Molecular Quantum Similarity Analysis of Estrogenic Activity

Ana Gallegos Saliner,*,† Lluís Amat,† Ramon Carbó-Dorca,† T. Wayne Schultz,‡ and Mark T. D. Cronin§

Institute of Computational Chemistry, University of Girona, Campus Montilivi, 17071 Girona, Spain, Department of Comparative Medicine, College of Veterinary Medicine, The University of Tennessee, 2407 River Drive, Knoxville, Tennessee 37996-4500, and School of Pharmacy and Chemistry, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, England

Received January 24, 2003

The main objective of this study was to evaluate the capability of 120 aromatic chemicals to bind to the human alpha estrogen receptor (hERα) by the use of quantum similarity methods. The experimental data were segregated into two categories, i.e., those compounds with and without estrogenicity activity (active and inactive). To identify potential ligands, semiquantitative structure−activity relationships were developed for the complete set correlating the presence or lack of binding affinity to the estrogen receptor with structural features of the molecules. The structure−activity relationships were based upon molecular similarity indices, which implicitly contain information related to changes in the electron distributions of the molecules, along with indicator variables, accounting for several structural features. In addition, the whole set was split into several chemical classes for modeling purposes. Models were validated by dividing the complete set into several training and test sets to allow for external predictions to be made.

INTRODUCTION

Endocrine disruptive (ED) chemicals[1] are present in the aquatic environment as pollutants and in food and water as antioxidants and metabolites of other anthropogenic chemicals.[2] Of the chemicals capable of endocrine disruption, environmental estrogens form a large group.[3] These may result either from naturally occurring compounds or synthetically produced chemicals and are derived from a variety of sources such as pesticides, plastics, combustion byproducts, and plants, i.e., phytoestrogens and agricultural products, among others. Such chemicals can bind to estrogen receptors (ER) and may thus interfere with genetic functions such as sexual development and reproductive fecundity.[4]

Due to the deleterious effects of endocrine disrupting chemicals on both human health and the environment,[5] the U.S. Environmental Protection Agency (EPA) has run a screening and testing program to identify such compounds. The test program takes into account the broad range of structurally diverse compounds with potential estrogenic activity. The Endocrine Disruptor Screening and Testing Advisory Committee (EDSTAC)[6] suggested quantitative structure−activity relationships (QSARs) to be a suitable tool to model ligand binding to the ER.[7−9] The appreciation of the relationship between chemical structure and estrogenic activity (i.e. structure−activity relationships (SARs)) dates back to the 1930s, when it was observed that nonsteroidal derivatives could elicit a biological response analogous to that of 17β-estradiol.[10] Since that time, the understanding of the structural features controlling the potency of estrogenic activity lead to the discovery of a wider pool of active chemicals. Subsequently SAR studies have been used extensively to predict ligand-hormone receptor binding affinities and, more specifically, to estimate the effects of estrogenic compounds for hazard identification, human health risk assessments,[11] and wildlife exposure studies.[12] These methods are effective in the prioritization of untested compounds for more intensive and costly empiric evaluations based on in vivo and in vitro bioassays.[13] Specially, as estrogenic activity is difficult to assess in vivo, it has been evaluated *in vitro* by a number of methods; included among them there are the well-known estrogen receptor competitive binding,[14] MCF-7 breast cell proliferation, and the transfected yeast assays.[15]

A number of QSAR and SAR methods are available to model the interactions between a ligand and a receptor. For instance, molecular similarity based methods applied to QSAR analysis offer the possibility to predict which compounds may act in the same receptor. To be successful, however, they require well-defined biological and toxicological data from chemicals representative of the diversity of structures for which predictions are to be made. In the literature, several 3D-QSAR similarity studies predicting hormone receptor binding affinity have been reported. The Comparative Molecular Field Analysis (CoMFA)[16,17] approach analyses steric and electrostatic interactions of aligned series of molecules and relates them to a biological activity. The Common Reactive Pattern (CoRePa)[18,19] approach defines the common reactive pattern across global and local reactivity descriptors associated with the specific biological endpoint. As such it takes into account the conformational flexibility of ligands and quantifies chemical similarities.

Other techniques to model ligand−receptor interactions are based on the assessment and quantification of molecular

* Corresponding author phone: + 34 972 41 83 67; fax: + 34 972 41 83 56; e-mail: annag@iqc.udg.es.
† University of Girona.
‡ The University of Tennessee.
§ Liverpool John Moores University.

10.1021/ci034014a CCC: $25.00 © 2003 American Chemical Society
Published on Web 06/27/2003

similarity. The effective application of any molecular similarity technique is founded on the underlying assumption that similar molecules possess similar properties.[20] Quantum similarity provides such a suitable quantification of the resemblance of molecular structures. It is a technique based on quantum-mechanical principles. Carbó et al.[21] demonstrated the formal connection between quantum similarity and QSAR, and, since then, this theory has been successfully applied to construct QSAR models, to both pharmacological[22−24] and toxicological[25,26] endpoints. In conjunction with QSAR, quantum similarity has been applied to several fields. Recent reviews of the underlying theory are available.[27−29] The development and evaluation of molecular similarities to predict a biological activity require that a knowledge base containing both training and evaluation tests of chemicals, whose modes of action and potency are well defined, be established. Given the plethora of high quality data currently available for endocrine disruption endpoints, and estrogenic activity in particular, the use of quantum similarity methods should be also applicable in this area.

In summary, several approaches to screen for endocrine disruption have described the main molecular features required for the pharmacophore that allows binding to the estrogenic receptor.[11,14,30,31] These 2D structural fragments assist in the comparison of the analoges with the structural features of the estradiol molecule, considered to be one of the most potent estrogens, and therefore provide the most detailed information regarding the pharmacophore. The general structural requirements relevant for a compound to bind to the estrogen receptor can be summarized as follows: 1. a hydroxy (phenolic) group at the C-3 position of the aromatic A-ring of a steroid, which is considered to be important for hydrogen bonding; 2. a ketone or alcohol functional group at site 17 of a steroid, which is able to accept a hydrogen bond; 3. four hydrophobic centers, corresponding to the A to D rings of estradiol, which form a relatively flat and rigid hydrocarbon core; 4. a hydrophobic group at the *para* position relative to the phenolic hydroxy group; and 5. the shape of the ligand must be constrained to fit the estrogen receptor pocket.

