MTD−PLS: A PLS Variant of the Minimal Topologic Difference Method. III. Mapping Interactions between Estradiol Derivatives and the Alpha Estrogenic Receptor

Ludovic Kurunczi,*[,†,‡] Edward Seclaman,[†,‡] Tudor I. Oprea,[§] Luminita Crisan,[‡] and Zeno Simon[‡]

University of Medicine and Pharmacy "V. Babes", 2 Eftimie Murgu, 300041 Timisoara, Romania, Institute of Chemistry "C. Dragulescu" Timisoara, 24 Mihai Vitezul, 300223 Timisoara, Romania, and Division of Biocomputing, University of New Mexico, School of Medicine, Albuquerque, New Mexico 87131

Received March 4, 2005

A homogeneous collection of 45 estrogen agonist derivatives with relative binding affinities measured to the estrogen receptor from *Ratus norvegicus* was used. The quantitative structure−activity relationships were derived using an improved minimal topologic difference (MTD) method in a partial least-squares (PLS) variant. The spatially assigned analysis of fragment properties can provide receptor site maps, within the limits of the existing series. A steric misfit was found for the steroidal position 2; benefic hydrophobic and van der Waals (enhanced by high polarizability) interactions were found for the 17α−CH═CH−X group. MTD-PLS mapping results are confirmed by the experimentally derived estradiol−estrogen receptor binding site contacts (based on X-ray crystallography). Our results suggest that this MTD-PLS method can yield useful results for interactions with receptors of unknown 3D structure and, generally, for the steric rigidity of receptor sites.

INTRODUCTION

An improved minimal topologic difference (MTD)[1] method, MTD-PLS,[2,3] seeks to correlate spatially positioned properties for ligand series with a given target property. In this method, atoms from different molecules that occupy vertexes of the (manually derived) hypermolecule are characterized by fragmental increments in volume, partial electric charge, polarizability, hydrophobicity, and H-bond donor or acceptor properties. Variations in the above properties that occur at given hypermolecule nodes are then related to changes in biological activity. Certain restrictions, based on chemical intuition, are also introduced: steric misfit is always attributed a detrimental effect, while hydrophobic interactions and those based on polarizability are initially considered beneficial. Using this spatially assigned analysis of fragment properties, MTD-PLS can provide receptor site maps, within the limits of the existing series. Our comparison of the MTD-PLS mapping results with the experimentally derived acetylcholine−acetylcholinesterase binding site interactions (on the basis of X-ray crystallography) were encouraging.[3]

Here, we continue to compare MTD-PLS results to experimentally determined ligand binding modes, by applying this method to a series of estradiol derivatives binding to the rat (*Ratus norvegicus*) uterus estrogen receptor alpha (ERα), then comparing the ligand-based map with the ligand−receptor contacts from X-ray crystallography.[4] We demonstrate that key interactions, predicted with MTD-PLS, are in agreement with those identified from PDB (the Protein Data Bank).[5] We further observe that the information extracted during initial PLS (partial least squares projection to latent structures)[6] modeling leads to the same interpretation, even when followed by further refinement.

* Corresponding author e-mail: dick@acad-icht.tm.edu.ro.
† University of Medicine and Pharmacy "V. Babes".
‡ Institute of Chemistry "C. Dragulescu" Timisoara.
§ University of New Mexico.

METHODS AND MATERIALS

Activity Processing and Ligand Selection. Biological activities, Y_i for agonists, were processed as log values of the relative binding affinities (RBAs), defined in eq 1:

$$\mathrm{RBA} = 100[\mathrm{E}]_{50}/[\mathrm{C}]_{50} \quad (1)$$

where $[\mathrm{E}]_{50}$ is the unlabeled estradiol concentration that reduces the 6,7-[^{3}H]-estradiol saturation of the receptor to 50% and $[\mathrm{C}]_{50}$ is the ligand concentration producing the same effect. Only binding affinities determined at 0−4 °C[7] were used in our study. We eliminated some molecules that featured large substituents in the 17α position and had low RBA values, to forestall the accumulation of inactive compounds in the series. At the same time, to remove the significant accumulation of single-molecule/single-vertex occupancy in the hypermolecule (which would result in a decreased signal-to-noise ratio), we try to avoid structures that introduce many supplementary single-occupied vertexes in the hypermolecule. The $N = 45$ compounds that passed our exclusion criteria are listed in Table 1 (compounds 1−45).

Hypermolecule Construction. Performed according to previously outlined principles,[2,3] the hypermolecule was derived starting with the 3D structure of 17β-estradiol, as modeled with HyperChem 5.11 Pro.[8] Conformational analysis was performed only for side chains; only structures having energy values with a maximum of 1 kcal/mol above the minimum were retained. The superposition started with the (rigid) estradiol, using, for any given ligand, all the monitored low-energy conformations but retaining only the ones that introduce a minimal number of new vertexes. "If the molecule has several low energy conformations, it will adopt the one which fits best to the receptor",[9] meaning the one that resembles the conformations adopted by other molecules from the series the most. The hypermolecule depicted in

