Journal of Computer-Aided Molecular Design, 11 (1997) 79–92
ESCOM

J-CAMD 371

MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids

Gianpaolo Bravi[a,*], Emanuela Gancia[a], Paolo Mascagni[a], Monica Pegna[a], Roberto Todeschini[b] and Andrea Zaliani[a]

[a]*Italfarmaco Research Centre, via Lavoratori 54, I-20092 Cinisello Balsamo (Milan), Italy*
[b]*Department of Environmental Sciences, University of Milan, via Emanueli 15, I-20126 Milan, Italy*

Received 25 April 1996
Accepted 17 July 1996

Keywords: CoMFA; Holistic description; Connolly surface; PCA; Experimental design; PLS

Summary

The recently proposed WHIM (Weighted Holistic Invariant Molecular) approach [Todeschini, R., Lasagni, M. and Marengo, E., J. Chemometrics, 8 (1994) 263] has been applied to molecular surfaces to derive new 3D theoretical descriptors, called MS-WHIM. To test their reliability, a 3D QSAR study has been performed on a series of steroids, comparing the MS-WHIM description to both the original WHIM indices and CoMFA fields. The analysis of the statistical models obtained shows that MS-WHIM descriptors provide meaningful quantitative structure–activity correlations. Thus, the results obtained agree well with those achieved using CoMFA fields. The concise number of indices, the ease of their calculation and their invariance to the coordinate system make MS-WHIM an attractive tool for 3D QSAR studies.

Introduction

The Comparative Molecular Field Analysis (CoMFA) [1] approach is one of the most widely used techniques for 3D QSAR studies. The possibility to merge structurally heterogeneous compounds, the accurate description in terms of steric and electrostatic fields and the easy interpretability of the statistical results are undoubtedly the main reasons for its success. On the other hand, the large number of descriptor variables to handle and the strict dependence of statistical results on molecular alignment are the major drawbacks of CoMFA. Holistic descriptors could in principle overcome these problems, as they allow to condense 3D chemical information into a brief numerical vector, describing each molecular structure per se (i.e., its 3D orientation with respect to any reference system does not have to be considered). Examples of this type of description include topological indices [2], autocorrelation function-based indices [3,4] and the recently proposed Weighted Holistic Invariant Molecular (WHIM) indices [5–7]. In particular, the latter consist of 12 statistical parameters, calculated from the x,y,z coordinates of a molecule within different weighting schemes, and contain information about the whole molecular structure in terms of size, shape, symmetry and atom distribution. These indices were successfully correlated to the toxicity of heterogeneous organic molecules [7] and to molecular properties such as total accessible surface area [5], log P, and boiling and melting points [6]. Although the results were shown to be superior to those obtained with traditional QSAR methods, WHIM indices have not been used to solve structure–activity problems involving highly specific biological interactions.

Although the WHIM approach efficiently vectorizes the atomic x,y,z coordinates of a given molecular structure and their related physicochemical features, it is in principle applicable to any set of coordinates weighted by any kind of property. Thus, for 3D QSAR purposes, a new approach (G-WHIM, Grid-Weighted Holistic Invariant Molecular descriptors) has been recently proposed,

*To whom correspondence should be addressed.

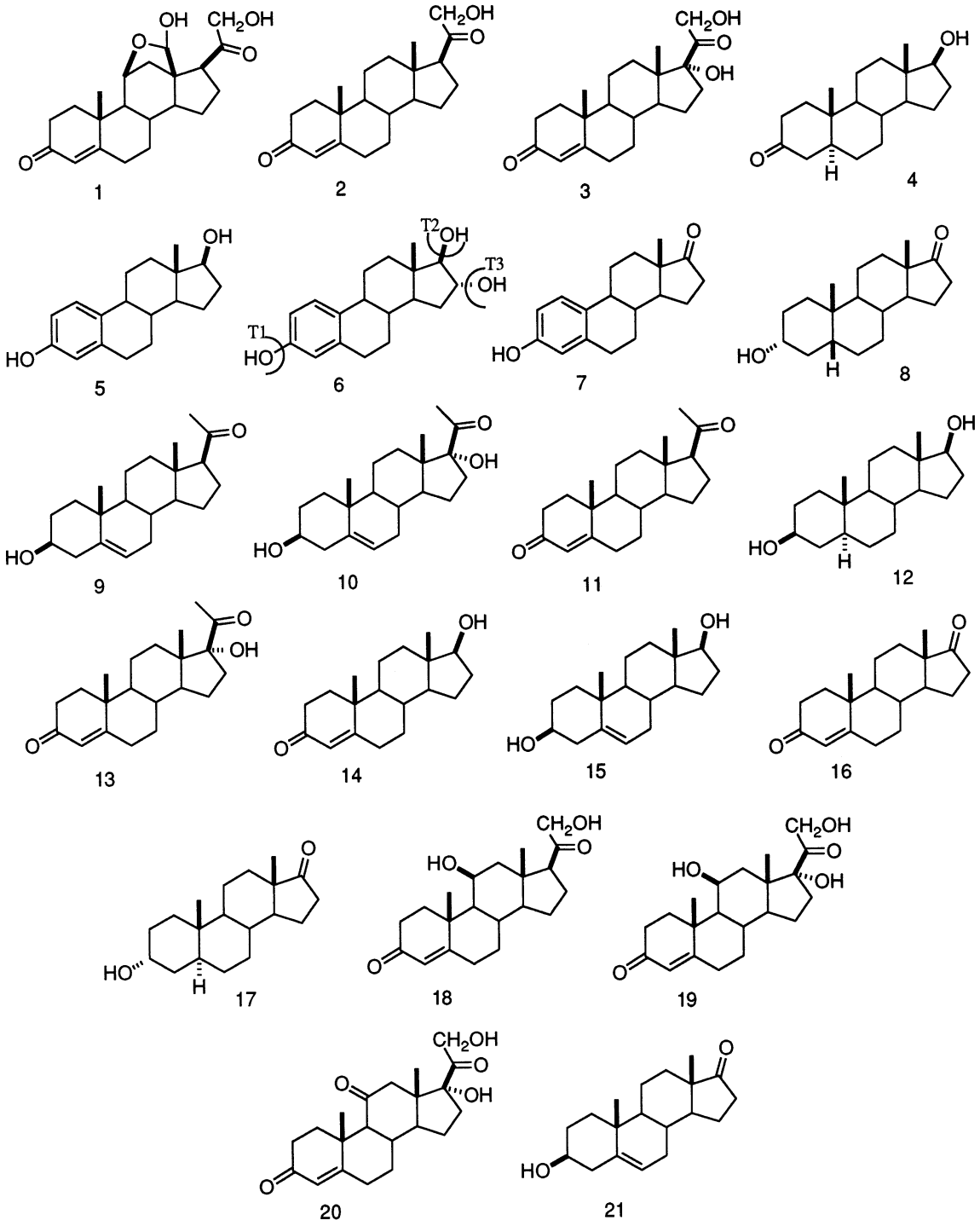

Fig. 1. Structures of the 21 steroids contained in the training set. T1, T2 and T3 represent the side-chain torsional angles that were rotated for estriol (compound 6).

based on scalar interaction fields computed on a grid of points where each molecule is embedded [8].

