Journal of Computer-Aided Molecular Design, 8 (1994) 97–112
ESCOM

J-CAMD 236

Effects of variable selection on CoMFA coefficient contour maps in a set of triazines inhibiting DHFR*

Giovanni Greco, Ettore Novellino, Maurizio Pellecchia, Carlo Silipo** and Antonio Vittoria***

Dipartimento di Chimica Farmaceutica e Tossicologica, Università degli Studi di Napoli,
Via D. Montesano 49, I-80131 Naples, Italy

Received 7 July 1993
Accepted 23 August 1993

Key words: Triazines; DHFR; QSAR; 3D QSAR; CoMFA

SUMMARY

An example of a CoMFA study is described with the aim to discuss one of the major problems of this 3D QSAR method: lack of variable selection. It is shown that the use of nonrelevant energy parameters might produce CoMFA contour maps which poorly reflect the actual nature of the binding site and are in part statistical artefacts. The data set employed in our analysis comprises triazine inhibitors of dihydrofolate reductase (DHFR), isolated from chicken liver, which have already been the object of a QSAR study by other authors. Since three-dimensional structures of triazine–DHFR complexes are known, it was possible not only to reduce ambiguities in the superimposition of the ligands, but also to compare the resulting CoMFA contour maps with the enzyme active site.

INTRODUCTION

Comparative Molecular Field Analysis (CoMFA) [1] is a 3D QSAR method which efficiently handles shape-dependent pharmacodynamic interactions. There are several advantages in using the CoMFA approach in structure–activity studies. Since no substituent constant values for steric and electrostatic descriptors are necessary to build up a CoMFA table, no data point is excluded due to the lack of tabulated parameters. Moreover, structurally heterogeneous compounds may be merged in the same model, because CoMFA descriptors do not depend on any reference compound. Lastly, CoMFA fields accurately and easily describe in three dimensions the steric and electrostatic features of the compounds.

*Supplementary material available: The Cartesian coordinates and the atomic charges of the PM3-optimized structures used in the CoMFA study are available as MOL2 files upon request.

**To whom correspondence should be addressed.

***To whose memory this paper is dedicated.

The main problems that may be encountered in CoMFA are the following: (a) establishing a sensible alignment rule and a receptor-recognized conformation for the ligands; (b) the large number of descriptors compared with the number of data points; this means that frequently there is a risk of using parameters which represent 'noise'.

In the present paper we describe how the results of a CoMFA study may heavily depend on the latter point, i.e., whether some form of variable selection (in addition to the simple standard deviation-based dropping of CoMFA columns) can lead to a better model.

Traditional QSAR analysis [2] has proved to be rather effective in highlighting substituent effects on biological activity in congeneric series. In such cases, one of the strengths of the Hansch approach is represented by the relatively small number of molecular descriptors needed to derive a correlation (much less than the hundreds of energy parameters generally required in CoMFA).

We thought it interesting to see the consequences of reducing variables in CoMFA by simply removing steric and electrostatic fields of substituents (borne by an invariant molecular frame) that a robust QSAR model had already established to be unimportant. The purpose of the study consisted, in other words, in evaluating the sensitivity of CoMFA statistics and coefficient contour plots to a variable selection procedure, guided by information obtained through classical QSAR.

The data set chosen for our study comprises a subset of 35 triazine inhibitors of the dihydrofolate reductase (DHFR), isolated from chicken liver, which were selected from a larger data set investigated by Hansch et al. [3]. The authors compared the QSAR models derived from 114 data points with the 3D structure of the enzyme active site determined by crystallography. These data were ideal for our purposes, as we had available the coordinates of a triazine–DHFR complex reported by Volz et al. [4]. In fact, we could define almost unambiguously the alignment rule and the bioactive conformation for the ligands. It is well known that the choice of the conformation and the criteria adopted in ligand superimposition are two major variables in CoMFA.

METHODS

Source of biological data

The structures of the analyzed compounds and their biological data are listed in Table 1. The biological activity of interest is the relative inhibitory activity, expressed as log $1/K_{i\ app}$ [5], determined by Hansch et al. on DHFR isolated from chicken liver [3]. The 35 compounds listed in Table 1 are a subset of 114 triazine derivatives, studied by the above authors, which present a low degree of conformational freedom. Thus, the rigidity of the selected ligands allowed us to significantly minimize problems related to the determination of the actual conformation of the ligand bound to the enzyme.

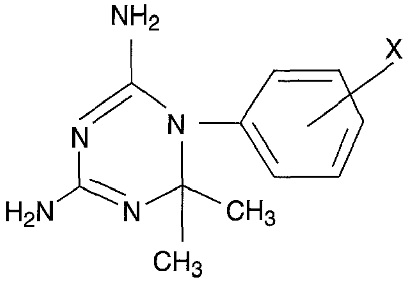
Molecular modeling

Molecular models of the ligands were constructed using standard bond distances and bond angles with the molecular modeling software package SYBYL [6], implemented on an Evans and Sutherland graphics system. The starting coordinates of the triazine–phenyl assembly were all taken from the crystallographic coordinates [4] of the triazine derivative **I**, oriented in the DHFR receptor cavity.

Optimized geometries and partial atomic charges were obtained by PM3 [7] calculations within the MOPAC program [8]. The resulting output geometries were in excellent agreement with the

