Binary Quantitative Structure−Activity Relationship (QSAR) Analysis of Estrogen Receptor Ligands†

Hua Gao,*,‡ Chris Williams,§ Paul Labute,§ and Jürgen Bajorath‡,¶

Computational Chemistry and Informatics, MDS Panlabs, 11804 North Creek Parkway South, Bothell, Washington 98011, Chemical Computing Group Inc., 1255 University Street, Suite 1600, Montreal, Quebec, Canada H3B 3X3, and Department of Biological Structure, University of Washington, Seattle, Washington 98195

Received August 15, 1998

The use of high throughput screening (HTS) to identify lead compounds has greatly challenged conventional quantitative structure−activity relationship (QSAR) techniques that typically correlate structural variations in similar compounds with continuous changes in biological activity. A new QSAR-like methodology that can correlate less quantitative assay data (i.e., "active" versus "inactive"), as initially generated by HTS, has been introduced. In the present study, we have, for the first time, applied this approach to a drug discovery problem; that is, the study of estrogen receptor ligands. The binding affinities of 463 estrogen analogues were transformed into a binary data format, and a predictive binary QSAR model was derived using 410 estrogen analogues as a training set. The model was applied to predict the activity of 53 estrogen analogues not included in the training set. An overall accuracy of 94% was obtained.

INTRODUCTION

Quantitative structure−activity relationship (QSAR)[1] analysis has been widely used to modify lead compounds, however identified, to optimize their biological activity, selectivity, and pharmacokinetic properties and to minimize the toxic effects.[2] It is often applied as part of the drug discovery process to better understand the interaction mechanisms between chemical compounds and biological targets.[3−5] A variety of molecular descriptors and statistical techniques have been developed to achieve better correlation between chemical structures and biological activities.[6−12] The fundamental hypothesis of QSAR is that biological activity is a function of molecular structure. Thus, molecules with similar structures exert similar biological activities, and changes in structure are thought to modulate biological activities.

In recent years, the use of high throughput screening (HTS) to identify lead compounds has greatly challenged commonly used QSAR techniques. HTS usually generates a large amount of assay data, which initially classifies compounds as active or inactive. In addition, compounds in screening libraries are typically noncongeneric (i.e., they do not share similar core structures). This attribute makes it difficult, if not impossible, to analyze HTS data by classical QSAR techniques and to predict active compounds. However, HTS data provide a substantial knowledge-base that correlates chemical structures with biological activities in a semiquantitative manner (active or inactive). Thus, if it would be possible to identify structural characteristics that render compounds "active", such compounds could be identified in virtual screening libraries and compound selection for testing could be rationalized (as an alternative to random testing).

To this end, a new QSAR-like methodology, termed "binary QSAR", has recently been introduced[13] and implemented.[14] Binary QSAR correlates compound structures, using molecular descriptors, with a "binary" expression of activity (i.e., 1 = active and 0 = inactive) and calculates a probability distribution for active and inactive compounds in a training set. This function can then be used to predict active compounds for a given target in a test set. This methodology is applied here, for the first time, to a drug discovery problem; that is, the analysis of estrogen receptor ligands.

The estrogen receptor is an extensively studied pharmaceutical target for which a large number of ligand analogues has been generated and characterized.[15−19] In addition, structural studies have elucidated the mechanism of the estrogen receptor−ligand interaction and identified the binding determinants. We have transformed estrogen receptor binding affinity data of estrogen analogues into a binary data format, derived a predictive binary QSAR model, and applied this model to a test set of other estrogen analogues. Both active and inactive analogues were predicted with high accuracy. The binary QSAR model was stable for a variety of binary activity cutoff values and quite insensitive to boundary effects.

METHODS

Binary QSAR. The binary QSAR analysis procedure used in this study is depicted in Figure 1. Binary QSAR estimates, from a training set, the probability density $\Pr(Y = 1/X = x)$ where Y is a Bernoulli random variable (i.e., Y takes on

* Corresponding author. Telephone: (425)487-8200 ext. 390. Fax: (425)-487-3787. E-mail: hgao@panlabs.com.

