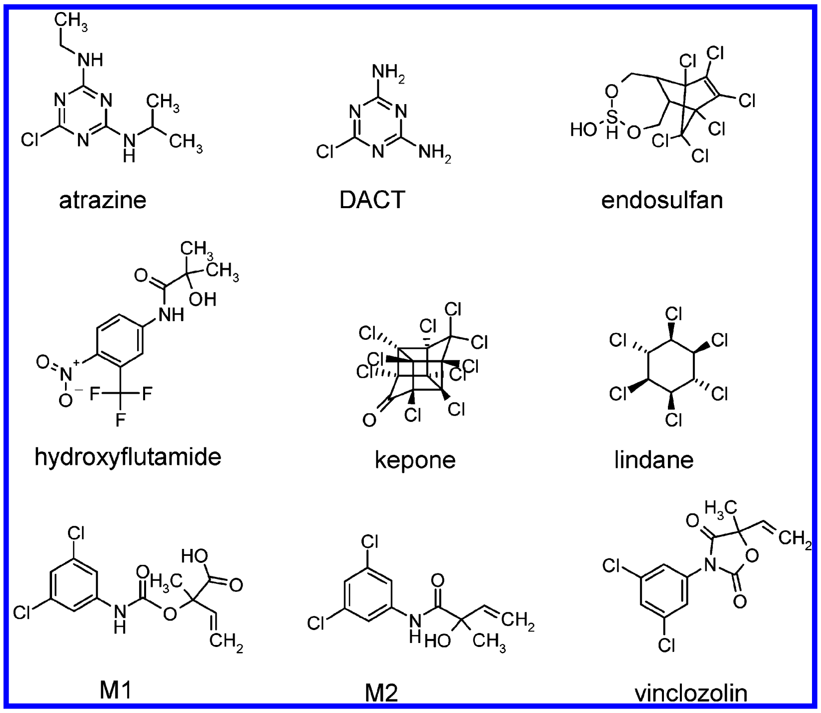

Figure 6. Pesticides.

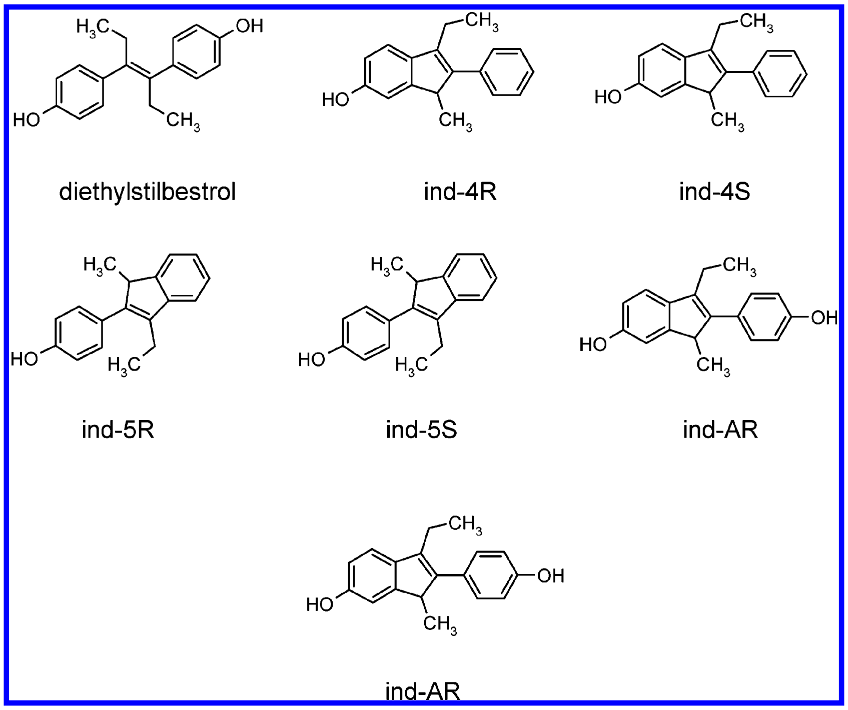

Figure 7. DESs.