TABLE 1

OBSERVED, PREDICTED AND CALCULATED log $1/K_{i\ app}$ VALUES

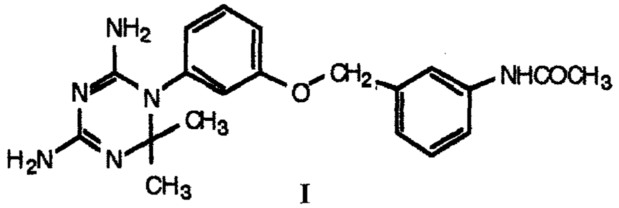


		log $1/K_{i\ app}$								
			Standard CoMFA[b]		PDVS-CoMFA[c]					
Compd	X	Obsd[a]	Pred.	Calcd	Pred.	Calcd	π_3	π_4	σ	ν
1	H	6.69	6.64	6.68	6.41	6.44	0.00	0.00	0.00	0.00
2	3-SO_2NH_2	5.00	4.81	4.88	5.49	5.24	−1.82	0.00	0.46	0.00
3	3-$CONH_2$	5.07	5.42	5.30	5.59	5.40	−1.49	0.00	0.28	0.00
4	3-$COCH_3$	5.56	5.80	5.71	6.14	6.00	−0.55	0.00	0.38	0.00
5	3-OH	5.57	6.42	6.25	6.07	5.95	−0.67	0.00	0.12	0.00
6	3-OCH_3	6.41	5.91	6.00	6.28	6.28	−0.02	0.00	0.12	0.00
7	3-$C(CH_3)_3$	6.75	7.59	6.85	8.04	7.51	1.98	0.00	−0.10	0.00
8	3-F	6.79	6.80	6.83	6.64	6.66	0.14	0.00	0.34	0.00
9	3-CN	6.94	6.32	6.54	6.09	6.34	−0.57	0.00	0.56	0.00
10	3-NO_2	6.95	6.45	6.60	6.42	6.70	−0.28	0.00	0.71	0.00
11	3-CH_2CH_3	7.00	6.66	6.70	7.10	7.07	1.03	0.00	−0.07	0.00
12	3-CF_3	7.01	7.13	7.10	7.32	7.24	0.88	0.00	0.43	0.00
13	3-CH_3	7.08	7.07	7.06	6.77	6.78	0.56	0.00	−0.07	0.00
14	3-Br	7.35	7.35	7.34	7.06	7.09	0.86	0.00	0.39	0.00
15	3-Cl	7.36	7.24	7.25	6.93	6.98	0.71	0.00	0.37	0.00
16	3-I	7.44	7.51	7.48	7.21	7.24	1.12	0.00	0.35	0.00
17	4-$COOC_2H_5$	4.45	5.41	4.51	4.87	4.21	0.00	0.51	0.00	1.51
18	4-$NHCOCH_3$	4.69	4.92	4.55	4.67	4.43	0.00	−0.97	0.00	0.39
19	4-SO_2NH_2	4.70	4.46	4.51	4.79	4.78	0.00	−1.82	0.00	0.99
20	4-$COOCH_3$	4.75	5.29	4.95	5.22	4.85	0.00	−0.01	0.00	1.51
21	4-CN	4.95	6.34	6.06	6.11	5.95	0.00	−0.57	0.00	0.40
22	4-$CONH_2$	4.95	5.15	5.05	5.07	5.02	0.00	−1.49	0.00	0.72
23	4-SO_2CH_3	5.25	4.47	4.88	4.75	5.12	0.00	−1.63	0.00	0.99
24	4-NH_2	5.67	5.49	5.51	5.43	5.48	0.00	−1.23	0.00	0.35
25	4-$COCH_3$	5.69	5.48	5.54	5.51	5.59	0.00	−0.55	0.00	0.72
26	4-OH	5.70	6.15	6.07	5.94	5.91	0.00	−0.67	0.00	0.32
27	4-CCH	6.05	6.22	6.17	6.22	6.17	0.00	0.40	0.00	0.58
28	4-OCH_3	6.48	6.20	6.32	6.13	6.27	0.00	−0.02	0.00	0.36
29	4-$C(CH_3)_3$	6.71	6.82	6.77	7.16	7.00	0.00	1.98	0.00	1.24
30	4-CF_3	6.77	6.58	6.71	6.88	6.90	0.00	0.88	0.00	0.91
31	4-F	6.89	6.69	6.76	6.47	6.51	0.00	0.14	0.00	0.27
32	4-I	6.93	7.17	7.17	7.16	7.14	0.00	1.12	0.00	0.78
33	4-Cl	6.95	6.79	6.85	6.78	6.82	0.00	0.71	0.00	0.55
34	4-CH_3	7.09	6.69	6.80	6.67	6.75	0.00	0.56	0.00	0.52
35	4-Br	7.12	6.93	7.00	6.88	6.94	0.00	0.86	0.00	0.65

[a] Data taken from Ref. 4.

[b] Values predicted and calculated through 3D Eq. 1.

[c] Values predicted and calculated through 3D Eq. 2.



I

X-ray structure of **I** at the level of the triazine–phenyl system. The root-mean-square (rms) distance between the non-hydrogen atoms shared by structure **I** and the PM3-optimized geometry of compound **1** (unsubstituted at the phenyl ring) was 0.1371 Å. Moreover, the torsion angle about the N-C single bond, defining the orientation of the phenyl ring attached to the triazine ring, was about 65° for all the investigated triazines as well as for **I**.

All the nonsymmetrical substituents, such as OCH_3, SO_2NH_2, $COCH_3$, etc., were oriented according to the following criteria: (a) the local dipole moment of the substituent had to be aligned as much as possible with that of the OCH_2 moiety featured by derivative **I**; (b) the steric bulk of the substituent had to be smallest in the direction of the triazine nucleus. Figure 1 shows the PM3 geometries of a few ligands as they were employed in CoMFA.

Molecular superimpositions were performed by minimizing the rms distance between all the heavy atoms in common with the triazine–phenyl system. This was done specifically through the SYBYL/FIT option.

The DHFR active site employed in this study was constructed with the following residues of the triazine **I**–DHFR complex crystallographic data [4]: Ser^6–Val^{10}, Ile^{16}–Ser^{39}, Asn^{48}–Trp^{57}, Ser^{59}–Asn^{72}, Trp^{113}–Tyr^{121}, Thr^{136}–Ile^{138}, Asp^{145} and Thr^{146}.

'Standard' CoMFA calculations

CoMFA was carried out using the QSAR option of SYBYL. The steric and electrostatic probe–ligand interaction energies (kcal/mol) were calculated with the Lennard-Jones and Coulomb potential functions within the TRIPOS force field [9], using a C_{sp^3} probe atom carrying a charge of +1.0. The steric energies were truncated at 5.0 kcal/mol; the electrostatic ones were not dropped, in correspondence to lattice points inside the union volume of the superimposed ligands.

The dimensions of the CoMFA lattice were determined through an automatic procedure, featured by the SYBYL/CoMFA routine, which insures that the lattice walls extend beyond the dimensions of each structure by 4.0 Å in all directions. The lattice spacing was set to a value of 2.0 Å.

In order to include the hydrophobicity of the substituents at the 3- and 4-positions of the benzene ring (X_3 and X_4, respectively), the π_3 and π_4 descriptors, used earlier by Hansch et al. in their QSAR study [3], were added to the CoMFA energy columns.

The inclusion of squared hydrophobic terms in the correlation was also considered. However, because of the limited range of π_3 and π_4 values in the set of 35 triazines chosen for the present study, the resulting parabolic model was not justified by statistics.

CoMFA with variable selection

The description of the hydrophobicity of the X_3 and X_4 substituents in the above reported 'standard' CoMFA procedure was position-dependent, while the steric and electrostatic fields were brought into the calculation as they were computed all over the ligand's structures. Here we describe how the hydrophobic *and* CoMFA fields were sampled according to the position of substitution.

A second type of CoMFA was carried out on the same data points by setting the steric field of all the 3-substituted derivatives (compounds **2–16** in Table 1) equal to that of compound **1**, where both X_3 and X_4 are hydrogens. Similarly, the electrostatic field of all the 4-substituted derivatives (compounds **17–35**) was set equal to that of the unsubstituted derivative **1**.