† This manuscript is dedicated to Professor Corwin Hansch on the occasion of his 80th birthday.

‡ Computational Chemistry and Informatics.

§ Chemical Computing Group Inc.

¶ Department of Biological Structure.

10.1021/ci980140g CCC: $18.00
Published on Web 12/16/1998

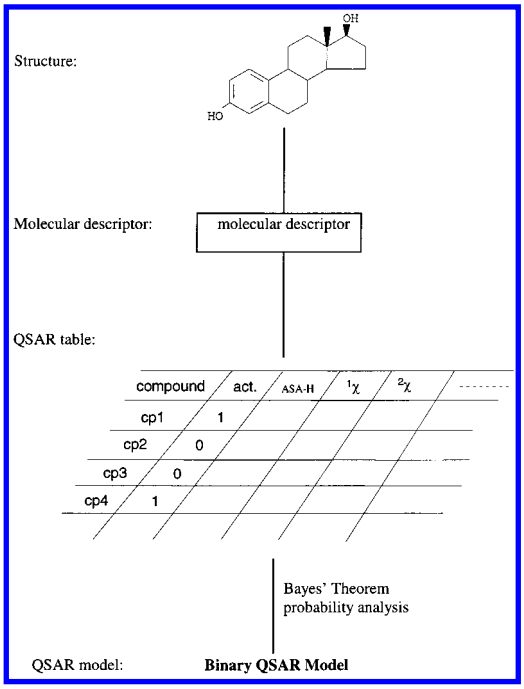


Figure 1. The flow chart of binary QSAR analysis in MOE.

values of 0 or 1) representing "active" or "inactive" and X is a random n-vector of real numbers (a random collection of molecular descriptors).

A principal components analysis (PCA) is conducted on the training set to calculate an n by p linear transform, Q, and an n-vector, u, such that the random p-vector $Z = Q(X - u)$ has mean and variance equal to the p by p identity matrix. The quantity p is referred to as the number of *principal components*.

The original molecular descriptors are transformed by Q and u to obtain a decorrelated and normalized set of descriptors. The desired probability density is then approximated by applying Bayes' theorem and assuming that the transformed descriptors are mutually independent:

$$\Pr(Y = 1|X = x) \approx \left[1 + \frac{\Pr(Y = 0)}{\Pr(Y = 1)} \prod_{i=1}^{p} \frac{\Pr(Z_i = z_i|Y = 0)}{\Pr(Z_i = z_i|Y = 1)}\right]^{-1}$$

$$Z = Q(\mathrm{X} - u) = (Z_1,...,Z_p)$$

Each probability density $\Pr(Z_i = z_i)$ is estimated by constructing a histogram. Conventional procedures for histogram construction are sensitive to bin boundaries because every observation, no matter how close to a bin boundary, is treated as though it fell in the center of the bin. To reduce this sensitivity, each observation is replaced with a Gaussian density with variance σ^2. This variance can be interpreted as an observation error or as a *smoothing parameter*.

Once all of the $2p + 2$ probability densities have been estimated from the training set, the desired density $\Pr(Y = 1/X = x)$ is constructed using the formula just presented. A thorough description of the specifics of the binary QSAR methodology was reported in ref 13.

Biological Data. The binding data of estrogen analogues to estrogen receptors of different species was collected from the literature.[15−20] There is little, if any, evidence for receptor−species difference in estrogen analogue structure−affinity relationships.[17,21] There are two subtypes of estrogen receptors, ER-α and ER-β. The data reported in this paper are presumed to come from ER-α because this subtype is the predominant one in uterine and breast tissue. The binding data was placed on a common "relative binding affinity" (RBA) scale. Values on this scale were calculated as a percentage of the ratio of IC_{50} values of test compounds to displace 50% of [^{3}H]estradiol from estrogen receptor binding. Thus, on the RBA scale, estradiol has a value of 100, with lower affinity analogues having lower values and higher affinity analogues higher RBA values. A total of 463 compounds were selected (tested for binding at 0 to 4 °C), 410 of which were used as a training set to derive a binary QSAR model, and 53 of which were used as a test set to evaluate the model by predicting active and inactive compounds. Table 1 shows the composition of estrogen analogues used in this analysis. The continuous biological activity data was expressed in binary form using a threshold criterion (log RBA). Any compounds with log RBA larger than or equal to this criterion were classified as active, and any compounds with lower log RBA values were classified as inactive.

