Journal of Computer-Aided Molecular Design, **14:** 293–306, 2000.
KLUWER/ESCOM

Global 3D-QSAR methods: MS-WHIM and autocorrelation

Emanuela Gancia*, Gianpaolo Bravi#, Paolo Mascagni & Andrea Zaliani
Italfarmaco Research Centre, via Lavoratori 54, I-20092 Cinisello Balsamo, Milan, Italy

Received 7 October 1998; Accepted 13 October 1999

Key words: Connolly surface, endothelin A, HIV-reverse transcriptase, holistic description, molecular electrostatic potential, PCA, PLS

Summary

The recently proposed MS-WHIM indices, a set of theoretical descriptors containing information about size, shape and electrostatic distribution of a molecule, have been further investigated. The main objectives of this work were: (i) to confirm the descriptive power of MS-WHIM in modelling specific biological interactions, (ii) to analyse the dependence of MS-WHIM on the type of atomic charges used for computing electrostatic potential and (iii) to compare the performances of MS-WHIM with those provided by other global 3D molecular descriptors. The spatial autocorrelation of atomic and molecular surface properties were selected for comparison purposes. WHIM-based and autocorrelation-based vectors were calculated for two molecular sets from the literature, namely a series of 18 HIV-1 reverse transcriptase inhibitors and a set of 36 sulphonamide endothelin inhibitors. PLS was adopted to derive statistical predictive models that were validated by means of cross-validation. The reported results confirmed that MS-WHIM indices are able to provide meaningful statistical correlations with biological activity. MS-WHIM descriptors are sensitive to the type of partial atomic charges applied and improved models were obtained using more accurate charges. Moreover for both the datasets, MS-WHIM results, in terms of fitting and predictive power of PLS models, were superior to those from autocorrelation. Finally, the strengths/weaknesses of global 3D-QSAR descriptors over local CoMFA-like methods, as well as the main differences between WHIM-based and autocorrelation-based vectors, are discussed.

Introduction

Quantitative Structure-Activity Relationships (QSAR) techniques are usually divided into two classes, classical-QSAR and 3D-QSAR methods. The difference between the two approaches depends on the type of molecular descriptors used to correlate molecular structure to biological activity. Classical QSAR methods rely on descriptors that are scalars and essentially measure only one dimension in the parameter space [1] (e.g., log P, molar refractivity, Hansch–Fujita lipophilic constants, Hammett electronic constants, Verloop sterimol parameters [2], etc.).

3D-QSAR methods are based on the detailed description of the local properties of each chemical structure. They are sensitive to the particular conformation adopted by a molecule as well as to its orientation with respect to the other molecules. Well-known examples include molecular fields (CoMFA [3], HINT [4], LUMO [5] and E-state fields [6]), similarity matrices [7], COMPASS distances [8] and receptor surface interaction energies [9].

An intermediate position between classic QSAR and local 3D-QSAR is occupied by procedures that can be defined as 'global' 3D-QSAR methods [10]: they are based on holistic descriptors, which condense into a brief numerical vector some of the chemical information contained in a 3D structure. Although these types of indices do not explicitly describe local spatial steric and electrostatic molecular features, they still contain all of this information in an aver-

*To whom correspondence should be addressed at: Celltech Chiroscience, Drug Design, Cambridge Science Park, Milton Road, Cambridge CB4 0WE, U.K. E-mail: emanuelagancia@chiroscience.com
#Present address: Glaxo Wellcome R&D, Medicines Research Centre, Gunnels Wood Rd, Stevenage SG5 2NY, Hertfordshire, U.K.

aged manner. Their main advantages with respect to traditional 3D-QSAR methods consist in the reduced number of indices necessary to describe the molecules and in their independence on molecular orientation. Examples include 3D topological indices [11], 3D autocorrelation function-based coefficients [12, 13], CoMMA descriptors [10], and Weighted Holistic Invariant Molecular (WHIM) indices [14–16]. WHIM descriptors consist of 12 statistical parameters, calculated from the *x-y-z* coordinates of a molecule within different weighting schemes (i.e., atomic properties). They contain information about the whole molecular structure in terms of size, shape, symmetry and atomic distribution.

We recently proposed new indices, named MS-WHIM [17], which are computed from Molecular Surface (MS) points [18], weighted by their Molecular Electrostatic Potential (MEP) [19] values, by applying a WHIM-based mathematical approach. They were developed in an attempt to consider the contribution arising from molecular surface recognition in specific ligand-receptor interactions. The MS-WHIM approach was successfully applied to a set of 31 steroids, which had been assayed for their binding affinity to Corticosteroid Binding Globulin. The results were shown to be superior to those obtained from the original 'atomic' WHIM indices and comparable to those derived from CoMFA fields.

Here we confirm the descriptive power of MS-WHIM in modelling specific biological interactions through its application to two molecular sets selected from the literature: (i) a set of 18 2-pyridinone derivatives that are HIV-1-specific Reverse Transcriptase inhibitors [20] and (ii) 36 sulphonamide endothelin inhibitors [21].

Furthermore, the dependence of MS-WHIM on the type of atomic charges used for computing MEP was analysed and its performance compared to those achievable by other global 3D descriptors, namely spatial autocorrelation functions. The autocorrelation technique was first applied to molecular description by Broto et al. [12]. They computed the autocorrelogram of atomic properties to describe the distribution of a given atomic property over a molecular structure. Recently, Wagener et al. [22] extended the autocorrelation concept to a molecular surface property (i.e., MEP) by following a reasoning analogous to that used in deriving MS-WHIM descriptors from the original WHIM descriptors. We calculated both atomic and molecular surface autocorrelation coefficients and compared their modelling power to MS-WHIM efficacy. Partial atomic charges were calculated at different levels of accuracy ranging from molecular mechanics to semiempirical methods. PLS [23] was adopted for modelling biological activities. Results were validated by means of cross-validation and repeated scrambling of the response variable.

Methods

Computational methods

Molecules were built within the SYBYL 6.2 [24] and QUANTA [25] molecular modelling packages. Mopac 6.0 [26] was used to optimise molecular structures and to calculate partial atomic charges. MEP fitted charges were instead computed using version 5 of the program. Connolly surfaces were generated using the MS program [18], and internally developed codes were employed to calculate WHIM, MS-WHIM and spatial autocorrelation (ACOR and MS-ACOR) descriptors. Statistical analyses were accomplished within Sybyl 6.2. All the calculations were performed on a Silicon Graphics Crimson workstation.

Molecular datasets and binding affinities

HIV-1 RT inhibitors. A set of 18 compounds, recently used to correlate some physico-chemical descriptors of heteroaromatic rings to biological activity [20], was used as the first training set. These molecules consist of N-substituted 3-aminopyridin-2(1H)-one derivatives and differ only in the type of aromatic ring present at one position. Only the variable portion of each inhibitor was therefore considered, and a methyl group was added at the substitution site. The molecular structures and the activity values [27] are reported in Figure 1. The different heterocyclic fragments were taken from the Sybyl fragment database or manually added, if necessary. Molecular structures were fully optimised by means of the TRIPOS force field [28]. Gasteiger-Hückel charges [29] were then computed. To evaluate the charge effect, the following semiempirical partial atomic charges were also calculated on fully optimised PM3 [30] and AM1 [31] structures: AM1/PM3 charges, AM1/PM3 molecular electrostatic potential fitted charges (ESP routine).