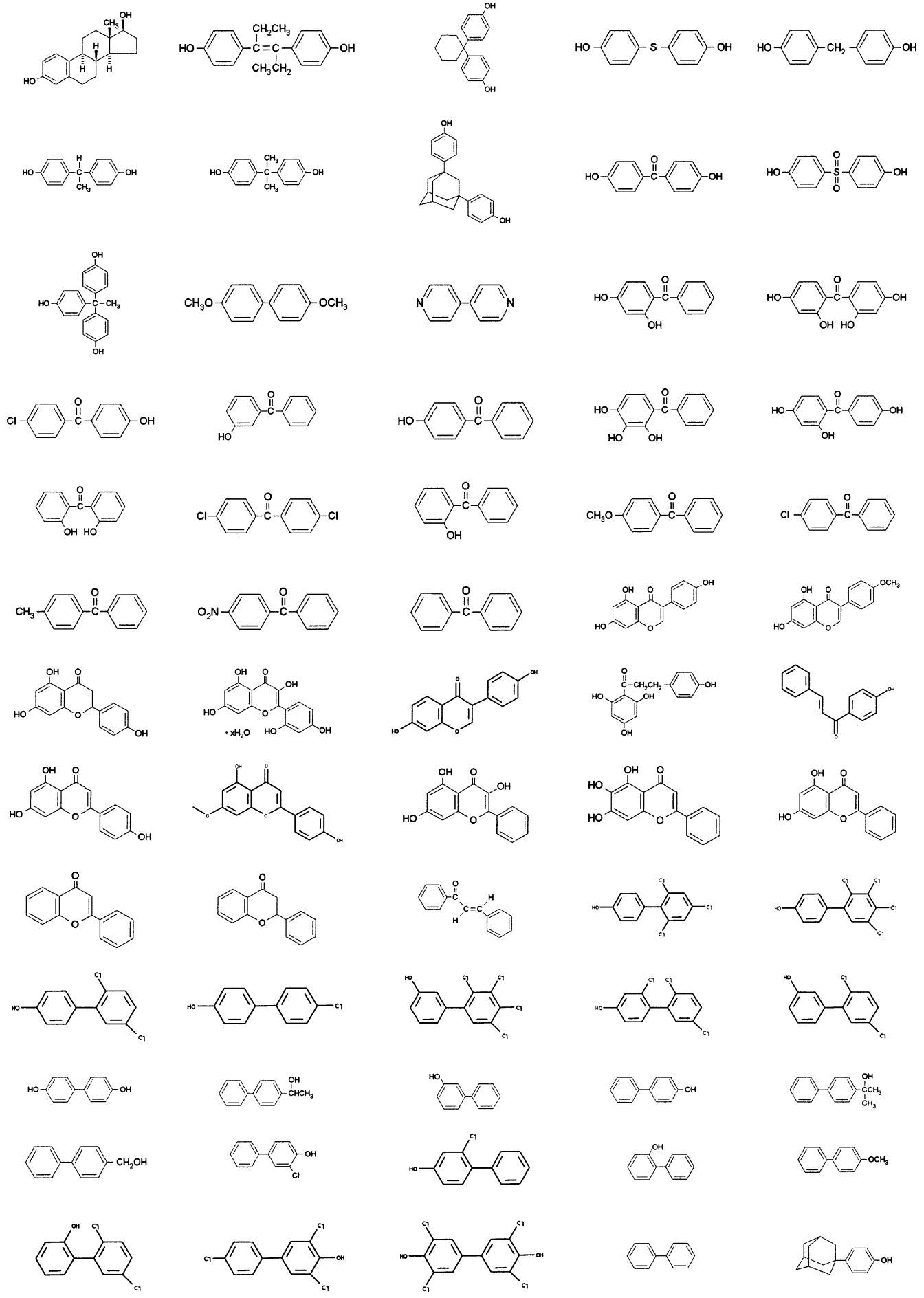
The aim of the current study was to develop a computer-based strategy for the prediction of estrogenic activity utilizing molecular quantum similarity methods. To achieve this, a high quality data set of in vitro estrogenicity data (gene expression in a yeast following binding to the human estrogen receptor) was analyzed.

MATERIALS AND METHODS

Data Set and Classification of Structures. The compounds studied were taken from the work of Schultz et al.,[32] who reported SARs for the estrogenic gene activation of an assorted group of 120 aromatic compounds. The structures are reported in Figure 1 and cover a wide range of chemical classes, which represent a broad degree of structural diversity and so allow for the comprehension of the structural requirements for estrogen receptor binding. For the development of SARs, Schultz et al.[32] divided aromatic compounds including bisphenols, benzophenones, flavonoids, biphenyls, phenols, and other aromatic and biaromatic chemicals, into five subsets according to the chemical structural characteristics and functional groups present in each molecule, as shown in Table 1. Therefore, five classes (of between 13 and 56 compounds) were analyzed created and used for analysis. However, the original classification of compounds is restricted as some compounds can be classified in more than a single class. Further, the size of the different classes was not uniform, ranging from about 10 to more than 50 compounds. Since such a distribution of data is not convenient for the comparison of results from the different subsets, and considering the hydrophobic region between the two aromatic rings of great importance to explain activity,[33] the previous classification criteria were replaced by criteria developed from a more fundamental structural basis to split the complete set into subsets. Hence, the compounds were classified into several categories, with approximately the same number of compounds, ranging from 23 to 37, on the basis of their fundamental structure. This classification was similar to that applied by Shi et al.[14] The different classes, presented in Table 2, were as follows: (1) chemicals with a single aromatic ring (e.g. alkylphenols, parabens, phenols, and miscellaneous); (2) compounds with two aromatic rings separated by a single bond (e.g. biphenyls and polychlorinated biphenyls); (3) compounds with two aromatic rings separated by a single bridging atom (e.g. diphenylmethanes, benzophenones, and bisphenol derivatives); and (4) compounds with two aromatic rings separated by two or three bridging atoms (e.g. DES and hexestrol derivatives, phytoestrogens, and flavonoids).

Biological Activity: Relative Binding Affinity. The recombinant yeast assay as reported by Routledge and Sumpter[34] and performed according to the protocol of Schultz et al.[35] was used in this study. The Glaxo Wellcome-derived *Saccharomyces cerevisiae* system is an in vitro assay used as a surrogate for estrogenicity with a 0.08 nM detection limit, 10 000-fold responsiveness, short duration, and low cost. The strain used in this assay expresses constitutively the gene for the α-human estrogen receptor, hERα. The responsive sequences control the expression of the L*ac*-Z reporter gene on a separate expression plasmid. Xenoestrogens interact with the bound receptor on the hybrid promoter to initiate transcription of the genes resulting in the production of β-galactosidase. Such an activity is measured colorimetrically when the yellow chromogen chlorophenol red-β-D-galactopyranoside is metabolized into phenol red, which is measured with a microplate reader as absorbance at 540 nm. The concentration eliciting an activity equal to 50% of the positive control 17β-estradiol was determined for each compound. These EC_{50} values were converted to molar units, and the relative gene expression was calculated by dividing the EC_{50} value of 17β-estradiol by the EC_{50} of the test chemical and multiplying by 100. Those compounds with a maximum gene expression at a concentration less than 50% of 17β-estradiol were noted as being detectable, but an EC_{50} could not be established. Test compounds that were found to be nonactive were evaluated twice with a minimum of three replicates.

Classification of Actives vs Inactives. For modeling purposes, experimental data were classified as active (A) or inactive (I), as reported previously.[24,26,36] The compounds considered and their activities are listed in Table 3.