The resulting models derived from analyses of the smaller training subsets provide a means to assess the external predictive ability of the models. As indicated by the average absolute error of predictions reported in Table 10, the CoMFA technique proved superior to the competing techniques for three of the eight external test sets including phenols, PCBs, and steroids.

The optimal HQSAR model for the total training set, as assessed by the lowest cross-validated standard error of predictions, required a hologram length of 199 bins as reported in Table 1. Five principal components were required to explain approximately 80% of the variance in the data. As was the case for the CoMFA QSAR models, the number of bins and statistical results varied with the composition and size of the training set. Better models were identified for certain of the training subsets; however, these values were achieved at the expense of structural diversity represented in the model.

The expanded QSAR equation for the model derived from the analysis on the total training set consists of 199 terms, one per bin, and is not reported here. A benefit of this technique is that it is possible to identify substructures in molecules that can be mapped into given bins and thus given a regression coefficient. This is useful in graphically identifying and color-coding beneficial or detrimental substructures in molecules of interest.

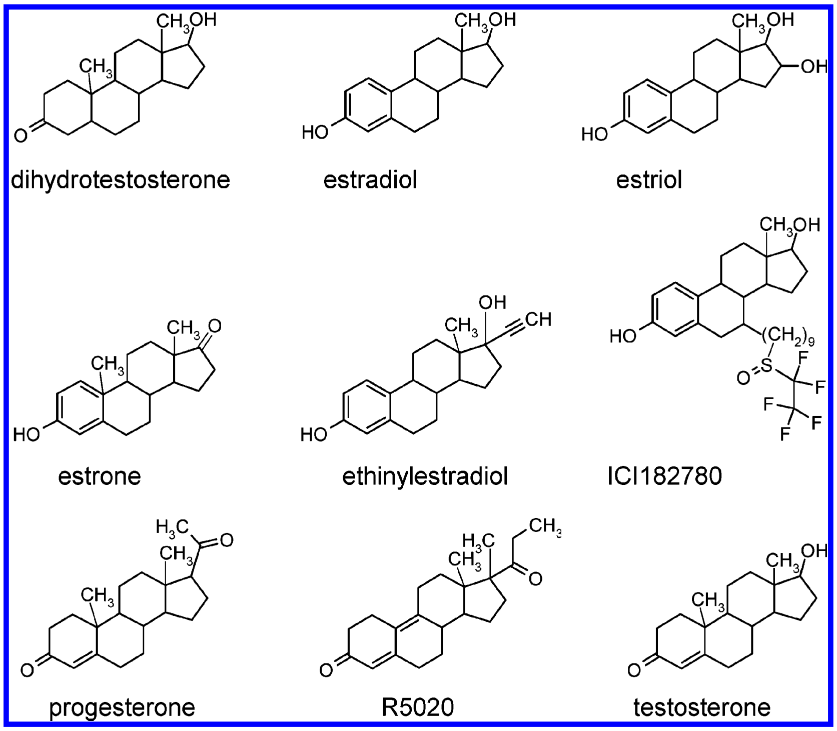

Figure 8. Steroids.

Table 1. Statistical Results

	CoMFA	HQSAR	FRED/SKEYS
q^2	0.580 (3)	0.578 (5)	0.700
r^2	0.893	0.805	0.783
s	0.657	0.905	1.008

The HQSAR technique was much less computationally intense than the CoMFA technique with respect to making predictions. As reported in Tables 2−9, the HQSAR technique yielded the best unbiased predictions for only one of the test subsets (DESs). It is probable that this substructure-based QSAR technique is not particularly ideal for test sets possessing substructures not represented in the training set. The potential field-based technique of CoMFA seems better suited for extrapolations to bioisosteric substructures in molecules.