Endothelin inhibitors. A series of 36 aryl sulphonamides assayed for endothelin receptor subtype A (ET_A) antagonism was selected as the second dataset. Their

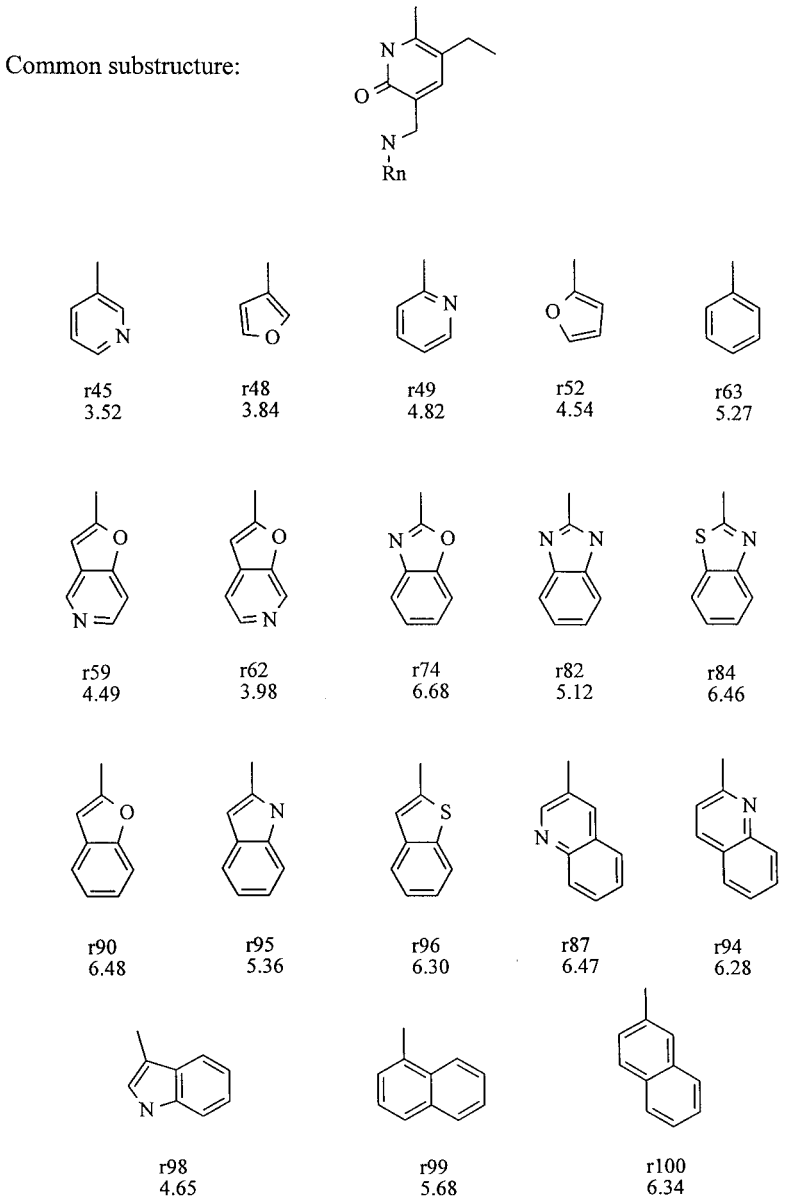

Figure 1. Structures and activity values (pIC_{50}) for the first dataset (HIV-1 RT inhibitors).

structures and activity values are reported in Figure 2. This molecular series has been previously analysed by S.R. Krystek et al. in a 3D-QSAR study involving CoMFA fields [21]. Molecular models were built in QUANTA CHARMm [32], as described in the original paper [21]. Only the aryl substituents were minimised for each molecule by using CHARMm [32] (1000 steps of conjugate gradient until the gradient was < 0.001, dielectric constant $= r$). To evaluate the charge effect, the following partial atomic charges were calculated on the previously described structures: Gasteiger-Hückel charges, AM1/PM3 charges, AM1/PM3 molecular electrostatic potential fitted charges (ESP routine). No geometry optimisation (SCF=1) was performed at the semiempirical level to preserve the input conformation.

For deriving structure-activity correlations, all the experimental binding affinities were transformed in the negative decimal logarithm of IC_{50}.

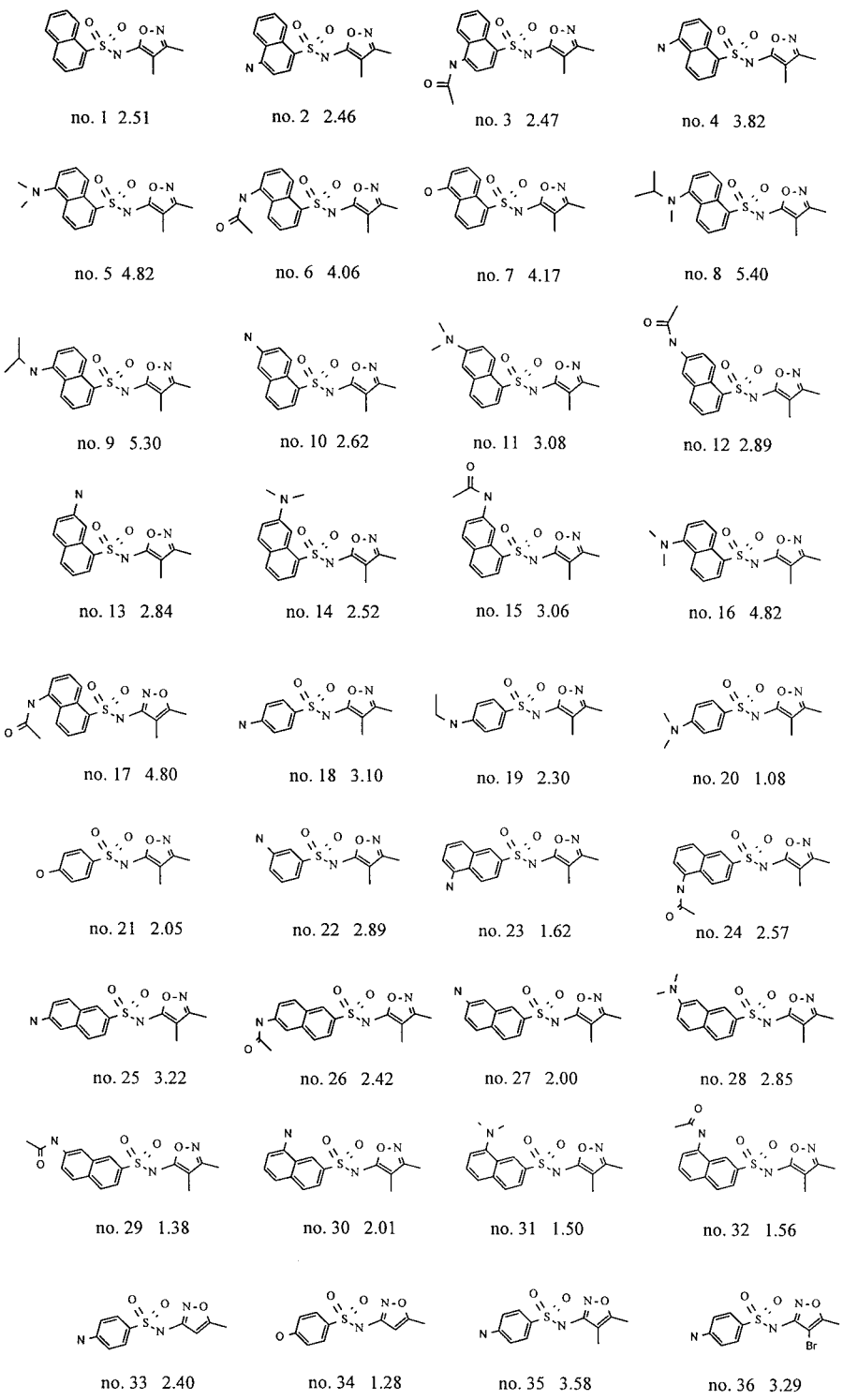

Figure 2. Structures and activity values (pIC_{50}) for the the second dataset (ET_A antagonists).

WHIM and MS-WHIM

WHIM indices are 12 statistical parameters computed starting from a coordinate matrix X_{ij} ($i = 1$, natoms; $j = 1, 3$) and a weight vector w_i.

This matrix is centred and a Principal Component Analysis (PCA) is performed leading to the score matrix T_{im} ($i = 1$, natoms; $m = 1, 3$) in the principal component space.