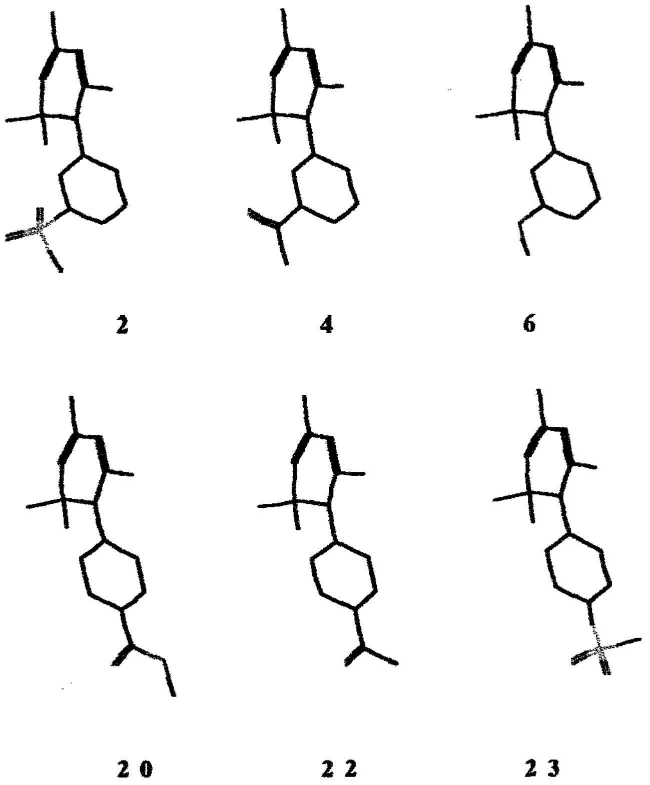
Before the regression analysis was carried out, all the columns containing a nearly constant energy value (characterized therefore by a nearly zero variance) were rejected by a standard deviation threshold set to 0.5 kcal/mol. The purpose of this operation was to exclude parameters considered as irrelevant (namely, the steric columns associated with variation of X_3 and the electrostatic columns associated with variation of X_4).

The procedure followed in SYBYL to build up the table for the CoMFA with variable selection (CoMFA-VS) was the following:

(1) a QSAR TABLE was CREATEd and a CoMFA COLUMN was APPENDed (steric plus electrostatic);

(2) the electrostatic fields of the 3-substituted derivatives **2–16** and the steric fields of the 4-substituted derivatives **17–35** were ENTERed in the CoMFA column;

(3) the empty cells of the table were filled by the electrostatic and steric fields of the unsub-



2 **4** **6**

20 **22** **23**

Fig. 1. Conformations of compounds **2**, **4**, **6**, **20**, **22** and **23**.

stituted derivative **1**; these two fields were previously EXTRACTed from the same table.

Note that this scheme could be followed only because the orientation of the phenyl group of compound **1** was practically identical to that of the phenyl groups of the remaining compounds. Of course, it would not make sense to use the above procedure in a data set where the ligands do not have the same type of orientation about the positions of substitution.

Regression analysis

Before proceeding to the Partial Least Squares (PLS) analyses [10], all the molecular descriptors associated with a standard deviation lower than 0.5 kcal/mol were removed. The analyses were performed either without scaling of the CoMFA fields and hydrophobic substituent constants, or with a scaling according to standard deviations, as proposed by Cramer et al. [11] (the CoMFA_STD keyword was used). By switching off the scaling option, the models became considerably worse (see below).

The cross-validation procedure [12] was applied to determine the model dimensionality associated with the highest predictive ability. The 'leave-one-out' method was adopted. The number of significant latent variables was selected through the use of the F-test on the 'prediction sum of squared residuals' ($\mathrm{PRESS} = \Sigma(Y_{obs} - Y_{pred})^2$).

Calibration models were derived by setting the number of cross-validation groups to zero. These models were converted into the original parametric space to yield 3D QSAR equations with coefficients associated with CoMFA lattice points. The 3D QSAR coefficients were in turn used to generate CoMFA contour maps, interpolated to user-specified values. As stated by Cramer et al. [1] in their first paper on CoMFA, coefficient contour maps should not be overinterpreted by considering them 'realistic' pictures of the receptor. The interpretation of such plots is possible only on a qualitative basis, since the system is statistically underdetermined because of the relatively high number of descriptors compared to the number of observations.

RESULTS AND DISCUSSION

Background

Our CoMFA study was based on a priori knowledge of the physicochemical factors which modulate activity as established in a previous paper [3]. The QSAR equations derived from the original set of triazines showed that the X_3 substituent played a hydrophobic and electrostatic role in the interaction with the enzyme (Eq. 1), while the effect of the X_4 substituent was of hydrophobic and steric type (Eq. 2).

QSAR for *meta*-triazines:

$$\log 1/K_{i\ app} = 1.01\,\pi'_3 - 1.16 \log(\beta\, 10^{\pi'_3} + 1) + 0.86\,\sigma + 6.33 \qquad (1)$$
$$n = 59,\ r = 0.906,\ s = 0.267,\ F_{1,54} = 9.16,\ \log\beta = -1.08,\ \pi_0 = 1.89$$

QSAR for *para*-triazines:

$$\log 1/K_{i\ app} = 0.73\,\pi'_4 - 1.40 \log(\beta\, 10^{\pi'_4} + 1) - 0.29\,\nu + 6.49 \qquad (2)$$
$$n = 32,\ r = 0.949,\ s = 0.280,\ F_{1,27} = 5.26,\ \log\beta = -2.40,\ \pi_0 = 2.44$$

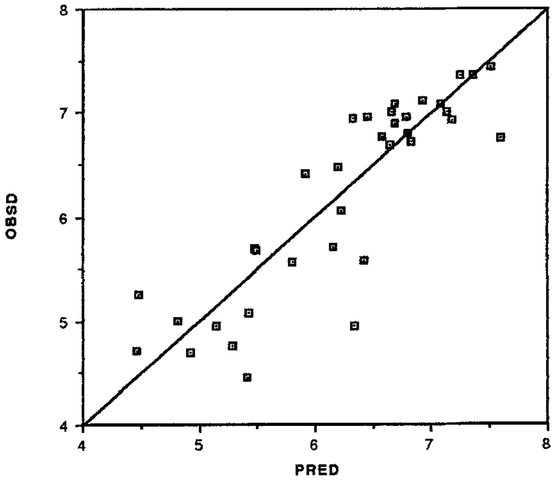


Fig. 2. Plot of observed versus predicted log $1/K_{i\ app}$ values according to 3D Eq. 1.

In these equations n represents the number of data points used to derive the equation, r is the correlation coefficient and s is the standard deviation from the regression. The β parameter is used to iteratively build the bilinear model.