The FRED/SKEYS analysis on the total data set yielded a model with a 10 variable QSAR equation: $\mathrm{pIC50} = -2.568 - (1.558)\mathrm{key25} - (1.009)\mathrm{key63} + (0.453)\mathrm{key75} + (1.674)\mathrm{key108} + (1.951)\mathrm{key113} + (0.556)\mathrm{key127} + (1.010)\mathrm{key131} - (0.441)\mathrm{key133} + (1.009)\mathrm{key144} - (1.002)\mathrm{key157}$. Additionally, the following keys were identified as being common among all molecules: keys 163 and 165. The keys can be interpreted as corresponding to a maximal common substructure of a six-membered ring system. From the QSAR equation, the presence of substructure keys 75 (an endocyclic nitrogen substituted with a heavy atom), 108 (six membered heavy atom fragment with terminal methyl group), 113 (a nonaromatic oxygen), 127 (an exocyclic oxygen atom), 131 (a heteroatom connected to a hydrogen), and 144 (a heavy atom not in an aromatic system) tends to lead to an increase of the potency, while the presence of substructure keys 25

Table 2. Experimental and Predicted Values for Alkylphenols

alkylphenols	actual p*K*i	CoMFA	residual	HQSAR	residual	FRED/SKEYS	residual
2-*tert*-butylphenol	−2.37	−1.31	1.06	−0.39	1.98	−0.41	1.96
3-*tert*-butylphenol	−2.60	−1.16	1.43	−0.46	2.14	−1.16	1.44
4-*tert*-butylphenol	−2.21	−1.48	0.73	−0.27	1.94	−1.16	1.05
4-*tert*-octylphenol	−0.12	0.34	0.46	0.23	0.35	−1.16	1.04
bisphenol A	−0.16	−1.38	1.21	−0.69	0.53	0.09	0.25
nonylphenol	0.08	−0.90	0.98	−0.51	0.59	0.59	0.51

Table 3. Experimental and Predicted Values for Phthalates

phthalates	actual p*K*i	CoMFA	residual	HQSAR	residual	FRED/SKEYS	residual
butylbenzylhthalate	−1.88	0.30	2.18	−0.30	1.58	1.44	3.32
di-*n*-butylphthalate	−2.00	−1.49	0.51	−1.38	0.62	−0.80	1.20

Table 4. Experimental and Predicted Values for Phytoestrogens

phytos	actual p*K*i	CoMFA	residual	HQSAR	residual	FRED/SKEYS	residual
coumestrol	1.03	−0.53	2.19	1.01	0.02	1.02	0.01
genistein	0.41	−0.04	0.44	0.47	0.06	1.02	0.61
zearalenone	2.22	−1.16	3.38	−0.73	2.95	1.91	0.31

Table 5. Experimental and Predicted Values for DDTs

DDTs	actual p*K*i	CoMFA	residual	HQSAR	residual	FRED/SKEYS	residual
HPTE	1.30	−0.25	1.55	−1.07	2.37	0.19	1.11
methoxychlor	−1.84	0.06	1.90	−1.26	0.58	0.90	2.74
o,p′-DDT	−0.46	0.30	0.76	−1.38	0.92	−1.34	0.88
p,p′-DDD	−3.00	−0.03	2.97	−1.35	1.65	−1.31	1.69
p,p′-DDE	−3.00	−1.11	1.89	0.35	3.35	−1.31	1.69
p,p′-DDT	−3.00	−0.31	2.70	−1.64	1.36	−1.31	1.69