10.1021/ci050077c CCC: $30.25
Published on Web 07/28/2005

Table 1. Studied Compounds (1−45, Training Set), Their Activities (0−4 °C, Rat ER), and the Occupied Vertexes in Hypermolecule (see also Figure 1) and, for Compounds 16, 22, 30−32, 40, 46−51, the RBAs Determined for Lamb ER (in Parentheses)

number	compound (L_i)	RBA %	Y_i = log RBA	occupied vertexes (j)
1	estradiol (E2)	100	2.00	7, 12
2	17-deoxy-E2	74	1.87	7
3	3-OCH_3−E2	0.06	−1.22	7, 8, 12
4	3-F−E2	0.43	−0.37	7, 12
5	17β-NH_2−E2	82	1.91	7, 12
6	17-NNH_2−E2	81	1.91	7, 33, 34
7	17β-NHOH−E2	13	1.11	7, 12, 13
8	17-NOH−E2	13	1.11	7, 33, 34
9	17β-OCH_3−E2	14	1.15	7, 12, 13
10	17β-$OCOCH_3$−E2	22	1.34	7, 12−15
11	2-OCH_2CH_3−E2	0.01	−2.00	2−4, 7, 12
12	2-$NHCH_2CH_3$−E2	0.35	−0.46	2−4, 7, 12
13	2-SCH_2CH_3−E2	0.29	−0.54	2−4, 7, 12
14	2-CH═$CHCH_3$−E2	0.01	−2.00	2, 4, 7, 12, 35
15	1-OH−E2	19	1.28	1, 7, 12
16	2-F−E2	51 *(100)*[a]	1.71 *(2.00)*[a]	2, 7, 12
17	2-Cl−E2	9.6	0.98	2, 7, 12
18	2-Br−E2	1.2	0.08	2, 7, 12
19	2-NO_2−E2	0.03	−1.52	2, 5−7, 12
20	2-NH_2−E2	10	1.00	2, 7, 12
21	4-CH_3−E2	10	1.00	7, 9, 12
22	4-F−E2	182 *(128)*[a]	2.26 *(2.11)*[a]	7, 9, 12
23	4-Cl−E2	110	2.04	7, 9, 12
24	4-NO_2−E2	6.2	0.79	7, 9−12
25	17α-CH_3−E2	73	1.86	7, 12, 16
26	17α-$CH_2CH_2CH_3$−E2	4.9	0.69	7, 12, 16−18
27	17α-$CH_2CH_2CH_2CH_3$−E2	4.7	0.67	7, 12, 16−19
28	17α-$CH_2CH_2CH_2OH$−E2	1.4	0.15	7, 12, 16−19
29	17α-C≡CCH_2OH−E2	6.5	0.81	7, 12, 16, 23−25
30	estrone	34 *(37)*[a]	1.53 *(1.57)*[a]	7, 33
31	17α-C≡CH−E2	104 *(70)*[a]	2.02 *(1.85)*[a]	7, 12, 16, 23
32	17α-C≡CCH_3−E2	32 *(44)*[a]	1.51 *(1.64)*[a]	7, 12, 16, 23, 24
33	17α-C≡$C(CH_2)_5CH_3$−E2	0.9	−0.05	7, 12, 16, 23−28
34	17α-C≡CI−E2	81	1.91	7, 12, 16, 23, 24
35	17α −C≡$C(CH_2)_3CH_2I$−E2	4.7	0.67	7, 12, 16, 23−28
36	17α-C≡$C(CH_2)_5CH_2I$−E2	0.8	−0.10	7, 12, 16, 23−30
37	17α-(E)CH═CHCl−E2	102	2.01	7, 12, 16, 20, 22
38	17α-(Z)CH═CHCl−E2	126	2.10	7, 12, 16, 20, 21
39	17α-(E)CH═CHBr−E2	78	1.89	7, 12, 16, 20, 22
40	17α-(Z)CH═CHBr−E2	195 *(117)*[a]	2.29 *(2.07)*[a]	7, 12, 16, 20, 21
41	17α-(E)CH═CHI−E2	77	1.89	7, 12, 16, 20, 22
42	17α-(Z)CH═CHI−E2	202	2.31	7, 12, 16, 20, 21
43	11β-OCH_3−E2	9.7	0.99	7, 12, 31, 32
44	11β-OCH_3-17α-C≡CH−E2	13.9	1.14	7, 12, 16, 23, 31, 32
45	11β-CH_2CH_3−E2	123	2.09	7, 12, 31, 32
46	*17α-(E)-CH═CHSPh−E2*	*24.5 (30.8)*[a]	*1.39 (1.49)*[a]	
47	*17α-(Z)-CH═CHSPh−E2*	*117 (81.3)*[a]	*2.07 (1.91)*[a]	
48	*17α-Ph−E2*	*30 (25)*[a]	*1.48 (1.4)*[a]	
49	*17α-C≡CtBu−E2*	*1.4 (1.5)*[a]	*0.15 (0.18)*[a]	
50	*17α-(Z)-CH═CHtBu−E2*	*3.4 (3)*[a]	*0.53 (0.48)*[a]	
51	*estriol [16α-OH−E2]*	*22 (22)*[a]	*1.34 (1.34)*[a]	

[a] In parentheses: the lamb ER RBA and activity values.[7] Compounds 46−51 were not used in the training set; these data will be used later in the model validation phase.

