Virtual Screening for Aryl Hydrocarbon Receptor Binding Prediction

Elena Lo Piparo,*,† Konrad Koehler,‡ Antonio Chana,† and Emilio Benfenati†

Istituto di Ricerche Farmacologiche "Mario Negri", Via Eritrea 62, 20157 Milano, Italy, and Karo Bio Computational & Medicinal Chemistry Novum, SE-141 57 Huddinge, Sweden

Received May 5, 2006

The overall goal of this study has been to validate computational models for predicting aryl hydrocarbon receptor (AhR) binding. Due to the unavailability of the AhR X-ray crystal structure we have decided to use QSARs models for the binding prediction virtual screening. We have built up CoMFA, Volsurf, and HQSAR models using as a training set 84 AhR ligands. Additionally, we have built a hybrid model combining two of the final selected models in order to give a single operational system. The results show that CoMFA, VolSurf, HQSAR, and the hybrid models gives good results (R^2 equal to 0.91, 0.79, 0.85, and 0.82 and q^2 0.62, 0.58, 0.62, and 0.70, respectively). Since the techniques analyzed show a good correlation and good prediction also for an external test set, particularly the HQSAR and the hybrid model, we can conclude that these models can be used for predicting AhR binding in virtual screening.

Introduction

Nuclear receptors are involved in the regulation of critical cellular processes such as regulation of cell growth, differentiation, and metabolic processes.[1] They constitute an important super family of transcription regulators that include the dioxin/aryl hydrocarbon receptor (AhR). Free AhR is located in the cytoplasm, associated with heat shock proteins. Ligand binding to AhR is presumed to produce conformational changes in the AhR protein, causing the translocation of the whole complex into the nucleus.[2,3] Within the nucleus, the AhR−ligand complex dissociates from associated proteins and dimerizes with ARNT (its nuclear partner) to reconstitute an active transcription factor that binds specific DNA sequences.[4] Other AhR ligands such as dibenzo-*p*-dioxin (TCDD) and coplanar polychlorinated biphenyls (PCBs) are potent toxicants widespread in the environment. Their resistance to metabolic breakdown along with their lipophility causes them to accumulate in the food chain, bringing about their relevant effects on human health.[5,6]

The enormous number of compounds within the human food supply makes it impracticable to screen all of them for nuclear receptor binding experimentally. However computational procedures are available that can rapidly assess the likelihood of a given compound to bind a given receptor.[7−9] These rapid in silico methods can be used to prioritize compounds for follow-up experimental verification of nuclear receptor binding.

Computational methods for affinity prediction may be broadly classified into two categories.[10] When a detailed 3D structure of the protein receptor is known, then receptor fitting approaches can be done by docking a candidate ligand into the receptor cavity and using either molecular mechanics or an empirical scoring function to estimate the interaction energy, hence the affinity between the ligand and the receptor. The receptor structure can be obtained experimentally (e.g., X-ray crystallographic or NMR). Alternatively, if the structure has not been determined, but experimental structure(s) of closely related proteins are available, a homology model can be created by threading the sequence of the target receptor through an experimental template and mutating the corresponding amino acid residues in the template to match those of the target receptor. If no experimental structure or homology model is available, then a second method, called *receptor* mapping, can be used, which attempts to build a model of the receptor based on what binds to it. A variant of receptor mapping is 3D-QSAR (quantitative structure−activity relationship) in which a series of ligands with known affinity is aligned and then the strengths of the electrostatic and steric potentials of each ligand at regular grid points surrounding the molecule are correlated with the affinity of the compound. Once a QSAR regression equation has been generated for a training set of molecules, it can be used to predict the affinity of molecules not included in the training set. Another alternative to *receptor mapping* is to use other QSAR techniques that do not need structures optimization and superimposition. They are fast, user-friendly, and do not need human supervision, so they fit the purpose of virtual screening very well.

To identify unknown endocrine disruptors in the food supply, the AhR is a critical receptor, since no crystallographic structure of this receptor is available upon which to base a virtual screen. For this reason we created a homology model of the ligand binding domain (LBD) of AhR based on the NMR structure of the C-terminal PAS domain of human HIF-2a (PDB code 1P97).[11] However because of the relatively low sequence homology (∼25%) between the target AhR and the experimental template HIF-2a, virtual screening using this homology model is not likely to be accurate enough. As an alternative, we used QSAR models based on the experimentally determined binding affinities of dioxin and other families of AhR ligands to predict the activity of new ligands. Using published data and 3D-QSAR models by Waller and McKinney,[12] we added further descriptors and we explored more sophisticated and automatic methods for ligands alignment (maximizing overlap of steric and electrostatic fields), which is the critical phase that determines the quality and the utilization of the resultant 3D-QSAR model for the prediction of a big library of compounds. Additionally, we explored alternative tools for virtual screening. One of the proposed within this work, VolSurf,[13] does not require the alignment of the molecules; the other alternative, Hologram

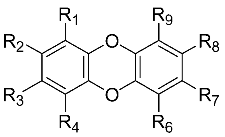
* To whom correspondence should be addressed. Present address: Bioinformatics Group, Department of Bioanalytical Science, Nestle Research Center, P.O. Box 44, CH-100000 Lausanne 26, Switzerland. Telephone: (0) 21 785 9530. Fax: (0) 21 785 9486. E-mail: elena.lopiparo@rdls.nestle.com.

† Istituto di Ricerche Farmacologiche "Mario Negri".

‡ Karo Bio Computational & Medicinal Chemistry Novum.