In this paper we propose new WHIM-based indices named MS-WHIM, which are derived directly from Molecular Surface (MS) properties. They were developed in an attempt to consider the contribution arising from molecular surface recognition in specific ligand–receptor interactions. To test the validity of this new approach, MS-WHIM was applied to a well-known biological problem and the results were compared with those from the CoMFA description and the original WHIM description. The steroid set previously analyzed by Cramer et al. [9] in the first publication on CoMFA, and subsequently used to propose other 3D QSAR methods such as MTD [10],

TABLE 1
BINDING AFFINITIES OF 21 STEROIDS TO HUMAN CORTICOSTEROID-BINDING GLOBULIN[a]

Molecule	CBG	CoMFA_FFD	WHIM_FD	MS-WHIM_FFD
1 Aldosterone	6.279	♦	♦	♦
2 Deoxycorticosterone	7.653	–	–	–
3 Deoxycortisol	7.881	♦	–	♦
4 Dihydrotestosterone	5.919	♦	–	–
5 Estradiol	5.000	–	–	♦
6 Estriol	5.000	♦	♦	♦
7 Estrone	5.000	♦	–	♦
8 Etiocholanolone	5.255	♦	♦	♦
9 Pregnenolone	5.255	♦	–	–
10 17-Hydroxypregnenolone	5.000	♦	♦	–
11 Progesterone	7.380	♦	–	♦
12 Androstanediol	5.000	–	–	♦
13 17-Hydroxyprogesterone	7.740	–	–	–
14 Testosterone	6.724	♦	–	♦
15 Androstenediol	5.000	♦	–	–
16 Androstendione	5.763	♦	♦	–
17 Androsterone	5.613	♦	♦	♦
18 Corticosterone	7.881	♦	–	–
19 Cortisol	7.881	–	♦	♦
20 Cortisone	6.892	♦	–	♦
21 Dehydroepiandrosterone	5.000	♦	–	♦

The symbol ♦ indicates the compounds selected by means of experimental design strategies within each description matrix (see the text for a more detailed explanation).
[a] Affinity data (log 1/k) from Ref. 14.

similarity matrices [11,12] and COMPASS [13], was employed. The training set comprises 21 molecules (Fig. 1), assayed for binding affinity (Table 1) [14] to corticosteroid-binding globulin (CBG). New CoMFA fields were calculated adopting, with respect to the original work, different criteria in selecting and aligning the starting geometries. The selected steroid structures were then used to compute WHIM and MS-WHIM indices. Each de-

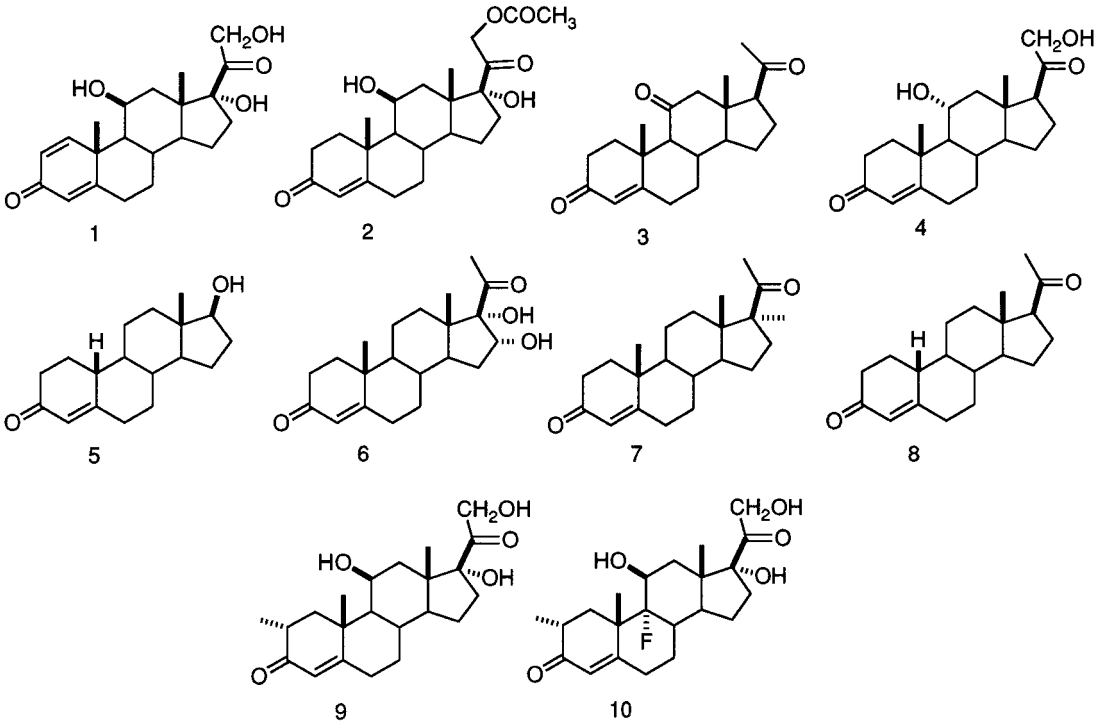

Fig. 2. Structures of the 10 steroids used as the test set.

scription matrix was analyzed by means of a chemometric strategy involving Principal Component Analysis (PCA) [15], molecule selection by means of design criteria [16,17] and Partial Least Squares (PLS) [18] regressions. Finally, the predictive capability of each derived statistical model was verified on an additional test set of 10 compounds (Fig. 2) [19].

WHIM from molecular surface (MS-WHIM)

Structure–activity data correlations require an accurate molecular description. Although atomic coordinates are themselves a sort of basic description, biological recognition is an event that occurs at the molecular surface level; thus, it is at the solvent-accessible surface level that the key forces (van der Waals interactions, hydrophobic effects and electrostatics) driving the binding process play their critical roles [20].

Based upon this consideration we developed new WHIM-based indices, computed from the x,y,z coordinates of Connolly surface points [21] instead of the atomic coordinates. The mathematical algorithm is common to both approaches, differing only in the starting coordinates and in the weighting schemes adopted.

WHIM descriptors

WHIM indices [5–7] are aimed at extracting and quantifying the information contained in the x,y,z atomic coordinates of a molecule. Two types of matrices are initially defined: a molecular matrix $M(n \times 3)$ containing the x,y,z coordinates of the n atoms, and four diagonal matrices $W(n \times n)$ containing the 'weights'. The latter are physicochemical properties associated with the n atoms of the molecule (see below). The following procedure [5] is then applied to each molecular conformation within each W matrix:

(1) The atomic x,y,z coordinates are centered with respect to their weighted mean.

(2) Weighted PCA is performed on the centered data to obtain the score matrix T in the three principal component axes.

(3) The following weighted statistical parameters are computed from the T matrix ($i = 1, \ldots, n$; $m = 1,2,3$): (i) variance (λ_m) = PCA eigenvalues; (ii) the eigenvalue proportion:

$$\theta_m = \lambda_m / \sum_m \lambda_m$$

(iii) the skewness:

$$\gamma_m = | [\sum_i (w_i t_{im}^3) / \sum_i w_i] | * 1/ \lambda_m^{3/2}$$

and (iv) the kurtosis:

$$\kappa_m = [\sum_i (w_i t_{im}^4) / \sum_i w_i] * 1/ \lambda_m^2$$

The PCA eigenvalues are correlated with the molecular size, as they refer to the coordinate extension. Eigenvalue proportions are easily related to molecular shape, as planar molecules will have only two components. Instead of using θ_3, the acentric factor $\omega = \theta_1 - \theta_3$ is used [6]; spherical molecules have a null acentric factor, while linear ones will have $\omega = \theta_1 = 1$. Skewness represents the molecular symmetry along each component. Since this is a third-order moment, it can assume negative values; to preserve the invariance to rotation, the absolute value is taken into account. The fourth-order moment, kurtosis, is related to the atomic distribution and the density around the center and along principal axes. To avoid problems related to infinite κ_3 values obtained when dealing with planar compounds, the reciprocal of this entity, $\eta_m = 1/\kappa_m$, was defined [7]; η_m may be viewed as the unfilled space per atom. A total of 12 WHIM indices are thus computed for each weight, i.e., $\lambda 1$, $\lambda 2$, $\lambda 3$, $\theta 1$, $\theta 2$, ω, $\gamma 1$, $\gamma 2$, $\gamma 3$, $\eta 1$, $\eta 2$ and $\eta 3$.

In relation to the kind of weights assigned to the atoms, different types of information can be obtained. From the original papers, the first weighting scheme applied is represented by the unitary case (i.e. $w_{ii} = 1$; $i = 1, \ldots, n$), where purely geometrical information can be extracted because different atom types are not distinguished. To achieve a physicochemical description, the following properties were introduced: (i) atomic mass; (ii) van der Waals atomic volume; and (iii) Mulliken atomic electronegativity. The information obtained within these schemes may be referred to (i) the mass distribution (in this case the three principal axes coincide with the directions of the inertia principal axes); (ii) the volume distribution; and, to a certain extent, (iii) the charge distribution. Scaling of the weights onto the carbon values is applied to assure comparable numbers for all the schemes.

Apart from the capability to condense 3D chemical information in a brief numerical vector, the WHIM approach appears a new promising tool for 3D QSAR studies as it provides a molecular description which is invariant to the roto-translation. Thus, step 1 (i.e., centering the atomic coordinates) assures invariance to translation, while step 2 (i.e., PCA) assures invariance to rotation. Consequently, the 3D orientation of a molecular structure with respect to either the coordinate system or any other molecule does not affect the WHIM description.

MS-WHIM descriptors

The 12 statistical parameters described above are computed starting from the x,y,z coordinates of Connolly surface points rather than from atomic coordinates and using different weighting schemes (i.e., properties associated with the surface points).