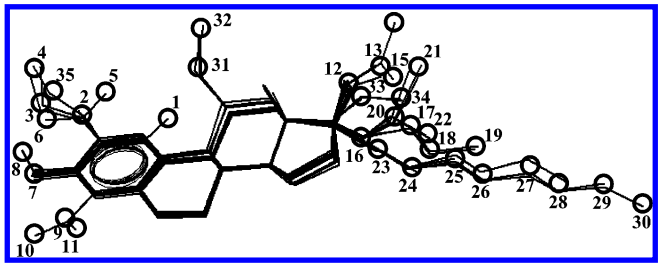

Figure 1. Hypermolecule construction and vertex numbering. The common rigid skeleton is used in superposition, and only the differing vertexes are numbered.

Figure 1 contains $M = 35$ vertexes. Table 1 contains also the numbers, j, of the vertexes occupied by each molecule.

A similar procedure to construct the hypermolecule, starting from representative coordinates and a cluster analysis, was developed by Kotani and Higashiura.[10]

Structural Parameters. Atomic partial charges (x_{ijS}; see also eq 2) were derived with the semiempirical AM1 method in HyperChem. Fragmental van der Waals volumes (x_{ijV}) were deduced using the procedure described by Olah.[11] Fragmental hydrophobicities (x_{ijH}) from Rekker[12] were adapted by introducing reasonable approximations to split some group contributions into atomic ones. For example, the pyridine N value (−1.06) was attributed to the aromatic −NO_2 nitrogen, and the aromatic −O− value (−0.433) was considered for the −C(═O)O group. The split of the corresponding C═O group was done by considering 0.195

Table 2. Fragmental Descriptors Used in This Work[a]

fragment	V (Å^3)	H (Rekker)	P (cm^3)
−OH	12.47	−0.343 (ar), −1.491 (al)	2.625 (2.13 in oxime)
−O− (ether)	8.17	−0.433 (ar), −1.581 (al)	1.525
−NH−	12.08	−0.964 (ar), −1.825 (al)	3.599
−S−	16.82	0.11 (ar)	7.69
═CH−	13.71	0.37 (ar)	4.352
−F	8.35	0.399 (ar)	0.997
−Cl	17.30	0.061 (ar), 0.061 (al)	5.967
−Br	22.11	1.131 (ar), 0.270 (al)	8.865
−I	27.19	0.587 (al)	13.900
N (nitro)	6.02	−1.06 (ar)	2.322
$-NH_2$	17.09	−0.854 (ar)	4.43 (5.58 in hydrazine)
$-CH_2-$	16.93	0.530	4.618
$-CH_3$	22.79	0.702	5.718
O (nitro)	9.32	0.491	2.234
═C<	7.56	0.195	3.252
−O− (ester)	8.17	−0.433 (ar)	1.643
═O (ceto)	10.47	−1.898 (al)	2.211
═O (ester)	10.47	−1.054 (al)	2.211
≡CH	18.37	0.452	4.611
≡C	12.23	0.277	3.511 (terminal), 3.72
═N−	9.03	−2.075	3.75 (hydrazine), 3.65 (oxime)

[a] ar, aromatic; al, aliphatic.

for the C_{sp2}; the same C_{sp2} value is used to estimate hydrophobicity for ═N−, taking into account the value for C═N (1.88) resulted from Ph_2C═NH. Finally, the aromatic −OH and $-NH_2$ values were used for the oxime and hydrazine derivatives. Fragmental polarizabilities ($x_{ij\mathrm{P}}$) were estimated exploiting bond and atomic refractivities from reference 13. The H-bond parameters (enthalpy donor factor, $x_{ij\mathrm{D}}$, and enthalpy acceptor factor, $x_{ij\mathrm{A}}$) are those recommended by Raevsky et al.[14] From the 347 molecules listed by Raevsky, we selected the ones in which the proton donor/acceptor group is in a chemical environment as close as possible to the one in the corresponding estrogen derivative from our series. Thus, for ligand 17 (2-chlorestradiol), the value $x_{ij\mathrm{D}} = E_{\mathrm{d}} = -2.20$ was selected for the −OH group, matching that of 2-chlorphenol in reference 14. These fragmental descriptors are listed in Table 2.

Mathematical Engine. For PLS modeling, we used SIMCA P 9.0.[15] To interpret the results, the PLS equation (with $\hat{Y}_i$, the calculated activity) was transformed as a function of the original variables $x_{ij\mu}$ (more exactly, coefficients $a_{j\mu}$), obtaining a relation of the type (μ: V, H, P, D, A, S)

$$\hat{Y}_i = a_0 + \sum_{j=1}^{M}(a_{j\mathrm{V}}x_{ij\mathrm{V}} + a_{j\mathrm{H}}x_{ij\mathrm{H}} + a_{j\mathrm{P}}x_{ij\mathrm{P}} + a_{j\mathrm{A}}x_{ij\mathrm{A}} + a_{j\mathrm{D}}x_{ij\mathrm{D}} + a_{j\mathrm{S}}x_{ij\mathrm{S}}) \quad (2)$$

From the theoretical number of 6M (M is the number of vertexes) descriptors, several are eliminated by the PLS procedure (small variance for the corresponding parameter columns). Other descriptors are nonexistent because of the nature of the atoms occupying the corresponding vertexes (for example, a lack of H-bond character). The corresponding $a_{\mu j}$ "regression" coefficients in eq 2 are zero.