This statistical protocol assures invariance to translation (centring) and to rotation (PCA), therefore the molecules do not need to be aligned before computing WHIM indices. The following statistical parameters are then calculated from matrix T along each component m:

(a) variance (i.e., PCA eigenvalues):

$$\lambda_m = \Sigma_i (w_i t_{im}^2) / \Sigma_i w_i \quad (1)$$

(b) eigenvalue proportion:

$$\theta_m = \lambda_m / \Sigma_m \lambda_m \quad (2)$$

(c) skewness:

$$\gamma_m = \left| \left[\Sigma_i (w_i t_{im}^3) / \Sigma_i w_i \right] \right| \cdot 1/\lambda_m^{3/2} \quad (3)$$

(d) kurtosis:

$$\kappa_m = \left[\Sigma_i (w_i t_{im}^4) / \Sigma_i w_i \right] \cdot 1/\lambda_m^2 \quad (4)$$

The PCA eigenvalues refer to the extension of the coordinates and are correlated with the molecular size. Eigenvalue proportions are easily related to molecular shape, as planar molecules will have only two components. The acentric factor $\omega = \theta_1 - \theta_3$ is used instead of θ_3 [15]. Skewness represents the molecular symmetry along each component. As it is a third order moment, it can assume negative values: to preserve the invariance to rotation the absolute value is considered. The fourth-order moment, kurtosis, is related to atomic distribution and density around the centre and along principal axes. The reciprocal of this entity, $\eta_m = 1/\kappa_m$, (emptiness) is actually used [15] to avoid problems related to infinite κ_3 values of planar compounds; η_m may be viewed as the unfilled space per atom. A total of 12 WHIM indices are thus computed for each weight:

$$\lambda_1, \lambda_2, \lambda_3, \theta_1, \theta_2, \omega, \gamma_1, \gamma_2, \gamma_3, \eta_1, \eta_2, \eta_3$$

The detailed computational procedure is described in the original papers [14–16].

In relation to the kind of weights assigned to the atoms, different types of information can be obtained. As in all the previous papers, in the present work WHIM indices were calculated from atomic *x*-*y*-*z* coordinates within 4 different weighting schemes: (1) unitary case, (2) atomic mass; (3) van der Waals atomic volume and (4) Mulliken atomic electronegativity. A total of 48 molecular descriptors were thus obtained.

MS-WHIM [17] indices were calculated by applying the above statistical procedure (centring + PCA) on Connolly surface points within 3 weighting schemes:

(1) unweighted case, (2) positive MEP and (3) absolute value of negative MEP, 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. MEP was computed onto the surface points by means of the classical Coulomb formula using a distance-dependent dielectric constant (ε):

$$V_p = \Sigma_i q_i / \varepsilon r_i \quad (5)$$

where V_p is MEP value relative to point p and r_i the distance between *p* and the *i*-th atom. MS-WHIM indices are sensitive to the surface point density applied. On the basis of our previous study [17], we assumed MS-WHIM descriptors independent on molecular orientation when computed on highly dense surfaces. In the present work all MS-WHIM indices were computed by using 10 points per Å^2 as density value. More details on MS-WHIM theory and calculation can be found in the first paper presenting MS-WHIM [17].

Autocorrelation

Topological autocorrelation of atomic properties. Each term of the autocorrelation vector of a given property p distributed over n atoms is defined as follows:

$$A_d = \Sigma_{i,j}\, p(i) * p(j) \quad (6)$$

where each coefficient differs in the way the atomic pairs *i* and *j* are selected.

The first term, A_0, is simply the sum of p squared over all the atoms ($i = j$), A_1 is the sum of the products of p over all the atom pairs separated by one bond, A_2 the sum of the products of p over all the atom pairs separated by two bonds, and so on. In other words, each coefficient contains pairs of atoms separated by the same number of bonds.

Spatial autocorrelation of atomic properties (ACOR). It is obtained by simply replacing the topological dis-

tance between atoms by the interatomic distance in the 3D space. Each coefficient has the following equation:

$$A(d_{\text{lower}}, d_{\text{upper}}) = \Sigma_{i,j}\, p(i) * p(j) \quad (7)$$

where the sum is extended over all the atom pairs, whose distance is between d_{lower} and d_{upper}. An important point is that two different conformations of the same molecule will show equal topological correlation, but different spatial correlation. In this study, we used only the spatial autocorrelation vectors. For homogeneity to WHIM indices we selected the following atomic properties: (1) unitary case, (2) atomic mass; (3) van der Waals atomic volume and (4) Mulliken atomic electronegativity. Three trials were done, using different steps to define the best distance: 0.5, 1.0, and 1.5 Å.

Spatial autocorrelation of molecular surface properties (MS-ACOR). Equation 7 is applied, by using MEP values on molecular surface points as property and dividing each coefficient for the number N of terms in the sum:

$$A(d_{\text{lower}}, d_{\text{upper}}) = (1/N) * \Sigma_{i,j}\, p(i) * p(j) \quad (8)$$

If the autocorrelation is calculated by using molecular surface points and MEP values instead of atomic coordinates and atomic properties, a more accurate description is expected. In this work, the molecular surface and the MEP values on each point were calculated using the same procedure adopted for MS-WHIM calculation: for each molecule a Connolly surface with a 1.5 Å probe radius and 10.0 $Å^2$ point density was calculated and the classic Coulomb approach was adopted to compute MEP values.

More details on autocorrelation of atomic and molecular surface properties applied to QSAR can be found in the original papers [11, 22].

Statistical methods

All the description matrices were autoscaled to assign unit variance to each descriptor. Partial Least Squares (PLS) [23] was used to derive linear regressions between each descriptor matrix and the experimental activities.

Two different cross-validation procedures were adopted to determine the optimal number of components of each PLS model: (1) Leave-One-Out (LOO) and (2) 5 random groups (5RG). The latter cross-validation method consists in dividing the dataset in 5 random groups and performing 5 PLS analyses, leaving out and predicting one group at a time. By following the suggestion of the authors of GOLPE [33], this protocol was repeated up to 100 times and the associated parameters represent mean values. The predictive power of each statistical model was evaluated by means of q^2 and s_{PRESS} computed as follows [34]:

$$q^2 = 1 - \Sigma(y_{\text{pred}} - y_{\text{obs}})^2 / \Sigma(y_{\text{obs}} - y_{\text{mean}})^2 \quad (9)$$

$$s_{\text{PRESS}} = (\Sigma(y_{\text{pred}} - y_{\text{obs}})^2/(n - c - 1))^{1/2} \quad (10)$$

where n = number of compounds and c = number of components.

A further statistical validation was provided by the so-called 'scrambling' of the Y variable (i.e., the activity values were mixed so that each value was no longer assigned to the right molecule). Hundreds of scrambled Y-vectors were created for both the data sets and associated to all the descriptor matrices. LOO PLS was then applied to investigate the possibility of a chance correlation between our descriptors and the randomly assigned activity values.

Results

WHIM-based and autocorrelation-based vectors were calculated for two molecular sets, taken from the literature: a series of 18 HIV-1 RT inhibitors [20] and a series of 36 endothelin inhibitors [21]. Partial atomic charges for MEP calculation were evaluated at different levels of accuracy. All the description matrices were then used to model the biological activities of the above molecules, by means of PLS regressions. The quality of the obtained models, the performances achievable by using different atomic charge types and the goodness of WHIM and autocorrelation methods are discussed below.