10.1021/jm060526f CCC: $33.50
Published on Web 08/22/2006

Table 1. Dibenzo-*p*-dioxins



compd[a]	R_1	R_2	R_3	R_4	R_6	R_7	R_8	R_9	pIC_{50}	log *P*	SE
1	H	Cl	Cl	H	H	Cl	Cl	H	9.144	6.35	0
2	Cl	Cl	Cl	H	H	Cl	Cl	H	8.118	6.84	0
3	H	Cl	Cl	H	Cl	Cl	H	H	7.768	6.22	0
4	H	Cl	Cl	H	Cl	H	H	H	7.61	5.74	0
5	Cl	Cl	Cl	Cl	H	Cl	Cl	H	7.49	7.32	0
6*	Cl	H	Cl	H	H	Cl	Cl	H	6.975	6.43	0
7*	Cl	Cl	H	Cl	H	Cl	Cl	H	6.811	6.92	0
8	Cl	Cl	Cl	Cl	H	H	H	H	6.728	5.84	0
9	H	Cl	Cl	H	H	Cl	H	H	8.171	5.74	0
10	H	Cl	H	H	H	H	Cl	H	6.281	5.12	0
11	Cl	Cl	Cl	Cl	H	Cl	H	H	5.937	6.71	0
12	Cl	Cl	H	Cl	H	H	H	H	5.585	5.43	0
13	Cl	Cl	Cl	Cl	Cl	Cl	Cl	Cl	5.715	8.30	0
14	Cl	H	H	H	H	H	H	H	4.572	4.10	0
15*	H	Br	Br	H	H	Br	Br	H	10.086	6.99	0
16*	H	Br	Br	H	H	Cl	Cl	H	10.093	6.67	0
17*	H	Br	Cl	H	H	Cl	Br	H	10.687	6.67	0
18*	H	Br	Cl	H	H	Cl	Cl	H	9.074	6.51	0
19	Br	H	Br	H	H	Br	H	Br	8.038	7.15	0
20	Br	H	Br	H	H	Br	Br	H	9.943	7.07	0
21	Br	Br	H	Br	H	Br	Br	H	8.881	7.72	0
22*	Br	Br	Br	H	H	Br	Br	H	9.35	7.64	0
23	H	Br	Br	H	H	Br	H	H	10.209	6.22	0
24	H	Br	H	H	H	Br	H	H	8.927	5.45	0
25	H	Br	H	H	H	H	H	H	7.464	4.41	0

[a] The compounds of the test set are marked with an asterisk.

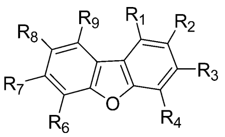
QSAR, is based on molecular fragments and also does not require structure optimization.

Experimental Details

Data Set. We used a set of 93 AhR ligands, splitted into a training and test sets, with experimental binding affinities as in the original article by Waller and McKinney,[12] considering the negative logarithm of the chemical molecular concentration necessary to displace 50% of radiolabeled 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) from the Ah receptor and reported as pIC_{50}. These data came from three different laboratories using 2,3,7,8-tetrachlorodibenzofuran (TCDF) as internal standard, to provide normalization for interlaboratory variability. All pIC_{50} values used to build up the models were normalized to a value of 8.444 for TCDF. From the original data set of 99 compounds it was decided to remove five of them, for which precise binding data were not available, and their removal greatly improved also the Waller et al. model.[12] Two compounds in the data set used by Waller et al. have identical structures and affinities (dibenzofurans called **57** and **63** in the original article), so we used only one, eliminating the duplication. Our models were through and applied for the screening of a list of artificial chemicals with high exposure risk identified given by the European Commission within CASCADE (EU contract no. FOOD-CT-2004-506319). Within the list of compounds to screen only nine have known binding activity, and we recognized that they were already included in the data set selected. Therefore, we excluded them from the training set and they have been used as the external set to validate our virtual screening method. Thus, in this work we used 84 compounds as a training set, plus the nine compounds in the external test set. They include the dibenzo-*p*-dioxins, dibenzofurans, biphenyls, naphthalenes, indolocarbazoles, and indolocarbazoles derivatives listed in Tables 1−7.

Molecular Modeling and Alignment Rules. The 3D atomic coordinates of the compounds were extracted from the SMILES using CORINA software.[14] A rough geometry optimization was performed using Schrodinger premin and Schrodinger Bmin MMFF/MCMM for a stochastic conformational search including full geometry optimization to find the global energy minimum.[15]

Table 2. Dibenzofurans



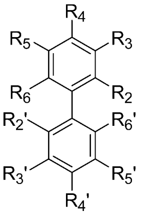
compd	R_1	R_2	R_3	R_4	R_6	R_7	R_8	R_9	pIC_{50}	log *P*	SE
26	H	Cl	H	H	H	H	H	H	4.061	4.48	0
27	H	H	Cl	H	H	H	H	H	5.003	4.48	0
28	H	H	H	Cl	H	H	H	H	3.429	4.48	0
29	H	Cl	Cl	H	H	H	H	H	6.088	5.04	0
30	H	Cl	H	H	Cl	H	H	H	4.125	5.12	0
31	H	Cl	H	H	H	H	Cl	H	4.103	5.12	0
32	Cl	H	Cl	H	Cl	H	H	H	6.123	5.82	0
33	Cl	H	Cl	H	H	H	Cl	H	4.653	5.82	0
34	H	Cl	Cl	Cl	H	H	H	H	5.396	5.59	0
35	H	Cl	Cl	H	H	H	Cl	H	6.858	5.67	0
36	H	Cl	H	H	Cl	Cl	H	H	7.255	5.67	0
37	H	Cl	Cl	Cl	Cl	H	H	H	7.379	6.23	0
38	H	Cl	Cl	Cl	H	H	Cl	H	7.657	6.23	0
39	Cl	H	Cl	H	Cl	H	Cl	H	7.61	6.53	0
40	H	Cl	Cl	H	H	Cl	Cl	H	8.444	6.23	0
41	Cl	Cl	H	Cl	H	H	Cl	H	5.715	6.42	0
42	Cl	Cl	H	Cl	Cl	Cl	H	H	8.194	6.98	0
43	Cl	Cl	H	Cl	H	Cl	H	Cl	5.371	7.13	0
44	Cl	Cl	Cl	Cl	H	H	Cl	H	7.911	6.83	0
45	Cl	Cl	Cl	H	H	Cl	Cl	H	8.147	6.79	0
46	Cl	Cl	H	Cl	H	Cl	Cl	H	6.728	6.98	0
47	H	Cl	Cl	Cl	H	Cl	Cl	H	8.943	6.79	0
48	Cl	Cl	Cl	Cl	H	Cl	Cl	H	7.587	7.39	0
49	Cl	Cl	Cl	H	Cl	Cl	Cl	H	7.508	7.34	0
50	Cl	Cl	H	Cl	Cl	Cl	Cl	H	5.808	7.53	0
51	H	Cl	Cl	Cl	Cl	Cl	Cl	H	8.376	7.34	0
52	H	Cl	Cl	H	Cl	H	Cl	H	7.61	6.38	0
53	Cl	Cl	Cl	H	Cl	H	H	H	7.379	6.23	0
54	Cl	Cl	Cl	H	H	Cl	H	H	7.954	6.23	0
55	Cl	H	Cl	Cl	H	Cl	Cl	H	7.657	6.98	0
56	H	Cl	Cl	Cl	H	Cl	H	Cl	7.657	6.93	0
57	Cl	Cl	Cl	H	H	Cl	H	Cl	7.313	6.93	0
58	H	H	H	H	H	H	H	H	3.429	3.84	0
59	H	Cl	Cl	Cl	H	Cl	H	H	8.689	6.23	0
60	Cl	Cl	Cl	H	H	Cl	H	H	7.954	6.23	0
61	Cl	H	Cl	Cl	H	Cl	Cl	H	7.623	6.98	0
62	H	Cl	Cl	Cl	H	Cl	H	Cl	7.623	6.93	0
63	Cl	Cl	H	Cl	Cl	H	Cl	H	6.297	7.13	0