Some restrictive conditions (described also in reference 2), that is, steric misfit is always detrimental and hydrophobicity- and polarizability-based interactions are always beneficial, are defined as below:

$$a_{j\mathrm{V}} < 0; \quad a_{j\mathrm{P}} > 0; \quad a_{j\mathrm{H}} > 0 \quad (3)$$

They were introduced manually; that is, for each constructed model, the parameter columns contradicting the above inequalities were eliminated. This is equivalent with a forced nil value for the corresponding $a_{j\mu}$ coefficients.

In the results analysis phase, only the "regression" coefficients $a_{j\mathrm{S}}$, $a_{j\mathrm{V}}$, and so forth [Coeff(A) in SIMCA, A being the number of significant principal components for the model] that presented VIP (variable influence on projection) values greater than 0.75 were considered as significant.[16]

Analysis of X-ray Data. The rat (*Ratus norvegicus*) uterus ERα, used in the RBA assay, has a 97% homology in the ligand-binding domain (LBD) to the human ERα, as shown in the application of BLAST[17] to the two proteins. None of the eight differing (although similar in polarities) amino acids from the two aligned LBD sequences are in direct contact with the ligands in the binding pocket. Therefore, we used the X-ray structure of the human ERα LBD complexed with 17β-estradiol (PDB access code 1ere).[4,5] CHIME and NCBF (Noncovalent Bond Finder) were used for structure visualization.[18] To determine the key ligand−receptor interactions, we superimposed the steroid skeleton of each ligand in the series, in the MTD-PLS-selected conformation, to the 17β-estradiol structure from 1ere inside the protein, and then, we deleted 17β-estradiol. Thus, we investigated the binding position proposed by MTD-PLS for almost each ligand, without any assumptions regarding protein flexibility. However, for the 17α-(*E*)- and 17α-(*Z*)-CH═CHI−E2 isomers, after replacing estradiol from the ligand binding pocket, a relaxation was allowed (in HyperChem) for the ligand and the amino acid residues in direct contact.

RESULTS AND DISCUSSIONS

PLS Results. Two approaches were used in PLS modeling. First, we used all ligands and $K = 99$ descriptors (K is the number of variables used in PLS), model M4. Applying distance-to-model criteria (elimination of the outliers laying too far from the model),[19] we derived the statistically sound model M47, with 28 compounds and 45 descriptors. Second, we started from M16 ($N = 39$, $K = 78$), in which all single-molecule/single-vertex occurrences were eliminated. The

Table 3. Statistical Characteristics of Four of the Deduced MTD-PLS Models[a]

model	R^2_X(CUM)	R^2_Y(CUM)	Q^2 (CUM)	A	eliminated compounds
M4	0.056	0.808	0.369	1	
M47	0.228	0.931	0.684	2	3, 4, 7, 10, 19, 24, 26−30, 33, 35, 36, 43−45
M16	0.069	0.807	0.450	1	3, 10, 15, 19, 24, 36
M44	0.414	0.942	0.757	2	2−10, 15, 19, 24, 26−36, 43−45

[a] R^2_X (CUM) and R^2_Y (CUM) are the cumulative sum of squares of all the X and Y values, respectively, explained by all extracted principal components; $Q2$ (CUM) is the fraction of the total variation of the Y values that can be predicted for all the A extracted principal components.

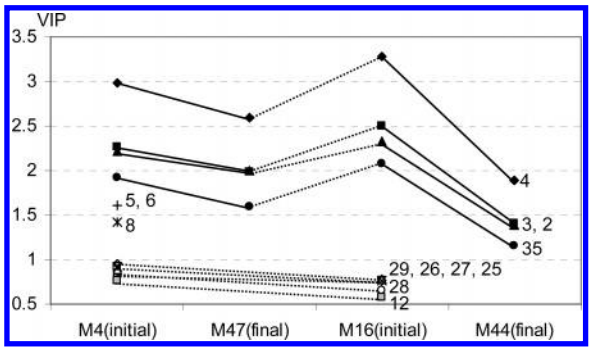

Figure 2. VIP values corresponding to significant x_{jV} (volume) contributions belonging to the four analyzed models: the labels correspond to the j vertexes. All the coefficients are negative (see eq 3), thus the influence on RBA is detrimental.