Table 1. Composition of Estrogen Receptor Ligands

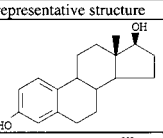
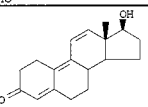
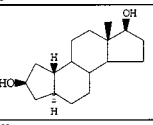
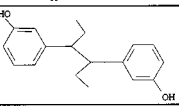
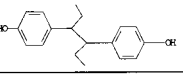
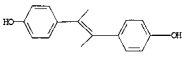
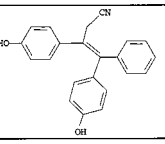
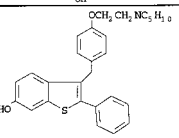
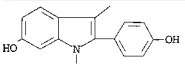
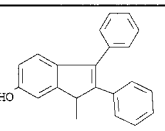
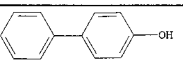
Category	representative structure	number of compounds
Estradiol derivatives		165
3-keto steroids		2
nonaromatic analogs		4
metahexstrol derivatives		15
hexestrol derivatives		50
diethylstilbestrol derivatives		10
tryphenylethylene analogs		40
2-phenylbezothiophene analogs		68
2-phenylindole analogs		61
indene analogs		45
Phenol and biphenols		3

Table 2. Data Profiles at Different Binary Threshold Values

threshold value (log RBA)	active compounds	inactive compounds	active%
−2.0	404	6	98%
−1.5	394	16	96%
−1.0	382	28	93%
0.0	307	103	75%
1.0	177	233	43%
1.2	146	264	36%
1.5	92	318	22%
1.7	62	348	15%
1.8	53	357	13%
2.0	27	383	7%

Different activity threshold values were used to alter the percentage of active compounds in the training set.

Binary QSAR Analysis. Molecular descriptors were calculated using 1998.03 version of MOE,[14] and binary QSAR analysis was carried out with the MOE binary QSAR function.

Performance of a binary QSAR model was measured as follows: let m_o represent the number of active compounds, m_1 the number of inactive compounds, c_o the number of active compounds correctly labeled by the QSAR model, and c_1 the number of inactive compounds correctly labeled by the QSAR model. Three parameters of performance were calculated: (1) accuracy on active compounds, c_o/m_o; (2) accuracy on inactive compounds, c_1/m_1; and (3) overall accuracy on all of the compounds, $(c_o + c_1)/(m_o + m_1)$. The derived binary QSAR model was cross-validated by the leave-one-out procedure.[22] In this procedure, only one object is eliminated at a time and the process is repeated until all objects have been eliminated once and only once. Accuracy was calculated for each step, and an average accuracy for all the steps was reported as a measure of the internal predictivity of the model within the training set.

RESULTS AND DISCUSSION

A set of 410 compounds was chosen to be a training set to derive a binary QSAR model. The range of the biological activities (log RBA) was −2.02 to 2.60. Table 2 shows the data profiles with different threshold values.

A value of 1.7 of log RBA, which corresponds to 50% of RBA, was selected as the threshold to derive the binary QSAR model. Based on this threshold criterion, 62 compounds were active and 348 compounds were inactive in the training set. A smoothing factor was introduced to minimize the sensitivity of the derived model to the selection of bin boundaries, as mentioned earlier. The binary QSAR model is also influenced by the number of principal components used. A 5 × 7 factor analysis was carried out to determine the effects of different smoothing factor values and principal component numbers on the binary QSAR analysis of the data set analyzed. Table 3 summarizes the results of the analysis.

