Spectroscopic QSAR Methods and Self-Organizing Molecular Field Analysis for Relating Molecular Structure and Estrogenic Activity

Arja Asikainen,[†] Juhani Ruuskanen,[†] and Kari Tuppurainen*,[‡]

Department of Environmental Sciences and Department of Chemistry, University of Kuopio, P.O. Box 1627, FIN-70211, Kuopio, Finland

Received June 3, 2003

The performance of three "spectroscopic" quantitative structure−activity relationship (QSAR) methods (eigenvalue (EVA), electronic eigenvalue (EEVA), and comparative spectra analysis (CoSA)) for relating molecular structure and estrogenic activity are critically evaluated. The methods were tested with respect to the relative binding affinities (RBA) of a diverse set of 36 estrogens previously examined in detail by the comparative molecular field analysis method. The CoSA method with ^{13}C chemical shifts appears to provide a predictive QSAR model for this data set. EEVA (i.e., molecular orbital energy in this context) is a borderline case, whereas the performances of EVA (i.e., vibrational normal mode) and CoSA with ^{1}H shifts are substandard and only semiquantitative. The CoSA method with ^{13}C chemical shifts provides an alternative and supplement to conventional 3D QSAR methods for rationalizing and predicting the estrogenic activity of molecules. If CoSA is to be applied to large data sets, however, it is desirable that the chemical shifts are available from common databases or, alternatively, that they can be estimated with sufficient accuracy using fast prediction schemes. Calculations of NMR chemical shifts by quantum mechanical methods, as in this case study, seem to be too time-consuming at this moment, but the situation is changing rapidly. An inherent shortcoming common to all spectroscopic QSAR methods is that they cannot take the chirality of molecules into account, at least as formulated at present. Moreover, the symmetry of molecules may cause additional problems. There are three pairs of enantiomers and nine symmetric (C_2 or C_{2v}) molecules present in the data set, so that the predictive ability of full 3D QSAR methods is expected to be better than that of spectroscopic methods. This is demonstrated with SOMFA (self-organizing molecular field analysis). In general, the use of external test sets with randomized data is encouraged as a validation tool in QSAR studies.

INTRODUCTION

Xenoestrogens. Intensive research has revealed that there is a plethora of xenoestrogens in our environment, i.e., both man-made and natural molecules that have been shown to bind to the estrogen receptor as either agonists or antagonists. Serious concern has recently arisen about the adverse effects of chemical compounds possessing estrogenic activity on humans and other species, but it is practically impossible to perform thorough toxicological tests on all of the more than 87 000 xenoestrogens that may ultimately need to be evaluated.[1] Thus there is an obvious need to develop alternative methods to predict the estrogenic activity of molecules with sufficient accuracy. These methods should ultimately facilitate the rapid screening of untested xenoestrogens, particularly in order to distinguish which molecules should have the highest priority for entry into expensive and stressful testing on animals. In this context, computational methods such as QSARs (quantitative structure−activity relationships) seem attractive.

QSARs are in widespread use in medicinal and pharmaceutical chemistry today, but they are not solely confined to drug development, as they can also be of assistance in the evaluation of the toxicological properties of molecules; for a review, see Schultz et al.[2] QSAR studies may be useful in toxicology because they can provide an insight into the metabolic activation mechanisms of toxins and into the nature of the interactions between receptor proteins and toxins. Furthermore, the predictive ability of QSARs should not be overlooked, although physicochemical interpretation of the correlative models may be difficult.

Recent findings corroborate that QSARs can be of valuable assistance in predicting the estrogenic activity of organic molecules; for a recent comprehensive review, see Fang et al.[3] and references therein. Most previous QSAR works have employed CoMFA (comparative molecular field analysis),[4] perhaps the most widely applied 3D QSAR method. Despite its success, CoMFA has some inherent shortcomings, in particular the fact that its performance is closely dependent on molecular alignment (superposition). This, together with its sensitivity to molecular conformations, makes it difficult to use with large, structurally diverse data sets. Until now, only a few works utilizing alignment-free QSAR methods have been reported. The Hansch method works well with congeneric data sets,[5] and the success achieved with *k*NN (*k*-nearest-neighbor) QSAR indicates that 2D methods are a promising alternative to more laborious 3D QSAR methods.[6] The COREPA (common reactivity pattern)[7] and CPSA (charged partial surface area)[8] approaches rely on descriptors

* Corresponding author phone: +358-17-162225; fax: +358-17-163259; e-mail: Kari.Tuppurainen@uku.fi.
[†] Department of Environmental Sciences.
[‡] Department of Chemistry.