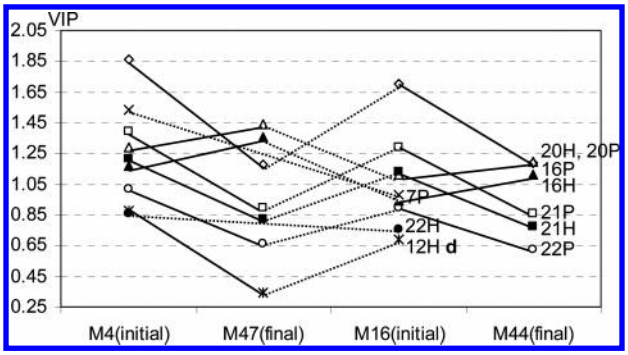

Figure 3. VIP values corresponding to significant x_{jH} (hydrophobicity) and x_{jP} (polarization) contributions belonging to the four analyzed models: the labels correspond to the j vertexes and the nature of the parameter. All the coefficients are positive (see eq 3), thus the influence on RBA is beneficial, except x_{12H}, which presents negative numeric values (noted with d, meaning detrimental).

same procedures yielded M44 ($N = 19$, $K = 27$), which was statistically the best model. Statistical characteristics of the four models are listed in Table 3. The restrictive conditions from eq 3 were applied to all models.

In the analysis for the above-mentioned models, we used the VIP values and the sign for the $a_{j\mu}$ coefficient. The VIP values are proportional to the influence on $\hat{Y}$ of every descriptor (column) $x_{j\mu}$, and we consider only those greater than 0.75 as important.[16] The contribution (enhancing or depleting) of a variable to RBA can be inferred by searching the sign of the $a_{\mu j}$ coefficients [Coeff(A) in SIMCA] together with the positive or negative value of $x_{ij\mu}$ or the mean of the $x_{j\mu}$ column (see eq 2). The principal results are presented in Figures 2−4.

The significant VIP values resulting from models M4, M47, M16, and M44 are presented in Figure 2 for the parameter volume (μ = V). Because the influence on RBA is always detrimental (the conditions from eq 3 force negative a_{jV} coefficients), it results very clearly from the models that

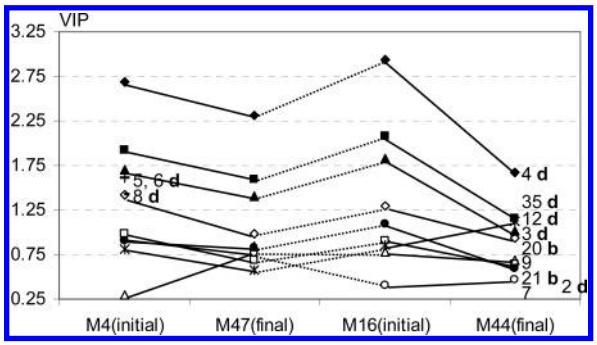

Figure 4. The VIP values corresponding to significant x_{jS} (charge) contributions for the four analyzed models: the labels correspond to the j vertexes and to RBA enhancing (b) or debilitating (d) effects.

any substituent in position 2 (the classical steroid skeleton numbering is adopted) will experience strong steric hindrance, diminishing the binding capacity of the corresponding ligand (see vertexes $j = 2-4$, 35, and also Figure 1). The presence here of vertexes $j = 5$ and 6 for the first model (M4) is in accordance with this conclusion. Although the compound elimination procedure used during the final model construction led to the disappearance of vertexes $j = 25-30$, one might conclude from the initial models M4 and M16 that the MTD-PLS results confirm the binding capacity depleting effect of the long chains in the 17α-C≡CR group (the appearance of steric interference with the receptor wall).

In Figure 3, the VIP values greater than 0.75 for the hydrophobicity (μ = H) and polarizability (μ = P) parameters are outlined. The corresponding $a_{j\mu}$ coefficients are positive, meaning that the atoms in vertexes with high VIP values are exposed to attractive, beneficial interactions. Exceptions are the negative x_{ijH} values, that is, hydrophilic groups, for which $a_{jH} > 0$ means the apparition of detrimental effects. The facts from the figure suggest that the binding ability of compounds containing 17α-CH═CHX substituents occupying vertexes $j = 16$ and 20−22 (compounds 37−42) is enhanced by nonpolar interactions with the neighbor amino acids from the protein. The detrimental behavior of hydrophilic vertex $j = 12$ will be discussed later.

From Figure 4, we can appreciate the importance of charge interactions between the ligands and receptor. The nature of these (b = beneficial, d = detrimental) was established by searching the sign of a_{jS} and x_{ijS}. Electrostatic repulsions are dominant for substituents in position 2 (vertexes $j = 2$, 3, 4, and 35, and also 5 and 6 for M4). On the other hand, attractive electrostatic interactions can be inferred for the atoms $j = 20$ (vinyl C) and 21 (halogen) from the 17α-CH═CHX substituent (see Figure 1). For the substituents in steroid position 4 ($j = 9$, compounds 21−24), all the a_{9S} values are negative. This means repulsive (detrimental) interactions for $x_{i9S} > 0$ (nitro-N and CH_3) and attractive interactions for $x_{i9S} < 0$ (F, Cl). Coefficient a_{7S} is only marginally significant.