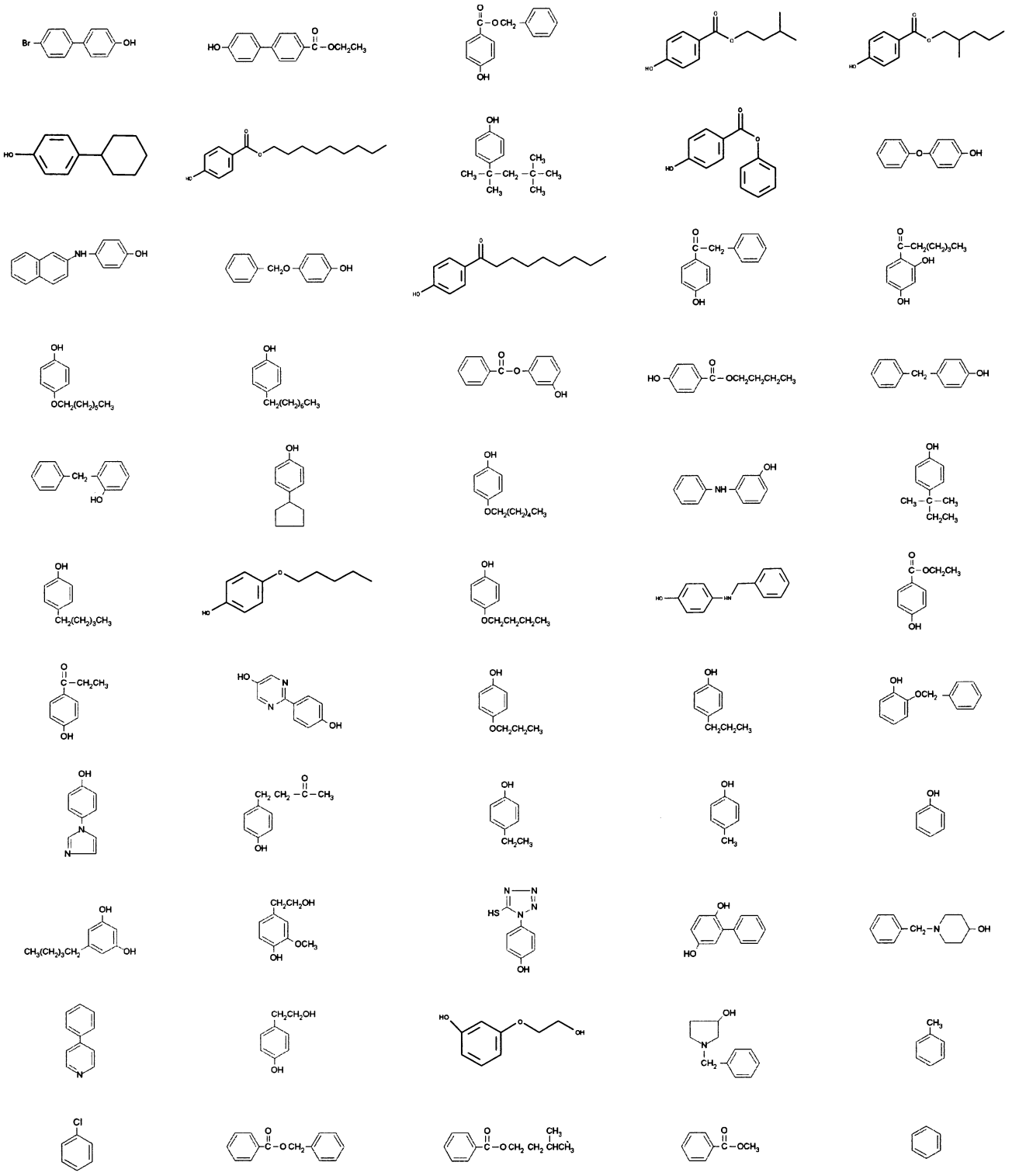


Figure 1. Structure for 120 compounds.

Molecular Quantum Similarity Measures (MQSM). Quantum similarity provides a quantitative measure of the resemblance between two quantum objects, A and B. A *molecular quantum similarity measure* (MQSM)[20] is defined by the simplest observable quantum chemical quantity, i.e., the electron distribution characterized by density functions. MQSMs are computed as the scalar product between the first-order molecular density functions $\rho_A(\mathbf{r}_1)$ and $\rho_B(\mathbf{r}_2)$ of the two molecular structures being compared, weighted by a nondifferential positive definite two-electron operator $\Omega(\mathbf{r}_1, \mathbf{r}_2)$. Analytically, MQSM are characterized by the value of the integral expressed in eq 1.

$$Z_{AB} = \int\int \rho_A(\mathbf{r}_1)\Omega(\mathbf{r}_1,\mathbf{r}_2)\rho_B(\mathbf{r}_2)\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 \qquad (1)$$

According to the form of the operator, different types of MQSM can be defined. In particular, the most extensively used MQSMs are the so-called *overlap MQSM*,[20] in which

Table 1. Functional Group Categories

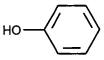
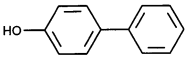
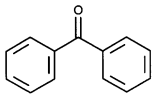
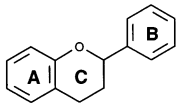
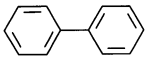
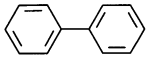
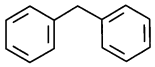
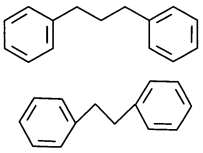
FUNCTIONAL GROUP CLASSIFICATION	STRUCTURES
Phenols	
Bisphenols	
Benzophenones	
Flavonoids	
Biphenyls	

Table 2. Fundamental Structural Basis for Classification

N. CLASS	CLASSIFICATION	STRUCTURES
1	A single aromatic ring	
2	Two directly bonded aromatic rings	
3	Two aromatic rings separated by one bridging atom	
4	Two aromatic rings two benzene rings separated by two or three bridging atoms	

the Ω operator is a Dirac's delta function, as shown in eq 2,

$$Z_{AB} = \int\int \rho_A(\mathbf{r}_1)\delta(\mathbf{r}_1 - \mathbf{r}_2)\rho_B(\mathbf{r}_2)\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 = \int \rho_A(\mathbf{r})\rho_B(\mathbf{r})\mathrm{d}\mathbf{r} \quad (2)$$

and the *Coulomb MQSM*,[20] where the Coulomb operator $\Omega = |\mathbf{r}_1 - \mathbf{r}_2|^{-1}$ is employed.

$$Z_{AB} = \int\int \rho_A(\mathbf{r}_1)\frac{1}{|\mathbf{r}_1 - \mathbf{r}_2|}\rho_B(\mathbf{r}_2)\mathrm{d}\mathbf{r}_1\mathrm{d}\mathbf{r}_2 \quad (3)$$

Overlap MQSMs are highly sensitive to exact atom superpositions, whereas the Coulomb operator acts as an attenuation factor for density distributions, along with being computationally more expensive, due to the greater complexity of the integral. The fact that all the compounds in the data set have at least one aromatic ring allows an exact atom−atom superposition for all the existing molecular pairs and, therefore, overlap MQSM provide similarity terms with reliable values; subsequently, all the models presented here correspond to this type of MQSM.

The use of this type of measure is in agreement with the study published by Gao et al.,[37] who reported that interactions between the estrogen receptor and the substituents of the ligands could be explained quantitatively in most cases simply by using steric parameters.

Quantum similarity matrices, $\mathbf{Z} = \{Z_{AB}\}$, constructed with the overall set of pairwise MQSMs for a given molecular family expressed in matrix form, may, therefore, be used as a source of molecular descriptors in QSAR analysis.

Computation of the Molecular Density Functions and Determination of the Molecular Alignment. To avoid computationally expensive calculations using ab initio density functions, accurate fitted first-order molecular density functions have been computed using the *Promolecular Atomic Shell Approximation* (ASA).[38,39] This approximation, which considers the molecular electron density as a sum of atomic density contributions, assumes the atomic densities to be linear combinations of 1*S* Gaussian functions. Atomic densities are fitted to an atomic ab initio calculation with a 3-21G basis set[38] and stored in a database. After that, the molecular densities are calculated by summing these atomic contributions. The coefficients and exponents for the 3-21G Promolecular ASA fitting can be downloaded from a Web site.[40]

Molecular geometries were optimized using the AM1 all-valence electron, semiempirical Hamiltonian of MOPAC 6.0.[41,42] To determine the minimum global energy conformer the following key words were specified "MMOK XYZ AM1 PRECISE NOINTER GEO-OK". As can be deduced from eqs 1−3, MQSMs depend on the relative orientation of the molecules compared. Thus, the alignment criterion chosen to overlay the structures appropriately is the maximization of the MQSM integral value, proposed by Constans et al.[43] To perform the density fitting, the matching procedure and the MQSM computations, the MOLSIMIL-97 software[44] was used.