Table 6. Experimental and Predicted Values for PCBs

PCBs	actual p*K*i	CoMFA	residual	HQSAR	residual	FRED/SKEYS	residual
2,4,6-trichloro-4′-biphenylol	1.32	1.33	0.01	−1.36	2.68	−0.63	1.95
2,3,4,5-tetrachloro-4′-biphenylol	1.35	0.06	1.29	−1.80	3.15	−0.63	1.98
2-chloro-4,4′-biphenyldiol	0.94	0.89	0.05	−0.57	1.51	0.59	0.35
2,6-dichloro-4′-biphenylol	0.52	1.38	0.86	−1.29	1.81	−1.60	2.12
2,5-dichloro-4′-biphenylol	0.45	0.84	0.39	−1.50	1.95	−1.60	2.05
3,3′,5,5′-tetrachloro-4,4′-biphenyldiol	−0.29	−1.06	0.77	−0.72	0.43	0.50	0.79
2-chloro-4-biphenylol	−0.51	−0.04	0.47	−1.21	0.70	−1.60	1.09
4-chloro-4′-biphenylol	−0.75	−1.24	0.49	−1.37	0.62	−1.60	0.85
2,3,5,6-tetrachloro-4,4′-biphenyldiol	−0.38	0.98	1.36	−0.71	0.33	0.59	0.97
2,2′,3,3′,6,6′-hexachloro-4-biphenylol	−0.85	−0.33	0.52	−1.53	0.68	−0.63	0.22
2,2′,3,4′,6,6′-hexachloro-4-biphenylol	−0.73	0.31	1.04	−1.43	0.70	−0.63	0.10
2,2′,3,6,6′-pentachloro-4′-biphenylol	−0.20	1.10	1.30	−1.45	1.25	−0.63	0.43
2,2′5,5′-tetrachlorobiphenyl	−0.79	−0.99	0.20	−2.45	1.66	−1.27	0.48
2,2′,4,4′,5,5′-hexachlorobiphenyl	−0.93	−1.65	0.72	−2.78	1.85	−1.27	0.34
2,2′,4,4′6,6′-hexcahlorobiphenyl	−0.12	0.02	0.13	−2.24	2.12	−1.27	1.15
2,2′3,3′,5,5′-hexachloro-6-biphenylol	−0.81	−2.17	1.36	−1.86	1.05	−0.63	0.18

Table 7. Experimental and Predicted Values for Pesticides

pesticides	actual p*K*i	CoMFA	residual	HQSAR	residual	FRED/SKEYS	residual
atrazine	−3.00	−0.29	2.71	−2.84	0.16	−0.16	2.84
DACT	−3.00	−1.77	1.23	−3.38	0.38	−0.16	2.84
endosulfan	−2.78	0.67	3.45	−1.35	1.43	−2.53	0.25
hydroxyflutamide	−3.00	−0.97	2.03	−3.36	0.36	−1.20	1.80
kepone	−0.15	−0.76	0.61	−18.49	18.34	−3.38	3.23
lindane	−3.00	−0.70	2.30	−2.47	0.53	−4.42	1.42
M1	−3.00	−1.23	1.77	−2.06	0.94	−1.20	1.80
M2	−3.00	−0.54	2.47	−2.12	0.88	−1.20	1.80
vinclozolin	−3.00	−0.57	2.43	−2.19	0.81	−1.01	1.99

Table 8. Experimental and Predicted Values for DESs

DESs	actual p*K*i	CoMFA	residual	HQSAR	residual	FRED/SKEYS	residual
diethylstilbestrol	3.16	0.68	2.48	0.31	2.85	−2.46	5.62
ind-4R	0.30	0.04	0.26	0.76	0.46	0.59	0.29
ind-4S	1.26	0.97	0.29	0.76	0.50	0.59	0.67
ind-5R	0.96	−0.10	1.06	0.72	0.24	0.59	0.37
ind-5S	1.75	1.65	0.10	0.72	1.03	0.59	1.16
ind-AR	2.36	−0.40	2.76	1.23	1.13	0.59	1.77
ind-AS	3.46	1.55	1.91	1.23	2.23	0.59	2.87