10.1021/ci034110b CCC: $25.00
Published on Web 10/28/2003

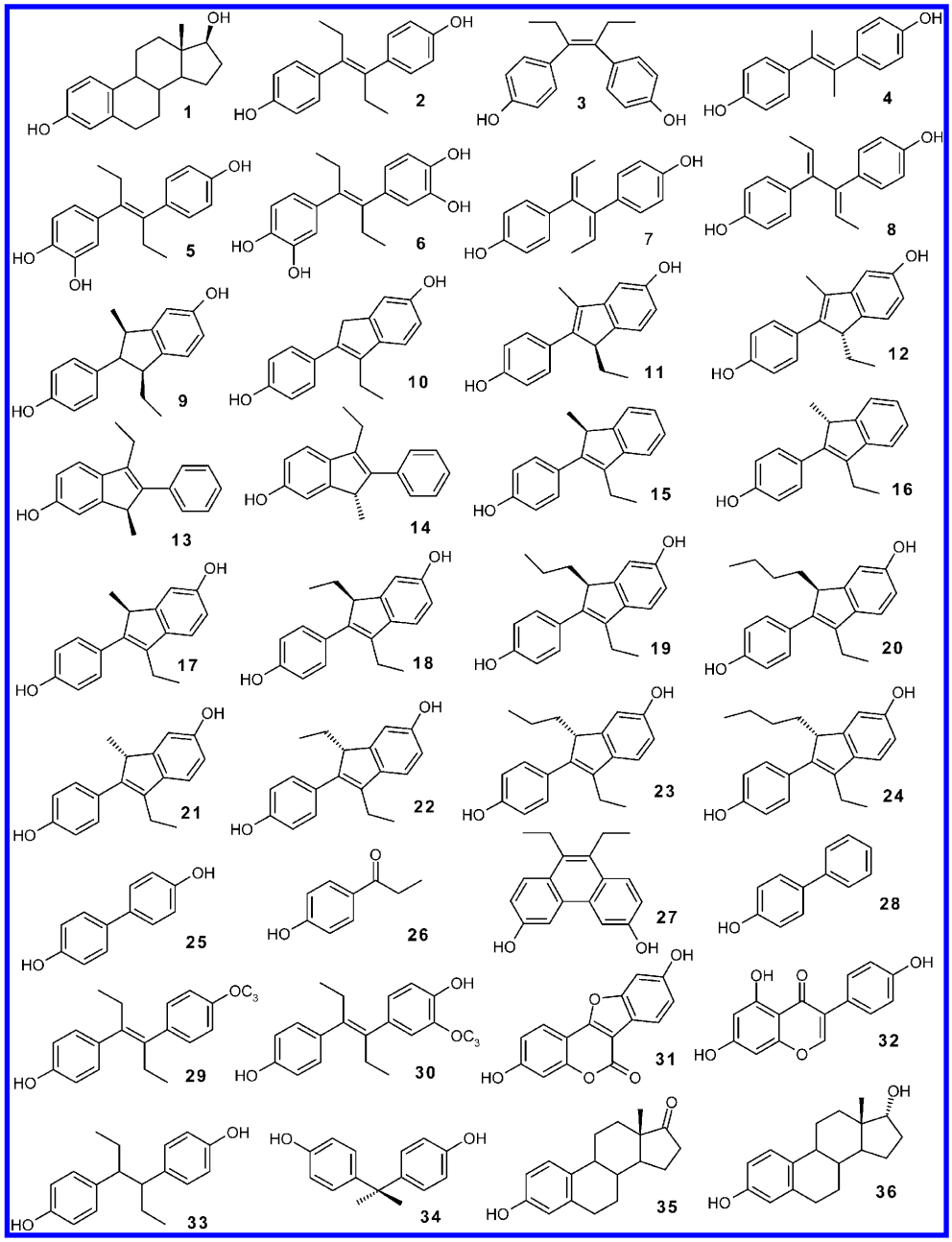

Figure 1. Structures of the 36 estrogens in the data set.

based on quantum mechanics, providing an alternative perspective on the ligand binding properties of the estrogen receptor.

The purpose of this work was to test the performance of some novel alignment-free QSAR methods for predicting the estrogenic activity of molecules employing the RBA (relative binding affinity) values of a diverse set of estrogens (Table 1, Figure 1). The QSAR methods considered here have their origin in the spectroscopic properties of molecules and their predictive ability is compared with that of a 3D QSAR method, self-organizing molecular field analysis (SOMFA),[9] which was tested with the same estrogen data set. In addition, comprehensive tests of comparative molecular field analysis (CoMFA) with the same data set have been reported.[10−12]

Spectroscopic QSAR Methods. A set of QSAR methods have recently been put forward in which vibration frequencies (EVA, eigenvalue, i.e., vibrational normal mode in this context),[13−17] MO energies (EEVA, electronic eigenvalue),[18−20] and NMR chemical shifts (CoSA, comparative spectra analysis[21] and its modifications[22−28]) are used to derive QSAR descriptors. It has been shown that spectroscopic

Table 1. Experimental Activities[10−12] (log(RBA) Values) of Compounds **1**−**36** (for Structures, See Figure 1)

compd	log(RBA)	compd	log(RBA)	compd	log(RBA)
1	2.00	**13**	0.26	**25**	−1.70
2	2.46	**14**	−0.70	**26**	−1.00
3	−0.10	**15**	0.75	**27**	−0.80
4	1.52	**16**	−0.05	**28**	−1.70
5	2.00	**17**	2.46	**29**	1.30
6	1.40	**18**	2.47	**30**	1.00
7	−0.52	**19**	2.36	**31**	1.97
8	1.30	**20**	2.25	**32**	0.70
9	0.30	**21**	1.11	**33**	2.25
10	1.14	**22**	1.04	**34**	−1.30
11	2.00	**23**	1.26	**35**	1.78
12	2.15	**24**	0.90	**36**	1.76

methods can provide robust, predictive QSAR models for a large number of biological data sets. In fact, the overall performance of spectroscopic QSAR methods seems to be comparable to that of CoMFA, despite the fact that they do not require structural alignment of the molecules. The success of spectroscopic methods in QSAR can be explained when we recall that all spectroscopic quantities reflect certain intrinsic physicochemical properties of molecules that are related to their 3D structure of molecules, at least implicitly. IR spectra, for example, reflect internal vibrations and the spatial arrangements of molecular functional groups. Moreover, the energies of normal molecular vibrations are in the range in which the recognition processes between molecules, crucial for all biological and biochemical processes, are taking place. The relationship between the ^{13}C NMR chemical shifts of a molecule and its 3D conformations is well-known and almost routinely employed. In general, these NMR shifts depend greatly on the electrostatic potential energy of the nucleus and the types of atomic orbitals surrounding the nucleus (i.e. its hybridization state), together with the disturbing effects of surrounding atoms. Thus the NMR chemical shifts are closely related to the substituent effect, the basis of all QSAR methods. MO energies, in turn, can give valuable information on the electronic structure of molecules. It should be emphasized that most spectroscopic properties of molecules can be calculated with a high degree of accuracy by molecular orbital (semiempirical and ab initio) or density functional methods, and theoretical values may therefore be used as surrogates for experimental ones to derive QSAR descriptors. Fast prediction schemes for many spectroscopic quantities, e.g., ^{13}C chemical shifts, have also been developed.

COMPUTATIONAL METHODS

Spectroscopic Descriptors. The computation of spectroscopic QSAR descriptors by means of Gaussian smoothing involves the following steps: (i) determination of the eigenvalues of a molecule (MO energies, vibration frequencies, and NMR chemical shifts) by a feasible computational method (see below), (ii) transformation of the eigenvalues to a bounded scale, (iii) placement of a Gaussian kernel of fixed standard deviation σ over each eigenvalue, and (iv) summation of the overlaid kernels at intervals of L (usually set at $\sigma/2$, eq 1):

$$Descriptor(x) = \sum_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-E_i)^2/2\sigma^2} \qquad (1)$$

where E_i is the ith eigenvalue of the molecule in question and N is the number of eigenvalues.