MS-WHIM sensitivity to charge type

Atomic charges were computed by the Gasteiger–Hückel method, at the semiempirical level using AM1 and PM3 hamiltonians, and through fitting to electrostatic potential (AM1 and PM3). The cross-validated MS-WHIM results are reported in Tables 1C and 2C for HIV-1 RT and endothelin inhibitors, respectively. Different charges provide different performances. LOO-q^2 ranged from 0.473 to 0.689 in the first set, and from 0.401 to 0.664 in the second set. In both the applications, the best results were

obtained by using charges fitted to the electrostatic potential (ESP). The found trend reflects the different accuracy of the methods in evaluating partial atomic charges: the 'best' charges (ESP charges are computationally expensive but usually provide better results than Mulliken charges) correspond to the best models. This study demonstrates that MS-WHIM indices, although condensing in 36 descriptors the information contained in thousands of molecular surface points, are sensitive to the type of charges applied. The choice of the proper method of computing charges represents a useful way to improve MS-WHIM derived models. ESP semiempirical charges provided better results than molecular mechanics and Mulliken atomic charges and allowed to improve the q^2 value by as much as 0.25.

WHIM and MS-WHIM

Sections A and C of Table 1 list the cross-validated parameters relative to the model of HIV-1 RT inhibitor activity obtained using WHIM and MS-WHIM indices, respectively. In the case of WHIM a LOO-q^2 as low as 0.399 was obtained. This value is rather low (even though q^2 greater than 0.3 are considered significant [35]) and it is inferior to the result from MS-WHIM, i.e., LOO-$q^2 = 0.557$, based on Gasteiger-Hückel charges. Moreover, as discussed above, MS-WHIM models are sensitive to the charge type and can be improved using semiempirical atomic charges instead of molecular mechanics. In the HIV-1 RT inhibitors application, when using MS-WHIM and PM3-ESP atomic charges, it was possible to obtain a very good model, both in terms of fitting ($r^2 = 0.932$) and predictive (LOO-$q^2 = 0.689$) power. These results are better than those obtained previously, which were based on a Principal Component Analysis of the physicochemical properties of a set of small heterocycles [20] (our r^2 values are in the range from 0.851 to 0.932 versus 0.79 from literature). The scatter plots of calculated/predicted versus experimental activities are reported in Figure 3 and Figure 4 for the WHIM and the best MS-WHIM models, respectively.

The results obtained for the 36 sulphonamides, assayed for ET_A antagonism, are summarised in Table 2, sections A and C. The WHIM model gave a good result (LOO-$q^2 = 0.652$), equivalent to the best MS-WHIM model (LOO-$q^2 = 0.664$, AM1-ESP charges). These results are comparable to those reported from a CoMFA study [21] (LOO-$q^2 = 0.7$, 6 components) and further attest the modelling ability of WHIM-based descriptors for QSAR studies. The scatter plots of calculated/predicted versus experimental activities for the WHIM model are given in Figure 5. The same plots for the best MS-WHIM model (charges: AM1-ESP) are reported in Figure 6. Being a 6-component model, it is not surprising that the fitting power of the MS-WHIM model is superior to that of the WHIM model.

As described in the original paper on ET_A receptor antagonists, we removed four molecules (one with high, one with low and two of medium activity) from the original dataset. The models obtained for the reduced dataset of 32 compounds were then used to predict the activities of the four molecules held out. In Table 3 calculated and assayed activities for compounds 27, 12, 4 and 16 are reported. The best predictions (corresponding to a sdep value of 0.382) were obtained by means of MS-WHIM descriptors.

Table 1. Cross-validated results for the 18 HIV-1 RT inhibitors[a]

Method	LOO-q^2	LOO-s_{PRESS}	5RG-q^2	5RG-s_{PRESS}
A. WHIM				
4 weights	0.399 (1)	0.825	0.367 (1)	0.845
B. Autocorrelation of atomic properties (ACOR)				
Step 0.5	0.526 (5)	0.846	0.353 (1)	0.855
Step 1.0	0.395 (2)	0.855	0.374 (2)	0.868
Step 1.5	0.420 (2)	0.837	0.401 (2)	0.849
C. MS-WHIM				
Gast–Hück	0.557 (2)	0.731	0.525 (2)	0.752
AM1	0.534 (3)	0.776	0.365 (3)	0.895
PM3	0.473 (2)	0.798	0.415 (1)	0.812
AM1-ESP	0.626 (4)	0.722	0.508 (3)	0.823
PM3-ESP	0.689 (3)	0.692	0.632 (3)	0.683
D. Autocorrelation of molecular surface properties (MS-ACOR)[b]				
Gast–Hück	0.518 (3)	0.790	0.265 (2)	0.929
AM1	0.427 (2)	0.856	0.334 (2)	0.894
PM3	0.461 (2)	0.807	0.428 (2)	0.830
AM1-ESP	0.393 (2)	0.856	0.334 (2)	0.894
PM3-ESP	0.481 (1)	0.767	0.471 (1)	0.773

[a]The data shown refer to Leave-One-Out (LOO) and 5 random groups repeated 100 times (5RG) cross-validation protocols. The optimum number of components is indicated in brackets.
[b]Distance interval = 1.0 Å.

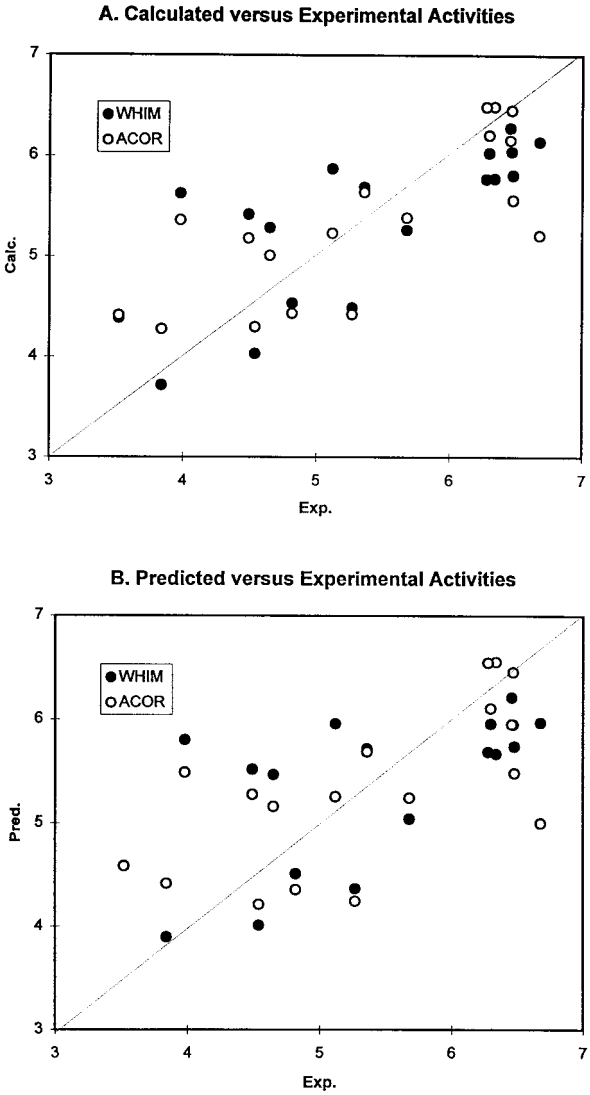

Figure 3. Scatter plots of experimental versus calculated activities (A) and experimental versus predicted (LOO) activities (B) for 18 HIV-1 RT inhibitors relative to models derived for WHIM (dark circles) and ACOR (white circles).