Table 9. Experimental and Predicted Values for Steroids

steroids	actual p*K*i	CoMFA	residual	HQSAR	residual	FRED/SKEYS	residual
dihydrotestosterone	−1.00	0.40	1.40	3.61	4.61	1.83	2.83
estradiol	2.59	0.99	1.60	1.66	0.93	2.41	0.18
estriol	1.85	1.38	0.47	1.91	0.06	2.41	0.56
estrone	2.36	0.56	1.80	0.89	1.47	2.41	0.05
ethinylestradiol	3.52	1.18	2.35	2.83	0.69	2.41	1.11
ICI182780	3.22	1.36	1.86	2.95	0.27	2.41	0.81
progesterone	−3.00	0.46	3.46	1.91	4.91	2.03	5.03
R5020	−0.07	0.84	0.91	−0.18	0.11	2.03	2.10
testosterone	−1.46	0.56	2.02	1.90	3.36	1.83	3.29

(guanidino), 133 (two heavy atoms in a ring system attached to a chain nitrogen in a chain), 63 (a nitrogen doubly bonded to an oxygen), and 157 (a carbon attached to an oxygen) tends to decrease the affinity of the compound for the estrogen receptor.

The FRED/SKEYS approach proved to be superior to CoMFA and HQSAR with respect to external predictive ability for three of the test subsets (phytoestrogens, DDTs, and pesticides). As was the case for HQSAR, these predictions required much less user time and input than the

Table 10. Average Absolute Errors of Predictions

	CoMFA	HQSAR	FRED/SKEYS
alkylphenols	0.98	1.25	1.04
phthalates	1.35	1.10	2.26
phytoestrogens	1.84	1.01	0.31
DDTs	1.96	1.70	1.63
PCBs	0.68	1.40	0.94
pesticides	2.11	2.65	2.00
DESs	1.26	1.20	1.82
steroids	1.76	1.82	1.77

corresponding CoMFA technique. It is interesting to note that while the predictions were generated on a time scale on par with HQSAR, the accuracy of the predictions was more comparable to the more intense CoMFA technique.

In summary, the CoMFA and HQSAR models based on the entire training set displayed similar cross-validated results in the range of 0.58, while the FRED/SKEYS approach yielded a significantly higher value of 0.700. These results indicate that, for this particular data set, the FRED/SKEYS model was more internally consistent with potentially greater predictive power. The non-cross-validated results for the HQSAR and FRED/SKEYS models were practically identical (approximately 0.8 for each). The CoMFA model displayed a slightly higher value of 0.893. This is primarily due to larger number of variables included in the CoMFA model. In general, the models displayed comparable abilities to predict the estrogen receptor binding affinities for diverse sets of compounds. It is interesting to note that the results for the CoMFA and HQSAR models were consistently similar, while the FRED/SKEYS was superior to both in its ability to predict the activities of the compounds in the phytoestrogen class. The converse situation exists for the compounds in the phthalate class, in that the CoMFA and HQSAR models outperformed the FRED/SKEYS model.

CONCLUSIONS

A variety of QSAR paradigms have been presented as possible computational tools to aid with the rapid assessment of endocrine disruption potential for environmentally relevant compounds. It has been demonstrated that it is possible to reduce the complexity and computational/person power required in order to produce QSAR models while maintaining a respectable level of predictive power. There are a number of publications in which the CoMFA and HQSAR techniques have been compared.[14−16] In this article, the novel QSAR paradigm, FRED/SKEYS, demonstrated the ability to generate predictive QSAR models either on par or superior to the previously compared techniques.

With respect to the resources required, the three-dimensional nature of the CoMFA technique necessitated that much time be spent generating reasonable structures in an appropriate alignment with information for various molecular fields parsed onto a grid. The actual CPU time expended during the regression component was negligible. In comparison, the HQSAR and FRED/SKEYS approaches do not require three-dimensional structures, alignments, or molecular fields. The generation of molecular descriptors is rapid for both of these techniques. The regression component of the FRED/SKEYS technique is the most intense of the trio since numerous models must be developed and evaluated per generation. However, due to the innovations in the variable selection routine implemented in FRED, the number of generations required in this particular evolutionary algorithm is much fewer than similar programs. The regression requirements of the HQSAR technique are similar to those of CoMFA. Therefore, in terms of the time required to build and validate QSAR models using the methods described herein, CoMFA ranks at the top being the most intensive followed by FRED/SKEYS, then HQSAR.