MO Energies and Vibrational Frequencies. Previous works indicate that both vibration frequencies[13−17] and MO energies[18−20] can be calculated with sufficient accuracy at the semiempirical level of theory. The three-dimensional structures of estrogens **1**−**36** (Figure 1) were modeled with the HYPERCHEM program package (Hypercube, Inc.) Since the molecular skeletons are relatively rigid in this case, a full conformational analysis was not performed, except for the flexible side chains. Appropriate dihedral angles were rotated in steps of 30° with subsequent minimizations, employing the MM+ force field, as implemented in HYPERCHEM. As expected, the side chains preferred the extended conformations. The structure of the lowest energy conformer of each compound was then minimized with the AMPAC program package (QCPE No. 506, version 2.11) and the AM1 Hamiltonian.[29] All the geometric parameters were fully optimized for each compound, keeping all the settings and options of AMPAC at their default values, except that the keywords PRECISE and FORCE were used. The FORTRAN code of AMPAC was translated automatically to the C/C++ programming language. An implementation of the AMPAC package in the LINUX operating system provided high computational speed and efficiency even in a microcomputer environment. The original code was slightly modified to output MO energies and vibration frequencies to separate disk files for further calculations. The eigenvalue ranges were 0−3600 cm^{-1} (EVA), −45 to +10 eV (EEVA), and 0−220 ppm (CoSA), respectively.

^{13}C and ^{1}H NMR Chemical Shifts. The gauge-invariant atomic orbitals (GIAO) method[30] has proved increasingly useful for predicting NMR chemical shifts, the calculation of which proceeded here in two phases. The first part involved obtaining sufficiently accurate determinations of the molecular geometries. We have shown previously that the 6-31G* basis set can provide reliable geometries for dibenzothiophenes,[31] camphenes,[32] and chloroterpenes,[33] and this method was also employed here. The second part consisted of a GIAO calculation for determining the chemical shifts, which were derived by calculating first the absolute NMR shielding for each nucleus and then taking the difference between the calculated value and the shielding value for the same nucleus of a suitable reference molecule, e.g., tetramethylsilane (TMS), calculated by reference to exactly the same model chemistry. Again the level is of crucial importance, as GIAO calculations employing density functional theory (DFT) have generally proved superior to corresponding calculations at the Hartree−Fock level of theory. Thus, in contrast to MO energies and vibration frequencies, high-level ab initio/density functional calculations must be employed to predict NMR chemical shifts correctly. As in our previous works,[31−33] a multiphase procedure (beginning from the MM+ optimized geometry and proceeding via AM1 and 3-21G* to the 6-31G* level) was adopted, in which the initial force constants were taken from a previous, lower level calculation. It appears that this approach has a clear advantage over direct calculations in which the MM+ optimized structure as such is used as an input for the 6-31G* calculation. Finally, the ^{13}C and ^{1}H chemical shifts were calculated by the GIAO method at the B3LYP/6-311G* level of theory with 6-31G* optimized geometries. To assess the quality of calculated ^{13}C chemical shifts, the experimental shifts of compounds **1** (estradiol-17β), **2** (diethylstilbestrol), **35** (estrone), and **36** (estradiol-17α), taken from the Aldrich Library of ^{13}C and ^{1}H NMR Spectra,[34] were compared with the calculated values. After a slight empirical scaling ($\sigma_{pred} = 0.9982\sigma_{calc} - 2.474$), the correlation between the calculated and experimental values is very good: $r^2 = 0.998$, SE = 1.93, and $n = 61$ (data not shown). All the ab initio and DFT/GIAO calculations were performed using the GAUSSIAN 98 program[35] running on a Silicon Graphics Origin2000 workstation.

SOMFA. Self-organizing molecular field analysis (SOMFA) is a grid-based, i.e., alignment-dependent, 3D QSAR

method. In contrast to most 3D QSAR methods, however, it avoids the use of complex statistical tools such as partial least-squares (PLS) and variable selection procedures such as simulated annealing or genetic algorithms; for details, the reader is referred to the original publication.[9] In the present instance, SOMFA was implemented using in-house MATLAB (MathWorks, Natick, MA) scripts written by the authors. In contrast with the original method, the field fit technique was used for molecular superposition. A novel descriptor, molecular polarizabilty, was tested as a supplement to the original SOMFA descriptors, i.e., molecular shape and electrostatic potential. It has been shown recently that molecular polarizablity fields, derived from semiempirically determined atomic polarizabilities, can provide highly predictive QSAR descriptors.[36] The AM1 atomic polarizabilities were calculated by the method proposed by Lewis,[37] the necessary modifications to the AMPAC code being performed by the authors.

Data Analysis. Spectroscopic QSAR models were derived employing a PLS algorithm with leave-one-out and leave-*n*-out cross-validation. All PLS analyses were performed using MATLAB scripts written by the authors. The scripts are based on an efficient modification of the PLS algorithm, SVDPLS (singular value decomposition PLS),[38] which facilitates very rapid cross-validation runs. The following nomenclature will be used in the PLS analyses. For the model building phase, CV = cross-validation, LOO = leave-one-out, PRESS = predictive residual sum of squares, S_{press} = cross-validated standard error of prediction (eq 2), and q^2 = cross-validated correlation coefficient (eq 3):

$$S_{press} = \sqrt{PRESS/(n - c - 1)} \tag{2}$$

$$q^2 = 1 - \frac{\sum(y_{obs} - y_{pred})^2}{\sum(y_{obs} - y_{mean})^2} = 1 - \frac{PRESS}{\sum(y_{obs} - y_{mean})^2} \tag{3}$$

where n is the number of compounds and c is the number of principal components. The S_{press} value is weighted so that it penalizes models with a large number of principal components. For fitted models, r^2 = conventional correlation coefficient, SE = standard error, and F = Fisher test for significance. For external test sets, the conventional squared correlation coefficient r^2_{ex}, mean absolute deviations ($|\Delta|_{ave}$), $SDEP$ (eq 4), and predictive r^2-scores (pr-r^2 in eq 5) were calculated:

$$SDEP = \sqrt{PRESS/n} \tag{4}$$

$$pr - r^2 = \frac{SD - PRESS}{SD} \tag{5}$$

where SD is the sum of squared deviations between the activities of molecules in the test set and the mean activity of the training set molecules.

RESULTS

Model Development with Spectroscopic Descriptors. In developing QSAR models, one should keep an eye on both internal and external predictability. As in our previous work,[20] we have put special emphasis on the latter in order to ensure the predictive ability of the models for new data sets that are not present in the model-building phase.