In this study, we have used two-dimensional (2D) molecular descriptors, which were shown to perform well in compound clustering.[23,24] In addition, Keir's shape indices were used, which contain implicit three-dimensional (3D) information. Explicit 3D descriptors were not considered to avoid bias of the analysis due to predicted conformational effects. We have systematically explored different combination of molecular descriptors to identify a set that captures structural characteristics of estrogen analogues and resulting activities well. This was done for the learning set similar to more conventional QSAR analysis. Table 3 shows that an optimal binary QSAR model was obtained by a combination of principal component numbers of 12 and a smoothing factor value of 0.12. Using this combination, the non-cross-validated accuracy is 85% on active compounds, 93% on inactive compounds, and 92% for all the compounds. The cross-validated accuracy is 76% on active compounds, 93% on inactive compounds, and 90% for all the compounds. Any departure from these parameter values decreased the non-cross-validated and/or cross-validated accuracy. Thirteen molecular descriptors were used to derive the binary QSAR model (Table 4) including four atomic connectivity indices, four molecular shape indices, one total hydrophobic accessible surface area descriptor, one charge descriptor, one aromatic bond descriptor, and two indicator variables for specific functional group and molecular structure. One of the descriptor used is I,es. A number of desthylstilbestrol (DES) analogues are found to be more potent estrogen receptor ligands than estradiol itself, despite their structure similarity (log RBA is 2.48 for DES versus 2.00 for estradiol). Because structure features that account for higher potency of DES analogues were not obvious, we have specifically included the indicator variable I,es to account for this effect. A phenolic OH group that resembles the 3-OH of estradiol molecule is required for tight binding to estrogen receptor. To account for this specific structural effect, an indicator variable, I,OH, was used.

Table 3. Effects of PCA Number and Smoothing Factor on Binary QSAR

PCA no.	smoothing factor 0.08	0.10	0.12	0.14	0.16	0.20	0.25
6	0.79	0.76	0.74	0.69	0.69	0.60	0.52
	0.63	0.61	0.60	0.60	0.55	0.48	0.45
8	0.81	0.77	0.77	0.76	0.76	0.73	0.66
	0.71	0.71	0.71	0.69	0.68	0.65	0.55
10	0.85	0.84	0.84	0.84	0.81	0.77	0.73
	0.68	0.68	0.66	0.68	0.69	0.68	0.65
12	0.85	0.85	***0.85***	0.82	0.81	0.81	0.79
	0.69	0.71	***0.76***	0.76	0.73	0.73	0.71
13	0.85	0.85	0.82	0.82	0.82	0.82	0.79
	0.63	0.66	0.69	0.69	0.69	0.66	0.68

Table 4. Molecular Descriptors Used in the Binary QSAR

symbol	description
b-ar	number of aromatic bonds[14]
ASA-H	total hydrophobic accessible surface area[14]
$^0\chi$	zero-order atomic connectivity index[7]
$^0\chi^v$	zero-order atomic valence connectivity index[7]
$^1\chi$	first-order atomic connectivity index[7]
$^1\chi^v$	first-order atomic valence connectivity index[7]
$^1\kappa$	Keir first shape index[8]
$^2\kappa$	Keir second shape index[8]
$^3\kappa$	Keir third shape index[8]
Φ	Keir molecular flexibility index[9]
Peoe-PC$^+$	total of positive charge in Gasteiger & Marsili charge model[14]
I,OH	indicator variable for phenolic hydroxy group; I,OH = 1 for compounds containing phenolic OH and 0 for other compounds
I,es	indicator variable for hexestrol derivatives; I,es = 1 for hexestrol compounds and 0 for other compounds.

Effects of Binary Threshold. We analyzed the effects of 10 different threshold values (log RBA values ranges from −2 to 2) on the binary QSAR model (Figure 2). Accuracy

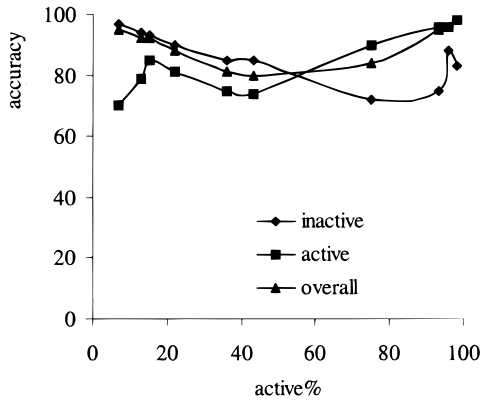


Figure 2. Accuracy versus active% compounds.