The unitary value and the Molecular Electrostatic Potential (MEP) [22] values computed at each point of

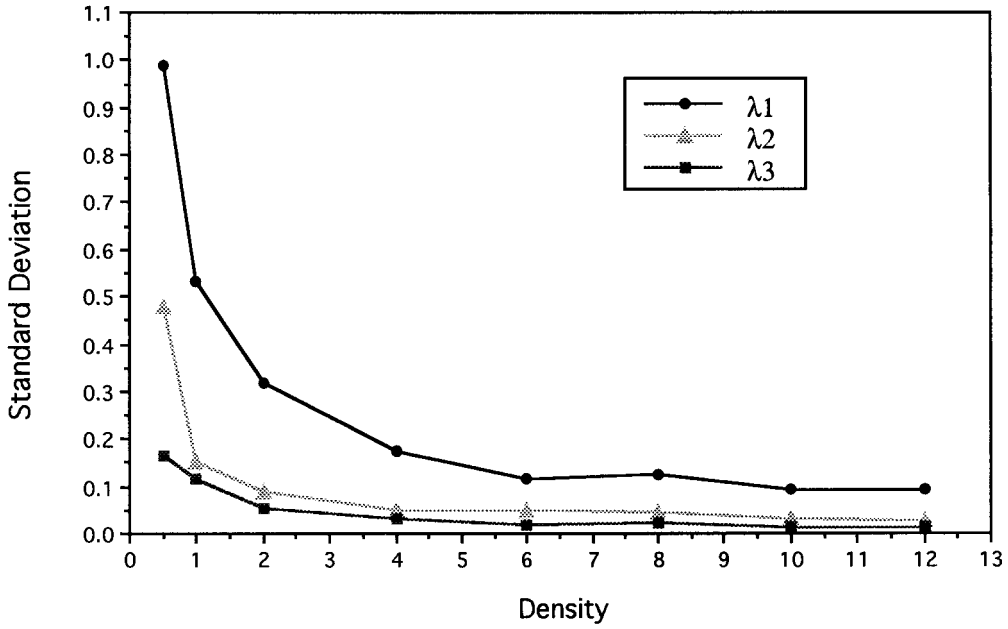

Fig. 3. Standard deviation of PCA eigenvalues (weight 1) evaluated within 20 random 3D orientations of deoxycortisol (compound 3 in Fig. 1) for different density levels (points per Å^2).

the Connolly surface are considered as weights. The unitary scheme maintains the geometrical information relative to the molecular surface shape, while the second weighting scheme provides the electrostatic information about the electron density distribution.

As the electrostatic potential can be either positive or negative, while statistical weights must be semipositive definite, MEP values are separated into two different matrices, one containing only positive MEP values and the other the absolute value of the negative ones, setting to zero the missing points of each matrix. Three definite weighting schemes for a total of 36 (12×3) indices are used.

Although the application of the WHIM approach to coordinates different from the atomic ones could in principle have some advantages (i.e., a reduced number of indices in spite of a larger number of collected points and a more accurate starting description), a new problem occurs. The sampling error inherent when representing a continuous surface as a collection of a finite number of points could affect MS-WHIM values in two ways. First, the indices are sensitive to the surface point density applied. Second, the Connolly algorithm samples points depending on the 3D orientation of the molecule examined. In this way, different spatial positions of the same molecular structure could lead to different surface point distributions and consequently to different MS-WHIM vectors, even if the same density value is applied. (It is worth to point out that the computation of the molecular surface, not the WHIM approach, is sensitive to the molecular orientation!) In principle, increasing the point density on the surface, MS-WHIM indices should converge to their theoretical value and the point distribution dependence on molecular orientation should become lower and lower. To find the minimum density value that allows a good compromise between result stability and computational time, several tests were executed on reference steroid structures by computing MS-WHIM indices for different density values and for different spatial orientations. The standard deviations of λ1, λ2 and λ3 (weight 1), evaluated within 20 different 3D orientations for a given conformation of deoxycortisol (compound 3 in Fig. 1), are represented in Fig. 3 as a function of density (from 0.5 to 12 points per Å^2). The relative positions of the curves are consistent with the calculation method (PCA): the standard deviation decreases on going from λ1 to λ3. The final result is as expected: the curves show common profiles and in all three cases the standard deviation lowers at higher density values. Similar trends were observed by considering all other descriptors and different steroid molecules as well. On the basis of these results, we can assume the MS-WHIM descriptors to be independent from molecular orientation when computed on highly dense surfaces. In the present work all MS-WHIM indices were computed by using a density value of 10 points per Å^2 (i.e., about 2500 points per molecule).

Methods

Computational methods

Conformational search, CoMFA and all statistical analyses were carried out within the SYBYL v. 6.03 molecular modelling package [23]. The cross-validation procedure suggested by the authors of GOLPE [24] and the

selection of the most informative molecules, following factorial and fractional factorial design strategies [16,17], were carried out by means of internally developed SPL (SYBYL Programming Language) macros. Similarity scores and molecular alignments were obtained by means of SEAL [25]. MS [21] was employed to generate Connolly surfaces, and an in-house developed program (written in the C programming language) was used to compute MEP, WHIM and MS-WHIM descriptors. All calculations were performed on a Silicon Graphics Crimson workstation.

Conformational analysis

Molecular models of 21 steroids were taken from the SYBYL CoMFA Tutorial. The additional 10 structures were built starting from the most similar steroid skeleton available and adding the proper functional groups. Each molecule was investigated by systematically varying the side-chain torsional angles on a 30° grid and further minimizing the structures obtained. The standard Tripos force field [26] was used, including the electrostatic contribution with Gasteiger–Marsili [27] partial atomic charges and a distance-dependent dielectric constant. Geometry optimization was carried out by means of the Powell algorithm, until the rms gradient was less than 0.01 kcal/mol Å. For each steroid, all minimum energy conformations within 4 kcal/mol of the global minimum were retained.

Structure selection and molecular alignment

As in the original work by Cramer et al. [9] and because of its highest affinity to CBG, deoxycortisol (compound 3 in Fig. 1) was taken as the lead compound. Furthermore, its global minimum conformer was chosen as the Template Structure (TS). The TS was then compared to the conformational minima of all molecules by means of a SEAL analysis. This allowed to select for each molecule, of both the training set and the test set, the most similar structure (i.e., the one characterized by the best SEAL score) and its optimal alignment to the TS [28]. SEAL allows a rapid pairwise comparison of dissimilar molecules through an alignment function, which comprises a double sum over all the possible atom pairs between two molecules; a similarity score A_F is computed as follows:

$$A_F = -\sum_i \sum_j w_{ij} \exp(-\alpha r_{ij})^2 \quad (1)$$

where r_{ij} is the distance between atom i of the first structure and atom j of the second structure and α is the attenuation range of this distance dependence. The w_{ij} preexponential factor is computed as a function of atomic partial charges and van der Waals radii:

$$w_{ij} = w_E\, q_i\, q_j + w_S\, v_i\, v_j \quad (2)$$

For every SEAL comparison, the steric and electrostatic proportions (w_E and w_S) were set to have the same weight.

CoMFA fields

The data set of selected structures, properly aligned with respect to the TS, was embedded in a regularly spaced (1 Å) grid of dimensions $20 \times 17 \times 17$ Å. Steric and electrostatic field energies were computed by means of the standard Tripos force field and Gasteiger–Marsili partial atomic charges, as in the original work by Cramer et al.; a C(sp^3) probe atom with a charge of +1 and a distance-dependent dielectric constant was used. An energy cutoff of ± 30 kcal/mol was applied for all interactions and the electrostatic contribution was ignored at sterically bad points (DROP YES option).

WHIM and MS-WHIM descriptors

The data set of selected conformers, not aligned but taken in their original 3D orientations, was used to compute WHIM and MS-WHIM descriptions. Moreover, to further highlight the invariance of MS-WHIM indices to roto-translation, they were also computed on nine additional data sets, each obtained by randomly varying the 3D orientation of each molecular structure.

WHIM indices were calculated from atomic x,y,z coordinates within four different weighting schemes: (i) the unitary case; (ii) atomic mass; (iii) van der Waals atomic volume; and (iv) Mulliken atomic electronegativity. A total of 48 molecular descriptors were thus obtained.