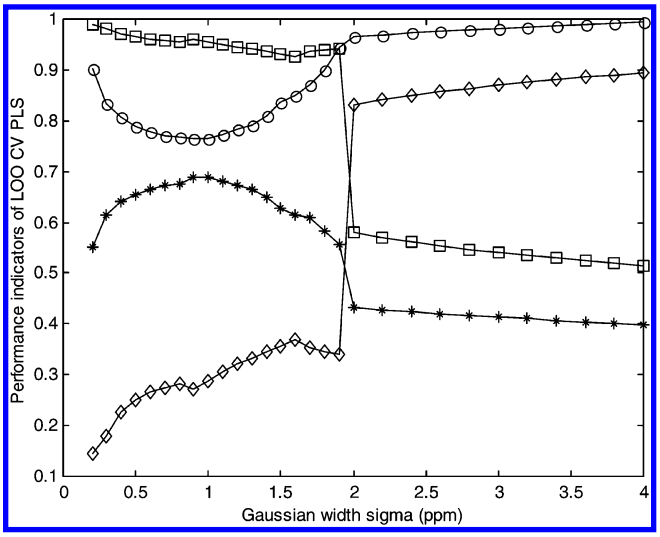

Figure 2. Variation in PLS performance indicators for the CoSA (^{13}C) method as a function of σ: S_{press} (○), q^2 (*), r^2 (□), and SE (◇).

There is an adjustable parameter, σ, which needs to be optimized for each spectroscopic QSAR model. It has been found previously that the EVA and EEVA methods may be sensitive to the value of σ, and the optimum value is a feature that is dependent on the data set. Thus development and validation of the model have been performed in three phases: (i) optimization of the Gaussian width σ in eq 1 employing LOO CV tests for a large number of reasonable σ values, as exemplified in Figure 2 (after this step, only qualified QSAR models, i.e., those with acceptably high q^2 and r^2 values, are selected for the more detailed validation tests); (ii) validation of the models employing the optimum value σ together with a large number of randomized training and test sets (i.e., by choosing two-thirds of the molecules for the training set at random and placing the remaining one-third in the test set and taking the validity to be good if all the performance indicators are in the acceptable range and their scatter is not very large); (iii) verification of the reliability of the model(s) by scrambling tests (i.e., mixing the activities of the training set compounds so that the values are no longer assigned to the right descriptors, repeating this calculation many times, and taking the reliability to be good if only a few random combinations yield statistics close to the correct ones).

In general, PLS models with $q^2 > 0.5$ and $r^2 > 0.9$ are accepted as statistically significant and internally consistent.[39] Applying these criteria, it is obvious that only CoSA with ^{13}C chemical shifts qualifies as a valid QSAR method for this data set (model 3, Table 2). EEVA (model 1) is a borderline case, whereas EVA (model 2) and CoSA with ^{1}H shifts (model 4) are substandard. Moreover, compound **31** (coumestrol) is an outlier in models 1, 2, and 4. Only a slightly better performance was achieved by dropping this molecule, however (data not shown).

Since different spectroscopic descriptors reflect different aspects of molecular structure, it seemed reasonable to test whether a combination of these would lead to an enhanced predictive ability. Model 3 was so superior to the others in this case, however, that models derived with various

Table 2. Internal Predictability of the Spectroscopic QSAR Models 1−4

model	descriptor	σ^a	PC[b]	S_{press}	q^2	r^2	SE	F
1	EEVA	0.075 eV	3	0.99	0.42	0.91	0.40	104.
2	EVA	10 cm^{-1}	2	1.14	0.22	0.47	0.94	14.5
3	CoSA (^{13}C)	1.0 ppm	6	0.76	0.69	0.96	0.29	105.
4	CoSA (^{1}H)	0.05 ppm	3	1.05	0.35	0.75	0.64	33.5

[a] Optimum Gaussian width. [b] Optimum number of principal components (PC).

Table 3. External[a] Predictability of the CoSA and SOMFA Models

model	S_{press}	q^2	r^2_{ex}	$\lvert\Delta\rvert_{av}$	SEDP	pr-r^2
CoSA (^{13}C)	0.90 (0.15)[b]	0.56 (0.15)	0.54 (0.20)	0.68 (0.15)	0.86 (0.16)	0.49 (0.22)
SOMFA1	0.64 (0.05)	0.74 (0.05)	0.77 (0.11)	0.51 (0.09)	0.62 (0.09)	0.74 (0.12)
SOMFA2	0.87 (0.06)	0.52 (0.07)	0.59 (0.14)	0.72 (0.11)	0.84 (0.12)	0.52 (0.17)
SOMFA3	0.75 (0.07)	0.64 (0.08)	0.70 (0.14)	0.63 (0.09)	0.71 (0.11)	0.65 (0.14)

[a] The external performance indicators are averages over 500 randomized runs. [b] The scatter of the indicators is given in terms of the standard deviations in parentheses.

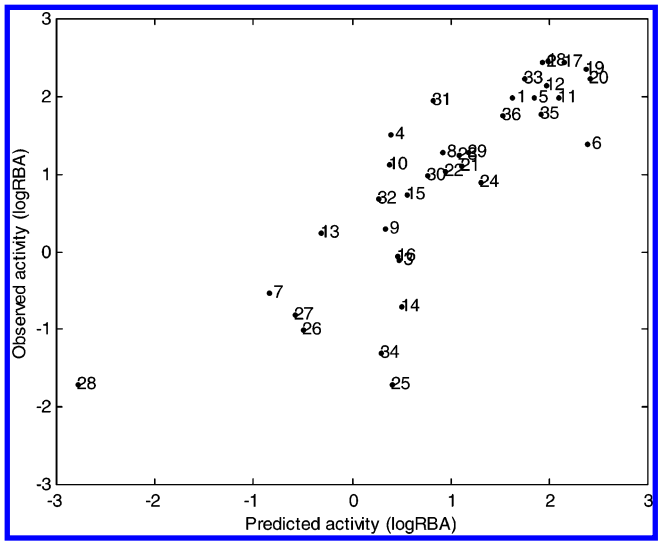

Figure 3. LOO CV CoSA (^{13}C) model for the estrogen data set.

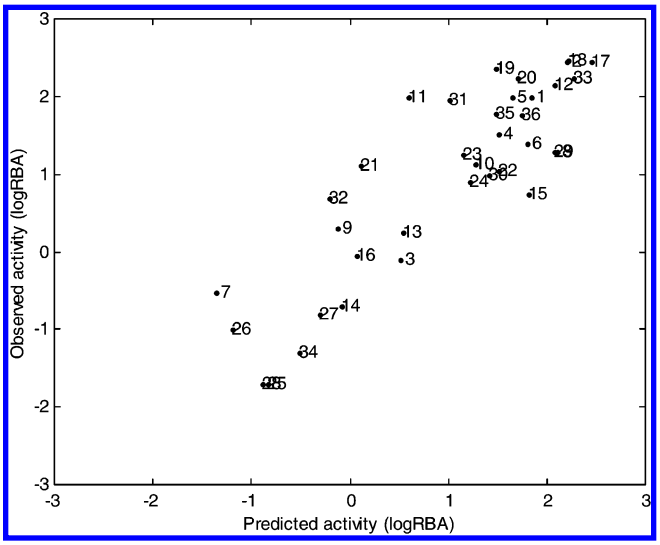

Figure 4. LOO CV SOMFA (property = molecular shape) model for the estrogen data set.

combinations of descriptors did not lead to improved performance (data not shown). Thus detailed consideration was given only to model 3, the predictive ability of which in the LOO CV test is exemplified in Figure 3.