Table 4. Compounds Used in Model Validation, with Actual and Predicted Activity Values; the Vertex Occupation Is Also Specified

number	compound	Y_i = log RBA (rat)	predicted Y_i M4	M47	M16	M44	occupied vertexes (j)
52	2-OCH_3−E2	−1.30	−0.10	−0.46	−0.41	−0.42	2, 3, 7, 12
53	2-OH−E2	1.47	0.96	0.80	0.90	0.58	2, 7, 12
54	3-deoxy−E2	−0.30	0.06	1.69	0.38	3.19	12
55	17α-C≡CCH_2F−E2	1.92	1.23	1.80	1.26	2.00	7, 12, 16, 23, 24, 25
56	17α-CH═CHMe−E2	2.00	2.46	2.13	2.46	2.25	7, 12, 16, 20, 21
57[a]	17α-C≡CCH_2Cl−E2	1.30	1.14	1.82	1.19	2.12	7, 12, 16, 23, 24, 25
58[a]	17α-CH═CH_2−E2	1.82	1.88	1.98	1.91	2.14	7, 12, 16, 20
59[a]	4-Br−E2	1.00	0.11	0.10	0.11	0.11	7, 9, 12
60[a]	4-I−E2	−1.07	−0.97	−1.38	−1.10	−1.38	7, 9, 12
61[a]	11β-CH_2Cl−E2	2.08	1.30	1.43	1.39	1.43	7, 12, 31, 32

[a] log RBA values for rat estrogen receptor were calculated using the regression equation rat−lamb ($Y_{rat} = 1.0374Y_{lamb} - 0.0355$; $n = 12$; $R^2 = 0.9477$) as described in text.

Because compounds 3 and 4, with extremely low activity, represent the highest variations in the x_{7S} column values, the sign of the coefficient oscillates as a function of the presence or absence of these compounds ($a_{7S} > 0$ for the initial and $a_{7S} < 0$ for the final models).

In our data, the most dramatic variation in the hydrogen-bond (HB) donor or acceptor character is the one given by the absence or presence of such substituents. Besides, reference 14 did not ascribe HB acceptor capacity to the phenolic OH (to vertex $j = 7$, which corresponds to $x_{i7A} = 0$, except for compound $i = 3$). Several ligands lacking the 17β-OH group ($j = 12$, compounds 2, 5, and 6) do not have significantly reduced biological activity. In these conditions, the PLS engine encounters difficulties in interpreting the input data. For substitution position 2 ($j = 2$), a_{2A} is highly significant (VIP values between 2.07 and 1.19 for the four models), but a_{7A} and a_{12A} are important only in the initial models M4 and M16. The analyzed models attribute activity-depleting effects to HB acceptor groups in all these positions. For the HB donors, only a_{7D} and a_{12D} are meaningful, and just in the initial models. Coefficient a_{7D} is negative for M4 and M16 and produces an activity enhancement for better donor groups (x_{ijD} are negative numbers); in the same models, $a_{12D} > 0$, meaning a detrimental effect exists for HB donors in this position. Our findings are not entirely in agreement with the hydrogen-bonding requirements proposed by Brzozowski et al.[20] for both substitution positions 3 and 17β. It is probable that this odd behavior of our models for vertex $j = 12$ also causes the presence of $a_{12H} > 0$ (this time with an activity decreasing effect) for the hydrophilic groups in substituent position 17β (see discussion of Figure 3).

Comparison with X-ray Crystallography. Ligand atoms occupying vertexes $j = 2-6$ and 35 are in a sterically crowded region of the receptor site, formed by amino acids Leu346, Leu349, Ala350, and Glu353. The 2-substituent−receptor atomic interdistances include short contacts between 2.19 and 2.47 Å. This is in accordance with the highly important unfavorable steric interactions detected by our method (see Figure 2 and its interpretation). The environment contains both polar (Glu353 OE2, and Ala350 backbone N) and nonpolar (Leu and Ala side chains) receptor atoms, explaining the detrimental charge interactions found in the analysis of Figure 4 and the lack of significant a_{jH} coefficients for this region. The absence of a neighboring, favorable protein HB donor group justifies the activity depleting effect of an HB acceptor group in $j = 2$ (see above).

Although in the final models, vertexes $j = 25-29$ are absent because of compound elimination and the lost structural information, their detrimental steric effect (Figure 2) as modeled in M4 and M16 can be justified by the protrusion of the long tail of substituent 17α-C≡$C(CH_2)_5$-CH_2I into the receptor cavity wall. This consists (including also the backbone atoms) of Met343, Met421, Val418, Glu339, and Cys417 for the all trans and Met342 and Ser338 for another low-energy, more bent chain conformation analyzed by us. Also, we note that for vertexes $j = 17-19$, which lie in the same region (compounds 26−28), the initial models give $a_{jV} < 0$ coefficients with VIP values between 0.72 and 0.43 (not represented in Figure 2), perceiving the same steric hindrance.

The 17α-CH═CHX substituents of ligands 37−42 are stuck in binding pockets containing the Met343 and Met528 side chains and the His524 ring edge for the *Z* isomers and the Met343 and Met421 side chains for the *E* isomers. The predominantly hydrophobic and polarizable receptor atoms are in favorable attractive interactions, mostly at van der Waals contact distances, with the halogens (I, Br, Cl), and other adjacent ligand atoms, explaining the large number of significant a_{jH} and a_{jP} coefficients in Figure 3 for $j = 16$ and 20−22. The corresponding neighbor ligand−receptor atomic charges justify the attractive electrostatic interactions for the same vertexes as represented in Figure 4.