MS-WHIM indices were calculated from the Connolly surface points within three weighting schemes: (i) the unitary case; (ii) positive MEP; and (iii) negative MEP values, yielding a total of 36 molecular descriptors. Connolly surfaces were generated using a 1.5 Å radius probe atom and a density of 10 points per $Å^2$. The MEP was computed onto the surface points by means of the classical Coulomb formula, using a distance-dependent dielectric constant (ε):

$$V_p = \sum_i q_i / \varepsilon\, |r_i - p| \quad (3)$$

where V_p is the MEP value relative to point p and r_i the distance between p and the *i*th atom.

Chemometric analysis

CoMFA columns were block-scaled with the CoMFA_STD scaling option, to assure that the total influence of each field on the PLS results was the same. To minimize the influence of noisy columns, all analyses were done with a column filter of 2 kcal/mol, i.e., any region point column having a standard deviation less than 2 kcal/mol was excluded from the PLS analysis. WHIM and MS-WHIM columns were autoscaled to assign unit variance to each descriptor.

The selection of the most informative structures in the steroid series was achieved through Factorial Design (FD) and Fractional Factorial Design (FFD) strategies. A PCA was performed on each data matrix to derive the scores in the n principal components space. The latent variable hyperspace can be divided into 2^n subspaces: one molecule for each subspace (FD) was selected if $n = 3$, while a total of $2^{(n-1)}$ molecules (FFD) were considered if $n > 3$. In the few cases where the subspace was empty, two molecules in the nearest subspace were considered if possible.

The optimum number of components in each PLS final model was determined through two cross-validation [29] procedures: (i) Leave-One-Out (LOO); and (ii) Five Random Groups (5RG). The latter protocol was repeated up to 100 times and the associated parameters represent mean values [24]. The predictive power of each statistical model was evaluated by means of q^2 and s_{PRESS}, computed as follows [30]:

$$q^2 = 1 - \sum(y_{pred} - y_{obs})^2 / \sum(y_{obs} - y_{mean})^2 \quad (4)$$

$$s_{PRESS} = (\sum(y_{pred} - y_{obs})^2 / (n - c - 1))^{1/2} \quad (5)$$

where n is the number of compounds and c is the number of components.

The SDEP [24,30] index was computed to check the quality of the external predictions:

$$SDEP = (\sum(y_{pred} - y_{obs})^2 / n)^{1/2} \quad (6)$$

The reliability of WHIM-based statistical models was further verified by scrambling several times the response variable [1] (i.e., the activities of the training set compounds were mixed so that each value was no longer assigned to the right molecule) and repeating the LOO PLS run.

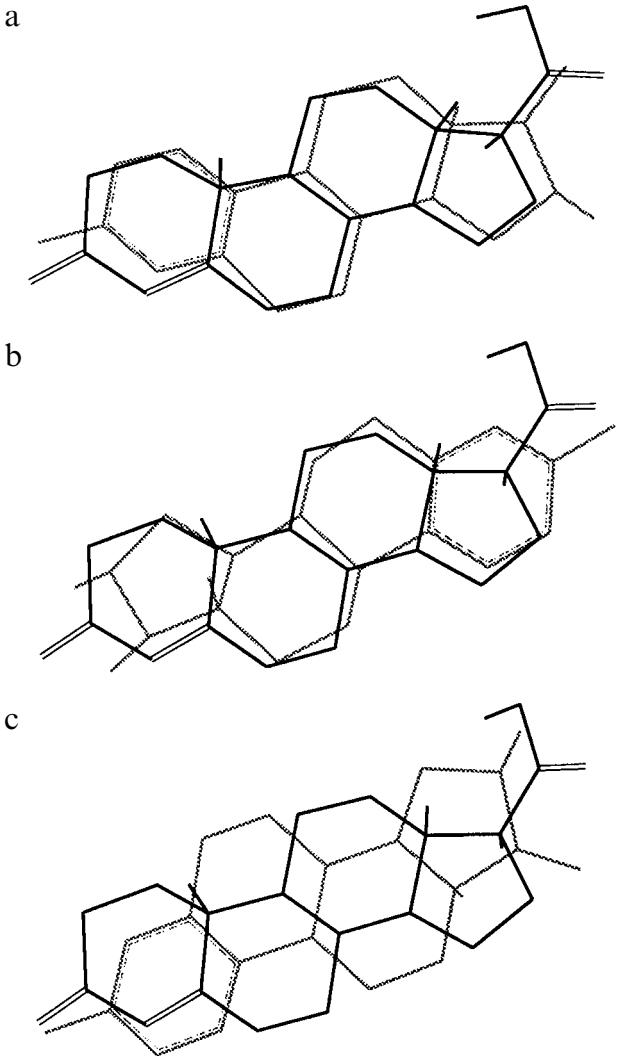

Fig. 4. Three molecular alignments found by SEAL for estriol (grey) versus deoxycortisol (black). (a) Molecular skeletons optimally aligned; (b) the estriol phenyl ring matched to the terminal five-membered ring of deoxycortisol; and (c) the estriol steroid nucleus flipped so that the 19-methyl groups of the matched molecules point toward opposite directions.

Results

Structure selection and molecular alignment

As more and more successful applications have been reported, it has become clear that what is critical in CoMFA is the self-consistency of the molecular conformations and their alignment within the data set. In other words, improved CoMFA models can be obtained if conformations are chosen and aligned so that their electrostatic and steric fields are as similar as possible [1]. In the series analyzed here, the relative rigidity of the steroid nucleus allows the conformational variable to be neglected; however, although the spatial orientation of an OH group does not heavily affect the steric field, the electrostatic field is strictly dependent on this orientation. Thus, in an attempt to maximize the efficiency of CoMFA, which was our main term of comparison, we investigated the torsional space of the 21 + 10 steroids and chose, for each molecule, the conformer having the largest steric and electrostatic similarity to the global minimum of the lead compound (TS). As described in the Methods section, all conformational minima were compared to the TS by means of the SEAL program and the conformer characterized by the best similarity score was selected for each molecule [28]. For all compounds, the best alignment consisted in the optimal steroid skeleton superposition. The only exceptions were the estrane analogues containing an aromatic ring, i.e., estradiol, estriol and estrone (compounds 5–7, Fig. 1). For each of these molecules, SEAL found three different alignments with the TS, which were characterized by close similarity scores. These alignments, illustrated in Fig. 4 for estriol, are: (i) optimally aligned molecular skeletons (Fig. 4a); (ii) the phenyl ring matching the five-membered terminal ring of deoxycortisol

TABLE 2
TORSIONAL VALUES[a] (Tn) AND SEAL SCORES RELATIVE TO ESTRIOL CONFORMERS COMPARED TO THE LOWEST ENERGY STRUCTURE OF DEOXYCORTISOL[b]

Conformer	T1[a]	T2[a]	T3[a]	ΔE[c]	SEAL1[d]	SEAL2[e]	SEAL3[f]
1	178	−63	180	0.00	**−76.21**	−74.92	−72.52
2	2	−63	180	0.04	−75.63	**−75.36**	−74.17
3	178	178	−63	0.24	−76.84	−74.33	**−71.33**
4	2	178	−63	0.28	−78.42	−77.10	**−71.59**
5	179	179	180	0.37	−72.77	−70.45	**−70.09**
6	2	179	180	0.40	−74.23	−73.10	**−70.75**
7	178	178	58	0.51	−75.56	−73.17	**−70.39**
8[g]	178	−64	−64	0.53	−78.31	**−77.21**	−75.76
9	2	178	58	0.54	−77.58	−76.70	**−71.28**
10	2	−64	−64	0.57	−79.46	−78.18	**−76.51**
11	178	−63	58	0.83	−77.01	**−76.55**	−74.63
12	2	−63	58	0.86	−78.55	−77.73	**−75.81**
13	178	65	179	1.33	−74.55	**−72.69**	−72.58
14	2	65	179	1.36	−74.11	−73.89	−73.69
15	179	66	−64	1.91	−76.58	−75.65	**−73.95**
16	1	66	−64	1.95	−78.04	−77.58	**−74.37**
17	178	66	59	2.04	−75.41	−74.46	**−73.05**
18	2	66	59	2.08	−77.49	**−76.67**	−72.11

[a] The selected three rotatable bonds are illustrated in Fig. 1.
[b] SEAL solutions which well align the steroid nucleus (type-i superposition; see text and Fig. 4a) are highlighted in bold.
[c] Difference energy values calculated by the Tripos force field with respect to the estriol lowest energy structure.
[d] SEAL score for the best alignment.
[e] SEAL score for the second best alignment.
[f] SEAL score for the third best alignment.
[g] Chosen conformer of estriol.