π_3' and π_4' are the normal hydrophobic constants [13], except for substituents of formula 3-$O(CH_2)_nCH_3$ and 4-$O(CH_2)_nCH_3$ for which the values of these descriptors are set to 0.0, because the authors found that the inhibitory activity was essentially unvaried, regardless of the length of the alkoxy group (the value of 0.0 did not greatly differ from the π constant value of –0.02 associated with the OCH_3 group).

TABLE 2
STATISTICS OF THE STANDARD CoMFA MODEL (3D Eq. 1)

N. LV	PRESS	r^2_{cv}	s_{cv}	$F_{1,k}$	p	CALSS	r^2_f	s_f	$F_{1,k}$	p
1	13.139	0.566	0.631	43.036	< 0.001	10.835	0.642	0.573	59.205	< 0.001
2	11.177	0.630	0.591	5.617	< 0.05	5.171	0.829	0.402	35.051	< 0.001
3	9.687	0.680	0.559	4.768	< 0.05	3.315	0.890	0.327	17.356	< 0.001
4[a]	7.321	0.758	0.494	9.695	< 0.005	3.015	0.901	0.317	2.985	< 0.1
5	6.963	0.770	0.490	1.491	> 0.1	2.339	0.922	0.284	8.381	< 0.01
6	8.256	0.727	0.543	–	–	1.708	0.944	0.247	10.344	< 0.005

Relative contributions of the weighted descriptors[b]

Steric field	48.2%
Electrostatic field	11.5%
π_3	18.0%
π_4	22.3%

N. LV is the number of latent variables; p is the probability level calculated through a stepwise F-test; PRESS = $\Sigma (y_{obs} - y_{pred})^2$; CALSS = $\Sigma (y_{obs} - y_{calc})^2$; r^2_{cv} = (SS – PRESS)/SS, where SS is the sum of the squared deviations of activity about the mean; r^2_f = (SS – CALSS)/SS; $s_{cv} = [PRESS/(N_{obs} - 1 - N_{var})]^{1/2}$, where N_{obs} is the number of observations and N_{var} is the number of independent variables used in the model; $s_f = [CALSS/N_{obs} - 1 - N_{var})]^{1/2}$.

[a] Number of components selected for the calibration model.

[b] The reported values refer to the calibration model.

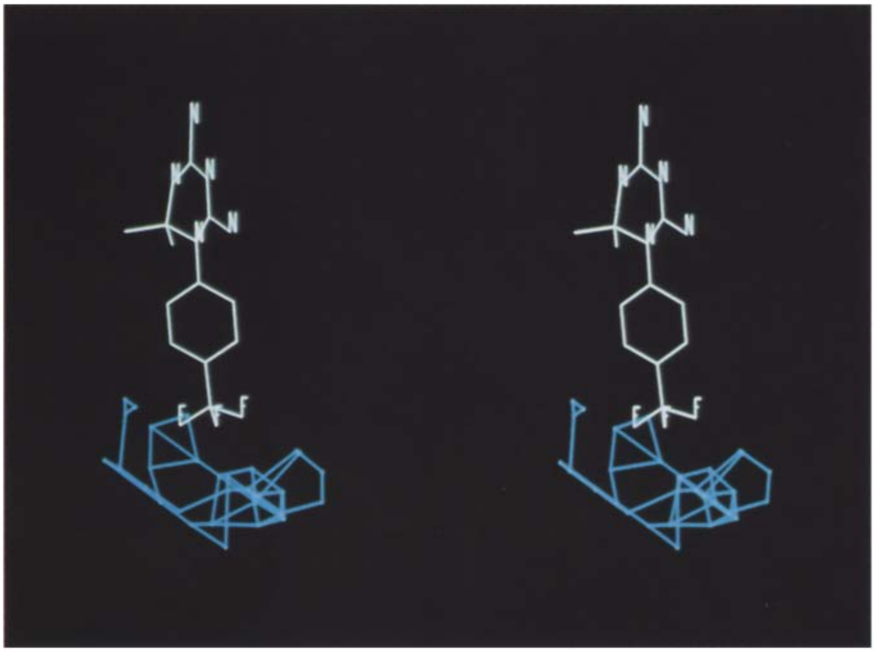


Fig. 3. Stereoview of CoMFA steric contours resulting from 3D Eq. 1 ('standard' CoMFA). The blue polyhedra were interpolated at –0.02 coefficient level. Compound **30** (white) is shown as a reference structure.

The σ term [13] of Eq. 2 (corresponding to meta-substituted benzoic acids) was thought to be related to the negative electrostatic potential of electron-withdrawing 3-substituents that may give rise to a favorable direct dipolar interaction with an electropositive enzyme region.

The Charton's steric parameter ν [14] was hypothesized to describe the steric hindrance exerted by 4-substituents in proximity of the Ile^{60} residue of the protein.

The results of the QSAR analysis were successively checked by observing at the graphics display the type of contacts occurring between some ligands and the enzyme.

Results of the 'standard' CoMFA

Table 2 lists the statistics of the CoMFA model (3D Eq. 1), derived from the 35 triazines reported in Table 1 by using the electrostatic and steric fields in combination with π_3 and π_4.

The dropping of variables, based upon the standard deviation cutoff, allowed us to use in the CoMFA model only 240 potentially relevant columns (49 steric, 189 electrostatic and 2 hydrophobic) of a total of 1602.

The rather satisfactory cross-validated values of r^2 and s ($r^2_{cv} = 0.758$ and $s_{cv} = 0.494$) indicate that the CoMFA model is characterized by good predictive ability. The predicted (cross-validation) and calculated (calibration) values of log $1/K_{i\,app}$ referring to 3D Eq. 1 are listed in Table 1. The plot of the observed versus predicted log $1/K_{i\,app}$ values is shown in Fig. 2.

The above results were obtained by making use of the scaling option CoMFA_STD within the SYBYL/CoMFA routine. When this option was switched off, the correlation became dramatically worse ($r^2_{cv} = 0.013$ on the first component; $r^2_{cv} = -0.124$ on the second component).