on active compounds ranges from 70 to 98%, with highest accuracy obtained for 98% active compounds and lowest for 7% active compounds. The overall accuracy remains stable at different threshold values (around 90%). Figure 2 shows that selected threshold values cause fluctuation of observed overall accuracy by approximately 10%. The minimum obtained overall accuracy is about 80%. Thus, on the basis of these findings, we conclude that the overall binary QSAR accuracy remains stable irrespective of the chosen threshold values.

Boundary Effects. Compounds with biological activity near the binary threshold value may fall into either the active or inactive category, which also depends on the experimental error. To analyze the influence of boundary effects on the binary QSAR model, compounds with log RBA values between 1.0 and 1.7 were omitted. Therefore, in these calculations, binary classification corresponds to largest difference in biological activities. This data set consisted of 292 inactive and 62 active (17.5%) compounds. In the resulting QSAR model, an accuracy of 87% on active, 95% on inactive, and 93% for all 354 compounds was achieved. The performance is only slightly better than that obtained for the original training set. These results indicate that boundary effects tested here have only marginal influence on the binary QSAR accuracy and suggest that the binary QSAR model is stable. Thus, obtained accuracy is not critically dependent on the binary classification of the observed activities with the data set analyzed, which is important with respect to the analysis of screening data.

Validation of the QSAR Model. To evaluate the predictive value of the binary QSAR model, 53 randomly selected estrogen analogues were tested. Seven out of 9 active compounds (78%) and 43 out of 44 inactive compounds (98%) were correctly predicted (overall accuracy of 94%), which is consistent with the cross-validation result. The percentage of active compounds in the test set was 15%. Thus, if the compounds were selected and tested based on the binary QSAR model, the "hit rate" of active compounds would be 5-fold higher than randomly selected compounds even for this small data set. This result illustrates the potential of binary QSAR for selection of active compounds based on initial (partial) screening data.

Correlation with X-ray Structural Data. The X-ray structures of the ligand binding domain of ER-α receptor in complex with estradiol and raloxifene have been reported.[25] The ligand is buried within the hydrophobic core of the ligand binding domain[26,27] but the polar ends of estradiol form hydrogen bonds to the only polar amino acid residues in the binding site. Glu353 forms a hydrogen bond to the A-ring phenolic hydroxyl group and His524 forms a hydrogen bond with 17β-hydroxyl group.[28] What are the critical components of the receptor−ligand interaction? The phenolic hydroxyl group is required for binding.[28,29] The 3-OH group on estradiol can act as a hydrogen bond donor or acceptor, but the hydrogen bond donor ability is more important than the acceptor ability in stabilizing the complex. 3-Keto and 3-methyl ether derivatives have much lower binding affinities because they lack a hydrogen bond donor.[17] The aromatic ring system is required for strong binding because analogues lacking aromatic moieties have only low binding affinity.[30] It follows that structural differences between active and inactive compounds are distinct but may be somewhat limited. We, therefore, consider estrogen analogues to be a challenging test case for binary QSAR analysis because of small structural modifications, which actually change binary activity in a more continuous way, are considered here to render compounds either active or inactive.

Comparison with Conventional QSAR. Estrogen analogues have also been studied by conventional QSAR techniques.[15] What are the distinguishing features of these approaches? Binary QSAR assigns a probability to a compound to be active in a particular test setting, but cannot predict specific modifications of lead compounds to improve their activity. Thus, binary QSAR is not an alternative to conventional QSAR analysis. In a drug discovery setting, these approaches should be complementary. After binary QSAR-guided selection of active compounds, conventional QSAR can be used to optimize their biological activity.