Application to large data sets is technically possible now with the introduction of faster QSAR techniques such as those presented herein. The rate-limiting factor to widespread implementation of these approaches as well as CoMFA and other QSAR methods remains validation of the external predictive ability. While the two newer QSAR paradigms presented herein possess predictive capabilities similar to CoMFA, extrapolation to dissimilar structures (i.e., those containing substructures not represented in the training set molecules) is troublesome. Ideally, the training set molecules would represent the chemical diversity of the molecules to be predicted. Practically, this is hampered by diversity analysis and experimental design concerns. At a minimum, the user must be informed of the degree of extrapolation required to make a given prediction. The CoMFA technique provides an indication of this uncertainty, yet HQSAR and FRED/SKEYS do not presently have this ability. Work is ongoing in my lab and others to overcome these limitations.

REFERENCES AND NOTES

(1) Kelce, W. R.; Stone, C. R.; Laws, S. C.; L. E. Gray, J.; Kemppainen, J. A. et al. Persistent DDT Metabolite p, p′-DDE is a Potent Androgen Receptor Antagonist. *Nature* **1995**, *375*, 581−585.

(2) Korach, K. S.; Sarver, P.; Chae, K.; McLachlan, J. A.; McKinney, J. D. Estrogen Receptor-Binding Activity of Polychlorinated Hydroxybiphenyls: Conformationally Restricted Structural Probes. *Mol. Pharmacol.* **1988**, *33*, 120−126.

(3) Waller, C. L.; Oprea, T. I.; Chae, K.; Park, H.-K.; Korach, K. S. et al. Ligand-Based Identification of Environmental Estrogens. *Chem. Res. Toxicol.* **1996**, *9*, 1240−1248.

(4) Cramer, R. D., III.; 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.

(5) SYBYL is available from Tripos, Inc., St. Louis, MO.

(6) MOPAC is available from Indiana University, Quantum Chemistry Program Exchange, #455.

(7) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(8) Kearsley, S. K.; Smith, G. M. An Alternative Method for the Alignment of Molecular Structures: Maximizing Electrostatic and Steric Overlap. *Tetrahedron Comput. Met.* **1990**, *3*, 615−633.

(9) HQSAR is available from Tripos, Inc., St. Louis, MO.

(10) Waller, C. L.; Bradley, M. P. Development and Validation of a Novel Variable Selection Technique with Application to Multidimensional QSAR Studies. *J. Chem. Info. Comput. Sci.* **1999**, *39*, 345−355.

(11) ISIS is available from MDL Information Systems, San Leandro, CA.

(12) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure−Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238−1244.

(13) Cvijovic, D.; Klinowski, J. Taboo Search: An Approach to the Multiple Minima Problem. *Science* **1995**, *267*, 664−666.

(14) Rodrigues, C. R.; Flahery, T. M.; Springer, C.; McKerrow, J. H.; Cohen, F. E. CoMFA and HQSAR of Acylhydrazide Cruzain Inhibitors. *Bioorg. Med. Chem Lett.* **2002**, *12*, 1537−1541.

(15) So, S. S.; Karplus, M. A Comparative Study of Ligand−Receptor Complex Binding Affinity Prediction Methods based on Glycogen Phosphorylase Inhibitors. *J. Comput. Aided. Mol. Des.* **1999**, *13*, 243−258.

(16) Tong, W.; Lowis, D. R.; Perkins, R.; Chen, Y.; Welsh, W. J. et al. 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.

CI0342526