In alignment processes, time efficiency is an important factor, due to the high number of pairwise superpositions that need to be computed, that is, *N(N-1)/2* superpositions for *N* molecular structures. Thus, in the present matching procedure, 7140 alignments were calculated, taking an average of 9 s and 17 s for Overlap and Coulomb MQSM, respectively, tested in a PIII with a 1000 MHz processor, 256 Mb RAM, and Win98 SE.

Statistical Methods: Treatment of Quantum Similarity Matrices. For purposes of statistical analysis, active compounds were assigned a value of 1, and inactive compounds a value of 0.

In the statistical treatment of the similarity matrices, several methods, implemented in the statistical package Minitab,[45] were used. To reduce the dimensionality of the quantum similarity matrices, *principal component analysis* (PCA) was performed upon MQSI. This method considers the molecules as points in a low-dimensional Euclidean space. Coordinates for these points are obtained in such a way that the interpoint distances match as well as possible the original similarities. Molecular coordinates in the new multidimensional space constitute the *principal coordinates* (PCs) of the system. The information encoded by the similarity matrices is reorganized, in such a way that the data variance explained by each PC is associated to the corresponding eigenvalue. So that to avoid the participation of PCs related to the transformation background noise, all those PCs accounting for less than 1% variance are neglected, and thus the matrix dimensionality is effectively reduced.

Table 3. Structures, CAS Number, and Relative Estrogenic Gene Activation for 120 Compounds

no. id	compound	CAS	measured activity	binary activity
1	17-β-estradiol	50-28-2	3.91e-11	A
2	4,4′-diethylethylene bisphenol	6898-97-1	9.11e-11	A
3	4,4′-cyclohexylidene bisphenol	843-55-0	4.43e-08	A
4	4,4′-thiodiphenol	2664-63-3	7.15e-08	A
5	bis(4-hydroxyphenyl)methane	620-92-8	8.15e-08	A
6	4,4′-ethylidene bisphenol	2081-08-5	2.28e-07	A
7	bisphenol A	80-05-7	4.28e-07	A
8	4,4′-(1,3-adamantanediyl) bisphenol	37677-93-3	6.10e-07	A
9	4,4′-dihydroxybenzophenone	611-99-4	6.33e-07	A
10	bis(4-hydroxyphenyl)sulfone	80-09-1	7.50e-05	A
11	1,1,1-tris(4-hydroxyphenyl)ethane	27955-94-8	1.03e-03	A
12	4,4′-dimethoxybiphenyl	2132-80-1	nonactive	I
13	4,4′-dipyridyl	553-26-4	nonactive	I
14	2,4-dihydroxybenzophenone	131-56-6	4.23e-08	A
15	2,2′,4,4′-tetrahydroxyl benzophenone	131-55-5	1.98e-07	A
16	4-chloro-4′-hydroxy benzophenone	42019-78-3	4.20e-07	A
17	3-hydroxybenzophenone	13020-57-0	4.93e-07	A
18	4-hydroxybenzophenone	1137-42-4	9.84e-07	A
19	2,3,4-trihydroxybenzophenone	1143-72-2	1.27e-06	A
20	2,4,4′-trihydroxybenzophenone	1470-79-7	1.41e-07	A
21	2,2′-dihydroxybenzophenone	835-11-0	nonactive	I
22	4,4′-dichlorobenzophenone	90-98-2	nonactive	I
23	2-hydroxybenzophenone	117-99-7	nonactive	I
24	4-methoxybenzophenone	611-94-9	nonactive	I
25	4-chlorobenzophenone	134-85-0	nonactive	I
26	4-methylbenzophenone	134-84-9	nonactive	I
27	4-nitrobenzophenone	1144-74-7	nonactive	I
28	benzophenone	119-61-9	nonactive	I
29	genistein (4′,5,7-trihydroxyisoflavone)	446-72-0	1.81e-07	A
30	biochanin A (5,7-dihydroxy-4′-methoxy isoflavone)	491-80-5	6.87e-07	A
31	naringenin (4′,5,7-trihydroxyflavanone)	480-41-1	2.30e-05	A
32	morin hydrate (3,3′,5,5′,7-pentahydroxyflavone)	480-16-0	8.80e-05	A
33	daidzein (4′,7-dihydroxyisoflavone)	486-66-8	4.92e-05	A
34	phloretin (2′,4,4′,6′-tetrahydroxychalcone)	60-82-2	1.80e-05	A
35	4′-hydroxychalcone	2657-25-2	1.40e-05	A
36	apigenin (4′,5,7-trihydroxyflavone)	520-36-5	no EC50	–
37	genkwanin (4′,5-dihydroxy-7-methoxyflavone)	437-64-9	no EC50	–
38	galangin (3,5,7-trihydroxyflavone)	548-83-4	nonactive	I
39	baicalein (5,6,7-trihydroxyflavone)	491-67-8	nonactive	I
40	chrysin (5,7-dihydroxyflavone)	480-40-0	nonactive	I
41	flavone	525-82-6	nonactive	I
42	flavanone	487-26-3	nonactive	I
43	trans-chalcone	614-47-1	nonactive	I
44	2′,4′,6′-trichloro-4-biphenylol	–	1.29e-09	A
45	2′,3′,4′,5′-tetrachloro-4-biphenylol	–	6.30e-09	A
46	2′,5′-dichloro-4-biphenylol	–	3.00e-08	A
47	4′-chloro-4-biphenylol	–	5.98e-08	A
48	2′,3′,4′,5′-tetrachloro-3-biphenylol	–	1.58e-07	A
49	2,2′,5′-trichloro-4-biphenylol	–	1.78e-07	A
50	2′,5′-dichloro-3-biphenylol	–	2.04e-07	A
51	4,4′-biphenyldiol	92-88-6	2.63e-07	A
52	4-(1-hydroxyethyl) biphenyl	3562-73-0	7.88e-06	A
53	3-hydroxybiphenyl	580-51-8	9.18e-06	A
54	4-hydroxybiphenyl	92-69-3	1.15e-06	A
55	4-(2-hydroxypropyl)biphenyl	34352-74-4	1.84e-06	A
56	4-biphenylmethanol	3597-91-9	2.12e-06	A
57	3-chloro-4-biphenylol	–	3.82e-06	A
58	2-chloro-4-biphenylol	–	3.82e-06	A
59	2-hydroxybiphenyl	90-43-7	1.84e-05	A
60	4-methoxybiphenyl	613-37-6	3.39e-05	A
61	2′,5′-dichloro-2-biphenylol	–	5.23e-05	A
62	3,4′,5-trichloro-4-biphenylol	–	nonactive	I
63	3,3′,5,5′-tetrachloro-4,4′-biphenyldiol	–	nonactive	I
64	biphenyl	92-52-4	nonactive	I
65	4-(1-adamantyl)phenol	29799-07-3	8.55e-09	A
66	4-(4-bromophenyl)phenol	29558-77-8	2.37e-08	A
67	ethyl-4′-hydroxy-4-biphenyl carboxylate	50670-76-3	5.03e-08	A
68	benzyl-4-hydroxybenzoate	94-18-8	1.07e-07	A
69	isoamyl-4-hydroxybenzoate	6521-30-8	1.17e-07	A
70	2-ethylhexyl-4′-hydroxy benzoate	5153-25-3	1.36e-07	A
71	4-cyclohexylphenol	1131-60-8	1.39e-07	A
72	nonyl-4-hydroxybenzoate	38713-56-3	1.65e-07	A
73	4-(tert-octyl)phenol	140-66-9	1.77e-07	A
74	phenyl-4-hydroxybenzoate	17696-62-7	2.28e-07	A
75	4-phenoxyphenol	831-82-3	2.62e-07	A