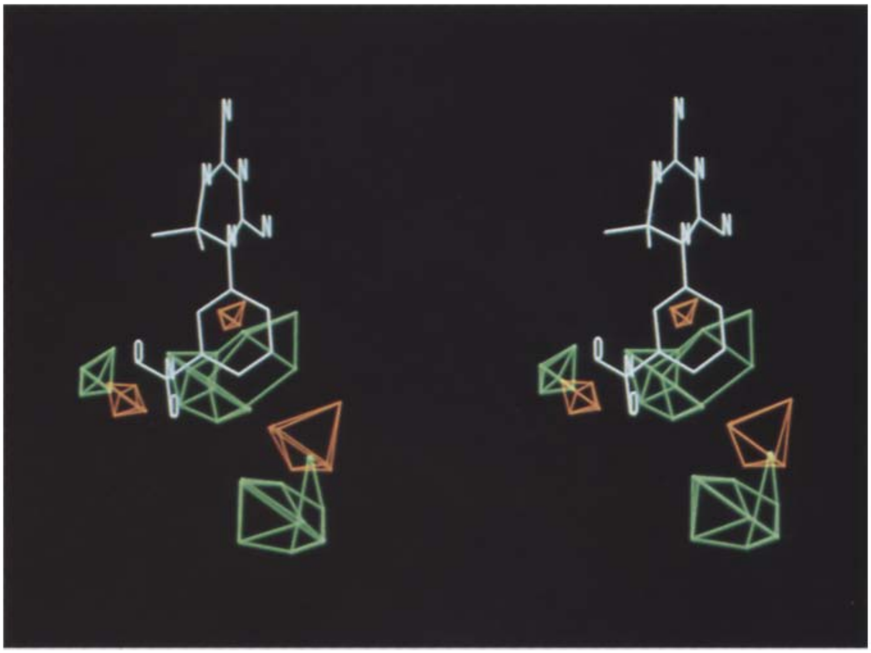


Fig. 4. Stereoview of CoMFA electrostatic contours resulting from 3D Eq. 1 ('standard' CoMFA). The green and orange polyhedra were interpolated at −0.0007 and +0.0007 coefficient levels, respectively. Compound **10** (white) is shown as a reference structure.

The inclusion of squared hydrophobic terms in the analysis did not improve the model. This result is not in contrast to the findings of Hansch et al. [3], who found a bilinear dependence of log $1/K_{1\ \mathrm{app}}$ from the hydrophobic substituent constants associated with the X_3 and X_4 substituents (see Eqs. 1 and 2). In fact, in the set of 35 triazines selected for the CoMFA study, the *t*-butyl group is the most hydrophobic substituent for both the 3- and 4-positions, with a π value of 1.98.

The relative contributions of the CoMFA descriptors to the model are listed in Table 1. These values constitute useful information to better understand to what amount the various steric, electronic and hydrophobic effects of the X_3 and X_4 substituents influence the binding to the enzyme. The fact that the π_3 contribution is significantly higher than the electrostatic one (18.0% versus 11.5%, respectively) means that the log $1/K_{i\ \mathrm{app}}$ values of 3-substituted derivatives characterized by similar electronic properties are mostly dependent on their hydrophobic constants. See e.g. the differences in biological activity between compounds **5** (3-OH, log $1/K_{i\ \mathrm{app}} = 5.57$) and **16** (3-I, log $1/K_{i\ \mathrm{app}} = 7.44$), bearing substituents with quite diverse values of π_3. It can also be noticed that the steric effect at the 4-position prevails over the hydrophobic one (48.2 and 22.3% are the corresponding relative contributions of the two descriptors). Thus, it is not surprising to find that compound **29**, with the most lipophilic and bulky X_4 substituent, *t*-butyl, and the unsubstituted derivative **1** exhibit almost identical log $1/K_{i\ \mathrm{app}}$ values (6.71 and 6.69, respectively).

CoMFA contour maps were generated by interpolating the coefficients of 3D Eq. 1 in the CoMFA lattice at a user-specified level (usually the interpolation is realized at a level close to the average value of the 3D QSAR coefficients of a given type of field).

Figure 3 shows compound **30** embedded in the steric contour (in blue), interpolated at –0.02 coefficient level. The blue contour describes regions of space where an increase in the size of X_4 leads to bad steric contacts between the ligand and the enzyme. There are evident analogies between this 'sterically unfavorable' blue contour and Charton's steric parameter associated with a negative coefficient in Eq. 2. As already mentioned, Hansch et al. [3] suggested that the reason of the negative coefficient of ν was the presence of the Ile^{60} residue very close to the 4-position of substitution.

No significant contour plot, associated with favorable steric interaction, could be obtained by interpolating the positive steric coefficients of 3D Eq. 1 at +0.02 coefficient level.

The electrostatic contour maps of 3D Eq. 1 are shown in Fig. 4, together with compound **10** as a reference structure. The green polyhedra are interpolated at a level of –0.0007 and indicate where high electron density increases activity. In contrast, the orange polyhedra, interpolated at a level of +0.0007, show regions where low electron density favors activity. The electrostatic contour maps obtained by 3D Eq. 1 will be discussed further in detail.

Results of CoMFA with variable selection (CoMFA-VS)

A larger number of columns was dropped from the CoMFA-VS input data table created by the 'nonstandard' sampling of the steric and electrostatic energy parameters (see Methods section). Specifically, only 159 columns survived (35 steric, 122 electrostatic and 2 hydrophobic), instead of the 240 columns preserved in the 'standard' CoMFA table.

Table 3 lists the statistics of the CoMFA-VS model (3D Eq. 2), obtained by using the CoMFA_STD keyword for the scaling of the parameters. Notice that the statistical indices of 3D Eqs. 1 and 2 are nearly coincident (r^2_{cv} is 0.758 and 0.755, respectively), although the latter model was derived from a lower number of probe–ligand intermolecular energy values.

The predicted and calculated activity values relative to 3D Eq. 2 are reported in Table 1. Only compounds **2** and **17** display predicted activity values diverging more than 0.5 log units from those corresponding to 3D Eq. 1. Figure 5 is a plot of observed versus predicted log $1/K_{i\ app}$ values according to 3D Eq. 2.

Concerning the relative contributions of the CoMFA descriptors to 3D Eq. 2, the performed variable selection did not lead to appreciable differences with respect to 3D Eq. 1.

The steric contour maps corresponding to the CoMFA-VS model, generated at –0.02 contour level, are shown in Fig. 6. The blue polyhedra, associated with unfavorable steric effects, are very similar to those generated from 3D Eq. 1 at the same contour level as shown in Fig. 3. The interpolation of the steric coefficients of 3D Eq. 2 at +0.02 coefficient level ('favorable' regions for occupancy by the ligand) yielded no contours.