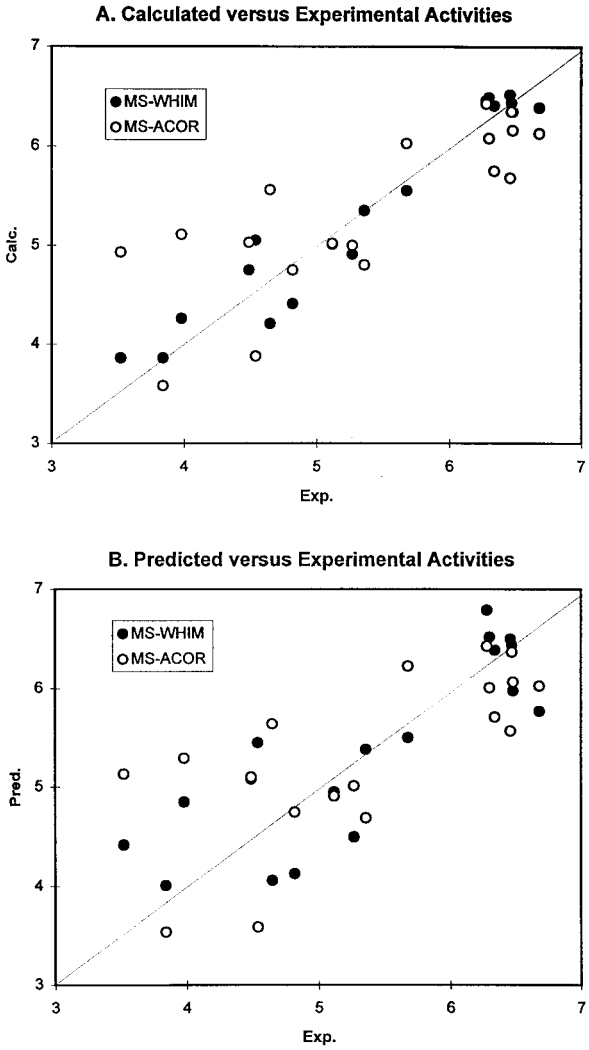

Figure 4. Scatter plots of experimental versus calculated activities (A) and experimental versus predicted (LOO) activities (B) for 18 HIV-1 RT inhibitors relative to models derived for MS-WHIM (dark circles) and MS-ACOR (white circles).

Autocorrelation

The statistical parameters referring to the models from autocorrelation of atomic properties (ACOR) of HIV-1 RT inhibitors are reported in Table 1, section B. Three analyses were performed, differing in the distance interval (0.5, 1.0 and 1.5 Å respectively) between the autocorrelation coefficients. When using a step of 0.5 Å, a good LOO result ($q^2 = 0.526$) did not correspond to a similarly good 5RG result ($q^2 = 0.353$); moreover, the optimal number of components for the different cross-validation protocols is not the same. The model obtained by using autocorrelation with step 1.5 Å shows a higher stability, giving a LOO-q^2 of 0.420 and a 5RG-q^2 of 0.401, both with 2 components.

When calculating the autocorrelation of molecular surface properties (MS-ACOR), the distance interval between indices was set to 1.0 Å, after preliminary runs exploring also 0.5 and 1.5 Å intervals. The distance of 1.0 Å provided the best results (data not shown). The same step of 1.0 Å has been used by Wagener et al., to compute autocorrelation vectors of a set of steroids [22] and they suggest the use of intervals equal to or lower than 1.0 Å.

Table 2. Cross-validated results for the 36 sulphonamide endothelin inhibitors[a]

Method	LOO-q^2	LOO-s_{PRESS}	5RG-q^2	5RG-s_{PRESS}
A. WHIM				
4 weights	0.652 (3)	0.705	0.622 (3)	0.732
B. Autocorrelation of atomic properties (ACOR)				
Step 0.5	0.491 (4)	0.865	0.430 (4)	0.913
Step 1.0	0.455 (4)	0.895	0.370 (4)	0.964
Step 1.5	0.362 (4)	0.969	0.286 (4)	1.022
C. MS-WHIM				
Gast–Hück	0.401 (2)	0.910	0.358 (2)	0.940
AM1	0.638 (5)	0.742	0.579 (5)	0.797
PM3	0.569 (6)	0.824	0.497 (6)	0.885
AM1-ESP	0.664 (6)	0.727	0.615 (6)	0.775
PM3-ESP	0.556 (5)	0.822	0.514 (5)	0.857
D. Autocorrelation of molecular surface properties (MS-ACOR)[b]				
Gast–Hück	0.189 (3)	1.075	0.118 (3)	1.120
AM1	0.200 (3)	1.068	0.145 (3)	1.101
PM3	0.180 (3)	1.081	0.138 (3)	1.106
AM1-ESP	0.174 (3)	1.085	0.130 (3)	1.111
PM3-ESP	0.205 (3)	1.064	0.135 (3)	1.108

[a]The data shown refer to Leave-One-Out (LOO) and 5 random groups repeated 100 times (5RG) cross-validation protocols. The optimum number of components is indicated in brackets.
[b]Distance interval = 1.0 Å.

Table 3. External predictions for 4 sulphonamide ET_A antagonists removed from the original dataset of 36 molecules

Cpd.	Assay	WHIM	ACOR	MS-WHIM
27	2.00	2.30	2.02	2.54
12	2.89	1.91	3.50	2.47
4	3.82	3.38	3.08	4.04
16	4.82	4.64	3.97	4.56
sdep[a]	–	0.566	0.639	0.382

[a]$\text{sdep} = \left(\Sigma(y_{\text{pred}} - y_{\text{obs}})^2/n\right)^{1/2}$.

As for MS-WHIM indices, we calculated MS-ACOR vectors using different atomic charge types. The results obtained for the 18 HIV-1 RT inhibitors by using MS-ACOR are only slightly better than those from the autocorrelation of atomic properties. In particular, when using PM3-ESP charges, we achieved a one component model with a LOO-q^2 of 0.481.

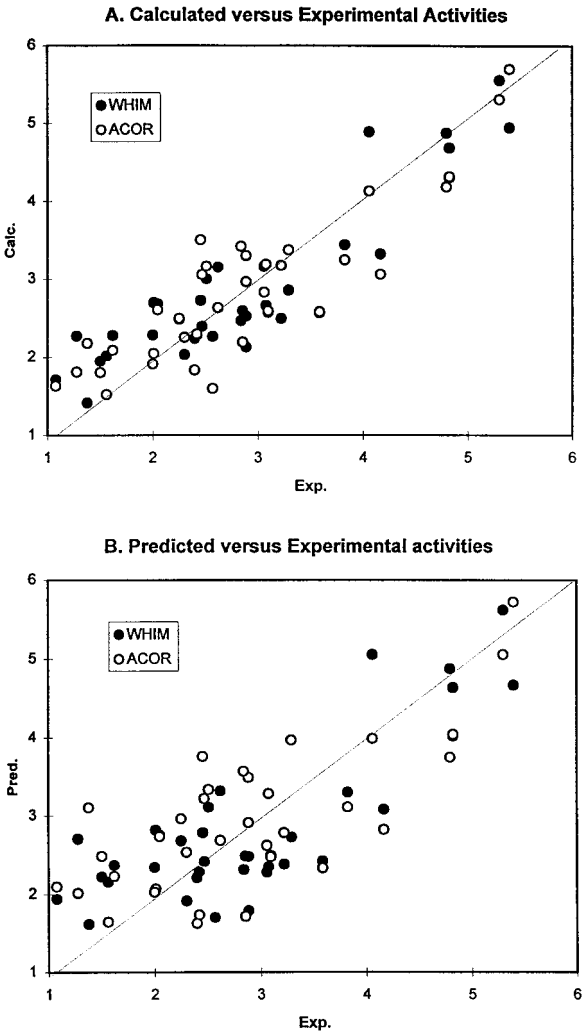

Figure 5. Scatter plots of experimental versus calculated activities (A) and experimental versus predicted (LOO) activities (B) for 36 ET_A antagonists relative to models derived from WHIM (dark circles) and ACOR (white circles).