Earlier QSAR studies[15] on estrogen analogues did not reveal a consistent positive hydrophobic contribution for receptor−ligand binding, except substituents at the 11-β position of estradiol derivatives, although hydrophobicity expressed as log P(o/w) differs significantly among the analogues. Similarly, in the binary QSAR model, log P(o/w) was not found to be a significant descriptor. In contrast, ASA-H (which does not strictly correlate with log P(o/w) ($r^2 = 0.62$)) was found to be significant. This finding suggests that the strength of van der Waals/hydrophobic interactions between ligands and receptor is more important than the differences in energy required to desolvate the hydrophobic ligands.

Comparison with other Classification/Clustering Approaches. Conventional QSAR based on regression techniques, such as multiple linear regression, partial least squares and, occasionally, neural networks, have been used to cluster compounds.[31−33] These methods seek to minimize the squared error between the model and the observed data. This optimization of the model parameters introduces sensitivity to errors in experiments and regression analyses. In contrast, binary QSAR does not use any form of regression analysis; that is, there is no attempt to minimize the model errors with regard to the model parameters. It is a nonlinear modeling method. Because no regression is used, the model estimation procedure is very fast, which is in contrast to neural networks that require a lengthy training phase. Therefore, binary QSAR can efficiently process large data sets such as HTS data.

Several other clustering methods have been tested to classify compounds into different clusters.[23,24,34] These methods are qualitative in that they are based on only

chemical structural information regardless of biological activities. Thus, compounds with similar structural features are clustered together. However, compounds with similar biological activities may appear in different clusters depending on their degree of structural similarity. In this case, identification of active clusters may be a nontrivial task. In contrast, binary QSAR takes both structure and activity information into account, and deduces a probability distribution function for novel compound to be either active or inactive.

In conclusion, we have shown that the binary QSAR approach yields promising results when applied to a data set that resembles a screening experiment. The utility of this approach can be tested by future acquisition of HTS data.

ACKNOWLEDGMENT

We thank Prof. Corwin Hansch of Pomona College and Prof. Eric Lien of University of Southern California for their detailed discussion and invaluable suggestions, and Jeff Godden and Ling Xue for help in this study.

REFERENCES AND NOTES

(1) Hansch, C.; Maloney, P. P.; Fujita, T.; Muir, R. M. Correlation of biological activity of phenoxyacetic acids with hammett substituent constant and partition coefficients. *Nature* **1962**, *194*, 178−280.

(2) Plummer, E. L. Successful application of the QSAR paradigm in discovery programs. In *Classical and three-dimensional QSAR in agrochemistry*; Hansch, C.; Fujita, T., Ed.; American Chemical Society: Washington D. C., 1995.

(3) Gao, H.; Hansch, C. QSAR of p450 oxidation: On the value of comparing kcat and Km with kcat/Km. *Drug. Metab. Rev.* **1996**, *28*, 513−526.

(4) Lien, E. J.; Gao, H. QSAR analysis of skin permeability of various drugs in man as compared in vivo and in vitro studies in rodents. *Pharm. Res.* **1995**, *12*, 583−587.

(5) Hansch, C.; Hoekaman, D.; Gao, H. Comparative QSAR: Toward a deeper understanding of chemicobiological interaction. *Chem. Rev.* **1996**, *96*, 1045−1075.

(6) Kubinyi, H. *QSAR: Hansch analysis and related approaches*; VCH: Weinheim, 1993.

(7) Kier, L. B.; Hall, L. H. The nature of structure−activity relationships and their relative to molecular connectivity. *Eur. J. Med. Chem.* **1977**, *12*, 307−312.

(8) Kier, L. B. Indexes of molecular shape from chemical graphs. *Med. Res. Rev.* **1987**, *7*, 417−440.

(9) Kier, L. B. An index of flexibility from molecular shape descriptors. *Prog. Clin. Biol. Res.* **1989**, *291*, 105−109.

(10) Hopfinger, A. J.; Potenzone, R. Ames test and antitumor activity of 1-(X-phenyl)-3,3-dialkyltriazenes. quantitative structure−activity studies based upon molecular shape analysis. *Mol. Pharmacol.* **1982**, *21*, 187−195.

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

(12) Klein, C. D.; Hopfinger, A. J. Pharmacological activity and membrane interactions of antiarrhythmics: 4D-QSAR/QSPR analysis. *Pharm. Res.* **1998**, *15*, 303−311.