In the second phase of testing, a large number of validation runs (500) were performed (data randomized as discussed above) with the optimum value of σ (1 ppm). The validation runs indicated that both the internal and external (Table 3) predictabilities of model 3 remain quite good irrespective of how compounds **1**−**36** are divided between the training and test sets. The scatter in the performance indicators (expressed as standard deviations) is considerable, however, a reminder of how important well-balanced teaching sets are in the QSAR method.

Finally, the reliability of model 3 was verified by scrambling the y variable 500 times. It appeared that no random combination yielded PLS statistics even close to the correct one: usually $q^2 < 0$ and only in very few cases were the q^2 values above zero (mean = −0.10).

Model Development with SOMFA. The alignment of molecules was performed by forming a best-matching fit of the shape and electrostatic fields with a common template, estradiol-17β, for each molecule. The simulated annealing method, implemented as a MATLAB code, was used for the minimization. A grid size of 22 Å^3 was used, and the grid resolution was two points per angstrom. The validity and reliability of the molecular shape (model SOMFA1; LOO CV statistics: $S_{press} = 0.63$, $q^2 = 0.76$), electrostatic potential (SOMFA2; LOO CV statistics: $S_{press} = 0.85$, $q^2 = 0.55$), and polarizability (SOMFA3; LOO CV statistics: $S_{press} = 0.73$, $q^2 = 0.67$) as QSAR descriptors were checked following steps i−iii discussed above, except that linear regression was used as a statistical tool instead of PLS. In the tests with external data sets, the master grid of SOMFA was calculated separately for each training set, after which the corresponding regression model was derived, i.e., the external tests were truly "blind".

The SOMFA results (Table 3) indicate that molecular shape outperforms both electrostatic potential and polarizability fields, irrespective of whether internal or external predictability is considered. The model SOMFA1 has no bad outliers, and at least the trends in the activities are correctly predicted. The absolute deviations between the observed and LOO CV-predicted values leave room for improvement, however, as exemplified in Figure 4.

Both the internal and external predictabilities of the models remain unchanged in the validation tests (Table 3), and all the SOMFA models passed the scrambling tests correctly (data not shown).

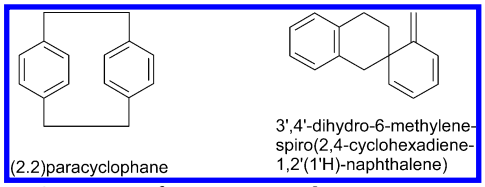

Figure 5. Structures of symmetric and nonsymmetric test molecules.

DISCUSSION

Before discussing the performance of the CoSA method in detail, it should be emphasized that there are both enantiomers and symmetric molecules in the present data set. This may cause problems for all spectroscopic QSAR methods, as will be discussed next.

Enantiomers, Molecular Symmetry, and Spectroscopic Descriptors. In many cases, enantiomers possess very different biological activities. On the other hand, all physical properties of enantiomers (except for their interaction with polarized light), including MO energies, IR frequencies, and NMR chemical shifts, are completely identical. As a consequence, all spectroscopic descriptors actually lack the information required to distinguish the different activities of the enantiomer pair. Further, if the enantiomers are represented by completely identical descriptors, a bias is introduced in the statistics, which may be reflected on the q^2 and S_{press} values or on the number of principal components. It would be possible to include only one enantiomer in the training set and to let the model predict the other in the test set. This approach was preliminarily tested for all spectroscopic descriptors, but without any real improvement in the LOO statistics (data not shown). It may be also possible to make chirality corrections to the spectroscopic descriptors, as discussed recently by Golbraikh and Tropsha in the case of topological descriptors.[40] However, it seems evident that the spectroscopic QSAR descriptors, at least as formulated at present, cannot draw a clear distinction between enantiomers.

Besides the problems caused by the enantiomers, the effect of molecular symmetry is another issue for all spectroscopic descriptors. In spectroscopy, a highly symmetric compound can be easily recognized from its IR or NMR spectrum by the small number of signals it presents. The molecular energy levels tend to degenerate with increasing symmetry, which may result in a considerable loss of information for highly symmetric molecules.

The influence of molecular symmetry on the spectroscopic descriptors was examined using two isoelectronic $C_{16}H_{16}$ hydrocarbons, (2.2)paracyclophane (high symmetry; presenting only three separate ^{13}C signals, i.e., 4-, 4-, and 8-fold degeneracy) and 3′,4′-dihydro-6-methylenespiro(2,4-cyclohexadiene-1,2′(1′H)-naphthalene (no symmetry; presenting 16 separate signals), as a test example (Figure 5).

As expected, it appeared that the information content of the ^{13}C CoSA descriptor of the symmetric compound is much smaller than that of the nonsymmetric one (Figure 6a). The symmetric compound produces a CoSA descriptor with spikelike peaks, and the differences in intensities between symmetric and nonsymmetric compounds are considerable. In contrast, the differences are not so clear for the EVA and EEVA descriptors (Figure 6b,c), but the intensities of some peaks are not commensurate. Large differences in the intensities may cause problems in PLS analyses, if autoscaling of the descriptors is not used. On the other hand, it has been previously found that autoscaling usually has a detrimental effect on the performance of the spectroscopic descriptors.[13,20]