Table 3 (Continued)

no. id	compound	CAS	measured activity	binary activity
76	N-(4-hyroxyphenyl)−2-naphthylamine	93-45-8	4.15e-07	A
77	4-(benzyloxy)phenol	103-16-2	5.43e-07	A
78	4-hydroxyoctanophenone	2589-73-3	8.85e-07	A
79	benzyl-4-hydroxyphenyl ketone	2491-32-9	9.20e-07	A
80	4-hexanoyl resorcinol	3144-54-5	9.38e-07	A
81	4-heptyloxyphenol	13037-86-0	1.88e-06	A
82	4-octylphenol	1806-26-4	1.89e-06	A
83	resorcinol monobenzoate	136-36-7	1.95e-06	A
84	butyl-4-hydroxybenzoate	94-26-8	2.01e-06	A
85	4-hydroxydiphenylmethane	101-53-1	2.12e-06	A
86	2-hydroxydiphenylmethane	28994-41-4	2.12e-06	A
87	4-cyclopentyl phenol	1518-83-8	2.41e-06	A
88	4-hexyloxyphenol	18979-55-0	4.02e-06	A
89	3-hydroxydiphenylamine	101-18-8	4.20e-06	A
90	4-(tert-pentyl)phenol	80-46-6	4.76e-06	A
91	4-*n*-pentylphenol	14938-35-3	9.50e-06	A
92	4-pentyloxyphenol	18979-53-8	1.73e-05	A
93	4-butoxyphenol	122-94-1	1.88e-05	A
94	N-benzyl-4-hydroxyaniline	103-14-0	6.27e-05	A
95	ethyl-4-hydroxybenzoate	120-47-8	7.52e-05	A
96	4-hydroxypropiophenone	70-70-2	8.32e-05	A
97	2-(4-hydroxyphenyl)−5-pyrimidinol	142172-97-2	1.33e-04	A
98	4-propoxyphenol	18979-50-5	1.64e-04	A
99	4-propylphenol	645-56-7	1.84e-04	A
100	2-(benzyloxy)phenol	6272-38-4	2.50e-04	A
101	4-(imidazol-1-yl)phenol	10041-02-8	1.25e-03	A
102	4-(4-hydroxyphenyl)−2-butanone	5471-51-2	1.22e-03	A
103	4-ethylphenol	123-07-9	no EC50	-
104	4-methylphenol	106-44-5	nonactive	I
105	phenol	108-95-2	nonactive	I
106	5-pentylresorcinol	500-66-3	nonactive	I
107	homovanillyl alcohol	2380-78-1	nonactive	I
108	1-(4-hydroxyphenyl)−1H-tetrazole-5-thiol	52431-78-4	nonactive	I
109	phenyl hydroquinone	1079-21-6	nonactive	I
110	1-benzyl-4-hydroxypiperidine	4727-72-4	nonactive	I
111	4-phenylpyridine	939-23-1	nonactive	I
112	2-(4-hydroxyphenyl)ethanol	501-94-0	nonactive	I
113	O−(2-hydroxyethyl) resorcinol	49650-88-6	nonactive	I
114	1-benzyl-3-pyrrolidinol	775-15-5	nonactive	I
115	toluene	108-88-3	nonactive	I
116	chlorobenzene	108-90-7	nonactive	I
117	benzyl benzoate	120-51-4	nonactive	I
118	isoamyl benzoate	94-46-2	nonactive	I
119	methyl benzoate	93-58-3	nonactive	I
120	benzene	71-43-2	nonactive	I

Afterward, the connections between the most descriptive variables and the binary property were examined by means of stepwise regression. The most predictive variables were subsequently entered into a *multidimensional discriminant analysis* (MDA),[46,47] to classify molecules into the two groups and investigate the contribution of predictors to the grouping.

The assignment of experimental data into discrete categories, as described by Gao et al.,[36] is useful in the use of *high throughput screening* (HTS) to identify lead compounds, especially in noncongeneric libraries where there is no common structure. The use of this qualitative approach allows for the correlation between chemical structures and discrete activities to be obtained, at least for preliminary compound selection. To develop QSARs, the data set was split into a training set and an external test set, which was removed before modeling the activities of the compounds. Activities for the test set were estimated to provide an estimate of predictivity. The goodness of fit of the models was quantified by the percentage of correct classifications from the model (% CC) and the leave-one-out cross-validation (% CC^{CV}).

The accuracy of the predictions for active compounds was calculated by dividing the number of active compounds correctly assigned by the model by the total number of active compounds; conversely, the accuracy on the predictions for inactive compounds was calculated as the proportion of correctly predicted inactive compounds, out of the total number of inactive compounds. Finally, the overall accuracy of the model was also calculated, considering all the compounds.

Several selection criteria were applied to obtain appropriate external test sets for validation. In all the selection criteria, the data set was split into two approximately equal sets, maintaining the same proportion of active and inactive compounds for both. Following this split, calculations were performed, and then the process was repeated using the test set as the training set and the training set as the test set. In the first choice of training and test sets, the selection was made according to the distribution of the three most predictive variables in a three-dimensional space. Compounds were selected in this way in order to ensure that training and evaluation sets contained chemicals representative of the

Table 4. Physico-Chemical Descriptors and Indicator Variables

descriptor	abbreviation
logarithm of octanol−water partition coefficient	Log P
molecular weight	MW
number of atoms	At
number of carbons	C
number of hydrogen bond donor groups	HB_don
number of hydrogen bond acceptor groups	HB_acc
number of hydroxyl groups	OH
number of hydroxyl groups in para-position	p-OH
number of rings	R
number of benzenes	Bz
number of phenols	Ph

Table 5. Results for the Optimal QSAR Model for the Entire Set[a]

selected descriptors	% CC	% CC^{CV}	no. descriptors
# Ph, PC3, # C	0.880	0.872	3

[a] % CC = percentage of correct classifications; % CC^{CV} = percentage of correct classifications for cross-validation.

diversity of structures for which predictions were to be made. In a second selection, and to ensure that the results were not conditioned for the distribution of the data, the training and test sets were selected randomly.

Molecular Similarity Measures and Physicochemical Descriptors. In this work, to correlate molecular structural information to experimentally measured biological properties, molecular quantum similarity indices (MQSI) as well as some indicator variables were used. MQSI mainly encode information regarding steric or electrostatic distribution on the surface of the molecule. However, to enhance the quality of the models for estrogenic activity some variables indicating the presence or absence of explicit structural features have been also included. To account for hydrophobicity, the logarithm of the octanol−water partition coefficient was calculated using the KOWWIN software.[48] Several indicator variables were also calculated with the TSAR for Windows software.[49] The complete pool of descriptors used is shown in Table 4.

RESULTS AND DISCUSSION

This study has attempted to develop structure-based methods to predict the potential of compounds to promote an estrogenic effect. From the 120 structurally diverse compounds, the activities of 117 molecules of the data set have been assessed, and models were developed on the basis of quantum measures of similarity.