(13) Labute, P. *Binary QSAR: A new method for the determination of quantitative structure−activity relationships*. Pacific Symposium On Biocomputing. '99, accepted for publication.

(14) Chemical Computing Group Inc., MOE 1998.03, 1255 University Street, Montreal, Quebec, Canada, H3B 3 × 3.

(15) Gao, H.; Katzenellenbogen, J. A.; Garg, R.; Hansch, C. Comparative QSAR analysis of estrogen receptor ligands. *Chem. Rev.*, in press.

(16) Sadler, B. R.; Cho, S. J.; Ishaq, R. 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.

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

(18) Lien, E. J.; Das, A.; Lien, L. L. Immunopharmacological and biochemical bases of Chinese herbal medicine. *Prog. Drug Res.* **1996**, *46*, 264−280.

(19) Lien, L. L.; Lien E. J. Hormone therapy and phytoestrogens. *J. Clin. Pharm. Ther.* **1996**, *21*, 101−111.

(20) Hardcastle, I. R.; Rowlands, M. G.; Grimshaw, R. M.; Houghton, J.; Jarman, M. Homologues of indoxifene variation of estrogen receptor binding and calmodulin antagonism with chain length. *J. Med. Chem.* **1996**, *39*, 999−1004.

(21) Kuiper, G. G.; Enmark, E.; Pelto-Huikko, M.; Nilsson, S.; Gustafsson, J. A. Cloning of a novel receptor expressed in rat prostate and ovary. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 5925−5930.

(22) Cramer, R. D.; Bunce, J. D.; Patterson, D. E.; Frank, I. E. Crossvalidation, Bootstrapping, and Partial Least Squares Compared with multiple regression in conventional QSAR Studies. *Quant. Struct.−Act. Relat.* **1988**, *7*, 18−25.

(23) Ajay, W.; Walters, P.; Murcko, M. A. Can we learn to distinguish between 'Drug-like" and "nondrug-like" molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.

(24) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare structure-base clustering methods and descriptors for use in compounds selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

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

(26) Wurtz, J.-M.; Bourguet, W.; Renaud, J.-P.; Vivat, V.; Chambon, P.; Moras, D.; Gronemeyer, H. A canonical structure for the ligand-binding domain of nuclear receptors. *Nat. Struct. Biol.* **1996**, *3*, 206.

(27) Bourguet, W.; Ruff, M.; Chambon, P.; Gronemeyer, H.; Moras, D. Crystal structure of the ligand-binding domain of the human nuclear receptor RXR-alpha. *Nature* **1995**, *375*, 377−382.

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

(29) Wurtz, J.-M.; Egner, H.; Heinrich, N.; Moras, D.; Mueller-Fahrnow, A. Three-dimensional models of estrogen receptor ligand binding domain complexes, based on related crystal structures and mutational and structure−activity relationship data. *J. Med. Chem.* **1998**, *41*, 1803−1814.

(30) Raynaud, J. P.; Ojasoo, T.; Bouton, M. M.; Bignon, E.; Pons, M.; Crastes de Paulet, A. Structure−activity relationships of steroidal estrogens. In *Estrogens in the environment II*; McLachlan, J. A., Ed.; Elsevier: New York, 1985, pp 24−42.

(31) So, S.-S.; Karplus, M. Three-dimensional quantitative structure−activity relationship from molecular similarity matrixes and genetic neural networks. 1. Method and validations. *J. Med. Chem.* **1997**, *40*, 4347−4359.

(32) Hosseini, M.; Maddalena, D. J.; Spence, I. Using artificial neural networks to classify the activity of capsaicin and its analogues. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1129−1137.

(33) Shi, L. M.; Fan, Y.; Myers, T. G.; O'Connor, P. M.; Paul, K. D.; Friend, S. H.; Weinstein, J. N. Mining the NCI anticancer drug discovery databases: Genetic function approximation for the QSAR study of anticancer ellipticine analogues. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 189−199.

(34) McGregor, M. J.; Pallai, P. V. Clustering of large database of compounds: Using the MDL "Keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443−448.

CI980140G