We also considered some constrains of the central torsion angle to 0° for nonplanar ligands (e.g., biphenyls) using MacroModel/MMFF and Maestro 7.0 graphical user interfaces and setting the torsion constraint to 4.182 kJ/mol (1 kcal/mol). Charges used within this work were based on the density functional B3LYP ab initio calculations at the 6-311G+ level.

The SEAL program that maximizes the overlap of steric and electrostatic fields[16] was used for aligning the ligands. Aligning the ligands using SEAL with the electrostatic parameter reduced the default value of 1.0−0.33, increases the alignment quality of the smaller ligands such as biphenyls that achieve better results, since steric components are now more heavily weighted than electrostatic ones.

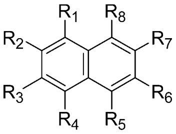
Calculation of Descriptors (CoMFA, VolSurf, Hologram, log *P*, SE). The CoMFA analysis was done on a Linux-based PC workstation using the software package SYBYL version 7.1.[17] The aligned molecules were placed in a three-dimensional grid space with the dimensions automatically set by the program and 1.5 Å (*x*, *y*, *z*) grid stepping. An absolute maximum of 30 kcal/mol for the steric and electrostatic energy calculated at each grid point was established experimentally. The CoMFA descriptors in terms of van der Waals (steric) and Coulombic (electrostatic) interactions were calculated using an sp^3 carbon atom with a +1 charge as a probe. Equal weights were assigned to steric and electrostatic fields using the CoMFA standard scaling procedure implemented in SYBYL. The important issue of the reduction of the number of descriptors was considered.[18] In the case of the CoMFA model,

Table 3. Biphenyls



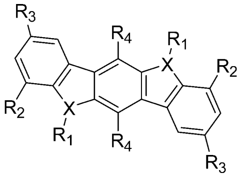
compd	R_2	R_3	R_4	R_5	R_6	$R_{2'}$	$R_{3'}$	$R_{4'}$	$R_{5'}$	pIC_{50}	log *P*	SE
64	H	Cl	Cl	H	H	H	Cl	Cl	H	7.028	6.41	1
65	H	Cl	Cl	Cl	H	H	H	Cl	H	5.204	6.37	1
66	H	Cl	Cl	Cl	H	H	Cl	Cl	H	7.871	6.99	1
67	Cl	H	H	H	H	H	Cl	Cl	Cl	5.584	6.24	1
68	Cl	Cl	Cl	H	H	H	Cl	Cl	H	6.134	6.82	1
69	Cl	H	Cl	Cl	H	H	Cl	Cl	H	5.762	6.99	1
70	Cl	Cl	Cl	Cl	H	H	H	Cl	H	6.157	6.78	1
71	Cl	Cl	Cl	Cl	H	H	Cl	Cl	H	6.057	7.40	1
72	Cl	H	Cl	Cl	H	H	Cl	Cl	Cl	5.482	7.57	1
73	Cl	Cl	Cl	Cl	H	H	Cl	Cl	Cl	5.885	7.98	1
74	Cl	H	Cl	H	H	Cl	H	Cl	H	4.442	6.41	1
75	Cl	H	Cl	Cl	H	Cl	H	Cl	Cl	4.689	7.57	1
76	Cl	Cl	Cl	Cl	H	H	H	H	H	4.405	6.10	1
77	Cl	H	Cl	H	Cl	H	Cl	Cl	Cl	4.577	7.57	1

Table 4. Naphthalenes



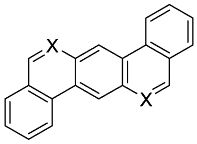
compd	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	pIC_{50}	log *P*	SE
78	H	Br	Br	H	H	H	H	H	5.616	4.67	0
79	H	Br	Br	H	H	Br	Br	H	7.668	6.19	0
80	Br	Br	H	Br	H	Br	Br	H	7.465	7.10	0
81	Br	Br	Br	Br	H	Br	Br	H	7.608	7.66	0
82	Br	Br	Br	H	Br	Br	Br	H	7.996	7.62	0

Table 5. Indolocarbazoles



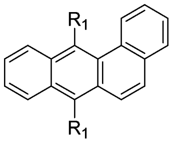
compd	X	R_1	R_2	R_3	R_4	pIC_{50}	log *P*	SE
83	N	CH_3	H	H	H	8.921	5.55	0
84	S	H	H	H	H	8.482	6.57	0
85	N	H	H	H	H	8.444	5.14	0
86	N	CH_2CH_3	H	H	H	8.051	6.48	0
87	N	$COCH_3$	H	H	H	7.951	3.64	0
88	N	H	CH_3	H	H	7.721	6.26	0
89	O	H	H	H	H	7.538	5.74	0

Table 6. Indocarbazole Derivates



compd	X	pIC_{50}	log *P*	SE
90	C	8.602	6.97	0
91	N	6.863	4.84	0

Table 7. Indocarbazole Derivates



compd[a]	R_1	pIC_{50}	log *P*	SE
92*	H	7.319	5.67	0
93*	CH_3	6.857	6.67	0

where thousands of descriptors are calculated during the analyses, the performance of different models considering the balance between grid stepping and molecular filtering was assayed. Finally, we selected a small grid stepping of 1.5 Å and an high column filtering of 2.5 to take into account most of the information and to avoid the problem of the computational time being intolerably long. This “column filtering” technique reduced the number of columns in the QSAR Molecular Spread Sheet to 435.