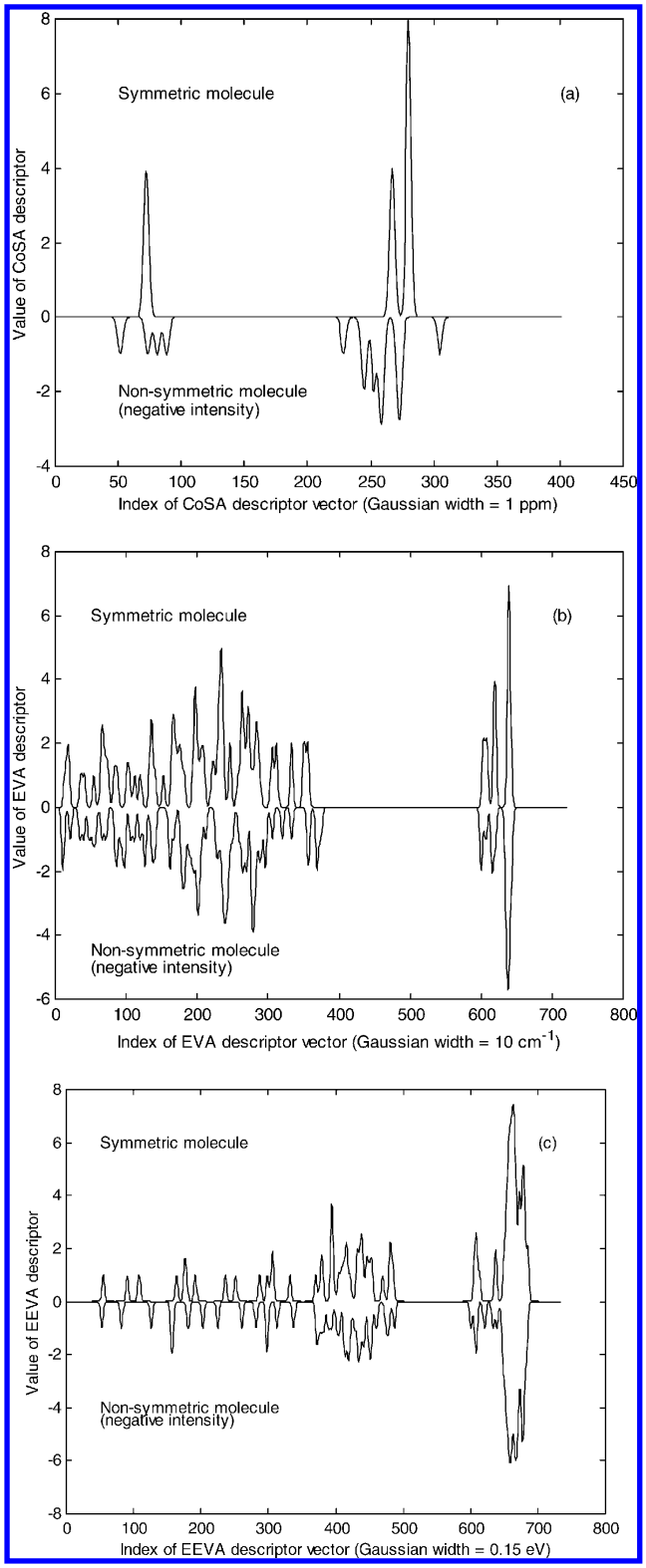

Figure 6. Spectroscopic descriptors of the symmetric and nonsymmetric test molecules: CoSA (a), EVA (b), and EEVA (c).

To summarize, all predictions for enantiomers and symmetric molecules with spectroscopic descriptors should be taken with due precaution, bearing in mind that any success may be fortuitous.

Performance of CoSA. In this case study, there are three pairs of enantiomers present in the data set (compounds **13/14**, **15/16**, and **17/21**). However, it should be emphasized that the differences in activities between the enantiomers are not very large: 0.96 (**13/14**), 0.80 (**15/16**), and 1.35 (**17/21**); cf. max = 2.47, min = −1.70, range = 4.17, mean = 0.95, and std = 1.25. This may partly explain the apparent success of the CoSA approach. Further, there are several compounds that possess C_2 or C_{2v} symmetry, and this will likely reduce the predictive ability of the spectroscopic descriptors, in particular for CoSA (cf. the LOO CV predictions for highly symmetric compounds **25** and **34**, Figure 3).

Bearing the above caveats in mind, the results indicate that the CoSA method with ^{13}C chemical shifts can still provide robust and predictive QSAR models for the estrogenic activity of molecules, corroborating the findings of previous CoSA tests on other estrogen data sets.[22,24] The results for the estrogen data set concerned here ($q^2 = 0.69$ and $S_{press} = 0.76$) are almost comparable to those reported in the original CoMFA work ($q^2 = 0.80$ and $S_{press} = 0.59$), but the results are clearly inferior to those achieved with a highly sophisticated CoMFA model that employs a receptor-based alignment of ligands and smart region definition (SRD) for variable selection ($q^2 = 0.92$ and $S_{press} = 0.35$). The results are not fully comparable, however, as large validation tests with randomized training and test sets, an essential feature of this work, were not performed in the previous cases.

Further, the reduced performance of CoSA is probably due to the fact that NMR-based QSAR descriptors are not sensitive enough for describing the small differences in features of molecular structure that determine the relative binding affinities of molecules in this fairly small data set. For larger data sets with more divergent structures, however, the performance of CoSA and other spectroscopic methods is probably more compelling among the 3D QSAR methods. Finally, it should be emphasized that the ^{13}C NMR data employed here were unassigned, i.e., no supplementary structural information was used. It has been shown previously that it is possible to achieve improved performance by adding assigned structural information to the one-dimensional CoSA-type QSAR model.[28]

In any case, the results indicate that the CoSA method can provide a promising alternative and supplement to conventional 3D QSAR methods, possessing many attractive features. It is computationally simple (especially if chemical shifts are available in advance), easy to use and invariant as to the alignment of the structures concerned. Even in the present case study with a small data set, however, a considerable amount of CPU time (about three weeks using a Silicon Graphics Origin2000 workstation) was needed to calculate the NMR chemical shifts for 36 organic molecules of moderate size. This may seem to rule out the use of CoSA with large data sets, although the NMR chemical shifts would only need to be calculated "once and for all", after which they would also be available for other potential applications. If the method is applied to large data sets involving large organic molecules, the experimental chemical shifts should preferably be available from common databases, although the problem can partially be circumvented, as ^{13}C shifts can be estimated with reasonable accuracy using a simple, fast prediction scheme. With ^{1}H shifts the situation is more complicated, as it is much more difficult to develop accurate predictors due to the strong influence of anisotropy effects. On the other hand, it can be anticipated that new theoretical advances, together with progress in computer technology and algorithms, will rapidly lead to a situation in which DFT/GIAO methods are a universal and powerful tool for computing NMR chemical shifts in organic molecules. This will make CoSA-type QSAR modeling more compelling.