The results obtained for the second dataset are reported in Table 2, sections B and D. The best model for ACOR was obtained by using an interval of 0.5 Å and was characterised by a LOO-q^2 of 0.491. In contrast, MS-ACOR provided worse results. The poor results in terms of q^2 and s_{PRESS} may be explained by the major influence of steric over electrostatic properties in determining the activity of these compounds. In the original CoMFA model [21] the ratio of steric and electrostatic contribution was 69:31. MS-ACOR, as defined in the Methods section, is aimed at extracting the electrostatic information contained in a molecular structure and it is not surprising that it fails in

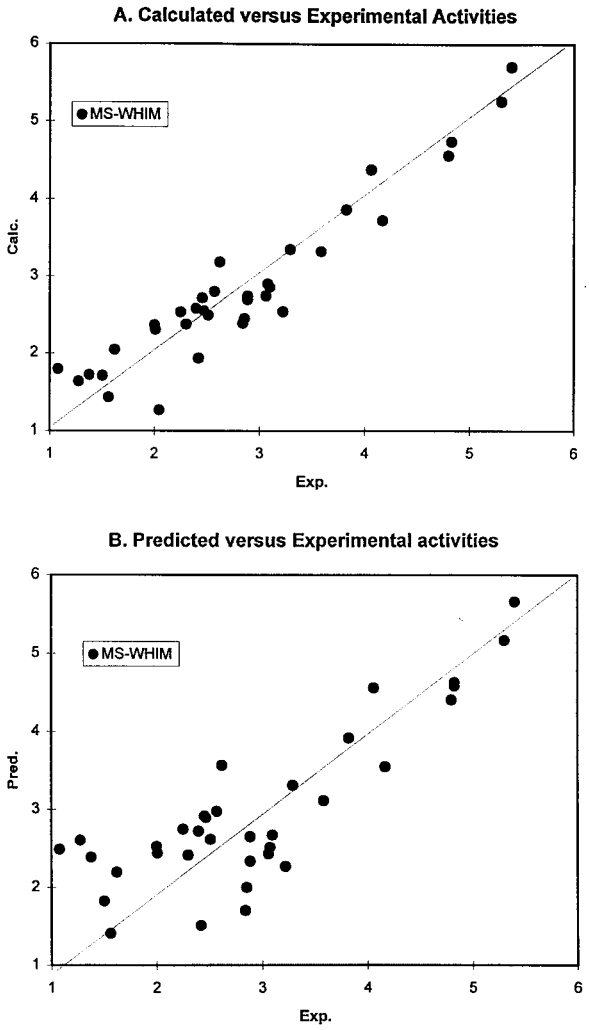

Figure 6. Scatter plots of experimental versus calculated activities (A) and experimental versus predicted (LOO) activities (B) for 36 ET_A antagonists relative to the MS-WHIM model.

modelling a biological activity mostly related to steric properties.

Discussion

We report the results from two statistical techniques, autocorrelation and WHIM, used to derive molecular descriptors for 3D-QSAR studies. Both the methods were applied either to atomic coordinates (named ACOR and WHIM) or to molecular surface points (named MS-WHIM and MS-ACOR). We tested these descriptors in two 3D-QSAR studies, the first one involving a set of HIV-1 RT inhibitors and the second one a series of endothelin inhibitors. The best results obtained for each method are summarised in Table 4.

Table 4. Summary of the best results obtained for each method

Method	LOO-q^2	LOO-s_{PRESS}	r^2	n. comp.
A. 18 HIV-1 RT inhibitors				
WHIM	0.399	0.825	0.551	1
ACOR	0.420	0.837	0.571	4
MS-WHIM	0.689	0.692	0.932	3
MS-ACOR	0.481	0.767	0.620	1
B. 36 endothelin inhibitors				
WHIM	0.652	0.705	0.802	3
ACOR	0.491	0.865	0.782	4
MS-WHIM	0.664	0.727	0.907	6
MS-ACOR	0.200	1.068	0.455	3

Macroscopically, autocorrelation- and WHIM-based indices have the same advantages and drawbacks, common to all 'global' molecular descriptors. The major strength is that they can be applied to molecules which do not belong to congeneric series or do not show common substructures. This is so because global 3D descriptors are insensitive to the molecular orientation or to the frame of reference and do not need to be aligned. Furthermore, global 3D descriptors are very fast to compute and the computational procedure can be easily automated. Drawbacks include: the inability to reconstruct the starting information from the descriptor vectors, a limited physical interpretation of each index and consequently of the derived statistical models.

Due to the restricted interpretation of these types of 3D-QSAR models, their validation is a major point; for this reason we chose to apply a more robust cross-validation (5RG, repeated 100 times) in addition to the standard leave-one-out technique. Moreover, we excluded the possibility of chance correlations by multiple scrambling of the biological activity. All the trials to correlate our descriptors to randomly permuted activity values failed, producing q^2 below or very close to zero (data not shown).

In general, WHIM and autocorrelation provide sound statistical models and the present study confirmed the power of holistic description in QSAR. The results on both molecular sets are comparable or better than those reported in the literature, the statistical parameters are significant (Tables 1–4) and the corre-

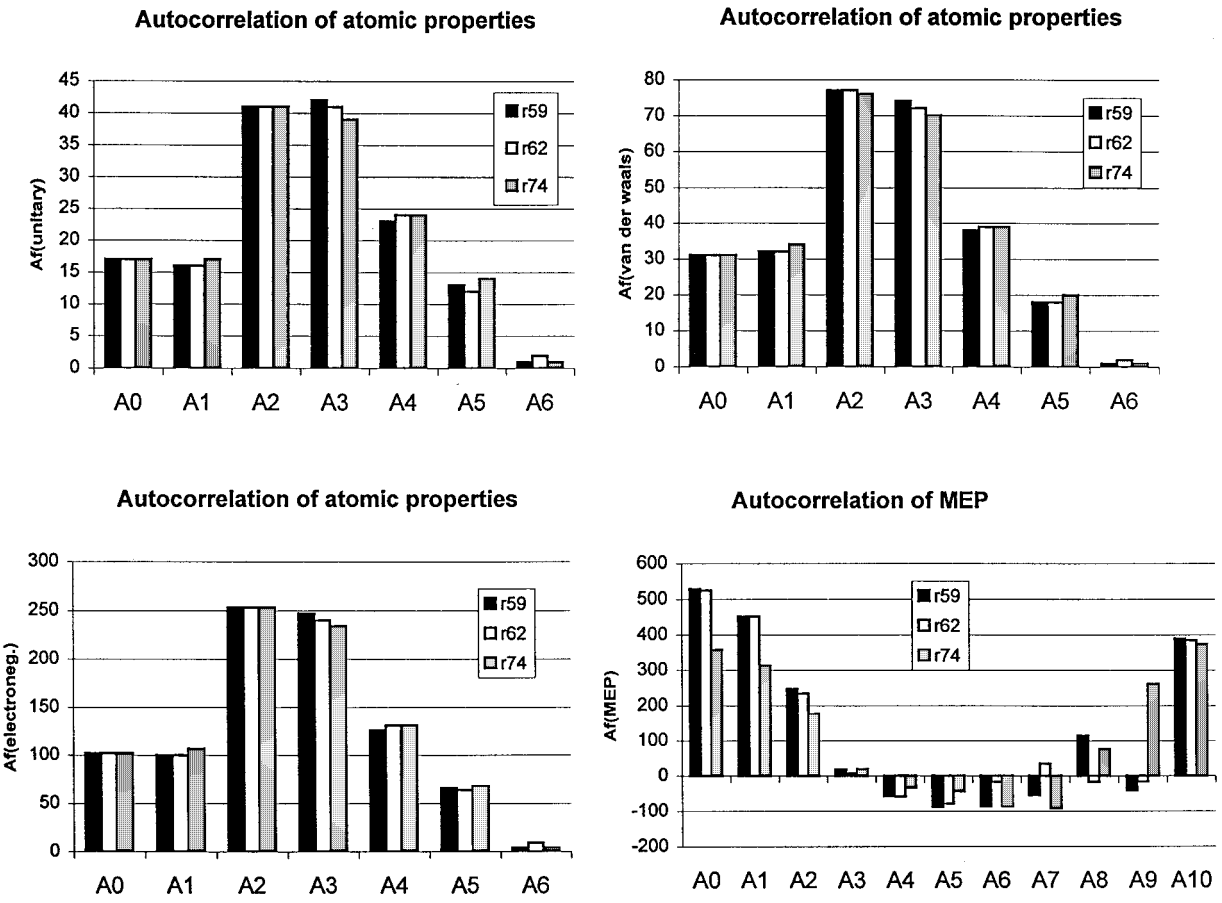

Figure 7. ACOR and MS-ACOR vectors of molecules *r*59, *r*62 and *r*74 (see Figure 1).

lation between predicted and actual biological activity is satisfactory (Figures 3–6).