Linear Discriminant Analysis of the Complete Data Set. First, the set of 117 compounds with available reported activity values was analyzed as a whole. The statistical coefficients for the resulting QSAR models obtained for the whole molecular set are gathered in Table 5.

As expected, the most important descriptors were some of the constitutional parameters accounting for structural features such as the number of phenols and the number of carbons, together with the third principal component extracted from the matrix of molecular quantum similarity indices. The presence of a phenolic OH group resembling the 3-hydroxyl group of the estradiol molecule seems to be essential for the effective binding to estrogen receptor. In addition, the number of carbons is indicative of the hydrophobic contribution from the ligand. Finally, the third principal component, accounting for the 4.3% of the total variation of the similarity indices, was found to be the most predictive.

Table 6. Misclassified Compounds for the Model Presented in Table 5

compound	binary activity	predicted activity	misclassification
21	0	1	false positive
23	0	1	false positive
52	1	0	false negative
55	1	0	false negative
56	1	0	false negative
60	1	0	false negative
62	0	1	false positive
63	0	1	false positive
66	1	0	false negative
80	1	0	false negative
84	1	0	false negative
95	1	0	false negative
109	0	1	false positive
112	0	1	false positive

With regard to the number of parameters included in the model, the objective was to formulate the most predictive model with the minimum number of parameters. Thus, although there were models having more descriptors and slightly improved better predictive capability, the small improvement achieved when adding more parameters to the model did not justify the increase in the complexity of the equation.

The percentage of correctly classified compounds for the model presented in Table 5 is very satisfactory. Inclusion of more than three-parameters in a model (which correctly classified the activity of 103 of 117 (88.0%) molecules) did not improve the statistics significantly; this was considered the best for modeling purposes. The misclassified compounds were molecules **21**, **23**, **52**, **55**, **56**, **60**, **62**, **63**, **66**, **80**, **84**, **95**, **109**, and **112** of Table 3. Seventy-three of 81 (90.1%) active compounds and 30 of 36 (83.3%) inactive compounds were correctly classified. To test the predictive capacity, or the capability of the model to interpolate within the training set population, the *leave-one-out* (LOO) cross-validation procedure was performed.[50,51] As before, the quality of the cross-validation predictions was quantified as the percentage of correct classifications. Thus, the model classified 102 of 117 (87.2%) molecules correctly when both groups are considered, and 72 out of 81 (88.9%) and 30 out of 36 (83.3%) compounds for active and inactive molecules, respectively.

Analysis of the misclassified compounds presented in Table 6 indicates that there may be plausible explanations, in terms of structural requirements, for the incorrect classifications. Compounds **21**, **23**, **62**, and **63** have been classified as being active as they appear to have all the implicit characteristics required for the pharmacophore. However, the model has not taken into account the possibility of formation of intramolecular hydrogen bonds. The hydroxy groups of molecules **21** and **23** are able to form intramolecular hydrogen bonds with an adjacent carbonyl group, whereas hydroxy groups of molecules **62** and **63** may have a weaker intramolecular hydrogen bond with a neighboring chlorine atom. Compounds **52**, **55**, **56**, and **60** are misclassified as inactive compounds due to the fact that they do not contain a phenolic group. It can be seen, though, that the oxygen atom present either as an ether or as an alcohol

Table 7. Selected Optimal Model for the Different Fundamental Structure Classes

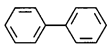
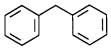
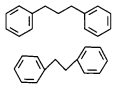
Fund. Struc. Class	Number of molecules	Selected Descriptors	% CC	% CC^{CV}	Misclassified Compounds	Predicted Activity
	37	p-OH, log P, PC6	97.3	91.9	104	1 False positive
	28	PC14, PC4, PC17	92.9	82.1	47, 57	2 False negatives
	28	p-OH, PC20	92.9	89.3	17, 86	2 False negatives
	23	Ph, PC1	100	100	-	-

Table 8. QSAR Results for the Training and Test Sets

training test			test set		
N.	% CC	% CC^{CV}	N.	% CC	misclassified compounds
59	77.9	76.3	58	87.9	23, 39, 40, 97, 98, 99
58	91.4	84.5	59	83.1	21, 24, 56, 60, 62, 78, 80, 95, 96, 109
59	83.1	79.7	58	62.0	38, 40, 42, 44, 46, 48, 50, 52, 54, 56, 58, 60, 69, 71, 73, 75, 83, 87, 91, 93, 99, 101
58	89.7	87.9	59	79.7	21, 23, 39, 41, 55, 63, 68, 90, 96, 98, 102, 106

group in the aromatic ring could act as a weaker hydrogen-bond acceptor. The unexplained misclassification observed for the compound **66** could be due to the presence of the bromine atom; this is the only molecule containing such a heavy atom and, thus, the result of the MQSM could have been biased for this reason. Finally, molecules **80**, **84**, and **95**, all containing a carbonyl group in the para position to the hydroxy group and a single aromatic ring, have been misclassified as inactive, possibly due to the lack of the appropriate hydrophobic area. There is no clear explanation for the incorrect assignment of the activities of molecules **109** and **112**.

Class-Based Models. The predictions for the classes developed from a fundamental structural basis are summarized in Table 7. In the first model, for compounds with only one aromatic ring, a three-parameter model was chosen as being optimal and led to a correct classification of 36 out of 37 (97.3%) molecules. The only misclassification was molecule **104**, 4-methylphenol, an inactive phenol predicted to be active. As expected, cross-validated results for almost all the models showed a slight decrease in classification rate. For the set of compounds with a biphenyl-like structure, belonging to class 2, the optimal model included only MQSM descriptors. This may possibly be due to the exact interatomic alignment of the rigid biphenyl common pattern. In this case, the activity for 26 of 28 compounds (92.9%) has been correctly assigned. The two misclassifications were both false negatives: compounds **47** and **57**, corresponding to 4′-chloro-4-biphenylol and 3-chloro-4-biphenylol, respectively. In the third group of compounds, composed of molecules with two aromatic rings separated with one bridging atom, the results are also satisfactory, again with two misclassified false negatives*:* 3-hydroxybenzophenone (**17)** and 2-hydroxydiphenylmethane (**86**). Finally, the model for the most heterogeneous subset that combines compounds with either two or three bridging atoms does not misclassify any of the molecules in either the adjustment or cross-validated results.

Comparison of results for the modeling of the complete data with those for the modeling of separate classes indicates that slightly poorer results were obtained for the complete data set. However, the results indicate that the model for the complete data set is biased to produce a greater percentage of false positives than false negatives. For drug discovery, false positives are of concern due to the unjustified cost of synthesizing a chemical with a low probability of being efficacious. Conversely, for regulatory purposes, the main goal is to minimize the rate of false negatives, to avoid any threat to public health. Hence, this could suggest the use of this type of model for health hazard assessment.

Training and External Test Sets. Models were validated using a *leave-many-out* procedure, in which a more significant number of chemicals was omitted from the modeling process in comparison with the LOO. To achieve this, the complete data set was split into a training and a test set prior to modeling. Test set selection was made with two criteria. First, the two sets were chosen on the basis of the distribution of the most significant descriptors (i.e. the number of phenol rings, the third principal component and the number of carbons). This complies with the hypothesis that in similarity-based QSAR approaches the training set should be as structurally diverse as possible. From a knowledge of the spatial distribution of the compounds in the 3-D descriptor space, they were selected to provide representative sample in both the training and test sets. Following the selection of the two sets, the model was redeveloped on the former test set and validated on the former training set. In a second approach to test set selection, the test set was extracted from the whole set randomly. For the sake of comparison with the previous assay, training and test sets were swapped and modeling and validation reperformed. The results of the validation exercises on the whole data set are summarized in Table 8.