(Fig. 4b); and (iii) matching of the steroid nuclei with the 19-methyl groups pointing in opposite directions (Fig. 4c).

In Table 2 the scores corresponding to these alignments are reported for each conformer of estriol. To get a consistent alignment for CoMFA field calculations within the entire set, only the first superposition (type-i superposition; see Fig. 4a) was considered, although it ranked as the second or third SEAL solution. Thus, although conformer no. 10 has the lowest score in the table, conformer no. 8 was selected because it is characterized by the best score within type-i alignments. The same criterion was adopted for estrone and estradiol.

PLS on the whole training set

The steroid structures selected for CoMFA field calculations were used to compute WHIM and MS-WHIM indices. PLS was then applied to each description matrix to search for a structure–activity correlation. Cross-validated results for the CoMFA, WHIM and MS-WHIM models are summarized in Table 3. The relative scatter plots of predicted versus actual binding affinities are reported in the left column of Fig. 5.

The first two columns of Table 3 list LOO parameters, while the last two show the mean q^2 and s_{PRESS} values obtained through 5RG made in 100 different ways. The original steroid CoMFA analysis (Cramer et al. [9]), which was cross-validated by forming four random groups only one time, is also shown. The results corresponding to the 1 Å grid-spacing analysis have been reported, since in our CoMFA refinement we used this step size. Moreover, the reported analysis is characterized by a higher q^2 value with respect to the standard parameter setting of one.

Since different cross-validation procedures were used, it is difficult to compare our CoMFA model with that

TABLE 3
CROSS-VALIDATED RESULTS FOR CoMFA, WHIM AND MS-WHIM ANALYSES ON THE WHOLE TRAINING SET[a]

Variable	LOO		5RG	
	q^2	s_{PRESS}	q^2	s_{PRESS}
CoMFA	0.840(3)	0.508	0.819(3)	0.536
			0.057	0.078
Cramer[b]	–	–	0.750(2)[c]	–
			–	–
WHIM	0.667(3)	0.735	0.602(3)	0.799
			0.092	0.087
MS-WHIM	0.631(2)	0.751	0.576(2)	0.801
			0.084	0.078

[a] The data shown refer to Leave-One-Out (LOO) and Five Random Groups repeated 100 times (5RG) cross-validation protocols. The optimum number of components is indicated in parentheses, the respective standard deviation values are reported below 5RG q^2 and 5RG s_{PRESS}.
[b] Values were taken from Table 5 of Ref. 9.
[c] Value obtained by means of four random cross-validation groups formed just one time.

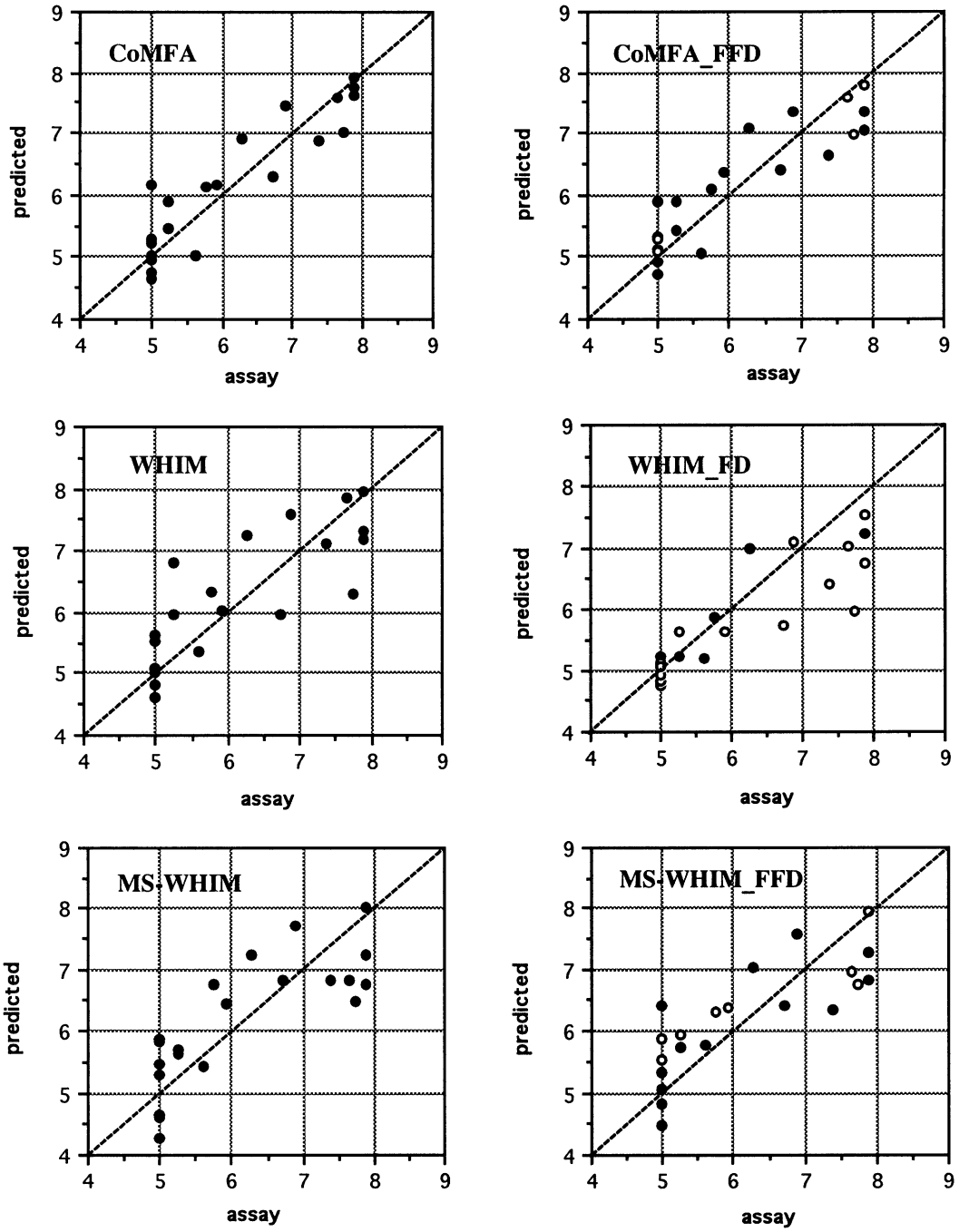

Fig. 5. Scatter plots of actual versus predicted activities for CoMFA, WHIM and MS-WHIM PLS models (LOO), computed both on the whole (left side) and on the design-reduced data set (right side). In the latter case, closed circles represent the steroid molecules that were included in the model, while open circles indicate the excluded compounds. As the optimal number of components for the WHIM_FD PLS model was evaluated by means of SDEP on the 14 compounds held out, closed circles in this case refer to fitted and not to predicted values.

obtained by Cramer et al.; in both cases, however, the q^2 values are high, 0.819 (5RG) and 0.750, respectively. WHIM and MS-WHIM analyses are characterized by good 5RG q^2, 0.602 and 0.576, but these values are lower than those from CoMFA. The low standard deviation values for 5RG q^2 and 5RG s_{PRESS} indicate that the PLS models for each description are stable and not affected significantly by the way in which the groups are formed. Interestingly, the standard deviation associated with CoMFA 5RG q^2 is the lowest; furthermore, CoMFA gave LOO q^2 and 5RG q^2 values of 0.840 and 0.819, respectively; these values are closer to each other than in the WHIM and MS-WHIM models. This may be a result from the different scaling options adopted: for the CoMFA matrix, the CoMFA_STD scaling procedure was applied only once before cross-validation, while for the

TABLE 4
OBSERVED AND PREDICTED ACTIVITY DATA FOR THE TEST SET OBTAINED BY USING THE WHOLE TRAINING SET[a]

Molecule	CoMFA	Cramer[b]	WHIM	MS-WHIM	Assay[c]
Test 1	8.084	**6.629**	7.617	7.228	7.512
Test 2	7.666	7.744	**11.775**	**8.573**	7.553
Test 3	6.538	6.594	7.460	7.061	6.779
Test 4	7.804	7.518	**8.562**	7.684	7.200
Test 5	6.396	6.650	6.019	6.218	6.114
Test 6	**7.346**	**7.409**	**7.278**	**7.112**	6.247
Test 7	7.010	**5.247**	**5.692**	6.487	7.120
Test 8	6.864	7.373	7.198	6.917	6.817
Test 9	7.970	7.908	**8.508**	7.810	7.688
Test 10	**8.005**	**7.800**	**8.352**	**7.499**	5.797
SDEP	0.837	1.022	1.750	0.742	

[a] Predicted values for which $|y_{pred} - y_{obs}| > 0.8$ are highlighted in bold.
[b] Values were taken from Table 5 of Ref. 9.
[c] Affinity data (log 1/k) from Ref. 19.