Performance of SOMFA. The SOMFA method with molecular shape provides robust, predictive QSAR models. In fact, the SOMFA1 model is comparable to that presented in the original CoMFA work.[10] On the other hand, the model is still clearly inferior to the best CoMFA model[11,12] discussed above. It is evident that receptor-based alignment of molecules is superior over the field fit alignment used here. The results indicate that the internal and external predictabilities of SOMFA fields are identical; i.e., the molecular property that gives the best fit in the LOO CV tests is also the best in the tests with external data sets.

External Tests. Finally, the importance of thorough validation tests for QSAR studies is worth discussing. In particular, the value of LOO CV as a model validation method has been subject to much debate in the recent literature (see Golbraikh and Tropsha[41] vs Hawkins et al.[42]), and the topic clearly needs further elucidation. In view of the above, it seems reasonable to suggest that a large number of validation tests, using a large number of randomized training and test sets, would be worth performing in each QSAR study. These tests can be of help in selecting the best-performing molecular property for the QSAR modeling. Of course, external tests will become quite meaningless if the data set is very small in size. The bottom line is that the use of LOO CV, perhaps supplemented with a single external test set—a common practice in the field of QSAR—can be hazardous, as the results (regardless of the goodness of the fit) are certainly not representative enough to assess the true predictive ability of the QSAR model or method.

CONCLUSIONS

The results indicate that both the CoSA and SOMFA methods can be used to derive robust, predictive QSAR models for the estrogen activity of organic molecules, providing an alternative and supplement to other (3D) QSAR methods. There is a growing trend toward developing expert systems for use in ranking estrogen receptor binding affinities for large data sets.[43−47] In this context, the potential of both methods considered here seems worth exploring in more detail. It seems that a full 3D technique such as CoMFA or SOMFA is the method of choice for accurate predictions, although the estrogen receptor is not very selective with regard to ligands. This is reflected in the large number of organic molecules that have been shown to bind to it with considerable affinity.

REFERENCES AND NOTES

(1) Patlak, M. A Testing Deadline for Endocrine Disrupters. *Environ. Sci. Technol.* **1996**, *30*, 540A−544A.
(2) Schultz, T. W.; Cronin, M. T. D.; Netzeva, T. I. The Present Status of QSAR in Toxicology. *J. Mol. Struct. (THEOCHEM)* **2003**, *622*, 23−38.
(3) Fang, H.; Tong, W.; Welsh, W. J.; Sheehan, D. M. QSAR Models in Receptor-Mediated Effects: The Nuclear Receptor Superfamily. *J. Mol. Struct. (THEOCHEM)* **2003**, *622*, 113−125.

(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) Gao, H.; Katzenellenbogen, J. A.; Garg, R.; Hansch, C. Comparative QSAR Analysis of Estrogen Receptor Ligands. *Chem. Rev*. **1999**, *99*, 723−744.

(6) Zheng, W.; Tropsha, A. Novel Variable Selection Quantitative Structure−Property Relationship Approach Based on the *k*-Nearest-Neighbor Principle. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 185−194.

(7) Bradbury, S.; Kamenska, V.; Schmieder, P.; Ankley, G.; Mekenyan, O. A Computationally Based Identification Algorithm for Estrogen Receptor Ligands: Part 1. Predicting hERα Binding Affinity. *Toxicol. Sci.* **2000**, *58*, 253−269.

(8) Stanton, D. T.; Dimitrov, S.; Grancharov, V.; Mekenyan, O. G. Charged Partial Surface Area (CPSA) Descriptors QSAR Applications. *SAR QSAR Environ. Res.* **2002**, *13*, 341−351.

(9) Robinson, D. D.; Winn, P. J.; Lyne, P. D.; Richards, W. G. Self-Organizing Molecular Field Analysis: A Tool for Structure−Activity Studies. *J. Med. Chem.* **1999**, *42*, 573−583.

(10) 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.

(11) Sippl, W. Receptor-Based 3D QSAR Analysis of Estrogen Receptor Ligands—Merging the Accuracy of Receptor-Based Alignments with the Computational Efficiency of Ligand-Based Methods. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 559−572.

(12) Sippl, W. Binding Affinity Prediction of Novel Estrogen Receptor Ligands Using Receptor-Based 3-D QSAR Methods. *Bioorg. Med. Chem*. **2002**, *10*, 3741−3755.

(13) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. Evaluation of a Novel Infrared Range Vibration-Based Descriptor (EVA) for QSAR Studies. 1. General Application. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 409−422.

(14) Ginn, C. M. R.; Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of the EVA Descriptor and Combination of Rankings Using Data Fusion. *J. Chem. Inf. Comput. Sci*. **1997**, *37*, 23−37.

(15) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Evaluation of a Novel Molecular Vibration-Based Descriptor (EVA) for QSAR Studies: 2. Model Validation Using a Benchmark Steroid Dataset. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 271−296.

(16) Turner, D. B.; Willett, P. Evaluation of the EVA Descriptor for QSAR Studies: 3. The Use of a Genetic Algorithm to Search Models with Enhanced Predictive Properties (EVA_GA). *J. Comput.-Aided Mol. Des.* **2000**, *14*, 1−21.

(17) Turner, D. B.; Willett, P. The EVA Spectral Descriptor. *Eur. J. Med. Chem*. **2000**, *35*, 367−375.

(18) Tuppurainen, K. EEVA (Electronic Eigenvalue): A New QSAR/QSPR Descriptor for Electronic Substituent Effects Based on Molecular Orbital Energies. *SAR QSAR Environ. Res.* **1999**, *10*, 39−46.

(19) Tuppurainen, K.; Ruuskanen, J. Electronic Eigenvalue (EEVA): A New QSAR/QSPR Descriptor for Electronic Substituent Effects Based on Molecular Orbital Energies. *Chemosphere* **2000**, *41*, 843−848.

(20) Tuppurainen, K.; Viisas, M.; Laatikainen, R.; Peräkylä, M. Evaluation of a Novel Electronic Eigenvalue (EEVA) Molecular Descriptor for QSAR/QSPR Studies: Validation Using a Benchmark Steroid Data Set. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 607−613.

(21) Bursi, R.; Dao, T.; van Wijk, T.; De Gooyer, M.; Kellenbach, E.; Verwer, P. Comparative Spectra Analysis (Cosa): Spectra as Three-Dimensional Molecular Descriptors for the Prediction of Biological Activities. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 861−867.

(22) Beger, R. D.; Freeman, J. P.; Lay, J. O., Jr.; Wilkes, J. G.; Miller, D. W. ^{13}C NMR and Electron Ionization Mass Spectrometric Data−Activity Relationship Model of Estrogen Receptor Binding. *Toxicol. Appl. Pharmacol*. **2000**, *169*, 17−25.