An internal comparison between WHIM and autocorrelation is particularly appropriate as both the approaches use the same starting molecular information and both are aimed at expressing how a chemical property is distributed over a molecular structure. The two methods differ only in the mathematical procedure adopted to condense such information into a numerical vector. In particular, at the atomic level, the same weights (or properties) were used, when computing WHIM-based and autocorrelation-based vectors: unitary (unweighted), van der Waals volume, electronegativity and atomic mass. Thus the observed difference in performances between WHIM and ACOR descriptors can be due only to differences in their efficiency in condensing the input information.

WHIM and ACOR models on HIV-1 RT inhibitors show similar q^2, as well as similar r^2 (Table 4A). The WHIM model on the endothelin inhibitors was instead better than the corresponding ACOR model (Table 4B). Anyway, both the results are worthy of consideration and show that the WHIM protocol can provide statistical models equal or better than autocorrelation at the atomic level.

When computing MS-WHIM and MS-ACOR descriptors, the comparison is less straightforward, as different weights were used. Either in the HIV-1 RT case (Table 4A) or on the endothelin inhibitors (Table 4B), the MS-WHIM approach provided better results than autocorrelation. MS-WHIM descriptors contain explicit information relative to the shape and size of the molecular surface, whereas MS-ACOR indices, as actually defined, take into account mainly the electrostatic contribution. The poor performance of MS-ACOR, at least on the second dataset, could be due to the higher contribution of steric over electrostatic properties in explaining the biological activity.

In summary, for the analysed datasets, the WHIM protocol, both at atomic and molecular surface level

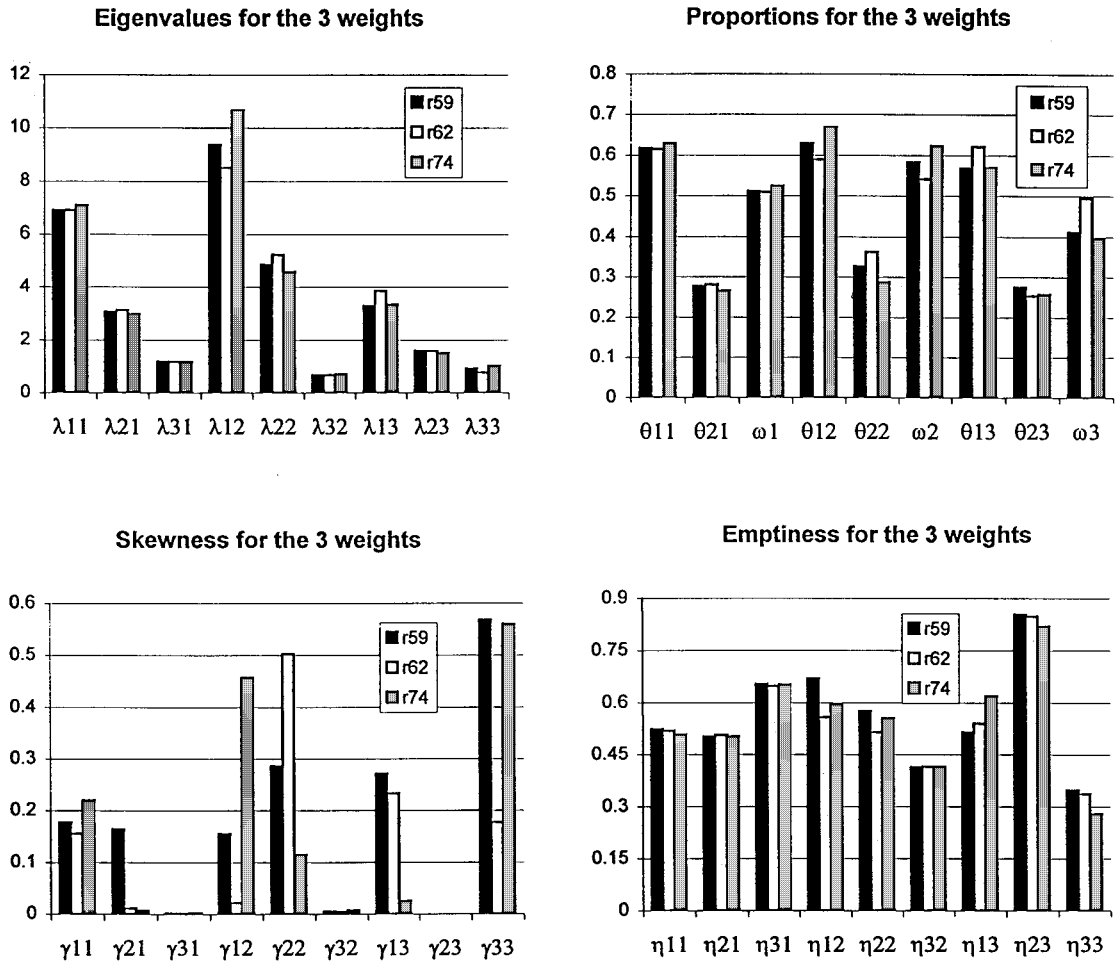

Figure 8. MS-WHIM indices of molecules *r*59, *r*62 and *r*74 (see Figure 1).

was more effective in condensing molecular information than autocorrelation. When dealing with thousands of compounds, aspects such as the calculation time or the number of indices can make the difference. The computational requirements are quite different for the four methods: the least expensive descriptors are ACOR vectors (2 s to code the database containing 36 endothelin antagonists on a Crimson workstation), followed by WHIM indices (3 s). The MS-WHIM protocol required 4 min for the same task. MS-ACOR proved to be the most expensive method with a computational time 4 times greater than MS-WHIM. One may wonder whether it is worth to apply MS-WHIM and autocorrelation techniques to molecular surface points (coding thousands of points) instead of simply use the atomic coordinates (coding tens of points). When the electrostatic properties appear to provide relevant QSAR descriptors (as observed for the HIV-1 RT inhibitors) the use of the molecular surface instead of the atomic coordinates has provided significantly better models for both the techniques. In the second series (endothelin inhibitors), the MS-WHIM model was also shown to be superior to the WHIM model (about the same LOO-q^2, but better external predictions for MS-WHIM), whereas the MS-ACOR gave a worse model than atomic autocorrelation. In summary, in three cases out of four, the superior amount of information contained in the molecular surface with respect to the atomic coordinates was conserved by the statistical treatment.

Moreover, MS-WHIM and MS-ACOR were found to be able to provide a description, which is more sensitive to subtle changes in the chemical structure with respect to WHIM and ACOR. HIV-1 RT inhibitors consist of simple molecular structures characterised by heteroaromatic moieties that, in some cases, are

very similar to each other. We investigated each single descriptor vector to verify how MS-WHIM and autocorrelation are able to distinguish among similar structures. We focused our attention on molecules *r*59, *r*62 and *r*74, which differ by just one atom type in one position. ACOR and MS-ACOR vectors of *r*59, *r*62 and *r*74 are reported in Figure 7, whereas their MS-WHIM vectors (charge types: PM3-ESP) are shown in Figure 8.

By using the autocorrelation of atomic properties, as well as atomic WHIM (data not shown), the 3 molecules are characterised by very similar descriptors; in particular it is almost impossible to distinguish between *r*59 and *r*62. On the other hand, when using MS-WHIM or MS-ACOR the two heterocycles show different indices. As expected *r*59 and *r*62 mainly differ in the skewness values. Molecules *r*74, *r*59 and *r*62 are characterised by different biological affinities (6.68, 4.49 and 3.98, respectively), so a QSAR method should be expected to distinguish among them in order to provide reliable results.