Figure 7 shows the electrostatic contour maps generated from the CoMFA-VS model. The green and orange polyhedra were obtained through interpolation at levels of –0.0007 and +0.0007, respectively. A visual comparison between these contour maps and the corresponding ones of the 'standard CoMFA' model (3D Eq. 1) reveals some similarities, but also some differences. Particularly, green and orange polyhedra surround the X_4 substituent in Fig. 4, while in Fig. 7 no similar contours exist around the same position of substitution. The vicinity of contours of opposite sign in proximity of the 4-position in Fig. 4 is difficult to be interpreted. Generally, the electrostatic contours generated via 'standard' CoMFA look more complex than those

TABLE 3
STATISTICS OF THE CoMFA-VS MODEL (3D Eq. 2)

N. LV	PRESS	r^2_{cv}	s_{cv}	$F_{1,k}$	p	CALSS	r^2_f	s_f	$F_{1,k}$	p
1	12.563	0.585	0.617	46.522	< 0.001	10.092	0.666	0.553	65.993	< 0.001
2	11.291	0.627	0.594	3.605	< 0.1	6.451	0.787	0.449	18.061	< 0.001
3	10.143	0.664	0.572	3.509	< 0.1	4.130	0.864	0.365	17.421	< 0.001
4[a]	7.440	0.755	0.498	10.899	< 0.005	3.781	0.875	0.355	2.769	> 0.1
5	7.963	0.737	0.524	–	–	2.524	0.917	0.295	14.442	< 0.001
6	8.044	0.734	0.536	–	–	2.133	0.929	0.276	5.133	< 0.05

Relative contributions of the weighted descriptors[b]	
Steric field	45.2%
Electrostatic field	11.5%
π_3	19.3%
π_4	24.0%

N. LV is the number of latent variables; p is the probability level calculated through a stepwise F-test; PRESS = $\Sigma (y_{obs} - y_{pred})^2$; CALSS = $\Sigma (y_{obs} - y_{calc})^2$; r^2_{cv} = (SS – PRESS)/SS, where SS is the sum of the squared deviations of activity about the mean; r^2_f = (SS – CALSS)/SS; $s_{cv} = [\text{PRESS}/(N_{obs} - 1 - N_{var})]^{1/2}$, where N_{obs} is the number of observations and N_{var} is the number of independent variables used in the model; $s_f = [\text{CALSS}/N_{obs} - 1 - N_{var})]^{1/2}$.

[a] Number of components selected for the calibration model.

[b] The reported values refer to the calibration model.

generated via CoMFA-VS. Since the removal of electrostatic energy parameters surrounding the X_4 substituent led to a correlation that was still good, it is likely that the green and orange polyhedra in Fig. 4 located around the X_4 substituent do not have any physicochemical meaning but are rather a statistical artefact.

The green contours in Fig. 7 suggest that activity can be increased by 3-substituents characterized by a local high electron density, while no electrostatic effect is exerted by substituents at the 4-position. These findings are in good agreement with the hypothesis made by Hansch et al. [3] about an involvement of only the 3-substituent in a direct polar interaction with a positively charged enzyme site.

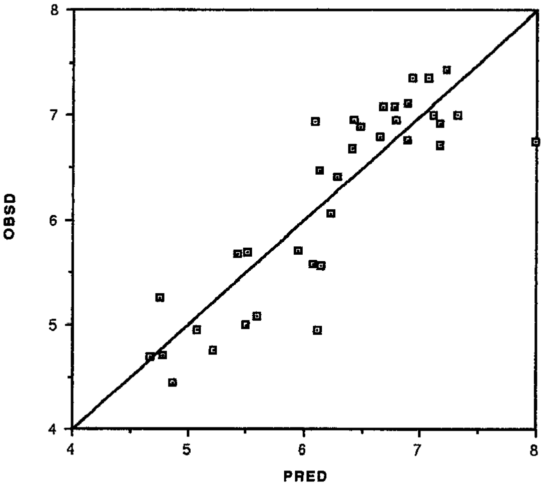


Fig. 5. Plot of observed versus predicted log $1/K_{i\ app}$ values according to 3D Eq. 2.

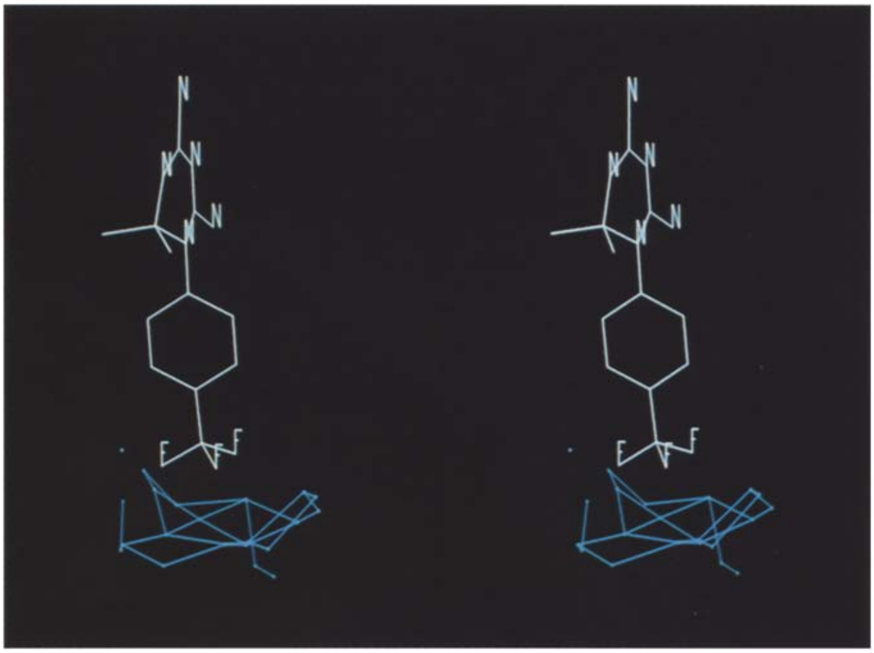


Fig. 6. Stereoview of CoMFA steric contours resulting from 3D Eq. 2 (CoMFA-VS). The blue polyhedra were interpolated at –0.02 coefficient level. Compound **30** (white) is shown as a reference structure.

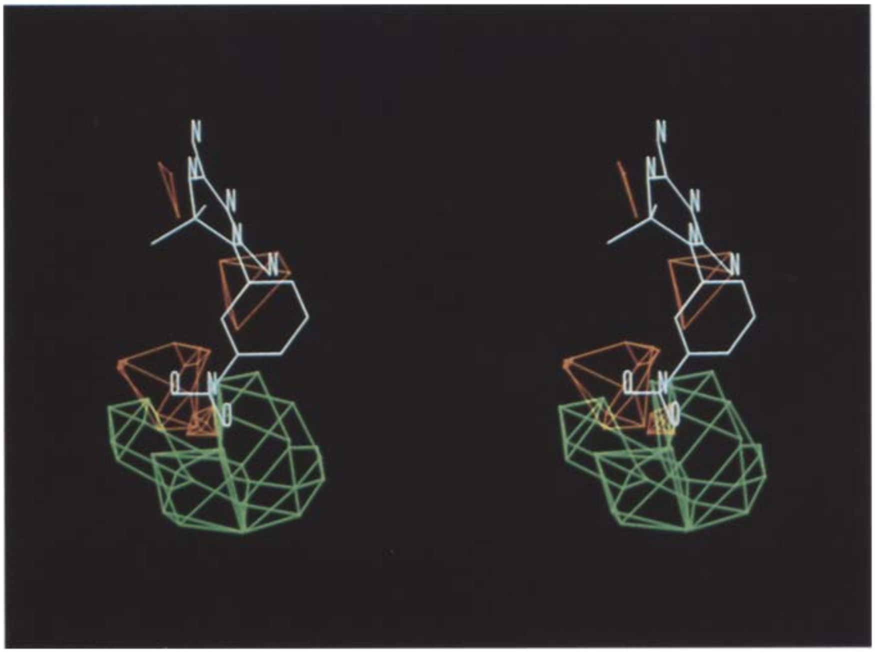


Fig. 7. Stereoview of CoMFA electrostatic contours resulting from 3D Eq. 2 (CoMFA-VS). The green and orange polyhedra were interpolated at –0.0007 and +0.0007 coefficient levels, respectively. Compound **10** (white) is shown as a reference structure.