TABLE 5
CROSS-VALIDATED RESULTS FOR MS-WHIM PLS ANALYSES COMPUTED WITHIN DIFFERENT SETS

Set	5RG q^2	5RG s_{PRESS}	ONC[b]
1[a]	0.576	0.801	2
2	0.562	0.814	2
3	0.582	0.795	2
4	0.513	0.859	2
5	0.565	0.811	2
6	0.505	0.866	2
7	0.515	0.857	2
8	0.516	0.856	2
9	0.557	0.819	2
10	0.530	0.844	2
Mean	0.542	0.832	
Standard deviation	0.029	0.027	

[a] The results obtained within this set have been reported in Table 3.
[b] Optimal number of components.

WHIM and MS-WHIM matrices the autoscaling procedure was repeated for each subgroup of molecules [1].

Test set predictions are listed in Table 4. A comparison between our CoMFA results and those of Cramer et al. indicates that considerable improvements were obtained. These are particularly evident for molecules test 1 and test 7, which are now well predicted; test 6 and test 10 are, however, still overpredicted.

With regard to the WHIM results, it should be noted that, in spite of a good q^2 value, test molecules 2, 4, 6, 7, 9 and 10 are poorly predicted, giving the highest final SDEP (1.750). The MS-WHIM approach gave results which, with the exception of molecule test 2, are comparable to those from the CoMFA analysis. Furthermore, the SDEP is the lowest in the table (0.742).

MS-WHIM invariance to the coordinate system

MS-WHIM indices, if computed on highly dense Connolly surfaces (see Fig. 3), are invariant to roto-translation. To further show that MS-WHIM does not necessitate any preliminary alignment of molecular structures, nine additional data sets were analyzed. Each set was obtained by randomly varying the 3D orientation of each molecular structure. The stability of the PLS models obtained was then verified. Table 5 reports the 5RG q^2 values obtained for all the analyses. The lowest and highest q^2 values associated with all analyses are 0.505 and 0.582, respectively. The standard deviation value, which is as low as 0.029, indicates high stability. Moreover, all these PLS models are two-component models. These results confirm that MS-WHIM descriptors are not affected by the 3D orientation of a given molecular conformation with respect to either the coordinate system or any other molecular structure. The consistency of the results was further confirmed by predicting the test set molecules by means of the PLS models obtained. Single predictions, mean and standard deviation values within the different data sets are reported in Table 6.

TABLE 6
EXTERNAL PREDICTIONS OBTAINED BY USING MS-WHIM WITHIN DIFFERENT SETS

Set	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8	Test 9	Test 10
1[a]	7.228	8.573	7.061	7.684	6.218	7.112	6.487	6.917	7.810	7.499
2	7.173	8.600	6.700	7.823	6.361	6.947	6.420	6.734	7.886	7.690
3	7.271	8.613	6.844	7.641	6.159	6.968	6.547	6.621	7.766	7.607
4	7.202	8.626	6.910	7.841	6.170	7.057	6.652	6.974	7.989	7.641
5	7.094	8.672	6.743	7.602	6.408	6.820	6.479	6.884	8.066	7.535
6	7.349	8.504	7.094	7.970	6.071	6.956	6.681	6.948	8.036	7.488
7	7.356	9.020	6.915	7.792	6.153	7.044	6.394	6.807	7.944	7.579
8	7.381	8.311	7.174	7.875	6.126	7.203	6.660	6.889	7.888	7.466
9	7.202	8.475	7.044	7.684	6.186	7.102	6.556	6.934	7.820	7.449
10	7.319	8.962	7.118	7.600	6.238	7.093	6.440	6.768	7.898	7.568
Mean	7.257	8.636	6.960	7.751	6.209	7.030	6.532	6.848	7.910	7.552
Standard deviation	0.093	0.213	0.163	0.127	0.104	0.109	0.105	0.112	0.098	0.079

[a] The results obtained within this set have been reported in Table 4.

TABLE 7
CROSS-VALIDATED RESULTS (LEAVE-ONE-OUT) FOR CoMFA, WHIM AND MS-WHIM ANALYSES[a] ON EXPERIMENTAL DESIGN-REDUCED TRAINING SETS

Variable	IN[b]	LOO		OUT[c]	SDEP
		q^2	s_{PRESS}		
CoMFA_FFD	16	0.729(3)	0.614	5	0.365
WHIM_FD[d]	7	–	–	14	0.717(2)
MS-WHIM_FFD	13	0.605(2)	0.795	8	0.661

[a] The optimum number of components is indicated in parentheses.
[b] Number of compounds included in the designed-reduced data sets.
[c] Number of compounds excluded by experimental design.
[d] Because of the low number of compounds, the optimum number of components for WHIM_FD was evaluated by means of SDEP on the 14 compounds held out.

PLS on design-reduced training sets

Each data matrix was further investigated by carrying out a PCA. Five relevant Principal Components (PCs) were retained for CoMFA and MS-WHIM matrices (72% and 78% of the explained variance, respectively), while only 3 PCs were sufficient to explain more than 80% of the total variance for the WHIM matrix. The relative score plots along the first three components are shown in Fig. 6. To obtain a better balanced set, FD was applied to the WHIM matrix while in the case of CoMFA and MS-WHIM matrices the latent variable hyperspace was explored by means of FFD (see Methods). The compounds selected within each data matrix are reported in Table 1; the reduced data sets contain 16, 7 and 13 molecules for CoMFA, WHIM and MS-WHIM, respectively.

Cross-validated results for analyses of reduced data sets are summarized in Table 7; the SDEP index computed for the compounds of the training set not included in the model is also given. Scatter plots of predicted versus actual activities are presented in the right column of Fig. 5. The LOO q^2 values for the CoMFA and MS-WHIM PLS models are slightly worse than the corresponding values in Table 3. The WHIM result is meaningless because of the low number of compounds included in the set; the optimal number of components for the WHIM PLS model was chosen through the SDEP index computed on the 14 steroids held out.

The improvement achieved using the experimental design strategy is supported by the test set predictions and the SDEP indices reported in Table 8, which are superior to those from Table 4. For comparison purposes, external predictions obtained by the COMPASS [13] method and representing, to date, the most accurate results are shown in Table 8. From these data it can be noticed that our CoMFA refinement gives comparable results. The only residual larger than 0.8 is relative to molecule test 10 and this is remarkable, considering that none of the previous studies could correctly predict this molecule, which is the only one possessing a fluorine substituent in position 9. WHIM analysis on the FD-reduced training set increased the number of well-predicted compounds from four to seven; molecules test 2, test 7 and test 10 remain poorly predicted.

Finally, considerable improvements can be observed for the MS-WHIM analysis, where all predictions are more accurate than those obtained with the whole data set. Like for CoMFA, test 10 is still overpredicted; however, the MS-WHIM model is again characterized by the lowest SDEP value (0.662).

Discussion

New 3D theoretical descriptors, MS-WHIM, computed on a Connolly molecular surface, were developed and applied to a set of steroids (the CBG case), originally studied by Cramer et al. [9] in the first application of CoMFA. To test the reliability of these MS-WHIM descriptors, new CoMFA fields were computed on 21 steroids and the previously described WHIM indices [5–7] were also evaluated from their x,y,z atomic coordinates. PLS regressions were then carried out for each description matrix, both on the whole set of steroids and on design-reduced training sets. Finally, cross-validated results and external predictions on 10 additional compounds were compared. The main findings are as follows.

The main differences in our CoMFA refinement with respect to the original work by Cramer et al. [9] reside in the strategy adopted in structure selection and in the alignment criterium. Unlike the approach of Cramer et al., where the lowest energy structure for each molecule arising from a grid search was selected and aligned by means of a geometrical fit of the steroid nucleus, we performed multiple molecular comparisons using SEAL.