We have calculated as well VolSurf descriptors, which do not need structures to be aligned. The program calculates energetically favorable interaction sites around the molecules to produce a 3-D molecular interaction fields (MIF)[19] grid map that is compressed into a few quantitative 2-D numerical descriptors that are physicochemically meaningful.[13,20] We used five probes (water, hydrophobic, carbonyl oxygen, carboxy oxygen, and amphipathic) to characterize the interaction sites around target molecules. Three-dimensional molecular field maps were transformed into 118 descriptors by VolSurf 4.0. Such descriptors were molecular volume (V), surface (S), molecular weight (MW), critical packing (CP), size of the hydrophilic (W) and hydrophobic (D) region, hydrogen-bonding properties (HB), integy moments and hydrophobic integy moment, and local interaction energy minima, which represent the energy of the best three local minima of interaction energies between the water probe and the compound. Integy moments are vectors pointing from the center of the mass to the center of hydrophilic and hydrophobic regions, respectively.

The HQSAR approach uses as molecular descriptors holograms that encode a fixed length array containing counts of a priori defined substructural fragments. This method uses only 2D structure information, thus avoiding the usual conformational flexibility and structure alignments problems. Holograms were generated using the standard parameters implemented in Sybyl 7.1. Molecular fragments were generated using the fragment size default (minimum 4, maximum 7) and the following fragment distinctions: atoms, bonds, and connections. The HQSAR analysis was done by screening the 12 default series of hologram length values from 53 to 401 bins. The fragment patterns counts from the training set compounds were then related to the measured biological activity, and the best HQSAR model hologram length found was 257 bins.

The logarithm of octanol−water partition coefficient, log *P*, was calculated using the Pallas 3.0[21] package.

An indicator variable that takes into account strain energy (SE) was also added as descriptor for all ligands that were not torsionally constrained and in fact are, in the global energy minimum, set SE = 0.0 and, for torsionally constrained ligands, set SE = 1.0.

Statistical Analysis. Statistical analysis was done using the partial least squares (PLS) method as employed in the QSAR module of SYBYL 7.1 and VolSurf 4 running on a Linux-based PC workstation. PLS is based on linear transformation of the descriptors' space, producing a new variable space based on a small number of orthogonal factors (latent variables), so there is no correlation. This method is particularly useful when the number of variables equals or exceeds the number of compounds (data points), because it leads to stable, correct, and highly predictive models, even for correlated descriptors.[22,23] To establish the maximal number of components contributing to the model with the lowest standard error of estimate (SEE), PLS analysis was combined with leave-one-out (LOO) cross-validation (CV). LOO-CV implies exclusion of each compound of the training set and the prediction of its activity by the model developed using the remaining compounds. To assess the goodness of the model we used the cross-validated coefficient q^2, which expresses the model's ability to reproduce the training set. It was calculated as follows:

$$q^2 = 1 - \frac{\sum (Y_{\text{pred}} - Y_{\text{obsd}})^2}{\sum (Y_{\text{obsd}} - Y_{\text{mean}})^2} \qquad (1)$$

where Y_{pred}, Y_{obsd}, and Y_{mean} are predicted, actual, and mean pIC$_{50}$, and $\sum(Y_{\text{pred}} - Y_{\text{obsd}})^2$ is the predictive sum of squares known as PRESS. For each model, the LOO-CV predictions were examined. The models were subsequently validated also using cross-validation with 10 groups. In this way, the data set is randomly divided into 10 sets with approximately equal size and class distributions. The model is trained afterward using all but one of the 10 groups and then tested on the unseen group. This procedure is repeated for each of the 10 groups. The cross-validation score is the average performance across each of the 10 training runs.

After establishing the optimal number of components, the PLS procedure was repeated without cross-validation while being given the exact number of components contributing to the final model as input. When additional columns (SE and logP descriptors) were added to the molecular spread sheet, the QSAR standard scaling method was set during the PLS analysis to weight the SE and log *P* columns as heavily as the CoMFA and Hologram descriptors.

Since the high LOO q^2 is the necessary condition but not a sufficient one for a model to have a high predictive power,[24] we also used cross-validation with 10 groups and an external test set of compounds never seen by the models. For the ideal model, the slope and the correlation coefficient is equal to 1 and intercept is equal to 0. Since a good QSAR model may have a high predictive ability if it is close to the ideal one, we have set the intercept of the test set plot to 0.

A hybrid model was also developed in order to improve the final results. To build combined models we used in-house software built as a PC-Windows Excel macro. We have selected the rule-based approach that consists of dividing the results interval into three main areas, where some of the values maximum, minimum, or mean from the selected models will rule the trend of the final model. In this way we have obtained a noncontinuous function that can be expressed as combinations of simple linear equations such as:

$$\text{pIC}_{50\ \text{calc}} = k_n[\text{Min,Mean,Max(models to combine)}] + a_n \qquad (2)$$

The result can be regarded as a new set of rules whose final expression is a hybrid system able to combine different models. The final expression found can be expressed in the following way.

If mean(models to combine) > 6.642

$$\text{pIC}_{50\ \text{calc}} = 1.049[\text{Min(models to combine)}] + 0.002$$

If mean(models to combine) > 3.604

$$\text{pIC}_{50\ \text{calc}} = 1.005[\text{Min(models to combine)}] - 0.009$$

Otherwise

$$\text{pIC}_{50\ \text{calc}} = 1.021[\text{Mean(models to combine)}] - 0.106$$

A systematic variation of the combinations of maximum, minimum, and mean values, a decision on which models to combine, and a later optimization of the values of V_1, V_2, k_1, a_1, k_2, a_2, k_3, a_3 to give a better value for r^2 were carried out. The optimization has been performed by means the downhill simplex method modified to search uphill for the higher r^2 values. A cross-validation has been also performed to obtain the q^2 value as described in eq 1 for the new hybrid model.