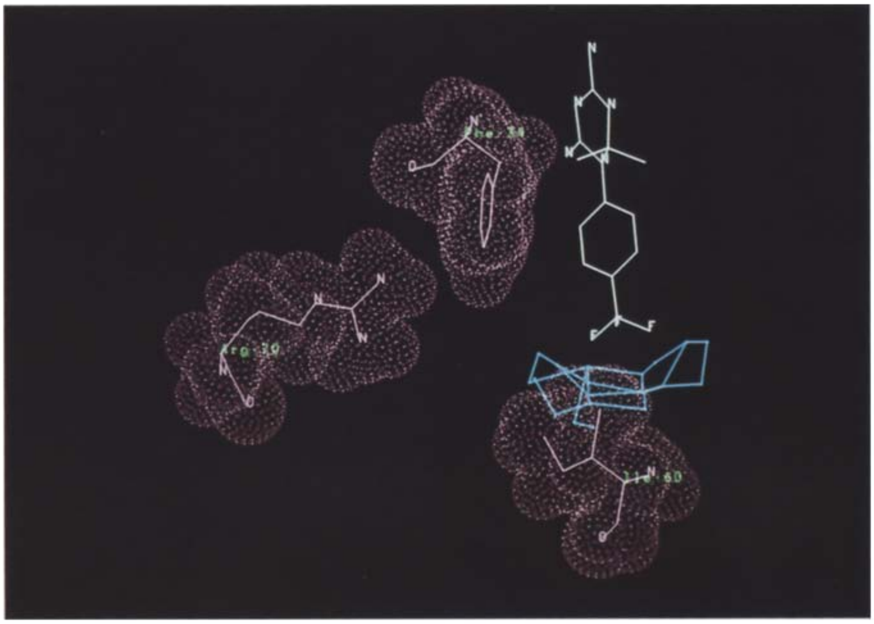


Fig. 8. CoMFA unfavorable steric contours (blue) resulting from 3D Eq. 2 (CoMFA-VS) oriented into the DHFR active site (magenta). Only Phe[34], Arg[70] and Ile[60] are shown. The dots in magenta are the van der Waals molecular surfaces of the three residues. Note that Ile[60] is in close proximity to the blue polyhedra. Compound **30** (white) is also shown as a reference structure.

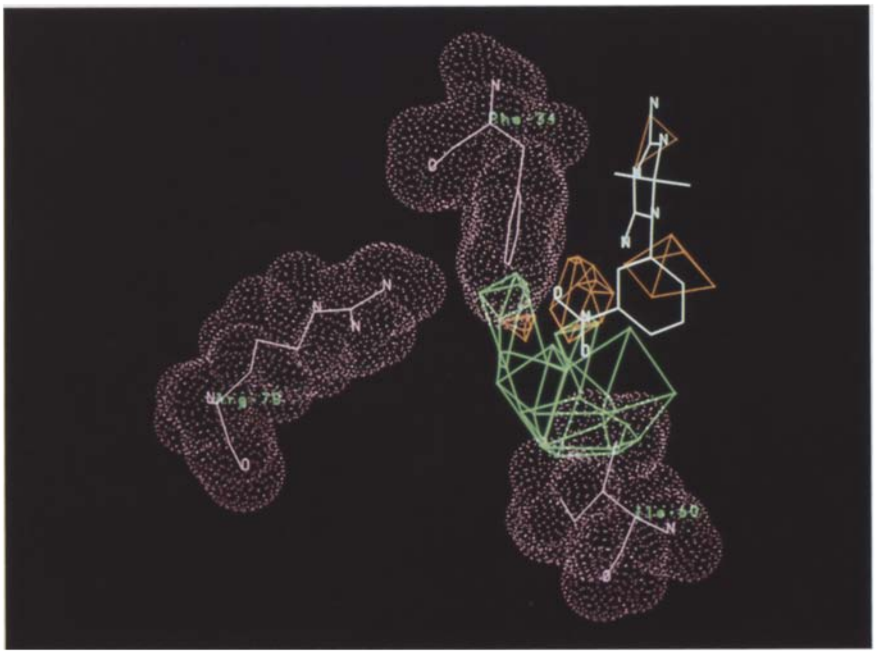


Fig. 9. CoMFA electrostatic contours resulting from 3D Eq. 2 (CoMFA-VS) within the DHFR active site (magenta). The green and orange polyhedra were interpolated at –0.0007 and +0.0007 coefficient levels, respectively. Only Phe[34], Arg[70] and Ile[60] are shown. The dots in magenta are the van der Waals molecular surfaces of the residues. Note that the Arg[70] and Phe[34] side chains approach respectively the green and orange polyhedra. Compound **10** (white) is shown as a reference structure.

The orange contours in Fig. 7 are mainly located above the plane of the substituted phenyl ring and also halfway between the 3- and the 2-position of the same ring. These polyhedra indicate regions of positive electrostatic potential of the ligand which increase the binding affinity. In analogy to what has been observed by Kim and Martin [15], who modeled pKa values in a set of benzoic acids, it is possible that all or part of the orange contours arise because of the requirement of electrical neutrality in the molecule as a whole. Similar orange polyhedra were found also for the standard CoMFA model 3D Eq. 1 (see Fig. 4).

Comparisons between CoMFA-VS coefficient contour maps with the enzyme active site

Figure 8 shows the CoMFA 'unfavorable' steric contour corresponding to 3D Eq. 2 (in blue) within the active site of DHFR (only the Phe^{34}, Arg^{70} and Ile^{60} residues are shown). Note that the blue coefficient contour is located in the proximity of the Ile^{60} side chain. This result is consistent with the previous observations of Hansch et al. [3], who rationalized the poor activity of triazines substituted at the 4-position with bulky groups in terms of bad steric contacts between the X_4 substituent and the Ile^{60} residue.

The CoMFA electrostatic 'favorable' contour maps generated from 3D Eq. 2, which are the regions where a negative potential is desirable for activity, are depicted in green in Fig. 9, together with the above three amino acid residues.