TABLE 8
OBSERVED AND PREDICTED ACTIVITY DATA FOR THE TEST SET BY USING DESIGN-REDUCED TRAINING SETS[a]

Molecule	CoMFA	COMPASS[b]	WHIM	MS-WHIM	Assay[c]
Test 1	7.883	7.062	7.244	7.300	7.512
Test 2	7.430	7.729	**11.592**	8.332	7.553
Test 3	6.642	6.462	6.869	6.821	6.779
Test 4	7.705	7.466	7.814	7.445	7.200
Test 5	6.495	5.994	5.533	6.121	6.114
Test 6	6.962	6.383	6.769	6.901	6.247
Test 7	6.848	6.625	**5.506**	6.532	7.120
Test 8	6.816	7.403	6.337	6.838	6.817
Test 9	7.767	7.741	7.935	7.860	7.688
Test 10	**7.793**	**7.779**	**7.826**	**7.491**	5.797
SDEP	0.716	0.705	1.563	0.662	

[a] Predicted values for which $|y_{pred} - y_{obs}| > 0.8$ are highlighted in bold.
[b] Values were taken from Table 4 of Ref. 13; to date these are the most accurate predictions on the 10 steroids series. The relative LOO q^2 on the 21 steroids training set is 0.89.
[c] Affinity data (log 1/k) from Ref. 19.

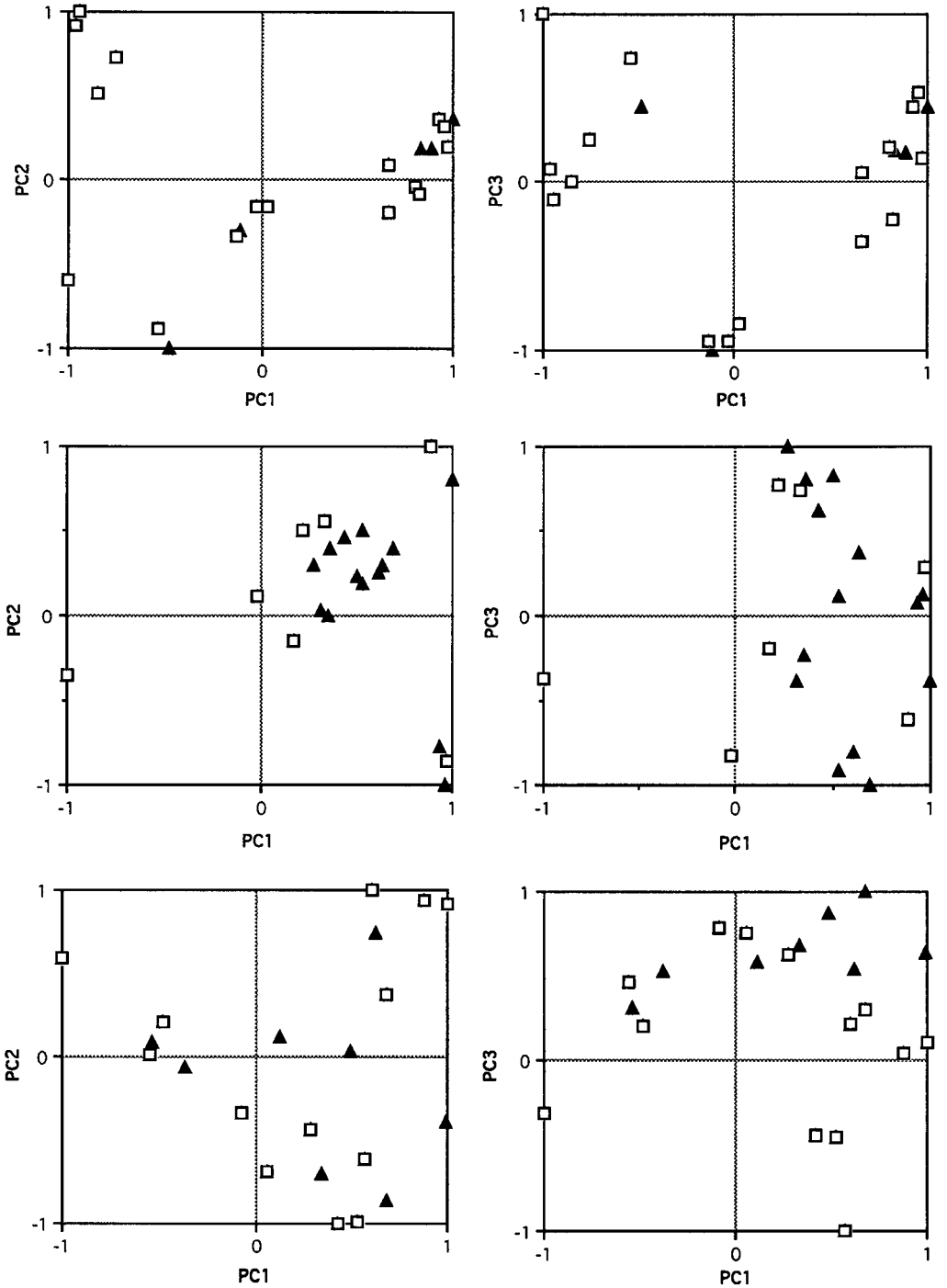

Fig. 6. Plots of object scores for CoMFA (top), WHIM (middle) and MS-WHIM (bottom). Left: first versus second principal component; right: first versus third principal component. The cumulative variances explained by the first three latent variables are 59%, 82% and 60% for CoMFA, WHIM and MS-WHIM, respectively. Squared boxes indicate the molecules chosen by means of experimental design.

The aim was to obtain, in one step only, a homogeneous data set in terms of steric and electrostatic properties. It is well accepted that a ligand tends to maximize its steric and electrostatic complementarity to the active site of the receptor. As no structural information on the CBG binding site is yet available, we chose for each molecule the conformer that had the largest steric and electrostatic similarity with the lead compound (deoxycortisol), with the assumption that the most active molecule in a series fits the steric and electronic requirements of the receptor binding site to a larger extent. The global minimum structure of deoxycortisol was selected as the template structure, since no information was available with regard to its structure when bound to the receptor. Although this represents an arbitrary choice, the reported results (see Tables 3 and 4), which show an improvement with respect

to the original work by Cramer et al., attest that the self-consistency of the molecular structures and their alignments are a crucial step in CoMFA [1].

The q^2 and external SDEP values of our CoMFA refinement are similar to those obtained by Jain et al. [13] with COMPASS (see Tables 7 and 8), which showed the highest q^2 and the most accurate predictions reported to date. It should be stressed, however, that the COMPASS approach relies on a nonlinear statistical tool like Neural Network, which is known to have overfitting and chance correlation problems [31,32]. Our results, reported in Tables 3 and 7 (q^2 and s_{PRESS}) and Tables 4 and 8 (external predictions), clearly show that linear PLS is sufficient to treat the steroids data set. It is noteworthy that our q^2 value is a consequence of the strategy adopted in creating the data set, while for COMPASS, q^2 represents the index to be optimized by means of an iterative choice of conformations and alignments. However, it should be emphasized that neither of the two methods guarantees that the real active site geometries have been analyzed or found, since the acquisition of a good q^2 value can be considered only a necessary condition, but not a sufficient one [1].

The reported results (see Tables 3 and 4 as well as Tables 7 and 8) attest that the combined use of PCA and design is a simple and useful tool to improve the predictive power of a statistical model. As pointed out from the score plots reported in Fig. 6, the 21 steroids appear not to be homogeneously spread over the pricipal component space and tend to be grouped into a few clusters. The trend appears to be common to the three description matrices, although it is less pronounced for MS-WHIM. By applying FD and FFD strategies, we obtained better balanced data sets. Thus, for each data matrix, the design produced (i) comparable predictions for the 21 steroid molecules (see the PLS plots of Fig. 5); and (ii) considerable improvements in the prediction of the 10 test set compounds (Table 8).

The PCA/PLS analyses performed on the steroids data set highlight that, for the analysis of highly specific ligand–receptor interactions, MS-WHIM descriptors provide new useful information with respect to the original WHIM indices. From PCA it is evident that the amount of information provided by MS-WHIM indices is better distributed over the principal component space (see Fig. 6). From PLS models, MS-WHIM indices show a higher predictive power, as the predictions carried out on the 10 test set molecules are more accurate both when considering the whole (Table 4) and the design-reduced (Table 8) training sets. Thus, the best external SDEP values were 1.563 and 0.662 for WHIM and MS-WHIM, respectively. These results may be expected, since the molecular surface coded by MEP is undoubtedly a more realistic representation, with respect to the molecular skeleton, of how a molecule is perceived by a biological system.