Results

Molecular Modeling and Alignment Rules. The constraint of the central torsion angle to 0° for nonplanar ligands (biphenyls) had very little effect on the torsion angle (<1°) for biphenyl without ortho substituents, because the potential energy surface is very steep. However, for biphenyls with one or more ortho and ortho-prime substituents, the constrained optimization had more effect on the order of 10−15° twisting relative to the global energy minimum (due to a shallower potential energy surface). The biphenyls are thus somewhat more planar, but the central torsion angles are still bigger than 45°. This approach does not force the biphenyls too much into a planar conformation. Although crystallographic structures of biphenyls are planar and the most potent AhR ligands are planar, it is also true that the crystallographic structure of biphenyls is planar because of the very large crystal packing force in small molecule crystals, but this is not necessarily true for protein/ligand complexes. Furthermore, the crystallographic structures show an average conformation. The true conformation may be a 50−50 mixture of two nonplanar conformations that when averaged appear planar. Finally, there are also small molecule crystallographic structures of biphenyl with one or more ortho subsituents that deviate significantly from planarity. In any case, the more toxic PCBs described, like PCB 126, do not have chlorine atoms in ortho positions.

Statistical Analysis. For the prediction of AhR binding affinity we applied CoMFA, Hologram, and VolSurf analysis to the pIC$_{50}$ of a data set of dioxins and dioxin-like compounds. Tables 1−7 list the experimental pIC$_{50}$, the log *P*, the strain energy descriptors and the corresponding structures of the compounds studied. Several CoMFA, HQSAR, and VolSurf models were developed, and the best models were selected according to the lowest number of components, the best statistical results after PLS analysis, and the greatest predictive power for the external test set. Table 8 gives a summary of the PLS analysis for the best models selected and reports the relative contributions of each model together with the optimal number of components, q^2 LOO, and q^2 CV using 10 groups and R^2.

log *P*, which is a measure of the compound's lipophilicity and a crude measure of desolvation energy, was added as an additional descriptor because numerous studies frequently have shown log *P* to be important in predictive QSAR models. In fact, it does generally well describes the bioavailability of a

Table 8. Summary of CoMFA, VolSurf, HQSAR Results for the Training Set

	CoMFA	HQSAR	VolSurf	hybrid
optimal no. of components	7	9	5	2
q^2 LOO	0.62	0.62	0.58	0.70
q^2 CV 10 groups	0.62	0.66	0.50	-
R^2	0.91	0.85	0.79	0.73
	Contributions			
CoMFA steric	0.31	N/A	N/A	N/A
CoMFA electrostatic	0.64	N/A	N/A	N/A
Hologram257	N/A	0.46	N/A	N/A
SE	0.05	0.03	N/A	N/A
log *P*	N/A	0.27	N/A	N/A
(log *P*)^2	N/A	0.23	N/A	N/A

chemical to the organism. Since octanol can represent the cell membrane, log *P* indicates the chemical's ability to permeate it and to be available for interaction with the organism. In general, the more hydrophobic the ligand, the higher the AhR binding affinity. The square of log *P* was also included to allow for a parabolic relationship between affinity and log *P*, since very hydrophobic ligands may have solubility problems and therefore lower apparent affinity. This was the case of the HQSAR model, which gave better results with this additional descriptor, so this model was selected and reported here.

The results for all models show good LOO regression coefficients, 0.62, 0.62, and 0.58, respectively, for CoMFA, HQSAR, and VolSurf. In all cases, no significant differences were observed using leave-more-out with 10 groups and the cross-validation regression coefficients (0.62, 0.66, and 0.5, respectively, for CoMFA, HQSAR, and VolSurf); this highlights the stability and predictivity of the models.

For practical reasons, the target of this work was to look for a general model with an adequate predictive power useful for virtual screening; thus, we have decided to carry out the combination of the models by means a rule-based approach. Such an approach was tested to analyze how prediction performances can be increased by combining individual models. The best result was obtained by the combination of the HQSAR and Volsurf models. The model obtained gave better results than the individual ones (q^2 LOO = 0.70, R^2 test set = 0.73).

The CoMFA model shows the importance of the alignment rules and the improvement with the use of the SEAL program, not only for the automatic, reproducible, and fast way of aligning the molecule, but also the easier application of the model in predicting external compounds. In fact, the CoMFA model is limited for fast virtual screening of a large compound libraries when very complicated and different alignment rules are applied to the dataset. Our CoMFA model also shows the importance of an additional descriptor (SE). The introduction of an indicator variable that takes account of strain energy (SE equal to 0 for all ligands that were not torsionally constrained and 1 for torsionally constrained ligands) adds information about the different energy minimization rules for the different families of compounds and confirms the importance of not forcing the biphenyls too much into a planar conformation. For the optimized CoMFA model the contribution parameters that depict the relevance of the descriptors are 0.31, 0.64, and 0.05 respectively from steric, electrostatic field, and log *P*, showing the more relevant electrostatic properties of these compounds.

The statistical CoMFA analyses can be confirmed visually on the contour plot (Figure 1) using the most active compound of the training set (**23**) for visualization. In this plot positive steric contributions are represented in green, negative contributions in yellow, positive electrostatic contributions in blue, and negative ones are red. The CoMFA contour map shows several red areas around the R$_2$, R$_3$, R$_7$, and R$_9$ dibenzo-*p*-dioxin substituents, representing regions where an electronegative environment would enhance the toxicity. Another red contour, near one of the dioxin ring oxygen atoms, indicates that high electron density may play a negative role in the toxicity of these compounds. The four blue contours encompassing the R$_1$, R$_3$, R$_4$, and R$_8$ dibenzo-*p*-dioxin substituents in the template molecules indicate regions where electropositive groups increase the activity. There is also a big green area around the R$_7$ substituent, indicating a sterically favorable region. The yellow steric region near the oxygen of the para dioxin ring and the substituents R$_2$, R$_3$, R$_4$, and R$_6$ suggests that the bulkier substituents may reduce the activity.

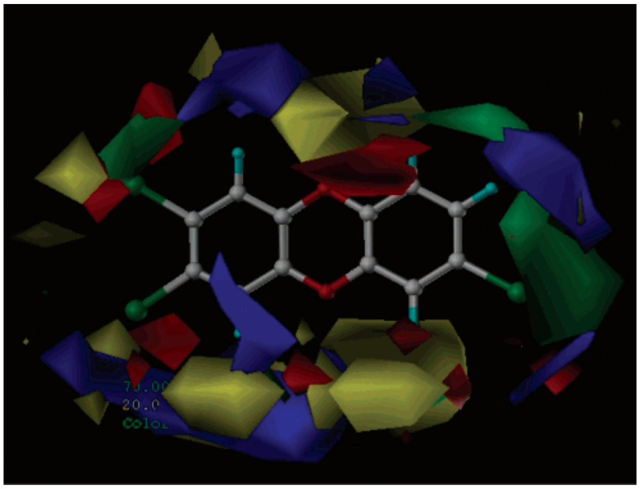


Figure 1. CoMFA contour plot map using the most active compound of the training set (**23**) for visualization: positive steric contributions are in green, negative contributions are yellow, positive electrostatic contributions in blue, and negative in red.