(23) Beger, R. D.; Freeman, J. P.; Lay, J. O., Jr.; Wilkes, J. G.; Miller, D. W. Producing ^{13}C NMR, Infrared Absorption, and Electron Ionization Mass Spectrometic Data Models of the Monodechlorination of Chlorobenzenes, Chlorophenols, and Chloroanilines. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 1449−1455.

(24) Beger, R. D.; Freeman, J. P.; Lay, J. O., Jr.; Wilkes, J. G.; Miller, D. W. Use of ^{13}C NMR Spectrometric Data To Produce a Predictive Model of Estrogen Receptor Binding Affinity. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 219−224.

(25) Beger, R. D.; Wilkes, J. G. Models of Polychlorinated Dibenzodioxins, Dibenzofurans, and Biphenyls Binding Affinity to the Aryl Hydrocarbon Receptor Developed Using ^{13}C NMR Data. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1322−1329.

(26) Beger, R. D.; Buzatu, D. A.; Wilkes, J. G.; Lay, J. O., Jr. ^{13}C NMR Quantitative Spectrometric Data−Activity Relationship (QSDAR) Models of Steroids Binding the Aromatase Enzyme. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1360−1366.

(27) Beger, R. D.; Wilkes, J. G. Developing ^{13}C NMR Quantitative Spectrometric Data−Activity Relationship (QSDAR) Models of Steroid Binding to the Corticosteroid Binding Globulin. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 659−669.

(28) Beger, R. D.; Buzatu, D. A.; Wilkes, J. G.; Lay, J. O., Jr. Comparative Structural Connectivity Spectra Analysis (CoSCoSA) Models of Steroid Binding to the Corticosteroid Binding Globulin. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 1123−1131.

(29) 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.

(30) Wolinski, K.; Hinton, J. F.; Pulay, P. Efficient Implementation of the Gauge-Independent Atomic Orbital Method for NMR Chemical Shift Calculations. *J. Am. Chem. Soc*. **1990**, *112*, 8251−8260.

(31) Kolehmainen, E.; Koivisto, J.; Nikiforov, V.; Peräkylä, M.; Tuppurainen, K.; Laihia, K.; Kauppinen, R.; Miltsov, S. A.; Karavan, V. S. NMR Spectroscopy in Environmental Chemistry: ^{1}H and ^{13}C NMR Chemical Shift Assignments of Chlorinated Dibenzothiophenes Based on Two-Dimensional NMR Techniques and *ab Initio* MO and DFT/GIAO Calculations. *Magn. Reson. Chem*. **1999**, *37*, 743−747.

(32) Koivisto, J.; Kolehmainen, E.; Nikiforov, V.; Nissinen, M.; Tuppurainen, K.; Peräkylä, M.; Miltsov, S.; Karavan, V. Synthesis, Structures and Spectroscopy of Polychlorinated Dihydrocamphenes. An Experimental and Theoretical Study. *ARKIVOC (iii)* **2001**, 95−113.

(33) Tuppurainen, K.; Ruuskanen, J. NMR and Molecular Modeling in Environmental Chemistry: Prediction of ^{13}C Chemical Shifts in Selected C_{10}-Chloroterpenes Employing DFT/GIAO Theory. *Chemosphere* **2003**, *50*, 603−609.

(34) *The Aldrich Library of ^{13}C and ^{1}H FT NMR Spectra*, 1st ed.; Pouchert, C. J., Behnke, J., Eds.; Aldrich Chemical Co.: Milwaukee, WI, 1993; Vol. 1−3.

(35) Frisch, M. J.; et al. *GAUSSIAN 98*, Revision A.7; Gaussian, Inc.: Pittsburgh, PA, 1998.

(36) Bradley, M.; Waller, C. L. Polarizability Fields for Use in Three-Dimensional Quantitative Structure−Activity Relationship (3D-QSAR). *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1301−1307.

(37) Lewis, D. F. V. The Calculation of Molecular Polarizabilities by the CNDO/2 Method: Correlation with the Hydrophobic Parameter, LogP. *J. Comput. Chem*. **1989**, *10*, 145−151.

(38) Wang, T. W.; Khettry, A.; Berry, M.; Batra, J. SVDPLS: An Efficient Algorithm for Performing PLS. *The First International Chemometrics InterNet Conference, INCINC'94*. (The MATLAB code of SVDPLS can be found from the WWW site http://www.emsl.pnl.gov:2080/docs/incinc/papers/wang_pls/sectionstar3_9.html).

(39) Cramer, R. D., III; DePriest, S. A.; Patterson, D. E.; Hecht, P. The Developing Practice of Comparative Molecular Field Analysis. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, The Netherlands, 1993.

(40) Golbraikh, A.; Tropsha, A. QSAR Modeling Using Chirality Descriptors Derived from Molecular Topology. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 144−154.

(41) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graphics Model.* **2002**, *20*, 269−276.

(42) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing Model Fit by Cross-Validation. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 579−586.

(43) Mekenyan, O. G.; Kamenska, V.; Serafimova, R.; Poellinger, L.; Brouwer, A.; Walker, J. Development and Validation of an Average Mammalian Estrogen Receptor-Based QSAR Model. *SAR QSAR Environ. Res*. **2002**, *13*, 579−595.

(44) Shi, L.; Tong, W.; Fang, H.; Perkins, R.; Wu, J.; Tu, M.; Blair, R. M.; Branham, W. S.; Waller, C.; Sheehan, D. M. An Integrated "Four-Phase" Approach for Endocrine Disruptor Priority Setting—Part 1: Phase I and II Predictions of Potential Estrogenic Endocrine Disruptors. *SAR QSAR Environ. Res*. **2002**, *13*, 69−88.

(45) Hong, H.; Tong, W.; Fang, H.; Shi, L.; Xie, Q.; Wu, J.; Perkins, R.; Walker, J. D.; Branham, W.; Sheehan, D. M. Prediction of Estrogen Receptor Binding for 58,000 Chemicals Using an Integrated System of a Tree-Based Model with Structural Alerts. *Environ. Health Perspect*. **2002**, *110*, 29−36.

(46) Suzuki, T.; Ide, K.; Ishida, M.; Shapiro, S. Classification of Environmental Estrogens by Physicochemical Properties Using Principal Component Analysis and Hierarchical Cluster Analysis. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 718−726.

(47) Tong, W.; Hong, H.; Fang, H.; Xie, Q.; Perkins, R. Decision Forest: Combining the Precictions of Multiple Independent Decision Tree Models. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 525−531.

CI034110B