Although the final MS-WHIM PLS model is characterized by a LOO q^2 value lower than that from CoMFA (0.605 versus 0.729), MS-WHIM predicts well all the test set molecules, giving a slightly better external SDEP (0.662 versus 0.716). The only exception is molecule test 10 which, however, is also not well predicted by CoMFA. Thus, the 36-column MS-WHIM matrix turns out to be as effective as the thousands of columns needed for CoMFA. Furthermore, comparable results were achieved by means of a great reduction of computational time. The 21-steroid set can be coded by the MS-WHIM procedure, using the considerable density of 10 points per $Å^2$, in as little as 150 s on a Silicon Graphics Crimson workstation, while PCA and PLS regression are more than immediate, also when applying heavy cross-validation procedures. Moreover, the concise number of MS-WHIM indices guarantees that complex and time-consuming statistical procedures like variable selection could be swiftly accomplished. As the GOLPE [24] procedure applied to CoMFA fields allowed significant improvements, it is likely that the high correlation within WHIM-based descriptors could be easily removed by using variable selection techniques.

In addition, the main advantage of MS-WHIM indices over CoMFA relies on the invariance to roto-translation, in that molecular structures do not need to be aligned if sufficiently dense molecular surfaces are used (Fig. 3 and Tables 5 and 6). In fact, the alignment procedure can be considered the major bias in CoMFA since statistical results are strictly dependent on how molecules are superimposed. Furthermore, as highlighted in Table 2, the alignment step often represents a multiple solution problem [33] (low energy value differences discriminate the first three SEAL solutions between estriol and deoxycortisol).

On the other hand, the mathematical procedure that underlies all the invariant molecular descriptors makes their physical interpretation difficult or almost impossible, apart from the amount of information provided. From this point of view, MS-WHIM descriptors are not different from the original WHIM or other known invariant descriptors, e.g. autocorrelation functions. Consequently, the derived statistical models do not allow any interpretation of specific 3D ligand–receptor interactions. For these reasons, WHIM-based models can be usefully applied to predict the activity of unknown molecules, but they cannot be used to suggest which type of punctual chemical modification should be applied to the molecules under examination to improve their biological efficacy.

Moreover, the limitation in the physical interpretation and the lack of a graphical display of the results require a careful statistical test of WHIM-based models. For instance, in this work we investigated the risk of finding a well-fitting, but meaningless model by scrambling the response variable several times. All the models thus ob-

tained gave negative q^2 values (data not shown). Considering this result and the good predictions obtained for the 10 test set steroids, we are confident that the goodness of the statistical models based on MS-WHIM descriptors is reliable and not due to chance correlation.

Finally, MS-WHIM, like CoMFA, cannot avoid the problem of conformational freedom; consequently, at present, for highly flexible compounds other tools are needed to choose the appropriate conformation before deriving regression models. Nevertheless, in the future we see possibilities to use MS-WHIM also for molecules with a large number of conformations and our efforts point toward this direction.

Conclusions

New WHIM-based 3D theoretical descriptors, called MS-WHIM, were derived from molecular surface properties. PCA/PLS analyses on a series of steroids clearly show that MS-WHIM indices contain meaningful chemical information, suitable for 3D QSAR studies. Thus, the statistical results obtained compare well with those achieved using CoMFA fields. The main limitation of MS-WHIM is that the information provided is highly condensed and cannot be extracted in order to interpret the statistical models obtained. On the other hand, the concise number of indices, the speed of calculation and the invariance to roto-translation, which avoids possible problems due to the molecular alignment, represent the main advantages of this new approach over CoMFA.

The analysis of additional weighting schemes like Molecular Lipophilicity Potential (MLP) [34] is in progress. Strategies employing MS-WHIM indices in the study of highly flexible compounds are also under investigation.

References

1 Cramer III, R.D., DePriest, S.A., Patterson, D.E. and Hecht, P., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, The Netherlands, 1993, pp. 443–485.

2 Bradshaw, J., Wynn, E.W., Salt, D.W. and Ford, M.G., In Wermuth, C.G. (Ed.) Trends in QSAR and Molecular Modelling 92 (Proceedings of the 9th European Symposium on Structure–Activity Relationships: QSAR and Molecular Modelling), ESCOM, Leiden, The Netherlands, 1993, pp. 220–224.

3 Broto, P., Moreau, G. and Vandycke, C., Eur. J. Med. Chem.–Chim. Ther., 19 (1984) 66.

4 Clementi, S., Cruciani, G., Riganelli, D., Valigi, R., Costantino, G., Baroni, M. and Wold, S., Pharm. Pharmacol. Lett., 3 (1993) 5.

5 Todeschini, R., Lasagni, M. and Marengo, E., J. Chemometr., 8 (1994) 263.

6 Todeschini, R., Gramatica, P., Provenzani, R. and Marengo, E., Chemometr. Intell. Lab. Syst., 27 (1995) 221.

7 Todeschini, R., Vighi, M., Provenzani, R., Finizio, A. and Gramatica, P., Chemosphere, 8 (1996) 1527.

8 Todeschini, R., Moro, G., Boggia, R., Bonati, L., Cosentino, U., Lasagni, M. and Pitea, D., Chemometr. Intell. Lab. Syst., in press.

9 Cramer III, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.

10 Oprea, T.I., Ciubotariu, D., Sulea, T.I. and Simon, Z., Quant. Struct.–Act. Relatsh., 12 (1993) 21.

11 Good, A.C., So, S. and Richards, W.G., J. Med. Chem., 36 (1993) 433.

12 Good, A.C., Peterson, S.J. and Richards, W.G., J. Med. Chem., 36 (1993) 2929.

13 Jain, N.A., Koile, K. and Chapman, D., J. Med. Chem., 37 (1994) 2315.

14 Dunn, J.F., Nisula, B.C. and Rodbard, D., J. Clin. Endocrin. Metab., 53 (1981) 58.

15 Joliffe, I.T., Principal Components Analysis, Springer, New York, NY, U.S.A., 1986.

16 Box, G.E.P., Hunter, W.G. and Hunter, J.S., Statistics for Experimenters, Wiley, New York, NY, U.S.A., 1978.

17 Clementi, S., Cruciani, G., Baroni, M. and Costantino, G., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, The Netherlands, 1993, pp. 570–572.

18 Wold, S., Albano, C., Dunn III, W.J., Edlund, U., Ebsen, K., Geladi, P., Hellberg, S., Johansson, E., Lindberg, W. and Sjostrom, M., In Kowalsky, B.R. (Ed.) Chemometrics, Reidel, Dordrecht, The Netherlands, 1984, p. 17.

19 Westphal, U., Steroid–Protein Interactions II, Springer, Berlin, Germany, 1986.

20 Roberts, S.M., Symposia in Print: Molecular Recognition in Chemistry and Biochemistry Problems, Royal Society, London, U.K., 1989.

21 Connolly, M., QCPE Bull., 1 (1981) 75.

22 Weiner, P., Langridge, R., Blaney, J.M., Schefer, R. and Kollman, P.A., Proc. Natl. Acad. Sci. USA, 79 (1982) 3754.

23 SYBYL molecular modeling system, available from Tripos Associates, Inc., St. Louis, MO, U.S.A.

24 Baroni, M., Costantino, G., Cruciani, G., Riganelli, D., Valigi, R. and Clementi, S., Quant. Struct.–Act. Relatsh., 12 (1993) 9.

25 Kearsley, S.K. and Smith, G.M., Tetrahedron Comput. Methodol., 3 (1990) 616.

26 Clark, M., Cramer III, R.D. and Van Opdenbosch, N., J. Comput. Chem., 7 (1986) 230.

27 Gasteiger, J. and Marsili, M., Tetrahedron, 36 (1980) 3219.

28 Klebe, G., Mietzner, T. and Weber, F., J. Comput.-Aided Mol. Design, 8 (1994) 751.

29 Wold, S., Technometrics, 20 (1978) 397.

30 Kubinyi, H. and Abraham, U., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, The Netherlands, 1993, pp. 717–728.

31 Manallack, D.T., Ellis, D.D. and Livingstone, D.J., J. Med. Chem., 37 (1994) 3758.

32 Manallack, D.T. and Livingstone, D.J., Med. Chem. Res., 2 (1992) 181.

33 Folkers, G., Merz, A. and Rognan, D., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, The Netherlands, 1993, pp. 583–618.

34 Audry, E., Dubost, J.P., Colleter, J.C. and Dallet, P., Eur. J. Med. Chem.–Chim. Ther., 21 (1986) 71.