The need for structural alignment camplicates 3D QSAR, but an interesting alternative is using VolSurf models. They are fast and completely independent of the alignment procedures, so they fit very well for fast virtual screening. Our VolSurf model gave good results with q^2 LOO being smaller than that for CoMFA, but using fewer components for the analysis. VolSurf descriptors quantitatively characterize size, shape, polarity, hydrophobicity, and the balance between them. We analyzed the VolSurf descriptors profile for compound **23** by means of the PLS coefficient plot. The activity increases particularly with high values of hydrophilic regions (WOH2); capacity factors (CwOH2), which measure the amount of hydrophilic regions per surface unit; local interaction energy minima distance of water probe (DOH2); hydrogen bonding (HB); log *P*; molecular weight; and the volume of interactions (WO). Hydrophobic regions (DDRY) and integy moments from the hydrophobic probe (IDDRY) are inversely related to activity, where the integy moment is a vector that measures the imbalance between the center of mass and the position of the hydrophobic regions around it. Chemically speaking, high hydrophobic integy moments mean that strong hydrophobic regions are concentrated in a few areas of the molecular surface. Summing up, it can be deduced that the toxicity for AhR binding increases with the hydrophilicity and therefore with the delocalization of the hydrophobicity.

We obtained a HQSAR model, based only on the 2D structures, comparable in predictive ability to that derived from CoMFA studies (q^2 LOO = 0.62). In HQSAR it is possible to visualize the individual contribution to activity of each atom in

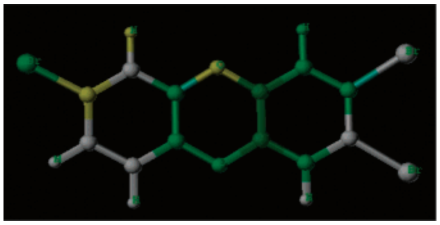


Figure 2. Individual atomic contributions for the most toxic compound of the training set (**23**): the colors at the red end of the spectrum (red, red-orange, and orange) reflect poor contributions, at the green end (yellow, green-blue, and green) favorable contributions, and atoms with intermediate contributions are white.

Table 9. External Test Set Prediction Results

compd	pIC_{50}	CoMFA	VolSurf	HQSAR	hybrid
6	6.97	7.95	8.43	7.61	7.98
7	6.81	7.586	7.63	6.93	7.27
15	10.09	9.29	10.19	10.39	10.7
16	10.09	9.121	9.76	9.57	10.05
17	10.69	9.065	9.89	9.03	9.48
18	9.07	9.045	9.15	8.86	9.3
22	9.35	8.326	10.18	10.10	10.6
92	7.32	7.533	5.91	7.05	6.2
93	6.86	7.577	6.8	7.33	7.14

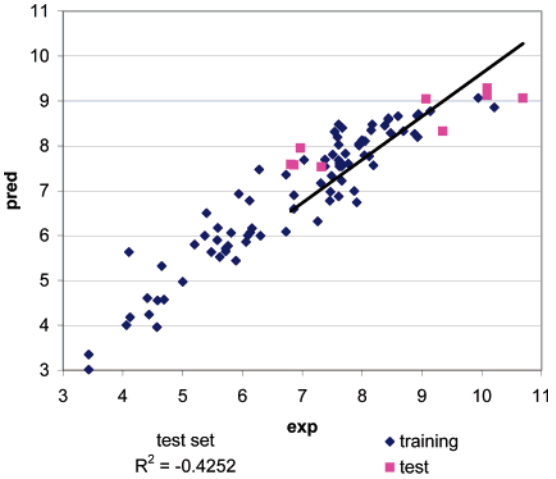


Figure 3. CoMFA plot of the experimental versus predicted pIC_{50} for the training and test set.

a given molecule of the data set by generating contribution maps. The HQSAR module implemented in Sybyl 7.1 uses the following color code to distinguish the main atomic contributions to activity: the colors at the red end of the spectrum (red, red-orange, and orange) reflect poor contributions and colors at the green end (yellow, green-blue, and green) reflect favorable contributions, while atoms with intermediate contributions are white. Figure 2 shows the individual atomic contributions for the most toxic compound of the training set (**23**). One fragment of this molecular structure, the dibenzo-*p*-dioxin moiety, seemed strongly related to the toxicity of this compound. The substituents R_4, R_6, and R_7 are yellow or green, indicating their positive contributions to the activity, while the other substituents are white, as they are invariant in the training set. Regions with intermediate or poor contribution in all molecules can be identified as potentially responsible in toxicity.

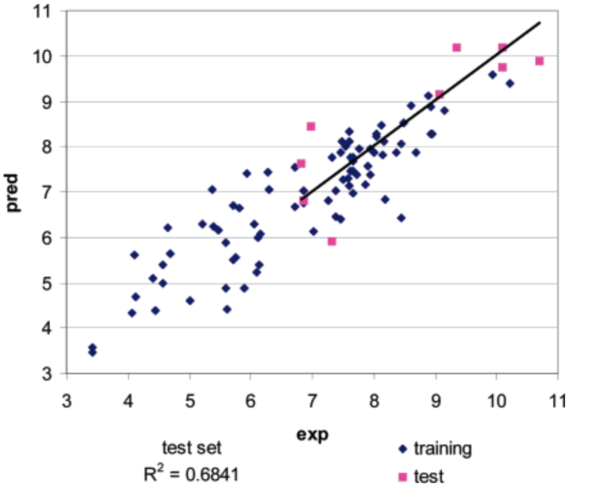


Figure 4. VolSurf plot of the experimental versus predicted pIC_{50} for the training and test set.