It can be easily seen that the green polyhedra approach the guanidine moiety of Arg^{70}. If the side chain of this amino acid is protonated under the biological test conditions, an ion-dipole attractive interaction could take place between the guanidinium cationic moiety and partially negatively charged X_3 substituents. We have already mentioned that a similar mechanism of interaction was postulated by Hansch et al. [3] to interpret the positive coefficient of the σ term in Eq. 1.

It is worth noting that a considerable portion of the orange electrostatic contours, also shown in Fig. 9, is very close to the phenyl moiety of Phe^{34}. Provided that the benzene ring of Phe^{34} produces a negative electrostatic potential above and under its plane, it seems reasonable to assume that a ligand featuring a complementary positive potential should bind more tightly to the enzyme. Such a hypothesis would be consistent with the spatial location of the orange polyhedra which indicate where high electron density in the ligand structure is detrimental for activity.

Differences in $\log 1/K_{i\,app}$ *values between isomers*

The biological data within each of seven pairs of isomers of the investigated set are nearly coincident (see in Table 1 the $\log 1/K_{i\,app}$ values of the pairs **3/22**, **4/25**, **5/26**, **6/28**, **7/29**, **8/31** and **13/34**). Bearing in mind that the regression analyses were carried out simultaneously on all the 35 triazines, the fact that 14 of the 35 data points show such behavior means that the CoMFA approach might discriminate the effects of the 3- and 4-substituents by chance to an amount of 40% (14/35). In order to check whether the equiactivity of the above listed seven pairs of isomers could affect the results of the CoMFA studies, we repeated the analyses by excluding these 14 data points. In the interest of brevity, only a short report of the results will be given.

The 'standard' CoMFA on the remaining 21 observations yielded the following statistical indices: 2 components, $r^2_{cv} = 0.705$, $s_{cv} = 0.614$. The calibration model with 2 components was used to predict the activity of the 14 compounds left out from the analysis. The standard error value of this prediction test turned out to be 0.562, that is, lower than the cross-validated one.

Analogous results were obtained with the already described procedure of position-dependent variable selection (CoMFA-VS): 2 components, $r^2_{cv} = 0.724$, $s_{cv} = 0.594$. The standard error associated with the prediction of the activity of the 14 left-out compounds was 0.584.

The CoMFA contour plots, generated from the two models based on 21 data points were found to be very similar to those shown so far in this paper. Thus, we had the confirmation that the equiactivity within each of the considered pairs of isomers did not affect the results of the 3D QSAR models 1 and 2.

CONCLUSIONS

In traditional QSAR analysis, the effects associated with two or more positions of substitutions can be described in the 'best' equation by using only those parameters whose inclusion is justified by statistics. Although the predictive ability of a CoMFA model is usually tested with cross-validation, there is still the possibility that the correlation is affected by noise (that is, non-relevant energy parameters associated with a sufficiently high variance). As a consequence, the resulting coefficient contour maps might not entirely reflect the actual nature of the ligand–receptor interaction. In this paper we have investigated whether a certain amount of noise in a 'standard' CoMFA could be related to the fact that the steric and electrostatic field of different substitution positions, in a congeneric series, are *both* brought into the calculation.

In the set of triazines under consideration, a variable selection procedure was adopted by taking into account the knowledge gained by an earlier developed QSAR model. In practice, the steric or electrostatic field associated with a given position of substitution, and believed to be irrelevant, was simply left out of the statistical analysis. The resulting model yielded a cross-validated r^2 value almost identical to that given by a 'standard' CoMFA run based on a higher number of energy parameters. However, the coefficient contour maps of CoMFA with variable selection were found to be in better agreement with previous QSAR results and crystallographic data.

The employed variable selection method should not be regarded as a general tool to handle problems in 3D QSAR related with a low signal-to-noise ratio. In fact, if preliminary results of classical QSAR were necessary to derive a 3D QSAR model, there would be no need to perform a 3D QSAR analysis at all. It is worth remarking that the QSAR equations and the X-ray crystallographic data were used just as a 'reliable' body of information, useful both to set up the reduction of variables and to evaluate how meaningful differently derived CoMFA contour maps could be.

One of the potential advantages of the CoMFA approach over conventional QSAR is to provide results that can be interpreted in terms of the physicochemistry involved in the biochemical process under study. As a consequence, it is very important that a CoMFA model offers not only the capability to forecast biological activity of new molecules, but also that the resulting coefficient contour plots reflect somehow the actual nature of the ligand–receptor interaction.

Our report is clearly not intended to criticize the CoMFA approach, but only to outline problems which should not be underestimated.

ACKNOWLEDGEMENTS

We thank Prof. Corwin Hansch for helpful discussions and Dr. Teri E. Klein for providing us with the crystallographic data. This work was supported by a grant from C.N.R. (Rome).

REFERENCES

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

2 Martin, Y.C., In Grunewald, G.L. (Ed.) Quantitative Drug Design: A Critical Introduction, Medicinal Research Series, Vol. 8, Marcel Dekker, New York, NY, 1978.

3 Hansch, C., Hathaway, B.A., Guo, Z., Selassie, C.D., Dietrich, S.W., Blaney, J.M., Langridge, R., Volz, K.W. and Kaufman, B.T., J. Med. Chem., 27 (1984) 129.

4 Volz, K.W., Matthews, D.A., Alden, R.A., Freer, S.T., Hansch, C., Kaufman, B.T. and Kraut, J., J. Biol. Chem., 257 (1982) 2528.

5 Dietrich, S.W., Blaney, J.M., Reynolds, M.A., Jow, P.Y.C. and Hansch, C., J. Med. Chem., 23 (1980) 1205.

6 SYBYL Molecular Modeling System (version 5.41), TRIPOS Associates, St. Louis, MO.

7 Stewart, J.J.P., J. Comput. Chem., 10 (1989) 209.

8 MOPAC (version 5.00), Quantum Chemistry Program Exchange, No. 455, 1989.

9 Vinter, J.G., Davis, A. and Saunders, M.R., J. Comput.-Aided Mol. Design, 1 (1987) 31.

10 Stahle, L. and Wold, S., Prog. Med. Chem., 25 (1988) 292.

11 Clark, M., Cramer III, R.D., Jones, D.M., Patterson, D.E. and Simeroth, P.E., Tetrahedron Comput. Methodol., 3 (1990) 47.

12 Cramer III, R.D., Bunce, J.D., Patterson, D.E. and Frank, I.E., Quant. Struct.-Act. Relatsh., 7 (1988) 18.

13 Hansch, C. and Leo, A., In Substituent Constants for Correlation Analysis in Chemistry and Biology, Wiley, New York, NY, 1979.

14 Charton, M., In Roche, E.B. (Ed.) Design of Biopharmaceutical Properties through Prodrugs and Analogues, American Pharmaceutical Association, Washington, DC, 1977, p. 228.

15 Kim, K.H. and Martin, Y.C., J. Org. Chem., 56 (1991) 2723.