It can be observed that, in relation to the predictive ability of the test sets selected from a knowledge of the 3-D distribution of the descriptors was comparable to that for the complete data set (87.9% and 83.1% correct classifications respectively). However, in the randomly chosen test sets, the correct classification rates were lower (62.0% and

79.7%). The decrease in the predictivity of the models due to the test set selection method is to be expected. Analysis of the results for the different test sets indicates that more than the half of misclassified compounds in the two former test sets coincides with the misclassified ones in the latter cases, i.e., the misclassified compounds **23**, **39**, **40**, **98**, and **99** from the first test set and **21**, **56**, **60**, and **96** from the second one coincide with the results of the third and fourth sets. For example, in the first case, the model fails to classify compounds **23**, **39**, **40**, **97**, **98**, and **99**. All those molecules, except molecule **97**, were incorrectly assigned in the other test subsets. This could be an indication of some intrinsic features in the remaining compounds hindering the structure−activity relationship; thus, some important information might be missing from these models and the selection of the test set. Conversely, for the misclassified compounds from the complete set, the classes based on fundamental structure and the test and training sets do not coincide. A reason for this behavior could be the different type of information encoded in the different descriptors used to build the models.

Another important remark concerning the predictivity of the models must also be made. Three compounds in Table 3 (**36**, **37**, and **103**) do not possess an experimentally measured estrogenic binding affinity. When the complete set of molecules is used to predict the binding ability of these ligands, the model suggested all three compounds to be unequivocally active. This is in agreement with results obtained in the literature, using other experimental procedures to measure the activity.[52−54]

CONCLUSIONS

This paper has focused on the use of quantum similarity to predict the binding affinity of estrogenic compounds. The structure−activity relationships developed have been constructed by correlating the structural information, described by quantum-chemical indices, physicochemical properties and indicator variables, with the biological activity. To achieve this, binary activity values have been treated by means of a qualitative statistical technique in order to classify the molecules into active or inactive groups and examine the contribution of the predictors to the categories. Additionally, the modeling procedures have been validated by the use of external test sets, for which predictions were made satisfactorily. Furthermore, predictions were made for a further three molecules of unknown activity, and the predicted activities were in agreement with other work reported in the literature.

With regard to the wide range of structurally diverse chemicals that constitute estrogenic endocrine disruptors, the SAR models for ER binding have been developed to predict if an estrogenic compound is active. Hence, such computer-based methods provide the possibility to screen for potential estrogens and predict their activity. Therefore, quantum similarity in conjunction with the judicious use of structural descriptors is a valuable tool for QSAR and computer-aided drug design.

The success of the different proposed approaches confirms that, whereas hydrophobicity and the presence of some functional groups play a central role in determining toxic potency, the electronic effects derived from quantum similarity theory are able to discriminate between active and inactive estrogen ligands. There is high confidence in the models presented as they are based on a small number of transparent and mechanistically determinable parameters.

ACKNOWLEDGMENT

This work has been partially supported by a pre-doctoral fellowship from the *Ministerio de Ciencia y Tecnología* (A. Gallegos Saliner). Finally, A.G.S. wishes to thank to Dr. Tanya Netzeva for her kindness and interesting comments.

REFERENCES AND NOTES

(1) Kavlock, R. J.; Daston, G. P.; DeRosa, C.; Fenner-Crisp, P.; Gray, L. E.; Kaattari, S.; Lucier, G.; Luster, M.; Mac, M. J.; Maczka, C.; Miller, R.; Moore, J.; Rolland, R.; Scott, G.; Sheehan, D. M.; Sinks, T.; Tilson, H. A. Research needed for the risk assessment of health and environmental effects of endocrine disruptors, a report of the U.S. EPA-sponsored workshop. *Environ. Health Perspect.* **1996**, *104*, 715−740.

(2) U.S. Congress. The Food Quality Protection Act (FQPA) and the Safe Drinking Water Sct. (SDWA), 1996.

(3) Lai, K. M.; Scrimshaw, M. D.; Lester, J. N. The effects of natural and synthetic steroid estrogens in relation to their environmental occurrence. *Crit. Rev. Toxicol.* **2002**, *32(2)*, 113−132.

(4) Kelce, W. R.; Stone, C. R.; Laws, S. C.; Gray, L. E.; Kemppainen, J. A.; and Wilson, E. W. Persistent DDT metabolite p, p′-DDE is a potent androgen receptor antagonist. *Nature* **1995**, *375*, 581−585.

(5) Colburn, T.; Dumanoski, D.; Myers, J. P. In *Our Stolen Future*; Plume*:* New York, 1996.

(6) EDSTAC http://www.epa.gov/opptintr/opptendo/finalrpt.htm.

(7) Bradbury, S.; Menkenyan, O. A. Quantitative structure−activity relationships for polychlorinated hydroxybiphenyl estrogen receptor binding affinity. An assessment of conformer flexibility. *Environ. Toxicol. Chem.* **1996**, *15*, 1945−1954.

(8) Tong, W. D.; Perkins, R.; Strelitz, R.; Collantes, E. R.; Keenan, S.; Welsh, W. J.; Branham, W. S.; Sheehan, D. M. Quantitative structure−activity relationships (QSARs) for estrogen binding to the estrogen receptor, predictions across species. *Environ. Health Perspect.* **1997**, *105*, 1116−1124.

(9) Tong, W. D.; Lowis, D. R.; Perkins, R.; Chen, Y.; Welsh, W. J.; Goddette, D. W.; Heritage, T. W.; Sheehan, D. M. Evaluation of Quantitative structure−activity relationships methods for large-escale prediction of chemicals binding to the estrogen receptor. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 669−677.

(10) Dodds, E. C.; Goldberg, L.; Lawson, W. Oestrogenic activity of alkylated stilboestrols. *Nature* **1938**, *142*, 34.

(11) Dodge, J. A. Natural and anthropogenic environmental oestrogens, the scientific basis for risk assessment. Structure/activity relationships. *Pure Appl. Chem.* **1998**, *70(9)*, 1725−1733.

(12) Colburn, T. Environmental estrogens, health implications for humans and wildlife. *Environ. Health Perspect.* **1995**, *103(7)*, 135−136.

(13) Sabih, J. Natural and anthropogenic environmental oestrogens, the scientific basis for risk assessment. Issues associated with the validation of in vitro and in vivo methods for assessing endocrine disrupting chemicals. *Pure Appl. Chem.* **1998**, *70(9)*, 1735−1745.

(14) Shi, L. M.*;* Fang, H.*;* Tong, W. D.*;* Wu, J.*;* Perkins, R.*;* Blair, R. M.*;* Branham, W. S.*;* Dial, S. L.*;* Moland, C. I.*;* Sheehan, D. M. QSAR models using a large diverse set of estrogens. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 186−195.

(15) Nishihara, T.*;* Nishikawa, J.*;* Kanayama, T.*;* Dakeyama, F.*;* Saito, K.*;* Imagawa, M.*;* Takatori, S.*;* Kitagawa, Y.*;* Hori, S.*;* Utsumi, H. Estrogenic activities of 517 chemicals by yeast two-hybrid assay. *J. Health. Sci.* **2000**, *46*, 282−298.