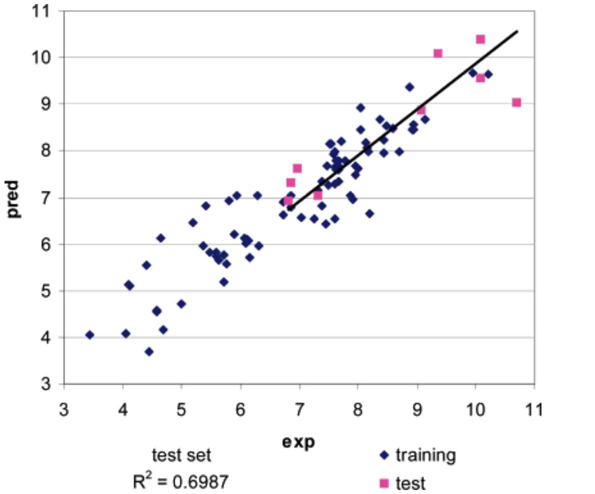


Figure 5. HQSAR plot of the experimental versus predicted pIC_{50} for the training and test set.

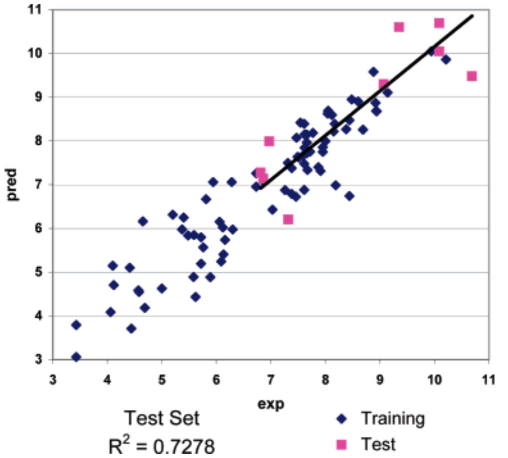


Figure 6. Hybrid model plot of the experimental versus predicted pIC_{50} for the training and test set.

The predictive power of the models was tested using an external set of nine compounds. The predicted values for the test set compounds are given in Table 9, while Figures 3−6 show the plot of experimental versus predicted of each model for the test set. We considered that these test set compounds were able to cover half of the activity on the whole model, the most active half. In our case, indeed, this is the most important part to analyze, because the aim of these models is to find fast and reliable methods to look for potential contaminants that can be present in food, and therefore, we want to be sure that they are able to predict and highlight the most toxic and dangerous compounds more than nontoxic ones. Moreover, the R^2 related to the external test set reported in this paper has been calculated by forcing the regression line to pass through zero; this is an extrapolation equivalent to adding a compound with activity equal to zero to the test set. It will decrease and underestimate our R^2_{test} results, but it can give a better idea of the external predictivity of our models. HQSAR was the fastest and most predictive model and can easily be used for prioritizing a list of potentially toxic compounds.

Discussion

Three QSAR methods (CoMFA, VolSurf, and HQSAR) were evaluated for predicting AhR binding. The QSAR models differ in how well the data fit the model. Since the regression coefficient (R^2) is greater for CoMFA analysis (0.91) than for VolSurf (0.79) and HQSAR (0.85), it can be inferred that the CoMFA model is the best. However, the cross-validation coefficient (q^2) for HQSAR is the same as the CoMFA one (0.62). The stability test (q^2 CV using 10 groups) showed that the exclusion of groups of compounds does not substantially change the results. Finally, the greater predictive ability of HQSAR was confirmed by the external test set prediction. Moreover, we also provided additional evidence that there is no correlation between the values of q^2 for the training set and the accuracy of prediction (R^2) for the test set.[24] Certainly in the CoMFA analyses the R^2_{test} value does not show very good values in spite of presenting high LOO q^2. This evidence very well underlines the contradiction between the q^2 and R^2_{test} results and guarantees that the HQSAR and VolSurf models will work for truly external compounds belong to the applicability domain defined.

An important role of a HQSAR model, besides predicting the activities of untested molecules, is to provide hints about what molecular fragments are directly related to biological activity. This information, combined with the CoMFA map, can give useful information about the toxicity of compounds not already tested. Figures 1 and 2 show the results using color codes for the contributions of CoMFA and HQSAR: they are in good agreement. Indeed R_4 and R_7 substituents make a positive contribution for both maps, R_4 with an electrostatic contribution and R_7 a steric one. The prevailing electrostatic contributions to the CoMFA model could be explained by the electrostatic nature of interactions and fit well with the results given by the VolSurf analysis that shows an increase of the toxicity with the hydrophobicity. Therefore, from a qualitative point of view the value of the CoMFA, VolSurf, and HQSAR analysis lies in the interpretation of the contributions of different descriptors on model responses, thus helping to understand the mechanism of action.

A rule-based model is an improvement in the final performance. Although it makes mandatory the calculation of activity values for the selected models, the hybrid model derived from HQSAR and Volsurf seems to be fairly robust and fast for consideration as a valid alternative in the virtual screening approach; mainly considering that the CoMFA approach is a rather complex one. Additionally, such methodology combines results from approaches with very different backgrounds, hence taking advantage of their qualities.

Other QSAR models published on the Ah receptor were considered[25−28] and thoroughly compared to the ones obtained in this paper. The major advantage found in our models is their usability, reproducibility, and quickness, facts that are of the greatest importance for virtual screening purposes. In fact, with equal performances, or an even slightly lower one, the easier model should be preferred.

Conclusion

This study developed a virtual screening method for fast prioritization of AhR binding predictions. Because of the relatively low sequence homology (∼25%) between the target AhR and the experimental template HIF-2a and subsequent difficulty in docking known high-affinity AhR ligands to this structure, virtual screening based on this homology model was not likely to be very accurate. As an alternative, 3D-QSAR models (using CoMFA, VolSurf, and HQSAR) were used based on the extensive literature of experimentally determined binding affinities of dioxin and other structural classes of AhR ligands.

The CoMFA model developed in this work was constructed by considering the issues of usability, reproducibility, and quickness, particularly taking into account the actual weakness related with the alignment, essential in the CoMFA approach. Actually, we have looked for an automatic and fast alignment rule that does not need any manual operation dealing with the individual structures. The SEAL alignment used is completely automatic and fast, making the CoMFA model very suitable for a large library of compounds to screen. Moreover, we tried to simplify our models in two different ways, the first one was going over the alignment rules using a computer program that does not need superimposition (VolSurf) and the second one was going over the optimization of the structure (HQSAR).

Both the HQSAR and the hybrid rule-based approaches resulted in mathematical models with greater predictive ability (q^2) than the already described models. Both are fast and highly predictive QSAR techniques that very rapidly generate models, making it applicable for screening a large amount of data. Moreover, the interpretations of the different families of descriptors show good agreement.