In conclusion, either for WHIM or for autocorrelation, a more accurate description and therefore better results can be expected when coding molecules through their molecular surfaces instead of their atomic coordinates. In addition, more flexibility is provided in the first case, as the proper choice of the atomic charge type for MEP calculation may improve the predictive power of PLS models. In an attempt to make WHIM and MS-WHIM indices more homogeneous, we also computed 36 atomic WHIM indices on HIV-1 RT inhibitors using the following 3 weights: (i) unitary values, (ii) positive atomic charges and (iii) absolute values of negative charges. We used the best atomic charges (ESP-PM3). The model obtained was not better than that achieved by using the 48 original WHIM indices (LOO-$q^2 = 0.332$).

Looking at the reported performances and requirements, MS-WHIM indices seem to allow a good trade-off between computational time, number of descriptors and reliability of the results of a QSAR study. On the other hand, atomic WHIM and ACOR can be usefully applied for quick screenings over thousands of compounds.

Conclusions

The descriptive power of recently developed MS-WHIM indices has been investigated through a comparative 3D-QSAR study on a set of 18 HIV-1 RT inhibitors and on a series of 36 endothelin inhibitors. The 3D autocorrelation vectors of molecular surface properties were chosen as a term of comparison. Several aspects of MS-WHIM descriptors were explored: from the sensitivity of the method to the type of partial atomic charges to the capacity to discern subtle chemical differences. The main findings are as follows.

The impact of the charge type on the quality of the MS-WHIM statistical model is considerable: in both the analysed datasets we obtained the same result: the more accurate the charges, the more accurate the models. Use of semiempirical atomic charges instead of molecular mechanics is an easy way to improve MS-WHIM-derived models. The analysis of MS-WHIM vectors obtained for very similar molecules reveals that MS-WHIM indices can monitor also small differences between two structures: molecules differing for only one position can be distinguished by MS-WHIM. This fact makes MS-WHIM a sensitive method of molecular description. Finally, the comparison between WHIM-based and autocorrelation-based molecular descriptors has demonstrated superior descriptive capabilities of MS-WHIM over 3D autocorrelation for the two analysed molecular sets.

We have focused our attention in monitoring MS-WHIM performances depending on the type of charge. Other ways to improve the MS-WHIM approach are under study. In particular the addition of new weights [36, 37] to the electrostatic weight could enhance the modelling power of the MS-WHIM technique. The explicit introduction of hydrophobic potential and of hydrogen-bonding acceptor/donor features is expected to further improve MS-WHIM performances as well as the interpretability of the derived QSAR models.

Acknowledgements

The authors are grateful to Dr David Manallack (Celltech Chiroscience, Cambridge, U.K.) for precious help in revising the paper.

References

1. Boyd, B.D., In Kent, A. and Williams, J.G. (Eds.) Encyclopedia of Computer Science and Technology, Vol. 33, suppl. 18, Marcel Dekker, New York, NY, 1995, p. 61.
2. Manhold, R., Krogsgaard-Larsen, P. and Timmermans, H. (Eds.) Methods and Principles on Medicinal Chemistry, Vol. 1, VCH publishers, Weinheim, 1993.

3. Cramer, R.D., III, Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
4. Abraham, D.J. and Kellogg, G.E., J. Comput.-Aided Mol. Design, 8 (1994) 41.
5. Poso, A., Juvonen, R. and Gynther, J., Quant. Struct.-Act. Relat., 14 (1995) 507.
6. Kellogg, G.E., Kier, L.B., Gaillard, P. and Hall, L.H., J. Comput.-Aided Mol. Design, 10 (1996) 513.
7. Good, A.C., So, S. and Richards, W.G., J. Med. Chem., 36 (1993) 433.
8. Jain, N.A., Koile, K. and Chapman, D., J. Med. Chem., 37 (1994) 2315.
9. Molecular Simulations Inc., Cerius^2 3.0 QSAR manual, 1997, p. 92.
10. Silverman, B.D. and Platt, D.E., J. Med. Chem., 39 (1996) 2129.
11. Bradshaw, J., Wynn, E.W., Salt, D.W. and Ford, M.G., In Wermuth, C.G. (Ed.) Trends in QSAR and Molecular Modelling 92, ESCOM, Leiden, 1993, pp. 220–224.
12. Broto, P., Moreau, G. and Vandycke C., Eur. J. Med. Chem.-Chim. Ther., 19 (1984) 66.
13. Clementi, S., Cruciani, G., Riganelli, D., Valigi, R., Costantino, G., Baroni, M. and Wold, S., Pharm. Pharmacol. Lett., 3 (1993) 5.
14. Todeschini, R., Lasagni. M. and Marengo, E., J. Chemometrics, 8 (1994) 263.
15. Todeschini, R., Gramatica, P., Provenzani, R. and Marengo, E., Chemometrics Intell. Lab. Syst., 27 (1995) 221.
16. Todeschini, R., Vighi, M., Provenzani, R., Finizio, A. and Gramatica, P., Chemosphere, 8 (1996) 1527.
17. Bravi, G., Gancia, E., Mascagni, P., Pegna, M., Todeschini, R. and Zaliani, A., J. Comput.-Aided Mol. Design, 11 (1997) 79.
18. Connolly, M., QCPE Bull., 1 (1981) 75.
19. Weiner, P., Langridge, R., Blaney, J.M., Schefer, R. and Kollman, P.A., Proc. Natl. Acad. Sci. USA, 79 (1982) 3754.
20. Gibson, S., McGuire R. and Rees, D., J. Med. Chem., 39 (1996) 4065.
21. Krysteck, S.R., Hunt, J.T., Stein, P.D. and Stouch, T.R., J. Med. Chem., 38 (1995) 659.
22. Wagener, M., Sadowsky, J. and Gasteiger, J., J. Am. Chem. Soc., 117 (1995) 7769.
23. Wold, S., Johansson, E. and Cocchi, M., In Kubinyi, H. (Ed.), 3D-QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, 1993, pp. 443–485.
24. SYBYL molecular modelling system, available from Tripos Associates, Inc., St. Louis, MO.
25. QUANTA (v. 4.0), Molecular Simulations Inc., Burlington, MA, 1994.
26. MOPAC: Quantum Chemistry Program Exchange no. 455.
27. Saari, W.S., Wai, J.S., Fisher, T.E., Thomas, C.M., Hoffman, J.M., Rooney, C.S., Smith, A.M., Jones, J.H., Banbenger, D.L., Goldman, M.E., O'Brien, J.A., Nunberg, J.H., Quintero, J.C., Schleif, W.A., Emini, E.A. and Anderson, P.S., J. Med. Chem., 35 (1992) 3792.
28. Clark, M., Cramer, R.D., III and Van Opdenbosch, N., J. Comput. Chem., 7 (1986) 230.
29. Sybyl 6.0 Theory Manual, Tripos Inc., St. Louis, MO, p. 2070.
30. Stewart, J.J.P., J. Comput. Chem., 10 (1989) 209.
31. Dewar, M.J.S., Zoebish, E.G., Healy, E.F. and Stewart, J.J.P., J. Am. Chem. Soc., 107 (1985) 3902.
32. Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., J. Comput. Chem., 4 (1983) 187.
33. Baroni, M., Costantino, G., Cruciani, G., Riganelli, D., Valigi, R. and Clementi, S., Quant. Struct.-Act. Relat., 12 (1993) 9.
34. Kubinyi, H. and Abraham, U., In Kubinyi, H. (Ed.) 3D-QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, 1993, pp. 717–728.
35. Cho, S.J. and Tropsha, A., J. Med. Chem., 38 (1995) 1060.
36. Bravi, G. and Wikel, J.H., Quant. Struct.-Act. Relat., in press.
37. Bravi, G. and Wikel, J.H., Quant. Struct.-Act. Relat., submitted.