(16) Cramer III, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect on Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(17) Gantchev, T. G.; Ali, H.; van Lier, J. E. Quantitative structure−activity relationships/comparative molecular field analysis (QSAR/CoMFA) for receptor binding properties of halogenated estradiol derivatives. *J. Med. Chem.* **1994**, *37*, 4164−4176.

(18) Mekenyan, O.; Nikolova, N.; Karabunarliev, S.; Bradbury, S. P.; Ankley, G. T.; Hansen, B. New developments in hazard identification algorithm for hormone receptor ligands. *Quant. Struct.-Act. Relat.* **1999**, *18*, 139−153.

(19) 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.

(20) Carbó, R.; Arnau, J.; Leyda, L. How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.* **1980**, *17*, 1185−1189.
(21) Carbó, R.; Besalú, E.; Amat, L.; Fradera, X. Quantum molecular similarity measures (QMSM) as a natural way leading towards a theoretical foundation of quantitative structure-properties relationships. *J. Math. Chem.* **1995**, *18*, 237−246.
(22) Fradera, X.; Amat, L.; Besalú, E.; Carbó-Dorca, R. Application of molecular quantum similarity to QSAR. *Quant. Struct.-Act. Relat.* **1997**, *16*, 25−32.
(23) Robert, D.; Amat, L.; Carbó-Dorca, R. Three-dimensional quantitative structure−activity relationships from tuned molecular quantum similarity measures. Prediction of the corticosteroid-binding globulin binding affinity for a steroid family. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 333−344.
(24) Robert, D.; Gironés, X.; Carbó-Dorca, R. Facet diagrams for quantum similarity data. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 597−610.
(25) Robert, D.; Carbó-Dorca, R. Aromatic compounds aquatic toxicity QSAR using quantum similarity measures. *SAR QSAR Environ. Res.* **1999**, *10*, 401−422.
(26) Gallegos, A.; Robert, D.; Gironés, X.; Carbó-Dorca, R. Structure-Toxicity Relationships of Polycyclic Aromatic Hydrocarbons using Molecular Quantum Similarity. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 67−80.
(27) Carbó-Dorca, R.; Besalú, E. A general survey of molecular quantum similarity. *J. Mol. Struct. (THEOCHEM)*, **1998**, *451*, 11−23.
(28) Carbó-Dorca, R.; Amat, L.; Besalú, E.; Lobato, M. In *Advances in Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; JAI Press*:* Greenwich, 1998; Vol. 2, pp 1−42.
(29) Carbó-Dorca, R.; Amat, L.; Besalú, E.; Gironés, X.; Robert, D. In *Proceedings of 4th Girona Seminar on Molecular Similarity*; Carbó-Dorca, R., Mezey, P. G., Eds.; Kluwer: Dordrecht, 2001; pp 187−320.
(30) Fang, H.; Tong, W.; Shi, L. M.; Blair, R.; Perkins, R.; Branham, W.; Hass, B. S.; Xie, Q.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. Structure−activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. *Chem. Res. Toxicol.* **2001**, *14*, 280−294.
(31) Shi, L.; Tong, W.; Fang, H.; Xie, Q.; Hong, R.; Perkins, R.; Wu, J.; Tu, M.; Blair, R. M.; Branham, W. S.; Waller, C.; Walker, J.; Sheehan, D. M. An integrated "4-phase" approach for setting endocrine disruption screening priorities-phase I and II predictions of estrogen receptor binding affinity. *SAR QSAR Environ. Res.* **2002**, *13(1)*, 69−88.
(32) Schultz, T. W.; Sinks, G. D.; Cronin, M. T. D. Structure−Activity Relationships for Gene Activation Oestrogenicity. Evaluation of a Diverse Set of Aromatic Chemicals. *Environ. Toxicol.* **2002**, *17*, 14−23.
(33) Danielian, P. S.; White, R.; Lees, J. A.; Parker, M. G. Identification of a conserved region required for hormone dependent transcriptional activation by steroid-hormone receptors. *EMBO J.* **1992**, *11*, 1025−1033.
(34) Routledge, E. J.; Sumpter, J. P. Estrogenic activity of surfactants and some of their degradation products assessed using a recombinant yeast. *Environ. Toxicol. Chem.* **1996**, *15*, 241−248.
(35) Schultz, T. W.; Sinks, G. D.; Cronin, M. T. D. Effect of substituent size and dimensionality on potency of phenolic xenoestrogens evaluated with a recombinant yeast assay. *Environ. Toxicol. Chem.* **2000**, *19*, 2637−2642.
(36) Gao, H.; Williams, C.; Labute, P.; Bajorath, J. Binary Quantitative-Activity Relationship (QSAR) Analysis of Estrogen Receptor Ligands. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 164−168.
(37) Gao, H.; Katzenellenbongen, J. A.; Garg, R.; Hansch, C. Comparative QSAR Analysis of Estrogen Receptor Ligands. *Chem. Rev*. **1999**, *99*, 723−744.
(38) Amat, L.; Carbó-Dorca, R. Quantum similarity measures under atomic shell approximation, first-order density fitting using elementary Jacobi rotations. *J. Comput. Chem*. **1997**, *18*, 2023−2039.
(39) Amat, L.; Carbó-Dorca, R. Fitted electronic density functions from H to Rn for use in quantum similarity measures, cis-diamminedichloroplatinum(II) complex as an application example. *J. Comput. Chem*. **1999**, *20*, 911−920.
(40) ASA coefficients and exponents for an assorted sample of atoms can be seen and downloaded from the Web site, http://iqc.udg.es/cat/similarity/ASA/funcset.html.
(41) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(42) Mopac 6.0, A General Molecular Orbitals Package; Steward, J. J. P. QCPE, 1990.
(43) Constans, P.; Amat, L.; Carbó-Dorca, R. Towards a Global Maximization of the Molecular Similarity Function: The Superposition of Two Molecules. *J. Comput. Chem*. **1997**, *18*, 826−846.
(44) Amat, L.; Constans, P.; Besalú, E.; Carbó-Dorca, R. MOLSIMIL 97, Institute of Computational Chemistry, University of Girona, Spain, 1997.
(45) Minitab 13; Minitab Inc., 2002.
(46) Tou, J. T.; Gonzalez, R. C. *Pattern Recognition Principles*; Addison-Wesley: Reading, 1974.
(47) Hand, D. J. *Discrimination and classification*; Wiley: Chichester, 1993.
(48) KOWWIN v1.66; Syracuse Research Corporation, 2000.
(49) TSAR v3.2; Oxford Molecular Ltd., Oxford, U.K.
(50) Wold, S. Cross-validatory estimation of the number of components in factor and principal component models. *Technometrics* **1978**, *20*, 397−405.
(51) Montgomery, D. C.; Peck, E. A. In *Introduction to linear regression analysis;* Wiley: New York, 1992.
(52) Klappa, P.; Freedman, R. B.; Langenbuch, M.; Lan, M. S.; Robinson, G. K.; Ruddock, L. W. The pancreas-specific protein disulphide-isomerase PDIp interacts with a hydroxyaryl group in ligands. *Biochem. J.* **2001**, *354*, 553−559.
(53) Bridgette, M.; Collins, J. A.; McLachlan, S. F. A. The estrogenic and antiestrogenic activities of phytochemicals with the human estrogen receptor expressed in yeast. *Steroids* **1997**, *62(4)*, 365−372.
(54) Zand, R. S.; Jenkins, D. J.; Diamandis, E. P. Steroid hormone activity of flavonoids and related compounds. *Breast Cancer Res. Treat.* **2000**, *62(1)*, 35−49.

CI034014A