The rigorous procedure adopted to test the methods ensures their applicability and reliability in predicting the binding affinity of not yet tested chemicals. Therefore, these methodologies can be used to select and limit the compounds for testing and will boost the chances of finding AhR ligands in food, saving time and money and focusing work on the most promising chemicals that act through this receptor.

Acknowledgment. The research described was part of the EC project CASCADE (*C*hemic*a*l*s* as *C*ont*a*min*a*nts in the Foo*d* Chain: A Network of Excellence for Research, Risk Assessment and Education). The project was carried out with financial support from the Commission of the European Communities under the Work program Food Quality and Safety. The authors are most grateful to Molecular Discovery, who generously enabled us to use VolSurf software for the purpose of our study.

References

(1) Gronemeyer H., G. J., Laudet V. Principles for modulation of the nuclear receptor superfamily (review). *Nature* **2004**, *3*, 950.

(2) Pocar, P.; Fischer, B.; Klonisch, T.; Hombach-Klonisch, S. Molecular interactions of the aryl hydrocarbon receptor and its biological and toxicological relevance for reproduction 10.1530/rep.1.00294. *Reproduction* **2005**, *129*, 379−389.

(3) Pollenz, R. S.; Sattler, C. A.; Poland, A. The aryl hydrocarbon receptor and aryl hydrocarbon receptor nuclear translocator protein show distinct subcellular localizations in Hepa 1c1c7 cells by immunofluorescence microscopy. *Mol. Pharmacol.* **1994**, *45*, 428−438.

(4) Wilson, C. L.; Safe, S. Mechanisms of ligand-induced aryl hydrocarbon receptor-mediated biochemical and toxic responses. *Toxicol. Pathol.* **1998**, *26*, 657−671.

(5) Fischer, B. Receptor-mediated effects of chlorinated hydrocarbons. *Andrologia* **2000**, *32*, 279−283.

(6) Stapleton, H. M.; Baker, J. E. Comparing polybrominated diphenyl ether and polychlorinated biphenyl bioaccumulation in a food web in Grand Traverse Bay, Lake Michigan. *Arch. Environ. Contam. Toxicol.* **2003**, *45*, 227−234.

(7) Kitchen, D. B.; Stahura, F. L.; Bajorath, J. Computational techniques for diversity analysis and compound classification. *Mini Rev. Med. Chem.* **2004**, *4*, 1029−1039.

(8) Stahura, F. L.; Bajorath, J. Virtual screening methods that complement HTS. *Comb. Chem. High Throughput Screen* **2004**, *7*, 259−269.

(9) Stahura, F. L.; Bajorath, J. Partitioning methods for the identification of active molecules. *Curr. Med. Chem.* **2003**, *10*, 707−715.

(10) Stahura, F. L.; Bajorath, J. New methodologies for ligand-based virtual screening. *Curr. Pharm. Des.* **2005**, *11*, 1189−1202.

(11) Erbel, P. J.; Card, P. B.; Karakuzu, O.; Bruick, R. K.; Gardner, K. H. Structural basis for PAS domain heterodimerization in the basic helix−loop−helix−PAS transcription factor hypoxia-inducible factor. *Proc. Natl. Acad. Sci. U S A* **2003**, *100*, 15504−15509.

(12) Waller, C. L.; McKinney, J. D. Three-dimensional quantitative structure−activity relationships of dioxins and dioxin-like compounds: Model validation and Ah receptor characterization. *Chem. Res. Toxicol.* **1995**, *8*, 847−858.

(13) Cruciani, G.; Pastor, M.; Guba, W. VolSurf: A new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* **2000**, *11*, S29−S39.

(14) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Comput. Methodol.* **1990**, *3*, 537−547.

(15) Mohamadi, F. R., N. G. J.; Guida, W. C.; Liskamp, R.; Lipton, M.; Caulfield, C.; Chang, G.; Hendrickson, T.; Still, W.C Macromodel an integrated software system for modeling organic and bioorganic molecules using molecular mechanics. *J. Comput. Chem.* **1990**, *11*, 440−467.

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

(17) SYBYL (Version 6.9), Tripos, Inc., 1699 South Hanley Road, St. Louis, MO 63144.

(18) Basak, S. C.; Natarajan, R.; Mills, D.; Hawkins, D. M.; Kraker, J. J. Quantitative structure−activity relationship modeling of juvenile hormone mimetic compounds for Culex pipiens larvae, with a discussion of descriptor-thinning methods. *J. Chem. Inf. Model* **2006**, *46*, 65−77.

(19) Cruciani, G., Ed.; R. M. S. E., Kubinyi, H., Series Ed.; Folkers, G., Series Ed.; *Molecular Interaction Fields: Applications in Drug Discovery and ADME Prediction*; ISBN: 3-527-31087-8, 2005.

(20) Cruciani, G.; Crivori, P.; Carrupt, P.-A.; Testa, B. Molecular fields in quantitative structure-permeation relationships: The VolSurf approach. *J. Mol. Struct.: THEOCHEM* **2000**, *503*, 17−30.

(21) ComGenex, Budapest, Hungary.

(22) Geladi, P. K. B. Partial least squares regression: A tutorial. *Anal. Chem. Acta* **1986**, *35*, 1−17.

(23) Höskuldsson, A. PLS regression methods. *J. Chemom.* **1988**, *2*, 211−228.

(24) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graph. Model.* **2002**, *20*, 269−276.

(25) Vedani, A.; Dobler, M.; Lill, M. A. In silico prediction of harmful effects triggered by drugs and chemicals. *Toxicology and Applied Pharmacology, Living in a Safe Chemical World. Proceedings of the 10th International Congress of Toxicology, 11−15 July, 2004, Tampere, Finland* **2005**, *207*, 398−407.

(26) Lill, M. A.; Dobler, M.; Vedani, A. In silico prediction of receptor-mediated environmental toxic phenomena-application to endocrine disruption. *SAR QSAR Environ. Res.* **2005**, *16*, 149−169.

(27) So, S. S.; Karplus, M. Three-dimensional quantitative structure−activity relationships from molecular similarity matrices and genetic neural networks. 2. Applications. *J. Med. Chem.* **1997**, *40*, 4360−4371.

(28) Lukacova, V.; Balaz, S. Multimode ligand binding in receptor site modeling: Implementation in CoMFA. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2093-2105.

JM060526F