In the end, the nature of the electrostatic interactions for compounds occupying vertex $j = 9$ (see the discussions for Figure 4 for substitution position 4) can be explained by the vicinity of these substituents with a positively charged guanidine side chain for Arg394.

Model Validation. To test the predictive ability of our models, we have explored the possibilities of identifying compounds with a similar vertex occupancy and known RBA values for the rat ER. We found only five compounds matching the above criteria, and that were unused in our training set (Table 4, compounds 52−56).[21] We also added compounds 57−61, for which the RBA values for lamb ER were known.[7] The unknown rat RBA values for these were calculated using a simple regression equation between the known lamb and rat values for compounds 16, 22, 30−32, 40, and 46−51 (see Table 1).

The predictive abilities of our models are tested using the externally predicted R^2 values (R^2_{pred}) and the Golbraikh−Tropsha set of criteria.[22] The results are presented in Table 5. R^2_{pred} values are uniformly high, if compound 54, 3-deoxy-

Table 5. Predictive Power Results for the External Test Set[a]

model	R^2_{pred}	R^2	R_0^2	$R'_0{}^2$	k	k'
M4	0.745	0.754	0.683	0.560	1.09	0.765
M4 without 54	0.725	0.737	0.655	0.534	1.1	0.769
M47	0.557	0.589	0.249	0.26	0.875	0.848
M47 without 54	0.798	0.807	0.229	0.5	1.01	0.874
M16	0.773	0.779	0.657	0.594	1.07	0.795
M16 without 54	0.781	0.797	0.643	0.614	1.09	0.803
M44	−0.049	0.309	−0.062	−0.142	0.607	0.870
M44 without 54	0.728	0.750	−0.115	0.282	0.922	0.915

[a] Criteria of Golbraikh and Tropsha[22] are used. The significance of R^2, R_0^2, $R'_0{}^2$, k, and k' are the same as that in ref 22; R^2_{pred} is the external R^2 from the same reference.

estradiol (only vertex 12 occupied), is eliminated. The poor predictivity for this derivative is not unexpected, considering that in our training set, there is no information concerning the lack of an OH substituent in position 3 (vertex 7). The $R^2 > 0.6$ condition is fulfilled by the models without 54. In the models with higher information content (M4 and M16), at least one R_0^2 is close enough to R^2. The slope values, k and k', in the majority of the models, are between 0.85 and 1.15. Thus, our models are acceptable from the point of view of the Golbraikh−Tropsha criteria.

DISCUSSIONS

Models M47 and M44 are significant according to cross-validation, whereas models M4 and M16 are marginally relevant. In exchange, the external validation confirms the value of the initial models. Since some molecules were eliminated from M16, compared to M4, we can assume that the structural information provided by the "missing" molecules is lost. Still, our comparative analysis with the X-ray data reveals that the statistically unsound "early" PLS models already contain much pertinent data. Thus, these models can be capitalized in the quantitative structure−activity relationship (QSAR) analysis, using especially those coefficients with sufficiently high VIP values. On the other hand, if the interpretations of the QSAR coefficients in the consecutive models present contradictions, a careful analysis is necessary concerning the motives.

Our data indicate that the PLS engine works well, and the results depend only on the quality of the input data. When insufficient or contradictory information is introduced, the resulting models become unstable or simply wrong. This was the case with 17β substitution: small variability in the substituents, combined with high activity for the OH-lacking derivatives. But if the input information is consistent, the model's answer is correct: the high activities of the 17α vinyl-halogen derivatives are very appealingly explained by the favorable hydrophobic and polarizability interactions.

CONCLUSIONS

Chance correlation considerations of Topliss and Costello,[23] that is, one cannot obtain more pertinent information out of a QSAR than the input information, certainly applies for modern 3D QSARs.

Our MTD-PLS procedure with the restrictive conditions concerning the detrimental effects of the volume parameter, and the enhancing effect of the hydrophobic and polarizability descriptor, which introduce supplementary information, certainly allows us to obtain interesting suggestions concerning the nature of ligand−receptor interactions. Good agreements between the QSAR coefficient interpretation and the ligand−receptor interaction can be expected either for MTD-PLS models, which are sound from a statistical point of view, or in weaker models for positions (vertexes) where a sufficient variety of substitution pattern is present. An important conclusion is the one concerning the possibility of utilization of the statistically weaker models, but containing more information, by taking into account the highly important coefficients (high VIP values) of the QSAR equation of these models. Also, the predictive ability of these initial models can be acceptably high.

Thus, in QSAR studies for series without the background of X-ray data results for the ligand−receptor complex, reliable data on the nature of the binding site can be obtained in some conditions. These are either (i) a sound PLS model or, at least, for (ii) vertexes with a high degree of variability in substitution. Even for cases with sufficient X-ray data, our MTD-PLS method could produce information concerning the rigidity or deformability of different regions of the binding site receptor cavity. Such data do not result directly from the X-ray crystallography.

ACKNOWLEDGMENT

We thank Dr. Erik Johansson (Umetrics, Sweden) for kindly providing the SIMCA program package and Dr. M. Mracec for access to the HYPERCHEM package.

Supporting Information Available: Calculated VIP (variable influence on projection) values and $a_{j\mu}$ coefficients. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Simon, Z.; Chiriac, A.; Holban, S.; Ciubotaru, D.; Mihalas, G. I. *Minimum Steric Difference*. *The MTD Method for QSAR Studies*; Research Studies Press Ltd.: Letchworth, U. K., 1984.

(2) Oprea, T. I.; Kurunczi, L.; Olah, M.; Simon, Z. MTD-PLS: A PLS-based variant of the MTD method. A 3-D QSAR analysis of receptor affinities for a series of halogenated dibenzoxin and biphenyl derivatives. *SAR QSAR Environ. Res.* **2001**, *12*, 75−92.

(3) Kurunczi, L.; Olah, M.; Oprea, T. I.; Bologa, C.; Simon, Z. MTD-PLS: A PLS-based Variant of the MTD Method. II. Mapping Ligand−Receptor Interactions. Enzymatic Acetic Acid Esters Hydrolysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 841−846.

(4) OCA Advanced Search website. http://oca.ebi.ac.uk/oca-bin/ocamain. Analysis of PDB entry 1ere.

(5) PDB Structure Explorer. http://www.rcsb.org/pdb/cgi/explore.cgi?pid=52851102327638&pdbId=1ERE. Brzozowski, A. M.; Pike, A. C.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; Engstrom, O.; Ohman, L.; Greene, G. L.; Gustafsson, J. A.; Carlquist, M. Human Estrogen Receptor Ligand-Binding Domain in Complex With 17beta-Estradiol.

(6) Wold, S.; Albano, C.; Dunn, W. J.; Edlund, U.; Esbensen, K.; Geladi, P.; Hellberg, S.; Johansson, E.; Lindberg, W.; Sjöström, M. Multivariate Data Analysis in Chemistry. In *Chemometrics: Mathematics and Statistics in Chemistry*; Kowalski, B. R., Ed.; D. Reidel: Dordrecht, The Netherlands, 1984. Wold, S.; Sjöström, M.; Eriksson, L. PLS in Chemistry. In *The Encyclopedia of Computational Chemistry*; Schleyer, P. V. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., III., Schreiner, P. R., Eds.; John Wiley and Sons: Chichester, U. K., 1999; pp 2006−2020.

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

(8) *HyperChem 5.11 Pro*; *ChemPlus 1.6*; HyperCube Inc.: Gainesville, FL. http://www.hyper.com.

(9) Ciubotariu, D.; Deretey, E.; Oprea, T. I.; Sulea, T.; Simon, Z.; Kurunczi, L.; Chiriac, A. Multiconformational Minimal Steric Dif-

ference. Structure−Acetylcholinesterase Hydrolysis Rates Relations for Acetic Acid Esters. *Quant. Struct.-Act. Relat.* **1993**, *12*, 367−372.
(10) Kotani, T.; Higashiura, K. Comparative Molecular Active Site Analysis (COMASA). 1. An Approach to Rapid Evaluation of 3D QSAR. *J. Med. Chem.* **2004**, *47*, 2732−2742.
(11) Olah, M. Molecular fragment volume calculation for QSAR studies. *Rev. Roum. Chim.* **2000**, *45* (12), 1123−1125.
(12) Rekker, R. *The Hydrophobic Fragmental Constant*; Elsevier: Amsterdam, 1977.
(13) Volkenstein, M. V. *Stroenie i fizicheskie svoistva molecul*; Izd. Akad. Nauk: Moscow, 1955; pp 230, 296.
(14) Raevsky, O. A.; Grigor'ev, V. Ju.; Kireev, D. B.; Zefirov, N. S. Complete Thermodynamic Description of H−Bonding in the Framework of Multiplicative Approach. *Quant. Struct.-Act. Relat*. **1992**, *11*, 49−63.
(15) SIMCA P, version 9.0; Umetrics AB: Umeå, Sweden. http://www.umetrics.com.
(16) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and Megavariate Data Analysis. Principles and Applications.* Umetrics AB: Umeå, Sweden, 2001; p 94−97.
(17) Basic Local Alignement Search Tool. http://au.expasy.org/tools/blast/.
(18) *MDL Chime*, version 2.6 SP4; MDL Information Systems, Inc.: San Leandro, CA (http://www.mdlchime.com/chime). Martz, E.; *Noncovalent Bond Finder (NCBF)*, 2002 (http://www.umass.edu/microbio/chime/find-ncb/).
(19) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and Megavariate Data Analysis. Principles and Applications.* Umetrics AB: Umeå, Sweden, 2001; pp 101−103, 507−509.
(20) Brzozowski, A. M.; Pike, A. C. W.; Dauter, Z.; Hubbard, R. E.; Bonn, T.; Engstrom, O.; Öhman, L.; Greene, G. L.; Gustafsson, J. Å.; Carlquist, M. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature* **1997**, *389*, 753−758.
(21) Endocrine disruptor knowledge base. http://edkb.fda.gov/.
(22) Golbraikh, A.; Tropsha, A. Beware of q^2! *J. Mol. Graph. Mod.* **2002**, *20*, 269−276.
(23) Topliss, J. G.; Costello, R. J. Chance correlations in structure−activity studies using multiple regression analysis. *J. Med. Chem.* **1972**, *15*, 1066−1069.

CI050077C